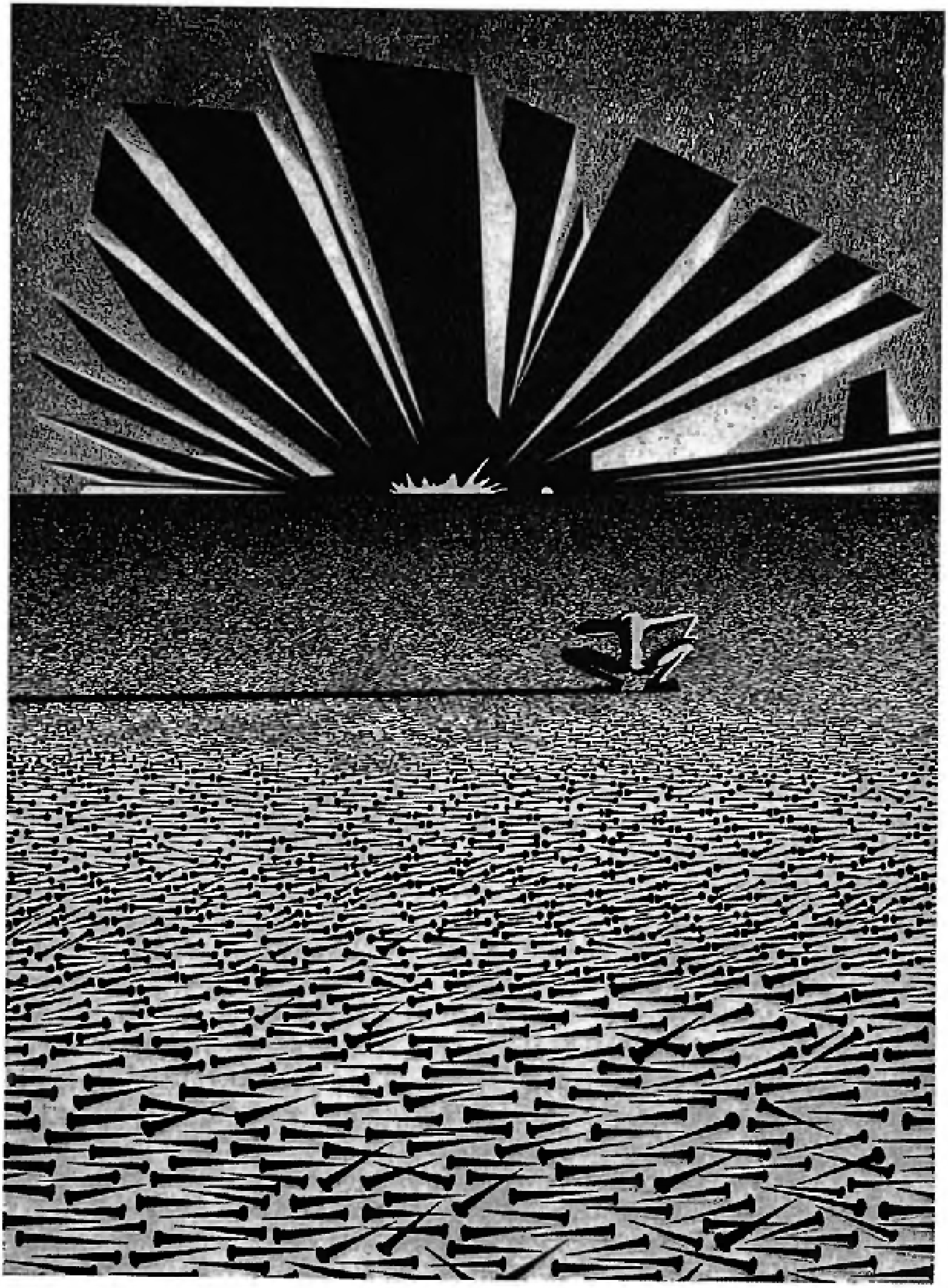
Graduate Texts in Mathematics

A.N. Shiryaev

Probability

Second Edition

Springer



"Order out of chaos"
(*Courtesy of Professor A. T. Fomenko of the Moscow State University*)

Preface to the Second Edition

In the Preface to the first edition, originally published in 1980, we mentioned that this book was based on the author's lectures in the Department of Mechanics and Mathematics of the Lomonosov University in Moscow, which were issued, in part, in mimeographed form under the title "Probability, Statistics, and Stochastic Processors, I, II" and published by that University. Our original intention in writing the first edition of this book was to divide the contents into three parts: probability, mathematical statistics, and theory of stochastic processes, which corresponds to an outline of a three-semester course of lectures for university students of mathematics. However, in the course of preparing the book, it turned out to be impossible to realize this intention completely, since a full exposition would have required too much space. In this connection, we stated in the Preface to the first edition that only *probability theory* and the *theory of random processes with discrete time* were really adequately presented.

Essentially all of the first edition is reproduced in this second edition. Changes and corrections are, as a rule, editorial, taking into account comments made by both Russian and foreign readers of the Russian original and of the English and German translations [S11]. The author is grateful to all of these readers for their attention, advice, and helpful criticisms.

In this second English edition, new material also has been added, as follows: in Chapter III, §5, §§7–12; in Chapter IV, §5; in Chapter VII, §§8–10. The most important addition is the third chapter. There the reader will find expositions of a number of problems connected with a deeper study of themes such as the distance between probability measures, metrization of weak convergence, and contiguity of probability measures. In the same chapter, we have added proofs of a number of important results on the rapidity of convergence in the central limit theorem and in Poisson's theorem on the approximation of the binomial by the Poisson distribution. These were merely stated in the first edition.

We also call attention to the new material on the probability of large deviations (Chapter IV, §5), on the central limit theorem for sums of dependent random variables (Chapter VII, §8), and on §§9 and 10 of Chapter VII.

During the last few years, the literature on probability published in Russia by Nauka has been extended by Sevastyanov [S10], 1982; Rozanov [R6], 1985; Borovkov [B4], 1986; and Gnedenko [G4], 1988. It appears that these publications, together with the present volume, being quite different and complementing each other, cover an extensive amount of material that is essentially broad enough to satisfy contemporary demands by students in various branches of mathematics and physics for instruction in topics in probability theory.

Gnedenko's textbook [G4] contains many well-chosen examples, including applications, together with pedagogical material and extensive surveys of the history of probability theory. Borovkov's textbook [B4] is perhaps the most like the present book in the style of exposition. Chapters 9 (Elements of Renewal Theory), 11 (Factorization of the Identity) and 17 (Functional Limit Theorems), which distinguish [B4] from this book and from [G4] and [R6], deserve special mention. Rozanov's textbook contains a great deal of material on a variety of mathematical models which the theory of probability and mathematical statistics provides for describing random phenomena and their evolution. The textbook by Sevastyanov is based on his two-semester course at the Moscow State University. The material in its last four chapters covers the minimum amount of probability and mathematical statistics required in a one-year university program. In our text, perhaps to a greater extent than in those mentioned above, a significant amount of space is given to set-theoretic aspects and mathematical foundations of probability theory.

Exercises and problems are given in the books by Gnedenko and Sevastyanov at the ends of chapters, and in the present textbook at the end of each section. These, together with, for example, the problem sets by A. V. Prokhorov and V. G. and N. G. Ushakov (*Problems in Probability Theory*, Nauka, Moscow, 1986) and by Zubkov, Sevastyanov, and Chistyakov (*Collected Problems in Probability Theory*, Nauka, Moscow, 1988) can be used by readers for independent study, and by teachers as a basis for seminars for students.

Special thanks to Harold Boas, who kindly translated the revisions from Russian to English for this new edition.

Moscow A. Shiryaev

Preface to the First Edition

This textbook is based on a three-semester course of lectures given by the author in recent years in the Mechanics–Mathematics Faculty of Moscow State University and issued, in part, in mimeographed form under the title *Probability, Statistics, Stochastic Processes, I, II* by the Moscow State University Press.

We follow tradition by devoting the first part of the course (roughly one semester) to the elementary theory of probability (Chapter I). This begins with the construction of probabilistic models with finitely many outcomes and introduces such fundamental probabilistic concepts as sample spaces, events, probability, independence, random variables, expectation, correlation, conditional probabilities, and so on.

Many probabilistic and statistical regularities are effectively illustrated even by the simplest random walk generated by Bernoulli trials. In this connection we study both classical results (law of large numbers, local and integral De Moivre and Laplace theorems) and more modern results (for example, the arc sine law).

The first chapter concludes with a discussion of dependent random variables generated by martingales and by Markov chains.

Chapters II–IV form an expanded version of the second part of the course (second semester). Here we present (Chapter II) Kolmogorov's generally accepted axiomatization of probability theory and the mathematical methods that constitute the tools of modern probability theory (σ-algebras, measures and their representations, the Lebesgue integral, random variables and random elements, characteristic functions, conditional expectation with respect to a σ-algebra, Gaussian systems, and so on). Note that two measure-theoretical results—Carathéodory's theorem on the extension of measures and the Radon–Nikodým theorem—are quoted without proof.

The third chapter is devoted to problems about weak convergence of probability distributions and the method of characteristic functions for proving limit theorems. We introduce the concepts of relative compactness and tightness of families of probability distributions, and prove (for the real line) Prohorov's theorem on the equivalence of these concepts.

The same part of the course discusses properties "with probability 1" for sequences and sums of independent random variables (Chapter IV). We give proofs of the "zero or one laws" of Kolmogorov and of Hewitt and Savage, tests for the convergence of series, and conditions for the strong law of large numbers. The law of the iterated logarithm is stated for arbitrary sequences of independent identically distributed random variables with finite second moments, and proved under the assumption that the variables have Gaussian distributions.

Finally, the third part of the book (Chapters V–VIII) is devoted to random processes with discrete parameters (random sequences). Chapters V and VI are devoted to the theory of stationary random sequences, where "stationary" is interpreted either in the strict or the wide sense. The theory of random sequences that are stationary in the strict sense is based on the ideas of ergodic theory: measure preserving transformations, ergodicity, mixing, etc. We reproduce a simple proof (by A. Garsia) of the maximal ergodic theorem; this also lets us give a simple proof of the Birkhoff–Khinchin ergodic theorem.

The discussion of sequences of random variables that are stationary in the wide sense begins with a proof of the spectral representation of the covariance fuction. Then we introduce orthogonal stochastic measures, and integrals with respect to these, and establish the spectral representation of the sequences themselves. We also discuss a number of statistical problems: estimating the covariance function and the spectral density, extrapolation, interpolation and filtering. The chapter includes material on the Kalman–Bucy filter and its generalizations.

The seventh chapter discusses the basic results of the theory of martingales and related ideas. This material has only rarely been included in traditional courses in probability theory. In the last chapter, which is devoted to Markov chains, the greatest attention is given to problems on the asymptotic behavior of Markov chains with countably many states.

Each section ends with problems of various kinds: some of them ask for proofs of statements made but not proved in the text, some consist of propositions that will be used later, some are intended to give additional information about the circle of ideas that is under discussion, and finally, some are simple exercises.

In designing the course and preparing this text, the author has used a variety of sources on probability theory. The Historical and Bibliographical Notes indicate both the historial sources of the results and supplementary references for the material under consideration.

The numbering system and form of references is the following. Each section has its own enumeration of theorems, lemmas and formulas (with no indication of chapter or section). For a reference to a result from a different section of the same chapter, we use double numbering, with the first number indicating the number of the section (thus, (2.10) means formula (10) of §2). For references to a different chapter we use triple numbering (thus, formula (II.4.3) means formula (3) of §4 of Chapter II). Works listed in the References at the end of the book have the form [Ln], where L is a letter and n is a numeral.

The author takes this opportunity to thank his teacher A. N. Kolmogorov, and B. V. Gnedenko and Yu. V. Prokhorov, from whom he learned probability theory and under whose direction he had the opportunity of using it. For discussions and advice, the author also thanks his colleagues in the Departments of Probability Theory and Mathematical Statistics at the Moscow State University, and his colleagues in the Section on probability theory of the Steklov Mathematical Institute of the Academy of Sciences of the U.S.S.R.

Moscow A. N. SHIRYAEV
Steklov Mathematical Institute

Translator's acknowledgement. I am grateful both to the author and to my colleague, C. T. Ionescu Tulcea, for advice about terminology.

R. P. B.

Contents

Introduction

The subject matter of probability theory is the mathematical analysis of random events, i.e., of those empirical phenomena which—under certain circumstance—can be described by saying that:

They do not have *deterministic regularity* (observations of them do not yield the same outcome);

whereas at the same time

They possess some *statistical regularity* (indicated by the statistical stability of their frequency).

We illustrate with the classical example of a "fair" toss of an "unbiased" coin. It is clearly impossible to predict with certainty the outcome of each toss. The results of successive experiments are very irregular (now "head," now "tail") and we seem to have no possibility of discovering any regularity in such experiments. However, if we carry out a large number of "independent" experiments with an "unbiased" coin we can observe a very definite statistical regularity, namely that "head" appears with a frequency that is "close" to $\frac{1}{2}$.

Statistical stability of a frequency is very likely to suggest a hypothesis about a possible quantitative estimate of the "randomness" of some event A connected with the results of the experiments. With this starting point, probability theory postulates that corresponding to an event A there is a definite number $\mathsf{P}(A)$, called the probability of the event, whose intrinsic property is that as the number of "independent" trials (experiments) increases the frequency of event A is approximated by $\mathsf{P}(A)$.

Applied to our example, this means that it is natural to assign the probability $\frac{1}{2}$ to the event A that consists of obtaining "head" in a toss of an "unbiased" coin.

There is no difficulty in multiplying examples in which it is very easy to obtain numerical values intuitively for the probabilities of one or another event. However, these examples are all of a similar nature and involve (so far) undefined concepts such as "fair" toss, "unbiased" coin, "independence," etc.

Having been invented to investigate the quantitative aspects of "randomness," probability theory, like every exact science, became such a science only at the point when the concept of a probabilistic model had been clearly formulated and axiomatized. In this connection it is natural for us to discuss, although only briefly, the fundamental steps in the development of probability theory.

Probability theory, as a science, originated in the middle of the seventeenth century with Pascal (1623–1662), Fermat (1601–1655) and Huygens (1629–1695). Although special calculations of probabilities in games of chance had been made earlier, in the fifteenth and sixteenth centuries, by Italian mathematicians (Cardano, Pacioli, Tartaglia, etc.), the first general methods for solving such problems were apparently given in the famous correspondence between Pascal and Fermat, begun in 1654, and in the first book on probability theory, *De Ratiociniis in Aleae Ludo* (*On Calculations in Games of Chance*), published by Huygens in 1657. It was at this time that the fundamental concept of "mathematical expectation" was developed and theorems on the addition and multiplication of probabilities were established.

The real history of probability theory begins with the work of James Bernoulli (1654–1705), *Ars Conjectandi* (*The Art of Guessing*) published in 1713, in which he proved (quite rigorously) the first limit theorem of probability theory, the law of large numbers; and of De Moivre (1667–1754), *Miscellanea Analytica Supplementum* (a rough translation might be *The Analytic Method* or *Analytic Miscellany*, 1730), in which the central limit theorem was stated and proved for the first time (for symmetric Bernoulli trials).

Bernoulli deserves the credit for introducing the "classical" definition of the concept of the *probability* of an event as the *ratio* of the number of possible outcomes of an experiment, that are favorable to the event, to the number of possible outcomes.

Bernoulli was probably the first to realize the importance of considering infinite sequences of random trials and to make a clear distinction between the probability of an event and the frequency of its realization.

De Moivre deserves the credit for defining such concepts as independence, mathematical expectation, and conditional probability.

In 1812 there appeared Laplace's (1749–1827) great treatise *Théorie Analytique des Probabilitiés* (*Analytic Theory of Probability*) in which he presented his own results in probability theory as well as those of his predecessors. In particular, he generalized De Moivre's theorem to the general (unsymmetric) case of Bernoulli trials, and at the same time presented De Moivre's results in a more complete form.

Laplace's most important contribution was the application of probabilistic methods to errors of observation. He formulated the idea of considering errors of observation as the cumulative results of adding a large number of independent elementary errors. From this it followed that under rather general conditions the distribution of errors of observation must be at least approximately normal.

The work of Poisson (1781–1840) and Gauss (1777–1855) belongs to the same epoch in the development of probability theory, when the center of the stage was held by limit theorems. In contemporary probability theory we think of Poisson in connection with the distribution and the process that bear his name. Gauss is credited with originating the theory of errors and, in particular, with creating the fundamental method of least squares.

The next important period in the development of probability theory is connected with the names of P. L. Chebyshev (1821–1894), A. A. Markov (1856–1922), and A. M. Lyapunov (1857–1918), who developed effective methods for proving limit theorems for sums of independent but arbitrarily distributed random variables.

The number of Chebyshev's publications in probability theory is not large—four in all—but it would be hard to overestimate their role in probability theory and in the development of the classical Russian school of that subject.

> "On the methodological side, the revolution brought about by Chebyshev was not only his insistence for the first time on complete rigor in the proofs of limit theorems, . . . but also, and principally, that Chebyshev always tried to obtain precise estimates for the deviations from the limiting regularities that are available for large but finite numbers of trials, in the form of inequalities that are valid unconditionally for any number of trials."
>
> (A. N. Kolmogorov [30])

Before Chebyshev the main interest in probability theory had been in the calculation of the probabilities of random events. He, however, was the first to realize clearly and exploit the full strength of the concepts of random variables and their mathematical expectations.

The leading exponent of Chebyshev's ideas was his devoted student Markov, to whom there belongs the indisputable credit of presenting his teacher's results with complete clarity. Among Markov's own significant contributions to probability theory were his pioneering investigations of limit theorems for sums of independent random variables and the creation of a new branch of probability theory, the theory of dependent random variables that form what we now call a Markov chain.

> "Markov's classical course in the calculus of probability and his original papers, which are models of precision and clarity, contributed to the greatest extent to the transformation of probability theory into one of the most significant

branches of mathematics and to a wide extension of the ideas and methods of Chebyshev."

(S. N. Bernstein [3])

To prove the central limit theorem of probability theory (the theorem on convergence to the normal distribution), Chebyshev and Markov used what is known as the method of moments. With more general hypotheses and a simpler method, the method of characteristic functions, the theorem was obtained by Lyapunov. The subsequent development of the theory has shown that the method of characteristic functions is a powerful analytic tool for establishing the most diverse limit theorems.

The modern period in the development of probability theory begins with its axiomatization. The first work in this direction was done by S. N. Bernstein (1880–1968), R. von Mises (1883–1953), and E. Borel (1871–1956). A. N. Kolmogorov's book *Foundations of the Theory of Probability* appeared in 1933. Here he presented the axiomatic theory that has become generally accepted and is not only applicable to all the classical branches of probability theory, but also provides a firm foundation for the development of new branches that have arisen from questions in the sciences and involve infinite-dimensional distributions.

The treatment in the present book is based on Kolmogorov's axiomatic approach. However, to prevent formalities and logical subtleties from obscuring the intuitive ideas, our exposition begins with the elementary theory of probability, whose elementariness is merely that in the corresponding probabilistic models we consider only experiments with finitely many outcomes. Thereafter we present the foundations of probability theory in their most general form.

The 1920s and '30s saw a rapid development of one of the new branches of probability theory, the theory of stochastic processes, which studies families of random variables that evolve with time. We have seen the creation of theories of Markov processes, stationary processes, martingales, and limit theorems for stochastic processes. Information theory is a recent addition.

The present book is principally concerned with stochastic processes with discrete parameters: random sequences. However, the material presented in the second chapter provides a solid foundation (particularly of a logical nature) for the study of the general theory of stochastic processes.

It was also in the 1920s and '30s that mathematical statistics became a separate mathematical discipline. In a certain sense mathematical statistics deals with inverses of the problems of probability: If the basic aim of probability theory is to calculate the probabilities of complicated events under a given probabilistic model, mathematical statistics sets itself the inverse problem: to clarify the structure of probabilistic–statistical models by means of observations of various complicated events.

Some of the problems and methods of mathematical statistics are also discussed in this book. However, all that is presented in detail here is probability theory and the theory of stochastic processes with discrete parameters.

CHAPTER I

Elementary Probability Theory

§1. Probabilistic Model of an Experiment with a Finite Number of Outcomes

1. Let us consider an experiment of which all possible results are included in a finite number of outcomes $\omega_1, \dots, \omega_N$. We do not need to know the nature of these outcomes, only that there are a finite number N of them.

We call $\omega_1, \dots, \omega_N$ *elementary events*, or *sample points*, and the finite set

$$\Omega = \{\omega_1, \dots, \omega_N\},$$

the *space of elementary events* or the *sample space*.

The choice of the space of elementary events is the *first step* in formulating a probabilistic model for an experiment. Let us consider some examples of sample spaces.

EXAMPLE 1. For a single toss of a coin the sample space Ω consists of two points:

$$\Omega = \{H, T\},$$

where H = "head" and T = "tail". (We exclude possibilities like "the coin stands on edge," "the coin disappears," etc.)

EXAMPLE 2. For n tosses of a coin the sample space is

$$\Omega = \{\omega : \omega = (a_1, \dots, a_n), a_i = H \text{ or } T\}$$

and the general number $N(\Omega)$ of outcomes is 2^n.

EXAMPLE 3. First toss a coin. If it falls "head" then toss a die (with six faces numbered 1, 2, 3, 4, 5, 6); if it falls "tail", toss the coin again. The sample space for this experiment is

$$\Omega = \{H1, H2, H3, H4, H5, H6, TH, TT\}.$$

We now consider some more complicated examples involving the selection of n balls from an urn containing M distinguishable balls.

2. EXAMPLE 4 (Sampling with replacement). This is an experiment in which after each step the selected ball is returned again. In this case each sample of n balls can be presented in the form $(a_1, \ldots, a_n)$, where a_i is the label of the ball selected at the ith step. It is clear that in sampling with replacement each a_i can have any of the M values $1, 2, \ldots, M$. The description of the sample space depends in an essential way on whether we consider samples like, for example, (4, 1, 2, 1) and (1, 4, 2, 1) as different or the same. It is customary to distinguish two cases: *ordered* samples and *unordered* samples. In the first case samples containing the same elements, but arranged differently, are considered to be different. In the second case the order of the elements is disregarded and the two samples are considered to be the same. To emphasize which kind of sample we are considering, we use the notation $(a_1, \ldots, a_n)$ for ordered samples and $[a_1, \ldots, a_n]$ for unordered samples.

Thus for ordered samples the sample space has the form

$$\Omega = \{\omega : \omega = (a_1, \ldots, a_n), a_i = 1, \ldots, M\}$$

and the number of (different) outcomes is

$$N(\Omega) = M^n. \tag{1}$$

If, however, we consider unordered samples, then

$$\Omega = \{\omega : \omega = [a_1, \ldots, a_n], a_i = 1, \ldots, M\}.$$

Clearly the number $N(\Omega)$ of (different) unordered samples is smaller than the number of ordered samples. Let us show that in the present case

$$N(\Omega) = C^n_{M+n-1}, \tag{2}$$

where $C_k^l \equiv k!/[l!\,(k-l)!]$ is the number of combinations of l elements, taken k at a time.

We prove this by induction. Let $N(M, n)$ be the number of outcomes of interest. It is clear that when $k \leq M$ we have

$$N(k, 1) = k = C_k^1.$$

Now suppose that $N(k, n) = C_{k+n-1}^{k}$ for $k \leq M$; we show that this formula continues to hold when n is replaced by $n + 1$. For the unordered samples $[a_1, \ldots, a_{n+1}]$ that we are considering, we may suppose that the elements are arranged in nondecreasing order: $a_1 \leq a_2 \leq \cdots \leq a_n$. It is clear that the number of unordered samples with $a_1 = 1$ is $N(M, n)$, the number with $a_1 = 2$ is $N(M - 1, n)$, etc. Consequently

$$\begin{aligned} N(M, n+1) &= N(M, n) + N(M-1, n) + \cdots + N(1, n) \\ &= C_{M+n-1}^{n} + C_{M-1+n-1}^{n} + \cdots C_{n}^{n} \\ &= (C_{M+n}^{n+1} - C_{M+n-1}^{n+1}) + (C_{M-1+n}^{n+1} - C_{M-1+n-1}^{n+1}) \\ &\quad + \cdots + (C_{n+1}^{n+1} - C_{n}^{n}) = C_{M+n}^{n+1}; \end{aligned}$$

here we have used the easily verified property

$$C_{k}^{l-1} + C_{k}^{l} = C_{k+1}^{l}$$

of the binomial coefficients.

Example 5 (Sampling without replacement). Suppose that $n \leq M$ and that the selected balls are not returned. In this case we again consider two possibilities, namely ordered and unordered samples.

For ordered samples without replacement the sample space is

$$\Omega = \{\omega: \omega = (a_1, \ldots, a_n), a_k \neq a_l, k \neq l, a_i = 1, \ldots, M\},$$

and the number of elements of this set (called *permutations*) is $M(M-1)\cdots(M-n+1)$. We denote this by $(M)_n$ or A_M^n and call it "the number of permutations of M things, n at a time").

For unordered samples (called *combinations*) the sample space

$$\Omega = \{\omega: \omega = [a_1, \ldots, a_n], a_k \neq a_l, k \neq l, a_i = 1, \ldots, M\}$$

consists of

$$N(\Omega) = C_M^n \tag{3}$$

elements. In fact, from each unordered sample $[a_1, \ldots, a_n]$ consisting of distinct elements we can obtain $n!$ ordered samples. Consequently

$$N(\Omega) \cdot n! = (M)_n$$

and therefore

$$N(\Omega) = \frac{(M)_n}{n!} = C_M^n.$$

The results on the numbers of samples of n from an urn with M balls are presented in Table 1.

Table 1

M^n	C^n_{M+n-1}	With replacement
$(M)_n$	C^n_M	Without replacement
Ordered	Unordered	Type \ Sample

For the case $M = 3$ and $n = 2$, the corresponding sample spaces are displayed in Table 2.

EXAMPLE 6 (Distribution of objects in cells). We consider the structure of the sample space in the problem of placing n objects (balls, etc.) in M cells (boxes, etc.). For example, such problems arise in statistical physics in studying the distribution of n particles (which might be protons, electrons, ...) among M states (which might be energy levels).

Let the cells be numbered $1, 2, \ldots, M$, and suppose first that the objects are distinguishable (numbered $1, 2, \ldots, n$). Then a distribution of the n objects among the M cells is completely described by an ordered set $(a_1, \ldots, a_n)$, where a_i is the index of the cell containing object i. However, if the objects are indistinguishable their distribution among the M cells is completely determined by the unordered set $[a_1, \ldots, a_n]$, where a_i is the index of the cell into which an object is put at the ith step.

Comparing this situation with Examples 4 and 5, we have the following correspondences:

$$\text{(ordered samples)} \leftrightarrow \text{(distinguishable objects)},$$

$$\text{(unordered samples)} \leftrightarrow \text{(indistinguishable objects)},$$

Table 2

(1, 1) (1, 2) (1, 3) (2, 1) (2, 2) (2, 3) (3, 1) (3, 2) (3, 3)	[1, 1] [2, 2] [3, 3] [1, 2] [1, 3] [2, 3]	With replacement
(1, 2) (1, 3) (2, 1) (2, 3) (3, 1) (3, 2)	[1, 2] [1, 3] [2, 3]	Without replacement
Ordered	Unordered	Type \ Sample

by which we mean that to an instance of an ordered (unordered) sample of n balls from an urn containing M balls there corresponds (one and only one) instance of distributing n distinguishable (indistinguishable) objects among M cells.

In a similar sense we have the following correspondences:

$$\text{(sampling with replacement)} \leftrightarrow \begin{pmatrix}\text{a cell may receive any number}\\ \text{of objects}\end{pmatrix},$$

$$\text{(sampling without replacement)} \leftrightarrow \begin{pmatrix}\text{a cell may receive at most}\\ \text{one object}\end{pmatrix}.$$

These correspondences generate others of the same kind:

$$\begin{pmatrix}\text{an unordered sample in}\\ \text{sampling without}\\ \text{replacement}\end{pmatrix} \leftrightarrow \begin{pmatrix}\text{indistinguishable objects in the}\\ \text{problem of distribution among cells}\\ \text{when each cell may receive at}\\ \text{most one object}\end{pmatrix}$$

etc.; so that we can use Examples 4 and 5 to describe the sample space for the problem of distributing distinguishable or indistinguishable objects among cells either with exclusion (a cell may receive at most one object) or without exclusion (a cell may receive any number of objects).

Table 3 displays the distributions of two objects among three cells. For distinguishable objects, we denote them by W (white) and B (black). For indistinguishable objects, the presence of an object in a cell is indicated by a +.

Table 3

[WB][ ][ ] [W][B][ ] [W][ ][B] [BW][ ][ ] [ ][WB][ ] [ ][W][B] [B][ ][W] [ ][B][W] [ ][ ][WB]	[++][ ][ ] [ ][++][ ] [ ][ ][++] [+][+][ ] [+][ ][+] [ ][+][+]	Without exclusion
[W][B][ ] [W][ ][B] [B][W][ ] [ ][W][B] [B][ ][W] [ ][B][W]	[+][+][ ] [+][ ][+] [ ][+][+]	With exclusion
Distinguishable objects	Indistinguishable objects	Kind of objects \ Distribution

Table 4

$N(\Omega)$ in the problem of placing n objects in M cells			
Kind of objects / Distribution	Distinguishable objects	Indistinguishable objects	
Without exclusion	M^n (Maxwell–Boltzmann statistics)	C^n_{M+n-1} (Bose–Einstein statistics)	With replacement
With exclusion	$(M)_n$	C^n_M (Fermi–Dirac statistics)	Without replacement
	Ordered samples	Unordered samples	Sample / Type
	$N(\Omega)$ in the problem of choosing n balls from an urn containing M balls		

The duality that we have observed between the two problems gives us an obvious way of finding the number of outcomes in the problem of placing objects in cells. The results, which include the results in Table 1, are given in Table 4.

In statistical physics one says that distinguishable (or indistinguishable, respectively) particles that are not subject to the Pauli exclusion principle† obey Maxwell–Boltzmann statistics (or, respectively, Bose–Einstein statistics). If, however, the particles are indistinguishable and are subject to the exclusion principle, they obey Fermi–Dirac statistics (see Table 4). For example, electrons, protons and neutrons obey Fermi–Dirac statistics. Photons and pions obey Bose–Einstein statistics. Distinguishable particles that are subject to the exclusion principle do not occur in physics.

3. In addition to the concept of sample space we now need the fundamental concept of *event*.

Experimenters are ordinarily interested, not in what particular outcome occurs as the result of a trial, but in whether the outcome belongs to some subset of the set of all possible outcomes. We shall describe as *events* all subsets $A \subset \Omega$ for which, under the conditions of the experiment, it is possible to say either "the outcome $\omega \in A$" or "the outcome $\omega \notin A$."

† At most one particle in each cell. (Translator)

For example, let a coin be tossed three times. The sample space Ω consists of the eight points

$$\Omega = \{\text{HHH, HHT}, \ldots, \text{TTT}\}$$

and if we are able to observe (determine, measure, etc.) the results of all three tosses, we say that the set

$$A = \{\text{HHH, HHT, HTH, THH}\}$$

is the event consisting of the appearance of at least two heads. If, however, we can determine only the result of the first toss, this set A cannot be considered to be an event, since there is no way to give either a positive or negative answer to the question of whether a specific outcome ω belongs to A.

Starting from a given collection of sets that are events, we can form new events by means of statements containing the logical connectives "or," "and," and "not," which correspond in the language of set theory to the operations "union," "intersection," and "complement."

If A and B are sets, their *union*, denoted by $A \cup B$, is the set of points that belong either to A or to B:

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}.$$

In the language of probability theory, $A \cup B$ is the event consisting of the realization either of A or of B.

The *intersection* of A and B, denoted by $A \cap B$, or by AB, is the set of points that belong to both A and B:

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}.$$

The event $A \cap B$ consists of the simultaneous realization of both A and B.

For example, if $A = \{\text{HH, HT, TH}\}$ and $B = \{\text{TT, TH, HT}\}$ then

$$A \cup B = \{\text{HH, HT, TH, TT}\} \quad (=\Omega),$$
$$A \cap B = \{\text{TH, HT}\}.$$

If A is a subset of Ω, its *complement*, denoted by $\bar{A}$, is the set of points of Ω that do not belong to A.

If $B \backslash A$ denotes the *difference* of B and A (i.e. the set of points that belong to B but not to A) then $\bar{A} = \Omega \backslash A$. In the language of probability, $\bar{A}$ is the event consisting of the nonrealization of A. For example, if $A = \{\text{HH, HT, TH}\}$ then $\bar{A} = \{\text{TT}\}$, the event in which two successive tails occur.

The sets A and $\bar{A}$ have no points in common and consequently $A \cap \bar{A}$ is empty. We denote the empty set by $\varnothing$. In probability theory, $\varnothing$ is called an *impossible* event. The set Ω is naturally called the *certain* event.

When A and B are disjoint ($AB = \varnothing$), the union $A \cup B$ is called the *sum* of A and B and written $A + B$.

If we consider a collection $\mathscr{A}_0$ of sets $A \subseteq \Omega$ we may use the set-theoretic operators $\cup$, $\cap$ and $\backslash$ to form a new collection of sets from the elements of

$\mathscr{A}_0$; these sets are again events. If we adjoin the certain and impossible events Ω and $\varnothing$ we obtain a collection $\mathscr{A}$ of sets which is an *algebra*, i.e. a collection of subsets of Ω for which

(1) $\Omega \in \mathscr{A}$,
(2) if $A \in \mathscr{A}$, $B \in \mathscr{A}$, the sets $A \cup B$, $A \cap B$, $A \setminus B$ also belong to $\mathscr{A}$.

It follows from what we have said that it will be advisable to consider collections of events that form algebras. In the future we shall consider only such collections.

Here are some examples of algebras of events:

(a) $\{\Omega, \varnothing\}$, the collection consisting of Ω and the empty set (we call this the *trivial* algebra);
(b) $\{A, \bar{A}, \Omega, \varnothing\}$, the collection generated by A;
(c) $\mathscr{A} = \{A: A \subseteq \Omega\}$, the collection consisting of *all* the subsets of Ω (including the empty set $\varnothing$).

It is easy to check that all these algebras of events can be obtained from the following principle.

We say that a collection

$$\mathscr{D} = \{D_1, \ldots, D_n\}$$

of sets is a *decomposition* of Ω, and call the D_i the *atoms* of the decomposition, if the D_i are not empty, are pairwise disjoint, and their sum is Ω:

$$D_1 + \cdots + D_n = \Omega.$$

For example, if Ω consists of three points, $\Omega = \{1, 2, 3\}$, there are five different decompositions:

$$\begin{aligned}
&\mathscr{D}_1 = \{D_1\} && \text{with} \quad D_1 = \{1, 2, 3\}; \\
&\mathscr{D}_2 = \{D_1, D_2\} && \text{with} \quad D_1 = \{1, 2\}, D_2 = \{3\}; \\
&\mathscr{D}_3 = \{D_1, D_2\} && \text{with} \quad D_1 = \{1, 3\}, D_2 = \{2\}; \\
&\mathscr{D}_4 = \{D_1, D_2\} && \text{with} \quad D_1 = \{2, 3\}, D_2 = \{1\}; \\
&\mathscr{D}_5 = \{D_1, D_2, D_3\} && \text{with} \quad D_1 = \{1\}, D_2 = \{2\}, D_3 = \{3\}.
\end{aligned}$$

(For the general number of decompositions of a finite set see Problem 2.)

If we consider all unions of the sets in $\mathscr{D}$, the resulting collection of sets, together with the empty set, forms an algebra, called the *algebra induced by* $\mathscr{D}$, and denoted by $\alpha(\mathscr{D})$. Thus the elements of $\alpha(\mathscr{D})$ consist of the empty set together with the sums of sets which are atoms of $\mathscr{D}$.

Thus if $\mathscr{D}$ is a decomposition, there is associated with it a specific algebra $\mathscr{B} = \alpha(\mathscr{D})$.

The converse is also true. Let $\mathscr{B}$ be an algebra of subsets of a finite space Ω. Then there is a unique decomposition $\mathscr{D}$ whose atoms are the elements of

$\mathscr{B}$, with $\mathscr{B} = \alpha(\mathscr{D})$. In fact, let $D \in \mathscr{B}$ and let D have the property that for every $B \in \mathscr{B}$ the set $D \cap B$ either coincides with D or is empty. Then this collection of sets D forms a decomposition $\mathscr{D}$ with the required property $\alpha(\mathscr{D}) = \mathscr{B}$. In Example (a), $\mathscr{D}$ is the trivial decomposition consisting of the single set $D_1 = \Omega$; in (b), $\mathscr{D} = \{A, \bar{A}\}$. The most fine-grained decomposition $\mathscr{D}$, which consists of the singletons $\{\omega_i\}$, $\omega_i \in \Omega$, induces the algebra in Example (c), i.e. the algebra of all subsets of Ω.

Let $\mathscr{D}_1$ and $\mathscr{D}_2$ be two decompositions. We say that $\mathscr{D}_2$ is finer than $\mathscr{D}_1$, and write $\mathscr{D}_1 \preccurlyeq \mathscr{D}_2$, if $\alpha(\mathscr{D}_1) \subseteq \alpha(\mathscr{D}_2)$.

Let us show that if Ω consists, as we assumed above, of a finite number of points $\omega_1, \ldots, \omega_N$, then the number $N(\mathscr{A})$ of sets in the collection $\mathscr{A}$ is equal to 2^N. In fact, every nonempty set $A \in \mathscr{A}$ can be represented as $A = \{\omega_{i_1}, \ldots, \omega_{i_k}\}$, where $\omega_{i_j} \in \Omega$, $1 \le k \le N$. With this set we associate the sequence of zeros and ones

$$(0, \ldots, 0, 1, 0, \ldots, 0, 1, \ldots),$$

where there are ones in the positions $i_1, \ldots, i_k$ and zeros elsewhere. Then for a given k the number of different sets A of the form $\{\omega_{i_1}, \ldots, \omega_{i_k}\}$ is the same as the number of ways in which k ones (k indistinguishable objects) can be placed in N positions (N cells). According to Table 4 (see the lower right-hand square) we see that this number is C_N^k. Hence (counting the empty set) we find that

$$N(\mathscr{A}) = 1 + C_N^1 + \cdots + C_N^N = (1 + 1)^N = 2^N.$$

4. We have now taken the first two steps in defining a probabilistic model of an experiment with a finite number of outcomes: we have selected a sample space and a collection $\mathscr{A}$ of subsets, which form an algebra and are called events. We now take the next step, to assign to each sample point (outcome) $\omega_i \in \Omega_i$, $i = 1, \ldots, N$, a *weight*. This is denoted by $p(\omega_i)$ and called the *probability* of the outcome ω_i; we assume that it has the following properties:

(a) $0 \le p(\omega_i) \le 1$ (nonnegativity),
(b) $p(\omega_1) + \cdots + p(\omega_N) = 1$ (normalization).

Starting from the given probabilities $p(\omega_i)$ of the outcomes ω_i, we define the probability $\mathsf{P}(A)$ of any event $A \in \mathscr{A}$ by

$$\mathsf{P}(A) = \sum_{\{i:\, \omega_i \in A\}} p(\omega_i). \tag{4}$$

Finally, we say that a triple

$$(\Omega, \mathscr{A}, \mathsf{P}),$$

where $\Omega = \{\omega_1, \ldots, \omega_N\}$, $\mathscr{A}$ is an algebra of subsets of Ω and

$$\mathsf{P} = \{\mathsf{P}(A);\ A \in \mathscr{A}\}$$

defines (or assigns) a *probabilistic model*, or a *probability space*, of experiments with a (finite) space Ω of outcomes and algebra $\mathscr{A}$ of events.

The following properties of probability follow from (4):

$$\mathsf{P}(\varnothing) = 0, \tag{5}$$

$$\mathsf{P}(\Omega) = 1, \tag{6}$$

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B). \tag{7}$$

In particular, if $A \cap B = \varnothing$, then

$$\mathsf{P}(A + B) = \mathsf{P}(A) + \mathsf{P}(B) \tag{8}$$

and

$$\mathsf{P}(\bar{A}) = 1 - \mathsf{P}(A). \tag{9}$$

5. In constructing a probabilistic model for a specific situation, the construction of the sample space Ω and the algebra $\mathscr{A}$ of events are ordinarily not difficult. In elementary probability theory one usually takes the algebra $\mathscr{A}$ to be the algebra of *all* subsets of Ω. Any difficulty that may arise is in assigning probabilities to the sample points. In principle, the solution to this problem lies outside the domain of probability theory, and we shall not consider it in detail. We consider that our fundamental problem is not the question of how to assign probabilities, but how to calculate the probabilities of complicated events (elements of $\mathscr{A}$) from the probabilities of the sample points.

It is clear from a mathematical point of view that for finite sample spaces we can obtain all conceivable (finite) probability spaces by assigning nonnegative numbers $p_1, \ldots, p_N$, satisfying the condition $p_1 + \cdots + p_N = 1$, to the outcomes $\omega_1, \ldots, \omega_N$.

The validity of the assignments of the numbers $p_1, \ldots, p_N$ can, in specific cases, be checked to a certain extent by using the law of large numbers (which will be discussed later on). It states that in a long series of "independent" experiments, carried out under identical conditions, the frequencies with which the elementary events appear are "close" to their probabilities.

In connection with the difficulty of assigning probabilities to outcomes, we note that there are many actual situations in which for reasons of symmetry it seems reasonable to consider all conceivable outcomes as equally probable. In such cases, if the sample space consists of points $\omega_1, \ldots, \omega_N$, with $N < \infty$, we put

$$p(\omega_1) = \cdots = p(\omega_N) = 1/N,$$

and consequently

$$\mathsf{P}(A) = N(A)/N \tag{10}$$

for every event $A \in \mathscr{A}$, where $N(A)$ is the number of sample points in A. This is called the classical method of assigning probabilities. It is clear that in this case the calculation of $\mathsf{P}(A)$ reduces to calculating the number of outcomes belonging to A. This is usually done by combinatorial methods, so that combinatorics, applied to finite sets, plays a significant role in the calculus of probabilities.

EXAMPLE 7 (Coincidence problem). Let an urn contain M balls numbered $1, 2, \ldots, M$. We draw an ordered sample of size n with replacement. It is clear that then

$$\Omega = \{\omega: \omega = (a_1, \ldots, a_n), a_i = 1, \ldots, M\}$$

and $N(\Omega) = M^n$. Using the classical assignment of probabilities, we consider the M^n outcomes equally probable and ask for the probability of the event

$$A = \{\omega: \omega = (a_1, \ldots, a_n), a_i \neq a_j, i \neq j\},$$

i.e., the event in which there is no repetition. Clearly $N(A) = M(M-1)\cdots(M-n+1)$, and therefore

$$\mathsf{P}(A) = \frac{(M)_n}{M^n} = \left(1 - \frac{1}{M}\right)\left(1 - \frac{2}{M}\right)\cdots\left(1 - \frac{n-1}{M}\right). \tag{11}$$

This problem has the following striking interpretation. Suppose that there are n students in a class. Let us suppose that each student's birthday is on one of 365 days and that all days are equally probable. The question is, what is the probability P_n that there are at least two students in the class whose birthdays coincide? If we interpret selection of birthdays as selection of balls from an urn containing 365 balls, then by (11)

$$P_n = 1 - \frac{(365)_n}{365^n}.$$

The following table lists the values of P_n for some values of n:

n	4	16	22	23	40	64
P_n	0.016	0.284	0.476	0.507	0.891	0.997

It is interesting to note that (unexpectedly!) the size of class in which there is probability $\frac{1}{2}$ of finding at least two students with the same birthday is not very large: only 23.

EXAMPLE 8 (Prizes in a lottery). Consider a lottery that is run in the following way. There are M tickets numbered $1, 2, \ldots, M$, of which n, numbered $1, \ldots, n$, win prizes ($M \geq 2n$). You buy n tickets, and ask for the probability (P, say) of winning at least one prize.

Since the order in which the tickets are drawn plays no role in the presence or absence of winners in your purchase, we may suppose that the sample space has the form

$$\Omega = \{\omega: \omega = [a_1, \ldots, a_n], a_k \neq a_l, k \neq l, a_i = 1, \ldots, M\}.$$

By Table 1, $N(\Omega) = C_M^n$. Now let

$$A_0 = \{\omega: \omega = [a_1, \ldots, a_n], a_k \neq a_l, k \neq l, a_i = n + 1, \ldots, M\}$$

be the event that there is no winner in the set of tickets you bought. Again by Table 1, $N(A_0) = C_{M-n}^n$. Therefore

$$\mathsf{P}(A_0) = \frac{C_{M-n}^n}{C_M^n} = \frac{(M-n)_n}{(M)_n}$$
$$= \left(1 - \frac{n}{M}\right)\left(1 - \frac{n}{M-1}\right)\cdots\left(1 - \frac{n}{M-n+1}\right)$$

and consequently

$$P = 1 - \mathsf{P}(A_0) = 1 - \left(1 - \frac{n}{M}\right)\left(1 - \frac{n}{M-1}\right)\cdots\left(1 - \frac{n}{M-n+1}\right).$$

If $M = n^2$ and $n \to \infty$, then $P(A_0) \to e^{-1}$ and

$$P \to 1 - e^{-1} \approx 0.632.$$

The convergence is quite fast: for $n = 10$ the probability is already $P = 0.670$.

6. Problems

1. Establish the following properties of the operators $\cap$ and $\cup$:

$$A \cup B = B \cup A, \quad AB = BA \quad \text{(commutativity)},$$
$$A \cup (B \cup C) = (A \cup B) \cup C, \quad A(BC) = (AB)C \quad \text{(associativity)},$$
$$A(B \cup C) = AB \cup AC, \quad A \cup (BC) = (A \cup B)(A \cup C) \quad \text{(distributivity)},$$
$$A \cup A = A, \quad AA = A \quad \text{(idempotency)}.$$

Show also that

$$\overline{A \cup B} = \bar{A} \cap \bar{B}, \quad \overline{AB} = \bar{A} \cup \bar{B}.$$

2. Let Ω contain N elements. Show that the number $d(N)$ of different decompositions of Ω is given by the formula

$$d(N) = e^{-1} \sum_{k=0}^{\infty} \frac{k^N}{k!}. \tag{12}$$

(Hint: Show that

$$d(N) = \sum_{k=0}^{N-1} C_{N-1}^k d(k), \quad \text{where} \quad d(0) = 1,$$

and then verify that the series in (12) satisfies the same recurrence relation.)

3. For any finite collection of sets $A_1, \ldots, A_n$,

$$\mathsf{P}(A_1 \cup \cdots \cup A_n) \leq \mathsf{P}(A_1) + \cdots + \mathsf{P}(A_n).$$

4. Let A and B be events. Show that $A\bar{B} \cup B\bar{A}$ is the event in which exactly one of A and B occurs. Moreover,

$$P(A\bar{B} \cup B\bar{A}) = \mathsf{P}(A) + \mathsf{P}(B) - 2\mathsf{P}(AB).$$

5. Let $A_1, \ldots, A_n$ be events, and define $S_0, S_1, \ldots, S_n$ as follows: $S_0 = 1$,

$$S_r = \sum_{J_r} \mathsf{P}(A_{k_1} \cap \cdots \cap A_{k_r}), \qquad 1 \leq r \leq n,$$

where the sum is over the unordered subsets $J_r = [k_1, \ldots, k_r]$ of $\{1, \ldots, n\}$.

Let B_m be the event in which each of the events $A_1, \ldots, A_n$ occurs exactly m times. Show that

$$\mathsf{P}(B_m) = \sum_{r=m}^{n} (-1)^{r-m} C_r^m S_r.$$

In particular, for $m = 0$

$$\mathsf{P}(B_0) = 1 - S_1 + S_2 - \cdots \pm S_n.$$

Show also that the probability that at least m of the events $A_1, \ldots, A_n$ occur simultaneously is

$$\mathsf{P}(B_1) + \cdots + \mathsf{P}(B_n) = \sum_{r=m}^{n} (-1)^{r-m} C_{r-1}^{m-1} S_r.$$

In particular, the probability that at least one of the events $A_1, \ldots, A_n$ occurs is

$$\mathsf{P}(B_1) + \cdots + \mathsf{P}(B_n) = S_1 - S_2 + \cdots \pm S_n.$$

§2. Some Classical Models and Distributions

1. Binomial distribution. Let a coin be tossed n times and record the results as an ordered set $(a_1, \ldots, a_n)$, where $a_i = 1$ for a head ("success") and $a_i = 0$ for a tail ("failure"). The sample space is

$$\Omega = \{\omega : \omega = (a_1, \ldots, a_n), a_i = 0, 1\}.$$

To each sample point $\omega = (a_1, \ldots, a_n)$ we assign the probability

$$p(\omega) = p^{\Sigma a_i} q^{n - \Sigma a_i},$$

where the nonnegative numbers p and q satisfy $p + q = 1$. In the first place, we verify that this assignment of the weights $p(\omega)$ is consistent. It is enough to show that $\sum_{\omega \in \Omega} p(\omega) = 1$.

We consider all outcomes $\omega = (a_1, \ldots, a_n)$ for which $\sum_i a_i = k$, where $k = 0, 1, \ldots, n$. According to Table 4 (distribution of k indistinguishable ones in n places) the number of these outcomes is C_n^k. Therefore

$$\sum_{\omega\in\Omega} p(\omega) = \sum_{k=0}^{n} C_n^k p^k q^{n-k} = (p+q)^n = 1.$$

Thus the space Ω together with the collection $\mathscr{A}$ of all its subsets and the probabilities $\mathsf{P}(A) = \sum_{\omega\in A} p(\omega)$, $A \in \mathscr{A}$, defines a probabilistic model. It is natural to call this the probabilistic model for n tosses of a coin.

In the case $n = 1$, when the sample space contains just the two points $\omega = 1$ ("success") and $\omega = 0$ ("failure"), it is natural to call $p(1) = p$ the probability of success. We shall see later that this model for n tosses of a coin can be thought of as the result of n "independent" experiments with probability p of success at each trial.

Let us consider the events

$$A_k = \{\omega: \omega = (a_1, \ldots, a_n), a_1 + \cdots + a_n = k\}, \qquad k = 0, 1, \ldots, n,$$

consisting of exactly k successes. It follows from what we said above that

$$\mathsf{P}(A_k) = C_n^k p^k q^{n-k}, \tag{1}$$

and $\sum_{k=0}^{n} \mathsf{P}(A_k) = 1$.

The set of probabilities $(\mathsf{P}(A_0), \ldots, \mathsf{P}(A_n))$ is called the *binomial distribution* (the number of successes in a sample of size n). This distribution plays an extremely important role in probability theory since it arises in the most diverse probabilistic models. We write $P_n(k) = \mathsf{P}(A_k)$, $k = 0, 1, \ldots, n$. Figure 1 shows the binomial distribution in the case $p = \frac{1}{2}$ (symmetric coin) for $n = 5, 10, 20$.

We now present a different model (in essence, equivalent to the preceding one) which describes the random walk of a "particle."

Let the particle start at the origin, and after unit time let it take a unit step upward or downward (Figure 2).

Consequently after n steps the particle can have moved at most n units up or n units down. It is clear that each path ω of the particle is completely specified by a set $(a_1, \ldots, a_n)$, where $a_i = +1$ if the particle moves up at the ith step, and $a_i = -1$ if it moves down. Let us assign to each path ω the weight $p(\omega) = p^{\nu(\omega)} q^{n-\nu(\omega)}$, where $\nu(\omega)$ is the number of $+1$'s in the sequence $\omega = (a_1, \ldots, a_n)$, i.e. $\nu(\omega) = [(a_1 + \cdots + a_n) + n]/2$, and the nonnegative numbers p and q satisfy $p + q = 1$.

Since $\sum_{\omega\in\Omega} p(\omega) = 1$, the set of probabilities $p(\omega)$ together with the space Ω of paths $\omega = (a_1, \ldots, a_n)$ and its subsets define an acceptable probabilistic model of the motion of the particle for n steps.

Let us ask the following question: What is the probability of the event A_k that after n steps the particle is at a point with ordinate k? This condition is satisfied by those paths ω for which $\nu(\omega) - (n - \nu(\omega)) = k$, i.e.

$$\nu(\omega) = \frac{n+k}{2}.$$

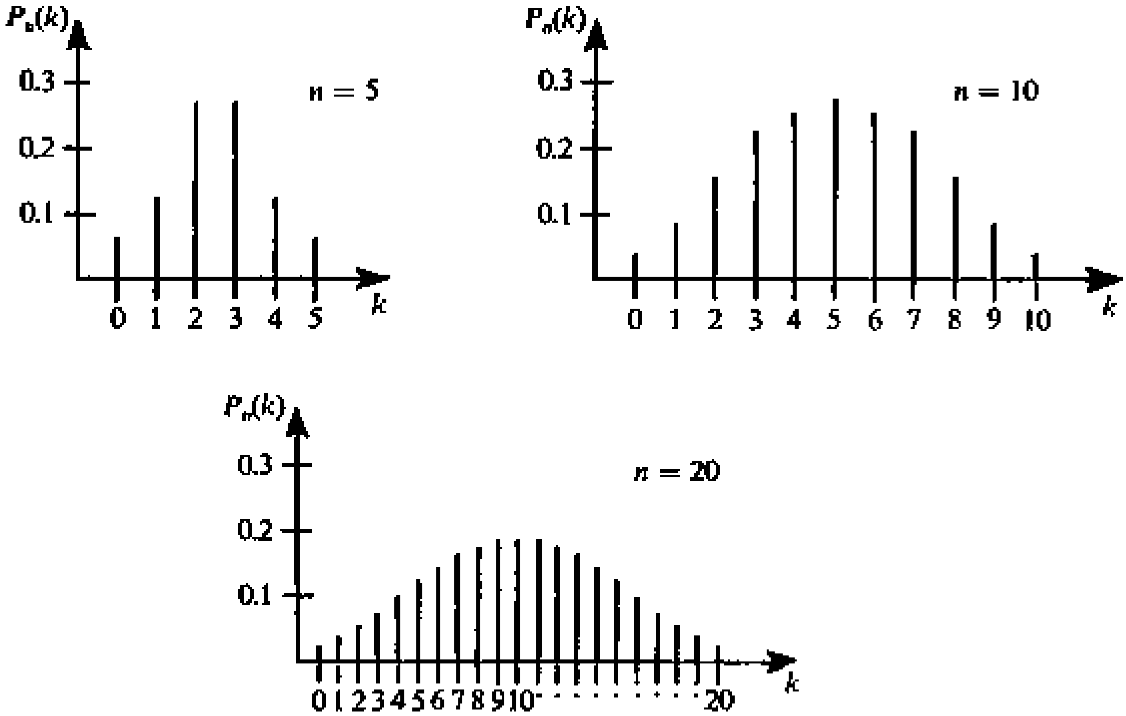


Figure 1. Graph of the binomial probabilities $P_n(k)$ for $n = 5, 10, 20$.

The number of such paths (see Table 4) is $C_n^{[n+k]/2}$, and therefore

$$\mathsf{P}(A_k) = C_n^{[n+k]/2} p^{[n+k]/2} q^{[n-k]/2}.$$

Consequently the binomial distribution $(\mathsf{P}(A_{-n}), \ldots, \mathsf{P}(A_0), \ldots, \mathsf{P}(A_n))$ can be said to describe the probability distribution for the position of the particle after n steps.

Note that in the symmetric case ($p = q = \frac{1}{2}$) when the probabilities of the individual paths are equal to 2^{-n},

$$\mathsf{P}(A_k) = C_n^{[n+k]/2} \cdot 2^{-n}.$$

Let us investigate the asymptotic behavior of these probabilities for large n.

If the number of steps is $2n$, it follows from the properties of the binomial coefficients that the largest of the probabilities $\mathsf{P}(A_k)$, $|k| \leq 2n$, is

$$\mathsf{P}(A_0) = C_{2n}^n \cdot 2^{-2n}.$$

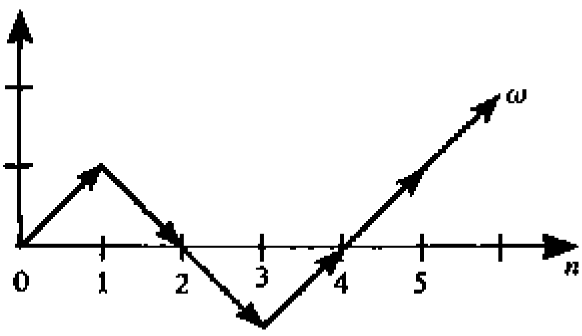


Figure 2

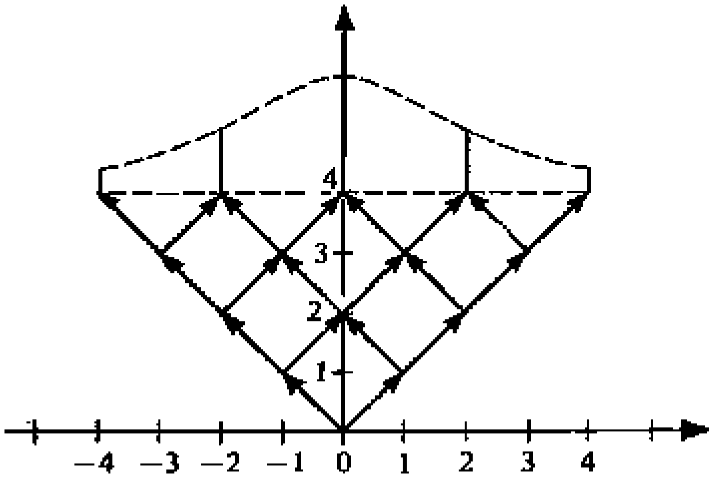


Figure 3. Beginning of the binomial distribution.

From Stirling's formula (see formula (6) in Section 4)

$$n! \sim \sqrt{2\pi n}\, e^{-n} n^n.†$$

Consequently

$$C_{2n}^n = \frac{(2n)!}{(n!)^2} \sim 2^{2n} \cdot \frac{1}{\sqrt{\pi n}}$$

and therefore for large n

$$\mathsf{P}(A_0) \sim \frac{1}{\sqrt{\pi n}}.$$

Figure 3 represents the beginning of the binomial distribution for $2n$ steps of a random walk (in contrast to Figure 2, the time axis is now directed upward).

2. Multinomial distribution. Generalizing the preceding model, we now suppose that the sample space is

$$\Omega = \{\omega: \omega = (a_1, \ldots, a_n),\, a_i = b_1, \ldots, b_r\},$$

where $b_1, \ldots, b_r$ are given numbers. Let $v_i(\omega)$ be the number of elements of $\omega = (a_1, \ldots, a_n)$ that are equal to b_i, $i = 1, \ldots, r$, and define the probability of ω by

$$p(\omega) = p_1^{v_1(\omega)} \cdots p_r^{v_r(\omega)},$$

where $p_i \geq 0$ and $p_1 + \cdots + p_r = 1$. Note that

$$\sum_{\omega \in \Omega} p(\omega) = \sum_{\left\{\substack{n_1 \geq, \ldots, n_r \geq 0, \\ n_1 + \cdots + n_r = n}\right\}} C_n(n_1, \ldots, n_r) p_1^{n_1} \cdots p_r^{n_r},$$

where $C_n(n_1, \ldots, n_r)$ is the number of (ordered) sequences $(a_1, \ldots, a_n)$ in which b_1 occurs n_1 times, ..., b_r occurs n_r times. Since n_1 elements b_1 can

† The notation $f(n) \sim g(n)$ means that $f(n)/g(n) \to 1$ as $n \to \infty$.

be distributed into n positions in $C_n^{n_1}$ ways; n_2 elements b_2 into $n - n_1$ positions in $C_{n-n_1}^{n_2}$ ways, etc., we have

$$\begin{aligned} C_n(n_1, \ldots, n_r) &= C_n^{n_1} \cdot C_{n-n_1}^{n_2} \cdots C_{n-(n_1+\cdots+n_{r-1})}^{n_r} \\ &= \frac{n!}{n_1!\,(n-n_1)!} \cdot \frac{(n-n_1)!}{n_2!\,(n-n_1-n_2)!} \cdots 1 \\ &= \frac{n!}{n_1! \cdots n_r!}. \end{aligned}$$

Therefore

$$\sum_{\omega \in \Omega} p(\omega) = \sum_{\substack{n_1 \geq 0, \ldots, n_r \geq 0, \\ n_1 + \cdots + n_r = n}} \frac{n!}{n_1! \cdots n_r!} p_1^{n_1} \cdots p_r^{n_r} = (p_1 + \cdots + p_r)^n = 1,$$

and consequently we have defined an acceptable method of assigning probabilities.

Let

$$A_{n_1, \ldots, n_r} = \{\omega: \nu_1(\omega) = n_1, \ldots, \nu_r(\omega) = n_r\}.$$

Then

$$\mathsf{P}(A_{n_1, \ldots, n_r}) = C_n(n_1, \ldots, n_r) p_1^{n_1} \cdots p_r^{n_r}. \tag{2}$$

The set of probabilities

$$\{\mathsf{P}(A_{n_1, \ldots, n_r})\}$$

is called the *multinomial* (or polynomial) distribution.

We emphasize that both this distribution and its special case, the binomial distribution, originate from problems about sampling *with replacement*.

3. The multidimensional hypergeometric distribution occurs in problems that involve sampling *without replacement*.

Consider, for example, an urn containing M balls numbered $1, 2, \ldots, M$, where M_1 balls have the color $b_1, \ldots, M_r$ balls have the color b_r, and $M_1 + \cdots + M_r = M$. Suppose that we draw a sample of size $n < M$ without replacement. The sample space is

$$\Omega = \{\omega: \omega = (a_1, \ldots, a_n), a_k \neq a_l, k \neq l, a_i = 1, \ldots, M\}$$

and $N(\Omega) = (M)_n$. Let us suppose that the sample points are equiprobable, and find the probability of the event $B_{n_1, \ldots, n_r}$ in which n_1 balls have color $b_1, \ldots, n_r$ balls have color b_r, where $n_1 + \cdots + n_r = n$. It is easy to show that

$$N(B_{n_1, \ldots, n_r}) = C_n(n_1, \ldots, n_r)(M_1)_{n_1} \cdots (M_r)_{n_r},$$

and therefore

$$\mathsf{P}(B_{n_1,\ldots,n_r}) = \frac{N(B_{n_1,\ldots,n_r})}{N(\Omega)} = \frac{C_{M_1}^{n_1} \cdots C_{M_r}^{n_r}}{C_M^n}. \tag{3}$$

The set of probabilities $\{\mathsf{P}(B_{n_1,\ldots,n_r})\}$ is called the *multidimensional hypergeometric distribution*. When $r = 2$ it is simply called the *hypergeometric* distribution because its "generating function" is a hypergeometric function.

The structure of the multidimensional hypergeometric distribution is rather complicated. For example, the probability

$$\mathsf{P}(B_{n_1,n_2}) = \frac{C_{M_1}^{n_1} C_{M_2}^{n_2}}{C_M^n}, \qquad n_1 + n_2 = n, \quad M_1 + M_2 = M, \tag{4}$$

contains nine factorials. However, it is easily established that if $M \to \infty$ and $M_1 \to \infty$ in such a way that $M_1/M \to p$ (and therefore $M_2/M \to 1 - p$) then

$$\mathsf{P}(B_{n_1,n_2}) \to C_{n_1+n_2}^{n_1} p^{n_1}(1-p)^{n_2}. \tag{5}$$

In other words, under the present hypotheses the hypergeometric distribution is approximated by the binomial; this is intuitively clear since when M and M_1 are large (but finite), sampling without replacement ought to give almost the same result as sampling with replacement.

Example. Let us use (4) to find the probability of picking six "lucky" numbers in a lottery of the following kind (this is an abstract formulation of the "sportloto," which is well known in Russia):

There are 49 balls numbered from 1 to 49; six of them are lucky (colored red, say, whereas the rest are white). We draw a sample of six balls, without replacement. The question is, What is the probability that all six of these balls are lucky? Taking $M = 49$, $M_1 = 6$, $n_1 = 6$, $n_2 = 0$, we see that the event of interest, namely

$$B_{6,0} = \{6 \text{ balls, all lucky}\}$$

has, by (4), probability

$$\mathsf{P}(B_{6,0}) = \frac{1}{C_{49}^6} \approx 7.2 \times 10^{-8}.$$

4. The numbers $n!$ increase extremely rapidly with n. For example,

$$10! = 3{,}628{,}800,$$

$$15! = 1{,}307{,}674{,}368{,}000,$$

and $100!$ has 158 digits. Hence from either the theoretical or the computational point of view, it is important to know Stirling's formula,

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left(\frac{\theta_n}{12n}\right), \qquad 0 < \theta_n < 1, \tag{6}$$

whose proof can be found in most textbooks on mathematical analysis (see also [69]).

5. Problems

1. Prove formula (5).

2. Show that for the multinomial distribution $\{\mathsf{P}(A_{n_1}, \ldots, A_{n_r})\}$ the maximum probability is attained at a point $(k_1, \ldots, k_r)$ that satisfies the inequalities $np_i - 1 < k_i \leq (n + r - 1)p_i$, $i = 1, \ldots, r$.

3. *One-dimensional Ising model.* Consider n particles located at the points $1, 2, \ldots, n$. Suppose that each particle is of one of two types, and that there are n_1 particles of the first type and n_2 of the second ($n_1 + n_2 = n$). We suppose that all $n!$ arrangements of the particles are equally probable.

   Construct a corresponding probabilistic model and find the probability of the event $A_n(m_{11}, m_{12}, m_{21}, m_{22}) = \{\nu_{11} = m_{11}, \ldots, \nu_{22} = m_{22}\}$, where ν_{ij} is the number of particles of type i following particles of type j $(i, j = 1, 2)$.

4. Prove the following inequalities by probabilistic reasoning:

$$\sum_{k=0}^{n} C_n^k = 2^n,$$

$$\sum_{k=0}^{n} (C_n^k)^2 = C_{2n}^n,$$

$$\sum_{k=0}^{n} (-1)^{n-k} C_m^k = C_{m-1}^n, \qquad m \geq n + 1,$$

$$\sum_{k=0}^{n} k(k-1) C_m^k = m(m-1)2^{m-2}, \qquad m \geq 2.$$

§3. Conditional Probability. Independence

1. The concept of probabilities of events lets us answer questions of the following kind: If there are M balls in an urn, M_1 white and M_2 black, what is the probability $\mathsf{P}(A)$ of the event A that a selected ball is white? With the classical approach, $\mathsf{P}(A) = M_1/M$.

The concept of *conditional probability*, which will be introduced below, lets us answer questions of the following kind: What is the probability that the second ball is white (event B) under the condition that the first ball was also white (event A)? (We are thinking of sampling without replacement.)

It is natural to reason as follows: if the first ball is white, then at the second step we have an urn containing $M - 1$ balls, of which $M_1 - 1$ are white and M_2 black; hence it seems reasonable to suppose that the (conditional) probability in question is $(M_1 - 1)/(M - 1)$.

We now give a definition of conditional probability that is consistent with our intuitive ideas.

Let $(\Omega, \mathscr{A}, \mathsf{P})$ be a (finite) probability space and A an event (i.e. $A \in \mathscr{A}$).

Definition 1. The *conditional probability* of event B assuming event A with $\mathsf{P}(A) > 0$ (denoted by $\mathsf{P}(B|A)$) is

$$\frac{\mathsf{P}(AB)}{\mathsf{P}(A)}. \tag{1}$$

In the classical approach we have $\mathsf{P}(A) = N(A)/N(\Omega)$, $\mathsf{P}(AB) = N(AB)/N(\Omega)$, and therefore

$$\mathsf{P}(B|A) = \frac{N(AB)}{N(A)}. \tag{2}$$

From Definition 1 we immediately get the following properties of conditional probability:

$$\begin{aligned} \mathsf{P}(A|A) &= 1, \\ \mathsf{P}(\varnothing|A) &= 0, \\ \mathsf{P}(B|A) &= 1, \qquad B \supseteq A, \\ \mathsf{P}(B_1 + B_2|A) &= \mathsf{P}(B_1|A) + \mathsf{P}(B_2|A). \end{aligned}$$

It follows from these properties that for a given set A the conditional probability $\mathsf{P}(\cdot|A)$ has the same properties on the space $(\Omega \cap A, \mathscr{A} \cap A)$, where $\mathscr{A} \cap A = \{B \cap A: B \in \mathscr{A}\}$, that the original probability $\mathsf{P}(\cdot)$ has on $(\Omega, \mathscr{A})$.

Note that

$$\mathsf{P}(B|A) + \mathsf{P}(\bar{B}|A) = 1;$$

however in general

$$\begin{aligned} \mathsf{P}(B|A) + \mathsf{P}(B|\bar{A}) &\neq 1, \\ \mathsf{P}(B|A) + \mathsf{P}(\bar{B}|\bar{A}) &\neq 1. \end{aligned}$$

Example 1. Consider a family with two children. We ask for the probability that both children are boys, assuming

(a) that the older child is a boy;
(b) that at least one of the children is a boy.

The sample space is

$$\Omega = \{\text{BB}, \text{BG}, \text{GB}, \text{GG}\},$$

where BG means that the older child is a boy and the younger is a girl, etc.

Let us suppose that all sample points are equally probable:

$$\mathsf{P}(\text{BB}) = \mathsf{P}(\text{BG}) = \mathsf{P}(\text{GB}) = \mathsf{P}(\text{GG}) = \tfrac{1}{4}.$$

Let A be the event that the older child is a boy, and B, that the younger child is a boy. Then $A \cup B$ is the event that at least one child is a boy, and AB is the event that both children are boys. In question (a) we want the conditional probability $\mathsf{P}(AB \mid A)$, and in (b), the conditional probability $\mathsf{P}(AB \mid A \cup B)$.

It is easy to see that

$$\mathsf{P}(AB \mid A) = \frac{\mathsf{P}(AB)}{\mathsf{P}(A)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2},$$

$$\mathsf{P}(AB \mid A \cup B) = \frac{\mathsf{P}(AB)}{\mathsf{P}(A \cup B)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

2. The simple but important formula (3), below, is called the formula for total probability. It provides the basic means for calculating the probabilities of complicated events by using conditional probabilities.

Consider a decomposition $\mathscr{D} = \{A_1, \ldots, A_n\}$ with $\mathsf{P}(A_i) > 0$, $i = 1, \ldots, n$ (such a decomposition is often called a complete set of disjoint events). It is clear that

$$B = BA_1 + \cdots + BA_n$$

and therefore

$$\mathsf{P}(B) = \sum_{i=1}^{n} \mathsf{P}(BA_i).$$

But

$$\mathsf{P}(BA_i) = \mathsf{P}(B \mid A_i)\mathsf{P}(A_i).$$

Hence we have the *formula for total probability*:

$$\mathsf{P}(B) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i)\mathsf{P}(A_i). \tag{3}$$

In particular, if $0 < \mathsf{P}(A) < 1$, then

$$\mathsf{P}(B) = \mathsf{P}(B \mid A)\mathsf{P}(A) + \mathsf{P}(B \mid \bar{A})\mathsf{P}(\bar{A}). \tag{4}$$

EXAMPLE 2. An urn contains M balls, m of which are "lucky." We ask for the probability that the second ball drawn is lucky (assuming that the result of the first draw is unknown, that a sample of size 2 is drawn without replacement, and that all outcomes are equally probable). Let A be the event that the first ball is lucky, B the event that the second is lucky. Then

$$\mathsf{P}(B|A) = \frac{\mathsf{P}(BA)}{\mathsf{P}(A)} = \frac{\dfrac{m(m-1)}{M(M-1)}}{\dfrac{m}{M}} = \frac{m-1}{M-1},$$

$$\mathsf{P}(B|\bar{A}) = \frac{\mathsf{P}(B\bar{A})}{\mathsf{P}(\bar{A})} = \frac{\dfrac{m(M-m)}{M(M-1)}}{\dfrac{M-m}{M}} = \frac{m}{M-1}$$

and

$$\begin{aligned}\mathsf{P}(B) &= \mathsf{P}(B|A)\mathsf{P}(A) + \mathsf{P}(B|\bar{A})\mathsf{P}(\bar{A}) \\ &= \frac{m-1}{M-1}\cdot\frac{m}{M} + \frac{m}{M-1}\cdot\frac{M-m}{M} = \frac{m}{M}.\end{aligned}$$

It is interesting to observe that $\mathsf{P}(A)$ is precisely m/M. Hence, when the nature of the first ball is unknown, it does not affect the probability that the second ball is lucky.

By the definition of conditional probability (with $\mathsf{P}(A) > 0$),

$$\mathsf{P}(AB) = \mathsf{P}(B|A)\mathsf{P}(A). \tag{5}$$

This formula, the *multiplication formula for probabilities*, can be generalized (by induction) as follows: If $A_1, \ldots, A_{n-1}$ are events with $\mathsf{P}(A_1 \cdots A_{n-1}) > 0$, then

$$\mathsf{P}(A_1 \cdots A_n) = \mathsf{P}(A_1)\mathsf{P}(A_2|A_1)\cdots\mathsf{P}(A_n|A_1 \cdots A_{n-1}) \tag{6}$$

(here $A_1 \cdots A_n = A_1 \cap A_2 \cap \cdots \cap A_n$).

3. Suppose that A and B are events with $\mathsf{P}(A) > 0$ and $\mathsf{P}(B) > 0$. Then along with (5) we have the parallel formula

$$\mathsf{P}(AB) = \mathsf{P}(A|B)\mathsf{P}(B). \tag{7}$$

From (5) and (7) we obtain *Bayes's formula*

$$\mathsf{P}(A|B) = \frac{\mathsf{P}(A)\mathsf{P}(B|A)}{\mathsf{P}(B)}. \tag{8}$$

If the events $A_1, \ldots, A_n$ form a decomposition of Ω, (3) and (8) imply *Bayes's theorem*:

$$\mathsf{P}(A_i|B) = \frac{\mathsf{P}(A_i)\mathsf{P}(B|A_i)}{\sum_{j=1}^n \mathsf{P}(A_j)\mathsf{P}(B|A_j)}. \tag{9}$$

In statistical applications, $A_1, \ldots, A_n$ $(A_1 + \cdots + A_n = \Omega)$ are often called hypotheses, and $\mathsf{P}(A_i)$ is called the *a priori*† probability of A_i. The conditional probability $\mathsf{P}(A_i|B)$ is considered as the *a posteriori* probability of A_i after the occurrence of event B.

EXAMPLE 3. Let an urn contain two coins: A_1, a fair coin with probability $\frac{1}{2}$ of falling H; and A_2, a biased coin with probability $\frac{1}{3}$ of falling H. A coin is drawn at random and tossed. Suppose that it falls head. We ask for the probability that the fair coin was selected.

Let us construct the corresponding probabilistic model. Here it is natural to take the sample space to be the set $\Omega = \{A_1\text{H}, A_1\text{T}, A_2\text{H}, A_2\text{T}\}$, which describes all possible outcomes of a selection and a toss (A_1H means that coin A_1 was selected and fell heads, etc.) The probabilities $p(\omega)$ of the various outcomes have to be assigned so that, according to the statement of the problem,

$$\mathsf{P}(A_1) = \mathsf{P}(A_2) = \tfrac{1}{2}$$

and

$$\mathsf{P}(\text{H}|A_1) = \tfrac{1}{2}, \qquad \mathsf{P}(\text{H}|A_2) = \tfrac{1}{3}.$$

With these assignments, the probabilities of the sample points are uniquely determined:

$$\mathsf{P}(A_1\text{H}) = \tfrac{1}{4}, \qquad \mathsf{P}(A_1\text{T}) = \tfrac{1}{4}, \qquad \mathsf{P}(A_2\text{H}) = \tfrac{1}{6}, \qquad \mathsf{P}(A_2\text{T}) = \tfrac{1}{3}.$$

Then by Bayes's formula the probability in question is

$$\mathsf{P}(A_1|\text{H}) = \frac{\mathsf{P}(A_1)\mathsf{P}(\text{H}|A_1)}{\mathsf{P}(A_1)\mathsf{P}(\text{H}|A_1) + \mathsf{P}(A_2)\mathsf{P}(\text{H}|A_2)} = \frac{3}{5},$$

and therefore

$$\mathsf{P}(A_2|\text{H}) = \tfrac{2}{5}.$$

4. In certain sense, the concept of *independence*, which we are now going to introduce, plays a central role in probability theory: it is precisely this concept that distinguishes probability theory from the general theory of measure spaces.

† *A priori*: before the experiment; *a posteriori*: after the experiment.

If A and B are two events, it is natural to say that B is independent of A if knowing that A has occurred has no effect on the probability of B. In other words, "B is independent of A" if

$$\mathsf{P}(B|A) = \mathsf{P}(B) \tag{10}$$

(we are supposing that $\mathsf{P}(A) > 0$).

Since

$$\mathsf{P}(B|A) = \frac{\mathsf{P}(AB)}{\mathsf{P}(A)},$$

it follows from (10) that

$$\mathsf{P}(AB) = \mathsf{P}(A)\mathsf{P}(B). \tag{11}$$

In exactly the same way, if $\mathsf{P}(B) > 0$ it is natural to say that "A is independent of B" if

$$\mathsf{P}(A|B) = \mathsf{P}(A).$$

Hence we again obtain (11), which is symmetric in A and B and still makes sense when the probabilities of these events are zero.

After these preliminaries, we introduce the following definition.

Definition 2. Events A and B are called *independent* or *statistically independent* (with respect to the probability P) if

$$\mathsf{P}(AB) = \mathsf{P}(A)\mathsf{P}(B).$$

In probability theory it is often convenient to consider not only independence of events (or sets) but also independence of collections of events (or sets).

Accordingly, we introduce the following definition.

Definition 3. Two algebras $\mathscr{A}_1$ and $\mathscr{A}_2$ of events (or sets) are called *independent* or *statistically independent* (with respect to the probability P) if all pairs of sets A_1 and A_2, belonging respectively to $\mathscr{A}_1$ and $\mathscr{A}_2$, are independent.

For example, let us consider the two algebras

$$\mathscr{A}_1 = \{A_1, \bar{A}_1, \varnothing, \Omega\} \quad \text{and} \quad \mathscr{A}_2 = \{A_2, \bar{A}_2, \varnothing, \Omega\},$$

where A_1 and A_2 are subsets of Ω. It is easy to verify that $\mathscr{A}_1$ and $\mathscr{A}_2$ are independent if and only if A_1 and A_2 are independent. In fact, the independence of $\mathscr{A}_1$ and $\mathscr{A}_2$ means the independence of the 16 events A_1 and A_2, A_1 and $\bar{A}_2, \dots, \Omega$ and Ω. Consequently A_1 and A_2 are independent. Conversely, if A_1 and A_2 are independent, we have to show that the other 15 pairs of events are independent. Let us verify, for example, the independence of A_1 and $\bar{A}_2$. We have

$$\begin{aligned} P(A_1\bar{A}_2) &= P(A_1) - P(A_1A_2) = P(A_1) - P(A_1)P(A_2) \\ &= P(A_1)\cdot(1 - P(A_2)) = P(A_1)P(\bar{A}_2). \end{aligned}$$

The independence of the other pairs is verified similarly.

5. The concept of independence of two sets or two algebras of sets can be extended to any finite number of sets or algebras of sets.

Thus we say that the sets $A_1, \ldots, A_n$ are collectively *independent* or *statistically independent* (with respect to the probability P) if for $k = 1, \ldots, n$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq n$

$$P(A_{i_1} \cdots A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k}). \tag{12}$$

The algebras $\mathscr{A}_1, \ldots, \mathscr{A}_n$ of sets are called *independent* or *statistically independent* (with respect to the probability P) if all sets $A_1, \ldots, A_n$ belonging respectively to $\mathscr{A}_1, \ldots, \mathscr{A}_n$ are independent.

Note that *pairwise independence* of events *does not imply* their independence. In fact if, for example, $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ and all outcomes are equiprobable, it is easily verified that the events

$$A = \{\omega_1, \omega_2\}, \qquad B = \{\omega_1, \omega_3\}, \qquad C = \{\omega_1, \omega_4\}$$

are pairwise independent, whereas

$$P(ABC) = \tfrac{1}{4} \neq (\tfrac{1}{2})^3 = P(A)P(B)P(C).$$

Also note that if

$$P(ABC) = P(A)P(B)P(C)$$

for events A, B and C, it by no means follows that these events are pairwise independent. In fact, let Ω consist of the 36 ordered pairs (i, j), where $i, j = 1, 2, \ldots, 6$ and all the pairs are equiprobable. Then if $A = \{(i, j): j = 1, 2 \text{ or } 5\}$, $B = \{(i, j): j = 4, 5 \text{ or } 6\}$, $C = \{(i, j): i + j = 9\}$ we have

$$\begin{aligned} P(AB) &= \tfrac{1}{6} \neq \tfrac{1}{4} = P(A)P(B), \\ P(AC) &= \tfrac{1}{36} \neq \tfrac{1}{18} = P(A)P(C), \\ P(BC) &= \tfrac{1}{12} \neq \tfrac{1}{18} = P(B)P(C), \end{aligned}$$

but also

$$P(ABC) = \tfrac{1}{36} = P(A)P(B)P(C).$$

6. Let us consider in more detail, from the point of view of independence, the classical model $(\Omega, \mathscr{A}, P)$ that was introduced in §2 and used as a basis for the binomial distribution.

In this model

$$\Omega = \{\omega: \omega = (a_1, \ldots, a_n), a_i = 0, 1\}, \qquad \mathscr{A} = \{A: A \subseteq \Omega\}$$

and

$$p(\omega) = p^{\Sigma a_i} q^{n - \Sigma a_i}. \tag{13}$$

Consider an event $A \subseteq \Omega$. We say that this event depends on a trial at time k if it is determined by the value a_k alone. Examples of such events are

$$A_k = \{\omega: a_k = 1\}, \qquad \bar{A}_k = \{\omega: a_k = 0\}.$$

Let us consider the sequence of algebras $\mathscr{A}_1, \mathscr{A}_2, \ldots, \mathscr{A}_n$, where $\mathscr{A}_k = \{A_k, \bar{A}_k, \varnothing, \Omega\}$ and show that under (13) these algebras are independent.

It is clear that

$$\begin{aligned}\mathsf{P}(A_k) &= \sum_{\{\omega: a_k = 1\}} p(\omega) = \sum_{\{\omega: a_k = 1\}} p^{\Sigma a_i} q^{n - \Sigma a_i} \\ &= p \sum_{(a_1, \ldots, a_{k-1}, a_{k+1}, \ldots, a_n)} p^{a_1 + \cdots + a_{k-1} + a_{k+1} + \cdots + a_n} \\ &\quad \times q^{(n-1) - (a_1 + \cdots + a_{k-1} + a_{k+1} + \cdots + a_n)} = p \sum_{l=0}^{n-1} C_{n-1}^l p^l q^{(n-1)-l} = p,\end{aligned}$$

and a similar calculation shows that $\mathsf{P}(\bar{A}_k) = q$ and that, for $k \neq l$,

$$\mathsf{P}(A_k A_l) = p^2, \qquad \mathsf{P}(A_k \bar{A}_l) = pq, \qquad \mathsf{P}(\bar{A}_k A_l) = q^2.$$

It is easy to deduce from this that $\mathscr{A}_k$ and $\mathscr{A}_l$ are independent for $k \neq l$.

It can be shown in the same way that $\mathscr{A}_1, \mathscr{A}_2, \ldots, \mathscr{A}_n$ are independent. This is the basis for saying that our model $(\Omega, \mathscr{A}, \mathsf{P})$ corresponds to "n independent trials with two outcomes and probability p of success." James Bernoulli was the first to study this model systematically, and established the law of large numbers (§5) for it. Accordingly, this model is also called the Bernoulli scheme with two outcomes (success and failure) and probability p of success.

A detailed study of the probability space for the Bernoulli scheme shows that it has the structure of a direct product of probability spaces, defined as follows.

Suppose that we are given a collection $(\Omega_1, \mathscr{B}_1, \mathsf{P}_1), \ldots, (\Omega_n, \mathscr{B}_n, \mathsf{P}_n)$ of finite probability spaces. Form the space $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$ of points $\omega = (a_1, \ldots, a_n)$, where $a_i \in \Omega_i$. Let $\mathscr{A} = \mathscr{B}_1 \otimes \cdots \otimes \mathscr{B}_n$ be the algebra of the subsets of Ω that consists of sums of sets of the form

$$A = B_1 \times B_2 \times \cdots \times B_n$$

with $B_i \in \mathscr{B}_i$. Finally, for $\omega = (a_1, \ldots, a_n)$ take $p(\omega) = p_1(a_1) \cdots p_n(a_n)$ and define $\mathsf{P}(A)$ for the set $A = B_1 \times B_2 \times \cdots \times B_n$ by

$$\mathsf{P}(A) = \sum_{\{a_1 \in B_1, \ldots, a_n \in B_n\}} p_1(a_1) \cdots p_n(a_n).$$

It is easy to verify that $\mathsf{P}(\Omega) = 1$ and therefore the triple $(\Omega, \mathscr{A}, \mathsf{P})$ defines a probability space. This space is called the *direct product of the probability spaces* $(\Omega_1, \mathscr{B}_1, \mathsf{P}_1), \ldots, (\Omega_n, \mathscr{B}_n, \mathsf{P}_n)$.

We note an easily verified property of the direct product of probability spaces: with respect to P, the events

$$A_1 = \{\omega: a_1 \in B_1\}, \ldots, A_n = \{\omega: a_n \in B_n\},$$

where $B_i \in \mathscr{B}_i$, are independent. In the same way, the algebras of subsets of Ω,

$$\mathscr{A}_1 = \{A_1: A_1 = \{\omega: a_1 \in B_1\}, B_1 \in \mathscr{B}_1\},$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\mathscr{A}_n = \{A_n: A_n = \{\omega: a_n \in B_n\}, B_n \in \mathscr{B}_n\}$$

are independent.

It is clear from our construction that the Bernoulli scheme

$$(\Omega, \mathscr{A}, \mathsf{P}) \quad \text{with} \quad \Omega = \{\omega: \omega = (a_1, \ldots, a_n), a_i = 0 \text{ or } 1\}$$
$$\mathscr{A} = \{A: A \subseteq \Omega\} \quad \text{and} \quad p(\omega) = p^{\Sigma a_i} q^{n - \Sigma a_i}$$

can be thought of as the direct product of the probability spaces $(\Omega_i, \mathscr{B}_i, \mathsf{P}_i)$, $i = 1, 2, \ldots, n$, where

$$\Omega_i = \{0, 1\}, \qquad \mathscr{B}_i = \{\{0\}, \{1\}, \varnothing, \Omega_i\},$$
$$\mathsf{P}_i(\{1\}) = p, \qquad \mathsf{P}_i(\{0\}) = q.$$

7. Problems

1. Give examples to show that in general the equations

$$\mathsf{P}(B \mid A) + \mathsf{P}(B \mid \bar{A}) = 1,$$
$$\mathsf{P}(B \mid A) + \mathsf{P}(\bar{B} \mid \bar{A}) = 1$$

are false.

2. An urn contains M balls, of which M_1 are white. Consider a sample of size n. Let B_j be the event that the ball selected at the jth step is white, and A_k the event that a sample of size n contains exactly k white balls. Show that

$$\mathsf{P}(B_j \mid A_k) = k/n$$

both for sampling with replacement and for sampling without replacement.

3. Let $A_1, \ldots, A_n$ be independent events. Then

$$\mathsf{P}\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \prod_{i=1}^{n} \mathsf{P}(\bar{A}_i).$$

4. Let $A_1, \ldots, A_n$ be independent events with $\mathsf{P}(A_i) = p_i$. Then the probability P_0 that neither event occurs is

$$P_0 = \prod_{i=1}^{n} (1 - p_i).$$

5. Let A and B be independent events. In terms of $\mathsf{P}(A)$ and $\mathsf{P}(B)$, find the probabilities of the events that exactly k, at least k, and at most k of A and B occur ($k = 0, 1, 2$).

6. Let event A be independent of itself, i.e. let A and A be independent. Show that $\mathsf{P}(A)$ is either 0 or 1.

7. Let event A have $\mathsf{P}(A) = 0$ or 1. Show that A and an arbitrary event B are independent.

8. Consider the electric circuit shown in Figure 4:

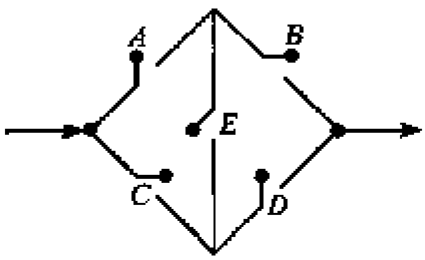


Figure 4

Each of the switches A, B, C, D, and E is independently open or closed with probabilities p and q, respectively. Find the probability that a signal fed in at "input" will be received at "output". If the signal is received, what is the conditional probability that E is open?

§4. Random Variables and Their Properties

1. Let $(\Omega, \mathscr{A}, \mathsf{P})$ be a probabilistic model of an experiment with a finite number of outcomes, $N(\Omega) < \infty$, where $\mathscr{A}$ is the algebra of all subsets of Ω. We observe that in the examples above, where we calculated the probabilities of various events $A \in \mathscr{A}$, the specific nature of the sample space Ω was of no interest. We were interested only in numerical properties depending on the sample points. For example, we were interested in the probability of some number of successes in a series of n trials, in the probability distribution for the number of objects in cells, etc.

The concept "random variable," which we now introduce (later it will be given a more general form) serves to define quantities that are subject to "measurement" in random experiments.

Definition 1. Any numerical function $\xi = \xi(\omega)$ defined on a (finite) sample space Ω is called a (simple) *random variable*. (The reason for the term "simple" random variable will become clear after the introduction of the general concept of random variable in §4 of Chapter II.)

EXAMPLE 1. In the model of two tosses of a coin with sample space $\Omega = \{\text{HH, HT, TH, TT}\}$, define a random variable $\xi = \xi(\omega)$ by the table

ω	HH	HT	TH	TT
$\xi(\omega)$	2	1	1	0

Here, from its very definition, $\xi(\omega)$ is nothing but the number of heads in the outcome ω.

Another extremely simple example of a random variable is the *indicator* (or *characteristic function*) of a set $A \in \mathscr{A}$:

$$\xi = I_A(\omega),$$

where†

$$I_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}$$

When experimenters are concerned with random variables that describe observations, their main interest is in the probabilities with which the random variables take various values. From this point of view they are interested, not in the distribution of the probability P over $(\Omega, \mathscr{A})$, but in its distribution over the range of a random variable. Since we are considering the case when Ω contains only a finite number of points, the range X of the random variable ξ is also finite. Let $X = \{x_1, \ldots, x_m\}$, where the (different) numbers $x_1, \ldots, x_m$ exhaust the values of ξ.

Let $\mathscr{X}$ be the collection of all subsets of X, and let $B \in \mathscr{X}$. We can also interpret B as an event if the sample space is taken to be X, the set of values of ξ.

On $(X, \mathscr{X})$, consider the probability $\mathsf{P}_\xi(\cdot)$ induced by ξ according to the formula

$$P_\xi(B) = \mathsf{P}\{\omega: \xi(\omega) \in B\}, \qquad B \in \mathscr{X}.$$

It is clear that the values of this probability are completely determined by the probabilities

$$P_\xi(x_i) = \mathsf{P}\{\omega: \xi(\omega) = x_i\}, \qquad x_i \in X.$$

The set of numbers $\{P_\xi(x_1), \ldots, P_\xi(x_m)\}$ is called the *probability distribution of the random variable* ξ.

† The notation $I(A)$ is also used. For frequently used properties of indicators see Problem 1.

EXAMPLE 2. A random variable ξ that takes the two values 1 and 0 with probabilities p ("success") and q ("failure"), is called a Bernoulli† random variable. Clearly

$$P_\xi(x) = p^x q^{1-x}, \qquad x = 0, 1. \tag{1}$$

A *binomial* (or binomially distributed) *random variable* ξ is a random variable that takes the $n + 1$ values $0, 1, \dots, n$ with probabilities

$$P_\xi(x) = C_n^x p^x q^{n-x}, \qquad x = 0, 1, \dots, n. \tag{2}$$

Note that here and in many subsequent examples we do not specify the sample spaces $(\Omega, \mathscr{A}, \mathsf{P})$, but are interested only in the values of the random variables and their probability distributions.

The probabilistic structure of the random variables ξ is completely specified by the probability distributions $\{P_\xi(x_i), i = 1, \dots, m\}$. The concept of distribution function, which we now introduce, yields an equivalent description of the probabilistic structure of the random variables.

Definition 2. Let $x \in R^1$. The function

$$F_\xi(x) = \mathsf{P}\{\omega: \xi(\omega) \le x\}$$

is called the *distribution function* of the random variable ξ.

Clearly

$$F_\xi(x) = \sum_{\{i:\, x_i \le x\}} P_\xi(x_i)$$

and

$$P_\xi(x_i) = F_\xi(x_i) - F_\xi(x_i -),$$

where $F_\xi(x-) = \lim_{y \uparrow x} F_\xi(y)$.

If we suppose that $x_1 < x_2 < \cdots < x_m$ and put $F_\xi(x_0) = 0$, then

$$P_\xi(x_i) = F_\xi(x_i) - F_\xi(x_{i-1}), \qquad i = 1, \dots, m.$$

The following diagrams (Figure 5) exhibit $P_\xi(x)$ and $F_\xi(x)$ for a binomial random variable.

It follows immediately from Definition 2 that the distribution $F_\xi = F_\xi(x)$ has the following properties:

(1) $F_\xi(-\infty) = 0$, $F_\xi(+\infty) = 1$;
(2) $F_\xi(x)$ is continuous on the right ($F_\xi(x+) = F_\xi(x)$) and piecewise constant.

† We use the terms "Bernoulli, binomial, Poisson, Gaussian, ..., random variables" for what are more usually called random variables with Bernoulli, binomial, Poisson, Gaussian, ..., distributions.

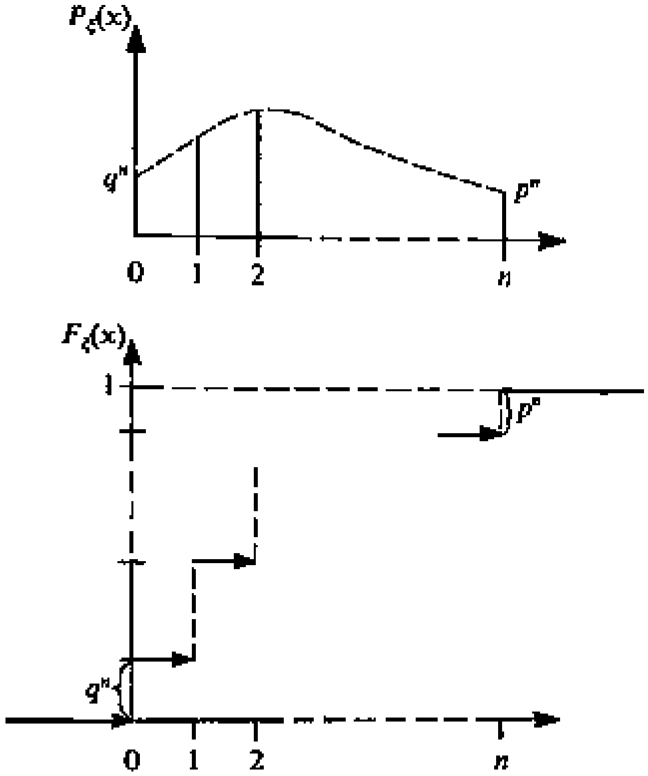


Figure 5

Along with random variables it is often necessary to consider *random vectors* $\xi = (\xi_1, \ldots, \xi_r)$ whose components are random variables. For example, when we considered the multinomial distribution we were dealing with a random vector $v = (v_1, \ldots, v_r)$, where $v_i = v_i(\omega)$ is the number of elements equal to b_i, $i = 1, \ldots, r$, in the sequence $\omega = (a_1, \ldots, a_n)$.

The set of probabilities

$$P_\xi(x_1, \ldots, x_r) = \mathsf{P}\{\omega: \xi_1(\omega) = x_1, \ldots, \xi_r(\omega) = x_r\},$$

where $x_i \in X_i$, the range of ξ_i, is called the *probability distribution of the random vector* ξ, and the function

$$F_\xi(x_1, \ldots, x_r) = \mathsf{P}\{\omega: \xi_1(\omega) \leq x_1, \ldots, \xi_r(\omega) \leq x_r\},$$

where $x_i \in R^1$, is called the *distribution function of the random vector* $\xi = (\xi_1, \ldots, \xi_r)$.

For example, for the random vector $v = (v_1, \ldots, v_r)$ mentioned above,

$$P_v(n_1, \ldots, n_r) = C_n(n_1, \ldots, n_r)p_1^{n_1} \cdots p_r^{n_r}$$

(see (2.2)).

2. Let $\xi_1, \ldots, \xi_r$ be a set of random variables with values in a (finite) set $X \subseteq R^1$. Let $\mathscr{X}$ be the algebra of subsets of X.

Definition 3. The random variables $\xi_1, \ldots, \xi_r$ are said to be *independent* (*collectively independent*) if

$$\mathsf{P}\{\xi_1 = x_1, \ldots, \xi_r = x_r\} = \mathsf{P}\{\xi_1 = x_1\} \cdots \mathsf{P}\{\xi_r = x_r\}$$

for all $x_1, \ldots, x_r \in X$; or, equivalently, if

$$\mathsf{P}\{\xi_1 \in B_1, \ldots, \xi_r \in B_r\} = \mathsf{P}\{\xi_1 \in B_1\} \cdots \mathsf{P}\{\xi_r \in B_r\}$$

for all $B_1, \ldots, B_r \in \mathscr{X}$.

We can get a very simple example of independent random variables from the Bernoulli scheme. Let

$$\Omega = \{\omega : \omega = (a_1, \ldots, a_n), a_i = 0, 1\}, \quad p(\omega) = p^{\Sigma a_i} q^{n - \Sigma a_i}$$

and $\xi_i(\omega) = a_i$ for $\omega = (a_1, \ldots, a_n)$, $i = 1, \ldots, n$. Then the random variables $\xi_1, \xi_2, \ldots, \xi_n$ are independent, as follows from the independence of the events

$$A_1 = \{\omega : a_1 = 1\}, \ldots, A_n = \{\omega : a_n = 1\},$$

which was established in §3.

3. We shall frequently encounter the problem of finding the probability distributions of random variables that are functions $f(\xi_1, \ldots, \xi_r)$ of random variables $\xi_1, \ldots, \xi_r$. For the present we consider only the determination of the distribution of a sum $\zeta = \xi + \eta$ of random variables.

If ξ and η take values in the respective sets $X = \{x_1, \ldots, x_k\}$ and $Y = \{y_1, \ldots, y_l\}$, the random variable $\zeta = \xi + \eta$ takes values in the set $Z = \{z : z = x_i + y_j, i = 1, \ldots, k; j = 1, \ldots, l\}$. Then it is clear that

$$P_\zeta(z) = \mathsf{P}\{\zeta = z\} = \mathsf{P}\{\xi + \eta = z\} = \sum_{\{(i, j) : x_i + y_j = z\}} \mathsf{P}\{\xi = x_i, \eta = y_j\}.$$

The case of independent random variables ξ and η is particularly important. In this case

$$\mathsf{P}\{\xi = x_i, \eta = y_j\} = \mathsf{P}\{\xi = x_i\}\mathsf{P}\{\eta = y_j\},$$

and therefore

$$P_\zeta(z) = \sum_{\{(i, j) : x_i + y_j = z\}} P_\xi(x_i)P_\eta(y_j) = \sum_{i=1}^{k} P_\xi(x_i)P_\eta(z - x_i) \tag{3}$$

for all $z \in Z$, where in the last sum $P_\eta(z - x_i)$ is taken to be zero if $z - x_i \notin Y$.

For example, if ξ and η are independent Bernoulli random variables, taking the values 1 and 0 with respective probabilities p and q, then $Z = \{0, 1, 2\}$ and

$$\begin{aligned} P_\zeta(0) &= P_\xi(0)P_\eta(0) = q^2, \\ P_\zeta(1) &= P_\xi(0)P_\eta(1) + P_\xi(1)P_\eta(0) = 2pq, \\ P_\zeta(2) &= P_\xi(1)P_\eta(1) = p^2. \end{aligned}$$

It is easy to show by induction that if $\xi_1, \xi_2, \ldots, \xi_n$ are independent Bernoulli random variables with $\mathsf{P}\{\xi_i = 1\} = p$, $\mathsf{P}\{\xi_i = 0\} = q$, then the random variable $\zeta = \xi_1 + \cdots + \xi_n$ has the binomial distribution

$$P_\zeta(k) = C_n^k p^k q^{n-k}, \qquad k = 0, 1, \ldots, n. \tag{4}$$

4. We now turn to the important concept of the expectation, or mean value, of a random variable.

Let $(\Omega, \mathscr{A}, \mathsf{P})$ be a (finite) probability space and $\xi = \xi(\omega)$ a random variable with values in the set $X = \{x_1, \ldots, x_k\}$. If we put $A_i = \{\omega : \xi = x_i\}$, $i = 1, \ldots, k$, then ξ can evidently be represented as

$$\xi(\omega) = \sum_{i=1}^{k} x_i I(A_i), \tag{5}$$

where the sets $A_1, \ldots, A_k$ form a decomposition of Ω (i.e., they are pairwise disjoint and their sum is Ω; see Subsection 3 of §1).

Let $p_i = \mathsf{P}\{\xi = x_i\}$. It is intuitively plausible that if we observe the values of the random variable ξ in "n repetitions of identical experiments", the value x_i ought to be encountered about $p_i n$ times, $i = 1, \ldots, k$. Hence the mean value calculated from the results of n experiments is roughly

$$\frac{1}{n}[np_1 x_1 + \cdots + np_k x_k] = \sum_{i=1}^{k} p_i x_i.$$

This discussion provides the motivation for the following definition.

Definition 4. The *expectation*† or *mean value* of the random variable $\xi = \sum_{i=1}^{k} x_i I(A_i)$ is the number

$$\mathsf{E}\xi = \sum_{i=1}^{k} x_i P(A_i). \tag{6}$$

Since $A_i = \{\omega : \xi(\omega) = x_i\}$ and $P_\xi(x_i) = \mathsf{P}(A_i)$, we have

$$\mathsf{E}\xi = \sum_{i=1}^{k} x_i P_\xi(x_i). \tag{7}$$

Recalling the definition of $F_\xi = F_\xi(x)$ and writing

$$\Delta F_\xi(x) = F_\xi(x) - F_\xi(x-),$$

we obtain $P_\xi(x_i) = \Delta F_\xi(x_i)$ and consequently

$$\mathsf{E}\xi = \sum_{i=1}^{k} x_i \Delta F_\xi(x_i). \tag{8}$$

† Also known as mathematical expectation, or expected value, or (especially in physics) expectation value. (Translator)

Before discussing the properties of the expectation, we remark that it is often convenient to use another representation of the random variable ξ, namely

$$\xi(\omega) = \sum_{j=1}^{l} x'_j I(B_j),$$

where $B_1 + \cdots + B_l = \Omega$, but some of the x'_j may be repeated. In this case $\mathsf{E}\xi$ can be calculated from the formula $\sum_{j=1}^{l} x'_j \mathsf{P}(B_j)$, which differs formally from (5) because in (5) the x_i are all different. In fact,

$$\sum_{\{j: x'_j = x_i\}} x'_j \mathsf{P}(B_j) = x_i \sum_{\{j: x'_j = x_i\}} \mathsf{P}(B_j) = x_i \mathsf{P}(A_i)$$

and therefore

$$\sum_{j=1}^{l} x'_j \mathsf{P}(B_j) = \sum_{i=1}^{k} x_i \mathsf{P}(A_i).$$

5. We list the basic properties of the expectation:

(1) *If* $\xi \geq 0$ *then* $\mathsf{E}\xi \geq 0$.
(2) $\mathsf{E}(a\xi + b\eta) = a\mathsf{E}\xi + b\mathsf{E}\eta$, where a and b are constants.
(3) *If* $\xi \geq \eta$ *then* $\mathsf{E}\xi \geq \mathsf{E}\eta$.
(4) $|\mathsf{E}\xi| \leq \mathsf{E}|\xi|$.
(5) If ξ *and* η *are independent, then* $\mathsf{E}\xi\eta = \mathsf{E}\xi \cdot \mathsf{E}\eta$.
(6) $(\mathsf{E}|\xi\eta|)^2 \leq \mathsf{E}\xi^2 \cdot \mathsf{E}\eta^2$ *(Cauchy–Bunyakovskii inequality)*.†
(7) *If* $\xi = I(A)$ *then* $\mathsf{E}\xi = \mathsf{P}(A)$.

Properties (1) and (7) are evident. To prove (2), let

$$\xi = \sum_i x_i I(A_i), \qquad \eta = \sum_j y_j I(B_j).$$

Then

$$\begin{aligned} a\xi + b\eta &= a\sum_{i,j} x_i I(A_i \cap B_j) + b\sum_{i,j} y_j I(A_i \cap B_j) \\ &= \sum_{i,j} (ax_i + by_j) I(A_i \cap B_j) \end{aligned}$$

and

$$\begin{aligned} \mathsf{E}(a\xi + b\eta) &= \sum_{i,j} (ax_i + by_j) \mathsf{P}(A_i \cap B_j) \\ &= \sum_i ax_i \mathsf{P}(A_i) + \sum_j by_j \mathsf{P}(B_j) \\ &= a\sum_i x_i \mathsf{P}(A_i) + b\sum_j y_j \mathsf{P}(B_j) = a\mathsf{E}\xi + b\mathsf{E}\eta. \end{aligned}$$

† Also known as the Cauchy–Schwarz or Schwarz inequality. (Translator)

Property (3) follows from (1) and (2). Property (4) is evident, since

$$|\mathsf{E}\xi| = \left|\sum_i x_i \mathsf{P}(A_i)\right| \le \sum_i |x_i| \mathsf{P}(A_i) = \mathsf{E}|\xi|.$$

To prove (5) we note that

$$\begin{aligned}\mathsf{E}\xi\eta &= \mathsf{E}\left(\sum_i x_i I(A_i)\right)\left(\sum_j y_j I(B_j)\right)\\ &= \mathsf{E}\sum_{i,j} x_i y_j I(A_i \cap B_j) = \sum_{i,j} x_i y_j \mathsf{P}(A_i \cap B_j)\\ &= \sum_{i,j} x_i y_j \mathsf{P}(A_i)\mathsf{P}(B_j)\\ &= \left(\sum_i x_i \mathsf{P}(A_i)\right)\cdot\left(\sum_j y_j \mathsf{P}(B_j)\right) = \mathsf{E}\xi\cdot\mathsf{E}\eta,\end{aligned}$$

where we have used the property that for independent random variables the events

$$A_i = \{\omega\colon \xi(\omega) = x_i\} \quad \text{and} \quad B_j = \{\omega\colon \eta(\omega) = y_j\}$$

are independent: $P(A_i \cap B_j) = \mathsf{P}(A_i)\mathsf{P}(B_j)$.

To prove property (6) we observe that

$$\xi^2 = \sum_i x_i^2 I(A_i), \qquad \eta^2 = \sum_j y_j^2 I(B_j)$$

and

$$\mathsf{E}\xi^2 = \sum_i x_i^2 \mathsf{P}(A_i), \qquad \mathsf{E}\eta^2 = \sum_j y_j^2 \mathsf{P}(B_j).$$

Let $\mathsf{E}\xi^2 > 0$, $\mathsf{E}\eta^2 > 0$. Put

$$\tilde{\xi} = \frac{\xi}{\sqrt{\mathsf{E}\xi^2}}, \qquad \tilde{\eta} = \frac{\eta}{\sqrt{\mathsf{E}\eta^2}}.$$

Since $2|\tilde{\xi}\tilde{\eta}| \le \tilde{\xi}^2 + \tilde{\eta}^2$, we have $2\mathsf{E}|\tilde{\xi}\tilde{\eta}| \le \mathsf{E}\tilde{\xi}^2 + \mathsf{E}\tilde{\eta}^2 = 2$. Therefore $\mathsf{E}|\tilde{\xi}\tilde{\eta}| \le 1$ and $(\mathsf{E}|\xi\eta|)^2 \le \mathsf{E}\xi^2 \cdot \mathsf{E}\eta^2$.

However, if, say, $\mathsf{E}\xi^2 = 0$, this means that $\sum_i x_i^2 \mathsf{P}(A_i) = 0$ and consequently the mean value of ξ is 0, and $\mathsf{P}\{\omega\colon \xi(\omega) = 0\} = 1$. Therefore if at least one of $\mathsf{E}\xi^2$ or $\mathsf{E}\eta^2$ is zero, it is evident that $\mathsf{E}|\xi\eta| = 0$ and consequently the Cauchy–Bunyakovskii inequality still holds.

Remark. Property (5) generalizes in an obvious way to any finite number of random variables: if $\xi_1, \ldots, \xi_r$ are independent, then

$$\mathsf{E}\xi_1 \cdots \xi_r = \mathsf{E}\xi_1 \cdots \mathsf{E}\xi_r.$$

The proof can be given in the same way as for the case $r = 2$, or by induction.

EXAMPLE 3. Let ξ be a Bernoulli random variable, taking the values 1 and 0 with probabilities p and q. Then

$$\mathsf{E}\xi = 1 \cdot \mathsf{P}\{\xi = 1\} + 0 \cdot \mathsf{P}\{\xi = 0\} = p.$$

EXAMPLE 4. Let $\xi_1, \dots, \xi_n$ be n Bernoulli random variables with $\mathsf{P}\{\xi_i = 1\} = p$, $\mathsf{P}\{\xi_i = 0\} = q$, $p + q = 1$. Then if

$$S_n = \xi_1 + \cdots + \xi_n$$

we find that

$$\mathsf{E}S_n = np.$$

This result can be obtained in a different way. It is easy to see that $\mathsf{E}S_n$ is not changed if we assume that the Bernoulli random variables $\xi_1, \dots, \xi_n$ are independent. With this assumption, we have according to (4)

$$\mathsf{P}(S_n = k) = C_n^k p^k q^{n-k}, \qquad k = 0, 1, \dots, n.$$

Therefore

$$\begin{aligned}
\mathsf{E}S_n &= \sum_{k=0}^n k\mathsf{P}(S_n = k) = \sum_{k=0}^n kC_n^k p^k q^{n-k} \\
&= \sum_{k=0}^n k \cdot \frac{n!}{k!\,(n-k)!} p^k q^{n-k} \\
&= np \sum_{k=1}^n \frac{(n-1)!}{(k-1)!\,((n-1)-(k-1))!} p^{k-1} q^{(n-1)-(k-1)} \\
&= np \sum_{l=0}^n \frac{(n-1)!}{l!((n-1)-l)!} p^l q^{(n-1)-l} = np.
\end{aligned}$$

However, the first method is more direct.

6. Let $\xi = \sum_i x_i I(A_i)$, where $A_i = \{\omega : \xi(\omega) = x_i\}$, and $\varphi = \varphi(\xi(\omega))$ is a function of $\xi(\omega)$. If $B_j = \{\omega : \varphi(\xi(\omega)) = y_j\}$, then

$$\varphi(\xi(\omega)) = \sum_j y_j I(B_j),$$

and consequently

$$\mathsf{E}\varphi = \sum_j y_j \mathsf{P}(B_j) = \sum_j y_j \mathsf{P}_\varphi(y_j). \tag{9}$$

But it is also clear that

$$\varphi(\xi(\omega)) = \sum_i \varphi(x_i) I(A_i).$$

Hence, as in (9), the expectation of the random variable $\varphi = \varphi(\xi)$ can be calculated as

$$\mathsf{E}\varphi(\xi) = \sum_i \varphi(x_i)P_\xi(x_i).$$

7. The important notion of the variance of a random variable ξ indicates the amount of scatter of the values of ξ around $\mathsf{E}\xi$.

Definition 5. The *variance* (also called the *dispersion*) of the random variable ξ (denoted by $\mathsf{V}\xi$) is

$$\mathsf{V}\xi = \mathsf{E}(\xi - \mathsf{E}\xi)^2.$$

The number $\sigma = +\sqrt{\mathsf{V}\xi}$ is called the standard deviation.

Since

$$\mathsf{E}(\xi - \mathsf{E}\xi)^2 = \mathsf{E}(\xi^2 - 2\xi\cdot\mathsf{E}\xi + (\mathsf{E}\xi)^2) = \mathsf{E}\xi^2 - (\mathsf{E}\xi)^2,$$

we have

$$\mathsf{V}\xi = \mathsf{E}\xi^2 - (\mathsf{E}\xi)^2.$$

Clearly $\mathsf{V}\xi \ge 0$. It follows from the definition that

$$\mathsf{V}(a + b\xi) = b^2\mathsf{V}\xi, \quad \text{where } a \text{ and } b \text{ are constants.}$$

In particular, $\mathsf{V}a = 0$, $\mathsf{V}(b\xi) = b^2\mathsf{V}\xi$.

Let ξ and η be random variables. Then

$$\begin{aligned}\mathsf{V}(\xi + \eta) &= \mathsf{E}((\xi - \mathsf{E}\xi) + (\eta - \mathsf{E}\eta))^2 \\ &= \mathsf{V}\xi + \mathsf{V}\eta + 2\mathsf{E}(\xi - \mathsf{E}\xi)(\eta - \mathsf{E}\eta).\end{aligned}$$

Write

$$\operatorname{cov}(\xi, \eta) = \mathsf{E}(\xi - \mathsf{E}\xi)(\eta - \mathsf{E}\eta).$$

This number is called the *covariance* of ξ and η. If $\mathsf{V}\xi > 0$ and $\mathsf{V}\eta > 0$, then

$$\rho(\xi, \eta) = \frac{\operatorname{cov}(\xi, \eta)}{\sqrt{\mathsf{V}\xi\cdot\mathsf{V}\eta}}$$

is called the *correlation coefficient* of ξ and η. It is easy to show (see Problem 7 below) that if $\rho(\xi, \eta) = \pm 1$, then ξ and η are linearly dependent:

$$\eta = a\xi + b,$$

with $a > 0$ if $\rho(\xi, \eta) = 1$ and $a < 0$ if $\rho(\xi, \eta) = -1$.

We observe immediately that if ξ and η are independent, so are $\xi - \mathsf{E}\xi$ and $\eta - \mathsf{E}\eta$. Consequently by Property (5) of expectations,

$$\operatorname{cov}(\xi, \eta) = \mathsf{E}(\xi - \mathsf{E}\xi)\cdot\mathsf{E}(\eta - \mathsf{E}\eta) = 0.$$

Using the notation that we introduced for covariance, we have

$$\mathsf{V}(\xi + \eta) = \mathsf{V}\xi + \mathsf{V}\eta + 2\operatorname{cov}(\xi, \eta); \tag{10}$$

if ξ and η are independent, the variance of the sum $\xi + \eta$ is equal to the sum of the variances,

$$\mathsf{V}(\xi + \eta) = \mathsf{V}\xi + \mathsf{V}\eta. \tag{11}$$

It follows from (10) that (11) is still valid under weaker hypotheses than the independence of ξ and η. In fact, it is enough to suppose that ξ and η are uncorrelated, i.e. $\operatorname{cov}(\xi, \eta) = 0$.

Remark. If ξ and η are uncorrelated, it does not follow in general that they are independent. Here is a simple example. Let the random variable α take the values 0, $\pi/2$ and π with probability $\frac{1}{3}$. Then $\xi = \sin \alpha$ and $\eta = \cos \alpha$ are uncorrelated; however, they are not only stochastically dependent (i.e., not independent with respect to the probability P):

$$\mathsf{P}\{\xi = 1, \eta = 1\} = 0 \neq \tfrac{1}{9} = \mathsf{P}\{\xi = 1\}\mathsf{P}\{\eta = 1\},$$

but even functionally dependent: $\xi^2 + \eta^2 = 1$.

Properties (10) and (11) can be extended in the obvious way to any number of random variables:

$$\mathsf{V}\left(\sum_{i=1}^{n} \xi_i\right) = \sum_{i=1}^{n} \mathsf{V}\xi_i + 2\sum_{i>j} \operatorname{cov}(\xi_i, \xi_j). \tag{12}$$

In particular, if $\xi_1, \ldots, \xi_n$ are pairwise independent (pairwise uncorrelated is sufficient), then

$$\mathsf{V}\left(\sum_{i=1}^{n} \xi_i\right) = \sum_{i=1}^{n} \mathsf{V}\xi_i. \tag{13}$$

EXAMPLE 5. If ξ is a Bernoulli random variable, taking the values 1 and 0 with probabilities p and q, then

$$\mathsf{V}\xi = \mathsf{E}(\xi - \mathsf{E}\xi)^2 = (\xi - p)^2 = (1 - p)^2 p + p^2 q = pq.$$

It follows that if $\xi_1, \ldots, \xi_n$ are independent identically distributed Bernoulli random variables, and $S_n = \xi_1 + \cdots + \xi_n$, then

$$\mathsf{V}S_n = npq. \tag{14}$$

8. Consider two random variables ξ and η. Suppose that only ξ can be observed. If ξ and η are correlated, we may expect that knowing the value of ξ allows us to make some inference about the values of the unobserved variable η.

Any function $f = f(\xi)$ of ξ is called an *estimator* for η. We say that an estimator $f^* = f^*(\xi)$ is *best in the mean-square sense* if

$$\mathsf{E}(\eta - f^*(\xi))^2 = \inf_f \mathsf{E}(\eta - f(\xi))^2.$$

Let us show how to find a best estimator in the class of linear estimators $\lambda(\xi) = a + b\xi$. We consider the function $g(a, b) = \mathsf{E}(\eta - (a + b\xi))^2$. Differentiating $g(a, b)$ with respect to a and b, we obtain

$$\frac{\partial g(a, b)}{\partial a} = -2\mathsf{E}[\eta - (a + b\xi)],$$

$$\frac{\partial g(a, b)}{\partial b} = -2\mathsf{E}[(\eta - (a + b\xi))\xi],$$

whence, setting the derivatives equal to zero, we find that the best mean-square linear estimator is $\lambda^*(\xi) = a^* + b^*\xi$, where

$$a^* = \mathsf{E}\eta - b^*\mathsf{E}\xi, \qquad b^* = \frac{\operatorname{cov}(\xi, \eta)}{\mathsf{V}\xi}. \tag{15}$$

In other words,

$$\lambda^*(\xi) = \mathsf{E}\eta + \frac{\operatorname{cov}(\xi, \eta)}{\mathsf{V}\xi}(\xi - \mathsf{E}\xi). \tag{16}$$

The number $\mathsf{E}(\eta - \lambda^*(\xi))^2$ is called the *mean-square error of observation*. An easy calculation shows that it is equal to

$$\Delta^* = \mathsf{E}(\eta - \lambda^*(\xi))^2 = \mathsf{V}\eta - \frac{\operatorname{cov}^2(\xi, \eta)}{\mathsf{V}\xi} = \mathsf{V}\eta[1 - \rho^2(\xi, \eta)]. \tag{17}$$

Consequently, the larger (in absolute value) the correlation coefficient $\rho(\xi, \eta)$ between ξ and η, the smaller the mean-square error of observation Δ^*. In particular, if $|\rho(\xi, \eta)| = 1$ then $\Delta^* = 0$ (cf. Problem 7). On the other hand, if ξ and η are uncorrelated ($\rho(\xi, \eta) = 0$), then $\lambda^*(\xi) = \mathsf{E}\eta$, i.e. in the absence of correlation between ξ and η the best estimate of η in terms of ξ is simply $\mathsf{E}\eta$ (cf. Problem 4).

9. Problems

1. Verify the following properties of indicators $I_A = I_A(\omega)$:

$$I_\varnothing = 0, \qquad I_\Omega = 1, \qquad I_A + I_{\bar{A}} = 1,$$

$$I_{AB} = I_A \cdot I_B,$$

$$I_{A \cup B} = I_A + I_B - I_{AB}.$$

The indicator of $\bigcup_{i=1}^n A_i$ is $1 - \prod_{i=1}^n (1 - I_{A_i})$, the indicator of $\overline{\bigcup_{i=1}^n A_i}$ is $\prod_{i=1}^n (1 - I_{A_i})$, and the indicator of $\sum_{i=1}^n A_i$ is $\sum_{i=1}^n I_{A_i}$.

$$I_{A \triangle B} = (I_A - I_B)^2,$$

where $A \triangle B$ is the *symmetric difference* of A and B, i.e. the set $(A \backslash B) \cup (B \backslash A)$.

2. Let $\xi_1, \ldots, \xi_n$ be independent random variables and

$$\xi_{\min} = \min(\xi_1, \ldots, \xi_n), \qquad \xi_{\max} = \max(\xi_1, \ldots, \xi_n).$$

Show that

$$\mathsf{P}\{\xi_{\min} \geq x\} = \prod_{i=1}^{n} \mathsf{P}\{\xi_i \geq x\},$$

$$\mathsf{P}\{\xi_{\max} < x\} = \prod_{i=1}^{n} \mathsf{P}\{\xi_i < x\}.$$

3. Let $\xi_1, \ldots, \xi_n$ be independent Bernoulli random variables such that

$$\mathsf{P}\{\xi_i = 0\} = 1 - \lambda_i \Delta,$$

$$\mathsf{P}\{\xi_i = 1\} = \lambda_i \Delta,$$

where Δ is a small number, $\Delta > 0$, $\lambda_i > 0$.

Show that

$$\mathsf{P}\{\xi_1 + \cdots + \xi_n = 1\} = \left(\sum_{i=1}^{n} \lambda_i\right)\Delta + O(\Delta^2),$$

$$\mathsf{P}\{\xi_1 + \cdots + \xi_n > 1\} = O(\Delta^2).$$

4. Show that $\inf_{-\infty<a<\infty} \mathsf{E}(\xi - a)^2$ is attained for $a = \mathsf{E}\xi$ and consequently

$$\inf_{-\infty<a<\infty} \mathsf{E}(\xi - a)^2 = \mathsf{V}\xi.$$

5. Let ξ be a random variable with distribution function $F_\xi(x)$ and let m_e be a median of $F_\xi(x)$, i.e. a point such that

$$F_\xi(m_e-) \leq \tfrac{1}{2} \leq F_\xi(m_e).$$

Show that

$$\inf_{-\infty<a<\infty} \mathsf{E}|\xi - a| = \mathsf{E}|\xi - m_e|.$$

6. Let $P_\xi(x) = \mathsf{P}\{\xi = x\}$ and $F_\xi(x) = \mathsf{P}(\xi \leq x\}$. Show that

$$P_{a\xi+b}(x) = P_\xi\left(\frac{x-b}{a}\right),$$

$$F_{a\xi+b}(x) = F_\xi\left(\frac{x-b}{a}\right)$$

for $a > 0$ and $-\infty < b < \infty$. If $y \geq 0$, then

$$F_{\xi^2}(y) = F_\xi(+\sqrt{y}) - F_\xi(-\sqrt{y}) + P_\xi(-\sqrt{y}).$$

Let $\xi^+ = \max(\xi, 0)$. Then

$$F_{\xi^+}(x) = \begin{cases} 0, & x < 0, \\ F_\xi(0), & x = 0, \\ F_\xi(x), & x > 0. \end{cases}$$

7. Let ξ and η be random variables with $V\xi > 0$, $V\eta > 0$, and let $\rho = \rho(\xi, \eta)$ be their correlation coefficient. Show that $|\rho| \le 1$. If $|\rho| = 1$, find constants a and b such that $\eta = a\xi + b$. Moreover, if $\rho = 1$, then

$$\frac{\eta - E\eta}{\sqrt{V\eta}} = \frac{\xi - E\xi}{\sqrt{V\xi}}$$

(and therefore $a > 0$), whereas if $\rho = -1$, then

$$\frac{\eta - E\eta}{\sqrt{V\eta}} = -\frac{\xi - E\xi}{\sqrt{V\xi}}$$

(and therefore $a < 0$).

8. Let ξ and η be random variables with $E\xi = E\eta = 0$, $V\xi = V\eta = 1$ and correlation coefficient $\rho = \rho(\xi, \eta)$. Show that

$$E\max(\xi^2, \eta^2) \le 1 + \sqrt{1 - \rho^2}.$$

9. Use the equation

$$\left(\text{Indicator of } \overline{\bigcup_{i=1}^{n} A_i}\right) = \prod_{i=1}^{n}(1 - I_{A_i}),$$

to deduce the formula $P(B_0) = 1 - S_1 + S_2 + \cdots \pm S_n$ from Problem 4 of §1.

10. Let $\xi_1, \ldots, \xi_n$ be independent random variables, $\varphi_1 = \varphi_1(\xi_1, \ldots, \xi_k)$ and $\varphi_2 = \varphi_2(\xi_{k+1}, \ldots, \xi_n)$, functions respectively of $\xi_1, \ldots, \xi_k$ and $\xi_{k+1}, \ldots, \xi_n$. Show that the random variables φ_1 and φ_2 are independent.

11. Show that the random variables $\xi_1, \ldots, \xi_n$ are independent if and only if

$$F_{\xi_1, \ldots, \xi_n}(x_1, \ldots, x_n) = F_{\xi_1}(x_1) \cdots F_{\xi_n}(x_n)$$

for all $x_1, \ldots, x_n$, where $F_{\xi_1, \ldots, \xi_n}(x_1, \ldots, x_n) = P\{\xi_1 \le x_1, \ldots, \xi_n \le x_n\}$.

12. Show that the random variable ξ is independent of itself (i.e., ξ and ξ are independent) if and only if $\xi = \text{const}$.

13. Under what hypotheses on ξ are the random variables ξ and $\sin \xi$ independent?

14. Let ξ and η be independent random variables and $\eta \ne 0$. Express the probabilities of the events $P\{\xi\eta \le z\}$ and $P\{\xi/\eta \le z\}$ in terms of the probabilities $P_\xi(x)$ and $P_\eta(y)$.

§5. The Bernoulli Scheme. I. The Law of Large Numbers

1. In accordance with the definitions given above, a triple

$$(\Omega, \mathscr{A}, P) \quad \text{with} \quad \Omega = \{\omega : \omega = (a_1, \ldots, a_n), a_i = 0, 1\},$$
$$\mathscr{A} = \{A : A \subseteq \Omega\}, \qquad p(\omega) = p^{\sum a_i} q^{n - \sum a_i}$$

is called a probabilistic model of n independent experiments with two outcomes, or a Bernoulli scheme.

In this and the next section we study some limiting properties (in a sense described below) for Bernoulli schemes. These are best expressed in terms of random variables and of the probabilities of events connected with them.

We introduce random variables $\xi_1, \ldots, \xi_n$ by taking $\xi_i(\omega) = a_i$, $i = 1, \ldots, n$, where $\omega = (a_1, \ldots, a_n)$. As we saw above, the Bernoulli variables $\xi_i(\omega)$ are independent and identically distributed:

$$\mathsf{P}\{\xi_i = 1\} = p, \qquad \mathsf{P}\{\xi_i = 0\} = q, \qquad i = 1, \ldots, n.$$

It is natural to think of ξ_i as describing the result of an experiment at the ith stage (or at time i).

Let us put $S_0(\omega) \equiv 0$ and

$$S_k = \xi_1 + \cdots + \xi_k, \qquad k = 1, \ldots, n.$$

As we found above, $\mathsf{E}S_n = np$ and consequently

$$\mathsf{E}\frac{S_n}{n} = p. \tag{1}$$

In other words, the mean value of the frequency of "success", i.e. S_n/n, coincides with the probability p of success. Hence we are led to ask how much the frequency S_n/n of success differs from its probability p.

We first note that we cannot expect that, for a sufficiently small $\varepsilon > 0$ and for sufficiently large n, the deviation of S_n/n from p is less than ε for all ω, i.e. that

$$\left|\frac{S_n(\omega)}{n} - p\right| \leq \varepsilon, \qquad \omega \in \Omega. \tag{2}$$

In fact, when $0 < p < 1$,

$$\mathsf{P}\left\{\frac{S_n}{n} = 1\right\} = \mathsf{P}\{\xi_1 = 1, \ldots, \xi_n = 1\} = p^n,$$
$$\mathsf{P}\left\{\frac{S_n}{n} = 0\right\} = \mathsf{P}\{\xi_1 = 0, \ldots, \xi_n = 0\} = q^n,$$

whence it follows that (2) is not satisfied for sufficiently small $\varepsilon > 0$.

We observe, however, that when n is large the probabilities of the events $\{S_n/n = 1\}$ and $\{S_n/n = 0\}$ are small. It is therefore natural to expect that the total probability of the events for which $|[S_n(\omega)/n] - p| > \varepsilon$ will also be small when n is sufficiently large.

We shall accordingly try to estimate the probability of the event $\{\omega: |[S_n(\omega)/n] - p| > \varepsilon\}$. For this purpose we need the following inequality, which was discovered by Chebyshev.

Chebyshev's inequality. *Let* $(\Omega, \mathscr{A}, \mathsf{P})$ *be a probability space and* $\xi = \xi(\omega)$ *a nonnegative random variable. Then*

$$\mathsf{P}\{\xi \geq \varepsilon\} \leq \mathsf{E}\xi/\varepsilon \tag{3}$$

for all $\varepsilon > 0$.

PROOF. We notice that

$$\xi = \xi I(\xi \geq \varepsilon) + \xi I(\xi < \varepsilon) \geq \xi I(\xi \geq \varepsilon) \geq \varepsilon I(\xi \geq \varepsilon),$$

where $I(A)$ is the indicator of A.

Then, by the properties of the expectation,

$$\mathsf{E}\xi \geq \varepsilon \mathsf{E} I(\xi \geq \varepsilon) = \varepsilon\, \mathsf{P}(\xi \geq \varepsilon),$$

which establishes (3).

Corollary. *If* ξ *is any random variable, we have for* $\varepsilon > 0$,

$$\begin{aligned} &\mathsf{P}\{|\xi| \geq \varepsilon\} \leq \mathsf{E}|\xi|/\varepsilon, \\ &\mathsf{P}\{|\xi| \geq \varepsilon\} = \mathsf{P}\{\xi^2 \geq \varepsilon^2\} \leq \mathsf{E}\xi^2/\varepsilon^2, \\ &\mathsf{P}\{|\xi - \mathsf{E}\xi| \geq \varepsilon\} \leq \mathsf{V}\xi/\varepsilon^2. \end{aligned} \tag{4}$$

In the last of these inequalities, take $\xi = S_n/n$. Then using (4.14), we obtain

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right\} \leq \frac{\mathsf{V}(S_n/n)}{\varepsilon^2} = \frac{\mathsf{V}S_n}{n^2\varepsilon^2} = \frac{npq}{n^2\varepsilon^2} = \frac{pq}{n\varepsilon^2}.$$

Therefore

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right\} \leq \frac{pq}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}, \tag{5}$$

from which we see that for large n there is rather small probability that the frequency S_n/n of success deviates from the probability p by more than ε.

For $n \geq 1$ and $0 \leq k \leq n$, write

$$P_n(k) = C_n^k p^k q^{n-k}.$$

Then

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right\} = \sum_{\{k : |(k/n) - p| \geq \varepsilon\}} P_n(k),$$

and we have actually shown that

$$\sum_{\{k : |(k/n) - p| \geq \varepsilon\}} P_n(k) \leq \frac{pq}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}, \tag{6}$$

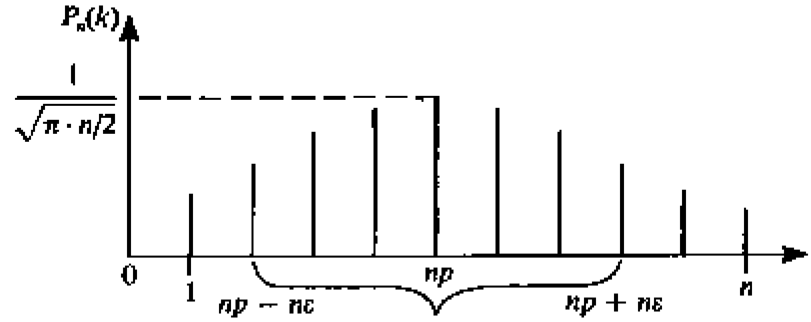


Figure 6

i.e. we have proved an inequality that could also have been obtained analytically, without using the probabilistic interpretation.

It is clear from (6) that

$$\sum_{\{k:|(k/n)-p|\geq\varepsilon\}} P_n(k) \to 0, \qquad n \to \infty. \tag{7}$$

We can clarify this graphically in the following way. Let us represent the binomial distribution $\{P_n(k), 0 \leq k \leq n\}$ as in Figure 6.

Then as n increases the graph spreads out and becomes flatter. At the same time the sum of $P_n(k)$, over k for which $np - n\varepsilon \leq k < np + n\varepsilon$, tends to 1.

Let us think of the sequence of random variables $S_0, S_1, \ldots, S_n$ as the *path* of a wandering particle. Then (7) has the following interpretation.

Let us draw lines from the origin of slopes kp, $k(p + \varepsilon)$, and $k(p - \varepsilon)$. Then on the average the path follows the kp line, and for every $\varepsilon > 0$ we can say that when n is sufficiently large there is a large probability that the point S_n specifying the position of the particle at time n lies in the interval $[n(p - \varepsilon), n(p + \varepsilon)]$; see Figure 7.

We would like to write (7) in the following form:

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right\} \to 0, \qquad n \to \infty, \tag{8}$$

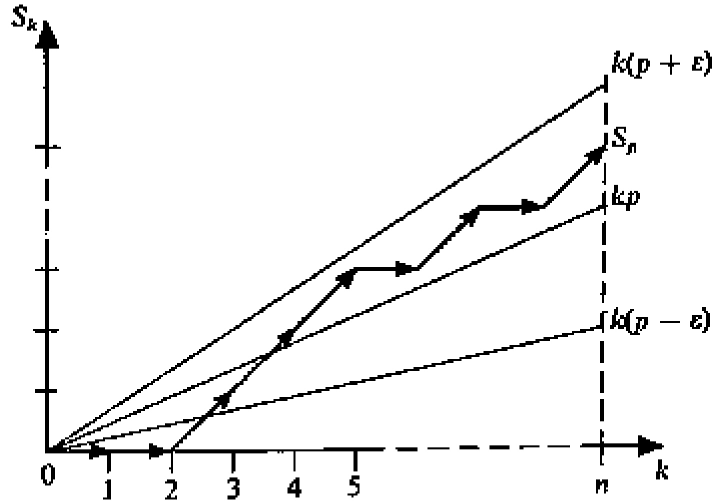


Figure 7

However, we must keep in mind that there is a delicate point involved here. Indeed, the form (8) is really justified only if P is a probability on a space $(\Omega, \mathscr{A})$ on which infinitely many sequences of independent Bernoulli random variables $\xi_1, \xi_2, \dots$, are defined. Such spaces can actually be constructed and (8) can be justified in a completely rigorous probabilistic sense (see Corollary 1 below, the end of §4, Chapter II, and Theorem 1, §9, Chapter II). For the time being, if we want to attach a meaning to the analytic statement (7), using the language of probability theory, we have proved only the following.

Let $(\Omega^{(n)}, \mathscr{A}^{(n)}, \mathsf{P}^{(n)})$, $n \ge 1$, be a sequence of Bernoulli schemes such that

$$\Omega^{(n)} = \{\omega^{(n)}: \omega^{(n)} = (a_1^{(n)}, \dots, a_n^{(n)}), a_i^{(n)} = 0, 1\},$$
$$\mathscr{A}^{(n)} = \{A: A \subseteq \Omega^{(n)}\},$$
$$p^{(n)}(\omega^{(n)}) = p^{\Sigma a_i^{(n)}} q^{n - \Sigma a_i^{(n)}}$$

and

$$S_k^{(n)}(\omega^{(n)}) = \xi_1^{(n)}(\omega^{(n)}) + \cdots + \xi_k^{(n)}(\omega^{(n)}),$$

where, for $n \le 1$, $\xi_1^{(n)}, \dots, \xi_n^{(n)}$ are sequences of independent identically distributed Bernoulli random variables.

Then

$$\mathsf{P}^{(n)}\left\{\omega^{(n)}: \left|\frac{S_n^{(n)}(\omega^{(n)})}{n} - p\right| \ge \varepsilon\right\} = \sum_{\{k: |(k/n) - p| \ge \varepsilon\}} P_n(k) \to 0, \qquad n \to \infty. \quad (9)$$

Statements like (7)–(9) go by the name of **James Bernoulli's law of large numbers.** We may remark that to be precise, Bernoulli's proof consisted in establishing (7), which he did quite rigorously by using estimates for the "tails" of the binomial probabilities $P_n(k)$ (for the values of k for which $|(k/n) - p| \ge \varepsilon$). A direct calculation of the sum of the tail probabilities of the binomial distribution $\sum_{\{k: |(k/n) - p| \ge \varepsilon\}} P_n(k)$ is rather difficult problem for large n, and the resulting formulas are ill adapted for actual estimates of the probability with which the frequencies S_n/n differ from p by less than ε. Important progress resulted from the discovery by De Moivre (for $p = \frac{1}{2}$) and then by Laplace (for $0 < p < 1$) of simple asymptotic formulas for $P_n(k)$, which led not only to new proofs of the law of large numbers but also to more precise statements of both local and integral limit theorems, the essence of which is that for large n and at least for $k \sim np$,

$$P_n(k) \sim \frac{1}{\sqrt{2\pi npq}} e^{-(k-np)^2/(2npq)},$$

and

$$\sum_{\{k: |(k/n) - p| \le \varepsilon\}} P_n(k) \sim \frac{1}{\sqrt{2\pi}} \int_{-\varepsilon\sqrt{n/pq}}^{\varepsilon\sqrt{n/pq}} e^{-x^2/2}\, dx.$$

2. The next section will be devoted to precise statements and proofs of these results. For the present we consider the question of the real meaning of the law of large numbers, and of its empirical interpretation.

Let us carry out a large number, say N, of series of experiments, each of which consists of "n independent trials with probability p of the event C of interest." Let S_n^i/n be the frequency of event C in the ith series and N_ε the number of series in which the frequency deviates from p by less than ε:

N_ε is the number of i's for which $|(S_n^i/n) - p| \le \varepsilon$. Then

$$N_\varepsilon/N \sim P_\varepsilon \tag{10}$$

where $P_\varepsilon = \mathsf{P}\{|(S_n^1/n) - p| \le \varepsilon\}$.

It is important to emphasize that an attempt to make (10) precise inevitably involves the introduction of some probability measure, just as an estimate for the deviation of S_n/n from p becomes possible only after the introduction of a probability measure P.

3. Let us consider the estimate obtained above,

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \ge \varepsilon\right\} = \sum_{\{k:|(k/n)-p|\ge\varepsilon\}} P_n(k) \le \frac{1}{4n\varepsilon^2}, \tag{11}$$

as an answer to the following question that is typical of mathematical statistics: what is the least number n of observations that is guaranteed to have (for arbitrary $0 < p < 1$)

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \le \varepsilon\right\} \ge 1 - \alpha, \tag{12}$$

where α is a given number (usually small)?

It follows from (11) that this number is the smallest integer n for which

$$n \ge \frac{1}{4\varepsilon^2\alpha}. \tag{13}$$

For example, if $\alpha = 0.05$ and $\varepsilon = 0.02$, then 12 500 observations guarantee that (12) will hold independently of the value of the unknown parameter p.

Later (Subsection 5, §6) we shall see that this number is much overstated; this came about because Chebyshev's inequality provides only a very crude upper bound for $\mathsf{P}\{|(S_n/n) - p| \ge \varepsilon\}$.

4. Let us write

$$C(n, \varepsilon) = \left\{\omega : \left|\frac{S_n(\omega)}{n} - p\right| \le \varepsilon\right\}.$$

From the law of large numbers that we proved, it follows that for every $\varepsilon > 0$ and for sufficiently large n, the probability of the set $C(n, \varepsilon)$ is close to 1. In this sense it is natural to call paths (realizations) ω that are in $C(n, \varepsilon)$ *typical* (or (n, ε)-typical).

We ask the following question: How many typical realizations are there, and what is the weight $p(\omega)$ of a typical realization?

For this purpose we first notice that the total number $N(\Omega)$ of points is 2^n, and that if $p = 0$ or 1, the set of typical paths $C(n, \varepsilon)$ contains only the single path $(0, 0, \dots, 0)$ or $(1, 1, \dots, 1)$. However, if $p = \frac{1}{2}$, it is intuitively clear that "almost all" paths (all except those of the form $(0, 0, \dots, 0)$ or $(1, 1, \dots, 1)$) are typical and that consequently there should be about 2^n of them.

It turns out that we can give a definitive answer to the question whenever $0 < p < 1$; it will then appear that both the number of typical realizations and the weights $p(\omega)$ are determined by a function of p called the entropy.

In order to present the corresponding results in more depth, it will be helpful to consider the somewhat more general scheme of Subsection 2 of §2 instead of the Bernoulli scheme itself.

Let $(p_1, p_2, \dots, p_r)$ be a finite probability distribution, i.e. a set of nonnegative numbers satisfying $p_1 + \cdots + p_r = 1$. The *entropy* of this distribution is

$$H = -\sum_{i=1}^{r} p_i \ln p_i, \tag{14}$$

with $0 \cdot \ln 0 = 0$. It is clear that $H \geq 0$, and $H = 0$ if and only if every p_i, with one exception, is zero. The function $f(x) = -x \ln x$, $0 \leq x \leq 1$, is convex upward, so that, as know from the theory of convex functions,

$$\frac{f(x_1) + \cdots + f(x_r)}{r} \leq f\left(\frac{x_1 + \cdots + x_r}{r}\right).$$

Consequently

$$H = -\sum_{i=1}^{r} p_i \ln p_i \leq -r \cdot \frac{p_1 + \cdots + p_r}{r} \cdot \ln\left(\frac{p_1 + \cdots + p_r}{r}\right) = \ln r.$$

In other words, the entropy attains its largest value for $p_1 = \cdots = p_r = 1/r$ (see Figure 8 for $H = H(p)$ in the case $r = 2$).

If we consider the probability distribution $(p_1, p_2, \dots, p_r)$ as giving the probabilities for the occurrence of events $A_1, A_2, \dots, A_r$, say, then it is quite clear that the "degree of indeterminancy" of an event will be different for

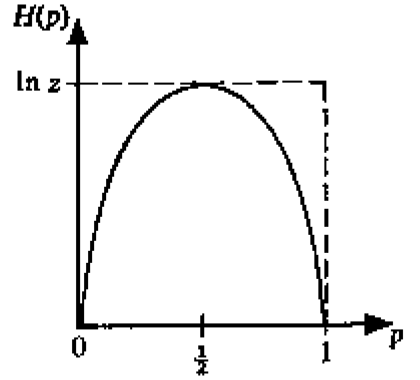


Figure 8. The function $H(p) = -p \ln p - (1 - p)\ln(1 - p)$.

different distributions. If, for example, $p_1 = 1$, $p_2 = \cdots = p_r = 0$, it is clear that this distribution does not admit any indeterminacy: we can say with complete certainty that the result of the experiment will be A_1. On the other hand, if $p_1 = \cdots = p_r = 1/r$, the distribution has maximal indeterminacy, in the sense that it is impossible to discover any preference for the occurrence of one event rather than another.

Consequently it is important to have a quantitative measure of the indeterminacy of different probability distributions, so that we may compare them in this respect. The entropy successfully provides such a measure of indeterminacy; it plays an important role in statistical mechanics and in many significant problems of coding and communication theory.

Suppose now that the sample space is

$$\Omega = \{\omega : \omega = (a_1, \ldots, a_n), a_i = 1, \ldots, r\}$$

and that $p(\omega) = p_1^{\nu_1(\omega)} \cdots p_r^{\nu_r(\omega)}$, where $\nu_i(\omega)$ is the number of occurrences of i in the sequence ω, and $(p_1, \ldots, p_r)$ is a probability distribution.

For $\varepsilon > 0$ and $n = 1, 2, \ldots$, let us put

$$C(n, \varepsilon) = \left\{\omega : \left|\frac{\nu_i(\omega)}{n} - p_i\right| < \varepsilon, i = 1, \ldots, r\right\}.$$

It is clear that

$$\mathsf{P}(C(n, \varepsilon)) \geq 1 - \sum_{i=1}^{r} \mathsf{P}\left\{\left|\frac{\nu_i(\omega)}{n} - p_i\right| \geq \varepsilon\right\},$$

and for sufficiently large n the probabilities $\mathsf{P}\{|(\nu_i(\omega)/n) - p_i| \geq \varepsilon\}$ are arbitrarily small when n is sufficiently large, by the law of large numbers applied to the random variables

$$\xi_k(\omega) = \begin{cases} 1, & a_k = i, \\ 0, & a_k \neq i \end{cases}, \qquad k = 1, \ldots, n.$$

Hence for large n the probability of the event $C(n, \varepsilon)$ is close to 1. Thus, as in the case $n = 2$, a path in $C(n, \varepsilon)$ can be said to be typical.

If all $p_i > 0$, then for every $\omega \in \Omega$

$$p(\omega) = \exp\left\{-n \sum_{k=1}^{r} \left(-\frac{\nu_k(\omega)}{n} \ln p_k\right)\right\}.$$

Consequently if ω is a typical path, we have

$$\left|\sum_{k=1}^{r} \left(-\frac{\nu_k(\omega)}{n} \ln p_k\right) - H\right| \leq -\sum_{k=1}^{r} \left|\frac{\nu_k(\omega)}{n} - p_k\right| \ln p_k \leq -\varepsilon \sum_{k=1}^{r} \ln p_k.$$

It follows that for typical paths the probability $p(\omega)$ is close to e^{-nH} and—since, by the law of large numbers, the typical paths "almost" exhaust Ω when n is large—the number of such paths must be of order e^{nH}. These considerations lead up to the following proposition.

Theorem (Macmillan). *Let $p_i > 0$, $i = 1, \ldots, r$ and $0 < \varepsilon < 1$. Then there is an $n_0 = n_0(\varepsilon; p_1, \ldots, p_r)$ such that for all $n > n_0$*

(a) $e^{n(H-\varepsilon)} \le N(C(n, \varepsilon_1)) \le e^{n(H+\varepsilon)}$;

(b) $e^{-n(H+\varepsilon)} \le p(\omega) \le e^{-n(H-\varepsilon)}, \qquad \omega \in C(n, \varepsilon_1)$;

(c) $\mathsf{P}(C(n, \varepsilon_1)) = \sum_{\omega \in C(n, \varepsilon_n)} p(\omega) \to 1, \qquad n \to \infty,$

where

$$\varepsilon_1 \text{ is the smaller of } \varepsilon \text{ and } \varepsilon/\left\{-2 \sum_{k=1}^{r} \ln p_k\right\}.$$

Proof. Conclusion (c) follows from the law of large numbers. To establish the other conclusions, we notice that if $\omega \in C(n, \varepsilon)$ then

$$np_k - \varepsilon_1 n < \nu_k(\omega) < np_k + \varepsilon_1 n, \qquad k = 1, \ldots, r,$$

and therefore

$$\begin{aligned} p(\omega) = \exp\{-\textstyle\sum \nu_k \ln p_k\} &< \exp\{-n \textstyle\sum p_k \ln p_k - \varepsilon_1 n \sum \ln p_k\} \\ &\le \exp\{-n(H - \tfrac{1}{2}\varepsilon)\}. \end{aligned}$$

Similarly

$$p(\omega) > \exp\{-n(H + \tfrac{1}{2}\varepsilon)\}.$$

Consequently (b) is now established.

Furthermore, since

$$\mathsf{P}(C(n, \varepsilon_1)) \ge N(C(n, \varepsilon_1)) \cdot \min_{\omega \in C(n, \varepsilon_1)} p(\omega),$$

we have

$$N(C(n, \varepsilon_1)) \le \frac{\mathsf{P}(C(n, \varepsilon_1))}{\min_{\omega \in C(n, \varepsilon_1)} p(\omega)} < \frac{1}{e^{-n(H+(1/2)\varepsilon)}} = e^{n(H+(1/2)\varepsilon)}$$

and similarly

$$N(C(n, \varepsilon_1)) \ge \frac{\mathsf{P}(C(n, \varepsilon_1))}{\max_{\omega \in C(n, \varepsilon_1)} p(\omega)} > \mathsf{P}(C(n, \varepsilon_1)) e^{n(H-(1/2)\varepsilon)}.$$

Since $P(C(n, \varepsilon_1)) \to 1$, $n \to \infty$, there is an n_1 such that $P(C(n, \varepsilon_1)) > 1 - \varepsilon$ for $n > n_1$, and therefore

$$\begin{aligned} N(C(n, \varepsilon_1)) &\ge (1 - \varepsilon) \exp\{n(H - \tfrac{1}{2})\} \\ &= \exp\{n(H - \varepsilon) + (\tfrac{1}{2} n\varepsilon + \ln(1 - \varepsilon))\}. \end{aligned}$$

Let n_2 be such that

$$\tfrac{1}{2}n\varepsilon + \ln(1-\varepsilon) > 0.$$

for $n > n_2$. Then when $n \geq n_0 = \max(n_1, n_2)$ we have

$$N(C(n, \varepsilon_1)) \geq e^{n(H-\varepsilon)}.$$

This completes the proof of the theorem.

5. The law of large numbers for Bernoulli schemes lets us give a simple and elegant proof of Weierstrass's theorem on the approximation of continuous functions by polynomials.

Let $f = f(p)$ be a continuous function on the interval [0, 1]. We introduce the polynomials

$$B_n(p) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) C_n^k p^k q^{n-k},$$

which are called Bernstein polynomials after the inventor of this proof of Weierstrass's theorem.

If $\xi_1, \ldots, \xi_n$ is a sequence of independent Bernoulli random variables with $\mathsf{P}\{\xi_i = 1\} = p$, $\mathsf{P}\{\xi_i = 0\} = q$ and $S_n = \xi_1 + \cdots + \xi_n$, then

$$\mathsf{E}f\left(\frac{S_n}{n}\right) = B_n(p).$$

Since the function $f = f(p)$, being continuous on [0, 1], is uniformly continuous, for every $\varepsilon > 0$ we can find $\delta > 0$ such that $|f(x) - f(y)| \leq \varepsilon$ whenever $|x - y| \leq \delta$. It is also clear that the function is bounded: $|f(x)| \leq M < \infty$.

Using this and (5), we obtain

$$\begin{aligned}
|f(p) - B_n(p)| &= \left|\sum_{k=0}^{n}\left[f(p) - f\left(\frac{k}{n}\right)\right] C_n^k p^k q^{n-k}\right| \\
&\leq \sum_{\{k:|(k/n)-p|\leq\delta\}} \left|f(p) - f\left(\frac{k}{n}\right)\right| C_n^k p^k q^{n-k} \\
&\quad + \sum_{\{k:|(k/n)-p|>\delta\}} \left|f(p) - f\left(\frac{k}{n}\right)\right| C_n^k p^k q^{n-k} \\
&\leq \varepsilon + 2M \sum_{\{k:|(k/n)-p|>\delta\}} C_n^k p^k q^{n-k} \leq \varepsilon + \frac{2M}{4n\delta^2} = \varepsilon + \frac{M}{2n\delta^2}.
\end{aligned}$$

Hence

$$\lim_{n\to\infty} \max_{0\leq p\leq 1} |f(p) - B_n(p)| = 0,$$

which is the conclusion of Weierstrass's theorem.

6. Problems

1. Let ξ and η be random variables with correlation coefficient ρ. Establish the following two-dimensional analog of Chebyshev's inequality:

$$\mathsf{P}\{|\xi - \mathsf{E}\xi| \geq \varepsilon\sqrt{\mathsf{V}\xi} \text{ or } |\eta - \mathsf{E}\eta| \geq \varepsilon\sqrt{\mathsf{V}\eta}\} \leq \frac{1}{\varepsilon^2}(1 + \sqrt{1 - \rho^2}).$$

(*Hint*: Use the result of Problem 8 of §4.)

2. Let $f = f(x)$ be a nonnegative even function that is nondecreasing for positive x. Then for a random variable ξ with $|\xi(\omega)| \leq C$,

$$\frac{\mathsf{E}f(\xi) - f(\varepsilon)}{f(C)} \leq \mathsf{P}\{|\xi - \mathsf{E}\xi| \geq \varepsilon\} \leq \frac{\mathsf{E}f(\xi - \mathsf{E}\xi)}{f(\varepsilon)}.$$

In particular, if $f(x) = x^2$,

$$\frac{\mathsf{E}\xi^2 - \varepsilon^2}{C^2} \leq \mathsf{P}\{|\xi - \mathsf{E}\xi| \geq \varepsilon\} \leq \frac{\mathsf{V}\xi}{\varepsilon^2}.$$

3. Let $\xi_1, \ldots, \xi_n$ be a sequence of independent random variables with $\mathsf{V}\xi_i \leq C$. Then

$$\mathsf{P}\left\{\left|\frac{\xi_1 + \cdots + \xi_n}{n} - \frac{\mathsf{E}(\xi_1 + \cdots + \xi_n)}{n}\right| \geq \varepsilon\right\} \leq \frac{C}{n\varepsilon^2}. \tag{15}$$

(With the same reservations as in (8), inequality (15) implies the validity of the law of large numbers in more general contexts than Bernoulli schemes.)

4. Let $\xi_1, \ldots, \xi_n$ be independent Bernoulli random variables with $\mathsf{P}\{\xi_i = 1\} = p > 0$, $\mathsf{P}\{\xi_i = -1\} = 1 - p$. Derive the following *inequality of Bernstein*: there is a number $a > 0$ such that

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - (2p - 1)\right| \geq \varepsilon\right\} \leq 2e^{-a\varepsilon^2 n},$$

where $S_n = \xi_1 + \cdots + \xi_n$ and $\varepsilon > 0$.

§6. The Bernoulli Scheme. II. Limit Theorems (Local, De Moivre–Laplace, Poisson)

1. As in the preceding section, let

$$S_n = \xi_1 + \cdots + \xi_n.$$

Then

$$\mathsf{E}\frac{S_n}{n} = p, \tag{1}$$

and by (4.14)

$$\mathsf{E}\left(\frac{S_n}{n} - p\right)^2 = \frac{pq}{n}. \tag{2}$$

It follows from (1) that $S_n/n \sim p$, where the equivalence symbol $\sim$ has been given a precise meaning in the law of large numbers as the assertion $\mathsf{P}\{|(S_n/n) - p| \geq \varepsilon\} \to 0$. It is natural to suppose that, in a similar way, the relation

$$\left|\frac{S_n}{n} - p\right| \sim \sqrt{\frac{pq}{n}} \tag{3}$$

which follows from (2), can also be given a precise probabilistic meaning involving, for example, probabilities of the form

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \leq x\sqrt{\frac{pq}{n}}\right\}, \qquad x \in R^1,$$

or equivalently

$$\mathsf{P}\left\{\left|\frac{S_n - \mathsf{E}S_n}{\sqrt{\mathsf{V}S_n}}\right| \leq x\right\}$$

(since $\mathsf{E}S_n = np$ and $\mathsf{V}S_n = npq$).

If, as before, we write

$$P_n(k) = C_n^k p^k q^{n-k}, \qquad 0 \leq k \leq n,$$

for $n \geq 1$, then

$$\mathsf{P}\left\{\left|\frac{S_n - \mathsf{E}S_n}{\sqrt{\mathsf{V}S_n}}\right| \leq x\right\} = \sum_{\{k:|(k-np)/\sqrt{npq}| \leq x\}} P_n(k). \tag{4}$$

We set the problem of finding convenient asymptotic formulas, as $n \to \infty$, for $P_n(k)$ and for their sum over the values of k that satisfy the condition on the right-hand side of (4).

The following result provides an answer not only for these values of k (that is, for those satisfying $|k - np| = O(\sqrt{npq})$) but also for those satisfying $|k - np| = o(npq)^{2/3}$.

Local Limit Theorem. *Let* $0 < p < 1$; *then*

$$P_n(k) \sim \frac{1}{\sqrt{2\pi npq}} e^{-(k-np)^2/(2npq)}, \tag{5}$$

uniformly for k *such that* $|k - np| = o(npq)^{2/3}$, *i.e. as* $n \to \infty$

$$\sup_{\{k:|k-np| \leq \varphi(n)\}} \left|\frac{P_n(k)}{\frac{1}{\sqrt{2\pi npq}} e^{-(k-np)^2/(2npq)}} - 1\right| \to 0,$$

where $\varphi(n) = o(npq)^{2/3}$.

The proof depends on Stirling's formula (2.6)

$$n! = \sqrt{2\pi n}\, e^{-n} n^n (1 + R(n)),$$

where $R(n) \to 0$ as $n \to \infty$.

Then if $n \to \infty$, $k \to \infty$, $n - k \to \infty$, we have

$$\begin{aligned}
C_n^k &= \frac{n!}{k!\,(n-k)!} \\
&= \frac{\sqrt{2\pi n}\, e^{-n} n^n}{\sqrt{2\pi k \cdot 2\pi(n-k)}\, e^{-k} k^k \cdot e^{-(n-k)}(n-k)^{n-k}} \frac{1 + R(n)}{(1 + R(k))(1 + R(n-k))} \\
&= \frac{1}{\sqrt{2\pi n \dfrac{k}{n}\left(1 - \dfrac{k}{n}\right)}} \cdot \frac{1 + \varepsilon(n, k, n-k)}{\left(\dfrac{k}{n}\right)^k \left(1 - \dfrac{k}{n}\right)^{n-k}},
\end{aligned}$$

where $\varepsilon = \varepsilon(n, k, n-k)$ is defined in an evident way and $\varepsilon \to 0$ as $n \to \infty$, $k \to \infty$, $n - k \to \infty$.

Therefore

$$P_n(k) = C_n^k p^k q^{n-k} = \frac{1}{\sqrt{2\pi n \dfrac{k}{n}\left(1 - \dfrac{k}{n}\right)}} \frac{p^k (1-p)^{n-k}}{\left(\dfrac{k}{n}\right)^k \left(1 - \dfrac{k}{n}\right)^{n-k}} (1 + \varepsilon).$$

Write $\hat{p} = k/n$. Then

$$\begin{aligned}
P_n(k) &= \frac{1}{\sqrt{2\pi n \hat{p}(1-\hat{p})}} \left(\frac{p}{\hat{p}}\right)^k \left(\frac{1-p}{1-\hat{p}}\right)^{n-k} (1 + \varepsilon) \\
&= \frac{1}{\sqrt{2\pi n \hat{p}(1-\hat{p})}} \exp\left\{k \ln \frac{p}{\hat{p}} + (n-k) \ln \frac{1-p}{1-\hat{p}}\right\} \cdot (1 + \varepsilon) \\
&= \frac{1}{\sqrt{2\pi n \hat{p}(1-\hat{p})}} \exp\left\{n\left[\frac{k}{n} \ln \frac{p}{\hat{p}} + \left(1 - \frac{k}{n}\right) \ln \frac{1-p}{1-\hat{p}}\right]\right\} (1 + \varepsilon) \\
&= \frac{1}{\sqrt{2\pi n \hat{p}(1-\hat{p})}} \exp\{-nH(\hat{p})\}(1 + \varepsilon),
\end{aligned}$$

where

$$H(x) = x \ln \frac{x}{p} + (1-x) \ln \frac{1-x}{1-p}.$$

We are considering values of k such that $|k - np| = o(npq)^{2/3}$, and consequently $p - \hat{p} \to 0$, $n \to \infty$.

Since, for $0 < x < 1$,

$$H'(x) = \ln\frac{x}{p} - \ln\frac{1-x}{1-p},$$

$$H''(x) = \frac{1}{x} + \frac{1}{1-x},$$

$$H'''(x) = -\frac{1}{x^2} + \frac{1}{(1-x)^2},$$

if we write $H(\hat{p})$ in the form $H(p + (\hat{p} - p))$ and use Taylor's formula, we find that for sufficiently large n

$$\begin{aligned} H(\hat{p}) &= H(p) + H'(p)(\hat{p} - p) + \tfrac{1}{2}H''(p)(\hat{p} - p)^2 + O(|\hat{p} - p|^3) \\ &= \frac{1}{2}\left(\frac{1}{p} + \frac{1}{q}\right)(\hat{p} - p)^2 + O(|\hat{p} - p|^3). \end{aligned}$$

Consequently

$$P_n(k) = \frac{1}{\sqrt{2\pi n\hat{p}(1-\hat{p})}} \exp\left\{-\frac{n}{2pq}(\hat{p} - p)^2 + nO(|\hat{p} - p|^3)\right\}(1 + \varepsilon).$$

Notice that

$$\frac{n}{2pq}(\hat{p} - p)^2 = \frac{n}{2pq}\left(\frac{k}{n} - p\right)^2 = \frac{(k - np)^2}{2npq}.$$

Therefore

$$P_n(k) = \frac{1}{\sqrt{2\pi npq}} e^{-(k-np)^2/(2npq)}(1 + \varepsilon'(n, k, n - k)),$$

where

$$1 + \varepsilon'(n, k, n - k) = (1 + \varepsilon(n, k, n - k))\exp\{n\,O(|p - \hat{p}|^3)\}\sqrt{\frac{p(1-p)}{\hat{p}(1-\hat{p})}}$$

and, as is easily seen,

$$\sup|\varepsilon'(n, k, n - k)| \to 0, \qquad n \to \infty,$$

if the sup is taken over the values of k for which

$$|k - np| \le \varphi(n), \qquad \varphi(n) = o(npq)^{2/3}.$$

This completes the proof.

Corollary. *The conclusion of the local limit theorem can be put in the following equivalent form: For all* $x \in \mathbb{R}^1$ *such that* $x = o(npq)^{1/6}$, *and for* $np + x\sqrt{npq}$ *an integer from the set* $\{0, 1, \ldots, n\}$,

$$P_n(np + x\sqrt{npq}) \sim \frac{1}{\sqrt{2\pi npq}} e^{-x^2/2}, \tag{7}$$

i.e. as $n \to \infty$,

$$\sup_{\{x:|x| \le \psi(n)\}} \left| \frac{P_n(np + x\sqrt{npq})}{\frac{1}{\sqrt{2\pi npq}} e^{-x^2/2}} - 1 \right| \to 0, \tag{8}$$

where $\psi(n) = o(npq)^{1/6}$.

With the reservations made in connection with formula (5.8), we can reformulate these results in probabilistic language in the following way:

$$\mathsf{P}\{S_n = k\} \sim \frac{1}{\sqrt{2\pi npq}} e^{-(k-np)^2/(2npq)}, \qquad |k - np| = o(npq)^{2/3}, \tag{9}$$

$$\mathsf{P}\left\{\frac{S_n - np}{\sqrt{npq}} = x\right\} \sim \frac{1}{\sqrt{2\pi npq}} e^{-x^2/2}, \qquad x = o(npq)^{1/6}. \tag{10}$$

(In the last formula $np + x\sqrt{npq}$ is assumed to have one of the values $0, 1, \ldots, n$.)

If we put $t_k = (k - np)/\sqrt{npq}$ and $\Delta t_k = t_{k+1} - t_k = 1/\sqrt{npq}$, the preceding formula assumes the form

$$\mathsf{P}\left\{\frac{S_n - np}{\sqrt{npq}} = t_k\right\} \sim \frac{\Delta t_k}{\sqrt{2\pi}} e^{-t_k^2/2}, \qquad t_k = o(npq)^{1/6}. \tag{11}$$

It is clear that $\Delta t_k = 1/\sqrt{npq} \to 0$ and the set of points $\{t_k\}$ as it were "fills" the real line. It is natural to expect that (11) can be used to obtain the integral formula

$$\mathsf{P}\left\{a < \frac{S_n - np}{\sqrt{npq}} \le b\right\} \sim \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx, \qquad -\infty < a \le b < \infty.$$

Let us now give a precise statement.

2. For $-\infty < a \le b < \infty$ let

$$P_n(a, b] = \sum_{a < x \le b} P_n(np + x\sqrt{npq}),$$

where the summation is over those x for which $np + x\sqrt{npq}$ is an integer.

It follows from the local theorem (see also (11)) that for all t_k defined by $k = np + t_k\sqrt{npq}$ and satisfying $|t_k| \leq T < \infty$,

$$P_n(np + t_k\sqrt{npq}) = \frac{\Delta t_k}{\sqrt{2\pi}} e^{-t_k^2/2}[1 + \varepsilon(t_k, n)], \tag{12}$$

where

$$\sup_{|t_k| \leq T} |\varepsilon(t_k, n)| \to 0, \qquad n \to \infty. \tag{13}$$

Consequently, if a and b are given so that $-T \leq a \leq b \leq T$, then

$$\sum_{a < t_k \leq b} P_n(np + t_k\sqrt{npq}) = \sum_{a < t_k \leq b} \frac{\Delta t_k}{\sqrt{2\pi}} e^{-t_k^2/2} + \sum_{a < t_k \leq b} \varepsilon(t_k, n) \frac{\Delta t_k}{\sqrt{2\pi}} e^{-t_k^2/2}$$

$$= \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx + R_n^{(1)}(a, b) + R_n^{(2)}(a, b), \tag{14}$$

where

$$R_n^{(1)}(a, b) = \sum_{a < t_k \leq b} \frac{\Delta t_k}{\sqrt{2\pi}} e^{-t_k^2/2} - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx,$$

$$R_n^{(2)}(a, b) = \sum_{a < t_k \leq b} \varepsilon(t_k, n) \frac{\Delta t_k}{\sqrt{2\pi}} e^{-t_k^2/2}.$$

From the standard properties of Riemann sums,

$$\sup_{-T \leq a \leq b \leq T} |R_n^{(1)}(a, b)| \to 0, \qquad n \to \infty. \tag{15}$$

It also clear that

$$\begin{aligned}
&\sup_{-T \leq a \leq b \leq T} |R_n^{(2)}(a, b)| \\
&\leq \sup_{|t_k| \leq T} |\varepsilon(t_k, n)| \cdot \sum_{|t_k| \leq T} \frac{\Delta t_k}{\sqrt{2\pi}} e^{-t_k^2/2} \\
&\leq \sup_{|t_k| \leq T} |\varepsilon(t_k, n)| \\
&\times \left[\frac{1}{\sqrt{2\pi}} \int_{-T}^{T} e^{-x^2/2}\, dx + \sup_{-T \leq a \leq b \leq T} |R_n^{(1)}(a, b)|\right] \to 0,
\end{aligned} \tag{16}$$

where the convergence of the right-hand side to zero follows from (15) and from

$$\frac{1}{\sqrt{2\pi}} \int_{-T}^{T} e^{-x^2/2}\, dx \leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = 1, \tag{17}$$

the value of the last integral being well known.

We write

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt.$$

Then it follows from (14)–(16) that

$$\sup_{-T \le a \le b \le T} |P_n(a, b] - (\Phi(b) - \Phi(a))| \to 0, \qquad n \to \infty. \tag{18}$$

We now show that this result holds for $T = \infty$ as well as for finite T. By (17), corresponding to a given $\varepsilon > 0$ we can find a finite $T = T(\varepsilon)$ such that

$$\frac{1}{\sqrt{2\pi}} \int_{-T}^{T} e^{-x^2/2}\, dx > 1 - \tfrac{1}{4}\, \varepsilon. \tag{19}$$

According to (18), we can find an N such that for all $n > N$ and $T = T(\varepsilon)$ we have

$$\sup_{-T \le a \le b \le T} |P_n(a, b] - (\Phi(b) - \Phi(a))| < \tfrac{1}{4}\, \varepsilon. \tag{20}$$

It follows from this and (19) that

$$P_n(-T, T] > 1 - \tfrac{1}{2}\, \varepsilon,$$

and consequently

$$P_n(-\infty, T] + P_n(T, \infty) \le \tfrac{1}{2}\, \varepsilon,$$

where $P_n(-\infty, T] = \lim_{S \downarrow -\infty} P_n(S, T]$ and $P_n(T, \infty) = \lim_{S \uparrow \infty} P_n(T, S]$.

Therefore for $-\infty \le a \le -T < T \le b \le \infty$,

$$\begin{aligned}
&\left| P_n(a, b] - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx \right| \\
&\quad \le \left| P_n(-T, T] - \frac{1}{\sqrt{2\pi}} \int_{-T}^{T} e^{-x^2/2}\, dx \right| \\
&\quad + \left| P_n(a, -T] - \frac{1}{\sqrt{2\pi}} \int_a^{-T} e^{-x^2/2}\, dx \right| + \left| P_n(T, b] - \frac{1}{\sqrt{2\pi}} \int_T^b e^{-x^2/2}\, dx \right| \\
&\quad \le \frac{1}{4}\varepsilon + P_n(-\infty, -T] + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-T} e^{-x^2/2}\, dx + P_n(T, \infty) \\
&\quad + \frac{1}{\sqrt{2\pi}} \int_T^{\infty} e^{-x^2/2}\, dx \le \frac{1}{4}\varepsilon + \frac{1}{2}\varepsilon + \frac{1}{8}\varepsilon + \frac{1}{8}\varepsilon = \varepsilon.
\end{aligned}$$

By using (18) it is now easy to see that $P_n(a, b]$ tends uniformly to $\Phi(b) - \Phi(a)$ for $-\infty \le a < b \le \infty$.

Thus we have proved the following theorem.

De Moivre–Laplace Integral Theorem. *Let* $0 < p < 1$,

$$P_n(k) = C_n^k p^k q^{n-k}, \qquad P_n(a, b] = \sum_{a < x \le b} P_n(np + x\sqrt{npq}).$$

Then

$$\sup_{-\infty \le a < b \le \infty} \left| P_n(a, b] - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx \right| \to 0, \qquad n \to \infty. \tag{21}$$

With the same reservations as in (5.8), (21) can be stated in probabilistic language in the following way:

$$\sup_{-\infty \le a < b \le \infty} \left| \mathsf{P}\left\{ a < \frac{S_n - \mathsf{E}S_n}{\sqrt{\mathsf{V}S_n}} \le b \right\} - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx \right| \to 0, \qquad n \to \infty.$$

It follows at once from this formula that

$$\mathsf{P}\{A < S_n \le B\} - \left[\Phi\left(\frac{B - np}{\sqrt{npq}} \right) - \Phi\left(\frac{A - np}{\sqrt{npq}} \right) \right] \to 0, \tag{22}$$

as $n \to \infty$, whenever $-\infty \le A < B \le \infty$.

Example. A true die is tossed 12 000 times. We ask for the probability P that the number of 6's lies in the interval (1800, 2100].

The required probability is

$$P = \sum_{1800 < k \le 2100} C_{12\,000}^k \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{12\,000 - k}.$$

An exact calculation of this sum would obviously be rather difficult. However, if we use the integral theorem we find that the probability P in question is ($n = 12\,000$, $p = \frac{1}{6}$, $a = 1800$, $b = 2100$)

$$\Phi\left(\frac{2100 - 2000}{\sqrt{12\,000 \cdot \frac{1}{6} \cdot \frac{5}{6}}} \right) - \Phi\left(\frac{1800 - 2000}{\sqrt{12\,000 \cdot \frac{1}{6} \cdot \frac{5}{6}}} \right) = \Phi(\sqrt{6}) - \Phi(-2\sqrt{6})$$

$$\approx \Phi(2.449) - \Phi(-4.898) \approx 0.992,$$

where the values of $\Phi(2.449)$ and $\Phi(-4.898)$ were taken from tables of $\Phi(x)$ (this is the normal distribution function; see Subsection 6 below).

3. We have plotted a graph of $P_n(np + x\sqrt{npq})$ (with x assumed such that $np + x\sqrt{npq}$ is an integer) in Figure 9.

Then the local theorem says that when $x = o(npq)^{1/6}$, the curve $(1/\sqrt{2\pi npq})e^{-x^2/2}$ provides a close fit to $P_n(np + x\sqrt{npq})$. On the other hand the integral theorem says that $P_n(a, b] = \mathsf{P}\{a\sqrt{npq} < S_n - np \le b\sqrt{npq}\} = \mathsf{P}\{np + a\sqrt{npq} < S_n \le np + b\sqrt{npq}\}$ is closely approximated by the integral $(1/\sqrt{2\pi})\int_a^b e^{-x^2/2}\, dx$.

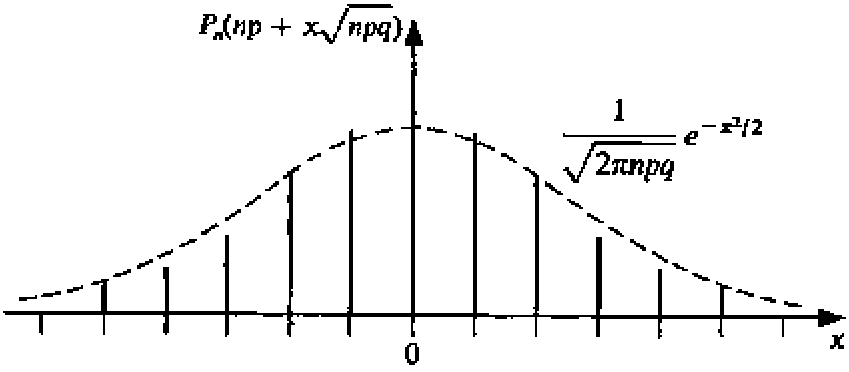


Figure 9

We write

$$F_n(x) = P_n(-\infty, x] \quad \left(= \mathsf{P}\left\{\frac{S_n - np}{\sqrt{npq}} \leq x\right\}\right).$$

Then it follows from (21) that

$$\sup_{-\infty \leq x \leq \infty} |F_n(x) - \Phi(x)| \to 0, \qquad n \to \infty. \tag{23}$$

It is natural to ask how rapid the approach to zero is in (21) and (23), as $n \to \infty$. We quote a result in this direction (a special case of the Berry–Esseen theorem: see §6 in Chapter III):

$$\sup_{-\infty \leq x \leq \infty} |F_n(x) - \Phi(x)| \leq \frac{p^2 + q^2}{\sqrt{npq}}. \tag{24}$$

It is important to recognize that the order of the estimate $(1/\sqrt{npq})$ cannot be improved; this means that the approximation of $F_n(x)$ by $\Phi(x)$ can be poor for values of p that are close to 0 or 1, even when n is large. This suggests the question of whether there is a better method of approximation for the probabilities of interest when p or q is small, something better than the normal approximation given by the local and integral theorems. In this connection we note that for $p = \frac{1}{2}$, say, the binomial distribution $\{P_n(k)\}$ is symmetric (Figure 10). However, for small p the binomial distribution is asymmetric (Figure 10), and hence it is not reasonable to expect that the normal approximation will be satisfactory.

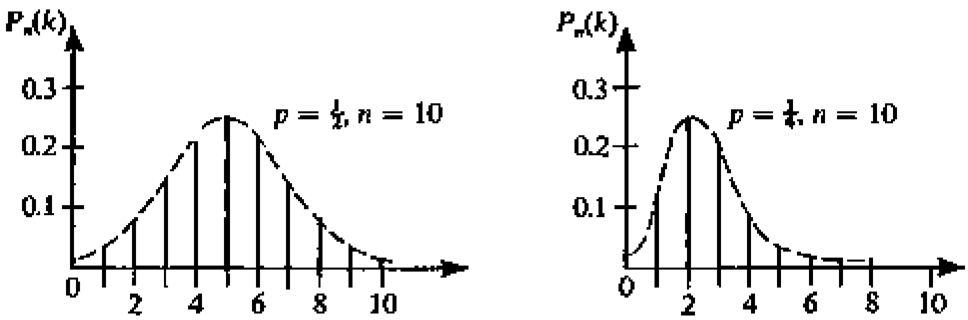


Figure 10

4. It turns out that for small values of p the distribution known as the Poisson distribution provides a good approximation to $\{P_n(k)\}$.

Let

$$P_n(k) = \begin{cases} C_n^k p^k q^{n-k}, & k = 0, 1, \ldots, n, \\ 0, & k = n+1, n+2, \ldots, \end{cases}$$

and suppose that p is a function $p(n)$ of n.

Poisson's Theorem. *Let $p(n) \to 0$, $n \to \infty$, in such a way that $np(n) \to \lambda$, where $\lambda > 0$. Then for $k = 1, 2, \ldots,$*

$$P_n(k) \to \pi_k, \qquad n \to \infty, \tag{25}$$

where

$$\pi_k = \frac{\lambda^k e^{-\lambda}}{k!}, \qquad k = 0, 1, \ldots. \tag{26}$$

The proof is extremely simple. Since $p(n) = (\lambda/n) + o(1/n)$ by hypothesis, for a given $k = 0, 1, \ldots$ and sufficiently large n,

$$\begin{aligned} P_n(k) &= C_n^k p^k q^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{k!}\left[\frac{\lambda}{n} + o\left(\frac{1}{n}\right)\right]^k \cdot \left[1 - \frac{\lambda}{n} + o\left(\frac{1}{n}\right)\right]^{n-k}. \end{aligned}$$

But

$$\begin{aligned} &n(n-1)\cdots(n-k+1)\left[\frac{\lambda}{n} + o\left(\frac{1}{n}\right)\right]^k \\ &\quad = \frac{n(n-1)\cdots(n-k+1)}{n^k}[\lambda + o(1)]^k \to \lambda^k, \qquad n \to \infty, \end{aligned}$$

and

$$\left[1 - \frac{\lambda}{n} + o\left(\frac{1}{n}\right)\right]^{n-k} \to e^{-\lambda}, \qquad n \to \infty,$$

which establishes (25).

The set of numbers $\{\pi_k,\ k = 0, 1, \ldots\}$ defines the *Poisson probability distribution* ($\pi_k \ge 0$, $\sum_{k=0}^{\infty} \pi_k = 1$). Notice that all the (discrete) distributions considered previously were concentrated at only a finite number of points. The Poisson distribution is the first example that we have encountered of a (discrete) distribution concentrated at a countable number of points.

The following result of Prokhorov exhibits the rapidity with which $P_n(k)$ converges to π_k as $n \to \infty$: if $np(n) = \lambda > 0$, then

$$\sum_{k=0}^{\infty} |P_n(k) - \pi_k| \le \frac{2\lambda}{n} \cdot \min(2, \lambda). \tag{27}$$

(A proof of a somewhat weaker result is given in §12, Chapter III.)

5. Let us return to the De Moivre–Laplace limit theorem, and show how it implies the law of large numbers (with the same reservation that was made in connection with (5.8)). Since

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \le \varepsilon\right\} = \mathsf{P}\left\{\left|\frac{S_n - np}{\sqrt{npq}}\right| \le \varepsilon\sqrt{\frac{n}{pq}}\right\},$$

it is clear from (21) that when $\varepsilon > 0$

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \le \varepsilon\right\} - \frac{1}{\sqrt{2\pi}}\int_{-\varepsilon\sqrt{n/pq}}^{\varepsilon\sqrt{n/pq}} e^{-x^2/2}\,dx \to 0, \qquad n \to \infty, \tag{28}$$

whence

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \le \varepsilon\right\} \to 1, \qquad n \to \infty,$$

which is the conclusion of the law of large numbers.

From (28)

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \le \varepsilon\right\} \sim \frac{1}{\sqrt{2\pi}}\int_{-\varepsilon\sqrt{n/pq}}^{\varepsilon\sqrt{n/pq}} e^{-x^2/2}\,dx, \qquad n \to \infty, \tag{29}$$

whereas Chebyshev's inequality yielded only

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \le \varepsilon\right\} \ge 1 - \frac{pq}{n\varepsilon^2}.$$

It was shown at the end of §5 that Chebyshev's inequality yielded the estimate

$$n \ge \frac{1}{4\varepsilon^2\alpha}$$

for the number of observations needed for the validity of the inequality

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \le \varepsilon\right\} \ge 1 - \alpha.$$

Thus with $\varepsilon = 0.02$ and $\alpha = 0.05$, 12 500 observations were needed. We can now solve the same problem by using the approximation (29).

We define the number $k(\alpha)$ by

$$\frac{1}{\sqrt{2\pi}}\int_{-k(\alpha)}^{k(\alpha)} e^{-x^2/2}\,dx = 1 - \alpha.$$

Since $\varepsilon\sqrt{(n/pq)} \ge 2\varepsilon\sqrt{n}$, if we define n as the smallest integer satisfying

$$2\varepsilon\sqrt{n} \ge k(\alpha) \tag{30}$$

we find that

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \le \varepsilon\right\} \gtrsim 1 - \alpha. \tag{31}$$

We find from (30) that the smallest integer n satisfying

$$n \geq \frac{k^2(\alpha)}{4\varepsilon^2}$$

guarantees that (31) is satisfied, and the accuracy of the approximation can easily be established by using (24).

Taking $\varepsilon = 0.02$, $\alpha = 0.05$, we find that in fact 2500 observations suffice, rather than the 12 500 found by using Chebyshev's inequality. The values of $k(\alpha)$ have been tabulated. We quote a number of values of $k(\alpha)$ for various values of α:

α	$k(a)$
0.50	0.675
0.3173	1.000
0.10	1.645
0.05	1.960
0.0454	2.000
0.01	2.576
0.0027	3.000

6. The function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt, \tag{32}$$

which was introduced above and occurs in the De Moivre–Laplace integral theorem, plays an exceptionally important role in probability theory. It is known as the *normal* or *Gaussian distribution* on the real line, with the (normal or Gaussian) density

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad x \in R^1.$$

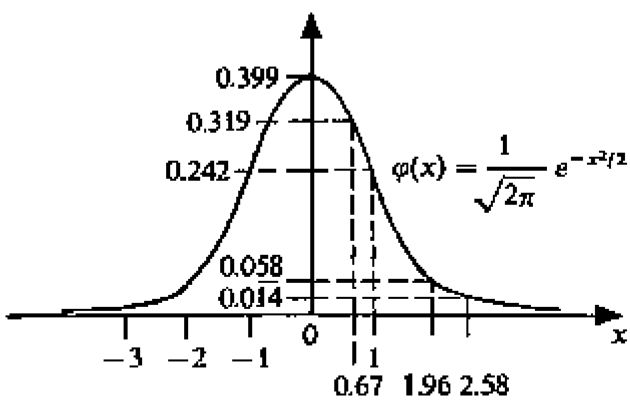


Figure 11. Graph of the normal probability density $\varphi(x)$.

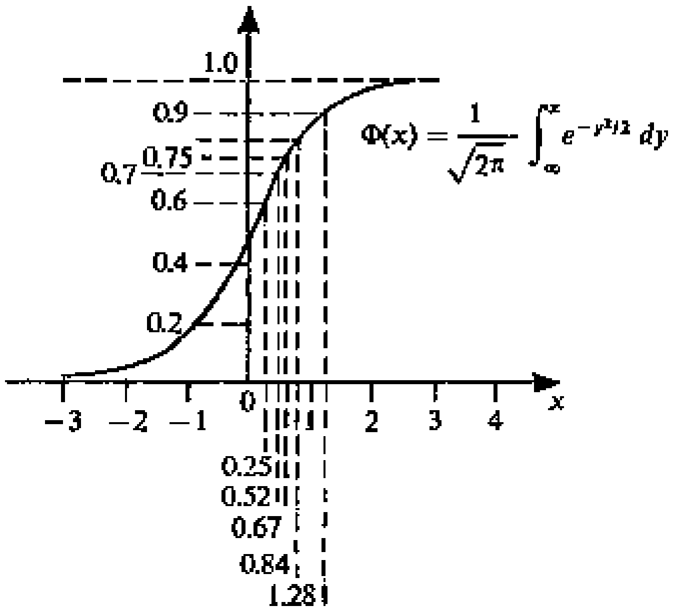


Figure 12. Graph of the normal distribution $\Phi(x)$.

We have already encountered (discrete) distributions concentrated on a finite or countable set of points. The normal distribution belongs to another important class of distributions that arise in probability theory. We have mentioned its exceptional role; this comes about, first of all, because under rather general hypotheses, sums of a large number of independent random variables (not necessarily Bernoulli variables) are closely approximated by the normal distribution (§4 of Chapter III). For the present we mention only some of the simplest properties of $\varphi(x)$ and $\Phi(x)$, whose graphs are shown in Figures 11 and 12.

The function $\varphi(x)$ is a symmetric bell-shaped curve, decreasing very rapidly with increasing $|x|$: thus $\varphi(1) = 0.24197$, $\varphi(2) = 0.053991$, $\varphi(3) = 0.004432$, $\varphi(4) = 0.000134$, $\varphi(5) = 0.000016$. Its maximum is attained at $x = 0$ and is equal to $(2\pi)^{-1/2} \approx 0.399$.

The curve $\Phi(x) = (1/\sqrt{2\pi})\int_{-\infty}^{x} e^{-t^2/2}\,dt$ approximates 1 very rapidly as x increases: $\Phi(1) = 0.841345$, $\Phi(2) = 0.977250$, $\Phi(3) = 0.998650$, $\Phi(4) = 0.999968$, $\Phi(4, 5) = 0.999997$.

For tables of $\varphi(x)$ and $\Phi(x)$, as well as of other important functions that are used in probability theory and mathematical statistics, see [A1].

7. At the end of subsection 3, §5, we noticed that the upper bound for the probability of the event $\{\omega\colon |(S_n/n) - p| \geq \varepsilon\}$, given by Chebyshev's inequality, was rather crude. That estimate was obtained from Chebyshev's inequality $\mathsf{P}\{X \geq \varepsilon\} \leq \mathsf{E}X^2/\varepsilon^2$ for nonnegative random variables $X \geq 0$. We may, however, use Chebyshev's inequality in the form

$$\mathsf{P}\{X \geq \varepsilon\} = \mathsf{P}\{X^{2k} \geq \varepsilon^{2k}\} \leq \frac{\mathsf{E}X^{2k}}{\varepsilon^{2k}}. \tag{33}$$

However, we can go further by using the "exponential form" of Chebyshev's inequality: if $X \geq 0$ and $\lambda > 0$, this states that

$$\mathsf{P}\{X \geq \varepsilon\} = \mathsf{P}\{e^{\lambda X} \geq e^{\lambda \varepsilon}\} \leq \mathsf{E}e^{\lambda(X-\varepsilon)}. \tag{34}$$

Since the positive number λ is arbitrary, it is clear that

$$\mathsf{P}\{X \geq \varepsilon\} \leq \inf_{\lambda > 0} \mathsf{E}e^{\lambda(X-\varepsilon)}. \tag{35}$$

Let us see what the consequences of this approach are in the case when $X = S_n/n$, $S_n = \xi_1 + \cdots + \xi_n$, $P(\xi_i = 1) = p$, $P(\xi_i = 0) = q$, $i \geq 1$.

Let us set $\varphi(\lambda) = \mathsf{E}e^{\lambda \xi_1}$. Then

$$\varphi(\lambda) = 1 - p + pe^{\lambda}$$

and, under the hypothesis of the independence of $\xi_1, \xi_2, \ldots, \xi_n$,

$$\mathsf{E}e^{\lambda S_n} = [\varphi(\lambda)]^n.$$

Therefore, $(0 < a < 1)$

$$\begin{aligned}\mathsf{P}\left\{\frac{S_n}{n} \geq a\right\} &\leq \inf_{\lambda > 0} \mathsf{E}e^{\lambda(S_n/n - a)} = \inf_{\lambda > 0} e^{-n[\lambda a/n - \ln \varphi(\lambda/n)]} \\ &= \inf_{s > 0} e^{-n[as - \ln \varphi(s)]} = e^{-n \sup_{s > 0}[as - \ln \varphi(s)]}. \end{aligned}\tag{36}$$

Similarly,

$$\mathsf{P}\left\{\frac{S_n}{n} \leq a\right\} \leq e^{-n \sup_{s<0}[as - \ln \varphi(s)]}. \tag{37}$$

The function $f(s) = as - \log[1 - p + pe^S]$ attains its maximum for $p \leq a \leq 1$ at the point $s_0(f'(s_0) = 0)$ determined by the equation

$$e^{s_0} = \frac{a(1-p)}{p(1-a)}.$$

Consequently,

$$\sup_{s > 0} f(s) = H(a),$$

where

$$H(a) = a \ln \frac{a}{p} + (1 - a) \ln \frac{1-a}{1-p}$$

is the function that was previously used in the proof of the local theorem (subsection 1).

Thus, for $p \leq a \leq 1$

$$\mathsf{P}\left\{\frac{S_n}{n} \geq a\right\} \leq e^{-nH(a)}, \tag{38}$$

and therefore, since $H(p + x) \geq 2x^2$ and $0 \leq p + x \leq 1$, we have, for $\varepsilon > 0$ and $0 \leq p \leq 1$,

$$\mathsf{P}\left\{\frac{S_n}{n} - p \geq \varepsilon\right\} \leq e^{-2n\varepsilon^2}. \tag{39}$$

We can establish similarly that for $a \leq p \leq 1$

$$\mathsf{P}\left\{\frac{S_n}{n} \leq a\right\} \leq e^{-nH(a)}, \tag{40}$$

and consequently, for every $\varepsilon > 0$ and $0 \leq p \leq 1$,

$$\mathsf{P}\left\{\frac{S_n}{n} - p \leq -\varepsilon\right\} \leq e^{-2n\varepsilon^2}. \tag{41}$$

Therefore,

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right\} \leq 2e^{-2n\varepsilon^2}. \tag{42}$$

Hence, it follows that the number $n_3(\alpha)$ of observations of the inequality

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \leq \varepsilon\right\} \geq 1 - \alpha, \tag{43}$$

that are guaranteed to be satisfied for every p, $0 < p < 1$, is determined by the formula

$$n_3(\alpha) = \left[\frac{\ln(2/\alpha)}{2\varepsilon^2}\right], \tag{44}$$

where $[x]$ is the integral part of x. If we neglect "integral parts" and compare $n_3(\alpha)$ with $n_1(\alpha) = [(4\alpha\varepsilon^2)^{-1}]$, we find that

$$\frac{n_1(\alpha)}{n_3(\alpha)} = \frac{1}{2\alpha \ln \dfrac{2}{\alpha}} \uparrow \infty, \qquad \alpha \downarrow 0.$$

It is clear from this that when $\alpha \downarrow 0$, an estimate of the smallest number of observations needed that can be obtained from the exponential Chebyshev inequality is more precise than the estimate obtained from the ordinary Chebyshev inequality, especially for small α.

There is no difficulty in applying the formula

$$\frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2}\, dy \sim \frac{1}{\sqrt{2\pi}x} e^{-x^2/2}, \qquad x \to \infty,$$

to show that $k^2(\alpha) \sim 2 \log(2/\alpha)$ when $\alpha \downarrow 0$. Therefore,

$$\frac{n_2(\alpha)}{n_3(\alpha)} \to 1, \qquad \alpha \downarrow 0.$$

Inequalities like (38)–(42) are known as *inequalities for the probability of large deviations*. This terminology can be explained in the following way.

The De Moivre–Laplace integral theorem makes it possible to estimate in a simple way the probabilities of the events $\{|S_n - np| \le x\sqrt{n}\}$ characterizing the "standard" deviation (up to order $\sqrt{n}$) of S_n from np. Even the inequalities (39), (41), and (42) provide an estimate of the probabilities of the events $\{\omega : |S_n - np| \le xn\}$, describing deviations of order greater than $\sqrt{n}$, in fact of order n.

We shall continue the discussion of probabilities of large deviations, in more general situations, in §5, chap. IV.

8. Problems

1. Let $n = 100$, $p = \frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \frac{4}{10}, \frac{5}{10}$. Using tables (for example, those in [A1]) of the binomial and Poisson distributions, compare the values of the probabilities

$$\mathsf{P}\{10 < S_{100} \le 12\}, \qquad \mathsf{P}\{20 < S_{100} \le 22\},$$
$$\mathsf{P}\{33 < S_{100} \le 35\}, \qquad \mathsf{P}\{40 < S_{100} \le 42\},$$
$$\mathsf{P}\{50 < S_{100} \le 52\}$$

with the corresponding values given by the normal and Poisson approximations.

2. Let $p = \frac{1}{2}$ and $Z_n = 2S_n - n$ (the excess of 1's over 0's in n trials). Show that

$$\sup_j |\sqrt{\pi n}\, P\{Z_{2n} = j\} - e^{-j^2/4n}| \to 0, \qquad n \to \infty.$$

3. Show that the rate of convergence in Poisson's theorem is given by

$$\sup_k \left| P_n(k) - \frac{\lambda^k e^{-\lambda}}{k!} \right| \le \frac{2\lambda^2}{n}.$$

§7. Estimating the Probability of Success in the Bernoulli Scheme

1. In the Bernoulli scheme $(\Omega, \mathscr{A}, \mathsf{P})$ with $\Omega = \{\omega : \omega = (x_1, \ldots, x_n), x_i = 0, 1)\}$, $\mathscr{A} = A : A \subseteq \Omega\}$,

$$p(\omega) = p^{\Sigma x_i} q^{n - \Sigma x_i},$$

we supposed that p (the probability of success) was known.

Let us now suppose that p is not known in advance and that we want to determine it by observing the outcomes of experiments; or, what amounts to the same thing, by observations of the random variables $\xi_1, \ldots, \xi_n$, where $\xi_i(\omega) = x_i$. This is a typical problem of mathematical statistics, and can be formulated in various ways. We shall consider two of the possible formulations: the problem of *estimation* and the problem of *constructing confidence intervals*.

In the notation used in mathematical statistics, the unknown parameter is denoted by θ, assuming *a priori* that θ belongs to the set $\Theta = [0, 1]$. We say that the set $(\Omega, \mathscr{A}, \mathsf{P}_\theta; \theta \in \Theta)$ with $p_\theta(\omega) = \theta^{\Sigma x_i}(1 - \theta)^{n - \Sigma x_i}$ is a probabilistic-statistical model (corresponding to "n independent trials" with probability of "success" $\theta \in \Theta$), and any function $T_n = T_n(\omega)$ with values in Θ is called an *estimator*.

If $S_n = \xi_1 + \cdots + \xi_n$ and $T_n^* = S_n/n$, it follows from the law of large numbers that T_n^* is *consistent*, in the sense that ($\varepsilon > 0$)

$$\mathsf{P}_\theta\{|T_n^* - \theta| \geq \varepsilon\} \to 0, \qquad n \to \infty. \tag{1}$$

Moreover, this estimator is *unbiased*: for every θ

$$\mathsf{E}_\theta T_n^* = \theta, \tag{2}$$

where E_θ is the expectation corresponding to the probability P_θ.

The property of being unbiased is quite natural: it expresses the fact that any reasonable estimate ought, at least "on the average," to lead to the desired result. However, it is easy to see that T_n^* is not the only unbiased estimator. For example, the same property is possessed by every estimator

$$T_n = \frac{b_1 x_1 + \cdots + b_n x_n}{n},$$

where $b_1 + \cdots + b_n = n$. Moreover, the law of large numbers (1) is also satisfied by such estimators (at least if $|b_i| \leq K < \infty$; see Problem 2, §3, Chapter III) and so these estimators T_n are just as "good" as T_n^*.

In this connection there arises the question of how to compare different unbiased estimators, and which of them to describe as best, or optimal.

With the same meaning of "estimator," it is natural to suppose that an estimator is better, the smaller its deviation from the parameter that is being estimated. On this basis, we call an estimator $\tilde{T}_n$ *efficient* (in the class of unbiased estimators T_n) if,

$$\mathsf{V}_\theta \tilde{T}_n = \inf_{T_n} \mathsf{V}_\theta T_n, \qquad \theta \in \Theta, \tag{3}$$

where $\mathsf{V}_\theta T_n$ is the dispersion of T_n, i.e. $\mathsf{E}_\theta(T_n - \theta)^2$.

Let us show that the estimator T_n^*, considered above, is efficient. We have

$$\mathsf{V}_\theta T_n^* = \mathsf{V}_\theta\left(\frac{S_n}{n}\right) = \frac{\mathsf{V}_\theta S_n}{n^2} = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}. \tag{4}$$

Hence to establish that T_n^* is efficient, we have only to show that

$$\inf_{T_n} \mathsf{V}_\theta T_n \geq \frac{\theta(1-\theta)}{n}. \tag{5}$$

This is obvious for $\theta = 0$ or 1. Let $\theta \in (0, 1)$ and

$$p_\theta(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}.$$

It is clear that

$$p_\theta(\omega) = \prod_{i=1}^n p_\theta(x_i).$$

Let us write

$$L_\theta(\omega) = \ln p_\theta(\omega).$$

Then

$$L_\theta(\omega) = \ln \theta \cdot \sum x_i + \ln(1-\theta)\sum(1-x_i)$$

and

$$\frac{\partial L_\theta(\omega)}{\partial \theta} = \frac{\sum(x_i - \theta)}{\theta(1-\theta)}.$$

Since

$$1 = \mathsf{E}_\theta 1 = \sum_\omega p_\theta(\omega),$$

and since T_n is unbiased,

$$\theta \equiv \mathsf{E}_\theta T_n = \sum_\omega T_n(\omega) p_\theta(\omega).$$

After differentiating with respect to θ, we find that

$$0 = \sum_\omega \frac{\partial p_\theta(\omega)}{\partial \theta} = \sum_\omega \frac{\left(\dfrac{\partial p_\theta(\omega)}{\partial \theta}\right)}{p_\theta(\omega)} p_\theta(\omega) = \mathsf{E}_\theta\left[\frac{\partial L_\theta(\omega)}{\partial \theta}\right],$$

$$1 = \sum_\omega T_n \frac{\left(\dfrac{\partial p_\theta(\omega)}{\partial \theta}\right)}{p_\theta(\omega)} p_\theta(\omega) = \mathsf{E}_\theta\left[T_n \frac{\partial L_\theta(\omega)}{\partial \theta}\right].$$

Therefore

$$1 = \mathsf{E}_\theta\left[(T_n - \theta)\frac{\partial L_\theta(\omega)}{\partial \theta}\right]$$

and by the Cauchy–Bunyakovskii inequality,

$$1 \le \mathsf{E}_\theta[T_n - \theta]^2 \cdot \mathsf{E}_\theta\left[\frac{\partial L_\theta(\omega)}{\partial \theta}\right]^2,$$

whence

$$\mathsf{E}_\theta[T_n - \theta]^2 \ge \frac{1}{I_n(\theta)}, \tag{6}$$

where

$$I_n(\theta) = \left[\frac{\partial L_\theta(\omega)}{\partial \theta}\right]^2$$

is known as *Fisher's information*.

From (6) we can obtain a special case of the Rao–Cramér inequality for unbiased estimators T_n:

$$\inf_{T_n} \mathsf{V}_\theta T_n \ge \frac{1}{I_n(\theta)}. \tag{7}$$

In the present case

$$I_n(\theta) = \mathsf{E}_\theta\left[\frac{\partial L_\theta(\omega)}{\partial\theta}\right]^2 = \mathsf{E}_\theta\left[\frac{\sum(\xi_i - \theta)}{\theta(1-\theta)}\right]^2 = \frac{n\theta(1-\theta)}{[\theta(1-\theta)]^2} = \frac{n}{\theta(1-\theta)},$$

which also establishes (5), from which, as we already noticed, there follows the efficiency of the unbiased estimator $T_n^* = S_n/n$ for the unknown parameter θ.

2. It is evident that, in considering T_n^* as a pointwise estimator for θ, we have introduced a certain amount of inaccuracy. It can even happen that the numerical value of T_n^* calculated from observations of $x_1, \ldots, x_n$ differs rather severely from the true value θ. Hence it would be advisable to determine the size of the error.

It would be too much to hope that $T_n^*(\omega)$ differs little from the true value θ for all sample points ω. However, we know from the law of large numbers that for every $\delta > 0$ and for sufficiently large n, the probability of the event $\{|\theta - T_n^*(\omega)| > \delta\}$ will be arbitrarily small.

By Chebyshev's inequality

$$\mathsf{P}_\theta\{|\theta - T_n^*| > \delta\} \le \frac{\mathsf{V}_\theta T_n^*}{\delta^2} = \frac{\theta(1-\theta)}{n\delta^2}$$

and therefore, for every $\lambda > 0$,

$$\mathsf{P}_\theta\left\{|\theta - T_n^*| \le \lambda\sqrt{\frac{\theta(1-\theta)}{n}}\right\} \ge 1 - \frac{1}{\lambda^2}.$$

If we take, for example, $\lambda = 3$, then with P_θ-probability greater than 0.888 $(1 - (1/3^2) = \frac{8}{9} \approx 0.8889)$ the event

$$|\theta - T_n^*| \le 3\sqrt{\frac{\theta(1-\theta)}{n}}$$

will be realized, and a fortiori the event

$$|\theta - T_n^*| \le \frac{3}{2\sqrt{n}},$$

since $\theta(1-\theta) \le \frac{1}{4}$.

Therefore

$$\mathsf{P}_\theta\left\{|\theta - T_n^*| \le \frac{3}{2\sqrt{n}}\right\} = \mathsf{P}_\theta\left\{T_n^* - \frac{3}{2\sqrt{n}} \le \theta \le T_n^* + \frac{3}{2\sqrt{n}}\right\} \ge 0.8888.$$

In other words, we can say with probability greater than 0.8888 that the exact value of θ is in the interval $[T_n^* - (3/2\sqrt{n}), T_n^* + (3/2\sqrt{n})]$. This statement is sometimes written in the symbolic form

$$\theta \simeq T_n^* \pm \frac{3}{2\sqrt{n}} \quad (\geq 88\%),$$

where "$\geq 88\%$" means "in more than 88% of all cases."

The interval $[T_n^* - (3/2\sqrt{n}), T_n^* + (3/2\sqrt{n})]$ is an example of what are called confidence intervals for the unknown parameter.

Definition. An interval of the form

$$[\psi_1(\omega), \psi_2(\omega)]$$

where $\psi_1(\omega)$ and $\psi_2(\omega)$ are functions of sample points, is called a *confidence interval of reliability* $1 - \delta$ (or of *significance level* δ) if

$$\mathsf{P}_\theta\{\psi_1(\omega) \leq \theta \leq \psi_2(\omega)\} \geq 1 - \delta.$$

for all $\theta \in \Theta$.

The preceding discussion shows that the interval

$$\left[T_n^* - \frac{\lambda}{2\sqrt{n}}, T_n^* + \frac{\lambda}{2\sqrt{n}}\right]$$

has reliability $1 - (1/\lambda^2)$. In point of fact, the reliability of this confidence interval is considerably higher, since Chebyshev's inequality gives only crude estimates of the probabilities of events.

To obtain more precise results we notice that

$$\left\{\omega: |\theta - T_n^*| \leq \lambda\sqrt{\frac{\theta(1-\theta)}{n}}\right\} = \{\omega: \psi_1(T_n^*, n) \leq \theta \leq \psi_2(T_n^*, n)\},$$

where $\psi_1 = \psi_1(T_n^*, n)$ and $\psi_2 = \psi_2(T_n^*, n)$ are the roots of the quadratic equation

$$(\theta - T_n^*)^2 = \frac{\lambda^2}{n}\theta(1 - \theta),$$

which describes an ellipse situated as shown in Figure 13.

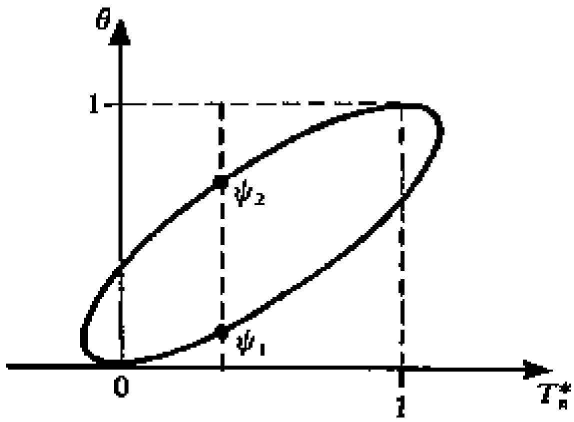


Figure 13

Now let

$$F_\theta^n(x) = \mathsf{P}_\theta\left\{\frac{S_n - n\theta}{\sqrt{n\theta(1-\theta)}} \le x\right\}.$$

Then by (6.24)

$$\sup_x |F_\theta^n(x) - \Phi(x)| \le \frac{1}{\sqrt{n\theta(1-\theta)}}.$$

Therefore if we know *a priori* that

$$0 < \Delta \le \theta \le 1 - \Delta < 1,$$

where Δ is a constant, then

$$\sup_x |F_\theta^n(x) - \Phi(x)| \le \frac{1}{\Delta\sqrt{n}}$$

and consequently

$$\begin{aligned}\mathsf{P}_\theta\{\psi_1(T_n^*, n) \le \theta \le \psi_2(T_n^*, n)\} &= \mathsf{P}_\theta\left\{|\theta - T_n^*| \le \lambda\sqrt{\frac{\theta(1-\theta)}{n}}\right\}\\ &= \mathsf{P}_\theta\left\{\frac{|S_n - n\theta|}{\sqrt{n\theta(1-\theta)}} \le \lambda\right\}\\ &\ge (2\Phi(\lambda) - 1) - \frac{2}{\Delta\sqrt{n}}.\end{aligned}$$

Let λ^* be the smallest λ for which

$$(2\Phi(\lambda) - 1) - \frac{2}{\Delta\sqrt{n}} \ge 1 - \delta^*,$$

where δ^* is a given significance level. Putting $\delta = \delta^* - (2/\Delta\sqrt{n})$, we find that λ^* satisfies the equation

$$\Phi(\lambda) = 1 - \frac{1}{2}\delta.$$

For large n we may neglect the term $2/\Delta\sqrt{n}$ and assume that λ^* satisfies

$$\Phi(\lambda^*) = 1 - \frac{1}{2}\delta^*.$$

In particular, if $\lambda^* = 3$ then $1 - \delta^* = 0.9973\ldots$. Then with probability approximately 0.9973

$$T_n^* - 3\sqrt{\frac{\theta(1-\theta)}{n}} \le \theta \le T_n^* + 3\sqrt{\frac{\theta(1-\theta)}{n}} \tag{8}$$

or, after iterating and then suppressing terms of order $O(n^{-3/4})$, we obtain

$$T_n^* - 3\sqrt{\frac{T_n^*(1 - T_n^*)}{n}} \leq \theta \leq T_n^* + 3\sqrt{\frac{T_n^*(1 - T_n^*)}{n}}. \tag{9}$$

Hence it follows that the confidence interval

$$\left[T_n^* - \frac{3}{2\sqrt{n}}, T_n^* + \frac{3}{2\sqrt{n}}\right] \tag{10}$$

has (for large n) reliability 0.9973 (whereas Chebyshev's inequality only provided reliability approximately 0.8889).

Thus we can make the following practical application. Let us carry out a large number N of series of experiments, in each of which we estimate the parameter θ after n observations. Then in about 99.73% of the N cases, in each series the estimate will differ from the true value of the parameter by at most $3/2\sqrt{n}$. (On this topic see also the end of §5.)

3. Problems

1. Let it be known *a priori* that θ has a value in the set $\Theta_0 \subseteq [0, 1]$. Construct an unbiased estimator for θ, taking values only in Θ_0.

2. Under the hypotheses of the preceding problem, find an analog of the Rao–Cramér inequality and discuss the problem of efficient estimators.

3. Under the hypotheses of the first problem, discuss the construction of confidence intervals for θ.

§8. Conditional Probabilities and Mathematical Expectations with Respect to Decompositions

1. Let $(\Omega, \mathscr{A}, \mathsf{P})$ be a finite probability space and

$$\mathscr{D} = \{D_1, \ldots, D_k\}$$

a decomposition of Ω ($D_i \in \mathscr{A}$, $\mathsf{P}(D_i) > 0$, $i = 1, \ldots, k$, and $D_1 + \cdots + D_k = \Omega$). Also let A be an event from $\mathscr{A}$ and $\mathsf{P}(A|D_i)$ the conditional probability of A with respect to D_i.

With a set of conditional probabilities $\{\mathsf{P}(A|D_i), i = 1, \ldots, k\}$ we may associate the random variable

$$\pi(\omega) = \sum_{i=1}^{k} \mathsf{P}(A|D_i) I_{D_i}(\omega) \tag{1}$$

(cf. (4.5)), that takes the values $\mathsf{P}(A|D_i)$ on the atoms of D_i. To emphasize that this *random variable* is associated specifically with the decomposition $\mathscr{D}$, we denote it by

$$\mathsf{P}(A|\mathscr{D}) \quad \text{or} \quad \mathsf{P}(A|\mathscr{D})(\omega)$$

and call it the *conditional probability of the event A with respect to the decomposition $\mathscr{D}$*.

This concept, as well as the more general concept of conditional probabilities with respect to a σ-algebra, which will be introduced later, plays an important role in probability theory, a role that will be developed progressively as we proceed.

We mention some of the simplest properties of conditional probabilities:

$$\mathsf{P}(A + B|\mathscr{D}) = \mathsf{P}(A|\mathscr{D}) + \mathsf{P}(B|\mathscr{D}); \tag{2}$$

if $\mathscr{D}$ is the trivial decomposition consisting of the single set Ω then

$$\mathsf{P}(A|\Omega) = \mathsf{P}(A). \tag{3}$$

The definition of $\mathsf{P}(A|\mathscr{D})$ as a random variable lets us speak of its expectation; by using this, we can write the *formula* (3.3) *for total probability* in the following compact form:

$$\mathsf{E}\mathsf{P}(A|\mathscr{D}) = \mathsf{P}(A). \tag{4}$$

In fact, since

$$\mathsf{P}(A|\mathscr{D}) = \sum_{i=1}^{k} \mathsf{P}(A|D_i)I_{D_i}(\omega),$$

then by the definition of expectation (see (4.5) and (4.6))

$$\mathsf{E}\mathsf{P}(A|\mathscr{D}) = \sum_{i=1}^{k} \mathsf{P}(A|D_i)\mathsf{P}(D_i) = \sum_{i=1}^{k} \mathsf{P}(AD_i) = \mathsf{P}(A).$$

Now let $\eta = \eta(\omega)$ be a random variable that takes the values $y_1, \dots, y_k$ with positive probabilities:

$$\eta(\omega) = \sum_{j=1}^{k} y_j I_{D_j}(\omega),$$

where $D_j = \{\omega: \eta(\omega) = y_j\}$. The decomposition $\mathscr{D}_\eta = \{D_1, \dots, D_k\}$ is called the decomposition induced by η. The conditional probability $\mathsf{P}(A|\mathscr{D}_\eta)$ will be denoted by $\mathsf{P}(A|\eta)$ or $\mathsf{P}(A|\eta)(\omega)$, and called the *conditional probability of A with respect to the random variable η*. We also denote by $\mathsf{P}(A|\eta = y_j)$ the conditional probability $\mathsf{P}(A|D_j)$, where $D_j = \{\omega: \eta(\omega) = y_j\}$.

Similarly, if $\eta_1, \eta_2, \dots, \eta_m$ are random variables and $\mathscr{D}_{\eta_1, \eta_2, \dots, \eta_m}$ is the decomposition induced by $\eta_1, \eta_2, \dots, \eta_m$ with atoms

$$D_{y_1, y_2, \dots, y_m} = \{\omega: \eta_1(\omega) = y_1, \dots, \eta_m(\omega) = y_m\},$$

then $\mathsf{P}(A|\mathscr{D}_{\eta_1, \eta_2, \dots, \eta_m})$ will be denoted by $\mathsf{P}(A|\eta_1, \eta_2, \dots, \eta_m)$ and called the conditional probability of A with respect to $\eta_1, \eta_2, \dots, \eta_m$.

EXAMPLE 1. Let ξ and η be independent identically distributed random variables, each taking the values 1 and 0 with probabilities p and q. For $k = 0, 1, 2$, let us find the conditional probability $\mathsf{P}(\xi + \eta = k|\eta)$ of the event $A = \{\omega: \xi + \eta = k\}$ with respect to η.

To do this, we first notice the following useful general fact: if ξ and η are independent random variables with respective values x and y, then

$$\mathsf{P}(\xi + \eta = z \,|\, \eta = y) = \mathsf{P}(\xi + y = z). \tag{5}$$

In fact,

$$\begin{aligned}\mathsf{P}(\xi + \eta = f \,|\, \eta = y) &= \frac{\mathsf{P}(\xi + \eta = z, \eta = y)}{\mathsf{P}(\eta = y)} \\ &= \frac{\mathsf{P}(\xi + y = z, \eta = y)}{\mathsf{P}(\eta = y)} = \frac{\mathsf{P}(\xi + y = z)\mathsf{P}(y = \eta)}{\mathsf{P}(\eta = y)} \\ &= \mathsf{P}(\xi + y = z).\end{aligned}$$

Using this formula for the case at hand, we find that

$$\begin{aligned}\mathsf{P}(\xi + \eta = k \,|\, \eta) &= \mathsf{P}(\xi + \eta = k \,|\, \eta = 0) I_{\{\eta = 0\}}(\omega) \\ &\quad + \mathsf{P}(\xi + \eta = k \,|\, \eta = 1) I_{\{\eta = 1\}}(\omega) \\ &= \mathsf{P}(\xi = k) I_{\{\eta = 0\}}(\omega) + \mathsf{P}\{\xi = k - 1\} I_{\{\eta = 1\}}(\omega).\end{aligned}$$

Thus

$$\mathsf{P}(\xi + \eta = k \,|\, \eta) = \begin{cases} q I_{\{\eta = 0\}}(\omega), & k = 0, \\ p I_{\{\eta = 0\}}(\omega) + q I_{\{\eta = 1\}}(\omega), & k = 1, \\ p I_{\{\eta = 1\}}(\omega), & k = 2, \end{cases} \tag{6}$$

or equivalently

$$\mathsf{P}(\xi + \eta = k \,|\, \eta) = \begin{cases} q(1 - \eta), & k = 0, \\ p(1 - \eta) + q\eta, & k = 1, \\ p\eta, & k = 2, \end{cases} \tag{7}$$

2. Let $\xi = \xi(\omega)$ be a random variable with values in the set $X = \{x_1, \ldots, x_n\}$:

$$\xi = \sum_{j=1}^{l} x_j I_{A_j}(\omega), \qquad A_j = \{\omega : \xi = x_j\}$$

and let $\mathscr{D} = \{D_1, \ldots, D_k\}$ be a decomposition. Just as we defined the expectation of ξ with respect to the probabilities $\mathsf{P}(A_j)$, $j = 1, \ldots, l$.

$$\mathsf{E}\xi = \sum_{j=1}^{l} x_j \mathsf{P}(A_j), \tag{8}$$

it is now natural to define the *conditional expectation of ξ with respect to $\mathscr{D}$* by using the conditional probabilities $\mathsf{P}(A_j \,|\, \mathscr{D})$, $j = 1, \ldots, l$. We denote this expectation by $\mathsf{E}(\xi \,|\, \mathscr{D})$ or $\mathsf{E}(\xi \,|\, \mathscr{D})(\omega)$, and define it by the formula

$$\mathsf{E}(\xi \,|\, \mathscr{D}) = \sum_{j=1}^{l} x_j \mathsf{P}(A_j \,|\, \mathscr{D}). \tag{9}$$

According to this definition the conditional expectation $\mathsf{E}(\xi \,|\, \mathscr{D})(\omega)$ is a random variable which, at all sample points ω belonging to the same atom

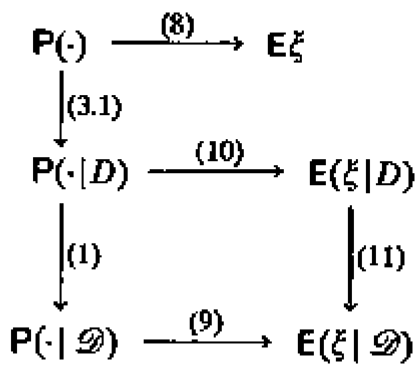


Figure 14

D_i, takes the same value $\sum_{j=1}^{l} x_j \mathsf{P}(A_j|D_i)$. This observation shows that the definition of $\mathsf{E}(\xi|\mathscr{D})$ could have been expressed differently. In fact, we could first define $\mathsf{E}(\xi|D_i)$, the conditional expectation of ξ with respect to D_i, by

$$\mathsf{E}(\xi|D_i) = \sum_{j=1}^{l} x_j \mathsf{P}(A_j|D_i) \left(= \frac{\mathsf{E}[\xi I_{D_i}]}{\mathsf{P}(D_i)} \right), \tag{10}$$

and then define

$$\mathsf{E}(\xi|\mathscr{D})(\omega) = \sum_{i=1}^{k} \mathsf{E}(\xi|D_i) I_{D_i}(\omega) \tag{11}$$

(see the diagram in Figure 14).

It is also useful to notice that $\mathsf{E}(\xi|D)$ and $\mathsf{E}(\xi|\mathscr{D})$ are independent of the representation of ξ.

The following properties of conditional expectations follow immediately from the definitions:

$$\mathsf{E}(a\xi + b\eta|\mathscr{D}) = a\mathsf{E}(\xi|\mathscr{D}) + b\mathsf{E}(\eta|\mathscr{D}), \qquad a \text{ and } b \text{ constants}; \tag{12}$$

$$\mathsf{E}(\xi|\Omega) = \mathsf{E}\xi; \tag{13}$$

$$\mathsf{E}(C|\mathscr{D}) = C, \qquad C \text{ constant}; \tag{14}$$

if $\xi = I_A(\omega)$ then

$$\mathsf{E}(\xi|\mathscr{D}) = \mathsf{P}(A|\mathscr{D}). \tag{15}$$

The last equation shows, in particular, that properties of conditional probabilities can be deduced directly from properties of conditional expectations.

The following important property generalizes the *formula for total probability* (5):

$$\mathsf{E}\mathsf{E}(\xi|\mathscr{D}) = \mathsf{E}\xi. \tag{16}$$

For the proof, it is enough to notice that by (5)

$$\mathsf{E}\mathsf{E}(\xi|\mathscr{D}) = \mathsf{E}\sum_{j=1}^{l} x_j P(A_j|\mathscr{D}) = \sum_{j=1}^{l} x_j \mathsf{E}\mathsf{P}(A_j|\mathscr{D}) = \sum_{j=1}^{l} x_j \mathsf{P}(A_j) = \mathsf{E}\xi.$$

Let $\mathscr{D} = \{D_1, \ldots, D_k\}$ be a decomposition and $\eta = \eta(\omega)$ a random variable. We say that η is measurable with respect to this decomposition, or $\mathscr{D}$-measurable, if $\mathscr{D}_\eta \preccurlyeq \mathscr{D}$, i.e. $\eta = \eta(\omega)$ can be represented in the form

$$\eta(\omega) = \sum_{i=1}^{k} y_i I_{D_i}(\omega),$$

where some y_i might be equal. In other words, a random variable is $\mathscr{D}$-measurable if and only if it takes constant values on the atoms of $\mathscr{D}$.

Example 2. If $\mathscr{D}$ is the trivial decomposition, $\mathscr{D} = \{\Omega\}$, then η is $\mathscr{D}$-measurable if and only if $\eta \equiv C$, where C is a constant. Every random variable η is measurable with respect to $\mathscr{D}_\eta$.

Suppose that the random variable η is $\mathscr{D}$-measurable. Then

$$\mathsf{E}(\xi\eta \mid \mathscr{D}) = \eta \mathsf{E}(\xi \mid \mathscr{D}) \tag{17}$$

and in particular

$$\mathsf{E}(\eta \mid \mathscr{D}) = \eta \qquad (\mathsf{E}(\eta \mid \mathscr{D}_\eta) = \eta). \tag{18}$$

To establish (17) we observe that if $\xi = \sum_{j=1}^{l} x_j I_{A_j}$, then

$$\xi\eta = \sum_{j=1}^{l} \sum_{i=1}^{k} x_j y_i I_{A_j D_i}$$

and therefore

$$\begin{aligned}
\mathsf{E}(\xi\eta \mid \mathscr{D}) &= \sum_{j=1}^{l} \sum_{i=1}^{k} x_j y_i \mathsf{P}(A_j D_i \mid \mathscr{D}) \\
&= \sum_{j=1}^{l} \sum_{i=1}^{k} x_j y_i \sum_{m=1}^{k} \mathsf{P}(A_j D_i \mid D_m) I_{D_m}(\omega) \\
&= \sum_{j=1}^{l} \sum_{i=1}^{k} x_j y_i \mathsf{P}(A_j D_i \mid D_i) I_{D_i}(\omega) \\
&= \sum_{j=1}^{l} \sum_{i=1}^{k} x_j y_i \mathsf{P}(A_j \mid D_i) I_{D_i}(\omega).
\end{aligned} \tag{19}$$

On the other hand, since $I_{D_i}^2 = I_{D_i}$ and $I_{D_i} \cdot I_{D_m} = 0$, $i \neq m$, we obtain

$$\begin{aligned}
\eta \mathsf{E}(\xi \mid \mathscr{D}) &= \left[\sum_{i=1}^{k} y_i I_{D_i}(\omega)\right] \cdot \left[\sum_{j=1}^{l} x_j \mathsf{P}(A_j \mid \mathscr{D})\right] \\
&= \left[\sum_{i=1}^{k} y_i I_{D_i}(\omega)\right] \cdot \sum_{m=1}^{k} \left[\sum_{j=1}^{l} x_j \mathsf{P}(A_j \mid D_m)\right] \cdot I_{D_m}(\omega) \\
&= \sum_{i=1}^{k} \sum_{j=1}^{l} y_i x_j \mathsf{P}(A_j \mid D_i) \cdot I_{D_i}(\omega),
\end{aligned}$$

which, with (19), establishes (17).

We shall establish another important property of conditional expectations. Let $\mathscr{D}_1$ and $\mathscr{D}_2$ be two decompositions, with $\mathscr{D}_1 \preccurlyeq \mathscr{D}_2$ ($\mathscr{D}_2$ is "finer" than $\mathscr{D}_1$). Then

$$\mathsf{E}[\mathsf{E}(\xi \mid \mathscr{D}_2) \mid \mathscr{D}_1] = \mathsf{E}(\xi \mid \mathscr{D}_1). \tag{20}$$

For the proof, suppose that

$$\mathscr{D}_1 = \{D_{11}, \ldots, D_{1m}\}, \qquad \mathscr{D}_2 = \{D_{21}, \ldots, D_{2n}\}.$$

Then if $\xi = \sum_{j=1}^{l} x_j I_{A_j}$, we have

$$\mathsf{E}(\xi \mid \mathscr{D}_2) = \sum_{j=1}^{l} x_j \mathsf{P}(A_j \mid \mathscr{D}_2),$$

and it is sufficient to establish that

$$\mathsf{E}[\mathsf{P}(A_j \mid \mathscr{D}_2) \mid \mathscr{D}_1] = \mathsf{P}(A_j \mid \mathscr{D}_1). \tag{21}$$

Since

$$\mathsf{P}(A_j \mid \mathscr{D}_2) = \sum_{q=1}^{n} \mathsf{P}(A_j \mid D_{2q}) I_{D_{2q}},$$

we have

$$\begin{aligned}
\mathsf{E}[\mathsf{P}(A_j \mid \mathscr{D}_2) \mid \mathscr{D}_1] &= \sum_{q=1}^{n} \mathsf{P}(A_j \mid D_{2q}) \mathsf{P}(D_{2q} \mid \mathscr{D}_1) \\
&= \sum_{q=1}^{n} \mathsf{P}(A_j \mid D_{2q}) \left[\sum_{p=1}^{m} \mathsf{P}(D_{2q} \mid D_{1p}) I_{D_{1p}} \right] \\
&= \sum_{p=1}^{m} I_{D_{1p}} \cdot \sum_{q=1}^{n} \mathsf{P}(A_j \mid D_{2q}) \mathsf{P}(D_{2q} \mid D_{1p}) \\
&= \sum_{p=1}^{m} I_{D_{1p}} \cdot \sum_{\{q:\, D_{2q} \subseteq D_{1p}\}} \mathsf{P}(A_j \mid D_{2q}) \mathsf{P}(D_{2q} \mid D_{1p}) \\
&= \sum_{p=1}^{m} I_{D_{1p}} \cdot \sum_{\{q:\, D_{2q} \subseteq D_{1p}\}} \frac{\mathsf{P}(A_j D_{2q})}{\mathsf{P}(D_{2q})} \cdot \frac{\mathsf{P}(D_{2q})}{\mathsf{P}(D_{1p})} \\
&= \sum_{p=1}^{m} I_{D_{1p}} \cdot \mathsf{P}(A_j \mid D_{1p}) = \mathsf{P}(A_j \mid \mathscr{D}_1),
\end{aligned}$$

which establishes (21).

When $\mathscr{D}$ is induced by the random variables $\eta_1, \ldots, \eta_k$ ($\mathscr{D} = \mathscr{D}_{\eta_1, \ldots, \eta_k}$), the conditional expectation $\mathsf{E}(\xi \mid \mathscr{D}_{\eta_1, \ldots, \eta_k})$ is denoted by $\mathsf{E}(\xi \mid \eta_1, \ldots, \eta_k)$, or $\mathsf{E}(\xi \mid \eta_1, \ldots, \eta_k)(\omega)$, and is called the *conditional expectation of* ξ *with respect to* $\eta_1, \ldots, \eta_k$.

It follows immediately from the definition of $\mathsf{E}(\xi \mid \eta)$ that if ξ and η are *independent*, then

$$\mathsf{E}(\xi \mid \eta) = \mathsf{E}\xi. \tag{22}$$

From (18) it also follows that

$$\mathsf{E}(\eta \mid \eta) = \eta. \tag{23}$$

Property (22) admits the following generalization. Let ξ be independent of $\mathscr{D}$ (i.e. for each $D_i \in \mathscr{D}$ the random variables ξ and I_{D_i} are independent). Then

$$\mathsf{E}(\xi \mid \mathscr{D}) = \mathsf{E}\xi. \tag{24}$$

As a special case of (20) we obtain the following useful formula:

$$\mathsf{E}[\mathsf{E}(\xi \mid \eta_1, \eta_2) \mid \eta_1] = \mathsf{E}(\xi \mid \eta_1). \tag{25}$$

EXAMPLE 3. Let us find $\mathsf{E}(\xi + \eta \mid \eta)$ for the random variables ξ and η considered in Example 1. By (22) and (23),

$$\mathsf{E}(\xi + \eta \mid \eta) = \mathsf{E}\xi + \eta = p + \eta.$$

This result can also be obtained by starting from (8):

$$\mathsf{E}(\xi + \eta \mid \eta) = \sum_{k=0}^{2} k\mathsf{P}(\xi + \eta = k \mid \eta) = p(1 - \eta) + q\eta + 2p\eta = p + \eta.$$

EXAMPLE 4. Let ξ and η be independent and identically distributed random variables. Then

$$\mathsf{E}(\xi \mid \xi + \eta) = \mathsf{E}(\eta \mid \xi + \eta) = \frac{\xi + \eta}{2}. \tag{26}$$

In fact, if we assume for simplicity that ξ and η take the values $1, 2, \ldots, m$, we find $(1 \le k \le m, 2 \le l \le 2m)$

$$\begin{aligned}\mathsf{P}(\xi = k \mid \xi + \eta = l) &= \frac{\mathsf{P}(\xi = k, \xi + \eta = l)}{\mathsf{P}(\xi + \eta = l)} = \frac{\mathsf{P}(\xi = k, \eta = l - k)}{\mathsf{P}(\xi + \eta = l)} \\ &= \frac{\mathsf{P}(\xi = k)P(\eta = l - k)}{\mathsf{P}(\xi + \eta = l)} = \frac{\mathsf{P}(\eta = k)\mathsf{P}(\xi = l - k)}{\mathsf{P}(\xi + \eta = l)} \\ &= \mathsf{P}(\eta = k \mid \xi + \eta = l).\end{aligned}$$

This establishes the first equation in (26). To prove the second, it is enough to notice that

$$2\mathsf{E}(\xi \mid \xi + \eta) = \mathsf{E}(\xi \mid \xi + \eta) + \mathsf{E}(\eta \mid \xi + \eta) = \mathsf{E}(\xi + \eta \mid \xi + \eta) = \xi + \eta.$$

3. We have already noticed in §1 that to each decomposition $\mathscr{D} = \{D_1, \ldots, D_k\}$ of the finite set Ω there corresponds an algebra $\alpha(\mathscr{D})$ of subsets of Ω. The converse is also true: every algebra $\mathscr{B}$ of subsets of the finite space Ω generates a decomposition $\mathscr{D}$ ($\mathscr{B} = \alpha(\mathscr{D})$). Consequently there is a one-to-one correspondence between algebras and decompositions of a finite space Ω. This should be kept in mind in connection with the concept, which will be introduced later, of conditional expectation with respect to the special systems of sets called σ-algebras.

For finite spaces, the concepts of algebra and σ-algebra coincide. It will turn out that if $\mathscr{B}$ is an algebra, the conditional expectation $\mathsf{E}(\xi|\mathscr{B})$ of a random variable ξ with respect to $\mathscr{B}$ (to be introduced in §7 of Chapter II) simply coincides with $\mathsf{E}(\xi|\mathscr{D})$, the expectation of ξ with respect to the decomposition $\mathscr{D}$ such that $\mathscr{B} = \alpha(\mathscr{D})$. In this sense we can, in dealing with finite spaces in the future, not distinguish between $\mathsf{E}(\xi|\mathscr{B})$ and $\mathsf{E}(\xi|\mathscr{D})$, understanding in each case that $\mathsf{E}(\xi|\mathscr{B})$ is simply defined to be $\mathsf{E}(\xi|\mathscr{D})$.

4. Problems

1. Give an example of random variables ξ and η which are not independent but for which

$$\mathsf{E}(\xi|\eta) = \mathsf{E}\xi.$$

(Cf. (22).)

2. The conditional variance of ξ with respect to $\mathscr{D}$ is the random variable

$$\mathsf{V}(\xi|\mathscr{D}) = \mathsf{E}[(\xi - \mathsf{E}(\xi|\mathscr{D}))^2|\mathscr{D}].$$

Show that

$$\mathsf{V}\xi = \mathsf{E}\mathsf{V}(\xi|\mathscr{D}) + \mathsf{V}\mathsf{E}(\xi|\mathscr{D}).$$

3. Starting from (17), show that for every function $f = f(\eta)$ the conditional expectation $\mathsf{E}(\xi|\eta)$ has the property

$$\mathsf{E}[f(\eta)\mathsf{E}(\xi|\eta)] = \mathsf{E}[\xi f(\eta)].$$

4. Let ξ and η be random variables. Show that $\inf_f \mathsf{E}(\eta - f(\xi))^2$ is attained for $f^*(\xi) = \mathsf{E}(\eta|\xi)$. (Consequently, the best estimator for η in terms of ξ, in the mean-square sense, is the conditional expectation $\mathsf{E}(\eta|\xi)$).

5. Let $\xi_1, \ldots, \xi_n, \tau$ be independent random variables, where $\xi_1, \ldots, \xi_n$ are identically distributed and τ takes the values $1, 2, \ldots, n$. Show that if $S_\tau = \xi_1 + \cdots + \xi_\tau$ is the sum of a random number of the random variables,

$$\mathsf{E}(S_\tau|\tau) = \tau\mathsf{E}\xi_1, \qquad \mathsf{V}(S_\tau|\tau) = \tau\mathsf{V}\xi_1$$

and

$$\mathsf{E}S_\tau = \mathsf{E}\tau \cdot \mathsf{E}\xi_1, \qquad \mathsf{V}S_\tau = \mathsf{E}\tau \cdot \mathsf{V}\xi_1 + \mathsf{V}\tau \cdot (\mathsf{E}\xi_1)^2.$$

6. Establish equation (24).

§9. Random Walk. I. Probabilities of Ruin and Mean Duration in Coin Tossing

1. The value of the limit theorems of §6 for Bernoulli schemes is not just that they provide convenient formulas for calculating probabilities $\mathsf{P}(S_n = k)$ and $\mathsf{P}(A < S_n \leq B)$. They have the additional significance of being of a universal nature, i.e. they remain useful not only for independent Bernoulli random variables that have only two values, but also for variables of much more general character. In this sense the Bernoulli scheme appears as the simplest model, on the basis of which we can recognize many probabilistic regularities which are inherent also in much more general models.

In this and the next section we shall discuss a number of new probabilistic regularities, some of which are quite surprising. The ones that we discuss are again based on the Bernoulli scheme, although many results on the nature of random oscillations remain valid for random walks of a more general kind.

2. Consider the Bernoulli scheme $(\Omega, \mathscr{A}, \mathsf{P})$, where $\Omega = \{\omega: \omega = (x_1, \ldots, x_n), x_i = \pm 1\}$, $\mathscr{A}$ consists of all subsets of Ω, and $p(\omega) = p^{\nu(\omega)}q^{n-\nu(\omega)}$, $\nu(\omega) = (\sum x_i + n)/2$. Let $\xi_i(\omega) = x_i$, $i = 1, \ldots, n$. Then, as we know, the sequence $\xi_1, \ldots, \xi_n$ is a sequence of independent Bernoulli random variables,

$$\mathsf{P}(\xi_i = 1) = p, \qquad \mathsf{P}(\xi_i = -1) = q, \qquad p + q = 1.$$

Let us put $S_0 = 0$, $S_k = \xi_1 + \cdots + \xi_k$, $1 \le k \le n$. The sequence $S_0, S_1, \ldots, S_n$ can be considered as the path of the random motion of a particle starting at zero. Here $S_{k+1} = S_k + \xi_k$, i.e. if the particle has reached the point S_k at time k, then at time $k + 1$ it is displaced either one unit up (with probability p) or one unit down (with probability q).

Let A and B be integers, $A \le 0 \le B$. An interesting problem about this random walk is to find the probability that after n steps the moving particle has left the interval (A, B). It is also of interest to ask with what probability the particle leaves (A, B) at A or at B.

That these are natural questions to ask becomes particularly clear if we interpret them in terms of a gambling game. Consider two players (first and second) who start with respective bankrolls $(-A)$ and B. If $\xi_i = +1$, we suppose that the second player pays one unit to the first; if $\xi_i = -1$, the first pays the second. Then $S_k = \xi_1 + \cdots + \xi_k$ can be interpreted as the amount won by the first player from the second (if $S_k < 0$, this is actually the amount lost by the first player to the second) after k turns.

At the instant $k \le n$ at which for the first time $S_k = B$ ($S_k = A$) the bankroll of the second (first) player is reduced to zero; in other words, that player is ruined. (If $k < n$, we suppose that the game ends at time k, although the random walk itself is well defined up to time n, inclusive.)

Before we turn to a precise formulation, let us introduce some notation. Let x be an integer in the interval $[A, B]$ and for $0 \le k \le n$ let $S_k^x = x + S_k$,

$$\tau_k^x = \min\{0 \le l \le k: S_l^x = A \text{ or } B\}, \tag{1}$$

where we agree to take $\tau_k^x = k$ if $A < S_l^x < B$ for all $0 \le l \le k$.

For each k in $0 \le k \le n$ and $x \in [A, B]$, the instant τ_k^x, called a *stopping time* (see §11), is an integer-valued random variable defined on the sample space Ω (the dependence of τ_k^x on Ω is not explicitly indicated).

It is clear that for all $l < k$ the set $\{\omega: \tau_k^x = l\}$ is the event that the random walk $\{S_i^x, 0 \le i \le k\}$, starting at time zero at the point x, leaves the interval (A, B) at time l. It is also clear that when $l \le k$ the sets $\{\omega: \tau_k^x = l, S_l^x = A\}$ and $\{\omega: \tau_k^x = l, S_l^x = B\}$ represent the events that the wandering particle leaves the interval (A, B) at time l through A or B respectively.

For $0 \le k \le n$, we write

$$\mathscr{A}_k^x = \sum_{0 \le l \le k} \{\omega: \tau_k^x = l, S_l^x = A\},$$
$$\mathscr{B}_k^x = \sum_{0 \le l \le k} \{\omega: \tau_k^x = l, S_l^x = B\}, \tag{2}$$

and let

$$\alpha_k(x) = \mathsf{P}(\mathscr{A}_k^x), \qquad \beta_k(x) = \mathsf{P}(\mathscr{B}_k^x)$$

be the probabilities that the particle leaves (A, B), through A or B respectively, during the time interval $[0, k]$. For these probabilities we can find recurrent relations from which we can successively determine $\alpha_1(x), \ldots, \alpha_n(x)$ and $\beta_1(x), \ldots, \beta_n(x)$.

Let, then, $A < x < B$. It is clear that $\alpha_0(x) = \beta_0(x) = 0$. Now suppose $1 \le k \le n$. Then by (8.5),

$$\begin{aligned}\beta_k(x) = \mathsf{P}(\mathscr{B}_k^x) &= \mathsf{P}(\mathscr{B}_k^x | S_1^x = x + 1)\mathsf{P}(\xi_1 = 1) \\ &\quad + \mathsf{P}(\mathscr{B}_k^x | S_1^x = x - 1)\mathsf{P}(\xi_1 = -1) \\ &= p\mathsf{P}(\mathscr{B}_k^x | S_1^x = x + 1) + q\mathsf{P}(\mathscr{B}_k^x | S_1^x = x - 1).\end{aligned} \tag{3}$$

We now show that

$$\mathsf{P}(\mathscr{B}_k^x | S_1^x = x + 1) = \mathsf{P}(\mathscr{B}_{k-1}^{x+1}), \qquad \mathsf{P}(\mathscr{B}_k^x | S_1^x = x - 1) = \mathsf{P}(\mathscr{B}_{k-1}^{x-1}).$$

To do this, we notice that $\mathscr{B}_k^x$ can be represented in the form

$$\mathscr{B}_k^x = \{\omega: (x, x + \xi_1, \ldots, x + \xi_1 + \cdots + \xi_k) \in B_k^x\},$$

where B_k^x is the set of paths of the form

$$(x, x + x_1, \ldots, x + x_1 + \cdots x_k)$$

with $x_1 = \pm 1$, which during the time $[0, k]$ first leave (A, B) at B (Figure 15).

We represent B_k^x in the form $B_k^{x,x+1} + B_k^{x,x-1}$, where $B_k^{x,x+1}$ and $B_k^{x,x-1}$ are the paths in B_k^x for which $x_1 = +1$ or $x_1 = -1$, respectively.

Notice that the paths $(x, x + 1, x + 1 + x_2, \ldots, x + 1 + x_2 + \cdots + x_k)$ in $B_k^{x,x+1}$ are in one-to-one correspondence with the paths

$$(x + 1, x + 1 + x_2, \ldots, x + 1 + x_2, \ldots, x + 1 + x_2 + \cdots + x_k)$$

in B_{k-1}^{x+1}. The same is true for the paths in $B_k^{x,x-1}$. Using these facts, together with independence, the identical distribution of $\xi_1, \ldots, \xi_k$, and (8.6), we obtain

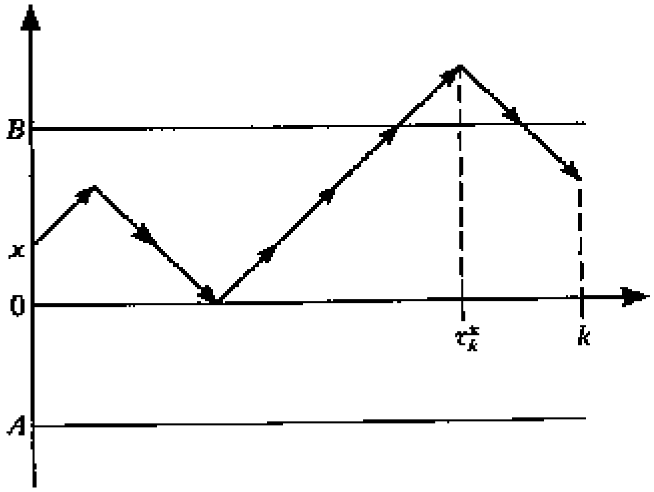


Figure 15. Example of a path from the set B_k^x.

$$\begin{aligned}
\mathsf{P}(\mathscr{B}_k^x | S_1^x = x + 1)\\
&= \mathsf{P}(\mathscr{B}_k^x | \xi_1 = 1)\\
&= \mathsf{P}\{(x, x + \xi_1, \ldots, x + \xi_1 + \cdots + \xi_k) \in B_k^x | \xi_1 = 1\}\\
&= \mathsf{P}\{(x + 1, x + 1 + \xi_2, \ldots, x + 1 + \xi_2 + \cdots + \xi_k) \in B_{k-1}^{x+1}\}\\
&= \mathsf{P}\{(x + 1, x + 1 + \xi_1, \ldots, x + 1 + \xi_1 + \cdots + \xi_{k-1}) \in B_{k-1}^{x+1}\}\\
&= \mathsf{P}(\mathscr{B}_{k-1}^{x+1}).
\end{aligned}$$

In the same way,

$$\mathsf{P}(\mathscr{B}_k^x | S_1^x = x - 1) = \mathsf{P}(\mathscr{B}_{k-1}^{x-1}).$$

Consequently, by (3) with $x \in (A, B)$ and $k \leq n$,

$$\beta_k(x) = p\beta_{k-1}(x + 1) + q\beta_{k-1}(x - 1), \tag{4}$$

where

$$\beta_l(B) = 1, \qquad \beta_l(A) = 0, \qquad 0 \leq l \leq n. \tag{5}$$

Similarly

$$\alpha_k(x) = p\alpha_{k-1}(x + 1) + q\alpha_{k-1}(x - 1) \tag{6}$$

with

$$\alpha_l(A) = 1, \qquad \alpha_l(B) = 0, \qquad 0 \leq l \leq n.$$

Since $\alpha_0(x) = \beta_0(x) = 0$, $x \in (A, B)$, these recurrent relations can (at least in principle) be solved for the probabilities

$$\alpha_1(x), \ldots, \alpha_n(x) \quad \text{and} \quad \beta_1(x), \ldots, \beta_n(x).$$

Putting aside any explicit calculation of the probabilities, we ask for their values for large n.

For this purpose we notice that since $\mathscr{B}_{k-1}^x \subset \mathscr{B}_k^x$, $k \leq n$, we have $\beta_{k-1}(x) \leq \beta_k(x) \leq 1$. It is therefore natural to expect (and this is actually the case; see Subsection 3) that for sufficiently large n the probability $\beta_n(x)$ will be close to the solution $\beta(x)$ of the equation

$$\beta(x) = p\beta(x+1) + q\beta(x-1) \tag{7}$$

with the boundary conditions

$$\beta(B) = 1, \qquad \beta(A) = 0, \tag{8}$$

that result from a formal approach to the limit in (4) and (5).

To solve the problem in (7) and (8), we first suppose that $p \neq q$. We see easily that the equation has the two particular solutions a and $b(q/p)^x$, where a and b are constants. Hence we look for a solution of the form

$$\beta(x) = a + b(q/p)^x. \tag{9}$$

Taking account of (8), we find that for $A \leq x \leq B$

$$\beta(x) = \frac{(q/p)^x - (q/p)^A}{(q/p)^B - (q/p)^A}. \tag{10}$$

Let us show that this is the *only* solution of our problem. It is enough to show that all solutions of the problem in (7) and (8) admit the representation (9).

Let $\tilde{\beta}(x)$ be a solution with $\tilde{\beta}(A) = 0$, $\tilde{\beta}(B) = 1$. We can always find constants $\tilde{a}$ and $\tilde{b}$ such that

$$\tilde{a} + \tilde{b}(q/p)^A = \tilde{\beta}(A), \qquad \tilde{a} + \tilde{b}(q/p)^{A+1} = \tilde{\beta}(A+1).$$

Then it follows from (7) that

$$\tilde{\beta}(A+2) = \tilde{a} + \tilde{b}(q/p)^{A+2}$$

and generally

$$\tilde{\beta}(x) = \tilde{a} + \tilde{b}(q/p)^x. \; .$$

Consequently the solution (10) is the only solution of our problem.

A similar discussion shows that the only solution of

$$\alpha(x) = p\alpha(x+1) + q\alpha(x-1), \qquad x \in (A, B) \tag{11}$$

with the boundary conditions

$$\alpha(A) = 1, \qquad \alpha(B) = 0 \tag{12}$$

is given by the formula

$$\alpha(x) = \frac{(p/q)^B - (q/p)^x}{(p/q)^B - (p/p)^A}, \qquad A \leq x \leq B. \tag{13}$$

If $p = q = \frac{1}{2}$, the only solutions $\beta(x)$ and $\alpha(x)$ of (7), (8) and (11), (12) are respectively

$$\beta(x) = \frac{x - A}{B - A} \tag{14}$$

and

$$\alpha(x) = \frac{B - x}{B - A}. \tag{15}$$

We note that

$$\alpha(x) + \beta(x) = 1 \tag{16}$$

for $0 \leq p \leq 1$.

We call $\alpha(x)$ and $\beta(x)$ the *probabilities of ruin for the first and second players*, respectively (when the first player's bankroll is $x - A$, and the second player's is $B - x$) under the assumption of infinitely many turns, which of course presupposes an infinite sequence of independent Bernoulli random variables $\xi_1, \xi_2, \ldots$, where $\xi_i = +1$ is treated as a gain for the first player, and $\xi_i = -1$ as a loss. The probability space $(\Omega, \mathscr{A}, \mathsf{P})$ considered at the beginning of this section turns out to be too small to allow such an infinite sequence of independent variables. We shall see later that such a sequence can actually be constructed and that $\beta(x)$ and $\alpha(x)$ are in fact the probabilities of ruin in an unbounded number of steps.

We now take up some corollaries of the preceding formulas.

If we take $A = 0, 0 \leq x \leq B$, then the definition of $\beta(x)$ implies that this is the probability that a particle starting at x arrives at B before it reaches 0. It follows from (10) and (14) (Figure 16) that

$$\beta(x) = \begin{cases} x/B, & p = q = \frac{1}{2}, \\ \dfrac{(q/p)^x - 1}{(q/p)^B - 1}, & p \neq q. \end{cases} \tag{17}$$

Now let $q > p$, which means that the game is unfavorable for the first player, whose limiting probability of being ruined, namely $\alpha = \alpha(0)$, is given by

$$\alpha = \frac{(q/p)^B - 1}{(q/p)^B - (q/p)^A}.$$

Next suppose that the rules of the game are changed: the original bankrolls of the players are still $(-A)$ and B, but the payoff for each player is now $\frac{1}{2}$,

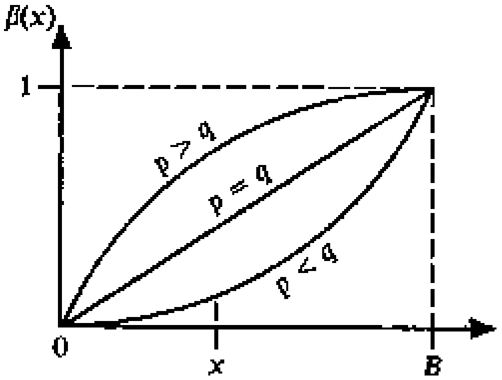


Figure 16. Graph of $\beta(x)$, the probability that a particle starting from x reaches B before reaching 0.

rather than 1 as before. In other words, now let $\mathsf{P}(\xi_i = \frac{1}{2}) = p$, $\mathsf{P}(\xi_i = -\frac{1}{2}) = q$. In this case let us denote the limiting probability of ruin for the first player by $\alpha_{1/2}$. Then

$$\alpha_{1/2} = \frac{(q/p)^{2B} - 1}{(q/p)^{2B} - (q/p)^{2A}},$$

and therefore

$$\alpha_{1/2} = \alpha \cdot \frac{(q/p)^B + 1}{(q/p)^B + (q/p)^A} < \alpha,$$

if $q > p$.

Hence we can draw the following conclusion: *if the game is unfavorable to the first player (i.e., $q > p$) then doubling the stake decreases the probability of ruin.*

3. We now turn to the question of how fast $\alpha_n(x)$ and $\beta_n(x)$ approach their limiting values $\alpha(x)$ and $\beta(x)$.

Let us suppose for simplicity that $x = 0$ and put

$$\alpha_n = \alpha_n(0), \qquad \beta_n = \beta_n(0), \qquad \gamma_n = 1 - (\alpha_n + \beta_n).$$

It is clear that

$$\gamma_n = \mathsf{P}\{A < S_k < B, 0 \le k \le n\},$$

where $\{A < S_k < B, 0 \le k \le n\}$ denotes the event

$$\bigcap_{0 \le k \le n} \{A < S_k < B\}.$$

Let $n = rm$, where r and m are integers and

$$\begin{aligned}
\zeta_1 &= \xi_1 + \cdots + \xi_m,\\
\zeta_2 &= \xi_{m+1} + \cdots + \xi_{2m},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
\zeta_r &= \xi_{m(r-1)+1} + \cdots + \xi_{rm}.
\end{aligned}$$

Then if $C = |A| + B$, it is easy to see that

$$\{A < S_k < B, 1 \le k \le rm\} \subseteq \{|\zeta_1| < C, \ldots, |\zeta_r| < C\},$$

and therefore, since $\zeta_1, \ldots, \zeta_r$ are independent and identically distributed,

$$\gamma_n \le \mathsf{P}\{|\zeta_1| < C, \ldots, |\zeta_r| < C\} = \prod_{i=1}^{r} \mathsf{P}\{|\zeta_i| < C\} = (\mathsf{P}\{|\zeta_1| < C\})^r. \quad (18)$$

We notice that $\mathsf{V}\zeta_1 = m[1 - (p - q)^2]$. Hence, for $0 < p < 1$ and sufficiently large m,

$$\mathsf{P}\{|\zeta_1| < C\} \le \varepsilon_1, \quad (19)$$

where $\varepsilon_1 < 1$, since $\mathsf{V}\zeta_1 \le C^2$ if $\mathsf{P}\{|\zeta_1| \le C\} = 1$.

If $p = 0$ or $p = 1$, then $\mathsf{P}\{|\zeta_1| < C\} = 0$ for sufficiently large m, and consequently (19) is satisfied for $0 \le p \le 1$.

It follows from (18) and (19) that for sufficiently large n

$$\gamma_n \le \varepsilon^n, \tag{20}$$

where $\varepsilon = \varepsilon_1^{1/m} < 1$.

According to (16), $\alpha + \beta = 1$. Therefore

$$(\alpha - \alpha_n) + (\beta - \beta_n) = \gamma_n,$$

and since $\alpha \ge \alpha_n$, $\beta \ge \beta_n$, we have

$$0 \le \alpha - \alpha_n \le \gamma_n \le \varepsilon^n,$$
$$0 \le \beta - \beta_n \le \gamma_n \le \varepsilon^n, \qquad \varepsilon < 1.$$

There are similar inequalities for the differences $\alpha(x) - \alpha_n(x)$ and $\beta(x) - \beta_n(x)$.

4. We now consider the question of the *mean duration* of the random walk.

Let $m_k(x) = \mathsf{E}\tau_k^x$ be the expectation of the stopping time $\tau_k^x, k \le n$. Proceeding as in the derivation of the recurrent relations for $\beta_k(x)$, we find that, for $x \in (A, B)$,

$$\begin{aligned}
m_k(x) = \mathsf{E}\tau_k^x &= \sum_{1 \le l \le k} l\mathsf{P}(\tau_k^x = l) \\
&= \sum_{1 \le l \le k} l \cdot [p\mathsf{P}(\tau_k^x = l \mid \xi_1 = 1) + q\mathsf{P}(\tau_k^x = l \mid \xi_1 = -1)] \\
&= \sum_{1 \le l \le k} l \cdot [p\mathsf{P}(\tau_{k-1}^{x+1} = l - 1) + q\mathsf{P}(\tau_{k-1}^{x-1} = l - 1)] \\
&= \sum_{0 \le l \le k-1} (l + 1)[p\mathsf{P}(\tau_{k-1}^{x+1} = l) + q\mathsf{P}(\tau_{k-1}^{x-1} = l)] \\
&= pm_{k-1}(x + 1) + qm_{k-1}(x - 1) \\
&\quad + \sum_{0 \le l \le k-1} [p\mathsf{P}(\tau_{k-1}^{x+1} = l) + q\mathsf{P}(\tau_{k-1}^{x-1} = l)] \\
&= pm_{k-1}(x + 1) + qm_{k-1}(x - 1) + 1.
\end{aligned}$$

Thus, for $x \in (A, B)$ and $0 \le k \le n$, the functions $m_k(x)$ satisfy the recurrent relations

$$m_k(x) = 1 + pm_{k-1}(x + 1) + qm_{k-1}(x - 1), \tag{21}$$

with $m_0(x) = 0$. From these equations together with the boundary conditions

$$m_k(A) = m_k(B) = 0, \tag{22}$$

we can successively find $m_1(x), \ldots, m_n(x)$.

Since $m_k(x) \le m_{k+1}(x)$, the limit

$$m(x) = \lim_{n \to \infty} m_n(x)$$

exists, and by (21) it satisfies the equation

$$m(x) = 1 + pm(x + 1) + qm(x - 1) \tag{23}$$

with the boundary conditions

$$m(A) = m(B) = 0. \tag{24}$$

To solve this equation, we first suppose that

$$m(x) < \infty, \qquad x \in (A, B). \tag{25}$$

Then if $p \neq q$ there is a particular solution of the form $x/(q - p)$ and the general solution (see (9)) can be written in the form

$$m(x) = \frac{x}{q - p} + a + b\left(\frac{q}{p}\right)^x.$$

Then by using the boundary conditions $m(A) = m(B) = 0$ we find that

$$m(x) = \frac{1}{p - q}(B\beta(x) + A\alpha(x) - x], \tag{26}$$

where $\beta(x)$ and $\alpha(x)$ are defined by (10) and (13). If $p = q = \frac{1}{2}$, the general solution of (23) has the form

$$m(x) = a + bx - x^2,$$

and since $m(A) = m(B) = 0$ we have

$$m(x) = (B - x)(x - A). \tag{27}$$

It follows, in particular, that if the players start with equal bankrolls ($B = -A$), then

$$m(0) = B^2.$$

If we take $B = 10$, and suppose that each turn takes a second, then the (limiting) time to the ruin of one player is rather long: 100 seconds.

We obtained (26) and (27) under the assumption that $m(x) < \infty, x \in (A, B)$. Let us now show that in fact $m(x)$ is finite for all $x \in (A, B)$. We consider only the case $x = 0$; the general case can be analyzed similarly.

Let $p = q = \frac{1}{2}$. We introduce the random variable S_{τ_n} defined in terms of the sequence $S_0, S_1, \ldots, S_n$ and the stopping time $\tau_n = \tau_n^0$ by the equation

$$S_{\tau_n} = \sum_{k=0}^{n} S_k I_{\{\tau_n = k\}}(\omega). \tag{28}$$

The descriptive meaning of S_{τ_n} is clear: it is the position reached by the random walk at the stopping time τ_n. Thus, if $\tau_n < n$, then $S_{\tau_n} = A$ or B; if $\tau_n = n$, then $A \leq S_{\tau_n} \leq B$.

Let us show that when $p = q = \frac{1}{2}$,

$$\mathsf{E}S_{\tau_n} = 0, \tag{29}$$

$$\mathsf{E}S_{\tau_n}^2 = \mathsf{E}\tau_n. \tag{30}$$

To establish the first equation we notice that

$$\begin{aligned}\mathsf{E}S_{\tau_n} &= \sum_{k=0}^{n} \mathsf{E}[S_k I_{\{\tau_n=k\}}(\omega)] \\ &= \sum_{k=0}^{n} \mathsf{E}[S_n I_{\{\tau_n=k\}}(\omega)] + \sum_{k=0}^{n} \mathsf{E}[(S_k - S_n) I_{\{\tau_n=k\}}(\omega)] \\ &= \mathsf{E}S_n + \sum_{k=0}^{n} \mathsf{E}[(S_k - S_n) I_{\{\tau_n=k\}}(\omega)],\end{aligned} \tag{31}$$

where we evidently have $\mathsf{E}S_n = 0$. Let us show that

$$\sum_{k=0}^{n} \mathsf{E}[(S_k - S_n) I_{\{\tau_n=k\}}(\omega)] = 0.$$

To do this, we notice that $\{\tau_n > k\} = \{A < S_1 < B, \ldots, A < S_k < B\}$ when $0 \le k < n$. The event $\{A < S_1 < B, \ldots, A < S_k < B\}$ can evidently be written in the form

$$\{\omega: (\xi_1, \ldots, \xi_k) \in A_k\}, \tag{32}$$

where A_k is a subset of $\{-1, +1\}^k$. In other words, this set is determined by just the values of $\xi_1, \ldots, \xi_k$ and does not depend on $\xi_{k+1}, \ldots, \xi_n$. Since

$$\{\tau_n = k\} = \{\tau_n > k - 1\} \setminus \{\tau_n > k\},$$

this is also a set of the form (32). It then follows from the independence of $\xi_1, \ldots, \xi_n$ and from Problem 10 of §4 that the random variables $S_n - S_k$ and $I_{\{\tau_n=k\}}$ are independent, and therefore

$$\mathsf{E}[(S_n - S_k) I_{\{\tau_n=k\}}] = \mathsf{E}[S_n - S_k] \cdot \mathsf{E}I_{\{\tau_n=k\}} = 0.$$

Hence we have established (29).

We can prove (30) by the same method:

$$\begin{aligned}\mathsf{E}S_{\tau_n}^2 &= \sum_{k=0}^{n} \mathsf{E}S_k^2 I_{\{\tau_n=k\}} = \sum_{k=0}^{n} \mathsf{E}([S_n + (S_k - S_n)]^2 I_{\{\tau_n=k\}}) \\ &= \sum_{k=0}^{n} [\mathsf{E}S_n^2 I_{\{\tau_n=k\}} + 2\mathsf{E}S_n(S_k - S_n) I_{\{\tau_n=k\}} \\ &\qquad + \mathsf{E}(S_n - S_k)^2 I_{\{\tau_n=k\}}] = \mathsf{E}S_n^2 - \sum_{k=0}^{n} \mathsf{E}(S_n - S_k)^2 I_{\{\tau_n=k\}} \\ &= n - \sum_{k=0}^{n} (n-k)\mathsf{P}(\tau_n = k) = \sum_{k=0}^{n} k\mathsf{P}(\tau_n = k) = \mathsf{E}\tau_n.\end{aligned}$$

Thus we have (29) and (30) when $p = q = \frac{1}{2}$. For general p and q $(p + q = 1)$ it can be shown similarly that

$$\mathsf{E}S_{\tau_n} = (p - q)\cdot\mathsf{E}\tau_n, \tag{33}$$

$$\mathsf{E}[S_{\tau_n} - \tau_n\cdot\mathsf{E}\xi_1]^2 = \mathsf{V}\xi_1\cdot\mathsf{E}\tau_n, \tag{34}$$

where $\mathsf{E}\xi_1 = p - q$, $\mathsf{V}\xi_1 = 1 - (p - q)^2$.

With the aid of the results obtained so far we can now show that $\lim_{n\to\infty} m_n(0) = m(0) < \infty$.

If $p = q = \frac{1}{2}$, then by (30)

$$\mathsf{E}\tau_n \le \max(A^2, B^2). \tag{35}$$

If $p \ne q$, then by (33),

$$\mathsf{E}\tau_n \le \frac{\max(|A|, B)}{|p - q|}, \tag{36}$$

from which it is clear that $m(0) < \infty$.

We also notice that when $p = q = \frac{1}{2}$

$$\mathsf{E}\tau_n = \mathsf{E}S_{\tau_n}^2 = A^2\cdot\alpha_n + B^2\cdot\beta_n + \mathsf{E}[S_n^2 I_{\{A<S_n<B\}}]$$

and therefore

$$A^2\cdot\alpha_n + B^2\cdot\beta_n \le \mathsf{E}\tau_n \le A^2\cdot\alpha_n + B^2\cdot\beta_n + \max(A^2, B^2)\cdot\gamma_n.$$

It follows from this and (20) that as $n \to \infty$, $\mathsf{E}\tau_n$ converges with exponential rapidity to

$$m(0) = A^2\alpha + B^2\beta = A^2\cdot\frac{B}{B - A} - B^2\cdot\frac{A}{B - A} = |AB|.$$

There is a similar result when $p \ne q$:

$$\mathsf{E}\tau_n \to m(0) = \frac{\alpha A + \beta B}{p - q}, \quad \text{exponentially fast.}$$

5. Problems

1. Establish the following generalizations of (33) and (34):

$$\mathsf{E}S_{\tau_n^x}^x = x + (p - q)\mathsf{E}\tau_n^x,$$

$$\mathsf{E}[S_{\tau_n^x} - \tau_n^x\cdot\mathsf{E}\xi_1]^2 = \mathsf{V}\xi_1\cdot\mathsf{E}\tau_n^x.$$

2. Investigate the limits of $\alpha(x)$, $\beta(x)$, and $m(x)$ when the level $A \downarrow -\infty$.

3. Let $p = q = \frac{1}{2}$ in the Bernoulli scheme. What is the order of $\mathsf{E}|S_n|$ for large n?

4. Two players each toss their own symmetric coins, independently. Show that the probability that each has the same number of heads after n tosses is $2^{-2n}\sum_{k=0}^{n}(C_n^k)^2$. Hence deduce the equation $\sum_{k=0}^{n}(C_n^k)^2 = C_{2n}^n$.

Let σ_n be the first time when the number of heads for the first player coincides with the number of heads for the second player (if this happens within n tosses; $\sigma_n = n+1$ if there is no such time). Find $\mathsf{E}\min(\sigma_n, n)$.

§10. Random Walk. II. Reflection Principle. Arcsine Law

1. As in the preceding section, we suppose that $\xi_1, \xi_2, \ldots, \xi_{2n}$ is a sequence of independent identically distributed Bernoulli random variables with

$$\mathsf{P}(\xi_i = 1) = p, \qquad \mathsf{P}(\xi_i = -1) = q,$$

$$S_k = \xi_1 + \cdots + \xi_k, \qquad 1 \le k \le 2n; \qquad S_0 = 0.$$

We define

$$\sigma_{2n} = \min\{1 \le k \le 2n \colon S_k = 0\},$$

putting $\sigma_{2n} = \infty$ if $S_k \neq 0$ for $1 \le k \le 2n$.

The descriptive meaning of σ_{2n} is clear: it is the time of first return to zero. Properties of this time are studied in the present section, where we assume that the random walk is symmetric, i.e. $p = q = \frac{1}{2}$.

For $0 \le k \le n$ we write

$$u_{2k} = \mathsf{P}(S_{2k} = 0), \qquad f_{2k} = \mathsf{P}(\sigma_{2n} = 2k). \tag{1}$$

It is clear that $u_0 = 1$ and

$$u_{2k} = C_{2k}^k \cdot 2^{-2k}.$$

Our immediate aim is to show that for $1 \le k \le n$ the probability f_{2k} is given by

$$f_{2k} = \frac{1}{2k} u_{2(k-1)}. \tag{2}$$

It is clear that

$$\{\sigma_{2n} = 2k\} = \{S_1 \neq 0, S_2 \neq 0, \ldots, S_{2k-1} \neq 0, S_{2k} = 0\}$$

for $1 \le k \le n$, and by symmetry

$$\begin{aligned} f_{2k} &= \mathsf{P}\{S_1 \neq 0, \ldots, S_{2k-1} \neq 0, S_{2k} = 0\} \\ &= 2\mathsf{P}\{S_1 > 0, \ldots, S_{2k-1} > 0, S_{2k} = 0\}. \end{aligned} \tag{3}$$

A sequence $(S_0, \ldots, S_k)$ is called a *path* of length k; we denote by $L_k(A)$ the number of paths of length k having some specified property A. Then

$$\begin{aligned} f_{2k} &= 2 \sum_{(a_{2k+1}, \ldots, a_n)} L_{2n}(S_1 > 0, \ldots, S_{2k-1} > 0, S_{2k} = 0, \\ &\qquad \text{and } S_{2k+1} = a_{2k+1}, \ldots, S_{2n} = a_{2k+1} + \cdots + a_{2n}) \cdot 2^{-2n} \\ &= 2L_{2k}(S_1 > 0, \ldots, S_{2k-1} > 0, S_{2k} = 0) \cdot 2^{-2k}, \end{aligned} \tag{4}$$

where the summation is over all sets $(a_{2k+1}, \ldots, a_{2n})$ with $a_i = \pm 1$.

Consequently the determination of the probability f_{2k} reduces to calculating the number of paths $L_{2k}(S_1 > 0, \ldots, S_{2k-1} > 0, S_{2k} = 0)$.

Lemma 1. *Let a and b be nonnegative integers, $a - b > 0$ and $k = a + b$. Then*

$$L_k(S_1 > 0, \ldots, S_{k-1} > 0, S_k = a - b) = \frac{a - b}{k} C_k^a. \tag{5}$$

PROOF. In fact,

$$\begin{aligned} &L_k(S_1 > 0, \ldots, S_{k-1} > 0, S_k = a - b) \\ &\quad = L_k(S_1 = 1, S_2 > 0, \ldots, S_{k-1} > 0, S_k = a - b) \\ &\quad = L_k(S_1 = 1, S_k = a - b) - L_k(S_1 = 1, S_k = a - b; \\ &\qquad \text{and } \exists\, i, 2 \le i \le k - 1, \text{ such that } S_i \le 0). \end{aligned} \tag{6}$$

In other words, the number of positive paths $(S_1, S_2, \ldots, S_k)$ that originate at $(1, 1)$ and terminate at $(k, a - b)$ is the same as the total number of paths from $(1, 1)$ to $(k, a - b)$ after excluding the paths that touch or intersect the time axis.*

We now notice that

$$\begin{aligned} L_k(S_1 = 1, S_k = a - b; \exists\, i, 2 \le i \le k - 1, \text{ such that } S_i \le 0) \\ = L_k(S_1 = -1, S_k = a - b), \end{aligned} \tag{7}$$

i.e. the number of paths from $\alpha = (1, 1)$ to $\beta = (k, a - b)$, neither touching nor intersecting the time axis, is equal to the total number of paths that connect $\alpha^* = (1, -1)$ with β. The proof of this statement, known as the *reflection principle*, follows from the easily established one-to-one correspondence between the paths $A = (S_1, \ldots, S_a, S_{a+1}, \ldots, S_k)$ joining α and β, and paths $B = (-S_1, \ldots, -S_a, S_{a+1}, \ldots, S_k)$ joining α^* and β (Figure 17); a is the first point where A and B reach zero.

* A path $(S_1, \ldots, S_k)$ is called *positive* (or nonnegative) if all $S_i > 0$ ($S_i \ge 0$); a path is said to *touch* the time axis if $S_j \ge 0$ or else $S_j \le 0$, for $1 \le j \le k$, and there is an i, $1 \le i \le k$, such that $S_i = 0$; and a path is said to *intersect* the time axis if there are two times i and j such that $S_i > 0$ and $S_j < 0$.

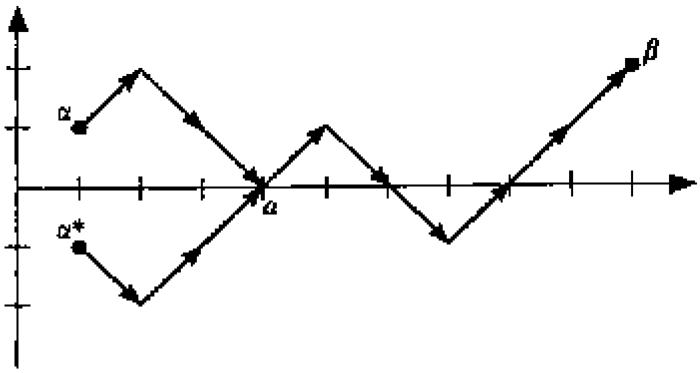


Figure 17. The reflection principle.

From (6) and (7) we find

$$\begin{aligned}
&L_k(S_1 > 0, \ldots, S_{k-1} > 0, S_k = a - b) \\
&\quad = L_k(S_1 = 1, S_k = a - b) - L_k(S_1 = -1, S_k = a - b) \\
&\quad = C_{k-1}^{a-1} - C_{k-1}^{a} = \frac{a-b}{k} C_k^a,
\end{aligned}$$

which establishes (5).

Turning to the calculation of f_{2k}, we find that by (4) and (5) (with $a = k$, $b = k - 1$),

$$\begin{aligned}
f_{2k} &= 2L_{2k}(S_1 > 0, \ldots, S_{2k-1} > 0, S_{2k} = 0) \cdot 2^{-2k} \\
&= 2L_{2k-1}(S_1 > 0, \ldots, S_{2k-1} = 1) \cdot 2^{-2k} \\
&= 2 \cdot 2^{-2k} \cdot \frac{1}{2k-1} C_{2k-1}^{k} = \frac{1}{2k} u_{2(k-1)}.
\end{aligned}$$

Hence (2) is established.

We present an alternative proof of this formula, based on the following observation. A straightforward verification shows that

$$\frac{1}{2k} u_{2(k-1)} = u_{2(k-1)} - u_{2k}, \tag{8}$$

At the same time, it is clear that

$$\begin{aligned}
\{\sigma_{2n} = 2k\} &= \{\sigma_{2n} > 2(k-1)\} \setminus \{\sigma_{2n} > 2k\}, \\
\{\sigma_{2n} > 2l\} &= \{S_1 \neq 0, \ldots, S_{2l} \neq 0\}
\end{aligned}$$

and therefore

$$\{\sigma_{2n} = 2k\} = \{S_1 \neq 0, \ldots, S_{2(k-1)} \neq 0\} \setminus \{S_1 \neq 0, \ldots, S_{2k} \neq 0\}.$$

Hence

$$f_{2k} = \mathsf{P}\{S_1 \neq 0, \ldots, S_{2(k-1)} \neq 0\} - \mathsf{P}\{S_1 \neq 0, \ldots, S_{2k} \neq 0\},$$

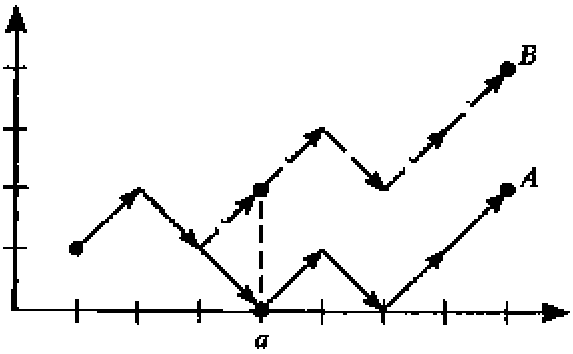


Figure 18

and consequently, because of (8), in order to show that $f_{2k} = (1/2k)u_{2(k-1)}$ it is enough to show only that

$$L_{2k}(S_1 \neq 0, \ldots, S_{2k} \neq 0) = L_{2k}(S_{2k} = 0). \tag{9}$$

For this purpose we notice that evidently

$$L_{2k}(S_1 \neq 0, \ldots, S_{2k} \neq 0) = 2L_{2k}(S_1 > 0, \ldots, S_{2k} > 0).$$

Hence to verify (9) we need only establish that

$$2L_{2k}(S_1 > 0, \ldots, S_{2k} > 0) = L_{2k}(S_1 \geq 0, \ldots, S_{2k} \geq 0) \tag{10}$$

and

$$L_{2k}(S_1 \geq 0, \ldots, S_{2k} \geq 0) = L_{2k}(S_{2k} = 0). \tag{11}$$

Now (10) will be established if we show that we can establish a one-to-one correspondence between the paths $A = (S_1, \ldots, S_{2k})$ for which at least one $S_i = 0$, and the positive paths $B = (S_1, \ldots, S_{2k})$.

Let $A = (S_1, \ldots, S_{2k})$ be a nonnegative path for which the first zero occurs at the point a (i.e., $S_a = 0$). Let us construct the path, starting at $(a, 2)$, $(S_a + 2, S_{a+1} + 2, \ldots, S_{2k} + 2)$ (indicated by the broken lines in Figure 18). Then the path $B = (S_1, \ldots, S_{a-1}, S_a + 2, \ldots, S_{2k} + 2)$ is positive.

Conversely, let $B = (S_1, \ldots, S_{2k})$ be a positive path and b the last instant at which $S_b = 1$ (Figure 19). Then the path

$$A = (S_1, \ldots, S_b, S_{b+1} - 2, \ldots, S_k - 2)$$

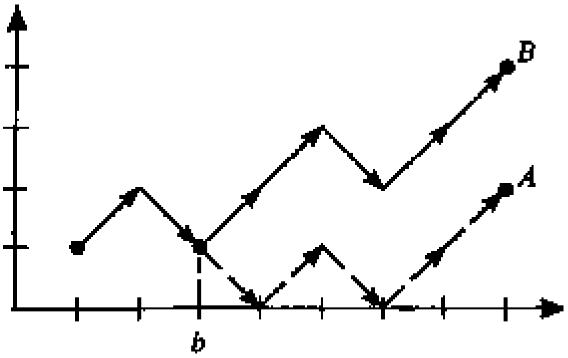


Figure 19

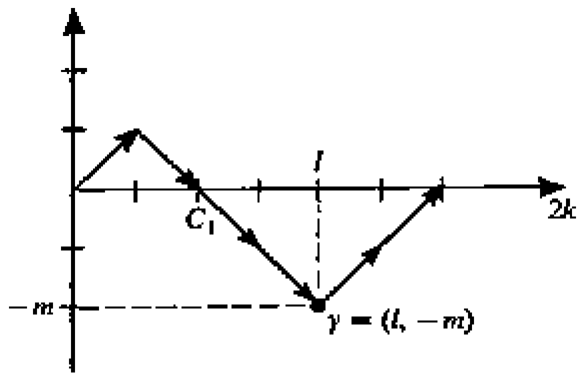


Figure 20

is nonnegative. It follows from these constructions that there is a one-to-one correspondence between the positive paths and the nonnegative paths with at least one $S_i = 0$. Therefore formula (10) is established.

We now establish (11). From symmetry and (10) it is enough to show that

$$L_{2k}(S_1 > 0, \ldots, S_{2k} > 0) + L_{2k}(S_1 \geq 0, \ldots, S_{2k} \geq 0 \text{ and } \exists\, i,$$
$$1 \leq i \leq 2k, \text{ such that } S_i = 0) = L_{2k}(S_{2k} = 0).$$

The set of paths $(S_{2k} = 0)$ can be represented as the sum of the two sets $\mathscr{C}_1$ and $\mathscr{C}_2$, where $\mathscr{C}_1$ contains the paths $(S_0, \ldots, S_{2k})$ that have just one minimum, and $\mathscr{C}_2$ contains those for which the minimum is attained at at least two points.

Let $C_1 \in \mathscr{C}_1$ (Figure 20) and let γ be the minimum point. We put the path $C_1 = (S_0, S_1, \ldots, S_{2k})$ in correspondence with the path C_1^* obtained in the following way (Figure 21). We reflect $(S_0, S_1, \ldots, S_l)$ around the vertical line through the point l, and displace the resulting path to the right and upward, thus releasing it from the point $(2k, 0)$. Then we move the origin to the point $(l, -m)$. The resulting path C_1^* will be positive.

In the same way, if $C_2 \in \mathscr{C}_2$ we can use the same device to put it into correspondence with a nonnegative path C_2^*.

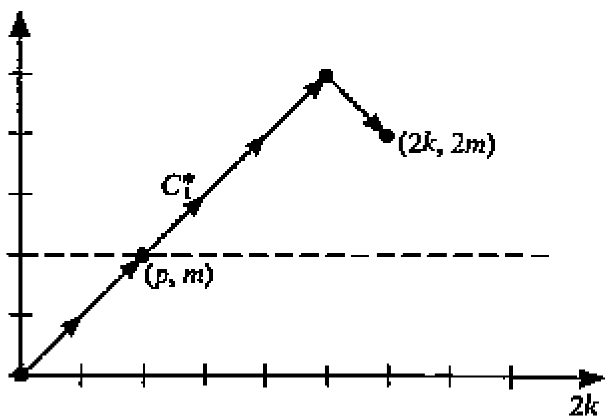


Figure 21

Conversely, let $C_1^* = (S_1 > 0, \ldots, S_{2k} > 0)$ be a positive path with $S_{2k} = 2m$ (see Figure 21). We make it correspond to the path C_1 that is obtained in the following way. Let p be the last point at which $S_p = m$. Reflect $(S_p, \ldots, S_{2m})$ with respect to the vertical line $x = p$ and displace the resulting path downward and to the left until its right-hand end coincides with the point (0, 0). Then we move the origin to the left-hand end of the resulting path (this is just the path drawn in Figure 20). The resulting path $C_1 = (S_0, \ldots, S_{2k})$ has a minimum at $S_{2k} = 0$. A similar construction applied to paths ($S_1 \geq 0, \ldots, S_{2k} \geq 0$ and $\exists\, i$, $1 \leq i \leq 2k$, with $S_i = 0$) leads to paths for which there are at least two minima and $S_{2k} = 0$. Hence we have established a one-to-one correspondence, which establishes (11).

Therefore we have established (9) and consequently also the formula $f_{2k} = u_{2(k-1)} - u_{2k} = (1/2k)u_{2(k-1)}$.

By Stirling's formula

$$u_{2k} = C_{2k}^k \cdot 2^{-2k} \sim \frac{1}{\sqrt{\pi k}}, \qquad k \to \infty.$$

Therefore

$$f_{2k} \sim \frac{1}{2\sqrt{\pi} k^{3/2}}, \qquad k \to \infty.$$

Hence it follows that the expectation of the first time when zero is reached, namely

$$\begin{aligned} \mathsf{E}\min(\sigma_{2n}, 2n) &= \sum_{k=1}^{n} 2k\mathsf{P}(\sigma_{2n} = 2k) + 2nu_{2n} \\ &= \sum_{k=1}^{n} u_{2(k-1)} + 2nu_{2n}, \end{aligned}$$

can be arbitrarily large.

In addition, $\sum_{k=1}^{\infty} u_{2(k-1)} = \infty$, and consequently the limiting value of the mean time for the walk to reach zero (in an unbounded number of steps) is ∞.

This property accounts for many of the unexpected properties of the symmetric random walk that we have been discussing. For example, it would be natural to suppose that after time $2n$ the number of zero net scores in a game between two equally matched players ($p = q = \frac{1}{2}$), i.e. the number of instants i at which $S_i = 0$, would be proportional to $2n$. However, in fact the number of zeros has order $\sqrt{2n}$ (see [F1] and (15) in §9, Chapter VII). Hence it follows, in particular, that, contrary to intuition, the "typical" walk $(S_0, S_1, \ldots, S_n)$ does not have a sinusoidal character (so that roughly half the time the particle would be on the positive side and half the time on the negative side), but instead must resemble a stretched-out wave. The precise formulation of this statement is given by the arcsine law, which we proceed to investigate.

2. Let $P_{2k,2n}$ be the probability that during the interval $[0, 2n]$ the particle spends $2k$ units of time on the positive side.*

Lemma 2. *Let $u_0 = 1$ and $0 \le k \le n$. Then*

$$P_{2k,2n} = u_{2k} \cdot u_{2n-2k} \tag{12}$$

PROOF. It was shown above that $f_{2k} = u_{2(k-1)} - u_{2k}$. Let us show that

$$u_{2k} = \sum_{r=1}^{k} f_{2r} \cdot u_{2(k-r)}. \tag{13}$$

Since $\{S_{2k} = 0\} \subseteq \{\sigma_{2n} \le 2k\}$, we have

$$\{S_{2k} = 0\} = \{S_{2k} = 0\} \cap \{\sigma_{2n} \le 2k\} = \sum_{1 \le l \le k} \{S_{2k} = 0\} \cap \{\sigma_{2n} = 2l\}.$$

Consequently

$$\begin{aligned} u_{2k} = \mathsf{P}(S_{2k} = 0) &= \sum_{1 \le l \le k} \mathsf{P}(S_{2k} = 0, \sigma_{2n} = 2l) \\ &= \sum_{1 \le l \le k} \mathsf{P}(S_{2k} = 0 \mid \sigma_{2k} = 2l)\mathsf{P}(\sigma_{2n} = 2l). \end{aligned}$$

But

$$\begin{aligned} \mathsf{P}(S_{2k} = 0 \mid \sigma_{2n} = 2l) &= \mathsf{P}(S_{2k} = 0 \mid S_1 \ne 0, \ldots, S_{2l-1} \ne 0, S_{2l} = 0) \\ &= \mathsf{P}(S_{2l} + (\xi_{2l+1} + \cdots + \xi_{2k}) = 0 \mid S_1 \ne 0, \ldots, S_{2l-1} \ne 0, S_{2l} = 0) \\ &= \mathsf{P}(S_{2l} + (\xi_{2l+1} + \cdots + \xi_{2k}) = 0 \mid S_{2l} = 0) \\ &= \mathsf{P}(\xi_{2l+1} + \cdots + \xi_{2k} = 0) = \mathsf{P}(S_{2(k-l)} = 0). \end{aligned}$$

Therefore

$$u_{2k} = \sum_{1 \le l \le k} \mathsf{P}(S_{2(k-l)} = 0)\mathsf{P}(\sigma_{2n} = 2l),$$

which establishes (13).

We turn now to the proof of (12). It is obviously true for $k = 0$ and $k = n$. Now let $1 \le k \le n - 1$. If the particle is on the positive side for exactly $2k$ instants, it must pass through zero. Let $2r$ be the time of first passage through zero. There are two possibilities: either $S_k \ge 0$, $k \le 2r$, or $S_k \le 0$, $k \le 2r$.

The number of paths of the first kind is easily seen to be

$$(\tfrac{1}{2} 2^{2r} f_{2r}) \cdot 2^{2(n-r)} P_{2(k-r),2(n-r)} = \tfrac{1}{2} \cdot 2^{2n} \cdot f_{2r} \cdot P_{2(k-r),2(n-r)}.$$

* We say that the particle is on the positive side in the interval $[m - 1, m]$ if one, at least, of the values S_{m-1} and S_m is positive.

The corresponding number of paths of the second kind is

$$\tfrac{1}{2}\cdot 2^{2n}\cdot f_{2r}\cdot P_{2k,\,2(n-r)}.$$

Consequently, for $1 \le k \le n-1$,

$$P_{2k,\,2n} = \frac{1}{2}\sum_{r=1}^{k} f_{2r}\cdot P_{2(k-r),\,2(n-r)} + \frac{1}{2}\sum_{r=1}^{k} f_{2r}\cdot P_{2k,\,2(n-r)}. \tag{14}$$

Let us suppose that $P_{2k,\,2m} = u_{2k}\cdot u_{2m-2k}$ holds for $m = 1, \ldots, n-1$. Then we find from (13) and (14) that

$$\begin{aligned} P_{2k,\,2n} &= \tfrac{1}{2}u_{2n-2k}\cdot\sum_{r=1}^{k} f_{2r}\cdot u_{2k-2r} + \tfrac{1}{2}u_{2k}\cdot\sum_{r=1}^{k} f_{2r}\cdot u_{2n-2r-2k} \\ &= \tfrac{1}{2}u_{2n-2k}\cdot u_{2k} + \tfrac{1}{2}u_{2k}\cdot u_{2n-2k} = u_{2k}\cdot u_{2n-2k}. \end{aligned}$$

This completes the proof of the lemma.

Now let $\gamma(2n)$ be the number of time units that the particle spends on the positive axis in the interval $[0, 2n]$. Then, when $x < 1$,

$$\mathsf{P}\left\{\frac{1}{2} < \frac{\gamma(2n)}{2n} \le x\right\} = \sum_{\{k,\,1/2<(2k/2n)\le x\}} P_{2k,\,2n}.$$

Since

$$u_{2k} \sim \frac{1}{\sqrt{\pi k}}$$

as $k \to \infty$, we have

$$P_{2k,\,2n} = u_{2k}\cdot u_{2(n-k)} \sim \frac{1}{\pi\sqrt{k(n-k)}},$$

as $k \to \infty$ and $n - k \to \infty$.

Therefore

$$\sum_{\{k:\,1/2<(2k/2n)\le x\}} P_{2k,\,2n} - \sum_{\{k:\,1/2<(2k/2n)\le x\}} \frac{1}{\pi n}\cdot\left[\frac{k}{n}\left(1-\frac{k}{n}\right)\right]^{-1/2} \to 0, \quad n \to \infty,$$

whence

$$\sum_{\{k:\,1/2<(2k/2n)\le x\}} P_{2k,\,2n} - \frac{1}{\pi}\int_{1/2}^{x} \frac{dt}{\sqrt{t(1-t)}} \to 0, \qquad n \to \infty.$$

But, by symmetry,

$$\sum_{\{k:\,k/n\le 1/2\}} P_{2k,\,2n} \to \tfrac{1}{2}$$

and

$$\frac{1}{\pi}\int_{1/2}^{x} \frac{dt}{\sqrt{t(1-t)}} = \frac{2}{\pi}\arcsin\sqrt{x} - \tfrac{1}{2}.$$

Consequently we have proved the following theorem.

Theorem (Arcsine Law). *The probability that the fraction of the time spent by the particle on the positive side is at most x tends to* $2\pi^{-1}\arcsin\sqrt{x}$:

$$\sum_{\{k:\,k/n\le x\}} P_{2k,2n} \to 2\pi^{-1}\arcsin\sqrt{x}. \tag{15}$$

We remark that the integrand $p(t)$ in the integral

$$\frac{1}{\pi}\int_{0}^{x} \frac{dt}{\sqrt{t(1-t)}}$$

represents a U-shaped curve that tends to infinity as $t \to 0$ or 1.

Hence it follows that, for large n,

$$\mathsf{P}\left\{0 < \frac{\gamma(2n)}{2n} \le \Delta\right\} > \mathsf{P}\left\{\frac{1}{2} < \frac{\gamma(2n)}{2n} \le \tfrac{1}{2} + \Delta\right\},$$

i.e., it is more likely that the fraction of the time spent by the particle on the positive side is close to zero or one, than to the intuitive value $\frac{1}{2}$.

Using a table of arcsines and noting that the convergence in (15) is indeed quite rapid, we find that

$$\mathsf{P}\left\{\frac{\gamma(2n)}{2n} \le 0.024\right\} \approx 0.1,$$

$$\mathsf{P}\left\{\frac{\gamma(2n)}{2n} \le 0.1\right\} \approx 0.2,$$

$$\mathsf{P}\left\{\frac{\gamma(2n)}{2n} \le 0.2\right\} \approx 0.3,$$

$$\mathsf{P}\left\{\frac{\gamma(2n)}{2n} \le 0.65\right\} \approx 0.6.$$

Hence if, say, $n = 1000$, then in about one case in ten, the particle spends only 24 units of time on the positive axis and therefore spends the greatest amount of time, 976 units, on the negative axis.

3. Problems

1. How fast does $\mathsf{E}\min(\sigma_{2n}, 2n) \to \infty$ as $n \to \infty$?

2. Let $\tau_n = \min\{1 \le k \le n: S_k = 1\}$, where we take $\tau_n = \infty$ if $S_k < 1$ for $1 \le k \le n$. What is the limit of $\mathsf{E}\min(\tau_n, n)$ as $n \to \infty$ for symmetric ($p = q = \frac{1}{2}$) and for unsymmetric ($p \ne q$) walks?

§11. Martingales. Some Applications to the Random Walk

1. The Bernoulli random walk discussed above was generated by a sequence $\xi_1, \ldots, \xi_n$ of *independent* random variables. In this and the next section we introduce two important classes of *dependent* random variables, those that constitute martingales and Markov chains.

The theory of martingales will be developed in detail in Chapter VII. Here we shall present only the essential definitions, prove a theorem on the preservation of the martingale property for stopping times, and apply this to deduce the "ballot theorem." In turn, the latter theorem will be used for another proof of proposition (10.5), which was obtained above by applying the reflection principle.

2. Let $(\Omega, \mathscr{A}, \mathsf{P})$ be a finite probability space and $\mathscr{D}_1 \preccurlyeq \mathscr{D}_2 \preccurlyeq \cdots \preccurlyeq \mathscr{D}_n$ a sequence of decompositions.

Definition 1. A sequence of random variables $\xi_1, \ldots, \xi_n$ is called a *martingale* (with respect to the decomposition $\mathscr{D}_1 \preccurlyeq \mathscr{D}_2 \preccurlyeq \cdots \preccurlyeq \mathscr{D}_n$) if

(1) ξ_k is $\mathscr{D}_k$-measurable,
(2) $\mathsf{E}(\xi_{k+1} | \mathscr{D}_k) = \xi_k$, $1 \le k \le n - 1$.

In order to emphasize the system of decompositions with respect to which the random variables form a martingale, we shall use the notation

$$\xi = (\xi_k, \mathscr{D}_k)_{1 \le k \le n}, \tag{1}$$

where for the sake of simplicity we often do not mention explicitly that $1 \le k \le n$.

When $\mathscr{D}_k$ is induced by $\xi_1, \ldots, \xi_n$, i.e.

$$\mathscr{D}_k = \mathscr{D}_{\xi_1, \ldots, \xi_k},$$

instead of saying that $\xi = (\xi_k, \mathscr{D}_k)$ is a martingale, we simply say that the sequence $\xi = (\xi_k)$ is a martingale.

Here are some examples of martingales.

Example 1. Let $\eta_1, \ldots, \eta_n$ be independent Bernoulli random variables with

$$\mathsf{P}(\eta_k = 1) = \mathsf{P}(\eta_k = -1) = \tfrac{1}{2},$$

$$S_k = \eta_1 + \cdots + \eta_k \quad \text{and} \quad \mathscr{D}_k = \mathscr{D}_{\eta_1, \ldots, \eta_k}.$$

We observe that the decompositions $\mathscr{D}_k$ have a simple structure:

$$\mathscr{D}_1 = \{D^+, D^-\},$$

where

$$D^+ = \{\omega: \eta_1 = +1\}, \qquad D^- = \{\omega: \eta_1 = -1\},$$
$$\mathscr{D}_2 = \{D^{++}, D^{+-}, D^{-+}, D^{--}\},$$

where

$$D^{++} = \{\omega: \eta_1 = +1, \eta_2 = +1\}, \ldots, D^{--} = \{\omega: \eta_1 = -1, \eta_2 = -1\},$$

etc.

It is also easy to see that $\mathscr{D}_{\eta_1,\ldots,\eta_k} = \mathscr{D}_{S_1,\ldots,S_k}$.

Let us show that $(S_k, \mathscr{D}_k)$ forms a martingale. In fact, S_k is $\mathscr{D}_k$-measurable, and by (8.12), (8.18) and (8.24),

$$\begin{aligned} \mathsf{E}(S_{k+1}|\mathscr{D}_k) &= \mathsf{E}(S_k + \eta_{k+1}|\mathscr{D}_k) \\ &= \mathsf{E}(S_k|\mathscr{D}_k) + \mathsf{E}(\eta_{k+1}|\mathscr{D}_k) = S_k + \mathsf{E}\eta_{k+1} = S_k. \end{aligned}$$

If we put $S_0 = 0$ and take $D_0 = \{\Omega\}$, the trivial decomposition, then the sequence $(S_k, \mathscr{D}_k)_{0 \le k \le n}$ also forms a martingale.

Example 2. Let $\eta_1, \ldots, \eta_n$ be independent Bernoulli random variables with $P(\eta_i = 1) = p$, $P(\eta_i = -1) = q$. If $p \neq q$, each of the sequences $\xi = (\xi_k)$ with

$$\xi_k = \left(\frac{q}{p}\right)^{S_k}, \qquad \xi_k = S_k - k(p-q), \qquad \text{where} \quad S_k = \eta_1 + \cdots + \eta_k,$$

is a martingale.

Example 3. Let η be a random variable, $\mathscr{D}_1 \preccurlyeq \cdots \preccurlyeq \mathscr{D}_n$, and

$$\xi_k = \mathsf{E}(\eta|\mathscr{D}_k). \tag{2}$$

Then the sequence $\xi = (\xi_k, \mathscr{D}_k)$ is a martingale. In fact, it is evident that $\mathsf{E}(\eta|\mathscr{D}_k)$ is $\mathscr{D}_k$-measurable, and by (8.20)

$$\mathsf{E}(\xi_{k+1}|\mathscr{D}_k) = \mathsf{E}[\mathsf{E}(\eta|\mathscr{D}_{k+1})\mathscr{D}_k] = \mathsf{E}(\eta|\mathscr{D}_k) = \xi_k.$$

In this connection we notice that if $\xi = (\xi_k, \mathscr{D}_k)$ is any martingale, then by (8.20)

$$\begin{aligned} \xi_k &= \mathsf{E}(\xi_{k+1}|\mathscr{D}_k) = \mathsf{E}[\mathsf{E}(\xi_{k+2}|\mathscr{D}_{k+1})|\mathscr{D}_k] \\ &= \mathsf{E}(\xi_{k+2}|\mathscr{D}_k) = \cdots = \mathsf{E}(\xi_n|\mathscr{D}_k). \end{aligned} \tag{3}$$

Consequently the set of martingales $\xi = (\xi_k, \mathscr{D}_k)$ is exhausted by the martingales of the form (2). (We note that for infinite sequences $\xi = (\xi_k, \mathscr{D}_k)_{k \ge 1}$ this is, in general, no longer the case; see Problem 7 in §1 of Chapter VII.)

EXAMPLE 4. Let $\eta_1, \ldots, \eta_n$ be a sequence of independent identically distributed random variables, $S_k = \eta_1 + \cdots + \eta_k$, and $\mathscr{D}_1 = \mathscr{D}_{S_n}$, $\mathscr{D}_2 = \mathscr{D}_{S_n, S_{n-1}}, \ldots,$ $\mathscr{D}_n = \mathscr{D}_{S_n, S_{n-1}, \ldots, S_1}$. Let us show that the sequence $\xi = (\xi_k, \mathscr{D}_k)$ with

$$\xi_1 = \frac{S_n}{n}, \xi_2 = \frac{S_{n-1}}{n-1}, \ldots, \xi_k = \frac{S_{n+1-k}}{n+1-k}, \ldots, \xi_n = S_1$$

is a martingale. In the first place, it is clear that $\mathscr{D}_k \preccurlyeq \mathscr{D}_{k+1}$ and ξ_k is $\mathscr{D}_k$-measurable. Moreover, we have by symmetry, for $j \le n - k + 1$,

$$\mathsf{E}(\eta_j | \mathscr{D}_k) = \mathsf{E}(\eta_1 | \mathscr{D}_k) \tag{4}$$

(compare (8.26)). Therefore

$$(n - k + 1)\mathsf{E}(\eta_1 | \mathscr{D}_k) = \sum_{j=1}^{n-k+1} \mathsf{E}(\eta_j | \mathscr{D}_k) = \mathsf{E}(S_{n-k+1} | \mathscr{D}_k) = S_{n-k+1},$$

and consequently

$$\xi_k = \frac{S_{n-k+1}}{n-k+1} = \mathsf{E}(\eta_1 | \mathscr{D}_k),$$

and it follows from Example 3 that $\xi = (\xi_k, \mathscr{D}_k)$ is a martingale.

Remark. From this martingale property of the sequence $\xi = (\xi_k, \mathscr{D}_k)_{1 \le k \le n}$, it is clear why we will sometimes say that the sequence $(S_k/k)_{1 \le k \le n}$ forms a *reversed martingale*. (Compare problem 6 in §1 of Chapter VII.)

EXAMPLE 5. Let $\eta_1, \ldots, \eta_n$ be independent Bernoulli random variables with

$$\mathsf{P}(\eta_i = +1) = \mathsf{P}(\eta_i = -1) = \tfrac{1}{2},$$

$S_k = \eta_1 + \cdots + \eta_k$. Let A and B be integers, $A < 0 < B$. Then with $0 < \lambda < \pi/2$, the sequence $\xi = (\xi_k, \mathscr{D}_k)$ with $\mathscr{D}_k = \mathscr{D}_{S_1, \ldots, S_k}$ and

$$\xi_k = (\cos \lambda)^{-k} \exp\left\{i\lambda\left(S_k - \frac{B + A}{2}\right)\right\} \tag{5}$$

is a complex martingale (i.e., the real and imaginary parts of ξ_k form martingales).

3. It follows from the definition of a martingale that the expectation $\mathsf{E}\xi_k$ is the same for every k:

$$\mathsf{E}\xi_k = \mathsf{E}\xi_1.$$

It turns out that this property persists if time k is replaced by a random time.

In order to formulate this property we introduce the following definition.

Definition 2. A random variable $\tau = \tau(\omega)$ that takes the values $1, 2, \ldots, n$ is called a *stopping time* (with respect to a decomposition $(\mathscr{D}_k)_{1 \le k \le n}$, $\mathscr{D}_1 \preccurlyeq \mathscr{D}_2 \preccurlyeq \cdots \preccurlyeq \mathscr{D}_n$) if, for $k = 1, \ldots, n$, the random variable $I_{\{\tau = k\}}(\omega)$ is $\mathscr{D}_k$-measurable.

If we consider $\mathscr{D}_k$ as the decomposition induced by observations for k steps (for example, $\mathscr{D}_k = \mathscr{D}_{\eta_1,\dots,\eta_k}$, the decomposition induced by the variables $\eta_1, \dots, \eta_k$), then the $\mathscr{D}_k$-measurability of $I_{\{\tau=k\}}(\omega)$ means that the realization or nonrealization of the event $\{\tau = k\}$ is determined only by observations for k steps (and is independent of the "future").

If $\mathscr{B}_k = \alpha(\mathscr{D}_k)$, then the $\mathscr{D}_k$-measurability of $I_{\{\tau=k\}}(\omega)$ is equivalent to the assumption that

$$\{\tau = k\} \in \mathscr{B}_k. \tag{6}$$

We have already introduced specific examples of stopping times: the times τ_k^x, σ_{2n} introduced in §§9 and 10. Those times are special cases of stopping times of the form

$$\begin{aligned} \tau^A &= \min\{0 < k \le n: \xi_k \in A\}, \\ \sigma^A &= \min\{0 \le k \le n: \xi_k \in A\}, \end{aligned} \tag{7}$$

which are the times (respectively the first time after zero and the first time) for a sequence $\xi_0, \xi_1, \dots, \xi_n$ to attain a point of the set A.

4. Theorem 1. *Let $\xi = (\xi_k, \mathscr{D}_k)_{1 \le k \le n}$ be a martingale and τ a stopping time with respect to the decomposition $(\mathscr{D}_k)_{1 \le k \le n}$. Then*

$$\mathsf{E}(\xi_\tau | \mathscr{D}_1) = \xi_1, \tag{8}$$

where

$$\xi_\tau = \sum_{k=1}^n \xi_k I_{\{\tau=k\}}(\omega) \tag{9}$$

and

$$\mathsf{E}\xi_\tau = \mathsf{E}\xi_1. \tag{10}$$

Proof (compare the proof of (9.29)). Let $D \in \mathscr{D}_1$. Using (3) and the properties of conditional expectations, we find that

$$\begin{aligned} \mathsf{E}(\xi_\tau | D) &= \frac{\mathsf{E}(\xi_\tau I_D)}{\mathsf{P}(D)} \\ &= \frac{1}{\mathsf{P}(D)} \cdot \sum_{l=1}^n \mathsf{E}(\xi_l \cdot I_{\{\tau=l\}} \cdot I_D) \\ &= \frac{1}{\mathsf{P}(D)} \sum_{l=1}^n \mathsf{E}[\mathsf{E}(\xi_n | \mathscr{D}_l) \cdot I_{\{\tau=l\}} \cdot I_D] \\ &= \frac{1}{\mathsf{P}(D)} \sum_{l=1}^n \mathsf{E}[\mathsf{E}(\xi_n I_{\{\tau=l\}} \cdot I_D | \mathscr{D}_l)] \\ &= \frac{1}{\mathsf{P}(D)} \sum_{l=1}^n \mathsf{E}[\xi_n I_{\{\tau=l\}} \cdot I_D] \\ &= \frac{1}{\mathsf{P}(D)} \mathsf{E}(\xi_n I_D) = \mathsf{E}(\xi_n | D), \end{aligned}$$

and consequently

$$\mathsf{E}(\xi_\tau|\mathscr{D}_1) = \mathsf{E}(\xi_n|\mathscr{D}_1) = \xi_1.$$

The equation $\mathsf{E}\xi_\tau = \mathsf{E}\xi_1$ then follows in an obvious way.

This completes the proof of the theorem.

Corollary. *For the martingale $(S_k, \mathscr{D}_k)_{1\le k\le n}$ of Example 1, and any stopping time τ (with respect to $(\mathscr{D}_k)$) we have the formulas*

$$\mathsf{E}S_\tau = 0, \qquad \mathsf{E}S_\tau^2 = \mathsf{E}\tau, \tag{11}$$

known as Wald's identities (cf. (9.29) and (9.30); see also Problem 1 and Theorem 3 in §2 of Chapter VII).

5. Let us use Theorem 1 to establish the following proposition.

Theorem 2 (Ballot Theorem). *Let $\eta_1, \ldots, \eta_n$ be a sequence of independent identically distributed random variables whose values are nonnegative integers, $S_k = \eta_1 + \cdots + \eta_k$, $1 \le k \le n$. Then*

$$\mathsf{P}\{S_k < k \text{ for all } k, 1 \le k \le n|S_n\} = \left(1 - \frac{S_n}{n}\right)^+, \tag{12}$$

where $a^+ = \max(a, 0)$.

Proof. On the set $\{\omega: S_n \ge n\}$ the formula is evident. We therefore prove (12) for the sample points at which $S_n < n$.

Let us consider the martingale $\xi = (\xi_k, \mathscr{D}_k)_{1\le k\le n}$ introduced in Example 4, with $\xi_k = S_{n+1-k}/(n+1-k)$ and $\mathscr{D}_k = \mathscr{D}_{S_{n+1-k},\ldots,S_n}$.

We define

$$\tau = \min\{1 \le k \le n: \xi_k \ge 1\},$$

taking $\tau = n$ on the set $\{\xi_k < 1$ for all k such that $1 \le k \le n\} = \{\max_{1\le l\le n}(S_l/l) < 1\}$. It is clear that $\xi_\tau = \xi_n = S_1 = 0$ on this set, and therefore

$$\left\{\max_{1\le l\le n} \frac{S_l}{l} < 1\right\} = \left\{\max_{1\le l\le n} \frac{S_l}{l} < 1, S_n < n\right\} \subseteq \{\xi_\tau = 0\}. \tag{13}$$

Now let us consider those outcomes for which simultaneously $\max_{1\le l\le n}(S_l/l) \ge 1$ and $S_n < n$. Write $\sigma = n + 1 - \tau$. It is easy to see that

$$\sigma = \max\{1 \le k \le n: S_k \ge k\}$$

and therefore (since $S_n < n$) we have $\sigma < n$, $S_\sigma \ge \sigma$, and $S_{\sigma+1} < \sigma + 1$. Consequently $\eta_{\sigma+1} = S_{\sigma+1} - S_\sigma < (\sigma + 1) - \sigma = 1$, i.e. $\eta_{\sigma+1} = 0$. Therefore $\sigma \le S_\sigma = S_{\sigma+1} < \sigma + 1$, and consequently $S_\sigma = \sigma$ and

$$\xi_\tau = \frac{S_{n+1-\tau}}{n+1-\tau} = \frac{S_\sigma}{\sigma} = 1.$$

Therefore

$$\left\{\max_{1\le l\le n}\frac{S_l}{l}\ge 1,\ S_n<n\right\}\subseteq\{\xi_\tau=1\}. \tag{14}$$

From (13) and (14) we find that

$$\left\{\max_{1\le l\le n}\frac{S_l}{l}\ge 1,\ S_n<n\right\}=\{\xi_\tau=1\}\cap\{S_n<n\}.$$

Therefore, on the set $\{S_n<n\}$, we have

$$\mathsf{P}\left\{\max_{1\le l\le n}\frac{S_l}{l}\ge 1\,|\,S_n\right\}=\mathsf{P}\{\xi_\tau=1|S_n\}=\mathsf{E}(\xi_\tau|S_n),$$

where the last equation follows because ξ_τ takes only the two values 0 and 1.

Let us notice now that $\mathsf{E}(\xi_\tau|S_n)=\mathsf{E}(\xi_\tau|\mathscr{D}_1)$, and (by Theorem 1) $\mathsf{E}(\xi_\tau|\mathscr{D}_1)=\xi_1=S_n/n$. Consequently, on the set $\{S_n<n\}$ we have $\mathsf{P}\{S_k<k$ *for all k such that* $1\le k\le n|S_n\}=1-(S_n/n)$.

This completes the proof of the theorem.

We now apply this theorem to obtain a different proof of Lemma 1 of §10, and explain why it is called the ballot theorem.

Let $\xi_1,\ldots,\xi_n$ be independent Bernoulli random variables with

$$\mathsf{P}(\xi_1=1)=\mathsf{P}(\xi_i=-1)=\tfrac{1}{2},$$

$S_k=\xi_1+\cdots+\xi_k$ and a, b nonnegative integers such that $a-b>0$, $a+b=n$. We are going to show that

$$\mathsf{P}\{S_1>0,\ldots,S_n>0|S_n=a-b\}=\frac{a-b}{a+b}. \tag{15}$$

In fact, by symmetry,

$$\begin{aligned}\mathsf{P}\{S_1&>0,\ldots,S_n>0|S_n=a-b\}\\&=\mathsf{P}\{S_1<0,\ldots,S_n<0|S_n=-(a-b)\}\\&=\mathsf{P}\{S_1+1<1,\ldots,S_n+n<n|S_n+n=n-(a-b)\}\\&=\mathsf{P}\{\eta_1<1,\ldots,\eta_1+\cdots+\eta_n<n|\eta_1+\cdots+\eta_n=n-(a-b)\}\\&=\left[1-\frac{n-(a-b)}{n}\right]^+=\frac{a-b}{n}=\frac{a-b}{a+b},\end{aligned}$$

where we have put $\eta_k=\xi_k+1$ and applied (12).

Now formula (10.5) follows from (15) in an evident way; the formula was also established in Lemma 1 of §10 by using the reflection principle.

Let us interpret $\xi_i = +1$ as a vote for candidate A and $\xi_i = -1$ as a vote for B. Then S_k is the difference between the numbers of votes cast for A and B at the time when k votes have been recorded, and

$$\mathsf{P}\{S_1 > 0, \ldots, S_n > 0 | S_n = a - b\}$$

is the probability that A was always ahead of B, with the understanding that A received a votes in all, B received b votes, and $a - b > 0$, $a + b = n$. According to (15) this probability is $(a - b)/n$.

6. Problems

1. Let $\mathscr{D}_0 \preccurlyeq \mathscr{D}_1 \preccurlyeq \cdots \preccurlyeq \mathscr{D}_n$ be a sequence of decompositions with $\mathscr{D}_0 = \{\Omega\}$, and let η_k be $\mathscr{D}_k$-measurable variables, $1 \le k \le n$. Show that the sequence $\xi = (\xi_k, \mathscr{D}_k)$ with

$$\xi_k = \sum_{l=1}^{k} [\eta_l - \mathsf{E}(\eta_l | \mathscr{D}_{l-1})]$$

is a martingale.

2. Let the random variables $\eta_1, \ldots, \eta_k$ satisfy $\mathsf{E}(\eta_k | \eta_1, \ldots, \eta_{k-1}) = 0$. Show that the sequence $\xi = (\xi_k)_{1 \le k \le n}$ with $\xi_1 = \eta_1$ and

$$\xi_{k+1} = \sum_{i=1}^{k} \eta_{i+1} f_i(\eta_1, \ldots, \eta_i),$$

where f_i are given functions, is a martingale.

3. Show that every martingale $\xi = (\xi_i, \mathscr{D}_k)$ has uncorrelated increments: if $a < b < c < d$ then

$$\operatorname{cov}(\xi_d - \xi_c, \xi_b - \xi_a) = 0.$$

4. Let $\xi = (\xi_1, \ldots, \xi_n)$ be a random sequence such that ξ_k is $\mathscr{D}_k$-measurable ($\mathscr{D} \preccurlyeq \mathscr{D}_2 \preccurlyeq \cdots \preccurlyeq \mathscr{D}_n$). Show that a necessary and sufficient condition for this sequence to be a martingale (with respect to the system $(\mathscr{D}_k)$) is that $\mathsf{E}\xi_\tau = \mathsf{E}\xi_1$ for every stopping time τ (with respect to $(\mathscr{D}_k)$). (The phrase "for every stopping time" can be replaced by "for every stopping time that assumes two values.")

5. Show that if $\xi = (\xi_k, \mathscr{D}_k)_{1 \le k \le n}$ is a martingale and τ is a stopping time, then

$$\mathsf{E}[\xi_n I_{\{\tau = k\}}] = \mathsf{E}[\xi_k I_{\{\tau = k\}}]$$

for every k.

6. Let $\xi = (\xi_k, \mathscr{D}_k)$ and $\eta = (\eta_k, \mathscr{D}_k)$ be two martingales, $\xi_1 = \eta_1 = 0$. Show that

$$\mathsf{E}\xi_n \eta_n = \sum_{k=2}^{n} \mathsf{E}(\xi_k - \xi_{k-1})(\eta_k - \eta_{k-1})$$

and in particular that

$$\mathsf{E}\xi_n^2 = \sum_{k=2}^{n} \mathsf{E}(\xi_k - \xi_{k-1})^2.$$

7. Let $\eta_1, \ldots, \eta_n$ be a sequence of independent identically distributed random variables with $\mathsf{E}\eta_i = 0$. Show that the sequence $\xi = (\xi_k)$ with

$$\xi_k = \left(\sum_{i=1}^{k} \eta_i\right)^2 - k\mathsf{E}\eta_1^2,$$

$$\xi_k = \frac{\exp \lambda(\eta_1 + \cdots + \eta_k)}{(\mathsf{E} \exp \lambda\eta_1)^k}$$

is a martingale.

8. Let $\eta_1, \ldots, \eta_n$ be a sequence of independent identically distributed random variables taking values in a finite set Y. Let $f_0(y) = \mathsf{P}(\eta_1 = y)$, $y \in Y$, and let $f_1(y)$ be a nonnegative function with $\sum_{y \in Y} f_1(y) = 1$. Show that the sequence $\xi = (\xi_k, \mathscr{D}_k^\eta)$ with $\mathscr{D}_k^\eta = D_{\eta_1, \ldots, \eta_k}$,

$$\xi_k = \frac{f_1(\eta_1) \cdots f_1(\eta_k)}{f_0(\eta_1) \cdots f_0(\eta_k)},$$

is a martingale. (The variables ξ_k, known as *likelihood ratios*, are extremely important in mathematical statistics.)

§12. Markov Chains. Ergodic Theorem. Strong Markov Property

1. We have discussed the Bernoulli scheme with

$$\Omega = \{\omega : \omega = (x_1, \ldots, x_n), x_i = 0, 1\},$$

where the probability $p(\omega)$ of each outcome is given by

$$p(\omega) = p(x_1) \cdots p(x_n), \tag{1}$$

with $p(x) = p^x q^{1-x}$. With these hypotheses, the variables $\xi_1, \ldots, \xi_n$ with $\xi_i(\omega) = x_i$ are *independent* and *identically distributed* with

$$\mathsf{P}(\xi_1 = x) = \cdots = \mathsf{P}(\xi_n = x) = p(x), \qquad x = 0, 1.$$

If we replace (1) by

$$p(\omega) = p_1(x_1) \cdots p_n(x_n),$$

where $p_i(x) = p_i^x(1 - p_i)$, $0 \leq p_i \leq 1$, the random variables $\xi_1, \ldots, \xi_n$ are still *independent*, but in general are *differently distributed*:

$$\mathsf{P}(\xi_1 = x) = p_1(x), \ldots, \mathsf{P}(\xi_n = x) = p_n(x).$$

We now consider a generalization that leads to *dependent* random variables that form what is known as a Markov chain.

Let us suppose that

$$\Omega = \{\omega : \omega = (x_0, x_1, \ldots, x_n), x_i \in X\},$$

where X is a finite set. Let there be given nonnegative functions $p_0(x)$, $p_1(x, y), \ldots, p_n(x, y)$ such that

$$\sum_{x \in X} p_0(x) = 1,$$
$$\sum_{y \in X} p_k(x, y) = 1, \qquad k = 1, \ldots, n; \quad y \in X. \tag{2}$$

For each $\omega = (x_0, x_1, \ldots, x_n)$, put

$$p(\omega) = p_0(x_0)p_1(x_0, x_1) \cdots p_n(x_{n-1}, x_n). \tag{3}$$

It is easily verified that $\sum_{\omega \in \Omega} p(\omega) = 1$, and consequently the set of numbers $p(\omega)$ together with the space Ω and the collection of its subsets defines a probabilistic model, which it is usual to call a *model of experiments that form a Markov chain.*

Let us introduce the random variables $\xi_0, \xi_1, \ldots, \xi_n$ with $\xi_i(\omega) = x_i$. A simple calculation shows that

$$\mathsf{P}(\xi_0 = a) = p_0(a),$$
$$\mathsf{P}(\xi_0 = a_0, \ldots, \xi_k = a_k) = p_0(a_0)p_1(a_0, a_1) \cdots p_k(a_{k-1}, a_k). \tag{4}$$

We now establish the validity of the following fundamental property of conditional probabilities:

$$\mathsf{P}\{\xi_{k+1} = a_{k+1} | \xi_k = a_k, \ldots, \xi_0 = a_0\} = \mathsf{P}\{\xi_{k+1} = a_{k+1} | \xi_k = a_k\} \tag{5}$$

(under the assumption that $\mathsf{P}(\xi_k = a_k, \ldots, \xi_0 = a_0) > 0$).

By (4),

$$\begin{aligned}\mathsf{P}\{\xi_{k+1} &= a_{k+1} | \xi_k = a_k, \ldots, \xi_0 = a_0\} \\ &= \frac{\mathsf{P}\{\xi_{k+1} = a_{k+1}, \ldots, \xi_0 = a_0\}}{\mathsf{P}\{\xi_k = a_k, \ldots, \xi_0 = a_0\}} \\ &= \frac{p_0(a_0)p_1(a_0, a_1) \cdots p_{k+1}(a_k, a_{k+1})}{p_0(a_0) \cdots p_k(a_{k-1}, a_k)} = p_{k+1}(a_k, a_{k+1}).\end{aligned}$$

In a similar way we verify

$$\mathsf{P}\{\xi_{k+1} = a_{k+1} | \xi_k = a_k\} = p_{k+1}(a_k, a_{k+1}), \tag{6}$$

which establishes (5).

Let $\mathscr{D}_k^\xi = \mathscr{D}_{\xi_0, \ldots, \xi_k}$ be the decomposition induced by $\xi_0, \ldots, \xi_k$, and $\mathscr{B}_k^\xi = \alpha(\mathscr{D}_k^\xi)$.

Then, in the notation introduced in §8, it follows from (5) that

$$\mathsf{P}\{\xi_{k+1} = a_{k+1} | \mathscr{B}_k^\xi\} = \mathsf{P}\{\xi_{k+1} = a_{k+1} | \xi_k\} \tag{7}$$

or

$$\mathsf{P}\{\xi_{k+1} = a_{k+1} | \xi_0, \ldots, \xi_k\} = \mathsf{P}\{\xi_{k+1} = a_{k+1} | \xi_k\}.$$

If we use the evident equation

$$P(AB|C) = P(A|BC)P(B|C),$$

we find from (7) that

$$P\{\xi_n = a_n, \ldots, \xi_{k+1} = a_{k+1}|\mathscr{B}_k^{\xi}\} = P\{\xi_n = a_n, \ldots, \xi_{k+1} = a_{k+1}|\xi_k\} \quad (8)$$

or

$$P\{\xi_n = a_n, \ldots, \xi_{k+1} = a_{k+1}|\xi_0, \ldots, \xi_k\} = P\{\xi_n = a_n, \ldots, \xi_{k+1} = a_{k+1}|\xi_k\}. \quad (9)$$

This equation admits the following intuitive interpretation. Let us think of ξ_k as the position of a particle "at present," $(\xi_0, \ldots, \xi_{k-1})$ as the "past," and $(\xi_{k+1}, \ldots, \xi_n)$ as the "future." Then (9) says that if the past and the present are given, the future depends only on the present and is independent of how the particle arrived at ξ_k, i.e. is independent of the past $(\xi_0, \ldots, \xi_{k-1})$.

Let $F = (\xi_n = a_n, \ldots, \xi_{k+1} = a_{k+1})$, $N = \{\xi_k = a_k\}$,

$$B = \{\xi_{k-1} = a_{k-1}, \ldots, \xi_0 = a_0\}.$$

Then it follows from (9) that

$$P(F|NB) = P(F|N),$$

from which we easily find that

$$P(FB|N) = P(F|N)P(B|N). \quad (10)$$

In other words, it follows from (7) that for a given present N, the future F and the past B are independent. It is easily shown that the converse also holds: if (10) holds for all $k = 0, 1, \ldots, n-1$, then (7) holds for every $k = 0, 1, \ldots, n-1$.

The property of the independence of future and past, or, what is the same thing, the lack of dependence of the future on the past when the present is given, is called the *Markov property*, and the corresponding sequence of random variables $\xi_0, \ldots, \xi_n$ is a *Markov chain*.

Consequently if the probabilities $p(\omega)$ of the sample points are given by (3), the sequence $(\xi_0, \ldots, \xi_n)$ with $\xi_i(\omega) = x_i$ forms a Markov chain.

We give the following formal definition.

Definition. Let $(\Omega, \mathscr{A}, P)$ be a (finite) probability space and let $\xi = (\xi_0, \ldots, \xi_n)$ be a sequence of random variables with values in a (finite) set X. If (7) is satisfied, the sequence $\xi = (\xi_0, \ldots, \xi_n)$ is called a (finite) *Markov chain*.

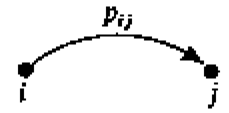
The set X is called the *phase space* or *state space* of the chain. The set of probabilities $(p_n(x))$, $x \in X$, with $p_0(x) = P(\xi_0 = x)$ is the *initial distribution*, and the matrix $\|p_k(x, y)\|$, $x, y \in X$, with $p(x, y) = P\{\xi_k = y|\xi_{k-1} = x\}$ is the *matrix of transition probabilities* (from state x to state y) at time $k = 1, \ldots, n$.

When the transition probabilities $p_k(x, y)$ are independent of k, that is, $p_k(x, y) = p(x, y)$, the sequence $\xi = (\xi_0, \ldots, \xi_n)$ is called a *homogeneous* Markov chain with transition matrix $\|p(x, y)\|$.

Notice that the matrix $\|p(x, y)\|$ is *stochastic*: its elements are nonnegative and the sum of the elements in each row is 1: $\sum_y p(x, y) = 1, x \in X$.

We shall suppose that the phase space X is a finite set of integers ($X = \{0, 1, \ldots, N\}$, $X = \{0, \pm 1, \ldots, \pm N\}$, etc.), and use the traditional notation $p_i = p_0(i)$ and $p_{ij} = p(i, j)$.

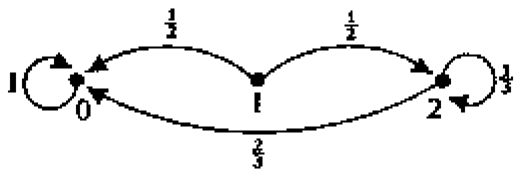
It is clear that the properties of homogeneous Markov chains completely determine the initial distributions p_i and the transition probabilities p_{ij}. In specific cases we describe the evolution of the chain, not by writing out the matrix $\|p_{ij}\|$ explicitly, but by a (directed) graph whose vertices are the states in X, and an arrow from state i to state j with the number p_{ij} over it indicates that it is possible to pass from point i to point j with probability p_{ij}. When $p_{ij} = 0$, the corresponding arrow is omitted.



EXAMPLE 1. Let $X = \{0, 1, 2\}$ and

$$\|p_{ij}\| = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{2}{3} & 0 & \frac{1}{3} \end{pmatrix}.$$

The following graph corresponds to this matrix:

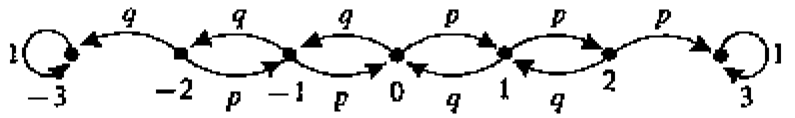


Here state 0 is said to be absorbing: if the particle gets into this state it remains there, since $p_{00} = 1$. From state 1 the particle goes to the adjacent states 0 or 2 with equal probabilities; state 2 has the property that the particle remains there with probability $\frac{1}{3}$ and goes to state 0 with probability $\frac{2}{3}$.

EXAMPLE 2. Let $X = \{0, \pm 1, \ldots, \pm N\}$, $p_0 = 1$, $p_{NN} = p_{(-N)(-N)} = 1$, and, for $|i| < N$,

$$p_{ij} = \begin{cases} p, & j = i + 1, \\ q, & j = i - 1, \\ 0 & \text{otherwise}. \end{cases} \tag{11}$$

The transitions corresponding to this chain can be presented graphically in the following way ($N = 3$):



This chain corresponds to the two-player game discussed earlier, when each player has a bankroll N and at each turn the first player wins $+1$ from the second with probability p, and loses (wins -1) with probability q. If we think of state i as the amount won by the first player from the second, then reaching state N or $-N$ means the ruin of the second or first player, respectively.

In fact, if $\eta_1, \eta_2, \dots, \eta_n$ are independent Bernoulli random variables with $\mathsf{P}(\eta_i = +1) = p$, $\mathsf{P}(\eta_i = -1) = q$, $S_0 = 0$ and $S_k = \eta_1 + \cdots + \eta_k$ the amounts won by the first player from the second, then the sequence S_0, $S_1, \dots, S_n$ is a Markov chain with $p_0 = 1$ and transition matrix (11), since

$$\begin{aligned}\mathsf{P}\{S_{k+1} = j \mid S_k = i_k, S_{k-1} = i_{k-1}, \dots\} \\ = \mathsf{P}\{S_k + \eta_{k+1} = j \mid S_k = i_k, S_{k-1} = i_{k-1}, \dots\} \\ = \mathsf{P}\{S_k + \eta_{k+1} = j \mid S_k = i_k\} = \mathsf{P}\{\eta_{k+1} = j - i_k\}.\end{aligned}$$

This Markov chain has a very simple structure:

$$S_{k+1} = S_k + \eta_{k+1}, \qquad 0 \le k \le n-1,$$

where $\eta_1, \eta_2, \dots, \eta_n$ is a sequence of independent random variables.

The same considerations show that if $\xi_0, \eta_1, \dots, \eta_n$ are independent random variables then the sequence $\xi_0, \xi_1, \dots, \xi_n$ with

$$\xi_{k+1} = f_k(\xi_k, \eta_{k+1}), \qquad 0 \le k \le n-1, \tag{12}$$

is also a Markov chain.

It is worth noting in this connection that a Markov chain constructed in this way can be considered as a natural probabilistic analog of a (deterministic) sequence $x = (x_0, \dots, x_n)$ generated by the recurrent equations

$$x_{k+1} = f_k(x_k).$$

We now give another example of a Markov chain of the form (12); this example arises in queueing theory.

Example 3. At a taxi stand let taxis arrive at unit intervals of time (one at a time). If no one is waiting at the stand, the taxi leaves immediately. Let η_k be the number of passengers who arrive at the stand at time k, and suppose that $\eta_1, \dots, \eta_n$ are independent random variables. Let ξ_k be the length of the waiting line at time k, $\xi_0 = 0$. Then if $\xi_k = i$, at the next time $k+1$ the length ξ_{k+1} of the waiting line is equal to

$$j = \begin{cases} \eta_{k+1} & \text{if } i = 0, \\ i - 1 + \eta_{k+1} & \text{if } i \geq 1. \end{cases}$$

In other words,

$$\xi_{k+1} = (\xi_k - 1)^+ + \eta_{k+1}, \qquad 0 \leq k \leq n-1,$$

where $a^+ = \max(a, 0)$, and therefore the sequence $\xi = (\xi_0, \ldots, \xi_n)$ is a Markov chain.

Example 4. This example comes from the theory of *branching processes*. A branching process with discrete times is a sequence of random variables $\xi_0, \xi_1, \ldots, \xi_n$, where ξ_k is interpreted as the number of particles in existence at time k, and the process of creation and annihilation of particles is as follows: each particle, independently of the other particles and of the "prehistory" of the process, is transformed into j particles with probability p_j, $j = 0, 1, \ldots, M$.

We suppose that at the initial time there is just one particle, $\xi_0 = 1$. If at time k there are ξ_k particles (numbered $1, 2, \ldots, \xi_k$), then by assumption ξ_{k+1} is given as a random sum of random variables,

$$\xi_{k+1} = \eta_1^{(k)} + \cdots + \eta_{\xi_k}^{(k)},$$

where $\eta_i^{(k)}$ is the number of particles produced by particle number i. It is clear that if $\xi_k = 0$ then $\xi_{k+1} = 0$. If we suppose that all the random variables $\eta_j^{(k)}$, $k \geq 0$, are independent of each other, we obtain

$$\begin{aligned} \mathsf{P}\{\xi_{k+1} = i_{k+1} | \xi_k = i_k, \xi_{k-1} = i_{k-1}, \ldots\} &= \mathsf{P}\{\xi_{k+1} = i_{k+1} | \xi_k = i_k\} \\ &= \mathsf{P}\{\eta_1^{(k)} + \cdots + \eta_{i_k}^{(k)} = i_{k+1}\}. \end{aligned}$$

It is evident from this that the sequence $\xi_0, \xi_1, \ldots, \xi_n$ is a Markov chain.

A particularly interesting case is that in which each particle either vanishes with probability q or divides in two with probability p, $p + q = 1$. In this case it is easy to calculate that

$$p_{ij} = \mathsf{P}\{\xi_{k+1} = j | \xi_k = i\}$$

is given by the formula

$$p_{ij} = \begin{cases} C_i^{j/2} p^{j/2} q^{i-j/2}, & j = 0, \ldots, 2i, \\ 0 & \text{in all other cases.} \end{cases}$$

2. Let $\xi = (\xi_k, \Pi, \mathbb{P})$ be a homogeneous Markov chain with starting vectors (rows) $\Pi = (p_i)$ and transition matrix $\mathbb{P} = \|p_{ij}\|$. It is clear that

$$p_{ij} = \mathsf{P}\{\xi_1 = j | \xi_0 = i\} = \cdots = \mathsf{P}\{\xi_n = j | \xi_{n-1} = i\}.$$

We shall use the notation

$$p_{ij}^{(k)} = \mathsf{P}\{\xi_k = j \mid \xi_0 = i\} \quad (= \mathsf{P}\{\xi_{k+l} = j \mid \xi_l = i\})$$

for the probability of a transition from state i to state j in k steps, and

$$p_j^{(k)} = \mathsf{P}\{\xi_k = j\}$$

for the probability of finding the particle at point j at time k. Also let

$$\Pi^{(k)} = \|p_i^{(k)}\|, \qquad \mathbb{P}^{(k)} = \|p_{ij}^{(k)}\|.$$

Let us show that the transition probabilities $p_{ij}^{(k)}$ satisfy the *Kolmogorov–Chapman equation*

$$p_{ij}^{(k+l)} = \sum_\alpha p_{i\alpha}^{(k)} p_{\alpha j}^{(l)}, \tag{13}$$

or, in matrix form,

$$\mathbb{P}^{(k+l)} = \mathbb{P}^{(k)} \cdot \mathbb{P}^{(l)} \tag{14}$$

The proof is extremely simple: using the formula for total probability and the Markov property, we obtain

$$\begin{aligned} p_{ij}^{(k+l)} &= \mathsf{P}(\xi_{k+l} = j \mid \xi_0 = i) = \sum_\alpha \mathsf{P}(\xi_{k+l} = j, \xi_k = \alpha \mid \xi_0 = i) \\ &= \sum_\alpha \mathsf{P}(\xi_{k+l} = j \mid \xi_k = \alpha)\mathsf{P}(\xi_k = \alpha \mid \xi_0 = i) = \sum_\alpha p_{\alpha j}^{(l)} p_{i\alpha}^{(k)}. \end{aligned}$$

The following two cases of (13) are particularly important:

the *backward equation*

$$p_{ij}^{(l+1)} = \sum_\nu p_{i\alpha} p_{\alpha j}^{(l)} \tag{15}$$

and the *forward equation*

$$p_{ij}^{(k+1)} = \sum_\alpha p_{i\alpha}^{(k)} p_{\alpha j} \tag{16}$$

(see Figures 22 and 23). The forward and backward equations can be written in the following matrix forms

$$\mathbb{P}^{(k+1)} = \mathbb{P}^{(k)} \cdot \mathbb{P}, \tag{17}$$

$$\mathbb{P}^{(k+1)} = \mathbb{P} \cdot \mathbb{P}^{(k)}. \tag{18}$$

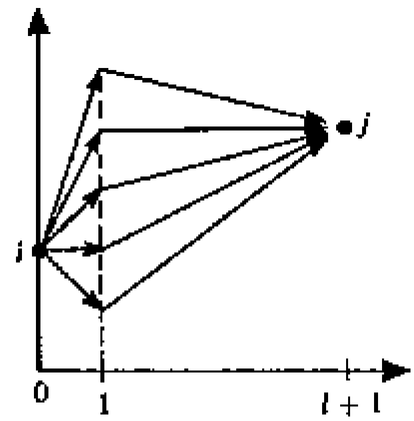


Figure 22. For the backward equation.

Similarly, we find for the (unconditional) probabilities $p_j^{(k)}$ that

$$p_j^{(k+l)} = \sum_\alpha p_\alpha^{(k)} p_{\alpha j}^{(l)}, \tag{19}$$

or in matrix form

$$\Pi^{(k+l)} = \Pi^{(k)} \cdot \mathbb{P}^{(l)}.$$

In particular,

$$\Pi^{(k+1)} = \Pi^{(k)} \cdot \mathbb{P}$$

(*forward equation*) and

$$\Pi^{(k+1)} = \Pi^{(1)} \cdot \mathbb{P}^{(k)}$$

(*backward equation*). Since $\mathbb{P}^{(1)} = \mathbb{P}, \Pi^{(1)} = \Pi$, it follows from these equations that

$$\mathbb{P}^{(k)} = \mathbb{P}^k, \qquad \Pi^{(k)} = \Pi^k.$$

Consequently for homogeneous Markov chains the k-step transition probabilities $p_{ij}^{(k)}$ are the elements of the kth powers of the matrix $\mathbb{P}$, so that many properties of such chains can be investigated by the methods of matrix analysis.

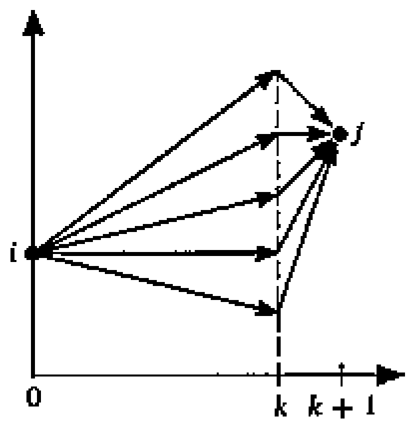


Figure 23. For the forward equation.

EXAMPLE 5. Consider a homogeneous Markov chain with the two states 0 and 1 and the matrix

$$\mathbb{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

It is easy to calculate that

$$\mathbb{P}^2 = \begin{pmatrix} p_{00}^2 + p_{01}p_{10} & p_{01}(p_{00} + p_{11}) \\ p_{10}(p_{00} + p_{11}) & p_{11}^2 + p_{01}p_{10} \end{pmatrix}$$

and (by induction)

$$\mathbb{P}^n = \frac{1}{2 - p_{00} - p_{11}} \begin{pmatrix} 1 - p_{11} & 1 - p_{00} \\ 1 - p_{11} & 1 - p_{00} \end{pmatrix} + \frac{(p_{00} + p_{11} - 1)^n}{2 - p_{00} - p_{11}} \begin{pmatrix} 1 - p_{00} & -(1 - p_{00}) \\ -(1 - p_{11}) & 1 - p_{11} \end{pmatrix}$$

(under the hypothesis that $|p_{00} + p_{11} - 1| < 1$).

Hence it is clear that if the elements of $\mathbb{P}$ satisfy $|p_{00} + p_{11} - 1| < 1$ (in particular, if all the transition probabilities p_{ij} are positive), then as $n \to \infty$

$$\mathbb{P}^n \to \frac{1}{2 - p_{00} - p_{11}} \begin{pmatrix} 1 - p_{11} & 1 - p_{00} \\ 1 - p_{11} & 1 - p_{00} \end{pmatrix}, \tag{20}$$

and therefore

$$\lim_n p_{i0}^{(n)} = \frac{1 - p_{11}}{2 - p_{00} - p_{11}}, \qquad \lim_n p_{i1}^{(n)} = \frac{1 - p_{00}}{2 - p_{00} - p_{11}}.$$

Consequently if $|p_{00} + p_{11} - 1| < 1$, such a Markov chain exhibits regular behavior of the following kind: the influence of the initial state on the probability of finding the particle in one state or another eventually becomes negligible ($p_{ij}^{(n)}$ approach limits π_j, *independent of* i and forming a probability distribution: $\pi_0 \geq 0$, $\pi_1 \geq 0$, $\pi_0 + \pi_1 = 1$); if also all $p_{ij} > 0$ then $\pi_0 > 0$ and $\pi_1 > 0$.

3. The following theorem describes a wide class of Markov chains that have the property called *ergodicity*: the limits $\pi_j = \lim_n p_{ij}$ not only exist, are independent of i, and form a probability distribution ($\pi_j \geq 0$, $\sum_j \pi_j = 1$), but also $\pi_j > 0$ for all j (such a distribution π_j is said to be *ergodic*).

Theorem 1 (Ergodic Theorem). *Let* $\mathbb{P} = \|p_{ij}\|$ *be the transition matrix of a chain with a finite state space* $X = \{1, 2, \ldots, N\}$.

(a) *If there is an* n_0 *such that*

$$\min_{i,j} p_{ij}^{(n_0)} > 0, \tag{21}$$

then there are numbers $\pi_1, \ldots, \pi_N$ *such that*

$$\pi_j > 0, \qquad \sum_j \pi_j = 1 \tag{22}$$

and

$$p_{ij}^{(n)} \to \pi_j, \qquad n \to \infty \tag{23}$$

for every $i \in X$.

(b) *Conversely, if there are numbers* $\pi_1, \ldots, \pi_N$ *satisfying* (22) *and* (23), *there is an* n_0 *such that* (21) *holds.*

(c) *The numbers* $(\pi_1, \ldots, \pi_N)$ *satisfy the equations*

$$\pi_j = \sum_\alpha \pi_\alpha p_{\alpha j}, \qquad j = 1, \ldots, N. \tag{24}$$

PROOF. (a) Let

$$m_j^{(n)} = \min_i p_{ij}^{(n)}, \qquad M_j^{(n)} = \max_i p_{ij}^{(n)}.$$

Since

$$p_{ij}^{(n+1)} = \sum_\alpha p_{i\alpha} p_{\alpha j}^{(n)}, \tag{25}$$

we have

$$m_j^{(n+1)} = \min_i p_{ij}^{(n+1)} = \min_i \sum_\alpha p_{i\alpha} p_{\alpha j}^{(n)} \ge \min_i \sum_\alpha p_{i\alpha} \min_\alpha p_{\alpha j}^{(n)} = m_j^{(n)},$$

whence $m_j^{(n)} \le m_j^{(n+1)}$ and similarly $M_j^{(n)} \ge M_j^{(n+1)}$. Consequently, to establish (23) it will be enough to prove that

$$M_j^{(n)} - m_j^{(n)} \to 0, \qquad n \to \infty, \quad j = 1, \ldots, N.$$

Let $\varepsilon = \min_{i,j} p_{ij}^{(n_0)} > 0$. Then

$$p_{ij}^{(n_0+n)} = \sum_\alpha p_{i\alpha}^{(n_0)} p_{\alpha j}^{(n)} = \sum_\alpha [p_{i\alpha}^{(n_0)} - \varepsilon p_{j\alpha}^{(n)}] p_{\alpha j}^{(n)} + \varepsilon \sum_\alpha p_{j\alpha}^{(n)} p_{\alpha j}^{(n)}$$
$$= \sum_\alpha [p_{i\alpha}^{(n_0)} - \varepsilon p_{j\alpha}^{(n)}] p_{\alpha j}^{(n)} + \varepsilon p_{jj}^{(2n)}.$$

But $p_{i\alpha}^{(n_0)} - \varepsilon p_{j\alpha}^{(n)} \ge 0$; therefore

$$p_{ij}^{(n_0+n)} \ge m_j^{(n)} \cdot \sum_\alpha [p_{i\alpha}^{(n_0)} - \varepsilon p_{j\alpha}^{(n)}] + \varepsilon p_{jj}^{(2n)} = m_j^{(n)}(1 - \varepsilon) + \varepsilon p_{jj}^{(2n)},$$

and consequently

$$m_j^{(n_0+n)} \ge m_j^{(n)}(1 - \varepsilon) + \varepsilon p_{jj}^{(2n)}.$$

In a similar way

$$M_j^{(n_0+n)} \le M_j^{(n)}(1 - \varepsilon) + \varepsilon p_{jj}^{(2n)}.$$

Combining these inequalities, we obtain

$$M_j^{(n_0+n)} - m_j^{(n_0+n)} \le (M_j^{(n)} - m_j^{(n)}) \cdot (1 - \varepsilon)$$

and consequently

$$M_j^{(kn_0+n)} - m_j^{(kn_0+n)} \le (M_j^{(n)} - m_j^{(n)})(1-\varepsilon)^k \downarrow 0, \qquad k \to \infty.$$

Thus $M_j^{(n_\beta)} - m_j^{(n_\beta)} \to 0$ for some subsequence n_β, $n_\beta \to \infty$. But the difference $M_j^{(n)} - m_j^{(n)}$ is monotonic in n, and therefore $M_j^{(n)} - m_j^{(n)} \to 0$, $n \to \infty$.

If we put $\pi_j = \lim_n m_j^{(n)}$, it follows from the preceding inequalities that

$$|p_{ij}^{(n)} - \pi_j| \le M_j^{(n)} - m_j^{(n)} \le (1-\varepsilon)^{[n/n_0]-1}$$

for $n \ge n_0$, that is, $p_{ij}^{(n)}$ converges to its limit π_j geometrically (i.e., as fast as a geometric progression).

It is also clear that $m_j^{(n)} \ge m_j^{(n_0)} \ge \varepsilon > 0$ for $n \ge n_0$, and therefore $\pi_j > 0$.

(b) Inequality (21) follows from (23) and (25).

(c) Equation (24) follows from (23) and (25).

This completes the proof of the theorem.

4. Equations (24) play a major role in the theory of Markov chains. A nonnegative solution $(\pi_1, \ldots, \pi_N)$ satisfying $\sum_\alpha \pi_\alpha = 1$ is said to be a *stationary* or *invariant* probability distribution for the Markov chain with transition matrix $\|p_{ij}\|$. The reason for this terminology is as follows.

Let us select an initial distribution $(\pi_1, \ldots, \pi_N)$ and take $p_j = \pi_j$. Then

$$p_j^{(1)} = \sum_\alpha \pi_\alpha p_{\alpha j} = \pi_j$$

and in general $p_j^{(n)} = \pi_j$. In other words, if we take $(\pi_1, \ldots, \pi_N)$ as the initial distribution, this distribution is unchanged as time goes on, i.e. for any k

$$\mathsf{P}(\xi_k = j) = \mathsf{P}(\xi_0 = j), \qquad j = 1, \ldots, N.$$

Moreover, with this initial distribution the Markov chain $\xi = (\xi, \Pi, \mathsf{P})$ is really *stationary*: the joint distribution of the vector $(\xi_k, \xi_{k+1}, \ldots, \xi_{k+l})$ is independent of k for all l (assuming that $k + l \le n$).

Property (21) guarantees both the existence of limits $\pi_j = \lim p_{ij}^{(n)}$, which are independent of i, and the existence of an ergodic distribution, i.e. one with $\pi_j > 0$. The distribution $(\pi_1, \ldots, \pi_N)$ is also a *stationary* distribution. Let us now show that the set $(\pi_1, \ldots, \pi_N)$ is the *only* stationary distribution.

In fact, let $(\tilde{\pi}_1, \ldots, \tilde{\pi}_N)$ be another stationary distribution. Then

$$\tilde{\pi}_j = \sum_\alpha \tilde{\pi}_\alpha p_{\alpha j} = \cdots = \sum_\alpha \tilde{\pi}_\alpha p_{\alpha j}^{(n)},$$

and since $p_{\alpha j}^{(n)} \to \pi_j$ we have

$$\tilde{\pi}_j = \sum_\alpha (\tilde{\pi}_\alpha \cdot \pi_j) = \pi_j.$$

These problems will be investigated in detail in Chapter VIII for Markov chains with countably many states as well as with finitely many states.

We note that a stationary probability distribution (even unique) may exist for a nonergodic chain. In fact, if

$$\mathbb{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

then

$$\mathbb{P}^{2n} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \mathbb{P}^{2n+1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and consequently the limits $\lim p_{ij}^{(n)}$ do not exist. At the same time, the system

$$\pi_j = \sum_{\alpha} \pi_\alpha p_{\alpha j}, \qquad j = 1, 2,$$

reduces to

$$\pi_1 = \pi_2,$$
$$\pi_2 = \pi_1,$$

of which the unique solution satisfying $\pi_1 + \pi_2 = 1$ is $(\frac{1}{2}, \frac{1}{2})$.

We also notice that for this example the system (24) has the form

$$\pi_0 = \pi_0 p_{00} + \pi_1 p_{10},$$
$$\pi_1 = \pi_0 p_{01} + \pi_1 p_{11},$$

from which, by the condition $\pi_0 = \pi_1 = 1$, we find that the unique stationary distribution (π_0, π_1) coincides with the one obtained above:

$$\pi_0 = \frac{1 - p_{11}}{2 - p_{00} - p_{11}}, \qquad \pi_1 = \frac{1 - p_{00}}{2 - p_{00} - p_{11}}.$$

We now consider some corollaries of the ergodic theorem.

Let A be a set of states, $A \subseteq X$ and

$$I_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

Consider

$$v_A(n) = \frac{I_A(\xi_0) + \cdots + I_A(\xi_n)}{n + 1}$$

which is the fraction of the time spent by the particle in the set A. Since

$$\mathsf{E}[I_A(\xi_k) \mid \xi_0 = i] = \mathsf{P}(\xi_k \in A \mid \xi_0 = i) = \sum_{j \in A} p_{ij}^{(k)} (= p_i^{(k)}(A)),$$

we have

$$\mathsf{E}[\nu_A(n) \mid \xi_0 = i] = \frac{1}{n+1} \sum_{k=0}^{n} p_i^{(k)}(A)$$

and in particular

$$\mathsf{E}[\nu_{\{j\}}(n) \mid \xi_0 = i] = \frac{1}{n+1} \sum_{k=0}^{n} p_{ij}^{(k)}.$$

It is known from analysis (see also Lemma 1 in §3 of Chapter IV) that if $a_n \to a$ then $(a_0 + \cdots + a_n)/(n+1) \to a$, $n \to \infty$. Hence if $p_{ij}^{(k)} \to \pi_j$, $k \to \infty$, then

$$\mathsf{E}\nu_{\{j\}}(n) \to \pi_j, \qquad \mathsf{E}\nu_A(n) \to \pi_A, \qquad \text{where} \quad \pi_A = \sum_{j \in A} \pi_j.$$

For ergodic chains one can in fact prove more, namely that the following result holds for $I_A(\xi_0), \ldots, I_A(\xi_n), \ldots$.

Law of Large Numbers. *If $\xi_0, \xi_1, \ldots$ form a finite ergodic Markov chain, then*

$$\mathsf{P}\{|\nu_A(n) - \pi_A| > \varepsilon\} \to 0, \qquad n \to \infty, \tag{26}$$

for every $\varepsilon > 0$ and every initial distribution.

Before we undertake the proof, let us notice that we cannot apply the results of §5 directly to $I_A(\xi_0), \ldots, I_A(\xi_n), \ldots$, since these variables are, in general, dependent. However, the proof can be carried through along the same lines as for independent variables if we again use Chebyshev's inequality, and apply the fact that for an ergodic chain with finitely many states there is a number ρ, $0 < \rho < 1$, such that

$$|p_{ij}^{(n)} - \pi_j| \le C \cdot \rho^n. \tag{27}$$

Let us consider states i and j (which might be the same) and show that, for $\varepsilon > 0$,

$$\mathsf{P}\{|\nu_{\{j\}}(n) - \pi_j| > \varepsilon \mid \xi_0 = i\} \to 0, \qquad n \to \infty. \tag{28}$$

By Chebyshev's inequality,

$$\mathsf{P}\{|\nu_{\{j\}}(n) - \pi_j| > \varepsilon \mid \xi_0 = i\} < \frac{\mathsf{E}\{|\nu_{\{j\}}(n) - \pi_j|^2 \mid \xi_0 = i\}}{\varepsilon^2}.$$

Hence we have only to show that

$$\mathsf{E}\{|\nu_{\{j\}}(n) - \pi_j|^2 \mid \xi_0 = i\} \to 0, \qquad n \to \infty.$$

A simple calculation shows that

$$\mathsf{E}\{|\nu_{\{j\}}(n) - \pi_j|^2 | \xi_0 = i\} = \frac{1}{(n+1)^2} \cdot \mathsf{E}\left\{\left[\sum_{k=0}^{n} (I_{\{j\}}(\xi_k) - \pi_j)\right]^2 \middle| \xi_0 = i\right\}$$
$$= \frac{1}{(n+1)^2} \sum_{k=0}^{n} \sum_{l=0}^{n} m_{ij}^{(k,l)},$$

where

$$m_{ij}^{(k,l)} = \mathsf{E}\{[I_{\{j\}}(\xi_k) I_{\{j\}}(\xi_l)] | \xi_0 = i\}$$
$$- \pi_j \cdot \mathsf{E}[I_{\{j\}}(\xi_k) | \xi_0 = i] - \pi_j \cdot \mathsf{E}[I_{\{j\}}(\xi_l) | \xi_0 = i] + \pi_j^2$$
$$= p_{ij}^{(s)} \cdot p_{jj}^{(t)} - \pi_j \cdot p_{ij}^{(k)} - \pi_j \cdot p_{ij}^{(l)} + \pi_j^2,$$
$$s = \min(k, l) \quad \text{and} \quad t = |k - l|.$$

By (27),

$$p_{ij}^{(n)} = \pi_j + \varepsilon_{ij}^{(n)}, \qquad |\varepsilon_{ij}^{(n)}| \le C\rho^n.$$

Therefore

$$|m_{ij}^{(k,l)}| \le C_1[\rho^s + \rho^t + \rho^k + \rho^l],$$

where C_1 is a constant. Consequently

$$\frac{1}{(n+1)^2} \sum_{k=0}^{n} \sum_{l=0}^{n} m_{ij}^{(k,l)} \le \frac{C_1}{(n+1)^2} \sum_{k=0}^{n} \sum_{l=0}^{n} [\rho^s + \rho^t + \rho^k + \rho^l]$$
$$\le \frac{4C_1}{(n+1)^2} \cdot \frac{2(n+1)}{1-\rho} = \frac{8C_1}{(n+1)(1-\rho)} \to 0, \quad n \to \infty.$$

Then (28) follows from this, and we obtain (26) in an obvious way.

5. In §9 we gave, for a random walk $S_0, S_1, \ldots$ generated by a Bernoulli scheme, recurrent equations for the probability and the expectation of the exit time at either boundary. We now derive similar equations for Markov chains.

Let $\xi = (\xi_0, \ldots, \xi_n)$ be a Markov chain with transition matrix $\|p_{ij}\|$ and phase space $X = \{0, \pm 1, \ldots, \pm N\}$. Let A and B be two integers, $-N \le A \le 0 \le B \le N$, and $x \in X$. Let $\mathscr{B}_{k+1}$ be the set of paths $(x_0, x_1, \ldots, x_k)$, $x_i \in X$, that leave the interval (A, B) for the first time at the upper end, i.e. leave (A, B) by going into the set $(B, B+1, \ldots, N)$.

For $A \le x \le B$, put

$$\beta_k(x) = \mathsf{P}\{(\xi_0, \ldots, \xi_k) \in \mathscr{B}_{k+1} | \xi_0 = x\}.$$

In order to find these probabilities (for the first exit of the Markov chain from (A, B) through the upper boundary) we use the method that was applied in the deduction of the backward equations.

We have

$$\beta_k(x) = \mathsf{P}\{(\xi_0, \dots, \xi_k) \in \mathscr{B}_{k+1} | \xi_0 = x\}$$
$$= \sum_y p_{xy} \cdot \mathsf{P}\{(\xi_0, \dots, \xi_k) \in \mathscr{B}_{k+1} | \xi_0 = x, \xi_1 = y\},$$

where, as is easily seen by using the Markov property and the homogeneity of the chain,

$$\mathsf{P}\{(\xi_0, \dots, \xi_k) \in \mathscr{B}_{k+1} | \xi_0 = x, \xi_1 = y\}$$
$$= \mathsf{P}\{(x, y, \xi_2, \dots, \xi_k) \in \mathscr{B}_{k+1} | \xi_0 = x, \xi_1 = y\}$$
$$= \mathsf{P}\{(y, \xi_2, \dots, \xi_k) \in \mathscr{B}_k | \xi_1 = y\}$$
$$= \mathsf{P}\{(y, \xi_1, \dots, \xi_{k-1}) \in \mathscr{B}_k | \xi_0 = y\} = \beta_{k-1}(y).$$

Therefore

$$\beta_k(x) = \sum_y p_{xy} \beta_{k-1}(y)$$

for $A < x < B$ and $1 \leq k \leq n$. Moreover, it is clear that

$$\beta_k(x) = 1, \qquad x = B, B + 1, \dots, N,$$

and

$$\beta_k(x) = 0, \qquad x = -N, \dots, A.$$

In a similar way we can find equations for $\alpha_k(x)$, the probabilities for first exit from (A, B) through the lower boundary.

Let $\tau_k = \min\{0 \leq l \leq k: \xi_l \notin (A, B)\}$, where $\tau_k = k$ if the set $\{\cdot\} = \varnothing$. Then the same method, applied to $m_k(x) = \mathsf{E}(\tau_k | \xi_0 = x)$, leads to the following recurrent equations:

$$m_k(x) = 1 + \sum_y m_{k-1}(y) p_{xy}$$

(here $1 \leq k \leq n$, $A < x < B$). We define

$$m_k(x) = 0, \qquad x \notin (A, B).$$

It is clear that if the transition matrix is given by (11) the equations for $\alpha_k(x)$, $\beta_k(x)$ and $m_k(x)$ become the corresponding equations from §9, where they were obtained by essentially the same method that was used here.

These equations have the most interesting applications in the limiting case when the walk continues for an unbounded length of time. Just as in §9, the corresponding equations can be obtained by a formal limiting process ($k \to \infty$).

By way of example, we consider the Markov chain with states $\{0, 1, \dots, B\}$ and transition probabilities

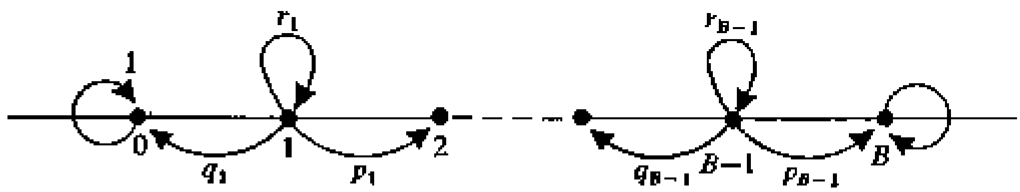
$$p_{00} = 1, \qquad p_{BB} = 1,$$

and

$$p_{ij} = \begin{cases} p_i > 0, & j = i + 1, \\ r_i, & j = i, \\ q_i > 0, & j = i - 1, \end{cases}$$

for $1 \le i \le B - 1$, where $p_i + q_i + r_i = 1$.

For this chain, the corresponding graph is



It is clear that states 0 and B are absorbing, whereas for every other state i the particle stays there with probability r_i, moves one step to the right with probability p_i, and to the left with probability q_i.

Let us find $\alpha(x) = \lim_{k\to\infty} \alpha_k(x)$, the limit of the probability that a particle starting at the point x arrives at state zero before reaching state B. Taking limits as $k \to \infty$ in the equations for $\alpha_k(x)$, we find that

$$\alpha(j) = q_j\alpha(j-1) + r_j\alpha(j) + p_j\alpha(j+1)$$

when $0 < j < B$, with the boundary conditions

$$\alpha(0) = 1, \qquad \alpha(B) = 0.$$

Since $r_j - 1 - q_j - p_j$, we have

$$p_j(\alpha(j+1) - \alpha(j)) = q_j(\alpha(j) - \alpha(j-1))$$

and consequently

$$\alpha(j+1) - \alpha(j) = \rho_j(\alpha(1) - 1),$$

where

$$\rho_j = \frac{q_1 \cdots q_j}{p_1 \cdots p_j}, \qquad \rho_0 = 1.$$

But

$$\alpha(j+1) - 1 = \sum_{i=0}^{j} (\alpha(i+1) - \alpha(i)).$$

Therefore

$$\alpha(j+1) - 1 = (\alpha(1) - 1)\cdot \sum_{i=0}^{j} \rho_i.$$

If $j = B - 1$, we have $\alpha(j + 1) = \alpha(B) = 0$, and therefore

$$\alpha(1) = 1 = -\frac{1}{\sum_{i=1}^{B-1} \rho_i},$$

whence

$$\alpha(1) = \frac{\sum_{i=1}^{B-1} \rho_i}{\sum_{i=0}^{B-1} \rho_i} \quad \text{and} \quad \alpha(j) = \frac{\sum_{i=j}^{B-1} \rho_i}{\sum_{i=1}^{B-1} \rho_i}, \qquad j = 1, \ldots, B.$$

(This should be compared with the results of §9.)

Now let $m(x) = \lim_k m_k(x)$, the limiting value of the average time taken to arrive at one of the states 0 or B. Then $m(0) = m(B) = 0$,

$$m(x) = 1 + \sum_y m(y)p_{xy}$$

and consequently for the example that we are considering,

$$m(j) = 1 + q_j m(j - 1) + r_j m(j) + p_j m(j + 1)$$

for $j = 1, 2, \ldots, B - 1$. To find $m(j)$ we put

$$M(j) = m(j) - m(j - 1), \qquad j = 0, 1, \ldots, B.$$

Then

$$p_j M(j + 1) = q_j M(j) - 1, \qquad j = 1, \ldots, B - 1,$$

and consequently we find that

$$M(j + 1) = \rho_j M(1) - R_j,$$

where

$$\rho_j = \frac{q_1 \cdots q_j}{p_1 \cdots p_j}, \qquad R_j = \frac{1}{p_j}\left[1 + \frac{q_j}{p_{j-1}} + \cdots + \frac{q_j \cdots q_2}{p_j \cdots p_1}\right].$$

Therefore

$$m(i) = m(j) - m(0) = \sum_{i=0}^{j-1} M(i + 1)$$
$$= \sum_{i=0}^{j-1} (\rho_i m(1) - R_i) = m(1) \sum_{i=0}^{j-1} \rho_i - \sum_{i=0}^{j-1} R_i.$$

It remains only to determine $m(1)$. But $m(B) = 0$, and therefore

$$m(1) = \frac{\sum_{i=0}^{B-1} R_i}{\sum_{i=0}^{B-1} \rho_i},$$

and for $1 < j \leq B$,

$$m(j) = \sum_{i=0}^{j-1} \rho_i \cdot \frac{\sum_{i=0}^{B-1} R_i}{\sum_{i=0}^{B-1} \rho_i} - \sum_{i=0}^{j-1} R_i.$$

(This should be compared with the results in §9 for the case $r_i = 0$, $p_i = p$, $q_i = q$.)

6. In this subsection we consider a stronger version of the Markov property (8), namely that it remains valid if time k is replaced by a random time (see also Theorem 2). The significance of this, the *strong Markov property*, can be illustrated in particular by the example of the derivation of the recurrent relations (38), which play an important role in the classification of the states of Markov chains (Chapter VIII).

Let $\xi = (\xi_1, \dots, \xi_n)$ be a homogeneous Markov chain with transition matrix $\|p_{ij}\|$; let $\mathscr{D}^\xi = (\mathscr{D}_k^\xi)_{0 \le k \le n}$ be a system of decompositions, $\mathscr{D}_k^\xi = \mathscr{D}_{\xi_0, \dots, \xi_k}$. Let $\mathscr{B}_k^\xi$ denote the algebra $\alpha(\mathscr{D}_k^\xi)$ generated by the decomposition $\mathscr{D}_k^\xi$.

We first put the Markov property (8) into a somewhat different form. Let $B \in \mathscr{B}_k^\xi$. Let us show that then

$$\begin{aligned} \mathsf{P}\{\xi_n = a_n, \dots, \xi_{k+1} = a_{k+1} | B \cap (\xi_k = a_k)\} \\ = \mathsf{P}\{\xi_n = a_n, \dots, \xi_{k+1} = a_{k+1} | \xi_k = a_k\} \end{aligned} \tag{29}$$

(assuming that $\mathsf{P}\{B \cap (\xi_k = a_k)\} > 0$). In fact, B can be represented in the form

$$B = \sum\nolimits^* \{\xi_0 = a_0^*, \dots, \xi_k = a_k^*\},$$

where $\sum^*$ extends over some set $(a_0^*, \dots, a_k^*)$. Consequently

$$\begin{aligned} &\mathsf{P}\{\xi_n = a_n, \dots, \xi_{k+1} = a_{k+1} | B \cap (\xi_k = a_k)\} \\ &= \frac{\mathsf{P}\{(\xi_n = a_n, \dots, \xi_k = a_k) \cap B\}}{\mathsf{P}\{(\xi_k = a_k) \cap B\}} \\ &= \frac{\sum^* \mathsf{P}\{(\xi_n = a_n, \dots, \xi_k = a_k) \cap (\xi_0 = a_0^*, \dots, \xi_k = a_k^*)\}}{\mathsf{P}\{(\xi_k = a_k) \cap B\}}. \end{aligned} \tag{30}$$

But, by the Markov property,

$$\begin{aligned} &\mathsf{P}\{(\xi_n = a_n, \dots, \xi_k = a_k) \cap (\xi_0 = a_0^*, \dots, \xi_k = a_k^*)\} \\ &= \begin{cases} \mathsf{P}\{\xi_n = a_n, \dots, \xi_{k+1} = a_{k+1} | \xi_0 = a_0^*, \dots, \xi_k = a_k^*\} \\ \qquad \times \mathsf{P}\{\xi_0 = a_0^*, \dots, \xi_k = a_k^*\} \quad \text{if } a_k = a_k^*, \\ 0 \quad \text{if } a_k \ne a_k^*, \end{cases} \\ &= \begin{cases} \mathsf{P}\{\xi_n = a_n, \dots, \xi_{k+1} = a_{k+1} | \xi_k = a_k\} \mathsf{P}\{\xi_0 = a_0^*, \dots, \xi_k = a_k^*\} \\ \quad \text{if } a_k = a_k^*, \\ 0 \quad \text{if } a_k \ne a_k^*, \end{cases} \\ &= \begin{cases} \mathsf{P}\{\xi_n = a_n, \dots, \xi_{k+1} = a_{k+1} | \xi_k = a_k\} \mathsf{P}\{(\xi_k = a_k) \cap B\} \\ \quad \text{if } a_k = a_k^*, \\ 0 \quad \text{if } a_k \ne a_k^*. \end{cases} \end{aligned}$$

Therefore the sum $\sum^*$ in (30) is equal to

$$\mathsf{P}\{\xi_n = a_n, \ldots, \xi_{k+1} = a_{k+1} | \xi_k = a_k\}\mathsf{P}\{(\xi_k = a_k) \cap B\}.$$

This establishes (29).

Let τ be a stopping time (with respect to the system $D^\xi = (D_k^\xi)_{0 \le k \le n}$; see Definition 2 in §11).

Definition. We say that a set B in the algebra $\mathscr{B}_n^\xi$ belongs to the system of sets $\mathscr{B}_\tau^\xi$ if, for each k, $0 \le k \le n$,

$$B \cap \{\tau = k\} \in \mathscr{B}_k^\xi. \tag{31}$$

It is easily verified that the collection of such sets B forms an algebra (called the algebra of events observed at time τ).

Theorem 2. *Let $\xi = (\xi_0, \ldots, \xi_n)$ be a homogeneous Markov chain with transition matrix $\|p_{ij}\|$, τ a stopping time (with respect to $\mathscr{D}^\xi$), $B \in \mathscr{B}_\tau^\xi$ and $A = \{\omega: \tau + l \le n\}$. Then if $P\{A \cap B \cap (\xi_\tau = a_0)\} > 0$, we have*

$$\begin{aligned}\mathsf{P}\{\xi_{\tau+l} = a_l, \ldots, \xi_{\tau+1} = a_1 | A \cap B \cap (\xi_\tau = a_0)\}\\ = \mathsf{P}\{\xi_{\tau+l} = a_l, \ldots, \xi_{\tau+1} = a_1 | A \cap (\xi_\tau = a_0)\},\end{aligned} \tag{32}$$

and if $P\{A \cap (\xi_\tau = a_0)\} > 0$ then

$$\mathsf{P}\{\xi_{\tau+l} = a_l, \ldots, \xi_{\tau+1} = a_1 | A \cap (\xi_\tau = a_0)\} = p_{a_0 a_1} \cdots p_{a_{l-1} a_l}. \tag{33}$$

For the sake of simplicity, we give the proof only for the case $l = 1$. Since $B \cap (\tau = k) \in \mathscr{B}_k^\xi$, we have, according to (29),

$$\begin{aligned}&\mathsf{P}\{\xi_{\tau+1} = a_1, A \cap B \cap (\xi_\tau = a_0)\}\\ &\quad= \sum_{k \le n-1} \mathsf{P}\{\xi_{k+1} = a_1, \xi_k = a_0, \tau = k, B\}\\ &\quad= \sum_{k \le n-1} \mathsf{P}\{\xi_{k+1} = a_1 | \xi_k = a_0, \tau = k, B\}\mathsf{P}\{\xi_k = a_0, \tau = k, B\}\\ &\quad= \sum_{k \le n-1} \mathsf{P}\{\xi_{k+1} = a_1 | \xi_k = a_0\}\mathsf{P}\{\xi_k = a_0, \tau = k, B\}\\ &\quad= p_{a_0 a_1} \cdot \sum_{k \le n-1} \mathsf{P}\{\xi_k = a_0, \tau = k, B\} = p_{a_0 a_1} \cdot \mathsf{P}\{A \cap B \cap (\xi_\tau = a_0)\},\end{aligned}$$

which simultaneously establishes (32) and (33) (for (33) we have to take $B = \Omega$).

Remark. When $l = 1$, the strong Markov property (32), (33) is evidently equivalent to the property that

$$\mathsf{P}\{\xi_{\tau+1} \in C | A \cap B \cap (\xi_\tau = a_0)\} = P_{a_0}(C), \tag{34}$$

for every $C \subseteq X$, where

$$P_{a_0}(C) = \sum_{a_1 \in C} p_{a_0 a_1}.$$

In turn, (34) can be restated as follows: on the set $A = \{\tau \le n - 1\}$,

$$\mathsf{P}\{\xi_{\tau+1} \in C \mid \mathscr{B}_\tau^\xi\} = P_{\xi_\tau}(C), \tag{35}$$

which is a form of the strong Markov property that is commonly used in the general theory of homogeneous Markov processes.

7. Let $\xi = (\xi_0, \ldots, \xi_n)$ be a homogeneous Markov chain with transition matrix $\|p_{ij}\|$, and let

$$f_{ii}^{(k)} = \mathsf{P}\{\xi_k = i, \xi_l \ne i, 1 \le l \le k - 1 \mid \xi_0 = i\} \tag{36}$$

and

$$f_{ij}^{(k)} = \mathsf{P}\{\xi_k = j, \xi_l \ne j, 1 \le l \le k - 1 \mid \xi_0 = i\} \tag{37}$$

for $i \ne j$ be respectively the probability of first return to state i at time k and the probability of first arrival at state j at time k.

Let us show that

$$p_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)}, \qquad \text{where} \quad p_{jj}^{(0)} = 1. \tag{38}$$

The intuitive meaning of the formula is clear: to go from state i to state j in n steps, it is necessary to reach state j for the first time in k steps $(1 \le k \le n)$ and then to go from state j to state j in $n - k$ steps. We now give a rigorous derivation.

Let j be given and

$$\tau = \min\{1 \le k \le n : \xi_k = j\},$$

assuming that $\tau = n + 1$ if $\{\cdot\} = \varnothing$. Then $f_{ij}^{(k)} = \mathsf{P}\{\tau = k \mid \xi_0 = i\}$ and

$$\begin{aligned} p_{ij}^{(n)} &= \mathsf{P}\{\xi_n = j \mid \xi_0 = i\} \\ &= \sum_{1 \le k \le n} \mathsf{P}\{\xi_n = j, \tau = k \mid \xi_0 = i\} \\ &= \sum_{1 \le k \le n} \mathsf{P}\{\xi_{\tau+n-k} = j, \tau = k \mid \xi_0 = i\}, \end{aligned} \tag{39}$$

where the last equation follows because $\xi_{\tau+n-k} = \xi_n$ on the set $\{\tau = k\}$. Moreover, the set $\{\tau = k\} = \{\tau = k, \xi_\tau = j\}$ for every k, $1 \le k \le n$. Therefore if $\mathsf{P}\{\xi_0 = i, \tau = k\} > 0$, it follows from Theorem 2 that

$$\begin{aligned} \mathsf{P}\{\xi_{\tau+n-k} = j \mid \xi_0 = i, \tau = k\} &= \mathsf{P}\{\xi_{\tau+n-k} = j \mid \xi_0 = i, \tau = k, \xi_\tau = j\} \\ &= \mathsf{P}\{\xi_{\tau+n-k} = j \mid \xi_\tau = j\} = p_{jj}^{(n-k)} \end{aligned}$$

and by (37)

$$p_{ij}^{(n)} = \sum_{k=1}^{n} \mathsf{P}\{\xi_{\tau+n-k} = j \mid \xi_0 = i, \tau = k\}\mathsf{P}\{\tau = k \mid \xi_0 = i\}$$

$$= \sum_{k=1}^{n} p_{jj}^{(n-k)} f_{ij}^{(k)},$$

which establishes (38).

8. Problems

1. Let $\xi = (\xi_0, \ldots, \xi_n)$ be a Markov chain with values in X and $f = f(x)$ $(x \in X)$ a function. Will the sequence $(f(\xi_0), \ldots, f(\xi_n))$ form a Markov chain? Will the "reversed" sequence

$$(\xi_n, \xi_{n-1}, \ldots, \xi_0)$$

form a Markov chain?

2. Let $\mathbb{P} = \|p_{ij}\|$, $1 \le i, j \le r$, be a stochastic matrix and λ an eigenvalue of the matrix, i.e. a root of the characteristic equation $\det\|\mathbb{P} - \lambda E\| = 0$. Show that $\lambda_0 = 1$ is an eigenvalue and that all the other eigenvalues have moduli not exceeding 1. If all the eigenvalues $\lambda_1, \ldots, \lambda_r$ are distinct, then $p_{ij}^{(k)}$ admits the representation

$$p_{ij}^{(k)} = \pi_j + a_{ij}(1)\lambda_1^k + \cdots + a_{ij}(r)\lambda_r^k,$$

where $\pi_j, a_{ij}(1), \ldots, a_{ij}(r)$ can be expressed in terms of the elements of $\mathbb{P}$. (It follows from this algebraic approach to the study of Markov chains that, in particular, when $|\lambda_1| < 1, \ldots, |\lambda_r| < 1$, the limit $\lim p_{ij}^{(k)}$ exists for every j and is independent of i.)

3. Let $\xi = (\xi_0, \ldots, \xi_n)$ be a homogeneous Markov chain with state space X and transition matrix $\mathbb{P} = \|p_{xy}\|$. Let

$$T\varphi(x) = \mathsf{E}[\varphi(\xi_1) \mid \xi_0 = x] \quad \left(= \sum_y \varphi(y)p_{xy}\right).$$

Let the nonnegative function φ satisfy

$$T\varphi(x) = \varphi(x), \qquad x \in X.$$

Show that the sequence of random variables

$$\zeta = (\zeta_k, \mathscr{D}_k^{\xi}) \qquad \text{with} \quad \zeta_k = \varphi(\xi_k)$$

is a martingale.

4. Let $\xi = (\xi_n, \Pi, \mathbb{P})$ and $\tilde{\xi} = (\tilde{\xi}_n, \tilde{\Pi}, \mathbb{P})$ be two Markov chains with different initial distributions $\Pi = (p_1, \ldots, p_r)$ and $\tilde{\Pi} = (\tilde{p}_1, \ldots, \tilde{p}_r)$. Show that if $\min_{i,j} p_{ij} \ge \varepsilon > 0$ then

$$\sum_{i=1}^{r} |\tilde{p}_i^{(n)} - p_i^{(n)}| \le 2(1 - \varepsilon)^n.$$

CHAPTER II

Mathematical Foundations of Probability Theory

§1. Probabilistic Model for an Experiment with Infinitely Many Outcomes. Kolmogorov's Axioms

1. The models introduced in the preceding chapter enabled us to give a probabilistic–statistical description of experiments with a finite number of outcomes. For example, the triple $(\Omega, \mathscr{A}, \mathsf{P})$ with

$$\Omega = \{\omega : \omega = (a_1, \ldots, a_n),\, a_i = 0, 1\},\ \mathscr{A} = \{A : A \subseteq \Omega\}$$

and $p(\omega) = p^{\Sigma a_i} q^{n - \Sigma a_i}$ is a model for the experiment in which a coin is tossed n times "independently" with probability p of falling head. In this model the number $N(\Omega)$ of outcomes, i.e. the number of points in Ω, is the finite number 2^n.

We now consider the problem of constructing a probabilistic model for the experiment consisting of an infinite number of independent tosses of a coin when at each step the probability of falling head is p.

It is natural to take the set of outcomes to be the set

$$\Omega = \{\omega : \omega = (a_1, a_2, \ldots),\, a_i = 0, 1\},$$

i.e. the space of sequences $\omega = (a_1, a_2, \ldots)$ whose elements are 0 or 1.

What is the cardinality $N(\Omega)$ of Ω? It is well known that every number $a \in [0, 1)$ has a unique binary expansion (containing an infinite number of zeros)

$$a = \frac{a_1}{2} + \frac{a_2}{2^2} + \cdots \qquad (a_i = 0, 1).$$

Hence it is clear that there is a one-to-one correspondence between the points ω of Ω and the points a of the set $[0, 1)$, and therefore Ω has the cardinality of the continuum.

Consequently if we wish to construct a probabilistic model to describe experiments like tossing a coin infinitely often, we must consider spaces Ω of a rather complicated nature.

We shall now try to see what probabilities ought reasonably to be assigned (or assumed) in a model of infinitely many independent tosses of a fair coin ($p + q = \frac{1}{2}$).

Since we may take Ω to be the set $[0, 1)$, our problem can be considered as the problem of choosing points at random from this set. For reasons of symmetry, it is clear that all outcomes ought to be equiprobable. But the set $[0, 1)$ is uncountable, and if we suppose that its probability is 1, then it follows that the probability $p(\omega)$ of each outcome certainly must equal zero. However, this assignment of probabilities ($p(\omega) = 0$, $\omega \in [0, 1)$) does not lead very far. The fact is that we are ordinarily not interested in the probability of one outcome or another, but in the probability that the result of the experiment is in one or another specified set A of outcomes (an event). In elementary probability theory we use the probabilities $p(\omega)$ to find the probability $\mathsf{P}(A)$ of the event A: $\mathsf{P}(A) = \sum_{\omega \in A} p(\omega)$. In the present case, with $p(\omega) = 0$, $\omega \in [0, 1)$, we cannot define, for example, the probability that a point chosen at random from $[0, 1)$ belongs to the set $[0, \frac{1}{2})$. At the same time, it is intuitively clear that this probability should be $\frac{1}{2}$.

These remarks should suggest that in constructing probabilistic models for uncountable spaces Ω we must assign probabilities, not to individual outcomes but to subsets of Ω. The same reasoning as in the first chapter shows that the collection of sets to which probabilities are assigned must be closed with respect to unions, intersections and complements. Here the following definition is useful.

Definition 1. Let Ω be a set of points ω. A system $\mathscr{A}$ of subsets of Ω is called an *algebra* if

(a) $\Omega \in \mathscr{A}$,

(b) $A, B \in \mathscr{A} \Rightarrow A \cup B \in \mathscr{A}, \qquad A \cap B \in \mathscr{A}$,

(c) $A \in \mathscr{A} \Rightarrow \bar{A} \in \mathscr{A}$

(Notice that in condition (b) it is sufficient to require only that either $A \cup B \in \mathscr{A}$ or that $A \cap B \in \mathscr{A}$, since $A \cup B = \overline{\bar{A} \cap \bar{B}}$ and $A \cap B = \overline{\bar{A} \cup \bar{B}}$.)

The next definition is needed in formulating the concept of a probabilistic model.

Definition 2. Let $\mathscr{A}$ be an algebra of subsets of Ω. A set function $\mu = \mu(A)$, $A \in \mathscr{A}$, taking values in $[0, \infty]$, is called a *finitely additive measure* defined on $\mathscr{A}$ if

$$\mu(A + B) = \mu(A) + \mu(B). \tag{1}$$

for every pair of disjoint sets A and B in $\mathscr{A}$.

A finitely additive measure μ with $\mu(\Omega) < \infty$ is called finite, and when $\mu(\Omega) = 1$ it is called a finitely additive probability measure, or a finitely additive probability.

2. We now define a probabilistic model (in the extended sense).

Definition 3. An ordered triple $(\Omega, \mathscr{A}, \mathsf{P})$, where

(a) Ω is a set of points ω;
(b) $\mathscr{A}$ is an algebra of subsets of Ω;
(c) P is a finitely additive probability on A,

is a *probabilistic model in the extended sense.*

It turns out, however, that this model is too broad to lead to a fruitful mathematical theory. Consequently we must restrict both the class of subsets of Ω that we consider, and the class of admissible probability measures.

Definition 4. A system $\mathscr{F}$ of subsets of Ω is a *σ-algebra* if it is an algebra and satisfies the following additional condition (stronger than (b) of Definition 1):

(b*) if $A_n \in \mathscr{F}$, $n = 1, 2, \ldots$, then

$$\bigcup A_n \in \mathscr{F}, \qquad \bigcap A_n \in \mathscr{F}$$

(it is sufficient to require either that $\bigcup A_n \in \mathscr{F}$ or that $\bigcap A_n \in \mathscr{F}$).

Definition 5. The space Ω together with a σ-algebra $\mathscr{F}$ of its subsets is a *measurable space*, and is denoted by $(\Omega, \mathscr{F})$.

Definition 6. A finitely additive measure μ defined on an algebra $\mathscr{A}$ of subsets of Ω is *countably additive* (or *σ-additive*), or simply a measure, if, for all pairwise disjoint subsets $A_1, A_2, \ldots$ of A with $\sum A_n \in \mathscr{A}$

$$\mu\left(\sum_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

A finitely additive measure μ is said to be σ-finite if Ω can be represented in the form

$$\Omega = \sum_{n=1}^{\infty} \Omega_n, \qquad \Omega_n \in \mathscr{A},$$

with $\mu(\Omega_n) < \infty$, $n = 1, 2, \ldots$.

If a countably additive measure P on the algebra A satisfies $\mathsf{P}(\Omega) = 1$, it is called a *probability measure* or a *probability* (defined on the sets that belong to the algebra $\mathscr{A}$).

Probability measures have the following properties.

If $\varnothing$ is the empty set then

$$\mathsf{P}(\varnothing) = 0.$$

If $A, B \in \mathscr{A}$ then

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B).$$

If $A, B \in \mathscr{A}$ and $B \subseteq A$ then

$$\mathsf{P}(B) \leq \mathsf{P}(A).$$

If $A_n \in \mathscr{A}$, $n = 1, 2, \ldots$, and $\bigcup A_n \in \mathscr{A}$, then

$$\mathsf{P}(A_1 \cup A_2 \cup \cdots) \leq \mathsf{P}(A_1) + \mathsf{P}(A_2) + \cdots.$$

The first three properties are evident. To establish the last one it is enough to observe that $\bigcup_{n=1}^{\infty} A_n = \sum_{n=1}^{\infty} B_n$, where $B_1 = A_1$, $B_n = \overline{A}_1 \cap \cdots \cap \overline{A}_{n-1} \cap A_n$, $n \geq 2$, $B_i \cap B_j = \varnothing$, $i \neq j$, and therefore

$$\mathsf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathsf{P}\left(\sum_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mathsf{P}(B_n) \leq \sum_{n=1}^{\infty} \mathsf{P}(A_n).$$

The next theorem, which has many applications, provides conditions under which a finitely additive set function is actually countably additive.

Theorem. *Let P be a finitely additive set function defined over the algebra $\mathscr{A}$, with $\mathsf{P}(\Omega) = 1$. The following four conditions are equivalent:*

(1) P *is σ-additive* (P *is a probability*);

(2) P *is continuous from below, i.e. for any sets $A_1, A_2, \ldots \in \mathscr{A}$ such that $A_n \subseteq A_{n+1}$ and $\bigcup_{n=1}^{\infty} A_n \in \mathscr{A}$,*

$$\lim_n \mathsf{P}(A_n) = \mathsf{P}\left(\bigcup_{n=1}^{\infty} A_n\right);$$

(3) P *is continuous from above, i.e. for any sets $A_1, A_2, \ldots$ such that $A_n \supseteq A_{n+1}$ and $\bigcap_{n=1}^{\infty} A_n \in \mathscr{A}$,*

$$\lim_n \mathsf{P}(A_n) = \mathsf{P}\left(\bigcap_{n=1}^{\infty} A_n\right);$$

(4) P *is continuous at* $\varnothing$, *i.e. for any sets* $A_1, A_2, \ldots \in \mathscr{A}$ *such that* $A_{n+1} \subseteq A_n$ *and* $\bigcap_{n=1}^{\infty} A_n = \varnothing$,

$$\lim_n \mathsf{P}(A_n) = 0.$$

PROOF. (1) $\Rightarrow$ (2). Since

$$\bigcup_{n=1}^{\infty} A_n = A_1 + (A_2 \backslash A_1) + (A_3 \backslash A_2) + \cdots,$$

we have

$$\begin{aligned}\mathsf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) &= \mathsf{P}(A_1) + \mathsf{P}(A_2 \backslash A_1) + \mathsf{P}(A_3 \backslash A_2) + \cdots \\ &= \mathsf{P}(A_1) + \mathsf{P}(A_2) - \mathsf{P}(A_1) + \mathsf{P}(A_3) - \mathsf{P}(A_2) + \cdots \\ &= \lim_n \mathsf{P}(A_n).\end{aligned}$$

(2) $\Rightarrow$ (3). Let $n \geq 1$; then

$$\mathsf{P}(A_n) = \mathsf{P}(A_1 \backslash (A_1 \backslash A_n)) = \mathsf{P}(A_1) - \mathsf{P}(A_1 \backslash A_n).$$

The sequence $\{A_1 \backslash A_n\}_{n \geq 1}$ of sets is nondecreasing (see the table in Subsection 3 below) and

$$\bigcup_{n=1}^{\infty} (A_1 \backslash A_n) = A_1 \backslash \bigcap_{n=1}^{\infty} A_n.$$

Then, by (2)

$$\lim_n \mathsf{P}(A_1 \backslash A_n) = \mathsf{P}\left(\bigcup_{n=1}^{\infty} (A_1 \backslash A_n)\right)$$

and therefore

$$\begin{aligned}\lim_n \mathsf{P}(A_n) &= \mathsf{P}(A_1) - \lim_n \mathsf{P}(A_1 \backslash A_n) \\ &= \mathsf{P}(A_1) - \mathsf{P}\left(\bigcup_{n=1}^{\infty} (A_1 \backslash A_n)\right) = \mathsf{P}(A_1) - \mathsf{P}\left(A_1 \backslash \bigcap_{n=1}^{\infty} A_n\right) \\ &= \mathsf{P}(A_1) - \mathsf{P}(A_1) + \mathsf{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \mathsf{P}\left(\bigcap_{n=1}^{\infty} A_n\right)\end{aligned}$$

(3) $\Rightarrow$ (4). Obvious.

(4) $\Rightarrow$ (1). Let $A_1, A_2, \ldots \in \mathscr{A}$ be pairwise disjoint and let $\sum_{n=1}^{\infty} A_n \in \mathscr{A}$. Then

$$\mathsf{P}\left(\sum_{i=1}^{\infty} A_i\right) = \mathsf{P}\left(\sum_{i=1}^{n} A_i\right) + \mathsf{P}\left(\sum_{i=n+1}^{\infty} A_i\right),$$

Table

Notation	Set-theoretic interpretation	Interpretation in probability theory
ω	element or point	outcome, sample point, elementary event
Ω	set of points	sample space; certain event
$\mathscr{F}$	σ-algebra of subsets	σ-algebra of events
$A \in \mathscr{F}$	set of points	event (if $\omega \in A$, we say that event A occurs)
$\bar{A} = \Omega \setminus A$	complement of A, i.e. the set of points ω that are not in A	event that A does not occur
$A \cup B$	union of A and B, i.e. the set of points ω belonging either to A or to B	event that either A or B occurs
$A \cap B$ (or AB)	intersection of A and B, i.e. the set of points ω belonging to both A and B	event that both A and B occur
$\varnothing$	empty set	impossible event
$A \cap B = \varnothing$	A and B are disjoint	events A and B are mutually exclusive, i.e. cannot occur simultaneously
$A + B$	sum of sets, i.e. union of disjoint sets	event that one of two mutually exclusive events occurs
$A \setminus B$	difference of A and B, i.e. the set of points that belong to A but not to B	event that A occurs and B does not
$A \triangle B$	symmetric difference of sets, i.e. $(A \setminus B) \cup (B \setminus A)$	event that A or B occurs, but not both
$\bigcup_{n=1}^{\infty} A_n$	union of the sets $A_1, A_2, \ldots$	event that at least one of $A_1, A_2, \ldots$ occurs

$\sum_{n=1}^{\infty} A_n$	sum, i.e. union of pairwise disjoint sets $A_1, A_2, \ldots$	event that one of the mutually exclusive events $A_1, A_2, \ldots$ occurs
$\bigcap_{n=1}^{\infty} A_n$	intersection of $A_1, A_2, \ldots$	event that all the events $A_1, A_2, \ldots$ occur
$A_n \uparrow A$ $\left(\text{or } A = \lim_n \uparrow A_n\right)$	the increasing sequence of sets A_n converges to A, i.e. $A_1 \subseteq A_2 \subseteq \cdots$ and $A = \bigcup_{n=1}^{\infty} A_n$	the increasing sequence of events converges to event A
$A_n \downarrow A$ $\left(\text{or } A = \lim_n \downarrow A_n\right)$	the decreasing sequence of sets A_n converges to A, i.e. $A_1 \supseteq A_2 \supseteq \cdots$ and $A = \bigcap_{n=1}^{\infty} A_n$	the decreasing sequence of events converges to event A
$\overline{\lim}\, A_n$ (or $\lim\sup A_n$ or* $\{A_n \text{ i.o.}\}$)	the set $\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$	event that infinitely many of events $A_1, A_2, \ldots$ occur
$\underline{\lim}_n\, A_n$ (or $\lim\inf A_n$)	the set $\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$	event that all the events $A_1, A_2, \ldots$ occur with the possible exception of a finite number of them

* i.o. = infinitely often.

and since $\sum_{i=n+1}^{\infty} A_i \downarrow \varnothing$, $n \to \infty$, we have

$$\sum_{i=1}^{\infty} \mathsf{P}(A_i) = \lim_n \sum_{i=1}^{n} \mathsf{P}(A_i) = \lim_n \mathsf{P}\left(\sum_{i=1}^{n} A_i\right)$$
$$= \lim_n \left[\mathsf{P}\left(\sum_{i=1}^{\infty} A_i\right) - \mathsf{P}\left(\sum_{i=n+1}^{\infty} A_i\right)\right]$$
$$= \mathsf{P}\left(\sum_{i=1}^{\infty} A_i\right) - \lim_n \mathsf{P}\left(\sum_{i=n+1}^{\infty} A_i\right) = \mathsf{P}\left(\sum_{i=1}^{\infty} A_i\right).$$

3. We can now formulate Kolmogorov's generally accepted axiom system, which forms the basis for the concept of a probability space.

Fundamental Definition. *An ordered triple* $(\Omega, \mathscr{F}, \mathsf{P})$ *where*

(a) Ω *is a set of points* ω,
(b) $\mathscr{F}$ *is a* σ*-algebra of subsets of* Ω,
(c) P *is a probability on* $\mathscr{F}$,

is called a probabilistic model or a probability space. Here Ω *is the sample space or space of elementary events, the sets* A *in* $\mathscr{F}$ *are events, and* $\mathsf{P}(A)$ *is the probability of the event* A.

It is clear from the definition that the axiomatic formulation of probability theory is based on set theory and measure theory. Accordingly, it is useful to have a table (pp. 136–137) displaying the ways in which various concepts are interpreted in the two theories. In the next two sections we shall give examples of the measurable spaces that are most important for probability theory and of how probabilities are assigned on them.

4. Problems

1. Let $\Omega = \{r: r \in [0, 1]\}$ be the set of rational points of $[0, 1]$, $\mathscr{A}$ the algebra of sets each of which is a finite sum of disjoint sets A of one of the forms $\{r: a < r < b\}$, $\{r: a \le r < b\}$, $\{r: a < r \le b\}$, $\{r: a \le r \le b\}$, and $\mathsf{P}(A) = b - a$. Show that $\mathsf{P}(A)$, $A \in \mathscr{A}$, is finitely additive set function but not countably additive.

2. Let Ω be a countable set and $\mathscr{F}$ the collection of all its subsets. Put $\mu(A) = 0$ if A is finite and $\mu(A) = \infty$ if A is infinite. Show that the set function μ is finitely additive but not countably additive.

3. Let μ be a finite measure on a σ-algebra $\mathscr{F}$, $A_n \in \mathscr{F}$, $n = 1, 2, \ldots$, and $A = \lim_n A_n$ (i.e., $A = \underline{\lim}_n A_n = \overline{\lim}_n A_n$). Show that $\mu(A) = \lim_n \mu(A_n)$.

4. Prove that $\mathsf{P}(A \triangle B) = \mathsf{P}(A) + \mathsf{P}(B) - 2\mathsf{P}(A \cap B)$.

5. Show that the "distances" $\rho_1(A, B)$ and $\rho_2(A, B)$ defined by

$$\rho_1(A, B) = \mathsf{P}(A \triangle B),$$

$$\rho_2(A, B) = \begin{cases} \dfrac{\mathsf{P}(A \triangle B)}{\mathsf{P}(A \cup B)} & \text{if } \mathsf{P}(A \cup B) \neq 0, \\ 0 & \text{if } \mathsf{P}(A \cup B) = 0 \end{cases}$$

satisfy the triangle inequality.

6. Let μ be a finitely additive measure on an algebra $\mathscr{A}$, and let the sets $A_1, A_2, \ldots \in \mathscr{A}$ be pairwise disjoint and satisfy $A = \sum_{i=1}^{\infty} A_i \in \mathscr{A}$. Then $\mu(A) \geq \sum_{i=1}^{\infty} \mu(A_i)$.

7. Prove that

$$\overline{\limsup A_n} = \liminf \bar{A}_n, \qquad \overline{\liminf A_n} = \limsup \bar{A}_n,$$

$$\liminf A_n \subseteq \limsup A_n, \qquad \limsup(A_n \cup B_n) = \limsup A_n \cup \limsup B_n,$$

$$\limsup A_n \cap \liminf B_n \subseteq \limsup(A_n \cap B_n) \subseteq \limsup A_n \cap \limsup B_n.$$

If $A_n \uparrow A$ or $A_n \downarrow A$, then

$$\liminf A_n = \limsup A_n.$$

8. Let $\{x_n\}$ be a sequence of numbers and $A_n = (-\infty, x_n)$. Show that $x = \limsup x_n$ and $A = \limsup A_n$ are related in the following way: $(-\infty, x) \subseteq A \subseteq (-\infty, x]$. In other words, A is equal to either $(-\infty, x)$ or to $(-\infty, x]$.

9. Give an example to show that if a measure takes the value $+\infty$, it does not follow in general that countable additivity implies continuity at $\varnothing$.

§2. Algebras and σ-Algebras. Measurable Spaces

1. Algebras and σ-algebras are the components out of which probabilistic models are constructed. We shall present some examples and a number of results for these systems.

Let Ω be a sample space. Evidently each of the collections of sets

$$\mathscr{F}_* = \{\varnothing, \Omega\}, \qquad \mathscr{F}^* = \{A : A \subseteq \Omega\}$$

is both an algebra and a σ-algebra. In fact, $\mathscr{F}_*$ is trivial, the "poorest" σ-algebra, whereas $\mathscr{F}^*$ is the "richest" σ-algebra, consisting of all subsets of Ω.

When Ω is a finite space, the σ-algebra $\mathscr{F}^*$ is fully surveyable, and commonly serves as the system of events in the elementary theory. However, when the space is uncountable the class $\mathscr{F}^*$ is much too large, since it is impossible to define "probability" on such a system of sets in any consistent way.

If $A \subseteq \Omega$, the system

$$\mathscr{F}_A = \{A, \bar{A}, \varnothing, \Omega\}$$

is another example of an algebra (and a σ-algebra), the algebra (or σ-algebra) generated by A.

This system of sets is a special case of the systems generated by decompositions. In fact, let

$$\mathscr{D} = \{D_1, D_2, \ldots\}$$

be a *countable* decomposition of Ω into nonempty sets:

$$\Omega = D_1 + D_2 + \cdots; \qquad D_i \cap D_j = \varnothing, \qquad i \neq j.$$

Then the system $\mathscr{A} = \alpha(\mathscr{D})$, formed by the sets that are unions of finite numbers of elements of the decomposition, is an algebra.

The following lemma is particularly useful since it establishes the important principle that there is a smallest algebra, or σ-algebra, containing a given collection of sets.

Lemma 1. *Let $\mathscr{E}$ be a collection of subsets of Ω. Then there are a smallest algebra $\alpha(\mathscr{E})$ and a smallest σ-algebra $\sigma(\mathscr{E})$ containing all the sets that are in $\mathscr{E}$.*

PROOF. The class $\mathscr{F}^*$ of all subsets of Ω is a σ-algebra. Therefore there are at least one algebra and one σ-algebra containing $\mathscr{E}$. We now define $\alpha(\mathscr{E})$ (or $\sigma(\mathscr{E})$) to consist of all sets that belong to every algebra (or σ-algebra) containing $\mathscr{E}$. It is easy to verify that this system is an algebra (or σ-algebra) and indeed the smallest.

Remark. The algebra $\alpha(E)$ (or $\sigma(E)$, respectively) is often referred to as the smallest algebra (or σ-algebra) generated by $\mathscr{E}$.

We often need to know what additional conditions will make an algebra, or some other system of sets, into a σ-algebra. We shall present several results of this kind.

Definition 1. A collection $\mathscr{M}$ of subsets of Ω is a *monotonic class* if $A_n \in \mathscr{M}$, $n = 1, 2, \ldots$, together with $A_n \uparrow A$ or $A_n \downarrow A$, implies that $A \in \mathscr{M}$.

Let $\mathscr{E}$ be a system of sets. Let $\mu(\mathscr{E})$ be the smallest monotonic class containing $\mathscr{E}$. (The proof of the existence of this class is like the proof of Lemma 1.)

Lemma 2. *A necessary and sufficient condition for an algebra $\mathscr{A}$ to be a σ-algebra is that it is a monotonic class.*

PROOF. A σ-algebra is evidently a monotonic class. Now let $\mathscr{A}$ be a monotonic class and $A_n \in \mathscr{A}$, $n = 1, 2, \ldots$. It is clear that $B_n = \bigcup_{i=1}^n A_i \in \mathscr{A}$ and $B_n \subseteq B_{n+1}$. Consequently, by the definition of a monotonic class, $B_n \uparrow \bigcup_{i=1}^\infty A_i \in \mathscr{A}$. Similarly we could show that $\bigcap_{i=1}^\infty A_i \in \mathscr{A}$.

By using this lemma, we can prove that, starting with an algebra $\mathscr{A}$, we can construct the σ-algebra $\sigma(\mathscr{A})$ by means of monotonic limiting processes.

Theorem 1. *Let $\mathscr{A}$ be an algebra. Then*

$$\mu(\mathscr{A}) = \sigma(\mathscr{A}). \tag{1}$$

PROOF. By Lemma 2, $\mu(\mathscr{A}) \subseteq \sigma(\mathscr{A})$. Hence it is enough to show that $\mu(\mathscr{A})$ is a σ-algebra. But $\mathscr{M} = \mu(\mathscr{A})$ is a monotonic class, and therefore, by Lemma 2 again, it is enough to show that $\mu(\mathscr{A})$ is an algebra.

Let $A \in \mathscr{M}$; we show that $\bar{A} \in \mathscr{M}$. For this purpose, we shall apply a principle that will often be used in the future, the *principle of appropriate sets*, which we now illustrate.

Let

$$\tilde{\mathscr{M}} = \{B: B \in \mathscr{M}, \bar{B} \in \mathscr{M}\}$$

be the sets that have the property that concerns us. It is evident that $\mathscr{A} \subseteq \tilde{\mathscr{M}} \subseteq \mathscr{M}$. Let us show that $\tilde{\mathscr{M}}$ is a monotonic class.

Let $B_n \in \tilde{\mathscr{M}}$; then $B_n \in \mathscr{M}$, $\bar{B}_n \in \mathscr{M}$, and therefore

$$\lim \uparrow B_n \in \mathscr{M}, \qquad \lim \uparrow \bar{B}_n \in \mathscr{M}, \qquad \lim \downarrow B_n \in \mathscr{M}, \qquad \lim \downarrow \bar{B}_n \in \mathscr{M}.$$

Consequently

$$\overline{\lim \uparrow B_n} = \lim \downarrow \bar{B}_n \in \mathscr{M}, \qquad \overline{\lim \downarrow B_n} = \lim \uparrow \bar{B}_n \in \mathscr{M},$$

$$\overline{\lim \uparrow \bar{B}_n} = \lim \downarrow B_n \in \mathscr{M}, \qquad \overline{\lim \downarrow \bar{B}_n} = \lim \uparrow B_n \in \mathscr{M},$$

and therefore $\tilde{\mathscr{M}}$ is a monotonic class. But $\tilde{\mathscr{M}} \subseteq \mathscr{M}$ and $\mathscr{M}$ is the smallest monotonic class. Therefore $\tilde{\mathscr{M}} = \mathscr{M}$, and if $A \in \mathscr{M} = \mu(\mathscr{A})$, then we also have $\bar{A} \in \mathscr{M}$, i.e. $\mathscr{M}$ is closed under the operation of taking complements.

Let us now show that $\mathscr{M}$ is closed under intersections.

Let $A \in \mathscr{M}$ and

$$\mathscr{M}_A = \{B: B \in \mathscr{M}, A \cap B \in \mathscr{M}\}.$$

From the equations

$$\lim \downarrow (A \cap B_n) = A \cap \lim \downarrow B_n,$$

$$\lim \uparrow (A \cap B_n) = A \cap \lim \uparrow B_n$$

it follows that $\mathscr{M}_A$ is a monotonic class.

Moreover, it is easily verified that

$$(A \in \mathscr{M}_B) \Leftrightarrow (B \in \mathscr{M}_A). \tag{2}$$

Now let $A \in \mathscr{A}$; then since $\mathscr{A}$ is an algebra, for every $B \in \mathscr{A}$ the set $A \cap B \in \mathscr{A}$ and therefore

$$\mathscr{A} \subseteq \mathscr{M}_A \subseteq \mathscr{M}.$$

But $\mathscr{M}_A$ is a monotonic class (since $\lim \uparrow AB_n = A \lim \uparrow B_n$ and $\lim \downarrow AB_n = A \lim \downarrow B_n$), and $\mathscr{M}$ is the smallest monotonic class. Therefore $\mathscr{M}_A = \mathscr{M}$ for all $A \in \mathscr{A}$. But then it follows from (2) that

$$(A \in \mathscr{M}_B) \Leftrightarrow (B \in \mathscr{M}_A = \mathscr{M}).$$

whenever $A \in \mathscr{A}$ and $B \in \mathscr{M}$. Consequently if $A \in \mathscr{A}$ then

$$A \in \mathscr{M}_B$$

for every $B \in \mathscr{M}$. Since A is any set in $\mathscr{A}$, it follows that

$$\mathscr{A} \subseteq \mathscr{M}_B \subseteq \mathscr{M}.$$

Therefore for every $B \in \mathscr{M}$

$$\mathscr{M}_B = \mathscr{M},$$

i.e. if $B \in \mathscr{M}$ and $C \in \mathscr{M}$ then $C \cap B \in \mathscr{M}$.

Thus $\mathscr{M}$ is closed under complementation and intersection (and therefore under unions). Consequently $\mathscr{M}$ is an algebra, and the theorem is established.

Definition 2. Let Ω be a space. A class $\mathscr{D}$ of subsets of Ω is a *d-system* if

(a) $\Omega \in \mathscr{D}$;
(b) $A, B, \in \mathscr{D}, A \subseteq B \Rightarrow B \setminus A \in \mathscr{D}$;
(c) $A_n \in \mathscr{D}, A_n \subseteq A_{n+1} \Rightarrow \bigcup A_n \in \mathscr{D}$.

If $\mathscr{E}$ is a collection of sets then $d(\mathscr{E})$ denotes the smallest d-system containing $\mathscr{E}$.

Theorem 2. *If the collection $\mathscr{E}$ of sets is closed under intersections, then*

$$d(\mathscr{E}) = \sigma(\mathscr{E}) \tag{3}$$

PROOF. Every σ-algebra is a d-system, and consequently $d(\mathscr{E}) \subseteq \sigma(\mathscr{E})$. Hence if we prove that $d(\mathscr{E})$ is closed under intersections, $d(\mathscr{E})$ must be a σ-algebra and then, of course, the opposite inclusion $\sigma(\mathscr{E}) \subseteq d(\mathscr{E})$ is valid.

The proof once again uses the principle of appropriate sets.

Let

$$\mathscr{E}_1 = \{B \in d(\mathscr{E}): B \cap A \in d(\mathscr{E}) \text{ for all } A \in \mathscr{E}\}.$$

If $B \in \mathscr{E}$ then $B \cap A \in \mathscr{E}$ for all $A \in \mathscr{E}$ and therefore $\mathscr{E} \subseteq \mathscr{E}_1$. But $\mathscr{E}_1$ is a d-system. Hence $d(\mathscr{E}) \subseteq \mathscr{E}_1$. On the other hand, $\mathscr{E}_1 \subseteq d(\mathscr{E})$ by definition. Consequently

$$\mathscr{E}_1 = d(\mathscr{E}).$$

Now let

$$\mathscr{E}_2 = \{B \in d(\mathscr{E}): B \cap A \in d(\mathscr{E}) \text{ for all } A \in d(\mathscr{E})\}.$$

Again it is easily verified that $\mathscr{E}_2$ is a d-system. If $B \in \mathscr{E}$, then by the definition of $\mathscr{E}_1$ we obtain that $B \cap A \in d(\mathscr{E})$ for all $A \in \mathscr{E}_1 = d(\mathscr{E})$. Consequently $\mathscr{E} \subseteq \mathscr{E}_2$ and $d(\mathscr{E}) \subseteq \mathscr{E}_2$. But $d(\mathscr{E}) \supseteq \mathscr{E}_2$; hence $d(\mathscr{E}) = \mathscr{E}_2$, and therefore whenever A and B are in $d(\mathscr{E})$, the set $A \cap B$ also belongs to $d(\mathscr{E})$, i.e. $d(\mathscr{E})$ is closed under intersections.

This completes the proof of the theorem.

We next consider some measurable spaces $(\Omega, \mathscr{F})$ which are extremely important for probability theory.

2. The measurable space $(R, \mathscr{B}(R))$. Let $R = (-\infty, \infty)$ be the real line and

$$(a, b] = \{x \in R : a < x \leq b\}$$

for all a and b, $-\infty \leq a < b < \infty$. The interval $(a, \infty]$ is taken to be (a, ∞). (This convention is required if the complement of an interval $(-\infty, b]$ is to be an interval of the same form, i.e. open on the left and closed on the right.)

Let $\mathscr{A}$ be the system of subsets of R which are finite sums of disjoint intervals of the form $(a, b]$:

$$A \in \mathscr{A} \text{ if } A = \sum_{i=1}^{n} (a_i, b_i], \qquad n < \infty.$$

It is easily verified that this system of sets, in which we also include the empty set $\varnothing$, is an algebra. However, it is not a σ-algebra, since if $A_n = (0, 1 - 1/n] \in \mathscr{A}$, we have $\bigcup_n A_n = (0, 1) \notin \mathscr{A}$.

Let $\mathscr{B}(R)$ be the smallest σ-algebra $\sigma(\mathscr{A})$ containing $\mathscr{A}$. This σ-algebra, which plays an important role in analysis, is called the *Borel* algebra of subsets of the real line, and its sets are called *Borel sets*.

If $\mathscr{I}$ is the system of intervals $\mathscr{I}$ of the form $(a, b]$, and $\sigma(\mathscr{I})$ is the smallest σ-algebra containing $\mathscr{I}$, it is easily verified that $\sigma(\mathscr{I})$ is the Borel algebra. In other words, we can obtain the Borel algebra from $\mathscr{I}$ without going through the algebra $\mathscr{A}$, since $\sigma(\mathscr{I}) = \sigma(\alpha(\mathscr{I}))$.

We observe that

$$(a, b) = \bigcup_{n=1}^{\infty} \left(a, b - \frac{1}{n}\right], \qquad a < b,$$

$$[a, b] = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, b\right], \qquad a < b,$$

$$\{a\} = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, a\right].$$

Thus the Borel algebra contains not only intervals $(a, b]$ but also the singletons $\{a\}$ and all sets of the six forms

$$(a, b), \quad [a, b], \quad [a, b), \quad (-\infty, b), \quad (-\infty, b], \quad (a, \infty). \tag{4}$$

Let us also notice that the construction of $\mathscr{B}(R)$ could have been based on any of the six kinds of intervals instead of on $(a, b]$, since all the minimal σ-algebras generated by systems of intervals of any of the forms (4) are the same as $\mathscr{B}(R)$.

Sometimes it is useful to deal with the σ-algebra $\mathscr{B}(\bar{R})$ of subsets of the extended real line $\bar{R} = [-\infty, \infty]$. This is the smallest σ-algebra generated by intervals of the form

$$(a, b] = \{x \in \bar{R} : a < x \le b\}, \qquad -\infty \le a < b \le \infty,$$

where $(-\infty, b]$ is to stand for the set $\{x \in \bar{R}: -\infty \le x \le b\}$.

Remark 1. The measurable space $(R, \mathscr{B}(R))$ is often denoted by $(R, \mathscr{B})$ or $(R^1, \mathscr{B}_1)$.

Remark 2. Let us introduce the metric

$$\rho_1(x, y) = \frac{|x - y|}{1 + |x - y|}$$

on the real line R (this is equivalent to the usual metric $|x - y|$) and let $\mathscr{B}_0(R)$ be the smallest σ-algebra generated by the open sets $S_\rho(x^0) = \{x \in R: \rho_1(x, x^0) < \rho\}$, $\rho > 0$, $x^0 \in R$. Then $\mathscr{B}_0(R) = \mathscr{B}(R)$ (see Problem 7).

3. The measurable space $(R^n, \mathscr{B}(R^n))$. Let $R^n = R \times \cdots \times R$ be the direct, or Cartesian, product of n copies of the real line, i.e. the set of ordered n-tuples $x = (x_1, \ldots, x_n)$, where $-\infty < x_k < \infty$, $k = 1, \ldots, n$. The set

$$I = I_1 \times \cdots \times I_n,$$

where $I_k = (a_k, b_k]$, i.e. the set $\{x \in R^n: x_k \in I_k,\ k = 1, \ldots, n\}$, is called a rectangle, and I_k is a side of the rectangle. Let $\mathscr{I}$ be the set of all rectangles I. The smallest σ-algebra $\sigma(\mathscr{I})$ generated by the system $\mathscr{I}$ is the *Borel algebra* of subsets of R^n and is denoted by $\mathscr{B}(R^n)$. Let us show that we can arrive at this Borel algebra by starting in a different way.

Instead of the rectangles $I = I_1 \times \cdots \times I_n$ let us consider the rectangles $B = B_1 \times \cdots \times B_n$ with Borel sides (B_k is the Borel subset of the real line that appears in the kth place in the direct product $R \times \cdots \times R$). The smallest σ-algebra containing all rectangles with Borel sides is denoted by

$$\mathscr{B}(R) \otimes \cdots \otimes \mathscr{B}(R)$$

and called the *direct product* of the σ-algebras $\mathscr{B}(R)$. Let us show that in fact

$$\mathscr{B}(R^n) = \mathscr{B}(R) \otimes \cdots \otimes \mathscr{B}(R).$$

In other words, the smallest σ-algebra generated by the rectangles $I = I_1 \times \cdots \times I_n$ and the (broader) class of rectangles $B = B_1 \times \cdots \times B_n$ with Borel sides are actually the same.

The proof depends on the following proposition.

Lemma 3. *Let $\mathscr{E}$ be a class of subsets of Ω, let $B \subseteq \Omega$, and define*

$$\mathscr{E} \cap B = \{A \cap B: A \in \mathscr{E}\}. \tag{5}$$

Then

$$\sigma(\mathscr{E} \cap B) = \sigma(\mathscr{E}) \cap B. \tag{6}$$

PROOF. Since $\mathscr{E} \subseteq \sigma(\mathscr{E})$, we have

$$\mathscr{E} \cap B \subseteq \sigma(\mathscr{E}) \cap B. \tag{7}$$

But $\sigma(\mathscr{E}) \cap B$ is a σ-algebra; hence it follows from (7) that

$$\sigma(\mathscr{E} \cap B) \subseteq \sigma(\mathscr{E}) \cap B.$$

To prove the conclusion in the opposite direction, we again use the principle of appropriate sets.

Define

$$\mathscr{C}_B = \{A \in \sigma(\mathscr{E}): A \cap B \in \sigma(\mathscr{E} \cap B)\}.$$

Since $\sigma(\mathscr{E})$ and $\sigma(\mathscr{E} \cap B)$ are σ-algebras, $\mathscr{C}_B$ is also a σ-algebra, and evidently

$$\mathscr{E} \subseteq \mathscr{C}_B \subseteq \sigma(\mathscr{E}),$$

whence $\sigma(\mathscr{E}) \subseteq \sigma(\mathscr{C}_B) = \mathscr{C}_B \subseteq \sigma(\mathscr{E})$ and therefore $\sigma(\mathscr{E}) = \mathscr{C}_B$. Therefore

$$A \cap B \in \sigma(\mathscr{E} \cap B)$$

for every $A \subseteq \sigma(\mathscr{E})$, and consequently $\sigma(\mathscr{E}) \cap B \subseteq \sigma(\mathscr{E} \cap B)$.

This completes the proof of the lemma.

Proof that $\mathscr{B}(R^n)$ and $\mathscr{B} \otimes \cdots \otimes \mathscr{B}$ are the same. This is obvious for $n = 1$. We now show that it is true for $n = 2$.

Since $\mathscr{B}(R^2) \subseteq \mathscr{B} \otimes \mathscr{B}$, it is enough to show that the Borel rectangle $B_1 \times B_2$ belongs to $\mathscr{B}(R^2)$.

Let $R^2 = R_1 \times R_2$, where R_1 and R_2 are the "first" and "second" real lines, $\tilde{\mathscr{B}}_1 = \mathscr{B}_1 \times R_2$, $\tilde{\mathscr{B}}_2 = R_1 \times \mathscr{B}_2$, where $\mathscr{B}_1 \times R_2$ (or $R_1 \times \mathscr{B}_2$) is the collection of sets of the form $B_1 \times R_2$ (or $R_1 \times B_2$), with $B_1 \in \mathscr{B}_1$ (or $B_2 \in \mathscr{B}_2$). Also let $\mathscr{I}_1$ and $\mathscr{I}_2$ be the sets of intervals in R_1 and R_2, and $\tilde{\mathscr{I}}_1 = \mathscr{I}_1 \times R_2$, $\tilde{\mathscr{I}}_2 = R_1 \times \mathscr{I}_2$. Then, by (6),

$$\begin{aligned} B_1 \times B_2 = \tilde{B}_1 \cap \tilde{B}_2 \in \tilde{\mathscr{B}}_1 \cap \tilde{\mathscr{B}}_2 &= \sigma(\tilde{\mathscr{I}}_1) \cap \tilde{B}_2 \\ &= \sigma(\tilde{\mathscr{I}}_1 \cap \tilde{B}_2) \subseteq \sigma(\tilde{\mathscr{I}}_1 \cap \tilde{\mathscr{I}}_2) \\ &= \sigma(\mathscr{I}_1 \times \mathscr{I}_2), \end{aligned}$$

as was to be proved.

The case of any $n, n > 2$, can be discussed in the same way.

Remark. Let $\mathscr{B}_0(R^n)$ be the smallest σ-algebra generated by the open sets

$$S_\rho(x^0) = \{x \in R^n: \rho_n(x, x^0) < \rho\}, \qquad x^0 \in R^n, \quad \rho > 0,$$

in the metric

$$\rho_n(x, x^0) = \sum_{k=1}^{n} 2^{-k} \rho_1(x_k, x_k^0),$$

where $x = (x_1, \dots, x_n)$, $x^0 = (x_1^0, \dots, x_n^0)$.

Then $\mathscr{B}_0(R_n) = \mathscr{B}(R^n)$ (Problem 7).

4. The measurable space $(R^\infty, \mathscr{B}(R^\infty))$ plays a significant role in probability theory, since it is used as the basis for constructing probabilistic models of experiments with infinitely many steps.

The space R^∞ is the space of *ordered* sequences of numbers,

$$x = (x_1, x_2, \dots), \qquad -\infty < x_k < \infty, \quad k = 1, 2, \dots$$

Let I_k and B_k denote, respectively, the intervals $(a_k, b_k]$ and the Borel subsets of the kth line (with coordinate x_k). We consider the *cylinder sets*

$$\mathscr{I}(I_1 \times \cdots \times I_n) = \{x: x = (x_1, x_2, \dots), x_1 \in I_1, \dots, x_n \in I_n\}, \tag{8}$$

$$\mathscr{I}(B_1 \times \cdots \times B_n) = \{x: x = (x_1, x_2 \cdots), x_1 \in B_1, \dots, x_n \in B_n\}, \tag{9}$$

$$\mathscr{I}(B^n) = \{x: (x_1, \dots, x_n) \in B^n\}, \tag{10}$$

where B^n is a Borel set in $\mathscr{B}(R^n)$. Each cylinder $\mathscr{I}(B_1 \times \cdots \times B_n)$, or $\mathscr{I}(B^n)$, can also be thought of as a cylinder with base in $R^{n+1}, R^{n+2}, \dots$, since

$$\mathscr{I}(B_1 \times \cdots \times B_n) = \mathscr{I}(B_1 \times \cdots \times B_n \times R),$$

$$\mathscr{I}(B^n) = \mathscr{I}(B^{n+1}),$$

where $B^{n+1} = B^n \times R$.

It follows that both systems of cylinders $\mathscr{I}(B_1 \times \cdots \times B_n)$ and $\mathscr{I}(B^n)$ are algebras. It is easy to verify that the unions of disjoint cylinders

$$\mathscr{I}(I_1 \times \cdots \times I_n)$$

also form an algebra. Let $\mathscr{B}(R^\infty)$, $\mathscr{B}_1(R^\infty)$ and $\mathscr{B}_2(R^\infty)$ be the smallest σ-algebras containing all the sets (8), (9) or (10), respectively. (The σ-algebra $\mathscr{B}_1(R^\infty)$ is often denoted by $\mathscr{B}(R) \otimes \mathscr{B}(R) \times \cdots$.) It is clear that $\mathscr{B}(R^\infty) \subseteq \mathscr{B}_1(R^\infty) \subseteq \mathscr{B}_2(R^\infty)$. As a matter of fact, all three σ-algebras are the same.

To prove this, we put

$$\mathscr{C}_n = \{A \in R^n: \{x: (x_1, \dots, x_n) \in A\} \in \mathscr{B}(R^\infty)\}$$

for $n = 1, 2, \dots$. Let $B^n \in \mathscr{B}(R^n)$. Then

$$B^n \in \mathscr{C}_n \subseteq \mathscr{B}(R^\infty).$$

But $\mathscr{C}_n$ is a σ-algebra, and therefore

$$\mathscr{B}(R^n) \subseteq \sigma(\mathscr{C}_n) = \mathscr{C}_n \subseteq \mathscr{B}(R^\infty);$$

consequently

$$\mathscr{B}_2(R^\infty) \subseteq \mathscr{B}(R^\infty).$$

Thus $\mathscr{B}(R^\infty) = \mathscr{B}_1(R^\infty) = \mathscr{B}_2(R^\infty)$.

From now on we shall describe sets in $\mathscr{B}(R^\infty)$ as Borel sets (in R^∞).

Remark. Let $\mathscr{B}_0(R^\infty)$ be the smallest σ-algebra generated by the *open* sets

$$S_\rho(x^0) = \{x \in R^\infty : \rho_\infty(x, x^0) < \rho\}, \qquad x^0 \in R^\infty, \quad \rho > 0,$$

in the metric

$$\rho_\infty(x, x^0) = \sum_{k=1}^{\infty} 2^{-k} \rho_1(x_k, x_k^0),$$

where $x = (x_1, x_2, \ldots)$, $x^0 = (x_1^0, x_2^0, \ldots)$. Then $\mathscr{B}(R^\infty) = \mathscr{B}_0(R^\infty)$ (Problem 7).

Here are some examples of Borel sets in R^∞:

(a) $\{x \in R^\infty : \sup x_n > a\}$,
$\{x \in R^\infty : \inf x_n < a\}$;

(b) $\{x \in R^\infty : \overline{\lim}\, x_n \le a\}$,
$\{x \in R^\infty : \underline{\lim}\, x_n > a\}$,

where, as usual,

$$\overline{\lim}\, x_n = \inf_n \sup_{m \ge n} x_m, \qquad \underline{\lim}\, x_n = \sup_n \inf_{m \ge n} x_m;$$

(c) $\{x \in R^\infty : x_n \to\}$, the set of $x \in R^\infty$ for which $\lim x_n$ exists and is finite;
(d) $\{x \in R^\infty : \lim x_n > a\}$;
(e) $\{x \in R^\infty : \sum_{n=1}^\infty |x_n| > a\}$;
(f) $\{x \in R^\infty : \sum_{k=1}^n x_k = 0 \text{ for at least one } n \ge 1\}$.

To be convinced, for example, that sets in (a) belong to the system $\mathscr{B}(R^\infty)$, it is enough to observe that

$$\{x : \sup x_n > a\} = \bigcup_n \{x : x_n > a\} \in \mathscr{B}(R^\infty),$$

$$\{x : \inf x_n < a\} = \bigcup_n \{x : x_n < a\} \in \mathscr{B}(R^\infty).$$

5. The measurable space $(R^T, \mathscr{B}(R^T))$, where T is an arbitrary set. The space R^T is the collection of real functions $x = (x_t)$ defined for $t \in T$†. In general we shall be interested in the case when T is an uncountable subset of the real

† We shall also use the notations $x = (x_t)_{t \in R^T}$ and $x = (x_t)$, $t \in R^T$, for elements of R^T.

line. For simplicity and definiteness we shall suppose for the present that $T = [0, \infty)$.

We shall consider three types of cylinder sets

$$\mathscr{I}_{t_1, \ldots, t_n}(I_1 \times \cdots \times I_n) = \{x : x_{t_1} \in I_1, \ldots, x_{t_n} \in I_1\}, \tag{11}$$

$$\mathscr{I}_{t_1, \ldots, t_n}(B_1 \times \cdots \times B_n) = \{x : x_{t_1} \in B_1, \ldots, x_{t_n} \in B_n\}, \tag{12}$$

$$\mathscr{I}_{t_1, \ldots, t_n}(B^n) = \{x : (x_{t_1}, \ldots, x_{t_n}) \in B^n\}, \tag{13}$$

where I_k is a set of the form $(a_k, b_k]$, B_k is a Borel set on the line, and B^n is a Borel set in R^n.

The set $\mathscr{I}_{t_1, \ldots, t_n}(I_1 \times \cdots \times I_n)$ is just the set of functions that, at times $t_1, \ldots, t_n$, "get through the windows" $I_1, \ldots, I_n$ and at other times have arbitrary values (Figure 24).

Let $\mathscr{B}(R^T)$, $\mathscr{B}_1(R^T)$ and $\mathscr{B}_2(R^T)$ be the smallest σ-algebras corresponding respectively to the cylinder sets (11), (12) and (13). It is clear that

$$\mathscr{B}(R^T) \subseteq \mathscr{B}_1(R^T) \subseteq \mathscr{B}_2(R^T). \tag{14}$$

As a matter of fact, all three of these σ-algebras are the same. Moreover, we can give a complete description of the structure of their sets.

Theorem 3. *Let T be any uncountable set. Then $\mathscr{B}(R^T) = \mathscr{B}_1(R^T) = \mathscr{B}_2(R^T)$, and every set $A \in \mathscr{B}(R^T)$ has the following structure: there are a countable set of points $t_1, t_2, \ldots$ of T and a Borel set B in $\mathscr{B}(R^\infty)$ such that*

$$A = \{x : (x_{t_1}, x_{t_2}, \ldots) \in B\}. \tag{15}$$

PROOF. Let $\mathscr{E}$ denote the collection of sets of the form (15) (for various aggregates $(t_1, t_2, \ldots)$ and Borel sets B in $\mathscr{B}(R^\infty)$). If $A_1, A_2, \ldots \in \mathscr{E}$ and the corresponding aggregates are $T^{(1)} = (t_1^{(1)}, t_2^{(1)}, \ldots)$, $T^{(2)} = (t_1^{(2)}, t_2^{(2)}, \ldots), \ldots,$

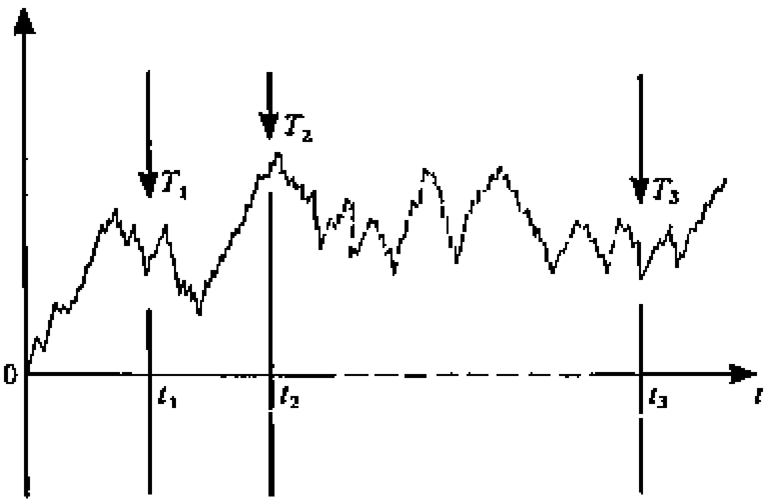


Figure 24

then the set $T^{(\infty)} = \bigcup_k T^{(k)}$ can be taken as a basis, so that every $A^{(i)}$ has a representation

$$A_i = \{x: (x_{\tau_1}, x_{\tau_2}, \ldots) \in B_i\},$$

where B_i is a set in one and the same σ-algebra $\mathscr{B}(R^\infty)$, and $\tau_i \in T^{(\infty)}$.

Hence it follows that the system $\mathscr{E}$ is a σ-algebra. Clearly this σ-algebra contains all cylinder sets of the form (1) and, since $\mathscr{B}_2(R^T)$ is the smallest σ-algebra containing these sets, and since we have (14), we obtain

$$\mathscr{B}(R^T) \subseteq \mathscr{B}_1(R^T) \subseteq \mathscr{B}_2(R^T) \subseteq \mathscr{E}. \tag{16}$$

Let us consider a set A from $\mathscr{E}$, represented in the form (15). For a given aggregate $(t_1, t_2, \ldots)$, the same reasoning as for the space $(R^\infty, \mathscr{B}(R^\infty))$ shows that A is an element of the σ-algebra generated by the cylinder sets (11). But this σ-algebra evidently belongs to the σ-algebra $\mathscr{B}(R^T)$; together with (16), this established both conclusions of the theorem.

Thus every Borel set A in the σ-algebra $\mathscr{B}(R^T)$ is determined by restrictions imposed on the functions $x = (x_t)$, $t \in T$, on an at most countable set of points $t_1, t_2, \ldots$. Hence it follows, in particular, that the sets

$$A_1 = \{x: \sup x_t < C \text{ for all } t \in [0, 1]\},$$

$$A_2 = \{x: x_t = 0 \text{ for at least one } t \in [0, 1]\},$$

$$A_3 = \{x: x_t \text{ is continuous at a given point } t_0 \in [0, 1]\},$$

which depend on the behavior of the function on an uncountable set of points, cannot be Borel sets. And indeed *none of these three sets belongs to* $\mathscr{B}(R^{[0,1]})$.

Let us establish this for A_1. If $A_1 \in \mathscr{B}(R^{[0,1]})$, then by our theorem there are a point $(t_1^0, t_2^0, \ldots)$ and a set $B^0 \in \mathscr{B}(R^\infty)$ such that

$$\left\{x: \sup_t x_t < C, t \in [0, 1]\right\} = \{x: (x_{t_1^0}, x_{t_2^0}, \ldots) \in B^0\}.$$

It is clear that the function $y_t \equiv C - 1$ belongs to A_1, and consequently $(y_{t_1^0}, \ldots) \in B^0$. Now form the function

$$z_t = \begin{cases} C - 1, & t \in (t_1^0, t_2^0, \ldots), \\ C + 1, & t \notin (t_1^0, t_2^0, \ldots). \end{cases}$$

It is clear that

$$(y_{t_1^0}, y_{t_2^0}, \ldots) = (z_{t_1^0}, z_{t_2^0}, \ldots),$$

and consequently the function $z = (z_t)$ belongs to the set $\{x: (x_{t_1^0}, \ldots\} \in B^0\}$. But at the same time it is clear that it does not belong to the set $\{x: \sup x_t < C\}$. This contradiction shows that $A_1 \notin \mathscr{B}(R^{[0,1]})$.

Since the sets A_1, A_2 and A_3 are nonmeasurable with respect to the σ-algebra $\mathscr{B}[R^{[0,1]})$ in the space of all functions $x = (x_t)$, $t \in [0, 1]$, it is natural to consider a smaller class of functions for which these sets are measurable. It is intuitively clear that this will be the case if we take the intial space to be, for example, the space of continuous functions.

6. The measurable space $(C, \mathscr{B}(C))$. Let $T = [0, 1]$ and let C be the space of continuous functions $x = (x_t)$, $0 \le t \le 1$. This is a metric space with the metric $\rho(x, y) = \sup_{t \in T} |x_t - y_t|$. We introduce two σ-algebras in C: $\mathscr{B}(C)$ is the σ-algebra generated by the cylinder sets, and $\mathscr{B}_0(C)$ is generated by the open sets (open with respect to the metric $\rho(x, y)$). Let us show that in fact these σ-algebras are the same: $\mathscr{B}(C) = \mathscr{B}_0(C)$.

Let $B = \{x: x_{t_0} < b\}$ be a cylinder set. It is easy to see that this set is open. Hence it follows that $\{x: x_{t_1} < b_1, \ldots, x_{t_n} < b_n\} \in \mathscr{B}_0(C)$, and therefore $\mathscr{B}(C) \subseteq \mathscr{B}_0(C)$.

Conversely, consider a set $B_\rho = \{y: y \in S_\rho(x^0)\}$ where x^0 is an element of C and $S_\rho(x^0) = \{x \in C: \sup_{t \in T} |x_t - x_t^0| < \rho\}$ is an open ball with center at x^0. Since the functions in C are continuous,

$$B_\rho = \{y \in C: y \in S_\rho(x^0)\} = \left\{y \in C: \max_t |y_t - x_t^0| < \rho\right\}$$
$$= \bigcap_{t_k} \{y \in C: |y_{t_k} - x_{t_k}^0| < \rho\} \in \mathscr{B}(C), \quad (17)$$

where t_k are the rational points of $[0, 1]$. Therefore $\mathscr{B}_0(C) \subseteq \mathscr{B}(C)$.

The following example is fundamental.

7. The measurable space $(D, \mathscr{B}(D))$, where D is the space of functions $x = (x_t)$, $t \in [0, 1]$, that are continuous on the right ($x_t = x_{t+}$ for all $t < 1$) and have limits from the left (at every $t > 0$).

Just as for C, we can introduce a metric $d(x, y)$ on D such that the σ-algebra $\mathscr{B}_0(D)$ generated by the open sets will coincide with the σ-algebra $\mathscr{B}(D)$ generated by the cylinder sets. This metric $d(x, y)$, which was introduced by Skorohod, is defined as follows:

$$d(x, y) = \inf\{\varepsilon > 0: \exists\, \lambda \in \Lambda: \sup_t |x_t - y_{\lambda(t)}| + \sup_t |t - \lambda(t)| \le \varepsilon\}, \quad (18)$$

where Λ is the set of strictly increasing functions $\lambda = \lambda(t)$ that are continuous on $[0, 1]$ and have $\lambda(0) = 0$, $\lambda(1) = 1$.

8. The measurable space $(\prod_{t \in T} \Omega_t, \boxtimes_{t \in T} \mathscr{F}_t)$. Along with the space $(R^T, \mathscr{B}(R^T))$, which is the direct product of T copies of the real line together with the system of Borel sets, probability theory also uses the measurable space $(\prod_{t \in T} \Omega_t, \boxtimes_{t \in T} \mathscr{F}_t)$, which is defined in the following way.

Let T be any set of indices and $(\Omega_t, \mathscr{F}_t)$ a measurable space, $t \in T$. Let $\Omega = \prod_{t \in T} \Omega_t$, the set of functions $\omega = (\omega_t)$, $t \in T$, such that $\omega_t \in \Omega_t$ for each $t \in T$.

The collection of cylinder sets

$$\mathscr{I}_{t_1, \dots, t_n}(B_1 \times \cdots \times B_n) = \{\omega : \omega_{t_1} \in B_1, \dots, \omega_{t_n} \in B_n\},$$

where $B_{t_i} \in \mathscr{F}_{t_i}$, is easily shown to be an algebra. The smallest σ-algebra containing all these cylinder sets is denoted by $\boxtimes_{t \in T} \mathscr{F}_t$, and the measurable space $(\prod \Omega_t, \boxtimes \mathscr{F}_t)$ is called the *direct product* of the measurable spaces $(\Omega_t, \mathscr{F}_t)$, $t \in T$.

9. Problems

1. Let $\mathscr{B}_1$ and $\mathscr{B}_2$ be σ-algebras of subsets of Ω. Are the following systems of sets σ-algebras?

$$\mathscr{B}_1 \cap \mathscr{B}_2 \equiv \{A : A \in \mathscr{B}_1 \text{ and } A \in \mathscr{B}_2\},$$

$$\mathscr{B}_1 \cup \mathscr{B}_2 \equiv \{A : A \in \mathscr{B}_1 \text{ or } A \in \mathscr{B}_2\}.$$

2. Let $\mathscr{D} = \{D_1, D_2, \dots\}$ be a countable decomposition of Ω and $\mathscr{B} = \sigma(\mathscr{D})$. Are there also only countably many sets in $\mathscr{B}$?

3. Show that

$$\mathscr{B}(R^n) \otimes \mathscr{B}(R) = \mathscr{B}(R^{n+1}).$$

4. Prove that the sets (b)–(f) (see Subsection 4) belong to $\mathscr{B}(R^\infty)$.

5. Prove that the sets A_2 and A_3 (see Subsection 5) do not belong to $\mathscr{B}(R^{[0,1]})$.

6. Prove that the function (15) actually defines a metric.

7. Prove that $\mathscr{B}_0(R^n) = \mathscr{B}(R^n)$, $n \ge 1$, and $\mathscr{B}_0(R^\infty) = \mathscr{B}(R^\infty)$.

8. Let $C = C[0, \infty)$ be the space of continuous functions $x = (x_t)$ defined for $t \ge 0$. Show that with the metric

$$\rho(x, y) = \sum_{n=1}^{\infty} 2^{-n} \min\left[\sup_{0 \le t \le n} |x_t - y_t|, 1\right], \qquad x, y \in C,$$

this is a complete separable metric space and that the σ-algebra $\mathscr{B}_0(C)$ generated by the open sets coincides with the σ-algebra $\mathscr{B}(C)$ generated by the cylinder sets.

§3. Methods of Introducing Probability Measures on Measurable Spaces

1. The measurable space $(R, \mathscr{B}(R))$. Let $\mathsf{P} = \mathsf{P}(A)$ be a probability measure defined on the Borel subsets A of the real line. Take $A = (-\infty, x]$ and put

$$F(x) = \mathsf{P}(-\infty, x], \qquad x \in R. \tag{1}$$

This function has the following properties:

(1) $F(x)$ is *nondecreasing*;
(2) $F(-\infty) = 0, F(+\infty) = 1$, *where*

$$F(-\infty) = \lim_{x \downarrow -\infty} F(x), \qquad F(+\infty) = \lim_{x \uparrow \infty} F(x);$$

(3) *$F(x)$ is continuous on the right and has a limit on the left at each $x \in R$.*

The first property is evident, and the other two follow from the continuity properties of probability measures.

Definition 1. Every function $F = F(x)$ satisfying conditions (1)–(3) is called a *distribution function* (on the real line R).

Thus to every probability measure P on $(R, \mathscr{B}(R))$ there corresponds (by (1)) a distribution function. It turns out that the converse is also true.

Theorem 1. *Let $F = F(x)$ be a distribution function on the real line R. There exists a unique probability measure P on $(R, \mathscr{B}(R))$ such that*

$$\mathsf{P}(a, b] = F(b) - F(a) \tag{2}$$

for all $a, b, -\infty \leq a < b < \infty$.

PROOF. Let $\mathscr{A}$ be the algebra of the subsets A of R that are finite sums of disjoint intervals of the form $(a, b]$:

$$A = \sum_{k=1}^{n} (a_k, b_k].$$

On these sets we define a set function P_0 by putting

$$\mathsf{P}_0(A) = \sum_{k=1}^{n} [F(b_k) - F(a_k)], \qquad A \in \mathscr{A}. \tag{3}$$

This formula defines, evidently uniquely, a finitely additive set function on $\mathscr{A}$. Therefore if we show that this function is also countably additive on this algebra, the existence and uniqueness of the required measure P on $\mathscr{B}(R)$ will follow immediately from a general result of measure theory (which we quote without proof).

Carathéodory's Theorem. *Let Ω be a space, $\mathscr{A}$ an algebra of its subsets, and $\mathscr{B} = \sigma(\mathscr{A})$ the smallest σ-algebra containing $\mathscr{A}$. Let μ_0 be a σ-additive measure on (Ω, A). Then there is a unique measure μ on $(\Omega, \sigma(\mathscr{A}))$ which is an extension of μ_0, i.e. satisfies*

$$\mu(A) = \mu_0(A), \qquad A \in \mathscr{A}.$$

We are now to show that P_0 is countably additive on $\mathscr{A}$. By a theorem from §1 it is enough to show that P_0 is continuous at $\varnothing$, i.e. to verify that

$$P_0(A_n) \downarrow 0, \qquad A_n \downarrow \varnothing, \qquad A_n \in \mathscr{A}.$$

Let $A_1, A_2, \ldots$ be a sequence of sets from $\mathscr{A}$ with the property $A_n \downarrow \varnothing$. Let us suppose first that the sets A_n belong to a closed interval $[-N, N]$, $N < \infty$. Since A is the sum of finitely many intervals of the form $(a, b]$ and since

$$P_0(a', b] = F(b) - F(a') \to F(b) - F(a) = P_0(a, b]$$

as $a' \downarrow a$, because $F(x)$ is continuous on the right, we can find, for every A_n, a set $B_n \in \mathscr{A}$ such that its closure $[B_n] \subseteq A_n$ and

$$P_0(A_n) - P_0(B_n) \le \varepsilon \cdot 2^{-n},$$

where ε is a preassigned positive number.

By hypothesis, $\bigcap A_n = \varnothing$ and therefore $\bigcap [B_n] = \varnothing$. But the sets $[B_n]$ are closed, and therefore there is a finite $n_0 = n_0(\varepsilon)$ such that

$$\bigcap_{n=1}^{n_0} [B_n] = \varnothing. \tag{4}$$

(In fact, $[-N, N]$ is compact, and the collection of sets $\{[-N, N]\setminus[B_n]\}_{n\ge 1}$ is an open covering of this compact set. By the Heine–Borel theorem there is a finite subcovering:

$$\bigcup_{n=1}^{n_0} ([-N, N]\setminus[B_n]) = [-N, N]$$

and therefore $\bigcap_{n=1}^{n_0} [B_n] = \varnothing$).

Using (4) and the inclusions $A_{n_0} \subseteq A_{n_0-1} \subseteq \cdots \subseteq A_1$, we obtain

$$\begin{aligned} P_0(A_{n_0}) &= P_0\left(A_{n_0}\setminus \bigcap_{k=1}^{n_0} B_k\right) + P_0\left(\bigcap_{k=1}^{n_0} B_k\right) \\ &= P_0\left(A_{n_0}\setminus \bigcap_{k=1}^{n_0} B_k\right) \le P_0\left(\bigcup_{k=1}^{n_0} (A_k\setminus B_k)\right) \\ &\le \sum_{k=1}^{n_0} P_0(A_k\setminus B_k) \le \sum_{k=1}^{n_0} \varepsilon \cdot 2^{-k} \le \varepsilon. \end{aligned}$$

Therefore $P_0(A_n) \downarrow 0$, $n \to \infty$.

We now abandon the assumption that $A_n \subseteq [-N, N]$ for some N. Take an $\varepsilon > 0$ and choose N so that $P_0[-N, N] > 1 - \varepsilon/2$. Then, since

$$A_n = A_n \cap [-N, N] + A_n \cap \overline{[-N, N]},$$

we have

$$\begin{aligned} P_0(A_n) &= P_0(A_n[-N, N] + P_0(A_n \cap \overline{[-N, N]}) \\ &\le P_0(A_n \cap [-N, N]) + \varepsilon/2 \end{aligned}$$

and, applying the preceding reasoning (replacing A_n by $A_n \cap [-N, N]$), we find that $\mathsf{P}_0(A_n \cap [-N, N]) \le \varepsilon/2$ for sufficiently large n. Hence once again $\mathsf{P}_0(A_n) \downarrow 0$, $n \to \infty$. This completes the proof of the theorem.

Thus there is a one-to-one correspondence between probability measures P on $(R, \mathscr{B}(R))$ and distribution functions F on the real line R. The measure P constructed from the function F is usually called the Lebesgue–Stieltjes probability measure corresponding to the distribution function F.

The case when

$$F(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \le x \le 1, \\ 1, & x > 1. \end{cases}$$

is particularly important. In this case the corresponding probability measure (denoted by λ) is *Lebesgue measure* on $[0, 1]$. Clearly $\lambda(a, b] = b - a$. In other words, the Lebesgue measure of $(a, b]$ (as well as of any of the intervals (a, b), $[a, b]$ or $[a, b)$) is simply its length $b - a$.

Let

$$\mathscr{B}([0, 1]) = \{A \cap [0, 1] : A \in \mathscr{B}(R)\}$$

be the collection of Borel subsets of $[0, 1]$. It is often necessary to consider, besides these sets, the Lebesgue measurable subsets of $[0, 1]$. We say that a set $\Lambda \subseteq [0, 1]$ belongs to $\overline{\mathscr{B}}([0, 1)]$ if there are Borel sets A and B such that $A \subseteq \Lambda \subseteq B$ and $\lambda(B \setminus A) = 0$. It is easily verified that $\overline{\mathscr{B}}([0, 1])$ is a σ-algebra. It is known as the system of *Lebesgue measurable subsets of* $[0, 1]$. Clearly $\mathscr{B}([0, 1]) \subseteq \overline{\mathscr{B}}([0, 1])$.

The measure λ, defined so far only for sets in $\mathscr{B}([0, 1])$, extends in a natural way to the system $\overline{\mathscr{B}}([0, 1])$ of Lebesgue measurable sets. Specifically, if $\Lambda \in \overline{\mathscr{B}}([0, 1])$ and $A \subseteq \Lambda \subseteq B$, where A and $B \in \overline{\mathscr{B}}([0, 1])$ and $\lambda(B \setminus A) = 0$, we define $\bar{\lambda}(\Lambda) = \lambda(A)$. The set function $\bar{\lambda} = \bar{\lambda}(\Lambda)$, $\Lambda \in \overline{\mathscr{B}}([0, 1])$, is easily seen to be a probability measure on $([0, 1], \overline{\mathscr{B}}([0, 1]))$. It is usually called *Lebesgue measure* (on the system of Lebesgue-measurable sets).

Remark. This process of completing (or extending) a measure can be applied, and is useful, in other situations. For example, let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space. Let $\overline{\mathscr{F}}^{\mathsf{P}}$ be the collection of all the subsets A of Ω for which there are sets B_1 and B_2 of $\mathscr{F}$ such that $B_1 \subseteq A \subseteq B_2$ and $\mathsf{P}(B_2 \setminus B_1) = 0$. The probability measure can be defined for sets $A \in \overline{\mathscr{F}}^{\mathsf{P}}$ in a natural way (by $\mathsf{P}(A) = \mathsf{P}(B_1)$). The resulting probability space is the completion of $(\Omega, \mathscr{F}, \mathsf{P})$ with respect to P.

A probability measure such that $\overline{\mathscr{F}}^{\mathsf{P}} = \mathscr{F}$ is called *complete*, and the corresponding space $(\Omega, \mathscr{F}, \mathsf{P})$ is a *complete probability space*.

The correspondence between probability measures P and distribution functions F established by the equation $\mathsf{P}(a, b] = F(b) - F(a)$ makes it

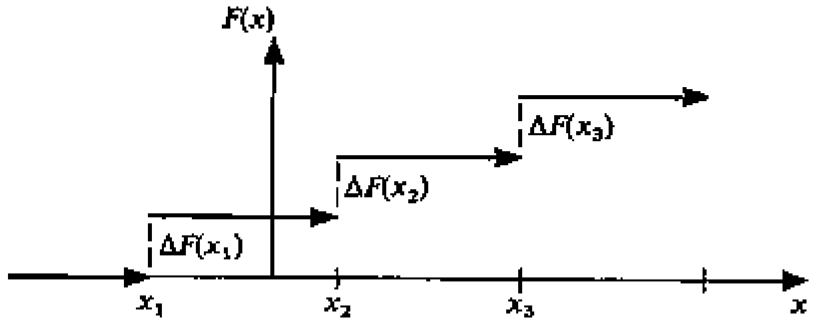


Figure 25

possible to construct various probability measures by obtaining the corresponding distribution functions.

Discrete measures are measures P for which the corresponding distributions $F = F(x)$ are piecewise constant (Figure 25), changing their values at the points $x_1, x_2, \ldots$ ($\Delta F(x_i) > 0$, where $\Delta F(x) = F(x) - F(x-)$. In this case the measure is concentrated at the points $x_1, x_2, \ldots$:

$$\mathsf{P}(\{x_k\}) = \Delta F(x_k) > 0, \qquad \sum_k \mathsf{P}(\{x_k\}) = 1.$$

The set of numbers $(p_1, p_2, \ldots)$, where $p_k = \mathsf{P}(\{x_k\})$, is called a *discrete probability distribution* and the corresponding distribution function $F = F(x)$ is called *discrete*.

We present a table of the commonest types of discrete probability distribution, with their names.

Table 1

Distribution	Probabilities p_k	Parameters
Discrete uniform	$1/N, \quad k = 1, 2, \ldots, N$	$N = 1, 2, \ldots$
Bernoulli	$p_1 = p, \quad p_0 = q$	$0 \le p \le 1, q = 1 - p$
Binomial	$C_n^k p^k q^{n-k}, \quad k = 0, 1, \ldots, n$	$0 \le p \le 1, \quad q = 1 - p,$ $n = 1, 2, \ldots$
Poisson	$e^{-\lambda}\lambda^k/k!, \quad k = 0, 1, \ldots$	$\lambda > 0$
Geometric	$q^{k-1}p, \quad k = 0, 1, \ldots$	$0 \le p \le 1, \quad q = 1 - p$
Negative binomial	$C_{k-1}^{r-1} p^r q^{k-r}, \quad k = r, r + 1, \ldots$	$0 \le p \le 1, \quad q = 1 - p,$ $r = 1, 2, \ldots$

Absolutely continuous measures. These are measures for which the corresponding distribution functions are such that

$$F(x) = \int_{-\infty}^{x} f(t)\, dt, \tag{5}$$

where $f = f(t)$ are nonnegative functions and the integral is at first taken in the Riemann sense, but later (see §6) in that of Lebesgue.

The function $f = f(x)$, $x \in R$, is the *density* of the distribution function $F = F(x)$ (or the density of the probability distribution, or simply the density) and $F = F(x)$ is called absolutely continuous.

It is clear that every nonnegative $f = f(x)$ that is Riemann integrable and such that $\int_{-\infty}^{\infty} f(x)\,dx = 1$ defines a distribution function by (5). Table 2 presents some important examples of various kinds of densities $f = f(x)$ with their names and parameters (a density $f(x)$ is taken to be zero for values of x not listed in the table).

Table 2

Distribution	Density	Parameters
Uniform on $[a, b]$	$1/(b-a), \quad a \le x \le b$	$a, b \in R; \quad a < b$
Normal or Gaussian	$(2\pi\sigma^2)^{-1/2} e^{-(x-m)^2/(2\sigma^2)}, \quad x \in R$	$m \in R, \sigma > 0$
Gamma	$\dfrac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}, \quad x \ge 0$	$\alpha > 0, \beta > 0$
Beta	$\dfrac{x^{r-1}(1-x)^{s-1}}{B(r, s)}, \quad 0 \le x \le 1$	$r > 0, s > 0$
Exponential (gamma with $\alpha = 1, \beta = 1/\lambda$)	$\lambda e^{-\lambda x}, \quad x \ge 0$	$\lambda > 0$
Bilateral exponential	$\frac{1}{2}\lambda e^{-\lambda\|x\|}, \quad x \in R$	$\lambda > 0$
Chi-squared, χ^2 (gamma with a $\alpha = n/2, \beta = 2$)	$2^{-n/2} x^{n/2-1} e^{-x/2}/\Gamma(n/2), \quad x \ge 0$	$n = 1, 2, \ldots$
Student, t	$\dfrac{\Gamma(\frac{1}{2}(n+1))}{(n\pi)^{1/2}\Gamma(n/2)}\left(1 + \dfrac{x^2}{n}\right)^{-(n+1)/2}, \quad x \in R$	$n = 1, 2, \ldots$
F	$\dfrac{(m/n)^{m/2}}{B(m/2, n/2)} \dfrac{x^{m/2-1}}{(1 + mx/n)^{(m+n)/2}}$	$m, n = 1, 2, \ldots$
Cauchy	$\dfrac{\theta}{\pi(x^2 + \theta^2)}, \quad x \in R$	$\theta > 0$

Singular measures. These are measures whose distribution functions are continuous but have all their points of increases on sets of *zero Lebesgue measure*. We do not discuss this case in detail; we merely give an example of such a function.

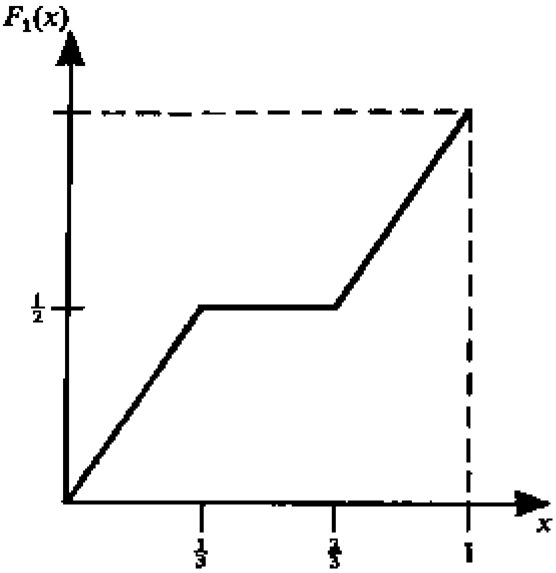


Figure 26

We consider the interval [0, 1] and construct $F(x)$ by the following procedure originated by Cantor.

We divide [0, 1] into thirds and put (Figure 26)

$$F_2(x) = \begin{cases} \frac{1}{2}, & x \in (\frac{1}{3}, \frac{2}{3}), \\ \frac{1}{4}, & x \in (\frac{1}{9}, \frac{2}{9}), \\ \frac{3}{4}, & x \in (\frac{7}{9}, \frac{8}{9}), \\ 0, & x = 0, \\ 1, & x = 1 \end{cases}$$

defining it in the intermediate intervals by linear interpolation.

Then we divide each of the intervals $[0, \frac{1}{3}]$ and $[\frac{2}{3}, 1]$ into three parts and define the function (Figure 27) with its values at other points determined by linear interpolation.

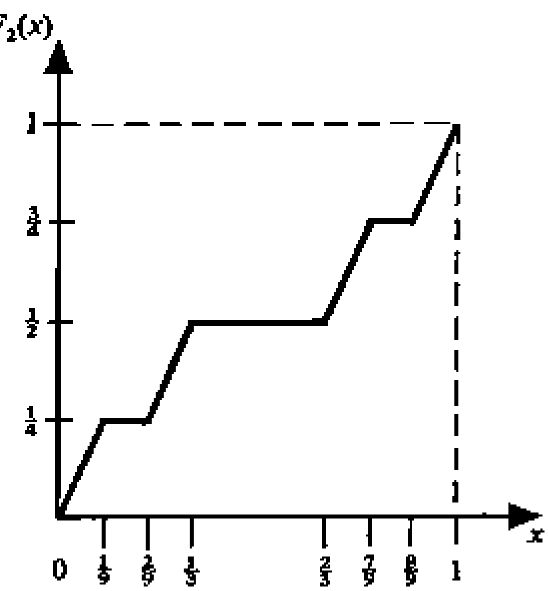


Figure 27

Continuing this process, we construct a sequence of functions $F_n(x)$, $n = 1, 2, \ldots$, which converges to a nondecreasing continuous function $F(x)$ (the Cantor function), whose points of increase (x is a point of increase of $F(x)$ if $F(x + \varepsilon) - F(x - \varepsilon) > 0$ for every $\varepsilon > 0$) form a set of Lebesgue measure zero. In fact, it is clear from the construction of $F(x)$ that the total length of the intervals $(\frac{1}{3}, \frac{2}{3})$, $(\frac{1}{9}, \frac{2}{9})$, $(\frac{7}{9}, \frac{8}{9})$, ... on which the function is constant is

$$\frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \cdots = \frac{1}{3}\sum_{n=0}^{\infty}\left(\frac{2}{3}\right)^n = 1. \tag{6}$$

Let $\mathscr{N}$ be the set of points of increase of the Cantor function $F(x)$. It follows from (6) that $\lambda(\mathscr{N}) = 0$. At the same time, if μ is the measure corresponding to the Cantor function $F(x)$, we have $\mu(\mathscr{N}) = 1$. (We then say that the measure is *singular* with respect to Lebesgue measure λ.)

Without any further discussion of possible types of distribution functions, we merely observe that in fact the *three types* that have been mentioned cover all possibilities. More precisely, every distribution function can be represented in the form $p_1F_1 + p_2F_2 + p_3F_3$, where F_1 is discrete, F_2 is absolutely continuous, and F_3 is singular, and p_i are nonnegative numbers, $p_1 + p_2 + p_3 = 1$.

2. Theorem 1 establishes a one-to-one correspondence between probability measures on $(R, \mathscr{B}(R))$ and distribution functions on R. An analysis of the proof of the theorem shows that in fact a stronger theorem is true, one that in particular lets us introduce Lebesgue measure on the real line.

Let μ be a σ-finite measure on $(\Omega, \mathscr{A})$, where $\mathscr{A}$ is an algebra of subsets of Ω. It turns out that the conclusion of Carathéodory's theorem on the extension of a measure and an algebra $\mathscr{A}$ to a minimal σ-algebra $\sigma(\mathscr{A})$ remains valid with a σ-finite measure; this makes it possible to generalize Theorem 1.

A *Lebesgue–Stieltjes measure* on $(R, \mathscr{B}(R))$ is a (countably additive) measure μ such that the measure $\mu(I)$ of every bounded interval I is finite. A *generalized distribution function* on the real line R is a nondecreasing function $G = G(x)$, with values on $(-\infty, \infty)$, that is continuous on the right.

Theorem 1 can be generalized to the statement that the formula

$$\mu(a, b] = G(b) - G(a), \qquad a < b,$$

again establishes a one-to-one correspondence between Lebesgue–Stieltjes measures μ and generalized distribution functions G.

In fact, if $G(+\infty) - G(-\infty) < \infty$, the proof of Theorem 1 can be taken over without any change, since this case reduces to the case when $G(+\infty) - G(-\infty) = 1$ and $G(-\infty) = 0$.

Now let $G(+\infty) - G(-\infty) = \infty$. Put

$$G_n(x) = \begin{cases} G(x), & |x| \le n, \\ G(n) & x = n, \\ G(-n), & x = -n. \end{cases}$$

On the algebra $\mathscr{A}$ let us define a finitely additive measure μ_0 such that $\mu_0(a, b] = G(b) - G(a)$, and let μ_n be the finitely additive measure previously constructed (by Theorem 1) from $G_n(x)$.

Evidently $\mu_n \uparrow \mu_0$ on $\mathscr{A}$. Now let $A_1, A_2, \ldots$ be disjoint sets in $\mathscr{A}$ and $A \equiv \sum A_n \in \mathscr{A}$. Then (Problem 6 of §1)

$$\mu_0(A) \geq \sum_{n=1}^{\infty} \mu_0(A_n).$$

If $\sum_{n=1}^{\infty} \mu_0(A_n) = \infty$ then $\mu_0(A) = \sum_{n=1}^{\infty} \mu_0(A_n)$. Let us suppose that $\sum \mu_0(A_n) < \infty$. Then

$$\mu_0(A) = \lim_n \mu_n(A) = \lim_n \sum_{k=1}^{\infty} \mu_n(A_k).$$

By hypothesis, $\sum \mu_0(A_n) < \infty$. Therefore

$$0 \leq \mu_0(A) - \sum_{k=1}^{\infty} \mu_0(A_k) = \lim_n \left[\sum_{k=1}^{\infty} (\mu_n(A_k) - \mu_0(A_k)) \right] \leq 0,$$

since $\mu_n \leq \mu_0$.

Thus a σ-finite finitely additive measure μ_0 is countably additive on $\mathscr{A}$, and therefore (by Carathéodory's theorem) it can be extended to a countably additive measure μ on $\sigma(\mathscr{A})$.

The case $G(x) = x$ is particularly important. The measure λ corresponding to this generalized distribution function is Lebesgue measure on $(R, \mathscr{B}(R))$. As for the interval $[0, 1]$ of the real line, we can define the system $\overline{\mathscr{B}}(R)$ by writing $A \in \overline{\mathscr{B}}(R)$ if there are Borel sets A and B such that $A \subseteq \Lambda \subseteq B$, $\lambda(B \setminus A) = 0$. Then Lebesgue measure $\bar{\lambda}$ on $\overline{\mathscr{B}}(R)$ is defined by $\bar{\lambda}(\Lambda) = \lambda(A)$ if $A \subseteq \Lambda \subseteq B$, $\lambda \in \overline{\mathscr{B}}(R)$ and $\Lambda(B \setminus A) = 0$.

3. The measurable space $(R^n, \mathscr{B}(R^n))$. Let us suppose, as for the real line, that P is a probability measure on $(R^n, \mathscr{B}(R^n))$.

Let us write

$$F_n(x_1, \ldots, x_n) = \mathsf{P}((-\infty, x_1] \times \cdots \times (-\infty, x_n]),$$

or, in a more compact form,

$$F_n(x) = \mathsf{P}(-\infty, x],$$

where $x = (x_1, \ldots, x_n)$, $(-\infty, x] = (-\infty, x_1] \times \cdots \times (-\infty, x_n]$.

Let us introduce the difference operator $\Delta_{a_i, b_i}: R^n \to R$, defined by the formula

$$\Delta_{a_i, b_i} F_n(x_1, \ldots, x_n) = F_n(x_1, \ldots, x_{i-1}, b_i, x_{i+1} \ldots) - F_n(x_1, \ldots, x_{i-1}, a_i, x_{i+1} \ldots)$$

where $a_i \le b_i$. A simple calculation shows that

$$\Delta_{a_1 b_1} \cdots \Delta_{a_n b_n} F_n(x_1 \cdots x_n) = \mathsf{P}(a, b], \tag{7}$$

where $(a, b] = (a_1, b_1] \times \cdots \times (a_n, b_n]$. Hence it is clear, in particular, that (in contrast to the one-dimensional case) $\mathsf{P}(a, b]$ is in general not equal to $F_n(b) - F_n(a)$.

Since $\mathsf{P}(a, b] \ge 0$, it follows from (7) that

$$\Delta_{a_1 b_1} \cdots \Delta_{a_n b_n} F_n(x_1, \dots, x_n) \ge 0 \tag{8}$$

for arbitrary $a = (a_1, \dots, a_n)$, $b = (b_1, \dots, b_n)$.

It also follows from the continuity of P that $F_n(x_1, \dots, x_n)$ is continuous on the right with respect to the variables collectively, i.e. if $x^{(k)} \downarrow x$, $x^{(k)} = (x_1^{(k)}, \dots, x_n^{(k)})$, then

$$F_n(x^{(k)}) \downarrow F_n(x), \qquad k \to \infty. \tag{9}$$

It is also clear that

$$F_n(+\infty, \dots, +\infty) = 1 \tag{10}$$

and

$$\lim_{x \downarrow y} F_n(x_1, \dots, x_n) = 0, \tag{11}$$

if at least one coordinate of y is $-\infty$.

Definition 2. An n-dimensional distribution function (on R^n) is a function $F = F(x_1, \dots, x_n)$ with properties (8)–(11).

The following result can be established by the same reasoning as in Theorem 1.

Theorem 2. *Let $F = F_n(x_1, \dots, x_n)$ be a distribution function on R^n. Then there is a unique probability measure P on $(R^n, \mathscr{B}(R^n))$ such that*

$$\mathsf{P}(a, b] = \Delta_{a_1 b_1} \cdots \Delta_{a_n b_n} F_n(x_1, \dots, x_n). \tag{12}$$

Here are some examples of n-dimensional distribution functions.

Let $F^1, \dots, F^n$ be one-dimensional distribution functions (on R) and

$$F_n(x_1, \dots, x_n) = F^1(x_1) \cdots F^n(x_n).$$

It is clear that this function is continuous on the right and satisfies (10) and (11). It is also easy to verify that

$$\Delta_{a_1 b_1} \cdots \Delta_{a_n b_n} F_n(x_1, \dots, x_n) = \prod [F^k(b_k) - F^k(a_k)] \ge 0.$$

Consequently $F_n(x_1, \dots, x_n)$ is a distribution function.

The case when

$$F^k(x_k) = \begin{cases} 0 & x_k < 0, \\ x_k, & 0 \le x_k \le 1, \\ 1, & x_k > 1 \end{cases}$$

is particularly important. In this case

$$F_n(x_1, \ldots, x_n) = x_1 \cdots x_n.$$

The probability measure corresponding to this n-dimensional distribution function is *n-dimensional Lebesgue measure on* $[0, 1]^n$.

Many n-dimensional distribution functions appear in the form

$$F_n(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_n(t_1, \ldots, t_n)\, dt_1 \cdots dt_n,$$

where $f_n(t_1, \ldots, t_n)$ is a nonnegative function such that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_n(t_1, \ldots, t_n)\, dt_1 \cdots dt_n = 1,$$

and the integrals are Riemann (more generally, Lebesgue) integrals. The function $f = f_n(t_1, \ldots, t_n)$ is called the *density* of the n-dimensional distribution function, the density of the n-dimensional probability distribution, or simply an n-dimensional density.

When $n = 1$, the function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/(2\sigma^2)}, \qquad x \in R,$$

with $\sigma > 0$ is the density of the (nondegenerate) *Gaussian* or *normal distribution*. There are natural analogs of this density when $n > 1$.

Let $\mathbb{R} = \|r_{ij}\|$ be a nonnegative definite symmetric $n \times n$ matrix:

$$\sum_{i,j=1}^{n} r_{ij}\lambda_i\lambda_j \ge 0, \qquad \lambda_i \in R, \quad i = 1, \ldots, n, \qquad r_{ij} = r_{ji}.$$

When $\mathbb{R}$ is a positive definite matrix, $|\mathbb{R}| = \det \mathbb{R} > 0$ and consequently there is an inverse matrix $A = \|a_{ij}\|$.

$$f_n(x_1, \ldots, x_n) = \frac{|A|^{1/2}}{(2\pi)^{n/2}} \exp\{-\tfrac{1}{2} \textstyle\sum a_{ij}(x_i - m_i)(x_j - m_j)\}, \tag{13}$$

where $m_i \in R$, $i = 1, \ldots, n$, has the property that its (Riemann) integral over the whole space equals 1 (this will be proved in §13) and therefore, since it is also positive, it is a density.

This function is the *density of the n-dimensional* (nondegenerate) *Gaussian* or *normal distribution* (with vector mean $m = (m_1, \ldots, m_n)$ and covariance matrix $\mathbb{R} = A^{-1}$).

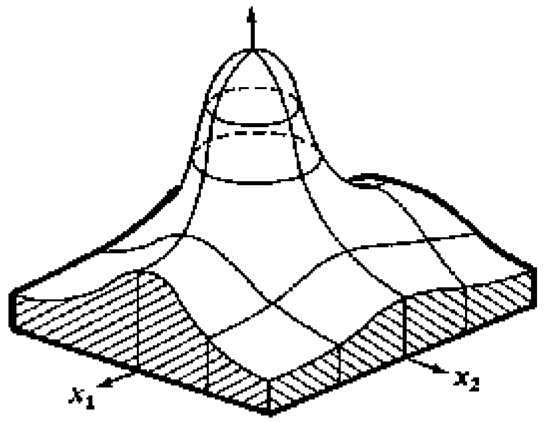


Figure 28. Density of the two-dimensional Gaussian distribution.

When $n = 2$ the density $f_2(x_1, x_2)$ can be put in the form

$$f_2(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-m_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-m_1)(x_2-m_2)}{\sigma_1\sigma_2} + \frac{(x_2-m_2)^2}{\sigma_2^2}\right]\right\}, \tag{14}$$

where $\sigma_i > 0$, $|\rho| < 1$. (The meanings of the parameters m_i, σ_i and ρ will be explained in §8.)

Figure 28 indicates the form of the two-dimensional Gaussian density.

Remark. As in the case $n = 1$, Theorem 2 can be generalized to (similarly defined) Lebesgue–Stieltjes measures on $(R^n, \mathscr{B}(R^n))$ and generalized distribution functions on R^n. When the generalized distribution function $G_n(x_1, \ldots, x_n)$ is $x_1 \cdots x_n$, the corresponding measure is Lebesgue measure on the Borel sets of R^n. It clearly satisfies

$$\lambda(a, b) = \prod_{i=1}^{n} (b_i - a_i),$$

i.e. the Lebesgue measure of the "rectangle"

$$(a, b] = (a_1, b_1] \times \cdots \times (a_n, b_n]$$

is its "content."

4. The measurable space $(R^\infty, \mathscr{B}(R^\infty))$. For the spaces R^n, $n \geq 1$, the probability measures were constructed in the following way: first for elementary sets (rectangles $(a, b]$), then, in a natural way, for sets $A = \sum (a_i, b_i]$, and finally, by using Carathéodory's theorem, for sets in $\mathscr{B}(R^n)$.

A similar construction for probability measures also works for the space $(R^\infty, \mathscr{B}(R^\infty))$.

Let

$$\mathscr{I}_n(B) = \{x \in R^\infty : (x_1, \ldots, x_n) \in B\}, \qquad B \in \mathscr{B}(R^n),$$

denote a cylinder set in R^∞ with base $B \in \mathscr{B}(R^n)$. We see at once that it is natural to take the cylinder sets as *elementary sets* in R^∞, with their probabilities defined by the probability measure on the sets of $\mathscr{B}(R^\infty)$.

Let P be a probability measure on $(R^\infty, \mathscr{B}(R^\infty))$. For $n = 1, 2, \ldots,$ we take

$$P_n(B) = \mathsf{P}(\mathscr{I}_n(B)), \qquad B \in \mathscr{B}(R^n). \tag{15}$$

The sequence of probability measures $P_1, P_2, \ldots$ defined respectively on $(R, \mathscr{B}(R)), (R^2, \mathscr{B}(R^2)), \ldots,$ has the following evident consistency property: for $n = 1, 2, \ldots$ and $B \in \mathscr{B}(R^n)$,

$$P_{n+1}(B \times R) = P_n(B). \tag{16}$$

It is noteworthy that the converse also holds.

Theorem 3 (Kolmogorov's Theorem on the Extension of Measures in $(R^\infty, \mathscr{B}(R^\infty))$). *Let $P_1, P_2, \ldots$ be a sequence of probability measures on $(R, \mathscr{B}(R)), (R^2, \mathscr{B}(R^2)), \ldots,$ possessing the consistency property* (16). *Then there is a unique probability measure P on $(R^\infty, \mathscr{B}(R^\infty))$ such that*

$$\mathsf{P}(\mathscr{I}_n(B)) = P_n(B), \qquad B \in \mathscr{B}(R^n). \tag{17}$$

for $n = 1, 2, \ldots$.

Proof. Let $B^n \in \mathscr{B}(R^n)$ and let $\mathscr{I}_n(B^n)$ be the cylinder with base B^n. We assign the measure $\mathsf{P}(\mathscr{I}_n(B^n))$ to this cylinder by taking $\mathsf{P}(\mathscr{I}_n(B^n)) = P_n(B^n)$.

Let us show that, in virtue of the consistency condition, this definition is consistent, i.e. the value of $\mathsf{P}(\mathscr{I}_n(B^n))$ is independent of the representation of the set $\mathscr{I}_n(B^n)$. In fact, let the same cylinder be represented in two way:

$$\mathscr{I}_n(B^n) = \mathscr{I}_{n+k}(B^{n+k}).$$

It follows that, if $(x_1, \ldots, x_{n+k}) \in R^{n+k}$, we have

$$(x_1, \ldots, x_n) \in B^n \Leftrightarrow (x_1, \ldots, x_{n+k}) \in B^{n+k}, \tag{18}$$

and therefore, by (16) and (18),

$$\begin{aligned} P_n(B^n) &= P_{n+1}((x_1, \ldots, x_{n+1}) : (x_1, \ldots, x_n) \in B^n) \\ &= \cdots = P_{n+k}((x_1, \ldots, x_{n+k}) : (x_1, \ldots, x_n) \in B^n) \\ &= P_{n+k}(B^{n+k}). \end{aligned}$$

Let $\mathscr{A}(R^\infty)$ denote the collection of all cylinder sets $\hat{B}^n = \mathscr{I}_n(B^n)$, $B^n \in \mathscr{B}(R^n)$, $n = 1, 2, \ldots$.

Now let $\hat{B}_1, \ldots, \hat{B}_k$ be disjoint sets in $\mathscr{A}(R^\infty)$. We may suppose without loss of generality that $\hat{B}_i = \mathscr{I}_n(B_i^n)$, $i = 1, \ldots, k$, for some n, where $B_1^n, \ldots, B_k^n$ are disjoint sets in $\mathscr{B}(R^n)$. Then

$$\mathsf{P}\left(\sum_{i=1}^k \hat{B}_i\right) = \mathsf{P}\left(\sum_{i=1}^k \mathscr{I}_n(B_i^n)\right) = P_n\left(\sum_{i=1}^k B_i^n\right) = \sum_{i=1}^n P_n(B_i^n) = \sum_{i=1}^n \mathsf{P}(\hat{B}_i),$$

i.e. the set function P is finitely additive on the algebra $\mathscr{A}(R^\infty)$.

Let us show that P is "continuous at zero," i.e. if the sequence of sets $\hat{B}_n \downarrow \varnothing$, $n \to \infty$, then $\mathsf{P}(\hat{B}_n) \to 0$, $n \to \infty$. Suppose the contrary, i.e. let $\lim \mathsf{P}(\hat{B}_n) = \delta > 0$. We may suppose without loss of generality that $\{\hat{B}_n\}$ has the form

$$\hat{B}_n = \{x : (x_1, \ldots, x_n) \in B_n\}, \qquad B_n \in \mathscr{B}(R^n).$$

We use the following property of probability measures P_n on $(R^n, \mathscr{B}(R^n))$ (see Problem 9): if $B_n \in \mathscr{B}(R^n)$, for a given $\delta > 0$ we can find a compact set $A_n \in \mathscr{B}(R^n)$ such that $A_n \subseteq B_n$ and

$$P_n(B_n \backslash A_n) \le \delta/2^{n+1}.$$

Therefore if

$$\hat{A}_n = \{x : (x_1, \ldots, x_n) \in A_n\},$$

we have

$$\mathsf{P}(\hat{B}_n \backslash \hat{A}_n) = P_n(B_n \backslash A_n) \le \delta/2^{n+1}.$$

Form the set $\hat{C}_n = \bigcap_{k=1}^n \hat{A}_k$ and let C_n be such that

$$\hat{C}_n = \{x : (x_1, \ldots, x_n) \in C_n\}.$$

Then, since the sets $\hat{B}_n$ decrease, we obtain

$$\mathsf{P}(\hat{B}_n \backslash \hat{C}_n) \le \sum_{k=1}^n P(\hat{B}_n \backslash \hat{A}_k) \le \sum_{k=1}^n \mathsf{P}(\hat{B}_k \backslash A_k) \le \delta/2.$$

But by assumption $\lim_n \mathsf{P}(\hat{B}_n) = \delta > 0$, and therefore $\lim_n \mathsf{P}(\hat{C}_n) \ge \delta/2 > 0$. Let us show that this contradicts the condition $\hat{C}_n \downarrow \varnothing$.

Let us choose a point $\hat{x}^{(n)} = (x_1^{(n)}, x_2^{(n)}, \ldots)$ in $\hat{C}_n$. Then $(x_1^{(n)}, \ldots, x_n^{(n)}) \in C_n$ for $n \ge 1$.

Let (n_1) be a subsequence of (n) such that $x_1^{(n_1)} \to x_1^0$, where x_1^0 is a point in C_1. (Such a sequence exists since $x_1^{(n)} \in C_1$ and C_1 is compact.) Then select a subsequence (n_2) of (n_1) such that $(x_1^{(n_2)}, x_2^{(n_2)}) \to (x_1^0, x_2^0) \in C_2$. Similarly let $(x_1^{(n_k)}, \ldots, x_k^{(n_k)}) \to (x_1^0, \ldots, x_k^0) \in C_k$. Finally form the diagonal sequence (m_k), where m_k is the kth term of (n_k). Then $x_i^{(m_k)} \to x_i^0$ as $m_k \to \infty$ for $i = 1, 2, \ldots$; and $(x_1^0, x_2^0, \ldots) \in \hat{C}_n$ for $n = 1, 2, \ldots$, which evidently contradicts the assumption that $\hat{C}_n \downarrow \varnothing$, $n \to \infty$. This completes the proof of the theorem.

Remark. In the present case, the space R^∞ is a countable product of lines, $R^\infty = R \times R \times \cdots$. It is natural to ask whether Theorem 3 remains true if $(R^\infty, \mathscr{B}(R^\infty))$ is replaced by a direct product of measurable spaces $(\Omega_i, \mathscr{F}_i)$, $i = 1, 2, \ldots$.

We may notice that in the preceding proof the only topological property of the real line that was used was that every set in $\mathscr{B}(R^n)$ contains a compact subset whose probability measure is arbitrarily close to the probability measure of the whole set. It is known, however, that this is a property not only of spaces $(R^n, \mathscr{B}(R^n))$, but also of arbitrary complete separable metric spaces with σ-algebras generated by the open sets.

Consequently Theorem 3 remains valid if we suppose that $P_1, P_2, \ldots$ is a sequence of consistent probability measures on $(\Omega_1, \mathscr{F}_1)$,

$$(\Omega_1 \times \Omega_2, \mathscr{F}_1 \otimes \mathscr{F}_2), \ldots,$$

where $(\Omega_i, \mathscr{F}_i)$ are complete separable metric spaces with σ-algebras $\mathscr{F}_i$ generated by open sets, and $(R^\infty, \mathscr{B}(R^\infty))$ is replaced by

$$(\Omega_1 \times \Omega_2 \times \cdots, \mathscr{F}_1 \otimes \mathscr{F}_2 \otimes \cdots).$$

In §9 (Theorem 2) it will be shown that the result of Theorem 3 remains valid for arbitrary measurable spaces $(\Omega_i, \mathscr{F}_i)$ if the measures P_n are concentrated in a particular way. However, Theorem 3 may fail in the general case (without any hypotheses on the topological nature of the measurable spaces or on the structure of the family of measures $\{P_n\}$). This is shown by the following example.

Let us consider the space $\Omega = (0, 1]$, which is evidently not complete, and construct a sequence $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \cdots$ of σ-algebras in the following way. For $n = 1, 2, \ldots$, let

$$\varphi_n(\omega) = \begin{cases} 1, & 0 < \omega < 1/n, \\ 0, & 1/n \le \omega \le 1, \end{cases}$$

$$\mathscr{C}_n = \{A \in \Omega: A = \{\omega: \varphi_n(\omega) \in B\}, B \in \mathscr{B}(R)\}$$

and let $\mathscr{F}_n = \sigma\{\mathscr{C}_1, \ldots, \mathscr{C}_n\}$ be the smallest σ-algebra containing the sets $\mathscr{C}_1, \ldots, \mathscr{C}_n$. Clearly $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \cdots$. Let $\mathscr{F} = \sigma(\bigcup \mathscr{F}_n)$ be the smallest σ-algebra containing all the $\mathscr{F}_n$. Consider the measurable space $(\Omega, \mathscr{F}_n)$ and define a probability measure P_n on it as follows:

$$P_n\{\omega: (\varphi_1(\omega), \ldots, \varphi_n(\omega)) \in B^n\} = \begin{cases} 1 & \textit{if } (1, \ldots, 1) \in B^n, \\ 0 & \textit{otherwise}, \end{cases}$$

where $B^n = \mathscr{B}(R^n)$. It is easy to see that the family $\{P_n\}$ is consistent: if $A \in \mathscr{F}_n$ then $P_{n+1}(A) = P_n(A)$. However, we claim that there is no probability measure P on $(\Omega, \mathscr{F})$ such that its *restriction* $\mathsf{P} | \mathscr{F}_n$ (i.e., the measure P

considered only on sets in $\mathscr{F}_n$) coincides with P_n for $n = 1, 2, \dots$. In fact, let us suppose that such a probability measure P exists. Then

$$\mathsf{P}\{\omega: \varphi_1(\omega) = \cdots = \varphi_n(\omega) = 1\} = P_n\{\omega: \varphi_1(\omega) = \cdots = \varphi_n(\omega) = 1\} = 1 \tag{19}$$

for $n = 1, 2, \dots$. But

$$\{\omega: \varphi_1(\omega) = \cdots = \varphi_n(\omega) = 1\} = (0, 1/n) \downarrow \varnothing,$$

which contradicts (19) and the hypothesis of countable additivity (and therefore continuity at the "zero" $\varnothing$) of the set function P.

We now give an example of a probability measure on $(R^\infty, \mathscr{B}(R^\infty))$. Let $F_1(x), F_2(x), \dots$ be a sequence of one-dimensional distribution functions. Define the functions $G(x) = F_1(x)$, $G_2(x_1, x_2) = F_1(x_1)F_2(x_2), \dots$, and denote the corresponding probability measures on $(R, \mathscr{B}(R))$, $(R^2, \mathscr{B}(R^2)), \dots$ by $P_1, P_2, \dots$. Then it follows from Theorem 3 that there is a measure P on $(R^\infty, \mathscr{B}(R^\infty))$ such that

$$P\{x \in R^\infty: (x_1, \dots, x_n) \in B\} = P_n(B), \qquad B \in \mathscr{B}(R^n)$$

and, in particular,

$$\mathsf{P}\{x \in R^\infty: x_1 \le a_1, \dots, x_n \le a_n\} = F_1(a_1) \cdots F_n(a_n).$$

Let us take $F_i(x)$ to be a Bernoulli distribution,

$$F_i(x) = \begin{cases} 0, & x < 0, \\ q, & 0 \le x < 1, \\ 1, & x \ge 1. \end{cases}$$

Then we can say that there is a probability measure P on the space Ω of sequences of numbers $x = (x_1, x_2, \dots)$, $x_i = 0$ or 1, together with the σ-algebra of its Borel subsets, such that

$$\mathsf{P}\{x: x_1 = a_1, \dots, x_n = a_n\} = p^{\Sigma a_i} q^{n - \Sigma a_i}.$$

This is precisely the result that was not available in the first chapter for stating the law of large numbers in the form (I.5.8).

5. The measurable space $(R^T, \mathscr{B}(R^T))$. Let T be a set of indices $t \in T$ and R_t a real line corresponding to the index t. We consider a finite unordered set $\tau = [t_1, \dots, t_n]$ of distinct indices t_i, $t_i \in T$, $n \ge 1$, and let P_τ be a probability measure on $(R^\tau, \mathscr{B}(R^\tau))$, where $R^\tau = R_{t_1} \times \cdots \times R_{t_n}$.

We say that the family $\{P_\tau\}$ of probability measures, where τ runs through all finite unordered sets, is *consistent* if, for all sets $\tau = [t_1, \dots, t_n]$ and $\sigma = [s_1, \dots, s_k]$ such that $\sigma \subseteq \tau$ we have

$$P_\sigma\{(x_{s_1}, \dots, x_{s_k}): (x_{s_1}, \dots, x_{s_k}) \in B\} = P_\tau\{(x_{t_1}, \dots, x_{t_n}): (x_{s_1}, \dots, x_{s_k}) \in B\} \tag{20}$$

for every $B \in \mathscr{B}(R^\sigma)$.

Theorem 4 (Kolmogorov's Theorem on the Extension of Measures in $(R^T, \mathscr{B}(R^T))$). *Let $\{P_\tau\}$ be a consistent family of probability measures on $(R^\tau, \mathscr{B}(R^\tau))$. Then there is a unique probability measure* P *on $(R^T, \mathscr{B}(R^T))$ such that*

$$\mathsf{P}(\mathscr{I}_\tau(B)) = P_\tau(B) \tag{21}$$

for all unordered sets $\tau = [t_1, \ldots, t_n]$ of different indices $t_i \in T$, $B \in \mathscr{B}(R^\tau)$ and $\mathscr{I}_\tau(B) = \{x \in R^T: (x_{t_1}, \ldots, x_{t_n}) \in B\}$.

Proof. Let the set $\hat{B} \in \mathscr{B}(R^T)$. By the theorem of §2 there is an at most countable set $S = \{s_1, s_2, \ldots\} \subseteq T$ such that $\hat{B} = \{x: (x_{s_1}, x_{s_2}, \ldots) \in B\}$, where $B \in \mathscr{B}(R^S)$, $R^S = R_{s_1} \times R_{s_2} \times \cdots$. In other words, $\hat{B} = \mathscr{I}_S(B)$ is a cylinder set with base $B \in \mathscr{B}(R^S)$.

We can define a set function P on such cylinder sets by putting

$$\mathsf{P}(\mathscr{I}_S(B)) = P_S(B), \tag{22}$$

where P_S is the probability measure whose existence is guaranteed by Theorem 3. We claim that P is in fact the measure whose existence is asserted in the theorem. To establish this we first verify that the definition (22) is consistent, i.e. that it leads to a unique value of $\mathsf{P}(\hat{B})$ for all possible representations of $\hat{B}$; and second, that this set function is countably additive.

Let $\hat{B} = \mathscr{I}_{S_1}(B_1)$ and $\hat{B} = \mathscr{I}_{S_2}(B_2)$. It is clear that then $\hat{B} = \mathscr{I}_{S_1 \cup S_2}(B_3)$ with some $B_3 \in \mathscr{B}(R^{S_1 \cup S_2})$; therefore it is enough to show that if $S \subseteq S'$ and $B \in \mathscr{B}(R^S)$, then $P_{S'}(B') = P_S(B)$, where

$$B' = \{(x_{s_1'}, x_{s_2'}, \ldots): (x_{s_1}, x_{s_2}, \ldots) \in B\}$$

with $S' = \{s_1', s_2', \ldots\}$, $S = \{s_1, s_2, \ldots\}$. But by the assumed consistency of (20) this equation follows immediately from Theorem 3. This establishes that the value of $\mathsf{P}(\hat{B})$ is independent of the representation of $\hat{B}$.

To verify the countable additivity of P, let us suppose that $\{\hat{B}_n\}$ is a sequence of pairwise disjoint sets in $\mathscr{B}(R^T)$. Then there is an at most countable set $S \subseteq T$ such that $\hat{B}_n = \mathscr{I}_S(B_n)$ for all $n \geq 1$, where $B_n \in \mathscr{B}(R^S)$. Since P_S is a probability measure, we have

$$\begin{aligned}\mathsf{P}(\textstyle\sum \hat{B}_n) &= \mathsf{P}(\textstyle\sum \mathscr{I}_S(B_n)) = P_S(\textstyle\sum B_n) = \textstyle\sum P_S(B_n) \\ &= \textstyle\sum \mathsf{P}(I_S(B_n)) = \textstyle\sum \mathsf{P}(\hat{B}_n).\end{aligned}$$

Finally, property (21) follows immediately from the way in which P was constructed.

This completes the proof.

Remark 1. We emphasize that T is *any* set of indices. Hence, by the remark after Theorem 3, the present theorem remains valid if we replace the real lines R_t by arbitrary complete separable metric spaces Ω_t (with σ-algebras generated by open sets).

Remark 2. The original probability measures $\{P_\tau\}$ were assumed defined on *unordered* sets $\tau = [t_1, \ldots, t_n]$ of different indices. It is also possible to start from a family of probability measures $\{P_\tau\}$ where τ runs through all *ordered* sets $\tau = (t_1, \ldots, t_n)$ of different indices. In this case, in order to have Theorem 4 hold we have to adjoin to (20) a further *consistency condition*:

$$P_{(t_1,\ldots,t_n)}(A_{t_1} \times \cdots \times A_{t_n}) = P_{(t_{i_1},\ldots,t_{i_n})}(A_{t_{i_1}} \times \cdots \times A_{t_{i_n}}), \tag{23}$$

where $(i, \ldots, i_n)$ is an arbitrary permutation of $(1, \ldots, n)$ and $A_{t_i} \in \mathscr{B}(R_{t_i})$. As a necessary condition for the existence of P this follows from (21) (with $P_{[t_1,\ldots,t_n]}(B)$ replaced by $P_{(t_1,\ldots,t_n)}(B)$).

From now on we shall assume that the sets τ under consideration are *unordered*. If T is a subset of the real line (or some completely ordered set), we may assume without loss of generality that the set $\tau = [t_1, \ldots, t_n]$ satisfies $t_1 < t_2 < \cdots < t_n$. Consequently it is enough to define "finite-dimensional" probabilities only for sets $\tau = [t_1, \ldots, t_n]$ for which $t_1 < t_2 < \cdots < t_n$.

Now consider the case $T = [0, \infty)$. Then R^T is the space of all real functions $x = (x_t)_{t \geq 0}$. A fundamental example of a probability measure on $(R^{[0,\infty)}, \mathscr{B}(R^{[0,\infty)}))$ is Wiener measure, constructed as follows.

Consider the family $\{\varphi_t(y|x)\}_{t \geq 0}$ of Gaussian densities (as functions of y for fixed x):

$$\varphi_t(y|x) = \frac{1}{\sqrt{2\pi t}} e^{-(y-x)^2/2t}, \qquad y \in R,$$

and for each $\tau = [t_1, \ldots, t_n]$, $t_1 < t_2 < \cdots < t_n$, and each set

$$B = I_1 \times \cdots \times I_n, \qquad I_k = (a_k, b_k),$$

construct the measure $P_\tau(B)$ according to the formula

$$P_\tau(I_1 \times \cdots \times I_n)$$
$$= \int_{I_1} \cdots \int_{I_n} \varphi_{t_1}(a_1|0)\varphi_{t_2-t_1}(a_2|a_1) \cdots \varphi_{t_n-t_{n-1}}(a_n|a_{n-1})\, da_1 \cdots da_n \tag{24}$$

(integration in the Riemann sense). Now we define the set function P for each cylinder set $\mathscr{I}_{t_1\ldots t_n}(I_1 \times \cdots \times I_n) = \{x \in R^T : x_{t_1} \in I_1, \ldots, x_{t_n} \in I_n\}$ by taking

$$\mathsf{P}(\mathscr{I}_{t_1\ldots t_n}(I_1 \times \cdots \times I_n)) = P_{[t_1\ldots t_n]}(I_1 \times \cdots \times I_n).$$

The intuitive meaning of this method of assigning a measure to the cylinder set $\mathscr{I}_{t_1\ldots t_n}(I_1 \times \cdots \times I_n)$ is as follows.

The set $\mathscr{I}_{t_1\ldots t_n}(I_1 \times \cdots \times I_n)$ is the set of functions that at times $t_1, \ldots, t_n$ pass through the "windows" $I_1, \ldots, I_n$ (see Figure 24 in §2). We shall interpret

$\varphi_{t_k - t_{k-1}}(a_k | a_{k-1})$ as the probability that a particle, starting at a_{k-1} at time $t_k - t_{k-1}$, arrives in a neighborhood of a_k. Then the product of densities that appears in (24) describes a certain independence of the increments of the displacements of the moving "particle" in the time intervals

$$[0, t_1], [t_1, t_2], \ldots, [t_{n-1}, t_n].$$

The family of measures $\{P_\tau\}$ constructed in this way is easily seen to be consistent, and therefore can be extended to a measure on $(R^{[0,\infty)}, \mathscr{B}(R^{[0,\infty)}))$. The measure so obtained plays an important role in probability theory. It was introduced by N. Wiener and is known as *Wiener measure*.

6. Problems

1. Let $F(x) = \mathsf{P}(-\infty, x]$. Verify the following formulas:

$$\mathsf{P}(a, b] = F(b) - F(a), \qquad \mathsf{P}(a, b) = F(b-) - F(a),$$
$$\mathsf{P}[a, b] = F(b) - F(a-), \qquad \mathsf{P}[a, b) = F(b-) - F(a-),$$
$$\mathsf{P}\{x\} = F(x) - F(x-),$$

where $F(x-) = \lim_{y \uparrow x} F(y)$.

2. Verify (7).

3. Prove Theorem 2.

4. Show that a distribution function $F = F(x)$ on R has at most a countable set of points of discontinuity. Does a corresponding result hold for distribution functions on R^n?

5. Show that each of the functions

$$G(x, y) = \begin{cases} 1, & x + y \geq 0, \\ 0, & x + y < 0, \end{cases}$$

$$G(x, y) = [x + y], \text{ the integral part of } x + y,$$

is continuous on the right, and continuous in each argument, but is not a (generalized) distribution function on R^2.

6. Let μ be the Lebesgue–Stieltjes measure generated by a continuous distribution function. Show that if the set A is at most countable, then $\mu(A) = 0$.

7. Let c be the cardinal number of the continuum. Show that the cardinal number of the collection of Borel sets in R^n is c, whereas that of the collection of Lebesgue measurable sets is 2^c.

8. Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space and $\mathscr{A}$ an algebra of subsets of Ω such that $\sigma(\mathscr{A}) = \mathscr{F}$. Using the principle of appropriate sets, prove that for every $\varepsilon > 0$ and $B \in \mathscr{F}$ there is a set $A \in \mathscr{A}$ such that

$$P(A \triangle B) \leq \varepsilon.$$

9. Let P be a probability measure on $(R^n, \mathscr{B}(R^n))$. Using Problem 8, show that, for every $\varepsilon > 0$ and $B \in \mathscr{B}(R^n)$, there is a compact subset A of $\mathscr{B}(R^n)$ such that $A \subseteq B$ and
$$\mathsf{P}(B \setminus A) \le \varepsilon.$$
(This was used in the proof of Theorem 1.)

10. Verify the consistency of the measure defined by (21).

§4. Random Variables. I

1. Let $(\Omega, \mathscr{F})$ be a measurable space and let $(R, \mathscr{B}(R))$ be the real line with the system $\mathscr{B}(R)$ of Borel sets.

Definition 1. A real function $\xi = \xi(\omega)$ defined on (Ω, F) is an *$\mathscr{F}$-measurable function*, or a *random variable*, if
$$\{\omega\colon \xi(\omega) \in B\} \in \mathscr{F} \tag{1}$$
for every $B \in \mathscr{B}(R)$; or, equivalently, if the inverse image
$$\xi^{-1}(B) \equiv \{\omega\colon \xi(\omega) \in B\}$$
is a measurable set in Ω.

When $(\Omega, \mathscr{F}) = (R^n, \mathscr{B}(R^n))$, the $\mathscr{B}(R^n)$-measurable functions are called *Borel functions*.

The simplest example of a random variable is the indicator $I_A(\omega)$ of an arbitrary (measurable) set $A \in \mathscr{F}$.

A random variable ξ that has a representation
$$\xi(\omega) = \sum_{i=1}^{\infty} x_i I_{A_i}(\omega), \tag{2}$$
where $\sum A_i = \Omega$, $A_i \in \mathscr{F}$, is called *discrete*. If the sum in (2) is finite, the random variable is called *simple*.

With the same interpretation as in §4 of Chapter I, we may say that a random variable is a numerical property of an experiment, with a value depending on "chance." Here the requirement (1) of measurability is fundamental, for the following reason. If a probability measure P is defined on $(\Omega, \mathscr{F})$, it then makes sense to speak of the probability of the event $\{\xi(\omega) \in B\}$ that the value of the random variable belongs to a Borel set B.

We introduce the following definitions.

Definition 2. A probability measure P_ξ on $(R, \mathscr{B}(R))$ with
$$P_\xi(B) = \mathsf{P}\{\omega\colon \xi(\omega) \in B\}, \qquad B \in \mathscr{B}(R),$$
is called the *probability distribution of* ξ on $(R, \mathscr{B}(R))$.

Definition 3. The function

$$F_\xi(x) = \mathsf{P}\{\omega: \xi(\omega) \le x\}, \qquad x \in R,$$

is called the *distribution function of* ξ.

For a discrete random variable the measure P_ξ is concentrated on an at most countable set and can be represented in the form

$$P_\xi(B) = \sum_{\{k:\, x_k \in B\}} p(x_k), \tag{3}$$

where $p(x_k) = \mathsf{P}\{\xi = x_k\} = \Delta F_\xi(x_k)$.

The converse is evidently true: If P_ξ is represented in the form (3) then ξ is a *discrete* random variable.

A random variable ξ is called *continuous* if its distribution function $F_\xi(x)$ is continuous for $x \in R$.

A random variable ξ is called *absolutely continuous* if there is a nonnegative function $f = f_\xi(x)$, called its density, such that

$$F_\xi(x) = \int_{-\infty}^{x} f_\xi(y)\, dy, \qquad x \in R, \tag{4}$$

(the integral can be taken in the Riemann sense, or more generally in that of Lebesgue; see §6 below).

2. To establish that a function $\xi = \xi(\omega)$ is a random variable, we have to verify property (1) for all sets $B \in \mathscr{F}$. The following lemma shows that the class of such "test" sets can be considerably narrowed.

Lemma 1. *Let $\mathscr{E}$ be a system of sets such that $\sigma(\mathscr{E}) = \mathscr{B}(R)$. A necessary and sufficient condition that a function $\xi = \xi(\omega)$ is $\mathscr{F}$-measurable is that*

$$\{\omega: \xi(\omega) \in E\} \in \mathscr{F} \tag{5}$$

for all $E \in \mathscr{E}$.

Proof. The necessity is evident. To prove the sufficiency we again use the principle of appropriate sets.

Let $\mathscr{D}$ be the system of those Borel sets D in $\mathscr{B}(R)$ for which $\xi^{-1}(D) \in \mathscr{F}$. The operation "form the inverse image" is easily shown to preserve the set-theoretic operations of union, intersection and complement:

$$\begin{gathered} \xi^{-1}\Big(\bigcup_\alpha B_\alpha\Big) = \bigcup_\alpha \xi^{-1}(B_\alpha), \\ \xi^{-1}\Big(\bigcap_\alpha B_\alpha\Big) = \bigcap_\alpha \xi^{-1}(B_\alpha), \\ \overline{\xi^{-1}(B_\alpha)} = \xi^{-1}\overline{(B_\alpha)}. \end{gathered} \tag{6}$$

It follows that $\mathscr{D}$ is a σ-algebra. Therefore

$$\mathscr{E} \subseteq \mathscr{D} \subseteq \mathscr{B}(R)$$

and

$$\sigma(\mathscr{E}) \subseteq \sigma(\mathscr{D}) = \mathscr{D} \subseteq \mathscr{B}(R).$$

But $\sigma(E) = \mathscr{B}(R)$ and consequently $\mathscr{D} = \mathscr{B}(R)$.

Corollary. *A necessary and sufficient condition for $\xi = \xi(\omega)$ to be a random variable is that*

$$\{\omega: \xi(\omega) < x\} \in \mathscr{F}$$

for every $x \in R$, or that

$$\{\omega: \xi(\omega) \leq x\} \in \mathscr{F}$$

for every $x \in R$.

The proof is immediate, since each of the systems

$$\mathscr{E}_1 = \{x: x < c, c \in R\},$$
$$\mathscr{E}_2 = \{x: x \leq c, c \in R\}$$

generates the σ-algebra $\mathscr{B}(R)$: $\sigma(E_1) = \sigma(E_2) = \mathscr{B}(R)$ (see §2).

The following lemma makes it possible to construct random variables as functions of other random variables.

Lemma 2. *Let $\varphi = \varphi(x)$ be a Borel function and $\xi = \xi(\omega)$ a random variable. Then the composition $\eta = \varphi \circ \xi$, i.e. the function $\eta(\omega) = \varphi(\xi(\omega))$, is also a random variable.*

The proof follows from the equations

$$\{\omega: \eta(\omega) \in B\} = \{\omega: \varphi(\xi(\omega)) \in B\} = \{\omega: \xi(\omega) \in \varphi^{-1}(B)\} \in \mathscr{F} \tag{7}$$

for $B \in \mathscr{B}(R)$, since $\varphi^{-1}(B) \in \mathscr{B}(R)$.

Therefore if ξ is a random variable, so are, for examples, ξ^n, $\xi^+ = \max(\xi, 0)$, $\xi^- = -\min(\xi, 0)$, and $|\xi|$, since the functions x^n, x^+, x^- and $|x|$ are Borel functions (Problem 4).

3. Starting from a given collection of random variables $\{\xi_n\}$, we can construct new functions, for example, $\sum_{k=1}^{\infty} |\xi_k|$, $\overline{\lim}\, \xi_n$, $\underline{\lim}\, \xi_n$, etc. Notice that in general such functions take values on the extended real line $\bar{R} = [-\infty, \infty]$. Hence it is advisable to extend the class of $\mathscr{F}$-measurable functions somewhat by allowing them to take the values $\pm\infty$.

Definition 4. A function $\xi = \xi(\omega)$ defined on $(\Omega, \mathscr{F})$ with values in $\bar{R} = [-\infty, \infty]$ will be called an *extended random variable* if condition (1) is satisfied for every Borel set $B \in \mathscr{B}(\bar{R})$.

The following theorem, despite its simplicity, is the key to the construction of the Lebesgue integral (§6).

Theorem 1.

(a) *For every random variable $\xi = \xi(\omega)$ (extended ones included) there is a sequence of simple random variables $\xi_1, \xi_2, \ldots$, such that $|\xi_n| \le |\xi|$ and $\xi_n(\omega) \to \xi(\omega)$, $n \to \infty$, for all $\omega \in \Omega$.*

(b) *If also $\xi(\omega) \ge 0$, there is a sequence of simple random variables $\xi_1, \xi_2, \ldots$, such that $\xi_n(\omega) \uparrow \xi(\omega)$, $n \to \infty$, for all $\omega \in \Omega$.*

PROOF. We begin by proving the second statement. For $n = 1, 2, \ldots$, put

$$\xi_n(\omega) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} I_{k,n}(\omega) + nI_{\{\xi(\omega) \ge n\}}(\omega),$$

where $I_{k,n}$ is the indicator of the set $\{(k-1)/2^n \le \xi(\omega) < k/2^n\}$. It is easy to verify that the sequence $\xi_n(\omega)$ so constructed is such that $\xi_n(\omega) \uparrow \xi(\omega)$ for all $\omega \in \Omega$. The first statement follows from this if we merely observe that ξ can be represented in the form $\xi = \xi^+ - \xi^-$. This completes the proof of the theorem.

We next show that the class of extended random variables is closed under pointwise convergence. For this purpose, we note first that if $\xi_1, \xi_2, \ldots$ is a sequence of extended random variables, then $\sup \xi_n$, $\inf \xi_n$, $\overline{\lim}\, \xi_n$ and $\underline{\lim}\, \xi_n$ are also random variables (possibly extended). This follows immediately from

$$\{\omega: \sup \xi_n > x\} = \bigcup_n \{\omega: \xi_n > x\} \in \mathscr{F},$$

$$\{\omega: \inf \xi_n < x\} = \bigcup_n \{\omega: \xi_n < x\} \in \mathscr{F},$$

and

$$\overline{\lim}\, \xi_n = \inf_n \sup_{m \ge n} \xi_m, \qquad \underline{\lim}\, \xi_n = \sup_n \inf_{m \ge n} \xi_m.$$

Theorem 2. *Let $\xi_1, \xi_2, \ldots$ be a sequence of extended random variables and $\xi(\omega) = \lim \xi_n(\omega)$. Then $\xi(\omega)$ is also an extended random variable.*

The proof follows immediately from the remark above and the fact that

$$\begin{aligned}\{\omega: \xi(\omega) < x\} &= \{\omega: \lim \xi_n(\omega) < x\} \\ &= \{\omega: \overline{\lim}\, \xi_n(\omega) = \underline{\lim}\, \xi_n(\omega)\} \cap \{\overline{\lim}\, \xi_n(\omega) < x\} \\ &= \Omega \cap \{\overline{\lim}\, \xi_n(\omega) < x\} = \{\overline{\lim}\, \xi_n(\omega) < x\} \in \mathscr{F}.\end{aligned}$$

4. We mention a few more properties of the simplest functions of random variables considered on the measurable space $(\Omega, \mathscr{F})$ and possibly taking values on the extended real line $\bar{R} = [-\infty, \infty]$.†

If ξ and η are random variables, $\xi + \eta$, $\xi - \eta$, $\xi\eta$, and ξ/η are also random variables (assuming that they are defined, i.e. that no indeterminate forms like $\infty - \infty$, ∞/∞, $a/0$ occur.

In fact, let $\{\xi_n\}$ and $\{\eta_n\}$ be sequences of random variables converging to ξ and η (see Theorem 1). Then

$$\xi_n \pm \eta_n \to \xi \pm \eta,$$

$$\xi_n \eta_n \to \xi\eta,$$

$$\frac{\xi_n}{\eta_n + \dfrac{1}{n} I_{\{\eta_n = 0\}}(\omega)} \to \frac{\xi}{\eta}.$$

The functions on the left-hand sides of these relations are simple random variables. Therefore, by Theorem 2, the limit functions $\xi \pm \eta$, $\xi\eta$ and ξ/η are also random variables.

5. Let ξ be a random variable. Let us consider sets from $\mathscr{F}$ of the form $\{\omega: \xi(\omega) \in B\}$, $B \in \mathscr{B}(R)$. It is easily verified that they form a σ-algebra, called the *σ-algebra generated by* ξ, and denoted by $\mathscr{F}_\xi$.

If φ is a Borel function, it follows from Lemma 2 that the function $\eta = \varphi \circ \xi$ is also a random variable, and in fact $\mathscr{F}_\xi$-measurable, i.e. such that

$$\{\omega: \eta(\omega) \in B\} \in \mathscr{F}_\xi, \quad B \in \mathscr{B}(R)$$

(see (7)). It turns out that the converse is also true.

Theorem 3. *Let η be a $\mathscr{F}_\xi$-measurable random variable. Then there is a Borel function φ such that $\eta = \varphi \circ \xi$, i.e. $\eta(\omega) = \varphi(\xi(\omega))$ for every $\omega \in \Omega$.*

PROOF. Let Φ^ξ be the class of $\mathscr{F}_\xi$-measurable functions $\eta = \eta(\omega)$ and $\tilde{\Phi}_\xi$ the class of $\mathscr{F}_\xi$-measurable functions representable in the form $\varphi \circ \xi$, where φ is a Borel function. It is clear that $\tilde{\Phi}_\xi \subseteq \Phi_\xi$. The conclusion of the theorem is that in fact $\tilde{\Phi}_\xi = \Phi_\xi$.

Let $A \in \mathscr{F}_\xi$ and $\eta(\omega) = I_A(\omega)$. Let us show that $\eta \in \tilde{\Phi}_\xi$. In fact, if $A \in \mathscr{F}_\xi$ there is a $B \in \mathscr{B}(R)$ such that $A = \{\omega: \xi(\omega) \in B\}$. Let

$$\chi_B(x) = \begin{cases} 1, & x \in B, \\ 0, & x \notin B. \end{cases}$$

Then $I_A(\omega) = \chi_B(\xi(\omega)) \in \tilde{\Phi}_\xi$. Hence it follows that every simple $\mathscr{F}_\xi$-measurable function $\sum_{i=1}^n c_i I_{A_i}(\omega)$, $A_i \in \mathscr{F}_\xi$, also belongs to $\tilde{\Phi}_\xi$.

† We shall assume the usual conventions about arithmetic operations in $\bar{R}$: if $a \in R$ then $a \pm \infty = \pm\infty$, $a/\pm\infty = 0$; $a \cdot \infty = \infty$ if $a > 0$, and $a \cdot \infty = -\infty$ if $a < 0$; $0 \cdot (\pm\infty) = 0$, $\infty + \infty = \infty$, $-\infty - \infty = -\infty$.

Now let η be an arbitrary $\mathscr{F}_\xi$-measurable function. By Theorem 1 there is a sequence of simple $\mathscr{F}_\xi$-measurable functions $\{\eta_n\}$ such that $\eta_n(\omega) \to \eta(\omega)$, $n \to \infty$, $\omega \in \Omega$. As we just showed, there are Borel functions $\varphi_n = \varphi_n(x)$ such that $\eta_n(\omega) = \varphi_n(\xi(\omega))$. Then $\varphi_n(\xi(\omega)) \to \eta(\omega)$, $n \to \infty$, $\omega \in \Omega$.

Let B denote the set $\{x \in R: \lim_n \varphi_n(x) \text{ exists}\}$. This is a Borel set. Therefore

$$\varphi(x) = \begin{cases} \lim\limits_n \varphi_n(x), & x \in B, \\ 0, & x \notin B \end{cases}$$

is also a Borel function (see Problem 7).

But then it is evident that $\eta(\omega) = \lim_n \varphi_n(\xi(\omega)) = \varphi(\xi(\omega))$ for all $\omega \in \Omega$. Consequently $\bar{\Phi}_\xi = \Phi_\xi$.

6. Consider a measurable space $(\Omega, \mathscr{F})$ and a *finite* or *countably* infinite decomposition $\mathscr{D} = \{D_1, D_2, \ldots\}$ of the space Ω: namely, $D_i \in \mathscr{F}$ and $\sum_i D_i = \Omega$. We form the algebra $\mathscr{A}$ containing the empty set $\varnothing$ and the sets of the form $\sum_\alpha D_\alpha$, where the sum is taken in the finite or countably infinite sense. It is evident that the system $\mathscr{A}$ is a monotonic class, and therefore, according to Lemma 2, §2, chap. II, the algebra $\mathscr{A}$ is at the same time a σ-algebra, denoted $\sigma(\mathscr{D})$ and called the σ-algebra generated by the decomposition $\mathscr{D}$. Clearly $\sigma(\mathscr{D}) \subseteq \mathscr{F}$.

Lemma 3. *Let $\xi = \xi(\omega)$ be a $\sigma(\mathscr{D})$-measurable random variable. Then ξ is representable in the form*

$$\xi(\omega) = \sum_{k=1}^{\infty} x_k I_{D_k}(\omega), \tag{8}$$

where $x_k \in R$, $k \geq 1$, i.e., $\xi(\omega)$ is constant on the elements D_k of the decomposition, $k \geq 1$.

Proof. Let us choose a set D_k and show that the $\sigma(\mathscr{D})$-measurable function ξ has a constant value on that set.

For this purpose, we write

$$x_k = \sup[c: D_k \cap \{\omega: \xi(\omega) < c\} = \varnothing].$$

Since $\{\omega: \xi(\omega) < x_k\} = \bigcup_{\substack{r < x_k \\ r \text{ rational}}} \{\omega: \xi(\omega) < r\}$, we have $D_k \cap \{\omega: \xi(\omega) < x_k\} = \varnothing$.

Now let $c > x_k$. Then $D_k \cap \{\omega: \xi(\omega) < c\} \neq \varnothing$, and since the set $\{\omega: \xi(\omega) < c\}$ has the form $\sum_\alpha D_\alpha$, where the sum is over a finite or countable collection of indices, we have

$$D_k \cap \{\omega: \xi(\omega) < c\} = D_k.$$

Hence, it follows that, for all $c > x_k$,

$$D_k \cap \{\omega: \xi(\omega) \geq c\} = \varnothing,$$

and since $\{\omega: \xi(\omega) > x_k\} = \bigcup_{\substack{r > x_k \\ r \text{ rational}}} \{\omega: \xi(\omega) \geq r\}$, we have

$$D_k \cap \{\omega: \xi(\omega) > x_k\} = \varnothing.$$

Consequently, $D_k \cap \{\omega: \xi(\omega) \neq x_k\} = \varnothing$, and therefore

$$D_k \subseteq \{\omega: \xi(\omega) = x_k\}$$

as required.

7. Problems

1. Show that the random variable ξ is continuous if and only if $\mathsf{P}(\xi = x) = 0$ for all $x \in R$.

2. If $|\xi|$ is $\mathscr{F}$-measurable, is it true that ξ is also $\mathscr{F}$-measurable?

3. Show that $\xi = \xi(\omega)$ is an extended random variable if and only if $\{\omega: \xi(\omega) \in \bar{B}\} \in \mathscr{F}$ for all $\bar{B} \in \mathscr{B}(\bar{R})$.

4. Prove that x^n, $x^+ = \max(x, 0)$, $x^- = -\min(x, 0)$, and $|x| = x^+ + x^-$ are Borel functions.

5. If ξ and η are $\mathscr{F}$-measurable, then $\{\omega: \xi(\omega) = \eta(\omega)\} \in \mathscr{F}$.

6. Let ξ and η be random variables on $(\Omega, \mathscr{F})$, and $A \in \mathscr{F}$. Then the function

$$\zeta(\omega) = \xi(\omega) \cdot I_A + \eta(\omega) I_{\bar{A}}$$

is also a random variable.

7. Let $\xi_1, \ldots, \xi_n$ be random variables and $\varphi(x_1, \ldots, x_n)$ a Borel function. Show that $\varphi(\xi_1(\omega), \ldots, \xi_n(\omega))$ is also a random variable.

8. Let ξ and η be random variables, both taking the values $1, 2, \ldots, N$. Suppose that $\mathscr{F}_\xi = \mathscr{F}$. Show that there is a permutation $(i_1, i_2, \ldots, i_N)$ of $(1, 2, \ldots, N)$ such that $\{\omega: \xi = j\} = \{\omega: \eta = i_j\}$ for $j = 1, 2, \ldots, N$.

§5. Random Elements

1. In addition to random variables, probability theory and its applications involve random objects of more general kinds, for example random points, vectors, functions, processes, fields, sets, measures, etc. In this connection it is desirable to have the concept of a random object of any kind.

Definition 1. Let $(\Omega, \mathscr{F})$ and $(E, \mathscr{E})$ be measurable spaces. We say that a function $X = X(\omega)$, defined on Ω and taking values in E, is *$\mathscr{F}/\mathscr{E}$-measurable*, or is a *random element* (with values in E), if

$$\{\omega: X(\omega) \in B\} \in \mathscr{F} \tag{1}$$

for every $B \in \mathscr{E}$. Random elements (with values in E) are sometimes called E-valued random variables.

Let us consider some special cases.

If $(E, \mathscr{E}) = (R, \mathscr{B}(R))$, the definition of a random element is the same as the definition of a random variable (§4).

Let $(E, \mathscr{E}) = (R^n, \mathscr{B}(R^n))$. Then a random element $X(\omega)$ is a "random point" in R^n. If π_k is the projection of R^n on the kth coordinate axis, $X(\omega)$ can be represented in the form

$$X(\omega) = (\xi_1(\omega), \dots, \xi_n(\omega)), \tag{2}$$

where $\xi_k = \pi_k \circ X$.

It follows from (1) that ξ_k is an ordinary random variable. In fact, for $B \in \mathscr{B}(R)$ we have

$$\begin{aligned}\{\omega: \xi_k(\omega) \in B\} &= \{\omega: \xi_1(\omega) \in R, \dots, \xi_{k-1} \in R, \xi_k \in B, \xi_{k+1} \in R, \dots\} \\ &= \{\omega: X(\omega) \in (R \times \cdots \times R \times B \times R \times \cdots \times R)\} \in \mathscr{F},\end{aligned}$$

since $R \times \cdots \times R \times B \times R \times \cdots \times R \in \mathscr{B}(R^n)$.

Definition 2. An ordered set $(\eta_1(\omega), \dots, \eta_n(\omega))$ of random variables is called an *n-dimensional random vector.*

According to this definition, every random element $X(\omega)$ with values in R^n is an n-dimensional random vector. The converse is also true: every random vector $X(\omega) = (\xi_1(\omega), \dots, \xi_n(\omega))$ is a random element in R^n. In fact, if $B_k \in \mathscr{B}(R)$, $k = 1, \dots, n$, then

$$\{\omega: X(\omega) \in (B_1 \times \cdots \times B_n)\} = \prod_{k=1}^{n} \{\omega: \xi_k(\omega) \in B_k\} \in \mathscr{F}.$$

But $\mathscr{B}(R^n)$ is the smallest σ-algebra containing the sets $B_1 \times \cdots \times B_n$. Consequently we find immediately, by an evident generalization of Lemma 1 of §4, that whenever $B \in \mathscr{B}(R^n)$, the set $\{\omega: X(\omega) \in B\}$ belongs to $\mathscr{F}$.

Let $(E, \mathscr{E}) = (Z, B(Z))$, where Z is the set of complex numbers $x + iy$, $x, y \in R$, and $B(Z)$ is the smallest σ-algebra containing the sets $\{z: z = x + iy, a_1 < x \le b_1, a_2 < y \le b_2\}$. It follows from the discussion above that a complex-valued random variable $Z(\omega)$ can be represented as $Z(\omega) = X(\omega) + iY(\omega)$, where $X(\omega)$ and $Y(\omega)$ are random variables. Hence we may also call $Z(\omega)$ a *complex random variable.*

Let $(E, \mathscr{E}) = (R^T, \mathscr{B}(R^T))$, where T is a subset of the real line. In this case every random element $X = X(\omega)$ can evidently be represented as $X = (\xi_t)_{t \in T}$ with $\xi_t = \pi_t \circ X$, and is called a random function with time domain T.

Definition 3. Let T be a subset of the real line. A set of random variables $X = (\xi_t)_{t \in T}$ is called a *random process*† *with time domain T.*

† Or stochastic process (Translator).

If $T = \{1, 2, \ldots\}$ we call $X = (\xi_1, \xi_2, \ldots)$ a *random process with discrete time*, or a *random sequence*.

If $T = [0, 1]$, $(-\infty, \infty)$, $[0, \infty), \ldots$, we call $X = (\xi_t)_{t \in T}$ a *random process with continuous time*.

It is easy to show, by using the structure of the σ-algebra $\mathscr{B}(R^T)$ (§2) that every random process $X = (\xi_t)_{t \in T}$ (in the sense of Definition 3) is also a random function on the space $(R^T, \mathscr{B}(R^T))$.

Definition 4. Let $X = (\xi_t)_{t \in T}$ be a random process. For each given $\omega \in \Omega$ the function $(\xi_t(\omega))_{t \in T}$ is said to be a *realization* or a *trajectory* of the process, corresponding to the outcome ω.

The following definition is a natural generalization of Definition 2 of §4.

Definition 5. Let $X = (\xi_t)_{t \in T}$ be a random process. The probability measure P_X on $(R^T, \mathscr{B}(R^T))$ defined by

$$P_X(B) = \mathsf{P}\{\omega: X(\omega) \in B\}, \qquad B \in \mathscr{B}(R^T),$$

is called the *probability distribution of* X. The probabilities

$$P_{t_1, \ldots, t_n}(B) \equiv \mathsf{P}\{\omega: (\xi_{t_1}, \ldots, \xi_{t_n}) \in B\}$$

with $t_1 < t_2 < \cdots < t_n$, $t_i \in T$, are called finite-dimensional probabilities (or probability distributions). The functions

$$F_{t_1, \ldots, t_n}(x_1, \ldots, x_n) \equiv \mathsf{P}\{\omega: \xi_{t_1} \leq x_1, \ldots, \xi_{t_n} \leq x_n\}$$

with $t_1 < t_2 < \cdots < t_n$, $t_i \in T$, are called finite-dimensional distribution functions.

Let $(E, \mathscr{E}) = (C, \mathscr{B}_0(C))$, where C is the space of continuous functions $x = (x_t)_{t \in T}$ on $T = [0, 1]$ and $\mathscr{B}_0(C)$ is the σ-algebra generated by the open sets (§2). We show that every random element X on $(C, \mathscr{B}_0(C))$ is also a random process with continuous trajectories in the sense of Definition 3.

In fact, according to §2 the set $A = \{x \in C: x_t < a\}$ is open in $\mathscr{B}_0(C)$. Therefore

$$\{\omega: \xi_t(\omega) < a\} = \{\omega: X(\omega) \in A\} \in \mathscr{F}.$$

On the other hand, let $X = (\xi_t(\omega))_{t \in T}$ be a random process (in the sense of Definition 3) whose trajectories are continuous functions for every $\omega \in \Omega$. According to (2.14),

$$\{x \in C: x \in S_\rho(x^0)\} = \bigcap_{t_k} \{x \in C: |x_{t_k} - x^0_{t_k}| < \rho\},$$

where t_k are the rational points of $[0, 1]$. Therefore

$$\{\omega: X(\omega) \in S_\rho(X^0\omega))\} = \bigcap_{t_k} \{\omega: |\xi_{t_k}(\omega) - \xi^0_{t_k}(\omega)| < \rho\} \in \mathscr{F},$$

and therefore we also have $\{\omega: X(\omega) \in B\} \in \mathscr{F}$ for every $B \in \mathscr{B}_0(C)$.

Similar reasoning will show that every random element of the space $(D, \mathscr{B}_0(D)$ can be considered as a random process with trajectories in the space of functions with no discontinuities of the second kind; and conversely.

2. Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space and $(E_\alpha, \mathscr{E}_\alpha)$ measurable spaces, where α belongs to an (arbitrary) set $\mathfrak{A}$.

Definition 6. We say that the $\mathscr{F}/\mathscr{E}_\alpha$-measurable functions $(X_\alpha(\omega))$, $\alpha \in \mathfrak{A}$, are independent (or collectively independent) if, for every finite set of indices $\alpha_1, \ldots, \alpha_n$ the random elements $X_{\alpha_1}, \ldots, X_{\alpha_n}$ are independent, i.e.

$$\mathsf{P}(X_{\alpha_1} \in B_{\alpha_1}, \ldots, X_{\alpha_n} \in B_{\alpha_n}) = \mathsf{P}(X_{\alpha_1} \in B_{\alpha_1}) \cdots \mathsf{P}(X_{\alpha_n} \in B_{\alpha_n}), \tag{3}$$

where $B_\alpha \in \mathscr{E}_\alpha$.

Let $\mathfrak{A} = \{1, 2, \ldots, n\}$, let ξ_α be random variables, let $\alpha \in \mathfrak{A}$ and let

$$F_\xi(x_1, \ldots, x_n) = \mathsf{P}(\xi_1 \leq x_1, \ldots, \xi_n \leq x_n)$$

be the n-dimensional distribution function of the random vector $\xi = (\xi_1, \ldots, \xi_n)$. Let $F_{\xi_i}(x_i)$ be the distribution functions of the random variables ξ_i, $i = 1, \ldots, n$.

Theorem. *A necessary and sufficient condition for the random variables $\xi_1, \ldots, \xi_n$ to be independent is that*

$$F_\xi(x_1, \ldots, x_n) = F_{\xi_1}(x_1) \cdots F_{\xi_n}(x_n) \tag{4}$$

for all $(x_1, \ldots, x_n) \in R^n$.

Proof. The necessity is evident. To prove the sufficiency we put $a = (a_1, \ldots, a_n)$, $b = (b_1, \ldots, b_n)$,

$$P_\xi(a, b] = \mathsf{P}\{\omega: a_1 < \xi_1 \leq b_1, \ldots, a_n < \xi_n \leq b_n\},$$
$$P_{\xi_i}(a_i, b_i] = \mathsf{P}\{a_i < \xi_i \leq b_i\}.$$

Then

$$P_\xi(a, b] = \prod_{i=1}^n [F_{\xi_i}(b_i) - F_{\xi_i}(a_i)] = \prod_{i=1}^n P_{\xi_i}(a_i, b_i]$$

by (4) and (3.7), and therefore

$$\mathsf{P}\{\xi_1 \in I_1, \ldots, \xi_n \in I_n\} = \prod_{i=1}^n \mathsf{P}\{\xi_i \in I_i\}, \tag{5}$$

where $I_i = (a_i, b_i]$.

We fix $I_2, \ldots, I_n$ and show that

$$\mathsf{P}\{\xi_1 \in B_1, \xi_2 \in I_2, \ldots, \xi_n \in I_n\} = \mathsf{P}\{\xi_1 \in B_1\} \prod_{i=2}^n \mathsf{P}\{\xi_i \in I_i\} \tag{6}$$

for all $B_1 \in \mathscr{B}(R)$. Let $\mathscr{M}$ be the collection of sets in $\mathscr{B}(R)$ for which (6) holds. Then $\mathscr{M}$ evidently contains the algebra $\mathscr{A}$ of sets consisting of sums of disjoint intervals of the form $I_1 = (a_1, b_1]$. Hence $\mathscr{A} \subseteq \mathscr{M} \subseteq \mathscr{B}(R)$. From the countable additivity (and therefore continuity) of probability measures it also follows that $\mathscr{M}$ is a monotonic class. Therefore (see Subsection 1 of §2)

$$\mu(\mathscr{A}) \subseteq \mathscr{M} \subseteq \mathscr{B}(R).$$

But $\mu(\mathscr{A}) = \sigma(\mathscr{A}) = \mathscr{B}(R)$ by Theorem 1 of §2. Therefore $\mathscr{M} = \mathscr{B}(R)$.

Thus (6) is established. Now fix $B_1, I_3, \ldots, I_n$; by the same method we can establish (6) with I_2 replaced by the Borel set B_2. Continuing in this way, we can evidently arrive at the required equation,

$$\mathsf{P}(\xi_1 \in B_1, \ldots, \xi_n \in B_n) = \mathsf{P}(\xi_1 \in B_1) \cdots \mathsf{P}(\xi_n \in B_n),$$

where $B_i \in \mathscr{B}(R)$. This completes the proof of the theorem.

3. Problems

1. Let $\xi_1, \ldots, \xi_n$ be discrete random variables. Show that they are independent if and only if

$$\mathsf{P}(\xi_1 = x_1, \ldots, \xi_n = x_n) = \prod_{i=1}^{n} \mathsf{P}(\xi_i = x_i)$$

for all real $x_1, \ldots, x_n$.

2. Carry out the proof that every random function (in the sense of Definition 1) is a random process (in the sense of Definition 3) and conversely.

3. Let $X_1, \ldots, X_n$ be random elements with values in $(E_1, \mathscr{E}_1), \ldots, (E_n, \mathscr{E}_n)$, respectively. In addition let $(E'_1, \mathscr{E}'_1), \ldots, (E'_n, \mathscr{E}'_n)$ be measurable spaces and let $g_1, \ldots, g_n$ be $\mathscr{E}_1/\mathscr{E}'_1, \ldots, \mathscr{E}_n/\mathscr{E}'_n$-measurable functions, respectively. Show that if $X_1, \ldots, X_n$ are independent, the random elements $g_1 \cdot X_1, \ldots, g_n \cdot X_n$ are also independent.

§6. Lebesgue Integral. Expectation

1. When $(\Omega, \mathscr{F}, \mathsf{P})$ is a finite probability space and $\xi = \xi(\omega)$ is a simple random variable,

$$\xi(\omega) = \sum_{k=1}^{n} x_k I_{A_k}(\omega), \tag{1}$$

the expectation $\mathsf{E}\xi$ was defined in §4 of Chapter I. The same definition of the expectation $\mathsf{E}\xi$ of a simple random variable ξ can be used for any probability space $(\Omega, \mathscr{F}, \mathsf{P})$. That is, we define

$$\mathsf{E}\xi = \sum_{k=1}^{n} x_k \mathsf{P}(A_k). \tag{2}$$

This definition is consistent (in the sense that $\mathsf{E}\xi$ is independent of the particular representation of ξ in the form (1)), as can be shown just as for finite probability spaces. The simplest properties of the expectation can be established similarly (see Subsection 5 of §4 of Chapter I).

In the present section we shall define and study the properties of the expectation $\mathsf{E}\xi$ of an arbitrary random variable. In the language of analysis, $\mathsf{E}\xi$ is merely the Lebesgue integral of the $\mathscr{F}$-measurable function $\xi = \xi(\omega)$ with respect to the measure P. In addition to $\mathsf{E}\xi$ we shall use the notation $\int_\Omega \xi(\omega)\mathsf{P}(d\omega)$ or $\int_\Omega \xi\, d\mathsf{P}$.

Let $\xi = \xi(\omega)$ be a nonnegative random variable. We construct a sequence of simple nonnegative random variables $\{\xi_n\}_{n\geq 1}$ such that $\xi_n(\omega) \uparrow \xi(\omega)$, $n \to \infty$, for each $\omega \in \Omega$ (see Theorem 1 in §4).

Since $\mathsf{E}\xi_n \leq \mathsf{E}\xi_{n+1}$ (cf. Property 3) of Subsection 5, §4, Chapter I), the limit $\lim_n \mathsf{E}\xi_n$ exists, possibly with the value $+\infty$.

Definition 1. The *Lebesgue integral* of the nonnegative random variable $\xi = \xi(\omega)$, or its *expectation*, is

$$\mathsf{E}\xi \equiv \lim_n \mathsf{E}\xi_n. \tag{3}$$

To see that this definition is consistent, we need to show that the limit is independent of the choice of the approximating sequence $\{\xi_n\}$. In other words, we need to show that if $\xi_n \uparrow \xi$ and $\eta_m \uparrow \xi$, where $\{\eta_m\}$ is a sequence of simple functions, then

$$\lim_n \mathsf{E}\xi_n = \lim_m \mathsf{E}\eta_m. \tag{4}$$

Lemma 1. *Let η and ξ_n be simple random variables, $n \geq 1$, and*

$$\xi_n \uparrow \xi \geq \eta.$$

Then

$$\lim_n \mathsf{E}\xi_n \geq \mathsf{E}\eta. \tag{5}$$

PROOF. Let $\varepsilon > 0$ and

$$A_n = \{\omega\colon \xi_n \geq \eta - \varepsilon\}.$$

It is clear that $A_n \uparrow \Omega$ and

$$\xi_n = \xi_n I_{A_n} + \xi_n I_{\bar{A}_n} \geq \xi_n I_{A_n} \geq (\eta - \varepsilon) I_{A_n}.$$

Hence by the properties of the expectations of simple random variables we find that

$$\begin{aligned}\mathsf{E}\xi_n \geq \mathsf{E}(\eta - \varepsilon)I_{A_n} &= \mathsf{E}\eta I_{A_n} - \varepsilon\mathsf{P}(A_n) \\ &= \mathsf{E}\eta - \mathsf{E}\eta I_{\bar{A}_n} - \varepsilon\mathsf{P}(A_n) \geq \mathsf{E}\eta - C\mathsf{P}(\bar{A}_n) - \varepsilon,\end{aligned}$$

where $C = \max_\omega \eta(\omega)$. Since ε is arbitrary, the required inequality (5) follows. It follows from this lemma that $\lim_n \mathsf{E}\xi_n \geq \lim_m \mathsf{E}\eta_m$ and by symmetry $\lim_m \mathsf{E}\eta_m \geq \lim_n \mathsf{E}\xi_n$, which proves (4).

The following remark is often useful.

Remark 1. The expectation $\mathsf{E}\xi$ of the nonnegative random variable ξ satisfies

$$\mathsf{E}\xi = \sup_{\{s \in S: s \leq \xi\}} \mathsf{E}s, \tag{6}$$

where $S = \{s\}$ is a set of simple random variables (Problem 1).

Thus the expectation is well defined for nonnegative random variables. We now consider the general case.

Let ξ be a random variable and $\xi^+ = \max(\xi, 0)$, $\xi^- = -\min(\xi, 0)$.

Definition 2. We say that the expectation $\mathsf{E}\xi$ of the random variable ξ *exists*, or *is defined*, if at least one of $\mathsf{E}\xi^+$ and $\mathsf{E}\xi^-$ is finite:

$$\min(\mathsf{E}\xi^+, \mathsf{E}\xi^-) < \infty.$$

In this case we *define*

$$\mathsf{E}\xi \equiv \mathsf{E}\xi^+ - \mathsf{E}\xi^-.$$

The *expectation* $\mathsf{E}\xi$ is also called the *Lebesgue integral* (of the function ξ with respect to the probability measure P).

Definition 3. We say that the *expectation of ξ is finite* if $\mathsf{E}\xi^+ < \infty$ and $\mathsf{E}\xi^- < \infty$.

Since $|\xi| = \xi^+ - \xi^-$, the finiteness of $\mathsf{E}\xi$, or $|\mathsf{E}\xi| < \infty$, is equivalent to $\mathsf{E}|\xi| < \infty$. (In this sense one says that the Lebesgue integral is absolutely convergent.)

Remark 2. In addition to the expectation $\mathsf{E}\xi$, significant numerical characteristics of a random variable ξ are the number $\mathsf{E}\xi^r$ (if defined) and $\mathsf{E}|\xi|^r$, $r > 0$, which are known as the *moment* of order r (or rth moment) and the *absolute moment* of order r (or absolute rth moment) of ξ.

Remark 3. In the definition of the Lebesgue integral $\int_\Omega \xi(\omega)\mathsf{P}(d\omega)$ given above, we suppose that P was a probability measure ($\mathsf{P}(\Omega) = 1$) and that the $\mathscr{F}$-measurable functions (random variables) ξ had values in $R = (-\infty, \infty)$. Suppose now that μ is any measure defined on a measurable space $(\Omega, \mathscr{F})$ and possibly taking the value $+\infty$, and that $\xi = \xi(\omega)$ is an $\mathscr{F}$-measurable function with values in $\bar{R} = [-\infty, \infty]$ (an extended random variable). In this case the Lebesgue integral $\int_\Omega \xi(\omega)\mu(d\omega)$ is defined in the same way: first, for nonnegative simple ξ (by (2) with P replaced by μ), then for arbitrary nonnegative ξ, and in general by the formula

$$\int_\Omega \xi(\omega)\mu(d\omega) = \int_\Omega \xi^+ \mu(d\omega) - \int_\Omega \xi^- \mu(d\omega),$$

provided that no indeterminacy of the form $\infty - \infty$ arises.

A case that is particularly important for mathematical analysis is that in which $(\Omega, F) = (R, \mathscr{B}(R))$ and μ is Lebesgue measure. In this case the integral $\int_R \xi(x)\mu(dx)$ is written $\int_R \xi(x)\,dx$, or $\int_{-\infty}^{\infty} \xi(x)\,dx$, or (L) $\int_{-\infty}^{\infty} \xi(x)\,dx$ to emphasize its difference from the Riemann integral (R) $\int_{-\infty}^{\infty} \xi(x)\,dx$. If the measure μ (Lebesgue–Stieltjes) corresponds to a generalized distribution function $G = G(x)$, the integral $\int_R \xi(x)\mu(dx)$ is also called a *Lebesgue–Stieltjes integral* and is denoted by (L–S) $\int_R \xi(x)G(dx)$, a notation that distinguishes it from the corresponding Riemann–Stieltjes integral

$$(\text{R–S}) \int_R \xi(x)G(dx)$$

(see Subsection 10 below).

It will be clear from what follows (Property **D**) that if $\mathsf{E}\xi$ is defined then so is the expectation $\mathsf{E}(\xi I_A)$ for every $A \in \mathscr{F}$. The notations $\mathsf{E}(\xi; A)$ or $\int_A \xi\,dP$ are often used for $\mathsf{E}(\xi I_A)$ or its equivalent, $\int_\Omega \xi I_A\,d\mathsf{P}$. The integral $\int_A \xi\,d\mathsf{P}$ is called the *Lebesgue integral of ξ with respect to P over the set A*.

Similarly, we write $\int_A \xi\,d\mu$ instead of $\int_\Omega \xi \cdot I_A\,d\mu$ for an arbitrary measure μ. In particular, if μ is an n-dimensional Lebesgue–Stieltjes measure, and $A = (a_1, b_1] \times \cdots \times (a_n, b_n]$, we use the notation

$$\int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} \xi(x_1, \ldots, x_n)\mu(dx_1 \cdots dx_n) \qquad \text{instead of} \qquad \int_A \xi\,d\mu.$$

If μ is Lebesgue measure, we write simply $dx_1 \cdots dx_n$ instead of $\mu(dx_1, \ldots, dx_n)$.

2. Properties of the expectation $\mathsf{E}\xi$ of the random variable ξ.

A. *Let c be a constant and let $\mathsf{E}\xi$ exist. Then $\mathsf{E}(c\xi)$ exists and*

$$\mathsf{E}(c\xi) = c\mathsf{E}\xi.$$

B. *Let $\xi \le \eta$; then*

$$\mathsf{E}\xi \le \mathsf{E}\eta$$

with the understanding that

$$\text{if} \quad -\infty < \mathsf{E}\xi \quad \text{then} \quad -\infty < \mathsf{E}\eta \quad \text{and} \quad \mathsf{E}\xi \le \mathsf{E}\eta$$

or

$$\text{if} \quad \mathsf{E}\eta < \infty \quad \text{then} \quad \mathsf{E}\xi < \infty \quad \text{and} \quad \mathsf{E}\xi \le \mathsf{E}\eta.$$

C. *If* $\mathsf{E}\xi$ *exists then*

$$|\mathsf{E}\xi| \le \mathsf{E}|\xi|.$$

D. *If* $\mathsf{E}\xi$ *exists then* $\mathsf{E}(\xi I_A)$ *exists for each* $A \in \mathscr{F}$; *if* $\mathsf{E}\xi$ *is finite,* $\mathsf{E}(\xi I_A)$ *is finite.*

E. *If* ξ *and* η *are nonnegative random variables, or such that* $\mathsf{E}|\xi| < \infty$ *and* $\mathsf{E}|\eta| < \infty$, *then*

$$\mathsf{E}(\xi + \eta) = \mathsf{E}\xi + \mathsf{E}\eta.$$

(See Problem 2 for a generalization.)

Let us establish **A–E**.

A. This is obvious for simple random variables. Let $\xi \ge 0$, $\xi_n \uparrow \xi$, where ξ_n are simple random variables and $c \ge 0$. Then $c\xi_n \uparrow c\xi$ and therefore

$$\mathsf{E}(c\xi) = \lim \mathsf{E}(c\xi_n) = c \lim \mathsf{E}\xi_n = c\mathsf{E}\xi.$$

In the general case we need to use the representation $\xi = \xi^+ - \xi^-$ and notice that $(c\xi)^+ = c\xi^+$, $(c\xi)^- = c\xi^-$ when $c \ge 0$, whereas when $c < 0$, $(c\xi)^+ = -c\xi^-$, $(c\xi)^- = -c\xi^+$.

B. If $0 \le \xi \le \eta$, then $\mathsf{E}\xi$ and $\mathsf{E}\eta$ are defined and the inequality $\mathsf{E}\xi \le \mathsf{E}\eta$ follows directly from (6). Now let $\mathsf{E}\xi > -\infty$; then $\mathsf{E}\xi^- < \infty$. If $\xi \le \eta$, we have $\xi^+ \le \eta^+$ and $\xi^- \ge \eta^-$. Therefore $\mathsf{E}\eta^- \le \mathsf{E}\xi^- < \infty$; consequently $\mathsf{E}\eta$ is defined and $\mathsf{E}\xi = \mathsf{E}\xi^+ - \mathsf{E}\xi^- \le \mathsf{E}\eta^+ - \mathsf{E}\eta^- = \mathsf{E}\eta$. The case when $\mathsf{E}\eta < \infty$ can be discussed similarly.

C. Since $-|\xi| \le \xi \le |\xi|$, Properties **A** and **B** imply

$$-\mathsf{E}|\xi| \le \mathsf{E}\xi \le \mathsf{E}|\xi|,$$

i.e. $|\mathsf{E}\xi| \le \mathsf{E}|\xi|$.

D. This follows from **B** and

$$(\xi I_A)^+ = \xi^+ I_A \le \xi^+, \qquad (\xi I_A)^- = \xi^- I_A \le \xi^-.$$

E. Let $\xi \ge 0$, $\eta \ge 0$, and let $\{\xi_n\}$ and $\}\eta_n\}$ be sequences of simple functions such that $\xi_n \uparrow \xi$ and $\eta_n \uparrow \eta$. Then $\mathsf{E}(\xi_n + \eta_n) = \mathsf{E}\xi_n + \mathsf{E}\eta_n$ and

$$\mathsf{E}(\xi_n + \eta_n) \uparrow \mathsf{E}(\xi + \eta), \quad \mathsf{E}\xi_n \uparrow \mathsf{E}\xi, \quad \mathsf{E}\eta_n \uparrow \mathsf{E}\eta$$

and therefore $\mathsf{E}(\xi + \eta) = \mathsf{E}\xi + \mathsf{E}\eta$. The case when $\mathsf{E}|\xi| < \infty$ and $\mathsf{E}|\eta| < \infty$ reduces to this if we use the facts that

$$\xi = \xi^+ - \xi^-, \quad \eta = \eta^+ - \eta^-, \quad \xi^+ \le |\xi|, \quad \xi^- \le |\xi|,$$

and

$$\eta^+ \le |\eta|, \qquad \eta^- \le |\eta|.$$

The following group of statements about expectations involve the notion of "P-almost surely." We say that a property holds "P-*almost surely*" *if there is a set* $\mathcal{N} \in \mathcal{F}$ *with* $\mathsf{P}(\mathcal{N}) = 0$ *such that the property holds for every point* ω *of* $\Omega \setminus \mathcal{N}$. Instead of "P-almost surely" we often say "P-almost everywhere" or simply "almost surely" (a.s.) or "almost everywhere" (a.e.).

F. If $\xi = 0$ (a.s.) then $\mathsf{E}\xi = 0$.

In fact, if ξ is a simple random variable, $\xi = \sum x_k I_{A_k}(\omega)$ and $x_k \neq 0$, we have $\mathsf{P}(A_k) = 0$ by hypothesis and therefore $\mathsf{E}\xi = 0$. If $\xi \geq 0$ and $0 \leq s \leq \xi$, where s is a simple random variable, then $s = 0$ (a.s.) and consequently $\mathsf{E}s = 0$ and $\mathsf{E}\xi = \sup_{\{s \in S: s \leq \xi\}} \mathsf{E}s = 0$. The general case follows from this by means of the representation $\xi = \xi^+ - \xi^-$ and the facts that $\xi^+ \leq |\xi|$, $\xi^- \leq |\xi|$, and $|\xi| = 0$ (a.s.).

G. If $\xi = \eta$ (a.s.) and $\mathsf{E}|\xi| < \infty$, then $\mathsf{E}|\eta| < \infty$ and $\mathsf{E}\xi = \mathsf{E}\eta$ (see also Problem 3).

In fact, let $\mathcal{N} = \{\omega: \xi \neq \eta\}$. Then $\mathsf{P}(\mathcal{N}) = 0$ and $\xi = \xi I_{\mathcal{N}} + \xi I_{\bar{\mathcal{N}}}$, $\eta = \eta I_{\mathcal{N}} + \eta I_{\bar{\mathcal{N}}} = \eta I_{\mathcal{N}} + \xi I_{\bar{\mathcal{N}}}$. By properties E and F, we have $\mathsf{E}\xi = \mathsf{E}\xi I_{\mathcal{N}} + \mathsf{E}\xi I_{\bar{\mathcal{N}}} = \mathsf{E}\eta I_{\bar{\mathcal{N}}}$. But $\mathsf{E}\eta I_{\mathcal{N}} = 0$, and therefore $\mathsf{E}\xi = \mathsf{E}\eta I_{\bar{\mathcal{N}}} + \mathsf{E}\eta I_{\mathcal{N}} = \mathsf{E}\eta$, by Property E.

H. Let $\xi \geq 0$ and $\mathsf{E}\xi = 0$. Then $\xi = 0$ (a.s).

For the proof, let $A = \{\omega: \xi(\omega) > 0\}$, $A_n = \{\omega: \xi(\omega) \geq 1/n\}$. It is clear that $A_n \uparrow A$ and $0 \leq \xi \cdot I_{A_n} \leq \xi \cdot I_A$. Hence, by Property **B**,

$$0 \leq \mathsf{E}\xi I_{A_n} \leq \mathsf{E}\xi = 0.$$

Consequently

$$0 = \mathsf{E}\xi I_{A_n} \geq \frac{1}{n}\mathsf{P}(A_n)$$

and therefore $\mathsf{P}(A_n) = 0$ for all $n \geq 1$. But $\mathsf{P}(A) = \lim \mathsf{P}(A_n)$ and therefore $\mathsf{P}(A) = 0$.

I. Let ξ and η be such that $\mathsf{E}|\xi| < \infty$, $\mathsf{E}|\eta| < \infty$ and $\mathsf{E}(\xi I_A) \leq \mathsf{E}(\eta I_A)$ for all $A \in \mathcal{F}$. Then $\xi \leq \eta$ (a.s.).

In fact, let $B = \{\omega: \xi(\omega) > \eta(\omega)\}$. Then $\mathsf{E}(\eta I_B) \leq \mathsf{E}(\xi I_B) \leq \mathsf{E}(\eta I_B)$ and therefore $\mathsf{E}(\xi I_B) = \mathsf{E}(\eta I_B)$. By Property E, we have $\mathsf{E}((\xi - \eta)I_B) = 0$ and by Property H we have $(\xi - \eta)I_B = 0$ (a.s.), whence $\mathsf{P}(B) = 0$.

J. Let ξ be an extended random variable and $\mathsf{E}|\xi| < \infty$. Then $|\xi| < \infty$ (a. s). In fact, let $A = \{\omega | \xi(\omega)| = \infty\}$ and $\mathsf{P}(A) > 0$. Then $\mathsf{E}|\xi| \geq \mathsf{E}(|\xi| I_A) = \infty \cdot \mathsf{P}(A) = \infty$, which contradicts the hypothesis $\mathsf{E}|\xi| < \infty$. (See also Problem 4.)

3. Here we consider the fundamental theorems on *taking limits* under the expectation sign (or the Lebesgue integral sign).

Theorem 1 (On Monotone Convergence). *Let η, ξ, ξ_1, $\xi_2, \ldots$ be random variables.*

(a) *If $\xi_n \geq \eta$ for all $n \geq 1$, $\mathsf{E}\eta > -\infty$, and $\xi_n \uparrow \xi$, then*

$$\mathsf{E}\xi_n \uparrow \mathsf{E}\xi.$$

(b) *If $\xi_n \leq \eta$ for all $n \geq 1$, $\mathsf{E}\eta < \infty$, and $\xi_n \downarrow \xi$, then*

$$\mathsf{E}\xi_n \downarrow \mathsf{E}\xi.$$

PROOF. (a) First suppose that $\eta \geq 0$. For each $k \geq 1$ let $\{\xi_k^{(n)}\}_{n\geq 1}$ be a sequence of simple functions such that $\xi_k^{(n)} \uparrow \xi_k$, $n \to \infty$. Put $\zeta^{(n)} = \max_{1\leq k\leq N} \xi_k^{(n)}$. Then

$$\zeta^{(n-1)} \leq \zeta^{(n)} = \max_{1\leq k\leq n} \xi_k^{(n)} \leq \max_{1\leq k\leq n} \xi_k = \xi_n.$$

Let $\zeta = \lim_n \zeta^{(n)}$. Since

$$\xi_k^{(n)} \leq \zeta^{(n)} \leq \xi_n$$

for $1 \leq k \leq n$, we find by taking limits as $n \to \infty$ that

$$\xi_k \leq \zeta \leq \xi$$

for every $k \geq 1$ and therefore $\xi = \zeta$.

The random variables $\zeta^{(n)}$ are simple and $\zeta^{(n)} \uparrow \zeta$. Therefore

$$\mathsf{E}\xi = \mathsf{E}\zeta = \lim \mathsf{E}\zeta^{(n)} \leq \lim \mathsf{E}\xi_n.$$

On the other hand, it is obvious, since $\xi_n \leq \xi_{n+1} \leq \xi$, that

$$\lim \mathsf{E}\xi_n \leq \mathsf{E}\xi.$$

Consequently $\lim \mathsf{E}\xi_n = \mathsf{E}\xi$.

Now let η be any random variable with $\mathsf{E}\eta > -\infty$.

If $\mathsf{E}\eta = \infty$ then $\mathsf{E}\xi_n = \mathsf{E}\xi = \infty$ by Property **B**, and our proposition is proved. Let $\mathsf{E}\eta < \infty$. Then instead of $\mathsf{E}\eta > -\infty$ we find $E|\eta| < \infty$. It is clear that $0 \leq \xi_n - \eta \uparrow \xi - \eta$ for all $\omega \in \Omega$. Therefore by what has been established, $\mathsf{E}(\xi_n - \eta) \uparrow \mathsf{E}(\xi - \eta)$ and therefore (by Property E and Problem 2)

$$\mathsf{E}\xi_n - \mathsf{E}_\eta \uparrow \mathsf{E}\xi - \mathsf{E}\eta.$$

But $\mathsf{E}|\eta| < \infty$, and therefore $\mathsf{E}\xi_n \uparrow \mathsf{E}_\xi$, $n \to \infty$.

The proof of (b) follows from (a) if we replace the original variables by their negatives.

Corollary. *Let $\{\eta_n\}_{n\geq 1}$ be a sequence of nonnegative random variables. Then*

$$\mathsf{E} \sum_{n=1}^{\infty} \eta_n = \sum_{n=1}^{\infty} \mathsf{E}\eta_n.$$

The proof follows from Property **E** (see also Problem 2), the monotone convergence theorem, and the remark that

$$\sum_{n=1}^{k} \eta_n \uparrow \sum_{n=1}^{\infty} \eta_n, \qquad k \to \infty.$$

Theorem 2 (Fatou's Lemma). *Let $\eta, \xi_1, \xi_2, \ldots$ be random variables.*

(a) *If $\xi_n \geq \eta$ for all $n \geq 1$ and $\mathsf{E}\eta > -\infty$, then*

$$\mathsf{E} \varliminf \xi_n \leq \varliminf \mathsf{E}\xi_n.$$

(b) *If $\xi_n \leq \eta$ for all $n \geq 1$ and $\mathsf{E}\eta < \infty$, then*

$$\varlimsup \mathsf{E}\xi_n \leq \mathsf{E} \varlimsup \xi_n.$$

(c) *If $|\xi_n| \leq \eta$ for all $n \geq 1$ and $\mathsf{E}\eta < \infty$, then*

$$\mathsf{E} \varliminf \xi_n \leq \varliminf \mathsf{E}\xi_n \leq \varlimsup \mathsf{E}\xi_n \leq \mathsf{E} \varlimsup \xi_n. \tag{7}$$

PROOF. (a) Let $\zeta_n = \inf_{m \geq n} \xi_m$; then

$$\varliminf \xi_n = \lim_n \inf_{m \geq n} \xi_m = \lim_n \zeta_n.$$

It is clear that $\zeta_n \uparrow \varliminf \xi_n$ and $\zeta_n \geq \eta$ for all $n \geq 1$. Then by Theorem 1

$$\mathsf{E} \varliminf \xi_n = \mathsf{E} \lim_n \zeta_n = \lim_n \mathsf{E}\zeta_n = \varliminf_n \mathsf{E}\zeta_n \leq \varliminf \mathsf{E}\xi_n,$$

which establishes (a). The second conclusion follows from the first. The third is a corollary of the first two.

Theorem 3 (Lebesgue's Theorem on Dominated Convergence). *Let $\eta, \xi, \xi_1, \xi_2, \ldots$ be random variables such that $|\xi_n| \leq \eta$, $\mathsf{E}\eta < \infty$ and $\xi_n \to \xi$ (a.s.). Then $\mathsf{E}|\xi| < \infty$,*

$$\mathsf{E}\xi_n \to \mathsf{E}\xi \tag{8}$$

and

$$\mathsf{E}|\xi_n - \xi| \to 0 \tag{9}$$

as $n \to \infty$.

PROOF. Formula (7) is valid by Fatou's lemma. By hypothesis, $\varliminf \xi_n = \varlimsup \xi_n = \xi$ (a.s.). Therefore by Property **G**,

$$\mathsf{E} \varliminf \xi_n = \varliminf \mathsf{E}\xi_n = \varlimsup \mathsf{E}\xi_n = \mathsf{E} \varlimsup \xi_n = \mathsf{E}\xi,$$

which establishes (8). It is also clear that $|\xi| \leq \eta$. Hence $\mathsf{E}|\xi| < \infty$.

Conclusion (9) can be proved in the same way if we observe that $|\xi_n - \xi| \leq 2\eta$.

Corollary. *Let η, ξ, $\xi_1, \ldots$ be random variables such that $|\xi_n| \le \eta$, $\xi_n \to \xi$ (a.s.) and $\mathsf{E}\eta^p < \infty$ for some $p > 0$. Then $\mathsf{E}|\xi|^p < \infty$ and $\mathsf{E}|\xi - \xi_n|^p \to 0$, $n \to \infty$.*

For the proof, it is sufficient to observe that

$$|\xi| \le \eta, |\xi - \xi_n|^p \le (|\xi| + |\xi_n|)^p \le (2\eta)^p.$$

The condition "$|\xi_n| \le \eta$, $\mathsf{E}\eta < \infty$" that appears in Fatou's lemma and the dominated convergence theorem, and ensures the validity of formulas (7)–(9), can be somewhat weakened. In order to be able to state the corresponding result (Theorem 4), we introduce the following definition.

Definition 4. A family $\{\xi_n\}_{n\ge1}$ of random variables is said to be *uniformly integrable* if

$$\sup_n \int_{\{|\xi_n| > c\}} |\xi_n| \mathsf{P}(d\omega) \to 0, \qquad c \to \infty, \tag{10}$$

or, in a different notation,

$$\sup_n \mathsf{E}[|\xi_n| I_{\{|\xi_n| > c\}}] \to 0, \qquad c \to \infty. \tag{11}$$

It is clear that if ξ_n, $n \ge 1$, satisfy $|\xi_n| \le \eta$, $\mathsf{E}\eta < \infty$, then the family $\{\xi_n\}_{n\ge1}$ is uniformly integrable.

Theorem 4. *Let $\{\xi_n\}_{n\ge1}$ be a uniformly integrable family of random variables. Then*

(a) $\mathsf{E} \varliminf \xi_n \le \varliminf \mathsf{E}\xi_n \le \varlimsup \mathsf{E}\xi_n \le \mathsf{E} \varlimsup \xi_n$.

(b) *If in addition $\xi_n \to \xi$ (a.s.) then ξ is integrable and*

$$\mathsf{E}\xi_n \to \mathsf{E}\xi, \qquad n \to \infty,$$

$$\mathsf{E}|\xi_n - \xi| \to 0, \qquad n \to \infty.$$

Proof. (a) For every $c > 0$

$$\mathsf{E}\xi_n = \mathsf{E}[\xi_n I_{\{\xi_n < -c\}}] + \mathsf{E}[\xi_n I_{\{\xi_n \ge -c\}}]. \tag{12}$$

By uniform integrability, for every $\varepsilon > 0$ we can take c so large that

$$\sup_n |\mathsf{E}[\xi_n I_{\{\xi_n < -c\}}]| < \varepsilon. \tag{13}$$

By Fatou's lemma,

$$\varliminf \mathsf{E}[\xi_n I_{\{\xi_n \ge -c\}}] \ge \mathsf{E}[\varliminf \xi_n I_{\{\xi_n \ge -c\}}].$$

But $\xi_n I_{\{\xi_n \ge -c\}} \ge \xi_n$ and therefore

$$\varliminf \mathsf{E}[\xi_n I_{\{\xi_n \ge -c\}}] \ge \mathsf{E}[\varliminf \xi_n]. \tag{14}$$

From (12)–(14) we obtain

$$\underline{\lim}\, \mathsf{E}\xi_n \geq \mathsf{E}[\underline{\lim}\, \xi_n] - \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, it follows that $\underline{\lim}\, \mathsf{E}\xi_n \geq \mathsf{E}\, \underline{\lim}\, \xi_n$. The inequality with upper limits, $\overline{\lim}\, \mathsf{E}\xi_n \leq \mathsf{E}\, \overline{\lim}\, \xi_n$, is proved similarly.

Conclusion (b) can be deduced from (a) as in Theorem 3.

The deeper significance of the concept of uniform integrability is revealed by the following theorem, which gives a necessary and sufficient condition for taking limits under the expectation sign.

Theorem 5. *Let* $0 \leq \xi_n \to \xi$ *and* $\mathsf{E}\xi_n < \infty$. *Then* $\mathsf{E}\xi_n \to \mathsf{E}\xi < \infty$ *if and only if the family* $\{\xi_n\}_{n\geq 1}$ *is uniformly integrable.*

Proof. The sufficiency follows from conclusion (b) of Theorem 4. For the proof of the necessity we consider the (at most countable) set

$$A = \{a \colon \mathsf{P}(\xi = a) > 0\}.$$

Then we have $\xi_n I_{\{\xi_n < a\}} \to \xi I_{\{\xi < a\}}$ for each $a \notin A$, and the family

$$\{\xi_n I_{\{\xi_n < a\}}\}_{n\geq 1}$$

is uniformly integrable. Hence, by the sufficiency part of the theorem, we have $\mathsf{E}\xi_n I_{\{\xi_n < a\}} \to \mathsf{E}\xi I_{\{\xi_n < a\}}$, $a \notin A$, and therefore

$$\mathsf{E}\xi_n I_{\{\xi_n \geq a\}} \to \mathsf{E}\xi I_{\{\xi \geq a\}}, \qquad a \notin A, \quad n \to \infty. \tag{15}$$

Take an $\varepsilon > 0$ and choose $a_0 \notin A$ so large that $\mathsf{E}\xi I_{\{\xi \geq a_0\}} < \varepsilon/2$; then choose N_0 so large that

$$\mathsf{E}\xi_n I_{\{\xi_n \geq a_0\}} \leq \mathsf{E}\xi I_{\{\xi \geq a_0\}} + \varepsilon/2$$

for all $n \geq N_0$, and consequently $\mathsf{E}\xi_n I_{\{\xi_n \geq a_0\}} \leq \varepsilon$. Then choose $a_1 \geq a_0$ so large that $\mathsf{E}\xi_{\{\xi_n \geq a_1\}} \leq \varepsilon$ for all $n \leq N_0$. Then we have

$$\sup_n \mathsf{E}\xi_n I_{\{\xi_n \geq a_1\}} \leq \varepsilon,$$

which establishes the uniform integrability of the family $\{\xi_n\}_{n\geq 1}$ of random variables.

4. Let us notice some tests for uniform integrability.

We first observe that if $\{\xi_n\}$ is a family of uniformly integrable random variables, then

$$\sup_n \mathsf{E}|\xi_n| < \infty. \tag{16}$$

In fact, for a given $\varepsilon > 0$ and sufficiently large $c > 0$

$$\sup_n \mathsf{E}|\xi_n| = \sup_n [\mathsf{E}(|\xi_n| I_{\{|\xi_n| \geq c\}}) + \mathsf{E}(|\xi_n| I_{\{|\xi_n| < c\}})]$$
$$\leq \sup_n \mathsf{E}(|\xi_n| I_{\{|\xi_n| \geq c\}}) + \sup_n \mathsf{E}(|\xi_n| I_{\{|\xi_n| < c\}}) \leq \varepsilon + c,$$

which establishes (16).

It turns out that (16) together with a condition of uniform continuity is necessary and sufficient for uniform integrability.

Lemma 2. *A necessary and sufficient condition for a family* $\{\xi_n\}_{n \geq 1}$ *of random variables to be uniformly integrable is that* $\mathsf{E}|\xi_n|$, $n \geq 1$, *are uniformly bounded (i.e.,* (16) *holds) and that* $\mathsf{E}\{|\xi_n| I_A\}, n \geq 1$, *are uniformly absolutely continuous (i.e.* $\sup \mathsf{E}\{|\xi_n| I_A\} \to 0$ *when* $\mathsf{P}(A) \to 0$).

PROOF. *Necessity.* Condition (16) was verified above. Moreover,

$$\mathsf{E}\{|\xi_n| I_A\} = \mathsf{E}\{|\xi_n| I_{A \cap \{|\xi_n| \geq c\}}\} + \mathsf{E}\{|\xi_n| I_{A \cap \{|\xi_n| < c\}}\}$$
$$\leq \mathsf{E}\{|\xi_n| I_{\{|\xi_n| \geq c\}}\} + c\mathsf{P}(A). \tag{17}$$

Take c so large that $\sup_n \mathsf{E}\{|\xi| I_{\{|\xi_n| \geq c\}}\} \leq \varepsilon/2$. Then if $\mathsf{P}(A) \leq \varepsilon/2c$, we have

$$\sup_n \mathsf{E}\{|\xi_n| I_A\} \leq \varepsilon$$

by (17). This establishes the uniform absolute continuity.

Sufficiency. Let $\varepsilon > 0$ and $\delta > 0$ be chosen so that $\mathsf{P}(A) < \delta$ implies that $\mathsf{E}(|\xi_n| I_A) \leq \varepsilon$, uniformly in n. Since

$$\mathsf{E}|\xi_n| \geq \mathsf{E}|\xi_n| I_{\{|\xi_n| \geq c\}} \geq c\mathsf{P}\{|\xi_n| \geq c\}$$

for every $c > 0$ (cf. Chebyshev's inequality), we have

$$\sup_n \mathsf{P}\{|\xi_n| \geq c\} \leq \frac{1}{c} \sup \mathsf{E}|\xi_n| \to 0, \qquad c \to \infty,$$

and therefore, when c is sufficiently large, any set $\{|\xi_n| \geq c\}$, $n \geq 1$, can be taken as A. Therefore $\sup \mathsf{E}(|\xi_n| I_{\{|\xi_n| \geq c\}}) \leq \varepsilon$, which establishes the uniform integrability. This completes the proof of the lemma.

The following proposition provides a simple sufficient condition for uniform integrability.

Lemma 3. *Let* $\xi_1, \xi_2, \ldots$ *be a sequence of integrable random variables and* $G = G(t)$ *a nonnegative increasing function, defined for* $t \geq 0$, *such that*

$$\lim_{t \to \infty} \frac{G(t)}{t} = \infty. \tag{18}$$

$$\sup_n \mathsf{E}[G(|\xi_n|)] < \infty. \tag{19}$$

Then the family $\{\xi_n\}_{n \geq 1}$ *is uniformly integrable.*

PROOF. Let $\varepsilon > 0$, $M = \sup_n \mathsf{E}[G(|\xi_n|)]$, $a = M/\varepsilon$. Take c so large that $G(t)/t \geq a$ for $t \geq c$. Then

$$\mathsf{E}[|\xi_n| I_{\{|\xi_n| \geq c\}}] \leq \frac{1}{a} \mathsf{E}[G(|\xi_n|) \cdot I_{\{|\xi_n| \geq c\}}] \leq \frac{M}{a} = \varepsilon$$

uniformly for $n \geq 1$.

5. If ξ and η are independent simple random variables, we can show, as in Subsection 5 of §4 of Chapter I, that $\mathsf{E}\xi\eta = \mathsf{E}\xi \cdot \mathsf{E}\eta$. Let us now establish a similar proposition in the general case (see also Problem 5).

Theorem 6. *Let ξ and η be independent random variables with $\mathsf{E}|\xi| < \infty$, $\mathsf{E}|\eta| < \infty$. Then $\mathsf{E}|\xi\eta| < \infty$ and*

$$\mathsf{E}\xi\eta = \mathsf{E}\xi \cdot \mathsf{E}\eta. \tag{20}$$

PROOF. First let $\xi \geq 0$, $\eta \geq 0$. Put

$$\xi_n = \sum_{k=0}^{\infty} \frac{k}{n} I_{\{k/n \leq \xi(\omega) < (k+1)/n\}},$$

$$\eta_n = \sum_{k=0}^{\infty} \frac{k}{n} I_{\{k/n \leq \eta(\omega) < (k+1)/n\}}.$$

Then $\xi_n \leq \xi$, $|\xi_n - \xi| \leq 1/n$ and $\eta_n \leq \eta$, $|\eta_n - \eta| \leq 1/n$. Since $\mathsf{E}\xi < \infty$ and $\mathsf{E}\eta < \infty$, it follows from Lebesgue's dominated convergence theorem that

$$\lim \mathsf{E}\xi_n = \mathsf{E}\xi, \qquad \lim \mathsf{E}\eta_n = \mathsf{E}\eta.$$

Moreover, since ξ and η are independent,

$$\begin{aligned}\mathsf{E}\xi_n\eta_n &= \sum_{k,l \geq 0} \frac{kl}{n^2} \mathsf{E} I_{\{k/n \leq \xi < (k+1)/n\}} I_{\{l/n \leq \eta < (l+1)/n\}} \\ &= \sum_{k,l \geq 0} \frac{kl}{n^2} \mathsf{E} I_{\{k/n \leq \xi < (k+1)/n\}} \cdot \mathsf{E} I_{\{l/n \leq \eta < (l+1)/n\}} = \mathsf{E}\xi_n \cdot \mathsf{E}\eta_n.\end{aligned}$$

Now notice that

$$\begin{aligned}|\mathsf{E}\xi\eta - \mathsf{E}\xi_n\eta_n| &\leq \mathsf{E}|\xi\eta - \xi_n\eta_n| \leq \mathsf{E}[|\xi| \cdot |\eta - \eta_n|] \\ &\quad + \mathsf{E}[|\eta_n| \cdot |\xi - \xi_n|] \leq \frac{1}{n}\mathsf{E}\xi + \frac{1}{n}\mathsf{E}\left(\eta + \frac{1}{n}\right) \to 0, \quad n \to \infty.\end{aligned}$$

Therefore $\mathsf{E}\xi\eta = \lim_n \mathsf{E}\xi_n\eta_n = \lim \mathsf{E}\xi_n \cdot \lim \mathsf{E}\eta_n = \mathsf{E}\xi \cdot \mathsf{E}\eta$, and $\mathsf{E}\xi\eta < \infty$.

The general case reduces to this one if we use the representations $\xi = \xi^+ - \xi^-$, $\eta = \eta^+ - \eta^-$, $\xi\eta = \xi^+\eta^+ - \xi^-\eta^+ - \xi^+\eta^- + \xi^-\eta^-$. This completes the proof.

6. The inequalities for expectations that we develop in this subsection are regularly used both in probability theory and in analysis.

Chebyshev's Inequality. *Let ξ be a nonnegative random variable. Then for every $\varepsilon > 0$*

$$P(\xi \geq \varepsilon) \leq \frac{E\xi}{\varepsilon}. \tag{21}$$

The proof follows immediately from

$$E\xi \geq E[\xi \cdot I_{(\xi \geq \varepsilon)}] \geq \varepsilon E I_{(\xi \geq \varepsilon)} = \varepsilon P(\xi \geq \varepsilon).$$

From (21) we can obtain the following variant of Chebyshev's inequality: If ξ is any random variable then

$$P(\xi \geq \varepsilon) \leq \frac{E\xi^2}{\varepsilon^2} \tag{22}$$

and

$$P(|\xi - E\xi| \geq \varepsilon) \leq \frac{V\xi^2}{\varepsilon^2}, \tag{23}$$

where $V\xi = E(\xi - E\xi)^2$ is the variance of ξ.

The Cauchy–Bunyakovskii Inequality. *Let ξ and η satisfy $E\xi^2 < \infty, E\eta^2 < \infty$. Then $E|\xi\eta| < \infty$ and*

$$(E|\xi\eta|)^2 \leq E\xi^2 \cdot E\eta^2. \tag{24}$$

PROOF. Suppose that $E\xi^2 > 0$, $E\eta^2 > 0$. Then, with $\tilde{\xi} = \xi/\sqrt{E\xi^2}$, $\tilde{\eta} = \eta/\sqrt{E\eta^2}$, we find, since $2|\tilde{\xi}\tilde{\eta}| \leq \tilde{\xi}^2 + \tilde{\eta}^2$, that

$$2E|\tilde{\xi}\tilde{\eta}| \leq E\tilde{\xi}^2 + E\tilde{\eta}^2 = 2,$$

i.e. $E|\tilde{\xi}\tilde{\eta}| \leq 1$, which establishes (24).

On the other hand if, say, $E\xi^2 = 0$, then $\xi = 0$ (a.s.) by Property **I**, and then $E\xi\eta = 0$ by Property **F**, i.e. (24) is still satisfied.

Jensen's Inequality. *Let the Borel function $g = g(x)$ be convex downward and $E|\xi| < \infty$. Then*

$$g(E\xi) \leq Eg(\xi). \tag{25}$$

PROOF. If $g = g(x)$ is convex downward, for each $x_0 \in R$ there is a number $\lambda(x_0)$ such that

$$g(x) \geq g(x_0) + (x - x_0) \cdot \lambda(x_0) \tag{26}$$

for all $x \in R$. Putting $x = \xi$ and $x_0 = E\xi$, we find from (26) that

$$g(\xi) \geq g(E\xi) + (\xi - E\xi) \cdot \lambda(E\xi),$$

and consequently $Eg(\xi) \geq g(E\xi)$.

A whole series of useful inequalities can be derived from Jensen's inequality. We obtain the following one as an example.

Lyapunov's Inequality. *If* $0 < s < t$,

$$(\mathsf{E}|\xi|^s)^{1/s} \leq (\mathsf{E}|\xi|^t)^{1/t}. \tag{27}$$

To prove this, let $r = t/s$. Then, putting $\eta = |\xi|^s$ and applying Jensen's inequality to $g(x) = |x|^r$, we obtain $|\mathsf{E}\eta|^r \leq \mathsf{E}|\eta|^r$, i.e.

$$(\mathsf{E}|\xi|^s)^{t/s} \leq \mathsf{E}|\xi|^t,$$

which establishes (27).

The following chain of inequalities among absolute moments in a consequence of Lyapunov's inequality:

$$\mathsf{E}|\xi| \leq (\mathsf{E}|\xi|^2)^{1/2} \leq \cdots \leq (\mathsf{E}|\xi|^n)^{1/n}. \tag{28}$$

Hölder's Inequality. *Let* $1 < p < \infty$, $1 < q < \infty$, *and* $(1/p) + (1/q) = 1$. *If* $\mathsf{E}|\xi|^p < \infty$ *and* $\mathsf{E}|\eta|^q < \infty$, *then* $\mathsf{E}|\xi\eta| < \infty$ *and*

$$\mathsf{E}|\xi\eta| \leq (\mathsf{E}|\xi|^p)^{1/p}(\mathsf{E}|\eta|^q)^{1/q}. \tag{29}$$

If $\mathsf{E}|\xi|^p = 0$ or $\mathsf{E}|\eta|^q = 0$, (29) follows immediately as for the Cauchy–Bunyakovskii inequality (which is the special case $p = q = 2$ of Hölder's inequality).

Now let $\mathsf{E}|\xi|^p > 0$, $\mathsf{E}|\eta|^q > 0$ and

$$\tilde{\xi} = \frac{\xi}{(\mathsf{E}|\xi|^p)^{1/p}}, \qquad \tilde{\eta} = \frac{\eta}{(\mathsf{E}|\eta|^q)^{1/q}}.$$

We apply the inequality

$$x^a y^b \leq ax + by, \tag{30}$$

which holds for positive x, y, a, b and $a + b = 1$, and follows immediately from the concavity of the logarithm:

$$\ln[ax + by] \geq a \ln x + b \ln y = \ln x^a y^b.$$

Then, putting $x = |\tilde{\xi}|^p$, $y = |\tilde{\eta}|^q$, $a = 1/p$, $b = 1/q$, we find that

$$|\tilde{\xi}\tilde{\eta}| \leq \frac{1}{p}|\tilde{\xi}|^p + \frac{1}{q}|\tilde{\eta}|^q,$$

whence

$$\mathsf{E}|\tilde{\xi}\tilde{\eta}| \leq \frac{1}{p}\mathsf{E}|\tilde{\xi}|^p + \frac{1}{q}\mathsf{E}|\tilde{\eta}|^q = \frac{1}{p} + \frac{1}{q} = 1.$$

This establishes (29).

Minkowski's Inequality. *If* $\mathsf{E}|\xi|^p < \infty$, $\mathsf{E}|\eta|^p < \infty$, $1 \le p < \infty$, *then we have* $\mathsf{E}|\xi + \eta|^p < \infty$ *and*

$$(\mathsf{E}|\xi + \eta|^p)^{1/p} \le (\mathsf{E}|\xi|^p)^{1/p} + (\mathsf{E}|\eta|^p)^{1/p}. \tag{31}$$

We begin by establishing the following inequality: if $a, b > 0$ and $p \ge 1$, then

$$(a + b)^p \le 2^{p-1}(a^p + b^p). \tag{32}$$

In fact, consider the function $F(x) = (a + x)^p - 2^{p-1}(a^p + x^p)$. Then

$$F'(x) = p(a + x)^{p-1} - 2^{p-1}px^{p-1},$$

and since $p \ge 1$, we have $F'(a) = 0$, $F'(x) > 0$ for $x < a$ and $F'(x) < 0$ for $x > a$. Therefore

$$F(b) \le \max F(x) = F(a) = 0,$$

from which (32) follows.

According to this inequality,

$$|\xi + \eta|^p \le (|\xi| + |\eta|)^p \le 2^{p-1}(|\xi|^p + |\eta|^p) \tag{33}$$

and therefore if $\mathsf{E}|\xi|^p < \infty$ and $\mathsf{E}|\eta|^p < \infty$ it follows that $\mathsf{E}|\xi + \eta|^p < \infty$. If $p = 1$, inequality (31) follows from (33).

Now suppose that $p > 1$. Take $q > 1$ so that $(1/p) + (1/q) = 1$. Then

$$|\xi + \eta|^p = |\xi + \eta| \cdot |\xi + \eta|^{p-1} \le |\xi| \cdot |\xi + \eta|^{p-1} + |\eta| |\xi + \eta|^{p-1}. \tag{34}$$

Notice that $(p - 1)q = p$. Consequently

$$\mathsf{E}(|\xi + \eta|^{p-1})^q = \mathsf{E}|\xi + \eta|^p < \infty,$$

and therefore by Hölder's inequality

$$\begin{aligned}\mathsf{E}(|\xi||\xi + \eta|^{p-1}) &\le (\mathsf{E}|\xi|^p)^{1/p}(\mathsf{E}|\xi + \eta|^{(p-1)q})^{1/q} \\ &= (\mathsf{E}|\xi|^p)^{1/p}(\mathsf{E}|\xi + \eta|^p)^{1/q} < \infty.\end{aligned}$$

In the same way,

$$\mathsf{E}(|\eta||\xi + \eta|^{p-1}) \le (\mathsf{E}|\eta|^p)^{1/p}(\mathsf{E}|\xi + \eta|^p)^{1/q}.$$

Consequently, by (34),

$$\mathsf{E}|\xi + \eta|^p \le (\mathsf{E}|\xi + \eta|^p)^{1/q}((\mathsf{E}|\xi|^p)^{1/p} + (\mathsf{E}|\eta|^p)^{1/p}). \tag{35}$$

If $\mathsf{E}|\xi + \eta|^p = 0$, the desired inequality (31) is evident. Now let $\mathsf{E}|\xi + \eta|^p > 0$. Then we obtain

$$(\mathsf{E}|\xi + \eta|^p)^{1-(1/q)} \le (\mathsf{E}|\xi|^p)^{1/p} + (\mathsf{E}|\eta|^p)^{1/p}$$

from (35), and (31) follows since $1 - (1/q) = 1/p$.

7. Let ξ be a random variable for which $\mathsf{E}\xi$ is defined. Then, by Property **D**, the set function

$$\mathsf{Q}(A) \equiv \int_A \xi \, d\mathsf{P}, \qquad A \in \mathscr{F}, \tag{36}$$

is well defined. Let us show that this function is countably additive.

First suppose that ξ is nonnegative. If $A_1, A_2, \ldots$ are pairwise disjoint sets from $\mathscr{F}$ and $A = \sum A_n$, the corollary to Theorem 1 implies that

$$\begin{aligned}\mathsf{Q}(A) &= \mathsf{E}(\xi \cdot I_A) = \mathsf{E}(\xi \cdot I_{\Sigma A_n}) = \mathsf{E}(\textstyle\sum \xi \cdot I_{A_n}) \\ &= \textstyle\sum \mathsf{E}(\xi \cdot I_{A_n}) = \sum \mathsf{Q}(A_n).\end{aligned}$$

If ξ is an arbitrary random variable for which $\mathsf{E}\xi$ is defined, the countable additivity of $\mathsf{Q}(A)$ follows from the representation

$$\mathsf{Q}(A) = \mathsf{Q}^+(A) - \mathsf{Q}^-(A), \tag{37}$$

where

$$\mathsf{Q}^+(A) = \int_A \xi^+ \, d\mathsf{P}, \qquad \mathsf{Q}^-(A) = \int_A \xi^- \, d\mathsf{P},$$

together with the countable additivity for nonnegative random variables and the fact that $\min(\mathsf{Q}^+(\Omega), \mathsf{Q}^-(\Omega)) < \infty$.

Thus if $\mathsf{E}\xi$ is defined, the set function $\mathsf{Q} = \mathsf{Q}(A)$ is a signed measure—a countably additive set function representable as $\mathsf{Q} = \mathsf{Q}_1 - \mathsf{Q}_2$, where at least one of the measures Q_1 and Q_2 is finite.

We now show that $\mathsf{Q} = \mathsf{Q}(A)$ has the following important property of *absolute continuity* with respect to P:

$$\textit{if} \quad \mathsf{P}(A) = 0 \qquad \textit{then} \quad \mathsf{Q}(A) = 0 \qquad (A \in \mathscr{F})$$

(this property is denoted by the abbreviation $\mathsf{Q} \ll \mathsf{P}$).

To prove the sufficiency we consider nonnegative random variables. If $\xi = \sum_{k=1}^n x_k I_{A_k}$ is a simple nonnegative random variable and $\mathsf{P}(A) = 0$, then

$$\mathsf{Q}(A) = \mathsf{E}(\xi \cdot I_A) = \sum_{k=1}^n x_k \mathsf{P}(A_k \cap A) = 0.$$

If $\{\xi_n\}_{n \geq 1}$ is a sequence of nonnegative simple functions such that $\xi_n \uparrow \xi \geq 0$, then the theorem on monotone convergence shows that

$$\mathsf{Q}(A) = \mathsf{E}(\xi \cdot I_A) = \lim \mathsf{E}(\xi_n \cdot I_A) = 0,$$

since $\mathsf{E}(\xi_n \cdot I_A) = 0$ for all $n \geq 1$ and A with $\mathsf{P}(A) = 0$.

Thus the Lebesgue integral $\mathsf{Q}(A) = \int_A \xi \, d\mathsf{P}$, considered as a function of sets $A \in \mathscr{F}$, is a signed measure that is absolutely continuous with respect to P ($\mathsf{Q} \ll \mathsf{P}$). It is quite remarkable that the converse is also valid.

Radon–Nikodým Theorem. *Let $(\Omega, \mathscr{F})$ be a measurable space, μ a σ-finite measure, and λ a signed measure (i.e., $\lambda = \lambda_1 - \lambda_2$, where at least one of the measures λ_1 and λ_2 is finite) which is absolutely continuous with respect to μ. Then there is an $\mathscr{F}$-measurable function $f = f(\omega)$ with values in $\bar{R} = [-\infty, \infty]$ such that*

$$\lambda(A) = \int_A f(\omega)\mu(d\omega), \qquad A \in \mathscr{F}. \tag{38}$$

The function $f(\omega)$ is unique up to sets of μ-measure zero: if $h = h(\omega)$ is another $\mathscr{F}$-measurable function such that $\lambda(A) = \int_A h(\omega)\mu(d\omega)$, $A \in \mathscr{F}$, then $\mu\{\omega: f(\omega) \neq h(\omega)\} = 0$.

If λ is a measure, then $f = f(\omega)$ has its values in $\bar{R}^+ = [0, \infty]$.

Remark. The function $f = f(\omega)$ in the representation (38) is called the *Radon–Nikodým derivative* or the *density* of the measure λ with respect to μ, and denoted by $d\lambda/d\mu$ or $(d\lambda/d\mu)(\omega)$.

The Radon–Nikodým theorem, which we quote without proof, will play a key role in the construction of conditional expectations (§7).

8. If $\xi = \sum_{i=1}^{n} x_i I_{A_i}$ is a simple random variable,

$$\mathsf{E}g(\xi) = \sum g(x_i)\mathsf{P}(A_i) = \sum g(x_i)\Delta F_\xi(x_i). \tag{39}$$

In other words, in order to calculate the expectation of a function of the (simple) random variable ξ it is unnecessary to know the probability measure P completely; it is enough to know the probability distribution P_ξ or, equivalently, the distribution function F_ξ of ξ.

The following important theorem generalizes this property.

Theorem 7 (Change of Variables in a Lebesgue Integral). *Let $(\Omega, \mathscr{F})$ and $(E, \mathscr{E})$ be measurable spaces and $X = X(\omega)$ an $\mathscr{F}/\mathscr{E}$-measurable function with values in E. Let P be a probability measure on $(\Omega, \mathscr{F})$ and P_X the probability measure on $(E, \mathscr{E})$ induced by $X = X(\omega)$:*

$$P_X(A) = \mathsf{P}\{\omega: X(\omega) \in A\}, \qquad A \in \mathscr{E}. \tag{40}$$

Then

$$\int_A g(x)P_X(dx) = \int_{X^{-1}(A)} g(X(\omega))\mathsf{P}(d\omega), \qquad A \in \mathscr{E}, \tag{41}$$

for every $\mathscr{E}$-measurable function $g = g(x)$, $x \in E$ (in the sense that if one integral exists, the other is well defined, and the two are equal).

Proof. Let $A \in \mathscr{E}$ and $g(x) = I_B(x)$, where $B \in \mathscr{E}$. Then (41) becomes

$$P_X(AB) = \mathsf{P}(X^{-1}(A) \cap X^{-1}(B)), \tag{42}$$

which follows from (40) and the observation that $X^{-1}(A) \cap X^{-1}(B) = X^{-1}(A \cap B)$.

It follows from (42) that (41) is valid for nonnegative simple functions $g = g(x)$, and therefore, by the monotone convergence theorem, also for all nonnegative $\mathscr{E}$-measurable functions.

In the general case we need only represent g as $g^+ - g^-$. Then, since (41) is valid for g^+ and g^-, if (for example) $\int_A g^+(x)P_X(dx) < \infty$, we have

$$\int_{X^{-1}(A)} g^+(X(\omega))P(d\omega) < \infty$$

also, and therefore the existence of $\int_A g(x)P_X(dx)$ implies the existence of $\int_{X^{-1}(A)} g(X(\omega))\mathsf{P}(d\omega)$.

Corollary. *Let $(E, \mathscr{E}) = (R, \mathscr{B}(R))$ and let $\xi = \xi(\omega)$ be a random variable with probability distribution P_ξ. Then if $g = g(x)$ is a Borel function and either of the integrals $\int_A g(x)P_\xi(dx)$ or $\int_{\xi^{-1}(A)} g(\xi(\omega))\mathsf{P}(d\omega)$ exists, we have*

$$\int_A g(x)P_\xi(dx) = \int_{\xi^{-1}(A)} g(\xi(\omega))\mathsf{P}(d\omega).$$

In particular, for $A = R$ we obtain

$$\mathsf{E}g(\xi(\omega)) = \int_\Omega g(\xi(\omega))\mathsf{P}(d\omega) = \int_R g(x)P_\xi(dx). \tag{43}$$

The measure P_ξ can be uniquely reconstructed from the distribution function F_ξ (Theorem 1 of §3). Hence the Lebesgue integral $\int_R g(x)P_\xi(dx)$ is often denoted by $\int_R g(x)F_\xi(dx)$ and called a *Lebesgue–Stieltjes integral* (with respect to the measure corresponding to the distribution function $F_\xi(x)$).

Let us consider the case when $F_\xi(x)$ has a density $f_\xi(x)$, i.e. let

$$F_\xi(x) = \int_{-\infty}^x f_\xi(y)\,dy, \tag{44}$$

where $f_\xi = f_\xi(x)$ is a nonnegative Borel function and the integral is a Lebesgue integral with respect to Lebesgue measure on the set $(-\infty, x]$ (see Remark 2 in Subsection 1). With the assumption of (44), formula (43) takes the form

$$\mathsf{E}g(\xi(\omega)) = \int_{-\infty}^{\infty} g(x)f_\xi(x)\,dx, \tag{45}$$

where the integral is the Lebesgue integral of the function $g(x)f_\xi(x)$ with respect to Lebesgue measure. In fact, if $g(x) = I_B(x)$, $B \in \mathscr{B}(R)$, the formula becomes

$$P_\xi(B) = \int_B f_\xi(x)\,dx, \qquad B \in \mathscr{B}(R); \tag{46}$$

its correctness follows from Theorem 1 of §3 and the formula

$$F_\xi(b) - F_\xi(a) = \int_a^b f_\xi(x)\, dx.$$

In the general case, the proof is the same as for Theorem 7.

9. Let us consider the special case of measurable spaces $(\Omega, \mathscr{F})$ with a measure μ, where $\Omega = \Omega_1 \times \Omega_2$, $\mathscr{F} = \mathscr{F}_1 \otimes \mathscr{F}_2$, and $\mu = \mu_1 \times \mu_2$ is the direct product of measures μ_1 and μ_2 (i.e., the measure on $\mathscr{F}$ such that

$$\mu_1 \times \mu_2(A \times B) = \mu_1(A_1)\mu_2(B), \qquad A \in \mathscr{F}_1, \quad B \in \mathscr{F}_2;$$

the existence of this measure follows from the proof of Theorem 8).

The following theorem plays the same role as the theorem on the reduction of a double Riemann integral to an iterated integral.

Theorem 8 (Fubini's Theorem). *Let $\xi = \xi(\omega_1, \omega_2)$ be an $\mathscr{F}_1 \otimes \mathscr{F}_2$-measurable function, integrable with respect to the measure $\mu_1 \times \mu_2$:*

$$\int_{\Omega_1 \times \Omega_2} |\xi(\omega_1, \omega_2)|\, d(\mu_1 \times \mu_2) < \infty. \tag{47}$$

Then the integrals $\int_{\Omega_1} \xi(\omega_1, \omega_2)\mu_1(d\omega_1)$ and $\int_{\Omega_2} \xi(\omega_1, \omega_2)\mu_2(d\omega_2)$

(1) *are defined for all ω_1 and ω_2;*

(2) *are respectively $\mathscr{F}_2$- and $\mathscr{F}_1$-measurable functions with*

$$\begin{aligned} \mu_2\left\{\omega_2 : \int_{\Omega_1} |\xi(\omega_1, \omega_2)|\, \mu_1(d\omega_1) = \infty\right\} &= 0, \\ \mu_1\left\{\omega_1 : \int_{\Omega_2} |\xi(\omega_1, \omega_2)|\, \mu_2(d\omega_2) = \infty\right\} &= 0 \end{aligned} \tag{48}$$

and (3)

$$\begin{aligned} \int_{\Omega_1 \times \Omega_2} \xi(\omega_1, \omega_2)\, d(\mu_1 \times \mu_2) &= \int_{\Omega_1}\left[\int_{\Omega_2} \xi(\omega_1, \omega_2)\mu_2(d\omega_2)\right]\mu_1(d\omega_1) \\ &= \int_{\Omega_2}\left[\int_{\Omega_1} \xi(\omega_1, \omega_2)\mu_1(d\omega_1)\right]\mu_2(d\omega_2). \end{aligned} \tag{49}$$

Proof. We first show that $\xi_{\omega_1}(\omega_2) = \xi(\omega_1, \omega_2)$ is $\mathscr{F}_2$-measurable with respect to ω_2, for each $\omega_1 \in \Omega_1$.

Let $F \in \mathscr{F}_1 \otimes \mathscr{F}_2$ and $\xi(\omega_1, \omega_2) = I_F(\omega_1, \omega_2)$. Let

$$F_{\omega_1} = \{\omega_2 \in \Omega_2 : (\omega_1, \omega_2) \in \mathscr{F}\}$$

be the *cross-section* of F at ω_1, and let $\mathscr{C}_{\omega_1} = \{F \in \mathscr{F}: F_{\omega_1} \in F_2\}$. We must show that $\mathscr{C}_{\omega_1} = \mathscr{F}$ for every ω_1.

If $F = A \times B$, $A \in \mathscr{F}_1$, $B \in \mathscr{F}_2$, then

$$(A \times B)_{\omega_1} = \begin{cases} B & \text{if } \omega_1 \in A, \\ \varnothing & \text{if } \omega_1 \notin A. \end{cases}$$

Hence rectangles with measurable sides belong to $\mathscr{C}_{\omega_1}$. In addition, if $F \in \mathscr{F}$, then $(\overline{F})_{\omega_1} = \overline{F_{\omega_1}}$, and if $\{F^n\}_{n \geq 1}$ are sets in $\mathscr{F}$, then $(\bigcup F^n)_{\omega_1} = \bigcup F^n_{\omega_1}$. It follows that $\mathscr{C}_{\omega_1} = \mathscr{F}$.

Now let $\xi(\omega_1, \omega_2) \geq 0$. Then, since the function $\xi(\omega_1, \omega_2)$ is $\mathscr{F}_2$-measurable for each ω_1, the integral $\int_{\Omega_2} \xi(\omega_1, \omega_2)\mu_2(d\omega_2)$ is defined. Let us show that this integral is an $\mathscr{F}_1$-measurable function and

$$\int_{\Omega_1}\left[\int_{\Omega_2} \xi(\omega_1, \omega_2)\mu_2(d\omega_2)\right]\mu_1(d\omega_1) = \int_{\Omega_1 \times \Omega_2} \xi(\omega_1, \omega_2)\, d(\mu_1 \times \mu_2). \quad (50)$$

Let us suppose that $\xi(\omega_1, \omega_2) = I_{A \times B}(\omega_1, \omega_2)$, $A \in \mathscr{F}_1$, $B \in \mathscr{F}_2$. Then since $I_{A \times B}(\omega_1, \omega_2) = I_A(\omega_1)I_B(\omega_2)$, we have

$$\int_{\Omega_2} I_{A \times B}(\omega_1, \omega_2)\mu_2(d\omega_2) = I_A(\omega_1) \int_{\Omega_2} I_B(\omega_2)\mu_2(d\omega_2) \quad (51)$$

and consequently the integral on the left of (51) is an $\mathscr{F}_1$-measurable function.

Now let $\xi(\omega_1, \omega_2) = I_F(\omega_1, \omega_2)$, $F \in \mathscr{F} = \mathscr{F}_1 \otimes \mathscr{F}_2$. Let us show that the integral $f(\omega_1) = \int_{\Omega_2} I_F(\omega_1, \omega_2)\mu_2(d\omega_2)$ is $\mathscr{F}$-measurable. For this purpose we put $\mathscr{C} = \{F \in \mathscr{F}: f(\omega_1) \text{ is } \mathscr{F}_1\text{-measurable}\}$. According to what has been proved, the set $A \times B$ belongs to $\mathscr{C}$ $(A \in \mathscr{F}_1, B \in \mathscr{F}_2)$ and therefore the algebra $\mathscr{A}$ consisting of finite sums of disjoint sets of this form also belongs to $\mathscr{C}$. It follows from the monotone convergence theorem that $\mathscr{C}$ is a monotonic class, $\mathscr{C} = \mu(\mathscr{C})$. Therefore, because of the inclusions $\mathscr{A} \subseteq \mathscr{C} \subseteq \mathscr{F}$ and Theorem 1 of §2, we have $\mathscr{F} = \sigma(\mathscr{A}) = \mu(\mathscr{A}) \subseteq \mu(\mathscr{C}) = \mathscr{C} \subseteq \mathscr{F}$, i.e. $\mathscr{C} = \mathscr{F}$.

Finally, if $\xi(\omega_1, \omega_2)$ is an arbitrary nonnegative $\mathscr{F}$-measurable function, the $\mathscr{F}_1$-measurability of the integral $\int_{\Omega_2} \xi(\omega_1, \omega_2)\mu_2(d\omega)$ follows from the monotone convergence theorem and Theorem 2 of §4.

Let us now show that the measure $\mu = \mu_1 \times \mu_2$ defined on $\mathscr{F} = \mathscr{F}_1 \otimes \mathscr{F}_2$, with the property $(\mu_1 \times \mu_2)(A \times B) = \mu_1(A) \cdot \mu_2(B)$, $A \in \mathscr{F}_1$, $B \in \mathscr{F}_2$, actually exists and is unique.

For $F \in \mathscr{F}$ we put

$$\mu(F) = \int_{\Omega_1}\left[\int_{\Omega_2} I_{F_{\omega_1}}(\omega_2)\mu_2(d\omega_2)\right]\mu_1(d\omega_1).$$

As we have shown, the inner integral is an $\mathscr{F}_1$-measurable function, and consequently the set function $\mu(F)$ is actually defined for $F \in \mathscr{F}$. It is clear that if $F = A \times B$, then $\mu(A \times B) = \mu_1(A)\mu_2(B)$. Now let $\{F^n\}$ be disjoint sets from $\mathscr{F}$. Then

$$\begin{aligned}\mu\left(\sum F^n\right) &= \int_{\Omega_1}\left[\int_{\Omega_2} I_{(\sum F^n)\omega_1}(\omega_2)\mu_2(d\omega_2)\right]\mu_1(d\omega_1)\\ &= \int_{\Omega_1}\sum_n\left[\int_{\Omega_2} I_{F^n_{\omega_1}}(\omega_2)\mu_2(d\omega_2)\right]\mu_1(d\omega_1)\\ &= \sum_n \int_{\Omega_1}\left[\int_{\Omega_2} I_{F^n_{\omega_1}}(\omega_2)\mu_2(d\omega_2)\right]\mu_1(d\omega_1) = \sum_n \mu(F^n),\end{aligned}$$

i.e. μ is a (σ-finite) measure on $\mathscr{F}$.

It follows from Carathéodory's theorem that this measure μ is the unique measure with the property that $\mu(A \times B) = \mu_1(A)\mu_2(B)$.

We can now establish (50). If $\xi(\omega_1, \omega_2) = I_{A\times B}(\omega_1, \omega_2)$, $A \in \mathscr{F}_1$, $B \in \mathscr{F}_2$, then

$$\int_{\Omega_1\times\Omega_2} I_{A\times B}(\omega_1, \omega_2)d(\mu_1 \times \mu_2) = (\mu_1 \times \mu_2)(A \times B), \tag{52}$$

and since $I_{A\times B}(\omega_1, \omega_2) = I_A(\omega_1)I_B(\omega_2)$, we have

$$\begin{aligned}&\int_{\Omega_1}\left[\int_{\Omega_2} I_{A\times B}(\omega_1, \omega_2)\mu_2(d\omega_2)\right]\mu_1(d\omega_1)\\ &\quad = \int_{\Omega_1}\left[I_A(\omega_1)\int_{\Omega_2} I_B(\omega_1, \omega_2)\mu_2(d\omega_2)\right]\mu_1(d\omega_1) = \mu_1(A)\mu_2(B).\end{aligned} \tag{53}$$

But, by the definition of $\mu_1 \times \mu_2$,

$$(\mu_1 \times \mu_2)(A \times B) = \mu_1(A)\mu_2(B).$$

Hence it follows from (52) and (53) that (50) is valid for $\xi(\omega_1, \omega_2) = I_{A\times B}(\omega_1, \omega_2)$.

Now let $\xi(\omega_1, \omega_2) = I_F(\omega_1, \omega_2)$, $F \in \mathscr{F}$. The set function

$$\lambda(F) = \int_{\Omega_2\times\Omega_2} I_F(\omega_1, \omega_2)d(\mu_1 \times \mu_2), \qquad F \in \mathscr{F},$$

is evidently a σ-finite measure. It is also easily verified that the set function

$$\nu(F) = \int_{\Omega_1}\left[\int_{\Omega_2} I_F(\omega_1, \omega_2)\mu_2(d\omega_2)\right]\mu_1(d\omega_1)$$

is a σ-finite measure. It will be shown below that λ and ν coincide on sets of the form $F = A \times B$, and therefore on the algebra $\mathscr{F}$. Hence it follows by Carathéodory's theorem that λ and ν coincide for all $F \in \mathscr{F}$.

We turn now to the proof of the full conclusion of Fubini's theorem. By (47),

$$\int_{\Omega_1 \times \Omega_2} \xi^+(\omega_1, \omega_2)\, d(\mu_1 \times \mu_2) < \infty, \qquad \int_{\Omega_1 \times \Omega_2} \xi^-(\omega_1, \omega_2) d(\mu_1 \times \mu_2) < \infty.$$

By what has already been proved, the integral $\int_{\Omega_2} \xi^+(\omega_1, \omega_2)\mu_2(d\omega_2)$ is an $\mathscr{F}_1$-measurable function of ω_1 and

$$\int_{\Omega_1} \left[\int_{\Omega_2} \xi^+(\omega_1, \omega_2)\mu_2(d\omega_2) \right] \mu_1(d\omega_1) = \int_{\Omega_1 \times \Omega_2} \xi^+(\omega_1, \omega_2)\, d(\mu_1 \times \mu_2) < \infty.$$

Consequently by Problem 4 (see also Property J in Subsection 2)

$$\int_{\Omega_2} \xi^+(\omega_1, \omega_2)\mu_2(d\omega_2) < \infty \quad (\mu_1\text{-a.s.}).$$

In the same way

$$\int_{\Omega_2} \xi^-(\omega_1, \omega_2)\mu_1(d\omega_1) < \infty \quad (\mu_1\text{-a.s.}),$$

and therefore

$$\int_{\Omega_2} |\xi(\omega_1, \omega_2)|\, \mu_2(d\omega_2) < \infty \quad (\mu_1\text{-a.s.}).$$

It is clear that, except on a set $\mathscr{N}$ of μ_1-measure zero,

$$\int_{\Omega_2} \xi(\omega_1, \omega_2)\mu_2(d\omega_2) = \int_{\Omega_2} \xi^+(\omega_1, \omega_2)\mu_2(d\omega_2) - \int_{\Omega_2} \xi^-(\omega_1, \omega_2)\mu_2(d\omega_2). \tag{54}$$

Taking the integrals to be zero for $\omega_1 \in \mathscr{N}$, we may suppose that (54) holds for all $\omega \in \Omega_1$. Then, integrating (54) with respect to μ_1 and using (50), we obtain

$$\begin{aligned}
\int_{\Omega_1} \left[\int_{\Omega_2} \xi(\omega_1, \omega_2)\mu_2(d\omega_2) \right] \mu_1(d\omega_1) &= \int_{\Omega_1} \left[\int_{\Omega_2} \xi^+(\omega_1, \omega_2)\mu_2(d\omega_2) \right] \mu_1(d\omega_1) \\
&\quad - \int_{\Omega_1} \left[\int_{\Omega_2} \xi^-(\omega_1, \omega_2)\mu_2(d\omega_2) \right] \mu_1(d\omega_1) \\
&= \int_{\Omega_1 \times \Omega_2} \xi^+(\omega_1, \omega_2) d(\mu_1 \times \mu_2) \\
&\quad - \int_{\Omega_1 \times \Omega_2} \xi^-(\omega_1, \omega_2)\, d(\mu_1 \times \mu_2) \\
&= \int_{\Omega_1 \times \Omega_2} \xi(\omega_1, \omega_2) d(\mu_1 \times \mu_2).
\end{aligned}$$

Similarly we can establish the first equation in (48) and the equation

$$\int_{\Omega_1 \times \Omega_2} \xi(\omega_1, \omega_2)\, d(\mu_1 \times \mu_2) = \int_{\Omega_2}\left[\int_{\Omega_1} \xi(\omega_1, \omega_2)\mu_1(d\omega_1)\right]\mu_2(d\omega_2).$$

This completes the proof of the theorem.

Corollary. *If* $\int_{\Omega_1} [\int_{\Omega_2} |\xi(\omega_1, \omega_2)|\mu_2(d\omega_2)]\mu_1 (d\omega_1) < \infty$, *the conclusion of Fubini's theorem is still valid.*

In fact, under this hypothesis (47) follows from (50), and consequently the conclusions of Fubini's theorem hold.

EXAMPLE. Let (ξ, η) be a pair of random variables whose distribution has a two-dimensional density $f_{\xi\eta}(x, y)$, i.e.

$$\mathsf{P}((\xi, \eta) \in B) = \int_B f_{\xi\eta}(x, y)\, dx\, dy, \qquad B \in \mathscr{B}(R^2),$$

where $f_{\xi\eta}(x, y)$ is a nonnegative $\mathscr{B}(R^2)$-measurable function, and the integral is a Lebesgue integral with respect to two-dimensional Lebesgue measure.

Let us show that the one-dimensional distributions for ξ and η have densities $f_\xi(x)$ and $f_\eta(y)$, and furthermore

$$f_\xi(x) = \int_{-\infty}^{\infty} f_{\xi\eta}(x, y)\, dy$$

and (55)

$$f_\eta(y) = \int_{-\infty}^{\infty} f_{\xi\eta}(x, y)\, dx.$$

In fact, if $A \in \mathscr{B}(R)$, then by Fubini's theorem

$$\mathsf{P}(\xi \in A) = \mathsf{P}((\xi, \eta) \in A \times R) = \int_{A \times R} f_{\xi\eta}(x, y)\, dx\, dy = \int_A\left[\int_R f_{\xi\eta}(x, y)\, dy\right]dx.$$

This establishes both the existence of a density for the probability distribution of ξ and the first formula in (55). The second formula is established similarly.

According to the theorem in §5, a necessary and sufficient condition that ξ and η are independent is that

$$F_{\xi\eta}(x, y) = F_\xi(x)F_\eta(y), \qquad (x, y) \in R^2.$$

Let us show that when there is a two-dimensional density $f_{\xi\eta}(x, y)$, the variables ξ and η are independent if and only if

$$f_{\xi\eta}(x, y) = f_\xi(x)f_\eta(y) \tag{56}$$

(where the equation is to be understood in the sense of holding almost surely with respect to two-dimensional Lebesgue measure).

In fact, in (56) holds, then by Fubini's theorem

$$F_{\xi\eta}(x, y) = \int_{(-\infty,x]\times(-\infty,y]} f_{\xi\eta}(u, v)\, du\, dv = \int_{(-\infty,x]\times(-\infty,y]} f_\xi(u) f_\eta(v)\, du\, dv$$
$$= \int_{(-\infty,x]} f_\xi(u)\, du \left(\int_{(-\infty,y]} f_\eta(v)\, dv\right) = F_\xi(x) F_\eta(y)$$

and consequently ξ and η are independent.

Conversely, if they are independent and have a density $f_{\xi\eta}(x, y)$, then again by Fubini's theorem

$$\int_{(-\infty,x]\times(-\infty,y]} f_{\xi\eta}(u, v)\, du\, dv = \left(\int_{(-\infty,x]} f_\xi(u)\, du\right)\left(\int_{(-\infty,y]} f_\eta(v)\, dv\right)$$
$$= \int_{(-\infty,x]\times(-\infty,y]} f_\xi(u) f_\eta(v)\, du\, dv.$$

It follows that

$$\int_B f_{\xi\eta}(x, y)\, dx\, dy = \int_B f_\xi(x) f_\eta(y)\, dx\, dy$$

for every $B \in \mathscr{B}(R^2)$, and it is easily deduced from Property I that (56) holds.

10. In this subsection we discuss the relation between the Lebesgue and Riemann integrals.

We first observe that the construction of the Lebesgue integral is independent of the measurable space $(\Omega, \mathscr{F})$ on which the integrands are given. On the other hand, the Riemann integral is not defined on abstract spaces in general, and for $\Omega = R^n$ it is defined sequentially: first for R^1, and then extended, with corresponding changes, to the case $n > 1$.

We emphasize that the constructions of the Riemann and Lebesgue integrals are based on different ideas. The first step in the construction of the Riemann integral is to group the points $x \in R^1$ according to their distances along the x axis. On the other hand, in Lebesgue's construction (for $\Omega = R^1$) the points $x \in R^1$ are grouped according to a different principle: by the distances between the values of the integrand. It is a consequence of these different approaches that the Riemann approximating sums have limits only for "mildly" discontinuous functions, whereas the Lebesgue sums converge to limits for a much wider class of functions.

Let us recall the definition of the Riemann–Stieltjes integral. Let $G = G(x)$ be a generalized distribution function on R (see subsection 2 of §3) and μ its corresponding Lebesgue–Stieltjes measure, and let $g = g(x)$ be a bounded function that vanishes outside $[a, b]$.

Consider a decomposition $\mathscr{P} = \{x_0, \dots, x_n\}$,

$$a = x_0 < x_1 < \cdots < x_n = b,$$

of $[a, b]$, and form the upper and lower sums

$$\overline{\sum_{\mathscr{P}}} = \sum_{i=1}^{n} \bar{g}_i[G(x_i) - G(x_{i-1})], \qquad \underline{\sum_{\mathscr{P}}} = \sum_{i=1}^{n} \underline{g}_i[G(x_i) - G(x_{i-1})]$$

where

$$\bar{g}_i = \sup_{x_{i-1} < y \le x_i} g(y), \qquad \underline{g}_i = \inf_{x_{i-1} < y \le x_i} g(y).$$

Define simple functions $\bar{g}_{\mathscr{P}}(x)$ and $\underline{g}_{\mathscr{P}}(x)$ by taking

$$\bar{g}_{\mathscr{P}}(x) = \bar{g}_i, \qquad \underline{g}_{\mathscr{P}}(x) = \underline{g}_i,$$

on $x_{i-1} < x \le x_i$, and define $\bar{g}_{\mathscr{P}}(a) = \underline{g}_{\mathscr{P}}(a) = g(a)$. It is clear that then

$$\overline{\sum_{\mathscr{P}}} = \text{(L-S)} \int_a^b \bar{g}_{\mathscr{P}}(x) G(dx)$$

and

$$\underline{\sum_{\mathscr{P}}} = \text{(L-S)} \int_a^b \underline{g}_{\mathscr{P}}(x) G(dx).$$

Now let $\{\mathscr{P}_k\}$ be a sequence of decompositions such that $\mathscr{P}_k \subseteq \mathscr{P}_{k+1}$. Then

$$\bar{g}_{\mathscr{P}_1} \ge \bar{g}_{\mathscr{P}_2} \ge \cdots \ge g \ge \cdots \ge \underline{g}_{\mathscr{P}_2} \ge \underline{g}_{\mathscr{P}_1},$$

and if $|g(x)| \le C$ we have, by the dominated convergence theorem,

$$\begin{aligned} \lim_{k\to\infty} \overline{\sum_{\mathscr{P}_k}} &= \text{(L-S)} \int_a^b \bar{g}(x) G(dx), \\ \lim_{k\to\infty} \underline{\sum_{\mathscr{P}_k}} &= \text{(L-S)} \int_a^b \underline{g}(x) G(dx), \end{aligned} \tag{57}$$

where $\bar{g}(x) = \lim_k \bar{g}_{\mathscr{P}_k}(x)$, $\underline{g}(x) = \lim_k \underline{g}_{\mathscr{P}_k}(x)$.

If the limits $\lim_k \overline{\sum}_{\mathscr{P}_k}$ and $\lim_k \underline{\sum}_{\mathscr{P}_k}$ are *finite and equal, and their common value is independent of the sequence of decompositions* $\{\mathscr{P}_k\}$, we say that $g = g(x)$ is *Riemann–Stieltjes integrable*, and the common value of the limits is denoted by

$$\text{(R-S)} \int_a^b g(x) G(dx). \tag{58}$$

When $G(x) = x$, the integral is called a Riemann integral and denoted by

$$\text{(R)} \int_a^b g(x)\, dx.$$

Now let (L-S) $\int_a^b g(x)G(dx)$ be the corresponding Lebesgue–Stieltjes integral (see Remark 2 in Subsection 2).

Theorem 9. *If $g = g(x)$ is continuous on $[a, b]$, it is Riemann–Stieltjes integrable and*

$$\text{(R-S)} \int_a^b g(x)G(dx) = \text{(L-S)} \int_a^b g(x)G(dx). \tag{59}$$

PROOF. Since $g(x)$ is continuous, we have $\bar{g}(x) = g(x) = \underline{g}(x)$. Hence by (57) $\lim_{k\to\infty} \overline{\sum}_{\mathscr{P}_k} = \lim_{k\to\infty} \underline{\sum}_{\mathscr{P}_k}$. Consequently $g = g(x)$ is Riemann–Stieltjes integral (again by (57)).

Let us consider in more detail the question of the correspondence between the Riemann and Lebesgue integrals for the case of Lebesgue measure on the line R.

Theorem 10. *Let $g(x)$ be a bounded function on $[a, b]$.*

(a) *The function $g = g(x)$ is Riemann integrable on $[a, b]$ if and only if it is continuous almost everywhere (with respect to Lebesgue measure $\bar{\lambda}$ on $\overline{\mathscr{B}}([a, b])$).*
(b) *If $g = g(x)$ is Riemann integrable, it is Lebesgue integrable and*

$$\text{(R)} \int_a^b g(x)\,dx = \text{(L)} \int_a^b g(x)\bar{\lambda}(dx). \tag{60}$$

PROOF. (a) Let $g = g(x)$ be Riemann integrable. Then, by (57),

$$\text{(L)} \int_a^b \bar{g}(x)\bar{\lambda}(dx) = \text{(L)} \int_a^b \underline{g}(x)\bar{\lambda}(dx).$$

But $\underline{g}(x) \le g(x) \le \bar{g}(x)$, and hence by Property **H**

$$\underline{g}(x) = g(x) = \bar{g}(x) \quad (\bar{\lambda}\text{-a.s.}), \tag{61}$$

from which it is easy to see that $g(x)$ is continuous almost everywhere (with respect to $\bar{\lambda}$).

Conversely, let $g = g(x)$ be continuous almost everywhere (with respect to $\bar{\lambda}$). Then (61) is satisfied and consequently $g(x)$ differs from the (Borel) measurable function $\bar{g}(x)$ only on a set $\mathscr{N}$ with $\bar{\lambda}(\mathscr{N}) = 0$. But then

$$\begin{aligned}\{x: g(x) \le c\} &= \{x: g(x) \le c\} \cap \overline{\mathscr{N}} + \{x: g(x) \le c\} \cap \mathscr{N} \\ &= \{x: \bar{g}(x) \le c\} \cap \overline{\mathscr{N}} + \{x: g(x) \le c\} \cap \mathscr{N}\end{aligned}$$

It is clear that the set $\{x: \bar{g}(x) \le c\} \cap \overline{\mathscr{N}} \in \overline{\mathscr{B}}([a, b])$, and that

$$\{x: g(x) \le c\} \cap \mathscr{N}$$

is a subset of $\mathcal{N}$ having Lebesgue measure λ equal to zero and therefore also belonging to $\overline{\mathscr{B}}([a, b])$. Therefore $g(x)$ is $\overline{\mathscr{B}}([a, b])$-measurable and, as a bounded function, is Lebesgue integrable. Therefore by Property **G**,

$$(\mathrm{L}) \int_a^b \bar{g}(x)\lambda(dx) = (\mathrm{L}) \int_a^b \underline{g}(x)\lambda(dx) = (\mathrm{L}) \int_a^b g(x)\lambda(dx),$$

which completes the proof of (a).

(b) If $g = g(x)$ is Riemann integrable, then according to (a) it is continuous (λ-a.s.). It was shown above than then $g(x)$ is Lebesgue integrable and its Riemann and Lebesgue integrals are equal.

This completes the proof of the theorem.

Remark. Let μ be a Lebesgue–Stieltjes measure on $\mathscr{B}([a, b])$. Let $\overline{\mathscr{B}}_\mu([a, b])$ be the system consisting of those subsets $\Lambda \subseteq [a, b]$ for which there are sets A and B in $\mathscr{B}([a, b])$ such that $A \subseteq \Lambda \subseteq B$ and $\mu(B \setminus A) = 0$. Let μ be an *extension* of μ to $\overline{\mathscr{B}}_\mu([a, b])$ ($\bar{\mu}(\Lambda) = \mu(A)$ for Λ such that $A \subseteq \Lambda \subseteq B$ and $\mu(B \setminus A) = 0$). Then the conclusion of the theorem remains valid if we consider $\bar{\mu}$ instead of Lebesgue measure λ, and the Riemann–Stieltjes and Lebesgue–Stieltjes measures with respect to $\bar{\mu}$ instead of the Riemann and Lebesgue integrals.

11. In this part we present a useful theorem on integration by parts for the Lebesgue–Stieltjes integral.

Let two generalized distribution functions $F = F(x)$ and $G = G(x)$ be given on $(R, \mathscr{B}(R))$.

Theorem 11. *The following formulas are valid for all real a and b, a < b:*

$$F(b)G(b) - F(a)G(a) = \int_a^b F(s-)dG(s) + \int_a^b G(s)\,dF(s), \tag{62}$$

or equivalently

$$F(b)G(b) - F(a)G(a) = \int_a^b F(s-)dG(s) + \int_a^b G(s-)\,dF(s) + \sum_{a<s\leq b} \Delta F(s) \cdot \Delta G(s), \tag{63}$$

where $F(s-) = \lim_{t \uparrow s} F(t)$, $\Delta F(s) = F(s) - F(s-)$.

Remark 1. Formula (62) can be written symbolically in "differential" form

$$d(FG) = F_-\,dG + G\,dF. \tag{64}$$

Remark 2. The conclusion of the theorem remains valid for functions F and G of bounded variation on $[a, b]$. (Every such function that is continuous on the right and has limits on the left can be represented as the difference of two monotone nondecreasing functions.)

PROOF. We first recall that in accordance with Subsection 1 an integral $\int_a^b (\cdot)$ means $\int_{(a,b]} (\cdot)$. Then (see formula (2) in §3)

$$(F(b) - F(a))(G(b) - G(a)) = \int_a^b dF(s) \cdot \int_a^b dG(t).$$

Let $F \times G$ denote the direct product of the measures corresponding to F and G. Then by Fubini's theorem

$$\begin{aligned}
(F(b) - F(a))(G(b) - G(a)) &= \int_{(a,b] \times (a,b]} d(F \times G)(s, t) \\
&= \int_{(a,b] \times (a,b]} I_{\{s \geq t\}}(s, t)\, d(F \times G)(s, t) + \int_{(a,b] \times (a,b]} I_{\{s < t\}}(s, t)\, d(F \times G)(s, t) \\
&= \int_{(a,b]} (G(s) - G(a))\, dF(s) + \int_{(a,b]} (F(t-) - F(a))\, dG(t) \\
&= \int_a^b G(s)\, dF(s) + \int_a^b F(s-)\, dG(s) - G(a)(F(b) - F(a)) - F(a)(G(b) - G(a)),
\end{aligned} \tag{65}$$

where I_A is the indicator of the set A.

Formula (62) follows immediately from (65). In turn, (63) follows from (62) if we observe that

$$\int_a^b (G(s) - G(s-))\, dF(s) = \sum_{a < s \leq b} \Delta G(s) \cdot \Delta F(s). \tag{66}$$

Corollary 1. *If $F(x)$ and $G(x)$ are distribution functions, then*

$$F(x)G(x) = \int_{-\infty}^x F(s-)\, dG(s) + \int_{-\infty}^x G(s)\, dF(s). \tag{67}$$

If also

$$F(x) = \int_{-\infty}^x f(s)\, ds,$$

then

$$F(x)G(x) = \int_{-\infty}^x F(s)\, dG(s) + \int_{-\infty}^x G(s) f(s)\, ds. \tag{68}$$

Corollary 2. *Let ξ be a random variable with distribution function $F(x)$ and $\mathsf{E}|\xi|^n < \infty$. Then*

$$\int_0^\infty x^n\,dF(x) = n\int_0^\infty x^{n-1}[1 - F(x)]\,dx, \tag{69}$$

$$\int_{-\infty}^0 |x|^n\,dF(x) = -\int_0^\infty x^n\,dF(-x) = n\int_0^\infty x^{n-1}F(-x)\,dx \tag{70}$$

and

$$\mathsf{E}|\xi|^n = \int_{-\infty}^\infty |x|^n\,dF(x) = n\int_0^\infty x^{n-1}[1 - F(x) + F(-x)]\,dx. \tag{71}$$

To prove (69) we observe that

$$\int_0^b x^n\,dF(x) = -\int_0^b x^n\,d(1 - F(x))$$

$$= -b^n(1 - F(b)) + n\int_0^b x^{n-1}(1 - F(x))\,dx. \tag{72}$$

Let us show that since $\mathsf{E}|\xi|^n < \infty$,

$$b^n(1 - F(b) + F(-b)) \le b^n\mathsf{P}(|\xi| \ge b) \to 0. \tag{73}$$

In fact,

$$\mathsf{E}|\xi|^n = \sum_{k=1}^\infty \int_{k-1}^k |x|^n\,dF(x) < \infty$$

and therefore

$$\sum_{k\ge b+1}\int_{k-1}^k |x|^n\,dF(x) \to 0, \qquad n \to \infty.$$

But

$$\sum_{k\ge b+1}\int_{k-1}^k |x|^n\,dF(x) \ge b^n\mathsf{P}(|\xi| \ge b),$$

which establishes (73).

Taking the limit as $b \to \infty$ in (72), we obtain (69).

Formula (70) is proved similarly, and (71) follows from (69) and (70).

12. Let $A(t)$, $t \ge 0$, be a function of locally bounded variation (i.e., of bounded variation on each finite interval $[a, b]$), which is continuous on the right and has limits on the left. Consider the equation

$$Z_t = 1 + \int_0^t Z_{s-}\,dA(s), \tag{74}$$

which can be written in differential form as

$$dZ = Z_-\, dA, \qquad Z_0 = 1. \tag{75}$$

The formula that we have proved for integration by parts lets us solve (74) explicitly in the class of functions of bounded variation.

We introduce the function

$$\mathscr{E}_t(A) = e^{A(t)-A(0)} \prod_{0 \le s \le t} (1 + \Delta A(s))e^{-\Delta A(s)}, \tag{76}$$

where $\Delta A(s) = A(s) - A(s-)$ for $s > 0$, and $\Delta A(0) = 0$.

The function $A(s)$, $0 \le s \le t$, has bounded variation and therefore has at most a countable number of discontinuities, and so the series $\sum_{0 \le s \le t} |\Delta A(s)|$ converges. It follows that

$$\prod_{0 \le s \le t} (1 + \Delta A(s))e^{-\Delta A(s)}$$

is a function of locally bounded variation.

If $A^c(t) = A(t) - \sum_{0 \le s \le t} \Delta A(s)$ is the continuous component of $A(t)$, we can rewrite (76) in the form

$$\mathscr{E}_t(A) = e^{A^c(t)-A^c(0)} \prod_{0 \le s \le t} (1 + \Delta A(s)). \tag{77}$$

Let us write

$$F(t) = e^{A^c(t)-A^c(0)}, \qquad G(t) = \prod_{0 \le s \le t} (1 + \Delta A(s)).$$

Then by (62)

$$\begin{aligned}
\mathscr{E}_t(A) = F(t)G(t) &= 1 + \int_0^t F(s)\, dG(s) + \int_0^t G(s-)\, dF(s) \\
&= 1 + \sum_{0 \le s \le t} F(s)G(s-)\Delta A(s) + \int_0^t G(s-)F(s)\, dA^c(s) \\
&= 1 + \int_0^t \mathscr{E}_{s-}(A)\, dA(s).
\end{aligned}$$

Therefore $\mathscr{E}_t(A)$, $t \ge 0$, is a (locally bounded) solution of (74). Let us show that this is the only locally bounded solution.

Suppose that there are two such solutions and let $Y = Y(t)$, $t \ge 0$, be their difference. Then

$$Y(t) = \int_0^t Y(s-)\, dA(s).$$

Put

$$T = \inf\{t \ge 0 : Y(t) \ne 0\},$$

where we take $T = \infty$ if $Y(t) = 0$ for $t \ge 0$.

Since $A(t)$ is a function of locally bounded variation, there are two generalized distribution functions $A_1(t)$ and $A_2(t)$ such that $A(t) = A_1(t) - A_2(t)$. If we suppose that $T < \infty$, we can find a finite T' such that

$$[A_1(T') + A_2(T')] - [A_1(T) + A_2(T)] \le \tfrac{1}{2}.$$

Then it follows from the equation

$$Y(t) = \int_T^t Y(s-)\, dA(s), \qquad t \ge T,$$

that

$$\sup_{t \le t'} |Y(t)| \le \tfrac{1}{2} \sup_{t \le T'} |Y(t)|$$

and since $\sup|Y(t)| < \infty$, we have $Y(t) = 0$ for $T < t \le T'$, contradicting the assumption that $T < \infty$.

Thus we have proved the following theorem.

Theorem 12. *There is a unique locally bounded solution of* (74), *and it is given by* (76).

13. Problems

1. Establish the representation (6).

2. Prove the following extension of Property **E**. Let ξ and η be random variables for which $\mathsf{E}\xi$ and $\mathsf{E}\eta$ are defined and the sum $\mathsf{E}\xi + \mathsf{E}\eta$ is meaningful (does not have the form $\infty - \infty$ or $-\infty + \infty$). Then

$$\mathsf{E}(\xi + \eta) = \mathsf{E}\xi + \mathsf{E}\eta.$$

3. Generalize Property **G** by showing that if $\xi = \eta$ (a.s.) and $\mathsf{E}\xi$ exists, then $\mathsf{E}\eta$ exists and $\mathsf{E}\xi = \mathsf{E}\eta$.

4. Let ξ be an extended random variable, μ a σ-finite measure, and $\int_\Omega |\xi|\, d\mu < \infty$. Show that $|\xi| < \infty$ (μ-a.s.) (cf. Property **J**).

5. Let μ be a σ-finite measure, ξ and η extended random variables for which $\mathsf{E}\xi$ and $\mathsf{E}\eta$ are defined. If $\int_A \xi\, d\mathsf{P} \le \int_A \eta\, d\mathsf{P}$ for all $A \in \mathscr{F}$, then $\xi \le \eta$ (μ-a.s.). (Cf. Property **I**.)

6. Let ξ and η be independent nonnegative random variables. Show that $\mathsf{E}\xi\eta = \mathsf{E}\xi \cdot \mathsf{E}\eta$.

7. Using Fatou's lemma, show that

$$\mathsf{P}(\underline{\lim}\, A_n) \le \underline{\lim}\, \mathsf{P}(A_n), \qquad \mathsf{P}(\overline{\lim}\, A_n) \ge \overline{\lim}\, \mathsf{P}(A_n).$$

8. Find an example to show that in general it is impossible to weaken the hypothesis "$|\xi_n| \le \eta$, $\mathsf{E}\eta < \infty$" in the dominated convergence theorem.

9. Find an example to show that in general the hypothesis "$\xi_n \le \eta$, $\mathsf{E}\eta > -\infty$" in Fatou's lemma cannot be omitted.

10. Prove the following variants of Fatou's lemma. Let the family $\{\xi_n^+\}_{n\ge 1}$ of random variables be uniformly integrable and let $\mathsf{E}\,\overline{\lim}\,\xi_n$ exist. Then

$$\overline{\lim}\,\mathsf{E}\xi_n \le \mathsf{E}\,\overline{\lim}\,\xi_n.$$

Let $\xi_n \le \eta_n$, $n \ge 1$, where the family $\{\xi_n^+\}_{n\ge 1}$ is uniformly integrable and η_n converges a.s. (or only in probability—see §10 below) to a random variable η. Then $\overline{\lim}\,\mathsf{E}\xi_n \le \mathsf{E}\,\overline{\lim}\,\xi_n$.

11. Dirichlet's function

$$d(x) = \begin{cases} 1, & x \text{ irrational}, \\ 0, & x \text{ rational}, \end{cases}$$

is defined on $[0, 1]$, Lebesgue integrable, but not Riemann integrable. Why?

12. Find an example of a sequence of Riemann integrable functions $\{f_n\}_{n\ge 1}$, defined on $[0, 1]$, such that $|f_n| \le 1$, $f_n \to f$ almost everywhere (with Lebesgue measure), but f is not Riemann integrable.

13. Let $(a_{i,j}; i, j \ge 1)$ be a sequence of real numbers such that $\sum_{i,j} |a_{i,j}| < \infty$. Deduce from Fubini's theorem that

$$\sum_{(i,j)} a_{ij} = \sum_i \left(\sum_j a_{ij} \right) = \sum_j \left(\sum_i a_{ij} \right). \tag{78}$$

14. Find an example of a sequence $(a_{ij}; i, j \ge 1)$ for which $\sum_{i,j} |a_{ij}| = \infty$ and the equation in (78) does not hold.

15. Starting from simple functions and using the theorem on taking limits under the Lebesgue integral sign, prove the following result on *integration by substitution*.

Let $h = h(y)$ be a nondecreasing continuously differentiable function on $[a, b]$, and let $f(x)$ be (Lebesgue) integrable on $[h(a), h(b)]$. Then the function $f(h(y))h'(y)$ is integrable on $[a, b]$ and

$$\int_{h(a)}^{h(b)} f(x)\,dx = \int_a^b f(h(y))h'(y)\,dy.$$

16. Prove formula (70).

17. Let $\xi, \xi_1, \xi_2, \ldots$ be nonnegative integrable random variables such that $\mathsf{E}\xi_n \to \mathsf{E}\xi$ and $\mathsf{P}(\xi - \xi_n > \varepsilon) \to 0$ for every $\varepsilon > 0$. Show that then $\mathsf{E}|\xi_n - \xi| \to 0$, $n \to \infty$.

18. Let ξ, η, ζ and ξ_n, η_n, ζ_n, $n \ge 1$, be random variables such that

$$\xi_n \xrightarrow{\mathsf{P}} \xi, \qquad \eta_n \xrightarrow{\mathsf{P}} \eta, \qquad \zeta_n \xrightarrow{\mathsf{P}} \zeta, \qquad \eta_n \le \xi_n \le \zeta_n, \quad n \ge 1,$$
$$\mathsf{E}\zeta_n \to \mathsf{E}\zeta, \qquad \mathsf{E}\eta_n \to \mathsf{E}\eta,$$

and the expectations $\mathsf{E}\xi$, $\mathsf{E}\eta$, $\mathsf{E}\zeta$ are finite. Show that then $\mathsf{E}\xi_n \to \mathsf{E}\xi$ (*Pratt's lemma*).

If also $\eta_n \le 0 \le \zeta_n$ then $\mathsf{E}|\xi_n - \xi| \to 0$.

Deduce that if $\xi_n \xrightarrow{\mathsf{P}} \xi$, $\mathsf{E}|\xi_n| \to \mathsf{E}|\xi|$ and $\mathsf{E}|\xi| < \infty$, then $\mathsf{E}|\xi_n - \xi| \to 0$.

§7. Conditional Probabilities and Conditional Expectations with Respect to a σ-Algebra

1. Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space, and let $A \in \mathscr{F}$ be an event such that $\mathsf{P}(A) > 0$. As for finite probability spaces, the *conditional probability of B with respect to A* (denoted by $\mathsf{P}(B|A)$) means $\mathsf{P}(BA)/\mathsf{P}(A)$, and the *conditional probability of B with respect to the finite or countable decomposition* $\mathscr{D} = \{D_1, D_2 \ldots\}$ with $\mathsf{P}(D_i) > 0, i \geq 1$ (denoted by $\mathsf{P}(B|\mathscr{D})$) is the random variable equal to $\mathsf{P}(B|D_i)$ for $\omega \in D_i$, $i \geq 1$:

$$\mathsf{P}(B|\mathscr{D}) = \sum_{i \geq 1} \mathsf{P}(B|D_i) I_{D_i}(\omega).$$

In a similar way, if ξ is a random variable for which $\mathsf{E}\xi$ is defined, the *conditional expectation of ξ with respect to the event A* with $\mathsf{P}(A) > 0$ (denoted by $\mathsf{E}(\xi|A)$) is $\mathsf{E}(\xi I_A)/\mathsf{P}(A)$ (cf. (I.8.10)).

The random variable $\mathsf{P}(B|\mathscr{D})$ is evidently measurable with respect to the σ-algebra $\mathscr{G} = \sigma(\mathscr{D})$, and is consequently also denoted by $\mathsf{P}(B|\mathscr{G})$ (see §8 of Chapter I).

However, in probability theory we may have to consider conditional probabilities with respect to events whose probabilities are *zero*.

Consider, for example, the following experiment. Let ξ be a random variable that is uniformly distributed on $[0, 1]$. If $\xi = x$, toss a coin for which the probability of head is x, and the probability of tail is $1 - x$. Let ν be the number of heads in n independent tosses of this coin. What is the "conditional probability $\mathsf{P}(\nu = k|\xi = x)$"? Since $\mathsf{P}(\xi = x) = 0$, the conditional probability $\mathsf{P}(\nu = k|\xi = x)$ is undefined, although it is intuitively plausible that "it ought to be $C_n^k x^k(1 - x)^{n-k}$."

Let us now give a general definition of conditional expectation (and, in particular, of conditional probability) with respect to a σ-algebra $\mathscr{G}$, $\mathscr{G} \subseteq \mathscr{F}$, and compare it with the definition given in §8 of Chapter I for finite probability spaces.

2. Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space, $\mathscr{G}$ a σ-algebra, $\mathscr{G} \subseteq \mathscr{F}$ ($\mathscr{G}$ is a σ-*subalgebra* of $\mathscr{F}$), and $\xi = \xi(\omega)$ a random variable. Recall that, according to §6, the expectation $\mathsf{E}\xi$ was defined in two stages: first for a nonnegative random variable ξ, then in the general case by

$$\mathsf{E}\xi = \mathsf{E}\xi^+ - \mathsf{E}\xi^-,$$

and only under the assumption that

$$\min(\mathsf{E}\xi^-, \mathsf{E}\xi^+) < \infty.$$

A similar two-stage construction is also used to define conditional expectations $\mathsf{E}(\xi|\mathscr{G})$.

Definition 1.

(1) The *conditional expectation of a nonnegative* random variable ξ *with respect to the σ-algebra $\mathscr{G}$* is a nonnegative extended random variable, denoted by $\mathsf{E}(\xi|\mathscr{G})$ or $\mathsf{E}(\xi|\mathscr{G})(\omega)$, such that

(a) $\mathsf{E}(\xi|\mathscr{G})$ is $\mathscr{G}$-measurable;
(b) for every $A \in \mathscr{G}$

$$\int_A \xi\, d\mathsf{P} = \int_A \mathsf{E}(\xi|\mathscr{G})\, d\mathsf{P}. \tag{1}$$

(2) *The conditional expectation* $\mathsf{E}(\xi|\mathscr{G})$, or $\mathsf{E}(\xi|\mathscr{G})(\omega)$, *of any* random variable ξ *with respect to the σ-algebra $\mathscr{G}$*, is considered to be defined if

$$\min(\mathsf{E}(\xi^+|\mathscr{G}), \mathsf{E}(\xi^-|\mathscr{G})) < \infty,$$

P-a.s., and it is given by the formula

$$\mathsf{E}(\xi|\mathscr{G}) \equiv \mathsf{E}(\xi^+|\mathscr{G}) - \mathsf{E}(\xi^-|\mathscr{G}),$$

where, on the set (of probability zero) of sample points for which $\mathsf{E}(\xi^+|\mathscr{G}) = \mathsf{E}(\xi^-|\mathscr{G}) = \infty$, the difference $\mathsf{E}(\xi^+|\mathscr{G}) - \mathsf{E}(\xi^-|\mathscr{G})$ is given an arbitrary value, for example zero.

We begin by showing that, for nonnegative random variables, $\mathsf{E}(\xi|\mathscr{G})$ actually exists. By (6.36) the set function

$$\mathsf{Q}(A) = \int_A \xi\, d\mathsf{P}, \qquad A \in \mathscr{G}, \tag{2}$$

is a measure on $(\Omega, \mathscr{G})$, and is absolutely continuous with respect to P (considered on $(\Omega, \mathscr{G})$, $\mathscr{G} \subseteq \mathscr{F}$). Therefore (by the Radon–Nikodým theorem) there is a nonnegative $\mathscr{G}$-measurable extended random variable $\mathsf{E}(\xi|\mathscr{G})$ such that

$$\mathsf{Q}(A) = \int_A \mathsf{E}(\xi|\mathscr{G})\, d\mathsf{P}. \tag{3}$$

Then (1) follows from (2) and (3).

Remark 1. In accordance with the Radon–Nikodým theorem, the conditional expectation $\mathsf{E}(\xi|\mathscr{G})$ is defined only up to sets of P-measure zero. In other words, $\mathsf{E}(\xi|\mathscr{G})$ can be taken to be any $\mathscr{G}$-measurable function $f(\omega)$ for which $\mathsf{Q}(A) = \int_A f(\omega)\, d\mathsf{P}$, $A \in \mathscr{G}$ (a "variant" of the conditional expectation).

Let us observe that, in accordance with the remark on the Radon–Nikodým theorem,

$$\mathsf{E}(\xi|\mathscr{G}) \equiv \frac{d\mathsf{Q}}{d\mathsf{P}}(\omega), \tag{4}$$

i.e. the conditional expectation is just the derivative of the Radon–Nikodým measure Q with respect to P (considered on $(\Omega, \mathscr{G})$).

Remark 2. In connection with (1), we observe that we cannot in general put $\mathsf{E}(\xi | \mathscr{G}) = \xi$, since ξ is *not necessarily* $\mathscr{G}$-measurable.

Remark 3. Suppose that ξ is a random variable for which $\mathsf{E}\xi$ does not exist. Then $\mathsf{E}(\xi | \mathscr{G})$ may be definable as a $\mathscr{G}$-measurable function for which (1) holds. This is usually just what happens. Our definition $\mathsf{E}(\xi | \mathscr{G}) \equiv \mathsf{E}(\xi^+ | \mathscr{G}) - \mathsf{E}(\xi^- | \mathscr{G})$ has the advantage that for the trivial σ-algebra $\mathscr{G} = \{\varnothing, \Omega\}$ it reduces to the definition of $\mathsf{E}\xi$ but does not presuppose the existence of $\mathsf{E}\xi$. (For example, if ξ is a random variable with $\mathsf{E}\xi^+ = \infty$, $\mathsf{E}\xi^- = \infty$, and $\mathscr{G} = \mathscr{F}$, then $\mathsf{E}\xi$ is not defined, but in terms of Definition 1, $\mathsf{E}(\xi | \mathscr{G})$ exists and is simply $\xi = \xi^+ - \xi^-$.

Remark 4. Let the random variable ξ have a conditional expectation $\mathsf{E}(\xi | \mathscr{G})$ with respect to the σ-algebra $\mathscr{G}$. The *conditional variance* (denoted by $\mathsf{V}(\xi | \mathscr{G})$ or $\mathsf{V}(\xi | \mathscr{G})(\alpha)$) of ξ is the random variable

$$\mathsf{V}(\xi | \mathscr{G}) \equiv \mathsf{E}[(\xi - \mathsf{E}(\xi | \mathscr{G}))^2 | \mathscr{G}].$$

(Cf. the definition of the conditional variance $\mathsf{V}(\xi | \mathscr{D})$ of ξ with respect to a decomposition $\mathscr{D}$, as given in Problem 2, §8, Chapter I.)

Definition 2. Let $B \in \mathscr{F}$. The conditional expectation $\mathsf{E}(I_B | \mathscr{G})$ is denoted by $\mathsf{P}(B | \mathscr{G})$, or $\mathsf{P}(B | \mathscr{G})(\omega)$, and is called the *conditional probability of the event B with respect to the σ-algebra $\mathscr{G}$, $\mathscr{G} \subseteq \mathscr{F}$.*

It follows from Definitions 1 and 2 that, for a given $B \in \mathscr{F}$, $\mathsf{P}(B | \mathscr{G})$ is a random variable such that

(a) $\mathsf{P}(B | \mathscr{G})$ is $\mathscr{G}$-measurable,

(b)
$$\mathsf{P}(A \cap B) = \int_A \mathsf{P}(B | \mathscr{G}) d\mathsf{P} \tag{5}$$
for every $A \in \mathscr{G}$.

Definition 3. Let ξ be a random variable and $\mathscr{G}_\eta$ the σ-algebra generated by a random element η. Then $\mathsf{E}(\xi | \mathscr{G}_\eta)$, if defined, means $\mathsf{E}(\xi | \eta$ or $\mathsf{E}(\xi | \eta)(\omega)$, and is called the *conditional expectation of ξ with respect to η.*

The conditional probability $\mathsf{P}(B | \mathscr{G}_\eta)$ is denoted by $\mathsf{P}(B | \eta)$ or $\mathsf{P}(B | \eta)(\omega)$, and is called the *conditional probability of B with respect to η.*

3. Let us show that the definition of $\mathsf{E}(\xi | \mathscr{G})$ given here agrees with the definition of conditional expectation in §8 of Chapter I.

Let $\mathscr{D} = \{D_1, D_2, \ldots\}$ be a finite or countable decomposition with atoms D_i with respect to the probability P (i.e. $\mathsf{P}(D_i) > 0$, and if $A \subseteq D_i$, then either $\mathsf{P}(A) = 0$ or $\mathsf{P}(D_i \backslash A) = 0$).

Theorem 1. *If $\mathscr{G} = \sigma(\mathscr{D})$ and ξ is a random variable for which $\mathsf{E}\xi$ is defined, then*

$$\mathsf{E}(\xi \mid \mathscr{G}) = \mathsf{E}(\xi \mid D_i) \quad (\mathsf{P}\text{-}a.s.\ on\ D_i) \tag{6}$$

or equivalently

$$\mathsf{E}(\xi \mid \mathscr{G}) = \frac{\mathsf{E}(\xi I_{D_i})}{\mathsf{P}(D_i)} \quad (\mathsf{P}\text{-}a.s.\ on\ D_i).$$

(The notation "$\xi = \eta$ (P-a.s. on A)," or

"$\xi = \eta(A;\ \mathsf{P}\text{-a.s.})$" means that $\mathsf{P}(A \cap \{\xi \neq \eta\}) = 0$.)

PROOF. According to Lemma 3 of §4, $\mathsf{E}(\xi \mid \mathscr{G}) = K_i$ on D_i, where K_i are constants. But

$$\int_{D_i} \xi \, d\mathsf{P} = \int_{D_i} \mathsf{E}(\xi \mid \mathscr{G}) \, d\mathsf{P} = K_i \mathsf{P}(D_i),$$

whence

$$K_i = \frac{1}{\mathsf{P}(D_i)} \int_{D_i} \xi \, d\mathsf{P} = \frac{\mathsf{E}(\xi I_{D_i})}{\mathsf{P}(D_i)} = \mathsf{E}(\xi \mid D_i).$$

This completes the proof of the theorem.

Consequently the concept of the conditional expectation $\mathsf{E}(\xi \mid \mathscr{D})$ with respect to a finite decomposition $\mathscr{D} = \{D_1, \ldots, D_n\}$, as introduced in Chapter I, is a special case of the concept of conditional expectation with respect to the σ-algebra $\mathscr{G} = \sigma(D)$.

4. Properties of conditional expectations. (We shall suppose that the expectations are defined for all the random variables that we consider and that $\mathscr{G} \subseteq \mathscr{F}$.)

A*. *If C is a constant and $\xi = C$ (a.s.), then $\mathsf{E}(\xi \mid \mathscr{G}) = C$ (a.s.).*
B*. *If $\xi \leq \eta$ (a.s.) then $\mathsf{E}(\xi \mid \mathscr{G}) \leq \mathsf{E}(\eta \mid \mathscr{G})$ (a.s.).*
C*. $|\mathsf{E}(\xi \mid \mathscr{G})| \leq \mathsf{E}(|\xi| \mid \mathscr{G})$ *(a.s.).*
D*. *If a, b are constants and $a\mathsf{E}\xi + b\mathsf{E}\eta$ is defined, then*

$$\mathsf{E}(a\xi + b\eta \mid \mathscr{G}) = a\mathsf{E}(\xi \mid \mathscr{G}) + b\mathsf{E}(\eta \mid \mathscr{G}) \quad (a.s.).$$

E*. *Let $\mathscr{F}_* = \{\varphi, \Omega\}$ be the trivial σ-algebra. Then*

$$\mathsf{E}(\xi \mid \mathscr{F}_*) = \mathsf{E}\xi \quad (a.s.).$$

F*. $\mathsf{E}(\xi|\mathscr{F}) = \xi$ *(a.s.)*.

G*. $\mathsf{E}(\mathsf{E}(\xi|\mathscr{G})) = \mathsf{E}\xi$.

H*. *If* $\mathscr{G}_1 \subseteq \mathscr{G}_2$ *then*

$$E[E(\xi|\mathscr{G}_2)|\mathscr{G}_1] = \mathsf{E}(\xi|\mathscr{G}_1) \quad (a.s.).$$

I*. *If* $\mathscr{G}_1 \supseteq \mathscr{G}_2$ *then*

$$\mathsf{E}[\mathsf{E}(\xi|\mathscr{G}_2)|\mathscr{G}_1) = \mathsf{E}(\xi|\mathscr{G}_2) \quad (a.s.).$$

J*. *Let a random variable ξ for which $\mathsf{E}\xi$ is defined be independent of the σ-algebra $\mathscr{G}$ (i.e., independent of I_B, $B \in \mathscr{G}$). Then*

$$\mathsf{E}(\xi|\mathscr{G}) = \mathsf{E}\xi \quad (a.s.).$$

K*. *Let η be a $\mathscr{G}$-measurable random variable, $\mathsf{E}|\xi| < \infty$ and $\mathsf{E}|\xi\eta| < \infty$. Then*

$$\mathsf{E}(\xi\eta|\mathscr{G}) = \eta\mathsf{E}(\xi|\mathscr{G}) \quad (a.s.).$$

Let us establish these properties.

A*. A constant function is measurable with respect to $\mathscr{G}$. Therefore we need only verify that

$$\int_A \xi d\mathsf{P} = \int_A C\, d\mathsf{P}, \qquad A \in \mathscr{G}.$$

But, by the hypothesis $\xi = C$ (a.s.) and Property **G** of §6, this equation is obviously satisfied.

B*. If $\xi \le \eta$ (a.s.), then by Property **B** of §6

$$\int_A \xi\, d\mathsf{P} \le \int_A \eta\, d\mathsf{P}, \qquad A \in \mathscr{G},$$

and therefore

$$\int_A \mathsf{E}(\xi|\mathscr{G})\, d\mathsf{P} \le \int_A \mathsf{E}(\eta|\mathscr{G})\, d\mathsf{P}, \qquad A \in \mathscr{G}.$$

The required inequality now follows from Property **I** (§6).

C*. This follows from the preceding property if we observe that $-|\xi| \le \xi \le |\xi|$.

D*. If $A \in \mathscr{G}$ then by Problem 2 of §6,

$$\int_A (a\xi + b\eta)\, d\mathsf{P} = \int_A a\xi\, d\mathsf{P} + \int_A b\eta\, d\mathsf{P} = \int_A a\mathsf{E}(\xi|\mathscr{G})\, d\mathsf{P}$$
$$+ \int_A b\mathsf{E}(\eta|\mathscr{G})\, d\mathsf{P} = \int_A [a\mathsf{E}(\xi|\mathscr{G}) + b\mathsf{E}(\eta|\mathscr{G})]\, d\mathsf{P},$$

which establishes **D***.

E*. This property follows from the remark that $\mathsf{E}\xi$ is an $\mathscr{F}_*$-measurable function and the evident fact that if $A = \Omega$ or $A = \varnothing$ then

$$\int_A \xi \, d\mathsf{P} = \int_A \mathsf{E}\xi \, d\mathsf{P}.$$

F*. Since ξ if $\mathscr{F}$-measurable and

$$\int_A \xi \, d\mathsf{P} = \int_A \xi \, d\mathsf{P}, \qquad A \in \mathscr{F},$$

we have $\mathsf{E}(\xi | \mathscr{F}) = \xi$ (a.s.).

G*. This follows from **E*** and **H*** by taking $\mathscr{G}_1 = \{\varnothing, \Omega\}$ and $\mathscr{G}_2 = \mathscr{G}$.

H*. Let $A \in \mathscr{G}_1$; then

$$\int_A \mathsf{E}(\xi | \mathscr{G}_1) \, d\mathsf{P} = \int_A \xi \, d\mathsf{P}.$$

Since $\mathscr{G}_1 \subseteq \mathscr{G}_2$, we have $A \in \mathscr{G}_2$ and therefore

$$\int_A \mathsf{E}[\mathsf{E}(\xi | \mathscr{G}_2) | \mathscr{G}_1] \, d\mathsf{P} = \int_A \mathsf{E}(\xi | \mathscr{G}_2) \, d\mathsf{P} = \int_A \xi \, d\mathsf{P}.$$

Consequently, when $A \in \mathscr{G}_1$,

$$\int_A \mathsf{E}(\xi | \mathscr{G}_1) \, d\mathsf{P} = \int_A \mathsf{E}[\mathsf{E}(\xi | \mathscr{G}_2) | \mathscr{G}_1] \, d\mathsf{P}$$

and by Property I (§6) and Problem 5 (§6)

$$\mathsf{E}(\xi | \mathscr{G}_1) = \mathsf{E}[\mathsf{E}(\xi | \mathscr{G}_2) | \mathscr{G}_1] \quad \text{(a.s.)}.$$

I*. If $A \in \mathscr{G}_1$, then by the definition of $\mathsf{E}[\mathsf{E}(\xi | \mathscr{G}_2) | \mathscr{G}_1]$

$$\int_A \mathsf{E}[\mathsf{E}(\xi | \mathscr{G}_2) | \mathscr{G}_1] d\mathsf{P} = \int_A \mathsf{E}(\xi | \mathscr{G}_2) \, d\mathsf{P}.$$

The function $\mathsf{E}(\xi | \mathscr{G}_2)$ is $\mathscr{G}_2$-measurable and, since $\mathscr{G}_2 \subseteq \mathscr{G}_1$, also $\mathscr{G}_1$-measurable. It follows that $\mathsf{E}(\xi | \mathscr{G}_2)$ is a variant of the expectation $\mathsf{E}[\mathsf{E}(\xi | \mathscr{G}_2) | \mathscr{G}_1]$, which proves Property **I***.

J*. Since $\mathsf{E}\xi$ is a $\mathscr{G}$-measurable function, we have only to verify that

$$\int_B \xi \, d\mathsf{P} = \int_B \mathsf{E}\xi \, d\mathsf{P},$$

i.e. that $\mathsf{E}[\xi \cdot I_B] = \mathsf{E}\xi \cdot \mathsf{E}I_B$. If $\mathsf{E}|\xi| < \infty$, this follows immediately from Theorem 6 of §6. The general case can be reduced to this by applying Problem 6 of §6.

The proof of Property **K*** will be given a little later; it depends on conclusion (a) of the following theorem.

Theorem 2 (On Taking Limits Under the Expectation Sign). *Let $\{\xi_n\}_{n\geq 1}$ be a sequence of extended random variables.*

(a) *If $|\xi_n| \leq \eta$, $\mathsf{E}\eta < \infty$ and $\xi_n \to \xi$ (a.s.), then*

$$\mathsf{E}(\xi_n|\mathscr{G}) \to \mathsf{E}(\xi|\mathscr{G}) \quad (a.s.)$$

and

$$\mathsf{E}(|\xi_n - \xi||\mathscr{G}) \to 0 \quad (a.s.).$$

(b) *If $\xi_n \geq \eta$, $\mathsf{E}\eta > -\infty$ and $\xi_n \uparrow \xi$ (a.s.), then*

$$\mathsf{E}(\xi_n|\mathscr{G}) \uparrow \mathsf{E}(\xi|\mathscr{G}) \quad (a.s.).$$

(c) *If $\xi_n \leq \eta$, $\mathsf{E}\eta < \infty$, and $\xi_n \downarrow \xi$ (a.s.), then*

$$\mathsf{E}(\xi_n|\mathscr{G}) \downarrow \mathsf{E}(\xi|\mathscr{G}) \quad (a.s.).$$

(d) *If $\xi_n \geq \eta$, $\mathsf{E}\eta > -\infty$, then*

$$\mathsf{E}(\underline{\lim}\, \xi_n|\mathscr{G}) \leq \underline{\lim}\, \mathsf{E}(\xi_n|\mathscr{G}) \quad \text{(a.s.)}.$$

(e) *If $\xi_n \leq \eta$, $\mathsf{E}\eta < \infty$, then*

$$\overline{\lim}\, \mathsf{E}(\xi_n|\mathscr{G}) \leq \mathsf{E}(\overline{\lim}\, \xi_n|\mathscr{G}) \quad (a.s.).$$

(f) *If $\xi_n \geq 0$ then*

$$\mathsf{E}(\sum \xi_n|\mathscr{G}) = \sum \mathsf{E}(\xi_n|\mathscr{G}) \quad (a.s.).$$

PROOF. (a) Let $\zeta_n = \sup_{m\geq n}|\xi_m - \xi|$. Since $\xi_n \to \xi$ (a.s.), we have $\zeta_n \downarrow 0$ (a.s.). The expectations $\mathsf{E}\xi_n$ and $\mathsf{E}\xi$ are finite; therefore by Properties **D*** and **C*** (a.s.)

$$|\mathsf{E}(\xi_n|\mathscr{G}) - \mathsf{E}(\xi|\mathscr{G})| = |\mathsf{E}(\xi_n - \xi|\mathscr{G})| \leq \mathsf{E}(|\xi_n - \xi||\mathscr{G}) \leq \mathsf{E}(\zeta_n|\mathscr{G}).$$

Since $\mathsf{E}(\zeta_{n+1}|\mathscr{G}) \leq \mathsf{E}(\zeta_n|\mathscr{G})$ (a.s.), the limit $h = \lim_n \mathsf{E}(\zeta_n|\mathscr{G})$ exists (a.s.). Then

$$0 \leq \int_\Omega h\, d\mathsf{P} \leq \int_\Omega \mathsf{E}(\zeta_n|\mathscr{G})\, d\mathsf{P} = \int_\Omega \zeta_n\, d\mathsf{P} \to 0, \qquad n \to \infty,$$

where the last statement follows from the dominated convergence theorem, since $0 \leq \zeta_n \leq 2\eta$, $\mathsf{E}\eta < \infty$. Consequently $\int_\Omega h\, d\mathsf{P} = 0$ and then $h = 0$ (a.s.) by Property **H**.

(b) First let $\eta \equiv 0$. Since $\mathsf{E}(\xi_n|\mathscr{G}) \leq \mathsf{E}(\xi_{n+1}|\mathscr{G})$ (a.s.) the limit $\zeta(\omega) = \lim_n \mathsf{E}(\xi_n|\mathscr{G})$ exists (a.s.). Then by the equation

$$\int_A \xi_n\, d\mathsf{P} = \int_A \mathsf{E}(\xi_n|\mathscr{G})\, d\mathsf{P}, \qquad A \in \mathscr{G},$$

and the theorem on monotone convergence,

$$\int_A \xi\, d\mathsf{P} = \int_A \zeta\, d\mathsf{P}, \qquad A \in \mathscr{G}.$$

Consequently $\xi = \zeta$ (a.s.) by Property I and Problem 5 of §6.

For the proof in the general case, we observe that $0 \leq \xi_n^+ \uparrow \xi^+$, and by what has been proved,

$$\mathsf{E}(\xi_n^+ | \mathscr{G}) \uparrow \mathsf{E}(\xi^+ | \mathscr{G}) \quad \text{(a.s.)}. \tag{7}$$

But $0 \leq \xi_n^- \leq \xi^-$, $\mathsf{E}\xi^- < \infty$, and therefore by (a)

$$\mathsf{E}(\xi_n^- | \mathscr{G}) \to \mathsf{E}(\xi^- | \mathscr{G}),$$

which, with (7), proves (b).

Conclusion (c) follows from (b).

(d) Let $\zeta_n = \inf_{m \geq n} \xi_m$; then $\zeta_n \uparrow \zeta$, where $\zeta = \underline{\lim}\, \xi_n$. According to (b), $\mathsf{E}(\zeta_n | \mathscr{G}) \uparrow \mathsf{E}(\zeta | \mathscr{G})$ (a.s.). Therefore (a.s.) $\mathsf{E}(\underline{\lim}\, \xi_n | \mathscr{G}) = \mathsf{E}(\zeta | \mathscr{G}) = \lim_n \mathsf{E}(\zeta_n | \mathscr{G}) = \underline{\lim}\, \mathsf{E}(\zeta_n | \mathscr{G}) \leq \underline{\lim}\, \mathsf{E}(\xi_n | \mathscr{G})$.

Conclusion (e) follows from (d).

(f) If $\xi_n \geq 0$, by Property **D*** we have

$$\mathsf{E}\left(\sum_{k=1}^n \xi_k | \mathscr{G}\right) = \sum_{k=1}^n \mathsf{E}(\xi_k | \mathscr{G}) \quad \text{(a.s.)}$$

which, with (b), establishes the required result.

This completes the proof of the theorem.

We can now establish Property **K***. Let $\eta = I_B$, $B \in \mathscr{G}$. Then, for every $A \in \mathscr{G}$,

$$\int_A \xi\eta \, d\mathsf{P} = \int_{A \cap B} \xi \, d\mathsf{P} = \int_{A \cap B} \mathsf{E}(\xi | \mathscr{G}) \, d\mathsf{P} = \int_A I_B \mathsf{E}(\xi | \mathscr{G}) \, d\mathsf{P} = \int_A \eta \mathsf{E}(\xi | \mathscr{G}) \, d\mathsf{P}.$$

By the additivity of the Lebesgue integral, the equation

$$\int_A \xi\eta \, d\mathsf{P} = \int_A \eta \mathsf{E}(\xi | \mathscr{G}) \, d\mathsf{P}, \qquad A \in \mathscr{G}, \tag{8}$$

remains valid for the simple random variables $\eta = \sum_{k=1}^n y_k I_{B_k}$, $B_k \in \mathscr{G}$. Therefore, by Property **I** (§6), we have

$$\mathsf{E}(\xi\eta | \mathscr{G}) = \eta \mathsf{E}(\xi | \mathscr{G}) \quad \text{(a.s.)} \tag{9}$$

for these random variables.

Now let η be any $\mathscr{G}$-measurable random variable with $\mathsf{E}|\eta| < \infty$, and let $\{\eta_n\}_{n \geq 1}$ be a sequence of simple $\mathscr{G}$-measurable random variables such that $|\eta_n| \leq \eta$ and $\eta_n \to \eta$. Then by (9)

$$\mathsf{E}(\xi\eta_n | \mathscr{G}) = \eta_n \mathsf{E}(\xi | \mathscr{G}) \quad \text{(a.s.)}.$$

It is clear that $|\xi\eta_n| \leq |\xi\eta|$, where $\mathsf{E}|\xi_\eta| < \infty$. Therefore $\mathsf{E}(\xi\eta_n | \mathscr{G}) \to \mathsf{E}(\xi\eta | \mathscr{G})$ (a.s.) by Property (a). In addition, since $\mathsf{E}|\xi| < \infty$, we have $\mathsf{E}(\xi | \mathscr{G})$ finite (a.s.) (see Property **C*** and Property **J** of §6). Therefore $\eta_n \mathsf{E}(\xi | \mathscr{G}) \to \eta \mathsf{E}(\xi | \mathscr{G})$ (a.s.). (The hypothesis that $\mathsf{E}(\xi | \mathscr{G})$ is finite, almost surely, is essential, since, according to the footnote on p. 172, $0 \cdot \infty = 0$, but if $\eta_n = 1/n$, $\eta \equiv 0$, we have $1/n \cdot \infty \not\to 0 \cdot \infty = 0$.)

5. Here we consider the more detailed structure of conditional expectations $\mathsf{E}(\xi\,|\,\mathscr{G}_\eta)$, which we also denote, as usual, by $\mathsf{E}(\xi\,|\,\eta)$.

Since $\mathsf{E}(\xi\,|\,\eta)$ is a $\mathscr{G}_\eta$-measurable function, then by Theorem 3 of §4 (more precisely, by its obvious modification for extended random variables) there is a Borel function $m = m(y)$ from $\bar{R}$ to $\bar{R}$ such that

$$m(\eta(\omega)) = \mathsf{E}(\xi\,|\,\eta)(\omega) \tag{10}$$

for all $\omega \in \Omega$. We denote this function $m(y)$ by $\mathsf{E}(\xi\,|\,\eta = y)$ and call it the *conditional expectation of ξ with respect to the event $\{\eta = y\}$, or the conditional expectation of ξ under the condition that $\eta = y$.*

Correspondingly we define

$$\int_A \xi\, d\mathsf{P} = \int_A \mathsf{E}(\xi\,|\,\eta)\, d\mathsf{P} = \int_A m(\eta)\, d\mathsf{P}, \qquad A \in \mathscr{G}_\eta. \tag{11}$$

Therefore by Theorem 7 of §6 (on change of variable under the Lebesgue integral sign)

$$\int_{\{\omega:\, \eta \in B\}} m(\eta)\, d\mathsf{P} = \int_B m(y) P_\eta(dy), \qquad B \in \mathscr{B}(\bar{R}), \tag{12}$$

where P_η is the probability distribution of η. Consequently $m = m(y)$ is a Borel function such that

$$\int_{\{\omega:\, \eta \in B\}} \xi\, d\mathsf{P} = \int_B m(y)\, d\mathsf{P}_\eta. \tag{13}$$

for every $B \in \mathscr{B}(R)$.

This remark shows that we can give a different definition of the conditional expectation $\mathsf{E}(\xi\,|\,\eta = y)$.

Definition 4. Let ξ and η be random variables (possible, extended) and let $\mathsf{E}\xi$ be defined. The conditional expectation of the random variable ξ under the condition that $\eta = y$ is any $\mathscr{B}(\bar{R})$-measurable function $m = m(y)$ for which

$$\int_{\{\omega:\, \eta \in B\}} \xi\, dP = \int_B m(y) P_\eta(dy), \qquad B \in \mathscr{B}(\bar{R}). \tag{14}$$

That such a function exists follows again from the Radon–Nikodým theorem if we observe that the set function

$$\mathsf{Q}(B) = \int_{\{\omega:\, \eta \in B\}} \xi\, d\mathsf{P}$$

is a signed measure absolutely continuous with respect to the measure P_η.

Now suppose that $m(y)$ is a conditional expectation in the sense of Definition 4. Then if we again apply the theorem on change of variable under the Lebesgue integral sign, we obtain

$$\int_{\{\omega:\eta\in B\}} \xi = \int_B m(y)P_\eta(dy) = \int_{\{\omega:\eta\in B\}} m(\eta), \qquad B \in \mathscr{B}(\bar{R}).$$

The function $m(\eta)$ is $\mathscr{G}_\eta$-measurable, and the sets $\{\omega: \eta \in B\}$, $B \in \mathscr{B}(\bar{R})$, exhaust the subsets of $\mathscr{G}_\eta$.

Hence it follows that $m(\eta)$ is the expectation $\mathsf{E}(\xi\,|\,\eta)$. Consequently if we know $\mathsf{E}(\xi\,|\,\eta = y)$ we can reconstruct $\mathsf{E}(\xi\,|\,\eta)$, and conversely from $\mathsf{E}(\xi\,|\,\eta)$ we can find $\mathsf{E}(\xi\,|\,\eta = y)$.

From an intuitive point of view, the conditional expectation $\mathsf{E}(\xi\,|\,\eta = y)$ is simpler and more natural than $\mathsf{E}(\xi\,|\,\eta)$. However, $\mathsf{E}(\xi\,|\,\eta)$, considered as a $\mathscr{G}_\eta$-measurable random variable, is more convenient to work with.

Observe that Properties **A***–**K*** above and the conclusions of Theorem 2 can easily be transferred to $\mathsf{E}(\xi\,|\,\eta = y)$ (replacing "almost surely" by "P_η-almost surely"). Thus, for example, Property **K*** transforms as follows: if $\mathsf{E}\,|\xi| < \infty$ and $\mathsf{E}\,|\xi f(\eta)| < \infty$, where $f = f(y)$ is a $\mathscr{B}(\bar{R})$ measurable function, then

$$\mathsf{E}(\xi f(\eta)\,|\,\eta = y) = f(y)\mathsf{E}(\xi\,|\,\eta = y) \quad (\mathsf{P}_\eta\text{-a.s.}). \tag{15}$$

In addition (cf. Property **J***), if ξ and η are independent, then

$$\mathsf{E}(\xi\,|\,\eta = y) = \mathsf{E}\xi \quad (\mathsf{P}_\eta\text{-a.s.}).$$

We also observe that if $B \in \mathscr{B}(R^2)$ and ξ and η are independent, then

$$\mathsf{E}[I_B(\xi, \eta)\,|\,\eta = y] = \mathsf{E}I_B(\xi, y) \quad (\mathsf{P}_\eta\text{-a.s.}), \tag{16}$$

and if $\varphi = \varphi(x, y)$ is a $\mathscr{B}(R^2)$-measurable function such that $\mathsf{E}\,|\varphi(\xi, \eta)| < \infty$, then

$$\mathsf{E}[\varphi(\xi, \eta)\,|\,\eta = y] = \mathsf{E}[\varphi(\xi, y)] \quad (\mathsf{P}_\eta\text{-a.s.}).$$

To prove (16) we make the following observation. If $B = B_1 \times B_2$, the validity of (16) will follow from

$$\int_{\{\omega:\eta\in A\}} I_{B_1\times B_2}(\xi, \eta)\mathsf{P}(d\omega) = \int_{\{y\in A\}} \mathsf{E}I_{B_1\times B_2}(\xi, y)P_\eta(dy).$$

But the left-hand side is $\mathsf{P}\{\xi \in B_1, \eta \in A \cap B_2\}$, and the right-hand side is $\mathsf{P}(\xi \in B_1)\mathsf{P}(\eta \in A \cap B_2)$; their equality follows from the independence of ξ and η. In the general case the proof depends on an application of Theorem 1, §2, on monotone classes (cf. the corresponding part of the proof of Fubini's theorem).

Definition 5. The conditional probability of the event $A \in \mathscr{F}$ under the condition that $\eta = y$ (notation: $\mathsf{P}(A\,|\,\eta = y)$) is $\mathsf{E}(I_A\,|\,\eta = y)$.

It is clear that $\mathsf{P}(A|\eta = y)$ can be defined as the $\mathscr{B}(\bar{R})$-measurable function such that

$$\mathsf{P}(A \cap \{\eta \in B\}) = \int_B \mathsf{P}(A|\eta = y)\mathsf{P}_\eta(dy), \qquad B \in \mathscr{B}(\bar{R}). \tag{17}$$

6. Let us calculate some examples of conditional probabilities and conditional expectations.

Example 1. Let η be a discrete random variable with $\mathsf{P}(\eta = y_k) > 0$, $\sum_{k=1}^{\infty} \mathsf{P}(\eta = y_k) = 1$. Then

$$\mathsf{P}(A|\eta = y_k) = \frac{\mathsf{P}(A \cap \{\eta = y_k\})}{\mathsf{P}(\eta = y_k)}, \qquad k \geq 1.$$

For $y \notin \{y_1, y_2, \ldots\}$ the conditional probability $\mathsf{P}(A|\eta = y)$ can be defined in any way, for example as zero.

If ξ is a random variable for which $\mathsf{E}\xi$ exists, then

$$\mathsf{E}(\xi|\eta = y_k) = \frac{1}{\mathsf{P}(\eta = y_k)} \int_{\{\omega:\eta = y_k\}} \xi \, d\mathsf{P}.$$

When $y \notin \{y_1, y_2, \ldots\}$ the conditional expectation $\mathsf{E}(\xi|\eta = y)$ can be defined in any way (for example, as zero).

Example 2. Let (ξ, η) be a pair of random variables whose distribution has a density $f_{\xi\eta}(x, y)$:

$$\mathsf{P}\{(\xi, \eta) \in B\} = \int_B f_{\xi\eta}(x, y)\, dx\, dy, \qquad B \in \mathscr{B}(R^2).$$

Let $f_\xi(x)$ and $f_\eta(y)$ be the densities of the probability distribution of ξ and η (see (6.46), (6.55) and (6.56)).

Let us put

$$f_{\xi|\eta}(x|y) = \frac{f_{\xi\eta}(x, y)}{f_\eta(y)}, \tag{18}$$

taking $f_{\xi|\eta}(x|y) = 0$ if $f_\eta(y) = 0$.

Then

$$\mathsf{P}(\xi \in C|\eta = y) = \int_C f_{\xi|\eta}(x|y)\, dx, \qquad C \in \mathscr{B}(R), \tag{19}$$

i.e. $f_{\xi|\eta}(x|y)$ is the density of a conditional probability distribution.

In fact, in order to prove (19) it is enough to verify (17) for $B \in \mathscr{B}(R)$, $A = \{\xi \in C\}$. By (6.43), (6.45) and Fubini's theorem,

$$\begin{aligned}\int_B \left[\int_C f_{\xi|\eta}(x|y)\,dx\right] P_\eta(dy) &= \int_B \left[\int_C f_{\xi|\eta}(x|y)\,dx\right] f_\eta(y)\,dy \\ &= \int_{C\times B} f_{\xi|\eta}(x|y) f_\eta(y)\,dx\,dy \\ &= \int_{C\times B} f_{\xi\eta}(x, y)\,dx\,dy \\ &= \mathsf{P}\{(\xi, \eta) \in C \times B\} = \mathsf{P}\{(\xi \in C) \cap (\eta \in B)\},\end{aligned}$$

which proves (17).

In a similar way we can show that if $\mathsf{E}\xi$ exists, then

$$\mathsf{E}(\xi | \eta = y) = \int_{-\infty}^{\infty} x f_{\xi|\eta}(x|y)\,dx. \tag{20}$$

EXAMPLE 3. Let the length of time that a piece of apparatus will continue to operate be described by a nonnegative random variable $\eta = \eta(\omega)$ whose distribution $F_\eta(y)$ has a density $f_\eta(y)$ (naturally, $F_\eta(y) = f_\eta(y) = 0$ for $y < 0$). Find the conditional expectation $\mathsf{E}(\eta - a | \eta \geq a)$, i.e. the average time for which the apparatus will continue to operate on the hypothesis that it has already been operating for time a.

Let $\mathsf{P}(\eta \geq a) > 0$. Then according to the definition (see Subsection 1) and (6.45),

$$\begin{aligned}\mathsf{E}(\eta - a|\eta \geq a) &= \frac{\mathsf{E}[(\eta - a)I_{\{\eta \geq a\}}]}{\mathsf{P}(\eta \geq a)} = \frac{\int_\Omega (\eta - a)I_{\{\eta \geq a\}}\mathsf{P}(d\omega)}{\mathsf{P}(\eta \geq a)} \\ &= \frac{\int_a^\infty (y - a) f_\eta(y)\,dy}{\int_a^\infty f_\eta(y)\,dy}.\end{aligned}$$

It is interesting to observe that if η is exponentially distributed, i.e.

$$f_\eta(y) = \begin{cases}\lambda e^{-\lambda y}, & y \geq 0, \\ 0 & y < 0,\end{cases} \tag{21}$$

then $\mathsf{E}\eta = \mathsf{E}(\eta|\eta \geq 0) = 1/\lambda$ and $\mathsf{E}(\eta - a|\eta \geq a) = 1/\lambda$ for every $a > 0$. In other words, in this case the average time for which the apparatus continues to operate, assuming that it has already operated for time a, is independent of a and simply equals the average time $\mathsf{E}\eta$.

Under the assumption (21) we can find the conditional distribution $\mathsf{P}(\eta - a \leq x|\eta \geq a)$.

We have

$$\begin{aligned}
\mathsf{P}(\eta - a \le x \mid \eta \ge a) &= \frac{\mathsf{P}(a \le \eta \le a + x)}{\mathsf{P}(\eta \ge a)} \\
&= \frac{F_\eta(a + x) - F_\eta(a) + \mathsf{P}(\eta = a)}{1 - F_\eta(a) + \mathsf{P}(\eta = a)} \\
&= \frac{[1 - e^{-\lambda(a+x)}] - [1 - e^{-\lambda a}]}{1 - [1 - e^{-\lambda a}]} \\
&= \frac{e^{-\lambda a}[1 - e^{-\lambda x}]}{e^{-\lambda a}} = 1 - e^{-\lambda x}.
\end{aligned}$$

Therefore the conditional distribution $P(\eta - a \le x \mid \eta \ge a)$ is the same as the unconditional distribution $P(\eta \le x)$. This remarkable property is unique to the exponential distribution: there are no other distributions that have densities and possess the property $P(\eta - a \le x \mid \eta \ge a) = P(\eta \le x)$, $a \ge 0, 0 \le x < \infty$.

EXAMPLE 4 (Buffon's needle). Suppose that we toss a needle of unit length "at random" onto a pair of parallel straight lines, a unit distance apart, in a plane. What is the probability that the needle will intersect at least one of the lines?

To solve this problem we must first define what it means to toss the needle "at random." Let ξ be the distance from the midpoint of the needle to the left-hand line. We shall suppose that ξ is uniformly distributed on $[0, 1]$, and (see Figure 29) that the angle θ is uniformly distributed on $[-\pi/2, \pi/2]$. In addition, we shall assume that ξ and θ are independent.

Let A be the event that the needle intersects one of the lines. It is easy to see that if

$$B = \{(a, x): |a| \le \frac{\pi}{2}, \qquad x \in [0, \tfrac{1}{2}\cos a] \cup [1 - \tfrac{1}{2}\cos a, 1]\},$$

then $A = \{\omega: (\theta, \xi) \in B\}$, and therefore the probability in question is

$$\mathsf{P}(A) = \mathsf{E} I_A(\omega) = \mathsf{E} I_B(\theta(\omega), \xi(\omega)).$$

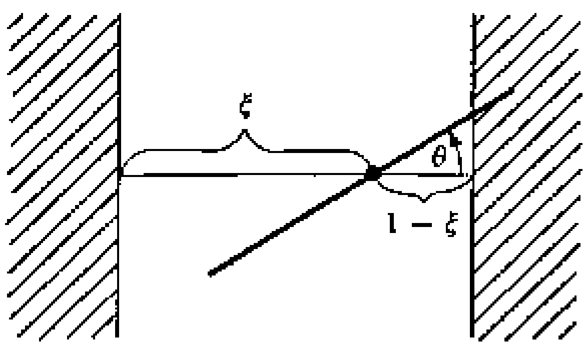


Figure 29

By Property G* and formula (16),

$$\begin{aligned}\mathsf{E}I_B(\theta(\omega), \xi(\omega)) &= \mathsf{E}(\mathsf{E}[I_B(\theta(\omega), \xi(\omega))\,|\,\theta(\omega)]) \\ &= \int_\Omega \mathsf{E}[I_B(\theta(\omega), \xi(\omega))\,|\,\theta(\omega)]\mathsf{P}(d\omega) \\ &= \int_{-\pi/2}^{\pi/2} \mathsf{E}[I_B(\theta(\omega), \xi(\omega))\,|\,\theta(\omega) = \alpha]P_\theta(da) \\ &= \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} \mathsf{E}I_B(a, \xi(\omega))\,da = \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} \cos a\,da = \frac{2}{\pi},\end{aligned}$$

where we have used the fact that

$$\mathsf{E}I_B(a, \xi(\omega)) = \mathsf{P}\{\xi \in [0, \tfrac{1}{2}\cos a] \cup [1 - \tfrac{1}{2}\cos a]\} = \cos a.$$

Thus the probability that a "random" toss of the needle intersects one of the lines is $2/\pi$. This result could be used as the basis for an experimental evaluation of π. In fact, let the needle be tossed N times independently. Define ξ_i to be 1 if the needle intersects a line on the ith toss, and 0 otherwise. Then by the law of large numbers (see, for example, (I.5.6))

$$\mathsf{P}\left\{\left|\frac{\xi_1 + \cdots + \xi_N}{N} - \mathsf{P}(A)\right| > \varepsilon\right\} \to 0, \qquad N \to \infty.$$

for every $\varepsilon > 0$.

In this sense the frequency satisfies

$$\frac{\xi_1 + \cdots + \xi_N}{N} \approx \mathsf{P}(A) = \frac{2}{\pi}$$

and therefore

$$\frac{2N}{\xi_1 + \cdots + \xi_N} \approx \pi.$$

This formula has actually been used for a statistical evaluation of π. In 1850, R. Wolf (an astronomer in Zurich) threw a needle 5000 times and obtained the value 3.1596 for π. Apparently this problem was one of the first applications (now known as Monte Carlo methods) of probabilistic–statistical regularities to numerical analysis.

7. If $\{\xi_n\}_{n\geq 1}$ is a sequence of nonnegative random variables, then according to conclusion (f) of Theorem 2,

$$\mathsf{E}(\sum \xi_n\,|\,\mathscr{G}) = \sum \mathsf{E}(\xi_n\,|\,\mathscr{G}) \quad \text{(a.s.)}$$

In particular, if $B_1, B_2, \ldots$ is a sequence of pairwise disjoint sets,

$$\mathsf{P}(\sum B_n\,|\,\mathscr{G}) = \sum \mathsf{P}(B_n\,|\,\mathscr{G}) \quad \text{(a.s.)} \tag{22}$$

It must be emphasized that this equation is satisfied only almost surely and that consequently the conditional probability $\mathsf{P}(B|\mathscr{G})(\omega)$ cannot be considered as a measure on B for given ω. One might suppose that, except for a set $\mathscr{N}$ of measure zero, $\mathsf{P}(\cdot|\mathscr{G})(\omega)$ would still be a measure for $\omega \in \overline{\mathscr{N}}$. However, in general this is not the case, for the following reason. Let $\mathscr{N}(B_1, B_2, \ldots)$ be the set of sample points ω such that the countable additivity property (22) fails for these $B_1, B_2, \ldots$. Then the excluded set $\mathscr{N}$ is

$$\mathscr{N} = \bigcup \mathscr{N}(B_1, B_2, \ldots), \tag{23}$$

where the union is taken over all $B_1, B_2, \ldots$ in $\mathscr{F}$. Although the P-measure of each set $\mathscr{N}(B_1, B_2, \ldots)$ is zero, the P-measure of $\mathscr{N}$ can be different from zero (because of an uncountable union in (23)). (Recall that the Lebesgue measure of a single point is zero, but the measure of the set $\mathscr{N} = [0, 1]$, which is an uncountable sum of the individual points $\{x\}$, is 1).

However, it would be convenient if the conditional probability $\mathsf{P}(\cdot|\mathscr{G})(\omega)$ were a measure for each $\omega \in \Omega$, since then, for example, the calculation of conditional probabilities $\mathsf{E}(\xi|\mathscr{G})$ could be carried out (see Theorem 3 below) in a simple way by averaging with respect to the measure $\mathsf{P}(\cdot|\mathscr{G})(\omega)$:

$$\mathsf{E}(\xi|\mathscr{G}) = \int_\Omega \xi(\omega)\mathsf{P}(d\omega|\mathscr{G}) \quad \text{(a.s.)}$$

(cf. (I.8.10)).

We introduce the following definition.

Definition 6. A function $\mathsf{P}(\omega; B)$, defined for all $\omega \in \Omega$ and $B \in \mathscr{F}$, is a *regular* conditional probability with respect to $\mathscr{G}$ if

(a) $\mathsf{P}(\omega; \cdot)$ is a probability measure on $\mathscr{F}$ for every $\omega \in \Omega$;
(b) For each $B \in \mathscr{F}$ the function $\mathsf{P}(\omega; B)$, as a function of ω, is a variant of the conditional probability $\mathsf{P}(B|\mathscr{G})(\omega)$, i.e. $\mathsf{P}(\omega: B) = \mathsf{P}(B|\mathscr{G})(\omega)$ (a.s.).

Theorem 3. *Let $P(\omega; B)$ be a regular conditional probability with respect to $\mathscr{G}$ and let ξ be an integrable random variable. Then*

$$\mathsf{E}(\xi|\mathscr{G})(\omega) = \int_\Omega \xi(\tilde{\omega})P(\omega; d\tilde{\omega}) \quad (a.s.). \tag{24}$$

PROOF. If $\xi = I_B$, $B \in \mathscr{F}$, the required formula (24) becomes

$$\mathsf{P}(B|\mathscr{G})(\omega) = \mathsf{P}(\omega; B) \quad \text{(a.s.)},$$

which holds by Definition 6(b). Consequently (24) holds for simple functions.

Now let $\xi \geq 0$ and $\xi_n \uparrow \xi$, where ξ_n are simple functions. Then by (b) of Theorem 2 we have $\mathsf{E}(\xi|\mathscr{G})(\omega) = \lim_n \mathsf{E}(\xi_n|\mathscr{G})(\omega)$ (a.s.). But since $P(\omega; \cdot)$ is a measure for every $\omega \in \Omega$, we have

$$\lim_n \mathsf{E}(\xi_n|\mathscr{G})(\omega) = \lim_n \int_\Omega \xi_n(\tilde{\omega})P(\omega; d\tilde{\omega}) = \int_\Omega \xi(\tilde{\omega})P(\omega; d\tilde{\omega})$$

by the monotone convergence theorem.

The general case reduces to this one if we use the representation $\xi = \xi^+ - \xi^-$.

This completes the proof.

Corollary. *Let $\mathscr{G} = \mathscr{G}_\eta$, where η is a random variable, and let the pair (ξ, η) have a probability distribution with density $f_{\xi\eta}(x, y)$. Let $\mathsf{E}|g(\xi)| < \infty$. Then*

$$\mathsf{E}(g(\xi)|\eta = y) = \int_{-\infty}^{\infty} g(x) f_{\xi|\eta}(x|y)\, dx,$$

where $f_{\xi|\eta}(x|y)$ is the density of the conditional distribution (see (18)).

In order to be able to state the basic result on the existence of regular conditional probabilities, we need the following definitions.

Definition 7. Let $(E, \mathscr{E})$ be a measurable space, $X = X(\omega)$ a random element with values in E, and $\mathscr{G}$ a σ-subalgebra of $\mathscr{F}$. A function $Q(\omega; B)$, defined for $\omega \in \Omega$ and $B \in \mathscr{E}$ is a *regular conditional distribution of X with respect to $\mathscr{G}$* if

(a) for each $\omega \in \Omega$ the function $Q(\omega; B)$ is a probability measure on $(E, \mathscr{E})$;
(b) for each $B \in \mathscr{E}$ the function $Q(\omega; B)$, as a function of ω, is a variant of the conditional probability $\mathsf{P}(X \in B|\mathscr{G})(\omega)$, i.e.

$$Q(\omega; B) = \mathsf{P}(X \in B|\mathscr{G})(\omega) \quad \text{(a.s.)}.$$

Definition 8. Let ξ be a random variable. A function $F = F(\omega; x)$, $\omega \in \Omega$, $x \in R$, is a *regular distribution function for ξ with respect to $\mathscr{G}$ if* :

(a) $F(\omega; x)$ is, for each $\omega \in \Omega$, a distribution function on R;
(b) $F(\omega; x) = \mathsf{P}(\xi \leq x|\mathscr{G})(\omega)$ (a.s.), for each $x \in R$.

Theorem 4. *A regular distribution function and a regular conditional distribution function always exist for the random variable ξ with respect to $\mathscr{G}$.*

PROOF. For each rational number $r \in R$, define $F_r(\omega) = \mathsf{P}(\xi \le r|\mathscr{G})(\omega)$, where $\mathsf{P}(\xi \le r|\mathscr{G})(\omega) = \mathsf{E}(I_{\{\xi \le r\}}|\mathscr{G})(\omega)$ is any variant of the conditional probability, with respect to $\mathscr{G}$, of the event $\{\xi \le r\}$. Let $\{r_i\}$ be the set of rational numbers in R. If $r_i < r_j$, Property $\mathbf{B}^*$ implies that $\mathsf{P}(\xi \le r_i|\mathscr{G}) \le \mathsf{P}(\xi \le r_j|\mathscr{G})$ (a.s.), and therefore if $A_{ij} = \{\omega: F_{r_j}(\omega) < F_{r_i}(\omega)\}$, $A = \bigcup A_{ij}$, we have $\mathsf{P}(A) = 0$. In other words, the set of points ω at which the distribution function $F_r(\omega)$, $r \in \{r_i\}$, fails to be monotonic has measure zero.

Now let

$$B_i = \left\{\omega: \lim_{n\to\infty} F_{r_i+(1/n)}(\omega) \ne F_{r_i}(\omega)\right\}, \qquad B = \bigcup_{i=1}^{\infty} B_i.$$

It is clear that $I_{\{\xi \le r_i+(1/n)\}} \downarrow I_{\{\xi \le r_i\}}$, $n \to \infty$. Therefore, by (a) of Theorem 2, $F_{r_i+(1/n)}(\omega) \to F_{r_i}(\omega)$ (a.s.), and therefore the set B on which continuity on the right fails (with respect to the rational numbers) also has measure zero, $\mathsf{P}(B) = 0$.

In addition, let

$$C = \left\{\omega: \lim_{n\to\infty} F_n(\omega) \ne 1\right\} \cup \left\{\omega: \lim_{n\to-\infty} F_n(\omega) > 0\right\}.$$

Then, since $\{\xi \le n\} \uparrow \Omega$, $n \to \infty$, and $\{\xi \le n\} \downarrow \varnothing$, $n \to -\infty$, we have $\mathsf{P}(C) = 0$.

Now put

$$F(\omega; x) = \begin{cases} \lim_{r \downarrow x} F_r(\omega), & \omega \notin A \cup B \cup C, \\ G(x), & \omega \in A \cup B \cup C, \end{cases}$$

where $G(x)$ is any distribution function on R; we show that $F(\omega; x)$ satisfies the conditions of Definition 8.

Let $\omega \notin A \cup B \cup C$. Then it is clear that $F(\omega; x)$ is a nondecreasing function of x. If $x < x' \le r$, then $F(\omega; x) \le F(\omega; x') \le F(\omega; r) = F_r(\omega) \downarrow F(\omega, x)$ when $r \downarrow x$. Consequently $F(\omega; x)$ is continuous on the right. Similarly $\lim_{x\to\infty} F(\omega; x) = 1$, $\lim_{x\to-\infty} F(\omega; x) = 0$. Since $F(\omega; x) = G(x)$ when $\omega \in A \cup B \cup C$, it follows that $F(\omega; x)$ is a distribution function on R for every $\omega \in \Omega$, i.e. condition (a) of Definition 8 is satisfied.

By construction, $\mathsf{P}(\xi \le r)|\mathscr{G})(\omega) = F_r(\omega) = F(\omega; r)$. If $r \downarrow x$, we have $F(\omega; r) \downarrow F(\omega; x)$ for all $\omega \in \Omega$ by the continuity on the right that we just established. But by conclusion (a) of Theorem 2, we have $\mathsf{P}(\xi \le r|\mathscr{G})(\omega) \to \mathsf{P}(\xi \le x|\mathscr{G})(\omega)$ (a.s.). Therefore $F(\omega; x) = \mathsf{P}(\xi \le x|G)(\omega)$ (a.s.), which establishes condition (b) of Definition 8.

We now turn to the proof of the existence of a regular conditional distribution of ξ with respect to $\mathscr{G}$.

Let $F(\omega; x)$ be the function constructed above. Put

$$Q(\omega; B) = \int_B F(\omega; dx),$$

where the integral is a Lebesgue–Stieltjes integral. From the properties of the integral (see §6, Subsection 7), it follows that $Q(\omega; B)$ is a measure on B for each given $\omega \in \Omega$. To establish that $Q(\omega; B)$ is a variant of the conditional probability $\mathsf{P}(\xi \in B \mid \mathscr{G})(\omega)$, we use the principle of appropriate sets.

Let $\mathscr{C}$ be the collection of sets B in $\mathscr{B}(R)$ for which $Q(\omega; B) = \mathsf{P}(\xi \in B \mid \mathscr{G})(\omega)$ (a.s.). Since $F(\omega; x) = \mathsf{P}(\xi \leq x \mid \mathscr{G})(\omega)$ (a.s.), the system $\mathscr{C}$ contains the sets B of the form $B = (-\infty, x]$, $x \in R$. Therefore $\mathscr{C}$ also contains the intervals of the form $(a, b]$, and the algebra $\mathscr{A}$ consisting of finite sums of disjoint sets of the form $(a, b]$. Then it follows from the continuity properties of $Q(\omega; B)$ (ω fixed) and from conclusion (b) of Theorem 2 that $\mathscr{C}$ is a monotone class, and since $\mathscr{A} \subseteq \mathscr{C} \subseteq \mathscr{B}(R)$, we have, from Theorem 1 of §2,

$$\mathscr{B}(R) = \sigma(\mathscr{A}) \subseteq \sigma(\mathscr{C}) = \mu(\mathscr{C}) = \mathscr{C} \subseteq \mathscr{B}(R),$$

whence $\mathscr{C} = \mathscr{B}(R)$.

This completes the proof of the theorem.

By using topological considerations we can extend the conclusion of Theorem 4 on the existence of a regular conditional distribution to random elements with values in what are known as Borel spaces. We need the following definition.

Definition 9. A measurable space $(E, \mathscr{E})$ is a *Borel space* if it is Borel equivalent to a Borel subset of the real line, i.e. there is a one-to-one mapping $\varphi = \varphi(e)$: $(E, \mathscr{E}) \to (R, \mathscr{B}(R))$ such that

(1) $\varphi(E) \equiv \{\varphi(e): e \in E\}$ is a set in $\mathscr{B}(R)$;
(2) φ is $\mathscr{E}$-measurable ($\varphi^{-1}(A) \in \mathscr{E}$, $A \in \varphi(E) \cap \mathscr{B}(R)$),
(3) φ^{-1} *is* $\mathscr{B}(R)/\mathscr{E}$-measurable ($\varphi(B) \in \varphi(E) \cap \mathscr{B}(R)$, $B \in \mathscr{E}$).

Theorem 5. *Let $X = X(\omega)$ be a random element with values in the Borel space $(E, \mathscr{E})$. Then there is a regular conditional distribution of X with respect to $\mathscr{G}$.*

Proof. Let $\varphi = \varphi(e)$ be the function in Definition 9. By (2), $\varphi(X(\omega))$ is a random variable. Hence, by Theorem 4, we can define the conditional distribution $Q(\omega; A)$ of $\varphi(X(\omega))$ with respect to $\mathscr{G}$, $A \in \varphi(E) \cap \mathscr{B}(R)$.

We introduce the function $\tilde{Q}(\omega; B) = Q(\omega; \varphi(B))$, $B \in \mathscr{E}$. By (3) of Definition 9, $\varphi(B) \in \varphi(E) \cap \mathscr{B}(R)$ and consequently $\tilde{Q}(\omega; B)$ is defined. Evidently $\tilde{Q}(\omega; B)$ is a measure on $B \in \mathscr{E}$ for every ω. Now fix $B \in \mathscr{E}$. By the one-to-one character of the mapping $\varphi = \varphi(e)$,

$$\tilde{Q}(\omega; B) = Q(\omega; \varphi(B)) = \mathsf{P}\{\varphi(X) \in \varphi(B) \mid \mathscr{G}\} = \mathsf{P}\{X \in B \mid \mathscr{G}\} \quad \text{(a.s.)}.$$

Therefore $\tilde{Q}(\omega; B)$ is a regular conditional distribution of X with respect to $\mathscr{G}$.

This completes the proof of the theorem.

Corollary. *Let $X = X(\omega)$ be a random element with values in a complete separable metric space $(E, \mathscr{E})$. Then there is a regular conditional distribution of X with respect to $\mathscr{G}$. In particular, such a distribution exists for the spaces $(R^n, \mathscr{B}(R^n))$ and $(R^\infty, \mathscr{B}(R^\infty))$.*

The proof follows from Theorem 5 and the well known topological result that such spaces are Borel spaces.

8. The theory of conditional expectations developed above makes it possible to give a generalization of Bayes's theorem; this has applications in statistics.

Recall that if $\mathscr{D} = \{A_1, \ldots, A_n\}$ is a partition of the space Ω with $\mathsf{P}(A_i) > 0$, Bayes's theorem (I.3.9) states that

$$\mathsf{P}(A_i \mid B) = \frac{\mathsf{P}(A_i)\mathsf{P}(B \mid A_i)}{\sum_{j=1}^n \mathsf{P}(A_j)\mathsf{P}(B \mid A_j)}. \tag{25}$$

for every B with $\mathsf{P}(B) > 0$. Therefore if $\theta = \sum_{i=1}^n a_i I_{A_i}$ is a discrete random variable then, according to (I.8.10),

$$\mathsf{E}[g(\theta) \mid B] = \frac{\sum_{i=1}^n g(a_i)\mathsf{P}(A_i)\mathsf{P}(B \mid A_i)}{\sum_{j=1}^n \mathsf{P}(A_j)\mathsf{P}(B \mid A_j)}, \tag{26}$$

or

$$\mathsf{E}[g(\theta) \mid B] = \frac{\int_{-\infty}^{\infty} g(a)\mathsf{P}(B \mid \theta = a)P_\theta(da)}{\int_{-\infty}^{\infty} \mathsf{P}(B \mid \theta = a)P_\theta(da)}. \tag{27}$$

On the basis of the definition of $\mathsf{E}[g(\theta) \mid B]$ given at the beginning of this section, it is easy to establish that (27) holds for all events B with $\mathsf{P}(B) > 0$, random variables θ and functions $g = g(a)$ with $\mathsf{E}\,|g(\theta)| < \infty$.

We now consider an analog of (27) for conditional expectations $\mathsf{E}[g(\theta) \mid \mathscr{G}]$ with respect to a σ-algebra $\mathscr{G}$, $\mathscr{G} \subseteq \mathscr{F}$.

Let

$$\mathsf{Q}(B) = \int_B g(\theta)\mathsf{P}(d\omega), \qquad B \in \mathscr{G}. \tag{28}$$

Then by (4)

$$\mathsf{E}[g(\theta) \mid \mathscr{G}] = \frac{d\mathsf{Q}}{d\mathsf{P}}(\omega). \tag{29}$$

We also consider the σ-algebra $\mathscr{G}_\theta$. Then, by (5),

$$\mathsf{P}(B) = \int_\Omega \mathsf{P}(B \mid \mathscr{G}_\theta)\, d\mathsf{P} \tag{30}$$

or, by the formula for change of variable in Lebesgue integrals,

$$\mathsf{P}(B) = \int_{-\infty}^{\infty} \mathsf{P}(B \mid \theta = a)\mathsf{P}_\theta(da). \tag{31}$$

Since

$$\mathsf{Q}(B) = \mathsf{E}[g(\theta)I_B] = \mathsf{E}[g(\theta)\cdot \mathsf{E}(I_B|\mathscr{G}_\theta)],$$

we have

$$\mathsf{Q}(B) = \int_{-\infty}^{\infty} g(a)\mathsf{P}(B|\theta = a)\mathsf{P}_\theta(da). \tag{32}$$

Now suppose that the conditional probability $\mathsf{P}(B|\theta = a)$ is regular and admits the representation

$$\mathsf{P}(B|\theta = a) = \int_B \rho(\omega; a)\lambda(d\omega), \tag{33}$$

where $\rho = \rho(\omega; a)$ is nonnegative and measurable in the two variables jointly, and λ is a σ-finite measure on $(\Omega, \mathscr{G})$.

Let $\mathsf{E}|g(\theta)| < \infty$. Let us show that (P-a.s.)

$$\mathsf{E}[g(\theta)|\mathscr{G}] = \frac{\int_{-\infty}^{\infty} g(a)\rho(\omega; a)\mathsf{P}_\theta(da)}{\int_{-\infty}^{\infty} \rho(\omega; a)P_\theta(da)} \tag{34}$$

(*generalized Bayes theorem*).

In proving (34) we shall need the following lemma.

Lemma. *Let $(\Omega, \mathscr{F})$ be a measurable space.*

(a) *Let μ and λ be σ-finite measures, and $f = f(\omega)$ an $\mathscr{F}$-measurable function. Then*

$$\int_\Omega f\, d\mu = \int_\Omega f \frac{d\mu}{d\lambda}\cdot d\lambda \tag{35}$$

(in the sense that if either integral exists, the other exists and they are equal).

(b) *If ν is a signed measure and μ, λ are σ-finite measures $\nu \ll \mu$, $\mu \ll \lambda$, then*

$$\frac{d\nu}{d\lambda} = \frac{d\nu}{d\mu}\cdot\frac{d\mu}{d\lambda} \quad (\lambda\text{-a.s.}) \tag{36}$$

and

$$\frac{d\nu}{d\mu} = \frac{d\nu}{d\lambda}\Big/\frac{d\mu}{d\lambda} \quad (\mu\text{-a.s.}) \tag{37}$$

PROOF. (a) Since

$$\mu(A) = \int_A \left(\frac{d\mu}{d\lambda}\right) d\lambda, \qquad A \in \mathscr{F},$$

(35) is evidently satisfied for simple functions $f = \sum f_i I_{A_i}$. The general case follows from the representation $f = f^+ - f^-$ and the monotone convergence theorem.

(b) From (a) with $f = d\nu/d\mu$ we obtain

$$\nu(A) = \int_A \left(\frac{d\nu}{d\mu}\right) d\mu = \int_A \left(\frac{d\nu}{d\mu}\right) \cdot \left(\frac{d\mu}{d\lambda}\right) \cdot d\lambda.$$

Then $\nu \ll \lambda$ and therefore

$$\nu(A) = \int_A \frac{d\nu}{d\lambda}\, d\lambda,$$

whence (36) follows since A is arbitrary, by Property **I** (§6).

Property (37) follows from (36) and the remark that

$$\mu\left\{\omega: \frac{d\mu}{d\lambda} = 0\right\} = \int_{\{\omega:\, d\mu/d\lambda = 0\}} \frac{d\mu}{d\lambda}\, d\lambda = 0$$

(on the set $\{\omega: d\mu/d\lambda = 0\}$ the right-hand side of (37) can be defined arbitrarily, for example as zero). This completes the proof of the lemma.

To prove (34) we observe that by Fubini's theorem and (33),

$$\mathsf{Q}(B) = \int_B \left[\int_{-\infty}^{\infty} g(a)\rho(\omega; a)P_\theta(da)\right]\lambda(d\omega), \tag{38}$$

$$\mathsf{P}(B) = \int_B \left[\int_{-\infty}^{\infty} \rho(\omega; a)P_\theta(da)\right]\lambda(d\omega). \tag{39}$$

Then by the lemma

$$\frac{d\mathsf{Q}}{d\mathsf{P}} = \frac{d\mathsf{Q}/d\lambda}{d\mathsf{P}/d\lambda} \quad (\mathsf{P}\text{-a.s.}).$$

Taking account of (38), (39) and (29), we have (34).

Remark. Formula (34) remains valid if we replace θ by a random element with values in some measurable space $(E, \mathscr{E})$ (and replace integration over R by integration over E).

Let us consider some special cases of (34).

Let the σ-algebra $\mathscr{G}$ be generated by the random variable ξ, $\mathscr{G} = \mathscr{G}_\xi$.

Suppose that

$$\mathsf{P}(\xi \in A \mid \theta = a) = \int_A q(x; a)\lambda(dx), \qquad A \in \mathscr{B}(R), \tag{40}$$

where $q = q(x; a)$ is a nonnegative function, measurable with respect to both variables jointly, and λ is a σ-finite measure on $(R, \mathscr{B}(R))$. Then we obtain

$$\mathsf{E}[g(\theta) \mid \xi = x] = \frac{\int_{-\infty}^{\infty} g(a)q(x; a)P_\theta(da)}{\int_{-\infty}^{\infty} q(x; a)P_\theta(da)}. \tag{41}$$

In particular, let (θ, ξ) be a pair of discrete random variables, $\theta = \sum a_i I_{A_i}$, $\xi = \sum x_j I_{B_j}$. Then, taking λ to be the *counting* measure ($\lambda(\{x_i\}) = 1$, $i = 1, 2, \ldots$) we find from (40) that

$$\mathsf{E}[g(\theta) \mid \xi = x_j] = \frac{\sum_i g(a_i) \mathsf{P}(\xi = x_j \mid \theta = a_i) \mathsf{P}(\theta = a_i)}{\sum_i \mathsf{P}(\xi = x_j \mid \theta = a_i) \mathsf{P}(\theta = a_i)}. \tag{42}$$

(Compare (26).)

Now let (θ, ξ) be a pair of absolutely continuous measures with density $f_{\theta,\xi}(a, x)$. Then by (19) the representation (40) applies with $q(x; a) = f_{\xi|\theta}(x \mid a)$ and Lebesgue measure λ. Therefore

$$\mathsf{E}[g(\theta) \mid \xi = x] = \frac{\int_{-\infty}^{\infty} g(a) f_{\xi|\theta}(x \mid a) f_\theta(a)\, da}{\int_{-\infty}^{\infty} f_{\xi|\theta}(x \mid a) f_\theta(a)\, da}. \tag{43}$$

9. Problems

1. Let ξ and η be independent identically distributed random variables with $\mathsf{E}\xi$ defined. Show that

$$\mathsf{E}(\xi \mid \xi + \eta) = \mathsf{E}(\eta \mid \xi + \eta) = \frac{\xi + \eta}{2} \quad \text{(a.s.)}.$$

2. Let $\xi_1, \xi_2, \ldots$ be independent identically distributed random variables with $\mathsf{E}|\xi_i| < \infty$. Show that

$$\mathsf{E}(\xi_1 \mid S_n, S_{n+1}, \ldots) = \frac{S_n}{n} \quad \text{(a.s.)},$$

where $S_n = \xi_1 + \cdots + \xi_n$.

3. Suppose that the random elements (X, Y) are such that there is a regular distribution $P_x(B) = \mathsf{P}(Y \in B \mid X = x)$. Show that if $\mathsf{E}|g(X, Y)| < \infty$ then

$$\mathsf{E}[g(X, Y) \mid X = x] = \int g(x, y) P_x(dy) \quad (P_x\text{-a.s.}).$$

4. Let ξ be a random variable with distribution function $F_\xi(x)$. Show that

$$\mathsf{E}(\xi \mid a < \xi \le b) = \frac{\int_a^b x\, dF_\xi(x)}{F_\xi(b) - F_\xi(a)}$$

(assuming that $F_\xi(b) - F_\xi(a) > 0$).

5. Let $g = g(x)$ be a convex Borel function with $\mathsf{E}|g(\xi)| < \infty$. Show that Jensen's inequality

$$g(\mathsf{E}(\xi \mid \mathscr{G})) \le \mathsf{E}(g(\xi) \mid \mathscr{G})$$

holds for the conditional expectations.

6. Show that a necessary and sufficient condition for the random variable ξ and the σ-algebra $\mathscr{G}$ to be independent (i.e., the random variables ξ and $I_B(\omega)$ are independent for every $B \in \mathscr{G}$) is that $\mathsf{E}(g(\xi)|\mathscr{G}) = \mathsf{E}g(\xi)$ for every Borel function $g(x)$ with $\mathsf{E}|g(\xi)| < \infty$.

7. Let ξ be a nonnegative random variable and $\mathscr{G}$ a σ-algebra, $\mathscr{G} \subseteq \mathscr{F}$. Show that $\mathsf{E}(\xi|\mathscr{G}) < \infty$ (a.s.) if and only if the measure Q, defined on sets $A \in \mathscr{G}$ by $\mathsf{Q}(A) = \int_A \xi\, d\mathsf{P}$, is σ-finite.

§8. Random Variables. II

1. In the first chapter we introduced characteristics of simple random variables, such as the variance, covariance, and correlation coefficient. These extend similarly to the general case. Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space and $\xi = \xi(\omega)$ a random variable for which $\mathsf{E}\xi$ is defined.

The variance of ξ is

$$\mathsf{V}\xi = \mathsf{E}(\xi - \mathsf{E}\xi)^2.$$

The number $\sigma = +\sqrt{\mathsf{V}\xi}$ is the *standard deviation*.

If ξ is a random variable with a Gaussian (normal) density

$$f_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-[(x-m)^2]/2\sigma^2}, \qquad \sigma > 0, \quad -\infty < m < \infty, \tag{1}$$

the parameters m and σ in (1) are very simple:

$$m = \mathsf{E}\xi, \qquad \sigma^2 = \mathsf{V}\xi.$$

Hence the probability distribution of this random variable ξ, which we call *Gaussian*, or *normally distributed*, is completely determined by its mean value m and variance σ^2. (It is often convenient to write $\xi \sim \mathscr{N}(m, \sigma^2)$.)

Now let (ξ, η) be a pair of random variables. Their covariance is

$$\operatorname{cov}(\xi, \eta) = \mathsf{E}(\xi - \mathsf{E}\xi)(\eta - \mathsf{E}\eta) \tag{2}$$

(assuming that the expectations are defined).

If $\operatorname{cov}(\xi, \eta) = 0$ we say that ξ and η are *uncorrelated*.

If $\mathsf{V}\xi > 0$ and $\mathsf{V}\eta > 0$, the number

$$\rho(\xi, \eta) \equiv \frac{\operatorname{cov}(\xi, \eta)}{\sqrt{\mathsf{V}\xi \cdot \mathsf{V}\eta}} \tag{3}$$

is the *correlation coefficient* of ξ and η.

The properties of variance, covariance, and correlation coefficient were investigated in §4 of Chapter I for simple random variables. In the general case these properties can be stated in a completely analogous way.

Let $\xi = (\xi_1, \ldots, \xi_n)$ be a random vector whose components have finite second moments. The *covariance matrix* of ξ is the $n \times n$ matrix $\mathbb{R} = \|R_{ij}\|$, where $R_{ij} = \operatorname{cov}(\xi_i, \xi_j)$. It is clear that $\mathbb{R}$ is *symmetric*. Moreover, it is *nonnegative definite*, i.e.

$$\sum_{i,j=1}^{n} R_{ij}\lambda_i\lambda_j \geq 0$$

for all $\lambda_i \in R$, $i = 1, \ldots, n$, since

$$\sum_{i,j}^{n} R_{ij}\lambda_i\lambda_j = \mathsf{E}\left[\sum_{i=1}^{n} (\xi_i - \mathsf{E}\xi_i)\lambda_i\right]^2 \geq 0.$$

The following lemma shows that the converse is also true.

Lemma. *A necessary and sufficient condition that an $n \times n$ matrix $\mathbb{R}$ is the covariance matrix of a vector $\xi = (\xi_1, \ldots, \xi_n)$ is that the matrix is symmetric and nonnegative definite, or, equivalently, that there is an $n \times k$ matrix A $(1 \leq k \leq n)$ such that*

$$\mathbb{R} = AA^{\mathrm{T}},$$

where T *denotes the transpose.*

PROOF. We showed above that every covariance matrix is symmetric and nonnegative definite.

Conversely, let $\mathbb{R}$ be a matrix with these properties. We know from matrix theory that corresponding to every symmetric nonnegative definite matrix $\mathbb{R}$ there is an orthogonal matrix $\mathcal{O}$ (i.e., $\mathcal{O}\mathcal{O}^{\mathrm{T}} = E$, the unit matrix) such that

$$\mathcal{O}^{\mathrm{T}}\mathbb{R}\mathcal{O} = D,$$

where

$$D = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{pmatrix}$$

is a diagonal matrix with nonnegative elements d_i, $i = 1, \ldots, n$.

It follows that

$$\mathbb{R} = \mathcal{O}D\mathcal{O}^{\mathrm{T}} = (\mathcal{O}B)(B^{\mathrm{T}}\mathcal{O}^{\mathrm{T}}),$$

where B is the diagonal matrix with elements $b_i = +\sqrt{d_i}$, $i = 1, \ldots, n$. Consequently if we put $A = \mathcal{O}B$ we have the required representation $\mathbb{R} = AA^{\mathrm{T}}$ for $\mathbb{R}$.

It is clear that every matrix AA^{T} is symmetric and nonnegative definite. Consequently we have only to show that $\mathbb{R}$ is the covariance matrix of some random vector.

Let $\eta_1, \eta_2, \ldots, \eta_n$ be a sequence of independent normally distributed random variables, $\mathcal{N}(0, 1)$. (The existence of such a sequence follows, for example, from Corollary 1 of Theorem 1, §9, and in principle could easily be derived from Theorem 2 of §3.) Then the random vector $\zeta = A\eta$ (vectors are thought of as column vectors) has the required properties. In fact,

$$\mathsf{E}\zeta\zeta^{\mathrm{T}} = \mathsf{E}(A\eta)(A\eta)^{\mathrm{T}} = A \cdot \mathsf{E}\eta\eta^{\mathrm{T}} \cdot A^{\mathrm{T}} = A\mathsf{E}A^{\mathrm{T}} = AA^{\mathrm{T}}.$$

(If $\zeta = \|\zeta_{ij}\|$ is a matrix whose elements are random variables, $\mathsf{E}\zeta$ means the matrix $\|\mathsf{E}\xi_{ij}\|$).

This completes the proof of the lemma.

We now turn our attention to the two-dimensional Gaussian (normal) density

$$f_{\xi\eta}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_1)^2}{\sigma_1^2} - 2\rho\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right]\right\}, \tag{4}$$

characterized by the five parameters m_1, m_2, σ_1, σ_2 and ρ (cf. (3.14)), where $|m_1| < \infty, |m_2| < \infty, \sigma_1 > 0, \sigma_2 > 0, |\rho| < 1$. An easy calculation identifies these parameters:

$$m_1 = \mathsf{E}\xi, \quad \sigma_1^2 = \mathsf{V}\xi,$$
$$m_2 = \mathsf{E}\eta, \quad \sigma_2^2 = \mathsf{V}\eta,$$
$$\rho = \rho(\xi, \eta).$$

In §4 of Chapter I we explained that if ξ and η are uncorrelated ($\rho(\xi, \eta) = 0$), it does not follow that they are independent. However, if the pair (ξ, η) is Gaussian, it does follow that if ξ and η are uncorrelated then they are independent.

In fact, if $\rho = 0$ in (4), then

$$f_{\xi\eta}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-[(x-m_1)2]/2\sigma_1^2} \cdot e^{-[(y-m_2)^2]/2\sigma_2^2}.$$

But by (6.55) and (4),

$$f_\xi(x) = \int_{-\infty}^{\infty} f_{\xi\eta}(x, y)\, dy = \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-[(x-m_1)^2]/2\sigma_1^2},$$

$$f_\eta(y) = \int_{-\infty}^{\infty} f_{\xi\eta}(x, y)\, dx = \frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-[(y-m_2)^2]/2\sigma_2^2}.$$

Consequently

$$f_{\xi\eta}(x, y) = f_\xi(x) \cdot f_\eta(y),$$

from which it follows that ξ and η are independent (see the end of Subsection 9 of §6).

2. A striking example of the utility of the concept of conditional expectation (introduced in §7) is its application to the solution of the following problem which is connected with *estimation theory* (cf. Subsection 8 of §4 of Chapter I).

Let (ξ, η) be a pair of random variables such that ξ is observable but η is not. We ask how the unobservable component η can be "estimated" from the knowledge of observations of ξ.

To state the problem more precisely, we need to define the concept of an *estimator*. Let $\varphi = \varphi(x)$ be a Borel function. We call the random variable $\varphi(\xi)$ an estimator of η in terms of ξ, and $\mathsf{E}[\eta - \varphi(\xi)]^2$ the (mean square) error of this estimator. An estimator $\varphi^*(\xi)$ is called *optimal* (in the mean-square sense) if

$$\Delta \equiv \mathsf{E}[\eta - \varphi^*(\xi)]^2 = \inf_{\varphi} \mathsf{E}[\eta - \varphi(\xi)]^2, \tag{5}$$

where inf is taken over all Borel functions $\varphi = \varphi(x)$.

Theorem 1. *Let $\mathsf{E}\eta^2 < \infty$. Then there is an optimal estimator $\varphi^* = \varphi^*(\xi)$ and $\varphi^*(x)$ can be taken to be the function*

$$\varphi^*(x) = \mathsf{E}(\eta \mid \xi = x). \tag{6}$$

PROOF. Without loss of generality we may consider only estimators $\varphi(\xi)$ for which $\mathsf{E}\varphi^2(\xi) < \infty$. Then if $\varphi(\xi)$ is such an estimator, and $\varphi^*(\xi) = \mathsf{E}(\eta \mid \xi)$, we have

$$\begin{aligned}
\mathsf{E}[\eta - \varphi(\xi)]^2 &= \mathsf{E}[(\eta - \varphi^*(\xi)) + (\varphi^*(\xi) - \varphi(\xi))]^2 \\
&= \mathsf{E}[\eta - \varphi^*(\xi)]^2 + \mathsf{E}[\varphi^*(\xi) - \varphi(\xi)]^2 \\
&\quad + 2\mathsf{E}[(\eta - \varphi^*(\xi))(\varphi^*(\xi) - \varphi(\xi))] \geq \mathsf{E}[\eta - \varphi^*(\xi)]^2,
\end{aligned}$$

since $\mathsf{E}[\varphi^*(\xi) - \varphi(\xi)]^2 \geq 0$ and, by the properties of conditional expectations,

$$\begin{aligned}
\mathsf{E}[(\eta - \varphi^*(\xi))(\varphi^*(\xi) - \varphi(\xi))] &= \mathsf{E}\{\mathsf{E}[(\eta - \varphi^*(\xi))(\varphi^*(\xi) - \varphi(\xi)) \mid \xi]\} \\
&= \mathsf{E}\{(\varphi^*(\xi) - \varphi(\xi))\mathsf{E}(\eta - \varphi^*(\xi) \mid \xi)\} = 0.
\end{aligned}$$

This completes the proof of the theorem.

Remark. It is clear from the proof that the conclusion of the theorem is still valid when ξ is not merely a random variable but any random element with values in a measurable space $(E, \mathscr{E})$. We would then assume that $\varphi = \varphi(x)$ is an $\mathscr{E}/\mathscr{B}(R)$-measurable function.

Let us consider the form of $\varphi^*(x)$ on the hypothesis that (ξ, η) is a Gaussian pair with density given by (4).

From (1), (4) and (7.10) we find that the density $f_{\eta|\xi}(y|x)$ of the conditional probability distribution is given by

$$f_{\eta|\xi}(y|x) = \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_2}} e^{[(y-m(x))^2]/[2\sigma_2^2(1-\rho^2)]}, \tag{7}$$

where

$$m(x) = m_2 + \frac{\sigma_2}{\sigma_1}\rho\cdot(x - m_1). \tag{8}$$

Then by the Corollary of Theorem 3, §7,

$$\mathsf{E}(\eta|\xi = x) = \int_{-\infty}^{\infty} y f_{\eta|\xi}(y|x)\,dy = m(x) \tag{9}$$

and

$$\begin{aligned} \mathsf{V}(\eta|\xi = x) &\equiv \mathsf{E}[(\eta - \mathsf{E}(\eta|\xi = x))^2|\xi = x] \\ &= \int_{-\infty}^{\infty} (y - m(x))^2 f_{\eta|\xi}(y|x)\,dy \\ &= \sigma_2^2(1-\rho^2). \end{aligned} \tag{10}$$

Notice that the conditional variance $\mathsf{V}(\eta|\xi = x)$ is independent of x and therefore

$$\Delta = \mathsf{E}[\eta - \mathsf{E}(\eta|\xi)]^2 = \sigma_2^2(1-\rho^2). \tag{11}$$

Formulas (9) and (11) were obtained under the assumption that $\mathsf{V}\xi > 0$ and $\mathsf{V}\eta > 0$. However, if $\mathsf{V}\xi > 0$ and $\mathsf{V}\eta = 0$ they are still evidently valid.

Hence we have the following result (cf. (I.4.16) and (I.4.17)).

Theorem 2. *Let (ξ, η) be a Gaussian vector with $\mathsf{V}\xi > 0$. Then the optimal estimator of η in terms of ξ is*

$$\mathsf{E}(\eta|\xi) = \mathsf{E}\eta + \frac{\operatorname{cov}(\xi,\eta)}{\mathsf{V}\xi}(\xi - \mathsf{E}\xi), \tag{12}$$

and its error is

$$\Delta \equiv \mathsf{E}[\eta - \mathsf{E}(\eta|\xi)]^2 = \mathsf{V}\eta - \frac{\operatorname{cov}^2(\xi,\eta)}{\mathsf{V}\xi}. \tag{13}$$

Remark. The curve $y(x) = \mathsf{E}(\eta|\xi = x)$ is the *curve of regression of η on ξ* or *of η with respect to ξ*. In the Gaussian case $\mathsf{E}(\eta|\xi = x) = a + bx$ and consequently the regression of η and ξ is linear. Hence it is not surprising that the right-hand sides of (12) and (13) agree with the corresponding parts of (I.4.6) and (I.4.17) for the optimal linear estimator and its error.

Corollary. *Let ε_1 and ε_2 be independent Gaussian random variables with mean zero and unit variance, and*

$$\xi = a_1\varepsilon_1 + a_2\varepsilon_2, \qquad \eta = b_1\varepsilon_1 + b_2\varepsilon_2.$$

Then $\mathsf{E}\xi = \mathsf{E}\eta = 0$, $\mathsf{V}\xi = a_1^2 + a_2^2$, $\mathsf{V}\eta = b_1^2 + b_2^2$, $\text{cov}(\xi, \eta) = a_1 b_1 + a_2 b_2$, *and if* $a_1^2 + a_2^2 > 0$, *then*

$$\mathsf{E}(\eta \mid \xi) = \frac{a_1 b_1 + a_2 b_2}{a_1^2 + a_2^2}\,\xi, \tag{14}$$

$$\Delta = \frac{(a_1 b_2 - a_2 b_1)^2}{a_1^2 + a_2^2}. \tag{15}$$

3. Let us consider the problem of determining the distribution functions of random variables that are functions of other random variables.

Let ξ be a random variable with distribution function $F_\xi(x)$ (and density $f_\xi(x)$, if it exists), let $\varphi = \varphi(x)$ be a Borel function and $\eta = \varphi(\xi)$. Letting $I_y = (-\infty, y)$, we obtain

$$F_\eta(y) = \mathsf{P}(\eta \le y) = \mathsf{P}(\varphi(\xi) \in I_y) = \mathsf{P}(\xi \in \varphi^{-1}(I_y)) = \int_{\varphi^{-1}(I_y)} F_\xi(dx), \tag{16}$$

which expresses the distribution function $F_\eta(y)$ in terms of $F_\xi(x)$ and φ.

For example, if $\eta = a\xi + b$, $a > 0$, we have

$$F_\eta(y) = \mathsf{P}\left(\xi \le \frac{y-b}{a}\right) = F_\xi\left(\frac{y-b}{a}\right). \tag{17}$$

If $\eta = \xi^2$, it is evident that $F_\eta(y) = 0$ for $y < 0$, while for $y \ge 0$

$$\begin{aligned} F_\eta(y) &= \mathsf{P}(\xi^2 \le y) = \mathsf{P}(-\sqrt{y} \le \xi \le \sqrt{y}) \\ &= F_\xi(\sqrt{y}) - F_\xi(-\sqrt{y}) + \mathsf{P}(\xi = -\sqrt{y}). \end{aligned} \tag{18}$$

We now turn to the problem of determining $f_\eta(y)$.

Let us suppose that the range of ξ is a (finite or infinite) open interval $I = (a, b)$, and that the function $\varphi = \varphi(x)$, with domain (a, b), is continuously differentiable and either strictly increasing or strictly decreasing. We also suppose that $\varphi'(x) \ne 0$, $x \in I$. Let us write $h(y) = \varphi^{-1}(y)$ and suppose for definiteness that $\varphi(x)$ is strictly increasing. Then when $y \in \varphi(I)$,

$$\begin{aligned} F_\eta(y) &= \mathsf{P}(\eta \le y) = \mathsf{P}(\varphi(\xi) \le y) = \mathsf{P}(\xi \le \varphi^{-1}(y)) \\ &= \mathsf{P}(\xi \le h(y)) = \int_{-\infty}^{h(y)} f_\xi(x)\,dx. \end{aligned} \tag{19}$$

By Problem 15 of §6,

$$\int_{-\infty}^{h(y)} f_\xi(x)\,dx = \int_{-\infty}^{y} f_\xi(h(z))h'(z)\,dz \tag{20}$$

and therefore

$$f_\eta(y) = f_\xi(h(y))h'(y). \tag{21}$$

Similarly, if $\varphi(x)$ is strictly decreasing,

$$f_\eta(y) = f_\xi(h(y))((-h'(y)).$$

Hence in either case

$$f_\eta(y) = f_\xi(h(y))|h'(y)|. \tag{22}$$

For example, if $\eta = a\xi + b$, $a \neq 0$, we have

$$h(y) = \frac{y-b}{a} \quad \text{and} \quad f_\eta(y) = \frac{1}{|a|} f_\xi\left(\frac{y-b}{a}\right).$$

If $\xi \sim \mathcal{N}(m, \sigma^2)$ and $\eta = e^\xi$, we find from (22) that

$$f_\eta(y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma y} \exp\left[-\dfrac{\ln(y/M)^2}{2\sigma^2}\right], & y > 0, \\ 0 & y \leq 0, \end{cases} \tag{23}$$

with $M = e^m$.

A probability distribution with the density (23) is said to be *lognormal* (logarithmically normal).

If $\varphi = \varphi(x)$ is neither strictly increasing nor strictly decreasing, formula (22) is inapplicable. However, the following generalization suffices for many applications.

Let $\varphi = \varphi(x)$ be defined on the set $\sum_{k=1}^{n} [a_k, b_k]$, continuously differentiable and either strictly increasing or strictly decreasing on each open interval $I_k = (a_k, b_k)$, and with $\varphi'(x) \neq 0$ for $x \in I_k$. Let $h_k = h_k(y)$ be the inverse of $\varphi(x)$ for $x \in I_k$. Then we have the following generalization of (22):

$$f_\eta(y) = \sum_{k=1}^{n} f_\xi(h_k(y))|h_k'(y)| \cdot I_{D_k}(y), \tag{24}$$

where D_k is the domain of $h_k(y)$.

For example, if $\eta = \xi^2$ we can take $I_1 = (-\infty, 0)$, $I_2 = (0, \infty)$, and find that $h_1(y) = -\sqrt{y}$, $h_2(y) = \sqrt{y}$, and therefore

$$f_\eta(y) = \begin{cases} \dfrac{1}{2\sqrt{y}} [f_\xi(\sqrt{y}) + f_\xi(-\sqrt{y})], & y > 0, \\ 0, & y \leq 0. \end{cases} \tag{25}$$

We can observe that this result also follows from (18), since $\mathrm{P}(\xi = -\sqrt{y}) = 0$. In particular, if $\xi \sim \mathcal{N}(0, 1)$,

$$f_{\xi^2}(y) = \begin{cases} \dfrac{1}{\sqrt{2\pi y}} e^{-y/2}, & y > 0, \\ 0, & y \le 0. \end{cases} \tag{26}$$

A straightforward calculation shows that

$$f_{|\xi|}(y) = \begin{cases} f_\xi(y) + f_\xi(-y), & y > 0, \\ 0, & y \le 0. \end{cases} \tag{27}$$

$$f_{+\sqrt{|\xi|}}(y) = \begin{cases} 2y(f_\xi(y^2) + f_\xi(-y^2)), & y > 0, \\ 0, & y \le 0. \end{cases} \tag{28}$$

4. We now consider functions of several random variables.

If ξ and η are random variables with joint distribution $F_{\xi\eta}(x, y)$, and $\varphi = \varphi(x, y)$ is a Borel function, then if we put $\zeta = \varphi(\xi, \eta)$ we see at once that

$$F_\zeta(z) = \int_{\{x,\, y:\, \varphi(x, y) \le z\}} dF_{\xi\eta}(x, y). \tag{29}$$

For example, if $\varphi(x, y) = x + y$, and ξ and η are independent (and therefore $F_{\xi\eta}(x, y) = F_\xi(x) \cdot F_\eta(y)$) then Fubini's theorem shows that

$$\begin{aligned} F_\zeta(z) &= \int_{\{x,\, y:\, x+y \le z\}} dF_\xi(x) \cdot dF_\eta(y) \\ &= \int_{R^2} I_{\{x+y \le z\}}(x, y)\, dF_\xi(x)\, dF_\eta(y) \\ &= \int_{-\infty}^{\infty} dF_\xi(x) \left\{ \int_{-\infty}^{\infty} I_{\{x+y \le z\}}(x, y)\, dF_\eta(y) \right\} = \int_{-\infty}^{\infty} F_\eta(z - x)\, dF_\xi(x) \end{aligned} \tag{30}$$

and similarly

$$F_\zeta(z) = \int_{-\infty}^{\infty} F_\xi(z - y)\, dF_\eta(y). \tag{31}$$

If F and G are distribution functions, the function

$$H(z) = \int_{-\infty}^{\infty} F(z - x)\, dG(x)$$

is denoted by $F * G$ and called the *convolution* of F and G.

Thus *the distribution function F_ζ of the sum of two independent random variables ξ and η is the convolution of their distribution functions F_ξ and F_η*:

$$F_\zeta = F_\xi * F_\eta.$$

It is clear that $F_\xi * F_\eta = F_\eta * F_\xi$.

Now suppose that the independent random variables ξ and η have densities f_ξ and f_η. Then we find from (31), with another application of Fubini's theorem, that

$$F_\zeta(z) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z-y} f_\xi(u)\,du\right] f_\eta(y)\,dy$$
$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z} f_\xi(u-y)\,du\right] f_\eta(y)\,dy = \int_{-\infty}^{z} \left[\int_{-\infty}^{\infty} f_\xi(u-y) f_\eta(y)\,dy\right] du,$$

whence

$$f_\zeta(z) = \int_{-\infty}^{\infty} f_\xi(z-y) f_\eta(y)\,dy, \tag{32}$$

and similarly

$$f_\zeta(z) = \int_{-\infty}^{\infty} f_\eta(z-x) f_\xi(x)\,dx. \tag{33}$$

Let us see some examples of the use of these formulas.

Let $\xi_1, \xi_2, \ldots, \xi_n$ be a sequence of independent identically distributed random variables with the uniform density on $[-1, 1]$:

$$f(x) = \begin{cases} \frac{1}{2}, & |x| \le 1, \\ 0, & |x| > 1. \end{cases}$$

Then by (32) we have

$$f_{\xi_1+\xi_2}(x) = \begin{cases} \dfrac{2-|x|}{4}, & |x| \le 2, \\ 0, & |x| > 2, \end{cases}$$

$$f_{\xi_1+\xi_2+\xi_3}(x) = \begin{cases} \dfrac{(3-|x|)^2}{16}, & 1 \le |x| \le 3, \\ \dfrac{3-x^2}{8}, & 0 \le |x| \le 1, \\ 0, & |x| > 3, \end{cases}$$

and by induction

$$f_{\xi_1+\cdots+\xi_n}(x) = \begin{cases} \dfrac{1}{2^n(n-1)!} \displaystyle\sum_{k=0}^{[(n+x)/2]} (-1)^k C_n^k (n+x-2k)^{n-1}, & |x| \le n, \\ 0, & |x| > n. \end{cases}$$

Now let $\xi \sim \mathcal{N}(m_1, \sigma_1^2)$ and $\eta \sim \mathcal{N}(m_2, \sigma_2^2)$. If we write

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

then

$$f_\xi(x) = \frac{1}{\sigma_1}\varphi\left(\frac{x - m_1}{\sigma_1}\right), \qquad f_\eta(x) = \frac{1}{\sigma_2}\varphi\left(\frac{x - m_2}{\sigma_2}\right),$$

and the formula

$$f_{\xi+\eta}(x) = \frac{1}{\sqrt{\sigma_1^2 + \sigma_2^2}}\varphi\left(\frac{x - (m_1 + m_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)$$

follows easily from (32).

Therefore the *sum of two independent Gaussian random variables is again a Gaussian random variable with mean* $m_1 + m_2$ *and variance* $\sigma_1^2 + \sigma_2^2$.

Let $\xi_1, \ldots, \xi_n$ be independent random variables each of which is normally distributed with mean 0 and variance 1. Then it follows easily from (26) (by induction) that

$$f_{\xi_1^2+\cdots+\xi_n^2}(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)}\, x^{(n/2)-1}e^{-x/2}, & x > 0, \\ 0, & x \leq 0. \end{cases} \tag{34}$$

The variable $\xi_1^2 + \cdots + \xi_n^2$ is usually denoted by χ_n^2, and its distribution (with density (30)) is the χ^2-*distribution* ("chi-square distribution") with n degrees of freedom (cf. Table 2 in §3).

If we write $\chi_n = +\sqrt{\chi_n^2}$, it follows from (28) and (34) that

$$f_{\chi_n}(x) = \begin{cases} \dfrac{2x^{n-1}e^{-x^2/2}}{2^{n/2}\Gamma(n/2)}, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{35}$$

The probability distribution with this density is the χ-distribution (chi-distribution) with n degrees of freedom.

Again let ξ and η be independent random variables with densities f_ξ and f_η. Then

$$F_{\xi\eta}(z) = \iint_{\{x,y:\, xy \leq z\}} f_\xi(x) f_\eta(y)\,dx\,dy,$$

$$F_{\xi/\eta}(z) = \iint_{\{x,y:\, x/y \leq z\}} f_\xi(x) f_\eta(y)\,dx\,dy.$$

Hence we easily obtain

$$f_{\xi\eta}(z) = \int_{-\infty}^{\infty} f_\xi\left(\frac{z}{y}\right) f_\eta(y)\,\frac{dy}{|y|} = \int_{-\infty}^{\infty} f_\eta\left(\frac{z}{x}\right) f_\xi(x)\,\frac{dx}{|x|} \tag{36}$$

and

$$f_{\xi/\eta}(z) = \int_{-\infty}^{\infty} f_\xi(zy) f_\eta(y)|y|\,dy. \tag{37}$$

Putting $\xi = \xi_0$ and $\eta = \sqrt{(\xi_1^2 + \cdots + \xi_n^2)/n}$, in (37), where $\xi_0, \xi_1, \ldots, \xi_n$ are independent Gaussian random variables with mean 0 and variance $\sigma^2 > 0$, and using (35), we find that

$$f_{\xi_0/[\sqrt{(1/n)(\xi_1^2 + \cdots + \xi_n^2)}]}(x) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \frac{1}{\left(1 + \frac{x^2}{n}\right)^{(n+1)/2}}. \tag{38}$$

The variable $\xi_0/[\sqrt{(1/n)(\xi_1^2 + \cdots + \xi_n^2)}]$ is denoted by t, and its distribution is the *t-distribution*, or *Student's distribution*, with n degrees of freedom (cf. Table 2 in §3). Observe that this distribution is independent of σ.

5. Problems

1. Verify formulas (9), (10), (24), (27), (28), and (34)–(38).

2. Let $\xi_1, \ldots, \xi_n$, $n \geq 2$, be independent identically distributed random variables with distribution function $F(x)$ (and density $f(x)$, if it exists), and let $\bar{\xi} = \max(\xi_1, \ldots, \xi_n)$, $\underline{\xi} = \min(\xi_1, \ldots, \xi_n)$, $\rho = \bar{\xi} - \underline{\xi}$. Show that

$$F_{\bar{\xi},\underline{\xi}}(y, x) = \begin{cases} (F(y))^n - (F(y) - F(x))^n, & y > x, \\ (F(y))^n, & y \leq x, \end{cases}$$

$$f_{\bar{\xi},\underline{\xi}}(y, x) = \begin{cases} n(n-1)[F(y) - F(x)]^{n-2} f(x) f(y), & y > x, \\ 0, & y < x, \end{cases}$$

$$F_\rho(x) = \begin{cases} n \int_{-\infty}^{\infty} [F(y) - F(y - x)]^{n-1} f(y)\, dy, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

$$f_\rho(x) = \begin{cases} n(n-1) \int_{-\infty}^{\infty} [F(y) - F(y - x)]^{n-2} f(y - x) f(y)\, dy, & x > 0, \\ 0, & x < 0. \end{cases}$$

3. Let ξ_1 and ξ_2 be independent Poisson random variables with respective parameters λ_1 and λ_2. Show that $\xi_1 + \xi_2$ has a Poisson distribution with parameter $\lambda_1 + \lambda_2$.

4. Let $m_1 = m_2 = 0$ in (4). Show that

$$f_{\xi/\eta}(z) = \frac{\sigma_1 \sigma_2 \sqrt{1 - \rho^2}}{\pi(\sigma_2^2 z - 2\rho\sigma_1\sigma_2 z + \sigma_1^2)}.$$

5. The *maximal correlation coefficient* of ξ and η is $\rho^*(\xi, \eta) = \sup_{u,v} \rho(u(\xi), v(\xi))$, where the supremum is taken over the Borel functions $u = u(x)$ and $v = v(x)$ for which the correlation coefficient $\rho(u(\xi), v(\xi))$ is defined. Show that ξ and η are independent if and only if $\rho^*(\xi, \eta) = 0$.

6. Let $\tau_1, \tau_2, \ldots, \tau_n$ be independent nonnegative identically distributed random variables with the exponential density

$$f(t) = \lambda e^{-\lambda t}, \qquad t \geq 0.$$

Show that the distribution of $\tau_1 + \cdots + \tau_k$ has the density

$$\frac{\lambda^k t^{k-1} e^{-\lambda t}}{(k-1)!}, \qquad t \geq 0, \quad 1 \leq k \leq n,$$

and that

$$\mathsf{P}(\tau_1 + \cdots + \tau_k > t) = \sum_{i=0}^{k-1} e^{-\lambda t} \frac{(\lambda t)^i}{i!}.$$

7. Let $\xi \sim \mathcal{N}(0, \sigma^2)$. Show that, for every $p \geq 1$,

$$\mathsf{E}|\xi|^p = C_p \sigma^p,$$

where

$$C_p = \frac{2^{p/2}}{\pi^{1/2}} \Gamma\left(\frac{p+1}{2}\right)$$

and $\Gamma(s) = \int_0^\infty e^{-x} x^{s-1}\, dx$ is the gamma function. In particular, for each integer $n \geq 1$,

$$\mathsf{E}\xi^{2n} = (2n-1)!!\, \sigma^{2n}.$$

§9. Construction of a Process with Given Finite-Dimensional Distribution

1. Let $\xi = \xi(\omega)$ be a random variable defined on the probability space $(\Omega, \mathscr{F}, \mathsf{P})$, and let

$$F_\xi(x) = \mathsf{P}\{\omega: \xi(\omega) \leq x\}$$

be its distribution function. It is clear that $F_\xi(x)$ is a distribution function on the real line in the sense of Definition 1 of §3.

We now ask the following question. Let $F = F(x)$ be a distribution function on R. Does there exist a random variable whose distribution function is $F(x)$?

One reason for asking this question is as follows. Many statements in probability theory begin, "Let ξ be a random variable with the distribution function $F(x)$; then ...". Consequently if a statement of this kind is to be meaningful we need to be certain that the object under consideration actually exists. Since to know a random variable we first have to know its domain $(\Omega, \mathscr{F})$, and in order to speak of its distribution we need to have a probability measure P on $(\Omega, \mathscr{F})$, a correct way of phrasing the question of the existence of a random variable with a given distribution function $F(x)$ is this:

Do there exist a probability space $(\Omega, \mathscr{F}, \mathsf{P})$ and a random variable $\xi = \xi(\omega)$ on it, such that

$$\mathsf{P}\{\omega: \xi(\omega) \leq x\} = F(x)?$$

Let us show that the answer is positive, and essentially contained in Theorem 1 of §1.

In fact, let us put

$$\Omega = R, \qquad \mathscr{F} = \mathscr{B}(R).$$

It follows from Theorem 1 of §1 that there is a probability measure P (and only one) on $(R, \mathscr{B}(R))$ for which $\mathsf{P}(a, b]] = F(b) - F(a)$, $a < b$.

Put $\xi(\omega) \equiv \omega$. Then

$$\mathsf{P}\{\omega\colon \xi(\omega) \le x\} = \mathsf{P}\{\omega\colon \omega \le x\} = \mathsf{P}(-\infty, x] = F(x).$$

Consequently we have constructed the required probability space and the random variable on it.

2. Let us now ask a similar question for random processes.

Let $X = (\xi_t)_{t \in T}$ be a random process (in the sense of Definition 3, §5) defined on the probability space $(\Omega, \mathscr{F}, \mathsf{P})$, with $t \in T \subseteq R$.

From a physical point of view, the most fundamental characteristic of a random process is the set $\{F_{t_1, \ldots, t_n}(x_1, \ldots, x_n)\}$ of its *finite-dimensional distribution functions*

$$F_{t_1, \ldots, t_n}(x_1, \ldots, x_n) = P\{\omega\colon \xi_{t_1} \le x_1, \ldots, \xi_{t_n} \le x_n\}, \tag{1}$$

defined for all sets $t_1, \ldots, t_n$ with $t_1 < t_2 < \cdots < t_n$.

We see from (1) that, for each set $t_1, \ldots, t_n$ with $t_1 < t_2 < \cdots < t_n$ the functions $F_{t_1, \ldots, t_n}(x_1, \ldots, x_n)$ are n-dimensional distribution functions (in the sense of Definition 2, §3) and that the collection $\{F_{t_1, \ldots, t_n}(x_1, \ldots, x_n)\}$ has the following *consistency* property:

$$\lim_{x_k \uparrow \infty} F_{t_1, \ldots, t_n}(x_1, \ldots, x_n) = F_{t_1, \ldots, \hat{t}_k, \ldots, t_n}(x_1, \ldots, \hat{x}_k, \ldots, x_n) \tag{2}$$

where ^ indicates an omitted coordinate.

Now it is natural to ask the following question: under what conditions can a given family $\{F_{t_1, \ldots, t_n}(x_1, \ldots, x_n)\}$ of distribution functions $F_{t_1, \ldots, t_n}(x_1, \ldots, x_n)$ (in the sense of Definition 2, §3) be the family of finite-dimensional distribution functions of a random process? It is quite remarkable that all such conditions are covered by the consistency condition (2).

Theorem 1 (Kolmogorov's Theorem on the Existence of a Process). *Let* $\{F_{t_1, \ldots, t_n}(x_1, \ldots, x_n)\}$, *with* $t_i \in T \subseteq R$, $t_1 < t_2 < \cdots < t_n$, $n \ge 1$, *be a given family of finite-dimensional distribution functions, satisfying the consistency condition* (2). *Then there are a probability space* $(\Omega, \mathscr{F}, \mathsf{P})$ *and a random process* $X = (\xi_t)_{t \in T}$ *such that*

$$\mathsf{P}\{\omega\colon \xi_{t_1} \le x_1, \ldots, \xi_{t_n} \le x_n\} = F_{t_1 \cdots t_n}(x_1 \cdots x_n). \tag{3}$$

PROOF. Put

$$\Omega = R^T, \qquad \mathscr{F} = \mathscr{B}(R^T),$$

i.e. take Ω to be the space of real functions $\omega = (\omega_t)_{t \in T}$ with the σ-algebra generated by the cylindrical sets.

Let $\tau = [t_1, \ldots, t_n]$, $t_1 < t_2 < \cdots < t_n$. Then by Theorem 2 of §3 we can construct on the space $(R^n, \mathscr{B}(R^n))$ a unique probability measure P_τ such that

$$P_\tau\{(\omega_{t_1}, \ldots, \omega_{t_n}): \omega_{t_1} \le x_1, \ldots, \omega_{t_n} \le x_n\} = F_{t_1 \cdots t_n}(x_1, \ldots, x_n). \tag{4}$$

It follows from the consistency condition (2) that the family $\{P_\tau\}$ is also consistent (see (3.20)). According to Theorem 4 of §3 there is a probability measure P on $(R^T, \mathscr{B}(R^T))$ such that

$$\mathsf{P}\{\omega: (\omega_{t_1}, \ldots, \omega_{t_n}) \in B\} = P_\tau(B)$$

for every set $\tau = [t_1, \ldots, t_n]$, $t_1 < \cdots < t_n$.

From this, it also follows that (4) is satisfied. Therefore the required random process $X = (\xi_t(\omega))_{t \in T}$ can be taken to be the process defined by

$$\xi_t(\omega) = \omega_t, \qquad t \in T. \tag{5}$$

This completes the proof of the theorem.

Remark 1. The probability space $(R^T, \mathscr{B}(R^T), \mathsf{P})$ that we have constructed is called *canonical*, and the construction given by (5) is called the *coordinate method* of constructing the process.

Remark 2. Let $(E_\alpha, \mathscr{E}_\alpha)$ be complete separable metric spaces, where α belongs to some set $\mathfrak{A}$ of indices. Let $\{P_\tau\}$ be a set of consistent finite-dimensional distribution functions P_τ, $\tau = [\alpha_1, \ldots, \alpha_n]$ on

$$(E_{\alpha_1} \times \cdots \times E_{\alpha_n}, \mathscr{E}_{\alpha_1} \otimes \cdots \otimes \mathscr{E}_{\alpha_n}).$$

Then there are a probability space $(\Omega, \mathscr{F}, \mathsf{P})$ and a family of $\mathscr{F}/\mathscr{E}_\alpha$-measurable functions $(X_\alpha(\omega))_{\alpha \in \mathfrak{A}}$ such that

$$\mathsf{P}\{(X_{\alpha_1}, \ldots, X_{\alpha_n}) \in B\} = P_\tau(B)$$

for all $\tau = [\alpha_1, \ldots, \alpha_n]$ and $B \in \mathscr{E}_{\alpha} \otimes \cdots \otimes \mathscr{E}_{\alpha_n}$.

This result, which generalizes Theorem 1, follows from Theorem 4 of §3 if we put $\Omega = \prod_\alpha E_\alpha$, $\mathscr{F} = \bigotimes_\alpha \mathscr{E}_\alpha$ and $X_\alpha(\omega) = \omega_\alpha$ for each $\omega = \omega(\omega_\alpha)$, $\alpha \in \mathfrak{A}$.

Corollary 1. *Let $F_1(x), F_2(x), \ldots$ be a sequence of one-dimensional distribution functions. Then there exist a probability space $(\Omega, \mathscr{F}, \mathsf{P})$ and a sequence of independent random variables $\xi_1, \xi_2, \ldots$ such that*

$$P\{\omega: \xi_i(\omega) \le x\} = F_i(x). \tag{6}$$

In particular, there is a probability space $(\Omega, \mathscr{F}, \mathsf{P})$ on which an infinite sequence of Bernoulli random variables is defined (in this connection see Subsection 2 of §5 of Chapter I). Notice that Ω can be taken to be the space

$$\Omega = \{\omega: \omega = (a_1, a_2, \ldots), a_i = 0, 1\}$$

(cf. also Theorem 2).

To establish the corollary it is enough to put $F_{1,\ldots,n}(x_1, \ldots, x_n) = F_1(x_1) \cdots F_n(x_n)$ and apply Theorem 1.

Corollary 2. *Let $T = [0, \infty)$ and let $\{p(s, x; t, B\}$ be a family of nonnegative functions defined for $s, t \in T, t > s, x \in R, B \in \mathscr{B}(R)$, and satisfying the following conditions*:

(a) $p(s, x; t, B)$ *is a* probability measure on B *for given* s, x *and* t;
(b) *for given* s, t *and* B, *the function* $p(s, x; t, B)$ *is a* Borel function of x;
(c) *for* $0 \le s < t < \tau$ *and* $B \in \mathscr{B}(R)$, *the* Kolmogorov–Chapman equation

$$p(s, x; \tau, B) = \int_R p(s, x; t, dy)p(t, y; \tau, B) \tag{7}$$

is satisfied.

Also let $\pi = \pi(B)$ be a probability measure on $(R, \mathscr{B}(R))$. Then there are a probability space $(\Omega, \mathscr{F}, \mathsf{P})$ and a random process $X = (\xi_t)_{t \ge 0}$ defined on it, such that

$$P\{\xi_{t_0} \le x_0, \xi_{t_1} \le x_1, \ldots, \xi_{t_n} \le x_n\} = \int_{-\infty}^{x_0} \pi(dy_0) \int_{-\infty}^{x_1} p(0, y_0; t_1, dy_1) \cdots \int_{-\infty}^{x_n} p(t_{n-1}, y_{n-1}; t_n, dy_n) \tag{8}$$

for $0 = t_0 < t_1 < \cdots < t_n$.

The process X so constructed is a *Markov process* with initial distribution π and transition probabilities $\{p(s, x; t, B\}$.

Corollary 3. *Let $T = \{0, 1, 2, \ldots\}$ and let $\{P_k(x; B)\}$ be a family of non-negative functions defined for $k \ge 1$, $x \in R$, $B \in \mathscr{B}(R)$, and such that $p_k(x; B)$ is a probability measure on B (for given k and x) and measurable in x (for given k and B). In addition, let $\pi = \pi(B)$ be a probability measure on $(R, \mathscr{B}(R))$.*

Then there is a probability space $(\Omega, \mathscr{F}, \mathsf{P})$ with a family of random variables $X = \{\xi_0, \xi_1, \ldots\}$ defined on it, such that

$$\mathsf{P}\{\xi_{t_0} \le x_0, \xi_{t_1} \le x_1, \ldots, \xi_n \le x_n\} = \int_{-\infty}^{x_0} \pi(dy_0) \int_{-\infty}^{x_1} p(0, y_0; t_1, dy_1) \cdots \int_{-\infty}^{x_n} p(t_{n-1}, y_{n-1}; t_n, dy_n)$$

3. In the situation of Corollary 1, there is a sequence of independent random variables $\xi_1, \xi_2, \ldots$ whose one-dimensional distribution functions are $F_1, F_2, \ldots$, respectively.

Now let $(E_1, \mathscr{E}_1)$, $(E_2, \mathscr{E}_2)$, ... be complete separable metric spaces and let $P_1, P_2, \ldots$ be probability measures on them. Then it follows from Remark 2 that there are a probability space $(\Omega, \mathscr{F}, \mathsf{P})$ and a sequence of independent elements $X_1, X_2, \ldots$ such that X_n is $\mathscr{F}/\mathscr{E}_n$-measurable and $\mathsf{P}(X_n \in B) = P_n(B)$, $B \in \mathscr{E}_n$.

It turns out that this result remains valid when the spaces $(E_n, \mathscr{E}_n)$ are *arbitrary measurable spaces.*

Theorem 2 (Ionescu Tulcea's Theorem on Extending a Measure and the Existence of a Random Sequence). *Let* $(\Omega_n, \mathscr{F}_n)$, $n = 1, 2, \ldots$, *be arbitrary measurable spaces and* $\Omega = \prod \Omega_n$, $\mathscr{F} = \bigotimes \mathscr{F}_n$. *Suppose that a probability measure* P_1 *is given on* $(\Omega_1, \mathscr{F}_1)$ *and that, for every set* $(\omega_1, \ldots, \omega_n) \in \Omega_1 \times \cdots \times \Omega_n, n \geq 1$, *probability measures* $P(\omega_1, \ldots, \omega_n; \cdot)$ *are given on* $(\Omega_{n+1}, \mathscr{F}_{n+1})$. *Suppose that for every* $B \in \mathscr{F}_{n+1}$ *the functions* $P(\omega_1, \ldots, \omega_n; B)$ *are Borel functions on* $(\omega_1, \ldots, \omega_n)$ *and let*

$$P_n(A_1 \times \cdots \times A_n) = \int_{A_1} P_1(d\omega_1) \int_{A_2} P(\omega_1; d\omega_2) \cdots \int_{A_n} P(\omega_1, \ldots, \omega_{n-1}; d\omega_n) \qquad A_i \in \mathscr{F}_i, \qquad n \geq 1. \quad (9)$$

Then there is a unique probability measure P *on* $(\Omega, \mathscr{F})$ *such that*

$$\mathsf{P}\{\omega : \omega_1 \in A_1, \ldots, \omega_n \in A_n\} = P_n(A_1 \times \cdots \times A_n) \qquad (10)$$

for every $n \geq 1$, *and there is a random sequence* $X = (X_1(\omega), X_2(\omega), \ldots)$ *such that*

$$\mathsf{P}\{\omega : X_1(\omega) \in A_1, \ldots, X_n(\omega) \in A_n\} = P_n(A_1 \times \cdots \times A_n), \qquad (11)$$

where $A_i \in \mathscr{E}_i$.

PROOF. The first step is to establish that for each $n > 1$ the set function P_n defined by (9) on the rectangle $A_1 \times \cdots \times A_n$ can be extended to the σ-algebra $\mathscr{F}_1 \otimes \cdots \otimes \mathscr{F}_n$.

For each $n \geq 2$ and $B \in \mathscr{F}_1 \otimes \cdots \otimes \mathscr{F}_n$ we put

$$P_n(B) = \int_{\Omega_1} P_1(d\omega_1) \int_{\Omega_2} P(\omega_1; d\omega_2) \int_{\Omega_{n-1}} P(\omega_1, \ldots, \omega_{n-2}; d\omega_{n-1}) \times \int_{\Omega_n} I_B(\omega_1, \ldots, \omega_n) P(\omega_1, \ldots, \omega_{n-1}; d\omega_n). \qquad (12)$$

It is easily seen that when $B = A_1 \times \cdots \times A_n$ the right-hand side of (12) is the same as the right-hand side of (9). Moreover, when $n = 2$ it can be shown, just as in Theorem 8 of §6, that P_2 is a measure. Consequently it is easily established by induction that P_n is a measure for all $n \geq 2$.

The next step is the same as in Kolmogorov's theorem on the extension of a measure in $(R^\infty, \mathscr{B}(R^\infty))$ (Theorem 3, §3). Thus for every cylindrical set $J_n(B) = \{\omega \in \Omega: (\omega_1, \ldots, \omega_n) \in B\}$, $B \in \mathscr{F}_1 \otimes \cdots \otimes \mathscr{F}_n$, we define the set function P by

$$\mathsf{P}(J_n(B)) = P_n(B). \tag{13}$$

If we use (12) and the fact that $P(\omega_1, \ldots, \omega_k; \cdot)$ are measures, it is easy to establish that the definition (13) is consistent, in the sense that the value of $\mathsf{P}(J_n(B))$ is independent of the representation of the cylindrical set.

It follows that the set function P defined in (13) for cylindrical sets, and in an obvious way on the algebra that contains all the cylindrical sets, is a finitely additive measure on this algebra. It remains to verify its countable additivity and apply Carathéodory's theorem.

In Theorem 3 of §3 the corresponding verification was based on the property of $(R^n, \mathscr{B}(R^n))$ that for every Borel set B there is a compact set $A \subseteq B$ whose probability measure is arbitrarily close to the measure of B. In the present case this part of the proof needs to be modified in the following way.

As in Theorem 3 of §3, let $\{\hat{B}_n\}_{n \geq 1}$ be a sequence of cylindrical sets

$$\hat{B}_n = \{\omega: (\omega_1, \ldots, \omega_n) \in B_n\},$$

that decrease to the empty set $\varnothing$, but have

$$\lim_{n \to \infty} \mathsf{P}(\hat{B}_n) > 0. \tag{14}$$

For $n > 1$, we have from (12)

$$\mathsf{P}(\hat{B}_n) = \int_{\Omega_1} f_n^{(1)}(\omega_1) P_1(d\omega_1),$$

where

$$f_n^{(1)}(\omega_1) = \int_{\Omega_2} P(\omega_1; d\omega_2) \cdots \int_{\Omega_n} I_{B_n}(\omega_1, \ldots, \omega_n) P(\omega_2, \ldots, \omega_{n-1}; d\omega_n).$$

Since $\hat{B}_{n+1} \subseteq \hat{B}_n$, we have $B_{n+1} \subseteq B_n \times \Omega_{n+1}$ and therefore

$$I_{B_{n+1}}(\omega_1, \ldots, \omega_{n+1}) \leq I_{B_n}(\omega_1, \ldots, \omega_n) I_{\Omega_{n+1}}(\omega_{n+1}).$$

Hence the sequence $\{f_n^{(1)}(\omega_1)\}_{n \geq 1}$ decreases. Let $f^{(1)}(\omega_1) = \lim_n f_n^{(1)}(\omega_1)$. By the dominated convergence theorem

$$\lim_n \mathsf{P}(\hat{B}_n) = \lim_n \int_{\Omega_1} f_n^{(1)}(\omega_1) P_1(d\omega_1) = \int_{\Omega_1} f^{(1)}(\omega_1) P_1(d\omega_1).$$

By hypothesis, $\lim_n \mathsf{P}(\hat{B}_n) > 0$. It follows that there is an $\omega_1^0 \in B$ such that $f^{(1)}(\omega_1^0) > 0$, since if $\omega_1 \notin B_1$ then $f_n^{(1)}(\omega_1) = 0$ for $n \geq 1$.

Moreover, for $n > 2$,

$$f_n^{(1)}(\omega_1^0) = \int_{\Omega_2} f_n^{(2)}(\omega_2) P(\omega_1^0; d\omega_2), \tag{15}$$

where

$$f_n^{(2)}(\omega_2) = \int_{\Omega} P(\omega_1^0, \omega_2; d\omega_3)$$

$$\cdots \int_{\Omega_n} I_{B_n}(\omega_1^0, \omega_2, \ldots, \omega_n) P(\omega_1^0, \omega_2, \ldots, \omega_{n-1}, d\omega_n).$$

We can establish, as for $\{f_n^{(1)}(\omega_1)\}$, that $\{f_n^{(2)}(\omega_2)\}$ is decreasing. Let $f^{(2)}(\omega_2) = \lim_{n\to\infty} f_n^{(2)}(\omega_2)$. Then it follows from (15) that

$$0 < f^{(1)}(\omega_1^0) = \int_{\Omega_2} f^{(2)}(\omega_2) P(\omega_1^0; d\omega_2),$$

and there is a point $\omega_2^0 \in \Omega_2$ such that $f^{(2)}(\omega_2^0) > 0$. Then $(\omega_1^0, \omega_2^0) \in B_2$. Continuing this process, we find a point $(\omega_1^0, \ldots, \omega_n^0) \in B_n$ for each n. Consequently $(\omega_1^0, \ldots, \omega_n^0, \ldots) \in \bigcap \hat{B}_n$, but by hypothesis we have $\bigcap \hat{B}_n = \varnothing$. This contradiction shows that $\lim_n \mathsf{P}(\hat{B}_n) = 0$.

Thus we have proved the part of the theorem about the existence of the probability measure P. The other part follows from this by putting $X_n(\omega) = \omega_n$, $n \geq 1$.

Corollary 1. *Let $(E_n, \mathscr{E}_n)_{n\geq 1}$ be any measurable spaces and $(P_n)_{n\geq 1}$, measures on them. Then there are a probability space $(\Omega, \mathscr{F}, \mathsf{P})$ and a family of independent random elements $X_1, X_2, \ldots$ with values in $(E_1, \mathscr{E}_1)$, $(E_2, \mathscr{E}_2), \ldots$, respectively, such that*

$$\mathsf{P}\{\omega: X_n(\omega) \in B\} = P_n(B), \qquad B \in \mathscr{E}_n, n \geq 1.$$

Corollary 2. *Let $E = \{1, 2, \ldots\}$, and let $\{p_k(x, y)\}$ be a family of nonnegative functions, $k \geq 1$, $x, y \in E$, such that $\sum_{y\in E} p_k(x; y) = 1$, $x \in E$, $k \geq 1$. Also let $\pi = \pi(x)$ be a probability distribution on E (that is, $\pi(x) \geq 0, \sum_{x\in E} \pi(x) = 1$).*

Then there are a probability space $(\Omega, \mathscr{F}, \mathsf{P})$ and a family $X = \{\xi_0, \xi_1, \ldots\}$ of random variables on it, such that

$$\mathsf{P}\{\xi_0 = x_0, \xi_1 = x_1, \ldots, \xi_n = x_n\} = \pi(x_0) p_1(x_0, x_1) \cdots p_n(x_{n-1}, x_n) \tag{16}$$

(cf. (I.12.4)) for all $x_i \in E$ and $n \geq 1$. We may take Ω to be the space

$$\Omega = \{\omega: \omega = (x_0, x_1, \ldots), x_i \in E\}.$$

A sequence $X = \{\xi_0, \xi_1, \ldots\}$ of random variables satisfying (16) is a *Markov chain* with a countable set E of states, transition matrix $\{p_k(x, y)\}$ and initial probability distribution π. (Cf. the definition in §12 of Chapter I.)

4. Problems

1. Let $\Omega = [0, 1]$, let $\mathscr{F}$ be the class of Borel subsets of $[0, 1]$, and let P be Lebesgue measure on $[0, 1]$. Show that the space $(\Omega, \mathscr{F}, \mathsf{P})$ is universal in the following sense. For every distribution function $F(x)$ on $(\Omega, \mathscr{F}, \mathsf{P})$ there is a random variable $\xi = \xi(\omega)$ such that its distribution function $F_\xi(x) = \mathsf{P}(\xi \le x)$ coincides with $F(x)$. (Hint. $\xi(\omega) = F^{-1}(\omega)$, $0 < \omega < 1$, where $F^{-1}(\omega) = \sup\{x: F(x) < \omega\}$, when $0 < \omega < 1$, and $\xi(0)$, $\xi(1)$ can be chosen arbitrarily.)

2. Verify the consistency of the families of distributions in the corollaries to Theorems 1 and 2.

3. Deduce Corollary 2, Theorem 2, from Theorem 1.

§10. Various Kinds of Convergence of Sequences of Random Variables

1. Just as in analysis, in probability theory we need to use various kinds of convergence of random variables. Four of these are particularly important: *in probability, with probability one, in mean of order p, in distribution.*

First some definitions. Let $\xi, \xi_1, \xi_2, \ldots$ be random variables defined on a probability space $(\Omega, \mathscr{F}, \mathsf{P})$.

Definition 1. The sequence $\xi_1, \xi_2, \ldots$ of random variables converges *in probability* to the random variable ξ (notation: $\xi_n \xrightarrow{\mathsf{P}} \xi$) if for every $\varepsilon > 0$

$$\mathsf{P}\{|\xi_n - \xi| > \varepsilon\} \to 0, \qquad n \to \infty. \tag{1}$$

We have already encountered this convergence in connection with the law of large numbers for a Bernoulli scheme, which stated that

$$\mathsf{P}\left(\left|\frac{S_n}{n} - p\right| > \varepsilon\right) \to 0, \qquad n \to \infty$$

(see §5 of Chapter I). In analysis this is known as *convergence in measure*.

Definition 2. The sequence $\xi_1, \xi_2, \ldots$ of random variables converges *with probability one* (*almost surely, almost everywhere*) to the random variable ξ if

$$\mathsf{P}\{\omega: \xi_n \not\to \xi\} = 0, \tag{2}$$

i.e. if the set of sample points ω for which $\xi_n(\omega)$ does not converge to ξ has probability zero.

This convergence is denoted by $\xi_n \to \xi$ (P-a.s.), or $\xi_n \xrightarrow{\text{a.s.}} \xi$ or $\xi_n \xrightarrow{\text{a.e.}} \xi$.

Definition 3. The sequence $\xi_1, \xi_2, \ldots$ of random variables converges *in mean of order p*, $0 < p < \infty$, to the random variable ξ if

$$\mathsf{E}|\xi_n - \xi|^p \to 0, \qquad n \to \infty. \tag{3}$$

In analysis this is known as *convergence in L^p*, and denoted by $\xi_n \xrightarrow{L^p} \xi$. In the special case $p = 2$ it is called *mean square convergence* and denoted by $\xi = \text{l.i.m. } \xi_n$ (for "limit in the mean").

Definition 4. The sequence $\xi_1, \xi_2, \ldots$ of random variables *converges in distribution* to the random variable ξ (notation: $\xi_n \xrightarrow{d} \xi$) if

$$\mathsf{E}f(\xi_n) \to \mathsf{E}f(\xi), \qquad n \to \infty, \tag{4}$$

for every bounded continuous function $f = f(x)$. The reason for the terminology is that, according to what will be proved in Chapter III, §1, condition (4) is equivalent to the convergence of the distribution $F_{\xi_n}(x)$ to $F_\xi(x)$ at each *point x of continuity* of $F_\xi(x)$. This convergence is denoted by $F_{\xi_n} \Rightarrow F_\xi$.

We emphasize that the convergence of random variables in distribution is defined only in terms of the convergence of their distribution functions. Therefore it makes sense to discuss this mode of convergence even when the random variables are defined on different probability spaces. This convergence will be studied in detail in Chapter III, where, in particular, we shall explain why in the definition of $F_{\xi_n} \Rightarrow F_\xi$ we require only convergence at points of continuity of $F_\xi(x)$ and not at all x.

2. In solving problems of analysis on the convergence (in one sense or another) of a given sequence of functions, it is useful to have the concept of a fundamental sequence (or Cauchy sequence). We can introduce a similar concept for each of the first three kinds of convergence of a sequence of random variables.

Let us say that a sequence $\{\xi_n\}_{n\ge 1}$ of random variables is *fundamental in probability*, or *with probability* 1, or *in mean of order*, p, $0 < p < \infty$, if the corresponding one of the following properties is satisfied: $P\{|\xi_n - \xi_m| > \varepsilon\} \to 0$, as $m, n \to \infty$ for every $\varepsilon > 0$; the sequence $\{\xi_n(\omega)\}_{n\ge1}$ is fundamental for almost all $\omega \in \Omega$; the sequence $\{\xi_n(\omega)\}_{n\ge1}$ is fundamental in L^p, i.e. $\mathsf{E}|\xi_n - \xi_m|^p \to 0$ as $n, m \to \infty$.

3. Theorem 1.

(a) *A necessary and sufficient condition that $\xi_n \to \xi$ (P-a.s.) is that*

$$\mathsf{P}\left\{\sup_{k\ge n}|\xi_k - \xi| \ge \varepsilon\right\} \to 0, \qquad n \to \infty. \tag{5}$$

for every $\varepsilon > 0$.

(b) *The sequence $\{\xi_n\}_{n\geq 1}$ is fundamental with probability* 1 *if and only if*

$$\mathsf{P}\left\{\sup_{\substack{k\geq n\\ l\geq n}}|\xi_k-\xi_l|\geq\varepsilon\right\}\to 0,\qquad n\to\infty, \tag{6}$$

for every $\varepsilon>0$; or equivalently

$$\mathsf{P}\left\{\sup_{k\geq 0}|\xi_{n+k}-\xi_n|\geq\varepsilon\right\}\to 0,\qquad n\to\infty. \tag{7}$$

PROOF. (a) Let $A_n^\varepsilon=\{\omega:|\xi_n-\xi|\geq\varepsilon\}$, $A^\varepsilon=\overline{\lim}A_n^\varepsilon\equiv\bigcap_{n=1}^\infty\bigcup_{k\geq n}A_k^\varepsilon$. Then

$$\{\omega:\xi_n\nrightarrow\xi\}=\bigcup_{\varepsilon\geq 0}A^\varepsilon=\bigcup_{m=1}^\infty A^{1/m}.$$

But

$$\mathsf{P}(A^\varepsilon)=\lim_n\mathsf{P}\left(\bigcup_{k\geq n}A_k^\varepsilon\right),$$

Hence (a) follows from the following chain of implications:

$$\begin{aligned}0=\mathsf{P}\{\omega:\xi_n\nrightarrow\xi\}=\mathsf{P}\left(\bigcup_{\varepsilon>0}A^\varepsilon\right)&\Leftrightarrow\mathsf{P}\left(\bigcup_{m=1}^\infty A^{1/m}\right)=0\\&\Leftrightarrow\mathsf{P}(A^{1/m})=0,\quad m\geq 1\Leftrightarrow\mathsf{P}(A^\varepsilon)=0,\quad\varepsilon>0,\\&\Leftrightarrow\mathsf{P}\left(\bigcup_{k\geq n}A_k^\varepsilon\right)\to 0,\quad n\to\infty\Leftrightarrow\mathsf{P}\left(\sup_{k\geq n}|\xi_k-\xi|\geq\varepsilon\right)\to 0,\\&\qquad n\to\infty.\end{aligned}$$

(b) Let

$$B_{k,l}^\varepsilon=\{\omega:|\xi_k-\xi_l|\geq\varepsilon\},\qquad B^\varepsilon=\bigcap_{n=1}^\infty\bigcup_{\substack{k\geq n\\ l\geq n}}B_{k,l}^\varepsilon.$$

Then $\{\omega:\{\xi_n(\omega)\}_{n\geq 1}$ *is not fundamental*$\}=\bigcup_{\varepsilon\geq 0}B^\varepsilon$, and it can be shown as in (a) that $\mathsf{P}\{\omega:\{\xi_n(\omega)\}_{n\geq 1}$ is not fundamental$\}=0\Leftrightarrow(6)$. The equivalence of (6) and (7) follows from the obvious inequalities

$$\sup_{k\geq 0}|\xi_{n+k}-\xi_n|\leq\sup_{\substack{k\geq 0\\ l\geq 0}}|\xi_{n+k}-\xi_{n+l}|\leq 2\sup_{k\geq 0}|\xi_{n+k}-\xi_n|.$$

This completes the proof of the theorem.

Corollary. *Since*

$$\mathsf{P}\left\{\sup_{k\geq n}|\xi_k-\xi|\geq\varepsilon\right\}=\mathsf{P}\left\{\bigcup_{k\geq n}(|\xi_k-\xi|\geq\varepsilon)\right\}\leq\sum_{k\geq n}\mathsf{P}\{|\xi_k-\xi|\geq\varepsilon\},$$

a sufficient condition for $\xi_n \xrightarrow{\text{a.s.}} \xi$ *is that*

$$\sum_{k=1}^{\infty} \mathsf{P}\{|\xi_k - \xi| \geq \varepsilon\} < \infty \tag{8}$$

is satisfied for every $\varepsilon > 0$.

It is appropriate to observe at this point that the reasoning used in obtaining (8) lets us establish the following simple but important result which is essential in studying properties that are satisfied with probability 1.

Let $A_1, A_2, \ldots$ be a sequence of events in F. Let (see the table in §1) $\{A_n \text{ i.o.}\}$ denote the event $\overline{\lim} A_n$ that consists in the realization of infinitely many of $A_1, A_2, \ldots$.

Borel–Cantelli Lemma.

(a) *If* $\sum \mathsf{P}(A_n) < \infty$ *then* $\mathsf{P}\{A_n \text{ i.o.}\} = 0$.
(b) *If* $\sum \mathsf{P}(A_n) = \infty$ *and* $A_1, A_2, \ldots$ *are independent, then* $\mathsf{P}\{A_n \text{ i.o.}\} = 1$.

PROOF. (a) By definition

$$\{A_n \text{ i.o.}\} = \overline{\lim}\, A_n = \bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_k.$$

Consequently

$$\mathsf{P}\{A_n \text{ i.o.}\} = \mathsf{P}\left\{\bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_k\right\} = \lim \mathsf{P}\left(\bigcup_{k \geq n} A_k\right) \leq \lim \sum_{k \geq n} \mathsf{P}(A_k),$$

and (a) follows.

(b) If $A_1, A_2, \ldots$ are independent, so are $\bar{A}_1, \bar{A}_2, \ldots$. Hence for $N \geq n$ we have

$$\mathsf{P}\left(\bigcap_{k=n}^{N} A_k\right) = \prod_{k=n}^{N} \mathsf{P}(\bar{A}_k),$$

and it is then easy to deduce that

$$\mathsf{P}\left(\bigcap_{k=n}^{\infty} \bar{A}_k\right) = \prod_{k=n}^{\infty} \mathsf{P}(\bar{A}_k). \tag{9}$$

Since $\log(1 - x) \leq -x$, $0 \leq x < 1$,

$$\log \prod_{k=n}^{\infty} [1 - \mathsf{P}(A_k)] = \sum_{k=n}^{\infty} \log[1 - \mathsf{P}(A_k)] \leq -\sum_{k=n}^{\infty} \mathsf{P}(A_k) = -\infty.$$

Consequently

$$\mathsf{P}\left(\bigcap_{k=n}^{\infty} \bar{A}_k\right) = 0$$

for all n, and therefore $\mathsf{P}(A_n \text{ i.o.}) = 1$.

This completes the proof of the lemma.

Corollary 1. *If* $A_n^\varepsilon = \{\omega: |\xi_n - \xi| \ge \varepsilon\}$ *then* (8) *shows that* $\sum_{n=1}^\infty \mathsf{P}(A_n) < \infty$, $\varepsilon > 0$, *and then by the Borel–Cantelli lemma we have* $\mathsf{P}(A^\varepsilon) = 0$, $\varepsilon > 0$, *where* $A^\varepsilon = \overline{\lim}\, A_n^\varepsilon$. *Therefore*

$$\sum \mathsf{P}\{|\xi_k - \xi| \ge \varepsilon\} < \infty, \varepsilon > 0 \Rightarrow \mathsf{P}(A^\varepsilon) = 0, \varepsilon > 0$$
$$\Rightarrow \mathsf{P}\{\omega: \xi_n \not\to \xi)\} = 0,$$

as we already observed above.

Corollary 2. *Let* $(\varepsilon_n)_{n\ge 1}$ *be a sequence of positive numbers such that* $\varepsilon_n \downarrow 0$, $n \to \infty$. *If*

$$\sum_{n=1}^\infty \mathsf{P}\{|\xi_n - \xi| \ge \varepsilon_n\} < \infty, \tag{10}$$

then $\xi_n \xrightarrow{\text{a.s.}} \xi$.

In fact, let $A_n = \{|\xi_n - \xi| \ge \varepsilon_n\}$. Then $\mathsf{P}(A_n \text{ i.o.}) = 0$ by the Borel–Cantelli lemma. This means that, for almost every $\omega \in \Omega$, there is an $N = N(\omega)$ such that $|\xi_n(\omega) - \xi(\omega)| \le \varepsilon_n$ for $n \ge N(\omega)$. But $\varepsilon_n \downarrow 0$, and therefore $\xi_n(\omega) \to \xi(\omega)$ for almost every $\omega \in \Omega$.

4. Theorem 2. *We have the following implications:*

$$\xi_n \xrightarrow{\text{a.s.}} \xi \Rightarrow \xi_n \xrightarrow{\mathsf{P}} \xi, \tag{11}$$

$$\xi_n \xrightarrow{L^p} \xi \Rightarrow \xi_n \xrightarrow{\mathsf{P}} \xi, \qquad p > 0, \tag{12}$$

$$\xi_n \xrightarrow{\mathsf{P}} \xi \Rightarrow \xi_n \xrightarrow{d} \xi. \tag{13}$$

PROOF. Statement (11) follows from comparing the definition of convergence in probability with (5), and (12) follows from Chebyshev's inequality.

To prove (13), let $f(x)$ be a continuous function, let $|f(x)| \le c$, let $\varepsilon > 0$, and let N be such that $\mathsf{P}(|\xi| > N) \le \varepsilon/4c$. Take δ so that $|f(x) - f(y)| \le \varepsilon/2c$ for $|x| < N$ and $|x - y| \le \delta$. Then (cf. the proof of Weierstrass's theorem in Subsection 5, §5, Chapter I)

$$\begin{aligned}
\mathsf{E}|f(\xi_n) - f(\xi)| &= \mathsf{E}(|f(\xi_n) - f(\xi)|; |\xi_n - \xi| \le \delta, |\xi| \le N)\\
&\quad + \mathsf{E}(|f(\xi_n) - f(\xi)|; |\xi_n - \xi| \le \delta, |\xi| > N)\\
&\quad + \mathsf{E}(|f(\xi_n) - f(\xi)|; |\xi_n - \xi| > \delta)\\
&\le \varepsilon/2 + \varepsilon/2 + 2c\mathsf{P}\{|\xi_n - \xi| > \delta\}\\
&= \varepsilon + 2c\mathsf{P}\{|\xi_n - \xi| > \delta\}.
\end{aligned}$$

But $\mathsf{P}\{|\xi_n - \xi| > \delta\} \to 0$, and hence $\mathsf{E}|f(\xi_n) - f(\xi)| \le 2\varepsilon$ for sufficiently large n; since $\varepsilon > 0$ is arbitrary, this establishes (13).

This completes the proof of the theorem.

We now present a number of examples which show, in particular, that the converses of (11) and (12) are false in general.

EXAMPLE 1 ($\xi_n \xrightarrow{P} \xi \not\Rightarrow \xi_n \xrightarrow{\text{a.s.}} \xi$; $\xi_n \xrightarrow{L^p} \xi \not\Rightarrow \xi_n \xrightarrow{\text{a.s.}} \xi$). Let $\Omega = [0, 1]$, $\mathscr{F} = \mathscr{B}([0, 1])$, P = Lebesgue measure. Put

$$A_n^i = \left[\frac{i-1}{n}, \frac{i}{n}\right], \qquad \xi_n^i = I_{A_n^i}(\omega), \qquad i = 1, 2, \ldots, n;\ n \geq 1.$$

Then the sequence

$$\{\xi_1^1;\ \xi_2^1, \xi_2^2;\ \xi_3^1, \xi_3^2, \xi_3^3; \cdots\}$$

of random variables converges both in probability and in mean of order $p > 0$, but does not converge at any point $\omega \in [0, 1]$.

EXAMPLE 2 ($\xi_n \xrightarrow{\text{a.s.}} \xi \Rightarrow \xi_n \xrightarrow{P} \xi \not\Rightarrow \xi_n \xrightarrow{L^p} \xi$, $p > 0$). Again let $\Omega = [0, 1]$, $\mathscr{F} = \mathscr{B}[0, 1]$, P = Lebesgue measure, and let

$$\xi_n(\omega) = \begin{cases} e^n, & 0 \leq \omega \leq 1/n, \\ 0, & \omega > 1/n. \end{cases}$$

Then $\{\xi_n\}$ converges with probability 1 (and therefore in probability) to zero, but

$$\mathsf{E}|\xi_n|^p = \frac{e^{np}}{n} \to \infty, \qquad n \to \infty,$$

for every $p > 0$.

EXAMPLE 3 ($\xi_n \xrightarrow{L^p} \xi \not\Rightarrow \xi_n \xrightarrow{\text{a.s.}} \xi$). Let $\{\xi_n\}$ be a sequence of independent random variables with

$$\mathsf{P}(\xi_n = 1) = p_n, \qquad \mathsf{P}(\xi_n = 0) = 1 - p_n.$$

Then it is easy to show that

$$\xi_n \xrightarrow{P} 0 \Leftrightarrow p_n \to 0, \qquad n \to \infty, \tag{14}$$

$$\xi_n \xrightarrow{L^p} 0 \Leftrightarrow p_n \to 0, \qquad n \to \infty, \tag{15}$$

$$\xi_n \xrightarrow{\text{a.s.}} 0 \Rightarrow \sum_{n=1}^{\infty} p_n < \infty. \tag{16}$$

In particular, if $p_n = 1/n$ then $\xi_n \xrightarrow{L^p} 0$ for every $p > 0$, but $\xi_n \overset{\text{a.s.}}{\not\to} 0$.

The following theorem singles out an interesting case when almost sure convergence implies convergence in L^1.

Theorem 3. *Let (ξ_n) be a sequence of nonnegative random variables such that $\xi_n \xrightarrow{\text{a.s.}} \xi$ and $\mathsf{E}\xi_n \to \mathsf{E}\xi < \infty$. Then*

$$\mathsf{E}|\xi_n - \xi| \to 0, \qquad n \to \infty. \tag{17}$$

PROOF. We have $\mathsf{E}\xi_n < \infty$ for sufficiently large n, and therefore for such n we have

$$\begin{aligned}\mathsf{E}|\xi - \xi_n| &= \mathsf{E}(\xi - \xi_n)I_{\{\xi \geq \xi_n\}} + \mathsf{E}(\xi_n - \xi)I_{\{\xi_n > \xi\}} \\ &= 2\mathsf{E}(\xi - \xi_n)I_{\{\xi \geq \xi_n\}} + \mathsf{E}(\xi_n - \xi).\end{aligned}$$

But $0 \leq (\xi - \xi_n)I_{\{\xi \geq \xi_n\}} \leq \xi$. Therefore, by the dominated convergence theorem, $\lim_n \mathsf{E}(\xi - \xi_n)I_{\{\xi \geq \xi_n\}} = 0$, which together with $\mathsf{E}\xi_n \to \mathsf{E}\xi$ proves (17).

Remark. The dominated convergence theorem also holds when almost sure convergence is replaced by convergence in probability (see Problem 1). Hence in Theorem 3 we may replace "$\xi_n \xrightarrow{\text{a.s.}} \xi$" by "$\xi_n \xrightarrow{\mathsf{P}} \xi$."

5. It is shown in analysis that every fundamental sequence (x_n), $x_n \in R$, is convergent (Cauchy criterion). Let us give a similar result for the convergence of a sequence of random variables.

Theorem 4 (Cauchy Criterion for Almost Sure Convergence). *A necessary and sufficient condition for the sequence $(\xi_n)_{n\geq 1}$ of random variables to converge with probability* 1 (*to a random variable ξ*) *is that it is fundamental with probability* 1.

PROOF. If $\xi_n \xrightarrow{\text{a.s.}} \xi$ then

$$\sup_{\substack{k\geq n\\ l\geq n}} |\xi_k - \xi_l| \leq \sup_{k\geq n} |\xi_k - \xi| + \sup_{l\geq n} |\xi_l - \xi|,$$

whence the necessity follows.

Now let $(\xi_n)_{n\geq 1}$ be fundamental with probability 1. Let $\mathcal{N} = \{\omega: (\xi_n(\omega))$ *is not fundamental*$\}$. Then whenever $\omega \in \Omega \setminus \mathcal{N}$ the sequence of numbers $(\xi_n(\omega))_{n\geq 1}$ is fundamental and, by Cauchy's criterion for sequences of numbers, $\lim \xi_n(\omega)$ exists. Let

$$\xi(\omega) = \begin{cases}\lim \xi_n(\omega), & \omega \in \Omega \setminus \mathcal{N}, \\ 0, & \omega \in \mathcal{N}.\end{cases} \tag{18}$$

The function so defined is a random variable, and evidently $\xi_n \xrightarrow{\text{a.s.}} \xi$.

This completes the proof.

Before considering the case of convergence in probability, let us establish the following useful result.

Theorem 5. *If the sequence (ξ_n) is fundamental* (*or convergent*) *in probability, it contains a subsequence (ξ_{n_k}) that is fundamental* (*or convergent*) *with probability* 1.

PROOF. Let (ξ_n) be fundamental in probability. By Theorem 4, it is enough to show that it contains a subsequence that converges almost surely.

Take $n_1 = 1$ and define n_k inductively as the smallest $n > n_{k-1}$ for which

$$P\{|\xi_t - \xi_s| > 2^{-k}\} < 2^{-k}.$$

for all $s \geq n$, $t \geq n$. Then

$$\sum_k P\{|\xi_{n_{k+1}} - \xi_{n_k}| > 2^{-k}\} < \sum 2^{-k} < \infty$$

and by the Borel–Cantelli lemma

$$P\{|\xi_{n_{k+1}} - \xi_{n_k}| > 2^{-k} \text{ i.o.}\} = 0.$$

Hence

$$\sum_{k=1}^{\infty} |\xi_{n_{k+1}} - \xi_{n_k}| < \infty$$

with probability 1.

Let $\mathcal{N} = \{\omega: \sum |\xi_{n_{k+1}} - \xi_{n_k}| = \infty\}$. Then if we put

$$\xi(\omega) = \begin{cases} \xi_{n_1}(\omega) + \sum_{k=1}^{\infty} (\xi_{n_{k+1}}^{(\omega)} - \xi_{n_k}(\omega)), & \omega \in \Omega \backslash \mathcal{N}, \\ 0, & \omega \in \mathcal{N}, \end{cases}$$

we obtain $\xi_{n_k} \xrightarrow{\text{a.s.}} \xi$.

If the original sequence converges in probability, then it is fundamental in probability (see also (19)), and consequently this case reduces to the one already considered.

This completes the proof of the theorem.

Theorem 6 (Cauchy Criterion for Convergence in Probability). *A necessary and sufficient condition for a sequence $(\xi_n)_{n \geq 1}$ of random variables to converge in probability is that it is fundamental in probability.*

PROOF. If $\xi_n \xrightarrow{P} \xi$ then

$$P\{|\xi_n - \xi_m| \geq \varepsilon\} \leq P\{|\xi_n - \xi| \geq \varepsilon/2\} + P\{|\xi_m - \xi| \geq \varepsilon/2\} \quad (19)$$

and consequently (ξ_n) is fundamental in probability.

Conversely, if (ξ_n) is fundamental in probability, by Theorem 5 there are a subsequence (ξ_{n_k}) and a random variable ξ such that $\xi_{n_k} \xrightarrow{\text{a.s.}} \xi$. But then

$$P\{|\xi_n - \xi| \geq \varepsilon\} \leq P\{|\xi_n - \xi_{n_k}| \geq \varepsilon/2\} + P\{|\xi_{n_k} - \xi| \geq \varepsilon/2\},$$

from which it is clear that $\xi_n \xrightarrow{P} \xi$. This completes the proof.

Before discussing convergence in mean of order p, we make some observations about L^p spaces.

We denote by $L^p = L^p(\Omega, \mathscr{F}, \mathsf{P})$ the space of random variables $\xi = \xi(\omega)$ with $\mathsf{E}|\xi|^p \equiv \int_\Omega |\xi|^p \, d\mathsf{P} < \infty$. Suppose that $p \geq 1$ and put

$$\|\xi\|_p = (\mathsf{E}|\xi|^p)^{1/p}.$$

It is clear that

$$\|\xi\|_p \geq 0, \tag{20}$$

$$\|c\xi\|_p = |c| \, \|\xi\|_p, \quad c \text{ constant}, \tag{21}$$

and by Minkowski's inequality (6.31)

$$\|\xi + \eta\|_p \leq \|\xi\|_p + \|\eta\|_p. \tag{22}$$

Hence, in accordance with the usual terminology of functional analysis, the function $\|\cdot\|_p$, defined on L^p and satisfying (20)–(22), is (for $p \geq 1$) a *semi-norm*.

For it to be a *norm*, it must also satisfy

$$\|\xi\|_p = 0 \Rightarrow \xi = 0. \tag{23}$$

This property is, of course, not satisfied, since according to Property H (§6) we can only say that $\xi = 0$ almost surely.

This fact leads to a somewhat different view of the space L^p. That is, we connect with every random variable $\xi \in L^p$ the class $[\xi]$ of random variables in L^p that are equivalent to it (ξ and η are *equivalent* if $\xi = \eta$ almost surely). It is easily verified that the property of equivalence is *reflexive*, *symmetric*, and *transitive*, and consequently the linear space L^p can be divided into disjoint equivalence classes of random variables. If we now think of $[L^p]$ as the collection of the classes $[\xi]$ of equivalent random variables $\xi \in L^p$, and define

$$[\xi] + [\eta] = [\xi + \eta],$$

$$a[\xi] = [a\xi], \qquad \text{where } a \text{ is a constant},$$

$$\|[\xi]\|_p = \|\xi\|_p,$$

then $[L^p]$ becomes a normed linear space.

In functional analysis, we ordinarily describe elements of a space $[L^p]$, not as *equivalence classes of functions*, but simply as *functions*. In the same way we do not actually use the notation $[L^p]$. From now on, we no longer think about sets of equivalence classes of functions, but simply about elements, functions, random variables, and so on.

It is a basic result of functional analysis that the spaces L^p, $p \geq 1$, are *complete*, i.e. that every fundamental sequence has a limit. Let us state and prove this in probabilistic language.

Theorem 7 (Cauchy Test for Convergence in Mean pth Power). *A necessary and sufficient condition that a sequence $(\xi_n)_{n \geq 1}$ of random variables in L^p*

convergences in mean of order p to a random variable in L^p is that the sequence is fundamental in mean of order p.

PROOF. The necessity follows from Minkowski's inequality. Let (ξ_n) be fundamental ($\|\xi_n - \xi_m\|_p \to 0$, $n, m \to \infty$). As in the proof of Theorem 5, we select a subsequence (ξ_{n_k}) such that $\xi_{n_k} \xrightarrow{\text{a.s.}} \xi$, where ξ is a random variable with $\|\xi\|_p < \infty$.

Let $n_1 = 1$ and define n_k inductively as the smallest $n > n_{k-1}$ for which

$$\|\xi_t - \xi_s\|_p < 2^{-2k}$$

for all $s \geq n$, $t \geq n$. Let

$$A_k = \{\omega: |\xi_{n_{k+1}} - \xi_{n_k}| \geq 2^{-k}\}.$$

Then by Chebyshev's inequality

$$\mathsf{P}(A_k) \leq \frac{\mathsf{E}|\xi_{n_{k+1}} - \xi_{n_k}|^r}{2^{-kr}} \leq \frac{2^{-2kr}}{2^{-kr}} = 2^{-kr} \leq 2^{-k}.$$

As in Theorem 5, we deduce that there is a random variable ξ such that $\xi_{n_k} \xrightarrow{\text{a.s.}} \xi$.

We now deduce that $\|\xi_n - \xi\|_p \to 0$ as $n \to \infty$. To do this, we fix $\varepsilon > 0$ and choose $N = N(\varepsilon)$ so that $\|\xi_n - \xi_m\|_p^p < \varepsilon$ for all $n \geq N$, $m \geq N$. Then for any fixed $n \geq N$, by Fatou's lemma,

$$\begin{aligned}\mathsf{E}|\xi_n - \xi|^p &= \mathsf{E}\left\{\lim_{n_k \to \infty} |\xi_n - \xi_{n_k}|^p\right\} = \mathsf{E}\left\{\varliminf_{n_k \to \infty} |\xi_n - \xi_{n_k}|^p\right\} \\ &\leq \varliminf_{n_k \to \infty} \mathsf{E}|\xi_n - \xi_{n_k}|^p = \varliminf_{n_k \to \infty} \|\xi_n - \xi_{n_k}\|_p^p \leq \varepsilon.\end{aligned}$$

Consequently $\mathsf{E}|\xi_n - \xi|^p \to 0$, $n \to \infty$. It is also clear that since $\xi = (\xi - \xi_n) + \xi_n$ we have $\mathsf{E}|\xi|^p < \infty$ by Minkowski's inequality.

This completes the proof of the theorem.

Remark 1. In the terminology of functional analysis a complete normed linear space is called a *Banach space*. Thus L^p, $p \geq 1$, is a Banach space.

Remark 2. If $0 < p < 1$, the function $\|\xi\|_p = (E|\xi|^p)^{1/p}$ does not satisfy the triangle inequality (22) and consequently is not a norm. Nevertheless the space (of equivalence classes) L^p, $0 < p < 1$, is complete in the metric $d(\xi, \eta) \equiv \mathsf{E}|\xi - \eta|^p$.

Remark 3. Let $L^\infty = L^\infty(\Omega, \mathscr{F}, \mathsf{P})$ be the space (of equivalence classes of) random variables $\xi = \xi(\omega)$ for which $\|\xi\|_\infty < \infty$, where $\|\xi\|_\infty$, the *essential supremum* of ξ, is defined by

$$\|\xi\|_\infty \equiv \operatorname{ess\,sup}|\xi| \equiv \inf\{0 \leq c \leq \infty: \mathsf{P}(|\xi| > c) = 0\}.$$

The function $\|\cdot\|_\infty$ is a norm, and L^∞ is complete in this norm.

6. Problems

1. Use Theorem 5 to show that almost sure convergence can be replaced by convergence in probability in Theorems 3 and 4 of §6.

2. Prove that L^∞ is complete.

3. Show that if $\xi_n \xrightarrow{P} \xi$ and also $\xi_n \xrightarrow{P} \eta$ then ξ and η are equivalent ($\mathsf{P}(\xi \neq \eta) = 0$).

4. Let $\xi_n \xrightarrow{P} \xi$, $\eta_n \xrightarrow{P} \eta$, and let ξ and η be equivalent. Show that

$$\mathsf{P}\{|\xi_n - \eta_n| \geq \varepsilon\} \to 0, \qquad n \to \infty,$$

for every $\varepsilon > 0$.

5. Let $\xi_n \xrightarrow{P} \xi$, $\eta_n \xrightarrow{P} \eta$. Show that $a\xi_n + b\eta_n \xrightarrow{P} a\xi + b\eta$ (a, b constants), $|\xi_n| \xrightarrow{P} |\xi|$, $\xi_n \eta_n \xrightarrow{P} \xi\eta$.

6. Let $(\xi_n - \xi)^2 \to 0$. Show that $\xi_n^2 \to \xi^2$.

7. Show that if $\xi_n \xrightarrow{d} C$, where C is a constant, then this sequence converges in probability:

$$\xi_n \xrightarrow{d} C \Rightarrow \xi_n \xrightarrow{P} C.$$

8. Let $(\xi_n)_{n \geq 1}$ have the property that $\sum_{n=1}^{\infty} \mathsf{E}|\xi_n|^p < \infty$ for some $p > 0$. Show that $\xi_n \to 0$ (P-a.s.).

9. Let $(\xi_n)_{n \geq 1}$ be a sequence of independent identically distributed random variables. Show that

$$\mathsf{E}|\xi_1| < \infty \Leftrightarrow \sum_{n=1}^{\infty} \mathsf{P}\{|\xi_1| > \varepsilon \cdot n\} < \infty$$

$$\Leftrightarrow \sum_{n=1}^{\infty} \mathsf{P}\left\{\left|\frac{\xi_n}{n}\right| > \varepsilon\right\} < \infty \Rightarrow \frac{\xi_n}{n} \to 0 \quad \text{(P-a.s.)}.$$

10. Let $(\xi_n)_{n \geq 1}$ be a sequence of random variables. Suppose that there are a random variable ξ and a sequence $\{n_k\}$ such that $\xi_{n_k} \to \xi$ (P-a.s.) and $\max_{n_{k-1} < l \leq n_k} |\xi_l - \xi_{n_{k-1}}| \to 0$ (P-a.s.) as $k \to \infty$. Show that then $\xi_n \to \xi$ (P-a.s.).

11. Let the d-metric on the set of random variables be defined by

$$d(\xi, \eta) = \mathsf{E} \frac{|\xi - \eta|}{1 + |\xi - \eta|}$$

and identify random variables that coincide almost surely. Show that convergence in probability is equivalent to convergence in the d-metric.

12. Show that there is no metric on the set of random variables such that convergence in that metric is equivalent to almost sure convergence.

§11. The Hilbert Space of Random Variables with Finite Second Moment

1. An important role among the Banach spaces L^p, $p \geq 1$, is played by the space $L^2 = L^2(\Omega, \mathscr{F}, \mathsf{P})$, the space of (equivalence classes of) random variables with finite second moments.

If ξ and $\eta \in L^2$, we put

$$(\xi, \eta) \equiv \mathsf{E}\xi\eta. \tag{1}$$

It is clear that if $\xi, \eta, \zeta \in L^2$ then

$$(a\xi + b\eta, \zeta) = a(\xi, \zeta) + b(\eta, \zeta), \qquad a, b \in R,$$
$$(\xi, \xi) \geq 0$$

and

$$(\xi, \xi) = 0 \Leftrightarrow \xi = 0.$$

Consequently (ξ, η) is a *scalar product*. The space L^2 is *complete* with respect to the norm

$$\|\xi\| = (\xi, \xi)^{1/2} \tag{2}$$

induced by this scalar product (as was shown in §10). In accordance with the terminology of functional analysis, a space with the scalar product (1) is a *Hilbert space.*

Hilbert space methods are extensively used in probability theory to study properties that depend only on the first two moments of random variables ("L^2-theory"). Here we shall introduce the basic concepts and facts that will be needed for an exposition of L^2-theory (Chapter VI).

2. Two random variables ξ and η in L^2 are said to be orthogonal ($\xi \perp \eta$) if $(\xi, \eta) \equiv \mathsf{E}\xi\eta = 0$. According to §8, ξ and η are uncorrelated if $\operatorname{cov}(\xi, \eta) = 0$, i.e. if

$$\mathsf{E}\xi\eta = \mathsf{E}\xi\mathsf{E}\eta.$$

It follows that the properties of being orthogonal and of being uncorrelated coincide for random variables with zero mean values.

A set $M \subseteq L^2$ is a *system of orthogonal random variables* if $\xi \perp \eta$ for every $\xi, \eta \in M$ ($\xi \neq \eta$).

If also $\|\xi\| = 1$ for every $\xi \in M$, then M is an *orthonormal system.*

3. Let $M = \{\eta_1, \ldots, \eta_n\}$ be an orthonormal system and ξ any random variable in L^2. Let us find, in the class of linear estimators $\sum_{i=1}^n a_i\eta_i$, the best mean-square estimator for ξ (cf. Subsection 2, §8).

A simple computation shows that

$$\begin{aligned}
\mathsf{E}\left|\xi - \sum_{i=1}^n \alpha_i\eta_i\right|^2 \equiv \left\|\xi - \sum_{i=1}^n a_i\eta_i\right\|^2 &= \left(\xi - \sum_{i=1}^n a_i\eta_i, \xi - \sum_{i=1}^n a_i\eta_i\right) \\
&= \|\xi\|^2 - 2\sum_{i=1}^n a_i(\xi, \eta_i) + \left(\sum_{i=1}^n a_i\eta_i, \sum_{i=1}^n a_i\eta_i\right) \\
&= \|\xi\|^2 - 2\sum_{i=1}^n a_i(\xi, \eta_i) + \sum_{i=1}^n a_i^2
\end{aligned}$$

$$= \|\xi\|^2 - \sum_{i=1}^{n} |(\xi, \eta_i)|^2 + \sum_{i=1}^{n} |a_i - (\xi, \eta_i)|^2$$

$$\geq \|\xi\|^2 - \sum_{i=1}^{n} |(\xi, \eta_i)|^2, \tag{3}$$

where we used the equation

$$a_i^2 - 2a_i(\xi, \eta_i) = |a_i - (\xi, \eta_i)|^2 - |(\xi, \eta_i)|^2.$$

It is now clear that the infimum of $\mathsf{E}|\xi - \sum_{i=1}^n a_i\eta_i|^2$ over all real $a_1, \ldots, a_n$ is attained for $a_i = (\xi, \eta_i)$, $i = 1, \ldots, n$.

Consequently the best (in the mean-square sense) estimator for ξ in terms of $\eta_1, \ldots, \eta_n$ is

$$\hat{\xi} = \sum_{i=1}^{n} (\xi, \eta_i)\eta_i. \tag{4}$$

Here

$$\Delta \equiv \inf \mathsf{E}\left|\xi - \sum_{i=1}^{n} a_i\eta_i\right|^2 = \mathsf{E}|\xi - \hat{\xi}|^2 = \|\xi\|^2 - \sum_{i=1}^{n} |(\xi, \eta_i)|^2 \tag{5}$$

(compare (I.4.17) and (8.13)).

Inequality (3) also implies *Bessel's inequality*: if $M = \{\eta_1, \eta_2, \ldots\}$ is an orthonormal system and $\xi \in L^2$, then

$$\sum_{i=1}^{\infty} |(\xi, \eta_i)|^2 \leq \|\xi\|^2; \tag{6}$$

and equality is attained if and only if

$$\xi = \underset{n}{\text{l.i.m.}} \sum_{i=1}^{n} (\xi, \eta_i)\eta_i. \tag{7}$$

The *best linear estimator* of ξ is often denoted by $\hat{\mathsf{E}}(\xi | \eta_1, \ldots, \eta_n)$ and called the *conditional expectation* (of ξ with respect to $\eta_1, \ldots, \eta_n$) in the *wide sense*.

The reason for the terminology is as follows. If we consider all estimators $\varphi = \varphi(\eta_1, \ldots, \eta_n)$ of ξ in terms of $\eta_1, \ldots, \eta_n$ (where φ is a Borel function), the best estimator will be $\varphi^* = \mathsf{E}(\xi | \eta_1, \ldots, \eta_n)$, i.e. the conditional expectation of ξ with respect to $\eta_1, \ldots, \eta_n$ (cf. Theorem 1, §8). Hence the best linear estimator is, by analogy, denoted by $\hat{\mathsf{E}}(\xi | \eta_1, \ldots, \eta_n)$ and called the conditional expectation in the wide sense. We note that if $\eta_1, \ldots, \eta_n$ form a Gaussian system (see §13 below), then $\mathsf{E}(\xi | \eta_1, \ldots, \eta_n)$ and $\hat{\mathsf{E}}(\xi | \eta_1, \ldots, \eta_n)$ are the same.

Let us discuss the *geometric meaning* of $\hat{\xi} = \hat{\mathsf{E}}(\xi | \eta_1, \ldots, \eta_n)$.

Let $\mathscr{L} = \mathscr{L}\{\eta_1, \ldots, \eta_n\}$ denote the *linear manifold* spanned by the orthonormal system of random variables $\eta_1, \ldots, \eta_n$ (i.e., the set of random variables of the form $\sum_{i=1}^n a_i\eta_i$, $a_i \in R$).

Then it follows from the preceding discussion that ξ admits the "orthogonal decomposition"

$$\xi = \hat{\xi} + (\xi - \hat{\xi}), \tag{8}$$

where $\hat{\xi} \in \mathscr{L}$ and $\xi - \hat{\xi} \perp \mathscr{L}$ in the sense that $\xi - \hat{\xi} \perp \lambda$ for every $\lambda \in \mathscr{L}$. It is natural to call $\hat{\xi}$ the *projection* of ξ on $\mathscr{L}$ (the element of $\mathscr{L}$ "closest" to ξ), and to say that $\xi - \hat{\xi}$ is *perpendicular* to $\mathscr{L}$.

4. The concept of orthonormality of the random variables $\eta_1, \ldots, \eta_n$ makes it easy to find the best linear estimator (the projection) $\hat{\xi}$ of ξ in terms of $\eta_1, \ldots, \eta_n$. The situation becomes complicated if we give up the hypothesis of orthonormality. However, the case of arbitrary $\eta_1, \ldots, \eta_n$ can in a certain sense be reduced to the case of orthonormal random variables, as will be shown below. We shall suppose for the sake of simplicity that all our random variables have zero mean values.

We shall say that the random variables $\eta_1, \ldots, \eta_n$ are *linearly independent* if the equation

$$\sum_{i=1}^{n} a_i \eta_i = 0 \quad (\text{P-a.s.})$$

is satisfied only when all a_i are zero.

Consider the covariance matrix

$$\mathbb{R} = \mathsf{E}\eta\eta^{\mathsf{T}}$$

of the vector $\eta = (\eta_1, \ldots, \eta_n)$. It is symmetric and nonnegative definite, and as noticed in §8, can be diagonalized by an orthogonal matrix $\mathcal{O}$:

$$\mathcal{O}^{\mathsf{T}} \mathbb{R} \mathcal{O} = D,$$

where

$$D = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{pmatrix}$$

has nonnegative elements d_i, the eigenvalues of $\mathbb{R}$, i.e. the zeros λ of the characteristic equation $\det(\mathbb{R} - \lambda E) = 0$.

If $\eta_1, \ldots, \eta_n$ are linearly independent, the Gram determinant ($\det \mathbb{R}$) is not zero and therefore $d_i > 0$. Let

$$B = \begin{pmatrix} \sqrt{d_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{d_n} \end{pmatrix}$$

and

$$\beta = B^{-1} \mathcal{O}^{\mathsf{T}} \eta. \tag{9}$$

Then the covariance matrix of β is

$$\mathsf{E}\beta\beta^{\mathsf{T}} = B^{-1} \mathcal{O}^{\mathsf{T}} \mathsf{E}\eta\eta^{\mathsf{T}} \mathcal{O} B^{-1} = B^{-1} \mathcal{O}^{\mathsf{T}} \mathbb{R} \mathcal{O} B^{-1} = E,$$

and therefore $\beta = (\beta_1, \ldots, \beta_n)$ consists of uncorrelated random variables.

It is also clear that

$$\eta = (\mathcal{O}B)\beta. \tag{10}$$

Consequently if $\eta_1, \ldots, \eta_n$ are linearly independent there is an orthonormal system such that (9) and (10) hold. Here

$$\mathscr{L}\{\eta_1, \ldots, \eta_n\} = \mathscr{L}\{\beta_1, \ldots, \beta_n\}.$$

This method of constructing an orthonormal system $\beta_1, \ldots, \beta_n$ is frequently inconvenient. The reason is that if we think of η_i as the value of the random sequence $(\eta_1, \ldots, \eta_n)$ at the instant i, the value β_i constructed above depends not only on the "past," $(\eta_1, \ldots, \eta_i)$, but also on the "future," $(\eta_{i+1}, \ldots, \eta_n)$. The *Gram–Schmidt orthogonalization process*, described below, does not have this defect, and moreover has the advantage that it can be applied to an infinite sequence of *linearly independent* random variables (i.e. to a sequence in which every finite set of the variables are linearly independent).

Let $\eta_1, \eta_2, \ldots$ be a sequence of linearly independent random variables in L^2. We construct a sequence $\varepsilon_1, \varepsilon_2, \ldots$ as follows. Let $\varepsilon_1 = \eta_1/\|\eta_1\|$. If $\varepsilon_1, \ldots, \varepsilon_{n-1}$ have been selected so that they are orthonormal, then

$$\varepsilon_n = \frac{\eta_n - \hat{\eta}_n}{\|\eta_n - \hat{\eta}_n\|}, \tag{11}$$

where $\hat{\eta}_n$ is the projection of η_n on the linear manifold $\mathscr{L}(\varepsilon_1, \ldots, \varepsilon_{n-1})$ generated by

$$\hat{\eta}_n = \sum_{k=1}^{n-1} (\eta_n, \varepsilon_k)\varepsilon_k. \tag{12}$$

Since $\eta_1, \ldots, \eta_n$ are linearly independent and $\mathscr{L}\{\eta_1, \ldots, \eta_{n-1}\} = \mathscr{L}\{\varepsilon_1, \ldots, \varepsilon_{n-1}\}$, we have $\|\eta_n - \hat{\eta}_n\| > 0$ and consequently ε_n is well defined.

By construction, $\|\varepsilon_n\| = 1$ for $n \geq 1$, and it is clear that $(\varepsilon_n, \varepsilon_k) = 0$ for $k < n$. Hence the sequence $\varepsilon_1, \varepsilon_2, \ldots$ is orthonormal. Moreover, by (11),

$$\eta_n = \hat{\eta}_n + b_n\varepsilon_n,$$

where $b_n = \|\eta_n - \hat{\eta}_n\|$ and $\hat{\eta}_n$ is defined by (12).

Now let $\eta_1, \ldots, \eta_n$ be any set of random variables (not necessarily linearly independent). Let $\det \mathbb{R} = 0$, where $\mathbb{R} \equiv \|r_{ij}\|$ is the covariance matrix of $(\eta_1, \ldots, \eta_n)$, and let

$$\operatorname{rank} \mathbb{R} = r < n.$$

Then, from linear algebra, the quadratic form

$$Q(a) = \sum_{i,j=1}^{n} r_{ij}a_i a_j, \qquad a = (a_1, \ldots, a_n),$$

has the property that there are $n - r$ linearly independent vectors $a^{(1)}, \ldots, a^{(n-r)}$ such that $Q(a^{(i)}) = 0$, $i = 1, \ldots, n - r$.

But

$$Q(a) = \mathsf{E}\left(\sum_{k=1}^{n} a_k \eta_k\right)^2.$$

Consequently

$$\sum_{k=1}^{n} a_k^{(i)} \eta_k = 0, \qquad i = 1, \ldots, n - r,$$

with probability 1.

In other words, there are $n - r$ linear relations among the variables $\eta_1, \ldots, \eta_n$. Therefore if, for example, $\eta_1, \ldots, \eta_r$ are linearly independent, the other variables $\eta_{r+1}, \ldots, \eta_n$ can be expressed linearly in terms of them, and consequently $\mathscr{L}\{\eta_1, \ldots, \eta_n\} = \mathscr{L}\{\varepsilon_1, \ldots, \varepsilon_r\}$. Hence it is clear that we can find r orthonormal random variables $\varepsilon_1, \ldots, \varepsilon_r$ such that $\eta_1, \ldots, \eta_n$ can be expressed linearly in terms of them and $\mathscr{L}\{\eta_1, \ldots, \eta_n\} = \mathscr{L}\{\varepsilon_1, \ldots, \varepsilon_r\}$.

5. Let $\eta_1, \eta_2, \ldots$ be a sequence of random variables in L^2. Let $\bar{\mathscr{L}} = \bar{\mathscr{L}}\{\eta_1, \eta_2, \ldots\}$ be the *linear manifold* spanned by $\eta_1, \eta_2, \ldots$, i.e. the set of random variables of the form $\sum_{i=1}^{n} a_i \eta_i$, $n \geq 1$, $a_i \in R$. Then $\bar{\mathscr{L}} = \bar{\mathscr{L}}\{\eta_1, \eta_2, \ldots\}$ denotes the *closed linear manifold* spanned by $\eta_1, \eta_2, \ldots$, i.e. the set of random variables in $\mathscr{L}$ together with their mean-square limits.

We say that a set $\eta_1, \eta_2, \ldots$ is a *countable orthonormal basis* (or a *complete orthonormal system*) if:

(a) $\eta_1, \eta_2, \ldots$ is an orthonormal system,
(b) $\bar{\mathscr{L}}\{\eta_1, \eta_2, \ldots\} = L^2$.

A Hilbert space with a countable orthonormal basis is said to be *separable*.

By (b), for every $\xi \in L^2$ and a given $\varepsilon > 0$ there are numbers $a_1, \ldots, a_n$ such that

$$\left\| \xi - \sum_{i=1}^{n} a_i \eta_i \right\| \leq \varepsilon.$$

Then by (3)

$$\left\| \xi - \sum_{i=1}^{n} (\xi, \eta_i) \eta_i \right\| \leq \varepsilon.$$

Consequently every element of a separable Hilbert space L^2 can be represented as

$$\xi = \sum_{i=1}^{\infty} (\xi, \eta_i) \cdot \eta_i, \tag{13}$$

or more precisely as

$$\xi = \operatorname*{l.i.m.}_{n} \sum_{i=1}^{n} (\xi, \eta_i) \eta_i.$$

We infer from this and (3) that *Parseval's equation* holds:

$$\|\xi\|^2 = \sum_{i=1}^{\infty} |(\xi, \eta_i)|^2, \qquad \xi \in L^2. \tag{14}$$

It is easy to show that the converse is also valid: if $\eta_1, \eta_2, \ldots$ is an orthonormal system and either (13) or (14) is satisfied, then the system is a basis.

We now give some examples of separable Hilbert spaces and their bases.

EXAMPLE 1. Let $\Omega = R$, $\mathscr{F} = \mathscr{B}(R)$, and let P be the Gaussian measure,

$$\mathsf{P}(-\infty, a] = \int_{-\infty}^{a} \varphi(x)\, dx, \qquad \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Let $D = d/dx$ and

$$H_n(x) = \frac{(-1)^n D^n \varphi(x)}{\varphi(x)}, \qquad n \geq 0. \tag{15}$$

We find easily that

$$\begin{aligned} D\varphi(x) &= -x\varphi(x), \\ D^2\varphi(x) &= (x^2 - 1)\varphi(x), \\ D^3\varphi(x) &= (3x - x^3)\varphi(x), \\ &\cdots\cdots\cdots\cdots \end{aligned} \tag{16}$$

It follows that $H_n(x)$ are polynomials (the *Hermite polynomials*). From (15) and (16) we find that

$$\begin{aligned} H_0(x) &= 1, \\ H_1(x) &= x, \\ H_2(x) &= x^2 - 1, \\ H_3(x) &= x^3 - 3x, \\ &\cdots\cdots\cdots \end{aligned}$$

A simple calculation shows that

$$\begin{aligned} (H_m, H_n) &= \int_{-\infty}^{\infty} H_m(x) H_n(x)\, d\mathsf{P} \\ &= \int_{-\infty}^{\infty} H_m(x) H_n(x) \varphi(x)\, dx = n!\, \delta_{mn}, \end{aligned}$$

where δ_{mn} is the Kronecker delta (0, if $m \neq n$, and 1 if $m = n$). Hence if we put

$$h_n(x) = \frac{H_n(x)}{\sqrt{n!}},$$

the system of *normalized Hermite polynomials* $\{h_n(x)\}_{n\ge 0}$ will be an orthonormal system. We know from functional analysis that if

$$\lim_{c\downarrow 0}\int_{-\infty}^{\infty} e^{c|x|}\mathsf{P}(dx) < \infty, \tag{17}$$

the system $\{1, x, x^2, \ldots\}$ is complete in L^2, i.e. every function $\xi = \xi(x)$ in L^2 can be represented either as $\sum_{i=1}^{n} a_i\eta_i(x)$, where $\eta_i(x) = x^i$, or as a limit of these functions (in the mean-square sense). If we apply the Gram–Schmidt orthogonalization process to the sequence $\eta_1(x), \eta_2(x), \ldots$, with $\eta_i(x) = x^i$, the resulting orthonormal system will be precisely the system of normalized Hermite polynomials. In the present case, (17) is satisfied. Hence $\{h_n(x)\}_{n\ge 0}$ is a basis and therefore every random variable $\xi = \xi(x)$ on this probability space can be represented in the form

$$\xi(x) = \underset{n}{\text{l.i.m.}} \sum_{i=0}^{n} (\xi, h_i)h_i(x). \tag{18}$$

EXAMPLE 2. Let $\Omega = \{0, 1, 2, \ldots\}$ and let $P = \{P_1, P_2, \ldots\}$ be the Poisson distribution

$$P_x = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, \ldots; \quad \lambda > 0.$$

Put $\Delta f(x) = f(x) - f(x-1)$ ($f(x) = 0$, $x < 0$), and by analogy with (15) define the *Poisson–Charlier polynomials*

$$\Pi_n(x) = \frac{(-1)^n\Delta^n P_x}{P_x}, \qquad n \ge 1, \quad \Pi_0 = 1. \tag{19}$$

Since

$$(\Pi_m, \Pi_n) = \sum_{x=0}^{\infty} \Pi_m(x)\Pi_n(x)P_x = c_n\delta_{mn},$$

where c_n are positive constants, the system of *normalized Poisson–Charlier polynomials* $\{\pi_n(x)\}_{n\ge 0}$, $\pi_n(x) = \Pi_n(x)/\sqrt{c_n}$, is an orthonormal system, which is a basis since it satisfies (17).

EXAMPLE 3. In this example we describe the Rademacher and Haar systems, which are of interest in function theory as well as in probability theory.

Let $\Omega = [0, 1]$, $\mathscr{F} = \mathscr{B}([0, 1])$, and let P be Lebesgue measure. As we mentioned in §1, every $x \in [0, 1]$ has a unique binary expansion

$$x = \frac{x_1}{2} + \frac{x_2}{2^2} + \cdots,$$

where $x_i = 0$ or 1. To ensure uniqueness of the expansion, we agree to consider only expansions containing an infinite number of zeros. Thus we

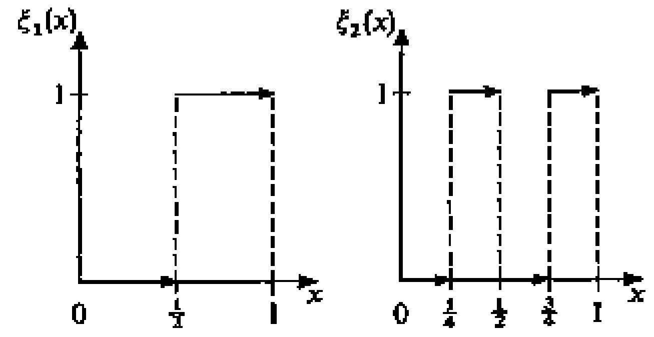


Figure 30

choose the first of the two expansions

$$\frac{1}{2} = \frac{1}{2} + \frac{0}{2^2} + \frac{0}{2^3} + \cdots = \frac{0}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots.$$

We define random variables $\xi_1(x)$, $\xi_2(x)$, ... by putting

$$\xi_n(x) = x_n.$$

Then for any numbers a_i, equal to 0 or 1,

$$\begin{aligned} \mathsf{P}\{x: \xi_1 = a_1, \ldots, \xi_n = a_n\} &= \mathsf{P}\left\{x: \frac{a_1}{2} + \frac{a_2}{2^2} + \cdots + \frac{a_n}{2^n} \le x < \frac{a_1}{2} + \frac{a_2}{2^2} + \cdots + \frac{a_n}{2^n} + \frac{1}{2^n}\right\} \\ &= \mathsf{P}\left\{x: x \in \left[\frac{a_1}{2} + \cdots + \frac{a_n}{2^n}, \frac{a_1}{2} + \cdots + \frac{a_n}{2^n} + \frac{1}{2^n}\right]\right\} = \frac{1}{2^n}. \end{aligned}$$

It follows immediately that $\xi_1, \xi_2, \ldots$ form a *sequence of independent Bernoulli random variables* (Figure 30 shows the construction of $\xi_1 = \xi_1(x)$ and $\xi_2 = \xi_2(x)$).

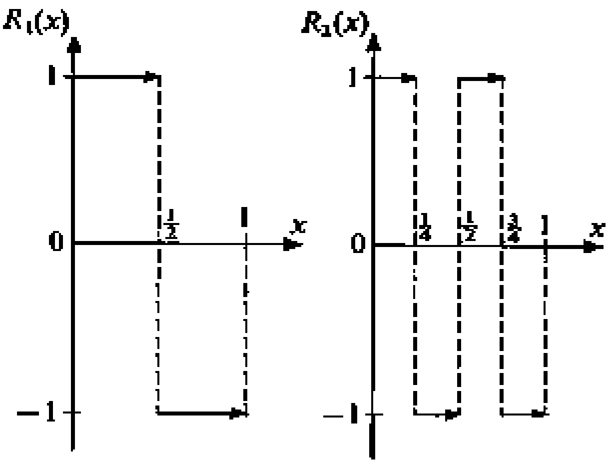


Figure 31. Rademacher functions.

If we now set $R_n(x) = 1 - 2\xi_n(x)$, $n \ge 1$, it is easily verified that $\{R_n\}$ (the *Rademacher functions*, Figure 31) are orthonormal:

$$\mathsf{E} R_n R_m = \int_0^1 R_n(x) R_m(x)\, dx = \delta_{nm}.$$

Notice that $(1, R_n) \equiv \mathsf{E} R_n = 0$. It follows that this system is not complete.

However, the Rademacher system can be used to construct the *Haar system*, which also has a simple structure and is both *orthonormal* and *complete*.

Again let $\Omega = [0, 1)$ and $\mathscr{F} = \mathscr{B}([0, 1))$. Put

$$H_1(x) = 1,$$

$$H_2(x) = R_1(x),$$

....

$$H_n(x) = \begin{cases} 2^{j/2} R_j(x) & \text{if } \dfrac{k-1}{2^j} \le x < \dfrac{k}{2^j}, \quad n = 2^j + k, \quad 1 \le k \le 2^j, j \ge 1, \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to see that $H_n(x)$ can also be written in the form

$$H_{2^m+1}(x) = \begin{cases} 2^{m/2}, & 0 \le x < 2^{-(m+1)}, \\ -2^{m/2}, & 2^{-(m+1)} \le x < 2^{-m}, \quad m = 1, 2, \ldots, \\ 0, & \text{otherwise,} \end{cases}$$

$$H_{2^m+j}(x) = H_{2^m+1}\left(x - \frac{j-1}{2^m}\right), \quad j = 1, \ldots, 2^m.$$

Figure 32 shows graphs of the first eight functions, to give an idea of the structure of the Haar functions.

It is easy to see that the Haar system is orthonormal. Moreover, it is complete both in L^1 and in L^2, i.e. if $f = f(x) \in L^p$ for $p = 1$ or 2, then

$$\int_0^1 |f(x) - \sum_{k=1}^n (f, H_k) H_k(x)|^p\, dx \to 0, \qquad n \to \infty.$$

The system also has the property that

$$\sum_{k=1}^n (f, H_k) H_k(x) \to f(x), \qquad n \to \infty,$$

with probability 1 (with respect to Lebesgue measure).

In §4, Chapter VII, we shall prove these facts by deriving them from general theorems on the convergence of martingales. This will, in particular, provide a good illustration of the application of martingale methods to the theory of functions.

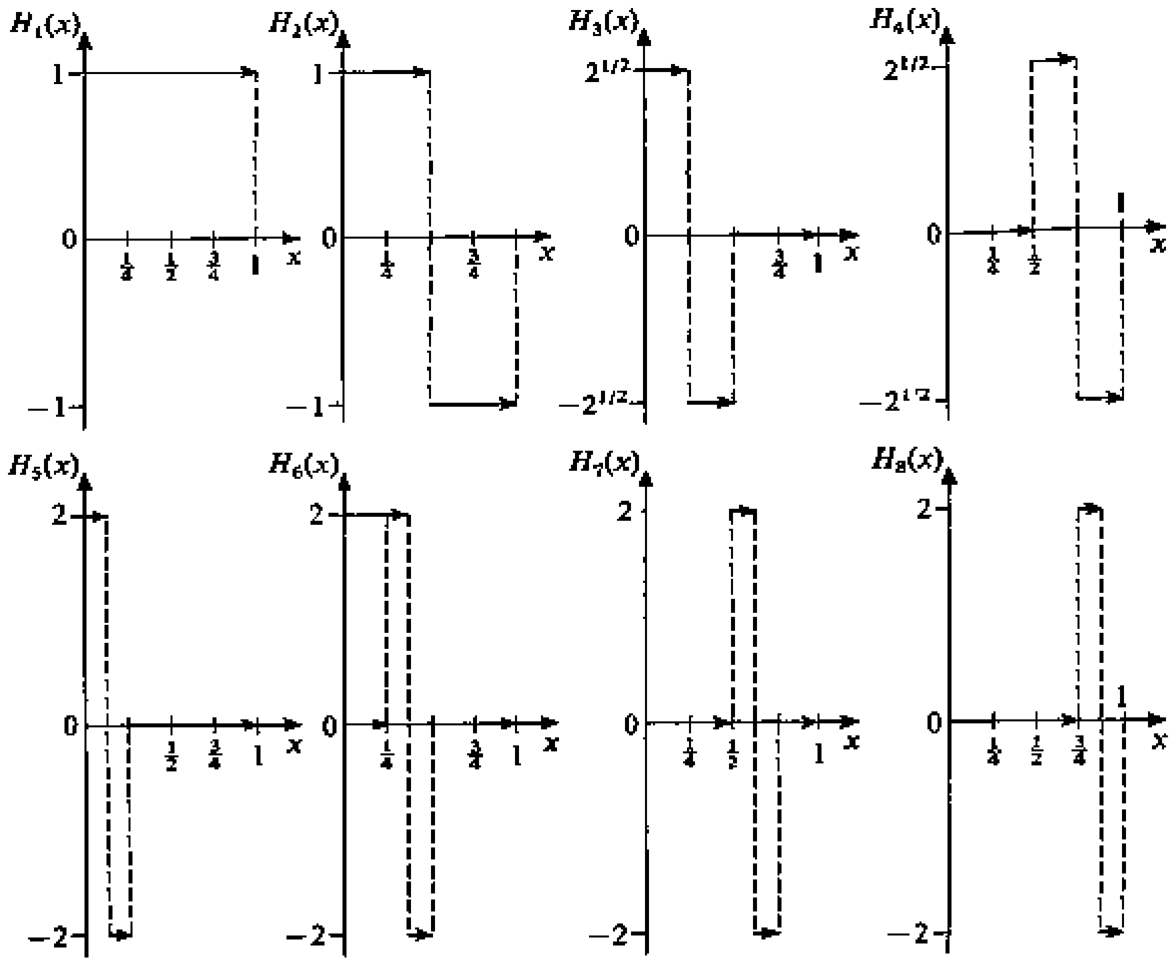


Figure 32. The Haar functions $H_1(x), \ldots, H_8(x)$.

6. If $\eta_1, \ldots, \eta_n$ is a finite orthonormal system then, as was shown above, for every random variable $\xi \in L^2$ there is a random variable $\hat{\xi}$ in the linear manifold $\mathscr{L} = \mathscr{L}\{\eta_1, \ldots, \eta_n\}$, namely the projection of ξ on $\mathscr{L}$, such that

$$\|\xi - \hat{\xi}\| = \inf\{\|\xi - \zeta\| : \zeta \in \mathscr{L}\{\eta_1, \ldots, \eta_n\}\}.$$

Here $\hat{\xi} = \sum_{i=1}^n (\xi, \eta_i)\eta_i$. This result has a natural generalization to the case when $\eta_1, \eta_2, \ldots$ is a countable orthonormal system (not necessarily a basis). In fact, we have the following result.

Theorem. *Let $\eta_1, \eta_2, \ldots$ be an orthonormal system of random variables, and $\bar{L} = \bar{L}\{\eta_1, \eta_2, \ldots\}$ the closed linear manifold spanned by the system. Then there is a unique element $\hat{\xi} \in \bar{L}$ such that*

$$\|\xi - \hat{\xi}\| = \inf\{\|\xi - \zeta\| : \zeta \in \bar{\mathscr{L}}\}. \tag{20}$$

Moreover,

$$\hat{\xi} = \underset{n}{\text{l.i.m.}} \sum_{i=1}^n (\xi, \eta_i)\eta_i \tag{21}$$

and $\xi - \hat{\xi} \perp \zeta$, $\zeta \in \bar{L}$.

PROOF. Let $d = \inf\{\|\xi - \zeta\| : \zeta \in \overline{\mathscr{L}}\}$ and choose a sequence $\zeta_1, \zeta_2, \ldots$ such that $\|\xi - \zeta_n\| \to d$. Let us show that this sequence is fundamental. A simple calculation shows that

$$\|\zeta_n - \zeta_m\|^2 = 2\|\zeta_n - \xi\|^2 + 2\|\zeta_m - \xi\|^2 - 4\left\|\frac{\zeta_n + \zeta_m}{2} - \xi\right\|^2.$$

It is clear that $(\zeta_n + \zeta_m)/2 \in \overline{\mathscr{L}}$; consequently $\|[(\zeta_n + \zeta_m)/2] - \xi\|^2 \geq d^2$ and therefore $\|\zeta_n - \zeta_m\|^2 \to 0$, $n, m \to \infty$.

The space L^2 is complete (Theorem 7, §10). Hence there is an element $\hat{\xi}$ such that $\|\zeta_n - \hat{\xi}\| \to 0$. But $\overline{\mathscr{L}}$ is closed, so $\hat{\xi} \in \overline{\mathscr{L}}$. Moreover, $\|\zeta_n - \xi\| \to d$, and consequently $\|\xi - \hat{\xi}\| = d$, which establishes the existence of the required element.

Let us show that $\hat{\xi}$ is the only element of $\overline{\mathscr{L}}$ with the required property. Let $\tilde{\xi} \in \overline{\mathscr{L}}$ and let

$$\|\xi - \hat{\xi}\| = \|\xi - \tilde{\xi}\| = d.$$

Then (by Problem 3)

$$\|\hat{\xi} + \tilde{\xi} - 2\xi\|^2 + \|\hat{\xi} - \tilde{\xi}\|^2 = 2\|\hat{\xi} - \xi\|^2 + 2\|\tilde{\xi} - \xi\|^2 = 4d^2.$$

But

$$\|\hat{\xi} + \tilde{\xi} - 2\xi\|^2 = 4\|\tfrac{1}{2}(\hat{\xi} + \tilde{\xi}) - \xi\|^2 \geq 4d^2.$$

Consequently $\|\hat{\xi} - \tilde{\xi}\|^2 = 0$. This establishes the uniqueness of the element of $\overline{\mathscr{L}}$ that is closest to ξ.

Now let us show that $\xi - \hat{\xi} \perp \zeta$, $\zeta \in \overline{\mathscr{L}}$. By (20),

$$\|\xi - \hat{\xi} - c\zeta\| \geq \|\xi - \hat{\xi}\|$$

for every $c \in R$. But

$$\|\xi - \hat{\xi} - c\zeta\|^2 = \|\xi - \hat{\xi}\|^2 + c^2\|\zeta\|^2 - 2(\xi - \hat{\xi}, c\zeta).$$

Therefore

$$c^2\|\zeta\|^2 \geq 2(\xi - \hat{\xi}, c\zeta). \tag{22}$$

Take $c = \lambda(\xi - \hat{\xi}, \zeta)$, $\lambda \in R$. Then we find from (22) that

$$(\xi - \hat{\xi}, \zeta)^2[\lambda^2\|\zeta\|^2 - 2\lambda] \geq 0.$$

We have $\lambda^2\|\zeta\|^2 - 2\lambda < 0$ if λ is a sufficiently small positive number. Consequently $(\xi - \hat{\xi}, \zeta) = 0$, $\zeta \in \overline{L}$.

It remains only to prove (21).

The set $\overline{\mathscr{L}} = \overline{\mathscr{L}}\{\eta_1, \eta_2, \ldots\}$ is a closed subspace of L^2 and therefore a Hilbert space (with the same scalar product). Now the system $\eta_1, \eta_2, \ldots$ is a basis for $\overline{\mathscr{L}}$ (Problem 5), and consequently

$$\hat{\xi} = \underset{n}{\text{l.i.m.}} \sum_{k=1}^{n} (\hat{\xi}, \eta_k)\eta_k. \tag{23}$$

But $\xi - \hat{\xi} \perp \eta_k$, $k \geq 1$, and therefore $(\hat{\xi}, \eta_k) = (\xi, \eta_k)$, $k \geq 0$. This, with (23) establishes (21).

This completes the proof of the theorem.

Remark. As in the finite-dimensional case, we say that $\hat{\xi}$ is the projection of ξ on $\bar{L} = \bar{L}\{\eta_1, \eta_2, \dots\}$, that $\xi - \hat{\xi}$ is perpendicular to $\bar{L}$, and that the representation

$$\xi = \hat{\xi} + (\xi - \hat{\xi})$$

is the orthogonal decomposition of ξ.

We also denote $\hat{\xi}$ by $\hat{\mathsf{E}}(\xi \mid \eta_1, \eta_2, \dots)$ and call it the *conditional expectation in the wide sense* (of ξ with respect to $\eta_1, \eta_2, \dots$). From the point of view of estimating ξ in terms of $\eta_1, \eta_2, \dots$, the variable $\hat{\xi}$ is the optimal linear estimator, with error

$$\Delta \equiv \mathsf{E}|\xi - \hat{\xi}|^2 \equiv \|\xi - \hat{\xi}\|^2 = \|\xi\|^2 - \sum_{i=1}^{\infty} |(\xi, \eta_i)|^2,$$

which follows from (5) and (23).

7. Problems

1. Show that if $\xi = \text{l.i.m. } \xi_n$ then $\|\xi_n\| \to \|\xi\|$.
2. Show that if $\xi = \text{l.i.m. } \xi_n$ and $\eta = \text{l.i.m. } \eta_n$ then $(\xi_n, \eta_n) \to (\xi, \eta)$.
3. Show that the norm $\|\cdot\|$ has the *parallelogram* property

$$\|\xi + \eta\|^2 + \|\xi - \eta\|^2 = 2(\|\xi\|^2 + \|\eta\|^2).$$

4. Let $(\xi_1, \dots, \xi_n)$ be a family of orthogonal random variables. Show that they have the *Pythagorean property*,

$$\left\|\sum_{i=1}^{n} \xi_i\right\|^2 = \sum_{i=1}^{n} \|\xi_i\|^2.$$

5. Let $\eta_1, \eta_2, \dots$ be an orthonormal system and $\bar{\mathscr{L}} = \bar{\mathscr{L}}\{\eta_1, \eta_2, \dots\}$ the closed linear manifold spanned by $\eta_1, \eta_2, \dots$. Show that the system is a basis for the (Hilbert) space $\bar{\mathscr{L}}$.
6. Let $\xi_1, \xi_2, \dots$ be a sequence of orthogonal random variables and $S_n = \xi_1 + \dots + \xi_n$. Show that if $\sum_{n=1}^{\infty} \mathsf{E}\xi_n^2 < \infty$ there is a random variable S with $\mathsf{E}S^2 < \infty$ such that $\text{l.i.m. } S_n = S$, i.e. $\|S_n - S\|^2 = \mathsf{E}|S_n - S|^2 \to 0$, $n \to \infty$.
7. Show that in the space $L^2 = L^2([-\pi, \pi], \mathscr{B}([-\pi, \pi]))$ with Lebesgue measure μ the system $\{(1/\sqrt{2\pi})e^{i\lambda n}, n = 0, \pm 1, \dots\}$ is an orthonormal basis.

§12. Characteristic Functions

1. The method of characteristic functions is one of the main tools of the analytic theory of probability. This will appear very clearly in Chapter III in the proofs of limit theorems and, in particular, in the proof of the central limit theorem, which generalizes the De Moivre–Laplace theorem. In the present section we merely define characteristic functions and present their basic properties.

First we make some general remarks.

Besides random variables which take real values, the theory of characteristic functions requires random variables that take complex values (see Subsection 1 of §5).

Many definitions and properties involving random variables can easily be carried over to the complex case. For example, the expectation $\mathsf{E}\zeta$ of a complex random variable $\zeta = \xi + i\eta$ will exist if the expectations $\mathsf{E}\xi$ and $\mathsf{E}\eta$ exist. In this case we define $\mathsf{E}\zeta = \mathsf{E}\xi + i\mathsf{E}\eta$. It is easy to deduce from the definition of the independence of random elements (Definition 6, §5) that the complex random variables $\zeta_1 = \xi_1 + i\eta_1$ and $\zeta_2 = \xi_2 + i\eta_2$ are independent if and only if the pairs (ξ_1, η_1) and (ξ_2, η_2) are independent; or, equivalently, the σ-algebras $\mathscr{L}_{\xi_1, \eta_1}$ and $\mathscr{L}_{\xi_2, \eta_2}$ are independent.

Besides the space L^2 of real random variables with finite second moment, we shall consider the Hilbert space of complex random variables $\zeta = \xi + i\eta$ with $\mathsf{E}|\zeta|^2 < \infty$, where $|\zeta|^2 = \xi^2 + \eta^2$ and the scalar product (ζ_1, ζ_2) is defined by $\mathsf{E}\zeta_1\bar{\zeta}_2$, where $\bar{\zeta}_2$ is the complex conjugate of ζ. The term "random variable" will now be used for both real and complex random variables, with a comment (when necessary) on which is intended.

Let us introduce some notation.

We consider a vector $a \in R^n$ to be a column vector,

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix},$$

and a^{T} to be a row vector, $a^{\mathrm{T}} = (a_1, \ldots, a_n)$. If a and $b \in R^n$ their scalar product (a, b) is $\sum_{i=1}^n a_i b_i$. Clearly $(a, b) = a^{\mathrm{T}}b$.

If $a \in R^n$ and $\mathbb{R} = \|r_{ij}\|$ is an n by n matrix,

$$(\mathbb{R}a, a) = a^{\mathrm{T}}\mathbb{R}a = \sum_{i,j=1}^n r_{ij}a_i a_j. \tag{1}$$

2. Definition 1. Let $F = F(x_1, \ldots, x_n)$ be an n-dimensional distribution function in $(R^n, \mathscr{B}(R^n))$. Its *characteristic function* is

$$\varphi(t) = \int_{R^n} e^{i(t,x)}\,dF(x), \qquad t \in R^n. \tag{2}$$

Definition 2. If $\xi = (\xi_1, \ldots, \xi_n)$ is a random vector defined on the probability space $(\Omega, \mathscr{F}, \mathsf{P})$ with values in R^n, its *characteristic function* is

$$\varphi_\xi(t) = \int_{R^n} e^{i(t,x)}\,dF_\xi(x), \qquad t \in R^n, \tag{3}$$

where $F_\xi = F_\xi(x_1, \ldots, x_n)$ is the distribution function of the vector $\xi = (\xi_1, \ldots, \xi_n)$.

If $F(x)$ has a density $f = f(x)$ then

$$\varphi(t) = \int_{R^n} e^{i(t,x)} f(x)\,dx.$$

In other words, in this case the characteristic function is just the Fourier transform of $f(x)$.

It follows from (3) and Theorem 6.7 (on change of variable in a Lebesgue integral) that the characteristic function $\varphi_\xi(t)$ of a random vector can also be defined by

$$\varphi_\xi(t) = \mathsf{E}e^{i(t,\xi)}, \qquad t \in R^n. \tag{4}$$

We now present some basic properties of characteristic functions, stated and proved for $n = 1$. Further important results for the general case will be given as problems.

Let $\xi = \xi(\omega)$ be a random variable, $F_\xi = F_\xi(x)$ its distribution function, and

$$\varphi_\xi(t) = \mathsf{E}e^{it\xi}$$

its characteristic function.

We see at once that if $\eta = a\xi + b$ then

$$\varphi_\eta(t) = \mathsf{E}e^{it\eta} = \mathsf{E}e^{it(a\xi+b)} = e^{itb}\mathsf{E}e^{iat\xi}.$$

Therefore

$$\varphi_\eta(t) = e^{itb}\varphi_\xi(at). \tag{5}$$

Moreover, if $\xi_1, \xi_2, \dots, \xi_n$ are independent random variables and $S_n = \xi_1 + \cdots + \xi_n$, then

$$\varphi_{S_n}(t) = \prod_{i=1}^{n} \varphi_{\xi_i}(t). \tag{6}$$

In fact,

$$\begin{aligned}\varphi_{S_n} &= \mathsf{E}e^{it(\xi_1 + \cdots + \xi_n)} = \mathsf{E}e^{it\xi_1} \cdots e^{it\xi_n} \\ &= \mathsf{E}e^{it\xi_1} \cdots \mathsf{E}e^{it\xi_n} = \prod_{j=1}^{n} \varphi_{\xi_j}(t),\end{aligned}$$

where we have used the property that the expectation of a product of independent (bounded) random variables (either real or complex; see Theorem 6 of §6, and Problem 1) is equal to the product of their expectations.

Property (6) is the key to the proofs of limit theorems for sums of independent random variables by the method of characteristic functions (see §3, Chapter III). In this connection we note that the distribution function F_{S_n} is expressed in terms of the distribution functions of the individual terms in a rather complicated way, namely $F_{S_n} = F_{\xi_1} * \cdots * F_{\xi_n}$ where $*$ denotes convolution (see §8, Subsection 4).

Here are some examples of characteristic functions.

EXAMPLE 1. Let ξ be a Bernoulli random variable with $\mathsf{P}(\xi = 1) = p$, $\mathsf{P}(\xi = 0) = q$, $p + q = 1$, $1 > p > 0$; then

$$\varphi_\xi(t) = pe^{it} + q.$$

If $\xi_1, \ldots, \xi_n$ are independent identically distributed random variables like ξ, then, writing $T_n = (S_n - n_p)/\sqrt{npq}$, we have

$$\begin{aligned}\varphi_{T_n}(t) = \mathsf{E}e^{iT_n t} &= e^{-it\sqrt{np/q}}[pe^{it/\sqrt{npq}} + q]^n \\ &= [pe^{it\sqrt{q/(np)}} + qe^{-it\sqrt{p/(nq)}}]^n.\end{aligned} \tag{7}$$

Notice that it follows that as $n \to \infty$

$$\varphi_{T_n}(t) \to e^{-t^2/2}, \qquad T_n = \frac{S_n - np}{\sqrt{npq}}. \tag{8}$$

EXAMPLE 2. Let $\xi \sim \mathcal{N}(m, \sigma^2)$, $|m| < \infty$, $\sigma^2 > 0$. Let us show that

$$\varphi_\xi(t) = e^{itm - t^2\sigma^2/2} \tag{9}$$

Let $\eta = (\xi - m)/\sigma$. Then $\eta \sim \mathcal{N}(0, 1)$ and, since

$$\varphi_\xi(t) = e^{itm}\varphi_\eta(\sigma t)$$

by (5), it is enough to show that

$$\varphi_\eta(t) = e^{-t^2/2}. \tag{10}$$

We have

$$\begin{aligned}\varphi_\eta(t) = \mathsf{E}e^{it\eta} &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{itx}e^{-x^2/2}\,dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\sum_{n=0}^{\infty}\frac{(itx)^n}{n!}e^{-x^2/2}\,dx = \sum_{n=0}^{\infty}\frac{(it)^n}{n!}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}x^n e^{-x^2/2}\,dx \\ &= \sum_{n=0}^{\infty}\frac{(it)^{2n}}{(2n)!}(2n-1)!! = \sum_{n=0}^{\infty}\frac{(it)^{2n}}{(2n)!}\frac{(2n)!}{2^n n!} \\ &= \sum_{n=0}^{\infty}\left(-\frac{t^2}{2}\right)^n \cdot \frac{1}{n!} = e^{-t^2/2},\end{aligned}$$

where we have used the formula (see Problem 7 in §8)

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^{2n}e^{-x^2/2}\,dx \equiv \mathsf{E}\eta^{2n} = (2n-1)!!.$$

EXAMPLE 3. Let ξ be a Poisson random variable,

$$\mathsf{P}(\xi = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k = 0, 1, \ldots.$$

Then

$$\mathsf{E}e^{it\xi} = \sum_{k=0}^{\infty} e^{itk}\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^{it})^k}{k!} = \exp\{\lambda(e^{it} - 1)\}. \tag{11}$$

3. As we observed in §9, Subsection 1, with every distribution function in $(R, \mathscr{B}(R))$ we can associate a random variable of which it is the distribution function. Hence in discussing the properties of characteristic functions (in the sense either of Definition 1 or Definition 2), we may consider only characteristic functions $\varphi(t) = \varphi_\xi(t)$ of random variables $\xi = \xi(\omega)$.

Theorem 1. *Let ξ be a random variable with distribution function $F = F(x)$ and*

$$\varphi(t) = \mathsf{E}e^{it\xi}$$

its characteristic function. Then φ has the following properties:

(1) $|\varphi(t)| \le \varphi(0) = 1$;
(2) *$\varphi(t)$ is uniformly continuous for $t \in R$;*
(3) $\varphi(t) = \overline{\varphi(-t)}$;
(4) *$\varphi(t)$ is real-valued if and only if F is symmetric ($\int_B dF(x) = \int_{-B} dF(x)$), $B \in \mathscr{B}(R)$, $-B = \{-x: x \in B\}$;*
(5) *if $\mathsf{E}|\xi|^n < \infty$ for some $n \ge 1$, then $\varphi^{(r)}(t)$ exists for every $r \le n$, and*

$$\varphi^{(r)}(t) = \int_R (ix)^r e^{itx}\, dF(x), \tag{12}$$

$$\mathsf{E}\xi^r = \frac{\varphi^{(r)}(0)}{i^r}, \tag{13}$$

$$\varphi(t) = \sum_{r=0}^{n} \frac{(it)^r}{r!}\mathsf{E}\xi^r + \frac{(it)^n}{n!}\varepsilon_n(t), \tag{14}$$

where $|\varepsilon_n(t)| \le 3\mathsf{E}|\xi|^n$ and $\varepsilon_n(t) \to 0$, $t \to 0$;
(6) *if $\varphi^{(2n)}(0)$ exists and is finite then $\mathsf{E}\xi^{2n} < \infty$;*
(7) *if $\mathsf{E}|\xi|^n < \infty$ for all $n \ge 1$ and*

$$\overline{\lim_n} \frac{(\mathsf{E}|\xi|^n)^{1/n}}{n} = \frac{1}{e \cdot R} < \infty,$$

then

$$\varphi(t) = \sum_{n=0}^{\infty} \frac{(it)^n}{n!}\mathsf{E}\xi^n. \tag{15}$$

for all $|t| < R$.

Proof. Properties (1) and (3) are evident. Property (2) follows from the inequality

$$|\varphi(t + h) - \varphi(t)| = |\mathsf{E}e^{it\xi}(e^{ih\xi} - 1)| \le \mathsf{E}|e^{ih\xi} - 1|$$

and the dominated convergence theorem, according to which $\mathsf{E}|e^{ih\xi} - 1| \to 0$, $h \to 0$.

Property (4). Let F be symmetric. Then if $g(x)$ is a bounded odd Borel function, we have $\int_R g(x)\, dF(x) = 0$ (observe that for simple odd functions this follows directly from the definition of the symmetry of F). Consequently $\int_R \sin tx \, dF(x) = 0$ and therefore

$$\varphi(t) = \mathsf{E} \cos t\xi.$$

Conversely, let $\varphi_\xi(t)$ be a real function. Then by (3)

$$\varphi_{-\xi}(t) = \varphi_\xi(-t) = \overline{\varphi_\xi(t)} = \varphi_\xi(t), \qquad t \in R.$$

Hence (as will be shown below in Theorem 2) the distribution functions $F_{-\xi}$ and F_ξ of the random variables $-\xi$ and ξ are the same, and therefore (by Theorem 3.1)

$$\mathsf{P}(\xi \in B) = \mathsf{P}(-\xi \in B) = \mathsf{P}(\xi \in -B)$$

for every $B \in \mathscr{B}(R)$.

Property (5). If $\mathsf{E}|\xi|^n < \infty$, we have $\mathsf{E}|\xi|^r < \infty$ for $r \leq n$, by Lyapunov's inequality (6.28).

Consider the difference quotient

$$\frac{\varphi(t+h) - \varphi(t)}{h} = \mathsf{E} e^{it\xi}\left(\frac{e^{ih\xi} - 1}{h}\right).$$

Since

$$\left|\frac{e^{ihx} - 1}{h}\right| \leq |x|,$$

and $\mathsf{E}|\xi| < \infty$, it follows from the dominated convergence theorem that the limit

$$\lim_{h \to 0} \mathsf{E} e^{it\xi}\left(\frac{e^{ih\xi} - 1}{h}\right)$$

exists and equals

$$\mathsf{E} e^{it\xi} \lim_{h \to 0}\left(\frac{e^{ih\xi} - 1}{h}\right) = i\mathsf{E}(\xi e^{it\xi}) = i \int_{-\infty}^{\infty} x e^{itx} \, dF(x). \tag{16}$$

Hence $\varphi'(t)$ exists and

$$\varphi'(t) = i(\mathsf{E}\xi e^{it\xi}) = i \int_{-\infty}^{\infty} x e^{itx} \, dF(x).$$

The existence of the derivatives $\varphi^{(r)}(t)$, $1 < r \leq n$, and the validity of (12), follow by induction.

Formula (13) follows immediately from (12). Let us now establish (14). Since

$$e^{iy} = \cos y + i \sin y = \sum_{k=0}^{n-1} \frac{(iy)^k}{k!} + \frac{(iy)^n}{n!}[\cos \theta_1 y + i \sin \theta_2 y]$$

for real y, with $|\theta_1| \le 1$ and $|\theta_2| \le 1$, we have

$$e^{it\xi} = \sum_{k=0}^{n-1} \frac{(it\xi)^k}{k!} + \frac{(it\xi)^n}{n!}[\cos \theta_1(\omega)t\xi + i \sin \theta_2(\omega)t\xi] \tag{17}$$

and

$$\mathsf{E}e^{it\xi} = \sum_{k=0}^{n-1} \frac{(it)^k}{k!}\mathsf{E}\xi^k + \frac{(it)^n}{n!}[\mathsf{E}\xi^n + \varepsilon_n(t)], \tag{18}$$

where

$$\varepsilon_n(t) = \mathsf{E}[\xi^n(\cos \theta_1(\omega)t\xi + i \sin \theta_2(\omega)t\xi - 1)].$$

It is clear that $|\varepsilon_n(t)| \le 3\mathsf{E}|\xi^n|$. The theorem on dominated convergence shows that $\varepsilon_n(t) \to 0$, $t \to 0$.

Property (6). We give a proof by induction. Suppose first that $\varphi''(0)$ exists and is finite. Let us show that in that case $\mathsf{E}\xi^2 < \infty$. By L'Hôpital's rule and Fatou's lemma,

$$\begin{aligned}
\varphi''(0) &= \lim_{h\to 0} \frac{1}{2}\left[\frac{\varphi'(2h) - \varphi'(0)}{2h} + \frac{\varphi'(0) - \varphi'(-2h)}{2h}\right] \\
&= \lim_{h\to 0} \frac{2\varphi'(2h) - 2\varphi'(-2h)}{8h} = \lim_{h\to 0} \frac{1}{4h^2}[\varphi(2h) - 2\varphi(0) + \varphi(-2h)] \\
&= \lim_{h\to 0} \int_{-\infty}^{\infty} \left(\frac{e^{ihx} - e^{-ihx}}{2h}\right)^2 dF(x) \\
&= -\lim_{h\to 0} \int_{-\infty}^{\infty} \left(\frac{\sin hx}{hx}\right)^2 x^2\, dF(x) \le -\int_{-\infty}^{\infty} \lim_{h\to 0}\left(\frac{\sin hx}{hx}\right)^2 x^2\, dF(x) \\
&= -\int_{-\infty}^{\infty} x^2\, dF(x).
\end{aligned}$$

Therefore,

$$\int_{-\infty}^{\infty} x^2\, dF(x) \le -\varphi''(0) < \infty.$$

Now let $\varphi^{(2k+2)}(0)$ exist, finite, and let $\int_{-\infty}^{+\infty} x^{2k}\, dF(x) < \infty$. If $\int_{-\infty}^{\infty} x^{2k} dF(x) = 0$, then $\int_{-\infty}^{\infty} x^{2k+2}\, dF(x) = 0$ also. Hence we may suppose that $\int_{-\infty}^{\infty} x^{2k}\, dF(x) > 0$. Then, by Property (5),

$$\varphi^{(2k)}(t) = \int_{-\infty}^{\infty} (ix)^{2k} e^{itx}\, dF(x)$$

and therefore,

$$(-1)^k \varphi^{(2k)}(t) = \int_{-\infty}^{\infty} e^{itx}\, dG(x),$$

where $G(x) = \int_{-\infty}^{x} u^{2k}\, dF(u)$.

Consequently the function $(-1)^k\varphi^{(2k)}(t)G(\infty)^{-1}$ is the characteristic function of the probability distribution $G(x)\cdot G^{-1}(\infty)$ and by what we have proved,

$$G^{-1}(\infty)\int_{-\infty}^{\infty} x^2\,dG(x) < \infty.$$

But $G^{-1}(\infty) > 0$, and therefore

$$\int_{-\infty}^{\infty} x^{2k+2}\,dF(x) = \int_{-\infty}^{\infty} x^2\,dG(x) < \infty.$$

Property (7). Let $0 < t_0 < R$. Then, by Stirling's formula we find that

$$\overline{\lim}\,\frac{(\mathsf{E}|\xi|^n)^{1/n}}{n} < \frac{1}{e\cdot t_0} \Rightarrow \overline{\lim}\,\frac{(\mathsf{E}|\xi|^n t_0^n)^{1/n}}{n} < \frac{1}{e} \Rightarrow \lim\left(\frac{\mathsf{E}|\xi|^n t_0^n}{n!}\right)^{1/n} < 1.$$

Consequently the series $\sum [\mathsf{E}|\xi|^n t_0^n/n!]$ converges by Cauchy's test, and therefore the series $\sum_{r=0}^{\infty} [(it)^r/r!]\mathsf{E}\xi^r$ converges for $|t| \le t_0$. But by (14), for $n \ge 1$,

$$\varphi(t) = \sum_{r=0}^{n} \frac{(it)^r}{r!}\mathsf{E}\xi^r + R_n(t),$$

where $|R_n(t)| \le 3(|t|^n/n!)\mathsf{E}|\xi|^n$. Therefore

$$\varphi(t) = \sum_{r=0}^{\infty} \frac{(it)^r}{r!}\mathsf{E}\xi^r$$

for all $|t| < R$. This completes the proof of the theorem.

Remark 1. By a method similar to that used for (14), we can establish that if $\mathsf{E}|\xi|^n < \infty$ for some $n \ge 1$, then

$$\varphi(t) = \sum_{k=0}^{n} \frac{i^k(t-s)^k}{k!}\int_{-\infty}^{\infty} x^k e^{isx}\,dF(x) + \frac{i^n(t-s)^n}{n!}\varepsilon_n(t-s), \qquad (19)$$

where $|\varepsilon_n(t-s)| \le 3\mathsf{E}|\xi^n|$, and $\varepsilon_n(t-s) \to 0$ as $t-s \to 0$.

Remark 2. With reference to the condition that appears in Property (7), see also Subsection 9, below, on the "uniqueness of the solution of the moment problem."

4. The following theorem shows that the characteristic function is uniquely determined by the distribution function.

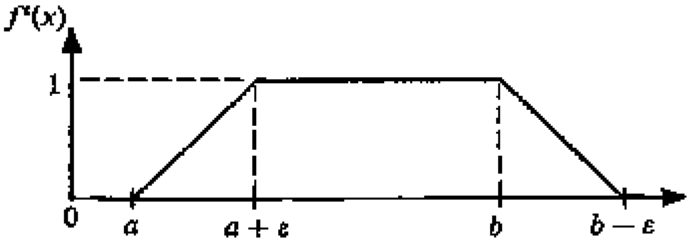


Figure 33

Theorem 2 (Uniqueness). *Let F and G be distribution functions with the same characteristic function, i.e.*

$$\int_{-\infty}^{\infty} e^{itx}\, dF(x) = \int_{-\infty}^{\infty} e^{itx}\, dG(x) \tag{20}$$

for all $t \in R$. *Then* $F(x) \equiv G(x)$.

PROOF. Choose a and $b \in R$, and $\varepsilon > 0$, and consider the function $f^\varepsilon = f^\varepsilon(x)$ shown in Figure 33. We show that

$$\int_{-\infty}^{\infty} f^\varepsilon(x)\, dF(x) = \int_{-\infty}^{\infty} f^\varepsilon(x)\, dG(x). \tag{21}$$

Let $n \geq 0$ be large enough so that $[a - \varepsilon, b + \varepsilon] \subseteq [-n, n]$, and let the sequence $\{\delta_n\}$ be such that $1 \geq \delta_n \downarrow 0$, $n \to \infty$. Like every continuous function on $[-n, n]$ that has equal values at the endpoints, $f^\varepsilon = f^\varepsilon(x)$ can be uniformly approximated by trigonometric polynomials (Weierstrass's theorem), i.e. there is a *finite* sum

$$f_n^\varepsilon(x) = \sum_k a_k \exp\left(i\pi x \frac{k}{n}\right) \tag{22}$$

such that

$$\sup_{-n \leq x \leq n} |f^\varepsilon(x) - f_n^\varepsilon(x)| \leq \delta_n. \tag{23}$$

Let us extend the periodic function $f_n(x)$ to all of R, and observe that

$$\sup_x |f_n^\varepsilon(x)| \leq 2.$$

Then, since by (20)

$$\int_{-\infty}^{\infty} f_n^\varepsilon(x)\, dF(x) = \int_{-\infty}^{\infty} f_n^\varepsilon(x)\, dG(x),$$

we have

$$\left|\int_{-\infty}^{\infty} f^\varepsilon(x)\,dF(x) - \int_{-\infty}^{\infty} f^\varepsilon(x)\,dG(x)\right| = \left|\int_{-n}^{n} f^\varepsilon\,dF - \int_{-n}^{n} f^\varepsilon\,dG\right|$$

$$\leq \left|\int_{-n}^{n} f_n^\varepsilon\,dF - \int_{-n}^{n} f_n^\varepsilon\,dG\right| + 2\delta_n$$

$$\leq \left|\int_{-\infty}^{\infty} f_n^\varepsilon\,dF - \int_{-\infty}^{\infty} f_n^\varepsilon\,dG\right| + 2\delta_n$$

$$+ 2F(\overline{[-n, n]}) + 2G(\overline{[-n, n]}), \tag{24}$$

where $F(A) = \int_A dF(x)$, $G(A) = \int_A dG(x)$. As $n \to \infty$, the right-hand side of (24) tends to zero, and this establishes (21).

As $\varepsilon \to 0$, we have $f^\varepsilon(x) \to I_{(a,b]}(x)$. It follows from (21) by the theorem on distribution functions' being the same.

$$\int_{-\infty}^{\infty} I_{(a,b]}(x)\,dF(x) = \int_{-\infty}^{\infty} I_{(a,b]}(x)\,dG(x),$$

i.e. $F(b) - F(a) = G(b) - G(a)$. Since a and b are arbitrary, it follows that $F(x) = G(x)$ for all $x \in R$.

This completes the proof of the theorem.

5. The preceding theorem says that a distribution function $F = F(x)$ is uniquely determined by its characteristic function $\varphi = \varphi(t)$. The next theorem gives an explicit representation of F in terms of φ.

Theorem 3 (Inversion Formula). *Let $F = F(x)$ be a distribution function and*

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx}\,dF(x)$$

its characteristic function.

(a) *For pairs of points a and b $(a < b)$ at which $F = F(x)$ is continuous,*

$$F(b) - F(a) = \lim_{c\to\infty} \frac{1}{2\pi}\int_{-c}^{c} \frac{e^{-ita} - e^{-itb}}{it}\varphi(t)\,dt; \tag{25}$$

(b) *If $\int_{-\infty}^{\infty} |\varphi(t)|\,dt < \infty$, the distribution function $F(x)$ has a density $f(x)$,*

$$F(x) = \int_{-\infty}^{x} f(y)\,dy \tag{26}$$

and

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi(t)\,dt. \tag{27}$$

Proof. We first observe that if $F(x)$ has density $f(x)$ then

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx, \tag{28}$$

and (27) is just the Fourier transform of the (integrable) function $\varphi(t)$. Integrating both sides of (27) and applying Fubini's theorem, we obtain

$$\begin{aligned} F(b) - F(a) = \int_a^b f(x)\, dx &= \frac{1}{2\pi} \int_a^b \left[\int_{-\infty}^{\infty} e^{-itx} \varphi(t)\, dt \right] dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi(t) \left[\int_a^b e^{-itx}\, dx \right] dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi(t) \frac{e^{-ita} - e^{-itb}}{it}\, dt. \end{aligned}$$

After these remarks, which to some extent clarify (25), we turn to the proof.

(a) We have

$$\begin{aligned} \Phi_c &\equiv \frac{1}{2\pi} \int_{-c}^{c} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt \\ &= \frac{1}{2\pi} \int_{-c}^{c} \frac{e^{-ita} - e^{-itb}}{it} \left[\int_{-\infty}^{\infty} e^{itx}\, dF(x) \right] dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_{-c}^{c} \frac{e^{-ita} - e^{-itb}}{it} e^{itx}\, dt \right] dF(x) \\ &= \int_{-\infty}^{\infty} \Psi_c(x)\, dF(x), \end{aligned} \tag{29}$$

where we have put

$$\Psi_c(x) = \frac{1}{2\pi} \int_{-c}^{c} \frac{e^{-ita} - e^{-itb}}{it} e^{itx}\, dt$$

and applied Fubini's theorem, which is applicable in this case because

$$\left| \frac{e^{-ita} - e^{-itb}}{it} \cdot e^{itx} \right| = \left| \frac{e^{-ita} - e^{-itb}}{it} \right| = \left| \int_a^b e^{-itx}\, dx \right| \le b - a$$

and

$$\int_{-c}^{c} \int_{-\infty}^{\infty} (b - a)\, dF(x) \le 2c(b - a) < \infty.$$

In addition,

$$\Psi_c(x) = \frac{1}{2\pi}\int_{-c}^{c} \frac{\sin t(x-a) - \sin t(x-b)}{t}\,dt$$

$$= \frac{1}{2\pi}\int_{-c(x-a)}^{c(x-a)} \frac{\sin v}{v}\,dv - \frac{1}{2\pi}\int_{-c(x-b)}^{c(x-b)} \frac{\sin u}{u}\,du. \tag{30}$$

The function

$$g(s, t) = \int_s^t \frac{\sin v}{v}\,dv$$

is uniformly continuous in s and t, and

$$g(s, t) \to \pi \tag{31}$$

as $s \downarrow -\infty$ and $t \uparrow \infty$. Hence there is a constant C such that $|\Psi_c(x)| < C < \infty$ for all c and x. Moreover, it follows from (30) and (31) that

$$\Psi_c(x) \to \Psi(x), \qquad c \to \infty,$$

where

$$\Psi(x) = \begin{cases} 0, & x < a, x > b, \\ \frac{1}{2}, & x = a, x = b, \\ 1, & a < x < b. \end{cases}$$

Let μ be a measure on $(R, \mathscr{B}(R))$ such that $\mu(a, b] = F(b) - F(a)$. Then if we apply the dominated convergence theorem and use the formulas of Problem 1 of §3, we find that, as $c \to \infty$,

$$\Phi_c = \int_{-\infty}^{\infty} \Psi_c(x)\,dF(x) \to \int_{-\infty}^{\infty} \Psi(x)\,dF(x)$$

$$= \mu(a, b) + \tfrac{1}{2}\mu\{a\} + \tfrac{1}{2}\mu\{b\}$$

$$= F(b-) - F(a) + \tfrac{1}{2}[F(a) - F(a-) + F(b) - F(b-)]$$

$$= \frac{F(b) + F(b-)}{2} - \frac{F(a) + F(a-)}{2} = F(b) - F(a),$$

where the last equation holds for all points a and b of continuity of $F(x)$.

Hence (25) is established.

(b) Let $\int_{-\infty}^{\infty} |\varphi(t)|\,dt < \infty$. Write

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi(t)\,dt.$$

It follows from the dominated convergence theorem that this is a continuous function of x and therefore is integrable on $[a, b]$. Consequently we find, applying Fubini's theorem again, that

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_a^b \frac{1}{2\pi}\left(\int_{-\infty}^{\infty} e^{-itx}\varphi(t)\,dt\right)dx \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\varphi(t)\left[\int_a^b e^{-itx}\,dx\right]dt = \lim_{c\to\infty}\frac{1}{2\pi}\int_{-c}^{c}\varphi(t)\left[\int_a^b e^{-itx}\,dx\right]dt \\ &= \lim_{c\to\infty}\frac{1}{2\pi}\int_{-c}^{c}\frac{e^{-ita}-e^{-itb}}{it}\varphi(t)\,dt = F(b)-F(a)\end{aligned}$$

for all points a and b of continuity of $F(x)$.

Hence it follows that

$$F(x) = \int_{-\infty}^{x} f(y)\,dy, \qquad x \in R,$$

and since $f(x)$ is continuous and $F(x)$ is nondecreasing, $f(x)$ is the density of $F(x)$.

This completes the proof of the theorem.

Corollary. *The inversion formula* (25) *provides a second proof of Theorem* 2.

Theorem 4. *A necessary and sufficient condition for the components of the random vector $\xi = (\xi_1, \ldots, \xi_n)$ to be independent is that its characteristic function is the product of the characteristic functions of the components:*

$$\mathsf{E}e^{i(t_1\xi_1+\cdots+t_n\xi_n)} = \prod_{k=1}^{n}\mathsf{E}e^{it_k\xi_k}, \qquad (t_1, \ldots, t_n) \in R^n.$$

PROOF. The necessity follows from Problem 1. To prove the sufficiency we let $F(x_1, \ldots, x_n)$ be the distribution function of the vector $\xi = (\xi_1, \ldots, \xi_n)$ and $F_k(x)$, the distribution functions of the ξ_k, $1 \le k \le n$. Put $G = G(x_1, \ldots, x_n) = F_1(x_1)\cdots F_n(x_n)$. Then, by Fubini's theorem, for all $(t_1, \ldots, t_n) \in R^n$,

$$\begin{aligned}\int_{R^n} e^{i(t_1x_1+\cdots+t_nx_n)}\,dG(x_1\cdots x_n) &= \prod_{k=1}^{n}\int_R e^{it_kx_k}\,dF_k(x) \\ &= \prod_{k=1}^{n}\mathsf{E}e^{it_k\xi_k} = \mathsf{E}e^{i(t_1\xi_1+\cdots+t_n\xi_n)} \\ &= \int_{R^n} e^{i(t_1x_1+\cdots+t_nx_n)}\,dF(x_1\cdots x_n).\end{aligned}$$

Therefore by Theorem 2 (or rather, by its multidimensional analog; see Problem 3) we have $F = G$, and consequently, by the theorem of §5, the random variables $\xi_1, \ldots, \xi_n$ are independent.

6. Theorem 1 gives us necessary conditions for a function to be a characteristic function. Hence if $\varphi = \varphi(t)$ fails to satisfy, for example, one of the first three conclusions of the theorem, that function cannot be a characteristic function. We quote without proof some results in the same direction.

Bochner–Khinchin Theorem. *Let $\varphi(t)$ be continuous, $t \in R$, with $\varphi(0) = 1$. A necessary and sufficient condition that $\varphi(t)$ is a characteristic function is that it is positive semi-definite, i.e. that for all real $t_1, \ldots, t_n$ and all complex $\lambda_1, \ldots, \lambda_n$, $n = 1, 2, \ldots,$*

$$\sum_{i,j=1}^{n} \varphi(t_i - t_j)\lambda_i \bar{\lambda}_j \geq 0. \tag{32}$$

The necessity of (32) is evident since if $\varphi(t) = \int_{-\infty}^{\infty} e^{itx}\, dF(x)$ then

$$\sum_{i,j=1}^{n} \varphi(t_i - t_j)\lambda_i \bar{\lambda}_j = \int_{-\infty}^{\infty} \left| \sum_{k=1}^{n} \lambda_k e^{it_k x} \right|^2 dF(x) \geq 0.$$

The proof of the sufficiency of (32) is more difficult.

Pólya's Theorem. *Let a continuous even function $\varphi(t)$ satisfy $\varphi(t) \geq 0$, $\varphi(0) = 1$, $\varphi(t) \to 0$ as $t \to \infty$ and let $\varphi(t)$ be convex on $0 \leq t < \infty$. Then $\varphi(t)$ is a characteristic function.*

This theorem provides a very convenient method of constructing characteristic functions. Examples are

$$\varphi_1(t) = e^{-|t|},$$

$$\varphi_2(t) = \begin{cases} 1 - |t|, & |t| \leq 1, \\ 0, & |t| > 1. \end{cases}$$

Another is the function $\varphi_3(t)$ drawn in Figure 34. On $[-a, a]$, the function $\varphi_3(t)$ coincides with $\varphi_2(t)$. However, the corresponding distribution functions F_2 and F_3 are evidently different. This example shows that in general two characteristic functions can be the same on a finite interval without their distribution functions' being the same.

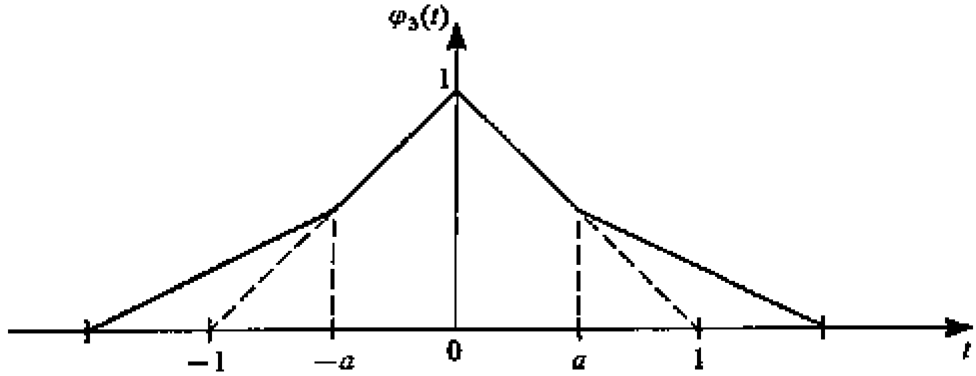


Figure 34

Marcinkiewicz's Theorem. *If a characteristic function $\varphi(t)$ is of the form $\exp \mathscr{P}(t)$, where $\mathscr{P}(t)$ is a polynomial, then this polynomial is of degree at most* 2.

It follows, for example, that e^{-t^4} is not a characteristic function.

7. The following theorem shows that a property of the characteristic function of a random variable can lead to a nontrivial conclusion about the nature of the random variable.

Theorem 5. *Let $\varphi_\xi(t)$ be the characteristic function of the random variable ξ.*

(a) *If $|\varphi_\xi(t_0)| = 1$ for some $t_0 \neq 0$, then ξ is concentrated at the points $a + nh$, $h = 2\pi/t_0$, for some a, that is,*

$$\sum_{n=-\infty}^{\infty} \mathsf{P}\{\xi = a + nh\} = 1, \tag{33}$$

where a is a constant.

(b) *If $|\varphi_\xi(t)| = |\varphi_\xi(\alpha t)| = 1$ for two different points t and αt, where α is irrational, then ξ is degenerate:*

$$\mathsf{P}\{\xi = a\} = 1,$$

where a is some number.

(c) *If $|\varphi_\xi(t)| \equiv 1$, then ξ is degenerate.*

PROOF. (a) If $|\varphi_\xi(t_0)| = 1$, $t_0 \neq 0$, there is a number a such that $\varphi(t_0) = e^{it_0 a}$. Then

$$e^{it_0 a} = \int_{-\infty}^{\infty} e^{it_0 x}\, dF(x) \Rightarrow 1 = \int_{-\infty}^{\infty} e^{it_0(x-a)}\, dF(x) \Rightarrow$$

$$1 = \int_{-\infty}^{\infty} \cos t_0(x-a)\, dF(x) \Rightarrow \int_{-\infty}^{\infty} [1 - \cos t_0(x-a)]\, dF(x) = 0.$$

Since $1 - \cos t_0(x - a) \geq 0$, it follows from property **H** (Subsection 2 of §6) that

$$1 = \cos t_0(\xi - a) \quad (\mathsf{P}\text{-a.s.}),$$

which is equivalent to (33).

(b) It follows from $|\varphi_\xi(t)| = |\varphi_\xi(\alpha t)| = 1$ and from (33) that

$$\sum_{n=-\infty}^{\infty} \mathsf{P}\left\{\xi = a + \frac{2\pi}{t} n\right\} = \sum_{m=-\infty}^{\infty} \mathsf{P}\left\{\xi = b + \frac{2\pi}{\alpha t} m\right\} = 1.$$

If ξ is not degenerate, there must be at least two pairs of common points:

$$a + \frac{2\pi}{t} n_1 = b + \frac{2\pi}{\alpha t} m_1, \qquad a + \frac{2\pi}{y} n_2 = b + \frac{2\pi}{\alpha t} m_2,$$

in the sets

$$\left\{a + \frac{2\pi}{t} n, n = 0, \pm 1, \ldots\right\} \quad \text{and} \quad \left\{b + \frac{2\pi}{\alpha t} m, m = 0, \pm 1, \ldots\right\},$$

whence

$$\frac{2\pi}{t}(n_1 - n_2) = \frac{2\pi}{\alpha t}(m_1 - m_2),$$

and this contradicts the assumption that α is irrational. Conclusion (c) follows from (b).

This completes the proof of the theorem.

8. Let $\xi = (\xi_1, \ldots, \xi_k)$ be a random vector,

$$\varphi_\xi(t) = \mathsf{E}e^{i(t, \xi)}, \qquad t = (t_1, \ldots, t_k),$$

its characteristic function. Let us suppose that $\mathsf{E}|\xi_i|^n < \infty$ for some $n \geq 1$, $i = 1, \ldots, k$. From the inequalities of Hölder (6.29) and Lyapunov (6.27) it follows that the (mixed) moments $\mathsf{E}(\xi_1^{\nu_1} \cdots \xi_k^{\nu_k})$ exist for all nonnegative $\nu_1, \ldots, \nu_k$ such that $\nu_1 + \cdots + \nu_k \leq n$.

As in Theorem 1, this implies the existence and continuity of the partial derivatives

$$\frac{\partial^{\nu_1 + \cdots + \nu_k}}{\partial t_1^{\nu_1} \cdots \partial t_k^{\nu_k}} \varphi_\xi(t_1, \ldots, t_k)$$

for $\nu_1 + \cdots + \nu_k \leq n$. Then if we expand $\varphi_\xi(t_1, \ldots, t_k)$ in a Taylor series, we see that

$$\varphi_\xi(t_1, \ldots, t_k) = \sum_{\nu_1 + \cdots + \nu_k \leq n} \frac{i^{\nu_1 + \cdots + \nu_k}}{\nu_1! \cdots \nu_k!} m_\xi^{(\nu_1, \ldots, \nu_k)} t_1^{\nu_1} \cdots t_k^{\nu_k} + o(|t|^n), \quad (34)$$

where $|t| = |t_1| + \cdots + |t_k|$ and

$$m_\xi^{(\nu_1, \ldots, \nu_k)} = \mathsf{E}\xi_1^{\nu_1} \cdots \xi_k^{\nu_k}$$

is the *mixed moment of order* $\nu = (\nu_1, \ldots, \nu_k)$.

Now $\varphi_\xi(t_1, \ldots, t_k)$ is continuous, $\varphi_\xi(0, \ldots, 0) = 1$, and consequently this function is different from zero in some neighborhood $|t| < \delta$ of zero. In this neighborhood the partial derivative

$$\frac{\partial^{\nu_1 + \cdots + \nu_k}}{\partial t_1^{\nu_1} \cdots \partial t_k^{\nu_k}} \ln \varphi_\xi(t_1, \ldots, t_k)$$

exists and is continuous, where $\ln z$ denotes the principal value of the logarithm (if $z = re^{i\theta}$, we take $\ln z$ to be $\ln r + i\theta$). Hence we can expand $\ln \varphi_\xi(t_1, \ldots, t_k)$ by Taylor's formula,

$$\ln \varphi_\xi(t_1 \cdots t_k) = \sum_{\nu_1 + \cdots + \nu_k \leq n} \frac{i^{\nu_1 + \cdots + \nu_k}}{v_1! \cdots \nu_k!} s_\xi^{(\nu_1, \ldots, \nu_k)} t_1^{\nu_1} \cdots t_k^{\nu_k} + o(|t|^n), \quad (35)$$

where the coefficients $s_\xi^{(\nu_1,\dots,\nu_k)}$ are the (*mixed*) *semi-invariants* or *cumulants* of order $\nu = \nu(\nu_1, \dots, \nu_k)$ of $\xi = \xi_1, \dots, \xi_k$.

Observe that if ξ and η are independent, then

$$\ln \varphi_{\xi+\eta}(t) = \ln \varphi_\xi(t) + \ln \varphi_\eta(t), \tag{36}$$

and therefore

$$s_{\xi+\eta}^{(\nu_1,\dots,\nu_k)} = s_\xi^{(\nu_1,\dots,\nu_k)} + s_\eta^{(\nu_1,\dots,\nu_k)}. \tag{37}$$

(It is this property that gives rise to the term "semi-invariant" for $s_\xi^{(\nu_1,\dots,\nu_k)}$.)

To simply the formulas and make (34) and (35) look "one-dimensional," we introduce the following notation.

If $\nu = (\nu_1, \dots, \nu_k)$ is a vector whose components are nonnegative integers, we put

$$\nu! = \nu_1! \cdots \nu_k!, \qquad |\nu| = \nu_1 + \cdots + \nu_k, \qquad t^\nu = t_1^{\nu_1} \cdots t_k^{\nu_k}.$$

We also put $s_\xi^{(\nu)} = s_\xi^{(\nu_1,\dots,\nu_k)}$, $m_\xi^{(\nu)} = m_\xi^{(\nu_1,\dots,\nu_k)}$.

Then (34) and (35) can be written

$$\varphi_\xi(t) = \sum_{|\nu| \le n} \frac{i^{|\nu|}}{\nu!} m_\xi^{(\nu)} t^\nu + o(|t|^n), \tag{38}$$

$$\ln \varphi_\xi(t) = \sum_{|\nu| \le n} \frac{i^{|\nu|}}{\nu!} s_\xi^{(\nu)} t^\nu + o(|t|^n). \tag{39}$$

The following theorem and its corollaries give *formulas that connect moments and semi-invariants*.

Theorem 6. *Let $\xi = (\xi_1, \dots, \xi_k)$ be a random vector with* $\mathsf{E}|\xi_i|^n < \infty$, *$i = 1, \dots, k$, $n \ge 1$. Then for $\nu = (\nu_1, \dots, \nu_k)$ such that $|\nu| \le n$*

$$m_\xi^{(\nu)} = \sum_{\lambda^{(1)} + \cdots + \lambda^{(q)} = \nu} \frac{1}{q!} \frac{\nu!}{\lambda^{(1)}! \cdots \lambda^{(q)}!} \prod_{p=1}^{q} s^{(\lambda^{(p)})}, \tag{40}$$

$$s_\xi^{(\nu)} = \sum_{\lambda^{(1)} + \cdots + \lambda^{(q)} = \nu} \frac{(-1)^{q-1}}{q} \frac{\nu!}{\lambda^{(1)}! \cdots \lambda^{(q)}!} \prod_{p=1}^{q} m_\xi^{(\lambda^{(p)})}, \tag{41}$$

where $\sum_{\lambda^{(1)} + \cdots + \lambda^{(q)} = \nu}$ indicates summation over all ordered sets of nonnegative integral vectors $\lambda^{(p)}$, $|\lambda^{(p)}| > 0$, whose sum is ν.

Proof. Since

$$\varphi_\xi(t) = \exp(\ln \varphi_\xi(t)),$$

if we expand the function exp by Taylor's formula and use (39), we obtain

$$\varphi_\xi(t) = 1 + \sum_{q=1}^{n} \frac{1}{q!} \left(\sum_{1 \le |\lambda| \le n} \frac{i^{|\lambda|}}{\lambda!} s_\xi^{(\lambda)} t^\lambda \right)^q + o(|t|^n). \tag{42}$$

Comparing terms in t^λ on the right-hand sides of (38) and (42), and using $|\lambda^{(1)}| + \cdots + |\lambda^{(q)}| = |\lambda^{(1)} + \cdots + \lambda^{(q)}|$, we obtain (40).

Moreover,

$$\ln \varphi_\xi(t) = \ln\left[1 + \sum_{1 \le |\lambda| \le n} \frac{i^{|\lambda|}}{\lambda!} m_\xi^{(\lambda)} t^\lambda + o(|t|^n)\right]. \tag{43}$$

For small z we have the expansion

$$\ln(1+z) = \sum_{q=1}^{n} \frac{(-1)^{q-1}}{q} z^q + o(z^q).$$

Using this in (43) and then comparing the coefficients of t^λ with the corresponding coefficients on the right-hand side of (38), we obtain (41).

Corollary 1. *The following formulas connect moments and semi-invariants*:

$$m_\xi^{(\nu)} = \sum_{\{r_1\lambda^{(1)} + \cdots + r_x\lambda^{(x)} = \nu\}} \frac{1}{r_1! \cdots r_x!} \frac{\nu!}{(\lambda^{(1)}!)^{r_1} \cdots (\lambda^{(x)}!)^{r_x}} \prod_{j=1}^{x} [s_\xi^{(\lambda^{(j)})}]^{r_j}, \tag{44}$$

$$s_\xi^{(\nu)} = \sum_{\{r_1\lambda^{(1)} + \cdots + r_x\lambda^{(x)} = \nu\}} \frac{(-1)^{q-1}(q-1)!}{r_1! \cdots r_x!} \frac{\nu!}{(\lambda^{(1)}!)^{r_1} \cdots (\lambda^{(x)}!)^{r_x}} \prod_{j=1}^{x} [m_\xi^{(\lambda^{(j)})}]^{r_j}, \tag{45}$$

where $\sum_{\{r_1\lambda^{(1)} + \cdots + r_x\lambda^{(x)} = \nu\}}$ *denotes summation over all unordered sets of different nonnegative integral vectors* $\lambda^{(j)}$, $|\lambda^{(j)}| > 0$, *and over all ordered sets of positive integral numbers* r_j *such that* $r_1\lambda^{(1)} + \cdots + r_x\lambda^{(x)} = \nu$.

To establish (44) we suppose that among all the vectors $\lambda^{(1)}, \ldots, \lambda^{(q)}$ that occur in (40), there are r_1 equal to $\lambda^{(i_1)}, \ldots, r_x$ equal to $\lambda^{(i_x)}$ ($r_j > 0$, $r_1 + \cdots + r_x = q$), where all the $\lambda^{(i_s)}$ are different. There are $q!/(r_1! \ldots r_x!)$ different sets of vectors, corresponding (except for order) with the set $\{\lambda^{(1)}, \ldots \lambda^{(q)}\}$). But if two sets, say, $\{\lambda^{(1)}, \ldots, \lambda^{(q)}\}$ and $\{\bar\lambda^{(1)}, \ldots, \bar\lambda^{(q)}\}$ differ only in order, then $\prod_{p=1}^{q} s_\xi^{(\lambda^{(p)})} = \prod_{p=1}^{q} s_\xi^{(\bar\lambda^{(p)})}$. Hence if we identify sets that differ only in order, we obtain (44) from (40).

Formula (45) can be deduced from (41) in a similar way.

Corollary 2. *Let us consider the special case when* $\nu = (1, \ldots, 1)$. *In this case the moments* $m_\xi^{(\nu)} \equiv \mathsf{E}\xi_1 \cdots \xi_k$, *and the corresponding semi-invariants, are called* simple.

Formulas connecting simple moments and simple semi-invariants can be read off from the formulas given above. However, it is useful to have them written in a different way.

For this purpose, we introduce the following notation.

Let $\xi = (\xi_1, \ldots, \xi_k)$ be a vector, and $I_\xi = \{1, 2, \ldots, k\}$ its set of indices. If $I \subseteq I_\xi$, let ξ_I denote the vector consisting of the components of ξ whose indices belong to I. Let $\chi(I)$ be the vector $\{\chi_1, \ldots, \chi_n\}$ for which $\chi_i = 1$ if $i \in I$, and $\chi_i = 0$ if $i \notin I$. These vectors are in one-to-one correspondence with the sets $I \subseteq I_\xi$. Hence we can write

$$m_\xi(I) = m_\xi^{(\chi(I))}, \qquad s_\xi(I) = s_\xi^{(\chi(I))}.$$

In other words, $m_\xi(I)$ and $s_\xi(I)$ are simple moments and semi-invariants of the subvector ξ_I of ξ.

In accordance with the definition given on p. 12, a *decomposition* of a set I is an unordered collection of disjoint nonempty sets I_p such that $\sum_p I_p = I$.

In terms of these definitions, we have the formulas

$$m_\xi(I) = \sum_{\sum_{p=1}^q I_p = I} \prod_{p=1}^{q} s_\xi(I_p), \tag{46}$$

$$s_\xi(I) = \sum_{\sum_{p=1}^q I_p = I} (-1)^{q-1}(q-1)! \prod_{p=1}^{q} m_\xi(I_p). \tag{47}$$

where $\sum_{\sum_{p=1}^q I_p = I}$ denotes summation over all decompositions of I, $1 \le q \le N(I)$, where $N(I)$ is the number of elements of the set I.

We shall derive (46) from (44). If $\nu = \chi(I)$ and $\lambda^{(1)} + \cdots + \lambda^{(q)} = \nu$, then $\lambda^{(p)} = \chi(I_p)$, $I_p \subseteq I$, where the $\lambda^{(p)}$ are all different, $\lambda^{(p)}! = \nu! = 1$, and every unordered set $\{\chi(I_1), \ldots, \chi(I_q)\}$ is in one-to-one correspondence with the decomposition $I = \sum_{p=1}^q I_p$. Consequently (46) follows from (44).

In a similar way, (47) follows from (35).

Example 1. Let ξ be a random variable ($k = 1$) and $m_n = m_\xi^{(n)} = \mathsf{E}\xi^n$, $s_n = s_\xi^{(n)}$. Then (40) and (41) imply the following formulas:

$$\begin{aligned}
m_1 &= s_1, \\
m_2 &= s_2 + s_1^2, \\
m_3 &= s_3 + 3s_1s_2 + s_1^3, \\
m_4 &= s_4 + 3s_2^2 + 4s_1s_3 + 6s_1^2s_2 + s_1^4, \\
&\cdots\cdots\cdots\cdots\cdots\cdots
\end{aligned} \tag{48}$$

and

$$\begin{aligned}
s_1 &= m_1 = \mathsf{E}\xi, \\
s_2 &= m_2 - m_1^2 = \mathsf{V}\xi, \\
s_3 &= m_3 - 3m_1m_2 + 2m_1^3, \\
s_4 &= m_4 - 3m_2^2 - 4m_1m_3 + 12m_1^2m_2 - 6m_1^4, \\
&\cdots\cdots\cdots\cdots\cdots\cdots
\end{aligned} \tag{49}$$

EXAMPLE 2. Let $\xi \sim \mathcal{N}(m, \sigma^2)$. Since, by (9),

$$\ln \varphi_\xi(t) = itm - \frac{t^2\sigma^2}{2},$$

we have $s_1 = m$, $s_2 = \sigma^2$ by (39), and all the semi-invariants, from the third on, are zero: $s_n = 0$, $n \geq 3$.

We may observe that by Marcinkiewicz's theorem a function $\exp \mathscr{P}(t)$, where $\mathscr{P}$ is a polynomial, can be a characteristic function only when the degree of that polynomial is at most 2. It follows, in particular, that the Gaussian distribution is the only distribution with the property that all its semi-invariants s_n are zero from a certain index onward.

EXAMPLE 3. If ξ is a Poisson random variable with parameter $\lambda > 0$, then by (11)

$$\ln \varphi_\xi(t) = \lambda(e^{it} - 1).$$

It follows that

$$s_n = \lambda \tag{50}$$

for all $n \geq 1$.

EXAMPLE 4. Let $\xi = (\xi_1, \ldots, \xi_n)$ be a random vector. Then

$$\begin{aligned} m_\xi(1) &= s_\xi(1), \\ m_\xi(1, 2) &= s_\xi(1, 2) + s_\xi(1)s_\xi(2), \\ m_\xi(1, 2, 3) &= s_\xi(1, 2, 3) + s_\xi(1, 2)s_\xi(3) + \\ &\quad + s_\xi(1, 3)s_\xi(2) + \\ &\quad + s_\xi(2, 3)s_\xi(1) + s_\xi(1)s_\xi(2)s_\xi(3) \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \end{aligned} \tag{51}$$

These formulas show that the simple moments can be expressed in terms of the simple semi-invariants in a very *symmetric* way. If we put $\xi_1 \equiv \xi_2 \equiv \cdots \equiv \xi_k$, we then, of course, obtain (48).

The group-theoretical origin of the coefficients in (48) becomes clear from (51). It also follows from (51) that

$$s_\xi(1, 2) = m_\xi(1, 2) - m_\xi(1)m_\xi(2) = \mathsf{E}\xi_1\xi_2 - \mathsf{E}\xi_1\mathsf{E}\xi_2, \tag{52}$$

i.e., $s_\xi(1, 2)$ is just the *covariance* of ξ_1 and ξ_2.

9. Let ξ be a random variable with distribution function $F = F(x)$ and characteristic function $\varphi(t)$. Let us suppose that all the moments $m_n = \mathsf{E}\xi^n$, $n \geq 1$, exist.

It follows from Theorem 2 that a characteristic function uniquely determines a probability distribution. Let us now ask the following question (uniqueness for the moment problem): Do the moments $\{m_n\}_{n\geq 1}$ determine the *probability distribution*?

More precisely, let F and G be distribution functions with the same moments, i.e.

$$\int_{-\infty}^{\infty} x^n \, dF(x) = \int_{-\infty}^{\infty} x^n \, dG(x) \tag{53}$$

for all integers $n \geq 0$. The question is whether F and G must be the same.

In general, the answer is "no." To see this, consider the distribution F with density

$$f(x) = \begin{cases} ke^{-\alpha x^\lambda}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

where $\alpha > 0, 0 < \lambda < \frac{1}{2}$, and k is determined by the condition $\int_0^\infty f(x)\, dx = 1$.

Write $\beta = \alpha \tan \lambda\pi$ and let $g(x) = 0$ for $x \leq 0$ and

$$g(x) = ke^{-\alpha x^\lambda}[1 + \varepsilon \sin(\beta x^\lambda)], \qquad |\varepsilon| < 1, \quad x > 0.$$

It is evident that $g(x) \geq 0$. Let us show that

$$\int_0^\infty x^n e^{-\alpha x^\lambda} \sin \beta x^\lambda \, dx = 0 \tag{54}$$

for all integers $n \geq 0$.

For $p > 0$ and complex q with $\operatorname{Re} q > 0$, we have

$$\int_0^\infty t^{p-1} e^{-qt} \, dt = \frac{\Gamma(p)}{q^p}.$$

Take $p = (n+1)/\lambda$, $q = \alpha + i\beta$, $t = x^\lambda$. Then

$$\begin{aligned}
\int_0^\infty x^{\lambda[((n+1)/\lambda)-1]} e^{-(\alpha+i\beta)x^\lambda} \lambda x^{\lambda-1} \, dx &= \lambda \int_0^\infty x^n e^{-(\alpha+i\beta)x^\lambda} \, dx \\
&= \lambda \int_0^\infty x^n e^{-\alpha x^\lambda} \cos\beta x^\alpha \, dx - i\lambda \int_0^\infty x^n e^{-\alpha x^\lambda} \sin \beta x^\lambda \, dx \\
&= \frac{\Gamma\left(\frac{n+1}{\lambda}\right)}{\alpha^{(n+1)/\lambda}(1 + i \tan \lambda\pi)^{(n+1)/\lambda}}.
\end{aligned} \tag{55}$$

But

$$\begin{aligned}
(1 + i \tan \lambda\pi)^{(n+1)/\lambda} &= (\cos \lambda\pi + i \sin \lambda\pi)^{(n+1)/\lambda}(\cos \lambda\pi)^{-(n+1)/\lambda} \\
&= e^{i\pi(n+1)}(\cos \lambda\pi)^{-(n+1)/\lambda} \\
&= \cos \pi(n+1) \cdot \cos(\lambda\pi)^{-(n+1)/\lambda},
\end{aligned}$$

since $\sin \pi(n+1) = 0$.

Hence right-hand side of (55) is real and therefore (54) is valid for all integral $n \geq 0$. Now let $G(x)$ be the distribution function with density $g(x)$. It follows from (54) that the distribution functions F and G have equal moments, i.e. (53) holds for all integers $n \geq 0$.

We now give some conditions that guarantee the uniqueness of the solution of the moment problem.

Theorem 7. *Let $F = F(x)$ be a distribution function and $\mu_n = \int_{-\infty}^{\infty} |x|^n \, dF(x)$. If*

$$\overline{\lim_{n\to\infty}} \frac{\mu_n^{1/n}}{n} < \infty, \tag{56}$$

the moments $\{m_n\}_{n\geq 1}$, where $m_n = \int_{-\infty}^{\infty} x^n \, dF(x)$, determine the distribution $F = F(x)$ uniquely.

Proof. It follows from (56) and conclusion (7) of Theorem 1 that there is a $t_0 > 0$ such that, for all $|t| \leq t_0$, the characteristic function

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx} \, dF(x)$$

can be represented in the form

$$\varphi(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} m_k$$

and consequently the moments $\{m_n\}_{n\geq 1}$ uniquely determine the characteristic function $\varphi(t)$ for $|t| \leq t_0$.

Take a point s with $|s| \leq t_0/2$. Then, as in the proof of (15), we deduce from (56) that

$$\varphi(t) = \sum_{k=0}^{\infty} \frac{i^k (t-s)^k}{k!} \varphi^{(k)}(s)$$

for $|t - s| \leq t_0$, where

$$\varphi^{(k)}(s) = i^k \int_{-\infty}^{\infty} x^k e^{isx} \, dF(x)$$

is uniquely determined by the moments $\{m_n\}_{n\geq 1}$. Consequently the moments determine $\varphi(t)$ uniquely for $|t| \leq \frac{3}{2} t_0$. Continuing this process, we see that $\{m_n\}_{n\geq 1}$ determines $\varphi(t)$ uniquely for all t, and therefore also determines $F(x)$.

This completes the proof of the theorem.

Corollary 1. *The moments completely determine the probability distribution if it is concentrated on a finite interval.*

Corollary 2. *A sufficient condition for the moment problem to have a unique solution is that*

$$\overline{\lim_{n\to\infty}} \frac{(m_{2n})^{1/2n}}{2n} < \infty. \tag{57}$$

For the proof it is enough to observe that the odd moments can be estimated in terms of the even ones, and then use (56).

EXAMPLE. Let $F(x)$ be the normal distribution function,

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-t^2/2\sigma^2}\, dt.$$

Then $m_{2n+1} = 0$, $m_{2n} = [(2n)!/2^n n!]\sigma^{2n}$, and it follows from (57) that these are the moments only of the normal distribution.

Finally we state, without proof:

Carleman's test for the uniqueness of the moment problem.

(a) *Let $\{m_n\}_{n\geq 1}$ be the moments of a probability distribution, and let*

$$\sum_{n=0}^{\infty} \frac{1}{(m_{2n})^{1/2n}} = \infty.$$

Then they determine the probability distribution uniquely.

(b) *If $\{m_n\}_{n\geq 1}$ are the moments of a distribution that is concentrated on $[0, \infty)$, then the solution will be unique if we require only that*

$$\sum_{n=0}^{\infty} \frac{1}{(m_n)^{1/2n}} = \infty.$$

10. Let $F = F(x)$ and $G = G(x)$ be distribution functions with characteristic functions $f = f(t)$ and $g = g(t)$, respectively. The following theorem, which we give without proof, makes it possible to estimate how close F and G are to each other (in the uniform metric) in terms of the closeness of f and g.

Theorem (Esseen's Inequality). *Let $G(x)$ have derivative $G'(x)$ with $\sup|G'(x)| \leq C$. Then for every $T > 0$*

$$\sup_x |F(x) - G(x)| \leq \frac{2}{\pi} \int_0^T \left| \frac{f(t) - g(t)}{t} \right| dt + \frac{24}{\pi T} \sup_x |G'(x)|. \tag{58}$$

(This will be used in §6 of Chapter III to prove a theorem on the rapidity of convergence in the central limit theorem.)

11. Problems

1. Let ξ and η be independent random variables, $f(x) = f_1(x) + if_2(x)$, $g(x) = g_1(x) + ig_2(x)$, where $f_k(x)$ and $g_k(x)$ are Borel functions, $k=1, 2$. Show that if $\mathsf{E}|f(\xi)| < \infty$ and $\mathsf{E}|g(\eta)| < \infty$, then
$$\mathsf{E}|f(\xi)g(\eta)| < \infty$$
and
$$\mathsf{E}f(\xi)g(\eta) = \mathsf{E}f(\xi) \cdot \mathsf{E}g(\eta).$$

2. Let $\xi = (\xi_1, \ldots, \xi_n)$ and $\mathsf{E}\|\xi\|^n < \infty$, where $\|\xi\| = +\sqrt{\sum \xi_i^2}$. Show that
$$\varphi_\xi(t) = \sum_{k=0}^{n} \frac{i^k}{k!} \mathsf{E}(t, \xi)^k + \varepsilon_n(t)\|t\|^n,$$
where $t = (t_1, \ldots, t_n)$ and $\varepsilon_n(t) \to 0$, $t \to 0$.

3. Prove Theorem 2 for n-dimensional distribution functions $F = F_n(x_1, \ldots, x_n)$ and $G_n(x_1, \ldots, x_n)$.

4. Let $F = F(x_1, \ldots, x_n)$ be an n-dimensional distribution function and $\varphi = \varphi(t_1, \ldots, t_n)$ its characteristic function. Using the notation of (3.12), establish the inversion formula
$$\mathsf{P}(a, b] = \lim_{c \to \infty} \frac{1}{(2\pi)^n} \int_{-c}^{c} \prod_{k=1}^{n} \frac{e^{it_k a_k} - e^{it_k b_k}}{it_k} \varphi(t_1, \ldots, t_k)\, dt_1 \cdots dt_k.$$
(We are to suppose that $(a, b]$ is an interval of continuity of $\mathsf{P}(a, b]$, i.e. for $k = 1, \ldots, n$ the points a_k, b_k are points of continuity of the marginal distribution functions $F_k(x_k)$ which are obtained from $F(x_1, \ldots, x_n)$ by taking all the variables except x_k equal to $+\infty$.)

5. Let $\varphi_k(t)$, $k \geq 1$, be a characteristic function, and let the nonnegative numbers λ_k, $k \geq 1$, satisfy $\sum \lambda_k = 1$. Show that $\sum \lambda_k \varphi_k(t)$ is a characteristic function.

6. If $\varphi(t)$ is a characteristic function, are $\operatorname{Re} \varphi(t)$ and $\operatorname{Im} \varphi(t)$ characteristic functions?

7. Let φ_1, φ_2 and φ_3 be characteristic functions, and $\varphi_1\varphi_2 = \varphi_1\varphi_3$. Does it follow that $\varphi_2 = \varphi_3$?

8. Construct the characteristic functions of the distributions given in Tables 1 and 2 of §3.

9. Let ξ be an integral-valued random variable and $\varphi_\xi(t)$ its characteristic function. Show that
$$\mathsf{P}(\xi = k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ikt} \varphi_\xi(t)\, dt, \qquad k = 0, \pm 1, \pm 2 \ldots.$$

§13. Gaussian Systems

1. Gaussian, or normal, distributions, random variables, processes, and systems play an extremely important role in probability theory and in mathematical statistics. This is explained in the first instance by the central

limit theorem (§4 of Chapter III and §8 of Chapter VII), of which the De Moivre–Laplace limit theorem is a special case (§6, Chapter I). According to this theorem, the normal distribution is universal in the sense that the distribution of the sum of a large number of random variables or random vectors, subject to some not very restrictive conditions, is closely approximated by this distribution.

This is what provides a theoretical explanation of the "law of errors" of applied statistics, which says that errors of measurement that result from large numbers of independent "elementary" errors obey the normal distribution.

A multidimensional Gaussian distribution is specified by a small number of parameters; this is a definite advantage in using it in the construction of simple probabilistic models. Gaussian random variables have finite second moments, and consequently they can be studied by Hilbert space methods. Here it is important that in the Gaussian case "uncorrelated" is equivalent to "independent," so that the results of L^2-theory can be significantly strengthened.

2. Let us recall that (see §8) a random variable $\xi = \xi(\omega)$ is Gaussian, or normally distributed, with parameters m and σ^2 ($\xi \sim \mathcal{N}(m, \sigma^2)$), $|m| < \infty$, $\sigma^2 > 0$, if its density $f_\xi(x)$ has the form

$$f_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2/2\sigma^2}, \tag{1}$$

where $\sigma = +\sqrt{\sigma^2}$.

As $\sigma \downarrow 0$, the density $f_\xi(x)$ "converges to the δ-function supported at $x = m$." It is natural to say that ξ is normally distributed with mean m and $\sigma^2 = 0$ ($\xi \sim \mathcal{N}(m, 0)$) if ξ has the property that $\mathsf{P}(\xi = m) = 1$.

We can, however, give a definition that applies both to the *nondegenerate* ($\sigma^2 > 0$) and the *degenerate* ($\sigma^2 = 0$) cases. Let us consider the characteristic function $\varphi_\xi(t) \equiv \mathsf{E}e^{it\xi}$, $t \in R$.

If $\mathsf{P}(\xi = m) = 1$, then evidently

$$\varphi_\xi(t) = e^{itm}, \tag{2}$$

whereas if $\xi \sim \mathcal{N}(m, \sigma^2)$, $\sigma^2 > 0$,

$$\varphi_\xi(t) = e^{itm-(1/2)t^2\sigma^2}. \tag{3}$$

It is obvious that when $\sigma^2 = 0$ the right-hand sides of (2) and (3) are the same. It follows, by Theorem 2 of §12, that the Gaussian random variable with parameters m and σ^2 ($|m| < \infty$, $\sigma^2 \geq 0$) must be the same as the random variable whose characteristic function is given by (3). This is an illustration of the "attraction of characteristic functions," a very useful technique in the multidimensional case.

Let $\xi = (\xi_1, \ldots, \xi_n)$ be a random vector and

$$\varphi_\xi(t) = \mathsf{E}e^{i(t,\xi)}, \qquad t = (t_1, \ldots, t_n) \in R^n, \tag{4}$$

its characteristic function (see Definition 2, §12).

Definition 1. A random vector $\xi = (\xi_1, \ldots, \xi_n)$ is *Gaussian*, or *normally distributed*, if its characteristic function has the form

$$\varphi_\xi(t) = e^{i(t,m) - (1/2)(\mathbb{R}t,t)}, \tag{5}$$

where $m = (m_1, \ldots, m_n)$, $|m_k| < \infty$ and $\mathbb{R} = \|r_{kl}\|$ is a symmetric nonnegative definite $n \times n$ matrix; we use the abbreviation $\xi \sim \mathcal{N}(m, \mathbb{R})$.

This definition immediately makes us ask whether (5) is in fact a characteristic function. Let us show that it is.

First suppose that $\mathbb{R}$ is nonsingular. Then we can define the inverse $A = \mathbb{R}^{-1}$ and the function

$$f(x) = \frac{|A|^{1/2}}{(2\pi)^{n/2}} \exp\{-\tfrac{1}{2}(A(x-m),(x-m))\}, \tag{6}$$

where $x = (x_1, \ldots, x_n)$ and $|A| = \det A$. This function is nonnegative. Let us show that

$$\int_{R^n} e^{i(t,x)} f(x)\,dx = e^{i(t,m) - (1/2)(\mathbb{R}t,t)},$$

or equivalently that

$$I_n \equiv \int_{R^n} e^{i(t,x-m)} \frac{|A|^{1/2}}{(2\pi)^{n/2}} e^{-(1/2)(A(x-m),(x-m))}\,dx = e^{-(1/2)(\mathbb{R}t,t)}. \tag{7}$$

Let us make the change of variable

$$x - m = \mathcal{O}u, \qquad t = \mathcal{O}v,$$

where $\mathcal{O}$ is an orthogonal matrix such that

$$\mathcal{O}^{\mathrm{T}}\mathbb{R}\mathcal{O} = D,$$

and

$$D = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{pmatrix}$$

is a diagonal matrix with $d_i \geq 0$ (see the proof of the lemma in §8). Since $|\mathbb{R}| = \det \mathbb{R} \neq 0$, we have $d_i > 0$, $i = 1, \ldots, n$. Therefore

$$|A| = |\mathbb{R}^{-1}| = d_1^{-1} \cdots d_n^{-1}. \tag{8}$$

Moreover (for notation, see Subsection 1, §12)

$$\begin{aligned} i(t, x - m) - \tfrac{1}{2}(A(x - m), x - m)) &= i(\mathcal{O}v, \mathcal{O}u) - \tfrac{1}{2}(A\mathcal{O}u, \mathcal{O}u) \\ &= i(\mathcal{O}v)^{\mathrm{T}}\mathcal{O}u - \tfrac{1}{2}(\mathcal{O}u)^{\mathrm{T}}A(\mathcal{O}u) \\ &= iv^{\mathrm{T}}u - \tfrac{1}{2}u^{\mathrm{T}}\mathcal{O}^{\mathrm{T}}A\mathcal{O}u \\ &= iv^{\mathrm{T}}u - \tfrac{1}{2}u^{\mathrm{T}}D^{-1}u. \end{aligned}$$

Together with (8) and (12.9), this yields

$$\begin{aligned} I_n &= (2\pi)^{-n/2}(d_1 \cdots d_n)^{-1/2} \int_{R^n} \exp(iv^{\mathrm{T}}u - \tfrac{1}{2}u^{\mathrm{T}}D^{-1}u)\, du \\ &= \prod_{k=1}^{n} (2\pi d_k)^{-1/2} \int_{-\infty}^{\infty} \exp\left(iv_k u_k - \frac{u_k^2}{2d_k}\right) du_k = \prod_{k=1}^{n} \exp(-\tfrac{1}{2}v_k^2 d_k) \\ &= \exp(-\tfrac{1}{2}v^{\mathrm{T}}Dv) = \exp(-\tfrac{1}{2}v^{\mathrm{T}}\mathcal{O}^{\mathrm{T}}\mathbb{R}\mathcal{O}v) = \exp(-\tfrac{1}{2}t^{\mathrm{T}}\mathbb{R}t) = \exp(-\tfrac{1}{2}(\mathbb{R}t, t)). \end{aligned}$$

It also follows from (6) that

$$\int_{R^n} f(x)\, dx = 1. \tag{9}$$

Therefore (5) is the characteristic function of a nondegenerate n-dimensional Gaussian distribution (see Subsection 3, §3).

Now let $\mathbb{R}$ be singular. Take $\varepsilon > 0$ and consider the positive definite symmetric matrix $\mathbb{R}^{\varepsilon} \equiv \mathbb{R} + \varepsilon E$. Then by what has been proved,

$$\varphi^{\varepsilon}(t) = \exp\{i(t, m) - \tfrac{1}{2}(\mathbb{R}^{\varepsilon}t, t)\}$$

is a characteristic function:

$$\varphi^{\varepsilon}(t) = \int_{R^n} e^{i(t,x)}\, dF_{\varepsilon}(x),$$

where $F_{\varepsilon}(x) = F_{\varepsilon}(x_1, \ldots, x_n)$ is an n-dimensional distribution function.

As $\varepsilon \to 0$,

$$\varphi^{\varepsilon}(t) \to \varphi(t) = \exp\{i(t, m) - \tfrac{1}{2}(\mathbb{R}^{\varepsilon}t, t)\}.$$

The limit function $\varphi(t)$ is continuous at $(0, \ldots, 0)$. Hence, by Theorem 1 and Problem 1 of §3 of Chapter III, it is a characteristic function.

We have therefore established Theorem 1.

3. Let us now discuss the significance of the vector m and the matrix $\mathbb{R} = \|r_{kl}\|$ that appear in (5).

Since

$$\ln \varphi_{\xi}(t) = i(t, m) - \tfrac{1}{2}(\mathbb{R}t, t) = i\sum_{k=1}^{n} t_k m_k - \frac{1}{2}\sum_{k,l=1}^{n} r_{kl}t_k t_l, \tag{10}$$

we find from (12.35) and the formulas that connect the moments and the semi-invariants that

$$m_1 = s_\xi^{(1,0,\dots,0)} = \mathsf{E}\xi_1, \dots, m_k = s_\xi^{(0,\dots,0,1)} = \mathsf{E}\xi_k.$$

Similarly

$$r_{11} = s_\xi^{(2,0,\dots,0)} = \mathsf{V}\xi_1, \quad r_{12} = s_\xi^{(1,1,0,\dots)} = \operatorname{cov}(\xi_1, \xi_2),$$

and generally

$$r_{kl} = \operatorname{cov}(\xi_k, \xi_l).$$

Consequently m is the *mean-value vector* of ξ and $\mathbb{R}$ is its covariance matrix.

If $\mathbb{R}$ is nonsingular, we can obtain this result in a different way. In fact, in this case ξ has a density $f(x)$ given by (6).

A direct calculation shows that

$$\mathsf{E}\xi_k = \int x_k f(x)\,dx = m_k, \tag{11}$$

$$\operatorname{cov}(\xi_k, \xi_l) = \int (x_k - m_k)(x_l - m_l) f(x)\,dx = r_{kl}.$$

4. Let us discuss some properties of Gaussian vectors.

Theorem 1

(a) *The components of a Gaussian vector are uncorrelated if and only if they are independent.*

(b) *A vector $\xi = (\xi_1, \dots, \xi_n)$ is Gaussian if and only if, for every vector $\lambda = (\lambda_1, \dots, \lambda_n)$, $\lambda_k \in R$, the random variable $(\xi, \lambda) = \lambda_1\xi_1 + \cdots + \lambda_n\xi_n$ has a Gaussian distribution.*

Proof. (a) If the components of $\xi = (\xi_1, \dots, \xi_n)$ are uncorrelated, it follows from the form of the characteristic function $\varphi_\xi(t)$ that it is a product of characteristic functions. Therefore, by Theorem 4 of §12, the components are independent.

The converse is evident, since independence always implies lack of correlation.

(b) If ξ is a Gaussian vector, it follows from (5) that

$$\mathsf{E}\exp\{it(\xi_1\lambda_1 + \cdots + \xi_n\lambda_n)\} = \exp\left\{it\left(\sum \lambda_k m_k\right) - \frac{t^2}{2}\left(\sum r_{kl}\lambda_k\lambda_l\right)\right\}, \quad t \in R,$$

and consequently

$$(\xi, \lambda) \sim \mathcal{N}\left(\sum \lambda_k m_k, \sum r_{kl}\lambda_k\lambda_l\right).$$

Conversely, to say that the random variable $(\xi, \lambda) = \xi_1\lambda_1 + \cdots + \xi_n\lambda_n$ is Gaussian means, in particular, that

$$\mathsf{E}e^{i(\xi,\lambda)} = \exp\left\{i\mathsf{E}(\xi, \lambda) - \frac{\mathsf{V}(\xi, \lambda)}{2}\right\} = \exp\{i \sum \lambda_k \mathsf{E}\xi_k - \tfrac{1}{2}\sum \lambda_k\lambda_l \operatorname{cov}(\xi_k, \xi_l)\}.$$

Since $\lambda_1, \ldots, \lambda_n$ are arbitrary it follows from Definition 1 that the vector $\xi = (\xi_1, \ldots, \xi_n)$ is Gaussian.

This completes the proof of the theorem.

Remark. Let (θ, ξ) be a Gaussian vector with $\theta = (\theta_1, \ldots, \theta_k)$ and $\xi = (\xi_1, \ldots, \xi_k)$. If θ and ξ are uncorrelated, i.e. $\operatorname{cov}(\theta_i, \xi_j) = 0$, $i = 1, \ldots, k$; $j = 1, \ldots, l$, they are independent.

The proof is the same as for conclusion (a) of the theorem.

Let $\xi = (\xi_1, \ldots, \xi_n)$ be a Gaussian vector; let us suppose, for simplicity, that its mean-value vector is zero. If rank $\mathbb{R} = r < n$, then (as was shown in §11), there are $n - r$ linear relations connecting $\xi_1, \ldots, \xi_n$. We may then suppose that, say, $\xi_1, \ldots, \xi_r$ are linearly independent, and the others can be expressed linearly in terms of them. Hence all the basic properties of the vector $\xi = \xi_1, \ldots, \xi_n$ are determined by the first r components $(\xi_1, \ldots, \xi_r)$ for which the corresponding covariance matrix is already known to be nonsingular.

Thus we may suppose that the original vector $\xi = (\xi_1, \ldots, \xi_n)$ had linearly independent components and therefore that $|\mathbb{R}| > 0$.

Let $\mathcal{O}$ be an orthogonal matrix that diagonalizes $\mathbb{R}$,

$$\mathcal{O}^T\mathbb{R}\mathcal{O} = D.$$

The diagonal elements of D are positive and therefore determine the inverse matrix. Put $B^2 = D$ and

$$\beta = B^{-1}\mathcal{O}^T\xi.$$

Then it is easily verified that

$$\mathsf{E}e^{i(t,\beta)} = \mathsf{E}e^{i\beta^T t} = e^{-(1/2)(Et,t)},$$

i.e. the vector $\beta = (\beta_1, \ldots, \beta_n)$ is a Gaussian vector with components that are uncorrelated and therefore (Theorem 1) independent. Then if we write $A = \mathcal{O}B$ we find that the original Gaussian vector $\xi = (\xi_1, \ldots, \xi_n)$ can be represented as

$$\xi = A\beta, \tag{12}$$

where $\beta = (\beta_1, \ldots, \beta_n)$ is a Gaussian vector with independent components, $\beta_k \sim \mathcal{N}(0, 1)$. Hence we have the following result. Let $\xi = (\xi_1, \ldots, \xi_n)$ be a vector with linearly independent components such that $\mathsf{E}\xi_k = 0$, $k = 1, \ldots, n$. This vector is Gaussian if and only if there are independent Gaussian variables $\beta_1, \ldots, \beta_n$, $\beta_k \sim \mathcal{N}(0, 1)$, and a nonsingular matrix A of order n such that $\xi = A\beta$. Here $\mathbb{R} = AA^{\mathsf{T}}$ is the covariance matrix of ξ.

If $|\mathbb{R}| \neq 0$, then by the Gram–Schmidt method (see §11)

$$\xi_k = \hat{\xi}_k + b_k\varepsilon_k, \qquad k = 1, \ldots, n, \tag{13}$$

where since $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_k) \sim \mathcal{N}(0, E)$ is a Gaussian vector,

$$\hat{\xi}_k = \sum_{l=1}^{k-1} (\xi_k, \varepsilon_l)\varepsilon_l, \tag{14}$$

$$b_k = \|\xi_k - \hat{\xi}_k\| \tag{15}$$

and

$$\mathscr{L}\{\xi_1, \ldots, \xi_k\} = \mathscr{L}\{\varepsilon_1, \ldots, \varepsilon_k\}. \tag{16}$$

We see immediately from the orthogonal decomposition (13) that

$$\hat{\xi}_k = \mathsf{E}(\xi_k | \xi_{k-1}, \ldots, \xi_1). \tag{17}$$

From this, with (16) and (14), it follows that in the Gaussian case the conditional expectation $\mathsf{E}(\xi_k | \xi_{k-1}, \ldots, \xi_1)$ is a linear function of $(\xi_1, \ldots, \xi_{k-1})$:

$$\mathsf{E}(\xi_k | \xi_{k-1}, \ldots, \xi_1) = \sum_{i=1}^{k=1} a_i\xi_i. \tag{18}$$

(This was proved in §8 for the case $k = 2$.)

Since, according to a remark made in Theorem 1 of §8, $\mathsf{E}(\xi_k | \xi_{k-1}, \ldots, \xi_1)$ is an optimal estimator (in the mean-square sense) for ξ_k in terms of $\xi_1, \ldots, \xi_{k-1}$, it follows from (18) that in the Gaussian case the optimal estimator is *linear*.

We shall use these results in looking for optimal estimators of $\theta = (\theta_1, \ldots, \theta_k)$ in terms of $\xi = (\xi_1, \ldots, \xi_l)$ under the hypothesis that (θ, ξ) is Gaussian. Let

$$m_\theta = \mathsf{E}\theta, \qquad m_\xi = \mathsf{E}\xi$$

be the column-vector mean values and

$$\mathsf{V}_{\theta\theta} \equiv \operatorname{cov}(\theta, \theta) \equiv \|\operatorname{cov}(\theta_i, \theta_j)\|, \qquad 1 \le i, j \le k,$$

$$\mathsf{V}_{\theta\xi} \equiv \operatorname{cov}(\theta, \xi) \equiv \|\operatorname{cov}(\theta_i, \xi_j)\|, \qquad 1 \le i \le k,\ 1 \le j \le l,$$

$$\mathsf{V}_{\xi\xi} \equiv \operatorname{cov}(\xi, \xi) \equiv \|\operatorname{cov}(\xi_i, \xi_j)\|, \qquad 1 \le i, j \le l$$

the covariance matrices. Let us suppose that $\mathsf{V}_{\xi\xi}$ has an inverse. Then we have the following theorem.

Theorem 2 (Theorem on Normal Correlation). *For a Gaussian vector* (θ, ξ), *the optimal estimator* $\mathsf{E}(\theta | \xi)$ *of* θ *in terms of* ξ, *and its error matrix*

$$\Delta = \mathsf{E}[\theta - \mathsf{E}(\theta | \xi)][\theta - \mathsf{E}(\theta(\xi)]^{\mathsf{T}}$$

are given by the formulas

$$\mathsf{E}(\theta\,|\,\xi) = m_\theta + \mathsf{V}_{\theta\xi}\mathsf{V}_{\xi\xi}^{-1}(\xi - m_\xi), \tag{19}$$

$$\Delta = \mathsf{V}_{\theta\theta} - \mathsf{V}_{\theta\xi}\mathsf{V}_{\xi\xi}^{-1}(\mathsf{V}_{\theta\xi})^{\mathrm{T}}. \tag{20}$$

PROOF. Form the vector

$$\eta = (\theta - m_\theta) - \mathsf{V}_{\theta\xi}\mathsf{V}_{\xi\xi}^{-1}(\xi - m_\xi). \tag{21}$$

We can verify at once that $\mathsf{E}\eta(\xi - m_\xi)^{\mathrm{T}} = 0$, i.e. η is not correlated with $(\xi - m_\xi)$. But since (θ, ξ) is Gaussian, the vector (η, ξ) is also Gaussian. Hence by the remark on Theorem 1, η and $\xi - m_\xi$ are independent. Therefore η and ξ are independent, and consequently $\mathsf{E}(\eta\,|\,\xi) = \mathsf{E}\eta = 0$. Therefore

$$\mathsf{E}[\theta - m_\theta\,|\,\xi] - \mathsf{V}_{\theta\xi}\mathsf{V}_{\xi\xi}^{-1}(\xi - m_\xi) = 0.$$

which establishes (19).

To establish (20) we consider the conditional covariance

$$\mathrm{cov}(\theta, \theta\,|\,\xi) \equiv \mathsf{E}[(\theta - \mathsf{E}(\theta\,|\,\xi))(\theta - \mathsf{E}(\theta\,|\,\xi))^{\mathrm{T}}\,|\,\xi]. \tag{22}$$

Since $\theta - \mathsf{E}(\theta\,|\,\xi) = \eta$, and η and ξ are independent, we find that

$$\begin{aligned}
\mathrm{cov}(\theta, \theta\,|\,\xi) &= \mathsf{E}(\eta\eta^{\mathrm{T}}\,|\,\xi) = \mathsf{E}\eta\eta^{\mathrm{T}} \\
&= \mathsf{V}_{\theta\theta} + \mathsf{V}_{\theta\xi}^{-1}\mathsf{V}_{\xi\xi}\mathsf{V}_{\xi\xi}^{-1}\mathsf{V}_{\theta\xi}^{\mathrm{T}} - 2\mathsf{V}_{\theta\xi}\mathsf{V}_{\xi\xi}^{-1}\mathsf{V}_{\xi\xi}\mathsf{V}_{\xi\xi}^{-1}\mathsf{V}_{\theta\xi}^{\mathrm{T}} \\
&= \mathsf{V}_{\theta\theta} - \mathsf{V}_{\theta\xi}\mathsf{V}_{\xi\xi}^{-1}\mathsf{V}_{\theta\xi}^{\mathrm{T}}.
\end{aligned}$$

Since $\mathrm{cov}(\theta, \theta\,|\,\xi)$ does not depend on "chance," we have

$$\Delta = \mathsf{E}\,\mathrm{cov}(\theta, \theta\,|\,\xi) = \mathrm{cov}(\theta, \theta\,|\,\xi),$$

and this establishes (20).

Corollary. *Let $(\theta, \xi_1, \ldots, \xi_n)$ be an $(n + 1)$-dimensional Gaussian vector, with $\xi_1, \ldots, \xi_n$ independent. Then*

$$\mathsf{E}(\theta\,|\,\xi_1, \ldots, \xi_n) = \mathsf{E}\theta + \sum_{i=1}^{n} \frac{\mathrm{cov}(\theta, \xi_i)}{\mathsf{V}\xi_i}(\xi_i - \mathsf{E}\xi_i),$$

$$\Delta = \mathsf{V}\theta - \sum_{i=1}^{n} \frac{\mathrm{cov}^2(\theta, \xi_i)}{\mathsf{V}\xi_i}$$

(cf. (8.12) and (8.13)).

5. Let $\xi_1, \xi_2, \ldots$ be a sequence of Gaussian random vectors that converge in probability to ξ. Let us show that ξ is also Gaussian.

In accordance with (b) of Theorem 1, it is enough to establish this only for random variables.

Let $m_n = \mathsf{E}\xi_n$, $\sigma_n^2 = \mathsf{V}\xi_n$. Then by Lebesgue's dominated convergence theorem

$$\lim_{n\to\infty} e^{itm_n-(1/2)\sigma_n^2t^2} = \lim_{n\to\infty} \mathsf{E}e^{it\xi_n} = \mathsf{E}e^{it\xi}.$$

It follows from the existence of the limit on the left-hand side that there are numbers m and σ^2 such that

$$m = \lim_{n\to\infty} m_n, \qquad \sigma^2 = \lim_{n\to\infty} \sigma_n^2.$$

Consequently

$$\mathsf{E}e^{it\xi} = e^{itm-(1/2)\sigma^2t^2},$$

i.e. $\xi \sim \mathcal{N}(m, \sigma^2)$.

It follows, in particular, that the closed linear manifold $\overline{\mathscr{L}}(\xi_1, \xi_2, \ldots\}$ generated by the Gaussian variables $\xi_1, \xi_2, \ldots$ (see §11, Subsection 5) consists of Gaussian variables.

6. We now turn to the concept of Gaussian systems in general.

Definition 2. A collection of random variables $\xi = (\xi_\alpha)$, where α belongs to some index set $\mathfrak{A}$, is a *Gaussian system* if the random vector $(\xi_{\alpha_1}, \ldots, \xi_{\alpha_n})$ is Gaussian for every $n \geq 1$ and all indices $\alpha_1, \ldots, \alpha_n$ chosen from $\mathfrak{A}$.

Let us notice some properties of Gaussian systems.

(a) If $\xi = (\xi_\alpha)$, $\alpha \in \mathfrak{A}$, is a Gaussian system, then every subsystem $\xi' = (\xi'_{\alpha'})$, $\alpha' \in \mathfrak{A}' \subseteq \mathfrak{A}$, is also Gaussian.

(b) If ξ_α, $\alpha \in \mathfrak{A}$, are independent Gaussian variables, then the system $\xi = (\xi_\alpha)$, $\alpha \in \mathfrak{A}$, is Gaussian.

(c) If $\xi = (\xi_\alpha)$, $\alpha \in \mathfrak{A}$, is a Gaussian system, the closed linear manifold $\overline{\mathscr{L}}(\xi)$, consisting of all variables of the form $\sum_{i=1}^n c_{\alpha_i}\xi_{\alpha_i}$, together with their mean-square limits, forms a Gaussian system.

Let us observe that the converse of (a) is false in general. For example, let ξ_1 and η_1 be independent and $\xi_1 \sim \mathcal{N}(0, 1)$, $\eta_1 \sim \mathcal{N}(0, 1)$. Define the system

$$(\xi, \eta) = \begin{cases} (\xi_1, |\eta_1|) & \text{if } \xi_1 \geq 0, \\ (\xi_1, -|\eta_1|) & \text{if } \xi_1 < 0. \end{cases} \tag{23}$$

Then it is easily verified that ξ and η are both Gaussian, but (ξ, η) is not.

Let $\xi = (\xi_\alpha)_{\alpha\in\mathfrak{A}}$ be a Gaussian system with mean-value vector $m = (m_\alpha)$, $\alpha \in \mathfrak{A}$, and covariance matrix $\mathbb{R} = (r_{\alpha\beta})_{\alpha,\beta\in\mathfrak{A}}$, where $m_\alpha = \mathsf{E}\xi_\alpha$. Then $\mathbb{R}$ is evidently symmetric ($r_{\alpha\beta} = r_{\beta\alpha}$) and nonnegative definite in the sense that for every vector $c = (c_\alpha)_{\alpha\in\mathfrak{A}}$ with values in $R^{\mathfrak{A}}$, and only a finite number of nonzero coordinates c_α,

$$(\mathbb{R}c, c) \equiv \sum_{\alpha,\beta} r_{\alpha\beta}c_\alpha c_\beta \geq 0. \tag{24}$$

We now ask the converse question. Suppose that we are given a parameter set $\mathfrak{A} = \{\alpha\}$, a vector $m = (m_\alpha)_{\alpha \in \mathfrak{A}}$ and a symmetric nonnegative definite matrix $\mathbb{R} = (r_{\alpha\beta})_{\alpha, \beta \in \mathfrak{A}}$. Do there exist a probability space $(\Omega, \mathscr{F}, \mathsf{P})$ and a Gaussian system of random variables $\xi = (\xi_\alpha)_{\alpha \in \mathfrak{A}}$ on it, such that

$$\mathsf{E}\xi_\alpha = m_\alpha,$$
$$\operatorname{cov}(\xi_\alpha, \xi_\beta) = r_{\alpha,\beta}, \qquad \alpha, \beta \in \mathfrak{A}?$$

If we take a finite set $\alpha_1, \ldots, \alpha_n$, then for the vector $\overline{m} = (m_{\alpha_1}, \ldots, m_{\alpha_n})$ and the matrix $\overline{\mathbb{R}} = (r_{\alpha\beta})$, $\alpha, \beta = \alpha_1, \ldots, \alpha_n$, we can construct in R^n the Gaussian distribution $F_{\alpha_1, \ldots, \alpha_n}(x_1, \ldots, x_n)$ with characteristic function

$$\varphi(t) = \exp\{i(t, \overline{m}) - \tfrac{1}{2}(\overline{\mathbb{R}}t, t)\}, \qquad t = (t_{\alpha_1}, \ldots, t_{\alpha_n}).$$

It is easily verified that the family

$$\{F_{\alpha_1, \ldots, \alpha_n}(x_1, \ldots, x_n);\ \alpha_i \in \mathfrak{A}\}$$

is consistent. Consequently by Kolmogorov's theorem (Theorem 1, §9, and Remark 2 on this) the answer to our question is positive.

7. If $\mathfrak{A} = \{1, 2, \ldots\}$, then in accordance with the terminology of §5 the system of random variables $\xi = (\xi_\alpha)_{\alpha \in \mathfrak{A}}$ is a *random sequence* and is denoted by $\xi = (\xi_1, \xi_2, \ldots)$. A Gaussian sequence is completely described by its mean-value vector $m = (m_1, m_2, \ldots)$ and covariance matrix $\mathbb{R} = \|r_{ij}\|$, $r_{ij} = \operatorname{cov}(\xi_i, \xi_j)$. In particular, if $r_{ij} = \sigma_i^2 \delta_{ij}$, then $\xi = (\xi_1, \xi_2, \ldots)$ is a Gaussian sequence of independent random variables with $\xi_i \sim \mathscr{N}(m_i, \sigma_i^2)$, $i \geq 1$.

When $\mathfrak{A} = [0, 1], [0, \infty), (-\infty, \infty), \ldots$, the system $\xi = (\xi_t)$, $t \in \mathfrak{A}$, is a *random process with continuous time*.

Let us mention some examples of Gaussian random processes. If we take their mean values to be zero, their probabilistic properties are completely described by the covariance matrices $\|r_{st}\|$. We write $r(s, t)$ instead of r_{st} and call it the *covariance function*.

Example 1. If $T = [0, \infty)$ and

$$r(s, t) = \min(s, t), \tag{25}$$

the Gaussian process $\xi = (\xi_t)_{t \geq 0}$ with this covariance function (see Problem 2) and $\xi_0 \equiv 0$ is a *Brownian motion* or *Wiener process*.

Observe that this process has *independent increments*; that is, for arbitrary $t_1 < t_2 < \cdots < t_n$ the random variables

$$\xi_{t_2} - \xi_{t_1}, \ldots, \xi_{t_n} - \xi_{t_{n-1}}$$

are independent. In fact, because the process is Gaussian it is enough to verify only that the increments are uncorrelated. But if $s < t < u < v$ then

$$\begin{aligned}\mathsf{E}[\xi_t - \xi_s][\xi_v - \xi_u] &= [r(t, v) - r(t, u)] - [r(s, v) - r(s, u)] \\ &= (t - t) - (s - s) = 0.\end{aligned}$$

EXAMPLE 2. The process $\xi = (\xi_t)$, $0 \le t \le 1$, with $\xi_0 \equiv 0$ and

$$r(s, t) = \min(s, t) - st \tag{26}$$

is a *conditional Wiener process* (observe that since $r(1, 1) = 0$ we have $\mathsf{P}(\xi_1 = 0) = 1$).

EXAMPLE 3. The process $\xi = (\xi_t)$, $-\infty < t < \infty$, with

$$r(s, t) = e^{-|t-s|} \tag{27}$$

is a *Gauss–Markov process.*

8. PROBLEMS

1. Let ξ_1, ξ_2, ξ_3 be independent Gaussian random variables, $\xi_i \sim \mathcal{N}(0, 1)$. Show that

$$\frac{\xi_1 + \xi_2\xi_3}{\sqrt{1 + \xi_3^2}} \sim \mathcal{N}(0, 1).$$

(In this case we encounter the interesting problem of describing the nonlinear transformations of independent Gaussian variables $\xi_1, \dots, \xi_n$ whose distributions are still Gaussian.)

2. Show that (25), (26) and (27) are nonnegative definite (and consequently are actually covariance functions).

3. Let A be an $m \times n$ matrix. An $n \times m$ matrix $A^{\oplus}$ is a *pseudoinverse* of A if there are matrices U and V such that

$$AA^{\oplus}A = A, \qquad A^{\oplus} = UA^{\mathsf{T}} = A^{\mathsf{T}}V.$$

Show that $A^{\oplus}$ exists and is unique.

4. Show that (19) and (20) in the theorem on normal correlation remains valid when $\mathsf{V}_{\xi\xi}$ is singular provided that $\mathsf{V}_{\xi\xi}^{-1}$ is replaced by $\mathsf{V}_{\xi\xi}^{\oplus}$.

5. Let $(\theta, \xi) = (\theta_1, \dots, \theta_k; \xi_1, \dots, \xi_l)$ be a Gaussian vector with nonsingular matrix $\Delta \equiv \mathsf{V}_{\theta\theta} - \mathsf{V}_{\xi\xi}^{\oplus}\mathsf{V}_{\theta\xi}^{*}$. Show that the distribution function

$$\mathsf{P}(\theta \le a | \xi) = \mathsf{P}(\theta_1 \le a_1, \dots, \theta_k \le a_k | \xi)$$

has (P-a.s.) the density $p(a_1, \dots, a_k | \xi)$ defined by

$$\frac{|\Delta^{-1/2}|}{(2\pi)^{k/2}} \exp\{-\tfrac{1}{2}(a - \mathsf{E}(\theta | \xi))^{\mathsf{T}}\Delta^{-1}(a - \mathsf{E}(\theta | \xi))\}.$$

6. (S. N. Bernstein). Let ξ and η be independent identically distributed random variables with finite variances. Show that if $\xi + \eta$ and $\xi - \eta$ are independent, then ξ and η are Gaussian.

CHAPTER III
Convergence of Probability Measures. Central Limit Theorem

§1. Weak Convergence of Probability Measures and Distributions

1. Many of the fundamental results in probability theory are formulated as *limit theorems*. Bernoulli's law of large numbers was formulated as a limit theorem; so was the De Moivre–Laplace theorem, which can fairly be called the origin of a genuine theory of probability and, in particular, which led the way to numerous investigations that clarified the conditions for the validity of the central limit theorem. Poisson's theorem on the approximation of the binomial distribution by the "Poisson" distribution in the case of rare events was formulated as a limit theorem. After the example of these propositions, and of results on the rapidity of convergence in the De Moivre–Laplace and Poisson theorems, it became clear that in probability it is necessary to deal with various kinds of convergence of distributions, and to establish the rapidity of convergence connected with the introduction of various "natural" measures of the distance between distributions. In the present chapter we shall discuss some general features of the convergence of probability distributions and of the distance between them. In this section we take up questions in the general theory of weak convergence of probability measures in metric spaces. The De Moivre–Laplace theorem, the progenitor of the central limit theorem, finds a natural place in this theory. From §3, it is clear that the method of characteristic functions is one of the most powerful means for proving limit theorems on the weak convergence of probability distributions in R^n. In §6, we consider questions of metrizability of weak convergence. Then, in §8, we turn our attention to a different kind of convergence of distributions (stronger than weak convergence), namely convergence in variation. Proofs of the simplest results on the rapidity of convergence in the central limit theorem and Poisson's theorem will be given in §§10 and 11.

2. We begin by recalling the statement of the law of large numbers (Chapter I, §5) for the Bernoulli scheme.

Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables with $\mathsf{P}(\xi_i = 1) = p$, $\mathsf{P}(\xi_i = 0) = q$, $p + q = 1$. In terms of the concept of convergence in probability (Chapter II, §10), Bernoulli's law of large numbers can be stated as follows:

$$\frac{S_n}{n} \xrightarrow{\mathsf{P}} p, \qquad n \to \infty, \tag{1}$$

where $S_n = \xi_1 + \cdots + \xi_n$. (It will be shown in Chapter IV that in fact we have convergence with probability 1.)

We put

$$\begin{aligned} F_n(x) &= \mathsf{P}\left\{\frac{S_n}{n} \leq x\right\}, \\ F(x) &= \begin{cases} 1, & x \geq p, \\ 0, & x < p, \end{cases} \end{aligned} \tag{2}$$

where $F(x)$ is the distribution function of the degenerate random variable $\xi \equiv p$. Also let P_n and P be the probability measures on $(R, \mathscr{B}(R))$ corresponding to the distributions F_n and F.

In accordance with Theorem 2 of §10, Chapter II, convergence in probability, $S_n/n \xrightarrow{\mathsf{P}} p$, implies convergence in distribution, $S_n/n \xrightarrow{d} p$, which means that

$$\mathsf{E}f\left(\frac{S_n}{n}\right) \to \mathsf{E}f(p), \qquad n \to \infty, \tag{3}$$

for every function $f = f(x)$ belonging to the class $\mathbb{C}(R)$ of bounded continuous functions on R.

Since

$$\mathsf{E}f\left(\frac{S_n}{n}\right) = \int_R f(x)\mathsf{P}_n(dx), \qquad \mathsf{E}f(p) = \int_R f(x)\mathsf{P}(dx),$$

(3) can be written in the form

$$\int_R f(x)\mathsf{P}_n(dx) \to \int_R f(x)\mathsf{P}(dx), \qquad f \in \mathbb{C}(R), \tag{4}$$

or (in accordance with §6 of Chapter II) in the form

$$\int_R f(x)\, dF_n(x) \to \int_R f(x)\, dF(x), \qquad f \in \mathbb{C}(R). \tag{5}$$

In analysis, (4) is called *weak convergence* (of P_n to P, $n \to \infty$) and written $\mathsf{P}_n \xrightarrow{w} \mathsf{P}$. It is also natural to call (5) weak convergence of F_n to F and denote it by $F_n \xrightarrow{w} F$.

Thus we may say that in a Bernoulli scheme

$$\frac{S_n}{n} \xrightarrow{\text{P}} p \Rightarrow F_n \xrightarrow{w} F. \tag{6}$$

It is also easy to see from (1) that, for the distribution functions defined in (2),

$$F_n(x) \to F(x), \qquad n \to \infty,$$

for all points $x \in R$ *except for the single point* $x = p$, where $F(x)$ has a discontinuity.

This shows that weak convergence $F_n \to F$ *does not imply* pointwise convergence of $F_n(x)$ to $F(x)$, $n \to \infty$, for *all* points $x \in R$. However, it turns out that, both for Bernoulli schemes and for arbitrary distribution functions, weak convergence is equivalent (see Theorem 2 below) to "convergence in general" in the sense of the following definition.

Definition 1. A sequence of distribution functions $\{F_n\}$, defined on the real line, converges *in general* to the distribution function F (notation: $F_n \Rightarrow F$) if as $n \to \infty$

$$F_n(x) \to F(x), \qquad x \in P_C(F),$$

where $P_C(F)$ is the set of points of continuity of $F = F(x)$.

For Bernoulli schemes, $F = F(x)$ is degenerate, and it is easy to see (see Problem 7 of §10, Chapter II) that

$$(F_n \Rightarrow F) \Rightarrow \left(\frac{S_n}{n} \xrightarrow{\text{P}} p\right).$$

Therefore, taking account of Theorem 2 below,

$$\left(\frac{S_n}{n} \xrightarrow{\text{P}} p\right) \Rightarrow (F_n \xrightarrow{w} F) \Leftrightarrow (F_n \Rightarrow F) \Rightarrow \left(\frac{S_n}{n} \xrightarrow{\text{P}} p\right) \tag{7}$$

and consequently the law of large numbers can be considered as a theorem on the weak convergence of the distribution functions defined in (2).

Let us write

$$F_n(x) = \text{P}\left\{\frac{S_n - np}{\sqrt{npq}} \le x\right\},$$

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du. \tag{8}$$

The De Moivre–Laplace theorem (§6, Chapter I) states that $F_n(x) \to F(x)$ for all $x \in R$, and consequently $F_n \Rightarrow F$. Since, as we have observed, weak convergence $F_n \xrightarrow{w} F$ and convergence in general, $F_n \Rightarrow F$, are equivalent, we may therefore say that the De Moivre–Laplace theorem is also a theorem on the weak convergence of the distribution functions defined by (8).

These examples justify the concept of weak convergence of probability measures that will be introduced below in Definition 2. Although, on the real line, weak convergence is equivalent to convergence in general of the corresponding distribution functions, it is preferable to use weak convergence from the beginning. This is because in the first place it is easier to work with, and in the second place it remains useful in more general spaces than the real line, and in particular for metric spaces, including the especially important spaces R^n, R^∞, C, and D (see §3 of Chapter II).

3. Let $(E, \mathscr{E}, \rho)$ be a metric space with metric $\rho = \rho(x, y)$ and σ-algebra $\mathscr{E}$ of Borel subsets generated by the open sets, and let $\mathsf{P}, \mathsf{P}_1, \mathsf{P}_2, \ldots$ be probability measures on $(E, \mathscr{E}, \rho)$.

Definition 2. A sequence of probability measures $\{\mathsf{P}_n\}$ *converges weakly* to the probability measure P (notation: $\mathsf{P}_n \xrightarrow{w} \mathsf{P}$) if

$$\int_E f(x)\mathsf{P}_n(dx) \to \int_E f(x)\mathsf{P}(dx) \tag{9}$$

for every function $f = f(x)$ in the class $\mathbb{C}(E)$ of continuous bounded functions on E.

Definition 3. A sequence of probability measures $\{\mathsf{P}_n\}$ *converges in general* to the probability measure P (notation: $\mathsf{P}_n \Rightarrow \mathsf{P}$) if

$$\mathsf{P}_n(A) \to \mathsf{P}(A) \tag{10}$$

for every set A of $\mathscr{E}$ for which

$$\mathsf{P}(\partial A) = 0. \tag{11}$$

(Here ∂A denotes the boundary of A:

$$\partial A = [A] \cap [\bar{A}],$$

where $[A]$ is the closure of A.)

The following fundamental theorem shows the equivalence of the concepts of weak convergence and convergence in general for probability measures, and contains still another equivalent statement.

Theorem 1. *The following statements are equivalent.*

(I) $\mathsf{P}_n \xrightarrow{w} \mathsf{P}$.
(II) $\overline{\lim}\, \mathsf{P}_n(A) \le \mathsf{P}(A)$, *$A$ closed.*
(III) $\underline{\lim}\, \mathsf{P}_n(A) \ge \mathsf{P}(A)$, *$A$ open.*
(IV) $\mathsf{P}_n \Rightarrow \mathsf{P}$.

PROOF. (I) $\Rightarrow$ (II). Let A be closed, $f(x) = I_A(x)$ and

$$f_\varepsilon(x) = g\left(\frac{1}{\varepsilon}\rho(x, A)\right), \qquad \varepsilon > 0,$$

where

$$\rho(x, A) = \inf\{\rho(x, y): y \in A\},$$

$$g(t) = \begin{cases} 1, & t \le 0, \\ 1 - t, & 0 \le t \le 1, \\ 0, & t \ge 1. \end{cases}$$

Let us also put

$$A_\varepsilon = \{x: \rho(x, A) < \varepsilon\}$$

and observe that $A_\varepsilon \downarrow A$ as $\varepsilon \downarrow 0$.

Since $f_\varepsilon(x)$ is bounded, continuous, and satisfies

$$\mathsf{P}_n(A) = \int_E I_A(x)\mathsf{P}_n(dx) \le \int_E f_\varepsilon(x)\mathsf{P}_n(dx),$$

we have

$$\overline{\lim_n}\, \mathsf{P}_n(A) \le \overline{\lim_n} \int_E f_\varepsilon(x)\mathsf{P}_n(dx) = \int_E f_\varepsilon(x)\mathsf{P}(dx) \le \mathsf{P}(A_\varepsilon) \downarrow \mathsf{P}(A),\ \varepsilon \downarrow 0,$$

which establishes the required implication.

The implications (II) $\Rightarrow$ (III) and (III) $\Rightarrow$ (II) become obvious if we take the complements of the sets concerned.

(III) $\Rightarrow$ (IV). Let $A^0 = A \backslash \partial A$ be the interior, and $[A]$ the closure, of A. Then from (II), (III), and the hypothesis $\mathsf{P}(\partial A) = 0$, we have

$$\overline{\lim_n}\, \mathsf{P}_n(A) \le \overline{\lim_n}\, \mathsf{P}_n([A]) \le \mathsf{P}([A]) = \mathsf{P}(A),$$

$$\underline{\lim_n}\, \mathsf{P}_n(A) \ge \underline{\lim_n}\, \mathsf{P}_n(A^0) \ge \mathsf{P}(A^0) = \mathsf{P}(A),$$

and therefore $\mathsf{P}_n(A) \to \mathsf{P}(A)$ for every A such that $\mathsf{P}(\partial A) = 0$.

(IV) $\to$ (I). Let $f = f(x)$ be a bounded continuous function with $|f(x)| \le M$. We put

$$D = \{t \in R: \mathsf{P}\{x: f(x) = t\} \ne 0\}$$

and consider a decomposition $T_k = (t_0, t_1, \ldots, t_k)$ of $[-M, M]$:

$$-M = t_0 < t_1 < \cdots < t_k = M, \quad k \ge 1,$$

with $t_i \notin D$, $i = 0, 1, \ldots, k$. (Observe that D is at most countable since the sets $f^{-1}\{t\}$ are disjoint and P is finite.)

Let $B_i = \{x: t_i \le f(x) < t_{i+1}\}$. Since $f(x)$ is continuous and therefore the set $f^{-1}(t_i, t_{i+1})$ is open, we have $\partial B_i \subseteq f^{-1}\{t_i\} \cup f^{-1}\{t_{i+1}\}$. The points $t_i, t_{i+1} \notin D$; therefore $\mathsf{P}(\partial B_i) = 0$ and, by (IV),

$$\sum_{i=0}^{k-1} t_i \mathsf{P}_n(B_i) \to \sum_{i=0}^{k-1} t_i \mathsf{P}(B_i). \tag{12}$$

But

$$\begin{aligned}\left|\int_E f(x)\mathsf{P}_n(dx) - \int_E f(x)\mathsf{P}(dx)\right| &\le \left|\int_E f(x)\mathsf{P}_n(dx) - \sum_{i=0}^{k-1} t_i \mathsf{P}_n(B_i)\right| \\ &\quad + \left|\sum_{i=0}^{k-1} t_i \mathsf{P}_n(B_i) - \sum_{i=0}^{k-1} t_i \mathsf{P}(B_i)\right| \\ &\quad + \left|\sum_{i=0}^{k-1} t_i \mathsf{P}(B_i) - \int_E f(x)\mathsf{P}(dx)\right| \\ &\le 2 \max_{0\le i\le k-1} (t_{i+1} - t_i) \\ &\quad + \left|\sum_{i=0}^{k-1} t_i \mathsf{P}_n(B_i) - \sum_{i=0}^{k-1} t_i \mathsf{P}(B_i)\right|,\end{aligned}$$

whence, by (12), since the T_k ($k \ge 1$) are arbitrary,

$$\lim_n \int_E f(x)\mathsf{P}_n(dx) = \int_E f(x)\mathsf{P}(dx).$$

This completes the proof of the theorem.

Remark 1. The functions $f(x) = I_A(x)$ and $f_\varepsilon(x)$ that appear in the proof that (I) $\Rightarrow$ (II) are respectively *upper semicontinuous* and *uniformly continuous*. Hence it is easy to show that each of the conditions of the theorem is equivalent to one of the following:

(V) $\int_E f(x)\mathsf{P}_n(x)\,dx \to \int_E f(x)\mathsf{P}(dx)$ *for all bounded uniformly continuous* $f(x)$;

(VI) $\overline{\lim} \int_E f(x)\mathsf{P}_n(dx) \le \int_E f(x)\mathsf{P}(dx)$ *for all bounded* $f(x)$ *that are upper semicontinuous* ($\overline{\lim} f(x_n) \le f(x)$, $x_n \to x$);

(VII) $\underline{\lim} \int_E f(x)\mathsf{P}_n(dx) \ge \int_E f(x)\mathsf{P}(dx)$ *for all bounded* $f(x)$ *that are lower semicontinuous* ($\underline{\lim}_n f(x_n) \ge f(x)$, $x_n \to x$).

Remark 2. Theorem 1 admits a natural generalization to the case when the probability measures P and P_n defined on $(E, \mathscr{E}, \rho)$ are replaced by *arbitrary* (not necessarily probability) *finite measures* μ and μ_n. For such measures we can introduce weak convergence $\mu_n \xrightarrow{w} \mu$ and convergence in general $\mu_n \Rightarrow \mu$ and, just as in Theorem 1, we can establish the equivalence of the following conditions:

(I*) $\mu_n \xrightarrow{w} \mu$;

(II*) $\overline{\lim}\, \mu_n(A) \le \mu(A)$, *where A is closed and* $\mu_n(E) \to \mu(E)$;

(III*) $\underline{\lim}\, \mu_n(A) \ge \mu(A)$, *where A is open and* $\mu_n(E) \to \mu(E)$;

(IV*) $\mu_n \Rightarrow \mu$.

Each of these is equivalent to any of (V*), (VI*), and (VII*), which are (V), (VI), and (VII) with P_n and P replaced by μ_n and μ.

4. Let $(R, \mathscr{B}(R))$ be the real line with the system $\mathscr{B}(R)$ of sets generated by the Euclidean metric $\rho(x, y) = |x - y|$ (compare Remark 2 of subsection 2 of §2 of Chapter II). Let P and $\mathsf{P}_n, n \geq 1$, be probability measures on $(R, \mathscr{B}(R))$ and let F and F_n, $n \geq 1$, be the corresponding distribution functions.

Theorem 2. *The following conditions are equivalent:*

(1) $\mathsf{P}_n \xrightarrow{w} \mathsf{P}$,
(2) $\mathsf{P}_n \Rightarrow \mathsf{P}$,
(3) $F_n \xrightarrow{w} F$,
(4) $F_n \Rightarrow F$.

PROOF. Since (2) $\Leftrightarrow$ (1) $\Leftrightarrow$ (3), it is enough to show that (2) $\Leftrightarrow$ (4).

If $\mathsf{P}_n \Rightarrow \mathsf{P}$, then in particular

$$\mathsf{P}_n(-\infty, x] \to \mathsf{P}(-\infty, x]$$

for all $x \in R$ such that $\mathsf{P}\{x\} = 0$. But this means that $F_n \Rightarrow F$.

Now let $F_n \Rightarrow F$. To prove that $\mathsf{P}_n \Rightarrow \mathsf{P}$ it is enough (by Theorem 1) to show that $\underline{\lim}_n \mathsf{P}_n(A) \geq \mathsf{P}(A)$ for every open set A.

If A is open, there is a countable collection of disjoint open intervals $I_1, I_2, \ldots$ (of the form (a, b)) such that $A = \sum_{k=1}^{\infty} I_k$. Choose $\varepsilon > 0$ and in each interval $I_k = (a_k, b_k)$ select a subinterval $I'_k = (a'_k, b'_k]$ such that a'_k, $b'_k \in P_C(F)$ and $P(I_k) \leq P(I'_k) + \varepsilon \cdot 2^{-k}$. (Since $F(x)$ has at most countably many discontinuities, such intervals I'_k, $k \geq 1$, certainly exist.) By Fatou's lemma,

$$\begin{aligned}\underline{\lim}_n \mathsf{P}_n(A) &= \underline{\lim}_n \sum_{k=1}^{\infty} \mathsf{P}_n(I_k) \geq \sum_{k=1}^{\infty} \underline{\lim}_n \mathsf{P}_n(I_k) \\ &\geq \sum_{k=1}^{\infty} \underline{\lim}_n \mathsf{P}_n(I'_k).\end{aligned}$$

But

$$\mathsf{P}_n(I'_k) = F_n(b'_k) - F_n(a'_k) \to F(b'_k) - F(a'_k) = \mathsf{P}(I'_k).$$

Therefore

$$\underline{\lim}_n \mathsf{P}_n(A) \geq \sum_{k=1}^{\infty} \mathsf{P}(I'_k) \geq \sum_{k=1}^{\infty} (\mathsf{P}(I_k) - \varepsilon \cdot 2^{-k}) = \mathsf{P}(A) - \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this shows that $\underline{\lim}_n P_n(A) \geq P(A)$ if A is open.

This completes the proof of the theorem.

5. Let $(E, \mathscr{E})$ be a measurable space. A collection $\mathscr{K}_0(E) \subseteq \mathscr{E}$ of subsets is a *determining class* whenever two probability measures P and Q on $(E, \mathscr{E})$

satisfy

$$\mathsf{P}(A) = \mathsf{Q}(A) \qquad \textit{for all} \quad A \in \mathscr{K}_0(E)$$

it follows that the measures are identical, i.e.,

$$\mathsf{P}(A) = \mathsf{Q}(A) \qquad \textit{for all} \quad A \in \mathscr{E}.$$

If $(E, \mathscr{E}, \rho)$ is a metric space, a collection $\mathscr{K}_1(E) \subseteq \mathscr{E}$ is a *convergence-determining class* whenever probability measures $\mathsf{P}, \mathsf{P}_1, \mathsf{P}_2, \ldots$ satisfy

$$\mathsf{P}_n(A) \to \mathsf{P}(A) \qquad \textit{for} \quad \textit{all} \quad A \in \mathscr{K}_1(E) \quad \text{with} \quad \mathsf{P}(\partial A) = 0$$

it follows that

$$\mathsf{P}_n(A) \to \mathsf{P}(A) \qquad \text{for all} \quad A \in E \quad \text{with} \quad \mathsf{P}(\partial A) = 0.$$

When $(E, \mathscr{E}) = (R, \mathscr{B}(R))$, we can take a determining class $\mathscr{K}_0(R)$ to be the class of "elementary" sets $\mathscr{K} = \{(-\infty, x], x \in R\}$ (Theorem 1, §3, Chapter II). It follows from the equivalence of (2) and (4) of Theorem 2 that this class $\mathscr{K}$ is also a convergence-determining class.

It is natural to ask about such determining classes in more general spaces.

For R^n, $n \geq 2$, the class $\mathscr{K}$ of "elementary" sets of the form $(-\infty, x] = (-\infty, x_1] \times \cdots \times (-\infty, x_n]$, where $x = (x_1, \ldots, x_n) \in R^n$, is both a determining class (Theorem 2, §3, Chapter II) and a convergence-determining class (Problem 2).

For R^∞ the cylindrical sets $\mathscr{K}_0(R^\infty)$ are the "elementary" sets whose probabilities uniquely determine the probabilities of the Borel sets (Theorem 3, §3, Chapter II). It turns out that in this case the class of cylindrical sets is also the class of convergence-determining sets (Problem 3). Therefore $\mathscr{K}_1(R^\infty) = \mathscr{K}_0(R^\infty)$.

We might expect that the cylindrical sets would still constitute determining classes in more general spaces. However, this is, in general, not the case.

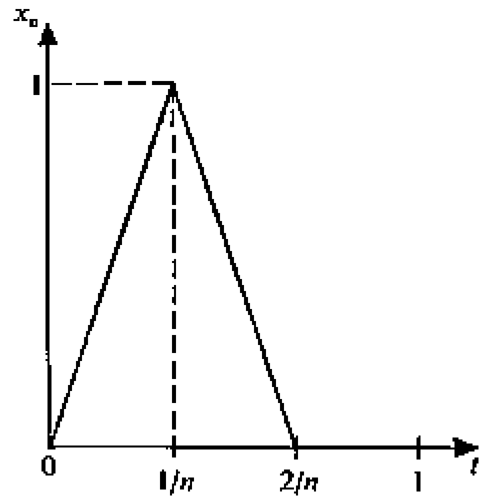


Figure 35

For example, consider the space $(C, \mathscr{B}_0(C), \rho)$ with the uniform metric ρ (see subsection 6, §2, Chapter II). Let P be the probability measure concentrated on the element $x = x(t) \equiv 0$, $0 \le t \le 1$, and let P_n, $n \ge 1$, be the probability measures each of which is concentrated on the element $x = x_n(t)$ shown in Figure 35. It is easy to see that $\mathsf{P}_n(A) \to \mathsf{P}(A)$ for all cylindrical sets A with $\mathsf{P}(\partial A) = 0$. But if we consider, for example, the set

$$A = \{\alpha \in C: |\alpha(t)| \le \tfrac{1}{2}, 0 \le t \le 1\} \in \mathscr{B}_0(C),$$

then $\mathsf{P}(\partial A) = 0$, $\mathsf{P}_n(A) = 0$, $\mathsf{P}(A) = 1$ and consequently $\mathsf{P}_n \not\Rightarrow \mathsf{P}$.

Therefore $\mathscr{K}_0(C) = \mathscr{B}_0(C)$ but $\mathscr{K}_0(C) \subset \mathscr{K}_1(C)$ (with strict inclusion).

6. Problems

1. Let us say that a function $F = F(x)$, defined on R^n, is *continuous at* $x \in R^n$ provided that, for every $\varepsilon > 0$, there is a $\delta > 0$ such that $|F(x) - F(y)| < \varepsilon$ for all $y \in R^n$ that satisfy

$$x - \delta e < y < x + \delta e,$$

where $e = (1, \ldots, 1) \in R^n$. Let us say that a sequence of distribution functions $\{F_n\}$ *converges in general* to the distribution function F ($F_n \Rightarrow F$) if $F_n(x) \to F(x)$, for all points $x \in R^n$ where $F = F(x)$ is continuous.

Show that the conclusion of Theorem 2 remains valid for R^n, $n > 1$. (See the remark on Theorem 2.)

2. Show that the class $\mathscr{K}$ of "elementary" sets in R^n is a convergence-determining class.

3. Let E be one of the spaces R^∞, C, or D. Let us say that a sequence $\{\mathsf{P}_n\}$ of problem measures (defined on the σ-algebra $\mathscr{E}$ of Borel sets generated by the open sets) *converges in general in the sense of finite-dimensional distributions* to the probability measure P (notation: $\mathsf{P}_n \xrightarrow{f} \mathsf{P}$) if $\mathsf{P}_n(A) \to \mathsf{P}(A)$, $n \to \infty$, for all *cylindrical* sets A with $\mathsf{P}(\partial A) = 0$.

For R^∞, show that

$$(\mathsf{P}_n \xrightarrow{f} \mathsf{P}) \Leftrightarrow (\mathsf{P}_n \Rightarrow \mathsf{P}).$$

4. Let F and G be distribution functions on the real line and let

$$L(F, G) = \inf\{h > 0: F(x - h) - h \le G(x) \le F(x + h) + h\}$$

be the *Lévy distance* (between F and G). Show that convergence in general is equivalent to convergence in the Lévy metric:

$$(F_n \Rightarrow F) \Leftrightarrow L(F_n, F) \to 0.$$

5. Let $F_n \Rightarrow F$ and let F be continuous. Show that in this case $F_n(x)$ converges uniformly to $F(x)$:

$$\sup_x |F_n(x) - F(x)| \to 0, \qquad n \to \infty.$$

6. Prove the statement in Remark 1 on Theorem 1.

7. Establish the equivalence of (I*)–(IV*) as stated in Remark 2 on Theorem 1.

8. Show that $\mathsf{P}_n \xrightarrow{w} \mathsf{P}$ if and only if every subsequence $\{\mathsf{P}_{n'}\}$ of $\{\mathsf{P}_n\}$ contains a subsequence $\{\mathsf{P}_{n''}\}$ such that $\mathsf{P}_{n''} \xrightarrow{w} \mathsf{P}$.

§2. Relative Compactness and Tightness of Families of Probability Distributions

1. If we are given a sequence of probability measures, then before we can consider the question of its (weak) convergence to some probability measure, we have of course to establish whether the sequence converges in general to some measure, or has at least one convergent subsequence.

For example, the sequence $\{\mathsf{P}_n\}$, where $\mathsf{P}_{2n} = \mathsf{P}$, $\mathsf{P}_{2n+1} = \mathsf{Q}$, and P and Q are different probability measures, is evidently not convergent, but has the two convergent subsequences $\{\mathsf{P}_{2n}\}$ and $\{\mathsf{P}_{2n+1}\}$.

It is easy to construct a sequence $\{\mathsf{P}_n\}$ of probability measures P_n, $n \geq 1$, that not only fails to converge, but contains no convergent subsequences at all. All that we have to do is to take P_n, $n \geq 1$, to be concentrated at $\{n\}$ (that is, $\mathsf{P}_n\{n\} = 1$). In fact, since $\lim_n \mathsf{P}_n(a, b] = 0$ whenever $a < b$, a limit measure would have to be identically zero, contradicting the fact that $1 = \mathsf{P}_n(R) \not\to 0$, $n \to \infty$. It is interesting to observe that in this example the corresponding sequence $\{F_n\}$ of distribution functions,

$$F_n(x) = \begin{cases} 1, & x \geq n, \\ 0, & x < n, \end{cases}$$

is evidently convergent: for every $x \in R$,

$$F_n(x) \to G(x) \equiv 0.$$

However, the limit function $G = G(x)$ is not a distribution function (in the sense of Definition 1 of §3, Chapter II).

This instructive example shows that the space of distribution functions is not compact. It also shows that if a sequence of distribution functions is to converge to a limit that is also a distribution function, we must have some hypothesis that will prevent mass from "escaping to infinity."

After these introductory remarks, which illustrate the kinds of difficulty that can arise, we turn to the basic definitions.

2. Let us suppose that all measures are defined on the metric space $(E, \mathscr{E}, \rho)$.

Definition 1. A family of probability measures $\mathscr{P} = \{\mathsf{P}_\alpha;\, \alpha \in \mathfrak{A}\}$ is *relatively compact* if every sequence of measures from $\mathscr{P}$ contains a subsequence which converges weakly to a probability measure.

We emphasize that in this definition the limit measure is to be a *probability measure*, although it need not belong to the original class $\mathscr{P}$. (This is why the word "relatively" appears in the definition.)

It is often far from simple to verify that a given family of probability measures is relatively compact. Consequently it is desirable to have simple and useable tests for this property. We need the following definitions.

Definition 2. A family of probability measures $\mathscr{P} = \{\mathsf{P}_\alpha; \alpha \in \mathfrak{A}\}$ is *tight* if, for every $\varepsilon > 0$, there is a compact set $K \subseteq E$ such that

$$\sup_{\alpha \in \mathfrak{A}} \mathsf{P}_\alpha(E \backslash K) \leq \varepsilon. \tag{1}$$

Definition 3. A family of distribution functions $F = \{F_\alpha; \alpha \in \mathfrak{A}\}$ defined on R^n, $n \geq 1$, is *relatively compact* (or *tight*) if the same property is possessed by the family $\mathscr{P} = \{\mathsf{P}_\alpha; \alpha \in \mathfrak{A}\}$ of probability measures, where P_α is the measure constructed from F_α.

3. The following result is fundamental for the study of weak convergence of probability measures.

Theorem 1 (Prokhorov's Theorem). *Let $\mathscr{P} = \{\mathsf{P}_\alpha; \alpha \in \mathfrak{A}\}$ be a family of probability measures defined on a complete separable metric space $(E, \mathscr{E}, \rho)$. Then $\mathscr{P}$ is relatively compact if and only if it is tight.*

We shall give the proof only when the space is the real line. (The proof can be carried over, almost unchanged, to arbitrary Euclidean spaces R^n, $n \geq 2$. Then the theorem can be extended successively to R^∞, to σ-compact spaces; and finally to general complete separable metric spaces, by reducing each case to the preceding one.)

Necessity. Let the family $\mathscr{P} = \{\mathsf{P}_\alpha: \alpha \in \mathfrak{A}\}$ of probability measures defined on $(R, \mathscr{B}(R))$ be relatively compact but not tight. Then there is an $\varepsilon > 0$ such that for every compact $K \subseteq R$

$$\sup_\alpha \mathsf{P}_\alpha(R \backslash K) > \varepsilon,$$

and therefore, for each interval $I = (a, b)$,

$$\sup_\alpha \mathsf{P}_\alpha(R \backslash I) > \varepsilon.$$

It follows that for every interval $I_n = (-n, n)$, $n \geq 1$, there is a measure P_{α_n} such that

$$\mathsf{P}_{\alpha_n}(R \backslash I_n) > \varepsilon.$$

Since the original family $\mathscr{P}$ is relatively compact, we can select from $\{\mathsf{P}_{\alpha_n}\}_{n \geq 1}$ a subsequence $\{\mathsf{P}_{\alpha_{n_k}}\}$ such that $\mathsf{P}_{\alpha_{n_k}} \xrightarrow{w} \mathsf{Q}$, where Q is a probability measure. Then, by the equivalence of conditions (I) and (II) in Theorem 1 of §1, we have

$$\overline{\lim_{k \to \infty}} \mathsf{P}_{\alpha_{n_k}}(R \backslash I_n) \leq \mathsf{Q}(R \backslash I_n) \tag{2}$$

for every $n \geq 1$. But $Q(R \setminus I_n) \downarrow 0$, $n \to \infty$, and the left side of (2) exceeds $\varepsilon > 0$. This contradiction shows that relatively compact sets are tight.

To prove the sufficiency we need a general result (Helly's theorem) on the *sequential compactness* of families of generalized distribution functions (Subsection 2 of §3 of Chapter II).

Let $\mathscr{J} = \{G\}$ be the collection of generalized distribution functions $G = G(x)$ that satisfy:

(1) $G(x)$ *is nondecreasing*;
(2) $0 \leq G(-\infty), G(+\infty) \leq 1$;
(3) $G(x)$ *is continuous on the right.*

Then $\mathscr{J}$ clearly contains the class of distribution functions $\mathscr{F} = \{F\}$ for which $F(-\infty) = 0$ and $F(+\infty) = 1$.

Theorem 2 (Helly's Theorem). *The class $\mathscr{J} = \{G\}$ of generalized distribution functions is sequentially compact, i.e., for every sequence $\{G_n\}$ of functions from $\mathscr{J}$ we can find a function $G \in \mathscr{J}$ and a sequence $\{n_k\} \subseteq \{n\}$ such that*

$$G_{n_k}(x) \to G(x), \qquad k \to \infty,$$

for every point x of the set $P_C(G)$ of points of continuity of $G = G(x)$.

Proof. Let $T = \{x_1, x_2, \ldots\}$ be a countable dense subset of R. Since the sequence of numbers $\{G_n(x_1)\}$ is *bounded*, there is a subsequence $N_1 = \{n_1^{(1)}, n_2^{(1)}, \ldots\}$ such that $G_{n_i^{(1)}}(x_1)$ approaches a limit g_1 as $i \to \infty$. Then we extract from N_1 a subsequence $N_2 = \{n_1^{(2)}, n_2^{(2)}, \ldots\}$ such that $G_{n_i^{(2)}}(x_2)$ approaches a limit g_2 as $i \to \infty$; and so on.

On the set $T \subseteq R$ we can define a function $G_T(x)$ by

$$G_T(x_i) = g_i, \qquad x_i \in T,$$

and consider the "Cantor" diagonal sequence $N = \{n_1^{(1)}, n_2^{(2)}, \ldots\}$. Then, for each $x_i \in T$, as $m \to \infty$, we have

$$G_{n_m^{(m)}}(x_i) \to G_T(x_i).$$

Finally, let us define $G = G(x)$ for all $x \in R$ by putting

$$G(x) = \inf\{G_T(y)\colon y \in T, y > x\}. \tag{3}$$

We claim that $G = G(x)$ is the required function and $G_{n_m^{(m)}}(x) \to G(x)$ at all points x of continuity of G.

Since all the functions G_n under consideration are nondecreasing, we have $G_{n_m^{(m)}}(x) \leq G_{n_m^{(m)}}(y)$ for all x and y that belong to T and satisfy the inequality $x \leq y$. Hence for such x and y,

$$G_T(x) \leq G_T(y).$$

It follows from this and (3) that $G = G(x)$ is nondecreasing.

Now let us show that it is continuous on the right. Let $x_k \downarrow x$ and $d = \lim_k G(x_k)$. Clearly $G(x) \le d$, and we have to show that actually $G(x) = d$. Suppose the contrary, that is, let $G(x) < d$. It follows from (3) that there is a $y \in T$, $x < y$, such that $G_T(y) < d$. But $x < x_k < y$ for sufficiently large k, and therefore $G(x_k) \le G_T(y) < d$ and $\lim G(x_k) < d$, which contradicts $d = \lim_k G(x_k)$. Thus we have constructed a function G that belongs to $\mathscr{J}$.

We now establish that $G_{n_{k_m}^{(m)}}(x^0) \to G(x^0)$ for every $x^0 \in P_C(G)$.

If $x^0 < y \in T$, then

$$\overline{\lim_m}\, G_{n_{k_m}^{(m)}}(x^0) \le \overline{\lim_m}\, G_{n_{k_m}^{(m)}}(y) = G_T(y),$$

whence

$$\overline{\lim_m}\, G_{n_{k_m}^{(m)}}(x^0) \le \inf\{G_T(y) \colon y > x^0, y \in T\} = G(x^0). \tag{4}$$

On the other hand, let $x^1 < y < x^0$, $y \in T$. Then

$$G(x^1) \le G_T(y) = \lim_m G_{n_{k_m}^{(m)}}(y) = \underline{\lim_m}\, G_{n_{k_m}^{(m)}}(y) \le \underline{\lim_m}\, G_{n_{k_m}^{(m)}}(x^0).$$

Hence if we let $x^1 \uparrow x^0$ we find that

$$G(x^0-) \le \underline{\lim_m}\, G_{n_{k_m}^{(m)}}(x^0). \tag{5}$$

But if $G(x^0-) = G(x^0)$ we can infer from (4) and (5) that $G_{n_{k_m}^{(m)}}(x^0) \to G(x^0)$, $m \to \infty$.

This completes the proof of the theorem.

We can now complete the proof of Theorem 1.

Sufficiency. Let the family $\mathscr{P}$ be tight and let $\{\mathsf{P}_n\}$ be a sequence of probability measures from $\mathscr{P}$. Let $\{F_n\}$ be the corresponding sequence of distribution functions.

By Helly's theorem, there are a subsequence $\{F_{n_k}\} \subseteq \{F_n\}$ and a generalized distribution function $G \in \mathscr{J}$ such that $F_{n_k}(x) \to G(x)$ for $x \in P_C(G)$. Let us show that because $\mathscr{P}$ was assumed tight, the function $G = G(x)$ is in fact a genuine distribution function ($G(-\infty) = 0$, $G(+\infty) = 1$).

Take $\varepsilon > 0$, and let $I = (a, b]$ be the interval for which

$$\sup_n \mathsf{P}_n(R \backslash I) < \varepsilon,$$

or, equivalently,

$$1 - \varepsilon \le \mathsf{P}_n(a, b], \qquad n \ge 1.$$

Choose points $a', b' \in P_C(G)$ such that $a' < a$, $b' > b$. Then $1 - \varepsilon \le \mathsf{P}_{n_k}(a, b] \le \mathsf{P}_{n_k}(a', b'] = F_{n_k}(b') - F_{n_k}(a') \to G(b') - G(a')$. It follows that $G(+\infty) - G(-\infty) = 1$, and since $0 \le G(-\infty) \le G(+\infty) \le 1$, we have $G(-\infty) = 0$ and $G(+\infty) = 1$.

Therefore the limit function $G = G(x)$ is a distribution function and $F_{n_k} \Rightarrow G$. Together with Theorem 2 of §1 this shows that $\mathsf{P}_{n_k} \xrightarrow{w} \mathsf{Q}$, where Q is the probability measure corresponding to the distribution function G.

This completes the proof of Theorem 1.

4. Problems

1. Carry out the proofs of Theorems 1 and 2 for R^n, $n \geq 2$.

2. Let P_α be a Gaussian measure on the real line, with parameters m_α and σ_α^2, $\alpha \in \mathfrak{A}$. Show that the family $\mathscr{P} = \{\mathsf{P}_\alpha; \alpha \in \mathfrak{A}\}$ is tight if and only if

$$|m_\alpha| \leq a, \qquad \sigma_\alpha^2 \leq b, \qquad \alpha \in \mathfrak{A}.$$

3. Construct examples of tight and nontight families $\mathscr{P} = \{\mathsf{P}_\alpha; \alpha \in \mathfrak{A}\}$ of probability measures defined on $(R^\infty, \mathscr{B}(R^\infty))$.

§3. Proofs of Limit Theorems by the Method of Characteristic Functions

1. The proofs of the first limit theorems of probability theory—the law of large numbers, and the De Moivre–Laplace and Poisson theorems for Bernoulli schemes—were based on direct analysis of the limit functions of the distributions F_n, which are expressed rather simply in terms of binomial probabilities. (In the Bernoulli scheme, we are adding random variables that take only two values, so that in principle we can find F_n explicitly.) However, it is practically impossible to apply a similar direct method to the study of more complicated random variables.

The first step in proving limit theorems for sums of arbitrarily distributed random variables was taken by Chebyshev. The inequality that he discovered, and which is now known as Chebyshev's inequality, not only makes it possible to give an elementary proof of James Bernoulli's law of large numbers, but also lets us establish very general conditions for this law to hold, when stated in the form

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - \frac{\mathsf{E}S_n}{n}\right| \geq \varepsilon\right\} \to 0, \qquad n \to \infty, \quad \text{every } \varepsilon > 0, \tag{1}$$

for sums $S_n = \xi_1 + \cdots + \xi_n$, $n \geq 1$, of independent random variables. (See Problem 2.)

Furthermore, Chebyshev created (and Markov perfected) the "method of moments" which made it possible to show that the conclusion of the De Moivre–Laplace theorem, written in the form

$$\mathsf{P}\left\{\frac{S_n - \mathsf{E}S_n}{\sqrt{\mathsf{V}S_n}} \leq x\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du, \tag{2}$$

is "universal," in the sense that it is valid under very general hypotheses concerning the nature of the random variables. For this reason it is known as the *central limit theorem* of probability theory.

Somewhat later Lyapunov proposed a different method for proving the central limit theorem, based on the idea (which goes back to Laplace) of the characteristic function of a probability distribution. Subsequent developments have shown that Lyapunov's method of characteristic functions is extremely effective for proving the most diverse limit theorems. Consequently it has been extensively developed and widely applied.

In essence, the method is as follows.

2. We already know (Chapter II, §12) that there is a one-to-one correspondence between distribution functions and characteristic functions. Hence we can study the properties of distribution functions by using the corresponding characteristic functions. It is a fortunate circumstance that weak convergence $F_n \xrightarrow{w} F$ of distributions is equivalent to pointwise convergence $\varphi_n \to \varphi$ of the corresponding characteristic functions. Moreover, we have the following result, which provides the basic method of proving theorems on weak convergence for distributions on the real line.

Theorem 1 (Continuity Theorem). *Let $\{F_n\}$ be a sequence of distribution functions $F_n = F_n(x)$, $x \in R$, and let $\{\varphi_n\}$ be the corresponding sequence of characteristic functions,*

$$\varphi_n(t) = \int_{-\infty}^{\infty} e^{itx}\, dF_n(x), \qquad t \in R.$$

(1) *If $F_n \xrightarrow{w} F$, where $F = F(x)$ is a distribution function, then $\varphi_n(t) \to \varphi(t)$, $t \in R$, where $\varphi(t)$ is the characteristic function of $F = F(x)$.*
(2) *If $\lim_n \varphi_n(t)$ exists for each $t \in R$ and $\varphi(t) = \lim_n \varphi_n(t)$ is continuous at $t = 0$, then $\varphi(t)$ is the characteristic function of a probability distribution $F = F(x)$, and*

$$F_n \xrightarrow{w} F.$$

The proof of conclusion (1) is an immediate consequence of the definition of weak convergence, applied to the functions $\operatorname{Re} e^{itx}$ and $\operatorname{Im} e^{itx}$.

The proof of (2) requires some preliminary propositions.

Lemma 1. *Let $\{\mathsf{P}_n\}$ be a tight family of probability measures. Suppose that every weakly convergent subsequence $\{\mathsf{P}_{n'}\}$ of $\{\mathsf{P}_n\}$ converges to the same probability measure P. Then the whole sequence $\{\mathsf{P}_n\}$ converges to P.*

Proof. Suppose that $\mathsf{P}_n \not\xrightarrow{w} \mathsf{P}$. Then there is a bounded continuous function $f = f(x)$ such that

$$\int_R f(x)\mathsf{P}_n(dx) \not\to \int_R f(x)\mathsf{P}(dx).$$

It follows that there exist $\varepsilon > 0$ and an infinite sequence $\{n'\} \subseteq \{n\}$ such that

$$\left| \int_R f(x)\mathsf{P}_{n'}(dx) - \int_R f(x)\mathsf{P}(dx) \right| \geq \varepsilon > 0. \tag{3}$$

By Prokhorov's theorem (§2) we can select a subsequence $\{\mathsf{P}_{n''}\}$ of $\{\mathsf{P}_{n'}\}$ such that $\mathsf{P}_{n''} \xrightarrow{w} \mathsf{Q}$, where Q is a probability measure.

By the hypotheses of the lemma, $\mathsf{Q} = \mathsf{P}$, and therefore

$$\int_R f(x)\mathsf{P}_{n''}(dx) \to \int_R f(x)\mathsf{P}(dx),$$

which leads to a contradiction with (3). This completes the proof of the lemma.

Lemma 2. *Let $\{\mathsf{P}_n\}$ be a tight family of probability measures on $(R, \mathscr{B}(R))$. A necessary and sufficient condition for the sequence $\{\mathsf{P}_n\}$ to converge weakly to a probability measure is that for each $t \in R$ the limit $\lim_n \varphi_n(t)$ exists, where $\varphi_n(t)$ is the characteristic function of P_n:*

$$\varphi_n(t) = \int_R e^{itx}\mathsf{P}_n(dx).$$

Proof. If $\{\mathsf{P}_n\}$ is tight, by Prohorov's theorem there is a subsequence $\{\mathsf{P}_{n'}\}$ and a probability measure P such that $\mathsf{P}_{n'} \xrightarrow{w} \mathsf{P}$. Suppose that the whole sequence $\{\mathsf{P}_n\}$ does not converge to P ($\mathsf{P}_n \not\xrightarrow{w} \mathsf{P}$). Then, by Lemma 1, there is a subsequence $\{P_{n''}\}$ and a probability measure Q such that $\mathsf{P}_{n''} \xrightarrow{w} \mathsf{Q}$, and $\mathsf{P} \neq \mathsf{Q}$.

Now we use the existence of $\lim_n \varphi_n(t)$ for each $t \in R$. Then

$$\lim_{n'} \int_R e^{itx}\mathsf{P}_{n'}(dx) = \lim_{n''} \int_R e^{itx}\mathsf{P}_{n''}(dx)$$

and therefore

$$\int_R e^{itx}\mathsf{P}(dx) = \int_R e^{itx}\mathsf{Q}(dx), \qquad t \in R.$$

But the characteristic function determines the distribution uniquely (Theorem 2, §12, Chapter II). Hence $\mathsf{P} = \mathsf{Q}$, which contradicts the assumption that $\mathsf{P}_n \not\xrightarrow{w} \mathsf{P}$.

The converse part of the lemma follows immediately from the definition of weak convergence.

The following lemma estimates the "tails" of a distribution function in terms of the behavior of its characteristic function in a neighborhood of zero.

Lemma 3. *Let $F = F(x)$ be a distribution function on the real line and let*

$\varphi = \varphi(t)$ be its characteristic function. Then there is a constant $K > 0$ such that for every $a > 0$

$$\int_{|x|\geq 1/a} dF(x) \leq \frac{K}{a}\int_0^a [1 - \operatorname{Re}\varphi(t)]\, dt. \tag{4}$$

PROOF. Since $\operatorname{Re}\varphi(t) = \int_{-\infty}^{\infty} \cos tx\, dF(x)$, we find by Fubini's theorem that

$$\begin{aligned}
\frac{1}{a}\int_0^a [1 - \operatorname{Re}\varphi(t)]\, dt &= \frac{1}{a}\int_0^a \left[\int_{-\infty}^{\infty}(1 - \cos tx)\, dF(x)\right] dt \\
&= \int_{-\infty}^{\infty}\left[\frac{1}{a}\int_0^a (1 - \cos tx)\, dt\right] dF(x) \\
&= \int_{-\infty}^{\infty}\left(1 - \frac{\sin ax}{ax}\right) dF(x) \\
&\geq \inf_{|y|\geq 1}\left(1 - \frac{\sin y}{y}\right)\cdot \int_{|ax|\geq 1} dF(x) \\
&= \frac{1}{K}\int_{|x|\geq 1/a} dF(x),
\end{aligned}$$

where

$$\frac{1}{K} = \inf_{|y|\geq 1}\left(1 - \frac{\sin y}{y}\right) = 1 - \sin 1 \geq \tfrac{1}{7},$$

so that (4) holds with $K = 7$. This establishes the lemma.

Proof of conclusion (2) of Theorem 1. Let $\varphi_n(t) \to \varphi(t)$, $n \to \infty$, where $\varphi(t)$ is continuous at 0. Let us show that it follows that the family of probability measures $\{\mathsf{P}_n\}$ is tight, where P_n is the measure corresponding to F_n.

By (4) and the dominated convergence theorem,

$$\begin{aligned}
\mathsf{P}_n\left\{R\setminus\left(-\frac{1}{a}, \frac{1}{a}\right)\right\} &= \int_{|x|\geq 1/a} dF_n(x) \leq \frac{K}{a}\int_0^a [1 - \operatorname{Re}\varphi_n(t)]\, dt \\
&\to \frac{K}{a}\int_0^a [1 - \operatorname{Re}\varphi(t)]\, dt
\end{aligned}$$

as $n \to \infty$.

Since, by hypothesis, $\varphi(t)$ is continuous at 0 and $\varphi(0) = 1$, for every $\varepsilon > 0$ there is an $a > 0$ such that

$$\mathsf{P}_n\left\{R\setminus\left(-\frac{1}{a}, \frac{1}{a}\right)\right\} \leq \varepsilon$$

for all $n \geq 1$. Consequently $\{\mathsf{P}_n\}$ is tight, and by Lemma 2 there is a probability measure P such that

$$\mathsf{P}_n \xrightarrow{w} \mathsf{P}.$$

Hence

$$\varphi_n(t) = \int_{-\infty}^{\infty} e^{itx}\mathsf{P}_n(dx) \to \int_{-\infty}^{\infty} e^{itx}\mathsf{P}(dx),$$

but also $\varphi_n(t) \to \varphi(t)$. Therefore $\varphi(t)$ is the characteristic function of P.

This completes the proof of the theorem.

Corollary. *Let $\{F_n\}$ be a sequence of distribution functions and $\{\varphi_n\}$ the corresponding sequence of characteristic functions. Also let F be a distribution function and φ its characteristic function. Then $F_n \xrightarrow{w} F$ if and only if $\varphi_n(t) \to \varphi(t)$ for all $t \in R$.*

Remark. Let $\eta, \eta_1, \eta_2, \ldots$ be random variables and $F_{\eta_n} \xrightarrow{w} F_\eta$. In accordance with the definition of §10 of Chapter II, we then say that *the random variables $\eta_1, \eta_2, \ldots$ converge to η in distribution, and write $\eta_n \xrightarrow{d} \eta$.*

Since this notation is self-explanatory, we shall frequently use it instead of $F_{\eta_n} \xrightarrow{w} F_\eta$ when stating limit theorems.

3. In the next section, Theorem 1 will be applied to prove the central limit theorem for independent but not identically distributed random variables. In the present section we shall merely apply the method of characteristic functions to prove some simple limit theorems.

Theorem 2 (Law of Large Numbers). *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables with $\mathsf{E}|\xi_1| < \infty$, $S_n = \xi_1 + \cdots + \xi_n$ and $\mathsf{E}\xi_1 = m$. Then $S_n/n \xrightarrow{\mathsf{P}} m$, that is, for every $\varepsilon > 0$*

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right\} \to 0, \qquad n \to \infty.$$

PROOF. Let $\varphi(t) = \mathsf{E}e^{it\xi_1}$ and $\varphi_{S_n/n}(t) = \mathsf{E}e^{itS_n/n}$. Since the random variables are independent, we have

$$\varphi_{S_n/n}(t) = \left[\varphi\left(\frac{t}{n}\right)\right]^n$$

by (II.12.6). But according to (II.12.14)

$$\varphi(t) = 1 + itm + o(t), \qquad t \to 0.$$

Therefore for each given $t \in R$

$$\varphi\left(\frac{t}{n}\right) = 1 + i\frac{t}{n}m + o\left(\frac{1}{n}\right), \qquad n \to \infty,$$

and therefore

$$\varphi_{S_n/n}(t) = \left[1 + i\frac{t}{n}m + o\left(\frac{1}{n}\right)\right]^n \to e^{itm}.$$

The function $\varphi(t) = e^{itm}$ is continuous at 0 and is the characteristic function of the degenerate probability distribution that is concentrated at m. Therefore

$$\frac{S_n}{n} \xrightarrow{d} m,$$

and consequently (see Problem 7, §10, Chapter II)

$$\frac{S_n}{n} \xrightarrow{P} m.$$

This completes the proof of the theorem.

Theorem 3 (Central Limit Theorem for *Independent Identically Distributed* Random Variables). *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed (nondegenerate) random variables with $\mathsf{E}\xi_1^2 < \infty$ and $S_n = \xi_1 + \cdots + \xi_n$. Then as $n \to \infty$*

$$\mathsf{P}\left\{\frac{S_n - \mathsf{E}S_n}{\sqrt{\mathsf{V}S_n}} \le x\right\} \to \Phi(x), \qquad x \in R, \tag{5}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du.$$

PROOF. Let $\mathsf{E}\xi_1 = m$, $\mathsf{V}\xi_1 = \sigma^2$ and

$$\varphi(t) = \mathsf{E}e^{it(\xi_1 - m)}.$$

Then if we put

$$\varphi_n(t) = \mathsf{E}\exp\left\{it\frac{S_n - \mathsf{E}S_n}{\sqrt{\mathsf{V}S_n}}\right\},$$

we find that

$$\varphi_n(t) = \left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

But by (II.12.14)

$$\varphi(t) = 1 - \frac{\sigma^2 t^2}{2} + o(t^2), \qquad t \to 0.$$

Therefore

$$\varphi_n(t) = \left[1 - \frac{\sigma^2 t^2}{2\sigma^2 n} + o\left(\frac{1}{n}\right)\right]^n \to e^{-t^2/2},$$

as $n \to \infty$ for fixed t.

The function $e^{-t^2/2}$ is the characteristic function of a random variable (denoted by $\mathcal{N}(0, 1)$) with mean zero and unit variance. This, by Theorem 1, also establishes (5). In accordance with the remark in Theorem 1, this can also be written in the form

$$\frac{S_n - \mathsf{E}S_n}{\sqrt{\mathsf{V}S_n}} \xrightarrow{d} \mathcal{N}(0, 1). \tag{6}$$

This completes the proof of the theorem.

The preceding two theorems have dealt with the behavior of the probabilities of (normalized and symmetrized) sums of independent and identically distributed random variables. However, in order to state Poisson's theorem (§6, Chapter I) we have to use a more general model.

Let us suppose that for each $n \geq 1$ we are given a sequence of independent random variables $\xi_{n1}, \ldots, \xi_{nn}$. In other words, let there be given a triangular array

$$\begin{pmatrix} \xi_{11} & & \\ \xi_{21}, & \xi_{22} & \\ \xi_{31}, & \xi_{32}, & \xi_{33} \\ \cdots & \cdots & \cdots \end{pmatrix}$$

of random variables, those in each row being independent. Put $S_n = \xi_{n1} + \cdots + \xi_{nn}$.

Theorem 4 (Poisson's Theorem). *For each $n \geq 1$ let the independent random variables $\xi_{n1}, \ldots, \xi_{nn}$ be such that*

$$\mathsf{P}(\xi_{nk} = 1) = p_{nk}, \qquad \mathsf{P}(\xi_{nk} = 0) = q_{nk},$$

$p_{nk} + q_{nk} = 1$. *Suppose that*

$$\max_{1 \leq k \leq n} p_{nk} \to 0, \qquad n \to \infty,$$

and $\sum_{k=1}^n p_{nk} \to \lambda > 0$, $n \to \infty$. Then, for each $m = 0, 1, \ldots,$

$$\mathsf{P}(S_n = m) \to \frac{e^{-\lambda}\lambda^m}{m!}, \qquad n \to \infty. \tag{7}$$

Proof. Since

$$\mathsf{E}e^{it\xi_{nk}} = p_{nk}e^{it} + q_{nk}$$

for $1 \le k \le n$, by our assumptions we have

$$\varphi_{S_n}(t) = \mathsf{E}e^{itS_n} = \prod_{k=1}^{n}(p_{nk}e^{it} + q_{nk})$$

$$= \prod_{k=1}^{n}(1 + p_{nk}(e^{it} - 1)) \to \exp\{\lambda(e^{it} - 1)\}, \qquad n \to \infty.$$

The function $\varphi(t) = \exp\{\lambda(e^{it} - 1)\}$ is the characteristic function of the Poisson distribution (II.12.11), so that (7) is established.

If $\pi(\lambda)$ denotes a Poisson random variable with parameter λ, then (7) can be written like (6), in the form

$$S_n \xrightarrow{d} \pi(\lambda).$$

This completes the proof of the theorem.

4. Problems

1. Prove Theorem 1 for R^n, $n \ge 2$.
2. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with finite means $\mathsf{E}|\xi_n|$ and variances $\mathsf{V}\xi_n$ such that $\mathsf{V}\xi_n \le K < \infty$, where K is a constant. Use Chebyshev's inequality to prove the law of large numbers (1).
3. Show, as a corollary to Theorem 1, that the family $\{\varphi_n\}$ is *uniformly continuous* and that $\varphi_n \to \varphi$ uniformly on every finite interval.
4. Let ξ_n, $n \ge 1$, be random variables with characteristic functions $\varphi_{\xi_n}(t)$, $n \ge 1$. Show that $\xi_n \xrightarrow{d} 0$ if and only if $\varphi_{\xi_n}(t) \to 1$, $n \to \infty$, in some neighborhood of $t = 0$.
5. Let $X_1, X_2, \ldots$ be a sequence of independent random vectors (with values in R^k) with mean zero and (finite) covariance matrix Γ. Show that

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \Gamma).$$

(Compare Theorem 3.)

§4. Central Limit Theorem for Sums of Independent Random Variables.

I. The Lindeberg Condition

1. In this section, the central limit theorem for (normalized and centralized) sums of independent random variables $\xi_1, \xi_2, \ldots$ will be proved under the traditional hypothesis that the *classical Lindeberg condition* is satisfied. In the next section, we shall consider a more general situation. First, the central limit theorem will be stated in the "series form" and, second, we shall prove it under the so-called *nonclassical hypotheses*.

Theorem 1. *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with finite second moments. Let $m_k = \mathsf{E}\xi_k$, $\sigma_k^2 = \mathsf{V}\xi_k > 0$, $S_n = \xi_1 + \cdots + \xi_n$, $D_n^2 = \sum_{k=1}^n \sigma_k^2$, and let $F_k = F_k(x)$ be the distribution function of the random variable ξ_k.*

Let us suppose that the Lindeberg condition is satisfied: for every $\varepsilon > 0$

$$\text{(L)} \quad \frac{1}{D_n^2} \sum_{k=1}^n \int_{\{x: |x-m_k| \geq \varepsilon D_n\}} (x - m_k)^2 \, dF_k(x) \to 0, \qquad n \to \infty. \tag{1}$$

Then

$$\frac{S_n - \mathsf{E}S_n}{\sqrt{\mathsf{V}S_n}} \xrightarrow{d} \mathcal{N}(0, 1). \tag{2}$$

PROOF. Without loss of generality we assume that $m_k = 0$ for $k \geq 1$. We set $\varphi_k(t) = \mathsf{E}e^{it\xi_k}$, $T_n = S_n/\sqrt{\mathsf{V}S_n} = S_n/D_n$, $\varphi_{S_n}(t) = \mathsf{E}e^{itS_n}$, $\varphi_{T_n}(t) = \mathsf{E}e^{itT_n}$.

Then

$$\varphi_{T_n}(t) = \mathsf{E}e^{itT_n} = \mathsf{E}e^{i(t/D_n)S_n} = \varphi_{S_n}\left(\frac{t}{D_n}\right) = \prod_{k=1}^n \varphi_k\left(\frac{t}{D_n}\right), \tag{3}$$

and for the proof of (2) it is sufficient (by Theorem 1 of §3) to establish that, for every $t \in R$,

$$\varphi_{T_n}(t) \to e^{-t^2/2}, \qquad n \to \infty. \tag{4}$$

We choose a $t \in R$ and suppose that it is fixed throughout the proof. By the representations

$$e^{iy} = 1 + iy + \frac{\theta_1 y^2}{2},$$

$$e^{iy} = 1 + iy - \frac{y^2}{2} + \frac{\theta_2 |y|^3}{3!},$$

which are valid for all real y, with $\theta_1 = \theta_1(y)$ and $\theta_2 = \theta_2(y)$, such that $|\theta_1| \leq 1$, $|\theta_2| \leq 1$, we obtain

$$\begin{aligned}\varphi_k(t) = \mathsf{E}e^{it\xi_k} &= \int_{-\infty}^{\infty} e^{itx} \, dF_k(x) = \int_{|x| \geq \varepsilon D_n} \left(1 + itx + \frac{\theta_1 (tx)^2}{2}\right) dF_k(x) \\ &\quad + \int_{|x| < \varepsilon D_n} \left(1 + itx - \frac{t^2 x^2}{2} + \frac{\theta_2 |tx|^3}{6}\right) dF_k(x) \\ &= 1 + \frac{t^2}{2} \int_{|x| \geq \varepsilon D_n} \theta_1 x^2 \, dF_k(x) - \frac{t^2}{2} \int_{|x| < \varepsilon D_n} x^2 \, dF_k(x) \\ &\quad + \frac{|t|^3}{6} \int_{|x| < \varepsilon D_n} \theta_2 |x|^3 \, dF_k(x)\end{aligned}$$

(here we have also used the fact that, by hypothesis, $m_k = \int_{-\infty}^{\infty} x \, dF_k(x) = 0$).

Consequently,

$$\varphi_k\left(\frac{1}{D_n}\right) = 1 - \frac{t^2}{2D_n^2}\int_{|x|<\varepsilon D_n} x^2\, dF_k(x) + \frac{t^2}{2D_n^2}\int_{|x|<\varepsilon D_n} \theta_1 x^2\, dF_k(x)$$
$$+ \frac{|t|^3}{6D_n^3}\int_{|x|<\varepsilon D_n} \theta_2 |x|^3\, dF_k(x). \tag{5}$$

Since

$$\left|\frac{1}{2}\int_{|x|\geq\varepsilon D_n} \theta_1 x^2\, dF_k(x)\right| \leq \frac{1}{2}\int_{|x|\geq\varepsilon D_n} x^2\, dF_k(x),$$

we have

$$\frac{1}{2}\int_{|x|\geq\varepsilon D_n} \theta_1 x^2\, dF_k(x) = \tilde{\theta}_1 \int_{|x|\geq\varepsilon D_n} x^2\, dF_k(x), \tag{6}$$

where $\tilde{\theta}_1 = \tilde{\theta}_1(t, k, n)$ and $|\tilde{\theta}_1| \leq 1/2$.

In the same way,

$$\left|\frac{1}{6}\int_{|x|<\varepsilon D_n} \theta_2 |x|^3\, dF_k(x)\right| \leq \frac{1}{6}\int_{|x|<\varepsilon D_n} \frac{\varepsilon D_n}{|x|}\cdot |x|^3\, dF_k(x) \leq \frac{1}{6}\int_{|x|<\varepsilon D_n} \varepsilon D_n x^2\, dF_k(x),$$

and therefore,

$$\frac{1}{6}\int_{|x|<\varepsilon D_n} \theta_2 |x|^3\, dF_k(x) = \tilde{\theta}_2 \int_{|x|<\varepsilon D_n} \varepsilon D_n x^2\, dF_k(x), \tag{7}$$

where $\tilde{\theta}_2 = \tilde{\theta}_2(t, k, n)$ and $|\tilde{\theta}_2| \leq 1/6$.

We now set

$$A_{kn} = \frac{1}{D_n^2}\int_{|x|<\varepsilon D_n} x^2\, dF_k(x),$$

$$B_{kn} = \frac{1}{D_n^2}\int_{|x|\geq\varepsilon D_n} x^2\, dF_k(x).$$

Then, by (5)–(7),

$$\varphi_k\left(\frac{t}{D_n}\right) = 1 - \frac{t^2 A_{kn}}{2} + t^2\tilde{\theta}_1 B_{kn} + |t|^3\varepsilon\tilde{\theta}_2 A_{kn} = 1 + C_{kn}. \tag{8}$$

We note that

$$\sum_{k=1}^{n} (A_{kn} + B_{kn}) = 1 \tag{9}$$

and by (1)

$$\sum_{k=1}^{n} B_{kn} \to 0, \qquad n \to \infty. \tag{10}$$

Consequently, for sufficiently large n,

$$\max_{1 \le k \le n} |C_{kn}| \le t^2\varepsilon^2 + \varepsilon|t|^3 \tag{11}$$

and

$$\sum_{k=1}^{n} |C_{kn}| \le t^2 + \varepsilon|t|^3. \tag{12}$$

We now appeal to the fact that, for any complex numbers z with $|z| \le 1/2$,

$$\ln(1 + z) = z + \theta|z|^2,$$

where $\theta = \theta(z)$ with $|\theta| \le 1$ and ln denotes the *principal value* of the logarithm.* Then, for sufficiently large n, it follows from (8) and (11) that, for sufficiently small E > 0,

$$\ln \varphi_k\left(\frac{t}{D_n}\right) = \ln(1 + C_{kn}) = C_{kn} + \theta_{kn}|C_{kn}|^2,$$

where $|\theta_{kn}| \le 1$. Consequently, by (3),

$$\frac{t^2}{2} + \ln \varphi_{T_n}(t) = \frac{t^2}{2} + \sum_{k=1}^{n} \ln \varphi_k\left(\frac{t}{D_n}\right) = \frac{t^2}{2} + \sum_{k=1}^{n} C_{kn} + \sum_{k=1}^{n} \theta_{kn}|C_{kn}|^2.$$

But

$$\frac{t^2}{2} + \sum_{k=1}^{n} C_{kn} = \frac{t^2}{2}\left(1 - \sum_{k=1}^{n} A_{kn}\right) + t^2 \sum_{k=1}^{n} \tilde{\theta}_1(t, k, n)B_{kn} + \varepsilon|t|^3 \sum_{k=1}^{n} \tilde{\theta}_2(t, k, n)A_{kn},$$

and by (9) and (10), for any $\delta > 0$ we can find numbers n_0 and $\varepsilon > 0$, with n_0 so large that for all $n \ge n_0$

$$\left|\frac{t^2}{2} + \sum_{k=1}^{n} C_{kn}\right| \le \frac{\delta}{2}.$$

In addition, by (11) and (12), we can find a positive number ε such that

$$\left|\sum_{k=1}^{n} \theta_{kn}|C_{kn}|^2\right| \le \max_{1 \le k \le n} |C_{kn}| \sum_{k=1}^{n} |C_{kn}| \le (t^2\varepsilon^2 + \varepsilon|t|^3)(t^2 + \varepsilon|t|^3).$$

Therefore, for sufficiently large n, we can choose $\varepsilon > 0$ so that

$$\left|\sum_{k=1}^{n} \theta_{kn}|C_{kn}|^2\right| \le \frac{\delta}{2}$$

and consequently,

$$\left|\frac{t^2}{2} + \ln \varphi_{T_n}(t)\right| \le \delta.$$

* The principal value $\ln z$ of the complex number z is defined by $\ln z = \ln|z| + i \arg z$, $-\pi < \arg z \le \pi$.

Therefore, for any real t,

$$\varphi_{T_n}(t)e^{t^2/2} \to 1, \qquad n \to \infty$$

and hence,

$$\varphi_{T_n}(t) \to e^{-t^2/2}, \qquad n \to \infty.$$

This completes the proof of the theorem.

2. We turn our attention to some special cases in which the Lindeberg condition (1) is satisfied and consequently, the central limit theorem is valid.

a) Let the "Lyapunov condition" be satisfied: for some $\delta > 0$

$$\frac{1}{D_n^{2+\delta}} \sum_{k=1}^{n} \mathsf{E}|\xi_k - m_k|^{2+\delta} \to 0, \qquad n \to \infty. \tag{13}$$

Let $\varepsilon > 0$; then

$$\begin{aligned}
\mathsf{E}|\xi_k - m_k|^{2+\delta} &= \int_{-\infty}^{\infty} |x - m_k|^{2+\delta}\, dF_k(x) \\
&\geq \int_{\{x:\, x|x-m_k| \geq \varepsilon D_n\}} |x - m_k|^{2+\delta}\, dF_k(x) \\
&\geq \varepsilon^{\delta} D_n^{\delta} \int_{\{x:\, |x-m_k| \geq \varepsilon D_n\}} (x - m_k)^2\, dF_k(x)
\end{aligned}$$

and therefore,

$$\frac{1}{D_n^2} \sum_{k=1} \int_{\{x:\, |x-m_k| \geq \varepsilon D_n\}} (x - m_k)^2\, dF_k(x) \leq \frac{1}{\varepsilon^{\delta}} \cdot \frac{1}{D_n^{2+\delta}} \sum_{k=1}^{n} \mathsf{E}|\xi_k - m_k|^{2+\delta}.$$

Consequently, the Lyapunov condition implies the Lindeberg condition.

b) Let $\xi_1, \xi_2, \ldots$ be independent identically distributed random variables with $m = \mathsf{E}\xi_1$ and variance $0 < \sigma^2 \equiv \mathsf{V}\xi_1 < \infty$. Then

$$\begin{aligned}
&\frac{1}{D_n^2} \sum_{k=1}^{n} \int_{\{x:\, |x-m| \geq \varepsilon D_n\}} |x - m|^2\, dF_k(x) \\
&\quad = \frac{n}{n\sigma^2} \int_{\{x:\, |x-m| \geq \varepsilon\sigma^2\sqrt{n}\}} |x - m|^2\, dF_1(x) \to 0,
\end{aligned}$$

since $\{x: |x - m| \geq \varepsilon\sigma^2\sqrt{n}\} \downarrow \varnothing$, $n \to \infty$, and $\sigma^2 = \mathsf{E}|\xi_1 - m|^2 < \infty$.

Therefore, the Lindeberg condition is satisfied and consequently, Theorem 3 of §3 follows from the proof of Theorem 1.

c) Let $\xi_1, \xi_2, \ldots$ be independent random variables such that for all $n \geq 1$

$$|\xi_k| \leq K < \infty,$$

where K is a constant and $D_n \to \infty$, $n \to \infty$.

Then by Chebyshev's inequality

$$\int_{\{x:|x-m_k|\geq\varepsilon D_n\}} |x-m_k|^2\,dF_k(x) = \mathsf{E}[(\xi_k-m_k)^2 I(|\xi_k-m_k|\geq\varepsilon D_n)]$$

$$\leq (2K)^2\mathsf{P}\{|\xi_k-m_k|\geq\varepsilon D_n\}\leq(2K)^2\frac{\sigma_k^2}{\varepsilon^2 D_n^2}$$

and therefore,

$$\frac{1}{D_n^2}\sum_{k=1}^n\int_{\{x:|x-m_k|\geq\varepsilon D_n\}}|x-m_k|^2\,dF_k(x)\leq\frac{(2K)^2}{\varepsilon^2 D_n^2}\to 0,\qquad n\to\infty.$$

Consequently, the Lindeberg condition is satisfied again and therefore, the central limit theorem is verified.

3. **Remark 1.** Let $T_n=(S_n-\mathsf{E}S_n)/D_n$ and $F_{T_n}(x)=\mathsf{P}(T_n\leq x)$. Then proposition (2) shows that for all $x\in R$

$$F_{T_n}(x)\to\Phi(x),\qquad n\to\infty.$$

Since $\Phi(x)$ is continuous, the convergence here is actually uniform (problem 5, §1):

$$\sup_{x\in R}|F_{T_n}(x)-\Phi(x)|\to 0,\qquad n\to\infty. \tag{14}$$

In particular, it follows that

$$\mathsf{P}\{S_n\leq x\}-\Phi\left(\frac{x-\mathsf{E}S_n}{D_n}\right)\to 0,\qquad n\to\infty.$$

This proposition is often expressed by the statement that for sufficiently large n the value S_n *is approximately normally distributed with mean* $\mathsf{E}S_n$ *and variance* $D_n^2\equiv\mathsf{V}S_n$.

Remark 2. Since, according to the preceding remarks, $F_{T_n}(x)\to\Phi(x)$ as $n\to\infty$, uniformly in x, it is natural to raise the question of the *rate of convergence* in (14). In the case when the numbers $\xi_1,\xi_2,\ldots$ are independent and uniformly distributed with $\mathsf{E}|\xi_1|^3<\infty$, this question is answered by the *Berry-Esseen inequality*:

$$\sup_x|F_{T_n}(x)-\Phi(x)|\leq C\frac{\mathsf{E}|\xi_1-\mathsf{E}\xi_1|^3}{\sigma^3\sqrt{n}}, \tag{15}$$

where the absolute constant C satisfies the inequality

$$1/\sqrt{(2\pi)}\leq C<0.8.$$

The proof of (15) will be given in §11.

Remark 3. We can state the Lindeberg condition in a somewhat different (and more compact) version which is especially convenient in the "series form."

Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables, with $m_k = \mathsf{E}\xi_k$, $\sigma_k^2 = \mathsf{V}\xi_k > 0$, $D_n^2 = \sum_{k=1}^n \sigma_k^2$, and $\xi_{nk} = (\xi_k - m_k)/D_n$. In this notation, condition (1) assumes the following form:

$$\text{(L)} \qquad \sum_{k=1}^n \mathsf{E}[\xi_{nk}^2 I(|\xi_{nk}| \geq \varepsilon)] \to 0, \qquad n \to \infty. \tag{16}$$

If $S_n = \xi_{n1} + \cdots + \xi_{nn}$, we have $\mathsf{V}S_n = 1$ and Theorem 1 can be given the following form: *if* (16) *is satisfied,* we have

$$S_n \xrightarrow{d} \mathscr{N}(0, 1).$$

In this form the central limit theorem is valid without the assumption that ξ_{nk} has the *special* form $(\xi_{nk} - m_k)/D_n$. In fact, we have the following result whose proof is word for word the same as that of Theorem 1.

Theorem 2. *For each $n \geq 1$ let*

$$\xi_{n1}, \xi_{n2}, \ldots, \xi_{nn}$$

be a sequence of independent random variables for which $\mathsf{E}\xi_{nk} = 0$ and $\mathsf{V}S_n = 1$, where $S_n = \xi_{n1} + \cdots + \xi_{nn}$.

Then the Lindeberg condition (16) *is a sufficient condition for the convergence $S_n \xrightarrow{d} \mathscr{N}(0, 1)$.*

4. Since

$$\max_{1 \leq k \leq n} \mathsf{E}\xi_{nk}^2 \leq \varepsilon^2 + \sum_{k=1}^n \mathsf{E}[\xi_{nk}^2 I(|\xi_{nk}| \geq \varepsilon)],$$

it is clear that the Lindeberg condition (16) implies that

$$\max_{1 \leq k \leq n} \mathsf{E}\xi_{nk}^2 \to 0, \qquad n \to \infty. \tag{17}$$

It is noteworthy that when this condition is satisfied, it follows automatically from the validity of the central limit theorem that the Lindeberg condition is satisfied (Lindberg–Feller theorem).

Theorem 3. *For each $n \geq 1$ let*

$$\xi_{n1}, \xi_{n2}, \ldots, \xi_{nn}$$

be a sequence of independent random variables for which $\mathsf{E}\xi_{nk} = 0$ and $\mathsf{V}S_n = 1$, where $S_n = \xi_{n1} + \cdots + \xi_{nn}$. Let (17) *be satisfied. Then the Lindeberg condition is necessary and sufficient for the validity of the central limit theorem, $S_n \to \mathscr{N}(0, 1)$.*

The sufficiency follows from Theorem 2. To establish the necessity we need the following lemma (compare Lemma 3, §3, Chapter III).

Lemma. *Let ξ be a random variable with distribution function $F = F(x)$, $\mathsf{E}\xi = 0$, $\mathsf{V}\xi = \gamma > 0$. Then for every $a > 0$*

$$\int_{|x| \geq 1/a} x^2 \, dF(x) \leq \frac{1}{a^2}[\operatorname{Re} f(\sqrt{6}a) - 1 + 3\gamma a^2], \tag{18}$$

where $f(t) = \mathsf{E}e^{it\xi}$ is the characteristic function of ξ.

PROOF. We have

$$\operatorname{Re} f(t) - 1 + \tfrac{1}{2}\gamma t^2 = \tfrac{1}{2}\gamma t^2 - \int_{-\infty}^{\infty} [1 - \cos tx] \, dF(x)$$

$$= \tfrac{1}{2}\gamma t^2 - \int_{|x| < 1/a} [1 - \cos tx] \, dF(x) - \int_{|x| \geq 1/a} [1 - \cos tx] \, dF(x)$$

$$\geq \tfrac{1}{2}\gamma t^2 - \tfrac{1}{2}t^2 \int_{|x| < 1/a} x^2 \, dF(x) - 2a^2 \int_{|x| \geq 1/a} x^2 \, dF(x)$$

$$= (\tfrac{1}{2}t^2 - 2a^2) \cdot \int_{|x| \geq 1/a} x^2 \, dF(x).$$

If we set $t = \sqrt{6}a$, we obtain (18), as required.

We now turn to the proof of the necessity in Theorem 3.

Let

$$F_{nk}(x) = \mathsf{P}(\xi_{nk} \leq x), \qquad f_{nk}(t) = \mathsf{E}e^{it\xi_{nk}},$$
$$\mathsf{E}\xi_{nk} = 0, \qquad \mathsf{V}\xi_{nk} = \gamma_{nk} > 0, \tag{19}$$
$$\sum_{k=1}^{n} \gamma_{nk} = 1, \qquad \max_{1 \leq k \leq n} \gamma_{nk} \to 0, \qquad n \to \infty.$$

Let $\ln z$ denote the *principal value* of the logarithm of the complex number z.

Then

$$\ln \prod_{k=1}^{n} f_{nk}(t) = \sum_{k=1}^{n} \ln f_{nk}(t) + 2\pi i m,$$

where $m = m(n, t)$ is an integer. Consequently,

$$\operatorname{Re} \ln \prod_{k=1}^{n} f_{nk}(t) = \operatorname{Re} \sum_{k=1}^{n} \ln f_{nk}(t). \tag{20}$$

Since

$$\prod_{k=1}^{n} f_{nk}(t) \to e^{-(1/2)t^2},$$

we have

$$\left| \prod_{k=1}^{n} f_{nk}(t) \right| \to e^{-(1/2)t^2}.$$

Therefore,

$$\operatorname{Re} \ln \prod_{k=1}^{n} f_{nk}(t) = \operatorname{Re} \ln \left| \prod_{k=1}^{n} f_{nk}(t) \right| \to -\tfrac{1}{2}t^2. \tag{21}$$

For $|z| < 1$

$$\ln(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \cdots \tag{22}$$

and for $|z| \le 1/2$

$$|\ln(1+z) - z| \le |z|^2. \tag{23}$$

By (19), for each fixed t, all sufficiently large n and all $k = 1, 2, \ldots, n$, we have

$$|f_{nk}(t) - 1| \le \tfrac{1}{2}\gamma_{nk} t^2 \le \tfrac{1}{2}. \tag{24}$$

Hence, we obtain from (23) and (24)

$$\left| \sum_{k=1}^{n} \{\ln[1 + (f_{nk}(t) - 1)] - (f_{nk}(t) - 1)\} \right| \le \sum_{k=1}^{n} |f_{nk}(t) - 1|^2$$

$$\le \frac{t^4}{4} \max_{1 \le k \le n} \gamma_{nk} \sum_{k=1}^{n} \gamma_{nk} = \frac{t^4}{4} \max_{1 \le k \le n} \gamma_{nk} \to 0, \qquad n \to \infty,$$

and consequently,

$$\left| \operatorname{Re} \sum_{k=1}^{n} \ln f_{nk}(t) - \operatorname{Re} \sum_{k=1}^{n} (f_{nk}(t) - 1) \right| \to 0, \qquad n \to \infty. \tag{25}$$

It follows from (20), (21), and (25) that

$$\operatorname{Re} \sum_{k=1}^{n} (f_{nk}(t) - 1) + \tfrac{1}{2}t^2 = \sum_{k=1}^{n} [\operatorname{Re} f_{nk}(t) - 1 + \tfrac{1}{2}t^2 \gamma_{nk}] \to 0, \qquad n \to \infty.$$

Setting $t = \sqrt{6}a$, we find that for each $a > 0$

$$\sum_{k=1}^{n} [\operatorname{Re} f_{nk}(\sqrt{6}a) - 1 + 3a^2 \gamma_{nk}] \to 0, \qquad n \to \infty. \tag{26}$$

Finally, from (18) with $a = 1/\varepsilon$ and (26), we obtain

$$\sum_{k=1}^{n} \mathsf{E}[\xi_{nk} I(|\xi_{nk}| \ge \varepsilon] = \sum_{k=1}^{n} \int_{|x| \ge \varepsilon} x^2 \, dF_{nk}(x)$$

$$\le \varepsilon^2 \sum_{k=1}^{n} [\operatorname{Re} f_{nk}(\sqrt{6}a) - 1 + 3a^2 \gamma_{nk}] \to 0, \qquad n \to \infty,$$

which shows that the Lindeberg condition is satisfied.

5. Problems

1. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables with $\mathsf{E}\xi_1 = 0$ and $\mathsf{E}\xi_1^2 = 1$. Show that

$$\max\left(\frac{|\xi_1|}{\sqrt{n}}, \ldots, \frac{|\xi_n|}{\sqrt{n}}\right) \xrightarrow{d} 0, \qquad n \to \infty.$$

2. Show that in the Bernoulli scheme the quantity $\sup_x |F_{T_n}(x) - \Phi(x)|$ is of order $1/\sqrt{n}$, $n \to \infty$.

§5. Central Limit Theorem for Sums of Independent Random Variables

II. Nonclassical Conditions

1. It was shown in §4 that the Lindeberg condition (4.16) implies that the condition

$$\max_{1 \le k \le n} \mathsf{E}\xi_{nk}^2 \to 0,$$

is satisfied. In turn, this implies the so-called condition of being *negligible in the limit* (*asymptotically infinitesimal*), that is, for every $\varepsilon > 0$,

$$\max_{1 \le k \le n} \mathsf{P}\{|\xi_{nk}| \ge \varepsilon\} \to 0, \qquad n \to \infty.$$

Consequently, we may say that Theorems 1 and 2 of §4 provide a condition of realization of the central limit theorem for sums of independent random variables under the hypothesis of negligibility in the limit. Limit theorems in which conditions of negligibility in the limit are imposed on individual terms are usually called theorems with a classical formulation. It is easy, however, to give examples of nondegenerate random variables for which neither the Lindeberg condition nor negligibility in the limit is satisfied, but nevertheless the central limit theorem is satisfied. Here is the simplest example.

Let $\xi_1, \xi_2, \ldots$ be a sequence of independent normally distributed random variables with $\mathsf{E}\xi_n = 0$, $\mathsf{V}\xi_1 = 1$, $\mathsf{V}\xi_k = 2^{k-2}$, $k \ge 2$. Let $S_n = \xi_{n1} + \cdots + \xi_{nn}$ with

$$\xi_{nk} = \xi_k \Big/ \sqrt{\sum_{k=1}^{n} \mathsf{V}\xi_i}.$$

It is easily verified that here neither the Lindeberg condition nor the condition of negligibility in the limit is satisfied, although the validity of the central limit theorem is evident, since S_n is normally distributed with $\mathsf{E}S_n = 0$ and $\mathsf{V}S_n = 1$.

Theorem 1 (below) provides a sufficient (and necessary) condition for the central limit theorem without assuming the "classical" condition of negligibility in the limit. In this sense, condition (Λ), presented below, is an example of "nonclassical" conditions which reflect the title of this section.

2. We shall suppose that for each $n \geq 1$ there is a given sequence ("series form") of independent random variables

$$\xi_{n1}, \xi_{n2}, \ldots, \xi_{nn}$$

with $\mathsf{E}\xi_{nk} = 0$, $\mathsf{V}\xi_{nk} = \sigma_{nk}^2 > 0$, $\sum_{k=1}^n \sigma_{nk}^2 = 1$. Let $S_n = \xi_{n1} + \cdots + \xi_{nn}$,

$$F_{nk}(x) = \mathsf{P}\{\xi_{nk} \leq x\}, \qquad \Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-y^2/2}\, dy, \qquad \Phi_{nk}(x) = \Phi\left(\frac{x}{\sigma_{nk}}\right).$$

Theorem 1. *To have*

$$S_n \xrightarrow{d} \mathcal{N}(0, 1), \tag{1}$$

it is sufficient (and necessary) that for every $\varepsilon > 0$ the condition

$$(\Lambda) \qquad \sum_{k=1}^{n} \int_{|x| > \varepsilon} |x| \, |F_{nk}(x) - \Phi_{nk}(x)|\, dx \to 0, \qquad n \to \infty \tag{2}$$

is satisfied.

The following theorem clarifies the connection between condition (Λ) and the classical Lindeberg condition

$$(\mathrm{L}) \qquad \sum_{k=1}^{n} \int_{|x| > \varepsilon} x^2 \, dF_{nk}(x) \to 0, \qquad n \to \infty. \tag{3}$$

Theorem 2. 1. *The Lindeberg condition implies that condition (Λ) is satisfied:*

$$(\mathrm{L}) \Rightarrow (\Lambda).$$

2. *If $\max_{1 \leq k \leq n} \mathsf{E}\xi_{nk}^2 \to 0$ as $n \to \infty$, the condition (Λ) implies the Lindeberg condition (L):*

$$(\Lambda) \Rightarrow (\mathrm{L}).$$

Proof of Theorem 1. The proof of the necessity of condition (Λ) is rather complicated. Here we only prove the sufficiency.

Let

$$f_{nk}(t) = \mathsf{E}e^{it\xi_{nk}}, \qquad f_n(t) = \mathsf{E}e^{itS_n},$$

$$\varphi_{nk}(t) = \int_{-\infty}^{\infty} e^{itx}\, d\Phi_{nk}(x), \qquad \varphi(t) = \int_{-\infty}^{\infty} e^{itx}\, d\Phi(x).$$

It follows from §12, Chapter II, that

$$\varphi_{nk}(t) = e^{-(t^2\sigma_{nk}^2)/2}, \qquad \varphi(t) = e^{-t^2/2}.$$

By the corollary of Theorem 1 of §3, we have $S_n \xrightarrow{d} \mathcal{N}(0, 1)$ if and only if $f_n(t) \to \varphi(t)$ as $n \to \infty$, for every real t.

We have

$$f_n(t) - \varphi(t) = \prod_{k=1}^{n} f_{nk}(t) - \prod_{k=1}^{n} \varphi_{nk}(t).$$

Since $|f_{nk}(t)| \le 1$ and $|\varphi_{nk}(t)| \le 1$, we have

$$\begin{aligned} |f_n(t) - \varphi(t)| &= \left| \prod_{k=1}^{n} f_{nk}(t) - \prod_{k=1}^{n} \varphi_{nk}(t) \right| \\ &\le \sum_{k=1}^{n} |f_{nk}(t) - \varphi_{nk}(t)| = \sum_{k=1}^{n} \left| \int_{-\infty}^{\infty} e^{itx} d(F_{nk} - \Phi_{nk}) \right| \\ &= \sum_{k=1}^{n} \left| \int_{-\infty}^{\infty} \left(e^{itx} - itx + \frac{t^2x^2}{2} \right) d(F_{nk} - \Phi_{nk}) \right|, \end{aligned} \tag{4}$$

where we have used the fact that

$$\int_{-\infty}^{\infty} x^k \, dF_{nk} = \int_{-\infty}^{\infty} x^k \, d\Phi_{nk} \qquad \text{for} \quad k = 1, 2.$$

If we apply the formula for integration by parts (Theorem 11, §6, Chapter II) to the integral

$$\int_a^b \left(e^{itx} - itx + \frac{t^2x^2}{2} \right) d(F_{nk} - \Phi_{nk}),$$

we obtain (taking account of the limits $x^2[1 - F_{nk}(x) + F_{nk}(-x)] \to 0$, and $x^2[1 - \Phi_{nk}(x) + \Phi_{nk}(-x)] \to 0$, $x \to \infty$)

$$\begin{aligned} &\int_{-\infty}^{\infty} \left(e^{itx} - itx + \frac{t^2x^2}{2} \right) d(F_{nk} - \Phi_{nk}) \\ &\qquad = it \int_{-\infty}^{\infty} (e^{itx} - 1 - itx)(F_{nk}(x) - \Phi_{nk}(x)) \, dx. \end{aligned} \tag{5}$$

From (4) and (5), we obtain

$$\begin{aligned} |f_n(t) - \varphi(t)| &\le \sum_{k=1}^{n} \left| t \int_{-\infty}^{\infty} (e^{itx} - 1 - itx)(F_{nk}(x) - \Phi_{nk}(x)) \right| \\ &\le \frac{|t|^3}{2} \varepsilon \sum_{k=1}^{n} \int_{|x| \ge \varepsilon} |x| \, |F_{nk}(x) - \Phi_{nk}(x)| \, dx \\ &\quad + 2t^2 \sum_{k=1}^{n} \int_{|x| > \varepsilon} |x| \, |F_{nk}(x) - \Phi_{nk}(x)| \, dx \\ &\le \varepsilon |t|^3 \sum_{k=1}^{n} \sigma_{nk}^2 + 2t^2 \sum_{k=1}^{n} \int_{|x| > \varepsilon} |x| \, |F_{nk}(x) - \Phi_{nk}(x)| \, dx, \end{aligned} \tag{6}$$

where we have used the inequality

$$\int_{|x|\le\varepsilon} |x||F_{nk}(x) - \Phi_{nk}(x)|\,dx \le 2\sigma_{nk}^2, \tag{7}$$

which is easily established by using (71), §6, Chapter II.

It follows from (6) that $f_n(t) \to \varphi(t)$ as $n \to \infty$, because ε is an arbitrary positive number and condition (Λ) is satisfied.

This completes the proof of the theorem.

PROOF OF THEOREM 2. 1. By §4 of the Lindeberg condition (L), it follows that $\max_{1\le k\le n}\sigma_{nk}^2 \to 0$. Hence, if we use the fact that $\sum_{k=1}^n \sigma_{nk}^2 = 1$, we obtain

$$\sum_{k=1}^n \int_{|x|>\varepsilon} x^2\,d\Phi_{nk}(x) \le \int_{|x|>\varepsilon/\sqrt{\max_{1\le k\le n}\sigma_{nk}^2}} x^2\,d\Phi(x) \to 0, \qquad n\to\infty. \tag{8}$$

Together with Condition (L), this shows that, for every $\varepsilon > 0$,

$$\sum_{k=1}^n \int_{|x|>\varepsilon} x^2 d[F_{nk}(x) + \Phi_{nk}(x)] \to 0, \qquad n\to\infty. \tag{9}$$

Let us fix $\varepsilon > 0$. Then there is a continuous differentiable even function $h = h(x)$ for which $|h(x)| \le x^2$, $|h'(x)| \le 4x$, and

$$h(x) = \begin{cases} x^2 & |x| > 2\varepsilon, \\ 0, & |x| \le \varepsilon. \end{cases}$$

For $h(x)$, we have by (9)

$$\sum_{k=1}^n \int_{|x|>\varepsilon} h(x) d[F_{nk}(x) + \Phi_{nk}(x)] \to 0, \qquad n\to\infty. \tag{10}$$

By integrating by parts in (10), we obtain

$$\sum_{k=1}^n \int_{x\ge\varepsilon} h'(x)[(1 - F_{nk}(x)) + (1 - \Phi_{nk}(x))]\,dx$$

$$= \sum_{k=1}^n \int_{x\ge\varepsilon} h(x)d[F_{nk} + \Phi_{nk}] \to 0,$$

$$\sum_{k=1}^n \int_{x\le-\varepsilon} h'(x)[F_{nk}(x) + \Phi_{nk}(x)]\,dx = \sum_{k=1}^n \int_{x\le-\varepsilon} h(x)d[F_{nk} + \Phi_{nk}] \to 0.$$

Since $h'(x) = 2x$ for $|x| \ge 2\varepsilon$, we obtain

$$\sum_{k=1}^n \int_{|x|\ge 2\varepsilon} |x||F_{nk}(x) - \Phi_{nk}(x)|\,dx \to 0, \qquad n\to\infty.$$

Therefore, since ε is an arbitrary positive number, we find that (L) $\Rightarrow$ (Λ).

2. For the function $h = h(x)$ introduced above, we find by (8) and the condition $\max_{1\le k\le n}\sigma_{nk}^2 \to 0$ that

$$\sum_{k=1}^n \int_{|x|>\varepsilon} h(x)\,d\Phi_{nk}(x) \le \sum_{k=1}^n \int_{|x|>\varepsilon} x^2\,d\Phi_{nk}(x) \to 0, \qquad n\to\infty. \tag{11}$$

If we integrate by parts, we obtain

$$\left|\sum_{k=1}^{n}\int_{|x|\geq\varepsilon} h(x)d[F_{nk}-\Phi_{nk}]\right| \leq \left|\sum_{k=1}^{n}\int_{x\geq\varepsilon} h(x)d[(1-F_{nk})-(1-\Phi_{nk})]\right|$$
$$+\left|\sum_{k=1}^{n}\int_{x\leq-\varepsilon} h(x)d[F_{nk}-\Phi_{nk}]\right|$$
$$\leq \sum_{k=1}^{n}\int_{x\geq\varepsilon} |h'(x)|\,|[(1-F_{nk})-(1-\Phi_{nk})]|\,dx$$
$$+\sum_{k=1}^{n}\int_{x\leq-\varepsilon} |h'(x)|\,|F_{nk}-\Phi_{nk}|\,dx$$
$$\leq 4\sum_{k=1}^{n}\int_{|x|\geq\varepsilon} |x|\,|F_{nk}(x)-\Phi_{nk}(x)|\,dx. \tag{12}$$

It follows from (11) and (12) that

$$\sum_{k=1}^{n}\int_{|x|\geq 2\varepsilon} x^2\,dF_{nk}(x) \leq \sum_{k=1}^{n}\int_{|x|\geq\varepsilon} h(x)\,dF_{nk}(x) \to 0, \qquad n\to\infty,$$

i.e., the Lindeberg condition (L) is satisfied.

This completes the proof of the theorem.

3. Problems

1. Establish formula (5).
2. Verify relations (10) and (12).

§6. Infinitely Divisible and Stable Distributions

1. In stating Poisson's theorem in §3 we found it necessary to use a triangular array, supposing that for each $n \geq 1$ there was a sequence of independent random variables $\{\xi_{n,k}\}$, $1 \leq k \leq n$.

Put

$$T_n = \xi_{n,1} + \cdots + \xi_{n,n}, \qquad n \geq 1. \tag{1}$$

The idea of an infinitely divisible distribution arises in the following problem: how can we determine all the distributions that can be expressed as limits of sequences of distributions of random variables T_n, $n \geq 1$?

Generally speaking, the problem of limit distributions is indeterminate in such great generality. Indeed, if ξ is a random variable and $\xi_{n,1} = \xi$, $\xi_{n,k} = 0$, $1 < k \leq n$, then $T_n \equiv \xi$ and consequently the limit distribution is the distribution of ξ, which can be arbitrary.

In order to have a more meaningful problem, we shall suppose in the present section that the variables $\xi_{n,1}, \ldots, \xi_{n,n}$ are, for each $n \geq 1$, not only independent, but also identically distributed.

Recall that this was the situation in Poisson's theorem (Theorem 4 of §3). The same framework also includes the central limit theorem (Theorem 3 of §3) for sums $S_k = \xi_1 + \cdots + \xi_n$, $n \geq 1$, of independent identically distributed random variables $\xi_1, \xi_2, \ldots$. In fact, if we put

$$\xi_{n,k} = \frac{\xi_k - \mathsf{E}\xi_k}{V_n}, \qquad V_n^2 = \mathsf{V}S_n,$$

then

$$T_n = \sum_{k=1}^{n} \xi_{n,k} = \frac{S_n - \mathsf{E}S_n}{V_n}.$$

Consequently both the normal and the Poisson distributions can be presented as limits in a triangular array. If $T_n \to T$, it is intuitively clear that since T_n is a sum of independent identically distributed random variables, the limit variable T must also be a sum of independent identically distributed random variables. With this in mind, we introduce the following definition.

Definition 1. A random variable T, its distribution F_T, and its characteristic function φ_T are said to be infinitely divisible if, for each $n \geq 1$, there are independent identically distributed random variables $\eta_1, \ldots, \eta_n$ such that† $T \overset{d}{=} \eta_1 + \cdots + \eta_n$ (or, equivalently, $F_T = F_{\eta_1} * \cdots * F_{\eta_n}$, or $\varphi_T = (\varphi_{\eta_1})^n$).

Theorem 1. *A random variable T can be a limit of sums $T_n = \sum_{k=1}^{n} \xi_{n,k}$ if and only if T is infinitely divisible.*

PROOF. If T is infinitely divisible, for each $n \geq 1$ there are independent identically distributed random variables $\xi_{n,1}, \ldots, \xi_{n,k}$ such that $T \overset{d}{=} \xi_{n,1} + \cdots + \xi_{n,k}$, and this means that $T \overset{d}{=} T_n$, $n \geq 1$.

Conversely, let $T_n \xrightarrow{d} T$. Let us show that T is infinitely divisible, i.e., for each k there are independent identically distributed random variables $\eta_1, \ldots, \eta_k$ such that $T \overset{d}{=} \eta_1 + \cdots + \eta_k$.

Choose a $k \geq 1$ and represent T_{nk} in the form $\zeta_n^{(1)} + \cdots + \zeta_n^{(k)}$, where

$$\zeta_n^{(1)} = \xi_{nk,1} + \cdots + \xi_{nk,n}, \ldots, \zeta_n^{(k)} = \xi_{nk,n(k-1)+1} + \cdots + \xi_{nk,nk}.$$

Since $T_{nk} \xrightarrow{d} T$, $n \to \infty$, the sequence of distribution functions corresponding to the random variables T_{nk}, $n \geq 1$, is relatively compact and therefore, by Prohorov's theorem, is tight. Moreover,

$$[\mathsf{P}(\zeta_n^{(1)} > z)]^k = \mathsf{P}(\zeta_n^{(1)} > z, \ldots, \zeta_n^{(k)} > z) \leq \mathsf{P}(T_{nk} > kz)$$

and

$$[\mathsf{P}(\zeta_n^{(1)} < -z)]^k = \mathsf{P}(\zeta_n^{(1)} < -z, \ldots, \zeta_n^{(k)} < -z) \leq \mathsf{P}(T_{nk} < -kz).$$

† The notation $\xi \overset{d}{=} \eta$ means that the random variables ξ and η agree in distribution, i.e., $F_\xi(x) = F_\eta(x)$, $x \in R$.

The family of distributions for $\zeta_n^{(1)}$, $n \geq 1$, is tight because of the preceding two inequalities and because the family of distributions for T_{nk}, $n \geq 1$, is tight. Therefore there is a subsequence $\{n_i\} \subseteq \{n\}$ and a random variable η_1 such that $\zeta_{n_i}^{(1)} \xrightarrow{d} \eta_1$ as $n_i \to \infty$. Since the variables $\zeta_n^{(1)}, \ldots, \zeta_n^{(k)}$ are identically distributed, we have $\zeta_{n_i}^{(2)} \xrightarrow{d} \eta_2, \ldots, \zeta_{n_i}^{(k)} \xrightarrow{d} \eta_k$, where $\eta_1 \overset{d}{=} \eta_2 \overset{d}{=} \cdots = \eta_k$. Since $\zeta_n^{(1)}, \ldots, \zeta_n^{(k)}$ are independent, it follows from the corollary to Theorem 1 of §3 that $\eta_1, \ldots, \eta_k$ are independent and

$$T_{n_ik} = \zeta_{n_i}^{(1)} + \cdots + \zeta_{n_i}^{(k)} \xrightarrow{d} \eta_1 + \cdots + \eta_k.$$

But $T_{n_ik} \xrightarrow{d} T$, therefore (Problem 1)

$$T \overset{d}{=} \eta_1 + \cdots + \eta_k.$$

This completes the proof of the theorem.

Remark. The conclusion of the theorem remains valid if we replace the hypothesis that $\xi_{n,1}, \ldots, \xi_{n,n}$ are identically distributed for each $n \geq 1$ by the hypothesis that they are uniformly asymptotically infinitesimal (4.2).

2. To test whether a given random variable T is infinitely divisible, it is simplest to begin with its characteristic function $\varphi(t)$. If we can find characteristic functions $\varphi_n(t)$ such that $\varphi(t) = [\varphi_n(t)]^n$ for every $n \geq 1$, then T is infinitely divisible.

In the Gaussian case,

$$\varphi(t) = e^{itm}e^{-(1/2)t^2\sigma^2},$$

and if we put

$$\varphi_n(t) = e^{itm/n}e^{-(1/2)t^2\sigma^2/n},$$

we see at once that $\varphi(t) = [\varphi_n(t)]^n$.

In the Poisson case,

$$\varphi(t) = \exp\{\lambda(e^{it} - 1)\},$$

and if we put $\varphi_n(t) = \exp\{(\lambda/n)(e^{it} - 1)\}$ then $\varphi(t) = [\varphi_n(t)]^n$.

If a random variable T has a Γ-distribution with density

$$f(x) = \begin{cases} \dfrac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

it is easy to show that its characteristic function is

$$\varphi(t) = \frac{1}{(1 - i\beta t)^\alpha}.$$

Consequently $\varphi(t) = [\varphi_n(t)]^n$ where

$$\varphi_n(t) = \frac{1}{(1 - i\beta t)^{\alpha/n}},$$

and therefore T is infinitely divisible.

We quote without proof the following result on the general form of the characteristic functions of infinitely divisible distributions.

Theorem 2 (Lévy–Khinchin Theorem). *A random variable T is infinitely divisible if and only if $\varphi(t) = \exp \psi(t)$ and*

$$\psi(t) = it\beta - \frac{t^2\sigma^2}{2} + \int_{-\infty}^{\infty} \left(e^{itx} - 1 - \frac{itx}{1+x^2}\right) \frac{1+x^2}{x^2}\, d\lambda(x), \tag{2}$$

where $\beta \in R$, $\sigma^2 \geq 0$ and λ is a finite measure on $(R, \mathscr{B}(R))$ with $\lambda\{0\} = 0$.

3. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables and $S_n = \xi_1 + \cdots + \xi_n$. Suppose that there are constants b_n and $a_n > 0$, and a random variable T, such that

$$\frac{S_n - b_n}{a_n} \xrightarrow{d} T. \tag{3}$$

We ask for a description of the distributions (random variables T) that can be obtained as limit distributions in (3).

If the independent identically distributed random variables $\xi_1, \xi_2, \ldots$ satisfy $0 < \sigma^2 \equiv \mathsf{V}\xi_1 < \infty$, then if we put $b_n = n\mathsf{E}\xi_1$ and $a_n = \sigma\sqrt{n}$, we find by §4 that T is the normal distribution $\mathscr{N}(0, 1)$.

If $f(x) = \theta/\pi(x^2 + \theta^2)$ is the Cauchy density (with parameter $\theta > 0$) and $\xi_1, \xi_2, \ldots$ are independent random variables with density $f(x)$, the characteristic functions $\varphi_{\xi_1}(t)$ are equal to $e^{-\theta|t|}$ and therefore $\varphi_{S_n/n}(t) = (e^{-\theta|t|/n})^n = e^{-\theta|t|}$, i.e., S_n/n also has a Cauchy distribution (with the same parameter θ).

Consequently there are other limit distributions besides the normal, for example the Cauchy distribution.

If we put $\xi_{nk} = (\xi_k/a_n) - (b_n/na_n)$, $1 \leq k \leq n$, we find that

$$\frac{S_n - b_n}{a_n} = \sum_{k=1}^{n} \xi_{n,k} \quad (= T_n).$$

Therefore all conceivable distributions for T that can conceivably appear as limits in (3) are necessarily (in agreement with Theorem 1) infinitely divisible. However, the specific characteristics of the variable $T_n = (S_n - b_n)/a_n$ may make it possible to obtain further information on the structure of the limit distributions that arise.

For this reason we introduce the following definition.

Definition 2. A random variable T, its distribution function $F(x)$, and its characteristic function $\varphi(t)$ are *stable* if, for every $n \geq 1$, there are constants $a_n > 0$, b_n, and independent random variables $\xi_1, \ldots, \xi_n$, distributed like T, such that

$$a_n T + b_n \overset{d}{=} \xi_1 + \cdots + \xi_n \tag{4}$$

or, equivalently, $F[(x - b_n)/a_n] = \underbrace{F * \cdots * F(x)}_{n\text{ times}}$, or

$$[\varphi(t)]^n = [\varphi(a_n t)]e^{ib_n t}. \tag{5}$$

Theorem 3. *A necessary and sufficient condition for the random variable T to be a limit in distribution of random variables $(S_n - b_n)/a_n$, $a_n > 0$, is that T is stable.*

PROOF. If T is stable, then by (4)

$$T \overset{d}{=} \frac{S_n - b_n}{a_n},$$

where $S_n = \xi_1 + \cdots + \xi_n$, and consequently $(S_n - b_n)/a_n \xrightarrow{d} T$.

Conversely, let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables, $S_n = \xi_1 + \cdots + \xi_n$ and $(S_n - b_n)/a_n \to T$, $a_n > 0$. Let us show that T is a stable random variable.

If T is degenerate, it is evidently stable. Let us suppose that T is nondegenerate.

Choose $k \geq 1$ and write

$$S_n^{(1)} = \xi_1 + \cdots + \xi_n, \ldots, \quad S_n^{(k)} = \xi_{(k-1)n+1} + \cdots + \xi_{kn},$$

$$T_n^{(1)} = \frac{S_n^{(1)} - b_n}{a_n}, \ldots, T_n^{(k)} = \frac{S_n^{(k)} - b_n}{a_n}.$$

It is clear that all the variables $T_n^{(1)}, \ldots, T_n^{(k)}$ have the same distribution and

$$T_n^{(i)} \xrightarrow{d} T, \qquad n \to \infty, \quad i = 1, \ldots, k.$$

Write

$$U_n^{(k)} = T_n^{(1)} + \cdots + T_n^{(k)}.$$

Then

$$U_n^{(k)} \xrightarrow{d} T^{(1)} + \cdots + T^{(k)},$$

where $T^{(1)} \overset{d}{=} \cdots \overset{d}{=} T^{(k)} \overset{d}{=} T$.

On the other hand,

$$\begin{aligned} U_n^{(k)} &= \frac{\xi_1 + \cdots + \xi_{kn} - kb_n}{a_n} \\ &= \frac{a_{kn}}{a_n}\left(\frac{\xi_1 + \cdots + \xi_{kn} - b_{kn}}{a_{kn}}\right) + \frac{b_{kn} - kb_n}{a_n} \\ &= \alpha_n^{(k)} V_{kn} + \beta_n^{(k)}, \end{aligned} \tag{6}$$

where

$$\alpha_n^{(k)} = \frac{a_{kn}}{a_n}, \qquad \beta_n^{(k)} = \frac{b_{kn} - kb_n}{a_n}$$

and

$$V_{kn} = \frac{\xi_1 + \cdots + \xi_{kn} - b_{kn}}{a_{kn}}.$$

It is clear from (6) that

$$V_{kn} = \frac{U_n^{(k)} - \beta_n^{(k)}}{\alpha_n^{(k)}},$$

where $V_{kn} \xrightarrow{d} T$, $U_n^{(k)} \xrightarrow{d} T^{(1)} + \cdots + T^{(k)}$, $n \to \infty$.

It follows from the lemma established below that there are constants $\alpha^{(k)} > 0$ and $\beta^{(k)}$ such that $\alpha_n^{(k)} \to \alpha^{(k)}$ and $\beta_n^{(k)} \to \beta^{(k)}$ as $n \to \infty$. Therefore

$$T \overset{d}{=} \frac{T^{(1)} + \cdots + T^{(k)} - \beta^{(k)}}{\alpha^{(k)}},$$

which shows that T is a stable random variable.

This completes the proof of the theorem.

We now state and prove the lemma that we used above.

Lemma. *Let $\xi_n \xrightarrow{d} \xi$ and let there be constants $a_n > 0$ and b_n such that*

$$a_n \xi_n + b_n \xrightarrow{d} \tilde{\xi},$$

where the random variables ξ and $\tilde{\xi}$ are not degenerate. Then there are constants $a > 0$ and b such that $\lim a_n = a$, $\lim b_n = b$, *and*

$$\tilde{\xi} = a\xi + b.$$

PROOF. Let φ_n, φ and $\tilde{\varphi}$ be the characteristic functions of ξ_n, ξ and $\tilde{\xi}$, respectively. Then $\varphi_{a_n\xi_n+b_n}(t)$, the characteristic function of $a_n\xi_n + b_n$, is equal to $e^{itb_n}\varphi_n(a_n t)$ and, by Theorem 1 and Problem 3 of §3,

$$e^{itb_n}\varphi_n(a_n t) \to \tilde{\varphi}(t), \tag{7}$$

$$\varphi_n(t) \to \varphi(t) \tag{8}$$

uniformly on every finite interval of length t.

Let $\{n_i\}$ be a subsequence of $\{n\}$ such that $a_{n_i} \to a$. Let us first show that $a < \infty$. Suppose that $a = \infty$. By (7),

$$\sup_{|t| \le c} \big| |\varphi_n(a_n t)| - |\tilde{\varphi}(t)| \big| \to 0, \qquad n \to \infty$$

for every $c > 0$. We replace t by t_0/a_{n_i}. Then, since $a_{n_i} \to \infty$, we have

$$\left| \varphi_{n_i}\left(a_{n_i} \frac{t_0}{a_{n_i}}\right) \right| - \left| \tilde{\varphi}\left(\frac{t_0}{a_{n_i}}\right) \right| \to 0$$

and therefore

$$|\varphi_{n_i}(t_0)| \to |\tilde{\varphi}(0)| = 1.$$

But $|\varphi_{n_i}(t_0)| \to |\varphi(t_0)|$. Therefore $|\varphi(t_0)| = 1$ for every $t_0 \in R$, and consequently, by Theorem 5, §12, Chapter II, the random variable ξ must be degenerate, which contradicts the hypotheses of the lemma.

Thus $a < \infty$. Now suppose that there are two subsequences $\{n_i\}$ and $\{n_i'\}$ such that $a_{n_i} \to a$, $a_{n_i'} \to a'$, where $a \neq a'$; suppose for definiteness that $0 \le a' < a$. Then by (7) and (8),

$$|\varphi_{n_i}(a_{n_i}t)| \to |\varphi(at)|, \qquad |\varphi_{n_i}(a_{n_i}t)| \to |\tilde{\varphi}(t)|$$

and

$$|\varphi_{n_i'}(a_{n_i'}t)| \to |\varphi(a't)|, \qquad |\varphi_{n_i'}(a_{n_i'}t)| \to |\tilde{\varphi}(t)|.$$

Consequently

$$|\varphi(at)| = |\varphi(a't)|,$$

and therefore, for all $t \in R$,

$$|\varphi(t)| = \left|\varphi\left(\frac{a'}{a}t\right)\right| = \cdots = \left|\varphi\left(\left(\frac{a'}{a}\right)^n t\right)\right| \to 1, \qquad n \to \infty.$$

Therefore $|\varphi(t)| \equiv 1$ and, by Theorem 5 of §12, Chapter II, it follows that ξ is a degenerate random variable. This contradiction shows that $a = a'$ and therefore that there is a finite limit $\lim a_n = a$, with $a \ge 0$.

Let us now show that there is a limit $\lim b_n = b$, and that $a > 0$. Since (8) is satisfied uniformly on each finite interval, we have

$$\varphi_n(a_n t) \to \varphi(at),$$

and therefore, by (7), the limit $\lim_{n\to\infty} e^{itb_n}$ exists for all t such that $\varphi(at) \neq 0$. Let $\delta > 0$ be such that $\varphi(at) \neq 0$ for all $|t| < \delta$. For such t, $\lim e^{itb_n}$ exists. Hence we can deduce (Problem 9) that $\overline{\lim}\, |b_n| < \infty$.

Let there be two sequences $\{n_i\}$ and $\{n_i'\}$ such that $\lim b_{n_i} = b$ and $\lim b_{n_i'} = b'$. Then

$$e^{itb} = e^{itb'},$$

for $|t| < \delta$, and consequently $b = b'$. Thus there is a finite limit $b = \lim b_n$ and, by (7),

$$\tilde{\varphi}(t) = e^{itb}\varphi(at),$$

which means that $\tilde{\xi} \overset{d}{=} a\xi + b$. Since $\tilde{\xi}$ is not degenerate, we have $a > 0$.

This completes the proof of the lemma.

4. We quote without proof a theorem on the general form of the characteristic functions of stable distributions.

Theorem 4 (Lévy–Khinchin Representation). *A random variable T is stable if and only if its characteristic function $\varphi(t)$ has the form $\varphi(t) = \exp \psi(t)$,*

$$\psi(t) = it\beta - d|t|^{\alpha}\left(1 + i\theta \frac{t}{|t|} G(t, \alpha)\right), \tag{9}$$

where $0 < \alpha < 2$, $\beta \in R$, $d \geq 0$, $|\theta| \leq 1$, $t/|t| = 0$ *for* $t = 0$, *and*

$$G(t, \alpha) = \begin{cases} \tan \frac{1}{2}\pi\alpha & \text{if } \alpha \neq 1, \\ (2/\pi) \log |t| & \text{if } \alpha = 1. \end{cases} \tag{10}$$

Observe that it is easy to exhibit characteristic functions of symmetric stable distributions:

$$\varphi(t) = e^{-d|t|^{\alpha}}, \tag{11}$$

where $0 < \alpha \leq 2$, $d \geq 0$.

5. Problems

1. Show that $\xi \overset{d}{=} \eta$ if $\xi_n \xrightarrow{d} \xi$ and $\xi_n \xrightarrow{d} \eta$.
2. Show that if φ_1 and φ_2 are infinitely divisible characteristic functions, so is $\varphi_1 \cdot \varphi_2$.
3. Let φ_n be infinitely divisible characteristic functions and let $\varphi_n(t) \to \varphi(t)$ for every $t \in R$, where $\varphi(t)$ is a characteristic function. Show that $\varphi(t)$ is infinitely divisible.
4. Show that the characteristic function of an infinitely divisible distribution cannot take the value 0.
5. Give an example of a random variable that is infinitely divisible but not stable.
6. Show that a stable random variable ξ always satisfies the inequality $\mathsf{E}|\xi|^r < \infty$ for all $r \in (0, \alpha)$.
7. Show that if ξ is a stable random variable with parameter $0 < \alpha \leq 1$, then $\varphi(t)$ is not differentiable at $t = 0$.
8. Prove that $e^{-d|t|^{\alpha}}$ is a characteristic function provided that $d \geq 0$, $0 < \alpha \leq 2$.
9. Let $(b_n)_{n \geq 1}$ be a sequence of numbers such that $\lim_n e^{itb_n}$ exists for all $|t| < \delta$, $\delta > 0$. Show that $\overline{\lim} |b_n| < \infty$.

§7. Metrizability of Weak Convergence

1. Let $(E, \mathscr{E}, \rho)$ be a metric space and $\mathscr{P}(E) = \{P\}$, a family of probability measures on $(E, \mathscr{E})$. It is natural to raise the question of whether it is possible to "metrize" the weak convergence $P_n \xrightarrow{w} P$ that was introduced in §1, that is, whether it is possible to introduce a distance $\mu(P, \tilde{P})$ between any two measures P and $\tilde{P}$ in $\mathscr{P}(E)$ in such a way that the limit $\mu(P_n, P) \to 0$ is equivalent to the limit $P_n \xrightarrow{w} P$.

In connection with this formulation of the problem, it is useful to recall that *convergence* of random variables *in probability*, $\xi_n \xrightarrow{P} \xi$, can be metrized by using, for example, the distance $d_{\mathsf{P}}(\xi, \eta) = \inf\{\varepsilon > 0: \mathsf{P}(|\xi - \eta| \geq \varepsilon) \leq \varepsilon\}$ or the distances $d(\xi, \eta) = \mathsf{E}(|\xi - \eta|/(1 + |\xi - \eta|))$, $d(\xi, \eta) = \mathsf{E}\min(1, |\xi - \eta|)$. (More generally, we can set $d(\xi, \eta) = \mathsf{E}g(|\xi - \eta|)$, where the function $g = g(x)$, $x \geq 0$, can be chosen as any nonnegative increasing Borel function that is continuous at zero and has the properties $g(x + y) \leq g(x) + g(y)$ for $x \geq 0$, $y \geq 0$, $g(0) = 0$, and $g(x) > 0$ for $x > 0$.) However, at the same time there is, in the space of random variables over $(\Omega, \mathscr{F}, \mathsf{P})$, no distance $d(\xi, \eta)$ such that $d(\xi_n, \xi) \to 0$ if and only if ξ_n converges to ξ *with probability one*. (In this connection, it is easy to find a sequence of random variables ξ_n, $n \geq 1$, that converges to ξ in probability but does not converge with probability one.) In other words, *convergence with probability one is not metrizable*. (See the statements of problems 11 and 12 in §10, Chapter II.)

The aim of this section is to obtain concrete instances of two metrics, $L(P, \tilde{P})$ and $\|P - \tilde{P}\|^*_{BL}$ in the space $\mathscr{P}(E)$ of measures, that metrize weak convergence:

$$P_n \xrightarrow{w} P \Leftrightarrow L(P_n, P) \to 0 \Leftrightarrow \|P_n - P\|^*_{BL} \to 0.$$

2. The Lévy-Prokhorov metric $L(P, \tilde{P})$. Let

$$\rho(x, A) = \inf\{\rho(x, y): y \in A\},$$

$$A^\varepsilon = \{x \in E: \rho(x, A) < \varepsilon\}, \qquad A \in \mathscr{E}.$$

For any two measures P and $\tilde{P} \in \mathscr{P}(E)$, we set

$$\sigma(P, \tilde{P}) = \inf\{\varepsilon > 0: P(F) \leq \tilde{P}(F^\varepsilon) + \varepsilon \text{ for all closed sets } F \in \mathscr{E}\} \tag{2}$$

and

$$L(P, \tilde{P}) = \max[\sigma(P, \tilde{P}), \sigma(\tilde{P}, P)]. \tag{3}$$

The following lemma shows that the function $L(P, \tilde{P}) \in \mathscr{P}(E)$, which is defined in this way, and is called the *Lévy–Prokhorov metric*, actually defines a metric.

Lemma 1. *The function $L(P, \tilde{P})$ has the following properties:*

(a) $L(P, \tilde{P}) = L(\tilde{P}, P)(= \sigma(P, \tilde{P}) = \sigma(\tilde{P}, P))$,
(b) $L(P, \tilde{P}) \leq L(P, \hat{P}) + L(\hat{P}, \tilde{P})$,
(c) $L(P, \tilde{P}) = 0$ *if and only if* $\tilde{P} = P$.

Proof. a) It is sufficient to show that (with $\alpha > 0$ and $\beta > 0$)

$$\text{“}P(F) \leq \tilde{P}(F^\alpha) + \beta \qquad \text{for all closed sets} \quad F \in \mathscr{E}\text{”} \tag{4}$$

if and only if

$$\text{“}\tilde{P}(F) \leq P(F^\alpha) + \beta \qquad \text{for all closed sets} \quad F \in \mathscr{E}\text{”}. \tag{5}$$

Let T be a closed subset of $\mathscr{E}$. Then the set T^α is open and it is easy to verify that $T \subseteq E \backslash (E \backslash T^\alpha)^\alpha$. If (4) is satisfied, then, in particular,

$$P(E\setminus T^{\alpha}) \le \tilde{P}((E\setminus T^{\alpha})^{\alpha}) + \beta$$

and therefore,

$$\tilde{P}(T) \le \tilde{P}(E\setminus(E\setminus T^{\alpha})^{\alpha}) \le P(T^{\alpha}) + \beta,$$

which establishes the equivalence of (4) and (5). Hence, it follows that

$$\sigma(P, \tilde{P}) = \sigma(\tilde{P}, P) \tag{6}$$

and therefore,

$$L(P, \tilde{P}) = \sigma(P, \tilde{P}) = \sigma(\tilde{P}, P) = L(\tilde{P}, P). \tag{7}$$

b) Let $L(P, \hat{P}) < \delta_1$ and $L(\hat{P}, \tilde{P}) < \delta_2$. Then for each closed set $F \in \mathscr{E}$

$$\tilde{P}(F) \le \hat{P}(F^{\delta_2}) + \delta_2 \le P((F^{\delta_2})^{\delta_1}) + \delta_1 + \delta_2 \le P(F^{\delta_1+\delta_2}) + \delta_1 + \delta_2$$

and therefore, $L(P, \tilde{P}) \le \delta_1 + \delta_2$. Hence, it follows that

$$L(P, \tilde{P}) \le L(P, \hat{P}) + L(\hat{P}, \tilde{P}).$$

c) If $L(P, \tilde{P}) = 0$, then for every closed set $F \in \mathscr{E}$ and every $\alpha > 0$

$$P(F) \le \tilde{P}(F^{\alpha}) + \alpha. \tag{8}$$

Since $F^{\alpha} \downarrow F$, $\alpha \downarrow 0$, we find, by taking the limit in (8) as $\alpha \downarrow 0$, that $P(F) \le \tilde{P}(F)$ and by symmetry $\tilde{P}(F) \le P(F)$. Hence, $P(F) = \tilde{P}(F)$ for all closed sets $F \in \mathscr{E}$. For each Borel set $A \in \mathscr{E}$ and every $\varepsilon > 0$, there is an open set $G_{\varepsilon} \supseteq A$ and a closed set $F_{\varepsilon} \subseteq A$ such that $P(G_{\varepsilon}\setminus F_{\varepsilon}) \le \varepsilon$. Hence, it follows that every probability measure P on a metric space $(E, \mathscr{E}, \rho)$ is completely determined by its values on closed sets. Consequently, it follows from the condition $\tilde{P}(F) = P(F)$ for all closed sets $F \in \mathscr{E}$ that $\tilde{P}(A) = P(A)$ for all Borel sets $A \in \mathscr{E}$.

Theorem 1. *The Lévy–Prokhorov metric $L(P, \tilde{P})$ metrizes weak convergence*:

$$L(P_n, P) \to 0 \Leftrightarrow P_n \xrightarrow{w} P. \tag{9}$$

Proof. ($\Rightarrow$) Let $L(P_n, P) \to 0$, $n \to \infty$. Then for every specified closed set $F \in \mathscr{E}$ and every $\varepsilon > 0$, we have, by (2) and equation a) of Lemma 1,

$$\overline{\lim_n}\, P_n(F) \le P(F^{\varepsilon}) + \varepsilon. \tag{10}$$

If we then let $\varepsilon \downarrow 0$, we find that

$$\overline{\lim_n}\, P_n(F) \le P(F).$$

According to Theorem 1 of §1, it follows that

$$P_n \xrightarrow{w} P. \tag{11}$$

The proof of the implication ($\Leftarrow$) will be based on a series of deep and powerful facts that illuminate the content of the concept of weak convergence and the method of establishing it, as well as methods of studying rates of convergence.

Thus, let $P_n \xrightarrow{w} P$. This means that for every bounded continuous function $f = f(x)$,

$$\int_E f(x)P_n(dx) \to \int_E f(x)P(dx). \tag{12}$$

Now suppose that $\mathscr{G}$ is a class of equicontinuous functions $g = g(x)$ (for every $\varepsilon > 0$ there is a $\delta > 0$ such that $|g(y) - g(x)| < \varepsilon$ if $\rho(x, y) < \delta$ for all $g \in \mathscr{G}$) and $|g(x)| \leq C$ for the same constant $C > 0$ (for all $x \in E$ and $g \in \mathscr{G}$). By Theorem 3, §8, the following condition, stronger than (12), is valid for $\mathscr{G}$:

$$P_n \xrightarrow{w} P \Rightarrow \sup_{g \in \mathscr{G}} \left| \int_E g(x)P_n(dx) - \int_E g(x)P(dx) \right| \to 0. \tag{13}$$

For each $A \in \mathscr{E}$ and $\varepsilon > 0$, we set (as in Theorem 1, §1)

$$f_A^\varepsilon(x) = \left[1 - \frac{\rho(x, A)}{\varepsilon}\right]^+. \tag{14}$$

It is clear that

$$I_A(x) \leq f_A^\varepsilon(x) \leq I_{A^\varepsilon}(x) \tag{15}$$

and

$$|f_A^\varepsilon(x) - f_A^\varepsilon(y)| \leq \varepsilon^{-1}|\rho(x, A) - \rho(y, A)| \leq \varepsilon^{-1}\rho(x, y).$$

Therefore, we have (13) for the class $\mathscr{G}^\varepsilon = \{f_A^\varepsilon(x), A \in \mathscr{E}\}$, i.e.,

$$\Delta_n \equiv \sup_{A \in \mathscr{E}} \left| \int_E f_A^\varepsilon(x)P_n(dx) - \int_E f_A^\varepsilon(x)P(dx) \right| \to 0, \qquad n \to \infty. \tag{16}$$

From this and (15) we conclude that, for every closed set $A \in \mathscr{E}$ and $\varepsilon > 0$,

$$P(A^\varepsilon) \geq \int_E f_A^\varepsilon(x)\, dP \geq \int_E f_A^\varepsilon(x)\, dP_n - \Delta_n \geq P_n(A) - \Delta_n. \tag{17}$$

We choose $n(\varepsilon)$ so that $\Delta_n \leq \varepsilon$ for $n \geq n(\varepsilon)$. Then, by (17), for $n \geq n(\varepsilon)$

$$P(A^\varepsilon) \geq P_n(A) - \varepsilon. \tag{18}$$

Hence, it follows from definitions (2) and (3) that $L(P_n, P) \leq \varepsilon$ as soon as $n \geq n(\varepsilon)$. Consequently,

$$P_n \xrightarrow{w} P \Rightarrow \Delta_n \to 0 \Rightarrow L(P_n, P) \to 0.$$

The theorem is now proved (up to (13)).

3. The metric $\|P - \tilde{P}\|_{BL}^*$. We denote by BL the set of bounded continuous functions $f = f(x)$, $x \in E$ (with $\|f\|_\infty = \sup_x |f(x)| < \infty$) that also satisfy the Lipschitz condition

$$\|f\|_L = \sup_{x \neq y} \frac{|f(x) - f(y)|}{\rho(x, y)} < \infty.$$

We set $\|f\|_{BL} = \|f\|_\infty + \|f\|_L$. The space BL with the norm $\|\cdot\|_{BL}$ is a Banach space.

We define the metric $\|P - \tilde{P}\|_{BL}^*$ by setting

$$\|P - \tilde{P}\|_{BL}^* = \sup_{f \in BL} \left\{ \left| \int f d(P - \tilde{P}) \right| : \|f\|_{BL} \leq 1 \right\}. \tag{19}$$

(We can verify that $\|P - \tilde{P}\|_{BL}^*$ actually satisfies the conditions for a matric; Problem 2.)

Theorem 2. *The metric* $\|P - \tilde{P}\|_{BL}^*$ *metrizes weak convergence:*

$$\|P_n - P\|_{BL}^* \to 0 \Leftrightarrow P_n \xrightarrow{w} P.$$

PROOF. The implication ($\Leftarrow$) follows directly from (13). To prove ($\Rightarrow$), it is enough to show that in the definition of weak convergence $P_n \xrightarrow{w} P$ as given by (12) for every continuous bounded function $f = f(x)$, it is enough to restrict consideration to the class of bounded functions that satisfy a Lipschitz condition. In other words, the implication ($\Rightarrow$) will be proved if we establish the following result.

Lemma 2. *Weak convergence* $P_n \xrightarrow{w} P$ *occurs if and only if property* (12) *is satisfied for every function* $f = f(x)$ *of class* BL.

PROOF. The proof is obvious in one direction. Let us now consider the functions $f_A^\varepsilon = f_A^\varepsilon(x)$ defined in (14). As was established above in the proof of Theorem 1, for each $\varepsilon > 0$ the class $\mathscr{G}^\varepsilon = \{f_A^\varepsilon(x), A \in \mathscr{G}\} \subseteq BL$. If we now analyze the proof of the implication (I) $\Rightarrow$ (II) in Theorem 1 of §1, we can observe that it actually establishes property (12) *not for all* bounded continuous functions but only for functions of class $\mathscr{G}^\varepsilon$, $\varepsilon > 0$. Since $\mathscr{G}^\varepsilon \subseteq BL$, $\varepsilon > 0$, it is evidently true that the satisfaction of (12) for functions of class BL implies proposition II of Theorem 1, §1, which is equivalent (by the same Theorem 1, §1) to the weak convergence $P_n \xrightarrow{w} P$.

Remark. The conclusion of Theorem 2 can be derived from Theorem 1 (the same as before) if we use the following inequalities between the metrics $L(P, \tilde{P})$ and $\|P - \tilde{P}\|_{BL}^*$, which are valid for the separable metric spaces $(E, \mathscr{E}, \rho)$:

$$\|P - \tilde{P}\|_{BL}^* \leq 2L(P, \tilde{P}), \tag{20}$$

$$\varphi(L(P, \tilde{P})) \leq \|P - \tilde{P}\|_{BL}^*, \tag{21}$$

where $\varphi(x) = 2x^2/(2 + x)$.

We notice that, for $x \geq 0$, we have $0 \leq \varphi \leq 2/3$ if and only if $x \leq 1$; and $(2/3)x^2 \leq \varphi(x)$ for $0 \leq x \leq 1$; we deduce from (20) and (21) that if $L(P, \tilde{P}) \leq 1$ or $\|P - \tilde{P}\|_{BL}^* \leq 2/3$, we have

$$\tfrac{2}{3}L^2(P, \tilde{P}) \leq \|P - \tilde{P}\|_{BL}^* \leq 2L(P, \tilde{P}). \tag{22}$$

4. Problems

1. Show that in case $E = R$ the Lévy–Prokhorov metric between the probability distributions P and $\tilde{P}$ becomes the *Lévy distance* $L(F, \tilde{F})$ between the distributions F and $\tilde{F}$ that correspond to P and $\tilde{P}$ (see Problem 4 in §1).
2. Show that formula (19) defines a metric on the space BL.
3. Establish the inequalities (20), (21), and (22).

§8. On the Connection of Weak Convergence of Measures with Almost Sure Convergence of Random Elements ("Method of a Single Probability Space")

1. Let us suppose that on the probability space $(\Omega, \mathscr{F}, \mathsf{P})$ there are given random elements $X = X(\omega)$, $X_n = X_n(\omega)$, $n \geq 1$, taking values in the metric space $(E, \mathscr{E}, \rho)$; see §5, Chapter II. We denote by P and P_n the probability distributions of X and X_n, i.e., let

$$P(A) = \mathsf{P}\{\omega : X(\omega) \in A\}, \qquad P_n(A) = \mathsf{P}\{\omega : X_n(\omega) \in A\}, \qquad A \in \mathscr{E}.$$

Generalizing the concept of convergence in distribution of random variables (see §10, chapter II), we introduce the following definition.

Definition 1. *A sequence of random elements X_n, $n \geq 1$, is said to converge in distribution, or in law (notation: $X_n \xrightarrow{\mathscr{D}} X$, or $X_n \xrightarrow{\mathscr{L}} X$), if $P_n \xrightarrow{w} P$.*

By analogy with the definitions of convergence of random variables in probability or with probability one (§10, Chapter II), it is natural to introduce the following definitions.

Definition 2. A sequence of random elements X_n, $n \geq 1$, is said to converge *in probability* to X if

$$\mathsf{P}\{\omega : \rho(X_n(\omega), X(\omega)) \geq \varepsilon\} \to 0, \qquad n \to \infty. \tag{1}$$

Definition 3. A sequence of random elements X_n, $n \geq 1$, is said to converge to X *with probability one* (*almost surely, almost everywhere*) if $\rho(X_n(\omega), X(\omega)) \xrightarrow{\text{a.s.}} 0$, $n \to \infty$.

Remark 1. Both of the preceding definitions make sense, of course, provided that $\rho(X_n(\omega), X(\omega))$ are, as functions of $\omega \in \Omega$, random variables, i.e., $\mathscr{F}$-measurable functions. This will certainly be the case if the space $(E, \mathscr{E}, \rho)$ is separable (Problem 1).

Remark 2. In connection with Definition 2, we note that our convergence in probability is metrized by the following metric that connects random elements X and Y (defined on $(\Omega, \mathscr{F}, \mathsf{P})$ with values in E):

$$d_{\mathsf{P}}(X, Y) = \inf\{\varepsilon > 0: \mathsf{P}\{\rho(X(\omega), Y(\omega)) \geq \varepsilon\} \leq \varepsilon\}. \tag{2}$$

Remark 3. If the definitions of convergence in probability and with probability one are defined for random elements on the same probability space, the definition $X_n \xrightarrow{\mathscr{D}} X$ of convergence in distribution is connected only with the convergence of distributions, and consequently, we may suppose that $X(\omega)$, $X_1(\omega)$, $X_2(\omega)$, ... have values in the same space E, but may be defined on "their own" probability spaces $(\Omega, \mathscr{F}, \mathsf{P})$, $(\Omega_1, \mathscr{F}_1, \mathsf{P}_1)$, $(\Omega_2, \mathscr{F}_2, \mathsf{P}_2)$, ... However, without loss of generality we may always suppose that they are defined on the same probability space, taken as the direct product of the preceding spaces and with the definitions $X(\omega, \omega_1, \omega_2, \ldots) = X(\omega)$, $X_1(\omega, \omega_1, \omega_2, \ldots) = X_1(\omega_1), \ldots$.

2. By Definition 1 and the theorem on change of variables under the Lebesgue integral sign (Theorem 7, §6, Chapter II)

$$X_n \xrightarrow{\mathscr{D}} X \Leftrightarrow \mathsf{E}f(X_n) \to \mathsf{E}f(X) \tag{3}$$

for every bounded continuous function $f = f(x)$, $x \in E$.

From (3) it is clear that, by Lebesgue's theorem on dominated convergence (Theorem 3, §6, Chapter II), the limit $X_n \xrightarrow{\text{a.s.}} X$ immediately implies the limit $X_n \xrightarrow{\mathscr{D}} X$, which is hardly surprising if we think of the situation when X and X_n are random variables (Theorem 2, §10, Chapter II). More unexpectedly, in a certain sense there is a converse result, the precise formulation and application we now turn to.

Preliminarily, we introduce a definition.

Definition 4. Random elements $X = X(\omega')$ and $Y = Y(\omega'')$, defined on probability spaces $(\Omega', \mathscr{F}', \mathsf{P}')$ and $(\Omega'', \mathscr{F}'', \mathsf{P}'')$ and with values in the same space E, are said to be *equivalent in distribution* (notation: $X \stackrel{\mathscr{D}}{=} Y$), if they have congruent probability distributions.

Theorem 1. *Let $(E, \mathscr{E}, \rho)$ be a separable metric space.*

1. *Let random elements X, X_n, $n \geq 1$, defined on a probability space $(\Omega, \mathscr{F}, \mathsf{P})$, and with values in E, have the property that $X_n \xrightarrow{\mathscr{D}} X$. Then we can find a probability space $(\Omega^*, \mathscr{F}^*, \mathsf{P}^*)$ and random elements X^*, X_n^*, $n \geq 1$, defined on it, with values in E, such that*

$$X_n^* \xrightarrow{\text{a.s.}} X^*$$

and

$$X^* \stackrel{\mathscr{D}}{=} X, \qquad X_n^* \stackrel{\mathscr{D}}{=} X_n, \qquad n \geq 1.$$

2. *Let P, P_n, $n \geq 1$, be probability measures on $(E, \mathscr{E}, \rho)$. Then there is a probability space $(\Omega^*, \mathscr{F}^*, \mathsf{P}^*)$ and random elements X^*, X_n^*, $n \geq 1$, defined on it, with values in E, such that*

$$X_n^* \xrightarrow{\text{a.s.}} X^*$$

and

$$P^* = P, \qquad P_n^* = P_n, \qquad n \geq 1,$$

where P^* and P_n^* are the probability distributions of X^* and X_n^*.

Before turning to the proof, we first notice that it is enough to prove only the second conclusion, since the first follows from it if we take P and P_n to be the distributions of X and X_n. Similarly, the second conclusion follows from the first. Second, we notice that a proof of the theorem in full generality is technically rather complicated. For this reason, here we give a proof only of the case $E = R$. This proof is rather transparent and moreover, provides a simple, clear construction of the required objectives. (Unfortunately, this construction does not work in the general case, even for $E = R^2$.)

PROOF OF THE THEOREM IN THE CASE $E = R$. Let $F = F(x)$ and $F_n = F_n(x)$ be distribution functions corresponding to the measures P and P_n on $(R, \mathscr{B}(R))$. We associate with a function $F = F(x)$ its corresponding *quantile function* $Q = Q(u)$, uniquely defined by the formula

$$Q(u) = \inf\{x : F(x) \geq u\}, \qquad 0 < u < 1. \tag{4}$$

It is easily verified that

$$F(x) \geq u \Leftrightarrow Q(u) \leq x. \tag{5}$$

We now take $\Omega^* = (0, 1)$, $\mathscr{F}^* = \mathscr{B}(0, 1)$, P^* to be Lebesgue measure, and $P^*(dx) = dx$. We also take $X^*(\omega^*) = Q(\omega^*)$ and $\omega^* \in \Omega^*$. Then

$$\mathsf{P}^*\{\omega^* : X^*(\omega^*) \leq x\} = \mathsf{P}^*\{\omega^* : Q(\omega^*) \leq x\} = \mathsf{P}^*\{\omega^* : \omega^* \leq F(x)\} = F(x),$$

i.e., the distribution of the random variable $X^*(\omega^*) = Q(\omega^*)$ coincides exactly with P. Similarly, the distribution of $X_n^*(\omega^*) = Q_n(\omega^*)$ coincides with P_n.

In addition, it is not difficult to show that the convergence of $F_n(x)$ to $F(x)$ at each point of continuity of the limit function $F = F(x)$ (equivalent, if $E = R$, to the limit $P_n \xrightarrow{w} P$; see Theorem 1 in §1) implies that the sequence of quantiles $Q_n(u)$, $n \geq 1$, also converges to $Q(u)$ at every point of continuity of the limit $Q = Q(u)$. Since the set of points of discontinuity of $Q = Q(u)$, $u \in (0, 1)$, is at most countable, its Lebesgue measure P^* is zero and therefore,

$$X_n^*(\omega^*) = Q_n(\omega^*) \xrightarrow{\text{a.s.}} X^*(\omega^*) = Q(\omega^*).$$

The theorem is established in the case of $E = R$.

This construction in Theorem 1 of a passage from given random elements X and X_n to new elements X^* and X_n^*, defined on the same probability space, explains the announcement in the heading of this section of the *method of a single probability space.*

We now turn to a number of propositions that are established very simply by using this method.

3. Let us assume that the random elements X and X_n, $n \geq 1$, are defined, for example, on a probability space $(\Omega, \mathscr{F}, \mathsf{P})$ with values in a separable metric space $(E, \mathscr{E}, \rho)$, so that $X_n \xrightarrow{\mathscr{D}} X$. Also let $h = h(x)$, $x \in E$, be a measurable mapping of $(E, \mathscr{E}, \rho)$ into another separable metric space $(E', \mathscr{E}', \rho')$. In probability and mathematical statistics it is often necessary to deal with the search for conditions under which we can say of $h = h(x)$ that the limit $X_n \xrightarrow{\mathscr{D}} X$ implies the limit $h(X_n) \xrightarrow{\mathscr{D}} h(X)$.

For example, let $\xi_1, \xi_2, \ldots$ be independent identically distributed random variables with $\mathsf{E}\xi_1 = m$, $\mathsf{V}\xi_1 = \sigma^2 > 0$. Let $\overline{X}_n = (\xi_1 + \cdots + \xi_n)/n$. The central limit theorem shows that

$$\frac{\sqrt{n}(\overline{X}_n - m)}{\sigma} \xrightarrow{d} \mathscr{N}(0, 1).$$

Let us ask, for what functions $h = h(x)$ can we guarantee that

$$h\left(\frac{\sqrt{n}(\overline{X}_n - m)}{\sigma}\right) \xrightarrow{d} h(\mathscr{N}(0, 1))?$$

(The *Mann–Wald theorem*, which is applicable to the present case, since it is satisfied for continuous functions $h = h(x)$, says that $n(\overline{X} - m)^2/\sigma^2 \xrightarrow{d} \chi_1^2$, where χ_1^2 is a random variable with a chi-squared distribution with one degree of freedom; see Table 2 in §3, Chapter I.)

A second example. If $X = X(t, \omega)$, $X_n = X_n(t, \omega)$, $t \in T$, are random processes (see §5, Chapter II) and $h(X) = \sup_{t \in T} |X(t, \omega)|$, $h(X_n) = \sup_{t \in T} |X_n(t, \omega)|$, our problem amounts to asking under what conditions on the convergence in distribution of the processes $X_n \xrightarrow{\mathscr{D}} X$ will there follow the convergence in distribution of their suprema, $h(X_n) \xrightarrow{\mathscr{D}} h(X)$.

A simple condition that guarantees the validity of the implication

$$X_n \xrightarrow{\mathscr{D}} X \Rightarrow h(X_n) \xrightarrow{\mathscr{D}} h(X),$$

is that the mapping $h = h(x)$ is continuous. In fact, if $f = f(x')$ *is a bounded continuous* function on E', the function $f(h(x))$ will also be a bounded continuous function on E. Consequently,

$$X_n \xrightarrow{\mathscr{D}} X \Rightarrow Ef(h(X_n)) \to Ef(h(X)).$$

The theorem given below shows that in fact the requirement of continuity of the function $h = h(x)$ can be somewhat weakened by using the limit properties of the random element X.

We denote by Δ_h the set $\{x \in E: h(x) \text{ is not } \rho\text{-continuous at } x\}$; i.e., let Δ_h be the set of points of discontinuity of the function $h = h(x)$. We note that $\Delta_h \in \mathscr{E}$ (problem 4).

Theorem 2.1. *Let $(E, \mathscr{E}, \rho)$ and $(E', \mathscr{E}', \rho')$ be separable metric spaces, and $X_n \xrightarrow{\mathscr{D}} X$. Let the mapping $h = h(x)$, $x \in E$, have the property that*

$$\mathsf{P}\{\omega: X(\omega) \in \Delta_h\} = 0. \tag{6}$$

Then $h(X_n) \xrightarrow{\mathscr{D}} h(X)$.

2. *Let P, P_n, $n \geq 1$, be probability distributions on the separable metric space $(E, \mathscr{E}, \rho)$ and $h = h(x)$ a measurable mapping of $(E, \mathscr{E}, \rho)$ on a separable metric space $(E', \mathscr{E}', \rho')$. Let*

$$P\{x: x \in \Delta_h\} = 0.$$

Then $P_n^h \xrightarrow{w} P^h$, where $P_n^h(A) = P_n\{h(x) \in A\}$, $P^h(A) = P\{h(x) \in A\}$, $A \in \mathscr{E}'$.

Proof. As in Theorem 1, it is enough to prove the validity of, for example, the first proposition.

Let X^* and X_n^*, $n \geq 1$, be random elements constructed by the "method of a single probability space," so that $X^* \stackrel{\mathscr{D}}{=} X$, $X_n^* \stackrel{\mathscr{D}}{=} X_n$, $n \geq 1$, and $X_n^* \xrightarrow{\text{a.s.}} X^*$. Let $A^* = \{\omega^*; \rho(X_n^*, X^*) \not\to 0\}$, $B^* = \{\omega^*: X^*(\omega^*) \in \Delta_n\}$. Then $\mathsf{P}^*(A^* \cup B^*) = 0$, and for $w^* \notin A^* \cup B^*$

$$h(X_n^*(\omega^*)) \to h(X^*(\omega^*)),$$

which implies that $h(X_n^*) \xrightarrow{\text{a.s.}} h(X^*)$. As we noticed in subsection 1, it follows that $h(X_n^*) \xrightarrow{\mathscr{D}} h(X^*)$. But $h(X_n^*) \stackrel{\mathscr{D}}{=} h(X_n)$ and $h(X^*) \stackrel{\mathscr{D}}{=} h(X)$. Therefore, $h(X_n^*) \xrightarrow{\mathscr{D}} h(X)$.

This completes the proof of the theorem.

4. In §6, in the proof of the implication ($\Leftarrow$) in Theorem 1, we used (13). We now give a proof that again relies on the "method of a single probability space."

Let $(E, \mathscr{E}, \rho)$ be a separable metric space, and $\mathscr{G}$ a class of equicontinuous functions $g = g(x)$ for which also $|g(x)| \leq C$ for all $x \in E$ and $g \in \mathscr{G}$.

Theorem 3. *Let P and P_n, $n \geq 1$, be probability measures on $(E, \mathscr{E}, \rho)$ for which $P_n \xrightarrow{w} P$. Then*

$$\sup_{g \in \mathscr{G}} \left| \int_E g(x) P_n(dx) - \int_E g(x) P(dx) \right| \to 0, \qquad n \to \infty. \tag{7}$$

Proof. Let (7) not occur. Then there are an $a > 0$ and functions $g_1, g_2 \ldots$ from $\mathscr{G}$ such that

$$\left| \int_E g_n(x) P_n(dx) - \int_E g(x) P(dx) \right| \geq a > 0 \tag{8}$$

for infinitely many values of n. Turning by the "method of a single probability space" to random elements X^* and X_n^* (see Theorem 1), we transform (8) to the form

$$|\mathsf{E}^* g_n(X_n^*) - \mathsf{E}^* g_n(X^*)| \geq a > 0 \tag{9}$$

for infinitely many values of n. But, by the properties of $\mathscr{G}$, for every $\varepsilon > 0$ there is a $\delta > 0$ for which $|g(y) - g(x)| < \varepsilon$ for all $g \in \mathscr{G}$, if $\rho(x, y) < \delta$. In addition, $|g(x)| \leq C$ for all $x \in E$ and $g \in \mathscr{G}$. Therefore,

$$\begin{aligned}|\mathsf{E}^* g_n(X_n^*) - \mathsf{E}^* g_n(X^*)| &\leq \mathsf{E}^*\{|g_n(X_n^*) - g_n(X^*)|; \rho(X_n^*, X^*) > \delta\} \\ &\quad + \mathsf{E}^*\{|g_n(X_n^*) - g_n(X^*)|; \rho(X_n^*, X^*) \leq \delta\} \\ &\leq 2C\mathsf{P}\{\rho(X_n^*, X^*) > \delta\} + \varepsilon.\end{aligned}$$

Since $X_n^* \xrightarrow{\text{a.s.}} X^*$, we have $\mathsf{P}^*\{\rho(X_n^*, X^*) > \delta\} \to 0$ as $n \to \infty$. Consequently, since $\varepsilon > 0$ is arbitrary,

$$\overline{\lim_n}\, |\mathsf{E}^* g_n(X_n^*) - \mathsf{E}^* g_n(X^*)| = 0,$$

which contradicts (9).

This completes the proof of the theorem.

5. In this section the idea of the "method of a single probability space" used in Theorem 1 will be applied to estimating upper bounds of the Lévy–Prokhorov metric $L(P, \tilde{P})$ between two probability distributions on a separable space $(E, \mathscr{E}, \rho)$.

Theorem 4. *For each pair P, $\tilde{P}$ of measures we can find a probability space $(\Omega^*, \mathscr{F}^*, \mathsf{P}^*)$ and random elements X and $\tilde{X}$ on it with values in E such that their distributions coincide respectively with P and $\tilde{P}$ and*

$$L(P, \tilde{P}) \leq d_{\mathsf{P}^*}(X, \tilde{X}) = \inf\{\varepsilon > 0\colon \mathsf{P}^*(\rho(X, \tilde{X}) \geq \varepsilon) \leq \varepsilon\}. \tag{10}$$

Proof. By Theorem 1, we can actually find a probability space $(\Omega^*, \mathscr{F}^*, \mathsf{P}^*)$ and random elements X and $\tilde{X}$ such that $\mathsf{P}^*(X \in A) = P(A)$ and $\mathsf{P}^*(\tilde{X} \in A) = \tilde{P}(A)$, $A \in \mathscr{E}$.

Let $\varepsilon > 0$ have the property that

$$\mathsf{P}^*(\rho(X, \tilde{X}) \geq \varepsilon) \leq \varepsilon. \tag{11}$$

Then for every $A \in \mathscr{E}$

$$\begin{aligned}\tilde{P}(A) = \mathsf{P}^*(\tilde{X} \in A) &= \mathsf{P}^*(\tilde{X} \in A, X \in A^\varepsilon) + \mathsf{P}^*(\tilde{X} \in A, X \notin A^\varepsilon) \\ &\leq \mathsf{P}^*(X \in A^\varepsilon) + \mathsf{P}^*(\rho(X, \tilde{X}) \geq \varepsilon) \leq P(A^\varepsilon) + \varepsilon.\end{aligned}$$

Hence, by the definition of the Lévy–Prokhorov metric (subsection 2, §6)

$$L(P, \tilde{P}) \leq \varepsilon. \tag{12}$$

From (11) and (12), if we take the infimum for $\varepsilon > 0$ we obtain the required assertion (10).

Corollary. *Let X and $\tilde{X}$ be random elements defined on a probability space $(\Omega, \mathscr{F}, \mathsf{P})$ with values in E. Let P_X and $P_{\tilde{X}}$ be their probability distributions. Then*

$$L(P_X, P_{\tilde{X}}) \le d_{\mathsf{P}}(X, \tilde{X}).$$

Remark 1. The preceding proof shows that in fact (10) is valid whenever we can exhibit on *any* probability space $(\Omega^*, \mathscr{F}^*, \mathsf{P}^*)$ random elements X and $\tilde{X}$ with values in E whose distributions coincide with P and $\tilde{P}$ and for which the set $\{\omega^* : \rho(X(\omega^*), \tilde{X}(\omega^*)) \ge \varepsilon\} \in \mathscr{F}^*$, $\varepsilon > 0$. Hence, the property of (10) depends in an essential way on how well, with respect to the measures P and $\tilde{P}$, the objects $(\Omega^*, \mathscr{F}^*, \mathsf{P}^*)$ and X, $\tilde{X}$ are constructed. (The procedure for constructing Ω^*, $\mathscr{F}^*$, P^* and X, $\tilde{X}$ as well as the measure P^*, is called *coupling* (joining, linking).) We could, for example, choose P^* equal to the direct product of the measures P and $\tilde{P}$, but this choice would, as a rule, not lead to a good estimate (10).

Remark 2. It is natural to raise the question of when there is equality in (10). In this connection we state the following result without proof: *Let P and $\tilde{P}$ be two probability measures on a separable metric space $(E, \mathscr{E}, \rho)$; then there are $(\Omega^*, \mathscr{F}^*, \mathsf{P}^*)$ and X, $\tilde{X}$, such that*

$$L(P, \tilde{P}) = d_{\mathsf{P}^*}(X, \tilde{X}) = \inf\{\varepsilon > 0 : \mathsf{P}^*(\rho(X, \tilde{X}) \ge \varepsilon) \le \varepsilon\}.$$

5. Problems

1. Prove that in the case of separable metric spaces the real function $\rho(X(\omega), Y(\omega))$ is a random variable for all random elements $X(\omega)$ and $Y(\omega)$ defined on a probability space $(\Omega, \mathscr{F}, \mathsf{P})$.

2. Prove that the function $d_{\mathsf{P}}(X, Y)$ defined in (2) is a metric in the space of random elements with values in E.

3. Establish (5).

4. Prove that the set $\Delta_h = \{x \in E : h(x) \text{ is not } \rho\text{-continuous at } x\} \in \mathscr{E}$.

§9. The Distance in Variation between Probability Measures. Kakutani–Hellinger Distance and Hellinger Integrals. Application to Absolute Continuity and Singularity of Measures

1. Let $(\Omega, \mathscr{F})$ be a measurable space and $\mathscr{P} = \{P\}$ a family of probability measures on it.

Definition 1. The *distance in variation* between measures P and $\tilde{P}$ in $\mathscr{P}$ (notation: $\|P - \tilde{P}\|$) is the total (signed) variation of $P - \tilde{P}$, i.e.,

$$\operatorname{Var}(P - \tilde{P}) \equiv \sup \left| \int_\Omega \varphi(\omega) d(P - \tilde{P}) \right|, \tag{1}$$

where the sup is over the class of all $\mathscr{F}$-measurable functions that satisfy the condition that $|\varphi(\omega)| \le 1$.

Lemma 1. *The distance in variation is given by*

$$\|P - \tilde{P}\| = 2 \sup_{A \in \mathscr{F}} |P(A) - \tilde{P}(A)|. \tag{2}$$

Proof. Since, for all $A \in \mathscr{F}$,

$$P(A) - \tilde{P}(A) = \tilde{P}(\bar{A}) - P(\bar{A}),$$

we have

$$2|P(A) - \tilde{P}(A)| = |P(A) - \tilde{P}(A)| + |P(\bar{A}) - \tilde{P}(\bar{A})| \le \|P - \tilde{P}\|,$$

where the last inequality follows from (1).

For the proof of the converse inequality we turn to the Hahn decomposition (see, for example, [K9] or [H1], p. 121) of a *signed measure* $\mu \equiv P - \tilde{P}$. In this decomposition the measure μ is represented in the form $\mu = \mu_+ - \mu_-$, where the nonnegative measures μ_+ and μ_- (the upper and lower variations of μ) are of the form

$$\mu_+(A) = \int_{A \cap M} d\mu, \; \mu_-(A) = -\int_{A \cap \bar{M}} d\mu, \qquad A \in \mathscr{F},$$

where M is a subset of $\mathscr{F}$. Here

$$\operatorname{Var} \mu = \operatorname{Var} \mu_+ + \operatorname{Var} \mu_- = \mu_+(\Omega) + \mu_-(\Omega).$$

Since

$$\mu_+(\Omega) = P(M) - \tilde{P}(M), \qquad \mu_-(\Omega) = \tilde{P}(\bar{M}) - P(\bar{M}),$$

we have

$$\|P - \tilde{P}\| = (P(M) - \tilde{P}(M)) + (\tilde{P}(\bar{M}) - P(\bar{M})) \le 2 \sup_{A \in \mathscr{F}} |P(A) - \tilde{P}(A)|.$$

This completes the proof of the lemma.

Definition 2. A sequence of probability measures P_n, $n \ge 1$, is said to be *convergent in variation* to the measure P if

$$\|P_n - P\| \to 0, \qquad n \to \infty. \tag{3}$$

From this definition and Theorem 1, §1, Chapter III, it is easily seen that convergence in variation of probability measures defined on a metric space $(\Omega, \mathscr{F}, \rho)$ implies their weak convergence.

The proximity in variation of distributions is, perhaps, the strongest form of closeness of probability distributions, since if two distributions are close in variation, then in practice, in specific situations, they can be considered indistinguishable. In this connection, the impression may be created that the study of distance in variation is not of much probabilistic interest. However, for example, in Poisson's theorem (§6, Chapter I) the convergence of the binomial to the Poisson distribution takes place in the sense of convergence in variation to the zero distribution. (Later, in §11, we shall obtain an upper bound for this distance.)

We also provide an example from the field of mathematical statistics, where the necessity of determining the distance in variation between measures P and $\tilde{P}$ arises in a natural way in connection with the problem of discrimination between the results of observations of two statistical hypotheses H (the true distribution is P) and $\tilde{H}$ (the true distribution is $\tilde{P}$) in connection with the question of whether the measure P or $\tilde{P}$, defined on $(\Omega, \mathscr{F})$, is more plausible. If $\omega \in \Omega$ is treated as the result of an observation, by a *test* (for different hypotheses H and $\tilde{H}$) we understand any $\mathscr{F}$-measurable function $\varphi = \varphi(\omega)$ with values in $[0, 1]$, the statistical meaning of which is that $\varphi(\omega)$ is "the probability with which hypothesis $\tilde{H}$ is accepted if the result of the observation is ω."

We shall characterize the quality of different hypotheses H and $\tilde{H}$ *by the probabilities of errors of the first and second kind*:

$$\alpha(\varphi) = \mathsf{E}\varphi(\omega) \quad (= \text{Prob } (\textit{accepting } \tilde{H} | H \textit{ is true})),$$

$$\beta(\varphi) = \tilde{\mathsf{E}}(1 - \varphi(\omega)) \quad (= \text{Prob } (\textit{accepting } H | \tilde{H} \textit{ is true})).$$

In the case when hypotheses H and $\tilde{H}$ are *equivalent*, the *optimum* is naturally to consider a test $\varphi^* = \varphi^*(\omega)$ (if there is such a test) that minimizes the sum $\alpha(\varphi) + \beta(\varphi)$ of the errors.

We set

$$\mathscr{E}r(P, \tilde{P}) = \inf_{\varphi} [\alpha(\varphi) + \beta(\varphi)]. \tag{4}$$

Let $Q = (P + \tilde{P})/2$ and $z = dP/dQ$, $\tilde{z} = d\tilde{P}/dQ$. Then

$$\begin{aligned}\mathscr{E}r(P, \tilde{P}) &= \inf_{\varphi} [\mathsf{E}\varphi + \tilde{\mathsf{E}}(1 - \varphi)] \\ &= \inf_{\varphi} \mathsf{E}_Q[z\varphi + \tilde{z}(1 - \varphi)] = 1 + \inf_{\varphi} \mathsf{E}_Q[\varphi(z - \tilde{z})],\end{aligned}$$

where E_Q is the expectation of the measure Q.

It is easy to see that the inf is attained by the function

$$\varphi^*(\omega) = I\{\tilde{z} < z\}$$

and, since $\mathsf{E}_Q(z - \tilde{z}) = 0$, that

$$\mathscr{E}r(P, \tilde{P}) = 1 - \tfrac{1}{2}\mathsf{E}_Q|z - \tilde{z}| = 1 - \tfrac{1}{2}\|P - \tilde{P}\|, \tag{5}$$

where the last equation will follow from Lemma 2, below. Therefore, it is clear from (5) that the quality of various hypotheses that characterize the function $\mathscr{E}r(P, \tilde{P})$ really depends on the degree of proximity of the measures P and $\tilde{P}$ in the sense of *distance in variation.*

Lemma 2. *Let Q be a σ-finite measure such that $P \ll Q$, $\tilde{P} \ll Q$ and $z = dP/dQ$, $\tilde{z} = d\tilde{P}/dQ$ are Radon–Nikodým measures of P and $\tilde{P}$ with respect to Q. Then*

$$\|P - \tilde{P}\| = \mathsf{E}_Q|z - \tilde{z}| \tag{6}$$

and if $Q = (P + \tilde{P})/2$, we have

$$\|P - \tilde{P}\| = \mathsf{E}_Q|z - \tilde{z}| = 2\mathsf{E}_Q|1 - z| = 2\mathsf{E}_Q|1 - \tilde{z}|. \tag{7}$$

PROOF. For all $\mathscr{F}$-measurable functions $\psi = \psi(\omega)$ with $|\psi(\omega)| \le 1$, we see from the definitions of z and $\tilde{z}$ that

$$|\mathsf{E}\psi - \tilde{\mathsf{E}}\psi| = |\mathsf{E}_Q\psi(z - \tilde{z})| \le \mathsf{E}_Q|\psi||z - \tilde{z}| \le \mathsf{E}_Q|z - \tilde{z}|. \tag{8}$$

Therefore,

$$\|P - \tilde{P}\| \le \mathsf{E}_Q|z - \tilde{z}|. \tag{9}$$

However, for the function

$$\psi = \operatorname{sgn}(\tilde{z} - z) = \begin{cases} 1, & \tilde{z} \ge z, \\ -1, & \tilde{z} < z, \end{cases}$$

we have

$$|\mathsf{E}\psi - \tilde{\mathsf{E}}\psi| = \mathsf{E}_Q|z - \tilde{z}|. \tag{10}$$

We obtain the required equation (6) from (9) and (10). Then (7) follows from (6) because $z + \tilde{z} = 2$ (Q-a.s.).

Corollary 1. *Let P and $\tilde{P}$ be two probability distributions on $(R, \mathscr{B}(R))$ with probability densities (with respect to Lebesgue measure dx) $p(x)$ and $\tilde{p}(x)$, $x \in R$. Then*

$$\|P - \tilde{P}\| = \int_{-\infty}^{\infty} |p(x) - \tilde{p}(x)|\, dx. \tag{11}$$

(*As the measure Q, we are to take Lebesgue measure on $(R, \mathscr{B}(R))$.*)

Corollary 2. *Let P and $\tilde{P}$ be two* discrete *measures, $P = (p_1, p_2, \ldots)$, $\tilde{P} = (\tilde{p}, \tilde{p}_2, \ldots)$, concentrated on a countable set of points $x_1, x_2, \ldots$. Then*

$$\|P - \tilde{P}\| = \sum_{i=1}^{\infty} |p_i - \tilde{p}_i|. \tag{12}$$

(*As the measure Q, we are to take the* counting measure, *i.e., that with $Q(\{x_i\}) = 1, i = 1, 2, \ldots$.*)

2. We now turn to still another measure of the proximity of two probability measures, from among many (as will follow later) related proximities of measures in variation.

Let P and $\tilde{P}$ be probability measures on $(\Omega, \mathscr{F})$ and Q, the third probability measure, *dominating* P and $\tilde{P}$, i.e., with the probabilities $P \ll Q$ and $\tilde{P} \ll Q$. We again use the notation

$$z = \frac{dP}{dQ}, \qquad \tilde{z} = \frac{d\tilde{P}}{dQ}.$$

Definition 3. The *Kakutani–Hellinger distance* between the measures P and $\tilde{P}$ is the nonnegative number $\rho(P, \tilde{P})$ such that

$$\rho^2(P, \tilde{P}) = \tfrac{1}{2}\mathsf{E}_Q[\sqrt{z} - \sqrt{\tilde{z}}]^2. \tag{13}$$

Since

$$\mathsf{E}_Q[\sqrt{z} - \sqrt{\tilde{z}}]^2 = \int_\Omega \left[\sqrt{\frac{dP}{dQ}} - \sqrt{\frac{d\tilde{P}}{dQ}}\right]^2 dQ, \tag{14}$$

it is natural to write $\rho^1(P, \tilde{P})$ symbolically in the form

$$\rho^2(P, \tilde{P}) = \tfrac{1}{2}\int_\Omega [\sqrt{dP} - \sqrt{d\tilde{P}}]^2. \tag{15}$$

If we set

$$H(P, \tilde{P}) = \mathsf{E}_Q\sqrt{z\tilde{z}}, \tag{16}$$

then, by analogy with (15), we may write symbolically

$$H(P, \tilde{P}) = \int_\Omega \sqrt{dP\, d\tilde{P}}. \tag{17}$$

From (13) and (16), as well as from (15) and (17), it is clear that

$$\rho^2(P, \tilde{P}) = 1 - H(P, \tilde{P}). \tag{18}$$

The number $H(P, \tilde{P})$ is called the *Hellinger integral* of the measures P and $\tilde{P}$. It turns out to be convenient, for many purposes, to consider the *Hellinger integrals* $H(\alpha; P, \tilde{P})$ *of order* $\alpha \in (0, 1)$, defined by the formula

$$H(\alpha; P, \tilde{P}) = \mathsf{E}_Q z^\alpha \tilde{z}^{1-\alpha}, \tag{19}$$

or, symbolically,

$$H(\alpha; P, \tilde{P}) = \int_\Omega (dP)^\alpha (d\tilde{P})^{1-\alpha}. \tag{20}$$

It is clear that $H(1/2; P, \tilde{P}) = H(P, \tilde{P})$.

For Definition 3 to be reasonable, we need to show that the number $\rho^2(P, \tilde{P})$ is independent of the choice of the dominating measure and that in fact $\rho(P, \tilde{P})$ satisfies the requirements of the concept of "distance."

Lemma 3.1. *The Hellinger integral of order* $\alpha \in (0, 1)$ *(and consequently also* $\rho(P, \tilde{P})$*) is independent of the choice of the dominating measure* Q.

2. *The function ρ defined in* (13) *is a metric on the set of probability measures.*

PROOF. 1. If the measure Q' dominates P and $\tilde{P}$, Q' also dominates $Q = (P + \tilde{P})/2$. Hence, it is enough to show that if $Q \ll Q'$, we have

$$\mathsf{E}_Q(z^\alpha \tilde{z}^{1-\alpha}) = \mathsf{E}_{Q'}(z')^\alpha(\tilde{z}')^{1-\alpha},$$

where $z' = dP/dQ'$ and $\tilde{z}' = d\tilde{P}/dQ'$.

Let us set $v = dQ/dQ'$. Then $z' = zv$, $\tilde{z}' = \tilde{z}v$, and

$$\mathsf{E}_Q(z^\alpha \tilde{z}^{1-\alpha}) = \mathsf{E}_{Q'}(vz^\alpha \tilde{z}^{1-\alpha}) = \mathsf{E}_{Q'}(z')^\alpha(\tilde{z}')^{1-\alpha},$$

which establishes the first assertion.

2. If $\rho(P, \tilde{P}) = 0$ we have $z = \tilde{z}$ (Q-a.e.) and hence, $P = \tilde{P}$. By symmetry, we evidently have $\rho(P, \tilde{P}) = \rho(\tilde{P}, P)$. Finally, let P, P', and P'' be three measures, $P \ll Q$, $P' \ll Q$, and $P'' \ll Q$, with $z = dP/dQ$, $z' = dP'/dQ$, and $z'' = dP''/dQ$. By using the validity of the triangle inequality for the norm in $L^2(\Omega, \mathscr{F}, Q)$, we obtain

$$[\mathsf{E}_Q(\sqrt{z} - \sqrt{z''})^2]^{1/2} \le [\mathsf{E}_Q(\sqrt{z} - \sqrt{z'})^2]^{1/2} + [\mathsf{E}_Q(\sqrt{z'} - \sqrt{z''})^2]^{1/2},$$

i.e.,

$$\rho(P, P'') \le \rho(P, P') + \rho(P', P'').$$

This completes the proof of the lemma.

By Definition (19) and Fubini's theorem (§6, Chapter II), it follows immediately that in the case when the measures P and $\tilde{P}$ are *direct products* of measures, $P = P_1 \times \cdots \times P_n$, $\tilde{P} = \tilde{P}_1 \times \cdots \times \tilde{P}_n$ (see subsection 9, §6, Chapter II), the Hellinger integral between the measures P and $\tilde{P}$ is equal to the product of the corresponding Hellinger integrals:

$$H(\alpha; P, \tilde{P}) = \prod_{i=1}^{n} H(\alpha; P_i, \tilde{P}_i).$$

The following theorem shows the connection between distance in variation and Kakutani–Hellinger distance (or, equivalently, the Hellinger integral). In particular, it shows that these distances define the same topology in the space of probability measures on $(\Omega, \mathscr{F})$.

Theorem 1. *We have the following inequalities*:

$$2[1 - H(P, \tilde{P})] \le \|P - \tilde{P}\| \le \sqrt{8[1 - H(P, \tilde{P})]}, \tag{21}$$

$$\|P - \tilde{P}\| \le 2\sqrt{1 - H^2(P, \tilde{P})}. \tag{22}$$

In particular,

$$2\rho^2(P, \tilde{P}) \le \|P - \tilde{P}\| \le \sqrt{8}\rho(P, \tilde{P}). \tag{23}$$

PROOF. Since $H(P, \tilde{P}) \le 1$ and $1 - x^2 \le 2(1 - x)$ for $0 \le x \le 1$, the right-hand inequality in (21) follows from (22), the proof of which is provided by the following chain of inequalities (where $Q = (1/2)(P + \tilde{P})$):

$$\begin{aligned}\tfrac{1}{2}\|P - \tilde{P}\| &= \mathsf{E}_Q|1 - z| \le \sqrt{\mathsf{E}_Q|1 - z|^2} = \sqrt{1 - \mathsf{E}_Q z(2 - z)} \\ &= \sqrt{1 - \mathsf{E}_Q z\tilde{z}} = \sqrt{1 - \mathsf{E}_Q(\sqrt{z\tilde{z}})^2} \le \sqrt{1 - (\mathsf{E}_Q\sqrt{z\tilde{z}})^2} \\ &= \sqrt{1 - H^2(P, \tilde{P})}.\end{aligned}$$

Finally, the first inequality in (21) follows from the fact that by the inequality

$$\tfrac{1}{2}[\sqrt{z} - \sqrt{2 - z}]^2 \le |z - 1|, \qquad z \in [0, 2],$$

and we have (again, $Q = (1/2)(P + \tilde{P})$)

$$1 - H(P, \tilde{P}) = \rho^2(P, \tilde{P}) = \tfrac{1}{2}\mathsf{E}_Q[\sqrt{z} - \sqrt{2 - z}]^2 \le \tfrac{1}{2}\mathsf{E}_Q|z - 1| = \tfrac{1}{2}\|P - \tilde{P}\|.$$

Remark. It can be shown in a similar way that, for every $\alpha \in (0, 1)$,

$$2[1 - H(\alpha; P, \tilde{P})] \le \|P - \tilde{P}\| \le \sqrt{c_\alpha(1 - H(\alpha; P, \tilde{P}))}, \tag{24}$$

where c_α is a constant.

Corollary 1. *Let P and P^n, $n \ge 1$, be probability measures on $(\Omega, \mathscr{F})$. Then (as $n \to \infty$)*

$$\|P^n - P\| \to 0 \Leftrightarrow H(P^n, P) \to 1 \Leftrightarrow \rho(P^n, P) \to 0,$$
$$\|P^n - P\| \to 2 \Leftrightarrow H(P^n, P) \to 0 \Leftrightarrow \rho(P^n, P) \to 1.$$

Corollary 2. *Since by* (5)

$$\mathscr{E}r(P, \tilde{P}) = 1 - \tfrac{1}{2}\|P - \tilde{P}\|,$$

we have, by (21) *and* (22),

$$\tfrac{1}{2}H^2(P, \tilde{P}) \le 1 - \sqrt{1 - H^2(P, \tilde{P})} \le \mathscr{E}r(P, \tilde{P}) \le H(P, \tilde{P}). \tag{25}$$

In particular, let

$$P^n = \underbrace{P \times \cdots \times P}_{n}, \qquad \tilde{P}^n = \underbrace{\tilde{P} \times \cdots \times \tilde{P}}_{n}$$

be direct products of measures. Then, since $H(P^n, \tilde{P}^n) = [H(P, \tilde{P})]^n = e^{-\lambda n}$ with $\lambda = -\ln H(P, \tilde{P}) \ge \rho^2(P, \tilde{P})$, we obtain from (25) the inequalities

$$\tfrac{1}{2}e^{-2\lambda n} \le \mathscr{E}r(P^n, \tilde{P}^n) \le e^{-\lambda n} \le e^{-n\rho^2(P, \tilde{P})}. \tag{26}$$

In connection with the problem, considered above, of *distinguishing two statistical hypotheses* from these inequalities, we have the following result.

Let $\xi_1, \xi_2, \ldots$ be independent identically distributed random elements, that have either the probability distribution P (Hypothesis H) or $\tilde{P}$ (Hypothesis $\tilde{H}$), with $\tilde{P} \neq P$, and therefore, $\rho^2(P, \tilde{P}) > 0$. Therefore, when $n \to \infty$, the function $\mathscr{E}r(P^n, \tilde{P}^n)$, which describes the quality of optimality of the hypotheses H and $\tilde{H}$ as observations of $\xi_1, \xi_2, \ldots$, decreases exponentially to zero.

4. In using Hellinger integrals of order α (described above), it will be convenient to introduce the notions of absolute continuity and singularity of probability measures.

Let P and $\tilde{P}$ be two probability measures defined on a measurable space $(\Omega, \mathscr{F})$. We say that $\tilde{P}$ is *absolutely continuous* with respect to P (notation: $\tilde{P} \ll P$) if $\tilde{P}(A) = 0$ whenever $P(A) = 0$ for $A \in \mathscr{F}$. If $\tilde{P} \ll P$ and $P \ll \tilde{P}$, we say that P and $\tilde{P}$ are *equivalent* ($\tilde{P} \sim P$). The measures P and $\tilde{P}$ are called *singular* or *orthogonal* ($\tilde{P} \perp P$), if there is an $A \in \mathscr{F}$ for which $P(A) = 1$ and $\tilde{P}(\bar{A}) = 1$ (i.e., P and $\tilde{P}$ "sit" on different sets).

Let Q be a probability measure, with $P \ll Q$, $\tilde{P} \ll Q$, $z = dP/dQ$, $\tilde{z} = d\tilde{P}/dQ$.

Theorem 2. *The following conditions are equivalent*:

(a) $\tilde{P} \ll P$,
(b) $\tilde{P}(z > 0) = 1$,
(c) $H(\alpha; P, \tilde{P}) \to 1$, $\alpha \downarrow 0$.

Theorem 3. *The following conditions are equivalent*:

(a) $\tilde{P} \perp P$,
(b) $\tilde{P}(z > 0) = 0$,
(c) $H(\alpha; P, \tilde{P}) \to 0$, $\alpha \downarrow 0$,
(d) $H(\alpha; P, \tilde{P}) = 0$ *for all* $\alpha \in (0, 1)$,
(e) $H(\alpha; P, \tilde{P}) = 0$ *for some* $\alpha \in (0, 1)$.

The proofs of these theorems will be given simultaneously. By the definitions of z and $\tilde{z}$,

$$P(z = 0) = \mathsf{E}_Q[zI(z = 0)] = 0, \tag{27}$$

$$\begin{aligned} \tilde{P}(A \cap \{z > 0\}) &= \mathsf{E}_Q[\tilde{z}I(A \cap \{z > 0\})] \\ &= \mathsf{E}_Q\left[\tilde{z}\frac{z}{z}I(A \cap \{z > 0\})\right] = \mathsf{E}\left[\frac{\tilde{z}}{z}I(A \cap \{z > 0\})\right] \\ &= \mathsf{E}\left[\frac{\tilde{z}}{z}I(A)\right]. \end{aligned} \tag{28}$$

Consequently, we have the *Lebesgue decomposition*

$$\tilde{P}(A) = \mathsf{E}\left[\frac{\tilde{z}}{z}I(A)\right] + \tilde{P}(A \cap \{z = 0\}), \qquad A \in \mathscr{F}, \tag{29}$$

in which $Z = \tilde{z}/z$ is called the *Lebesgue derivative of* $\tilde{P}$ *with respect to* P and denoted by $d\tilde{P}/dP$ (compare the remark on the Radon–Nikodým theorem, §6, Chapter III).

Hence, we immediately obtain the equivalence of (a) and (b) in both theorems.

Moreover, since

$$z^\alpha \tilde{z}^{1-\alpha} \to \tilde{z} I(z > 0), \qquad \alpha \downarrow 0,$$

and for $\alpha \in (0, 1)$

$$0 \le z^\alpha \tilde{z}^{1-\alpha} \le \alpha z + (1-\alpha)\tilde{z} \le z + \tilde{z}$$

with $\mathsf{E}_Q(z + \tilde{z}) = 2$, we have, by Lebesgue's dominated convergence theorem,

$$\lim_{\alpha \downarrow 0} H(\alpha; P, \tilde{P}) = \mathsf{E}_Q \tilde{z} I(z > 0) = \tilde{P}(z > 0)$$

and therefore, (b)$\Leftrightarrow$(c) in both theorems.

Finally, let us show that in the second theorem (c)$\Leftrightarrow$(d)$\Leftrightarrow$(e). For this, we need only note that $H(\alpha; P, \tilde{P}) = \tilde{\mathsf{E}}(z/\tilde{z})^\alpha I(\tilde{z} > 0)$ and $\tilde{P}(\tilde{z} > 0) = 1$. Hence, for each $\alpha \in (0, 1)$ we have $\tilde{P}(z > 0) = 0 \Leftrightarrow H(\alpha; P, \tilde{P}) = 0$, from which there follows the implication (c)$\Leftrightarrow$(d)$\Leftrightarrow$(e).

Example 1. Let $P = P_1 \times P_2 \times \ldots$, $\tilde{P} = \tilde{P}_1 \times \tilde{P}_2 \ldots$, where P_k and $\tilde{P}_k$ are Gaussian measures on $(R, \mathscr{B}(R))$ with densities

$$p_k(x) = \frac{1}{\sqrt{2\pi}} e^{(x-a_k)^2/2}, \qquad \tilde{p}_k(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\tilde{a}_k)^2/2}.$$

Since

$$H(\alpha; P, \tilde{P}) = \prod_{k=1}^{\infty} H(\alpha; P_k, \tilde{P}_k),$$

where a simple calculation shows that

$$H(\alpha; P_k, \tilde{P}_k) = \int_{-\infty}^{\infty} p_k^\alpha(x) \tilde{p}_k^{1-\alpha}(x)\, dx = e^{-(\alpha(1-\alpha)/2)(a_k - \tilde{a}_k)^2},$$

we have

$$H(\alpha; P, \tilde{P}) = e^{-\alpha(1-\alpha)/2 \sum_{k=1}^{\infty} (a_k - \tilde{a}_k)^2}.$$

From Theorems 2 and 3, we find that

$$\tilde{P} \ll P \Leftrightarrow P \ll \tilde{P} \Leftrightarrow \tilde{P} \sim P \Leftrightarrow \sum_{k=1}^{\infty} (a_k - \tilde{a}_k)^2 < \infty,$$

$$\tilde{P} \perp P \Leftrightarrow \sum_{k=1}^{\infty} (a_k - \tilde{a}_k)^2 = \infty.$$

Example 2. Again let $P = P_1 \times P_2 \times \ldots$, $\tilde{P} = \tilde{P}_1 \times \tilde{P}_2 \times \ldots$, where P_k and $\tilde{P}_k$ are Poisson distributions with respective parameters $\lambda_k > 0$ and $\tilde{\lambda}_k > 0$. Then it is easily shown that

$$\begin{gathered}\tilde{P} \ll P \Leftrightarrow P \ll \tilde{P} \Leftrightarrow \tilde{P} \sim P \Leftrightarrow \sum_{k=1}^{\infty} (\sqrt{\lambda_k} - \sqrt{\tilde{\lambda}_k})^2 < \infty, \\ \tilde{P} \perp P \Leftrightarrow \sum_{k=1}^{\infty} (\sqrt{\lambda_k} - \sqrt{\tilde{\lambda}_k})^2 = \infty. \end{gathered} \tag{30}$$

5. Problems

1. In the notation of Lemma 2, set

$$P \wedge \tilde{P} = \mathsf{E}_Q(z \wedge \tilde{z}),$$

where $z \wedge \tilde{z} = \min(z, \tilde{z})$. Show that

$$\|P - \tilde{P}\| = 2(1 - P \wedge \tilde{P})$$

(and consequently, $\mathscr{E}r(P, \tilde{P}) = P \wedge \tilde{P}$).

2. Let P, P_n, $n \geq 1$, be probability measures on $(R, \mathscr{B}(R))$ with densities (with respect to Lebesgue measure) $p(x)$, $p_n(x)$, $n \geq 1$. Let $p_n(x) \to p(x)$ for almost all x (with respect to Lebesgue measure). Show that then

$$\|P - P_n\| = \int_{-\infty}^{\infty} |p(x) - p_n(x)|\, dx \to 0, \qquad n \to \infty$$

(compare Problem 17 in §6, Chapter II).

3. Let P and $\tilde{P}$ be two probability measures. We define Kullback information $K(P, \tilde{P})$ as information by using P against $\tilde{P}$, by the equation

$$K(P, \tilde{P}) = \begin{cases} \mathsf{E} \ln(dP/d\tilde{P}) & \text{if } P \ll \tilde{P}, \\ \infty & \text{otherwise.} \end{cases}$$

Show that

$$K(P, \tilde{P}) \geq -2 \ln(1 - \rho^2(P, \tilde{P})) \geq 2\rho^2(P, \tilde{P}).$$

4. Establish formulas (11) and (12).

5. Prove inequalities (24).

6. Let P, $\tilde{P}$, and Q be probability measures on $(R, \mathscr{B}(R))$; $P * Q$ and $\tilde{P} * Q$, their convolutions (see subsection 4, §8, Chapter II). Then

$$\|P * Q - \tilde{P} * Q\| \leq \|P - \tilde{P}\|.$$

7. Prove (30).

§10. Contiguity and Entire Asymptotic Separation of Probability Measures

1. These concepts play a fundamental role in the asymptotic theory of mathematical statistics, being natural extensions of the concepts of absolute continuity and singularity of two measures in the case of *sequences* of pairs of measures.

Let us begin with definitions.

Let $(\Omega^n, \mathscr{F}^n)_{n\geq 1}$ be a sequence of measurable spaces; let $(P^n)_{n\geq 1}$ and $(\tilde{P}^n)_{n\geq 1}$ be sequences of probability measures with P^n and $\tilde{P}^n$ defined on $(\Omega^n, \mathscr{F}^n)$, $n \geq 1$.

Definition 1. We say that a sequence $(\tilde{P}^n)$ of measures is *contiguous* to the sequence (P^n) (notation: $(\tilde{P}^n) \lhd (P^n)$) if, for all $A^n \in \mathscr{F}^n$ such that $P^n(A^n) \to 0$ as $n \to \infty$, we have $\tilde{P}^n(A^n) \to 0$, $n \to \infty$.

Definition 2. We say that sequences $(\tilde{P}^n)$ and (P^n) of measures are *entirely* (*asymptotically*) *separated* (or for short: $(\tilde{P}^n) \triangle (P^n)$), if there is a subsequence $n_k \uparrow \infty$, $k \to \infty$, and sets $A^{n_k} \in \mathscr{F}^{n_k}$ such that

$$P^{n_k}(A^{n_k}) \to 1 \qquad \text{and} \qquad \tilde{P}^{n_k}(A^{n_k}) \to 0, \qquad k \to \infty.$$

We notice immediately that entire separation is a *symmetric* concept: $(\tilde{P}^n) \triangle (P^n) \Leftrightarrow (P^n) \triangle (\tilde{P}^n)$. Contiguity does not have this property. If $(\tilde{P}^n) \lhd (P^n)$ and $(P^n) \lhd (\tilde{P}^n)$, we write $(\tilde{P}^n) \lhd \rhd (P^n)$ and say that the sequences (P^n) and $(\tilde{P}^n)$ of measures are *mutually contiguous*.

We notice that in the case when $(\Omega^n, \mathscr{F}^n) = (\Omega, \mathscr{F})$, $P^n = P$, $\tilde{P}^n = \tilde{P}$ for all $n \geq 1$, we have

$$(\tilde{P}^n) \lhd (P^n) \Leftrightarrow \tilde{P} \ll P, \tag{1}$$

$$(\tilde{P}^n) \lhd \rhd (P^n) \Leftrightarrow \tilde{P} \sim P, \tag{2}$$

$$(\tilde{P}^n) \triangle (P^n) \Leftrightarrow \tilde{P} \perp P. \tag{3}$$

These properties and the definitions given above explain why contiguity and entire asymptotic separation are often thought of as "asymptotic absolute continuity" and "asymptotic singularity" for sequences $(\tilde{P}^n)$ and (P^n).

2. Theorems 1 and 2 presented below are natural extensions of Theorems 2 and 3 of §8 to sequences of measures.

Let $(\Omega^n, \mathscr{F}^n)_{n\geq 1}$ be a sequence of measurable spaces; Q^n, a probability measure on $(\Omega^n, \mathscr{F}^n)$; and ξ^n, a random variable (generally speaking, extended; see §4, Chapter II) on $(\Omega^n, \mathscr{F}^n)$, $n \geq 1$.

Definition 3. A sequence (ξ^n) of random variables is *tight* with respect to a sequence of measures (Q^n) (notation: $(\xi^n | Q^n)$ is *tight*) if

$$\lim_{N\uparrow\infty} \varlimsup_{n} Q^n(|\xi^n| > N) = 0. \tag{4}$$

(Compare the corresponding definition of tightness of a family of probability measures in §2.)

We shall always set

$$Q^n = \frac{P^n + \tilde{P}^n}{2}, \qquad z^n = \frac{dP^n}{dQ^n}, \quad \tilde{z}^n = \frac{d\tilde{P}^n}{dQ^n}.$$

We shall also use the notation

$$Z^n = \tilde{z}^n / z^n \tag{5}$$

for the Lebesgue derivative of $\tilde{P}^n$ with respect to P^n (see (29) in §9), taking $2/0 = \infty$. We note that if $\tilde{P}^n \ll P^n$, Z^n is precisely one of the versions of the density $d\tilde{P}^n/dP^n$ of the measure $\tilde{P}^n$ with respect to P^n (see §6, Chapter II).

For later use it is convenient to note that since

$$P^n\left(z^n \le \frac{1}{N}\right) = \mathsf{E}_{Q^n}\left(z^n I\left(z^n \le \frac{1}{N}\right)\right) \le \frac{1}{N} \tag{6}$$

and $Z^n \le 2/z^n$, we have

$$((1/z^n)|P^n) \qquad \text{tight}, \qquad (Z^n|P^n) \qquad \text{tight}. \tag{7}$$

Theorem 1. *The following statements are equivalent*:

(a) $(\tilde{P}^n) \lhd (P^n)$,

(b) $(\tilde{z}^{-n}|\tilde{P}^n)$ *is tight*,

(b′) $(Z^n|\tilde{P}^n)$ *is tight*,

(c) $\lim_{\alpha \downarrow 0} \underline{\lim}_n H(\alpha; P^n, \tilde{P}^n) = 1$.

Theorem 2. *The following statements are equivalent*:

(a) $(\tilde{P}^n) \triangle (P^n)$,

(b) $\underline{\lim}_n \tilde{P}^n(z^n \ge \varepsilon) = 0$ *for every* $\varepsilon > 0$,

(b′) $\overline{\lim}_n \tilde{P}^n(Z^n \le N) = 0$ *for every* $N > 0$,

(c) $\lim_{\alpha \downarrow 0} \underline{\lim}_n H(\alpha; P^n, \tilde{P}^n) = 0$,

(d) $\underline{\lim}_n H(\alpha; P^n, \tilde{P}^n) = 0$ *for all* $\alpha \in (0, 1)$,

(e) $\underline{\lim}_n H(\alpha; P^n, \tilde{P}^n) = 0$ *for some* $\alpha \in (0, 1)$.

Proof of Theorem 1.

(a) $\Rightarrow$ (b). If (b) is not satisfied, there are an $\varepsilon > 0$ and a sequence $n \uparrow \infty$ such that $\tilde{P}^{n_k}(z^{n_k} < 1/n_k) \ge \varepsilon$. But by (6), $P^{n_k}(z^{n_k} < 1/n_k) \le 1/n_k$, $k \to \infty$, which contradicts the assumption that $(\tilde{P}^n) \lhd (P^n)$.

(b) $\Leftrightarrow$ (b′). We have only to note that $Z^n = 2z^{-k} - 1$.

(b) $\Rightarrow$ (a). Let $A^n \in \mathscr{F}^n$ and $P^n(A^n) \to 0$, $n \to \infty$. We have

$$\begin{aligned}\tilde{P}^n(A^n) &\le \tilde{P}^n(z^n \le \varepsilon) + \mathsf{E}_{Q^n}(\tilde{z}^n I(A^n \cap \{z^n > \varepsilon\})) \\ &\le \tilde{P}^n(z^n \le \varepsilon) + \frac{2}{\varepsilon}\mathsf{E}_{Q^n}(z^n I(A^n)) = \tilde{P}^n(z^n \le \varepsilon) + \frac{2}{\varepsilon}P^n(A^n).\end{aligned}$$

Therefore,

$$\overline{\lim_n}\, \tilde{P}^n(A^n) \le \overline{\lim_n}\, \tilde{P}^n(z^n \le \varepsilon), \qquad \varepsilon > 0.$$

Proposition (b) is equivalent to saying that $\lim_{\varepsilon\downarrow 0} \overline{\lim}_n \tilde{P}^n(z^n \le \varepsilon) = 0$. Therefore, $\tilde{P}^n(A^n) \to 0$, i.e., (b) ⇒ (a).

(b) ⇒ (c). Let $\varepsilon > 0$. Then

$$\begin{aligned} H(\alpha; P^n, \tilde{P}^n) &= \mathsf{E}_{Q^n}[(z^n)^\alpha(\tilde{z}^n)^{1-\alpha}] \\ &\ge \mathsf{E}_{Q^n}\left[\left(\frac{z^n}{\tilde{z}^n}\right)^\alpha I(z^n \ge \varepsilon) I(\tilde{z}^n > 0)\tilde{z}^n\right] \\ &= \mathsf{E}_{\tilde{P}^n}\left[\left(\frac{z^n}{\tilde{z}^n}\right)^\alpha I(z^n \ge \varepsilon)\right] \ge \left(\frac{\varepsilon}{2}\right)^\alpha \tilde{P}^n(z^n \ge \varepsilon), \end{aligned} \tag{8}$$

since $z^n + \tilde{z}^n = 2$. Therefore, for $\varepsilon > 0$,

$$\underline{\lim}_{\alpha\downarrow 0}\, \underline{\lim}_n\, H(\alpha; P^n, \tilde{P}^n) \ge \underline{\lim}_{\alpha\downarrow 0} \left(\frac{\varepsilon}{2}\right)^\alpha \underline{\lim}_n\, \tilde{P}^n(z^n \ge \varepsilon) = \underline{\lim}_n\, \tilde{P}^n(z^n \ge \varepsilon). \tag{9}$$

By (b), $\underline{\lim}_{\varepsilon\downarrow 0}\, \underline{\lim}_n \tilde{P}^n(z^n \ge \varepsilon) = 1$. Hence, (c) follows from (9) and the fact that $H(\alpha; P^n, \tilde{P}^n) \le 1$.

(c) ⇒ (b). Let $\delta \in (0, 1)$. Then

$$\begin{aligned} H(\alpha; P^n, \tilde{P}^n) &= \mathsf{E}_{Q^n}[(z^n)^\alpha(\tilde{z}^n)^{1-\alpha} I(z^n < \varepsilon)] \\ &\quad + \mathsf{E}_{Q^n}[(z^n)^\alpha(\tilde{z}^n)^{1-\alpha} I(z^n \ge \varepsilon, \tilde{z}^n \le \delta)] \\ &\quad + \mathsf{E}_{Q^n}[(z^n)^\alpha(\tilde{z}^n)^{1-\alpha} I(z^n \ge \varepsilon, \tilde{z}^n > \delta)] \\ &\le 2\varepsilon^\alpha + 2\delta^{1-\alpha} + \mathsf{E}_{Q^n}\left[\tilde{z}^n \left(\frac{z^n}{\tilde{z}^n}\right)^\alpha I(z^n \ge \varepsilon, \tilde{z}^n > \delta)\right] \\ &\le 2\varepsilon^\alpha + 2\delta^{1-\alpha} + \left(\frac{2}{\delta}\right)^\alpha \tilde{P}^n(z^n \ge \varepsilon). \end{aligned} \tag{10}$$

Consequently,

$$\underline{\lim}_{\varepsilon\downarrow 0}\, \underline{\lim}_n\, \tilde{P}^n(z^n \ge \varepsilon) \ge \left(\frac{\delta}{2}\right)^\alpha \underline{\lim}_n\, H(\alpha; P^n, \tilde{P}^n) - \frac{2}{2^\alpha}\delta$$

for all $\alpha \in (0, 1)$, $\delta \in (0, 1)$. If we first let $\alpha \downarrow 0$, use (c), and then let $\delta \downarrow 0$, we obtain

$$\underline{\lim}_{\varepsilon\downarrow 0}\, \underline{\lim}_n\, \tilde{P}^n(z^n \ge \varepsilon) \ge 1,$$

from which (b) follows.

Proof of Theorem 2.

(a) ⇒ (b). Let $(\tilde{P}^n) \triangle (P^n)$, $n_k \uparrow \infty$, and let $A^{n_k} \in \mathscr{F}^{n_k}$ have the property that $P^{n_k}(A^{n_k}) \to 1$ and $\tilde{P}^{n_k}(A^{n_k}) \to 0$. Then, since $z^n + \tilde{z}^n = 2$, we have

$$\begin{aligned}\tilde{P}^{n_k}(z^{n_k} \ge \varepsilon) &\le \tilde{P}^{n_k}(A^{n_k}) + \mathsf{E}_{Q^{n_k}}\left\{z^{n_k} \cdot \frac{\tilde{z}^{n_k}}{z^{n_k}} I(\bar{A}^{n_k}) I(z^{n_k} \ge \varepsilon)\right\}\\ &= \tilde{P}^{n_k}(A^{n_k}) + \mathsf{E}_{P^{n_k}}\left\{\frac{\tilde{z}^{n_k}}{z^{n_k}} I(\bar{A}^{n_k}) I(z^{n_k} \ge \varepsilon)\right\}\\ &\le \tilde{P}^{n_k}(A^{n_k}) + \frac{2}{\varepsilon} P^{n_k}(\bar{A}^{n_k}).\end{aligned}$$

Consequently, $\tilde{P}^{n_k}(z^{n_k} \ge \varepsilon) \to 0$ and therefore, (b) is satisfied.

(b) $\Rightarrow$ (a). If (b) is satisfied, there is a sequence $n_k \uparrow \infty$ such that

$$\tilde{P}^{n_k}\left(z^{n_k} \ge \frac{1}{k}\right) \le \frac{1}{k} \to 0, \qquad k \to \infty.$$

Hence, having observed (see (6)) that $P^{n_k}(z^{n_k} \ge 1/k) \ge 1 - (1/k)$, we obtain (a).

(b) $\Rightarrow$ (b′). We have only to observe that $Z^n = (2/z^n) - 1$.

(b) $\Rightarrow$ (d). By (10) and (b),

$$\varliminf_n H(\alpha; P^n, \tilde{P}^n) \le 2\varepsilon^\alpha + 2\delta^{1-\alpha}$$

for arbitrary ε and δ on the interval (0, 1). Therefore, (d) is satisfied.

(d) $\Rightarrow$ (c) and (d) $\Rightarrow$ (e) are evident.

Finally, from (8) we have

$$\varliminf_n \tilde{P}^n(z^n \ge \varepsilon) \le \left(\frac{2}{\varepsilon}\right)^\alpha \varliminf_n H(\alpha; P^n, \tilde{P}^n).$$

Therefore, (c) $\Rightarrow$ (b) and (e) $\Rightarrow$ (b), since $(2/\varepsilon)^\alpha \to 1$, $\alpha \downarrow 0$.

3. We now consider a special case corresponding to the method of independent observations, where the calculation of the integrals $H(\alpha; P^n, \tilde{P}^n)$ and application of Theorems 1 and 2 do not present much difficulty.

Let us suppose that the measures P^n and $\tilde{P}^n$ are direct products of measures:

$$P^n = P_1 \times \cdots \times P_n, \qquad \tilde{P}^n = \tilde{P}_1 \times \cdots \times \tilde{P}_n, \qquad n \ge 1,$$

where P_k and $\tilde{P}_k$ are given on $(\Omega_k, \mathscr{F}_k)$, $k \ge 1$.

Since in this case

$$H(\alpha; P^n, \tilde{P}^n) = \prod_{k=1}^n H(\alpha; P_k, \tilde{P}_k) = e^{\sum_{k=1}^n \ln[1-(1-H(\alpha; P_k, \tilde{P}_k))]},$$

we obtain the following result from Theorems 1 and 2:

$$(\tilde{P}^n) \lhd (P^n) \Leftrightarrow \lim_{\alpha \downarrow 0} \varlimsup_n \sum_{k=1}^n [1 - H(\alpha; P_k, \tilde{P}_k)] = 0, \tag{11}$$

$$(\tilde{P}^n) \triangle (P^n) \Leftrightarrow \varlimsup_n \sum_{k=1}^n [1 - H(\alpha; P_k, \tilde{P}_k)] = \infty. \tag{12}$$

EXAMPLE. Let $(\Omega_k, \mathscr{F}_k) = (R, \mathscr{B}(R))$, $a_k \in [0, 1)$,

$$P_k(dx) = I_{[0,1]}(x)\,dx, \qquad \tilde{P}_k(dx) = \frac{1}{1-a_k} I_{[a_k,1]}(x)\,dx.$$

Since here $H(\alpha; P_k, \tilde{P}_k) = (1-a_k)^\alpha$, $\alpha \in (0, 1)$, from (11) and the fact that $H(\alpha; P_k, \tilde{P}_k) = H(1-\alpha; \tilde{P}_k, P_k)$, we obtain

$$(\tilde{P}^n) \vartriangleleft (P^n) \Leftrightarrow \overline{\lim_n}\, na_n < \infty, \quad \text{i.e.,} \quad a_n = O\left(\frac{1}{n}\right),$$

$$(P^n) \vartriangleleft (\tilde{P}^n) \Leftrightarrow \overline{\lim_n}\, na_n = 0, \quad \text{i.e.,} \quad a_n = o\left(\frac{1}{n}\right),$$

$$(\tilde{P}^n) \triangle (P^n) \Leftrightarrow \overline{\lim_n}\, na_n = \infty.$$

4. PROBLEMS

1. Let $P^n = P_1^n \times \cdots \times P_n^n$, $\tilde{P}^n = \tilde{P}_1^n \times \cdots \times \tilde{P}_n^n$, $n \geq 1$, where P_k^n and $\tilde{P}_k^n$ are *Gaussian measures* with parameters $(a_k^n, 1)$ and $(\tilde{a}_k^n, 1)$. Find conditions on (a_k^n) and $(\tilde{a}_k^n)$ under which $(\tilde{P}^n) \vartriangleleft (P^n)$ and $(\tilde{P}^n) \triangle (P^n)$.

2. Let $P^n = P_1^n \times \cdots \times P_n^n$ and $\tilde{P}^n = \tilde{P}_1^n \times \cdots \times \tilde{P}_n^n$, where P_k^n and $\tilde{P}_k^n$ are probability measures on $(R, \mathscr{B}(R))$ for which $P_k^n(dx) = I_{[0,1]}(x)\,dx$ and $\tilde{P}_k^n(dx) = I_{[a_n, 1+a_n]}(dx)$, $0 \leq a_n \leq 1$. Show that $H(\alpha; P_k^n, \tilde{P}_k^n) = 1 - a_n$ and

$$(\tilde{P}^n) \vartriangleleft (P^n) \Leftrightarrow (P^n) \vartriangleleft (\tilde{P}^n) \Leftrightarrow \overline{\lim_n}\, na_n = 0, \qquad (\tilde{P}^n) \triangle (P^n) \Leftrightarrow \overline{\lim_n}\, na_n = \infty.$$

§11. Rapidity of Convergence in the Central Limit Theorem

1. Let $\xi_{n1}, \ldots, \xi_{nn}$ be a sequence of independent random variables, $S_n = \xi_{n1} + \cdots + \xi_{nn}$, $F_n(x) = \mathsf{P}(S_n \leq x)$. If $S_n \to \mathscr{N}(0, 1)$, then $F_n(x) \to \Phi(x)$ for every $x \in R$. Since $\Phi(x)$ is continuous, the convergence here is actually uniform (Problem 5 in §1):

$$\sup_x |F_n(x) - \Phi(x)| \to 0, \qquad n \to \infty. \tag{1}$$

It is natural to ask how rapid the convergence in (1) is. We shall establish a result for the case when

$$S_n = \frac{\xi_1 + \cdots + \xi_n}{\sigma\sqrt{n}}, \qquad n \geq 1,$$

where $\xi_1, \xi_2, \ldots$ is a sequence of independent identically distributed random variables with $\mathsf{E}\xi_k = 0$, $\mathsf{V}\xi_k = \sigma^2$ and $\mathsf{E}|\xi_1|^3 < \infty$.

Theorem (Berry and Esseen). *We have the bound*

$$\sup_x |F_n(x) - \Phi(x)| \le \frac{CE|\xi_1|^3}{\sigma^3\sqrt{n}}, \tag{2}$$

where C is an absolute constant ($(2\pi)^{-1/2} \le C < 0.8$).

PROOF. For simplicity, let $\sigma^2 = 1$ and $\beta_3 = \mathsf{E}|\xi_1|^3$. By Esseen's inequality (Subsection 10, §12, Chapter II)

$$\sup_x |F_n(x) - \Phi(x)| \le \frac{2}{\pi}\int_0^T \left|\frac{f_n(t) - \varphi(t)}{t}\right| dt + \frac{24}{\pi T}\frac{1}{\sqrt{2\pi}} \tag{3}$$

where $\varphi(t) = e^{-t^2/2}$ and

$$f_n(t) = [f(t/\sqrt{n})]^n,$$

with $f(t) = \mathsf{E}e^{it\xi_1}$.

In (3) we may take T *arbitrarily*. Let us choose

$$T = \sqrt{n}/(5\beta_3).$$

We are going to show that for this T,

$$|f_n(t) - \varphi(t)| \le \frac{7}{6}\frac{\beta_3}{\sqrt{n}}|t|^3 e^{-t^2/4}, \qquad |t| \le T. \tag{4}$$

The required estimate (2), with C an absolute constant, will follow immediately from (3) by means of (4). (A more detailed analysis shows that $C < 0.8$.)

We now turn to the proof of (4).

By formula (18) from §2, Chapter II ($n = 3$, $\mathsf{E}\xi_1 = 0$, $\mathsf{E}\xi_1^2 = 1$, $\mathsf{E}|\xi_1|^3 < \infty$) we obtain

$$f(t) = \mathsf{E}e^{it\xi_1} = 1 - \frac{t^2}{2} + \frac{(it)^3}{6}[\mathsf{E}\xi_1^3(\cos\theta_1 t\xi_1 + i\sin\theta_2 t\xi_1)], \tag{5}$$

where $|\theta_1| \le 1$, $|\theta_2| \le 1$. Consequently,

$$f\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + \frac{(it)^3}{6n^{3/2}}\left[\mathsf{E}\xi_1^3\left(\cos\theta_1\frac{t}{\sqrt{n}}\xi_1 + i\sin\theta_1\frac{t}{\sqrt{n}}\xi_1\right)\right].$$

If $|t| \le T = \sqrt{n}/5\beta_3$, we find, by using the inequality $\beta_3 \ge \sigma^3 = 1$ (see (28), §6, Chapter II), that

$$1 - \left|f\left(\frac{t}{\sqrt{n}}\right)\right| \le \left|1 - f\left(\frac{t}{\sqrt{n}}\right)\right| \le \frac{t^2}{2n} + \frac{|t|^3\beta_3}{3n^{3/2}} \le \frac{1}{25}.$$

Consequently, for $|t| \le T$ it is possible to have the representation

$$\left[f\left(\frac{t}{\sqrt{n}}\right)\right]^n = e^{n \ln f(t/\sqrt{n})}, \tag{6}$$

where ln z means the principal value of the logarithm of the complex number z ($\ln z = \ln|z| + i \arg z$, $-\pi < \arg z \le \pi$).

Since $\beta_3 < \infty$, we obtain from Taylor's theorem with the Lagrange remainder (compare (35) in §12, Chapter II)

$$\ln f\left(\frac{t}{\sqrt{n}}\right) = \frac{it}{\sqrt{n}} s_{\xi_1}^{(1)} + \frac{(it)^2}{2n} s_{\xi_1}^{(2)} + \frac{(it)^3}{6n^{3/2}} (\ln f)'''\left(\theta \frac{t}{\sqrt{n}}\right)$$

$$= -\frac{t^2}{2n} + \frac{(it)^3}{6n^{3/2}} (\ln f)'''\left(\theta \frac{t}{\sqrt{n}}\right), \qquad |\theta| \le 1, \tag{7}$$

since the semi-invariants are $s_{\xi_1}^{(1)} = \mathsf{E}\xi_1 = 0$, $s_{\xi_1}^{(2)} = \sigma^2 = 1$.

In addition,

$$(\ln f(s))''' = \frac{f'''(s) \cdot f^2(s) - 3f''(s)f'(s)f(s) + 2(f'(s))^3}{f^3(s)}$$

$$= \frac{\mathsf{E}[(i\xi_1)^3 e^{i\xi_1 s}] f^2(s) - 3\mathsf{E}[(i\xi_1)^2 e^{i\xi_1 s}]\mathsf{E}[(i\xi_1)e^{i\xi_1 s}] f(s) + 2\mathsf{E}[(i\xi_1)e^{i\xi_1 s}]^3}{f^3(s)}.$$

From this, taking into account that $|f(t/\sqrt{n})| \ge 24/25$ for $|t| \le T$ and $|f(s)| \le 1$, we obtain

$$\left|(\ln f)'''\left(\theta \frac{t}{\sqrt{n}}\right)\right| \le \frac{\beta_3 + 3\beta_1 \cdot \beta_2 + 2\beta_1^3}{(\frac{24}{25})^3} \le 7\beta_3 \tag{8}$$

($\beta_k = \mathsf{E}|\xi_1|^k$, $k = 1, 2, 3$; $\beta_1 \le \beta_2^{1/2} \le \beta_3^{1/3}$; see (28), §6, Chapter II).

From (6)–(8), using the inequality $|e^z - 1| \le |z|e^{|z|}$, we find for $|t| \le T = \sqrt{n}/5\beta_3$ that

$$\left|\left[f\left(\frac{t}{\sqrt{n}}\right)\right]^n - e^{t^2/2}\right| = |e^{n \ln f(t/\sqrt{n})} - e^{-t^2/2}|$$

$$\le \left(\frac{7}{6}\right)\frac{\beta_3|t|^3}{\sqrt{n}} \exp\left\{-\frac{t^2}{2} + \left(\frac{7}{6}\right)|t|^3 \frac{\beta_3}{\sqrt{n}}\right\} \le \frac{7}{6}\frac{\beta_3|t|^3}{\sqrt{n}} e^{-t^2/4}.$$

This completes the proof of the theorem.

Remark. We observe that unless we make some supplementary hypothesis about the behavior of the random variables that are added, (2) cannot be improved. In fact, let $\xi_1, \xi_2, \ldots$ be independent identically distributed Bernoulli random variables with

$$\mathsf{P}(\xi_k = 1) = \mathsf{P}(\xi_k = -1) = \tfrac{1}{2}.$$

It is evident by symmetry that

$$2\mathsf{P}\left(\sum_{k=1}^{2n} \xi_k < 0\right) + \mathsf{P}\left(\sum_{k=1}^{2n} \xi_k = 0\right) = 1,$$

and hence, by Stirling's formula ((6), §2, chap. I)

$$\left|\mathsf{P}\left(\sum_{k=1}^{2n}\xi_k < 0\right) - \tfrac{1}{2}\right| = \tfrac{1}{2}\mathsf{P}\left(\sum_{k=1}^{2n}\xi_k = 0\right)$$

$$= \tfrac{1}{2}C_{2n}^{n}\cdot 2^{-2n} \sim \frac{1}{2\sqrt{\pi n}} = \frac{1}{\sqrt{(2\pi)\cdot(2n)}}.$$

It follows, in particular, that the constant C in (2) cannot be less than $(2\pi)^{-1/2}$.

2. Problems

1. Prove (8).

2. Let $\xi_1, \xi_2, \dots$ be independent identically distributed random variables with $\mathsf{E}\xi_k = 0$, $\mathsf{V}\xi_k = \sigma^2$ and $\mathsf{E}|\xi_1|^3 < \infty$.

   It is known that the following *nonuniform inequality* holds: for all $x \in R$,

$$|F_n(x) - \Phi(x)| \le \frac{C\mathsf{E}|\xi_1|^3}{\sigma^3\sqrt{n}}\cdot\frac{1}{(1+|x|)^3}.$$

   Prove this, at least for Bernoulli random variables.

§12. Rapidity of Convergence in Poisson's Theorem

1. Let $\xi_1, \xi_2, \dots, \xi_n$ be independent Bernoulli random variables that take the values 1 and 0 with probabilities

$$\mathsf{P}(\xi_k = 1) = p_k,\ \mathsf{P}(\xi_k = 0) = q_k (= 1 - p_k), \qquad 1 \le k \le n.$$

We set $S = \xi_1 + \cdots + \xi_n$; let $B = (B_0, B_1, \dots, B_n)$ be the binomial distribution of probabilities of the sum S, where $B_k = \mathsf{P}(S = k)$. Also let $\Pi = (\Pi_0, \Pi_1, \dots)$ be the Poisson distribution with parameter λ, where

$$\Pi_k = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k \ge 0.$$

We noticed in §6, Chapter I, that if

$$p_1 = \cdots = p_n, \qquad \lambda = np, \tag{1}$$

there is the following estimate (Prokhorov) for the *distance in variation* between the measures B and Π ($B_{n+1} = B_{n+2} = \cdots = 0$):

$$\|B - \Pi\| = \sum_{k=0}^{\infty}|B_k - \Pi_k| \le C_1(\lambda)p = C_1(\lambda)\cdot\frac{\lambda}{n}, \tag{2}$$

where

$$C_1(\lambda) = 2\min(2, \lambda). \tag{3}$$

For the case when p_k are not necessarily equal, but satisfy $\sum_{k=1}^n p_k = \lambda$, LeCam showed that

$$\|B - \Pi\| = \sum_{k=0}^{\infty} |B_k - \Pi_k| \le C_2(\lambda) \max_{1 \le k \le n} p_k, \tag{4}$$

where

$$C_2(\lambda) = 2\min(9, \lambda). \tag{5}$$

A theorem to be presented below will imply the estimate

$$\|B - \Pi\| \le C_3(\lambda) \max_{1 \le k \le n} p_k, \tag{6}$$

in which

$$C_3(\lambda) = 2\lambda. \tag{7}$$

Although $C_2(\lambda) < C_3(\lambda)$ for $\lambda > 9$, i.e., (6) is worse than (4), we nevertheless have preferred to give a proof of (6), since this proof is essentially elementary, whereas an emphasis on obtaining a "good" constant $C_2(\lambda)$ in (4) greatly complicates the proof.

2. Theorem. *Let* $\lambda = \sum_{k=1}^n p_k$. *Then*

$$\|B - \Pi\| = \sum_{k=0}^{\infty} |B_k - \Pi_k| \le 2 \sum_{k=1}^{n} p_k^2. \tag{8}$$

PROOF. We use the fact that each of the distributions B and Π is a *convolution* of distributions:

$$\begin{aligned} B &= B(p_1) * B(p_2) * \cdots * B(p_n), \\ \Pi &= \Pi(p_1) * \Pi(p_2) * \cdots * \Pi(p_n), \end{aligned} \tag{9}$$

understood as a convolution of the corresponding distribution functions (see subsection 4, §8, Chapter II), where $B(p_k) = (1 - p_k, p_k)$ is a Bernoulli distribution on the points 0 and 1, and $\Pi(p_k)$ is a Poisson distribution supported on the points 0, 1, ... with the parameters p_k.

It is easy to show that the difference $B - \Pi$ can be represented in the form

$$B - \Pi = R_1 + \cdots + R_n, \tag{10}$$

where

$$R_k = (B(p_k) - \Pi(p_k)) * F_k \tag{11}$$

with

$$F_1 = \Pi(p_2) * \cdots * \Pi(p_n),$$

....

$$F_k = B(p_1) * \cdots * B(p_{k-1}) * \Pi(p_{k+1}) * \cdots * \Pi(p_n), \qquad 2 \le k \le n-1,$$

$$F_n = B(p_1) * \cdots * B(p_{n-1})\Pi.$$

By problem 6 in §9, we have $\|R_k\| \le \|B(p_k) - \Pi(p_k)\|$. Consequently, we see immediately from (10) that

$$\|B - \Pi\| \le \sum_{k=1}^{n} \|B(p_k) - \Pi(p_k)\|. \tag{12}$$

By formula (12) in §9, we see that there is no difficulty in calculating the variation $\|B(p_k) - \Pi(p_k)\|$:

$$\begin{aligned}
&\|B(p_k) - \Pi(p_k)\| \\
&= |(1 - p_k) - e^{-p_k}| + |p_k - p_k e^{-p_k}| + \sum_{j \ge 2} \frac{p_k^j e^{-p_k}}{j!} \\
&= |(1 - p_k) - e^{-p_k}| + |p_k - p_k e^{-p_k}| + 1 - e^{-p_k} - p_k e^{-p_k} \\
&= 2p_k(1 - e^{-p_k}) \le 2p_k^2.
\end{aligned}$$

From this, together with (12), we obtain the required inequality (8).

This completes the proof of the theorem.

Corollary. *Since* $\sum_{k=1}^{n} p_k^2 \le \lambda \max_{1 \le k \le n} p_k$, *we obtain* (6).

3. Problems

1. Show that, if $\lambda_k = -\ln(1 - p_k)$,

$$\|B(p_k) - \Pi(\lambda_k)\| = 2(1 - e^{-\lambda_k} - \lambda_k e^{-\lambda_k}) \le \lambda_k^2$$

and consequently, $\|B - \Pi\| \le \sum_{k=1}^{n} \lambda_k^2$.

2. Establish the representations (9) and (10).

CHAPTER IV

Sequences and Sums of Independent Random Variables

§1. Zero-or-One Laws

1. The series $\sum_{n=1}^{\infty} (1/n)$ diverges and the series $\sum_{n=1}^{\infty} (-1)^n(1/n)$ converges. We ask the following question. What can we say about the convergence or divergence of a series $\sum_{n=1}^{\infty} (\xi_n/n)$, where $\xi_1, \xi_2, \ldots$ is a sequence of independent identically distributed Bernoulli random variables with $\mathsf{P}(\xi_1 = +1) = \mathsf{P}(\xi_1 = -1) = \frac{1}{2}$? In other words, what can be said about the convergence of a series whose general term is $\pm 1/n$, where the signs are chosen in a random manner, according to the sequence $\xi_1, \xi_2, \ldots$?

Let

$$A_1 = \left\{\omega: \sum_{n=1}^{\infty} \frac{\xi_n}{n} \text{ converges}\right\}$$

be the set of sample points for which $\sum_{n=1}^{\infty} (\xi_n/n)$ converges (to a finite number) and consider the probability $\mathsf{P}(A_1)$ of this set. It is far from clear, to begin with, what values this probability might have. However, it is a remarkable fact that we are able to say that the probability can have only two values, 0 or 1. This is a corollary of Kolmogorov's "zero-one law," whose statement and proof form the main content of the present section.

2. Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space, and let $\xi_1, \xi_2, \ldots$ be a sequence of random variables. Let $\mathscr{F}_n^{\infty} = \sigma(\xi_n, \xi_{n+1}, \ldots)$ be the σ-algebra generated by $\xi_n, \xi_{n+1}, \ldots$, and write

$$\mathscr{X} = \bigcap_{n=1}^{\infty} \mathscr{F}_n^{\infty}.$$

Since an intersection of σ-algebras is again a σ-algebra, $\mathscr{X}$ is a σ-algebra. It is called a *tail algebra* (or terminal or asymptotic algebra), because every event $A \in \mathscr{X}$ is independent of the values of $\xi_1, \ldots, \xi_n$ for every finite number n, and is determined, so to speak, only by the behavior of the infinitely remote values of $\xi_1, \xi_2, \ldots$.

Since, for every $k \geq 1$,

$$A_1 \equiv \left\{\sum_{n=1}^{\infty} \frac{\xi_n}{n} \text{ converges}\right\} = \left\{\sum_{n=k}^{\infty} \frac{\xi_n}{n} \text{ converges}\right\} \in \mathscr{F}_k^{\infty},$$

we have $A_1 \in \bigcap_k \mathscr{F}_k^{\infty} \equiv \mathscr{X}$. In the same way, if $\xi_1, \xi_2, \ldots$ is any sequence,

$$A_2 = \left\{\sum_n \xi_n \text{ converges}\right\} \in \mathscr{X}.$$

The following events are also tail events:

$$A_3 = \{\xi_n \in I_n \text{ for infinitely many } n\},$$

where $I_n \in \mathscr{B}(R)$, $n \geq 1$;

$$A_4 = \left\{\overline{\lim_n}\, \xi_n < \infty\right\};$$

$$A_5 = \left\{\overline{\lim_n}\, \frac{\xi_1 + \cdots + \xi_n}{n} < \infty\right\};$$

$$A_6 = \left\{\overline{\lim_n}\, \frac{\xi_1 + \cdots + \xi_n}{n} < c\right\};$$

$$A_7 = \left\{\frac{S_n}{n} \text{ converges}\right\};$$

$$A_8 = \left\{\overline{\lim_n}\, \frac{S_n}{\sqrt{2n \log n}} = 1\right\}.$$

On the other hand,

$$B_1 = \{\xi_n = 0 \text{ for all } n \geq 1\},$$

$$B_2 = \left\{\lim_n(\xi_1 + \cdots + \xi_n) \text{ exists and is less than } c\right\}$$

are examples of events that do not belong to $\mathscr{X}$.

Let us now suppose that our random variables are *independent*. Then by the Borel–Cantelli lemma it follows that

$$\mathsf{P}(A_3) = 0 \Leftrightarrow \sum \mathsf{P}(\xi_n \in I_n) < \infty,$$

$$\mathsf{P}(A_3) = 1 \Leftrightarrow \sum \mathsf{P}(\xi_n \in I_n) = \infty.$$

Therefore the probability of A_3 can take only the values 0 or 1 according to the convergence or divergence of $\sum \mathsf{P}(\xi_n \in I_n)$. This is Borel's zero-one law.

Theorem 1 (Kolmogorov's Zero-One Law). *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables and let $A \in \mathscr{X}$. The $\mathsf{P}(A)$ can only have one of the values zero or one.*

PROOF. The idea of the proof is to show that every tail event A is independent of itself and therefore $\mathsf{P}(A \cap A) = \mathsf{P}(A) \cdot \mathsf{P}(A)$, i.e., $\mathsf{P}(A) = \mathsf{P}^2(A)$, so that $\mathsf{P}(A) = 0$ or 1.

If $A \in \mathscr{X}$ then $A \in \mathscr{F}_1^\infty = \sigma\{\xi_1, \xi_2, \ldots\} = \sigma(\bigcup_n \mathscr{F}_1^n)$, where $\mathscr{F}_1^n = \sigma\{\xi_1, \ldots, \xi_n\}$, and we can find (Problem 8, §3, Chapter II) sets $A_n \in \mathscr{F}_1^n$, $m \geq 1$, such that $\mathsf{P}(A \triangle A_n) \to 0$, $n \to \infty$. Hence

$$\mathsf{P}(A_n) \to \mathsf{P}(A), \quad \mathsf{P}(A_n \cap A) \to \mathsf{P}(A). \tag{1}$$

But if $A \in \mathscr{X}$, the events A_n and A are independent for every $n \geq 1$. Hence it follows from (1) that $\mathsf{P}(A) = \mathsf{P}^2(A)$ and therefore $\mathsf{P}(A) = 0$ or 1.

This completes the proof of the theorem.

Corollary. *Let η be a random variable that is measurable with respect to the tail σ-algebra $\mathscr{X}$, i.e., $\{\eta \in B\} \in \mathscr{X}$, $B \in \mathscr{B}(R)$. Then η is degenerate, i.e., there is a constant c such that $\mathsf{P}(\eta = c) = 1$.*

3. Theorem 2 below provides an example of a nontrivial application of Kolmogorov's zero-one law.

Let $\xi_1, \xi_2, \ldots$ be a sequence of independent Bernoulli random variables with $\mathsf{P}(\xi_n = 1) = p$, $\mathsf{P}(\xi_n = -1) = q$, $p + q = 1$, $n \geq 1$, and let $S_n = \xi_1 + \cdots + \xi_n$. It seems intuitively clear that in the symmetric case ($p = \frac{1}{2}$) a "typical" path of the random walk S_n, $n \geq 1$, will cross zero infinitely often, whereas when $p \neq \frac{1}{2}$ it will go off to infinity. Let us give a precise formulation.

Theorem 2. (a) *If $p = \frac{1}{2}$ then* $\mathsf{P}(S_n = 0 \text{ i.o.}) = 1$.
(b) *If $p \neq \frac{1}{2}$, then* $\mathsf{P}(S_n = 0 \text{ i.o.}) = 0$.

PROOF. We first observe that the event $B = (S_n = 0 \text{ i.o.})$ is not a tail event, i.e., $B \notin \mathscr{X} = \bigcap \mathscr{F}_n^\infty$, $\mathscr{F}_n^\infty = \sigma\{\xi_n, \xi_{n+1}, \ldots\}$. Consequently it is, in principle, not clear that B should have only the values 0 or 1.

Statement (b) is easily proved by applying (the first part of) the Borel–Cantelli lemma. In fact, if $B_{2n} = \{S_{2n} = 0\}$, then by Stirling's formula

$$\mathsf{P}(B_{2n}) = C_{2n}^n p^n q^n \sim \frac{(4pq)^n}{\sqrt{\pi n}}$$

and therefore $\sum \mathsf{P}(B_{2n}) < \infty$. Consequently $\mathsf{P}(S_n = 0 \text{ i.o.}) = 0$.

To prove (a), it is enough to prove that the event

$$A = \left\{\overline{\lim} \frac{S_n}{\sqrt{n}} = \infty, \underline{\lim} \frac{S_n}{\sqrt{n}} = -\infty\right\}$$

has probability 1 (since $A \subseteq B$).

Let

$$A_c = \left\{\overline{\lim} \frac{S_n}{\sqrt{n}} > c\right\} \cap \left\{\underline{\lim} \frac{S_n}{\sqrt{n}} < -c\right\} (= A'_c \cap A''_c).$$

Then $A_c \downarrow A$, $c \to \infty$, and all the events A, A_c, A'_c, A''_c are tail events. Let us show that $\mathsf{P}(A'_c) = \mathsf{P}(A''_c) = 1$ for each $c > 0$. Since $A'_c \in \mathscr{X}$ and $A''_c \in \mathscr{X}$, it is sufficient to show only that $\mathsf{P}(A'_c) > 0$, $\mathsf{P}(A''_c) > 0$. But by Problem 5

$$\mathsf{P}\left(\underline{\lim} \frac{S_n}{\sqrt{n}} < -c\right) = \mathsf{P}\left(\overline{\lim} \frac{S_n}{\sqrt{n}} > c\right) \geq \overline{\lim}\, \mathsf{P}\left(\frac{S_n}{\sqrt{n}} > c\right) > 0,$$

where the last inequality follows from the De Moivre–Laplace theorem.

Thus $P(A_c) = 1$ for all $c > 0$ and therefore $P(A) = \lim_{c\to\infty} P(A_c) = 1$.

This completes the proof of the theorem.

4. Let us observe again that $B = \{S_n = 0 \text{ i.o.}\}$ is not a tail event. Nevertheless, it follows from Theorem 2 that, for a Bernoulli scheme, the probability of this event, just as for tail events, takes only the values 0 and 1. This phenomenon is not accidental: it is a corollary of the Hewitt–Savage zero-one law, which for independent identically distributed random variables extends the result of Theorem 1 to the class of "symmetric" events (which includes the class of tail events).

Let us give the essential definitions. A one-to-one mapping $\pi = (\pi_1, \pi_2, \ldots)$ of the set $(1, 2, \ldots)$ on itself is said to be a finite permutation if $\pi_n = n$ for every n with a finite number of exceptions.

If $\xi = \xi_1, \xi_2, \ldots$ is a sequence of random variables, $\pi(\xi)$ denotes the sequence $(\xi_{\pi_1}, \xi_{\pi_2}, \ldots)$. If A is the event $\{\xi \in B\}$, $B \in \mathscr{B}(R^\infty)$, then $\pi(A)$ denotes the event $\{\pi(\xi) \in B\}$, $B \in \mathscr{B}(R^\infty)$.

We call an event $A = \{\xi \in B\}$, $B \in \mathscr{B}(R^\infty)$, *symmetric* if $\pi(A)$ coincides with A for every finite permutation π.

An example of a symmetric event is $A = \{S_n = 0 \text{ i.o.}\}$, where $S_n = \xi_1 + \cdots + \xi_n$. Moreover, we may suppose (Problem 4) that every event in the tail σ-algebra $\mathscr{X}(S) = \bigcap \mathscr{F}_n^\infty(S) = \sigma\{\omega: S_n, S_{n+1}, \ldots\}$ generated by $S_1 = \xi_1, S_2 = \xi_1 + \xi_2, \ldots$ is symmetric.

Theorem 3 (Hewitt–Savage Zero-One Law). *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables, and*

$$A = \{\omega: (\xi_1, \xi_2, \ldots) \in B\}$$

a symmetric event. Then $P(A) = 0$ or 1.

PROOF. Let $A = \{\xi \in B\}$ be a symmetric event. Choose sets $B_n \in \mathscr{B}(R^n)$ such that, for $A_n = \{\omega: (\xi_1, \ldots, \xi_n) \in B_n\}$,

$$P(A \triangle A_n) \to 0, \qquad n \to \infty. \tag{2}$$

Since the random variables $\xi_1, \xi_2, \ldots$ are independent and identically distributed, the probability distributions $P_\xi(B) = P(\xi \in B)$ and $P_{\pi_n(\xi)}(B) = P(\pi_n(\xi) \in B)$ coincide. Therefore

$$\mathsf{P}(A \triangle A_n) = P_\xi(B \triangle B_n) = P_{\pi_n(\xi)}(B \triangle B_n). \tag{3}$$

Since A is symmetric, we have

$$A \equiv \{\xi \in B\} = \pi_n(A) \equiv \{\pi_n(\xi) \in B\}.$$

Therefore

$$\begin{aligned} P_{\pi_n(\xi)}(B \triangle B_n) &= \mathsf{P}\{\pi_n(\xi) \in B) \triangle (\pi_n(\xi) \in B_n)\} \\ &= \mathsf{P}\{(\xi \in B) \triangle (\pi_n(\xi) \in B_n)\} = \mathsf{P}\{A \triangle \pi_n(A_n)\}. \end{aligned} \tag{4}$$

Hence, by (3) and (4),

$$\mathsf{P}(A \triangle A_n) = \mathsf{P}(A \triangle \pi_n(A_n)). \tag{5}$$

It then follows from (2) that

$$\mathsf{P}(A \triangle (A_n \cap \pi_n(A_n))) \to 0, \qquad n \to \infty. \tag{6}$$

Hence, by (2), (5), and (6), we obtain

$$\begin{gathered} \mathsf{P}(A_n) \to \mathsf{P}(A), \qquad \mathsf{P}(\pi_n(A)) \to \mathsf{P}(A), \\ \mathsf{P}(A_n \cap \pi_n(A_n)) \to \mathsf{P}(A). \end{gathered} \tag{7}$$

Moreover, since ξ_1 and ξ_2 are independent,

$$\begin{aligned} \mathsf{P}(A_n \cap \pi_n(A_n)) &= \mathsf{P}\{(\xi_1, \ldots, \xi_n) \in B_n, (\xi_{n+1}, \ldots, \xi_{2n}) \in B_n\} \\ &= \mathsf{P}\{(\xi_1, \ldots, \xi_n) \in B_n\} \cdot \mathsf{P}\{(\xi_{n+1}, \ldots, \xi_{2n}) \in B_n\} \\ &= \mathsf{P}(A_n)\mathsf{P}(\pi_n(A_n)), \end{aligned}$$

whence by (7)

$$\mathsf{P}(A) = \mathsf{P}^2(A)$$

and therefore $\mathsf{P}(A) = 0$ or 1.

This completes the proof of the theorem.

5. Problems

1. Prove the corollary to Theorem 1.

2. Show that if (ξ_n) is a sequence of independent random variables, the random variables $\overline{\lim}\, \xi_n$ and $\underline{\lim}\, \xi_n$ are degenerate.

3. Let (ξ_n) be a sequence of independent random variables, $S_n = \xi_1 + \cdots + \xi_n$, and let the constants b_n satisfy $0 < b_n \uparrow \infty$. Show that the random variables $\overline{\lim}(S_n/b_n)$ and $\underline{\lim}(S_n/b_n)$ are degenerate.

4. Let $S_n = \xi_1 + \cdots + \xi_n$, $n \geq 1$, and $\mathscr{X}(S) = \bigcap \mathscr{F}_n^\infty(S)$, $\mathscr{F}_n^\infty(S) = \sigma\{\omega: S_n, S_{n+1}, \ldots\}$. Show that every event in $\mathscr{X}(S)$ is symmetric.

5. Let (ξ_n) be a sequence of random variables. Show that $\{\overline{\lim}\, \xi_n > c\} \supseteq \overline{\lim}\{\xi_n > c\}$ for each $c > 0$.

§2. Convergence of Series

1. Let us suppose that $\xi_1, \xi_2, \ldots$ is a sequence of independent random variables, $S_n = \xi_1 + \cdots + \xi_n$, and let A be the set of sample points ω for which $\sum \xi_n(\omega)$ converges to a finite limit. It follows from Kolmogorov's zero-one law that $\mathsf{P}(A) = 0$ or 1, i.e., the series $\sum \xi_n$ converges or diverges with probability 1. The object of the present section is to give criteria that will determine whether a sum of independent random variables converges or diverges.

Theorem 1 (Kolmogorov and Khinchin).
(a) *Let* $\mathsf{E}\xi_n = 0$, $n \geq 1$. *Then if*

$$\sum \mathsf{E}\xi_n^2 < \infty, \tag{1}$$

the series $\sum \xi_n$ *converges with probability* 1.
(b) *If the random variables* ξ_n, $n \geq 1$, *are uniformly bounded* (i.e., $\mathsf{P}(|\xi_n| \leq c) = 1$, $c < \infty$), *the converse is true: the convergence of* $\sum \xi_n$ *with probability* 1 *implies* (1).

The proof depends on

Kolmogorov's Inequality

(a) *Let* $\xi_1, \xi_2, \ldots, \xi_n$ *be independent random variables with* $\mathsf{E}\xi_i = 0$, $\mathsf{E}\xi_i^2 < \infty$, $i \leq n$. *Then for every* $\varepsilon > 0$

$$\mathsf{P}\left\{\max_{1 \leq k \leq n} |S_k| \geq \varepsilon\right\} \leq \frac{\mathsf{E}S_n^2}{\varepsilon^2}. \tag{2}$$

(b) *If also* $\mathsf{P}(|\xi_i| \leq c) = 1$, $i \leq n$, *then*

$$\mathsf{P}\left\{\max_{1 \leq k \leq n} |S_k| \geq \varepsilon\right\} \geq 1 - \frac{(c + \varepsilon)^2}{\mathsf{E}S_n^2}. \tag{3}$$

PROOF. (a) Put

$$A = \{\max |S_k| \geq \varepsilon\},$$
$$A_k = \{|S_i| < \varepsilon, i = 1, \ldots, k-1, |S_k| \geq \varepsilon\}, \qquad 1 \leq k \leq n.$$

Then $A = \sum A_k$ and

$$\mathsf{E}S_n^2 \geq \mathsf{E}S_n^2 I_A = \sum \mathsf{E}S_n^2 I_{A_k}.$$

But

$$\begin{aligned}\mathsf{E}S_n^2 I_{A_k} &= \mathsf{E}(S_k + (\xi_{k+1} + \cdots + \xi_n))^2 I_{A_k}\\ &= \mathsf{E}S_k^2 I_{A_k} + 2\mathsf{E}S_k(\xi_{k+1} + \cdots + \xi_n)I_{A_k} + \mathsf{E}(\xi_{k+1} + \cdots + \xi_n)^2 I_{A_k}\\ &\geq \mathsf{E}S_k^2 I_{A_k},\end{aligned}$$

since

$$\mathsf{E}S_k(\xi_{k+1} + \cdots + \xi_n)I_{A_k} = \mathsf{E}S_k I_{A_k} \cdot \mathsf{E}(\xi_{k+1} + \cdots + \xi_n) = 0$$

because of independence and the conditions $\mathsf{E}\xi_i = 0, i \leq n$. Hence

$$\mathsf{E}S_n^2 \geq \sum \mathsf{E}S_k^2 I_{A_k} \geq \varepsilon^2 \sum \mathsf{P}(A_k) = \varepsilon^2 \mathsf{P}(A),$$

which proves the first inequality.

(b) To prove (3), we observe that

$$\mathsf{E}S_n^2 I_A = \mathsf{E}S_n^2 - \mathsf{E}S_n^2 I_{\bar{A}} \geq \mathsf{E}S_n^2 - \varepsilon^2 \mathsf{P}(\bar{A}) = \mathsf{E}S_n^2 - \varepsilon^2 + \varepsilon^2 \mathsf{P}(A). \tag{4}$$

On the other hand, on the set A_k

$$|S_{k-1}| \leq \varepsilon, \qquad |S_k| \leq |S_{k-1}| + |\xi_k| \leq \varepsilon + c$$

and therefore

$$\begin{aligned}\mathsf{E}S_n^2 I_A &= \sum_k \mathsf{E}S_k^2 I_{A_k} + \sum_k \mathsf{E}(I_{A_k}(S_n - S_k)^2)\\ &\leq (\varepsilon + c)^2 \sum_k \mathsf{P}(A_k) + \sum_{k=1}^n \mathsf{P}(A_k) \sum_{j=k+1}^n \mathsf{E}\xi_j^2\\ &\leq \mathsf{P}(A)\left[(\varepsilon + c)^2 + \sum_{j=1}^n \mathsf{E}\xi_j^2\right] = \mathsf{P}(A)[(\varepsilon + c)^2 + \mathsf{E}S_n^2].\end{aligned} \tag{5}$$

From (4) and (5) we obtain

$$\mathsf{P}(A) \geq \frac{\mathsf{E}S_n^2 - \varepsilon^2}{(\varepsilon + c)^2 + \mathsf{E}S_n^2 - \varepsilon^2} = 1 - \frac{(\varepsilon + c)^2}{(\varepsilon + c)^2 + \mathsf{E}S_n^2 - \varepsilon^2} \geq 1 - \frac{(\varepsilon + c)^2}{\mathsf{E}S_n^2}.$$

This completes the proof of (3).

Proof of Theorem 1. (a) By Theorem 4 of §10, Chapter II, the sequence (S_n), $n \geq 1$, converges with probability 1, if and only if it is fundamental with probability 1. By Theorem 1 of §10, Chapter II, the sequence (S_n), $n \geq 1$, is fundamental (P-a.s.) if and only if

$$\mathsf{P}\left\{\sup_{k\geq 1} |S_{n+k} - S_n| \geq \varepsilon\right\} \to 0, \qquad n \to \infty. \tag{6}$$

By (2),

$$\begin{aligned}\mathsf{P}\left\{\sup_{k\geq 1} |S_{n+k} - S_n| \geq \varepsilon\right\} &= \lim_{N\to\infty} \mathsf{P}\left\{\max_{1\leq k\leq N} |S_{n+k} - S_n| \geq \varepsilon\right\}\\ &\leq \lim_{N\to\infty} \frac{\sum_{k=n}^{n+N} \mathsf{E}\xi_k^2}{\varepsilon^2} = \frac{\sum_{k=n}^{\infty} \mathsf{E}\xi_k^2}{\varepsilon^2}.\end{aligned}$$

Therefore (6) is satisfied if $\sum_{k=1}^{\infty} \mathsf{E}\xi_k^2 < \infty$, and consequently $\sum \xi_k$ converges with probability 1.

(b) Let $\sum \xi_k$ converge. Then, by (6), for sufficiently large n,

$$\mathsf{P}\left\{\sup_{k\geq 1}|S_{n+k}-S_n|\geq \varepsilon\right\}<\tfrac{1}{2}. \tag{7}$$

By (3),

$$\mathsf{P}\left\{\sup_{k\geq 1}|S_{n+k}-S_n|\geq \varepsilon\right\}\geq 1-\frac{(c+\varepsilon)^2}{\sum_{k=n}^{\infty}\mathsf{E}\xi_k^2}.$$

Therefore if we suppose that $\sum_{k=1}^{\infty}\mathsf{E}\xi_k^2=\infty$, we obtain

$$\mathsf{P}\left\{\sup_{k\geq 1}|S_{n+k}-S_n|\geq \varepsilon\right\}=1,$$

which contradicts (7).

This completes the proof of the theorem.

Example. If $\xi_1, \xi_2, \ldots$ is a sequence of independent Bernoulli random variables with $\mathsf{P}(\xi_n=+1)=\mathsf{P}(\xi_n=-1)=\frac{1}{2}$, then the series $\sum \xi_n a_n$, with $|a_n|\leq c$, converges with probability 1, if and only if $\sum a_n^2<\infty$.

2. Theorem 2 (Two-Series Theorem). *A sufficient condition for the convergence of the series $\sum \xi_n$ of independent random variables, with probability* 1, *is that both series $\sum \mathsf{E}\xi_n$ and $\sum \mathsf{V}\xi_n$ converge. If $\mathsf{P}(|\xi_n|\leq c)=1$, the condition is also necessary.*

Proof. If $\sum \mathsf{V}\xi_n<\infty$, then by Theorem 1 the series $\sum(\xi_n-\mathsf{E}\xi_n)$ converges (P-a.s.). But by hypothesis the series $\sum \mathsf{E}\xi_n$ converges; hence $\sum \xi_n$ converges (P-a.s.).

To prove the necessity we use the following symmetrization method. In addition to the sequence $\xi_1, \xi_2, \ldots$ we consider a different sequence $\tilde{\xi}_1, \tilde{\xi}_2, \ldots$ of independent random variables such that $\tilde{\xi}_n$ has the same distribution as ξ_n, $n\geq 1$. (When the original sample space is sufficiently rich, the existence of such a sequence follows from Theorem 1 of §9, Chapter II. We can also show that this assumption involves no loss of generality.)

Then if $\sum \xi_n$ converges (P-a.s.), the series $\sum \tilde{\xi}_n$ also converges, and hence so does $\sum(\xi_n-\tilde{\xi}_n)$. But $\mathsf{E}(\xi_n-\tilde{\xi}_n)=0$ and $\mathsf{P}(|\xi_n-\tilde{\xi}_n|\leq 2c)=1$. Therefore $\sum \mathsf{V}(\xi_n-\tilde{\xi}_n)<\infty$ by Theorem 1. In addition,

$$\sum \mathsf{V}\xi_n=\tfrac{1}{2}\sum \mathsf{V}(\xi_n-\tilde{\xi}_n)<\infty.$$

Consequently, by Theorem 1, $\sum(\xi_n-\mathsf{E}\xi_n)$ converges with probability 1, and therefore $\sum \mathsf{E}\xi_n$ converges.

Thus if $\sum \xi_n$ converges (P-a.s.) (and $\mathsf{P}(|\xi_n|\leq c)=1$, $n\geq 1$) it follows that both $\sum \mathsf{E}\xi_n$ and $\sum \mathsf{V}\xi_n$ converge.

This completes the proof of the theorem.

3. The following theorem provides a necessary and sufficient condition for the convergence of $\sum \xi_n$ without any boundedness condition on the random variables.

Let c be a constant and

$$\xi^c = \begin{cases} \xi, & |\xi| \le c, \\ 0, & |\xi| > c. \end{cases}$$

Theorem 3 (Kolmogorov's Three-Series Theorem). *Let $\xi_1, \xi_2, \dots$ be a sequence of independent random variables. A necessary condition for the convergence of $\sum \xi_n$ with probability 1 is that the series*

$$\sum \mathsf{E}\xi_n^c, \qquad \sum \mathsf{V}\xi_n^c, \qquad \sum \mathsf{P}(|\xi_n| \ge c)$$

converge for every $c > 0$; a sufficient condition is that these series converge for some $c > 0$.

Proof. *Sufficiency.* By the two-series theorem, $\sum \xi_n^c$ converges with probability 1. But if $\sum \mathsf{P}(|\xi_n| \ge c) < \infty$, then by the Borel–Cantelli lemma we have $\sum I(|\xi_n| \ge c) < \infty$ with probability 1. Therefore $\xi_n = \xi_n^c$ for all n with at most finitely many exceptions. Therefore $\sum \xi_n$ also converges (P-a.s.).

Necessity. If $\sum \xi_n$ converges (P-a.s.) then $\xi_n \to 0$ (P-a.s.), and therefore, for every $c > 0$, at most a finite number of the events $\{|\xi_n| \ge c\}$ can occur (P-a.s.). Therefore $\sum I(|\xi_n| \ge c) < \infty$ (P-a.s.), and, by the second part of the Borel–Cantelli lemma, $\sum \mathsf{P}(|\xi_n| > c) < \infty$. Moreover, the convergence of $\sum \xi_n$ implies the convergence of $\sum \xi_n^c$. Therefore, by the two-series theorem, both of the series $\sum \mathsf{E}\xi_n^c$ and $\sum \mathsf{V}\xi_n^c$ converge.

This completes the proof of the theorem.

Corollary. *Let $\xi_1, \xi_2, \dots$ be independent variables with $\mathsf{E}\xi_n = 0$. Then if*

$$\sum \mathsf{E}\frac{\xi_n^2}{1 + |\xi_n|} < \infty,$$

the series $\sum \xi_n$ converges with probability 1.

For the proof we observe that

$$\sum \mathsf{E}\frac{\xi_n^2}{1 + |\xi_n|} < \infty \Leftrightarrow \sum \mathsf{E}[\xi_n^2 I(|\xi_n| \le 1) + |\xi_n| I(|\xi_n| > 1)] < \infty.$$

Therefore if $\xi_n^1 = \xi_n I(|\xi_n| \le 1)$, we have

$$\sum \mathsf{E}(\xi_n^1)^2 < \infty.$$

Since $\mathsf{E}\xi_n = 0$, we have

$$\begin{aligned} \sum |\mathsf{E}\xi_n^1| &= \sum |\mathsf{E}\xi_n I(|\xi_n| \le 1)| = \sum |\mathsf{E}\xi_n I(|\xi_n| > 1)| \\ &\le \sum \mathsf{E}|\xi_n| I(|\xi_n| > 1) < \infty. \end{aligned}$$

Therefore both $\sum \mathsf{E}\xi_n^1$ and $\sum \mathsf{V}\xi_n^1$ converge. Moreover, by Chebyshev's inequality,

$$\mathsf{P}\{|\xi_n| > 1\} = \mathsf{P}\{|\xi_n| I(|\xi_n| > 1) > 1\} \le \mathsf{E}(|\xi_n| I(|\xi_n| > 1)).$$

Therefore $\sum \mathsf{P}(|\xi_n| > 1) < \infty$. Hence the convergence of $\sum \xi_n$ follows from the three-series theorem.

4. Problems

1. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables, $S_n = \xi_1, \ldots, \xi_n$. Show, using the three-series theorem, that
   (a) if $\sum \xi_n^2 < \infty$ (P-a.s.) then $\sum \xi_n$ converges with probability 1, if and only if $\sum \mathsf{E}\, \xi_i I(|\xi_i| \le 1)$ converges;
   (b) if $\sum \xi_n$ converges (P-a.s.) then $\sum \xi_n^2 < \infty$ (P-a.s.) if and only if
$$\sum (\mathsf{E}\,|\xi_n| I(|\xi_n| \le 1))^2 < \infty.$$
2. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables. Show that $\sum \xi_n^2 < \infty$ (P-a.s.) if and only if
$$\sum \mathsf{E}\, \frac{\xi_n^2}{1 + \xi_n^2} < \infty.$$
3. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables. Show that $\sum \xi_n$ converges (P-a.s.) if and only if it converges in probability.

§3. Strong Law of Large Numbers

1. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with finite second moments; $S_n = \xi_1 + \cdots + \xi_n$. By Problem 2, §3, Chapter III, if the numbers $\mathsf{V}\xi_i$ are uniformly bounded, we have the law of large numbers:
$$\frac{S_n - \mathsf{E}S_n}{n} \xrightarrow{\mathsf{P}} 0, \qquad n \to \infty. \tag{1}$$

A *strong law of large numbers* is a proposition in which convergence in probability is replaced by *convergence with probability* 1.

One of the earliest results in this direction is the following theorem.

Theorem 1 (Cantelli). *Let $\xi_1, \xi_2, \ldots$ be independent random variables with finite fourth moments and let*
$$\mathsf{E}\,|\xi_n - \mathsf{E}\xi_n|^4 \le C, \qquad n \ge 1,$$
for some constant C. Then as $n \to \infty$
$$\frac{S_n - \mathsf{E}S_n}{n} \to 0 \quad (\mathsf{P}\text{-}a.s.). \tag{2}$$

Proof. Without loss of generality, we may assume that $\mathsf{E}\xi_n = 0$ for $n \ge 1$. By the corollary to Theorem 1, §10, Chapter II, we will have $S_n/n \to 0$ (P-a.s.) provided that
$$\sum \mathsf{P}\left\{\left|\frac{S_n}{n}\right| \ge \varepsilon\right\} < \infty$$
for every $\varepsilon > 0$. In turn, by Chebyshev's inequality, this will follow from
$$\sum \mathsf{E}\left|\frac{S_n}{n}\right|^4 < \infty.$$

Let us show that this condition is actually satisfied under our hypotheses.

We have

$$S_n^4 = (\xi_1 + \cdots + \xi_n)^4 = \sum_{i=1}^n \xi_i^4 - \sum_{\substack{i,j \\ i<j}} \frac{4!}{2!2!}\xi_i^2\xi_j^2$$
$$+ \sum_{\substack{i\neq j \\ i\neq k \\ j<k}} \frac{4!}{2!1!1!}\xi_i^2\xi_j\xi_k + \sum_{i<j<k<l} 4!\,\xi_i\xi_j\xi_k\xi_l$$
$$+ \sum_{i\neq j} \frac{4!}{3!1!}\xi_i^3\xi_j.$$

Remembering that $\mathsf{E}\xi_k = 0$, $k \leq n$, we then obtain

$$\mathsf{E}S_n^4 = \sum_{i=1}^n \mathsf{E}\xi_i^4 + 6\sum_{i,j=1}^n \mathsf{E}\xi_i^2\mathsf{E}\xi_j^2 \leq nC + 6\sum_{\substack{i,j=1 \\ i<j}} \sqrt{\mathsf{E}\xi_i^4 \cdot \mathsf{E}\xi_j^4}$$
$$\leq nC + \frac{6n(n-1)}{2}C = (3n^2 - 2n)C < 3n^2C.$$

Consequently

$$\sum \mathsf{E}\left(\frac{S_n}{n}\right)^4 \leq 3C\sum\frac{1}{n^2} < \infty.$$

This completes the proof of the theorem.

2. The hypotheses of Theorem 1 can be considerably weakened by the use of more precise methods. In this way we obtain a stronger law of large numbers.

Theorem 2 (Kolmogorov). *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with finite second moments, and let there be positive numbers b_n such that $b_n \uparrow \infty$ and*

$$\sum \frac{\mathsf{V}\xi_n}{b_n^2} < \infty. \tag{3}$$

Then

$$\frac{S_n - \mathsf{E}S_n}{b_n} \to 0 \quad (\mathsf{P}\text{-}a.s.). \tag{4}$$

In particular, if

$$\sum \frac{\mathsf{V}\xi_n}{n^2} < \infty \tag{5}$$

then

$$\frac{S_n - \mathsf{E}S_n}{n} \to 0 \quad (\mathsf{P}\text{-}a.s.). \tag{6}$$

For the proof of this, and of Theorem 2 below, we need two lemmas.

Lemma 1 (Toeplitz). *Let $\{a_n\}$ be a sequence of nonnegative numbers, $b_n = \sum_{i=1}^n a_i$, $b_n > 0$ for $n \geq 1$, and $b_n \uparrow \infty$, $n \to \infty$. Let $\{x_n\}$ be a sequence of numbers converging to x. Then*

$$\frac{1}{b_n}\sum_{j=1}^n a_j x_j \to x. \tag{7}$$

In particular, if $a_n = 1$ then

$$\frac{x_1 + \cdots + x_n}{n} \to x. \tag{8}$$

PROOF. Let $\varepsilon > 0$ and let $n_0 = n_0(\varepsilon)$ be such that $|x_n - x| \leq \varepsilon/2$ for all $n \geq n_0$. Choose $n_1 > n_0$ so that

$$\frac{1}{b_{n_1}}\sum_{j=1}^{n_0} |x_j - x| < \varepsilon/2.$$

Then, for $n > n_1$,

$$\begin{aligned}
\left|\frac{1}{b_n}\sum_{j=1}^n a_j x_j - x\right| &\leq \frac{1}{b_n}\sum_{j=1}^n a_j|x_j - x| \\
&= \frac{1}{b_n}\sum_{j=1}^{n_0} a_j|x_j - x| + \frac{1}{b_n}\sum_{j=n_0+1}^{n} a_j|x_j - x| \\
&\leq \frac{1}{b_{n_1}}\sum_{j=1}^{n_0} a_j|x_j - x| + \frac{1}{b_n}\sum_{j=n_0+1}^{n} a_j|x_j - x| \\
&\leq \frac{\varepsilon}{2} + \frac{b_n - b_{n_0}}{b_n}\frac{\varepsilon}{2} \leq \varepsilon.
\end{aligned}$$

This completes the proof of the lemma.

Lemma 2 (Kronecker). *Let $\{b_n\}$ be a sequence of positive increasing numbers, $b_n \uparrow \infty$, $n \to \infty$, and let $\{x_n\}$ be a sequence of numbers such that $\sum x_n$ converges. Then*

$$\frac{1}{b_n}\sum_{j=1}^n b_j x_j \to 0, \qquad n \to \infty. \tag{9}$$

In particular, if $b_n = n$, $x_n = y_n/n$ and $\sum (y_n/n)$ converges, then

$$\frac{y_1 + \cdots + y_n}{n} \to 0, \qquad n \to \infty. \tag{10}$$

PROOF. Let $b_0 = 0$, $S_0 = 0$, $S_n = \sum_{j=1}^n x_j$. Then (by summation by parts)

$$\sum_{j=1}^n b_j x_j = \sum_{j=1}^n b_j(S_j - S_{j-1}) = b_n S_n - b_0 S_0 - \sum_{j=1}^n S_{j-1}(b_j - b_{j-1})$$

and therefore

$$\frac{1}{b_n}\sum_{j=1}^{n} b_j x_j = S_n - \frac{1}{b_n}\sum_{j=1}^{n} S_{j-1}a_j \to 0,$$

since, if $S_n \to x$, then by Toeplitz's lemma,

$$\frac{1}{b_n}\sum_{j=1}^{n} S_{j-1}a_j \to x.$$

This establishes the lemma.

PROOF OF THEOREM 1. Since

$$\frac{S_n - \mathsf{E}S_n}{b_n} = \frac{1}{b_n}\sum_{k=1}^{n} b_k\left(\frac{\xi_k - \mathsf{E}\xi_k}{b_k}\right),$$

a sufficient condition for (4) is, by Kronecker's lemma, that the series $\sum [(\xi_k - \mathsf{E}\xi_k)/b_k]$ converges (P-a.s.). But this series does converge by (3) of Theorem 1, §2.

This completes the proof of the theorem.

EXAMPLE 1. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent Bernoulli random variables with $\mathsf{P}(\xi_n = 1) = P(\xi_n = -1) = \frac{1}{2}$. Then, since $\sum [1/(n \log^2 n)] < \infty$, we have

$$\frac{S_n}{\sqrt{n}\log n} \to 0 \quad (\text{P-a.s.}). \tag{11}$$

3. In the case when the variables $\xi_1, \xi_2, \ldots$ are not only independent but also identically distributed, we can obtain a strong law of large numbers without requiring (as in Theorem 2) the existence of the second moment, provided that the first absolute moment exists.

Theorem 3 (Kolmogorov). *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables with $\mathsf{E}|\xi_1| < \infty$. Then*

$$\frac{S_n}{n} \to m \quad (\text{P-a.s.}) \tag{12}$$

where $m = \mathsf{E}\xi_1$.

We need the following lemma.

Lemma 3. *Let ξ be a nonnegative random variable. Then*

$$\sum_{n=1}^{\infty} \mathsf{P}(\xi \ge n) \le \mathsf{E}\xi \le 1 + \sum_{n=1}^{\infty} \mathsf{P}(\xi \ge n). \tag{13}$$

The proof consists of the following chain of inequalities:

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathsf{P}(\xi \geq n) &= \sum_{n=1}^{\infty} \sum_{k \geq n} \mathsf{P}(k \leq \xi < k+1) \\
&= \sum_{k=1}^{\infty} k\mathsf{P}(k \leq \xi < k+1) = \sum_{k=0}^{\infty} \mathsf{E}[kI(k \leq \xi < k+1)] \\
&\leq \sum_{k=0}^{\infty} \mathsf{E}[\xi I(k \leq \xi < k+1)] \\
&= \mathsf{E}\xi \leq \sum_{k=0}^{\infty} \mathsf{E}[(k+1)I(k \leq \xi < k+1)] \\
&= \sum_{k=0}^{\infty} (k+1)\mathsf{P}(k \leq \xi < k+1) \\
&= \sum_{n=1}^{\infty} \mathsf{P}(\xi \geq n) + \sum_{k=0}^{\infty} \mathsf{P}(k \leq \xi < k+1) = \sum_{n=1}^{\infty} \mathsf{P}(\xi \geq n) + 1.
\end{aligned}$$

Proof of Theorem 3. By Lemma 3 and the Borel–Cantelli lemma,

$$\begin{aligned}
\mathsf{E}|\xi_1| < \infty &\Leftrightarrow \sum \mathsf{P}\{|\xi_1| \geq n\} < \infty \\
&\Leftrightarrow \sum \mathsf{P}\{|\xi_n| \geq n\} < \infty \Leftrightarrow \mathsf{P}\{|\xi_n| \geq n \text{ i.o.}\} = 0.
\end{aligned}$$

Hence $|\xi_n| < n$, except for a finite number of n, with probability 1.

Let us put

$$\tilde{\xi}_n = \begin{cases} \xi_n, & |\xi_n| < n, \\ 0, & |\xi_n| \geq n, \end{cases}$$

and suppose that $\mathsf{E}\xi_n = 0$, $n \geq 1$. Then $(\xi_1 + \cdots + \xi_n)/n \to 0$ (P-a.s.), if and only if $(\tilde{\xi}_1 + \cdots + \tilde{\xi}_n)/n \to 0$ (P-a.s.). Note that in general $\mathsf{E}\tilde{\xi}_n \neq 0$ but

$$\mathsf{E}\tilde{\xi}_n = \mathsf{E}\xi_n I(|\xi_n| < n) = \mathsf{E}\xi_1 I(|\xi_1| < n) \to \mathsf{E}\xi_1 = 0.$$

Hence by Toeplitz's lemma

$$\frac{1}{n} \sum_{k=1}^{n} \mathsf{E}\tilde{\xi}_k \to 0, \qquad n \to \infty,$$

and consequently $(\xi_1 + \cdots + \xi_n)/n \to 0$ (P-a.s.), if and only if

$$\frac{(\tilde{\xi}_1 - \mathsf{E}\tilde{\xi}_1) + \cdots + (\tilde{\xi}_n - \mathsf{E}\tilde{\xi}_n)}{n} \to 0, \qquad n \to \infty \quad (\mathsf{P}\text{-a.s.}), \quad n \to \infty. \quad (14)$$

Write $\bar{\xi}_n = \tilde{\xi}_n - \mathsf{E}\tilde{\xi}_n$. By Kronecker's lemma, (14) will be established if $\sum (\bar{\xi}_n/n)$ converges (P-a.s.). In turn, by Theorem 1 of §2, this will follow if we show that, when $\mathsf{E}|\xi_1| < \infty$, the series $\sum (\mathsf{V}\bar{\xi}_n/n^2)$ converges.

We have

$$\sum \frac{\mathsf{V}\xi_n}{n^2} \le \sum_{n=1}^{\infty} \frac{\mathsf{E}\xi_n^2}{n^2} = \sum_{n=1}^{\infty} \frac{1}{n^2} \mathsf{E}[\xi_n I(|\xi_n| < n)]^2$$

$$= \sum_{n=1}^{\infty} \frac{1}{n^2} \mathsf{E}[\xi_1^2 I(|\xi_1| < n)] = \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{k=1}^{n} \mathsf{E}[\xi_1^2 I(k-1) \le |\xi_1| < k)]$$

$$= \sum_{k=1}^{\infty} \mathsf{E}[\xi_1^2 I(k-1 \le |\xi_1| < k)] \cdot \sum_{n=k}^{\infty} \frac{1}{n^2}$$

$$\le 2 \sum_{k=1}^{\infty} \frac{1}{k} \mathsf{E}[\xi_1^2 I(k-1 \le |\xi_1| < k)]$$

$$\le 2 \sum_{k=1}^{\infty} \mathsf{E}[|\xi_1| I(k-1 \le |\xi_1| < k)] = 2\mathsf{E}|\xi_1| < \infty.$$

This completes the proof of the theorem.

Remark 1. The theorem admits a converse in the following sense. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables such that

$$\frac{\xi_1 + \cdots + \xi_n}{n} \to C,$$

with probability 1, where C is a (finite) constant. Then $\mathsf{E}|\xi_1| < \infty$ and $C = \mathsf{E}\xi_1$.

In fact, if $S_n/n \to C$ (P-a.s.) then

$$\frac{\xi_n}{n} = \frac{S_n}{n} - \left(\frac{n-1}{n}\right)\frac{S_{n-1}}{n-1} \to 0 \quad \text{(P-a.s.)}$$

and therefore $\mathsf{P}(|\xi_n| > n \text{ i.o.}) = 0$. By the Borel–Cantelli lemma,

$$\sum \mathsf{P}(|\xi_1| > n) < \infty,$$

and by Lemma 3 we have $\mathsf{E}|\xi_1| < \infty$. Then it follows from the theorem that $C = \mathsf{E}\xi_1$.

Consequently for independent identically distributed random variables the condition $\mathsf{E}|\xi_1| < \infty$ is necessary and sufficient for the convergence (with probability 1) of the ratio S_n/n to a finite limit.

Remark 2. If the expectation $m = \mathsf{E}\xi_1$ exists but is not necessarily finite, the conclusion (10) of the theorem remains valid.

In fact, let, for example, $\mathsf{E}\xi_1^- < \infty$ and $\mathsf{E}\xi_1^+ = \infty$. With $C > 0$, put

$$S_n^C = \sum_{i=1}^{n} \xi_i I(\xi_i \le C).$$

Then (P-a.s.).

$$\varliminf_n \frac{S_n}{n} \geq \lim_n \frac{S_n^C}{n} = \mathsf{E}\xi_1 I(\xi_1 \leq C).$$

But as $C \to \infty$,

$$\mathsf{E}\xi_1 I(\xi_1 \leq C) \to \mathsf{E}\xi_1 = \infty;$$

therefore $S_n/n \to +\infty$ (P-a.s.).

4. Let us give some applications of the strong law of large numbers.

Example 1 (Application to number theory). Let $\Omega = [0, 1)$, let $\mathscr{B}$ be the algebra of Borel subsets of Ω and let P be Lebesgue measure on $[0, 1)$. Consider the binary expansions $\omega = 0\,.\,\omega_1\omega_2\ldots$ of numbers $\omega \in \Omega$ (with infinitely many 0's) and define random variables $\xi_1(\omega), \xi_2(\omega), \ldots$ by putting $\xi_n(\omega) = \omega_n$. Since, for all $n \geq 1$ and all $x_1, \ldots, x_n$ taking the values 0 or 1,

$$\{\omega: \xi_1(\omega) = x_1, \ldots, \xi_n(\omega) = x_n\}$$
$$= \left\{\omega: \frac{x_1}{2} + \frac{x_2}{2^2} + \cdots + \frac{x_n}{2^n} \leq \omega < \frac{x_1}{2} + \cdots + \frac{x_n}{2^n} + \frac{1}{2^n}\right\},$$

the P-measure of this set is $1/2^n$. It follows that $\xi_1, \xi_n, \ldots$ is a sequence of independent identically distributed random variables with

$$\mathsf{P}(\xi_1 = 0) = \mathsf{P}(\xi_1 = 1) = \tfrac{1}{2}.$$

Hence, by the strong law of large numbers, we have the following result of Borel: *almost every number in* $[0, 1)$ *is normal, in the sense that with probability* 1 *the proportion of zeros and ones in its binary expansion tends to* $\frac{1}{2}$, *i.e.*,

$$\frac{1}{n}\sum_{k=1}^{n} I(\xi_k = 1) \to \tfrac{1}{2} \quad (\text{P-a.s.}).$$

Example 2 (The Monte Carlo method). Let $f(x)$ be a continuous function defined on $[0, 1]$, with values on $[0, 1]$. The following idea is the foundation of the statistical method of calculating $\int_0^1 f(x)\,dx$ (the "Monte Carlo method").

Let $\xi_1, \eta_1, \xi_2, \eta_2, \ldots$ be a sequence of independent random variables, uniformly distributed on $[0, 1]$. Put

$$\rho_i = \begin{cases} 1 & \text{if } f(\xi_i) > \eta_i, \\ 0 & \text{if } f(\xi_i) < \eta_i. \end{cases}$$

It is clear that

$$\mathsf{E}\rho_1 = \mathsf{P}\{f(\xi_1) > \eta_1\} = \int_0^1 f(x)\,dx.$$

By the strong law of large numbers (Theorem 3)

$$\frac{1}{n}\sum_{i=1}^{n}\rho_i \to \int_0^1 f(x)\,dx \quad (\text{P-a.s.}).$$

Consequently we can approximate an integral $\int_0^1 f(x)\,dx$ by taking a simulation consisting of a pair of random variables (ξ_i, η_i), $i \geq 1$, and then calculating ρ_i and $(1/n)\sum_{i=1}^{n}\rho_i$.

5. **Problems**

1. Show that $\mathrm{E}\xi^2 < \infty$ if and only if $\sum_{n=1}^{\infty} n\mathrm{P}(|\xi| > n) < \infty$.

2. Supposing that $\xi_1, \xi_2, \ldots$ are independent and identically distributed, show that if $\mathrm{E}|\xi_1|^\alpha < \infty$ for some α, $0 < \alpha < 1$, then $S_n/n^{1/\alpha} \to 0$ (P-a.s.), and if $\mathrm{E}|\xi_1|^\beta < \infty$ for some β, $1 \leq \beta < 2$, then $(S_n - n\mathrm{E}\xi_1)/n^{1/\beta} \to 0$ (P-a.s.).

3. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables and let $\mathrm{E}|\xi_1| = \infty$. Show that

$$\overline{\lim_n}\left|\frac{S_n}{n} - a_n\right| = \infty \quad (\text{P-a.s.})$$

for every sequence of constants $\{a_n\}$.

4. Show that a rational number on $[0, 1)$ is never normal (in the sense of Example 1, Subsection 4).

§4. Law of the Iterated Logarithm

1. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent Bernoulli random variables with $\mathrm{P}(\xi_n = 1) = \mathrm{P}(\xi_n = -1) = \frac{1}{2}$; let $S_n = \xi_1 + \cdots + \xi_n$. It follows from the proof of Theorem 2, §1, that

$$\overline{\lim}\,\frac{S_n}{\sqrt{n}} = +\infty, \qquad \underline{\lim}\,\frac{S_n}{\sqrt{n}} = -\infty, \tag{1}$$

with probability 1. On the other hand, by (3.11),

$$\frac{S_n}{\sqrt{n}\log n} \to 0 \quad (\text{P-a.s.}). \tag{2}$$

Let us compare these results.

It follows from (1) that with probability 1 the paths of $(S_n)_{n\geq 1}$ intersect the "curves" $\pm\varepsilon\sqrt{n}$ infinitely often for any given ε; but at the same time (2) shows that they only finitely often leave the region bounded by the curves $\pm\varepsilon\sqrt{n}\log n$. These two results yield useful information on the amplitude of the oscillations of the symmetric random walk $(S_n)_{n\geq 1}$. The law of the iterated logarithm, which we present below, improves this picture of the amplitude of the oscillations of $(S_n)_{n\geq 1}$.

Let us introduce the following definition. We call a function $\varphi^* = \varphi^*(n)$, $n \geq 1$, *upper* (for $(S_n)_{n\geq 1}$) if, with probability 1, $S_n \leq \varphi^*(n)$ for all n from $n = n_0(\omega)$ on.

We call a function $\varphi_* = \varphi_*(n), n \geq 1$, *lower* (for $(S_n)_{n\geq 1}$) if, with probability 1, $S_n > \varphi_*(n)$ for infinitely many n.

Using these definitions, and appealing to (1) and (2), we can say that every function $\varphi^* = \varepsilon\sqrt{n}\log n$, $\varepsilon > 0$, is upper, whereas $\varphi_* = \varepsilon\sqrt{n}$ is lower, $\varepsilon > 0$.

Let $\varphi = \varphi(n)$ be a function and $\varphi_\varepsilon^* = (1+\varepsilon)\varphi$, $\varphi_{*\varepsilon} = (1-\varepsilon)\varphi$, where $\varepsilon > 0$. Then it is easily seen that

$$\begin{aligned}\left\{\overline{\lim}\,\frac{S_n}{\varphi(n)} \leq 1\right\} &= \left\{\lim_n \left[\sup_{m\geq n} \frac{S_m}{\varphi(m)}\right] \leq 1\right\}\\ &\Leftrightarrow \left\{\sup_{m\geq n_1(\varepsilon)} \frac{S_m}{\varphi(m)} \leq 1+\varepsilon \text{ for every } \varepsilon > 0, \text{ from some } n_1(\varepsilon) \text{ on}\right\}\\ &\Leftrightarrow \{S_m \leq (1+\varepsilon)\varphi(m) \text{ for every } \varepsilon > 0, \text{ from some } n_1(\varepsilon) \text{ on}\}.\end{aligned} \tag{3}$$

In the same way,

$$\begin{aligned}\left\{\overline{\lim}\,\frac{S_n}{\varphi(n)} \geq 1\right\} &= \left\{\lim_n \left[\sup_{m\geq n} \frac{S_m}{\varphi(m)}\right] \geq 1\right\}\\ &\Leftrightarrow \left\{\sup_{m\geq n_2(\varepsilon)} \frac{S_m}{\varphi(m)} \leq 1+\varepsilon \textit{ for every } \varepsilon > 0, \textit{ from some } n_1(\varepsilon) \textit{ on}\right\}\\ &\Leftrightarrow \begin{array}{l}\{S_m \geq (1-\varepsilon)\varphi(m) \text{ for every } \varepsilon > 0 \text{ and for infinitely}\\ \text{many } m \text{ larger than some } n_3(\varepsilon) \geq n_2(\varepsilon)\}.\end{array}\end{aligned} \tag{4}$$

It follows from (3) and (4) that in order to verify that each function $\varphi_\varepsilon^* = (1+\varepsilon)\varphi$, $\varepsilon > 0$, is upper, we have to show that

$$\mathsf{P}\left\{\overline{\lim}\,\frac{S_n}{\varphi(n)} \leq 1\right\} = 1. \tag{5}$$

But to show that $\varphi_{*\varepsilon} = (1-\varepsilon)\varphi$, $\varepsilon > 0$, is lower, we have to show that

$$\mathsf{P}\left\{\overline{\lim}\,\frac{S_n}{\varphi(n)} \geq 1\right\} = 1. \tag{6}$$

2. Theorem 1 (Law of the Iterated Logarithm). *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables with* $\mathsf{E}\xi_i = 0$ *and* $\mathsf{E}\xi_i^2 = \sigma^2 > 0$. *Then*

$$\mathsf{P}\left\{\overline{\lim}\, \frac{S_n}{\psi(n)} = 1\right\} = 1, \tag{7}$$

where

$$\psi(n) = \sqrt{2\sigma^2 n \log\log n}. \tag{8}$$

For uniformly bounded random variables, the law of the iterated logarithm was established by Khinchin (1924). In 1929 Kolmogorov generalized this result to a wide class of independent variables. Under the conditions of Theorem 1, the law of the iterated logarithm was established by Hartman and Wintner (1941).

Since the proof of Theorem 1 is rather complicated, we shall confine ourselves to the special case when the random variables ξ_n are normal, $\xi_n \sim \mathcal{N}(0, 1)$, $n \geq 1$.

We begin by proving two auxiliary results.

Lemma 1. *Let $\xi_1, \ldots, \xi_n$ be independent random variables that are symmetrically distributed* ($\mathsf{P}(\xi_k \in B) = \mathsf{P}(-\xi_k \in B)$ *for every* $B \in \mathscr{B}(R)$, $k \leq n$). *Then for every real number* a

$$\mathsf{P}\left(\max_{1 \leq k \leq n} S_k > a\right) \leq 2\mathsf{P}(S_n > a). \tag{9}$$

Proof. Let $A = \{\max_{1 \leq k \leq n} S_k > a\}$, $A_k = \{S_i \leq a, i \leq k - 1; S_k > a\}$ and $B = \{S_n > a\}$. Since $S_n > a$ on A_k (because $S_k \leq S_n$), we have

$$\begin{aligned}\mathsf{P}(B \cap A_k) &\geq \mathsf{P}(A_k \cap \{S_n \geq S_k\}) = \mathsf{P}(A_k)\mathsf{P}(S_n \geq S_k) \\ &= \mathsf{P}(A_k)\mathsf{P}(\xi_{k+1} + \cdots + \xi_n \geq 0).\end{aligned}$$

By the symmetry of the distributions of the random variables $\xi_1, \ldots, \xi_n$, we have

$$\mathsf{P}(\xi_{k+1} + \cdots + \xi_n > 0) = \mathsf{P}(\xi_{k+1} + \cdots + \xi_n < 0).$$

Hence $\mathsf{P}(\xi_{k+1} + \cdots + \xi_n > 0) \geq \frac{1}{2}$, and therefore

$$\mathsf{P}(B) \geq \sum_{k=1}^{n} \mathsf{P}(A_k \cap B) \geq \frac{1}{2} \sum_{k=1}^{n} \mathsf{P}(A_k) = \frac{1}{2}\mathsf{P}(A),$$

which establishes (9).

Lemma 2. *Let* $S_n \sim \mathcal{N}(0, \sigma^2(n))$, $\sigma^2(n) \uparrow \infty$, *and let* $a(n)$, $n \geq 1$, *satisfy* $a(n)/\sigma(n) \to \infty$, $n \to \infty$. *Then*

$$\mathsf{P}(S_n > a(n)) \sim \frac{\sigma(n)}{\sqrt{2\pi}\, a(n)} \exp\{-\tfrac{1}{2}a^2(n)/\sigma^2(n)\}. \tag{10}$$

The proof follows from the asymptotic formula

$$\frac{1}{\sqrt{2\pi}}\int_x^{\infty} e^{-y^2/2}\,dy \sim \frac{1}{\sqrt{2\pi}x} e^{-x^2/2}, \qquad x \to \infty,$$

since $S_n/\sigma(n) \sim \mathcal{N}(0, 1)$.

PROOF OF THEOREM 1 (for $\xi_i \sim \mathcal{N}(0, 1)$).

Let us first establish (5). Let $\varepsilon > 0$, $\lambda = 1 + \varepsilon$, $n_k = \lambda^k$, where $k \geq k_0$, and k_0 is chosen so that $\ln \ln k_0$ is defined. We also define

$$A_k = \{S_n > \lambda\psi(n) \text{ for some } n \in (n_k, n_{k+1}]\}, \tag{11}$$

and put

$$A = \{A_k \text{ i.o.}\} = \{S_n > \lambda\psi(n) \text{ for infinitely many } n\}.$$

In accordance with (3), we can establish (5) by showing that $\mathsf{P}(A) = 0$.

Let us show that $\sum \mathsf{P}(A_k) < \infty$. Then $\mathsf{P}(A) = 0$ by the Borel–Cantelli lemma.

From (11), (9), and (10) we find that

$$\begin{aligned}
\mathsf{P}(A_k) &\leq \mathsf{P}\{S_n > \lambda\psi(n_k) \text{ for some } n \in (n_k, n_{k+1})\} \\
&\leq \mathsf{P}\{S_n > \lambda\psi(n_k) \text{ for some } n \leq n_{k+1}\} \\
&\leq 2P\{S_{n_{k+1}} > \lambda\psi(n_k)\} \sim \frac{2\sqrt{n_k}}{\sqrt{2\pi}\,\lambda\psi(n_k)} \exp\{-\tfrac{1}{2}\lambda^2[\psi(n_k)/\sqrt{n_k}]^2\} \\
&\leq C_1 \exp(-\lambda \ln \ln \lambda^k) \leq Ce^{-\lambda \ln k} = C_2 k^{-\lambda},
\end{aligned}$$

where C_1 and C_2 are constants. But $\sum_{k=1}^{\infty} k^{-\lambda} < \infty$, and therefore

$$\sum \mathsf{P}(A_k) < \infty.$$

Consequently (5) is established.

We turn now to the proof of (6). In accordance with (4) we must show that, with $\lambda = 1 - \varepsilon$, $\varepsilon > 0$, we have with probability 1 that $S_n \geq \lambda\psi(n)$ for infinitely many n.

Let us apply (5), which we just proved, to the sequence $(-S_n)_{n\geq 1}$. Then we find that for all n, with finitely many exceptions, $-S_n \leq 2\psi(n)$ (P-a.s.). Consequently if $n_k = N^k$, $N > 1$, then for sufficiently large k, either

$$S_{n_{k-1}} \geq -2\psi(n_{k-1})$$

or

$$S_{n_k} \geq Y_k - 2\psi(n_{k-1}), \tag{12}$$

where $Y_k = S_{n_k} - S_{n_{k-1}}$.

Hence if we show that for infinitely many k

$$Y_k > \lambda\psi(n_k) + 2\psi(n_{k-1}), \tag{13}$$

this and (12) show that (P-a.s.) $S_{n_k} > \lambda\psi(n_k)$ for infinitely many k. Take some $\lambda' \in (\lambda, 1)$. Then there is an $N > 1$ such that for all k

$$\lambda'[2(N^k - N^{k-1}) \ln \ln N^k]^{1/2} > \lambda(2N^k \ln \ln N^k)^{1/2} + 2(2N^{k-1} \ln \ln N^{k-1})^{1/2} \equiv \lambda\psi(N^k) + 2\psi(N^{k-1}).$$

It is now enough to show that

$$Y_k > \lambda'[2(N^k - N^{k-1}) \ln \ln N^k]^{1/2} \tag{14}$$

for infinitely many k. Evidently $Y_k \sim \mathcal{N}(0, N^k - N^{k-1})$. Therefore, by Lemma 2,

$$\mathsf{P}\{Y_k > \lambda'[2(N^k - N^{k-1}) \ln \ln N^k]^{1/2}\} \sim \frac{1}{\sqrt{2\pi}\lambda'(2 \ln \ln N^k)^{1/2}} e^{-(\lambda')^2 \ln \ln N^k}$$

$$\geq \frac{C_1}{(\ln k)^{1/2}} k^{-(\lambda')^2} \geq \frac{C_2}{k \ln k}.$$

Since $\sum (1/k \ln k) = \infty$, it follows from the second part of the Borel–Cantelli lemma that, with probability 1, inequality (14) is satisfied for infinitely many k, so that (6) is established.

This completes the proof of the theorem.

Remark 1. Applying (7) to the random variables $(-S_n)_{n\geq 1}$, we find that

$$\underline{\lim} \frac{S_n}{\varphi(n)} = -1. \tag{15}$$

It follows from (7) and (15) that the law of the iterated logarithm can be put in the form

$$\mathsf{P}\left\{\overline{\lim} \frac{|S_n|}{\varphi(n)} = 1\right\} = 1. \tag{16}$$

Remark 2. The law of the iterated logarithm says that for every $\varepsilon > 0$ each function $\psi_\varepsilon^* = (1 + \varepsilon)\psi$ is upper, and $\psi_{*\varepsilon} = (1 - \varepsilon)\psi$ is lower.

The conclusion (7) is also equivalent to the statement that, for each $\varepsilon > 0$,

$$\mathsf{P}\{|S_n| \geq (1 - \varepsilon)\psi(n) \text{ i.o.}\} = 1,$$

$$\mathsf{P}\{|S_n| \geq (1 + \varepsilon)\psi(n) \text{ i.o.}\} = 0.$$

3. Problems

1. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with $\xi_n \sim \mathcal{N}(0, 1)$. Show that

$$\mathsf{P}\left\{\overline{\lim} \frac{\xi_n}{\sqrt{2 \ln n}} = 1\right\} = 1.$$

2. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables, distributed according to Poisson's law with parameter $\lambda > 0$. Show that (independently of λ)

$$\mathsf{P}\left\{\overline{\lim} \frac{\xi_n \ln \ln n}{\ln n} = 1\right\} = 1.$$

3. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables with

$$\mathsf{E}e^{it\xi_1} = e^{-|t|^\alpha}, \qquad 0 < \alpha < 2.$$

Show that

$$\mathsf{P}\left\{\overline{\lim} \left|\frac{S_n}{n^{1/\alpha}}\right|^{1/(\ln \ln n)} = e^{1/\alpha}\right\} = 1.$$

4. Establish the following generalization of (9). Let $\xi_1, \ldots, \xi_n$ be independent random variables. *Lévy's inequality*

$$\mathsf{P}\left\{\max_{0 \le k \le n} [S_k + \mu(S_n - S_k)] > a\right\} \le 2\mathsf{P}(S_n > a), \qquad S_0 = 0,$$

holds for every real a, where $\mu(\xi)$ is the median of ξ, i.e. a constant such that

$$\mathsf{P}(\xi \ge \mu(\xi)) \ge \tfrac{1}{2}, \qquad \mathsf{P}(\xi \le \mu(\xi)) \ge \tfrac{1}{2}.$$

§5. Rapidity of Convergence in the Strong Law of Large Numbers and in the Probabilities of Large Deviations

1. By the results of §6, Chapter I, we have the following estimate for the Bernoulli scheme:

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - p\right| \ge \varepsilon\right\} \le 2e^{-2n\varepsilon^2} \tag{1}$$

(see (42), subsection 7, §6, Chapter I). From this, of course, there follows the inequalities

$$\mathsf{P}\left\{\sup_{m \ge n}\left|\frac{S_m}{m} - p\right| \ge \varepsilon\right\} \le \sum_{m \ge n} \mathsf{P}\left\{\left|\frac{S_m}{m} - p\right| \ge \varepsilon\right\} \le \frac{2}{1 - e^{-2\varepsilon^2}} e^{-2n\varepsilon^2}, \tag{2}$$

which provide an approximation of the rate of convergence to p by the quantity S_n/n with probability 1.

We now consider the question of the validity of formulas of the types (1) and (2) in some more general situations, when $S_n = \xi_1 + \cdots + \xi_n$ is a sum of independent identically distributed random variables.

2. Let $\xi, \xi_1, \xi_2, \ldots$ be a sequence of independent random variables. We say that a random variable satisfies *Cramér's condition* if there is a $\lambda > 0$ for which

$$\varphi(\lambda) = \mathsf{E}e^{\lambda|\xi|} < \infty \tag{3}$$

(it can be shown that this condition is equivalent to an exponential decrease of $\mathsf{P}(|\xi| > x)$, $x \to \infty$).

Let

$$\Lambda = \{\lambda \in R: \varphi(\lambda) < \infty\}, \tag{4}$$

where we suppose that Λ contains a neighborhood of the point $\lambda = 0$, i.e., Cramér's condition (3) is satisfied for some $\lambda > 0$.

On the set Λ the function

$$\psi(\lambda) = \ln \varphi(\lambda) \tag{5}$$

is convex (from below), and strictly convex if the random variable ξ is not degenerate. We also notice that

$$\psi(0) = 0, \quad \psi'(0) = m(= \mathsf{E}\xi), \qquad \psi''(\lambda) \geq 0.$$

We extend $\psi(\lambda)$ to all $\lambda \in R$ by setting $\psi(\lambda) = \infty$ for $\lambda \notin \Lambda$.

We define the function

$$H(a) = \sup_{\lambda} [a\lambda - \psi(\lambda)], \qquad a \in R, \tag{6}$$

called the *Cramér transform* (of the distribution function $F = F(x)$ of the random variable ξ). The function $H(a)$ is also convex (from below) and its minimum is zero, attained at $\lambda = m$.

If $a > m$, we have

$$H(a) = \sup_{\lambda > 0} [a\lambda - \psi(\lambda)].$$

Then

$$\mathsf{P}\{\xi \geq a\} \leq \inf_{\lambda > 0} \mathsf{E}e^{\lambda(\xi - a)} = \inf_{\lambda > 0} e^{-[a\lambda - \psi(\lambda)]} = e^{-H(a)}. \tag{7}$$

Similarly, for $a < m$ we have $H(a) = \sup_{\lambda < 0} [a\lambda - \psi(\lambda)]$ and

$$\mathsf{P}\{\xi \leq a\} \leq e^{-H(a)}. \tag{8}$$

Consequently (compare (42), §6, Chapter 1),

$$\mathsf{P}\{|\xi - m| \geq \varepsilon\} \leq e^{-\min\{H(m-\varepsilon), H(m+\varepsilon)\}}. \tag{9}$$

If $\xi, \xi_1, \ldots \xi_n$ are independent identically distributed random variables that satisfy Cramér's condition (3), $S_n = \xi_1 + \cdots + \xi_n$, $\psi_n(\lambda) = \ln \mathsf{E} \exp(\lambda S_n/n)$, $\psi(\lambda) = \ln \mathsf{E}e^{\lambda\xi}$, and

$$H_n(a) = \sup_{\lambda} [a\lambda - \psi_n(\lambda)], \tag{10}$$

then

$$H_n(a) = nH(a)\left(= n \sup_{\lambda} [a\lambda - \psi(\lambda)]\right)$$

and the inequalities (7), (8), and (9) assume the following forms:

$$\mathsf{P}\left\{\frac{S_n}{n} \geq a\right\} \leq e^{-nH(a)}, \qquad a > m, \tag{11}$$

$$\mathsf{P}\left\{\frac{S_n}{n} \leq a\right\} \leq e^{-nH(a)}, \qquad a < m, \tag{12}$$

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right\} \leq 2e^{-\min\{H(m-\varepsilon), H(m+\varepsilon)\}\cdot n}. \tag{13}$$

Remark. Results of the type

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right\} \leq ae^{-bn}, \tag{14}$$

where $a > 0$ and $b > 0$, indicate exponential convergence "adjusted" by the constants a and b. In the theory of *large deviations*, results are often presented in a somewhat different, "cruder," form:

$$\overline{\lim_n} \frac{1}{n} \ln \mathsf{P}\left\{\left|\frac{S_n}{n} - m\right| \geq \varepsilon\right\} < 0, \tag{15}$$

that clearly arises from (14) and refers to the "exponential" rapidity of convergence, but without specifying the values of the constants a and b.

Now we turn to the question of upper bounds for the probabilities

$$\mathsf{P}\left\{\sup_{k \geq n} \frac{S_k}{k} > a\right\}, \qquad \mathsf{P}\left\{\inf_{k \geq n} \frac{S_k}{k} < a\right\}, \qquad \mathsf{P}\left\{\sup_{k \geq n} \left|\frac{S_k}{k} - m\right| > \varepsilon\right\},$$

which can provide definite bounds *on the rapidity of convergence in the strong law of large numbers.*

Let us suppose that the independent identically distributed nondegenerate random variables $\xi, \xi_1, \xi_2, \ldots$ satisfy Cramér's condition, i.e., $\varphi(\lambda) = \mathsf{E}e^{\lambda|\xi|} < \infty$ for some $\lambda > 0$.

We fix $n \geq 1$ and set

$$\kappa = \inf\left\{k \geq n : \frac{S_k}{k} > a\right\},$$

taking $\kappa = \infty$ if $S_k/k < a$ for $k \geq n$.

In addition, let a and $\lambda > 0$ satisfy

$$\lambda a - \ln \varphi(\lambda) \geq 0. \tag{16}$$

Then

$$\mathsf{P}\left\{\sup_{k \geq n} \frac{S_k}{k} > a\right\} = \mathsf{P}\left\{\bigcup_{k \geq n}\left\{\frac{S_k}{k} > a\right\}\right\}$$

$$= \mathsf{P}\left\{\frac{S_\kappa}{\kappa} > a, \kappa < \infty\right\} = \mathsf{P}\{e^{\lambda S_\kappa} > e^{\lambda a \kappa}, \kappa < \infty\}$$

$$\begin{aligned}
&= \mathsf{P}\{e^{\lambda S_\kappa - \kappa \ln \varphi(\lambda)} > e^{\kappa(\lambda a - \ln \varphi(\lambda))}, \kappa < \infty\} \\
&\le \mathsf{P}\{e^{\lambda S_\kappa - \kappa \ln \varphi(\lambda)} > e^{n(\lambda a - \ln \varphi(\lambda))}, \kappa < \infty\} \\
&\le \mathsf{P}\left\{\sup_{k \ge n} e^{\lambda S_k - k \ln \varphi(\lambda)} \ge e^{n(\lambda a - \ln \varphi(\lambda))}\right\}. \qquad (17)
\end{aligned}$$

To take the final step, we notice that the sequence of random variables

$$e^{\lambda S_k - k \ln \varphi(\lambda)}, \qquad k \ge 1,$$

with respect to the flow of σ-algebras $\mathscr{F}_k = \sigma\{\xi_1, \ldots, \xi_k\}$, $k \ge 1$, forms a *martingale*. (For more details, see Chapter VII and, in particular, Example 1 in §1). Then it follows from inequality (8) in §3, Chapter VII, that

$$\mathsf{P}\left\{\sup_{k \ge n} e^{\lambda S_k - k \ln \varphi(\lambda)} \ge e^{n(\lambda a - \ln \varphi(\lambda))}\right\} \le e^{-n(\lambda a - \ln \varphi(\lambda))},$$

and consequently, (by (16)) we obtain the inequality

$$\mathsf{P}\left\{\sup_{k \ge n} \frac{S_k}{k} > a\right\} \le e^{-n(\lambda a - \ln \varphi(\lambda))}. \qquad (18)$$

Let $a > m$. Since the function $f(\lambda) = a\lambda - \ln \varphi(\lambda)$ has the properties $f(0) = 0$, $f'(0) > 0$, there is a $\lambda > 0$ for which (16) is satisfied, and consequently, we obtain from (18) that if $a > m$ we have

$$\mathsf{P}\left\{\sup_{k \ge n} \frac{S_k}{k} > a\right\} \le e^{-n \sup_{\lambda > 0}[\lambda a - \ln \varphi(\lambda)]} = e^{-nH(a)}. \qquad (19)$$

Similarly, if $a < m$, we have

$$\mathsf{P}\left\{\sup_{k \ge n} \frac{S_k}{k} < a\right\} \le e^{-n \sup_{\lambda < 0}[\lambda a - \ln \varphi(\lambda)]} = e^{-nH(a)}. \qquad (20)$$

From (19) and (20), we obtain

$$\mathsf{P}\left\{\sup_{k \ge n} \left|\frac{S_k}{k} - m\right| > \varepsilon\right\} \le 2e^{-\min[H(m-\varepsilon), H(m+\varepsilon)] \cdot n}. \qquad (21)$$

Remark. Combining the right-hand sides of the inequalities (11) and (19) leads us to suspect that this situation is not random. In fact, this expectation is concealed in the fact that the sequences $(S_k/k)_{n \le k \le N}$ form, for every $n \le N$, *reversed martingales* (see Problem 6 in §1, Chapter VII, and Example 4 in §11, Chapter I).

2. Problems

1. Carry out the proof of inequalities (8) and (20).
2. Investigate the properties of $H(a)$.

CHAPTER V

Stationary (Strict Sense) Random Sequences and Ergodic Theory

§1. Stationary (Strict Sense) Random Sequences. Measure-Preserving Transformations

1. Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space and $\xi = (\xi_1, \xi_2, \ldots)$ a sequence of random variables or, as we say, a *random sequence*. Let $\theta_k \xi$ denote the sequence $(\xi_{k+1}, \xi_{k+2}, \ldots)$.

Definition 1. A random sequence ξ is *stationary* (*in the strict sense*) if the probability distributions of $\theta_k \xi$ and ξ are the same for every $k \geq 1$:

$$\mathsf{P}((\xi_1, \xi_2, \ldots) \in B) = \mathsf{P}((\xi_{k+1}, \xi_{k+2}, \ldots) \in B), \qquad B \in \mathscr{B}(R^{\infty}).$$

The simplest example is a sequence $\xi = (\xi_1, \xi_2, \ldots)$ of independent identically distributed random variables. Starting from such a sequence, we can construct a broad class of stationary sequences $\eta = (\eta_1, \eta_2, \ldots)$ by choosing any Borel function $g(x_1, \ldots, x_n)$ and setting $\eta_k = g(\xi_k, \xi_{k+1}, \ldots, \xi_{k+1})$.

If $\xi = (\xi_1, \xi_2, \ldots)$ is a sequence of independent identically distributed random variables with $\mathsf{E}|\xi_1| < \infty$ and $\mathsf{E}\xi_1 = m$, the law of large numbers tells us that, with probability 1,

$$\frac{\xi_1 + \cdots + \xi_n}{n} \to m, \qquad n \to \infty.$$

In 1931 Birkhoff obtained a remarkable generalization of this fact for the case of stationary sequences. The present chapter consists mainly of a proof of Birkhoff's theorem.

The following presentation is based on the idea of measure-preserving transformations, something that brings us in contact with an interesting branch of analysis (ergodic theory), and at the same time shows the connection between this theory and stationary random proceesses.

Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space.

Definition 2. A transformation T of Ω into Ω is *measurable* if, for every $A \in \mathscr{F}$,

$$T^{-1}A = \{\omega: T\omega \in A\} \in \mathscr{F}.$$

Definition 3. A measurable transformation T is a *measure-preserving transformation* (or morphism) if, for every $A \in F$,

$$\mathsf{P}(T^{-1}A) = \mathsf{P}(A).$$

Let T be a measure-preserving transformation, T^n its nth iterate, and $\xi_1 = \xi_1(\omega)$ a random variable. Put $\xi_k(\omega) = \xi_1(T^{n-1}\omega)$, $n \geq 2$, and consider the sequence $\xi = (\xi_1, \xi_2, \ldots)$. We claim that this sequence is stationary.

In fact, let $A = \{\omega: \xi \in B\}$ and $A_1 = \{\omega: \theta_1\xi \in B\}$, where $B \in \mathscr{B}(R^\infty)$. Since $A = \{\omega: (\xi_1(\omega), \xi_1(T\omega), \ldots) \in B\}$, and $A_1 = \{\omega: (\xi_1(T\omega), \xi_1(T^2\omega), \ldots) \in B\}$, we have $\omega \in A_1$ if and only if either $T\omega \in A$ or $A_1 = T^{-1}A$. But $\mathsf{P}(T^{-1}A) = \mathsf{P}(A)$, and therefore $\mathsf{P}(A_1) = \mathsf{P}(A)$. Similarly $\mathsf{P}(A_k) = \mathsf{P}(A)$ for every $A_k = \{\omega: \theta_k\xi \in B\}, k \geq 2$.

Thus we can use measure-preserving transformations to construct stationary (strict sense) random variables.

In a certain sense, there is a converse result: for every stationary sequence ξ considered on $(\Omega, \mathscr{F}, \mathsf{P})$ we can construct a new probability space $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{\mathsf{P}})$, a random variable $\tilde{\xi}_1(\tilde{\omega})$ and a measure-preserving transformation $\tilde{T}$, such that the distribution of $\tilde{\xi} = \{\tilde{\xi}_1(\tilde{\omega}), \tilde{\xi}_1(\tilde{T}\tilde{\omega}), \ldots\}$ coincides with the distribution of ξ.

In fact, take $\tilde{\Omega}$ to be the coordinate space R^∞ and put $\tilde{\mathscr{F}} = \mathscr{B}(R^\infty)$, $\tilde{\mathsf{P}} = P_\xi$, where $P_\xi(B) = \mathsf{P}\{\omega: \xi \in B\}$, $B \in \mathscr{B}(R^\infty)$. The action of $\tilde{T}$ on $\tilde{\Omega}$ is given by

$$\tilde{T}(x_1, x_2, \ldots) = (x_2, x_3, \ldots).$$

If $\tilde{\omega} = (x_1, x_2, \ldots)$, put

$$\tilde{\xi}_1(\tilde{\omega}) = x_1, \tilde{\xi}_n(\tilde{\omega}) = \tilde{\xi}_1(\tilde{T}^{n-1}\tilde{\omega}), \qquad n \geq 2.$$

Now let $A = \{\tilde{\omega}: (x_1, \ldots, x_k) \in B\}$, $B \in \mathscr{B}(R^k)$, and

$$\tilde{T}^{-1}A = \{\tilde{\omega}: (x_2, \ldots, x_{k+1}) \in B\}.$$

Then the property of being stationary means that

$$\tilde{\mathsf{P}}(A) = \mathsf{P}\{\omega: (\xi_1, \ldots, \xi_k) \in B\} = \mathsf{P}\{\omega: (\xi_2, \ldots, \xi_{k+1}) \in B\} = \tilde{\mathsf{P}}(\tilde{T}^{-1}A),$$

i.e. T is a measure-preserving transformation. Since $\tilde{\mathsf{P}}\{\tilde{\omega}: (\tilde{\xi}_1, \ldots, \tilde{\xi}_k) \in B\} = \tilde{\mathsf{P}}\{\omega: (\xi_1, \ldots, \xi_k) \in B\}$ for every k, it follows that ξ and $\tilde{\xi}$ have the same distribution.

Here are some examples of measure-preserving transformations.

EXAMPLE 1. Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ consist of n points (a finite number), $n \geq 2$, let $\mathscr{F}$ be the collection of its subsets, and let $T\omega_i = \omega_{i+1}$, $1 \leq i \leq n-1$, and $T_{\omega_n} = \omega_1$. If $\mathsf{P}(\omega_i) = 1/n$, the transformation T is measure-preserving.

EXAMPLE 2. If $\Omega = [0, 1)$, $\mathscr{F} = \mathscr{B}([0, 1))$, P is Lebesgue measure, $\lambda \in [0, 1)$, then $Tx = (x + \lambda) \bmod 1$ and $T = 2x \bmod 1$ are both measure-preserving transformations.

2. Let us pause to consider the physical hypotheses that led to the consideration of measure-preserving transformations.

Let us suppose that Ω is the phase space of a system that evolves (in discrete time) according to a given law of motion. If ω is the state at instant $n = 1$, then $T^n\omega$, where T is the translation operator induced by the given law of motion, is the state attained by the system after n steps. Moreover, if A is some set of states ω then $T^{-1}A = \{\omega: T\omega \in A\}$ is, by definition, the set of states ω that lead to A in one step. Therefore if we interpret Ω as an incompressible fluid, the condition $P(T^{-1}A) = P(A)$ can be thought of as the rather natural condition of conservation of volume. (For the classical conservative Hamiltonian systems, Liouville's theorem asserts that the corresponding transformation T preserves Lebesgue measure.)

3. One of the earliest results on measure-preserving transformations was Poincaré's recurrence theorem (1912).

Theorem 1. *Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space, let T be a measure-preserving transformation, and let $A \in F$. Then, for almost every point $\omega \in A$, we have $T^n\omega \in A$ for infinitely many $n \geq 1$.*

PROOF. Let $C = \{\omega \in A: T^n\omega \notin A, \text{ for all } n \geq 1\}$. Since $C \cap T^{-n}C = \varnothing$ for all $n \geq 1$, we have $T^{-m}C \cap T^{-(m+n)}C = T^{-m}(C \cap T^{-n}C) = \varnothing$. Therefore the sequence $\{T^{-n}C\}$ consists of disjoint sets of equal measure. Therefore $\sum_{n=0}^{\infty} \mathsf{P}(C) = \sum_{n=0}^{\infty} \mathsf{P}(T^{-n}C) \leq \mathsf{P}(\Omega) = 1$ and consequently $\mathsf{P}(C) = 0$. Therefore, for almost every point $\omega \in A$, for at least one $n \geq 1$, we have $T^n\omega \in A$. It follows that $T^n\omega \in A$ for infinitely many n.

Let us apply the preceding result to T^k, $k \geq 1$. Then for every $\omega \in A \backslash N$, where N is a set of probability zero, the union of the corresponding sets corresponding to the various values of k, there is an n_k such that $(T^k)^{n_k}\omega \in A$. It is then clear that $T^n\omega \in A$ for infinitely many n. This completes the proof of the theorem.

Corollary. *Let $\xi(\omega) \geq 0$. Then*

$$\sum_{k=0}^{\infty} \xi(T^k\omega) = \infty \quad (\mathsf{P}\text{-}a.s.)$$

on the set $\{\omega: \xi(\omega) > 0\}$.

In fact, let $A_n = \{\omega: \xi(\omega) \geq 1/n\}$. Then, according to the theorem, $\sum_{k=0}^{\infty} \xi(T^k\omega) = \infty$ (P-a.s.) on A_n, and the required result follows by letting $n \to \infty$.

Remark. The theorem remains valid if we replace the probability measure P by any finite measure μ with $\mu(\Omega) < \infty$.

4. Problems

1. Let T be a measure-preserving transformation and $\xi = \xi(\omega)$ a random variable whose expectation $E\xi(\omega)$ exists. Show that $E\xi(\omega) = E\xi(T\omega)$.
2. Show that the transformations in Examples 1 and 2 are measure-preserving.
3. Let $\Omega = [0, 1)$, $F = \mathscr{B}([0, 1))$ and let P be a measure whose distribution function is continuous. Show that the transformations $Tx = \lambda x, 0 < \lambda < 1$, and $Tx = x^2$ are not measure-preserving.

§2. Ergodicity and Mixing

1. In the present section T denotes a measure-preserving transformation on the probability space $(\Omega, \mathscr{F}, \mathsf{P})$.

Definition 1. A set $A \in \mathscr{F}$ is *invariant* if $T^{-1}A = A$. A set $A \in \mathscr{F}$ is *almost invariant* if A and $T^{-1}A$ differ only by a set of measure zero, i.e. $\mathsf{P}(A \triangle T^{-1}A) = 0$.

It is easily verified that the classes $\mathscr{I}$ and $\mathscr{I}^*$ of invariant or almost invariant sets, respectively, are σ-algebras.

Definition 2. A measure-preserving transformation T is *ergodic* (or *metrically transitive*) if every invariant set A has measure either zero or one.

Definition 3. A random variable $\xi = \xi(\omega)$ is *invariant* (or *almost invariant*) if $\xi(\omega) = \xi(T\omega)$ for all $\omega \in \Omega$ (or for almost all $\omega \in \Omega$).

The following lemma establishes a connection between invariant and almost invariant sets.

Lemma 1. *If A is almost invariant, there is an invariant set B such that $P(A \triangle B) = 0$.*

Proof. Let $B = \overline{\lim}\, T^{-n}A$. Then $T^{-1}B = \overline{\lim}\, T^{-(n+1)}A = B$, i.e. $B \in \mathscr{I}$. It is easily seen that $A \triangle B \subseteq \bigcup_{k=0}^{\infty} (T^{-k}A \triangle T^{-(k+1)}A)$. But

$$\mathsf{P}(T^{-k}A \triangle T^{-(k+1)}A) = \mathsf{P}(A \triangle T^{-1}A) = 0.$$

Hence $\mathsf{P}(A \triangle B) = 0$.

Lemma 2. *A transformation T is ergodic if and only if every almost invariant set has measure zero or one.*

Proof. Let $A \in \mathscr{I}^*$; then according to Lemma 1 there is an invariant set B such that $\mathsf{P}(A \triangle B) = 0$. But T is ergodic and therefore $\mathsf{P}(B) = 0$ or 1. Therefore $\mathsf{P}(A) = 0$ or 1. The converse is evident, since $\mathscr{I} \subseteq \mathscr{I}^*$. This completes the proof of the lemma.

Theorem 1. *Let T be a measure-preserving transformation. Then the following conditions are equivalent:*

(1) *T is ergodic;*
(2) *every almost invariant random variable is* (P-*a.s.*) *constant;*
(3) *every invariant random variable is* (P-*a.s.*) *constant.*

Proof. (1) $\Leftrightarrow$ (2). Let T be ergodic and ξ almost invariant, i.e. (P-a.s.) $\xi(\omega) = \xi(T\omega)$. Then for every $c \in R$ we have $A_c = \{\omega: \xi(\omega) \le c\} \in \mathscr{I}^*$, and then $\mathsf{P}(A_c) = 0$ or 1 by Lemma 2. Let $C = \sup\{c: \mathsf{P}(A_c) = 0\}$. Since $A_c \uparrow \Omega$ as $c \uparrow \infty$ and $A_c \downarrow \varnothing$ as $c \downarrow -\infty$, we have $|C| < \infty$. Then

$$\mathsf{P}\{\omega: \xi(\omega) < C\} = \mathsf{P}\left\{\bigcup_{n=1}^{\infty} \left\{\xi(\omega) \le C - \frac{1}{n}\right\}\right\} = 0$$

and similarly $\mathsf{P}\{\omega: \xi(\omega) > C\} = 0$. Consequently $\mathsf{P}\{\omega: \xi(\omega) = C\} = 1$.

(2) $\Rightarrow$ (3). Evident.

(3) $\Rightarrow$ (1). Let $A \in \mathscr{I}$; then I_A is an invariant random variable and therefore, (P-a.s.), $I_A = 0$ or $I_A = 1$, whence $\mathsf{P}(A) = 0$ or 1.

Remark. The conclusion of the theorem remains valid in the case when "random variable" is replaced by "bounded random variable".

We illustrate the theorem with the following example.

Example. Let $\Omega = [0, 1)$, $\mathscr{F} = \mathscr{B}([0, 1))$, let P be Lebesgue measure and let $T\omega = (\omega + \lambda) \bmod 1$. Let us show that T is ergodic if and only if λ is irrational.

Let $\xi = \xi(\omega)$ be a random variable with $\mathsf{E}\xi^2(\omega) < \infty$. Then we know that the Fourier series $\sum_{n=-\infty}^{\infty} c_n e^{2\pi i n\omega}$ of $\xi(\omega)$ converges in the mean square sense, $\sum |c_n|^2 < \infty$, and, because T is a measure-preserving transformation (Example 2, §1), we have (Problem 1, §1) that for the random variable ξ

$$\begin{aligned} c_n \mathsf{E}\xi(\omega)e^{2\pi i n\xi(\omega)} &= \mathsf{E}\xi(T\omega)e^{2\pi i n T\omega} = e^{2\pi i n\lambda}\mathsf{E}\xi(T\omega)e^{2\pi i n\omega} \\ &= e^{2\pi i n\lambda}\mathsf{E}\xi(\omega)e^{2\pi i n\omega} = c_n e^{2\pi i n\lambda}. \end{aligned}$$

So $c_n(1 - e^{2\pi i n\lambda}) = 0$. By hypothesis, λ is irrational and therefore $e^{2\pi i n\lambda} \ne 1$ for all $n \ne 0$. Therefore $c_n = 0$, $n \ne 0$, $\xi(\omega) = c_0$ (P-a.s.), and T is ergodic by Theorem 1.

On the other hand, let λ be rational, i.e. $\lambda = k/m$, where k and m are integers. Consider the set

$$A = \bigcup_{k=0}^{2m-2} \left\{\omega : \frac{k}{2m} \le \omega < \frac{k+1}{2m}\right\}.$$

It is clear that this set is invariant; but $P(A) = \frac{1}{2}$. Consequently T is not ergodic.

2. Definition 4. A measure-preserving transformation is *mixing* (or has the mixing property) if, for all A and $B \in F$,

$$\lim_{n\to\infty} \mathsf{P}(A \cap T^{-n}B) = \mathsf{P}(A)\mathsf{P}(B). \tag{1}$$

The following theorem establishes a connection between ergodicity and mixing.

Theorem 2. *Every mixing transformation T is ergodic.*

PROOF. Let $A \in \mathscr{F}$, $B \in \mathscr{I}$. Then $B = T^{-n}B$, $n \ge 1$, and therefore

$$\mathsf{P}(A \cap T^{-n}B) = \mathsf{P}(A \cap B)$$

for all $n \ge 1$. Because of (1), $\mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B)$. Hence we find, when $A = B$, that $\mathsf{P}(B) = \mathsf{P}^2(B)$, and consequently $\mathsf{P}(B) = 0$ or 1. This completes the proof.

3. PROBLEMS

1. Show that a random variable ξ is invariant if and only if it is $\mathscr{I}$-measurable.
2. Show that a set A is almost invariant if and only if either

$$\mathsf{P}(T^{-1}A \setminus A) = 0 \quad \text{or} \quad \mathsf{P}(A \setminus T^{-1}A) = 0.$$

3. Show that the transformation considered in the example of Subsection 1 of the present section is not mixing.
4. Show that a transformation is mixing if and only if, for all random variables ξ and η with $\mathsf{E}\xi^2 < \infty$ and $\mathsf{E}\eta^2 < \infty$,

$$\mathsf{E}\xi(T^n\omega)\eta(\omega) \to \mathsf{E}\xi(\omega)\mathsf{E}\eta(\omega), \qquad n \to \infty.$$

§3. Ergodic Theorems

1. Theorem 1 (Birkhoff and Khinchin). *Let T be a measure-preserving transformation and $\xi = \xi(\omega)$ a random variable with* $\mathsf{E}|\xi| < \infty$. *Then* (P-a.s.)

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} \xi(T^k\omega) = \mathsf{E}(\xi \mid \mathscr{I}). \tag{1}$$

If also T is ergodic then (P-*a.s.*)

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} \xi(T^k\omega) = \mathsf{E}\xi. \tag{2}$$

The proof given below is based on the following proposition, whose simple proof was given by A. Garsia (1965).

Lemma (Maximal Ergodic Theorem). *Let T be a measure-preserving transformation, let ξ be a random variable with $\mathsf{E}|\xi| < \infty$, and let*

$$S_k(\omega) = \xi(\omega) + \xi(T\omega) + \cdots + \xi(T^{k-1}\omega),$$

$$M_k(\omega) = \max\{0, S_1(\omega), \ldots, S_k(\omega)\}.$$

Then

$$\mathsf{E}[\xi(\omega)I_{\{M_n>0\}}(\omega)] \geq 0$$

for every $n \geq 1$.

Proof. If $n \geq k$, we have $M_n(T\omega) \geq S_k(T\omega)$ and therefore $\xi(\omega) + M_n(T\omega) \geq \xi(\omega) + S_k(T\omega) = S_{k+1}(\omega)$. Since it is evident that $\xi(\omega) \geq S_1(\omega) - M_n(T\omega)$, we have

$$\xi(\omega) \geq \max\{S_1(\omega), \ldots, S_n(\omega)\} - M_n(T\omega).$$

Therefore

$$\mathsf{E}[\xi(\omega)I_{\{M_n>0\}}(\omega)] \geq \mathsf{E}(\max(S_1(\omega), \ldots, S_n(\omega)) - M_n(T\omega)),$$

But $\max(S_1, \ldots, S_n) = M_n$ on the set $\{M_n > 0\}$. Consequently,

$$\begin{aligned}\mathsf{E}[\xi(\omega)I_{\{M_n>0\}}(\omega)] &\geq \mathsf{E}\{(M_n(\omega) - M_n(T\omega))I_{\{M_n(\omega)>0\}}\} \\ &\geq \mathsf{E}\{M_n(\omega) - M_n(T\omega)\} = 0,\end{aligned}$$

since if T is a measure-preserving transformation we have $\mathsf{E}M_n(\omega) = \mathsf{E}M_n(T\omega)$ (Problem 1, §1).

This completes the proof of the lemma.

Proof of the Theorem. Let us suppose that $\mathsf{E}(\xi|\mathscr{J}) = 0$ (otherwise replace ξ by $\xi - \mathsf{E}(\xi|\mathscr{J})$).

Let $\bar{\eta} = \overline{\lim}(S_n/n)$ and $\underline{\eta} = \underline{\lim}(S_n/n)$. It will be enough to establish that (P-a.s.)

$$0 \leq \underline{\eta} \leq \bar{\eta} \leq 0.$$

Consider the random variable $\bar{\eta} = \bar{\eta}(\omega)$. Since $\bar{\eta}(\omega) = \bar{\eta}(T\omega)$, the variable $\bar{\eta}$ is invariant and consequently, for every $\varepsilon > 0$, the set $A_\varepsilon = \{\bar{\eta}(\omega) > \varepsilon\}$ is also invariant. Let us introduce the new random variable

$$\xi^*(\omega) = (\xi(\omega) - \varepsilon)I_A(\omega),$$

and put

$$S_k^*(\omega) = \xi^*(\omega) + \cdots + \xi^*(T^{k-1}\omega), \qquad M_k^*(\omega) = \max(0, S_1^*, \ldots, S_k^*).$$

Then, by the lemma,

$$\mathsf{E}[\xi^* I_{\{M_n^* > 0\}}] \geq 0$$

for every $n \geq 1$. But as $n \to \infty$,

$$\{M_n^* > 0\} = \left\{\max_{1 \leq k \leq n} S_k^* > 0\right\} \uparrow \left\{\sup_{k \geq 1} S_k^* > 0\right\} = \left\{\sup_{k \geq 1} \frac{S_k^*}{k} > 0\right\}$$
$$= \left\{\sup_{k \geq 1} \frac{S_k}{k} > \varepsilon\right\} \cap A_\varepsilon = A_\varepsilon,$$

where the last equation follows because $\sup_{k \geq 1}(S_k^*/k) \geq \bar{\eta}$, and $A_\varepsilon = \{\omega : \bar{\eta} > \varepsilon\}$.

Moreover, $\mathsf{E}|\xi^*| \leq \mathsf{E}|\xi| + \varepsilon$. Hence, by the dominated convergence theorem,

$$0 \leq \mathsf{E}[\xi^* I_{\{M_n^* > 0\}}] \to \mathsf{E}[\xi^* I_A].$$

Thus

$$0 \leq \mathsf{E}[\xi^* I_{A_\varepsilon}] = \mathsf{E}[(\xi - \varepsilon) I_{A_\varepsilon}] = \mathsf{E}[\xi I_{A_\varepsilon}] - \varepsilon \mathsf{P}(A_\varepsilon)$$
$$= \mathsf{E}[\mathsf{E}(\xi | \mathscr{I}) I_{A_\varepsilon}] - \varepsilon \mathsf{P}(A_\varepsilon) = -\varepsilon \mathsf{P}(A_\varepsilon),$$

so that $\mathsf{P}(A_\varepsilon) = 0$ and therefore $\mathsf{P}(\bar{\eta} \leq 0) = 1$.

Similarly, if we consider $-\xi(\omega)$ instead of $\xi(\omega)$, we find that

$$\overline{\lim}\left(-\frac{S_n}{n}\right) = -\underline{\lim}\frac{S_n}{n} = -\underline{\eta}$$

and $\mathsf{P}(-\underline{\eta} \leq 0) = 1$, i.e. $\mathsf{P}(\underline{\eta} \geq 0) = 1$. Therefore $0 \leq \underline{\eta} \leq \bar{\eta} \leq 0$ (P-a.s.) and the first part of the theorem is established.

To prove the second part, we observe that since $\mathsf{E}(\xi | \mathscr{I})$ is an invariant random variable, we have $\mathsf{E}(\xi | \mathscr{I}) = \mathsf{E}\xi$ (P-a.s.) in the ergodic case.

This completes the proof of the theorem.

Corollary. *A measure-preserving transformation T is ergodic if and only if, for all A and $B \in \mathscr{F}$,*

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} \mathsf{P}(A \cap T^{-k}B) = \mathsf{P}(A)\mathsf{P}(B). \tag{3}$$

To prove the ergodicity of T we use $A = B \in \mathscr{I}$ in (3). Then $A \cap T^{-k}B = B$ and therefore $\mathsf{P}(B) = \mathsf{P}^2(B)$, i.e. $\mathsf{P}(B) = 0$ or 1. Conversely, let T be ergodic. Then if we apply (2) to the random variable $\xi = I_B(\omega)$, where $B \in \mathscr{F}$, we find that (P-a.s.)

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} I_{T^{-k}B}(\omega) = \mathsf{P}(B).$$

If we now integrate both sides over $A \in \mathscr{F}$ and use the dominated convergence theorem, we obtain (3) as required.

2. We now show that, under the hypotheses of Theorem 1, there is not only almost sure convergence in (1) and (2), but also convergence in mean. (This result will be used below in the proof of Theorem 3.)

Theorem 2. *Let T be a measure-preserving transformation and let $\xi = \xi(\omega)$ be a random variable with $\mathsf{E}|\xi| < \infty$. Then*

$$\mathsf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1}\xi(T^k\omega) - \mathsf{E}(\xi\,|\,\mathscr{I})\right| \to 0, \qquad n \to \infty. \tag{4}$$

If also T is ergodic, then

$$\mathsf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1}\xi(T^k\omega) - \mathsf{E}\xi\right| \to 0, \qquad n \to \infty. \tag{5}$$

PROOF. For every $\varepsilon > 0$ there is a bounded random variable $\eta(|\eta(\omega)| \le M)$ such that $\mathsf{E}|\xi - \eta| \le \varepsilon$. Then

$$\begin{aligned}\mathsf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1}\xi(T^k\omega) - \mathsf{E}(\xi\,|\,\mathscr{I})\right| &\le \mathsf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1}(\xi(T^k\omega) - \eta(T^k\omega))\right| \\ &\quad + \mathsf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1}\eta(T^k\omega) - \mathsf{E}(\eta\,|\,\mathscr{I})\right| + \mathsf{E}|\mathsf{E}(\xi\,|\,\mathscr{I}) - \mathsf{E}(\eta\,|\,\mathscr{I})|. \end{aligned} \tag{6}$$

Since $|\eta| \le M$, then by the dominated convergence theorem and by using (1) we find that the second term on the right of (6) tends to zero as $n \to \infty$. The first and third terms are each at most ε. Hence for sufficiently large n the left-hand side of (6) is less than 2ε, so that (4) is proved. Finally, if T is ergodic, then (5) follows from (4) and the remark that $\mathsf{E}(\xi\,|\,I) = \mathsf{E}\xi$ (P-a.s.).

This completes the proof of the theorem.

3. We now turn to the question of the validity of the ergodic theorem for stationary (strict sense) random sequences $\xi = (\xi_1, \xi_2, \ldots)$ defined on a probability space $(\Omega, \mathscr{F}, \mathsf{P})$. In general, $(\Omega, \mathscr{F}, \mathsf{P})$ need not carry any measure-preserving transformations, so that it is not possible to apply Theorem 1 directly. However, as we observed in §1, we can construct a coordinate probability space $(\tilde{\Omega}, \tilde{\mathscr{F}}, \tilde{\mathsf{P}})$, random variables $\tilde{\xi} = (\tilde{\xi}_1, \tilde{\xi}_2, \ldots)$, and a measure-preserving transformation $\tilde{T}$ such that $\tilde{\xi}_n(\tilde{\omega}) = \tilde{\xi}_1(\tilde{T}^{n-1}\tilde{\omega})$ and the distributions of ξ and $\tilde{\xi}$ are the same. Since such properties as almost sure convergence and convergence in the mean are defined only for probability distributions, from the convergence of $(1/n)\sum_{k=1}^{n}\tilde{\xi}_1(\tilde{T}^{k-1}\tilde{\omega})$ (P-a.s. and in mean) to a random variable $\tilde{\eta}$ it follows that $(1/n)\sum_{k=1}^{n}\xi_k(\omega)$ also converges (P-a.s. and in mean) to a random variable η such that $\eta \overset{d}{=} \tilde{\eta}$. It

follows from Theorem 1 that if $\tilde{\mathsf{E}}|\xi_1| < \infty$ then $\quad = \tilde{E}(\xi_1|\tilde{\mathscr{I}})$, where $\tilde{\mathscr{I}}$ is a collection of invariant sets ($\tilde{\mathsf{E}}$ is the average with respect to the measure $\tilde{\mathsf{P}}$). We now describe the structure of η.

Definition 1. A set $A \in \mathscr{F}$ is *invariant* with respect to the sequence ξ if there is a set $B \in \mathscr{B}(R\infty)$ such that for $n \geq 1$

$$A = \{\omega: (\xi_n, \xi_{n+1}, \ldots) \in B\}.$$

The collection of all such invariant sets is a σ-algebra, denoted by $\mathscr{I}_\xi$.

Definition 2. A stationary sequence ξ is *ergodic* if the measure of every invariant set is either 0 or 1.

Let us now show that the random variable η can be taken equal to $\mathsf{E}(\xi_1|\mathscr{I}_\xi)$. In fact, let $A \in \mathscr{I}_\xi$. Then since

$$\mathsf{E}\left|\frac{1}{n}\sum_{k=1}^{n-1}\xi_k - \eta\right| \to 0,$$

we have

$$\frac{1}{n}\sum_{k=1}^{n}\int_A \xi_k\, dP \to \int_A \eta\, d\mathsf{P}. \tag{7}$$

Let $B \in \mathscr{B}(R^\infty)$ be such that $A = \{\omega: (\xi_k, \xi_{k+1}, \ldots) \in B\}$ for all $k \geq 1$. Then since ξ is stationary,

$$\int_A \xi_k\, d\mathsf{P} = \int_{\{\omega: (\xi_k, \xi_{k+1}, \ldots) \in B\}} \xi_k\, d\mathsf{P} = \int_{\{\omega: (\xi_1, \xi_2, \ldots) \in B\}} \xi_1\, d\mathsf{P} = \int_A \xi_1\, d\mathsf{P}.$$

Hence it follows from (7) that for all $A \in \mathscr{I}_\xi$, which implies (see §7, Chapter II) that $\eta = \mathsf{E}(\xi_1|\mathscr{I}_\xi)$. Here $\mathsf{E}(\xi_1|\mathscr{I}_\xi) = \mathsf{E}\xi_1$ if ξ is ergodic.

Therefore we have proved the following theorem.

Theorem 3 (Ergodic Theorem). *Let $\xi = (\xi_1, \xi_2, \ldots)$ be a stationary* (*strict sense*) *random sequence with* $\mathsf{E}|\xi_1| < \infty$. *Then* ($\mathsf{P}$-*a.s., and in the mean*)

$$\lim \frac{1}{n}\sum_{k=1}^{n}\xi_k(\omega) = \mathsf{E}(\xi_1|\mathscr{I}_\xi).$$

If ξ is also an ergodic sequence, then (P-*a.s., and in the mean*)

$$\lim \frac{1}{n}\sum_{k=1}^{n}\xi_k(\omega) = \mathsf{E}\xi_1.$$

4. Problems

1. Let $\xi = (\xi_1, \xi_2, \ldots)$ be a Gaussian stationary sequence with $\mathsf{E}\xi_n = 0$ and covariance function $R(n) = \mathsf{E}\xi_{k+n}\xi_k$. Show that $R(n) \to 0$ is a sufficient condition for ξ to be ergodic.

2. Show that every sequence $\xi = (\xi_1, \xi_2, \ldots)$ of independent identically distributed random variables is ergodic.

3. Show that a stationary sequence ξ is ergodic if and only if

$$\frac{1}{n}\sum_{i=1}^{n} I_B(\xi_i, \ldots, \xi_{i+k}) \to \mathsf{P}((\xi_1, \ldots, \xi_{1+k}) \in B) \quad (\mathsf{P}\text{-a.s.})$$

for every $B \in \mathscr{B}(R)$, $k = 1, 2, \ldots$.

CHAPTER VI

Stationary (Wide Sense) Random Sequences. L^2-Theory

§1. Spectral Representation of the Covariance Function

1. According to the definition given in the preceding chapter, a random sequence $\xi = (\xi_1, \xi_2, \ldots)$ is stationary in the strict sense if, for every set $B \in \mathscr{B}(R^\infty)$ and every $n \geq 1$,

$$\mathsf{P}\{(\xi_1, \xi_2, \ldots) \in B\} = \mathsf{P}\{(\xi_{n+1}, \xi_{n+2}, \ldots) \in B\}. \tag{1}$$

It follows, in particular, that if $\mathsf{E}\xi_1^2 < \infty$ then $\mathsf{E}\xi_n$ is independent of n:

$$\mathsf{E}\xi_n = \mathsf{E}\xi_1, \tag{2}$$

and the covariance $\text{cov}(\xi_{n+m}\xi_n) = \mathsf{E}(\xi_{n+m} - \mathsf{E}\xi_{n+m})(\xi_n - \mathsf{E}\xi_n)$ depends only on m:

$$\text{cov}(\xi_{n+m}, \xi_n) = \text{cov}(\xi_{1+m}, \xi_1). \tag{3}$$

In the present chapter we study sequences that are stationary in the wide sense (and have finite second moments), namely those for which (1) is replaced by the (weaker) conditions (2) and (3).

The random variables ξ_n are understood to be defined for $n \in \mathbb{Z} = \{0, \pm 1, \ldots\}$ and to be complex-valued. The latter assumption not only does not complicate the theory, but makes it more elegant. It is also clear that results for real random variables can easily be obtained as special cases of the corresponding results for complex random variables.

Let $H^2 = H^2(\Omega, \mathscr{F}, \mathsf{P})$ be the space of (complex) random variables $\xi = \alpha + i\beta$, α, $\beta \in R$, with $\mathsf{E}|\xi|^2 < \infty$, where $|\xi|^2 = \alpha^2 + \beta^2$. If ξ and $\eta \in H^2$, we put

$$(\xi, \eta) = \mathsf{E}\xi\bar{\eta}, \tag{4}$$

where $\bar{\eta} = \alpha - i\beta$ is the complex conjugate of $\eta = \alpha + i\beta$ and

$$\|\xi\| = (\xi, \xi)^{1/2}. \tag{5}$$

As for real random variables, the space H^2 (more precisely, the space of equivalence classes of random variables; compare §§10 and 11 of Chapter II) is complete under the scalar product (ξ, η) and norm $\|\xi\|$. In accordance with the terminology of functional analysis, H^2 is called the complex (or unitary) Hilbert space (of random variables considered on the probability space $(\Omega, \mathscr{F}, \mathsf{P})$).

If $\xi, \eta \in H^2$ their *covariance* is

$$\operatorname{cov}(\xi, \eta) = \mathsf{E}(\xi - \mathsf{E}\xi)\overline{(\eta - \mathsf{E}\eta)}. \tag{6}$$

It follows from (4) and (6) that if $\mathsf{E}\xi = \mathsf{E}\eta = 0$ then

$$\operatorname{cov}(\xi, \eta) = (\xi, \eta). \tag{7}$$

Definition. A sequence of complex random variables $\xi = (\xi_n)_{n \in \mathbb{Z}}$ with $\mathsf{E}|\xi_n|^2 < \infty$, $n \in \mathbb{Z}$, is *stationary* (*in the wide sense*) if, for all $n \in \mathbb{Z}$,

$$\mathsf{E}\xi_n = \mathsf{E}\xi_0,$$
$$\operatorname{cov}(\xi_{k+n}, \xi_k) = \operatorname{cov}(\xi_n, \xi_0), \qquad k \in \mathbb{Z}. \tag{8}$$

As a matter of convenience, we shall always suppose that $\mathsf{E}\xi_0 = 0$. This involves no loss of generality, but does make it possible (by (7)) to identify the covariance with the scalar product and hence to apply the methods and results of the theory of Hilbert spaces.

Let us write

$$R(n) = \operatorname{cov}(\xi_n, \xi_0), \qquad n \in \mathbb{Z}, \tag{9}$$

and (assuming $R(0) = \mathsf{E}|\xi_0|^2 \neq 0$)

$$\rho(n) = \frac{R(n)}{R(0)}, \qquad n \in \mathbb{Z}. \tag{10}$$

We call $R(n)$ the *covariance function*, and $\rho(n)$, the *correlation function*, of the sequence ξ (assumed stationary in the wide sense).

It follows immediately from (9) that $R(n)$ is nonnegative-definite, i.e. for all complex numbers $a_1, \ldots, a_m$ and $t_1, \ldots, t_m \in \mathbb{Z}$, $m \geq 1$, we have

$$\sum_{i,j=1}^{m} a_i \bar{a}_j R(t_i - t_j) \geq 0. \tag{11}$$

It is then easy to deduce (either from (11) or directly from (9)) the following properties of the covariance function (see Problem 1):

$$R(0) \geq 0, \qquad R(-n) = \overline{R(n)}, \qquad |R(n)| \leq R(0),$$
$$|R(n) - R(m)|^2 \leq 2R(0)[R(0) - \operatorname{Re} R(n - m)]. \tag{12}$$

2. Let us give some examples of stationary sequences $\xi = (\xi_n)_{n \in \mathbb{Z}}$. (From now on, the words "in the wide sense" and the statement $n \in \mathbb{Z}$ will both be omitted.)

EXAMPLE 1. Let $\xi_n = \xi_0 \cdot g(n)$, where $\mathsf{E}\xi_0 = 0$, $\mathsf{E}\xi_0^2 = 1$ and $g = g(n)$ is a function. The sequence $\xi = (\xi_n)$ will be stationary if and only if $g(k + n)\overline{g(k)}$ depends only on n. Hence it is easy to see that there is a λ such that

$$g(n) = g(0)e^{i\lambda n}.$$

Consequently the sequence of random variables

$$\xi_n = \xi_0 \cdot g(0)e^{i\lambda n}$$

is stationary with

$$R(n) = |g(0)|^2 e^{i\lambda n}.$$

In particular, the random "constant" $\xi \equiv \xi_0$ is a stationary sequence.

EXAMPLE 2. *An almost periodic sequence.* Let

$$\xi_n = \sum_{k=1}^{N} z_k e^{i\lambda_k n}, \tag{13}$$

where $z_1, \dots, z_N$ are orthogonal ($\mathsf{E} z_i \bar{z}_j = 0, i \neq j$) random variables with zero means and $\mathsf{E}|z_k|^2 = \sigma_k^2 > 0$; $-\pi \le \lambda_k < \pi$, $k = 1, \dots, N$; $\lambda_i \neq \lambda_j$, $i \neq j$. The sequence $\xi = (\xi_n)$ is stationary with

$$R(n) = \sum_{k=1}^{N} \sigma_k^2 e^{i\lambda_k n}. \tag{14}$$

As a generalization of (13) we now suppose that

$$\xi_n = \sum_{k=-\infty}^{\infty} z_k e^{i\lambda_k n}, \tag{15}$$

where z_k, $k \in \mathbb{Z}$, have the same properties as in (13). If we suppose that $\sum_{k=-\infty}^{\infty} \sigma_k^2 < \infty$, the series on the right of (15) converges in mean-square and

$$R(n) = \sum_{k=-\infty}^{\infty} \sigma_k^2 e^{i\lambda_k n}. \tag{16}$$

Let us introduce the function

$$F(\lambda) = \sum_{\{k:\, \lambda_k \le \lambda\}} \sigma_k^2. \tag{17}$$

Then the covariance function (16) can be written as a Lebesgue–Stieltjes integral,

$$R(n) = \int_{-\pi}^{\pi} e^{i\lambda n}\, dF(\lambda). \tag{18}$$

The stationary sequence (15) is represented as a sum of "harmonics" $e^{i\lambda_k n}$ with "frequencies" λ_k and random "amplitudes" z_k of "intensities" $\sigma_k^2 = \mathsf{E}|z_k|^2$. Consequently the values of $F(\lambda)$ provide complete information on the "spectrum" of the sequence ξ, i.e. on the intensity with which each frequency appears in (15). By (18), the values of $F(\lambda)$ also completely determine the structure of the covariance function $R(n)$.

Up to a constant multiple, a (nondegenerate) $F(\lambda)$ is evidently a distribution function, which in the examples considered so far has been piecewise constant. It is quite remarkable that the covariance function of every stationary (wide sense) random sequence can be represented (see the theorem in Subsection 3) in the form (18), where $F(\lambda)$ is a distribution function (up to normalization), whose support is concentrated on $[-\pi, \pi)$, i.e. $F(\lambda) = 0$ for $\lambda < -\pi$ and $F(\lambda) = F(\pi)$ for $\lambda > \pi$.

The result on the integral representation of the covariance function, if compared with (15) and (16), suggests that every stationary sequence also admits an "integral" representation. This is in fact the case, as will be shown in §3 by using what we shall learn to call stochastic integrals with respect to orthogonal stochastic measures (§2).

EXAMPLE 3 (White noise). Let $\varepsilon = (\varepsilon_n)$ be an orthonormal sequence of random variables, $\mathsf{E}\varepsilon_n = 0$, $\mathsf{E}\varepsilon_i\bar{\varepsilon}_j = \delta_{ij}$, where δ_{ij} is the Kronecker delta. Such a sequence is evidently stationary, and

$$R(n) = \begin{cases} 1 & n = 0, \\ 0, & n \neq 0. \end{cases}$$

Observe that $R(n)$ can be represented in the form

$$R(n) = \int_{-\pi}^{\pi} e^{i\lambda n}\, dF(\lambda), \tag{19}$$

where

$$F(\lambda) = \int_{-\pi}^{\lambda} f(v)\, dv; \qquad f(\lambda) = \frac{1}{2\pi}, \qquad -\pi \leq \lambda < \pi. \tag{20}$$

Comparison of the spectral functions (17) and (20) shows that whereas the spectrum in Example 2 is discrete, in the present example it is absolutely continuous with constant "spectral density" $f(\lambda) \equiv \frac{1}{2}\pi$. In this sense we can say that the sequence $\varepsilon = (\varepsilon_n)$ "consists of harmonics of equal intensities." It is just this property that has led to calling such a sequence $\varepsilon = (\varepsilon_n)$ "white noise" by analogy with white light, which consists of different frequencies with the same intensities.

EXAMPLE 4 (Moving averages) Starting from the white noise $\varepsilon = (\varepsilon_n)$ introduced in Example 3, let us form the new sequence

$$\xi_n = \sum_{k=-\infty}^{\infty} a_k \varepsilon_{n-k}, \tag{21}$$

where a_k are complex numbers such that $\sum_{k=-\infty}^{\infty} |a_k|^2 < \infty$. By Parseval's equation,

$$\operatorname{cov}(\xi_{n+m}, \xi_m) = \operatorname{cov}(\xi_n, \xi_0) = \sum_{k=-\infty}^{\infty} a_{n+k}\bar{a}_k,$$

so that $\xi = (\xi_k)$ is a stationary sequence, which we call the sequence obtained from $\varepsilon = (\varepsilon_k)$ by a (*two-sided*) *moving average*.

In the special case when the a_k of negative index are zero, i.e.

$$\xi_n = \sum_{k=0}^{\infty} a_k \varepsilon_{n-k},$$

the sequence $\xi = (\xi_n)$ is a *one-sided moving average*. If, in addition, $a_k = 0$ for $k > p$, i.e. if

$$\xi_n = a_0\varepsilon_n + a_1\varepsilon_{n-1} + \cdots + a_p\varepsilon_{n-p}, \tag{22}$$

then $\xi = (\xi_n)$ is a *moving average of order p*.

We can show (Problem 5) that (22) has a covariance function of the form $R(n) = \int_{-\pi}^{\pi} e^{i\lambda n} f(\lambda)\, d\lambda$, where the spectral density is

$$f(\lambda) = \frac{1}{2\pi} |P(e^{-i\lambda})|^2 \tag{23}$$

with

$$P(z) = a_0 + a_1 z + \cdots + a_p z^p.$$

EXAMPLE 5 (Autoregression). Again let $\varepsilon = (\varepsilon_n)$ be white noise. We say that a random sequence $\xi = (\xi_n)$ is described by an *autoregressive model* of order q if

$$\xi_n + b_1\xi_{n-1} + \cdots + b_q\xi_{n-q} = \varepsilon_n. \tag{24}$$

Under what conditions on $b_1, \ldots, b_n$ can we say that (24) has a stationary solution? To find an answer, let us begin with the case $q = 1$:

$$\xi_n = \alpha\xi_{n-1} + \varepsilon_n, \tag{25}$$

where $\alpha = -b_1$. If $|\alpha| < 1$, it is easy to verify that the stationary sequence $\tilde{\xi} = (\tilde{\xi}_n)$ with

$$\tilde{\xi}_n = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{n-j} \tag{26}$$

is a solution of (25). (The series on the right of (26) converges in mean-square.) Let us now show that, in the class of stationary sequences $\xi = (\xi_n)$ (with finite second moments) this is the only solution. In fact, we find from (25), by successive iteration, that

$$\xi_n = \alpha\xi_{n-1} + \varepsilon_n = \alpha[\alpha\xi_{n-2} + \varepsilon_{n-1}] + \varepsilon_n = \cdots = \alpha^k\xi_{n-k} + \sum_{j=0}^{k-1} \alpha^j \varepsilon_{n-j}.$$

Hence it follows that

$$\mathsf{E}\left[\xi_n - \sum_{j=0}^{k-1} \alpha^j \varepsilon_{n-j}\right]^2 = \mathsf{E}[\alpha^k \xi_{n-k}]^2 = \alpha^{2k}\mathsf{E}\xi_{n-k}^2 = \alpha^{2k}\mathsf{E}\xi_0^2 \to 0, \qquad k \to \infty.$$

Therefore when $|\alpha| < 1$ a stationary solution of (25) exists and is representable as the one-sided moving average (26).

There is a similar result for every $q > 1$: if all the zeros of the polynomial

$$Q(z) = 1 + b_1 z + \cdots + b_q z^q \tag{27}$$

lie outside the unit disk, then the autoregression equation (24) has a unique stationary solution, which is representable as a one-sided moving average (Problem 2). Here the covariance function $R(n)$ can be represented (Problem 5) in the form

$$R(n) = \int_{-\pi}^{\pi} e^{i\lambda n}\, dF(\lambda), \qquad F(\lambda) = \int_{-\pi}^{\lambda} f(\nu)\, d\nu, \tag{28}$$

where

$$f(\lambda) = \frac{1}{2\pi} \cdot \frac{1}{|Q(e^{-i\lambda})|^2}. \tag{29}$$

In the special case $q = 1$, we find easily from (25) that $\mathsf{E}\xi_0 = 0$,

$$\mathsf{E}\xi_0^2 = \frac{1}{1 - |\alpha|^2},$$

and

$$R(n) = \frac{\alpha^n}{1 - |\alpha|^2}, \qquad n \geq 0$$

(when $n < 0$ we have $R(n) = \overline{R(-n)}$). Here

$$f(\lambda) = \frac{1}{2\pi} \cdot \frac{1}{|1 - \alpha e^{-i\lambda}|^2}.$$

EXAMPLE 6. This example illustrates how autoregression arises in the construction of probabilistic models in hydrology. Consider a body of water; we try to construct a probabilistic model of the deviations of the level of the water from its average value because of variations in the inflow and evaporation from the surface.

If we take a year as the unit of time and let H_n denote the water level in year n, we obtain the following *balance equation*:

$$H_{n+1} = H_n - KS(H_n) + \Sigma_{n+1}, \tag{30}$$

where Σ_{n+1} is the inflow in year $(n + 1)$, $S(H)$ is the area of the surface of the water at level H, and K is the coefficient of evaporation.

Let $\xi_n = H_n - \bar{H}$ be the deviation from the mean level (which is obtained from observations over many years) and suppose that $S(H) = S(\bar{H}) + c(H - \bar{H})$. Then it follows from the balance equation that ξ_n satisfies

$$\xi_{n+1} = \alpha\xi_n + \varepsilon_{n+1} \tag{31}$$

with $\alpha = 1 - cK$, $\varepsilon_n = \Sigma_n - KS(\bar{H})$. It is natural to assume that the random variables ε_n have zero means and are identically distributed. Then, as we showed in Example 5, equation (31) has (for $|\alpha| < 1$) a unique stationary solution, which we think of as the steady-state solution (with respect to time in years) of the oscillations of the level in the body of water.

As an example of practical conclusions that can be drawn from a (theoretical) model (31), we call attention to the possibility of predicting the level for the following year from the results of the observations of the present and preceding years. It turns out (see also Example 2 in §6) that (in the mean-square sense) the optimal linear estimator of ξ_{n+1} in terms of the values of $\ldots, \xi_{n-1}, \xi_n$ is simply $\alpha\xi_n$.

Example 7 (Autoregression and moving average (mixed model)). If we suppose that the right-hand side of (24) contains $\alpha_0\varepsilon_n + \alpha_1\varepsilon_{n-1} + \cdots + \alpha_p\varepsilon_{n-p}$ instead of ε_n, we obtain a mixed model with autoregression and moving average of order (p, q):

$$\xi_n + b_1\xi_{n-1} + \cdots + b_q\xi_{n-q} = a_0\varepsilon_n + a_1\varepsilon_{n-1} + \cdots + a_p\varepsilon_{n-p}. \tag{32}$$

Under the same hypotheses as in Example 5 on the zeros it will be shown later (Corollary 2 to Theorem 3 of §3) that (32) has the stationary solution $\xi = (\xi_n)$ for which the covariance function is $R(n) = \int_{-\pi}^{\pi} e^{i\lambda n}\, dF(\lambda)$ with $F(\lambda) = \int_{-\pi}^{\lambda} f(v)\, dv$, where

$$f(\lambda) = \frac{1}{2\pi} \cdot \left| \frac{P(e^{-i\lambda})}{Q(e^{-i\lambda})} \right|^2.$$

3. Theorem (Herglotz). *Let $R(n)$ be the covariance function of a stationary (wide sense) random sequence with zero mean. Then there is, on*

$$([-\pi, \pi), \mathscr{B}([-\pi, \pi))),$$

a finite measure $F = F(B)$, $B \in \mathscr{B}([-\pi, \pi))$, such that for every $n \in \mathbb{Z}$

$$R(n) = \int_{-\pi}^{\pi} e^{i\lambda n}\, F(d\lambda). \tag{33}$$

Proof. For $N \geq 1$ and $\lambda \in [-\pi, \pi]$, put

$$f_N(\lambda) = \frac{1}{2\pi N} \sum_{k=1}^{N} \sum_{l=1}^{N} R(k - l)e^{-ik\lambda}e^{il\lambda}. \tag{34}$$

Since $R(n)$ is nonnegative definite, $f_N(\lambda)$ is nonnegative. Since there are $N - |m|$ pairs (k, l) for which $k - l = m$, we have

$$f_N(\lambda) = \frac{1}{2\pi} \sum_{|m|<N} \left(1 - \frac{|m|}{N}\right) R(m) e^{-im\lambda}. \tag{35}$$

Let

$$F_N(B) = \int_B f_N(\lambda)\, d\lambda, \qquad B \in \mathscr{B}([-\pi, \pi)).$$

Then

$$\int_{-\pi}^{\pi} e^{i\lambda n} F_N(d\lambda) = \int_{-\pi}^{\pi} e^{i\lambda n} f_N(\lambda)\, d\lambda = \begin{cases} \left(1 - \dfrac{|n|}{N}\right) R(n), & |n| < N, \\ 0, & |n| \geq N. \end{cases} \tag{36}$$

The measures F_N, $N \geq 1$, are supported on the interval $[-\pi, \pi]$ and $F_N([-\pi, \pi]) = R(0) < \infty$ for all $N \geq 1$. Consequently the family of measures $\{F_N\}$, $N \geq 1$, is tight, and by Prokhorov's theorem (Theorem 1 of §2 of Chapter III) there are a sequence $\{N_k\} \subseteq \{N\}$ and a measure F such that $F_{N_k} \xrightarrow{w} F$. (The concepts of tightness, relative compactness, and weak convergence, together with Prokhorov's theorem, can be extended in an obvious way from probability measures to any finite measures.)

It then follows from (36) that

$$\int_{-\pi}^{\pi} e^{i\lambda n} F(d\lambda) = \lim_{N_k \to \infty} \int_{-\pi}^{\pi} e^{i\lambda n} F_{N_k}(d\lambda) = R(n).$$

The measure F so constructed is supported on $[-\pi, \pi]$. Without changing the integral $\int_{-\infty}^{\infty} e^{i\lambda n} F(d\lambda)$, we can redefine F by transferring the "mass" $F(\{\pi\})$, which is concentrated at π, to $-\pi$. The resulting new measure (which we again denote by F) will be supported on $[-\pi, \pi)$.

This completes the proof of the theorem.

Remark 1. The measure $F = F(B)$ involved in (33) is known as the spectral measure, and $F(\lambda) = F([-\pi, \lambda])$ as the spectral function, of the stationary sequence with covariance function $R(n)$.

In Example 2 above the spectral measure was discrete (concentrated at λ_k, $k = 0, \pm 1, \ldots$). In Examples 3–6 the spectral measures were absolutely continuous.

Remark 2. The spectral measure F is uniquely defined by the covariance function. In fact, let F_1 and F_2 be two spectral measures and let

$$\int_{-\pi}^{\pi} e^{i\lambda n} F_1(d\lambda) = \int_{-\pi}^{\pi} e^{i\lambda n} F_2(d\lambda), \qquad n \in \mathbb{Z}.$$

Since every bounded continuous function $g(\lambda)$ can be uniformly approximated on $[-\pi, \pi)$ by trigonometric polynomials, we have

$$\int_{-\pi}^{\pi} g(\lambda)F_1(d\lambda) = \int_{-\pi}^{\pi} g(\lambda)F_2(d\lambda).$$

It follows (compare the proof in Theorem 2, §12, Chapter II) that $F_1(B) = F_2(B)$ for all $B \in \mathscr{B}([-\pi, \pi))$.

Remark 3. If $\xi = (\xi_n)$ is a stationary sequence of real random variables ξ_n, then

$$R(n) = \int_{-\pi}^{\pi} \cos \lambda n\, F(d\lambda).$$

4. Problems

1. Derive (12) from (11).

2. Show that the autoregression equation (24) has a stationary solution if all the zeros of the polynomial $Q(z)$ defined by (27) lie outside the unit disk.

3. Prove that the covariance function (28) admits the representation (29) with spectral density given by (30).

4. Show that the sequence $\xi = (\xi_n)$ of random variables, where

$$\xi_n = \sum_{k=1}^{\infty} (\alpha_k \sin \lambda_k n + \beta_k \cos \lambda_k n)$$

and α_k and β_k are real random variables, can be represented in the form

$$\xi_n = \sum_{k=-\infty}^{\infty} z_k e^{i\lambda_k n}$$

with $z_k = \frac{1}{2}(\beta_k - i\alpha_k)$ for $k \geq 0$ and $z_k = \bar{z}_{-k}$, $\lambda_k = -\lambda_{-k}$ for $k < 0$.

5. Show that the spectral functions of the sequences (22) and (24) have densities given respectively by (23) and (29).

6. Show that if $\sum |R(n)| < \infty$, the spectral function $F(\lambda)$ has density $f(\lambda)$ given by

$$f(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-i\lambda n} R(n).$$

§2. Orthogonal Stochastic Measures and Stochastic Integrals

1. As we observed in §1, the integral representation of the covariance function and the example of a stationary sequence

$$\xi_n = \sum_{k=-\infty}^{\infty} z_k e^{i\lambda_k n} \tag{1}$$

with pairwise orthogonal random variables z_k, $k \in \mathbb{Z}$, suggest the possibility of representing an arbitrary stationary sequence as a corresponding integral generalization of (1).

If we put

$$Z(\lambda) = \sum_{\{k:\, \lambda_k \le \lambda\}} z_k, \tag{2}$$

we can rewrite (1) in the form

$$\xi_n = \sum_{k=-\infty}^{\infty} e^{i\lambda_k n} \Delta Z(\lambda_k), \tag{3}$$

where $\Delta Z(\lambda_k) \equiv Z(\lambda_k) - Z(\lambda_k -) = z_k$.

The right-hand side of (3) reminds us of an approximating sum for an integral $\int_{-\pi}^{\pi} e^{i\lambda n}\, dZ(\lambda)$ of Riemann–Stieltjes type. However, in the present case $Z(\lambda)$ is a random function (it also depends on ω). Hence it is clear that for an integral representation of a general stationary sequence we need to use functions $Z(\lambda)$ that do not have bounded variation for each ω. Consequently the simple interpretation of $\int_{-\pi}^{\pi} e^{i\lambda n}\, dZ(\lambda)$ as a Riemann–Stieltjes integral for each ω is inapplicable.

2. By analogy with the general ideas of the Lebesgue, Lebesgue–Stieltjes and Riemann–Stieltjes integrals (§6, Chapter II), we begin by defining stochastic measure.

Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space, and let E be a subset, with an algebra $\mathscr{E}_0$ of subsets and the σ-algebra $\mathscr{E}$ generated by $\mathscr{E}_0$.

Definition 1. A complex-valued function $Z(\Delta) = Z(\omega; \Delta)$, defined for $\omega \in \Omega$ and $\Delta \in \mathscr{E}_0$, is a *finitely additive stochastic measure* if

(1) $\mathsf{E}|Z(\Delta)|^2 < \infty$ for every $\Delta \in \mathscr{E}_0$;

(2) for every pair Δ_1 and Δ_2 of disjoint sets in $\mathscr{E}_0$,

$$Z(\Delta_1 + \Delta_2) = Z(\Delta_1) + Z(\Delta_2) \quad (\mathsf{P}\text{-a.s.}) \tag{4}$$

Definition 2. A finitely additive stochastic measure $Z(\Delta)$ is an *elementary stochastic measure* if, for all disjoint sets $\Delta_1, \Delta_2, \ldots$ of $\mathscr{E}_0$ such that $\Delta = \sum_{k=1}^{\infty} \Delta_k \in \mathscr{E}_0$,

$$\mathsf{E}\left| Z(\Delta) - \sum_{k=1}^{n} Z(\Delta_k) \right|^2 \to 0, \qquad n \to \infty. \tag{5}$$

Remark 1. In this definition of an elementary stochastic measure on subsets of $\mathscr{E}_0$, it is assumed that its values are in the Hilbert space $H^2 = H^2(\Omega, \mathscr{F}, \mathsf{P})$, and that countable additivity is understood in the mean-square sense (5). There are other definitions of stochastic measures, without the requirement of the existence of second moments, where countable additivity is defined (for example) in terms of convergence in probability or with probability one.

Remark 2. In analogy with nonstochastic measures, one can show that for finitely additive stochastic measures the condition (5) of countable additivity (in the mean-square sense) is equivalent to continuity (in the mean-square sense) at "zero":

$$\mathsf{E}|Z(\Delta_n)|^2 \to 0, \qquad \Delta_n \downarrow \varnothing, \qquad \Delta_n \in \mathscr{E}_0. \tag{6}$$

A particularly important class of elementary stochastic measures consists of those that are orthogonal according to the following definition.

Definition 3. An elementary stochastic measure $Z(\Delta)$, $\Delta \in \mathscr{E}_0$, is *orthogonal* (or a *measure with orthogonal values*) if

$$\mathsf{E}Z(\Delta_1)\overline{Z(\Delta_2)} = 0 \tag{7}$$

for every pair of disjoint sets Δ_1 and Δ_2 in $\mathscr{E}_0$; or, equivalently, if

$$\mathsf{E}Z(\Delta_1)\overline{Z}(\Delta_2) = \mathsf{E}|Z(\Delta_1 \cap \Delta_2)|^2 \tag{8}$$

for all Δ_1 and Δ_2 in $\mathscr{E}_0$.

We write

$$m(\Delta) = \mathsf{E}|Z(\Delta)|^2, \qquad \Delta \in \mathscr{E}_0. \tag{9}$$

For elementary orthogonal stochastic measures, the set function $m = m(\Delta)$, $\Delta \in \mathscr{E}_0$, is, as is easily verified, a finite measure, and consequently by Carathéodory's theorem (§3, Chapter II) it can be extended to $(E, \mathscr{E})$. The resulting measure will again be denoted by $m = m(\Delta)$ and called the *structure function* (of the elementary orthogonal stochastic measure $Z = Z(\Delta)$, $\Delta \in \mathscr{E}_0$).

The following question now arises naturally: since the set function $m = m(\Delta)$ defined on $(E, \mathscr{E}_0)$ admits an extension to $(E, \mathscr{E})$, where $\mathscr{E} = \sigma(\mathscr{E}_0)$, cannot an elementary orthogonal stochastic measure $Z = Z(\Delta)$, $\Delta \in \mathscr{E}_0$, be extended to sets Δ in E in such a way that $\mathsf{E}|Z(\Delta)|^2 = m(\Delta)$, $\Delta \in \mathscr{E}$?

The answer is affirmative, as follows from the construction given below. This construction, at the same time, leads to the stochastic integral which we need for the integral representation of stationary sequences.

3. Let $Z = Z(\Delta)$ be an elementary orthogonal stochastic measure, $\Delta \in \mathscr{E}_0$, with structure function $m = m(\Delta)$, $\Delta \in \mathscr{E}$. For every function

$$f(\lambda) = \sum f_k I_{\Delta_k}, \qquad \Delta_k \in \mathscr{E}_0, \tag{10}$$

with only a finite number of different (complex) values, we define the random variable

$$\mathscr{I}(f) = \sum f_k Z(\Delta_k).$$

Let $L^2 = L^2(E, \mathscr{E}, m)$ be the Hilbert space of complex-valued functions with the scalar product

$$\langle f, g \rangle = \int_E f(\lambda)\bar{g}(\lambda)m(d\lambda)$$

and the norm $\|f\| = \langle f, f \rangle^{1/2}$, and let $H^2 = H^2(\Omega, \mathscr{F}, \mathsf{P})$ be the Hilbert space of complex-valued random variables with the scalar product

$$(\xi, \eta) = \mathsf{E}\xi\bar{\eta}$$

and the norm $\|\xi\| = (\xi, \xi)^{1/2}$.

Then it is clear that, for every pair of functions f and g of the form (10),

$$(\mathscr{I}(f), \mathscr{I}(g)) = \langle f, g \rangle$$

and

$$\|\mathscr{I}(f)\|^2 = \|f\|^2 = \int_E |f(\lambda)|^2 m(d\lambda).$$

Now let $f \in L^2$ and let $\{f_n\}$ be functions of type (10) such that $\|f - f_n\| \to 0$, $n \to \infty$ (the existence of such functions follows from Problem 2). Consequently

$$\|\mathscr{I}(f_n) - \mathscr{I}(f_m)\| = \|f_n - f_m\| \to 0, \qquad n, m \to \infty.$$

Therefore the sequence $\{\mathscr{I}(f_n)\}$ is fundamental in the mean-square sense and by Theorem 7, §10, Chapter II, there is a random variable (denoted by $\mathscr{I}(f)$) such that $\mathscr{I}(f) \in H^2$ and $\|\mathscr{I}(f_n) - \mathscr{I}(f)\| \to 0$, $n \to \infty$.

The random variable $\mathscr{I}(f)$ constructed in this way is uniquely defined (up to stochastic equivalence) and is independent of the choice of the approximating sequence $\{f_n\}$. We call it the *stochastic integral* of $f \in L^2$ with respect to the elementary orthogonal stochastic measure Z and denote it by

$$\mathscr{I}(f) = \int_E f(\lambda)Z(d\lambda).$$

We note the following basic properties of the stochastic integral $\mathscr{I}(f)$; these are direct consequences of its construction (Problem 1). Let g, f, and $f_n \in L^2$. Then

$$(\mathscr{I}(f), \mathscr{I}(g)) = \langle f, g \rangle; \tag{11}$$

$$\|\mathscr{I}(f)\| = \|f\|; \tag{12}$$

$$\mathscr{I}(af + bg) = a\mathscr{I}(f) + b\mathscr{I}(g) \quad (\mathsf{P}\text{-a.s.}) \tag{13}$$

where a and b are constants;

$$\|\mathscr{I}(f_n) - \mathscr{I}(f)\| \to 0, \tag{14}$$

if $\|f_n - f\| \to 0$, $n \to \infty$.

4. Let us use the preceding definition of the stochastic integral to *extend* the elementary stochastic measure $Z(\Delta)$, $\Delta \in \mathscr{E}_0$, to sets in $\mathscr{E} = \sigma(\mathscr{E}_0)$.

Since m is assumed to be finite, we have $I_\Delta = I_\Delta(\lambda) \in L^2$ for all $\Delta \in \mathscr{E}$. Write $\tilde{Z}(\Delta) = \mathscr{I}(I_\Delta)$. It is clear that $\tilde{Z}(\Delta) = Z(\Delta)$ for $\Delta \in \mathscr{E}_0$. It follows from (13) that if $\Delta_1 \cap \Delta_2 = \varnothing$ for Δ_1 and $\Delta_2 \in \mathscr{E}$, then

$$\tilde{Z}(\Delta_1 + \Delta_2) = \tilde{Z}(\Delta_1) + \tilde{Z}(\Delta_2) \quad (\text{P-a.s.})$$

and it follows from (12) that

$$\mathsf{E}\,|\tilde{Z}(\Delta)|^2 = m(\Delta), \qquad \Delta \in \mathscr{E}.$$

Let us show that the random set function $\tilde{Z}(\Delta)$, $\Delta \in \mathscr{E}$, is countably additive in the mean-square sense. In fact, let $\Delta_k \in \mathscr{E}$ and $\Delta = \sum_{k=1}^{\infty} \Delta_k$. Then

$$\tilde{Z}(\Delta) - \sum_{k=1}^{n} \tilde{Z}(\Delta_k) = \mathscr{I}(g_n),$$

where

$$g_n(\lambda) = I_\Delta(\lambda) - \sum_{k=1}^{n} I_{\Delta_k}(\lambda) = I_{\Sigma_n}(\lambda), \qquad \Sigma_n = \sum_{k=n+1}^{\infty} \Delta_k.$$

But

$$\mathsf{E}\,|\mathscr{I}(g_n)|^2 = \|g_n\|^2 = m(\Sigma_n) \downarrow 0, \qquad n \to \infty,$$

i.e.

$$\mathsf{E}\,\Big|\tilde{Z}(\Delta) - \sum_{k=1}^{n} \tilde{Z}(\Delta_k)\Big|^2 \to 0, \qquad n \to \infty.$$

It also follows from (11) that

$$\mathsf{E}\,\tilde{Z}(\Delta_1)\overline{\tilde{Z}(\Delta_2)} = 0$$

when $\Delta_1 \cap \Delta_2 = \varnothing$, $\Delta_1, \Delta_2 \in \mathscr{E}$.

Thus our function $\tilde{Z}(\Delta)$, defined on $\Delta \in \mathscr{E}$, is countably additive in the mean-square sense and coincides with $Z(\Delta)$ on the sets $\Delta \in \mathscr{E}_0$. We shall call $\tilde{Z}(\Delta)$, $\Delta \in \mathscr{E}$, an orthogonal stochastic measure (since it is an extension of the elementary orthogonal stochastic measure $Z(\Delta)$) with respect to the structure function $m(\Delta)$, $\Delta \in \mathscr{E}$; and we call the integral $\mathscr{I}(f) = \int_E f(\lambda)\tilde{Z}(d\lambda)$, defined above, a stochastic integral with respect to this measure.

5. We now consider the case $(E, \mathscr{E}) = (R, \mathscr{B}(R))$, which is the most important for our purposes. As we know (§3, Chapter II), there is a one-to-one correspondence between finite measures $m = m(\Delta)$ on $(R, \mathscr{B}(R))$ and certain (generalized) distribution functions $G = G(x)$, with $m(a, b] = G(b) - G(a)$.

It turns out that there is something similar for orthogonal stochastic measures. We introduce the following definition.

Definition 4. A set of (complex-valued) random variables $\{Z_\lambda\}$, $\lambda \in R$, defined on $(\Omega, \mathscr{F}, \mathsf{P})$, is a *random process with orthogonal increments* if

(1) $\mathsf{E}|Z_\lambda|^2 < \infty, \lambda \in R$;
(2) for every $\lambda \in R$

$$\mathsf{E}|Z_\lambda - Z_{\lambda_n}|^2 \to 0, \qquad \lambda_n \downarrow \lambda, \qquad \lambda_n \in R;$$

(3) whenever $\lambda_1 < \lambda_2 < \lambda_3 < \lambda_4$,

$$\mathsf{E}(Z_{\lambda_4} - Z_{\lambda_3})(\overline{Z_{\lambda_2} - Z_{\lambda_1}}) = 0.$$

Condition (3) is the condition of orthogonal increments. Condition (1) means that $Z_\lambda \in H^2$. Finally, condition (2) is included for technical reasons; it is a requirement of continuity on the right (in the mean-square sense) at each $\lambda \in R$.

Let $Z = Z(\Delta)$ be an orthogonal stochastic measure with respect to the structure function $m = m(\Delta)$, of finite measure, with the (generalized) distribution function $G(\lambda)$. Let us put

$$Z_\lambda = Z(-\infty, \lambda].$$

Then

$$\mathsf{E}|Z_\lambda|^2 = m(-\infty, \lambda] = G(\lambda) < \infty, \qquad \mathsf{E}|Z_\lambda - Z_{\lambda_n}|^2 = m(\lambda_n, \lambda] \downarrow 0, \lambda_n \downarrow \lambda,$$

and (evidently) 3) is satisfied also. Then this process $\{Z_\lambda\}$ is called a process *with orthogonal increments.*

On the other hand, if $\{Z_\lambda\}$ is such a process with $\mathsf{E}|Z_\lambda|^2 = G(\lambda)$, $G(-\infty) = 0$, $G(+\infty) < \infty$, we put

$$Z(\Delta) = Z_b - Z_a$$

when $\Delta = (a, b]$. Let $\mathscr{E}_0$ be the algebra of sets

$$\Delta = \sum_{k=1}^{n} (a_k, b_k] \quad \text{and} \quad Z(\Delta) = \sum_{k=1}^{n} Z(a_k, b_k].$$

It is clear that

$$\mathsf{E}|Z(\Delta)|^2 = m(\Delta),$$

where $m(\Delta) = \sum_{k=1}^{n} [G(b_k) - G(a_k)]$ and

$$\mathsf{E}Z(\Delta_1)\overline{Z}(\Delta_2) = 0$$

for disjoint intervals $\Delta_1 = (a_1, b_1]$ and $\Delta_2 = (a_2, b_2]$.

Therefore $Z = Z(\Delta)$, $\Delta \in \mathscr{E}_0$, is an elementary stochastic measure with orthogonal values. The set function $m = m(\Delta)$, $\Delta \in \mathscr{E}_0$, has a unique extension to a measure on $\mathscr{E} = \mathscr{B}(R)$, and it follows from the preceding constructions that $Z = Z(\Delta)$, $\Delta \in \mathscr{E}_0$, can also be extended to the set $\Delta \in \mathscr{E}$, where $\mathscr{E} = \mathscr{B}(R)$, and $\mathsf{E}|Z(\Delta)|^2 = m(\Delta)$, $\Delta \in B(\mathscr{R})$.

Therefore there is a one-to-one correspondence between processes $\{Z_\lambda\}$, $\lambda \in R$, with orthogonal increments and $\mathsf{E}|Z_\lambda|^2 = G(\lambda)$, $G(-\infty) = 0$, $G(+\infty) < \infty$, and orthogonal stochastic measures $Z = Z(\Delta)$, $\Delta \in \mathscr{B}(R)$, with structure functions $m = m(\Delta)$. The correspondence is given by

$$Z_\lambda = Z(-\infty, \lambda], \qquad G(\lambda) = m(-\infty, \lambda]$$

and

$$Z(a, b] = Z_b - Z_a, \qquad m(a, b] = G(b) - G(a).$$

By analogy with the usual notation of the theory of Riemann–Stieltjes integration, the stochastic integral $\int_R f(\lambda)\, dZ_\lambda$, where $\{Z_\lambda\}$ is a process with orthogonal increments, means the stochastic integral $\int_R f(\lambda)Z(d\lambda)$ with respect to the corresponding process with an orthogonal stochastic measure.

6. Problems

1. Prove the equivalence of (5) and (6).

2. Let $f \in L^2$. Using the results of Chapter II (Theorem 1 of §4, the Corollary to Theorem 3 of §6, and Problem 9 of §3), prove that there is a sequence of functions f_n of the form (10) such that $\|f - f_n\| \to 0$, $n \to \infty$.

3. Establish the following properties of an orthogonal stochastic measure $Z(\Delta)$ with structure function $m(\Delta)$:

$$\mathsf{E}|Z(\Delta_1) - Z(\Delta_2)|^2 = m(\Delta_1 \triangle \Delta_2),$$
$$Z(\Delta_1 \setminus \Delta_2) = Z(\Delta_1) - Z(\Delta_1 \cap \Delta_2) \quad (\mathsf{P}\text{-a.s.}),$$
$$Z(\Delta_1 \triangle \Delta_2) = Z(\Delta_1) + Z(\Delta_2) - 2Z(\Delta_1 \cap \Delta_2) \quad (\mathsf{P}\text{-a.s.}).$$

§3. Spectral Representation of Stationary (Wide Sense) Sequences

1. If $\xi = (\xi_n)$ is a stationary sequence with $\mathsf{E}\xi_n = 0$, $n \in \mathbb{Z}$, then by the theorem of §1, there is a finite measure $F = F(\Delta)$ on $([-\pi, \pi), \mathscr{B}([-\pi, \pi)))$ such that its covariance function $R(n) = \operatorname{cov}(\xi_{k+n}, \xi_k)$ admits the spectral representation

$$R(n) = \int_{-\pi}^{\pi} e^{i\lambda n} F(d\lambda). \tag{1}$$

The following result provides the corresponding spectral representation of the sequence $\xi = (\xi_n)$, $n \in \mathbb{Z}$, itself.

Theorem 1. *There is an orthogonal stochastic measure* $Z = Z(\Delta)$, $\Delta \in \mathscr{B}([-\pi, \pi))$, *such that for every* $n \in \mathbb{Z}$ (P-*a.s.*)

$$\xi_n = \int_{-\pi}^{\pi} e^{i\lambda n} Z(d\lambda). \tag{2}$$

Moreover, $\mathsf{E}\,|Z(\Delta)|^2 = F(\Delta)$.

The simplest proof is based on properties of Hilbert spaces.

Let $L^2(F) = L^2(E, \mathscr{E}, F)$ be a Hilbert space of complex functions, $E = [-\pi, \pi)$, $\mathscr{E} = \mathscr{B}([-\pi, \pi))$, with the scalar product

$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(\lambda)\bar{g}(\lambda)F(d\lambda), \tag{3}$$

and let $L_0^2(F)$ be the linear manifold ($L_0^2(F) \subseteq L^2(F)$) spanned by $e_n = e_n(\lambda)$, $n \in \mathbb{Z}$, where $e_n(\lambda) = e^{i\lambda n}$.

Observe that since $E = [-\pi, \pi)$ and F is finite, the closure of $L_0^2(F)$ coincides (Problem 1) with $L^2(F)$:

$$\overline{L_0^2}(F) = L^2(F).$$

Also let $L_0^2(\xi)$ be the linear manifold spanned by the random variables ξ_n, $n \in \mathbb{Z}$, and let $L^2(\xi)$ be its closure in the mean-square sense (with respect to P).

We establish a one-to-one correspondence between the elements of $L_0^2(F)$ and $L_0^2(\xi)$, denoted by "$\leftrightarrow$", by setting

$$e_n \leftrightarrow \xi_n, \qquad n \in \mathbb{Z}, \tag{4}$$

and defining it for elements in general (more precisely, for equivalence classes of elements) by linearity:

$$\sum \alpha_n e_n \leftrightarrow \sum \alpha_n \xi_n \tag{5}$$

(here we suppose that only finitely many of the complex numbers α_n are different from zero).

Observe that (5) is a consistent definition, in the sense that $\sum \alpha_n e_n = 0$ almost everywhere with respect to F if and only if $\sum \alpha_n \xi_n = 0$ (P-a.s.).

The correspondence "$\leftrightarrow$" is an *isometry*, i.e. it preserves scalar products. In fact, by (3),

$$\langle e_n, e_m \rangle = \int_{-\pi}^{\pi} e_n(\lambda)\overline{e_m}(\lambda)F(d\lambda) = \int_{-\pi}^{\pi} e^{i\lambda(n-m)}F(d\lambda) = R(n-m)$$
$$= \mathsf{E}\xi_n\bar{\xi}_m = (\xi_n, \xi_m)$$

and similarly

$$\langle \sum \alpha_n e_n, \sum \beta_n e_n \rangle = (\sum \alpha_n \xi_n, \sum \beta_n \xi_n). \tag{6}$$

Now let $\eta \in L^2(\xi)$. Since $L^2(\xi) = \bar{L}_0^2(\xi)$, there is a sequence $\{\eta_n\}$ such that $\eta_n \in L_0^2(\xi)$ and $\|\eta_n - \eta\| \to 0$, $n \to \infty$. Consequently $\{\eta_n\}$ is a fundamental sequence and therefore so is the sequence $\{f_n\}$, where $f_n \in L_0^2(F)$ and $f_n \leftrightarrow \eta_n$. The space $L^2(F)$ is complete and consequently there is an $f \in L^2(F)$ such that $\|f_n - f\| \to 0$.

There is an evident converse: if $f \in L^2(F)$ and $\|f - f_n\| \to 0$, $f_n \in L_0^2(F)$, there is an element η of $L^2(\xi)$ such that $\|\eta - \eta_n\| \to 0$, $\eta_n \in L_0^2(\xi)$ and $\eta_n \leftrightarrow f_n$.

Up to now the isometry "$\leftrightarrow$" has been defined only as between elements of $L_0^2(\xi)$ and $L_0^2(F)$. We extend it by continuity, taking $f \leftrightarrow \eta$ when f and η are the elements considered above. It is easily verified that the correspondence obtained in this way is one-to-one (between classes of equivalent random variables and of functions), is linear, and preserves scalar products.

Consider the function $f(\lambda) = I_\Delta(\lambda)$, where $\Delta \in \mathscr{B}([-\pi, \pi))$, and let $Z(\Delta)$ be the element of $L^2(\xi)$ such that $I_\Delta(\lambda) \leftrightarrow Z(\lambda)$. It is clear that $\|I_\Delta(\lambda)\|^2 = F(\Delta)$ and therefore $\mathsf{E}|Z(\Delta)|^2 = F(\Delta)$. Moreover, if $\Delta_1 \cap \Delta_2 = \varnothing$, we have $\mathsf{E}Z(\Delta_1)\overline{Z(\Delta_2)} = 0$ and $\mathsf{E}|Z(\Delta) - \sum_{k=1}^n Z(\Delta_k)|^2 \to 0$, $n \to \infty$, where $\Delta = \sum_{k=1}^\infty \Delta_k$.

Hence the family of elements $Z(\Delta)$, $\Delta \in \mathscr{B}([-\pi, \pi))$, form an orthogonal stochastic measure, with respect to which (according to §2) we can define the stochastic integral

$$\mathscr{I}(f) = \int_{-\pi}^{\pi} f(\lambda) Z(d\lambda), \qquad f \in L^2(F).$$

Let $f \in L^2(F)$ and $\eta \leftrightarrow f$. Denote the element η by $\Phi(f)$ (more precisely, select single representatives from the corresponding equivalence classes of random variables or functions). Let us show that (P-a.s.)

$$\mathscr{I}(f) = \Phi(f). \tag{7}$$

In fact, if

$$f(\lambda) = \sum \alpha_k I_{\Delta_k}(\lambda) \tag{8}$$

is a finite linear combination of functions $I_{\Delta_k}(\lambda)$, $\Delta_k = (a_k, b_k]$, then, by the very definition of the stochastic integral, $\mathscr{I}(f) = \sum \alpha_k Z(\Delta_k)$, which is evidently equal to $\Phi(f)$. Therefore (7) is valid for functions of the form (8). But if $f \in L^2(F)$ and $\|f_n - f\| \to 0$, where f_n are functions of the form (8), then $\|\Phi(f_n) - \Phi(f)\| \to 0$ and $\|\mathscr{I}(f_n) - \mathscr{I}(f)\| \to 0$ (by (2.14)). Therefore $\Phi(f) = \mathscr{I}(f)$ (P-a.s.).

Consider the function $f(\lambda) = e^{i\lambda n}$. Then $\Phi(e^{i\lambda n}) = \xi_n$ by (4), but on the other hand $\mathscr{I}(e^{i\lambda n}) = \int_{-\pi}^{\pi} e^{i\lambda n} Z(d\lambda)$. Therefore

$$\xi_n = \int_{-\pi}^{\pi} e^{i\lambda n} Z(d\lambda), \qquad n \in \mathbb{Z} \quad \text{(P-a.s.)}$$

by (7). This completes the proof of the theorem.

Corollary 1. *Let $\xi = (\xi_n)$ be a stationary sequence of real random variables ξ_n, $n \in \mathbb{Z}$. Then the stochastic measure $Z = Z(\Delta)$ involved in the spectral representation* (2) *has the property that*

$$Z(\Delta) = \overline{Z(-\Delta)} \tag{9}$$

for every $\Delta = \mathscr{B}([-\pi, \pi))$, where $-\Delta = \{\lambda: -\lambda \in \Delta\}$.

In fact, let $f(\lambda) = \sum \alpha_k e^{i\lambda k}$ and $\eta = \sum \alpha_k \xi_k$ (finite sums). Then $f \leftrightarrow \eta$ and therefore

$$\bar{\eta} = \sum \bar{\alpha}_k \xi_k \leftrightarrow \sum \bar{\alpha}_k e^{i\lambda k} = \overline{f(-\lambda)}. \tag{10}$$

Since $\mathscr{I}_\Delta(\lambda) \leftrightarrow Z(\Delta)$, it follows from (10) that either $\mathscr{I}_\Delta(-\lambda) \leftrightarrow \bar{Z}(\Delta)$ or $\mathscr{I}_{-\Delta}(\lambda) \leftrightarrow \bar{Z}(\Delta)$. On the other hand, $\mathscr{I}_{-\Delta}(\lambda) \leftrightarrow Z(-\Delta)$. Therefore $\bar{Z}(\Delta) = Z(-\Delta)$ (P-a.s.).

Corollary 2. *Again let $\xi = (\xi_n)$ be a stationary sequence of real random variables ξ_n and $Z(\Delta) = Z_1(\Delta) + iZ_2(\Delta)$. Then*

$$\mathsf{E}Z_1(\Delta_1)Z_2(\Delta_2) = 0 \tag{11}$$

for every Δ_1 and Δ_2; and if $\Delta_1 \cap \Delta_2 = \varnothing$ then

$$\mathsf{E}Z_1(\Delta_1)Z_1(\Delta_2) = 0, \qquad \mathsf{E}Z_2(\Delta_1)Z_2(\Delta_2) = 0. \tag{12}$$

In fact, since $Z(\Delta) = \bar{Z}(-\Delta)$, we have

$$Z_1(-\Delta) = Z_1(\Delta), \qquad Z_2(-\Delta) = -Z_2(\Delta). \tag{13}$$

Moreover, since $\mathsf{E}Z(\Delta_1)\bar{Z}(\Delta_2) = \mathsf{E}|Z(\Delta_1 \cap \Delta_2)|^2$, we have $\operatorname{Im} \mathsf{E}Z(\Delta_1)\bar{Z}(\Delta_2) = 0$, i.e.

$$\mathsf{E}Z_1(\Delta_1)Z_2(\Delta_2) + \mathsf{E}Z_2(\Delta_1)Z_1(\Delta_2) = 0. \tag{14}$$

If we take the interval $-\Delta_1$ instead of Δ_1 we therefore obtain

$$\mathsf{E}Z_1(-\Delta_1)Z_2(\Delta_2) + \mathsf{E}Z_2(-\Delta_1)Z_1(\Delta_2) = 0,$$

which, by (13), can be transformed into

$$\mathsf{E}Z_1(\Delta_1)Z_2(\Delta_2) - \mathsf{E}Z_2(\Delta_1)Z_1(\Delta_2) = 0. \tag{15}$$

Then (11) follows from (14) and (15).

On the other hand, if $\Delta_1 \cap \Delta_2 = \varnothing$ then $\mathsf{E}Z(\Delta_1)\bar{Z}(\Delta_2) = 0$, whence $\operatorname{Re} \mathsf{E}Z(\Delta_1)\bar{Z}(\Delta_2) = 0$ and $\operatorname{Re} \mathsf{E}Z(-\Delta_1)\bar{Z}(\Delta_2) = 0$, which, with (13), provides an evident proof of (12).

Corollary 3. *Let $\xi = (\xi_n)$ be a Gaussian sequence. Then, for every family $\Delta_1, \dots, \Delta_k$, the vector $(Z_1(\Delta_1), \dots, Z_1(\Delta_k), Z_2(\Delta_1), \dots, Z_2(\Delta_k))$ is normally distributed.*

In fact, the linear manifold $L_0^2(\xi)$ consists of (complex-valued) Gaussian random variables η, i.e. the vector $(\operatorname{Re}\eta, \operatorname{Im}\eta)$ has a Gaussian distribution. Then, according to Subsection 5, §13, Chapter II, the closure of $L_0^2(\xi)$ also consists of Gaussian variables. It follows from Corollary 2 that, when $\xi = (\xi_n)$ is a Gaussian sequence, the real and imaginary parts of Z_1 and Z_2 are independent in the sense that the families of random variables $(Z_1(\Delta_1), \ldots, Z_1(\Delta_k))$ and $(Z_2(\Delta_1), \ldots, Z_2(\Delta_k))$ are independent. It also follows from (12) that when the sets $\Delta_1, \ldots, \Delta_k$ are disjoint, the random variables $Z_i(\Delta_1), \ldots, Z_i(\Delta_k)$ are collectively independent, $i = 1, 2$.

Corollary 4. *If $\xi = (\xi_n)$ is a stationary sequence of real random variables, then* (P-*a.s.*)

$$\xi_n = \int_{-\pi}^{\pi} \cos \lambda n \, Z_1(d\lambda) + \int_{-\pi}^{\pi} \sin \lambda n \, Z_2(d\lambda). \tag{16}$$

Remark. If $\{Z_\lambda\}$, $\lambda \in [-\pi, \pi)$, is a process with orthogonal increments, corresponding to an orthogonal stochastic measure $Z = Z(\Delta)$, then in accordance with §2 the spectral representation (2) can also be written in the following form:

$$\xi_n = \int_{-\pi}^{\pi} e^{i\lambda n} \, dZ_\lambda, \qquad n \in \mathbb{Z}. \tag{17}$$

2. Let $\xi = (\xi_n)$ be a stationary sequence with the spectral representation (2) and let $\eta \in L^2(\xi)$. The following theorem describes the structure of such random variables.

Theorem 2. *If $\eta \in L^2(\xi)$, there is a function $\varphi \in L^2(F)$ such that* (P-*a.s.*)

$$\eta = \int_{-\pi}^{\pi} \varphi(\lambda) Z(d\lambda). \tag{18}$$

PROOF. If

$$\eta_n = \sum_{|k| \le n} \alpha_k \xi_k, \tag{19}$$

then by (2)

$$\eta_n = \int_{-\pi}^{\pi} \left(\sum_{|k| \le n} \alpha_k e^{i\lambda k} \right) Z(d\lambda), \tag{20}$$

i.e. (18) is satisfied with

$$\varphi_n(\lambda) = \sum_{|k| \le n} \alpha_n e^{i\lambda k}. \tag{21}$$

In the general case, when $\eta \in L^2(\xi)$, there are variables η_n of type (19) such that $\|\eta - \eta_n\| \to 0$, $n \to \infty$. But then $\|\varphi_n - \varphi_m\| = \|\eta_n - \eta_m\| \to 0$, $n, m \to \infty$.

Consequently $\{\varphi_n\}$ is fundamental in $L^2(F)$ and therefore there is a function $\varphi \in L^2(F)$ such that $\|\varphi - \varphi_n\| \to 0$, $n \to \infty$.

By property (2.14) we have $\|\mathscr{I}(\varphi_n) - \mathscr{I}(\varphi)\| \to 0$, and since $\eta_n = \mathscr{I}(\varphi_n)$ we also have $\eta = \mathscr{I}(\varphi)$ (P-a.s.).

This completes the proof of the theorem.

Remark. Let $H_0(\xi)$ and $H_0(F)$ be the respective closed linear manifolds spanned by the variables ξ_n and by the functions e_n when $n \le 0$. Then if $\eta \in H_0(\xi)$ there is a function $\varphi \in H_0(F)$ such that (P-a.s.) $\eta = \int_{-\pi}^{\pi} \varphi(\lambda) Z(d\lambda)$.

3. Formula (18) describes the structure of the random variables that are obtained from ξ_n, $n \in \mathbb{Z}$, by linear transformations, i.e. in the form of finite sums (19) and their mean-square limits.

A special but important class of such linear transformations are defined by means of what are known as (linear) *filters*. Let us suppose that, at instant m, a system (filter) receives as input a signal x_m, and that the output of the system is, at instant n, the signal $h(n - m)x_m$, where $h = h(s)$, $s \in \mathbb{Z}$, is a complex valued function called the *impulse response* (of the filter).

Therefore the total signal obtained from the input can be represented in the form

$$y_n = \sum_{m=-\infty}^{\infty} h(n - m)x_m. \tag{22}$$

For physically realizable systems, the values of the input at instant n are determined only by the "past" values of the signal, i.e. the values x_m for $m \le n$. It is therefore natural to call a filter with the impulse response $h(s)$ *physically realizable* if $h(s) = 0$ for all $s < 0$, in other words if

$$y_n = \sum_{m=-\infty}^{n} h(n - m)x_m = \sum_{m=0}^{\infty} h(m)x_{n-m}. \tag{23}$$

An important *spectral characteristic* of a filter with the impulse response h is its Fourier transform

$$\varphi(\lambda) = \sum_{m=-\infty}^{\infty} e^{-i\lambda m} h(m), \tag{24}$$

known as the *frequency characteristic* or *transfer function* of the filter.

Let us now take up conditions, about which nothing has been said so far, for the convergence of the series in (22) and (24). Let us suppose that the input is a stationary random sequence $\xi = (\xi_n)$, $n \in \mathbb{Z}$, with covariance function $R(n)$ and spectral decomposition (2). Then if

$$\sum_{k,l=-\infty}^{\infty} h(k)R(k - l)\bar{h}(l) < \infty, \tag{25}$$

the series $\sum_{m=-\infty}^{\infty} h(n-m)\xi_m$ converges in mean-square and therefore there is a stationary sequence $\eta = (\eta_n)$ with

$$\eta_n = \sum_{m=-\infty}^{\infty} h(n-m)\xi_m = \sum_{m=-\infty}^{\infty} h(m)\xi_{n-m}. \tag{26}$$

In terms of the spectral measure, (25) is evidently equivalent to saying that $\varphi(\lambda) \in L^2(F)$, i.e.

$$\int_{-\pi}^{\pi} |\varphi(\lambda)|^2 F(d\lambda) < \infty. \tag{27}$$

Under (25) or (27), we obtain the spectral representation

$$\eta_n = \int_{-\pi}^{\pi} e^{i\lambda n}\varphi(\lambda)Z(d\lambda). \tag{28}$$

of η from (26) and (2). Consequently the covariance function $R_\eta(n)$ of η is given by the formula

$$R_\eta(n) = \int_{-\pi}^{\pi} e^{i\lambda n}|\varphi(\lambda)|^2 F(d\lambda). \tag{29}$$

In particular, if the input to a filter with frequency characteristic $\varphi = \varphi(\lambda)$ is taken to be white noise $\varepsilon = (\varepsilon_n)$, the output will be a stationary sequence (moving average)

$$\eta_n = \sum_{m=-\infty}^{\infty} h(m)\varepsilon_{n-m} \tag{30}$$

with spectral density

$$f_\eta(\lambda) = \frac{1}{2\pi}|\varphi(\lambda)|^2.$$

The following theorem shows that there is a sense in which every stationary sequence with a spectral density is obtainable by means of a moving average.

Theorem 3. *Let $\eta = (\eta_n)$ be a stationary sequence with spectral density $f_\eta(\lambda)$. Then (possibly at the expense of enlarging the original probability space) we can find a sequence $\varepsilon = (\varepsilon_n)$ representing white noise, and a filter, such that the representation* (30) *holds.*

PROOF. For a given (nonnegative) function $f_\eta(\lambda)$ we can find a function $\varphi(\lambda)$ such that $f_\eta(\lambda) = (1/2\pi)|\varphi(\lambda)|^2$. Since $\int_{-\pi}^{\pi} f_\eta(\lambda)\, d\lambda < \infty$, we have $\varphi(\lambda) \in L^2(\mu)$, where μ is Lebesgue measure on $[-\pi, \pi)$. Hence φ can be represented as a Fourier series (24) with $h(m) = (1/2\pi)\int_{-\pi}^{\pi} e^{im\lambda}\varphi(\lambda)\, d\lambda$, where convergence is understood in the sense that

$$\int_{-\pi}^{\pi} \left|\varphi(\lambda) - \sum_{|m|\le n} e^{-i\lambda m}h(m)\right|^2 d\lambda \to 0, \qquad n \to \infty.$$

Let

$$\eta_n = \int_{-\pi}^{\pi} e^{i\lambda n} Z(d\lambda), \qquad n \in \mathbb{Z}.$$

Besides the measure $Z = Z(\Delta)$ we introduce another independent orthogonal stochastic measure $\tilde{Z} = \tilde{Z}(\Delta)$ with $\mathsf{E}|\tilde{Z}(a, b]|^2 = (b - a)/2\pi$. (The possibility of constructing such a measure depends, in general, on having a sufficiently "rich" original probability space.) Let us put

$$\overline{Z}(\Delta) = \int_\Delta \varphi^\oplus(\lambda) Z(d\lambda) + \int_\Delta [1 - \varphi^\oplus(\lambda)\varphi(\lambda)]\tilde{Z}(d\lambda),$$

where

$$a^\oplus = \begin{cases} a^{-1}, & \text{if } a \neq 0, \\ 0, & \text{if } a = 0. \end{cases}$$

The stochastic measure $\overline{Z} = \overline{Z}(\Delta)$ is a measure with orthogonal values, and for every $\Delta = (a, b]$ we have

$$\mathsf{E}|\overline{Z}(\Delta)|^2 = \frac{1}{2\pi}\int_\Delta |\varphi^\oplus(\lambda)|^2 |\varphi(\lambda)|^2\, d\lambda + \frac{1}{2\pi}\int_\Delta |1 - \varphi^\oplus(\lambda)\varphi(\lambda)|^2\, d\lambda = \frac{|\Delta|}{2\pi},$$

where $|\Delta| = b - a$. Therefore the stationary sequence $\varepsilon = (\varepsilon_n)$, $n \in \mathbb{Z}$, with

$$\varepsilon_n = \int_{-\pi}^{\pi} e^{i\lambda n} \overline{Z}(d\lambda),$$

is a white noise.

We now observe that

$$\int_{-\pi}^{\pi} e^{i\lambda n}\varphi(\lambda)\overline{Z}(d\lambda) = \int_{-\pi}^{\pi} e^{in\lambda} Z(d\lambda) = \eta_n \tag{31}$$

and, on the other hand, by property (2.14) (P-a.s.)

$$\int_{-\pi}^{\pi} e^{in\lambda}\varphi(\lambda)\overline{Z}(d\lambda) = \int_{-\pi}^{\pi} e^{i\lambda n}\left(\sum_{m=-\infty}^{\infty} e^{-i\lambda m} h(m)\right)\overline{Z}(d\lambda)$$

$$= \sum_{m=-\infty}^{\infty} h(m) \int_{-\pi}^{\pi} e^{i\lambda(n-m)}\overline{Z}(d\lambda) = \sum_{m=-\infty}^{\infty} h(m)\varepsilon_{n-m},$$

which, together with (31), establishes the representation (30).

This completes the proof of the theorem.

Remark. If $f_\eta(\lambda) > 0$ (almost everywhere with respect to Lebesgue measure), the introduction of the auxiliary measure $\tilde{Z} = \tilde{Z}(\Delta)$ becomes unnecessary (since then $1 - \varphi^\oplus(\lambda)\varphi(\lambda) = 0$ almost everywhere with respect to Lebesgue measure), and the reservation concerning the necessity of extending the original probability space can be omitted.

Corollary 1. *Let the spectral density $f_\eta(\lambda) > 0$ (almost everywhere with respect to Lebesgue measure) and*

$$f_\eta(\lambda) = \frac{1}{2\pi}|\varphi(\lambda)|^2,$$

where

$$\varphi(\lambda) = \sum_{k=0}^{\infty} e^{-i\lambda k}h(k), \qquad \sum_{k=0}^{\infty}|h(k)|^2 < \infty.$$

Then the sequence η admits a representation as a one-sided moving average,

$$\eta_n = \sum_{m=0}^{\infty} h(m)\varepsilon_{n-m}.$$

In particular, let $P(z) = a_0 + a_1 z + \cdots + a_p z^p$ be a polynomial that has no zeros on $\{z: |z| = 1\}$. Then the sequence $\eta = (\eta_n)$ with spectral density

$$f_\eta(\lambda) = \frac{1}{2\pi}|P(e^{-i\lambda})|^2$$

can be represented in the form

$$\eta_n = a_0\varepsilon_n + a_1\varepsilon_{n-1} + \cdots + a_p\varepsilon_{n-p}.$$

Corollary 2. *Let $\xi = (\xi_n)$ be a sequence with rational spectral density*

$$f_\xi(\lambda) = \frac{1}{2\pi}\left|\frac{P(e^{-i\lambda})}{Q(e^{-i\lambda})}\right|^2, \tag{32}$$

where $P(z) = a_0 + a_1 z + \cdots + a_p z^p$, $Q(z) = 1 + b_1 z + \cdots + b_q z^q$.

Let us show that if $P(z)$ and $Q(z)$ have no zeros on $\{z: |z| = 1\}$, there is a white noise $\varepsilon = \varepsilon(n)$ such that (P-a.s.)

$$\xi_n + b_1\xi_{n-1} + \cdots + b_q\xi_{n-q} = a_0\varepsilon_n + a_1\varepsilon_{n-1} + \cdots + a_p\varepsilon_{n-p}. \tag{33}$$

Conversely, every stationary sequence $\xi = (\xi_n)$ that satisfies this equation with some white noise $\varepsilon = (\varepsilon_n)$ and some polynomial $Q(z)$ with no zeros on $\{z: |z| = 1\}$ has a spectral density (32).

In fact, let $\eta_n = \xi_n + b_1\xi_{n-1} + \cdots + b_q\xi_{n-q}$. Then $f_\eta(\lambda) = (1/2\pi)|P(e^{-i\lambda})|^2$ and the required representation follows from Corollary 1.

On the other hand, if (33) holds and $F_\xi(\lambda)$ and $F_\eta(\lambda)$ are the spectral functions of ξ and η, then

$$F_\eta(\lambda) = \int_{-\pi}^{\lambda}|Q(e^{-i\nu})|^2\,dF_\xi(\nu) = \frac{1}{2\pi}\int_{-\pi}^{\lambda}|P(e^{-i\nu})|^2\,d\nu.$$

Since $|Q(e^{-i\nu})|^2 > 0$, it follows that $F_\xi(\lambda)$ has a density defined by (32).

4. The following mean-square ergodic theorem can be thought of as an analog of the law of large numbers for stationary (wide sense) random sequences.

Theorem 4. *Let* $\xi = (\xi_n)$, $n \in \mathbb{Z}$, *be a stationary sequence with* $\mathsf{E}\xi_n = 0$, *covariance function* (1), *and spectral resolution* (2). *Then*

$$\frac{1}{n}\sum_{k=0}^{n-1} \xi_k \xrightarrow{L^2} Z(\{0\}) \tag{34}$$

and

$$\frac{1}{n}\sum_{k=0}^{n-1} R(k) \to F(\{0\}). \tag{35}$$

PROOF. By (2),

$$\frac{1}{n}\sum_{k=0}^{n-1} \xi_k = \int_{-\pi}^{\pi} \frac{1}{n}\sum_{k=0}^{n-1} e^{ik\lambda} Z(d\lambda) = \int_{-\pi}^{\pi} \varphi_n(\lambda) Z(d\lambda),$$

where

$$\varphi_n(\lambda) = \frac{1}{n}\sum_{k=0}^{n-1} e^{ik\lambda}. \tag{36}$$

It is clear that

$$|\varphi_n(\lambda)| \le 1.$$

Moreover, $\varphi_n(\lambda) \xrightarrow{L^2(F)} I_{\{0\}}(\lambda)$ and therefore by (2.14)

$$\int_{-\pi}^{\pi} \varphi_n(\lambda) Z(d\lambda) \xrightarrow{L^2} \int_{-\pi}^{\pi} I_{\{0\}}(\lambda) Z(d\lambda) = Z(\{0\}),$$

which establishes (34).

Relation (35) can be proved in a similar way.

This completes the proof of the theorem.

Corollary. *If the spectral function is continuous at zero, i.e.* $F(\{0\}) = 0$, *then* $Z(\{0\}) = 0$ (P-*a.s.*) *and by* (34) *and* (35),

$$\frac{1}{n}\sum_{k=0}^{n-1} R(k) \to 0 \Rightarrow \frac{1}{n}\sum_{k=0}^{n-1} \xi_k \xrightarrow{L^2} 0.$$

Since

$$\left|\frac{1}{n}\sum_{k=0}^{n-1} R(k)\right|^2 = \left|\mathsf{E}\left(\frac{1}{n}\sum_{k=0}^{n-1} \xi_k\right)\xi_0\right|^2 \le \mathsf{E}|\xi_0|^2 \mathsf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1} \xi_k\right|^2,$$

the converse implication also holds:

$$\frac{1}{n}\sum_{k=0}^{n-1}\xi_k \xrightarrow{L^2} 0 \Rightarrow \frac{1}{n}\sum_{k=0}^{n-1}R(k) \to 0.$$

Therefore the condition $(1/n)\sum_{k=0}^{n-1}R(k) \to 0$ is necessary and sufficient for the convergence (in the mean-square sense) of the arithmetic means $(1/n)\sum_{k=0}^{n-1}\xi_k$ to zero. It follows that if the original sequences $\xi = (\xi_n)$ has expectation m (that is, $\mathsf{E}\xi_0 = m$), then

$$\frac{1}{n}\sum_{k=0}^{n-1}R(k) \to 0 \Leftrightarrow \frac{1}{n}\sum_{k=0}^{n-1}\xi_k \xrightarrow{L^2} m, \tag{37}$$

where $R(n) = \mathsf{E}(\xi_n - \mathsf{E}\xi_n)(\overline{\xi_0 - \mathsf{E}\xi_0})$.

Let us also observe that if $Z(\{0\}) \neq 0$ (P-a.s.) and $m = 0$, then ξ_n "contains a random constant α":

$$\xi_n = \alpha + \eta_n,$$

where $\alpha = Z(\{0\})$; and in the spectral representation $\eta_n = \int_{-\pi}^{\pi} e^{i\lambda n} Z_\eta(d\lambda)$ the measure $Z_\eta = Z_\eta(\Delta)$ is such that $Z_\eta(\{0\}) = 0$ (P-a.s.). Conclusion (34) means that the arithmetic mean converges in mean-square to precisely this random constant α.

5. Problems

1. Show that $\overline{L_0^2}(F) = L^2(F)$ (for the notation see the proof of Theorem 1).

2. Let $\xi = (\xi_n)$ be a stationary sequence with the property that $\xi_{n+N} = \xi_n$ for some N and all n. Show that the spectral representation of such a sequence reduces to (1.13).

3. Let $\xi = (\xi_n)$ be a stationary sequence such that $\mathsf{E}\xi_n = 0$ and

$$\frac{1}{N^2}\sum_{k=0}^{N}\sum_{l=0}^{N}R(k-l) = \frac{1}{N}\sum_{|k|\le N-1}R(k)\left[1 - \frac{|k|}{N}\right] \le CN^{-\alpha}$$

for some $C > 0$, $\alpha > 0$. Use the Borel–Cantelli lemma to show that then

$$\frac{1}{N}\sum_{k=0}^{N}\xi_k \to 0 \quad (\text{P-a.s.})$$

4. Let the spectral density $f_\xi(\lambda)$ of the sequence $\xi = (\xi_n)$ be rational,

$$f_\xi(\lambda) = \frac{1}{2\pi}\frac{|P_{n-1}(e^{-i\lambda})|}{|Q_n(e^{-i\lambda})|}, \tag{38}$$

where $P_{n-1}(z) = a_0 + a_1 z + \cdots + a_{n-1}z^{n-1}$ and $Q_n(z) = 1 + b_1 z + \cdots + b_n z^n$, and all the zeros of these polynomials lie outside the unit disk.

Show that there is a white noise $\varepsilon = (\varepsilon_m)$, $m \in \mathbb{Z}$, such that the sequence (ξ_m) is a component of an n-dimensional sequence $(\xi_m^1, \xi_m^2, \ldots, \xi_m^n)$, $\xi_m^1 = \xi_m$, that satisfies the system of equations

$$\xi^i_{m+1} = \xi^{i+1}_m + \beta_i \varepsilon_{m+1}, \qquad i = 1, \ldots, n-1,$$
$$\xi^n_{m+1} = -\sum_{j=0}^{n-1} b_{n-j}\xi^{j+1}_m + \beta_n \varepsilon_{m+1}, \tag{39}$$

where $\beta_1 = a_0$, $\beta_i = a_{i-1} - \sum_{k=1}^{i-1} \beta_k b_{i-k}$.

§4. Statistical Estimation of the Covariance Function and the Spectral Density

1. Problems of the statistical estimation of various characteristics of the probability distributions of random sequences arise in the most diverse branches of science (geophysics, medicine, economics, etc.) The material presented in this section will give the reader an idea of the concepts and methods of estimation, and of the difficulties that are encountered.

To begin with, let $\xi = (\xi_n)$, $n \in \mathbb{Z}$, be a sequence, stationary in the wide sense (for simplicity, real) with expectation $\mathsf{E}\xi_n = m$ and covariance $R(n) = \int_{-\pi}^{\pi} e^{i\lambda n} F(d\lambda)$.

Let $x_0, x_1, \ldots, x_{N-1}$ be the results of observing the random variables $\xi_0, \xi_1, \ldots, \xi_{N-1}$. How are we then to construct a "good" estimator of the (unknown) mean value m?

Let us put

$$m_N(x) = \frac{1}{N} \sum_{k=0}^{N-1} x_k. \tag{1}$$

Then it follows from the elementary properties of the expectation that this is a "good" estimator of m in the sense that it is *unbiased* "in the mean over all kinds of data $x_0, \ldots, x_{N-1}$", i.e.

$$\mathsf{E} m_N(\xi) = \mathsf{E}\left(\frac{1}{N} \sum_{k=0}^{N-1} \xi_k\right) = m. \tag{2}$$

In addition, it follows from Theorem 4 of §3 that when $(1/N) \sum_{k=0}^{N} R(k) \to 0$, $N \to \infty$, our estimator is *consistent* (in mean-square), i.e.

$$\mathsf{E}\,|m_N(\xi) - m|^2 \to 0, \qquad N \to \infty. \tag{3}$$

Next we take up the problem of estimating the covariance function $R(n)$, the spectral function $F(\lambda) = F([-\pi, \lambda])$, and the spectral density $f(\lambda)$, all under the assumption that $m = 0$.

Since $R(n) = \mathsf{E}\xi_{n+k}\xi_k$, it is natural to estimate this function on the basis of N observations $x_0, x_1, \ldots, x_{N-1}$ (when $0 \le n < N$) by

$$\hat{R}_N(n; x) = \frac{1}{N-n} \sum_{k=0}^{N-n-1} x_{n+k} x_k.$$

It is clear that this estimator is unbiased in the sense that

$$\mathsf{E}\hat{R}_N(n; \xi) = R(n), \qquad 0 \le n < N.$$

Let us now consider the question of its consistency. If we replace ξ_k in (3.37) by $\xi_{n+k}\xi_k$ and suppose that the sequence $\xi = (\xi_n)$ under consideration has a fourth moment ($\mathsf{E}\xi_0^4 < \infty$), we find that the condition

$$\frac{1}{N} \sum_{k=0}^{N-1} \mathsf{E}[\xi_{n+k}\xi_k - R(n)][\xi_n\xi_0 - R(n)] \to 0, \qquad N \to \infty, \tag{4}$$

is necessary and sufficient for

$$\mathsf{E}|\hat{R}_N(n; \xi) - R(n)|^2 \to 0, \qquad N \to \infty. \tag{5}$$

Let us suppose that the original sequence $\xi = (\xi_n)$ is Gaussian (with zero mean and covariance $R(n)$). Then by (II.12.51)

$$\begin{aligned}\mathsf{E}[\xi_{n+k}\xi_k - R(n)][\xi_n\xi_0 - R(n)] &= \mathsf{E}\xi_{n+k}\xi_k\xi_n\xi_0 - R^2(n)\\ &= \mathsf{E}\xi_{n+k}\xi_k \cdot \mathsf{E}\xi_n\xi_0 + \mathsf{E}\xi_{n+k}\xi_n \cdot \mathsf{E}\xi_k\xi_0\\ &\quad + \mathsf{E}\xi_{n+k}\xi_0 \cdot \mathsf{E}\xi_k\xi_n - R^2(n)\\ &= R^2(k) + R(n+k)R(n-k).\end{aligned}$$

Therefore in the Gaussian case condition (4) is equivalent to

$$\frac{1}{N} \sum_{k=0}^{N-1} [R^2(k) + R(n+k)R(n-k)] \to 0, \qquad N \to \infty. \tag{6}$$

Since $|R(n+k)R(n-k)| \le |R(n+k)|^2 + |R(n-k)|^2$, the condition

$$\frac{1}{N} \sum_{k=0}^{N-1} R^2(k) \to 0, \qquad N \to \infty, \tag{7}$$

implies (6). Conversely, if (6) holds for $n = 0$, then (7) is satisfied.

We have now established the following theorem.

Theorem. *Let $\xi = (\xi_n)$ be a Gaussian stationary sequence with $\mathsf{E}\xi_n = 0$ and covariance function $R(n)$. Then (7) is a necessary and sufficient condition that, for every $n \ge 0$, the estimator $\hat{R}_N(n; x)$ is mean-square consistent,* (i.e. *that* (5) *is satisfied*).

Remark. If we use the spectral representation of the covariance function, we obtain

$$\frac{1}{N}\sum_{k=0}^{N-1} R^2(k) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{1}{N}\sum_{k=0}^{N-1} e^{i(\lambda-\nu)k} F(d\lambda)F(d\nu)$$

$$= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f_N(\lambda, \nu)F(d\lambda)F(d\nu),$$

where (compare (3.35))

$$f_N(\lambda, \nu) = \begin{cases} 1, & \lambda = \nu, \\ \dfrac{1 - e^{i(\lambda-\nu)N}}{N[1 - e^{i(\lambda-\nu)}]}, & \lambda \neq \nu. \end{cases}$$

But as $N \to \infty$

$$f_N(\lambda, \nu) \to f(\lambda, \nu) = \begin{cases} 1, & \lambda = \nu, \\ 0, & \lambda \neq \nu. \end{cases}$$

Therefore

$$\frac{1}{N}\sum_{k=0}^{N-1} R^2(k) \to \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(\lambda, \nu)F(d\lambda)F(d\nu)$$

$$= \int_{-\pi}^{\pi} F(\{\lambda\})F(d\lambda) = \sum_{\lambda} F^2(\{\lambda\}),$$

where the sum over λ contains at most a countable number of terms since the measure F is finite.

Hence (7) is equivalent to

$$\sum_{\lambda} F^2(\{\lambda\}) = 0, \tag{8}$$

which means that the spectral function $F(\lambda) = F([-\pi, \lambda])$ is *continuous.*

2. We now turn to the problem of finding estimators for the spectral function $F(\lambda)$ and the spectral density $f(\lambda)$ (under the assumption that they exist).

A method that naturally suggests itself for estimating the spectral density follows from the proof of Herglotz' theorem that we gave earlier. Recall that the function

$$f_N(\lambda) = \frac{1}{2\pi}\sum_{|n|<N}\left(1 - \frac{|n|}{N}\right)R(n)e^{-i\lambda n} \tag{9}$$

introduced in §1 has the property that the function

$$F_N(\lambda) = \int_{-\pi}^{\lambda} f_N(\nu)\,d\nu$$

converges on the whole (Chapter III, §1) to the spectral function $F(\lambda)$. Therefore if $F(\lambda)$ has a density $f(\lambda)$, we have

$$\int_{-\pi}^{\lambda} f_N(v)\,dv \to \int_{-\pi}^{\lambda} f(v)\,dv \tag{10}$$

for each $\lambda \in [-\pi, \pi)$.

Starting from these facts and recalling that an estimator for $R(n)$ (on the basis of the observations $x_0, x_1, \ldots, x_{N-1}$) is $\hat{R}_N(n; x)$, we take as an estimator for $f(\lambda)$ the function

$$\hat{f}_N(\lambda; x) = \frac{1}{2\pi} \sum_{|n|<N} \left(1 - \frac{|n|}{N}\right) \hat{R}_N(n; x) e^{-i\lambda n}, \tag{11}$$

putting $\hat{R}_N(n; x) = \hat{R}_N(|n|; x)$ for $|n| < N$.

The function $\hat{f}_N(\lambda; x)$ is known as a *periodogram*. It is easily verified that it can also be represented in the following more convenient form:

$$\hat{f}_N(\lambda; x) = \frac{1}{2\pi N} \left| \sum_{n=0}^{N-1} x_n e^{-i\lambda n} \right|^2. \tag{12}$$

Since $\mathsf{E}\hat{R}_N(n; \xi) = R(n)$, $|n| < N$, we have

$$\mathsf{E}\hat{f}_N(\lambda; \xi) = f_N(\lambda).$$

If the spectral function $F(\lambda)$ has density $f(\lambda)$, then, since $f_N(\lambda)$ can also be written in the form (1.34), we find that

$$\begin{aligned} f_N(\lambda) &= \frac{1}{2\pi N} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} \int_{-\pi}^{\pi} e^{iv(k-l)} e^{i\lambda(l-k)} f(v)\,dv \\ &= \int_{-\pi}^{\pi} \frac{1}{2\pi N} \left| \sum_{k=0}^{N-1} e^{i(v-\lambda)k} \right|^2 f(v)\,dv. \end{aligned}$$

The function

$$\Phi_N(\lambda) = \frac{1}{2\pi N} \left| \sum_{k=0}^{N-1} e^{i\lambda k} \right|^2 = \frac{1}{2\pi N} \left| \frac{\sin \frac{\lambda}{2} N}{\sin \lambda/2} \right|^2$$

is the Fejér kernel. It is known, from the properties of this function, that for almost every λ (with respect to Lebesgue measure)

$$\int_{-\pi}^{\pi} \Phi_N(\lambda - v) f(v)\,dv \to f(\lambda). \tag{13}$$

Therefore for almost every $\lambda \in [-\pi, \pi)$

$$\mathsf{E}\hat{f}_N(\lambda; \xi) \to f(\lambda); \tag{14}$$

in other words, the estimator $\hat{f}_N(\lambda; x)$ of $f(\lambda)$ on the basis of $x_0, x_1, \ldots, x_{N-1}$ is *asymptotically unbiased*.

In this sense the estimator $\hat{f}_N(\lambda; x)$ can be considered to be "good." However, at the individual observed values $x_0, \ldots, x_{N-1}$ the values of the periodogram $\hat{f}_N(\lambda; x)$ usually turn out to be far from the actual values $f(\lambda)$. In fact, let $\xi = (\xi_n)$ be a stationary sequence of independent Gaussian random variables, $\xi_n \sim \mathcal{N}(0, 1)$. Then $f(\lambda) \equiv 1/2\pi$ and

$$\hat{f}_N(\lambda; \xi) = \frac{1}{2\pi}\left|\frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}\xi_k e^{-i\lambda k}\right|^2.$$

Then at the point $\lambda = 0$ we have $\hat{f}_N(0, \xi)$ coinciding in distribution with the square of the Gaussian random variable $\eta \sim \mathcal{N}(0, 1)$. Hence, for every N,

$$\mathsf{E}\,|\hat{f}_N(0; \xi) - f(0)|^2 = \frac{1}{4\pi^2}\,\mathsf{E}\,|\eta^2 - 1|^2 > 0.$$

Moreover, an easy calculation shows that if $f(\lambda)$ is the spectral density of a stationary sequence $\xi = (\xi_n)$ that is constructed as a moving average:

$$\xi_n = \sum_{k=0}^{\infty} a_k \varepsilon_{n-k} \tag{15}$$

with $\sum_{k=0}^{\infty} |a_k| < \infty$, $\sum_{k=0}^{\infty} |a_k|^2 < \infty$, where $\varepsilon = (\varepsilon_n)$ is white noise with $\mathsf{E}\varepsilon_0^4 < \infty$, then

$$\lim_{N\to\infty} \mathsf{E}\,|\hat{f}_N(\lambda; \xi) - f(\lambda)|^2 = \begin{cases} 2f^2(0), & \lambda = 0, \pm\pi, \\ f^2(\lambda), & \lambda \neq 0, \pm\pi. \end{cases} \tag{16}$$

Hence it is clear that the periodogram cannot be a satisfactory estimator of the spectral density. To improve the situation, one often uses an estimator for $f(\lambda)$ of the form

$$f_N^W(\lambda; x) = \int_{-\pi}^{\pi} W_N(\lambda - \nu)\hat{f}_N(\nu; x)\, d\nu, \tag{17}$$

which is obtained from the periodogram $\hat{f}_N(\lambda; x)$ and a smoothing function $W_N(\lambda)$, and which we call a *spectral window*. Natural requirements on $W_N(\lambda)$ are:

(a) $W_N(\lambda)$ has a sharp maximum at $\lambda = 0$;
(b) $\int_{-\pi}^{\pi} W_N(\lambda)\, d\lambda = 1$;
(c) $\mathsf{P}\,|\hat{f}_N^W(\lambda; \xi) - f(\lambda)|^2 \to 0$, $\quad N \to \infty$, $\quad \lambda \in [-\pi, \pi)$.

By (14) and (b) the estimators $\hat{f}_N^W(\lambda; \xi)$ are asymptotically unbiased. Condition (c) is the condition of asymptotic consistency in mean-square, which, as we showed above, is violated for the periodogram. Finally, condition (a) ensures that the required frequency λ is "picked out" from the periodogram.

Let us give some examples of estimators of the form (17).

Bartlett's estimator is based on the spectral window

$$W_N(\lambda) = a_N B(a_N \lambda),$$

where $a_N \uparrow \infty$, $a_N/N \to 0$, $N \to \infty$, and

$$B(\lambda) = \frac{1}{2\pi}\left|\frac{\sin(\lambda/2)}{\lambda/2}\right|^2.$$

Parzen's estimator takes the spectral window to be

$$W_N(\lambda) = a_N P(a_N \lambda),$$

where a_N are the same as before and

$$P(\lambda) = \frac{3}{8\pi}\left|\frac{\sin(\lambda/4)}{\lambda/4}\right|^4.$$

Zhurbenko's estimator is constructed from a spectral window of the form

$$W_N(\lambda) = a_N Z(a_N \lambda)$$

with

$$Z(\lambda) = \begin{cases} -\dfrac{\alpha+1}{2\alpha}|\lambda|^\alpha + \dfrac{\alpha+1}{2\alpha}, & |\lambda| \le 1, \\ 0, & |\lambda| > 1, \end{cases}$$

where $0 < \alpha \le 2$ and the a_N are selected in a particular way.

We shall not spend any more time on problems of estimating spectral densities; we merely note that there is an extensive statistical literature dealing with the construction of spectral windows and the comparison of the corresponding estimators $\hat{f}_N^W(\lambda; x)$.

3. We now consider the problem of estimating the spectral function $F(\lambda) = F([-\pi, \lambda])$. We begin by defining

$$F_N(\lambda) = \int_{-\pi}^{\lambda} f_N(v)\, dv, \qquad \hat{F}_N(\lambda; x) = \int_{-\pi}^{\lambda} \hat{f}_N(v; x)\, dv,$$

where $\hat{f}_N(v; x)$ is the periodogram constructed with $(x_0, x_1, \ldots, x_{N-1})$.

It follows from the proof of Herglotz' theorem (§1) that

$$\int_{-\pi}^{\pi} e^{i\lambda n}\, dF_N(\lambda) \to \int_{-\pi}^{\pi} e^{i\lambda n}\, dF(\lambda)$$

for every $n \in \mathbb{Z}$. Hence it follows (compare the corollary to Theorem 1, §3, Chapter III) that $F_N \Rightarrow F$, i.e. $F_N(\lambda)$ converges to $F(\lambda)$ at each point of continuity of $F(\lambda)$.

Observe that

$$\int_{-\pi}^{\pi} e^{i\lambda n}\, d\hat{F}_N(\lambda; \xi) = \hat{R}_N(n; \xi)\left(1 - \frac{|n|}{N}\right)$$

for all $|n| < N$. Therefore if we suppose that $\hat{R}_N(n;\xi)$ converges to $R(n)$ with probability one as $N \to \infty$, we have

$$\int_{-\pi}^{\pi} e^{i\lambda n}\, d\hat{F}_N(\lambda;\xi) \to \int_{-\pi}^{\pi} e^{i\lambda n}\, dF(\lambda) \quad (\text{P-a.s.})$$

and therefore $\hat{F}_N(\lambda;\xi) \Rightarrow F(\lambda)$ (P-a.s.).

It is then easy to deduce (if necessary, passing from a sequence to a subsequence) that if $\hat{R}_N(n;\xi) \to R(n)$ in probability, then $\hat{F}_N(\lambda;\xi) \Rightarrow F(\lambda)$ in probability.

4. Problems

1. In (15) let $\varepsilon_n \sim \mathcal{N}(0, 1)$. Show that

$$(N - n)\mathsf{V}\hat{R}_N(n, \xi) \to 2\pi \int_{-\pi}^{\pi} (1 + e^{2in\lambda}) f^2(\lambda)\, d\lambda$$

for every n, as $N \to \infty$.

2. Establish (16) and the following generalization:

$$\lim_{N\to\infty} \operatorname{cov}(\hat{f}_N(\lambda;\xi), \hat{f}_N(\nu;\xi)) = \begin{cases} 2f^2(0), & \lambda = \nu = 0, \pm\pi, \\ f^2(\lambda), & \lambda = \nu \neq 0, \pm\pi, \\ 0, & \lambda \neq \pm\nu. \end{cases}$$

§5. Wold's Expansion

1. In contrast to the representation (3.2) which gives an expansion of a stationary sequence in the *frequency* domain, Wold's expansion operates in the *time* domain. The main point of this expansion is that a stationary sequence $\xi = (\xi_n)$, $n \in \mathbb{Z}$, can be represented as the sum of two stationary sequences, one of which is completely predictable (in the sense that its values are completely determined by its "past"), whereas the second does not have this property.

We begin with some definitions. Let $H_n(\xi) = \overline{L}^2(\xi^n)$ and $H(\xi) = \overline{L}^2(\xi)$ be closed linear manifolds, spanned respectively by $\xi^n = (\ldots, \xi_{n-1}, \xi_n)$ and $\xi = (\cdots \xi_{n-1}, \xi_n, \ldots)$. Let

$$S(\xi) = \bigcap_n H_n(\xi).$$

For every $\eta \in H(\xi)$, denote by

$$\hat{\pi}_n(\eta) = \hat{E}(\eta \,|\, H_n(\xi))$$

the projection of η on the subspace $H_n(\xi)$ (see §11, Chapter II). We also write

$$\hat{\pi}_{-\infty}(\eta) = \hat{E}(\eta \,|\, S(\xi)).$$

Every element $\eta \in H(\xi)$ can be represented as

$$\eta = \hat{\pi}_{-\infty}(\eta) + (\eta - \hat{\pi}_{-\infty}(\eta)),$$

where $\eta - \hat{\pi}_{-\infty}(\eta) \perp \hat{\pi}_{-\infty}(\eta)$. Therefore $H(\xi)$ is represented as the orthogonal sum

$$H(\xi) = S(\xi) \oplus R(\xi),$$

where $S(\xi)$ consists of the elements $\hat{\pi}_{-\infty}(\eta)$ with $\eta \in H(\xi)$, and $R(\xi)$ consists of the elements of the form $\eta - \hat{\pi}_{-\infty}(\eta)$.

We shall now assume that $\mathsf{E}\xi_n = 0$ and $\mathsf{V}\xi_n > 0$. Then $H(\xi)$ is automatically nontrivial (contains elements different from zero).

Definition 1. A stationary sequence $\xi = (\xi_n)$ is *regular* if

$$H(\xi) = R(\xi)$$

and *singular* if

$$H(\xi) = S(\xi).$$

Remark. Singular sequences are also called *deterministic* and regular sequences are called *purely* or *completely nondeterministic*. If $S(\xi)$ is a proper subspace of $H(\xi)$ we just say that ξ is *nondeterministic*.

Theorem 1. *Every stationary (wide sense) random sequence ξ has a unique decomposition*

$$\xi_n = \xi_n^r + \xi_n^s, \tag{1}$$

where $\xi^r = (\xi_n^r)$ is regular and $\xi^s = (\xi_n^s)$ is singular. Here ξ^r and ξ^s are orthogonal ($\xi_n^r \perp \xi_m^s$ for all n and m).

PROOF. We define

$$\xi_n^s = \hat{E}(\xi_n/S(\xi)), \qquad \xi_n^r = \xi_n - \xi_n^s.$$

Since $\xi_n^r \perp S(\xi)$, for every n, we have $S(\xi^r) \perp S(\xi)$. On the other hand, $S(\xi^r) \subseteq S(\xi)$ and therefore $S(\xi^r)$ is trivial (contains only random sequences that coincide almost surely with zero). Consequently ξ^r is regular.

Moreover, $H_n(\xi) \subseteq H_n(\xi^s) \oplus H_n(\xi^r)$ and $H_n(\xi^s) \subseteq H_n(\xi)$, $H_n(\xi^r) \subseteq H_n(\xi)$. Therefore $H_n(\xi) = H_n(\xi^s) \oplus H_n(\xi^r)$ and hence

$$S(\xi) \subseteq H_n(\xi^s) \oplus H_n(\xi^r) \tag{2}$$

for every n. Since $\xi_n^r \perp S(\xi)$ it follows from (2) that

$$S(\xi) \subseteq H_n(\xi^s),$$

and therefore $S(\xi) \subseteq S(\xi^s) \subseteq H(\xi^s)$. But $\xi_n^s \subseteq S(\xi)$; hence $H(\xi^s) \subseteq S(\xi)$ and consequently

$$S(\xi) = S(\xi^s) = H(\xi^s),$$

which means that ξ^s is singular.

The orthogonality of ξ^s and ξ^r follows in an obvious way from $\xi_n^s \in S(\xi)$ and $\xi_n^r \perp S(\xi)$.

Let us now show that (1) is unique. Let $\xi_n = \eta_n^r + \eta_n^s$, where η^r and η^s are regular and singular orthogonal sequences. Then since $H_n(\eta^s) = H(\eta^s)$, we have

$$H_n(\xi) = H_n(\eta^r) \oplus H_n(\eta^s) = H_n(\eta^r) \oplus H(\eta^s),$$

and therefore $S(\xi) = S(\eta^r) \oplus H(\eta^s)$. But $S(\eta^r)$ is trivial, and therefore $S(\xi) = H(\eta^s)$.

Since $\eta_n^s \in H(\eta^s) = S(\xi)$ and $\eta_n^r \perp H(\eta^s) = S(\xi)$, we have $\hat{\mathsf{E}}(\xi_n | S(\xi)) = \hat{\mathsf{E}}(\eta_n^r + \eta_n^s | S(\xi)) = \eta_n^s$, i.e. η_n^s coincides with ξ_n^s; this establishes the uniqueness of (1).

This completes the proof of the theorem.

2. Definition 2. Let $\xi = (\xi_n)$ be a nondegenerate stationary sequence. A random sequence $\varepsilon = (\varepsilon_n)$ is an *innovation* sequence (for ξ) if

(a) $\varepsilon = (\varepsilon_n)$ consists of pairwise orthogonal random variables with $\mathsf{E}\varepsilon_n = 0$, $\mathsf{E}|\varepsilon_n|^2 = 1$;
(b) $H_n(\xi) = H_n(\varepsilon)$ for all $n \in \mathbb{Z}$.

Remark. The reason for the term "innovation" is that ε_{n+1} provides, so to speak, new "information" not contained in $H_n(\xi)$ (in other words, "innovates" in $H_n(\xi)$ the information that is needed for forming $H_{n+1}(\xi)$).

The following fundamental theorem establishes a connection between one-sided moving averages (Example 4, §1) and regular sequences.

Theorem 2. *A necessary and sufficient condition for a nondegenerate sequence ξ to be regular is that there are an innovation sequence $\varepsilon = (\varepsilon_n)$ and a sequence (a_n) of complex numbers, $n \geq 0$, with $\sum_{n=0}^{\infty} |a_n|^2 < \infty$, such that*

$$\xi_n = \sum_{k=0}^{\infty} a_k \varepsilon_{n-k} \quad (\mathsf{P}\text{-}a.s.) \tag{3}$$

Proof. *Necessity.* We represent $H_n(\xi)$ in the form

$$H_n(\xi) = H_{n-1}(\xi) \oplus B_n.$$

Since $H_n(\xi)$ is spanned by elements of $H_{n-1}(\xi)$ and elements of the form $\beta \cdot \xi_n$, where β is a complex number, the dimension (dim) of B_n is either zero or one. But the space $H_n(\xi)$ cannot coincide with $H_{n-1}(\xi)$ for any value of n.

In fact, if B_n is trivial for some n, then by stationarity B_k is trivial for all k, and therefore $H(\xi) = S(\xi)$, contradicting the assumption that ξ is regular. Thus B_n has the dimension $\dim B_n = 1$. Let η_n be a nonzero element of B_n. Put

$$\varepsilon_n = \frac{\eta_n}{\|\eta_n\|},$$

where $\|\eta_n\|^2 = \mathsf{E}|\eta_n|^2 > 0$.

For given n and $k \geq 0$, consider the decomposition

$$H_n(\xi) = H_{n-k}(\xi) \oplus B_{n-k+1} \oplus \cdots \oplus B_n.$$

Then $\varepsilon_{n-k}, \ldots, \varepsilon_n$ is an orthogonal basis in $B_{n-k+1} \oplus \cdots \oplus B_n$ and

$$\xi_n = \sum_{j=0}^{k-1} a_j \varepsilon_{n-j} + \hat{\pi}_{n-k}(\xi_n), \tag{4}$$

where $a_j = \mathsf{E}\xi_n \bar{\varepsilon}_{n-j}$.

By Bessel's inequality (II.11.16)

$$\sum_{j=0}^{\infty} |a_j|^2 \leq \|\xi_n\|^2 < \infty.$$

It follows that $\sum_{j=0}^{\infty} a_i \varepsilon_{n-j}$ converges in mean square, and then, by (4), equation (3) will be established as soon as we show that $\hat{\pi}_{n-k}(\xi_n) \xrightarrow{L^2} 0$, $k \to \infty$.

It is enough to consider the case $n = 0$. Since

$$\hat{\pi}_{-k} = \hat{\pi}_0 + \sum_{i=0}^{k} [\hat{\pi}_{-i} - \hat{\pi}_{-i+1}],$$

and the terms that appear in this sum are orthogonal, we have for every $k \geq 0$

$$\begin{aligned} \sum_{i=0}^{k} \|\hat{\pi}_{-i} - \hat{\pi}_{-i+1}\|^2 &= \left\| \sum_{i=0}^{k} (\hat{\pi}_{-i} - \hat{\pi}_{-i+1}) \right\|^2 \\ &= \|\hat{\pi}_{-k} - \hat{\pi}_0\|^2 \leq 4\|\xi_0\|^2 < \infty. \end{aligned}$$

Therefore the limit $\lim_{k\to\infty} \hat{\pi}_{-k}$ exists (in mean square). Now $\hat{\pi}_{-k} \in H_{-k}(\xi)$ for each k, and therefore the limit in question must belong to $\bigcap_{k\geq 0} H_k(\xi) = S(\xi)$. But, by assumption, $S(\xi)$ is trivial, and therefore $\hat{\pi}_{-k} \xrightarrow{L^2} 0$, $k \to \infty$.

Sufficiency. Let the nondegenerate sequence ξ have a representation (3), where $\varepsilon = (\varepsilon_n)$ is an orthonormal system (not necessarily satisfying the condition $H_n(\xi) = H_n(\varepsilon)$, $n \in \mathbb{Z}$). Then $H_n(\xi) \subseteq H_n(\varepsilon)$ and therefore $S(\xi) = \bigcap_k H_k(\xi) \subseteq H_n(\varepsilon)$ for every n. But $\varepsilon_{n+1} \perp H_n(\varepsilon)$, and therefore $\varepsilon_{n+1} \perp S(\xi)$ and at the same time $\varepsilon = (\varepsilon_n)$ is a basis in $H(\xi)$. It follows that $S(\xi)$ is trivial, and consequently ξ is regular.

This completes the proof of the theorem.

Remark. It follows from the proof that a nondegenerate sequence ξ is regular if and only if it admits a representation as a one-sided moving average

$$\xi_n = \sum_{k=0}^{\infty} \tilde{a}_k \tilde{\varepsilon}_{n-k}, \tag{5}$$

where $\tilde{\varepsilon} = \tilde{\varepsilon}_n$ is an orthonormal system which (it is important to emphasize this!) does not necessarily satisfy the condition $H_n(\xi) = H_n(\tilde{\varepsilon})$, $n \in \mathbb{Z}$. In this sense the conclusion of Theorem 2 says more, and specifically that for a regular sequence ξ there exist $a = (a_n)$ and an orthonormal system $\varepsilon = (\varepsilon_n)$ such that not only (5), but also (3), is satisfied, with $H_n(\xi) = H_n(\varepsilon)$, $n \in \mathbb{Z}$.

The following theorem is an immediate corollary of Theorems 1 and 2.

Theorem 3 (Wold's Expansion). *If $\xi = (\xi_n)$ is a nondegenerate stationary sequence, then*

$$\xi_n = \xi_n^s + \sum_{k=0}^{\infty} a_k \varepsilon_{n-k}, \tag{6}$$

where $\sum_{k=0}^{\infty} |a_k|^2 < \infty$ and $\varepsilon = (\varepsilon_n)$ is an innovation sequence (for ξ^r).

3. The significance of the concepts introduced here (regular and singular sequences) becomes particularly clear if we consider the following (linear) extrapolation problem, for whose solution the Wold expansion (6) is especially useful.

Let $H_0(\xi) = \bar{L}^2(\xi^0)$ be the closed linear manifold spanned by the variables $\xi^0 = (\ldots, \xi_{-1}, \xi_0)$. Consider the problem of constructing an *optimal* (least-squares) *linear estimator* $\hat{\xi}_n$ of ξ_n in terms of the "past" $\xi^0 = (\ldots, \xi_{-1}, \xi_0)$. It follows from §11, Chapter II, that

$$\hat{\xi}_n = \hat{\mathsf{E}}(\xi_n | H_0(\xi)). \tag{7}$$

(In the notation of Subsection 1, $\hat{\xi}_n = \hat{\pi}_0(\xi_n)$.) Since ξ^r and ξ^s are orthogonal and $H_0(\xi) = H_0(\xi^r) \oplus H_0(\xi^s)$, we obtain, by using (6),

$$\begin{aligned}
\xi_n &= \hat{\mathsf{E}}(\xi_n^s + \xi_n^r | H_0(\xi)) = \hat{\mathsf{E}}(\xi_n^s | H_0(\xi)) + \hat{\mathsf{E}}(\xi_n^r | H_0(\xi)) \\
&= \hat{\mathsf{E}}(\xi_n^s | H_0(\xi^r) \oplus H_0(\xi^s)) + \hat{\mathsf{E}}(\xi_n^r | H_0(\xi^r) \oplus H_0(\xi^s)) \\
&= \hat{\mathsf{E}}(\xi_n^s | H_0(\xi^s)) + \hat{\mathsf{E}}(\xi_n^r | H_0(\xi^r)) \\
&= \xi_n^s + \hat{\mathsf{E}}\left(\sum_{k=0}^{\infty} a_k \varepsilon_{n-k} | H_0(\xi^r)\right).
\end{aligned}$$

In (6), the sequence $\varepsilon = (\varepsilon_n)$ is an innovation sequence for $\xi^r = (\xi_n^r)$ and therefore $H_0(\xi^r) = H_0(\varepsilon)$. Therefore

$$\hat{\xi}_n = \xi_n^s + \hat{\mathsf{E}}\left(\sum_{k=0}^{\infty} a_k \varepsilon_{n-k} | H_0(\varepsilon)\right) = \xi_n^s + \sum_{k=n}^{\infty} a_k \varepsilon_{n-k} \tag{8}$$

and the mean-square error of predicting ξ_n by $\xi_0 = (\ldots, \xi_{-1}, \xi_0)$ is

$$\sigma_n^2 = \mathsf{E}|\xi_n - \hat{\xi}_n|^2 = \sum_{k=0}^{n-1} |a_k|^2. \tag{9}$$

We can draw two important conclusions.

(a) If ξ is *singular*, then for every $n \geq 1$ the error (in the extrapolation) σ_n^2 is zero; in other words, we can predict ξ_n without error from its "past" $\xi^0 = (\ldots, \xi_{-1}, \xi_0)$.
(b) If ξ is *regular*, then $\sigma_n^2 \leq \sigma_{n+1}^2$ and

$$\lim_{n\to\infty} \sigma_n^2 = \sum_{k=0}^{\infty} |a_k|^2. \tag{10}$$

Since

$$\sum_{k=0}^{\infty} |a_k|^2 = \mathsf{E}|\xi_n|^2,$$

it follows from (10) and (9) that

$$\hat{\xi}_n \xrightarrow{L^2} 0, \qquad n \to \infty;$$

i.e. as n increases, the prediction of ξ_n in terms of $\xi_0 = (\ldots, \xi_{-1}, \xi_0)$ becomes trivial (reducing simply to $\mathsf{E}\xi_n = 0$).

4. Let us suppose that ξ is a nondegenerate regular stationary sequence. According to Theorem 2, every such sequence admits a representation as a one-sided moving average

$$\xi_n = \sum_{k=0}^{\infty} a_k \varepsilon_{n-k}, \tag{11}$$

where $\sum_{k=0}^{\infty} |a_k|^2 < \infty$ and the orthonormal sequence $\varepsilon = (\varepsilon_n)$ has the important property that

$$H_n(\xi) = H_n(\varepsilon), \qquad n \in \mathbb{Z}. \tag{12}$$

The representation (11) means (see Subsection 3, §3) that ξ_n can be interpreted as the output signal of a physically realizable filter with impulse response $a = (a_k)$, $k \geq 0$, when the input is $\varepsilon = (\varepsilon_n)$.

Like any sequence of two-sided moving averages, a regular sequence has a spectral density $f(\lambda)$. But since a regular sequence admits a representation as a one-sided moving average it is possible to obtain additional information about properties of the spectral density.

In the first place, it is clear that

$$f(\lambda) = \frac{1}{2\pi} |\varphi(\lambda)|^2,$$

where

$$\varphi(\lambda) = \sum_{k=0}^{\infty} e^{-i\lambda k} a_k, \qquad \sum_{k=0}^{\infty} |a_k|^2 < \infty. \tag{13}$$

Put

$$\Phi(z) = \sum_{k=0}^{\infty} a_k z^k. \tag{14}$$

This function is analytic in the open domain $|z| < 1$ and since $\sum_{k=0}^{\infty} |a_k|^2 < \infty$ it belongs to the *Hardy class* H^2, the class of functions $g = g(z)$, analytic in $|z| < 1$, satisfying

$$\sup_{0 \le r < 1} \frac{1}{2\pi} \int_{-\pi}^{\pi} |g(re^{i\theta})|^2 \, d\theta < \infty. \tag{15}$$

In fact,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |\Phi(re^{i\theta})|^2 \, d\theta = \sum_{k=0}^{\infty} |a_k|^2 r^{2k}$$

and

$$\sup_{0 \le r < 1} \sum |a_k|^2 r^{2k} \le \sum |a_k|^2 < \infty.$$

It is shown in the theory of functions of a complex variable that the boundary function $\Phi(e^{i\lambda})$, $-\pi \le \lambda < \pi$, of $\Phi \in H^2$, not identically zero, has the property that

$$\int_{-\pi}^{\pi} \ln |\Phi(e^{-i\lambda})| \, d\lambda > -\infty. \tag{16}$$

In our case

$$f(\lambda) = \frac{1}{2\pi} |\Phi(e^{-i\lambda})|^2,$$

where $\Phi \in H^2$. Therefore

$$\ln f(\lambda) = -\ln 2\pi + 2 \ln |\Phi(e^{-i\lambda})|,$$

and consequently the spectral density $f(\lambda)$ of a regular process satisfies

$$\int_{-\pi}^{\pi} \ln f(\lambda) \, d\lambda > -\infty. \tag{17}$$

On the other hand, let the spectral density $f(\lambda)$ satisfy (17). It again follows from the theory of functions of a complex variable that there is then a function $\Phi(z) = \sum_{k=0}^{\infty} a_k z^k$ in the Hardy class H^2 such that (almost everywhere with respect to Lebesgue measure)

$$f(\lambda) = \frac{1}{2\pi} |\Phi(e^{-i\lambda})|^2.$$

Therefore if we put $\varphi(\lambda) = \Phi(e^{-i\lambda})$ we obtain

$$f(\lambda) = \frac{1}{2\pi}|\varphi(\lambda)|^2,$$

where $\varphi(\lambda)$ is given by (13). Then it follows from the corollary to Theorem 3, §3, that ξ admits a representation as a one-sided moving average (11), where $\varepsilon = (\varepsilon_n)$ is an orthonormal sequence. From this and from the Remark on Theorem 2, it follows that ξ is regular.

Thus we have the following theorem.

Theorem 4 (Kolmogorov). *Let ξ be a nondegenerate regular stationary sequence. Then there is a spectral density $f(\lambda)$ such that*

$$\int_{-\pi}^{\pi} \ln f(\lambda)\, d\lambda > -\infty. \tag{18}$$

In particular, $f(\lambda) > 0$ (almost everywhere with respect to Lebesgue measure).

Conversely, if ξ is a stationary sequence with a spectral density satisfying (18), *the sequence is regular.*

5. Problems

1. Show that a stationary sequence with discrete spectrum (piecewise-constant spectral function $F(\lambda)$) is singular.

2. Let $\sigma_n^2 = \mathsf{E}|\xi_n - \hat{\xi}_n|^2$, $\hat{\xi}_n = \hat{\mathsf{E}}(\xi_n | H_0(\xi))$. Show that if $\sigma_n^2 = 0$ for some $n \geq 1$, the sequence is singular; if $\sigma_n^2 \to R(0)$ as $n \to \infty$, the sequence is regular.

3. Show that the stationary sequence $\xi = (\xi_n)$, $\xi_n = e^{in\varphi}$, where φ is a uniform random variable on $[0, 2\pi]$, is regular. Find the estimator $\hat{\xi}_n$ and the number σ_n^2, and show that the *nonlinear* estimator

$$\tilde{\xi}_n = \left(\frac{\xi_0}{\xi_{-1}}\right)^n$$

provides a correct estimate of ξ_n by the "past" $\xi^0 = (\ldots, \xi_{-1}, \xi_0)$, i.e.

$$\mathsf{E}|\tilde{\xi}_n - \xi_n|^2 = 0, \qquad n \geq 1.$$

§6. Extrapolation, Interpolation and Filtering

1. Extrapolation. According to the preceding section, a singular sequence admits an error-free prediction (extrapolation) of ξ_n, $n \geq 1$, in terms of the "past," $\xi^0 = (\ldots, \xi_{-1}, \xi_0)$. Consequently it is reasonable, when considering the problem of extrapolation for arbitrary stationary sequences, to begin with the case of regular sequences.

According to Theorem 2 of §5, every regular sequence $\xi = (\xi_n)$ admits a representation as a one-sided moving average,

$$\xi_n = \sum_{k=0}^{\infty} a_k \varepsilon_{n-k} \tag{1}$$

with $\sum_{k=0}^{\infty} |a_k|^2 < \infty$ and some innovation sequence $\varepsilon = (\varepsilon_n)$. It follows from §5 that the representation (1) solves the problem of finding the optimal (linear) estimator $\hat{\xi} = \hat{\mathsf{E}}(\xi_n | H_0(\xi))$ since, by (5.8),

$$\hat{\xi}_n = \sum_{k=n}^{\infty} a_k \varepsilon_{n-k} \tag{2}$$

and

$$\sigma_n^2 = \mathsf{E}|\xi_n - \hat{\xi}_n|^2 = \sum_{k=0}^{n-1} |a_k|^2. \tag{3}$$

However, this can be considered only as a theoretical solution, for the following reasons.

The sequences that we consider are ordinarily not given to us by means of their representations (1), but by their covariance functions $R(n)$ or the spectral densities $f(\lambda)$ (which exist for regular sequences). Hence a solution (2) can only be regarded as satisfactory if the coefficients a_k are given in terms of $R(n)$ or of $f(\lambda)$, and ε_k are given by their values $\cdots \xi_{k-1}, \xi_k$.

Without discussing the problem in general, we consider only the special case (of interest in applications) when the spectral density has the form

$$f(\lambda) = \frac{1}{2\pi} |\Phi(e^{-i\lambda})|^2, \tag{4}$$

where $\Phi(z) = \sum_{k=0}^{\infty} b_k z^k$ has radius of convergence $r > 1$ and has no zeros in $|z| \le 1$.

Let

$$\xi_n = \int_{-\pi}^{\pi} e^{i\lambda n} Z(d\lambda) \tag{5}$$

be the spectral representation of $\xi = (\xi_n)$, $n \in \mathbb{Z}$.

Theorem 1. *If the spectral density of ξ has the density* (4), *then the optimal (linear) estimator $\hat{\xi}_n$ of ξ_n in terms of $\xi^0 = (\ldots, \xi_{-1}, \xi_0)$ is given by*

$$\hat{\xi}_n = \int_{-\pi}^{\pi} \hat{\varphi}_n(\lambda) Z(d\lambda), \tag{6}$$

where

$$\hat{\varphi}_n(\lambda) = e^{i\lambda n} \frac{\Phi_n(e^{-i\lambda})}{\Phi(e^{-i\lambda})} \tag{7}$$

and

$$\Phi_n(z) = \sum_{k=n}^{\infty} b_k z^k.$$

PROOF. According to the remark on Theorem 2 of §3, every variable $\tilde{\xi}_n \in H_0(\xi)$ admits a representation in the form

$$\tilde{\xi}_n = \int_{-\pi}^{\pi} \tilde{\varphi}_n(\lambda) Z(d\lambda), \qquad \tilde{\varphi}_n \in H_0(F), \tag{8}$$

where $H_0(F)$ is the closed linear manifold spanned by the functions $e_n = e^{i\lambda n}$ for $n \leq 0$ $(F(\lambda) = \int_{-\pi}^{\lambda} f(v)\, dv)$.

Since

$$\begin{aligned} \mathsf{E}|\xi_n - \tilde{\xi}_n|^2 &= \mathsf{E}\left|\int_{-\pi}^{\pi} (e^{i\lambda n} - \tilde{\varphi}_n(\lambda)) Z(d\lambda)\right|^2 \\ &= \int_{-\pi}^{\pi} |e^{i\lambda n} - \tilde{\varphi}_n(\lambda)|^2 f(\lambda)\, d\lambda, \end{aligned}$$

the proof that (6) is optimal reduces to proving that

$$\inf_{\tilde{\varphi}_n \in H_0(F)} \int_{-\pi}^{\pi} |e^{i\lambda n} - \tilde{\varphi}_n(\lambda)|^2 f(\lambda)\, d\lambda = \int_{-\pi}^{\pi} |e^{i\lambda n} - \hat{\varphi}_n(\lambda)|^2 f(\lambda)\, d\lambda. \tag{9}$$

It follows from Hilbert-space theory (§11, Chapter II) that the optimal function $\hat{\varphi}_n(\lambda)$ (in the sense of (9)) is determined by the two conditions

(1) $\hat{\varphi}_n(\lambda) \in H_0(F)$, (10)
(2) $e^{i\lambda n} - \hat{\varphi}_n(\lambda) \perp H_0(F)$.

Since

$$e^{i\lambda n}\Phi_n(e^{-i\lambda}) = e^{i\lambda n}[b_n e^{-i\lambda n} + b_{n+1} e^{-i\lambda(n+1)} + \cdots] \in H_0(F)$$

and in a similar way $1/\Phi(e^{-i\lambda}) \in H_0(F)$, the function $\hat{\varphi}_n(\lambda)$ defined in (7) belongs to $H_0(F)$. Therefore in proving that $\hat{\varphi}_n(\lambda)$ is optimal it is sufficient to verify that, for every $m \geq 0$,

$$e^{i\lambda n} - \hat{\varphi}_n(\lambda) \perp e^{i\lambda m},$$

i.e.

$$I_{n,m} \equiv \int_{-\pi}^{\pi} [e^{i\lambda n} - \hat{\varphi}_n(\lambda)] e^{-i\lambda m} f(\lambda)\, d\lambda = 0, \qquad m \geq 0.$$

The following chain of equations shows that this is actually the case:

$$\begin{aligned} I_{n,m} &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\lambda(n-m)} \left[1 - \frac{\Phi_n(e^{-i\lambda})}{\Phi(e^{-i\lambda})}\right] |\Phi(e^{-i\lambda})|^2\, d\lambda \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\lambda(n-m)} [\Phi(e^{-i\lambda}) - \Phi_n(e^{-i\lambda})] \overline{\Phi(e^{-i\lambda})}\, d\lambda \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\lambda(n-m)} \left(\sum_{k=0}^{n-1} b_k e^{-i\lambda k}\right) \left(\sum_{l=0}^{\infty} \bar{b}_l e^{i\lambda l}\right) d\lambda \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-i\lambda m} \left(\sum_{k=0}^{n-1} b_k e^{i\lambda(n-k)}\right) \left(\sum_{l=0}^{\infty} \bar{b}_l e^{i\lambda l}\right) d\lambda = 0, \end{aligned}$$

where the last equation follows because, for $m \geq 0$ and $r > 1$,

$$\int_{-\pi}^{\pi} e^{-i\lambda m} e^{i\lambda r}\, d\lambda = 0.$$

This completes the proof of the theorem.

Remark 1. Expanding $\hat{\varphi}_n(\lambda)$ in a Fourier series, we find that the predicted value $\hat{\xi}_n$ of ξ_n, $n \geq 1$, in terms of the past, $\xi^0 = (\ldots, \xi_{-1}, \xi_0)$, is given by the formula

$$\hat{\xi}_n = C_0\xi_0 + C_{-1}\xi_{-1} + C_{-2}\xi_{-2} + \cdots.$$

Remark 2. A typical example of a spectral density represented in the form (4) is the *rational* function

$$f(\lambda) = \frac{1}{2\pi}\left|\frac{P(e^{-i\lambda})}{Q(e^{-i\lambda})}\right|^2,$$

where the polynomials $P(z) = a_0 + a_1 z + \cdots + a_p z^p$ and $Q(z) = 1 + b_1 z + \cdots + b_q z^q$ have no zeros in $\{z: |z| \leq 1\}$.

In fact, in this case it is enough to put $\Phi(z) = P(z)/Q(z)$. Then $\Phi(z) = \sum_{k=0}^{\infty} C_k z^k$ and the radius of convergence of this series is greater than one.

Let us illustrate Theorem 1 with two examples.

Example 1. Let the spectral density be

$$f(\lambda) = \frac{1}{2\pi}(5 + 4\cos\lambda).$$

The corresponding covariance function $R(n)$ has the shape of a triangle with

$$R(0) = 5, \qquad R(\pm 1) = 2, \qquad R(n) = 0 \quad \text{for} \quad |n| \geq 2. \tag{11}$$

Since this spectral density can be represented in the form

$$f(\lambda) = \frac{1}{2\pi}|2 + e^{-i\lambda}|^2,$$

we may apply Theorem 1. We find easily that

$$\hat{\varphi}_1(\lambda) = e^{i\lambda}\frac{e^{-i\lambda}}{2 + e^{-i\lambda}}, \qquad \hat{\varphi}_n(\lambda) = 0 \quad \text{for} \quad n \geq 2. \tag{12}$$

Therefore $\hat{\xi}_n = 0$ for all $n \geq 2$, i.e. the (linear) prediction of ξ_n in terms of $\xi^0 = (\ldots, \xi_{-1}, \xi_0)$ is trivial, which is not at all surprising if we observe that, by (11), the correlation between ξ_n and any of $\xi_0, \xi_{-1}, \ldots$ is zero for $n \geq 2$.

For $n = 1$ we find from (6) and (12) that

$$\hat{\xi}_1 = \int_{-\pi}^{\pi} e^{i\lambda} \frac{e^{-i\lambda}}{2 + e^{-i\lambda}} Z(d\lambda)$$

$$= \frac{1}{2} \int_{-\pi}^{\pi} \frac{1}{\left(1 + \frac{e^{-i\lambda}}{2}\right)} Z(d\lambda) = \sum_{k=0}^{\infty} \frac{(-1)^k}{2^{k+1}} \int_{-\pi}^{\pi} e^{-ik\lambda} Z(d\lambda)$$

$$= \sum_{k=0}^{\infty} \frac{(-1)^k \xi_k}{2^{k+1}} = \tfrac{1}{2}\xi_0 - \tfrac{1}{4}\xi_{-1} + \cdots.$$

EXAMPLE 2. Let the covariance function be

$$R(n) = a^n, \qquad |a| < 1.$$

Then (see Example 5 in §1)

$$f(\lambda) = \frac{1}{2\pi} \frac{1 - |a|^2}{|1 - ae^{-i\lambda}|^2},$$

i.e.

$$f(\lambda) = \frac{1}{2\pi} |\Phi(e^{-i\lambda})|^2,$$

where

$$\Phi(z) = \frac{(1 - |a|^2)^{1/2}}{1 - az} = (1 - |a|^2)^{1/2} \sum_{k=0}^{\infty} (az)^k,$$

from which $\varphi_n(\lambda) = a^n$ and therefore

$$\hat{\xi}_n = \int_{-\pi}^{\pi} a^n Z(d\lambda) = a^n \xi_0.$$

In other words, in order to predict the value of ξ_n from the observations $\xi^0 = (\ldots, \xi_{-1}, \xi_0)$ it is sufficient to know only the last observation ξ_0.

Remark 3. It follows from the Wold expansion of the regular sequence $\xi = (\xi_n)$ with

$$\xi_n = \sum_{k=0}^{\infty} a_k \xi_{n-k} \tag{13}$$

that the spectral density $f(\lambda)$ admits the representation

$$f(\lambda) = \frac{1}{2\pi} |\Phi(e^{-i\lambda})|^2, \tag{14}$$

where

$$\Phi(z) = \sum_{k=0}^{\infty} a_k z^k. \tag{15}$$

It is evident that the converse also holds, that is, if $f(\lambda)$ admits the representation (14) with a function $\Phi(z)$ of the form (15), then the Wold expansion of ξ_n has the form (13). Therefore the problem of representing the spectral density in the form (14) and the problem of determining the coefficients a_k in the Wold expansion are equivalent.

The assumptions that $\Phi(z)$ in Theorem 1 has no zeros for $|z| \leq 1$ and that $r > 1$ are in fact not essential. In other words, if the spectral density of a regular sequence is represented in the form (14), then the optimal estimator $\hat{\xi}_n$ (in the mean square sense) for ξ_n in terms of $\xi^0 = (\ldots, \xi_{-1}, \xi_0)$ is determined by formulas (6) and (7).

Remark 4. Theorem 1 (with the preceding remark) solves the prediction problem for regular sequences. Let us show that in fact the same answer remains valid for arbitrary stationary sequences. More precisely, let

$$\xi_n = \xi_n^s + \xi_n^r, \qquad \xi_n = \int_{-\pi}^{\pi} e^{i\lambda n} Z(d\lambda), \qquad F(\Delta) = \mathsf{E}|Z(\Delta)|^2,$$

and let $f^r(\lambda) = (1/2\pi)|\Phi(e^{-i\lambda})|^2$ be the spectral density of the regular sequence $\xi^r = (\xi_n^r)$. Then $\hat{\xi}_n$ is determined by (6) and (7).

In fact, let (see Subsection 3, §5)

$$\hat{\xi}_n = \int_{-\pi}^{\pi} \hat{\varphi}_n(\lambda) Z(d\lambda), \qquad \hat{\xi}_n^r = \int_{-\pi}^{\pi} \hat{\varphi}_n^r(\lambda) Z^r(d\lambda),$$

where $Z^r(\Delta)$ is the orthogonal stochastic measure in the representation of the regular sequence ξ^r. Then

$$\begin{aligned}
\mathsf{E}|\xi_n - \hat{\xi}_n|^2 &= \int_{-\pi}^{\pi} |e^{i\lambda n} - \hat{\varphi}_n(\lambda)|^2 F(d\lambda) \\
&\geq \int_{-\pi}^{\pi} |e^{i\lambda n} - \hat{\varphi}_n(\lambda)|^2 f^r(\lambda)\, d\lambda \geq \int_{-\pi}^{\pi} |e^{i\lambda n} - \hat{\varphi}_n^r(\lambda)|^2 f^r(\lambda)\, d\lambda \\
&= \mathsf{E}|\xi_n^r - \hat{\xi}_n^r|^2. \qquad (16)
\end{aligned}$$

But $\xi_n - \hat{\xi}_n = \xi_n^r - \hat{\xi}^r$. Hence $\mathsf{E}|\xi_n - \hat{\xi}_n|^2 = \mathsf{E}|\xi_n^r - \hat{\xi}_n^r|^2$, and it follows from (16) that we may take $\hat{\varphi}_n(\lambda)$ to be $\hat{\varphi}_n^r(\lambda)$.

2. Interpolation. Suppose that $\xi = (\xi_n)$ is a regular sequence with spectral density $f(\lambda)$. The simplest interpolation problem is the problem of constructing the optimal (mean-square) linear estimator from the results of the measurements $\{\xi_n, n = \pm 1, \pm 2, \ldots\}$ omitting ξ_0.

Let $H^0(\xi)$ be the closed linear manifold spanned by ξ_n, $n \neq 0$. Then according to the results of Theorem 2, §3, every random variable $\eta \in H^0(\xi)$ can be represented in the form

$$\eta = \int_{-\pi}^{\pi} \varphi(\lambda) Z(d\lambda),$$

where φ belongs to $H^0(F)$, the closed linear manifold spanned by the functions $e^{i\lambda n}$, $n \neq 0$. The estimator

$$\check{\xi}_0 = \int_{-\pi}^{\pi} \check{\varphi}(\lambda) Z(d\lambda) \tag{17}$$

will be optimal if and only if

$$\inf_{\eta \in H^0(\xi)} \mathsf{E}|\xi_0 - \eta|^2 = \inf_{\varphi \in H^0(F)} \int_{-\pi}^{\pi} |1 - \varphi(\lambda)|^2 F(d\lambda)$$
$$= \int_{-\pi}^{\pi} |1 - \check{\varphi}(\lambda)|^2 F(d\lambda) = \mathsf{E}|\xi_0 - \check{\xi}_0|^2.$$

It follows from the perpendicularity properties of the Hilbert space $H^0(F)$ that $\check{\varphi}(\lambda)$ is completely determined (compare (10)) by the two conditions

$$\begin{aligned} &(1) \quad \check{\varphi}(\lambda) \in H^0(F), \\ &(2) \quad 1 - \check{\varphi}(\lambda) \perp H^0(F). \end{aligned} \tag{18}$$

Theorem 2 (Kolmogorov). *Let $\xi = (\xi_n)$ be a regular sequence such that*

$$\int_{-\pi}^{\pi} \frac{d\lambda}{f(\lambda)} < \infty. \tag{19}$$

Then

$$\check{\varphi}(\lambda) = 1 - \frac{\alpha}{f(\lambda)}, \tag{20}$$

where

$$\alpha = \frac{2\pi}{\displaystyle\int_{-\pi}^{\pi} \frac{d\lambda}{f(\lambda)}}, \tag{21}$$

and the interpolation error $\delta^2 = \mathsf{E}|\xi_0 - \check{\xi}_0|^2$ is given by $\delta^2 = 2\pi \cdot \alpha$.

Proof. We shall give the proof only under very stringent hypotheses on the spectral density, specifically that

$$0 < c \leq f(\lambda) \leq C < \infty. \tag{22}$$

It follows from (2) and (18) that

$$\int_{-\pi}^{\pi} [1 - \check{\varphi}(\lambda)] e^{in\lambda} f(\lambda)\, d\lambda = 0 \tag{23}$$

for every $n \neq 0$. By (22), the function $[1 - \check{\varphi}(\lambda)] f(\lambda)$ belongs to the Hilbert space $L^2([-\pi, \pi], \mathscr{B}[-\pi, \pi], \mu)$ with Lebesgue measure μ. In this space the functions $\{e^{in\lambda}/\sqrt{2\pi}, n = 0, \pm 1, \ldots\}$ form an orthonormal basis (Problem 7, §11, Chapter II). Hence it follows from (23) that $[1 - \check{\varphi}(\lambda)] f(\lambda)$ is a constant, which we denote by α. Thus the second condition in (18) leads to the conclusion that

$$\tilde{\varphi}(\lambda) = 1 - \frac{\alpha}{f(\lambda)}. \tag{24}$$

Starting from the first condition (18), we now determine α.

By (22), $\tilde{\varphi} \in L^2$ and the condition $\tilde{\varphi} \in H^0(F)$ is equivalent to the condition that $\tilde{\varphi}$ belongs to the closed (in the L^2 norm) linear manifold spanned by the functions $e^{i\lambda n}$, $n \neq 0$. Hence it is clear that the zeroth coefficient in the expansion of $\tilde{\varphi}(\lambda)$ must be zero. Therefore

$$0 = \int_{-\pi}^{\pi} \tilde{\varphi}(\lambda)\, d\lambda = 2\pi - \alpha \int_{-\pi}^{\pi} \frac{d\lambda}{f(\lambda)}$$

and hence α is determined by (21).

Finally,

$$\delta^2 = \mathsf{E}|\xi_0 - \tilde{\xi}_0|^2 = \int_{-\pi}^{\pi} |1 - \tilde{\varphi}(\lambda)|^2 f(\lambda)\, d\lambda$$

$$= |\alpha|^2 \int_{-\pi}^{\pi} \frac{f(\lambda)}{f^2(\lambda)}\, d\lambda = \frac{4\pi^2}{\displaystyle\int_{-\pi}^{\pi} \frac{d\lambda}{f(\lambda)}}.$$

This completes the proof (under condition (22)).

Corollary. *If*

$$\tilde{\varphi}(\lambda) = \sum_{0<|k|\le N} c_k e^{i\lambda k},$$

then

$$\tilde{\xi}_0 = \sum_{0<|k|\le N} c_k \int_{-\pi}^{\pi} e^{i\lambda k} Z(d\lambda) = \sum_{0<|k|\le N} c_k \xi_k.$$

Example 3. Let $f(\lambda)$ be the spectral density in Example 2 above. Then an easy calculation shows that

$$\tilde{\xi}_0 = \int_{-\pi}^{\pi} \frac{a}{1 + |a|^2} [e^{i\lambda} + e^{-i\lambda}] Z(d\lambda) = \frac{a}{1 + |a|^2} [\xi_1 + \xi_{-1}],$$

and the interpolation error is

$$\delta^2 = \frac{1 - |\alpha|^2}{1 + |\alpha|^2}.$$

3. Filtering. Let $(\theta, \xi) = ((\theta_n), (\xi_n))$, $n \in \mathbb{Z}$, be a *partially observed sequence*, where $\theta = (\theta_n)$ and $\xi = (\xi_n)$ are respectively the unobserved and the observed components.

Each of the sequences θ and ξ will be supposed stationary (wide sense) with zero mean; let the spectral densities be

$$\theta_n = \int_{-\pi}^{\pi} e^{i\lambda n} Z_\theta(d\lambda), \quad \text{and} \quad \xi_n = \int_{-\pi}^{\pi} e^{i\lambda n} Z_\xi(d\lambda).$$

We write

$$F_\theta(\Delta) = \mathsf{E}\,|Z_\theta(\Delta)|^2, \qquad F_\xi(\Delta) = \mathsf{E}\,|Z_\xi(\Delta)|^2$$

and

$$F_{\theta\xi}(\Delta) = \mathsf{E} Z_\theta(\Delta)\overline{Z}_\xi(\Delta).$$

In addition, we suppose that θ and ξ are *connected in a stationary way*, i.e. that their covariance functions $\operatorname{cov}(\theta_n, \xi_m) = \mathsf{E}\theta_n\bar{\xi}_m$ depend only on the differences $n - m$. Let $R_{\theta\xi}(n) = \mathsf{E}\theta_n\bar{\xi}_0$; then

$$R_{\theta\xi}(n) = \int_{-\pi}^{\pi} e^{i\lambda n} F_{\theta\xi}(d\lambda).$$

The filtering problem that we shall consider is the construction of the optimal (mean-square) linear estimator $\hat{\theta}_n$ of θ_n in terms of some observation of the sequence ξ.

The problem is easily solved under the assumption that θ_n is to be constructed from *all* the values ξ_m, $m \in \mathbb{Z}$. In fact, since $\hat{\theta}_n = \hat{\mathsf{E}}(\theta_n \mid H(\xi))$ there is a function $\hat{\varphi}_n(\lambda)$ such that

$$\hat{\theta}_n = \int_{-\pi}^{\pi} \hat{\varphi}_n(\lambda) Z_\xi(d\lambda). \tag{25}$$

As in Subsections 1 and 2, the conditions to impose on the optimal $\hat{\varphi}_n(\lambda)$ are that

(1) $\hat{\varphi}_n(\lambda) \in H(F_\xi)$,
(2) $\theta_n - \hat{\theta}_n \perp H(\xi)$.

From the latter condition we find

$$\int_{-\pi}^{\pi} e^{i\lambda(n-m)} F_{\theta\xi}(d\lambda) - \int_{-\pi}^{\pi} e^{-i\lambda m} \hat{\varphi}_n(\lambda) F_\xi(d\lambda) = 0 \tag{26}$$

for every $m \in \mathbb{Z}$. Therefore if we suppose that $F_{\theta\xi}(\lambda)$ and $F_\xi(\lambda)$ have densities $f_{\theta\xi}(\lambda)$ and $f_\xi(\lambda)$, we find from (26) that

$$\int_{-\pi}^{\pi} e^{i\lambda(n-m)} [f_{\theta\xi}(\lambda) - e^{-i\lambda n} \hat{\varphi}_n(\lambda) f_\xi(\lambda)]\, d\lambda = 0.$$

If $f_\xi(\lambda) > 0$ (almost everywhere with respect to Lebesgue measure) we find immediately that

$$\hat{\varphi}_n(\lambda) = e^{i\lambda n} \hat{\varphi}(\lambda), \tag{27}$$

where

$$\varphi(\lambda) = f_{\theta\xi}(\lambda) \cdot f_\xi^{\oplus}(\lambda)$$

and $f_\xi^{\oplus}(\lambda)$ is the "pseudotransform" of $f_\xi(\lambda)$, i.e.

$$f_\xi^{\oplus}(\lambda) = \begin{cases} f_\xi^{-1}(\lambda), & f_\xi(\lambda) > 0, \\ 0, & f_\xi(\lambda) = 0. \end{cases}$$

Then the filtering error is

$$\mathsf{E}|\theta_n - \hat{\theta}_n|^2 = \int_{-\pi}^{\pi} [f_\theta(\lambda) - f_{\theta\xi}^2(\lambda) f_\xi^{\oplus}(\lambda)]\, d\lambda. \tag{28}$$

As is easily verified, $\varphi \in H(F_\xi)$, and consequently the estimator (25), with the function (27), is optimal.

EXAMPLE 4. *Detection of a signal in the presence of noise.* Let $\xi_n = \theta_n + \eta_n$, where the signal $\theta = (\theta_n)$ and the noise $\eta = (\eta_n)$ are uncorrelated sequences with spectral densities $f_\theta(\lambda)$ and $f_\eta(\lambda)$. Then

$$\hat{\theta}_n = \int_{-\pi}^{\pi} e^{i\lambda n} \varphi(\lambda) Z_\xi(d\lambda),$$

where

$$\varphi(\lambda) = f_\theta(\lambda)[f_\theta(\lambda) + f_\eta(\lambda)]^{\oplus},$$

and the filtering error is

$$\mathsf{E}|\theta_n - \hat{\theta}_n|^2 = \int_{-\pi}^{\pi} [f_\theta(\lambda) f_\eta(\lambda)][f_\theta(\lambda) + f_\eta(\lambda)]^{\oplus}\, d\lambda.$$

The solution (25) obtained above can now be used to construct an optimal estimator $\tilde{\theta}_{n+m}$ of θ_{n+m} as a result of observing ξ_k, $k \le n$, where m is a given element of $\mathbb{Z}$. Let us suppose that $\xi = (\xi_n)$ is regular, with spectral density

$$f(\lambda) = \frac{1}{2\pi} |\Phi(e^{-i\lambda})|^2,$$

where $\Phi(z) = \sum_{k=0}^{\infty} a_k z^k$. By the Wold expansion,

$$\xi_n = \sum_{k=0}^{\infty} a_k \varepsilon_{n-k},$$

where $\varepsilon = (\varepsilon_k)$ is white noise with the spectral resolution

$$\varepsilon_n = \int_{-\pi}^{\pi} e^{i\lambda n} Z_\varepsilon(d\lambda).$$

Since

$$\tilde{\theta}_{n+m} = \hat{\mathsf{E}}[\theta_{n+m} | H_n(\xi)] = \hat{\mathsf{E}}[\hat{\mathsf{E}}[\theta_{n+m} | H(\xi)] | H_n(\xi)] = \hat{\mathsf{E}}[\hat{\theta}_{n+m} | H_n(\xi)]$$

and

$$\hat{\theta}_{n+m} = \int_{-\pi}^{\pi} e^{i\lambda(n+m)}\hat{\varphi}(\lambda)\Phi(e^{-i\lambda})Z_\varepsilon(d\lambda) = \sum_{k \le n+m} \hat{a}_{n+m-k}\varepsilon_k,$$

where

$$\hat{a}_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\lambda k}\hat{\varphi}(\lambda)\Phi(e^{-i\lambda})\, d\lambda, \tag{29}$$

then

$$\tilde{\theta}_{n+m} = \hat{\mathsf{E}}\left[\sum_{k \le n+m} \hat{a}_{n+m-k}\varepsilon_k \,|\, H_n(\xi)\right].$$

But $H_n(\xi) = H_n(\varepsilon)$ and therefore

$$\tilde{\theta}_{n+m} = \sum_{k \le n} \hat{a}_{n+m-k}\varepsilon_k = \int_{-\pi}^{\pi}\left[\sum_{k \le n} \hat{a}_{n+m-k}e^{i\lambda k}\right]Z_\varepsilon(d\lambda)$$

$$= \int_{-\pi}^{\pi} e^{i\lambda n}\left[\sum_{l=0}^{\infty} \hat{a}_{l+m}e^{-i\lambda l}\right]\Phi^{\oplus}(e^{-i\lambda})Z_\xi(d\lambda),$$

where $\Phi^{\oplus}$ is the pseudotransform of Φ.

We have therefore established the following theorem.

Theorem 3. *If the sequence $\xi = (\xi_n)$ under observation is regular, then the optimal (mean-square) linear estimator $\tilde{\theta}_{n+m}$ of θ_{n+m} in terms of ξ_k, $k \le n$, is given by*

$$\tilde{\theta}_{n+m} = \int_{-\pi}^{\pi} e^{i\lambda n}H_m(e^{-i\lambda})Z_\xi(d\lambda), \tag{30}$$

where

$$H_m(e^{-i\lambda}) = \sum_{l=0}^{\infty} \hat{a}_{l+m}e^{-i\lambda l}\Phi^{\oplus}(e^{-i\lambda}) \tag{31}$$

and the coefficients a_k are defined by (29).

4. Problems

1. Let ξ be a nondegenerate regular sequence with spectral density (4). Show that $\Phi(z)$ has no zeros for $|z| \le 1$.
2. Show that the conclusion of Theorem 1 remains valid even without the hypotheses that $\Phi(z)$ has radius of convergence $r > 1$ and that the zeros of $\Phi(z)$ all lie in $|z| > 1$.

3. Show that, for a regular process, the function $\Phi(z)$ introduced in (4) can be represented in the form
$$\Phi(z) = \sqrt{2\pi}\exp\left\{\tfrac{1}{2}c_0 + \sum_{k=1}^{\infty} c_k z^k\right\}, \qquad |z| < 1,$$
where
$$c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\lambda} \ln f(\lambda)\, d\lambda.$$
Deduce from this formula and (5.9) that the one-step prediction error $\sigma_1^2 = \mathsf{E}|\hat{\xi}_1 - \xi_1|^2$ is given by the Szegö–Kolmogorov formula
$$\sigma_1^2 = 2\pi \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln f(\lambda)\, d\lambda\right\}.$$

4. Prove Theorem 2 without using (22).

5. Let a signal θ and a noise η, not correlated with each other, have spectral densities
$$f_\theta(\lambda) = \frac{1}{2\pi}\cdot\frac{1}{|1 + b_1 e^{-i\lambda}|^2} \quad\text{and}\quad f_\eta(\lambda) = \frac{1}{2\pi}\cdot\frac{1}{|1 + b_2 e^{-i\lambda}|^2}.$$
Using Theorem 3, find an estimator $\tilde{\theta}_{n+m}$ for θ_{n+m} in terms of ξ_k, $k \le n$, where $\xi_k = \theta_k + \eta_k$. Consider the same problem for the spectral densities
$$f_\theta(\lambda) = \frac{1}{2\pi}|2 + e^{-i\lambda}|^2 \quad\text{and}\quad f_\eta(\lambda) = \frac{1}{2\pi}.$$

§7. The Kalman–Bucy Filter and Its Generalizations

1. From a computational point of view, the solution presented above for the problem of filtering out an unobservable component θ by means of observations of ξ is not practical, since, because it is expressed in terms of the spectrum, it has to be carried out by spectral methods. In the method proposed by Kalman and Bucy, the synthesis of the optimal filter is carried out recursively; this makes it possible to do it with a digital computer. There are also other reasons for the wide use of the Kalman–Bucy filter, one being that it still "works" even without the assumption that the sequence (θ, ξ) is *stationary*.

We shall present not only the usual Kalman–Bucy method, but also a generalization in which the recurrent equations determined by (θ, ξ) have coefficients that depend on all the data observed in the past.

Thus, let us suppose that $(\theta, \xi) = ((\theta_n), (\xi_n))$ is a partially observed sequence, and let
$$\theta_n = (\theta_1(n), \ldots, \theta_k(n)) \quad\text{and}\quad \xi_n = (\xi_1(n), \ldots, \xi_l(n))$$
be governed by the recurrent equations
$$\begin{aligned}\theta_{n+1} &= a_0(n, \xi) + a_1(n, \xi)\theta_n + b_1(n, \xi)\varepsilon_1(n+1) + b_2(n, \xi)\varepsilon_2(n+1),\\ \xi_{n+1} &= A_0(n, \xi) + A_1(n, \xi)\theta_n + B_1(n, \xi)\varepsilon_1(n+1) + B_2(n, \xi)\varepsilon_2(n+1).\end{aligned} \tag{1}$$

Here

$$\varepsilon_1(n) = (\varepsilon_{11}(n), \ldots, \varepsilon_{1k}(n)) \quad \text{and} \quad \varepsilon_2(n) = (\varepsilon_{21}(n), \ldots, \varepsilon_{2l}(n))$$

are independent Gaussian vectors with independent components, each of which is normally distributed with parameters 0 and 1; $a_0(n, \xi) = (a_{01}(n, \xi), \ldots, a_{0k}(n, \xi))$ and $A_0(n, \xi) = (A_{01}(n, \xi), \ldots, A_{0l}(n, \xi))$ are vector functions, where the dependence on $\xi = \{\xi_0, \ldots, \xi_n\}$ is determined without looking ahead, i.e. for a given n the functions $a_0(n, \xi), \ldots, A_{0l}(n, \xi)$ depend only on $\xi_0, \ldots, \xi_n$; the matrix functions

$$b_1(n, \xi) = \|b_{ij}^{(1)}(n, \xi)\|, \qquad b_2(n, \xi) = \|b_{ij}^{(2)}(n, \xi)\|,$$

$$B_1(n, \xi) = \|B_{ij}^{(1)}(n, \xi)\|, \qquad B_2(n, \xi) = \|B_{ij}^{(2)}(n, \xi)\|,$$

$$a_1(n, \xi) = \|a_{ij}^{(1)}(n, \xi)\|, \qquad A_1(n, \xi) = \|A_{ij}^{(1)}(n, \xi)\|$$

have orders $k \times k$, $k \times l$, $l \times k$, $l \times l$, $k \times k$, $l \times k$, respectively, and also depend on ξ without looking ahead. We also suppose that the initial vector (θ_0, ξ_0) is independent of the sequences $\varepsilon_1 = (\varepsilon_1(n))$ and $\varepsilon_2 = (\varepsilon_2(n))$.

To simplify the presentation, we shall frequently not indicate the dependence of the coefficients on ξ.

So that the system (1) will have a solution with finite second moments, we assume that $\mathsf{E}(\|\theta_0\|^2 + \|\xi_0\|^2) < \infty$

$$\left(\|x\|^2 = \sum_{i=1}^{k} x_i^2,\ x = (x_1, \ldots, x_k)\right), \qquad |a_{ij}^{(1)}(n, \xi)| \le C, \qquad |A_{ij}^{(1)}(n, \xi)| \le C,$$

and if $g(n, \xi)$ is any of the functions a_{0i}, A_{0j}, $b_{ij}^{(1)}$, $b_{ij}^{(2)}$, $B_{ij}^{(1)}$ or $B_{ij}^{(2)}$ then $\mathsf{E}|g(n, \xi)|^2 < \infty$, $n = 0, 1, \ldots$. With these assumptions, (θ, ξ) has $\mathsf{E}(\|\theta_n\|^2 + \|\xi_n\|^2) < \infty$, $n \ge 0$.

Now let $\mathscr{F}_n^\xi = \sigma\{\omega: \xi_0, \ldots, \xi_n\}$ be the smallest σ-algebra generated by $\xi_0, \ldots, \xi_n$ and

$$m_n = \mathsf{E}(\theta_n | \mathscr{F}_n^\xi), \qquad \gamma_n = \mathsf{E}[(\theta_n - m_n)(\theta_n - m_n)^* | \mathscr{F}_n^\xi].$$

According to Theorem 1, §8, Chapter II, $m_n = (m_1(n), \ldots, m_k(n))$ is an optimal estimator (in the mean square sense) for the vector $\theta_n = (\theta_1(n), \ldots, \theta_k(n))$, and $\mathsf{E}\gamma_n = \mathsf{E}[(\theta_n - m_n)(\theta_n - m_n)^*]$ is the matrix of errors of observation. To determine these matrices for arbitrary sequences (θ, ξ) governed by equations (1) is a very difficult problem. However, there is a further supplementary condition on (θ_0, ξ_0) that leads to a system of recurrent equations for m_n and γ_m that still contains the Kalman–Bucy filter. This is the condition that the conditional distribution $\mathsf{P}(\theta_0 \le a | \xi_0|)$ is Gaussian,

$$\mathsf{P}(\theta_0 \le a | \xi_0) = \frac{1}{\sqrt{2\pi\gamma_0}} \int_{-\infty}^{a} \exp\left\{-\frac{(x - m_0)^2}{2\gamma_0^2}\right\} dx, \tag{2}$$

with parameters $m_0 = m_0(\xi_0)$, $\gamma_0 = \gamma_0(\xi_0)$.

To begin with, let us establish an important auxiliary result.

Lemma 1. *Under the assumptions made above about the coefficients of* (1), *together with* (2), *the sequence* (θ, ξ) *is conditionally Gaussian, i.e. the conditional distribution function*

$$\mathsf{P}\{\theta_0 \le a_0, \ldots, \eta_n \le a_n | \mathscr{F}_n^{\xi}\}$$

is (P-a.s.) *the distribution function of an n-dimensional Gaussian vector whose mean and covariance matrix depend on* $(\xi_0, \ldots, \xi_n)$.

PROOF. We prove only the Gaussian character of $\mathsf{P}(\theta_n \le a | \mathscr{F}_n^{\xi})$; this is enough to let us obtain equations for m_n and γ_n.

First we observe that (1) implies that the conditional distribution

$$\mathsf{P}(\theta_{n+1} \le a_1, \xi_{n+1} \le x | \mathscr{F}_n^{\xi}, \theta_n = b)$$

is Gaussian with mean-value vector

$$\mathbb{A}_0 + \mathbb{A}_1 b = \begin{pmatrix} a_0 + a_1 b \\ A_0 + A_1 b \end{pmatrix}$$

and covariance matrix

$$\mathbb{B} = \begin{pmatrix} b \circ b & b \circ B \\ (b \circ B)^* & B \circ B \end{pmatrix},$$

where $b \circ b = b_1 b_1^* + b_2 b_2^*$, $b \circ B = b_1 B_1^* + b_2 B_2^*$, $B \circ B = B_1 B_1^* + B_2 B_2^*$.

Let $\zeta_n = (\theta_n, \xi_n)$ and $t = (t_1, \ldots, t_{k+l})$. Then

$$\mathsf{E}[\exp(it^* \zeta_{n+1}) | \mathscr{F}_n^{\xi}, \theta_n] = \exp\{it^*(\mathbb{A}_0(n, \xi) + \mathbb{A}_1(n, \xi)\theta_n) - \tfrac{1}{2} t^* \mathbb{B}(n, \xi) t\}. \tag{3}$$

Suppose now that the conclusion of the lemma holds for some $n \ge 0$. Then

$$\mathsf{E}[\exp(it^* \mathbb{A}_1(n, \xi)\theta_n) | \mathscr{F}_n^{\xi}] = \exp(it^* \mathbb{A}_1(n, \xi) m_n - \tfrac{1}{2} t^*(\mathbb{A}_1(n, \xi)\gamma_n \mathbb{A}_1^*(n, \xi))t. \tag{4}$$

Let us show that (4) is also valid when n is replaced by $n + 1$.

From (3) and (4), we have

$$\begin{aligned} \mathsf{E}[\exp(it^* \zeta_{n+1}) | \mathscr{F}_n^{\xi}] = \exp\{&it^*(\mathbb{A}_0(n, \xi) + \mathbb{A}_1(n, \xi) m_n) \\ &- \tfrac{1}{2} t^* \mathbb{B}(n, \xi) t - \tfrac{1}{2} t^*(\mathbb{A}_1(n, \xi)\gamma_n \mathbb{A}_1^*(n, \xi))t\}. \end{aligned}$$

Hence the conditional distribution

$$\mathsf{P}(\theta_{n+1} \le a, \xi_{n+1} \le x | \mathscr{F}_n^{\xi}) \tag{5}$$

is Gaussian.

As in the proof of the theorem on normal correlation (Theorem 2, §13, Chapter II) we can verify that there is a matrix C such that the vector

$$\eta = [\theta_{n+1} - \mathsf{E}(\theta_{n+1} | \mathscr{F}_n^{\xi})] - C[\xi_{n+1} - \mathsf{E}(\xi_{n+1} | \mathscr{F}_n^{\xi})]$$

has the property that (P-a.s.)

$$\mathsf{E}[\eta(\xi_{n+1} - \mathsf{E}(\xi_{n+1} | \mathscr{F}_n^{\xi}))^* | \mathscr{F}_n^{\xi}] = 0.$$

It follows that the conditionally-Gaussian vectors η and ξ_{n+1}, considered under the condition $\mathscr{F}_n^\xi$, are independent, i.e.

$$\mathsf{P}(\eta \in A, \xi_{n+1} \in B \mid \mathscr{F}_n^\xi) = \mathsf{P}(\eta \in A \mid \mathscr{F}_n^\xi) \cdot \mathsf{P}(\xi_{n+1} \in B \mid \mathscr{F}_n^\xi)$$

for all $A \in \mathscr{B}(R^k)$, $B \in \mathscr{B}(R^l)$.

Therefore if $s = (s_1, \dots, s_n)$ then

$$\begin{aligned}
\mathsf{E}[\exp(is^*\theta_{n+1}) \mid \mathscr{F}_n^\xi, \xi_{n+1}]
&= \mathsf{E}\{\exp(is^*[\mathsf{E}(\theta_{n+1} \mid \mathscr{F}_n^\xi) + \eta + C[\xi_{n+1} - \mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^\xi)]]) \mid \mathscr{F}_n^\xi, \xi_{n+1}\} \\
&= \exp\{is^*[\mathsf{E}(\theta_{n+1} \mid \mathscr{F}_n^\xi) + C[\xi_{n+1} - \mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^\xi)]]\} \\
&\quad \times \mathsf{E}[\exp(is^*\eta) \mid \mathscr{F}_n^\xi, \xi_{n+1}] \\
&= \exp\{is^*[\mathsf{E}(\theta_{n+1} \mid \mathscr{F}_n^\xi)] + C[\xi_{n+1} - \mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^\xi)]]\} \\
&\quad \times \mathsf{E}(\exp(is^*\eta) \mid \mathscr{F}_n^\xi]. \qquad (6)
\end{aligned}$$

By (5), the conditional distribution $\mathsf{P}(\eta \le y \mid \mathscr{F}_n^\xi)$ is Gaussian. With (6), this shows that the conditional distribution $\mathsf{P}(\theta_{n+1} \le a \mid \mathscr{F}_{n+1}^\xi)$ is also Gaussian.

This completes the proof of the lemma.

Theorem 1. *Let (θ, ξ) be a partial observation of a sequence that satisfies the system* (1) *and condition* (2). *Then (m_n, γ_n) obey the following recursion relations*:

$$\begin{aligned}
m_{n+1} &= [a_0 + a_1 m_n] + [b \circ B + a_1 \gamma_n A_1^*][B \circ B + A_1 \gamma_n A_1^*]^\oplus \\
&\quad \times [\xi_{n+1} - A_0 - A_1 m_n], \qquad (7)
\end{aligned}$$

$$\begin{aligned}
\gamma_{n+1} &= [a_1 \gamma_n a_1^* + b \circ b] - [b \circ B + a_1 \gamma_n A_1^*][B \circ B + A_1 \gamma_n A_1^*]^\oplus \\
&\quad \times [b \circ B + a_1 \gamma_n A_1^*]^*. \qquad (8)
\end{aligned}$$

PROOF. From (1),

$$\mathsf{E}(\theta_{n+1} \mid \mathscr{F}_n^\xi) = a_0 + a_1 m_n, \qquad \mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^\xi) = A_0 + A_1 m_n \qquad (9)$$

and

$$\begin{aligned}
\theta_{n+1} - \mathsf{E}(\theta_{n+1} \mid \mathscr{F}_n^\xi) &= a_1[\theta_n - m_n] + b_1 \varepsilon_1(n+1) + b_2 \varepsilon_2(n+1), \\
\xi_{n+1} - \mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^\xi) &= A_1[\theta_n - m_n] + B_1 \varepsilon_1(n+1) + B_2 \varepsilon_2(n+1).
\end{aligned} \qquad (10)$$

Let us write

$$\begin{aligned}
d_{11} &= \operatorname{cov}(\theta_{n+1}, \theta_{n+1} \mid \mathscr{F}_n^\xi) \\
&= \mathsf{E}\{[\theta_{n+1} - \mathsf{E}(\theta_{n+1} \mid \mathscr{F}_n^\xi)][\theta_{n+1} - \mathsf{E}(\theta_{n+1} \mid \mathscr{F}_n^\xi)]^* / \mathscr{F}_n^\xi\}, \\
d_{12} &= \operatorname{cov}(\theta_{n+1}, \xi_{n+1} \mid \mathscr{F}_n^\xi) \\
&= \mathsf{E}\{[\theta_{n+1} - \mathsf{E}[\theta_{n+1} \mid \mathscr{F}_n^\xi)][\xi_{n+1} - \mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^\xi)]^* / \mathscr{F}_n^\xi\}, \\
d_{22} &= \operatorname{cov}(\xi_{n+1}, \xi_{n+1} \mid \mathscr{F}_n^\xi) \\
&= \mathsf{E}\{[\xi_{n+1} - \mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^\xi)][\xi_{n+1} - \mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^\xi)]^* / \mathscr{F}_n^\xi\}.
\end{aligned}$$

Then, by (10),

$$d_{11} = a_1\gamma_n a_1^* + b \circ b, \qquad d_{12} = a_1\gamma_n A_1^* + b \circ B, \qquad d_{22} = A_1\gamma_n A_1^* + B \circ B. \tag{11}$$

By the theorem on normal correlation (see Theorem 2 and Problem 4, §13, Chapter II),

$$m_{n+1} = \mathsf{E}(\theta_{n+1}\,|\,\mathscr{F}_n^\xi, \xi_{n+1}) = \mathsf{E}(\theta_{n+1}\,|\,\mathscr{F}_n^\xi) + d_{12}d_{22}^{\oplus}(\xi_{n+1} - \mathsf{E}(\xi_{n+1}\,|\,\mathscr{F}_n^\xi))$$

and

$$\gamma_{n+1} = \operatorname{cov}(\theta_{n-1}, \theta_{n+1}\,|\,\mathscr{F}_n^\xi, \xi_{n+1}) = d_{11} - d_{12}d_{22}^{\oplus}d_{12}^*.$$

If we then use the expressions from (9) for $\mathsf{E}(\theta_{n+1}\,|\,\mathscr{F}_n^\xi)$ and $\mathsf{E}(\xi_{n+1}\,|\,\mathscr{F}_n^\xi)$ and those for d_{11}, d_{12}, d_{22} from (11), we obtain the required recursion formulas (7) and (8).

This completes the proof of the theorem.

Corollary 1. *If the coefficients $a_0(n, \xi), \ldots, B_2(n, \xi)$ in* (1) *are independent of ξ the corresponding method is known as the Kalman–Bucy method, and equations* (7) *and* (8) *for m_n and γ_n describe the* Kalman–Bucy filter. *It is important to observe that in this case the conditional and unconditional error matrices γ_n agree, i.e.*

$$\gamma_n \equiv \mathsf{E}\gamma_n = \mathsf{E}[(\theta_n - m_n)(\theta_n - m_n)^*].$$

Corollary 2. *Suppose that a partially observed sequence (θ_n, ξ_n) has the property that θ_n satisfies the first equation* (1), *and that ξ_n satisfies the equation*

$$\begin{aligned}\xi_n = \tilde{A}_0(n-1, \xi) &+ \tilde{A}_1(n-1, \xi)\theta_n \\ &+ \tilde{B}_1(n-1, \xi)\varepsilon_1(n) + \tilde{B}_2(n-1, \xi)\varepsilon_2(n).\end{aligned} \tag{12}$$

Then evidently

$$\begin{aligned}\xi_{n+1} = \tilde{A}_0(n, \xi) &+ \tilde{A}_1(n, \xi)[a_0(n, \xi) + a_1(n, \xi)\theta_n \\ &+ b_1(n, \xi)\varepsilon_1(n+1) + b_2(n, \xi)\varepsilon_2(n+1)] + \tilde{B}_1(n, \xi)\varepsilon_1(n+1) \\ &+ \tilde{B}_2(n, \xi)\varepsilon_2(n+1),\end{aligned}$$

and with the notation

$$\begin{aligned} A_0 &= A_0 + \tilde{A}_1 a_0, & A_1 &= \tilde{A}_1 a_1, \\ B_1 &= A_1 b_1 + \tilde{B}_1, & B_2 &= \tilde{A}_1 b_2 + \tilde{B}_2, \end{aligned}$$

we find that the case under consideration also depends on the model (1), *and that m_n and γ_n satisfy* (7) *and* (8).

2. We now consider a *linear* model (compare (1))

$$\begin{aligned}\theta_{n+1} &= a_0 + a_1\theta_n + a_2\xi_n + b_1\varepsilon_1(n+1) + b_2\varepsilon_2(n+1), \\ \xi_{n+1} &= A_0 + A_1\theta_n + A_2\xi_n + B_1\varepsilon_1(n+1) + B_2\varepsilon_2(n+1),\end{aligned} \tag{13}$$

where the coefficients $a_0, \ldots, B_n$ may depend on n (but not on ξ), and $\varepsilon_{ij}(n)$ are independent Gaussian random variables with $\mathsf{E}\varepsilon_{ij}(n) = 0$ and $\mathsf{E}\varepsilon_{ij}^2(n) = 1$.

Let (13) be solved for the initial values (θ_0, ξ_0) so that the conditional distribution $\mathsf{P}(\theta_0 \le a \,|\, \xi_0)$ is Gaussian with parameters $m_0 = \mathsf{E}(\theta_0, \xi_0)$ and $\gamma = \mathrm{cov}(\theta_0, \theta_0 \,|\, \xi_0) = \mathsf{E}\gamma_0$. Then, by the theorem on normal correlation and (7) and (8), the optimal estimator $m_n = \mathsf{E}(\theta_n \,|\, \mathscr{F}_n^\xi)$ is a linear function of $\xi_0, \xi_1, \ldots, \xi_n$.

This remark makes it possible to prove the following important statement about the structure of the optimal linear filter without the assumption that it is Gaussian.

Theorem 2. *Let $(\theta, \xi) = (\theta_n, \xi_n)_{n \ge 0}$ be a partially observed sequence that satisfies* (13), *where $\varepsilon_{ij}(n)$ are uncorrelated random variables with $\mathsf{E}\varepsilon_{ij}(n) = 0$, $\mathsf{E}\varepsilon_{ij}^2(n) = 1$, and the components of the initial vector (θ_0, ξ_0) have finite second moments. Then the optimal linear estimator $\hat{m}_n = \hat{E}(\theta_n \,|\, \xi_0, \ldots, \xi_n)$ satisfies* (7) *with $a_0(n, \xi) = a_0(n) + a_2(n)\xi_n$, $A_0(n, \xi) = A_0(n) + A_2(n)\xi_n$, and the error matrix $\hat{\gamma}_n = \mathsf{E}[(\theta_n - \theta_n)(\theta_n - m_n)^*]$ satisfies* (8) *with initial values*

$$\begin{aligned} \hat{m}_0 &= \mathrm{cov}(\theta_0, \xi_0)\mathrm{cov}^{\oplus}(\xi_0, \xi_0) \cdot \xi_0, \\ \hat{\gamma}_0 &= \mathrm{cov}(\theta_0, \theta_0) - \mathrm{cov}(\theta_0, \xi_0)\,\mathrm{cov}^{\oplus}(\xi_0, \xi_0)\mathrm{cov}^*(\theta_0, \xi_0). \end{aligned} \tag{14}$$

For the proof of this lemma, we need the following lemma, which reveals the role of the Gaussian case in determining optimal linear estimators.

Lemma 2. *Let (α, β) be a two-dimensional random vector with $\mathsf{E}(\alpha^2 + \beta^2) < \infty$, $a(\tilde{\alpha}, \tilde{\beta})$ a two-dimensional Gaussian vector with the same first and second moments as (α, β), i.e.*

$$\mathsf{E}\tilde{\alpha}^i = \mathsf{E}\alpha^i, \qquad \mathsf{E}\tilde{\beta}^i = \mathsf{E}\beta^i, \qquad i = 1, 2; \qquad \mathsf{E}\tilde{\alpha}\tilde{\beta} = \mathsf{E}\alpha\beta.$$

Let $\lambda(b)$ be a linear function of b such that

$$\lambda(b) = \mathsf{E}(\tilde{\alpha} \,|\, \tilde{\beta} = b).$$

Then $\lambda(\beta)$ is the optimal (in the mean square sense) linear estimator of α in terms of β, i.e.

$$\hat{\mathsf{E}}(\alpha \,|\, \beta) = \lambda(\beta).$$

Here $\mathsf{E}\lambda(\beta) = \mathsf{E}\alpha$.

PROOF. We first observe that the existence of a linear function $\lambda(b)$ coinciding with $\mathsf{E}(\tilde{\alpha} \,|\, \tilde{\beta} = b)$ follows from the theorem on normal correlation. Moreover, let $\bar{\lambda}(b)$ be any other linear estimator. Then

$$\mathsf{E}[\tilde{\alpha} - \bar{\lambda}(\tilde{\beta})]^2 \ge \mathsf{E}[\tilde{\alpha} - \lambda(\tilde{\beta})]^2$$

and since $\bar{\lambda}(b)$ and $\lambda(b)$ are linear and the hypotheses of the lemma are satisfied, we have

$$\mathsf{E}[\alpha - \bar{\lambda}(\beta)]^2 = \mathsf{E}[\tilde{\alpha} - \bar{\lambda}(\tilde{\beta})]^2 \geq \mathsf{E}[\tilde{\alpha} - \lambda(\tilde{\beta})]^2 = \mathsf{E}[\alpha - \lambda(\beta)]^2,$$

which shows that $\lambda(\beta)$ is optimal in the class of linear estimators. Finally,

$$\mathsf{E}\lambda(\beta) = \mathsf{E}\lambda(\tilde{\beta}) = \mathsf{E}[\mathsf{E}(\tilde{\alpha}|\tilde{\beta})] = \mathsf{E}\tilde{\alpha} = \mathsf{E}\alpha.$$

This completes the proof of the lemma.

Proof of Theorem 2. We consider, besides (13), the system

$$\begin{aligned} \tilde{\theta}_{n+1} &= a_0 + a_1\tilde{\theta}_n + a_2\tilde{\xi}_n + b_1\tilde{\varepsilon}_{11}(n+1) + b_2\tilde{\varepsilon}_{12}(n+1), \\ \tilde{\xi}_{n+1} &= A_0 + A_1\tilde{\theta}_n + A_2\tilde{\xi}_n + B_1\tilde{\varepsilon}_{21}(n+1) + B_2\tilde{\varepsilon}_{22}(n+1), \end{aligned} \tag{15}$$

where $\tilde{\varepsilon}_{ij}(n)$ are independent Gaussian random variables with $\mathsf{E}\tilde{\varepsilon}_{ij}(n) = 0$ and $\mathsf{E}\tilde{\varepsilon}_{ij}^2(n) = 1$. Let $(\tilde{\theta}_0, \tilde{\xi}_0)$ also be a Gaussian vector which has the same first moment and covariance as (θ_0, ξ_0) and is independent of $\tilde{\varepsilon}_{ij}(n)$. Then since (15) is linear, the vector $(\tilde{\theta}_0, \ldots, \tilde{\theta}_n, \tilde{\xi}_0, \ldots, \tilde{\xi}_n)$ is Gaussian and therefore the conclusion of the theorem follows from Lemma 2 (more precisely, from its multidimensional analog) and the theorem on normal covariance.

This completes the proof of the theorem.

3. Let us consider some illustrations of Theorems 1 and 2.

Example 1. Let $\theta = (\theta_n)$ and $\eta = (\eta_n)$ be two stationary (wide sense) uncorrelated random sequences with $\mathsf{E}\theta_n = \mathsf{E}\eta_n = 0$ and spectral densities

$$f_\theta(\lambda) = \frac{1}{2\pi|1 + b_1e^{-i\lambda}|^2} \quad \text{and} \quad f_\eta(\lambda) = \frac{1}{2\pi}\cdot\frac{1}{|1 + b_2e^{-i\lambda}|^2},$$

where $|b_1| < 1, |b_2| < 1$.

We are going to interpret θ as a useful signal and η as noise, and suppose that observation produces a sequence $\xi = (\xi_n)$ with

$$\xi_n = \theta_n + \eta_n.$$

According to Corollary 2 to Theorem 3 of §3 there are (mutually uncorrelated) white noises $\varepsilon_1 = (\varepsilon_1(n))$ and $\varepsilon_2 = (\varepsilon_2(n))$ such that

$$\theta_{n+1} + b_1\theta_n = \varepsilon_1(n+1), \qquad \eta_{n+1} + b_2\eta_n = \varepsilon_2(n+1).$$

Then

$$\begin{aligned} \xi_{n+1} &= \theta_{n+1} + \eta_{n+1} = -b_1\theta_n - b_2\eta_n + \varepsilon_1(n+1) + \varepsilon_2(n+1) \\ &= -b_2(\theta_n + \eta_n) - \theta_n(b_1 - b_2) + \varepsilon_1(n+1) + \varepsilon_2(n+1) \\ &= -b_2\xi_n - (b_1 - b_2)\theta_n + \varepsilon_1(n+1) + \varepsilon_2(n+1). \end{aligned}$$

Hence θ and ξ satisfy the recursion relations

$$\begin{aligned}\theta_{n+1} &= -b_1\theta_n + \varepsilon_1(n+1),\\ \xi_{n+1} &= -(b_1 - b_2)\theta_n - b_2\xi_n + \varepsilon_1(n+1) + \varepsilon_2(n+1),\end{aligned}\tag{16}$$

and, according to Theorem 2, $m_n = \hat{\mathsf{E}}(\theta_n | \xi_0, \ldots, \xi_n)$ and $\gamma_n = \mathsf{E}(\theta_n - m_n)^2$ satisfy the following system of recursion equations for optimal linear filtering:

$$\begin{aligned}m_{n+1} &= -b_1 m_n + \frac{b_1(b_1 - b_2)\gamma_n}{2 + (b_1 - b_2)^2\gamma_n}[\xi_{n+1} + (b_1 - b_2)m_n + b_2\xi_n],\\ \gamma_{n+1} &= b_1^2\gamma_n + 1 - \frac{[1 + b_1(b_1 - b_2)\gamma_n]^2}{2 + (b_1 - b_2)^2\gamma_n}.\end{aligned}\tag{17}$$

Let us find the initial conditions under which we should solve this system. Write $d_{11} = \mathsf{E}\theta_n^2$, $d_{12} = \mathsf{E}\theta_n\xi_n$, $d_{22} = \mathsf{E}\xi_n^2$. Then we find from (16) that

$$\begin{aligned}d_{11} &= b_1^2 d_{11} + 1,\\ d_{12} &= b_1(b_1 - b_2)d_{11} + b_1 b_2 d_{12} + 1,\\ d_{22} &= (b_1 - b_2)^2 d_{11} + b_2^2 d_{22} + 2b_2(b_1 - b_2)d_{12} + 2,\end{aligned}$$

from which

$$d_{11} = \frac{1}{1 - b_1^2}, \qquad d_{12} = \frac{1}{1 - b_1^2}, \qquad d_{22} = \frac{2 - b_1^2 - b_2^2}{(1 - b_1^2)(1 - b_2^2)},$$

which, by (14), leads to the following initial values:

$$\begin{aligned}m_0 &= \frac{d_{12}}{d_{22}}\xi_0 = \frac{1 - b_2^2}{2 - b_1^2 - b_2^2}\xi_0,\\ \gamma_0 &= d_{11} - \frac{d_{12}^2}{d_{22}} = \frac{1}{1 - b_1^2} - \frac{1 - b_2^2}{(1 - b_1^2)(2 - b_1^2 - b_2^2)} = \frac{1}{2 - b_1^2 - b_2^2}.\end{aligned}\tag{18}$$

Thus the optimal (in the least squares sense) linear estimators m_n for the signal θ_n in terms of $\xi_0, \ldots, \xi_n$ and the mean-square error are determined by the system of recurrent equations (17), solved under the initial conditions (18). Observe that the equation for γ_n does not contain any random components, and consequently the number γ_n, which is needed for finding m_n, can be calculated in advance, before the filtering problem has been solved.

Example 2. This example is instructive because it shows that the result of Theorem 2 can be applied to find the optimal linear filter in a case where the sequence (θ, ξ) is described by a (nonlinear) system which is different from (13).

Let $\varepsilon_1 = (\varepsilon_1(n))$ and $\varepsilon_2 = (\varepsilon_2(n))$ be two independent Gaussian sequences of independent random variables with $\mathsf{E}\varepsilon_i(n) = 0$ and $\mathsf{E}\varepsilon_i^2(n) = 1$, $n \geq 1$. Consider a pair of sequences $(\theta, \xi) = (\theta_n, \xi_n)$, $n \geq 0$, with

$$\begin{aligned} \theta_{n+1} &= a\theta_n + (1 + \theta_n)\varepsilon_1(n+1), \\ \xi_{n+1} &= A\theta_n + \varepsilon_2(n+1). \end{aligned} \tag{19}$$

We shall suppose that θ_0 is independent of $(\varepsilon_1, \varepsilon_2)$ and that $\theta_0 \sim \mathcal{N}(m_0, \gamma_0)$.

The system (19) is *nonlinear*, and Theorem 2 is not immediately applicable. However, if we put

$$\tilde{\varepsilon}_1(n+1) = \frac{1 + \theta_n}{\sqrt{\mathsf{E}(1 + \theta_n)^2}}\, \varepsilon_1(n+1),$$

we can observe that $\mathsf{E}\tilde{\varepsilon}_1(n) = 0$, $\mathsf{E}\tilde{\varepsilon}_1(n)\tilde{\varepsilon}_1(m) = 0$, $n \neq m$, $\hat{\mathsf{E}}\varepsilon_1^2(n) = 1$. Hence we have reduced (19) to a linear system

$$\begin{aligned} \theta_{n+1} &= a_1\theta_n + b_1\tilde{\varepsilon}_1(n+1), \\ \xi_{n+1} &= A_1\theta_n + \varepsilon_2(n+1), \end{aligned} \tag{20}$$

where $b_1 = \sqrt{\mathsf{E}(1 + \theta_n)^2}$, and $\{\tilde{\varepsilon}_1(n)\}$ is a sequence of uncorrelated random variables.

Now (20) is a linear system of the same type as (13), and consequently the optimal linear estimator $\hat{m}_n = \hat{\mathsf{E}}(\theta_n | \xi_0, \ldots, \xi_n)$ and its error $\hat{\gamma}_n$ can be determined from (7) and (8) via Theorem 2, applied in the following form in the present case:

$$m_{n+1} = a_1 m_n + \frac{a_1 A_1 \gamma_n}{1 + A_1^2\gamma_n}[\xi_{n+1} - A_1 m_n],$$

$$\gamma_{n+1} = (a_1^2\gamma_n + b_1^2) - \frac{(a_1 A_1 \gamma_n)^2}{1 + A_1^2\gamma_n},$$

where $b_1 = \sqrt{\mathsf{E}(1 + \theta_n)^2}$ must be found from the first equation in (19).

Example 3. *Estimators for parameters.* Let $\theta = (\theta_1, \ldots, \theta_k)$ be a Gaussian vector with $\mathsf{E}\theta = m$ and $\operatorname{cov}(\theta, \theta) = \gamma$. Suppose that (with known m and ν) we want the optimal estimator of θ in terms of observations on an l-dimensional sequence $\xi = (\xi_n)$, $n \geq 0$, with

$$\xi_{n+1} = A_0(n, \xi) + A_1(n, \xi)\theta + B_1(n, \xi)\varepsilon_1(n+1), \qquad \xi_0 = 0, \tag{21}$$

where ε_1 is as in (1).

Then from (7) and (8), with $m_n = \mathsf{E}(\theta | \mathscr{F}_n^{\xi})$ and γ_n, we find that

$$\begin{aligned} m_{n+1} &= m_n + \gamma_n A_1^*(n, \xi)[(B_1B_1^*)(n, \xi) + A_1(n, \xi)\gamma_n A_1^*(n, \xi)]^{\oplus} \\ &\quad \times [\xi_{n+1} - A_0(n, \xi) - A_1(n, \xi)m_n], \\ \gamma_{n+1} &= \gamma_n - \gamma_n A_1^*(n, \xi)[(B_1B_1^*)(n, \xi) + A_1(n, \xi)\gamma_n A_1^*(n, \xi)]^{\oplus} A_1(n, \xi)\gamma_n \end{aligned} \tag{22}$$

If the matrices $B_1 B_1^*$ are nonsingular, the solution of (22) is given by

$$m_{n+1} = \left[E + \gamma \sum_{m=0}^{n} A_1^*(m, \xi)(B_1 B_1^*)^{-1}(m, \xi) A_1^*(m, \xi)\right]^{-1}$$

$$\times \left[m + \gamma \sum_{m=0}^{n} A_1^*(m, \xi)(B_1 B_1^*)^{-1}(m, \xi)(\xi_{m+1} - A_0(m, \xi))\right],$$

$$\gamma_{n+1} = \left[E + \gamma \sum_{m=0}^{n} A_1^*(m, \xi)(B_1 B_1^*)^{-1}(m, \xi) A_1(m, \xi)\right]^{-1} \gamma, \qquad (23)$$

where E is a unit matrix.

4. Problems

1. Show that the vectors m_n and $\theta_n - m_n$ in (1) are uncorrelated:

$$\mathsf{E}[m_n^*(\theta - m_n)] = 0.$$

2. In (1), let γ and the coefficients other than $a_0(n, \xi)$ and $A_0(n, \xi)$ be independent of "chance" (i.e. of ξ). Show that then the conditional covariance γ_n is independent of "chance": $\gamma_n = \mathsf{E}\gamma_n$.

3. Show that the solution of (22) is given by (23).

4. Let $(\theta, \xi) = (\theta_n, \xi_n)$ be a Gaussian sequence satisfying the following special case of (1):

$$\theta_{n+1} = a\theta_n + b\varepsilon_1(n+1), \qquad \xi_{n+1} = A\theta_n + B\varepsilon_2(n+1).$$

Show that if $A \neq 0$, $b \neq 0$, $B \neq 0$, the limiting error of filtering, $\gamma = \lim_{n\to\infty} \gamma_n$, exists and is determined as the positive root of the equation

$$\gamma^2 + \left[\frac{B^2(1 - a^2)}{A^2} - b^2\right]\gamma - \frac{b^2 B^2}{A^2} = 0.$$

CHAPTER VII

Sequences of Random Variables That Form Martingales

§1. Definitions of Martingales and Related Concepts

1. The study of the dependence of random variables arises in various ways in probability theory. In the theory of stationary (wide sense) random sequences, the basic indicator of dependence is the covariance function, and the inferences made in this theory are determined by the properties of that function. In the theory of Markov chains (§12 of Chapter I; Chapter VIII) the basic dependence is supplied by the transition function, which completely determines the development of the random variables involved in Markov dependence.

In the present chapter (see also §11, Chapter I), we single out a rather wide class of sequences of random variables (martingales and their generalizations) for which dependence can be studied by methods based on a discussion of the properties of conditional expectations.

2. Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a given probability space, and let $(\mathscr{F}_n)$ be a family of σ-algebras $\mathscr{F}_n$, $n \geq 0$, such that $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \cdots \subseteq \mathscr{F}$.

Let $X_0, X_1, \ldots$ be a sequence of random variables defined on $(\Omega, \mathscr{F}, \mathsf{P})$. If, for each $n \geq 0$, the variable X_n is *$\mathscr{F}_n$-measurable*, we say that the set $X = (X_n, \mathscr{F}_n)$, $n \geq 0$, or simply $X = (X_n, \mathscr{F}_n)$, is a *stochastic sequence*.

If a stochastic sequence $X = (X_n, \mathscr{F}_n)$ has the property that, for each $n \geq 1$, the variable X_n is *$\mathscr{F}_{n-1}$-measurable*, we write $X = (X_n, \mathscr{F}_{n-1})$, taking $F_{-1} = F_0$, and call X a *predictable sequence*. We call such a sequence *increasing* if $X_0 = 0$ and $X_n \leq X_{n+1}$ (P-a.s.).

Definition 1. A stochastic sequence $X = (X_n, \mathscr{F}_n)$ is a *martingale*, or a *submartingale*, if, for all $n \geq 0$,

$$\mathsf{E}|X_n| < \infty \tag{1}$$

and, respectively,

$$\mathsf{E}(X_{n+1}|\mathscr{F}_n) = X_n \quad (\text{P-a.s.}) \text{ (martingale)}$$

or (2)

$$\mathsf{E}(X_{n+1}|\mathscr{F}_n) \geq X_n \quad (\text{P-a.s.}) \text{ (submartingale)}.$$

A stochastic sequence $X = (X_n \mathscr{F}_n)$ is a *supermartingale* if the sequence $-X = (-X_n, \mathscr{F}_n)$ is a submartingale.

In the special case when $\mathscr{F}_n = \mathscr{F}_n^X$, where $\mathscr{F}_n^X = \sigma\{\omega: X_0, \ldots, X_n\}$, and the stochastic sequence $X = (X_n, \mathscr{F}_n)$ is a martingale (or submartingale), we say that the sequence $(X_n)_{n\geq 0}$ itself is a martingale (or submartingale).

It is easy to deduce from the properties of conditional expectations that (2) is equivalent to the property that, for every $n \geq 0$ and $A \in \mathscr{F}_n$,

$$\int_A X_{n+1}\, d\,\mathsf{P} = \int_A X_n\, d\mathsf{P}$$

or (3)

$$\int_A X_{n+1}\, d\mathsf{P} \geq \int_A X_n\, d\,\mathsf{P}.$$

EXAMPLE 1. If $(\xi_n)_{n\geq 0}$ is a sequence of independent random variables with $\mathsf{E}\xi_n = 0$ and $X_n = \xi_0 + \cdots + \xi_n$, $\mathscr{F}_n = \sigma\{\omega: \xi_0, \ldots, \xi_n\}$, the stochastic sequence $X = \{X_n, \mathscr{F}_n)$ is a martingale.

EXAMPLE 2. If $(\xi_n)_{n\geq 0}$ is a sequence of independent random variables with $\mathsf{E}\xi_n = 1$, the stochastic sequence $(X_n, \mathscr{F}_n)$ with $X_n = \prod_{k=0}^n \xi_k$, $\mathscr{F}_n = \sigma\{\omega: \xi_0, \ldots, \xi_n\}$ is also a martingale.

EXAMPLE 3. Let ξ be a random variable with $\mathsf{E}|\xi| < \infty$ and

$$\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \cdots \subseteq \mathscr{F}.$$

Then the sequence $X = (X_n, \mathscr{F}_n)$ with $X_n = \mathsf{E}(\xi|\mathscr{F}_n)$ is a martingale.

EXAMPLE 4. If $(\xi_n)_{n\geq 0}$ is a sequence of nonnegative integrable random variables, the sequence (X_n) with $X_n = \xi_0 + \cdots + \xi_n$ is a submartingale.

EXAMPLE 5. If $X = (X_n, \mathscr{F}_n)$ is a martingale and $g(x)$ is convex downward with $\mathsf{E}|g(X_n)| < \infty$, $n \geq 0$, then the stochastic sequence $(g(X_n), \mathscr{F}_n)$ is a submartingale (as follows from Jensen's inequality).

If $X = (X_n, \mathscr{F}_n)$ is a submartingale and $g(x)$ is convex downward and nondecreasing, with $\mathsf{E}|g(X_n)| < \infty$ for all $n \geq 0$, then $(g(X_n), \mathscr{F}_n)$ is also a submartingale.

Assumption (1) in Definition 1 ensures the existence of the conditional expectations $\mathsf{E}(X_{n+1}|\mathscr{F}_n)$, $n \geq 0$. However, these expectations can also exist without the assumption that $\mathsf{E}|X_{n+1}| < \infty$. Recall that by §7 of Chapter

II, $\mathsf{E}(X_{n+1}^+|\mathscr{F}_n)$ and $\mathsf{E}(X_{n+1}^-|\mathscr{F}_n)$ are always defined. Let us write $A = B$ (P-a.s.) when $\mathsf{P}(A \triangle B) = 0$. Then if

$$\{\omega: \mathsf{E}(X_{n+1}^+|\mathscr{F}_n) < \infty\} \cup \{\omega: \mathsf{E}(X_n^-|\mathscr{F}_n) < \infty\} = \Omega \quad (\text{P-a.s.})$$

we say that $\mathsf{E}(X_{n+1}|\mathscr{F}_n)$ is also defined and is given by

$$\mathsf{E}(X_{n+1}|\mathscr{F}_n) = \mathsf{E}(X_{n+1}^+|\mathscr{F}_n) - \mathsf{E}(X_{n+1}^-|\mathscr{F}_n).$$

After this, the following definition is natural.

Definition 2. A stochastic sequence $X = (X_n, \mathscr{F}_n)$ is a *generalized martingale* (or *submartingale*) if the conditional expectations $\mathsf{E}(X_{n+1}|\mathscr{F}_n)$ are defined for every $n \geq 0$ and (2) is satisfied.

Notice that it follows from this definition that $\mathsf{E}(X_{n+1}^-|\mathscr{F}_n) < \infty$for a generalized submartingale, and the $\mathsf{E}(|X_{n+1}||\mathscr{F}_n) < \infty$ (P-a.s.) for a generalized martingale.

3. In the following definition we introduce the concept of a Markov time, which plays a very important role in the subsequent theory.

Definition 3. A random variable $\tau = \tau(\omega)$ with values in the set $\{0, 1, \ldots, +\infty\}$ is a *Markov time* (with respect to $(\mathscr{F}_n)$) (or a *random variable independent of the future*) if, for each $n \geq 0$,

$$\{\tau = n\} \in \mathscr{F}_n. \tag{4}$$

When $\mathsf{P}(\tau < \infty) = 1$, a Markov time τ is called a *stopping time*.

Let $X = (X_n, \mathscr{F}_n)$ be a stochastic sequence and let τ be a Markov time (with respect to $(\mathscr{F}_n)$). We write

$$X_\tau = \sum_{n=0}^{\infty} X_n I_{\{\tau=n\}}(\omega)$$

(hence $X_\tau = 0$ on the set $\{\omega: \tau = \infty\}$).

Then for every $B \in \mathscr{B}(R)$,

$$\{\omega: X_\tau \in B\} = \sum_{n=0}^{\infty} \{X_n \in B, \tau = n\} \in \mathscr{F},$$

and consequently X_τ is a random variable.

EXAMPLE 6. Let $X = (X_n, \mathscr{F}_n)$ be a stochastic sequence and let $B \in \mathscr{B}(R)$. Then the time of first hitting the set B, that is,

$$\tau_B = \inf\{n \geq 0: X_n \in B\}$$

(with $\tau_B = +\infty$ if $\{\cdot\} = \varnothing$) is a Markov time, since

$$\{\tau_B = n\} = \{X_0 \notin B, \dots, X_{n-1} \notin B, X_n \in B\} \in \mathscr{F}_n$$

for every $n \geq 0$.

EXAMPLE 7. Let $X = (X_n, \mathscr{F}_n)$ be a martingale (or submartingale) and τ a Markov time (with respect to $(\mathscr{F}_n)$). Then the "stopped" process $X^\tau = (X_{n \wedge \tau}, \mathscr{F}_n)$ is also a martingale (or submartingale).

In fact, the equation

$$X_{n \wedge \tau} = \sum_{m=0}^{n-1} X_m I_{\{\tau = m\}} + X_n I_{\{\tau \geq n\}}$$

implies that the variables $X_{n \wedge \tau}$ are $\mathscr{F}_n$-measurable, are integrable, and satisfy

$$X_{(n+1) \wedge \tau} - X_{n \wedge \tau} = I_{\{\tau > n\}}(X_{n+1} - X_n),$$

whence

$$\mathsf{E}[X_{(n+1) \wedge \tau} - X_{n \wedge \tau} | \mathscr{F}_n] = I_{\{\tau > n\}} \mathsf{E}[X_{n+1} - X_n | \mathscr{F}_n] = 0 \quad (\text{or} \geq 0).$$

Every system $(\mathscr{F}_n)$ and Markov time τ corresponding to it generate a collection of sets

$$\mathscr{F}_\tau = \{A \in \mathscr{F} : A \cap \{\tau = n\} \in \mathscr{F}_n \text{ for all } n \geq 0\}.$$

It is clear that $\Omega \in \mathscr{F}_\tau$ and $\mathscr{F}_\tau$ is closed under countable unions. Moreover, if $A \in \mathscr{F}_\tau$, then $\bar{A} \cap \{\tau = n\} = \{\tau = n\} \setminus (A \cap \{\tau = n\}) \in \mathscr{F}_n$ and therefore $\bar{A} \in \mathscr{F}_\tau$. Hence it follows that $\mathscr{F}_\tau$ is a σ-algebra.

If we think of $\mathscr{F}_n$ as a collection of events observed up to time n (inclusive), then $\mathscr{F}_\tau$ can be thought of as a collection of events observed at the "random" time τ.

It is easy to show (Problem 3) that the random variables τ and X_τ are $\mathscr{F}_\tau$-measurable.

4. Definition 4. A stochastic sequence $X = (X_n, \mathscr{F}_n)$ is a *local martingale* (or *submartingale*) if there is a (localizing) sequence $(\tau_k)_{k \geq 1}$ of Markov times such that $\tau_k \leq \tau_{k+1}$ (P-a.s.), $\tau_k \uparrow \infty$ (P-a.s.) as $k \to \infty$, and every "stopped" sequence $X^{\tau_k} = (X_{\tau_k \wedge n} \cdot I_{\{\tau_k > 0\}}, \mathscr{F}_n)$ is a martingale (or submartingale).

In Theorem 1 below, we show that in fact the class of local martingales coincides with the class of generalized martingales. Moreover, every local martingale can be obtained by a "martingale transformation" from a martingale and a predictable sequence.

Definition 5. Let $Y = (Y_n, \mathscr{F}_n)$ be a stochastic sequence and let $V = (V_n, \mathscr{F}_{n-1})$ be a predictable sequence ($\mathscr{F}_{-1} = \mathscr{F}_0$). The stochastic sequence $V \cdot Y = ((V \cdot Y)_n, \mathscr{F}_n)$ with

$$(V \cdot Y)_n = V_0 Y_0 + \sum_{i=1}^{n} V_i \Delta Y_i, \tag{5}$$

where $\Delta Y_i = Y_i - Y_{i-1}$, is called the *transform of* Y *by* V. If, in addition, Y is a martingale, we say that $V \cdot Y$ is a *martingale transform*.

Theorem 1. *Let $X = (X_n, \mathscr{F}_n)_{n \geq 0}$ be a stochastic sequence and let $X_0 = 0$ (P-a.s.). The following conditions are equivalent:*

(a) *X is a local martingale;*
(b) *X is a generalized martingale;*
(c) *X is a martingale transform, i.e., there are a predictable sequence $V = (V_n, \mathscr{F}_{n-1})$ with $V_0 = 0$ and a martingale $Y = (Y_n, \mathscr{F}_n)$ with $Y_0 = 0$ such that $X = V \cdot Y$.*

PROOF. (a) $\Rightarrow$ (b). Let X be a local martingale and let (τ_k) be a local sequence of Markov times for X. Then for every $m \geq 0$

$$\mathsf{E}[|X_{m \wedge \tau_k}| I_{\{\tau_k > 0\}}] < \infty, \tag{6}$$

and therefore

$$\mathsf{E}[|X_{(n+1) \wedge \tau_k}| I_{\{\tau_k > n\}}] = \mathsf{E}[|X_{n+1}| I_{\{\tau_k > n\}}] < \infty. \tag{7}$$

The random variable $I_{\{\tau_k > n\}}$ is $\mathscr{F}_n$-measurable. Hence it follows from (7) that

$$\mathsf{E}[|X_{n+1}| I_{\{\tau_k > n\}} | \mathscr{F}_n] = I_{\{\tau_k > n\}} \mathsf{E}[|X_{n+1}| | \mathscr{F}_n] < \infty \quad \text{(P-a.s.)}.$$

Here $I_{\{\tau_k > n\}} \to 1$ (P-a.s.), $k \to \infty$, and therefore

$$\mathsf{E}[|X_{n+1}| | \mathscr{F}_n] < \infty \quad \text{(P-a.s.)}. \tag{8}$$

Under this condition, $\mathsf{E}[X_{n+1} | \mathscr{F}_n]$ is defined, and it remains only to show that $\mathsf{E}[X_{n+1} | \mathscr{F}_n] = X_n$ (P-a.s.).

To do this, we need to show that

$$\int_A X_{n+1}\, d\mathsf{P} = \int_A X_n\, d\mathsf{P}$$

for $A \in \mathscr{F}_n$. By Problem 7, §7, Chapter II, we have $\mathsf{E}[|X_{n+1}| | \mathscr{F}_n] < \infty$ (P-a.s.) if and only if the measure $\int_A |X_{n+1}|\, d\mathsf{P}$, $A \in \mathscr{F}_n$, is σ-finite. Let us show that the measure $\int_A |X_n|\, d\mathsf{P}$, $A \in \mathscr{F}_n$, is also σ-finite.

Since X^{τ_k} is a martingale, $|X^{\tau_k}| = (|X_{\tau_k \wedge n}| I_{\{\tau_k > 0\}}, \mathscr{F}_n)$ is a submartingale, and therefore (since $\{\tau_k > n\} \in \mathscr{F}_n$)

$$\int_{A\cap\{\tau_k>n\}} |X_n|\,d\mathsf{P} = \int_{A\cap\{\tau_k>n\}} |X_{n\wedge\tau_k}| I_{\{\tau_k>0\}}\,d\mathsf{P}$$

$$\leq \int_{A\cap\{\tau_k>n\}} |X_{(n+1)\wedge\tau_k}| I_{\{\tau_k>0\}}\,d\mathsf{P} = \int_{A\cap\{\tau_k>n\}} |X_{n+1}|\,d\mathsf{P}.$$

Letting $k \to \infty$, we have

$$\int_A |X_n|\,d\mathsf{P} \leq \int_A |X_{n+1}|\,d\mathsf{P},$$

from which there follows the required σ-finiteness of the measures $\int_A |X_n|\,d\mathsf{P}$, $A \in \mathscr{F}_n$.

Let $A \in \mathscr{F}_n$ have the property $\int_A |X_{n+1}|\,d\mathsf{P} < \infty$. Then, by Lebesgue's theorem on dominated convergence, we may take limits in the relation

$$\int_{A\cap\{\tau_k>n\}} X_n\,d\mathsf{P} = \int_{A\cap\{\tau_k>n\}} X_{n+1}\,d\mathsf{P},$$

which is valid since X is a local martingale. Therefore,

$$\int_A X_n\,d\mathsf{P} = \int_A X_{n+1}\,d\mathsf{P}$$

for all $A \in \mathscr{F}_n$ such that $\int_A |X_{n+1}|\,d\mathsf{P} < \infty$. It then follows that the preceding relation also holds for every $A \in \mathscr{F}_n$, and therefore, $\mathsf{E}(X_{n+1}|\mathscr{F}_n) = X_n$ (P-a.s.).

b) $\Rightarrow$ c). Let $\Delta X_n = X_n - X_{n-1}$, $X_0 = 0$, and $V_0 = 0$, $V_n = \mathsf{E}[|\Delta X_n||\mathscr{F}_{n-1}]$, $n \geq 1$. We set

$$W_n = V_n^{\oplus}\left(= \begin{cases} V_n^{-1}, & V_n \neq 0 \\ 0, & V_n = 0 \end{cases}\right), \qquad Y_0 = 0$$

and $Y_n = \sum_{i=1}^n W_i \Delta X_i$, $n \geq 1$. It is clear that

$$\mathsf{E}[|\Delta Y_n||\mathscr{F}_{n-1}] \leq 1, \qquad \mathsf{E}[\Delta Y_n|\mathscr{F}_{n-1}] = 0,$$

and consequently, $Y = (Y_n, \mathscr{F}_n)$ is a martingale.

Consequently, $Y = (Y_n, \mathscr{F}_n)$ is a martingale. Moreover, $X_0 = V_0 \cdot Y_0 = 0$ and $\Delta(V \cdot Y)_n = \Delta X_n$. Therefore

$$X = V \cdot Y.$$

(c) $\Rightarrow$ (a). Let $X = V \cdot Y$ where V is a predictable sequence, Y is a martingale and $V_0 = Y_0 = 0$. Put

$$\tau_k = \inf\{n \geq 0 : |V_{n+1}| > k\},$$

and suppose that $\tau_k = \infty$ if the set $\{\cdot\} = \varnothing$. Since V_{n+1} is $\mathscr{F}_n$-measurable, the variables τ_k are Markov times for every $k \geq 1$.

Consider a "stopped" sequence $X^{\tau_k} = ((V \cdot Y)_{n\wedge\tau_k} I_{\{\tau_k>0\}}, \mathscr{F}_n)$. On the set $\{\tau_k > 0\}$, the inequality $|V_{n\wedge\tau_k}| \leq k$ is in effect. Hence it follows that $\mathsf{E}|(V \cdot Y)_{n\wedge\tau_k} I_{\{\tau_k>0\}}| < \infty$ for every $n \geq 1$. In addition, for $n \geq 1$,

$$\mathsf{E}\{[(V\cdot Y)_{(n+1)\wedge\tau_k} - (V\cdot Y)_{n\wedge\tau_k}]I_{\{\tau_k>0\}}|\mathscr{F}_n\}$$
$$= I_{\{\tau_k>0\}}\cdot V_{(n+1)\wedge\tau_k}\cdot\mathsf{E}\{Y_{(n+1)\wedge\tau_k} - Y_{n\wedge\tau_k}|\mathscr{F}_n\} = 0$$

since (see Example 7) $\mathsf{E}\{Y_{(n+1)\wedge\tau_k} - Y_{n\wedge\tau_k}|\mathscr{F}_n\} = 0$.

Thus for every $k \geq 1$ the stochastic sequences X^{τ_k} are martingales, $\tau_k \uparrow \infty$ (P-a.s.), and consequently X is a local martingale.

This completes the proof of the theorem.

5. EXAMPLE 8. Let $(\eta_n)_{n\geq 1}$ be a sequence of independent identically distributed Bernoulli random variables and let $\mathsf{P}(\eta_n = 1) = p$, $\mathsf{P}(\eta_n = -1) = q$, $p + q = 1$. We interpret the event $\{\eta_n = 1\}$ as success (gain) and $\{\eta_n = -1\}$ as failure (loss) of a player at the nth turn. Let us suppose that the player's stake at the nth turn is V_n. Then the player's total gain through the nth turn is

$$X_n = \sum_{i=1}^{n} V_i\eta_i = X_{n-1} + V_n\eta_n, \qquad X_0 = 0.$$

It is quite natural to suppose that the amount V_n at the nth turn may depend on the results of the preceding turns, i.e., on $V_1, \ldots, V_{n-1}$ and on $\eta_1, \ldots, \eta_{n-1}$. In other words, if we put $F_0 = \{\varnothing, \Omega\}$ and $F_n = \sigma\{\omega: \eta_1, \ldots, \eta_n\}$, then V_n is an $\mathscr{F}_{n-1}$-measurable random variable, i.e., the sequence $V = (V_n, \mathscr{F}_{n-1})$ that determines the player's "strategy" is predictable. Putting $Y_n = \eta_1 + \cdots + \eta_n$, we find that

$$X_n = \sum_{i=1}^{n} V_i\Delta Y_i,$$

i.e., the sequence $X = (X_n, \mathscr{F}_n)$ with $X_0 = 0$ is the transform of Y by V.

From the player's point of view, the game in question is *fair* (or *favorable*, or *unfavorable*) if, at every stage, the conditional expectation

$$\mathsf{E}(X_{n+1} - X_n|\mathscr{F}_n) = 0 \text{ (or } \geq 0 \text{ or } \leq 0).$$

Moreover, it is clear that the game is

fair if $p = q = \frac{1}{2}$,
favorable if $p > q$,
unfavorable, if $p < q$.

Since $X = (X_n, \mathscr{F}_n)$ is a

martingale if $p = q = \frac{1}{2}$,
submartingale if $p > q$,
supermartingale if $p < q$,

we can say that the assumption that the game is fair (or favorable, or unfavorable) corresponds to the assumption that the sequence X is a martingale (or submartingale, or supermartingale).

Let us now consider the special class of strategies $V = (V_n, \mathscr{F}_{n-1})_{n\geq 1}$ with $V_1 = 1$ and (for $n > 1$)

$$V_n = \begin{cases} 2^{n-1} & \text{if } \eta_1 = -1, \ldots, \eta_{n-1} = -1, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

In such a strategy, a player, having started with a stake $V_1 = 1$, doubles the stake after a loss and drops out of the game immediately after a win.

If $\eta_1 = -1, \ldots, \eta_n = -1$, the total loss to the player after n turns will be

$$\sum_{i=1}^{n} 2^{i-1} = 2^n - 1.$$

Therefore if also $\eta_{n+1} = 1$, we have

$$X_{n+1} = X_n + V_{n+1} = -(2^n - 1) + 2^n = 1.$$

Let $\tau = \inf\{n \geq 1 : X_n = 1\}$. If $p = q = \frac{1}{2}$, i.e., the game in question is fair, then $\mathsf{P}(\tau = n) = (\frac{1}{2})^n$, $\mathsf{P}(\tau < \infty) = 1$, $\mathsf{P}(X_\tau = 1) = 1$, and $\mathsf{E}X_\tau = 1$. Therefore even for a fair game, by applying the strategy (9), a player can in a finite time (with probability unity) complete the game "successfully," increasing his capital by one unit ($\mathsf{E}X_\tau = 1 > X_0 = 0$).

In gambling practice, this system (doubling the stakes after a loss and dropping out of the game after a win) is called a martingale. This is the origin of the mathematical term "martingale."

Remark. When $p = q = \frac{1}{2}$, the sequence $X = (X_n, \mathscr{F}_n)$ with $X_0 = 0$ is a martingale and therefore

$$\mathsf{E}X_n = \mathsf{E}X_0 = 0 \qquad \text{for every} \quad n \geq 1.$$

We may therefore expect that this equation is preserved if the instant n is replaced by a random instant τ. It will appear later (Theorem 1, §2) that $\mathsf{E}X_\tau = \mathsf{E}X_0$ in "typical" situations. Violations of this equation (as in the game discussed above) arise in what we may describe as physically unrealizable situations, when either τ or $|X_n|$ takes values that are much too large. (Note that the game discussed above would be physically unrealizable, since it supposes an unbounded time for playing and an unbounded initial capital for the player.)

6. Definition 6. A stochastic sequence $\xi = (\xi_n, \mathscr{F}_n)$ is a *martingale-difference* if $\mathsf{E}|\xi| < \infty$ for all $n \geq 0$ and

$$\mathsf{E}(\xi_{n+1} | \mathscr{F}_n) = 0 \quad (\mathsf{P}\text{-a.s.}). \tag{10}$$

The connection between martingales and martingale-differences is clear from Definitions 1 and 6. Thus if $X = (X_n, \mathscr{F}_n)$ is a martingale, then $\xi = (\xi_n, \mathscr{F}_n)$ with $\xi_0 = X_0$ and $\xi_n = \Delta X_n$, $n \geq 1$, is a martingale-difference. In turn, if $\xi = (\xi_n, \mathscr{F}_n)$ is a martingale-difference, then $X = (X_n, \mathscr{F}_n)$ with $X_n = \xi_0 + \cdots + \xi_n$ is a martingale.

In agreement with this terminology, every sequence $\xi = (\xi_n)_{n\geq 0}$ of independent integrable random variables with $\mathsf{E}\xi_n = 0$ is a martingale-difference (with $\mathscr{F}_n = \sigma\{\omega: \xi_0, \xi_1, \ldots, \xi_n\}$).

7. The following theorem elucidates the structure of submartingales (or supermartingales).

Theorem 2 (Doob). *Let $X = (X_n, \mathscr{F}_n)$ be a submartingale. Then there is a martingale $m = (m_n, \mathscr{F}_n)$ and a predictable increasing sequence $A = (A_n, \mathscr{F}_{n-1})$ such that, for every $n \geq 0$, Doob's decomposition*

$$X_n = m_n + A_n \quad (\text{P-a.s.}) \tag{11}$$

holds. A decomposition of this kind is unique.

PROOF. Let us put $m_0 = X_0$, $A_0 = 0$ and

$$m_n = m_0 + \sum_{j=0}^{n-1} [X_{j+1} - \mathsf{E}(X_{j+1}|\mathscr{F}_j)], \tag{12}$$

$$A_n = \sum_{j=0}^{n-1} [\mathsf{E}(X_{j+1}|\mathscr{F}_j) - X_j]. \tag{13}$$

It is evident that m and A, defined in this way, have the required properties. In addition, let $X_n = m_n' + A_n'$, where $m' = (m_n', \mathscr{F}_n)$ is a martingale and $A' = (A_n', F_n)$ is a predictable increasing sequence. Then

$$A_{n+1}' - A_n' = (A_{n+1} - A_n) + (m_{n+1} - m_n) - (m_{n+1}' - m_n'),$$

and if we take conditional expectations on both sides, we find that (P-a.s.) $A_{n+1}' - A_n' = A_{n+1} - A_n$. But $A_0 = A_0' = 0$, and therefore $A_n = A_n'$ and $m_n = m_n'$ (P-a.s.) for all $n \geq 0$.

This completes the proof of the theorem.

It follows from (11) that the sequence $A = (A_n, F_{n-1})$ compensates $X = (X_n, F_n)$ so that it becomes a martingale. This observation is justified by the following definition.

Definition 7. A predictable increasing sequence $A = (A_n, \mathscr{F}_{n-1})$ appearing in the Doob decomposition (11) is called a *compensator* (of the submartingale X).

The Doob decomposition plays a key role in the study of square integrable martingales $M = (M_n, F_n)$ i.e., martingales for which $\mathsf{E}M_n^2 < \infty$, $n \geq 0$; this depends on the observation that the stochastic sequence $M^2 = (M^2, \mathscr{F}_n)$ is a submartingale. According to Theorem 2 there is a martingale $m = (m_n, \mathscr{F}_n)$ and a predictable increasing sequence $\langle M\rangle = (\langle M\rangle_n, \mathscr{F}_{n-1})$ such that

$$M_n^2 = m_n + \langle M\rangle_n. \tag{14}$$

The sequence $\langle M \rangle$ is called the *predictable quadratic variation* or the *quadratic characteristic* of M and, in many respects, determines its structure and properties.

It follows from (12) that

$$\langle M \rangle_n = \sum_{j=1}^{n} \mathsf{E}[(\Delta M_j)^2 | \mathscr{F}_{j-1}] \tag{15}$$

and, for all $l \leq k$,

$$\mathsf{E}[(M_k - M_l)^2 | \mathscr{F}_l] = \mathsf{E}[M_k^2 - M_l^2 | \mathscr{F}_l] = \mathsf{E}[\langle M \rangle_k - \langle M \rangle_l | \mathscr{F}_l]. \tag{16}$$

In particular, if $M_0 = 0$ (P-a.s.) then

$$\mathsf{E}M_k^2 = \mathsf{E}\langle M \rangle_k. \tag{17}$$

It is useful to observe that if $M_0 = 0$ and $M_n = \xi_1 + \cdots + \xi_n$, where (ξ_n) is a sequence of independent random variables with $\mathsf{E}\xi_i = 0$ and $\mathsf{E}\xi_i^2 < \infty$, the quadratic variation

$$\langle M \rangle_n = \mathsf{E}M_n^2 = \mathsf{V}\xi_1 + \cdots + \mathsf{V}\xi_n \tag{18}$$

is not random, and indeed coincides with the variance.

If $X = (X_n, \mathscr{F}_n)$ and $Y = (Y_n, \mathscr{F}_n)$ are square integrable martingales, we put

$$\langle X, Y \rangle_n = \tfrac{1}{4}[\langle X + Y \rangle_n - \langle X - Y \rangle_n]. \tag{19}$$

It is easily verified that $(X_n Y_n - \langle X, Y \rangle_n, \mathscr{F}_n)$ is a martingale and therefore, for $l \leq k$,

$$\mathsf{E}[(X_k - X_l)(Y_k - Y_l) | \mathscr{F}_l] = \mathsf{E}[\langle X, Y \rangle_k - \langle X, Y \rangle_l | \mathscr{F}_l]. \tag{20}$$

In the case when $X_n = \xi_1 + \cdots + \xi_n$, $Y_n = \eta_1 + \cdots + \eta_n$, where (ξ_n) and (η_n) are sequences of independent random variables with $\mathsf{E}\xi_i = \mathsf{E}\eta_i = 0$ and $E\xi_i^2 < \infty$, $\mathsf{E}\eta_i^2 < \infty$, the variable $\langle X, Y \rangle_n$ is given by

$$\langle X, Y \rangle_n = \sum^{n} \operatorname{cov}(\xi_i, \eta_i).$$

The sequence $\langle X, Y \rangle = (\langle X, Y \rangle_n, \mathscr{F}_{n-1})$ defined in (19) is often called the *mutual characteristic* of the (square integrable) martingales X and Y.

It is easy to show (compare with (15)) that

$$\langle X, Y \rangle_N = \sum_{i=1}^{n} \mathsf{E}[\Delta X_i \Delta Y_i | \mathscr{F}_{i-1}].$$

In the theory of martingales, an important role is also played by the *quadratic covariation*,

$$[X, Y]_n = \sum_{i=1}^{n} \Delta X_i \Delta Y_i,$$

and the *quadratic variation*,

$$[X]_n = \sum_{i=1}^{n} (\Delta X_i)^2,$$

which can be defined for all random sequences $X = (X_n)_{n\geq 1}$ and $Y = (Y_n)_{n\geq 1}$.

8. Problems

1. Show that (2) and (3) are equivalent.

2. Let σ and τ be Markov times. Show that $\tau + \sigma$, $\tau \wedge \sigma$, and $\tau \vee \sigma$ are also Markov times; and if $\mathsf{P}(\sigma \leq \tau) = 1$, then $\mathscr{F}_\sigma \subseteq \mathscr{F}_\tau$.

3. Show that τ and X_τ are $\mathscr{F}_\tau$-measurable.

4. Let $Y = (Y_n, \mathscr{F}_n)$ be a martingale (or submartingale), let $V = (V_n, \mathscr{F}_{n-1})$ be a predictable sequence, and let $(V \cdot Y)_n$ be integrable random variables, $n \geq 0$. Show that $V \cdot Y$ is a martingale (or submartingale).

5. Let $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \cdots$ be a nondecreasing family of σ-algebras and ξ an integrable random variable. Show that $(X_n)_{n\geq 1}$ with $X_n = \mathsf{E}(\xi|\mathscr{F}_n)$ is a martingale.

6. Let $\mathscr{G}_1 \supseteq \mathscr{G}_2 \supseteq \cdots$ be a nonincreasing family of σ-algebras and let ξ be an integrable random variable. Show that $(X_n)_{n\geq 1}$ with $X_n = \mathsf{E}(\xi|\mathscr{G}_n)$ is a *reversed* martingale, i.e.,

$$\mathsf{E}(X_n|X_{n+1}, X_{n+2}, \ldots) = X_{n+1} \quad (\mathsf{P}\text{-a.s.})$$

for every $n \geq 1$.

7. Let $\xi_1, \xi_2, \xi_3, \ldots$ be independent random variables, $\mathsf{P}(\xi_i = 0) = \mathsf{P}(\xi_i = 2) = \frac{1}{2}$ and $X_n = \prod_{i=1}^{n} \xi_i$. Show that there does not exist an integrable random variable ξ and a nondecreasing family $(\mathscr{F}_n)$ of σ-algebras such that $X_n = \mathsf{E}(\xi|\mathscr{F}_n)$. This example shows that not every martingale $(X_n)_{n\geq 1}$ can be represented in the form $(\mathsf{E}(\xi|\mathscr{F}_n))_{n\geq 1}$ (compare Example 3, §11, Chapter I.)

§2. Preservation of the Martingale Property Under Time Change at a Random Time

1. If $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ is a martingale, we have

$$\mathsf{E}X_n = \mathsf{E}X_0 \tag{1}$$

for every $n \geq 1$. Is this property preserved if the time n is replaced by a Markov time τ? Example 8 of the preceding section shows that, in general, the answer is "no": there exist a martingale X and a Markov time τ (finite with probability 1) such that

$$\mathsf{E}X_\tau \neq \mathsf{E}X_0. \tag{2}$$

The following basic theorem describes the "typical" situation, in which, in particular, $\mathsf{E}X_\tau = \mathsf{E}X_0$.

Theorem 1 (Doob). *Let $X = (X_n, \mathscr{F}_n)$ be a martingale (or submartingale), and τ_1 and τ_2, stopping times for which*

$$\mathsf{E}|X_{\tau_i}| < \infty, \qquad i = 1, 2, \tag{3}$$

$$\varliminf_{n\to\infty} \int_{\{\tau_i > n\}} |X_n| \, d\mathsf{P} = 0, \qquad i = 1, 2. \tag{4}$$

Then

$$\mathsf{E}(X_{\tau_2} | \mathscr{F}_{\tau_1}) \underset{(\geq)}{=} X_{\tau_1} \quad (\{\tau_2 \geq \tau_1\}; \mathsf{P}\text{-}a.s.). \tag{5}$$

If also $\mathsf{P}(\tau_1 \leq \tau_2) = 1$, *then*

$$\mathsf{E}X_{\tau_2} \underset{(\geq)}{=} \mathsf{E}X_{\tau_1}. \tag{6}$$

(Here and in the formulas below, read the upper symbol for martingales and the lower symbol for submartingales.)

Proof. It is sufficient to show that, for every $A \in \mathscr{F}_{\tau_1}$,

$$\int_{A \cap \{\tau_2 \geq \tau_1\}} X_{\tau_2} \, d\mathsf{P} \underset{(\geq)}{=} \int_{A \cap \{\tau_2 \geq \tau_1\}} X_{\tau_1} \, d\mathsf{P}. \tag{7}$$

For this, in turn, it is sufficient to show that, for every $n \geq 0$,

$$\int_{A \cap \{\tau_2 \geq \tau_1\} \cap \{\tau_1 = n\}} X_{\tau_2} \, d\mathsf{P} \underset{(\geq)}{=} \int_{A \cap \{\tau_2 \geq \tau_1\} \cap \{\tau_1 = n\}} X_{\tau_1} \, d\mathsf{P},$$

or, what amounts to the same thing,

$$\int_{B \cap \{\tau_2 \geq n\}} X_{\tau_2} \, d\mathsf{P} \underset{(\geq)}{=} \int_{B \cap \{\tau_2 \geq n\}} X_n \, d\mathsf{P}, \tag{8}$$

where $B = A \cap \{\tau_1 = n\} \in \mathscr{F}_n$.

We have

$$\begin{aligned}
\int_{B \cap \{\tau_2 \geq n\}} X_n \, d\mathsf{P} &= \int_{B \cap \{\tau_2 = n\}} X_n \, d\mathsf{P} + \int_{B \cap \{\tau_2 > n\}} X_n \, d\mathsf{P} \underset{(\leq)}{=} \int_{B \cap \{\tau_2 = n\}} X_n \, d\mathsf{P} \\
&\quad + \int_{B \cap \{\tau_2 > n\}} \mathsf{E}(X_{n+1} | \mathscr{F}_n) \, d\mathsf{P} = \int_{B \cap \{\tau_2 = n\}} X_{\tau_2} \, d\mathsf{P} + \int_{B \cap \{\tau_2 \geq n+1\}} X_{n+1} \, d\mathsf{P} \\
&\underset{(\leq)}{=} \int_{B \cap \{n \leq \tau_2 \leq n+1\}} X_{\tau_2} \, d\mathsf{P} + \int_{B \cap \{\tau_2 \geq n+2\}} X_{n+2} \, d\mathsf{P} \underset{(\leq)}{=} \cdots \\
&\underset{(\leq)}{=} \int_{B \cap \{n \leq \tau_2 \leq m\}} X_{\tau_2} \, d\mathsf{P} + \int_{B \cap \{\tau_2 > m\}} X_m \, d\mathsf{P},
\end{aligned}$$

whence

$$\int_{B \cap \{n \leq \tau_2 \leq m\}} X_{\tau_2} \, d\mathsf{P} \underset{(\geq)}{=} \int_{B \cap \{n \leq \tau_2\}} X_n \, d\mathsf{P} - \int_{B \cap \{m < \tau_2\}} X_m \, d\mathsf{P}$$

and since $X_m = 2X_m^+ - |X_m|$, we have, by (4),

$$\int_{B\cap\{\tau_2\geq n\}} X_{\tau_2}\,dP \underset{(\geq)}{=} \overline{\lim_{m\to\infty}}\left[\int_{B\cap\{n\leq\tau_2\}} X_n\,dP - \int_{B\cap\{m<\tau_2\}} X_m\,dP\right]$$

$$= \int_{B\cap\{n\leq\tau_2\}} X_n\,dP - \underline{\lim_{m\to\infty}}\int_{B\cap\{m<\tau_2\}} X_m\,dP = \int_{B\cap\{\tau_2\geq n\}} X_n\,dP,$$

which establishes (8), and hence (5). Finally, (6) follows from (5).

This completes the proof of the theorem.

Corollary 1. *If there is a constant N such that $P(\tau_1 \leq N) = 1$ and $P(\tau_2 \leq N) = 1$, then (3) and (4) are satisfied. Hence if, in addition, $P(\tau_1 \leq \tau_2) = 1$ and X is a martingale, then*

$$EX_0 = EX_{\tau_1} = EX_{\tau_2} = EX_N. \tag{9}$$

Corollary 2. *If the random variables $\{X_n\}$ are uniformly integrable (in particular, if $|X_n| \leq C < \infty$, $n \geq 0$), then (3) and (4) are satisfied.*

In fact, $P(\tau_i > n) \to 0$, $n \to \infty$, and hence (4) follows from Lemma 2, §6, Chapter II. In addition, since the family $\{X_n\}$ is uniformly integrable, we have (see II.6.(16))

$$\sup E|X_N| < \infty. \tag{10}$$

If τ is a stopping time and X is a submartingale, then by Corollary 1, applied to the bounded time $\tau_N = \tau \wedge N$,

$$EX_0 \leq EX_{\tau_N}.$$

Therefore

$$E|X_{\tau_N}| = 2EX_{\tau_N}^+ - EX_{\tau_N} \leq 2EX_{\tau_N}^+ - EX_0. \tag{11}$$

The sequence $X^+ = (X_n^+, \mathscr{F}_n)$ is a submartingale (Example 5, §1) and therefore

$$EX_{\tau_N}^+ = \sum_{j=0}^{N}\int_{\{\tau_N=j\}} X_j^+\,dP + \int_{\{\tau>N\}} X_N^+\,dP \leq \sum_{j=0}^{N}\int_{\{\tau_N=j\}} X_N^+\,dP$$

$$+ \int_{\{\tau>N\}} X_N^+\,dP = EX_N^+ \leq E|X_N| \leq \sup_N E|X_N|.$$

From this and (11) we have

$$E|X_{\tau_N}| \leq 3\sup_N E|X_N|,$$

and hence by Fatou's lemma

$$E|X_\tau| \leq 3\sup_N E|X_N|.$$

Therefore if we take $\tau = \tau_i$, $i = 1, 2$, and use (10), we obtain $\mathsf{E}|X_{\tau_i}| < \infty$, $i = 1, 2$.

Remark. In Example 8 of the preceding section,

$$\int_{\{\tau > n\}} |X_n|\, d\mathsf{P} = (2^n - 1)\mathsf{P}\{\tau > n\} = (2^n - 1)\cdot 2^{-n} \to 1, \qquad n \to \infty,$$

and consequently (4) is violated (for $\tau_2 = \tau$).

2. The following proposition, which we shall deduce from Theorem 1, is often useful in applications.

Theorem 2. *Let $X = (X_n)$ be a martingale (or submartingale) and τ a stopping time (with respect to $(\mathscr{F}_n^X)$, where $\mathscr{F}_n^X = \sigma\{\omega: X_0, \ldots, X_n\}$). Suppose that*

$$\mathsf{E}\tau < \infty,$$

and that for some $n \ge 0$ and some constant C

$$\mathsf{E}\{|X_{n+1} - X_n| \,|\, \mathscr{F}_n^X\} \le C \quad (\{\tau \ge n\};\ \mathsf{P}\text{-a.s.}).$$

Then

$$\mathsf{E}|X_\tau| < \infty$$

and

$$\mathsf{E}X_\tau \underset{(\ge)}{=} \mathsf{E}X_0. \tag{12}$$

We first verify that hypotheses (3) and (4) of Theorem 1 are satisfied with $\tau_2 = \tau$.

Let

$$Y_0 = |X_0|, \qquad Y_j = |X_j - X_{j-1}|, \qquad j \ge 1.$$

Then $|X_\tau| \le \sum_{j=0}^{\tau} Y_j$ and

$$\mathsf{E}|X_\tau| \le \mathsf{E}\left(\sum_{j=0}^{\tau} Y_j\right) = \int_\Omega \left(\sum_{j=0}^{\tau} Y_j\right) d\mathsf{P} = \sum_{n=0}^{\infty} \int_{\{\tau = n\}} \sum_{j=0}^{n} Y_j\, d\mathsf{P}$$

$$= \sum_{n=0}^{\infty} \sum_{j=0}^{n} \int_{\{\tau = n\}} Y_j\, d\mathsf{P} = \sum_{j=0}^{\infty} \sum_{n=j}^{\infty} \int_{\{\tau = n\}} Y_j\, d\mathsf{P} = \sum_{j=0}^{\infty} \int_{\{\tau \ge j\}} Y_j\, d\mathsf{P}.$$

The set $\{\tau \ge j\} = \Omega \backslash \{\tau < j\} \in \mathscr{F}_{j-1}^X$, $j \ge 1$. Therefore

$$\int_{\{\tau \ge j\}} Y_j\, d\mathsf{P} = \int_{\{\tau \ge j\}} \mathsf{E}[Y_j | X_0, \ldots, X_{j-1}]\, d\mathsf{P} \le C\mathsf{P}\{\tau \ge j\}$$

for $j \geq 1$; and

$$\mathsf{E}|X_\tau| \leq \mathsf{E}\left(\sum_{j=0}^{\tau} Y_j\right) \leq \mathsf{E}|X_0| + C\sum_{j=1}^{\infty} \mathsf{P}\{\tau \geq j\} = \mathsf{E}|X_0| + C\mathsf{E}\tau < \infty. \tag{13}$$

Moreover, if $\tau > n$, then

$$\sum_{j=0}^{n} Y_j \leq \sum_{j=0}^{\tau} Y_j,$$

and therefore

$$\int_{\{\tau > n\}} |X_n|\, d\mathsf{P} \leq \int_{\{\tau > n\}} \sum_{j=0}^{\tau} Y_j\, d\mathsf{P}.$$

Hence since (by (13)) $\mathsf{E}\sum_{j=0}^{\tau} Y_j < \infty$ and $\{\tau > n\} \downarrow \varnothing, n \to \infty$, the dominated convergence theorem yields

$$\varliminf_{n\to\infty} \int_{\{\tau > n\}} |X_n|\, d\mathsf{P} \leq \varliminf_{n\to\infty} \int_{\{\tau > n\}} \left(\sum_{j=0}^{\tau} Y_j\right) d\mathsf{P} = 0.$$

Hence the hypotheses of Theorem 1 are satisfied, and (12) follows as required.

This completes the proof of the theorem.

3. Here we present some applications of the preceding theorems.

Theorem 3 (Wald's Identities). *Let $\xi_1, \xi_2, \ldots$ be independent identically distributed random variables with $\mathsf{E}|\xi_i| < \infty$ and τ a stopping time (with respect to $\mathscr{F}_n^\xi$), where $\mathscr{F}_n^\xi = \sigma\{\omega: \xi_1, \ldots, \xi_n\}$, $\tau \geq 1$), and $\mathsf{E}\tau < \infty$. Then*

$$\mathsf{E}(\xi_1 + \cdots + \xi_\tau) = \mathsf{E}\xi_1 \cdot \mathsf{E}\tau. \tag{14}$$

If also $\mathsf{E}\xi_i^2 < \infty$ then

$$\mathsf{E}\{(\xi_1 + \cdots + \xi_\tau) - \tau\mathsf{E}\xi_1\}^2 = \mathsf{V}\xi_1 \cdot \mathsf{E}\tau. \tag{15}$$

Proof. It is clear that $X = (X_n, \mathscr{F}_n^\xi)_{n\geq 1}$ with $X_n = (\xi_1 + \cdots + \xi_n) - n\mathsf{E}\xi_1$ is a martingale with

$$\begin{aligned}\mathsf{E}[|X_{n+1} - X_n| \mid X_1, \ldots, X_n] &= \mathsf{E}[|\xi_{n+1} - \mathsf{E}\xi_1| \mid \xi_1, \ldots, \xi_n] \\ &= \mathsf{E}|\xi_{n+1} - \mathsf{E}\xi_1| \leq 2\mathsf{E}|\xi_1| < \infty.\end{aligned}$$

Therefore $\mathsf{E}X_\tau = \mathsf{E}X_0 = 0$, by Theorem 2, and (14) is established.

Similar considerations applied to the martingale $Y = (Y_n, \mathscr{F}_n^\xi)$ with $Y_n = X_n^2 - n\mathsf{V}\xi_1$ lead to a proof of (15).

Corollary. *Let $\xi_1, \xi_2, \ldots$ be independent identically distributed random variables with*

$$\mathsf{P}(\xi_i = 1) = \mathsf{P}(\xi_i = -1) = \tfrac{1}{2}, S_n = \xi_1 + \cdots + \xi_n$$

and $\tau = \inf\{n \geq 1 : S_n = 1\}$. *Then* $\mathsf{P}\{\tau < \infty\} = 1$ *(see, for example,* (I.9.20)) *and therefore* $\mathsf{P}(S_\tau = 1) = 1, \mathsf{E}S_\tau = 1$. *Hence it follows from* (14) *that* $\mathsf{E}\tau = \infty$.

Theorem 4 (Wald's Fundamental Identity). *Let* $\xi_1, \xi_2, \ldots,$ *be a sequence of independent identically distributed random variables,* $S_n = \xi_1 + \cdots + \xi_n$, *and* $n \geq 1$. *Let* $\varphi(t) = \mathsf{E}e^{t\xi_1}$, $t \in R$, *and for some* $t_0 \neq 0$ *let* $\varphi(t_0)$ *exist and* $\varphi(t_0) \geq 1$.

If τ *is a stopping time (with respect to* $(\mathscr{F}_n^\xi)$, $\mathscr{F}_n^\xi = \sigma\{\omega : \xi_1, \ldots, \xi_n\}$, $\tau \geq 1$), *such that* $|S_n| \leq C$ ($\{\tau \geq n\}$; P*-a.s.) and* $\mathsf{E}\tau < \infty$, *then*

$$\mathsf{E}\left[\frac{e^{t_0 S_\tau}}{(\varphi(t_0))^\tau}\right] = 1. \tag{16}$$

Proof. Take

$$Y_n = e^{t_0 S_n}(\varphi(t_0))^{-n}.$$

Then $Y = (Y_n, \mathscr{F}_n^\xi)_{n \geq 1}$ is a martingale with $\mathsf{E}Y_n = 1$ and, on the set $\{\tau \geq n\}$,

$$\begin{aligned}\mathsf{E}\{|Y_{n+1} - Y_n| \mid Y_1, \ldots, Y_n\} &= Y_n \mathsf{E}\left\{\left|\frac{e^{t_0\xi_{n+1}}}{\varphi(t_0)} - 1\right| \middle| \xi_1, \ldots, \xi_n\right\} \\ &= Y_n \cdot \mathsf{E}\{|e^{t_0\xi_1}\varphi^{-1}(t_0) - 1|\} \leq B < \infty,\end{aligned}$$

where B is a constant. Therefore Theorem 2 is applicable, and (16) follows since $\mathsf{E}Y_1 = 1$.

This completes the proof.

Example 1. This example will let us illustrate the use of the preceding examples to find the probabilities of ruin and of mean duration in games (see §9, Chapter I).

Let $\xi_1, \xi_2, \ldots$ be a sequence of independent Bernoulli random variables with $\mathsf{P}(\xi_i = 1) = p$, $\mathsf{P}(\xi_i = -1) = q$, $p + q = 1$, $S = \xi_1 + \cdots + \xi_n$, and

$$\tau = \inf\{n \geq 1 : S_n = B \text{ or } A\}, \tag{17}$$

where $(-A)$ and B are positive integers.

It follows from (I.9.20) that $\mathsf{P}(\tau < \infty) = 1$ and $\mathsf{E}\tau < \infty$. Then if $\alpha = \mathsf{P}(S_\tau = A)$, $\beta = \mathsf{P}(S_\tau = B)$, we have $\alpha + \beta = 1$. If $p = q = \frac{1}{2}$, we obtain

$$0 = \mathsf{E}S_\tau = \alpha A + \beta B, \quad \text{from (14)},$$

whence

$$\alpha = \frac{B}{B + |A|}, \qquad \beta = \frac{|A|}{B + |A|}.$$

Applying (15), we obtain

$$\mathsf{E}_\tau = \mathsf{E}S_\tau^2 = \alpha A^2 + \beta B^2 = |AB|.$$

However, if $p \neq q$ we find, by considering the martingale $((q/p)^{S_n})_{n\geq 1}$, that

$$\mathsf{E}\left(\frac{q}{p}\right)^{S_\tau} = \mathsf{E}\left(\frac{q}{p}\right)^{S_1} = 1,$$

and therefore

$$\alpha\left(\frac{q}{p}\right)^{A} + \beta\left(\frac{q}{p}\right)^{B} = 1.$$

Together with the equation $\alpha + \beta = 1$ this yields

$$\alpha = \frac{\left(\frac{q}{p}\right)^{B} - 1}{\left(\frac{q}{p}\right)^{B} - \left(\frac{q}{p}\right)^{|A|}}, \qquad \beta = \frac{1 - \left(\frac{q}{p}\right)^{|A|}}{\left(\frac{q}{p}\right)^{B} - \left(\frac{q}{p}\right)^{|A|}}. \tag{18}$$

Finally, since $\mathsf{E}S_\tau = (p - q)\mathsf{E}\tau$, we find

$$\mathsf{E}\tau = \frac{\mathsf{E}S_\tau}{p-q} = \frac{\alpha A + \beta B}{p - q},$$

where α and β are defined by (18).

EXAMPLE 2. In the example considered above, let $p = q = \frac{1}{2}$. Let us show that for every λ in $0 < \lambda < \pi/(B + |A|)$ and every time τ defined in (17),

$$\mathsf{E}(\cos\lambda)^{-\tau} = \frac{\cos\lambda\cdot\dfrac{B+A}{2}}{\cos\lambda\cdot\dfrac{B+|A|}{2}}. \tag{19}$$

For this purpose we consider the martingale $X = (X_n, \mathscr{F}_n^S)_{n\geq 0}$ with

$$X_n = (\cos\lambda)^{-n}\cos\lambda\left(S_n - \frac{B+A}{2}\right) \tag{20}$$

and $S_0 = 0$. It is clear that

$$\mathsf{E}X_n = \mathsf{E}X_0 = \cos\lambda\frac{B+A}{2}. \tag{21}$$

Let us show that the family $\{X_{n\wedge\tau}\}$ is uniformly integrable. For this purpose we observe that, by Corollary 1 to Theorem 1 for $0 < \lambda < \pi/(B + |A|)$,

$$\begin{aligned}\mathsf{E}X_0 = \mathsf{E}X_{n\wedge\tau} &= \mathsf{E}\,(\cos\lambda)^{-(n\wedge\tau)}\cos\lambda\left(S_{n\wedge\tau} - \frac{B+A}{2}\right)\\ &\geq \mathsf{E}\,(\cos\lambda)^{-(n\wedge\tau)}\cos\lambda\frac{B-A}{2}.\end{aligned}$$

Therefore, by (21),

$$\mathsf{E}(\cos\lambda)^{-(n\wedge\tau)} \leq \frac{\cos\lambda\dfrac{B+A}{2}}{\cos\lambda\dfrac{B+|A|}{2}},$$

and consequently by Fatou's lemma,

$$\mathsf{E}(\cos\lambda)^{-\tau} \leq \frac{\cos\lambda\dfrac{B+A}{2}}{\cos\lambda\dfrac{B+|A|}{2}}. \tag{22}$$

Consequently, by (20),

$$|X_{n\wedge\tau}| \leq (\cos\lambda)^{-\tau}.$$

With (22), this establishes the uniform integrability of the family $\{X_{n\wedge\tau}\}$. Then, by Corollary 2 to Theorem 1,

$$\cos\lambda\frac{B+A}{2} = \mathsf{E}X_0 = \mathsf{E}X_\tau = \mathsf{E}(\cos\lambda)^{-\tau}\cos\lambda\frac{B-A}{2},$$

from which the required inequality (19) follows.

4. Problems

1. Show that Theorem 1 remains valid for submartingales if (4) is replaced by

$$\varliminf_{n\to\infty}\int_{\{\tau_i>n\}} X_n^+\,d\mathsf{P} = 0, \qquad i = 1, 2.$$

2. Let $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ be a square-integrable martingale, τ a stopping time and

$$\varliminf_{n\to\infty}\int_{\{\tau>n\}} X_n^2\,d\mathsf{P} = 0,$$

$$\varliminf_{n\to\infty}\int_{\{\tau>n\}} |X_n|\,d\mathsf{P} = 0.$$

Show that then

$$\mathsf{E}X_\tau^2 = \mathsf{E}\langle X\rangle_\tau \left(= \mathsf{E}\sum_{j=0}^{\tau}(\Delta X_j)^2\right),$$

where $\Delta X_0 = X_0$, $\Delta X_j = X_j - X_{j-1}$, $j \geq 1$.

3. Show that

$$\mathsf{E}|X_\tau| \leq \varliminf_{n\to\infty}\mathsf{E}|X_n|$$

for every martingale or nonnegative submartingale $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ and every stopping time τ.

4. Let $X = (X_n, \mathscr{F}_n)_{n\ge 0}$ be a submartingale such that $X_n \ge \mathsf{E}(\xi|\mathscr{F}_n)$ (P-a.s.), $n \ge 0$, where $\mathsf{E}|\xi| < \infty$. Show that if τ_1 and τ_2 are stopping times with $\mathsf{P}(\tau_1 \le \tau_2) = 1$, then

$$X_{\tau_1} \ge \mathsf{E}(X_{\tau_2}|\mathscr{F}_{\tau_1}) \quad \text{(P-a.s.)}.$$

5. Let $\xi_1, \xi_2, \dots$ be a sequence of independent random variables with $\mathsf{P}(\xi_i = 1) = \mathsf{P}(\xi_i = -1) = \frac{1}{2}$, a and b positive numbers, $b > a$,

$$X_n = a \sum_{k=1}^{n} I(\xi_k = +1) - b \sum_{k=1}^{n} I(\xi_k = -1)$$

and

$$\tau = \inf\{n \ge 1 : X_n \le -r\}, \qquad r > 0.$$

Show that $\mathsf{E}e^{\lambda\tau} < \infty$ for $\lambda \le \alpha_0$ and $\mathsf{E}e^{\lambda\tau} = \infty$ for $\lambda > \alpha_0$, where

$$\alpha_0 = \frac{b}{a+b}\ln\frac{2b}{a+b} + \frac{a}{a+b}\ln\frac{2a}{a+b}.$$

6. Let $\xi_1, \xi_2, \dots$ be a sequence of independent random variables with $\mathsf{E}\xi_i = 0$, $\mathsf{V}\xi_i = \sigma_i^2$, $S_n = \xi_1 + \cdots + \xi_n$, $\mathscr{F}_n^\xi = \sigma\{\omega : \xi_1, \dots, \xi_n\}$. Prove the following generalizations of Wald's identities (14) and (15): If $\mathsf{E}\sum_{j=1}^{\tau} \mathsf{E}|\xi_j| < \infty$ then $\mathsf{E}S_\tau = 0$; if $\mathsf{E}\sum_{j=1}^{\tau} \mathsf{E}\xi_j^2 < \infty$, then

$$\mathsf{E}S_\tau^2 = \mathsf{E}\sum_{j=1}^{\tau} \xi_j^2 = \mathsf{E}\sum_{j=1}^{\tau} \sigma_j^2. \tag{23}$$

§3. Fundamental Inequalities

1. Let $X = (X_n, \mathscr{F}_n)_{n\ge 0}$ be a stochastic sequence,

$$X_n^* = \max_{0\le j\le n} |X_j|, \; \|X_n\|_p = (\mathsf{E}|X_n|^p)^{1/p}, \qquad p > 0.$$

In Theorems 1–3 below, we present Doob's fundamental "maximal inequalities for probabilities" and "maximal inequalities in L^p," for submartingales, supermartingales and martingales.

Theorem 1. I. *Let $X = (X_n, \mathscr{F}_n)_{n\ge 0}$ be a submartingale. Then for all $\lambda > 0$*

$$\lambda \mathsf{P}\left\{\max_{k\le n} X_k \ge \lambda\right\} \le \mathsf{E}\left[X_n^+ I\left(\max_{k\le n} X_k \ge \lambda\right)\right] \le \mathsf{E}X_n^+, \tag{1}$$

$$\lambda \mathsf{P}\left\{\min_{k\le n} X_k \le -\lambda\right\} \le \mathsf{E}\left[X_n I\left(\min_{k\le n} X_k > -\lambda\right)\right] - \mathsf{E}X_0 \le \mathsf{E}X_n^+ - \mathsf{E}X_0, \tag{2}$$

$$\lambda \mathsf{P}\left\{\max_{k\le n} |X_k| \ge \lambda\right\} \le 3 \max_{k\le n} \mathsf{E}|X_k|. \tag{3}$$

II. *Let $Y = (Y_n, \mathscr{F}_n)_{n\geq 0}$ be a supermartingale. Then for all $\lambda > 0$*

$$\lambda \mathsf{P}\left\{\max_{k\leq n} Y_k \geq \lambda\right\} \leq \mathsf{E} Y_0 - \mathsf{E}\left[Y_n I\left(\max_{k\leq n} Y_k < \lambda\right)\right] \leq \mathsf{E} Y_0 + \mathsf{E} Y_n^-, \tag{4}$$

$$\lambda \mathsf{P}\left\{\min_{k\leq n} Y_k \leq -\lambda\right\} \leq -\mathsf{E}\left[Y_n I\left(\min_{k\leq n} Y_k \leq -\lambda\right)\right] \leq \mathsf{E} Y_n^-, \tag{5}$$

$$\lambda \mathsf{P}\left\{\max_{k\leq n} |Y_k| \geq \lambda\right\} \leq 3 \max_{k\leq n} \mathsf{E}|Y_k|. \tag{6}$$

III. *Let $Y = (Y_n, \mathscr{F}_n)_{n\geq 0}$ be a nonnegative supermartingale. Then for all $\lambda > 0$*

$$\lambda \mathsf{P}\left\{\max_{k\leq n} Y_k \geq \lambda\right\} \leq \mathsf{E} Y_0, \tag{7}$$

$$\lambda \mathsf{P}\left\{\sup_{k\geq n} Y_k \geq \lambda\right\} \leq \mathsf{E} Y_n. \tag{8}$$

Theorem 2. *Let $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ be a nonnegative submartingale. Then for $p \geq 1$ we have the following inequalities:*

if $p > 1$,

$$\|X_n\|_p \leq \|X_n^*\|_p \leq \frac{p}{p-1}\|X_n\|_p; \tag{9}$$

if $p = 1$,

$$\|X_n\| \leq \|X_n^*\|_1 \leq \frac{e}{e-1}\{1 + \|X_n \ln^+ X_n\|_1\}. \tag{10}$$

Theorem 3. *Let $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ be a martingale, $\lambda > 0$ and $p \geq 1$. Then*

$$\mathsf{P}\left\{\max_{k\leq n} |X_k| \geq \lambda\right\} \leq \frac{\mathsf{E}|X_n|^p}{\lambda^p} \tag{11}$$

and if $p > 1$

$$\|X_n\|_p \leq \|X_n^*\|_p \leq \frac{p}{p-1}\|X_n\|_p. \tag{12}$$

In particular, if $p = 2$

$$\mathsf{P}\left\{\max_{k\leq n} |X_k| \geq \lambda\right\} \leq \frac{\mathsf{E}|X_n|^2}{\lambda^2}, \tag{13}$$

$$\mathsf{E}\left[\max_{k\leq n} X_k^2\right] \leq 4\mathsf{E} X_n^2. \tag{14}$$

PROOF OF THEOREM 1. Since a submartingale with the opposite sign is a supermartingale, (1)–(3) follow from (4)–(6). Therefore, we consider the case of a supermartingale $Y = (Y_n, \mathscr{F}_n)_{n\ge 0}$.

Let us set $\tau = \inf\{k \le n: Y_k \ge \lambda\}$ with $\tau = n$ if $\max_{k\le n} Y_k < \lambda$. Then, by (2.6),

$$\mathsf{E}Y_0 \ge \mathsf{E}Y_\tau = \mathsf{E}\left[Y_\tau; \max_{k\le n} Y_k \ge \lambda\right] + \mathsf{E}\left[Y_\tau; \max_{k\le n} Y_k < \lambda\right]$$
$$\ge \lambda\mathsf{P}\left\{\max_{k\le n} Y_k \ge \lambda\right\} + \mathsf{E}\left[Y_n; \max_{k\le n} Y_k < \lambda\right],$$

which proves (4).

Now let us set $\sigma = \inf\{k \le n: Y_k \le -\lambda\}$, and take $\sigma = n$ if $\min_{k\le n} Y_k > -\lambda$. Again, by (2.6),

$$\mathsf{E}Y_n \le \mathsf{E}Y_\tau = \mathsf{E}\left[Y_\tau; \min_{k\le n} Y_k \le -\lambda\right] + \mathsf{E}\left[Y_\tau; \min_{k\le n} Y_k > -\lambda\right]$$
$$\le \lambda\mathsf{P}\left\{\min_{k\le n} Y_k \le -\lambda\right\} + \mathsf{E}\left[Y_n; \min_{k\le n} Y_k > -\lambda\right].$$

Hence,

$$\lambda\mathsf{P}\left\{\min_{k\le n} Y_k \le -\lambda\right\} \le -\mathsf{E}\left[Y_n; \min_{k\le n} Y_k \le -\lambda\right] \le \mathsf{E}Y_n^-$$

which proves (5).

To prove (6), we notice that $Y^- = (-Y)^+$ is a submartingale. Then, by (4) and (1),

$$\mathsf{P}\left\{\max_{k\le n} |Y_k| \ge \lambda\right\} \le \mathsf{P}\left\{\max_{k\le n} Y_k^+ \ge \lambda\right\} + \mathsf{P}\left\{\max_{k\le n} Y_k^- \ge \lambda\right\}$$
$$= \mathsf{P}\left\{\max_{k\le n} Y_k \ge \lambda\right\} + \mathsf{P}\left\{\max_{k\le n} Y_k^- \ge \lambda\right\}$$
$$\le \mathsf{E}Y_0 + 2\mathsf{E}Y_n^- \le 3\max_{k\le n} \mathsf{E}|Y_k|.$$

Inequality (7) follows from (4).

To prove (8), we set $\gamma = \inf\{k \ge n: Y_k \ge \lambda\}$, taking $\gamma = \infty$ if $Y_k < \lambda$ for all $k \ge n$. Now let $n < N < \infty$. Then, by (2.6),

$$\mathsf{E}Y_n \ge \mathsf{E}Y_{\gamma\wedge N} \ge \mathsf{E}[Y_{\gamma\wedge N} I(\gamma \le N)] \ge \lambda\mathsf{P}\{\gamma \le N\},$$

from which, as $N \to \infty$,

$$\mathsf{E}Y_n \ge \lambda\mathsf{P}\{\gamma < \infty\} = \lambda\mathsf{P}\left\{\sup_{k\ge n} Y_k \ge \lambda\right\}.$$

PROOF OF THEOREM 2.

The first inequalities in (9) and (10) are evident.

To prove the second inequality in (9), we first suppose that

$$\|X_n^*\|_p < \infty, \tag{15}$$

and use the fact that, for every nonnegative random variable ξ and for $r > 0$,

$$\mathsf{E}\xi^r = r\int_0^\infty t^{r-1}\mathsf{P}(\xi \geq t)\,dt. \tag{16}$$

Then we obtain, by (1) and Fubini's theorem, that for $p > 1$

$$\begin{aligned}\mathsf{E}(X_n^*)^p &= p\int_0^\infty t^{p-1}\mathsf{P}\{X_n^* \geq t\}\,dt \leq p\int_0^\infty t^{p-2}\left(\int_{\{X_n^*\geq t\}} X_n\,d\mathsf{P}\right)dt\\ &= p\int_0^\infty t^{p-2}\left[\int_\Omega X_n I\{X_n^* \geq t\}\,d\mathsf{P}\right]dt\\ &= p\int_\Omega X_n\left[\int_0^{X_n^*} t^{p-2}\,dt\right]d\mathsf{P} = \frac{p}{p-1}\,\mathsf{E}[X_n(X_n^*)^{p-1}].\end{aligned} \tag{17}$$

Hence, by Hölder's inequality,

$$\mathsf{E}(X_n^*)^p \leq q\|X_n\|_p \cdot \|(X_n^*)^{p-1}\|_q = q\|X_n\|_p[\mathsf{E}(X_n^*)^p]^{1/q}, \tag{18}$$

where $q = p/(p-1)$.

If (15) is satisfied, we immediately obtain the second inequality in (9) from (18).

However, if (15) is not satisfied, we proceed as follows. In (17), instead of X_n^* we consider $(X_n^* \wedge L)$, where L is a constant. Then we obtain

$$\mathsf{E}(X_n^* \wedge L)^p \leq q\mathsf{E}[X_n(X_n^* \wedge L)^{p-1}] \leq q\|X_n\|_p[\mathsf{E}(X_n^* \wedge L)^p]^{1/q},$$

from which it follows, by the inequality $E(X_n^* \wedge L)^p \leq L^p < \infty$, that

$$\mathsf{E}(X_n^* \wedge L)^p \leq q^p\mathsf{E}X_n^p = q^p\|X_n\|_p^p$$

and therefore,

$$\mathsf{E}(X_n^*)^p = \lim_{L\to\infty} \mathsf{E}(X_n^* \wedge L)^p \leq q^p\|X_n\|_p^p.$$

We now prove the second inequality in (10).

Again applying (1), we obtain

$$\begin{aligned}\mathsf{E}X_n^* - 1 &\leq \mathsf{E}(X_n^* - 1)^+ = \int_0^\infty \mathsf{P}\left\{X_n^* - 1 \geq t\right\}dt\\ &\leq \int_0^\infty \frac{1}{1+t}\left[\int_{\{X_n^*\geq 1+t\}} X_n\,d\mathsf{P}\right]dt = \mathsf{E}X_n \int_0^{X_n^*-1}\frac{dt}{1+t} = \mathsf{E}X_n \ln X_n^*.\end{aligned}$$

Since, for arbitrary $a \geq 0$ and $b > 0$,

$$a \ln b \leq a \ln^+ a + be^{-1}, \tag{19}$$

we have

$$\mathsf{E}X_n^* - 1 \leq \mathsf{E}X_n \ln X_n^* \leq \mathsf{E}X_n \ln^+ X_n + e^{-1}\mathsf{E}X_n^*.$$

If $\mathsf{E}X_n^* < \infty$, we immediately obtain the second inequality (10).

However, if $\mathsf{E}X^* = \infty$, we proceed, as above, by replacing X_n^* by $X_n^* \wedge L$.

This proves the theorem.

The proof of Theorem 3 follows from the remark that $|X|^p$, $p \geq 1$, is a nonnegative submartingale (if $\mathsf{E}|X_n|^p < \infty$, $n \geq 0$), and from inequalities (1) and (9).

Corollary of Theorem 3. *Let* $X_n = \xi_0 + \cdots + \xi_n$, $n \geq 0$, *where* $(\xi_k)_{k \geq 0}$ *is a sequence of independent random variables with* $\mathsf{E}\xi_k = 0$ *and* $\mathsf{E}\xi_k^2 < \infty$. *Then inequality* (13) *becomes* Kolmogorov's inequality (§2, Chapter IV).

2. Let $X = (X_n, \mathscr{F}_n)$ be a nonnegative submartingale and

$$X_n = M_n + A_n,$$

its Doob decomposition. Then, since $\mathsf{E}M_n = 0$, it follows from (1) that

$$\mathsf{P}\{X_n^* \geq \varepsilon\} \leq \frac{\mathsf{E}A_n}{\varepsilon}.$$

Theorem 4, below, shows that this inequality is valid, not only for submartingales, but also for the wider class of sequences that have the property of domination in the following sense.

Definition. Let $X = (X_n, \mathscr{F}_n)$ be a nonnegative stochastic sequence, and $A = (A_n, \mathscr{F}_{n-1})$ an increasing predictable sequence. We shall say that X is *dominated* by the sequence A if

$$\mathsf{E}X_\tau \leq \mathsf{E}A_\tau \tag{20}$$

for every stopping time τ.

Theorem 4. *If* $X = (X_n, \mathscr{F}_n)$ *is a nonnegative stochastic sequence dominated by an increasing predictable sequence* $A = (A_n, \mathscr{F}_{n-1})$, *then for* $\lambda > 0$, $a > 0$, *and any stopping time* τ,

$$\mathsf{P}\{X_\tau^* \geq \lambda\} \leq \frac{\mathsf{E}A_\tau}{\lambda}, \tag{21}$$

$$\mathsf{P}\{X_\tau^* \geq \lambda\} \leq \frac{1}{\lambda}\mathsf{E}(A_\tau \wedge a) + \mathsf{P}(A_\tau \geq a), \tag{22}$$

$$\|X_\tau^*\|_p \le \left(\frac{2-p}{1-p}\right)^{1/p} \|A_\tau\|_p, \qquad 0 < p < 1. \tag{23}$$

PROOF. We set

$$\sigma_n = \min\{j \le \tau \wedge n: X_j \ge \lambda\},$$

taking $\sigma_n = \tau \wedge n$, if $\{\cdot\} = \varnothing$. Then

$$\mathsf{E}A_\tau \ge \mathsf{E}A_{\sigma_n} \ge \mathsf{E}X_{\sigma_n} \ge \int_{\{X_{\tau\wedge n}^* > \lambda\}} X_{\sigma_n}\, d\mathsf{P} \ge \lambda \mathsf{P}\{X_{\tau\wedge n}^* > \lambda\},$$

from which

$$\mathsf{P}\{X_{\tau\wedge n}^* > \lambda\} \le \frac{1}{\lambda}\mathsf{E}A_\tau,$$

and we obtain (21) by Fatou's lemma.

For the proof of (22), we introduce the time

$$\gamma = \inf\{j: A_{j+1} \ge a\},$$

setting $\gamma = \infty$ if $\{\cdot\} = \varnothing$. Then

$$\begin{aligned}\mathsf{P}\{X_\tau^* \ge \lambda\} &= \mathsf{P}\{X_\tau^* \ge \lambda, A_\tau < a\} + \mathsf{P}\{X_\tau^* \ge \lambda, A_\tau \ge a\} \\ &\le \mathsf{P}\{I_{\{A_\tau < a\}} X_\tau^* \ge \lambda\} + \mathsf{P}\{A_\tau \ge a\} \\ &\le \mathsf{P}\{X_{\tau\wedge\gamma}^* \ge \lambda\} + \mathsf{P}\{A_\tau \ge a\} \le \frac{1}{\lambda}\mathsf{E}A_{\tau\wedge\gamma} + \mathsf{P}A_\tau \ge a\} \\ &\le \frac{1}{\lambda}\mathsf{E}(A_\tau \wedge a) + \mathsf{P}(A_\tau \ge a),\end{aligned}$$

where we used (21) and the inequality $I_{\{A_\tau < a\}} X_\tau^* \le X_{\tau\wedge\gamma}^*$. Finally, by (22),

$$\begin{aligned}\|X_\tau^*\|_p^p = \mathsf{E}(X_\tau^*)^p &= \int_0^\infty \mathsf{P}\{(X_\tau^*)^p \ge t\}\, dt = \int_0^\infty \mathsf{P}\{X_\tau^* \ge t^{1/p}\}\, dt \\ &\le \int_0^\infty t^{-1/p}\mathsf{E}[A_\tau \wedge t^{1/p}]\, dt + \int_0^\infty \mathsf{P}\{A_\tau^p \ge t\}\, dt \\ &= \mathsf{E}\int_0^{A_\tau^p} dt + \mathsf{E}\int_{A_\tau^p}^\infty (A_\tau t^{-1/p})\, dt + \mathsf{E}A_\tau^p = \frac{2-p}{1-p}\mathsf{E}A_\tau^p.\end{aligned}$$

This completes the proof.

Remark. Let us suppose that the hypotheses of Theorem 4 are satisfied, except that the sequence $A = (A_n, \mathscr{F}_n)_{n\ge 0}$ is not necessarily predictable, but has the property that for some positive constant c

$$\mathsf{P}\left\{\sup_{k\ge 1} |\Delta A_k| \le c\right\} = 1,$$

where $\Delta A_k = A_k - A_{k-1}$. Then the following inequality is satisfied (compare (22)):

$$\mathsf{P}\{X_\tau^* \geq \lambda\} \leq \frac{1}{\lambda}\mathsf{E}[A_\tau \wedge (a+c)] + \mathsf{P}\{A_\tau \geq a\}. \tag{24}$$

The proof is analogous to that of (22). We have only to replace the time $\gamma = \inf\{j: A_{j+1} \geq a\}$ by $\gamma = \inf\{j: A_j \geq a\}$ and notice that $A_\gamma \leq a + c$.

Corollary. *Let the sequences $X^k = (X_n^k, \mathscr{F}_n^k)$ and $A^k = (A_n^k, \mathscr{F}_n^k)$, $n \geq 0$, $k \geq 1$ satisfy the hypotheses of Theorem 4 or the remark. Also, let $(\tau^k)_{k\geq 1}$ be a sequence of stopping times (with respect to $\mathscr{F}^k = (\mathscr{F}_n^k)$) and $A_{\tau_k}^k \xrightarrow{\mathsf{P}} 0$. Then $(X^k)_{\tau_k}^* \xrightarrow{\mathsf{P}} 0$.*

3. In this subsection we present (without proofs, but with applications) a number of significant inequalities for martingales. These generalize the inequalities of Khinchin and of Marcinkiewicz and Zygmund for sums of independent random variables.

Khinchin's Inequalities. *Let $\xi_1, \xi_2, \ldots$ be independent identically distributed Bernoulli random variables with $\mathsf{P}(\xi_i = 1) = \mathsf{P}(\xi_i = -1) = \frac{1}{2}$ and let $(c_n)_{n\geq 1}$ be a sequence of numbers.*

Then for every p, $0 < p < \infty$, there are universal constants A_p and B_p (independent of (c_n)) such that

$$A_p\left(\sum_{j=1}^n c_j^2\right)^{1/2} \leq \left\| \sum_{j=1}^n c_j\xi_j \right\|_p \leq B_p\left(\sum_{j=1}^n c_j^2\right)^{1/2} \tag{25}$$

for every $n \geq 1$.

The following result generalizes these inequalities (for $p \geq 1$).

Marcinkiewicz and Zygmund's Inequalities. *If $\xi_1, \xi_2, \ldots$ is a sequence of independent integrable random variables with $\mathsf{E}\xi_i = 0$, then for $p \geq 1$ there are universal constants A_p and B_p (independent of (ξ_n)) such that*

$$A_p\left\|\left(\sum_{j=1}^n \xi_j^2\right)^{1/2}\right\|_p \leq \left\|\sum_{j=1}^n \xi_j\right\|_p \leq B_p\left\|\left(\sum_{j=1}^n \xi_j^2\right)^{1/2}\right\|_p \tag{26}$$

for every $n \geq 1$.

In (25) and (26) the sequences $X = (X_n)$ with $X_n = \sum_{j=1}^n c_j\xi_j$ and $X_n = \sum_{j=1}^n \xi_j$ are martingales. It is natural to ask whether the inequalities can be extended to arbitrary martingales.

The first result in this direction was obtained by Burkholder.

Burkholder's Inequalities. *If $X = (X_n, \mathscr{F}_n)$ is a martingale, then for every $p > 1$ there are universal constants A_p and B_p (independent of X) such that*

$$A_p \|\sqrt{[X]_n}\|_p \le \|X_n\|_p \le B_p \|\sqrt{[X]_n}\|_p, \tag{27}$$

for every $n \ge 1$, where $[X]_n$ is the quadratic variation of X_n,

$$[X]_n = \sum_{j=1}^{n} (\Delta X_j)^2, \qquad X_0 = 0. \tag{28}$$

The constants A_p and B_p can be taken to have the values

$$A_p = [18p^{3/2}/(p-1)]^{-1}, \qquad B_p = 18p^{3/2}/(p-1)^{1/2}.$$

It follows from (17), by using (2), that

$$A_p \|\sqrt{[X]_n}\|_p \le \|X_n^*\|_p \le B_p^* \|\sqrt{[X]_n}\|_p, \tag{29}$$

where

$$A_p = [18p^{3/2}/(p-1)]^{-1}, \qquad B_p^* = 18p^{5/2}/(p-1)^{3/2}.$$

Burkholder's inequalities (27) hold for $p > 1$, whereas the Marcinkiewicz–Zygmund inequalities (26) also hold when $p = 1$. What can we say about the validity of (27) for $p = 1$? It turns out that a direct generalization to $p = 1$ is impossible, as the following example shows.

Example. Let $\xi_1, \xi_2, \ldots$ be independent Bernoulli random variables with $\mathsf{P}(\xi_i = 1) = \mathsf{P}(\xi_i = -1) = \frac{1}{2}$ and let

$$X_n = \sum_{j=1}^{n \wedge \tau} \xi_j,$$

where

$$\tau = \inf\left\{n \ge 1 : \sum_{i=1}^{n} \xi_j = 1\right\}.$$

The sequence $X = (X_n, \mathscr{F}_n^{\xi})$ is a martingale with

$$\|X_n\|_1 = \mathsf{E}|X_n| = 2\mathsf{E}X_n^+ \to 2, \qquad n \to \infty.$$

But

$$\|\sqrt{[X]_n}\|_1 = \mathsf{E}\sqrt{[X]_n} = \mathsf{E}\left(\sum_{j=1}^{\tau \wedge n} 1\right)^{1/2} = \mathsf{E}\sqrt{\tau \wedge n} \to \infty.$$

Consequently the first inequality in (27) fails.

It turns out that when $p = 1$ we must generalize not (27), but (29) (which is equivalent when $p > 1$).

Davis's Inequality. *If $X = (X_n, \mathscr{F}_n)$ is a martingale, there are universal* *constants A and B, $0 < A < B < \infty$, such that*

$$A\|\sqrt{[X]_n}\|_1 \le \|X_n^*\|_1 \le B\|\sqrt{[X]_n}\|_1, \tag{30}$$

i.e.,

$$A\mathsf{E}\sqrt{\sum_{j=1}^{n} (\Delta X_j)^2} \le \mathsf{E}\left[\max_{1\le j\le n} |X_n|\right] \le B\mathsf{E}\sqrt{\sum_{j=1}^{n} (\Delta X_j)^2}.$$

Corollary 1. *Let $\xi_1, \xi_2, \ldots$ be independent identically distributed random variables; $S_n = \xi_1 + \cdots + \xi_n$. If $\mathsf{E}|\xi_1| < \infty$ and $\mathsf{E}\xi_1 = 0$, then according to Wald's inequality* (2.14) *we have*

$$\mathsf{E}S_\tau = 0 \tag{31}$$

for every stopping time τ (with respect to $(\mathscr{F}_n^\xi)$) for which $\mathsf{E}\tau < \infty$.

It turns out that (31) is still valid under a weaker hypothesis than $\mathsf{E}\tau < \infty$ if we impose stronger conditions on the random variables. In fact, if

$$\mathsf{E}|\xi_1|^r < \infty,$$

where $1 < r \le 2$, the condition $\mathsf{E}\tau^{1/r} < \infty$ is a sufficient condition for $\mathsf{E}S_\tau = 0$.

For the proof, we put $\tau_n = \tau \wedge n$, $Y = \sup_n |S_{\tau_n}|$, and let $m = [t^r]$ (integral part of t^r) for $t > 0$. By Corollary 1 to Theorem 1, §2, we have $\mathsf{E}S_{\tau_n} = 0$. Therefore a sufficient condition for $\mathsf{E}S_\tau = 0$ is (by the dominated convergence theorem) that $\mathsf{E}\sup_n|S_{\tau_n}| < \infty$.

Using (1) and (27), we obtain

$$\begin{aligned}
\mathsf{P}(Y \ge t) &= \mathsf{P}(\tau \ge t^r, Y \ge t) + \mathsf{P}(\tau < t^r, Y \ge t)\\
&\le \mathsf{P}(\tau \ge t^r) + \mathsf{P}\left\{\max_{1\le j\le m} |S_{\tau_j}| \ge t\right\}\\
&\le \mathsf{P}(\tau \ge t^r) + t^{-r}\mathsf{E}|S_{\tau_m}|^r\\
&\le \mathsf{P}(\tau \ge t^r) + t^{-r}B_r\mathsf{E}\left(\sum_{j=1}^{\tau_m} \xi_j^2\right)^{r/2}\\
&\le \mathsf{P}(\tau \ge t^r) + t^{-r}B_r\mathsf{E}\sum_{j=1}^{\tau_m} |\xi_j|^r.
\end{aligned}$$

Notice that (with $\mathscr{F}_0^\xi = \{\varnothing, \Omega\}$)

$$\begin{aligned}
\mathsf{E}\sum_{j=1}^{\tau_m} |\xi_j|^r &= \mathsf{E}\sum_{j=1}^{\infty} I(j \le \tau_m)|\xi_j|^r\\
&= \sum_{j=1}^{\infty} \mathsf{E}\mathsf{E}[I(j \le \tau_m)|\xi_j|^r|\mathscr{F}_{j-1}^\xi]\\
&= \mathsf{E}\sum_{j=1}^{\infty} I(j \le \tau_m)\mathsf{E}[|\xi|^r|\mathscr{F}_{j-1}^\xi] = \mathsf{E}\sum_{j=1}^{\tau_m} \mathsf{E}|\xi_j|^r = \mu_r\mathsf{E}\tau_m,
\end{aligned}$$

where $\mu_r = \mathsf{E}|\xi_1|^r$. Consequently

$$\begin{aligned}\mathsf{P}(Y \geq t) &\leq \mathsf{P}(\tau \geq t^r) + t^{-r}B_r\mu_r\mathsf{E}\tau_m \\ &= \mathsf{P}(\tau \geq t^r) + B_r\mu_r t^{-r}\left[m\mathsf{P}(\tau \geq t^r) + \int_{\{\tau < t^r\}} \tau\, d\mathsf{P}\right] \\ &\leq (1 + B_r\mu_r)\mathsf{P}(\tau \geq t^r) + B_r\mu_r t^{-r}\int_{\{\tau < t^r\}} \tau\, d\mathsf{P}\end{aligned}$$

and therefore

$$\begin{aligned}\mathsf{E}Y\int_0^\infty \mathsf{P}(Y \geq t)\,dt &\leq (1 + B_r\mu_r)\mathsf{E}\tau^{1/r} + B_r\mu_r\int_0^\infty t^{-r}\left[\int_{\{\tau < t^r\}} \tau\, d\mathsf{P}\right]dt \\ &= (1 + B_r\mu_r)\mathsf{E}\tau^{1/r} + B_r\mu_r\int_\Omega \tau\left[\int_{\tau^{1/r}}^\infty t^{-r}\,dt\right]d\mathsf{P} \\ &= \left(1 + B_r\mu_r + \frac{B_r\mu_r}{r-1}\right)\mathsf{E}\tau^{1/r} < \infty.\end{aligned}$$

Corollary 2. *Let $M = (M_n)$ be a martingale with $\mathsf{E}|M_n|^{2r} < \infty$ for some $r \geq 1$ and such that (with $M_0 = 0$)*

$$\sum_{n=1}^\infty \frac{\mathsf{E}|\Delta M_n|^{2r}}{n^{1+r}} < \infty. \tag{32}$$

Then (compare Theorem 2 of §3, Chapter IV) we have the strong law of large numbers:

$$\frac{M_n}{n} \to 0 \quad (\mathsf{P}\text{-a.s.}), \qquad n \to \infty. \tag{33}$$

When $r = 1$ the proof follows the same lines as the proof of Theorem 2, §3, Chapter IV. In fact, let

$$m_n = \sum_{k=1}^n \frac{\Delta M_k}{k}.$$

Then

$$\frac{M_n}{n} = \frac{\sum_{k=1}^n \Delta M_k}{n} = \frac{1}{n}\sum_{k=1}^n k\Delta m_k$$

and, by Kronecker's lemma (§3, Chapter IV) a sufficient condition for the limit relation (P-a.s.)

$$\frac{1}{n}\sum_{k=1}^n k\Delta m_k \to 0, \qquad n \to \infty,$$

is that the limit $\lim_n m_n$ exists and is finite (P-a.s.) which in turn (Theorems 1 and 4, §10, Chapter II) is true if and only if

$$\mathsf{P}\left\{\sup_{k\geq 1}|m_{n+k} - m_n| \geq \varepsilon\right\} \to 0, \qquad n \to \infty. \tag{34}$$

By (1),

$$\mathsf{P}\left\{\sup_{k\geq 1}|m_{n+k}-m_n|\geq \varepsilon\right\}\leq \varepsilon^{-2}\sum_{k=n}^{\infty}\frac{\mathsf{E}(\Delta M_k)^2}{k^2}.$$

Hence the required result follows from (32) and (34).

Now let $r > 1$. Then the statement (33) is equivalent (Theorem 1, §10, Chapter II) to the statement that

$$\varepsilon^{2r}\mathsf{P}\left\{\sup_{j\geq n}\frac{|M_j|}{j}\geq \varepsilon\right\}\to 0, \qquad n\to\infty \tag{35}$$

for every $\varepsilon > 0$. By inequality (53) of Problem 1,

$$\varepsilon^{2r}\mathsf{P}\left\{\sup_{j\geq n}\frac{|M_j|}{j}\geq \varepsilon\right\} = \varepsilon^{2r}\lim_{m\to\infty}\mathsf{P}\left\{\max_{n\leq j\leq m}\frac{|M_j|^{2r}}{j^{2r}}\geq \varepsilon^{2r}\right\}$$
$$\leq \frac{1}{n^{2r}}\mathsf{E}|M_n|^{2r}+\sum_{j\geq n+1}\frac{1}{j^{2r}}\mathsf{E}(|M_j|^{2r}-|M_{j-1}|^{2r}).$$

It follows from Kronecker's lemma and (32) that

$$\lim_{n\to\infty}\frac{1}{n^{2r}}\mathsf{E}|M_n|^{2r}=0.$$

Hence to prove (35) we need only prove that

$$\sum_{j\geq 2}\frac{1}{j^{2r}}\mathsf{E}(|M_j|^{2r}-|M_{j-1}|^{2r})<\infty. \tag{36}$$

We have

$$I_N=\sum_{j=2}^{N}\frac{1}{j^{2r}}[\mathsf{E}|M_j|^{2r}-\mathsf{E}|M_{j-1}|^{2r}]$$
$$\leq \sum_{j=3}^{N}\left[\frac{1}{(j-1)^{2r}}-\frac{1}{j^{2r}}\right]\mathsf{E}|M_{j-1}|^{2r}+\frac{\mathsf{E}|M_N|^{2r}}{N^{2r}}.$$

By Burkholder's inequality (27) and Hölder's inequality,

$$\mathsf{E}|M_j|^{2r}\leq \mathsf{E}\left[\sum_{i=1}^{j}(\Delta M_i)^2\right]^r\leq \mathsf{E}j^{r-1}\sum_{i=1}^{j}|\Delta M_i|^{2r}.$$

Hence

$$I_N\leq \sum_{j=2}^{N-1}\left[\frac{1}{j^{2r}}-\frac{1}{(j+1)^{2r}}\right]j^{r-1}\sum_{i=1}^{j}\mathsf{E}|\Delta M_i|^{2r}$$
$$\leq C_1\sum_{j=2}^{N-1}\frac{1}{j^{r+2}}\sum_{i=1}^{j}\mathsf{E}|\Delta M_i|^{2r}\leq C_2\sum_{j=2}^{N}\frac{\mathsf{E}|\Delta M_j|^{2r}}{j^{r+1}}+C_3$$

(C_i are constants). By (32), this establishes (36).

4. The sequence of random variables $\{X_n\}_{n\geq 1}$ has a limit $\lim X_n$ (finite or infinite) with probability 1, if and only if the number of "oscillations between two arbitrary rational numbers a and b, $a < b$" is finite with probability 1. Theorem 5, below, provides an upper bound for the number of "oscillations" for submartingales. In the next section, this will be applied to prove the fundamental result on their convergence.

Let us choose two numbers a and b, $a < b$, and define the following *times* in terms of the stochastic sequence $X = (X_n, \mathscr{F}_n)$:

$$\begin{aligned} \tau_0 &= 0, \\ \tau_1 &= \min\{n > 0: X_n \leq a\}, \\ \tau_2 &= \min\{n > \tau_1: X_n \geq b\}, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ \tau_{2m-1} &= \min\{n > \tau_{2m-2}: X_n \leq a\}, \\ \tau_{2m} &= \min\{n > \tau_{2m-1}: X_n \geq b\}, \end{aligned}$$

taking $\tau_k = \infty$ if the corresponding set $\{\cdot\}$ is empty.

In addition, for each $n \geq 1$ we define the random variables

$$\beta_n(a, b) = \begin{cases} 0, & \text{if } \tau_2 > n, \\ \max\{m: \tau_{2m} \leq n\} & \text{if } \tau_2 \leq n. \end{cases}$$

In words, $\beta_n(a, b)$ is the *number of upcrossings* of $[a, b]$ by the sequence $X_1, \ldots, X_n$.

Theorem 5 (Doob). *Let $X = (X_n, \mathscr{F}_n)_{n\geq 1}$ be a submartingale. Then, for every $n \geq 1$,*

$$\mathsf{E}\beta_n(a, b) \leq \frac{\mathsf{E}[X_n - a]^+}{b - a}. \tag{37}$$

Proof. The number of intersections of $X = (X_n, \mathscr{F}_n)$ with $[a, b]$ is equal to the number of intersections of the nonnegative submartingale $X^+ = ((X_n - a)^+, \mathscr{F}_n)$ with $[0, b - a]$. Hence it is sufficient to suppose that X is nonnegative with $a = 0$, and show that

$$\mathsf{E}\beta_n(0, b) \leq \frac{\mathsf{E}X_n}{b}. \tag{38}$$

Put $X_0 = 0$, $\mathscr{F}_0 = \{\varnothing, \Omega\}$, and for $i = 1, 2, \ldots$, let

$$\varphi_i = \begin{cases} 1 & \text{if } \tau_m < i \leq \tau_{m+1} \text{ for some odd } m, \\ 0 & \text{if } \tau_m < i \leq \tau_{m+1} \text{ for some even } m. \end{cases}$$

It is easily seen that

$$b\beta_n(0, b) \leq \sum_{i=1}^n \varphi_i[X_i - X_{i-1}]$$

and

$$\{\varphi_i = 1\} = \bigcup_{\text{odd } m} [\{\tau_m < i\} \setminus \{\tau_{m+1} < i\}] \in \mathscr{F}_{i-1}.$$

Therefore

$$\begin{aligned} b\mathsf{E}\beta_n(0, b) &\le \mathsf{E}\sum_{i=1}^{n} \varphi_i[X_i - X_{i-1}] = \sum_{i=1}^{n} \int_{\{\varphi_i = 1\}} (X_i - X_{i-1})\, d\mathsf{P} \\ &= \sum_{i=1}^{n} \int_{\{\varphi_i = 1\}} \mathsf{E}(X_i - X_{i-1} \mid \mathscr{F}_{i-1})\, d\mathsf{P} \\ &= \sum_{i=1}^{n} \int_{\{\varphi_i = 1\}} [\mathsf{E}(X_i \mid \mathscr{F}_{i-1}) - X_{i-1}]\, d\mathsf{P} \\ &\le \sum_{i=1}^{n} \int_{\Omega} [\mathsf{E}(X_i \mid \mathscr{F}_{i-1}) - X_{i-1}]\, d\mathsf{P} = \mathsf{E}X_n, \end{aligned}$$

which establishes (38).

5. In this subsection we discuss some of the simplest inequalities for the probability of large deviations for martingales of integrable square.

Let $M = (M_n, \mathscr{F}_n)_{n \ge 0}$ be a martingale of integrable square with quadratic variation $\langle M \rangle = (\langle M \rangle_n, \mathscr{F}_{n-1})$. If we apply inequality (22) to $X_n = M_n$, $A_n = \langle M \rangle_n$, we find that for $a > 0$ and $b > 0$

$$\begin{aligned} \mathsf{P}\left\{\max_{k \le n} |M_k| \ge an\right\} &= \mathsf{P}\left\{\max_{k \le n} M_k^2 \ge (an)^2\right\} \\ &\le \frac{1}{(an)^2}\mathsf{E}[\langle M \rangle_n \wedge (bn)] + \mathsf{P}\{\langle M \rangle_n \ge an\}. \end{aligned} \tag{39}$$

In fact, at least in the case when $|\Delta M_n| \le C$ for all n and $\omega \in \Omega$, this inequality can be substantially improved by using the ideas explained in §5, Chapter IV for estimating the probability of large deviations for sums of independent identically distributed random variables.

Let us recall that in §5, Chapter IV, when we introduced the corresponding inequalities, the essential point was to use the property that the sequence

$$(e^{\lambda S_n}/[\varphi(\lambda)]^n, \mathscr{F}_n)_{n \ge 1}, \qquad \mathscr{F}_n = \sigma\{\xi_1, \ldots, \xi_n\}, \tag{40}$$

formed a nonnegative martingale, to which we could apply the inequality (8). If we now take M_n instead of S_n, by analogy with (40), the martingale

$$(e^{\lambda M_n}/\mathscr{E}_n(\lambda), \mathscr{F}_n)_{n \ge 1},$$

will be nonnegative, where

$$\mathscr{E}_n(\lambda) = \prod_{j=1}^{n} \mathsf{E}(e^{\lambda \Delta M_j} \mid \mathscr{F}_{j-1}) \tag{41}$$

is called the *stochastic exponential*.

This expression is rather complicated. At the same time, in using (8) it is not necessary for the sequence to be a *martingale*. It is enough for it to be a nonnegative *supermartingale*. Here we can arrange this by forming a sequence $(Z_n(\lambda), \mathscr{F}_n)$ ((43), below), which sufficiently depends simply on M_n and $\langle M\rangle_n$, and to which we can apply the method used in §5, Chapter IV.

Lemma 1. *Let $M = (M_n, \mathscr{F}_n)_{n\geq 0}$ be a square-integrable martingale, $M_0 = 0$, $\Delta M_0 = 0$, and $|\Delta M_n(\omega)| \leq c$ for all n and ω. Let $\lambda > 0$,*

$$\psi_c(\lambda) = \begin{cases} \dfrac{e^{\lambda c} - 1 - \lambda^c}{c^2}, & c > 0, \\ \dfrac{\lambda^2}{2}, & c = 0, \end{cases} \tag{42}$$

and

$$Z_n(\lambda) = e^{\lambda M_n - \psi_c(\lambda)\langle M\rangle_n}. \tag{43}$$

Then for every $c \geq 0$ the sequence $Z(\lambda) = (Z_n(\lambda), \mathscr{F}_n)_{n\geq 0}$ is a non-negative supermartingale.

Proof. For $|x| \leq c$,

$$e^{\lambda x} - 1 - \lambda x = (\lambda x)^2 \sum_{m\geq 2} \frac{(\lambda x)^{m-2}}{m!} \leq (\lambda x)^2 \sum_{m\geq 2} \frac{(\lambda c)^{m-2}}{m!} \leq x^2 \psi_c(\lambda).$$

Using this inequality and the following representation ($Z_n = Z_n(\lambda)$)

$$\Delta Z_n = Z_{n-1}[(e^{\lambda\Delta M_n} - 1)e^{-\Delta\langle M\rangle_n \psi_c(\lambda)} + (e^{-\Delta\langle M\rangle_n\psi_c(\lambda)} - 1)],$$

we find that

$$\begin{aligned} &\mathsf{E}(\Delta Z_n|\mathscr{F}_{n-1}) \\ &= Z_{n-1}[\mathsf{E}(e^{\lambda\Delta M_n} - 1|\mathscr{F}_{n-1})e^{-\Delta\langle M\rangle_n\psi_c(\lambda)} + (e^{-\Delta\langle M\rangle_n\psi_c(\lambda)} - 1)] \\ &= Z_{n-1}[\mathsf{E}(e^{\lambda\Delta M_n} - 1 - \lambda\Delta M_n|\mathscr{F}_{n-1})e^{-\Delta\langle M\rangle_n\psi_c(\lambda)} + (e^{-\Delta\langle M\rangle_n\psi_c(\lambda)} - 1)] \\ &\leq Z_{n-1}[\psi_c(\lambda)\mathsf{E}((\Delta M_n)^2|\mathscr{F}_{n-1})e^{-\Delta\langle M\rangle_n\psi_c(\lambda)} + (e^{-\Delta\langle M\rangle_n\psi_c(\lambda)} - 1)] \\ &= Z_{n-1}[\psi_c(\lambda)\Delta\langle M\rangle_n e^{-\Delta\langle M\rangle_n\psi_c(\lambda)} + (e^{-\Delta\langle M\rangle_n\psi_c(\lambda)} - 1)] \leq 0, \end{aligned} \tag{44}$$

where we have also used the fact that, for $x \geq 0$,

$$xe^{-x} + (e^{-x} - 1) \leq 0.$$

We see from (44) that

$$\mathsf{E}(Z_n|\mathscr{F}_{n-1}) \leq Z_{n-1},$$

i.e., $Z(\lambda) = (Z_n(\lambda), \mathscr{F}_n)$ is a supermartingale.

This establishes the lemma.

Let the hypotheses of the lemma be satisfied. Then we can always find $\lambda > 0$ for which, for given $a > 0$ and $b > 0$, we have $a\lambda - b\psi_c(\lambda) > 0$. From this, we obtain

$$\begin{aligned}
\mathsf{P}\left\{\max_{k\le n} M_k \ge an\right\} &= \mathsf{P}\left\{\max_{k\le n} e^{\lambda M_k} \ge e^{\lambda an}\right\} \\
&\le \mathsf{P}\left\{\max_{k\le n} e^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} \ge e^{\lambda an - \psi_c(\lambda)\langle M\rangle_n}\right\} \\
&= \mathsf{P}\left\{\max_{k\le n} e^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} \ge e^{\lambda an - \psi_c(\lambda)\langle M\rangle_n}, \langle M\rangle_n \le bn\right\} \\
&\quad + \mathsf{P}\left\{\max_{k\le n} e^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} \ge e^{\lambda an - \psi_c(\lambda)\langle M\rangle_n}, \langle M\rangle_n > bn\right\} \\
&\le \mathsf{P}\left\{\max_{k\le n} e^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} \ge e^{\lambda an - \psi_c(\lambda)bn}\right\} \qquad (45)
\end{aligned}$$

where the last inequality follows from (7).

Let us write

$$H_c(a, b) = \sup_{\lambda > 0} [a\lambda - b\psi_c(\lambda)].$$

Then it follows from (45) that

$$\mathsf{P}\left\{\max_{k\le n} M_k \ge an\right\} \le \mathsf{P}\{\langle M\rangle_n > bn\} + e^{-nH_c(a,b)}. \qquad (46)$$

Passing from M to $-M$, we find that the right-hand side of (46) also provides an upper bound for the probability $\mathsf{P}\{\min_{k\le n} M_k \le -an\}$. Consequently,

$$\mathsf{P}\left\{\max_{k\le n} |M_k| \ge an\right\} \le 2\mathsf{P}\{\langle M\rangle_n > bn\} + 2e^{-nH_c(a,b)}. \qquad (47)$$

Thus, we have proved the following theorem.

Theorem 6. *Let $M = (M_n, \mathscr{F}_n)$ be a martingale with uniformly bounded steps, i.e., $|\Delta M_n| \le c$ for some constant $c > 0$ and all n and ω. Then for every $a > 0$ and $b > 0$, we have the inequalities* (46) *and* (47).

Remark 2.

$$H_c(a, b) = \frac{1}{c}\left(a + \frac{b}{c}\right)\ln\left(1 + \frac{ac}{b}\right) - \frac{a}{c}. \qquad (48)$$

6. Under the hypotheses of Theorem 6, we now consider the question of estimates of probabilities of the type

$$\mathsf{P}\left\{\sup_{k\ge n} \frac{M_k}{\langle M\rangle_k} > a\right\},$$

which characterize, in particular, the rapidity of convergence in the strong law of large numbers for martingales (also see Theorem 4 in §5).

Proceeding as in §5, Chapter IV, we find that for every $a > 0$ there is a $\lambda > 0$ for which $a\lambda - \psi_c(\lambda) > 0$. Then, for every $b > 0$,

$$\begin{aligned}\mathsf{P}\left\{\sup_{k\geq n}\frac{M_k}{\langle M\rangle_k} > a\right\} &\leq \mathsf{P}\left\{\sup_{k\geq n} e^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} > e^{[a\lambda - \psi_c(\lambda)]\langle M\rangle_n}\right\}\\ &\leq \mathsf{P}\left\{\sup_{k\geq n} e^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} > e^{[a\lambda - \psi_c(\lambda)]bn}\right\}\\ &\quad + \mathsf{P}\{\langle M\rangle_n < bn\} \leq e^{-bn[a\lambda - \psi_c(\lambda)]} + \mathsf{P}\{\langle M\rangle_n < bn\},\end{aligned} \tag{49}$$

from which

$$\mathsf{P}\left\{\sup_{k\geq n}\frac{M_k}{\langle M\rangle_k} > a\right\} \leq \mathsf{P}\{\langle M\rangle_n < bn\} + e^{-nH_c(ab,b)} \tag{50}$$

$$\mathsf{P}\left\{\sup_{k\geq n}\left|\frac{M_k}{\langle M\rangle_k}\right| > a\right\} \leq 2\mathsf{P}\{\langle M\rangle_n < bn\} + 2e^{-nH_c(ab,b)}. \tag{51}$$

We have therefore proved the following theorem.

Theorem 7. *Let the hypotheses of the preceding theorem be satisfied. Then inequalities* (50) *and* (51) *are satisfied for all* $a > 0$ *and* $b > 0$.

Remark 3. Comparison of (51) with the estimate (21) in §5, Chapter IV, for the case of a Bernoulli scheme, $p = 1/2$, $M_n = S_n - (n/2)$, $b = 1/4$, $c = 1/2$, shows that for small $\varepsilon > 0$ it leads to the same result

$$\mathsf{P}\left\{\sup_{k\geq n}\left|\frac{M_k}{\langle M\rangle_k}\right| > \varepsilon\right\} = \mathsf{P}\left\{\sup_{k\geq n}\left|\frac{S_k - k/2}{k}\right| > \frac{\varepsilon}{4}\right\} \leq 2e^{-4\varepsilon^2 n}.$$

7. Problems

1. Let $X = (X_n, \mathscr{F}_n)$ be a nonnegative submartingale and let $V = (V_n, \mathscr{F}_{n-1})$ be a predictable sequence such that $0 \leq V_{n+1} \leq V_n \leq C$ (P-a.s.), where C is a constant. Establish the following generalization of (1):

$$\varepsilon\mathsf{P}\left\{\max_{i\leq j\leq n} V_j X_j \geq \varepsilon\right\} + \int_{\{\max_{1\leq j\leq n} V_j X_j < \varepsilon\}} V_n X_n \, d\mathsf{P} \leq \sum_{j=1}^{n} \mathsf{E} V_j \Delta X_j. \tag{52}$$

2. Establish *Krickeberg's decomposition*: every martingale $X = (X_n, \mathscr{F}_n)$ with $\sup \mathsf{E}|X_n| < \infty$ can be represented as the difference of two nonnegative martingales.

3. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables, $S_n = \xi_1 + \cdots + \xi_n$ and $S_{m,n} = \sum_{j=m+1}^{n} \xi_j$. Establish *Ottaviani's inequality*:

$$\mathsf{P}\left\{\max_{1\leq j\leq n} |S_j| > 2\varepsilon\right\} \leq \frac{\mathsf{P}\{|S_n| > \varepsilon\}}{\min_{1\leq j\leq n}\mathsf{P}\{|S_{j,n}| \leq \varepsilon\}}$$

and deduce that

$$\int_0^\infty \mathsf{P}\left\{\max_{1\le j\le n} |S_j| > 2t\right\} dt \le 2\mathsf{E}|S_n| + 2\int_{2\mathsf{E}|S_n|}^\infty \mathsf{P}\{|S_n| > t\}\, dt. \tag{53}$$

4. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with $\mathsf{E}\xi_i = 0$. Use (53) to show that in this case we can strengthen inequality (10) to

$$\mathsf{E}S_n^* \le 8\mathsf{E}|S_n|.$$

5. Verify formula (16).

6. Establish inequality (19).

7. Let the σ-algebra $\mathscr{F}_0, \ldots, \mathscr{F}_n$ be such that $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \cdots \subseteq \mathscr{F}_n$ and let the events $A_k \in \mathscr{F}_k$, $k = 1, \ldots, n$. Use (22) to establish *Dvoretzky's inequality*: for each $\varepsilon > 0$,

$$\mathsf{P}\left[\bigcup_{k=1}^n A_k \middle| \mathscr{F}_0\right] \le \varepsilon + \mathsf{P}\left[\sum_{k=1}^n \mathsf{P}(A_k|\mathscr{F}_{k-1}) > \varepsilon \middle| \mathscr{F}_0\right] \quad (\mathsf{P}\text{-a.s.}).$$

§4. General Theorems on the Convergence of Submartingales and Martingales

1. The following result, which is fundamental for all problems about the convergence of submartingales, can be thought of as an analog of the fact that in real analysis a bounded monotonic sequence of numbers has a (finite) limit.

Theorem 1 (Doob). *Let $X = (X_n, \mathscr{F}_n)$ be a submartingale with*

$$\sup_n \mathsf{E}|X_n| < \infty. \tag{1}$$

Then with probability 1, *the limit* $\lim X_n = X_\infty$ *exists and* $\mathsf{E}|X_\infty| < \infty$.

Proof. Suppose that

$$\mathsf{P}(\overline{\lim}\, X_n > \underline{\lim}\, X_n) > 0. \tag{2}$$

Then since

$$\{\overline{\lim}\, X_n > \underline{\lim}\, X_n\} = \bigcup_{a<b} \{\overline{\lim}\, X_n > b > a > \underline{\lim}\, X_n\}$$

(here a and b are rational numbers), there are values a and b such that

$$\mathsf{P}\{\overline{\lim}\, X_n > b > a > \underline{\lim}\, X_n\} > 0. \tag{3}$$

Let $\beta_n(a, b)$ be the number of upcrossings of (a, b) by the sequence $X_1, \ldots, X_n$, and let $\beta_\infty(a, b) = \lim_n \beta_n(a, b)$. By (3.27),

$$\mathsf{E}\beta_n(a, b) \leq \frac{\mathsf{E}[X_n - a]^+}{b - a} \leq \frac{\mathsf{E}X_n^+ + |a|}{b - a}$$

and therefore

$$\mathsf{E}\beta_\infty(a, b) = \lim_n \mathsf{E}\beta_n(a, b) \leq \frac{\sup_n \mathsf{E}X_n^+ + |a|}{b - a} < \infty,$$

which follows from (1) and the remark that

$$\sup_n \mathsf{E}|X_n| < \infty \Leftrightarrow \sup_n \mathsf{E}X_n^+ < \infty$$

for submartingales (since $\mathsf{E}X_n^+ \leq \mathsf{E}|X_n| = 2\mathsf{E}X_n^+ - \mathsf{E}X_n \leq 2\mathsf{E}X_n^+ - \mathsf{E}X_1$). But the condition $\mathsf{E}\beta_\infty(a, b) < \infty$ contradicts assumption (3). Hence $\lim X_n = X_\infty$ exists with probability 1, and then by Fatou's lemma

$$\mathsf{E}|X_\infty| \leq \sup_n \mathsf{E}|X_n| < \infty.$$

This completes the proof of the theorem.

Corollary 1. *If X is a nonpositive submartingale, then with probability* 1 *the limit* $\lim X_n$ *exists and is finite.*

Corollary 2. *If $X = (X_n, \mathscr{F}_n)_{n\geq 1}$ is a nonpositive submartingale, the sequence $\overline{X} = (X_n, \mathscr{F}_n)$ with $1 \leq n \leq \infty$, $X_\infty = \lim X_n$ and $\mathscr{F}_\infty = \sigma\{\bigcup \mathscr{F}_n\}$ is a (nonpositive) submartingale.*

In fact, by Fatou's lemma

$$\mathsf{E}X_\infty = \mathsf{E}\lim X_n \geq \overline{\lim}\, \mathsf{E}X_n \geq \mathsf{E}X_1 > -\infty$$

and (P-a.s.)

$$\mathsf{E}(X_\infty | \mathscr{F}_m) = \mathsf{E}(\lim X_n | \mathscr{F}_m) \geq \overline{\lim}\, \mathsf{E}(X_n | \mathscr{F}_m) \geq X_m.$$

Corollary 3. *If $X = (X_n, \mathscr{F}_n)$ is a nonnegative martingale, then* $\lim X_n$ *exists with probability* 1.

In fact, in that case

$$\sup \mathsf{E}|X_n| = \sup \mathsf{E}X_n = \mathsf{E}X_1 < \infty,$$

and Theorem 1 is applicable.

2. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with $\mathsf{P}(\xi_i = 0) = \mathsf{P}(\xi_i = 2) = \frac{1}{2}$. Then $X = (X_n, \mathscr{F}_n^\xi)$, with $X_n = \prod_{i=1}^n \xi_i$ and $\mathscr{F}_n^\xi = \sigma\{\omega: \xi_1, \ldots, \xi_n\}$ is a martingale with $\mathsf{E}X_n = 1$ and $X_n \to X_\infty \equiv 0$ (P-a.s.). At the same time, it is clear that $\mathsf{E}|X_n - X_\infty| = 1$ and therefore $X_n \overset{L^1}{\nrightarrow} X_\infty$. Therefore condition (1) does not in general guarantee the convergence of X_n to X_∞ in the L^1 sense.

Theorem 2 below shows that if hypothesis (1) is strengthened to uniform integrability of the family $\{X_n\}$ (from which (1) follows by Subsection 4,

§6, Chapter II), then besides almost sure convergence we also have convergence in L^1.

Theorem 2. *Let $X = \{X_n, \mathscr{F}_n\}$ be a uniformly integrable submartingale (that is, the family $\{X_n\}$ is uniformly integrable). Then there is a random variable X_∞ with $\mathsf{E}|X_\infty| < \infty$, such that as $n \to \infty$*

$$X_n \to X_\infty \quad (\mathsf{P}\text{-a.s.}), \tag{4}$$

$$X_n \xrightarrow{L^1} X_\infty. \tag{5}$$

Moreover, the sequence $\bar{X} = (X_n, \mathscr{F}_n)$, $1 \le n \le \infty$, with $\mathscr{F}_\infty = \sigma(\bigcup \mathscr{F}_n)$, is also a submartingale.

PROOF. Statement (4) follows from Theorem 1, and (5) follows from (4) and Theorem 4, §6, Chapter II.

Moreover, if $A \in \mathscr{F}_n$ and $m \ge n$, then

$$\mathsf{E} I_A |X_m - X_\infty| \to 0, \qquad m \to \infty,$$

and therefore

$$\lim_{m \to \infty} \int_A X_m \, d\mathsf{P} = \int_A X_\infty \, d\mathsf{P}.$$

The sequence $\{\int_A X_m \, d\mathsf{P}\}_{m \ge n}$ is nondecreasing and therefore

$$\int_A X_n \, d\mathsf{P} \le \int_A X_m \, d\mathsf{P} \le \int_A X_\infty \, d\mathsf{P},$$

whence $X_n \le \mathsf{E}(X_\infty | \mathscr{F}_n)$ (P-a.s.) for $n \ge 1$.

This completes the proof of the theorem.

Corollary. *If $X = (X_n, \mathscr{F}_n)$ is a submartingale and, for some $p > 1$,*

$$\sup_n \mathsf{E}|X_n|^p < \infty, \tag{6}$$

then there is an integrable random variable X_∞ for which (4) and (5) are satisfied.

For the proof, it is enough to observe that, by Lemma 3 of §6 of Chapter II, condition (6) guarantees the uniform integrability of the family $\{X_n\}$.

3. We now present a theorem on the continuity properties of conditional expectations. This was one of the very first results concerning the convergence of martingales.

Theorem 3 (P. Lévy). *Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space, and let $(\mathscr{F}_n)_{n \ge 1}$ be a nondecreasing family of σ-algebras, $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \cdots \subseteq \mathscr{F}$. Let ξ be a random variable with $\mathsf{E}|\xi| < \infty$ and $\mathscr{F}_\infty = \sigma(\bigcup_n \mathscr{F}_n)$. Then, both P-a.s. and in the L^1 sense,*

$$\mathsf{E}(\xi | F_n) \to \mathsf{E}(\xi | F_\infty), \qquad n \to \infty. \tag{7}$$

PROOF. Let $X_n = \mathsf{E}(\xi|\mathscr{F}_n)$, $n \geq 1$. Then, with $a > 0$ and $b > 0$,

$$\begin{aligned}\int_{\{|X_i|\geq a\}} |X_i|\, d\mathsf{P} &\leq \int_{\{|X_i|\geq a\}} \mathsf{E}(|\xi|\,|F_i)\, d\mathsf{P} = \int_{\{|X_i|\geq a\}} |\xi|\, d\mathsf{P}\\ &\leq \int_{\{|X_i|\geq a\}\cap\{|\xi|\leq b\}} |\xi|\, d\mathsf{P} + \int_{\{|X_i|\geq a\}\cap\{|\xi|>b\}} |\xi|\, d\mathsf{P}\\ &\leq b\mathsf{P}\{|X_i|\geq a\} + \int_{\{|\xi|>b\}} |\xi|\, d\mathsf{P}\\ &\leq \frac{b}{a}\mathsf{E}|X_i| + \int_{\{|\xi|>b\}} |\xi|\, d\mathsf{P}\\ &\leq \frac{b}{a}\mathsf{E}|\xi| + \int_{\{|\xi|>b\}} |\xi|\, d\mathsf{P}.\end{aligned}$$

Letting $a \to \infty$ and then $b \to \infty$, we obtain

$$\lim_{a\to\infty} \sup_i \int_{\{|X_i|\geq a\}} |X_i|\, d\mathsf{P} = 0,$$

i.e., the family $\{X_n\}$ is uniformly integrable.

Therefore, by Theorem 2, there is a random variable X_∞ such that $X_n = \mathsf{E}(\xi|F_n) \to X_\infty$ ((P-a.s.) and in the L^1 sense). Hence we only have to show that

$$X_\infty = \mathsf{E}(\xi|\mathscr{F}_\infty) \quad (\text{P-a.s.}).$$

Let $m \geq n$ and $A \in \mathscr{F}_n$. Then

$$\int_A X_m\, d\mathsf{P} = \int_A X_n\, d\mathsf{P} = \int_A \mathsf{E}(\xi|F_n)\, d\mathsf{P} = \int_A \xi\, d\mathsf{P}.$$

Since the family $\{X_n\}$ is uniformly integrable and since, by Theorem 5, §6, Chapter II, we have $\mathsf{E}I_A|X_m - X_\infty| \to 0$ as $m \to \infty$, it follows that

$$\int_A X_\infty\, d\mathsf{P} = \int_A \xi\, d\mathsf{P}. \tag{8}$$

This equation is satisfied for all $A \in \mathscr{F}_n$ and therefore for all $A \in \bigcup_{n=1}^\infty \mathscr{F}_n$. Since $\mathsf{E}|X_\infty| < \infty$ and $\mathsf{E}|\xi| < \infty$, the left-hand and right-hand sides of (8) are σ-additive measures; possibly taking negative as well as positive values, but finite and agreeing on the algebra $\bigcup_{n=1}^\infty \mathscr{F}_n$. Because of the uniqueness of the extension of a σ-additive measure to an algebra over the smallest σ-algebra containing it (Carathéodory's theorem, §3, Chapter II, equation (8) remains valid for sets $A \in F_\infty = \sigma(\bigcup F_n)$. Thus,

$$\int_A X_\infty\, d\mathsf{P} = \int_A \xi\, d\mathsf{P} = \int_A \mathsf{E}(\xi|\mathscr{F}_\infty)\, d\mathsf{P}, \qquad A \in \mathscr{F}_\infty. \tag{9}$$

Since X_∞ and $\mathsf{E}(\xi|\mathscr{F}_\infty)$ are $\mathscr{F}_\infty$-measurable, it follows from Property I of Subsection 2, §6, Chapter II, and from (9), that $X_\infty = \mathsf{E}(\xi|\mathscr{F}_\infty)$ (P-a.s.).

This completes the proof of the theorem.

Corollary. *A stochastic sequence $X = (X_n, \mathscr{F}_n)$ is a uniformly integrable martingale if and only if there is a random variable ξ with $\mathsf{E}|\xi| < \infty$ such that $X_n = \mathsf{E}(\xi | F_n)$ for all $n \geq 1$. Here $X_n \to \mathsf{E}(\xi | F_\infty)$ (both P-a.s. and in the L^1 sense) as $n \to \infty$.*

In fact, if $X = (X_n, \mathscr{F}_n)$ is a uniformly integrable martingale, then by Theorem 2 there is an integrable random variable X_∞ such that $X_n \to X_\infty$ (P-a.s. and in the L^1 sense) and $X_n = \mathsf{E}(X_\infty | F_n)$. As the random variable ξ we may take the $\mathscr{F}_\infty$-measurable variable X_∞.

The converse follows from Theorem 3.

4. We now turn to some applications of these theorems.

Example 1. *The "zero or one" law.* Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables, $\mathscr{F}_n^\xi = \sigma\{\omega: \xi_1, \ldots, \xi_n\}$ and let X be the σ-algebra of the "tail" events. By Theorem 3, we have $\mathsf{E}(I_A | F_n^\xi) \to \mathsf{E}(I_A | F_\infty^\xi) = I_A$ (P-a.s.). But I_A and $(\xi_1, \ldots, \xi_n)$ are independent. Since $\mathsf{E}(I_A | F_n^\xi) = \mathsf{E} I_A$ and therefore $I_A = \mathsf{E} I_A$ (P-a.s.), we find that either $\mathsf{P}(A) = 0$ or $\mathsf{P}(A) = 1$.

The next two examples illustrate possible applications of the preceding results to convergence theorems in analysis.

Example 2. If $f = f(x)$ satisfies a Lipschitz condition on $[0, 1)$, it is absolutely continuous and, as is shown in courses in analysis, there is a (Lebesgue) integrable function $g = g(x)$ such that

$$f(x) - f(0) = \int_0^x g(y)\, dy. \tag{10}$$

(In this sense, $g(x)$ is a "derivative" of $f(x)$.)

Let us show how this result can be deduced from Theorem 1.

Let $\Omega = [0, 1)$, $\mathscr{F} = \mathscr{B}([0, 1))$, and let P denote Lebesgue measure. Put

$$\xi_n(x) = \sum_{k=1}^{2^n} \frac{k-1}{2^n} I\left\{\frac{k-1}{2^n} \leq x < \frac{k}{2^n}\right\},$$

$\mathscr{F}_n = \sigma\{x: \xi_1, \ldots, \xi_n\} = \sigma\{x: \xi_n\}$, and

$$X_n = \frac{f(\xi_n + 2^{-n}) - f(\xi_n)}{2^{-n}}.$$

Since for a given ξ_n the random variable ξ_{n+1} takes only the values ξ_n and $\xi_n + 2^{-(n+1)}$ with conditional probabilities equal to $\frac{1}{2}$, we have

$$\begin{aligned}\mathsf{E}[X_{n+1} | \mathscr{F}_n] &= \mathsf{E}[X_{n+1} | \xi_n] = 2^{n+1}\mathsf{E}[f(\xi_{n+1} + 2^{-(n+1)}) - f(\xi_{n+1}) | \xi_n] \\ &= 2^{n+1}\{\tfrac{1}{2}[f(\xi_n + 2^{-(n+1)}) - f(\xi_n)] + \tfrac{1}{2}[f(\xi_n + 2^{-n}) - f(\xi_n + 2^{-n+1})]\} \\ &= 2^n\{f(\xi_n + 2^{-n}) - f(\xi_n)\} = X_n.\end{aligned}$$

It follows that $X = (X_n, \mathscr{F}_n)$ is a martingale, and it is uniformly integrable since $|X_n| \le L$, where L is the Lipschitz constant: $|f(x) - f(y)| \le L|x - y|$. Observe that $\mathscr{F} = \mathscr{B}([0, 1)) = \sigma(\bigcup \mathscr{F}_n)$. Therefore, by the corollary to Theorem 3, there is an $\mathscr{F}$-measurable function $g = g(x)$ such that $X_n \to g$ (P-a.s.) and

$$X_n = \mathsf{E}[g|\mathscr{F}_n]. \tag{11}$$

Consider the set $B = [0, k/2^n]$. Then by (11)

$$f\left(\frac{k}{2^n}\right) - f(0) = \int_0^{k/2^n} X_n\,dx = \int_0^{k/2^n} g(x)\,dx,$$

and since n and k are arbitrary, we obtain the required equation (10).

EXAMPLE 3. Let $\Omega = [0, 1)$, $\mathscr{F} = \mathscr{B}([0, 1))$ and let P denote Lebesgue measure. Consider the Haar system $\{H_n(x)\}_{n \ge 1}$, as defined in Example 3 of §11, Chapter II. Put $\mathscr{F}_n = \sigma\{x: H_1, \dots, H_n\}$ and observe that $\sigma(\bigcup \mathscr{F}_n) = \mathscr{F}$. From the properties of conditional expectations and the structure of the Haar functions, it is easy to deduce that

$$\mathsf{E}[f(x)|\mathscr{F}_n] = \sum_{k=1}^{n} a_k H_k(x) \quad (\mathsf{P}\text{-a.s.}), \tag{12}$$

for every Borel function $f \in L$, where

$$a_k = (f, H_k) = \int_0^1 f(x) H_k(x)\,dx.$$

In other words, the conditional expectation $\mathsf{E}[f(x)|\mathscr{F}_n]$ is a partial sum of the Fourier series of $f(x)$ in the Haar system. Then if we apply Theorem 3 to the martingale we find that, as $n \to \infty$,

$$\sum_{k=1}^{n} (f, H_k) H_k(x) \to f(x) \quad (\mathsf{P}\text{-a.s.})$$

and

$$\int_0^1 \left| \sum_{k=1}^{n} (f, H_k) H_k(x) - f(x) \right| dx \to 0.$$

EXAMPLE 4. Let $(\xi_n)_{n \ge 1}$ be a sequence of random variables. By Theorem 2, §10, Chapter II, the P-a.e. convergence of the series $\sum \xi_n$ implies its convergence in probability and in distribution. It turns out that if the random variables $\xi_1, \xi_2, \dots$ are independent, the converse is also valid: the convergence in distribution of the series $\sum \xi_n$ of independent random variables implies its convergence in probability and with probability one.

Let $S_n = \xi_1 + \cdots + \xi_n$, $n \ge 1$ and $S_n \xrightarrow{d} S$. Then $\mathsf{E}e^{itS_n} \to \mathsf{E}e^{itS}$ for every real number t. It is clear that there is a $\delta > 0$ such that $|\mathsf{E}e^{itS}| > 0$ for all $|t| < \delta$. Choose t_0 so that $|t_0| < \delta$. Then there is an $n_0 = n_0(t_0)$ such that $|\mathsf{E}e^{it_0 S_n}| \ge c > 0$ for all $n \ge n_0$, where c is a constant.

For $n \ge n_0$, we form the sequence $X = (X_n, \mathscr{F}_n)$ with

$$X_n = \frac{e^{it_0 S_n}}{\mathsf{E}e^{it_0 S_n}}, \qquad \mathscr{F}_n = \sigma\{\xi_1, \dots, \xi_n\}.$$

Since $\xi_1, \xi_2, \ldots$ were assumed to be independent, the sequence $X = (X_n, \mathscr{F}_n)$ is a martingale with

$$\sup_{n \geq n_0} \mathsf{E}|X_n| \leq c^{-1} < \infty.$$

Then it follows from Theorem 1 that with probability one the limit $\lim_n X_n$ exists and is finite. Therefore, the $\text{limit}_{n\to\infty} e^{it_0 S_n}$ also exists with probability one. Consequently, we can assert that there is a $\delta > 0$ such that for each t in the set $T = \{t: |t| < \delta\}$ the limit $\lim_n e^{itS_n}$ exists with probability one.

Let $T \times \Omega = \{(t, \omega): t \in T, \omega \in \Omega\}$, let $\bar{\mathscr{B}}(T)$ be the σ-algebra of Lebesgue sets on T and let λ be Lebesgue measure on $(T, \bar{\mathscr{B}}(T))$. Also, let

$$C = \left\{(t, \omega) \in T \times \Omega: \lim_n e^{itS_n(\omega)} \text{ exists}\right\}.$$

It is clear that $C \in \bar{\mathscr{B}}(T) \otimes \mathscr{F}$.

It was shown above that $P(C_t) = 1$ for every $t \in T$, where $C_t = \{\omega \in \Omega: (t, \omega) \in C\}$ is the section of C at the point t. By Fubini's theorem (Theorem 8, §6, Chapter II)

$$\int_{T\times\Omega} I_C(t, \omega) d(\lambda \times \mathsf{P}) = \int_T \left(\int_\Omega I_C(t, \omega)\, d\mathsf{P}\right) d\lambda$$

$$= \int_T \mathsf{P}(C_t)\, d\lambda = \lambda(T) = 2\delta > 0.$$

On the other hand, again by Fubini's theorem,

$$\lambda(T) = \int_{T\times\Omega} I_C(t, \omega) d(\lambda \times \mathsf{P}) = \int_\Omega d\mathsf{P}\left(\int_T I_C(t, \omega)\, d\lambda\right) = \int_\Omega \lambda(C_\omega)\, d\mathsf{P},$$

where $C_\omega = \{t: (t, \omega) \in C\}$.

Hence, it follows that there is a set $\tilde{\Omega}$ with $P(\tilde{\Omega}) = 1$ such that $\lambda(C_\omega) = \lambda(T) = 2\delta > 0$ for all $\omega \in \tilde{\Omega}$.

Consequently, we may say that for every $\omega \in \tilde{\Omega}$ the limit $\lim_n e^{itS_n}$ exists for all $t \in C_\omega$. In addition, the measure of C_ω is positive. From this and Problem 8, it follows that the limit $\lim_n S_n(\omega)$ exists and is finite for $\omega \in \tilde{\Omega}$. Since $P(\tilde{\Omega}) = 1$, the limit $\lim_n S_n(\omega)$ exists and is finite with probability one.

5. Problems

1. Let $\{\mathscr{G}_n\}$ be a nonincreasing family of σ-algebras, $\mathscr{G}_1 \supseteq \mathscr{G}_2 \supseteq \cdots \mathscr{G}_\infty = \bigcap \mathscr{G}_n$, and let η be an integrable random variable. Establish the following analog of Theorem 3: as $n \to \infty$,

$$\mathsf{E}(\eta | G_n) \to \mathsf{E}(\eta | G_\infty) \quad (\mathsf{P}\text{-a.s. and in the } L^1 \text{ sense}).$$

2. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables with $\mathsf{E}|\xi_1| < \infty$ and $\mathsf{E}\xi_1 = m$; let $S_n = \xi_1 + \cdots + \xi_n$. Having shown (see Problem 2, §7, Chapter II) that

$$\mathsf{E}(\xi_1 | S_n, S_{n+1}, \ldots) = \mathsf{E}(\xi_1 | S_n) = \frac{S_n}{n} \quad (\mathsf{P}\text{-a.s.}),$$

deduce from Problem 1 a stronger form of the law of large numbers: as $n \to \infty$,

$$\frac{S_n}{n} \to m \quad \text{(P-a.s. and in the } L^1 \text{ sense).}$$

3. Establish the following result, which combines Lebesgue's dominated convergence theorem and P. Lévy's theorem. Let $\{\xi_n\}_{n \geq 1}$ be a sequence of random variables such that $\xi_n \to \xi$ (P-a.s.), $|\xi_n| \leq \eta$, $\mathsf{E}\eta < \infty$ and $\{\mathscr{F}_m\}_{m \geq 1}$ is a nondecreasing family of σ-algebras, with $\mathscr{F}_\infty = \sigma(\bigcup \mathscr{F}_n)$. Then

$$\lim_{\substack{m \to \infty \\ n \to \infty}} \mathsf{E}(\xi_n | \mathscr{F}_m) = \mathsf{E}(\xi | \mathscr{F}_\infty) \quad \text{(P-a.s.)}.$$

4. Establish formula (12).

5. Let $\Omega = [0, 1]$, $\mathscr{F} = \mathscr{B}([0, 1))$, let P denote Lebesgue measure, and let $f = f(x) \in L^1$. Put

$$f_n(x) = 2^n \int_{k2^{-n}}^{(k+1)2^{-n}} f(y)\, dy, \qquad k2^{-n} \leq x < (k+1)2^{-n}.$$

Show that $f_n(x) \to f(x)$ (P-a.s.).

6. Let $\Omega = [0, 1)$, $\mathscr{F} = \mathscr{B}([0, 1))$, let P denote Lebesgue measure and let $f = f(x) \in L^1$. Continue this function periodically on $[0, 2)$ and put

$$f_n(x) = \sum_{i=1}^{2^n} 2^{-n} f(x + i2^{-n}).$$

Show that $f_n(x) \to f(x)$ (P-a.s.).

7. Prove that Theorem 1 remains valid for generalized submartingales $X = (X_n, \mathscr{F}_n)$, if $\inf_m \sup_{n \geq m} \mathsf{E}(X_n^+ | \mathscr{F}_m) < \infty$ (P-a.s.).

8. Let a_n, $n \geq 1$, be a sequence of real numbers such that for all real numbers t with $|t| < \delta$, $\delta > 0$, the limit $\lim_n e^{ita_n}$ exists. Prove that then the limit $\lim a_n$ exists and is finite.

§5. Sets of Convergence of Submartingales and Martingales

1. Let $X = (X_n, \mathscr{F}_n)$ be a stochastic sequence. Let us denote by $\{X_n \to\}$, or $\{-\infty < \lim X_n < \infty\}$, the set of sample points for which $\lim X_n$ *exists* and is *finite*. Let us also write $A \subseteq B$ (P-a.s.) if $\mathsf{P}(I_A \leq I_B) = 1$.

If X is a submartingale and $\sup \mathsf{E}|X_n| < \infty$ (or, equivalently, if $\sup \mathsf{E}X_n^+ < \infty$), then according to Theorem 1 of §4 we have

$$\{X_n \to\} = \Omega \quad \text{(P-a.s.)}.$$

Let us consider the structure of sets $\{X_n \to\}$ of convergence for submartingales when the hypothesis $\sup \mathsf{E}|X_n| < \infty$ is not satisfied.

Let $a > 0$, and $\tau_a = \inf\{n \geq 1 : X_n > a\}$ with $\tau_a = \infty$ if $\{\cdot\} = \varnothing$.

Definition. A stochastic sequence $X = (X_n, \mathscr{F}_n)$ belongs to *class* $\boldsymbol{C}^+$ $(X \in \boldsymbol{C}^+)$ if

$$\mathsf{E}(\Delta X_{\tau_a})^+ I\{\tau_a < \infty\} < \infty \tag{1}$$

for every $a > 0$, where $\Delta X_n = X_n - X_{n-1}$, $X_0 = 0$.

It is evident that $X \in \boldsymbol{C}^+$ if

$$\mathsf{E} \sup_n |\Delta X_n| < \infty \tag{2}$$

or, all the more so, if

$$|\Delta X_n| \le C < \infty \quad (\mathsf{P}\text{-a.s.}), \tag{3}$$

for all $n \ge 1$.

Theorem 1. *If the submartingale* $X \in \boldsymbol{C}^+$ *then*

$$\{\sup X_n < \infty\} = \{X_n \to\} \quad (\mathsf{P}\text{-}a.s.). \tag{4}$$

Proof. The inclusion $\{X_n \to\} \subseteq \{\sup X_n < \infty\}$ is evident. To establish the inclusion in the opposite direction, we consider the stopped submartingale $X^{\tau_a} = (X_{\tau_a \wedge n}, \mathscr{F}_n)$. Then, by (1),

$$\begin{aligned} \sup_n \mathsf{E} X^+_{\tau_a \wedge n} &\le a + \mathsf{E}[X^+_{\tau_a} \cdot I\{\tau_a < \infty\}] \\ &\le 2a + \mathsf{E}[(\Delta X_{\tau_a})^+ \cdot I\{\tau_a < \infty\}] < \infty, \end{aligned} \tag{5}$$

and therefore by Theorem 1 of §4,

$$\{\tau_a = \infty\} \subseteq \{X_n \to\} \quad (\mathsf{P}\text{-a.s.}).$$

But $\bigcup_{a>0}\{\tau_a = \infty\} = \{\sup \; X_n < \infty\}$; hence $\{\sup \; X_n < \infty\} \subseteq \{X_n \to\}$ (P-a.s.).

This completes the proof of the theorem.

Corollary. *Let X be a martingale with* $\mathsf{E} \sup|\Delta X_n| < \infty$. *Then* ($\mathsf{P}$-*a.s.*)

$$\{X_n \to\} \cup \{\underline{\lim}\, X_n = -\infty, \overline{\lim}\, X_n = +\infty\} = \Omega. \tag{6}$$

In fact, if we apply Theorem 1 to X and to $-X$, we find that (P-a.s.)

$$\{\overline{\lim}\, X_n < \infty\} = \{\sup X_n < \infty\} = \{X_n \to\},$$

$$\{\underline{\lim}\, X_n > -\infty\} = \{\inf X_n > -\infty\} = \{X_n \to\}.$$

Therefore (P-a.s.)

$$\{\overline{\lim}\, X_n < \infty\} \cup \{\underline{\lim}\, X_n > -\infty\} = \{X_n \to\},$$

which establishes (6).

Statement (6) means that, provided that $\mathsf{E} \sup |\Delta X_n| < \infty$, either almost all trajectories of the martingale M have finite limits, or all behave very badly, in the sense that $\overline{\lim}\, X_n = +\infty$ and $\underline{\lim}\, X_n = -\infty$.

2. If $\xi_1, \xi_2, \ldots$ is a sequence of independent random variables with $\mathsf{E}\xi_i = 0$ and $|\xi_i| \le c < \infty$, then by Theorem 1 of §2, Chapter IV, the series $\sum \xi_i$

converges (P-a.s.) if and only if $\sum \mathsf{E}\xi_i^2 < \infty$. The sequence $X = (X_n, \mathscr{F}_n)$ with $X_n = \xi_1 + \cdots + \xi_n$ and $\mathscr{F}_n = \sigma\{\omega: \xi_1, \ldots, \xi_n\}_n$ is a square-integrable martingale with $\langle X \rangle_n = \sum_{i=1}^n \mathsf{E}\xi_i^2$, and the proposition just stated can be interpreted as follows:

$$\{\langle X \rangle_\infty < \infty\} = \{X_n \to\} = \Omega \quad (\text{P-a.s.}),$$

where $\langle X \rangle_\infty = \lim_n \langle X \rangle_n$.

The following proposition generalizes this result to more general martingales and submartingales.

Theorem 2. *Let $X = (X_n, \mathscr{F}_n)$ be a submartingale and*

$$X_n = m_n + A_n$$

its Doob decomposition.

(a) *If X is a nonnegative submartingale, then* (P-*a.s.*)

$$\{A_\infty < \infty\} \subseteq \{X_n \to\} \subseteq \{\sup X_n < \infty\}. \tag{7}$$

(b) *If $X \in C^+$ then* (P-*a.s.*)

$$\{X_n \to\} = \{\sup X_n < \infty\} \subseteq \{A_\infty < \infty\}. \tag{8}$$

(c) *If X is a nonnegative submartingale and $X \in C^+$, then* (P-*a.s.*)

$$\{X_n \to\} = \{\sup X_n < \infty\} = \{A_\infty < \infty\}. \tag{9}$$

Proof. (a) The second inclusion in (7) is obvious. To establish the first inclusion we introduce the *times*

$$\sigma_a = \inf\{n \geq 1: A_{n+1} > a\}, \qquad a > 0,$$

taking $\sigma_a = +\infty$ if $\{\cdot\} = \varnothing$. Then $A_{\sigma_a} \leq a$ and by Corollary 1 to Theorem 1 of §2, we have

$$\mathsf{E}X_{n \wedge \sigma_a} = \mathsf{E}A_{n \wedge \sigma_a} \leq a.$$

Let $Y_n^a = X_{n \wedge \sigma_a}$. Then $Y^a = (Y_n^a, \mathscr{F}_n)$ is a submartingale with $\sup \mathsf{E}Y_n^a \leq a < \infty$. Since the martingale is nonnegative, it follows from Theorem 1, §4, that (P-a.s.)

$$\{A_\infty \leq a\} = \{\sigma_a = \infty\} \subseteq \{X_n \to\}.$$

Therefore (P-a.s.)

$$\{A_\infty < \infty\} = \bigcup_{a>0} \{A_\infty \leq a\} \subseteq \{X_n \to\}.$$

(b) The first equation follows from Theorem 1. To prove the second, we notice that, in accordance with (5),

$$\mathsf{E}A_{\tau_a \wedge n} = \mathsf{E}X_{\tau_a \wedge n} \leq \mathsf{E}X_{\tau_a \wedge n}^+ \leq 2a + \mathsf{E}[(\Delta X_{\tau_a})^+ I\{\tau_a < \infty\}]$$

and therefore

$$\mathsf{E}A_{\tau_a} = \mathsf{E}\lim_n A_{\tau_a \wedge n} < \infty.$$

Hence $\{\tau_a = \infty\} \subseteq \{A_\infty < \infty\}$ and we obtain the required conclusion since $\bigcup_{a>0} \{\tau_a = \infty\} = \{\sup X_n < \infty\}$.

(c) This is an immediate consequence of (a) and (b).

This completes the proof of the theorem.

Remark. The hypothesis that X is nonnegative can be replaced by the hypothesis $\sup_n \mathsf{E} X_n^- < \infty$.

Corollary 1. *Let* $X_n = \xi_1 + \cdots + \xi_n$, *where* $\xi_i \geq 0$, $\mathsf{E}\xi_i < \infty$, ξ_i *are* $\mathscr{F}_i$-*measurable, and* $\mathscr{F}_0 = \{\varnothing, \Omega\}$. *Then* (P-*a.s.*)

$$\left\{\sum_{n=1}^{\infty} \mathsf{E}(\xi_n | \mathscr{F}_{n-1}) < \infty\right\} \subseteq \{X_n \to\}, \tag{10}$$

and if, in addition, $\mathsf{E} \sup_n \xi_n < \infty$ *then* (P-*a.s.*)

$$\left\{\sum_{n=1}^{\infty} \mathsf{E}(\xi_n | \mathscr{F}_{n-1}) < \infty\right\} = \{X_n \to\}. \tag{11}$$

Corollary 2 (Borel–Cantelli–Lévy Lemma). *If the events* $B_n \in \mathscr{F}_n$, *then if we put* $\xi_n = I_{B_n}$ *in* (11), *we find that*

$$\left\{\sum_{n=1}^{\infty} \mathsf{P}(B_n | \mathscr{F}_{n-1}) < \infty\right\} = \left\{\sum_{n=1}^{\infty} I_{B_n} < \infty\right\}. \tag{12}$$

3. Theorem 3. *Let* $M = (M_n, \mathscr{F}_n)_{n\geq 1}$ *be a square-integrable martingale. Then* (P-*a.s.*)

$$\{\langle M \rangle_\infty < \infty\} \subseteq \{M_n \to\}. \tag{13}$$

If also $\mathsf{E} \sup |\Delta M_n|^2 < \infty$, *then* (P-*a.s.*)

$$\{\langle M \rangle_\infty < \infty\} = \{M_n \to\}, \tag{14}$$

where

$$\langle M \rangle_\infty = \sum_{n=1}^{\infty} \mathsf{E}((\Delta M_n)^2 | \mathscr{F}_{n-1}) \tag{15}$$

with $M_0 = 0$, $\mathscr{F}_0 = \{\varnothing, \Omega\}$.

PROOF. Consider the two submartingales $M^2 = (M_n^2, \mathscr{F}_n)$ and $(M + 1)^2 = ((M + 1)^2, \mathscr{F}_n)$. Let their Doob decompositions be

$$M_n^2 = m_n' + A_n', \qquad (M_n + 1)^2 = m_n'' + A_n''.$$

Then A_n' and A_n'' are the same, since

$$A_n' = \sum_{k=1}^{n} \mathsf{E}(\Delta M_k^2 | \mathscr{F}_{k-1}) = \sum_{k=1}^{n} \mathsf{E}((\Delta M_k)^2 | \mathscr{F}_{k-1})$$

and

$$A''_n = \sum_{k=1}^{n} \mathsf{E}(\Delta(M_k + 1)^2 | \mathscr{F}_{k-1}) = \sum_{k=1}^{n} \mathsf{E}(\Delta M_k^2 | \mathscr{F}_{k-1})$$
$$= \sum_{k=1}^{n} \mathsf{E}((\Delta M_k)^2 | \mathscr{F}_{k-1}).$$

Hence (7) implies that (P-a.s.)

$$\{\langle M \rangle_\infty < \infty\} = \{A'_\infty < \infty\} \subseteq \{M_n^2 \to\} \cap \{(M_n + 1)^2 \to\} = \{M_n \to\}.$$

Because of (9), equation (14) will be established if we show that the condition $\mathsf{E} \sup |\Delta M_n|^2 < \infty$ guarantees that M^2 belongs to C^+.

Let $\tau_a = \inf\{n \geq 1 : M_n^2 > a\}$, $a > 0$. Then, on the set $\{\tau_a < \infty\}$,

$$|\Delta M_{\tau_a}^2| = |M_{\tau_a}^2 - M_{\tau_a - 1}^2| \leq |M_{\tau_a} - M_{\tau_a - 1}|^2$$
$$+ 2|M_{\tau_a - 1}| \cdot |M_{\tau_a} - M_{\tau_a - 1}| \leq (\Delta M_{\tau_a})^2 + 2a^{1/2}|\Delta M_{\tau_a}|,$$

whence

$$\mathsf{E}|\Delta M_{\tau_a}^2| I\{\tau_a < \infty\} \leq \mathsf{E}(\Delta M_{\tau_a})^2 I\{\tau_a < \infty\} + 2a^{1/2}\sqrt{\mathsf{E}(\Delta M_{\tau_a})^2 I\{\tau_a < \infty\}}$$
$$\leq \mathsf{E} \sup |\Delta M_n|^2 + 2a^{1/2}\sqrt{\mathsf{E} \sup |\Delta M_n|^2} < \infty.$$

This completes the proof of the theorem.

As an illustration of this theorem, we present the following result, which can be considered as a distinctive version of the strong law of large numbers for square-integrable martingales (compare Theorem 2 of §3, Chapter IV and Corollary 2 of Subsection 3 of §3).

Theorem 4. *Let $M = (M_n, \mathscr{F}_n)$ be a square-integrable martingale and let $A = (A_n, \mathscr{F}_{n-1})$ be a predictable increasing sequence with $A_1 \geq 1$, $A_\infty = \infty$ (P-a.s.).*

If (P-a.s.)

$$\sum_{i=1}^{\infty} \frac{\mathsf{E}[(\Delta M_i)^2 | F_{i-1}]}{A_i^2} < \infty, \tag{16}$$

then

$$M_n/A_n \to 0, \qquad n \to \infty, \tag{17}$$

with probability 1.

In particular, if $\langle M \rangle = (M_n, \mathscr{F}_n)$ is the quadratic characteristic of the square-integrable martingale, $M = (M_n, \mathscr{F}_n)$ and $\langle M \rangle_\infty = \infty$ (P-a.s.), then with probability 1

$$\frac{M_n}{\langle M \rangle_n} \to 0, \qquad n \to \infty. \tag{18}$$

Proof. Consider the square-integrable martingale $m = (m_n, \mathscr{F}_n)$ with

$$m_n = \sum_{i=1}^{n} \frac{\Delta M_i}{A_i}.$$

Then

$$\langle m \rangle_n = \sum_{i=1}^{n} \frac{\mathsf{E}[(\Delta M_i)^2 | \mathscr{F}_{i-1}]}{A_i^2}. \tag{19}$$

Since

$$\frac{M_n}{A_n} = \frac{\sum_{k=1}^{n} A_k \Delta m_k}{A_n},$$

we have, by Kronecker's lemma (§3, Chapter IV), $M_n/A_n \to 0$ (P-a.s.) if the limit $\lim_n m_n$ exists (finite) with probability 1. By (13),

$$\{\langle m \rangle_\infty < \infty\} \subseteq \{m_n \to\}. \tag{20}$$

Therefore it follows from (19) that (16) is a sufficient condition for (17).

If now $A_n = \langle M \rangle_n$, then (16) is automatically satisfied (see Problem 6) and consequently we have

$$\frac{M_n}{\langle M \rangle_n} \to 0 \quad (\text{P-a.s.}).$$

This completes the proof of the theorem.

Example. Consider a sequence $\xi_1, \xi_2, \ldots$ of independent random variables with $\mathsf{E}\xi_i = 0$, $\mathsf{V}\xi_i = V_i > 0$, and let the sequence $X = \{X_n\}_{n \geq 0}$ be defined recursively by

$$X_{n+1} = \theta X_n + \xi_{n+1}, \tag{21}$$

where X_0 is independent of $\xi_1, \xi_2, \ldots$ and θ is an unknown parameter, $-\infty < \theta < \infty$.

We interpret X_n as the result of an observation made at time n and ask for an estimator of the unknown parameter θ. As an estimator of θ in terms of $X_0, X_1, \ldots, X_n$, we take

$$\hat{\theta}_n = \frac{\displaystyle\sum_{k=0}^{n-1} \frac{X_k X_{k+1}}{V_{k+1}}}{\displaystyle\sum_{k=0}^{n-1} \frac{X_k^2}{V_{k+1}}}, \tag{22}$$

taking this to be 0 if the denominator is 0. (The number θ_n is the *least-squares* estimator.)

It is clear from (21) and (22) that

$$\hat{\theta} = \theta + \frac{M_n}{A_n},$$

where

$$M_n = \sum_{k=0}^{n-1} \frac{X_k \xi_{k+1}}{V_{k+1}}, \qquad A_n = \langle M \rangle_n = \sum_{k=0}^{n-1} \frac{X_k^2}{V_{k+1}}.$$

Therefore if the true value of the unknown parameter is θ, then

$$\mathsf{P}(\hat{\theta}_n \to \theta) = 1, \tag{23}$$

when (P-a.s.)

$$\frac{M_n}{A_n} \to 0, \qquad n \to \infty. \tag{24}$$

(An estimator θ_n with property (23) is said to be *strongly consistent*; compare the notion of consistency in §7, Chapter I.) Let us show that the conditions

$$\sup_n \frac{V_{n+1}}{V_n} < \infty, \qquad \sum_{n=1}^{\infty} \mathsf{E}\left(\frac{\xi_n^2}{V_n} \wedge 1\right) = \infty \tag{25}$$

are sufficient for (24), and therefore sufficient for (23).

We have

$$\sum_{n=1}^{\infty} \left(\frac{\xi_n^2}{V_n} \wedge 1\right) \le \sum_{n=1}^{\infty} \frac{\xi_n^2}{V_n} = \sum_{n=1}^{\infty} \frac{(X_n - \theta X_{n-1})^2}{V_n}$$

$$\le 2\left[\sum_{n=1}^{\infty} \frac{X_n^2}{V_n} + \theta^2 \sum_{n=1}^{\infty} \frac{X_{n-1}^2}{V_n}\right] \le 2\left[\sup \frac{V_{n+1}}{V_n} + \theta^2\right] \langle M \rangle_\infty.$$

Therefore

$$\left\{\sum_{n=1}^{\infty} \left(\frac{\xi_n^2}{V_n} \wedge 1\right) = \infty\right\} \subseteq \{\langle M \rangle_\infty = \infty\}.$$

By the three-series theorem (Theorem 3 of §2, Chapter IV) the divergence of $\sum_{n=1}^{\infty} \mathsf{E}((\xi_n^2/V_n) \wedge 1)$ guarantees the divergence (P-a.s.) of $\sum_{n=1}^{\infty} ((\xi^2/V_n) \wedge 1)$. Therefore $\mathsf{P}\{\langle M \rangle_\infty = \infty\} = 1$. Moreover, if

$$m_n = \sum_{i=1}^{n} \frac{\Delta M_i}{\langle M \rangle_i},$$

then

$$\langle m \rangle_n = \sum_{i=1}^{n} \frac{\Delta \langle M \rangle_i}{\langle M \rangle_i^2}$$

and (see Problem 6) $\mathsf{P}(m_\infty < \infty) = 1$. Hence (24) follows directly from Theorem 4.

(In Subsection 5 of the next section we continue the discussion of this example for *Gaussian* variables $\xi_1, \xi_2, \ldots$.)

Theorem 5. *Let $X = (X_n, \mathscr{F}_n)$ be a submartingale, and let*

$$X_n = m_n + A_n$$

be its Doob decomposition. If $|\Delta X_n| \le C$, then (P-a.s.)

$$\{\langle m \rangle_\infty + A_\infty < \infty\} = \{X_n \to\}, \tag{26}$$

or equivalently,

$$\left\{\sum_{n=1}^{\infty} \mathsf{E}[\Delta X_n + (\Delta X_n)^2 | \mathscr{F}_{n-1}] < \infty\right\} = \{X_n \to\}. \tag{27}$$

PROOF. Since

$$A_n = \sum_{k=1}^{n} \mathsf{E}(\Delta X_k | \mathscr{F}_{k-1}), \tag{28}$$

and

$$m_n = \sum_{k=1}^{n} [\Delta X_k - \mathsf{E}(\Delta X_k | \mathscr{F}_{k-1})], \tag{29}$$

it follows from the assumption that $|\Delta X_k| \le C$ that the martingale $m = (m_n, \mathscr{F}_n)$ is square-integrable with $|\Delta m_n| \le 2C$. Then by (13)

$$\{\langle m \rangle_\infty + A_\infty < \infty\} \subseteq \{X_n \to\} \tag{30}$$

and according to (8)

$$\{X_n \to\} \subseteq \{A_\infty < \infty\}.$$

Therefore, by (14) and (20),

$$\begin{aligned}\{X_n \to\} &= \{X_n \to\} \cap \{A_\infty < \infty\} = \{X_n \to\} \cap \{A_\infty < \infty\} \cap \{m_n \to\} \\ &= \{X_n \to\} \cap \{A_\infty < \infty\} \cap \{\langle m \rangle_\infty < \infty\} \\ &= \{X_n \to\} \cap \{A_\infty + \langle m \rangle_\infty < \infty\} = \{A_\infty + \langle m \rangle_\infty < \infty\}.\end{aligned}$$

Finally, the equivalence of (26) and (27) follows because, by (29),

$$\langle m \rangle_n = \sum \{\mathsf{E}[(\Delta X_k)^2 | \mathscr{F}_{k-1}] - [\mathsf{E}(\Delta X_k | \mathscr{F}_{k-1})]^2\},$$

and the convergence of the series $\sum_{k=1}^{\infty} \mathsf{E}(\Delta X_k | \mathscr{F}_{k-1})$ of nonnegative terms implies the convergence of $\sum_{k=1}^{\infty} [\mathsf{E}(\Delta X_k | \mathscr{F}_{k-1})]^2$. This completes the proof.

4. Kolmogorov's three-series theorem (Theorem 3 of §2, Chapter IV) gives a necessary and sufficient condition for the convergence, with probability 1, of a series $\sum \xi_n$ of independent random variables. The following theorems, whose proofs are based on Theorems 2 and 3, describe sets of convergence of $\sum \xi_n$ without the assumption that the random variables $\xi_1, \xi_2, \ldots$ are independent.

Theorem 6. *Let $\xi = (\xi_n, \mathscr{F}_n)$, $n \geq 1$, be a stochastic sequence, let $\mathscr{F}_0 = \{\varnothing, \Omega\}$, and let c be a positive constant. Then the series $\sum \xi_n$ converges on the set A of sample points for which the three series*

$$\sum \mathsf{P}(|\xi_n| \geq c | \mathscr{F}_{n-1}), \qquad \sum \mathsf{E}(\xi_n^c | \mathscr{F}_{n-1}), \qquad \sum \mathsf{V}(\xi_n^c | \mathscr{F}_{n-1})$$

converge, where $\xi_n^c = \xi_n I(|\xi_n| \leq c)$.

PROOF. Let $X_n = \sum_{k=1}^n \xi_k$. Since the series $\sum \mathsf{P}(|\xi_n| \geq c | \mathscr{F}_{n-1})$ converges, by Corollary 2 of Theorem 2, and by the convergence of the series $\sum \mathsf{E}(\xi_n^c | \mathscr{F}_{n-1})$, we have

$$A \cap \{X_n \to\} = A \cap \left\{ \sum_{k=1}^n \xi_k I(|\xi_k| \leq c) \to \right\}$$

$$= A \cap \left\{ \sum_{k=1}^n [\xi_k I(|\xi_k| \leq c) - \mathsf{E}(\xi_k I(|\xi_k| \leq c) | \mathscr{F}_{k-1})] \to \right\}. \tag{31}$$

Let $\eta_k = \xi_k I(|\xi_k| \leq c) - \mathsf{E}(\xi_k I(|\xi_k| \leq c) | \mathscr{F}_{k-1})$ and let $Y_n = \sum_{k=1}^n \eta_k$. Then $Y = (Y_n, \mathscr{F}_n)$ is a square-integrable martingale with $|\eta_k| \leq 2c$. By Theorem 3, we have

$$A \subseteq \{\sum \mathsf{V}(\xi_n^c | \mathscr{F}_{n-1}) < \infty\} = \{\langle Y \rangle_\infty < \infty\} = \{Y_n \to\}. \tag{32}$$

Then it follows from (31) that

$$A \cap \{X_n \to\} = A,$$

and therefore $A \subseteq \{X_n \to\}$. This completes the proof.

5. Problems

1. Show that if a submartingale $X = (X_n, \mathscr{F}_n)$ satisfies $\mathsf{E} \sup_n |X_n| < \infty$, then it belongs to class C^+.

2. Show that Theorems 1 and 2 remain valid for generalized submartingales.

3. Show that generalized submartingales satisfy (P-a.s.) the inclusion

$$\left\{ \inf_m \sup_{n \geq m} \mathsf{E}(X_n^+ | \mathscr{F}_m) < \infty \right\} \subseteq \{X_n \to\}.$$

4. Show that the corollary to Theorem 1 remains valid for generalized martingales.

5. Show that every generalized submartingale of class C^+ is a local submartingale.

6. Let $a_n > 0$, $n \geq 1$, and let $b_n = \sum_{k=1}^n a_k$. Show that

$$\sum_{n=1}^\infty \frac{a_n}{b_n^2} < \infty.$$

§6. Absolute Continuity and Singularity of Probability Distributions

1. Let $(\Omega, \mathscr{F})$ be a measurable space on which there is defined a family $(\mathscr{F}_n)_{n\geq 1}$ of σ-algebras such that $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \cdots \subseteq \mathscr{F}$ and

$$\mathscr{F} = \sigma\left(\bigcup_{n=1}^{\infty} \mathscr{F}_n\right). \tag{1}$$

Let us suppose that two probability measures P and $\tilde{\mathsf{P}}$ are given on $(\Omega, \mathscr{F})$. Let us write

$$\mathsf{P}_n = \mathsf{P}\,|\,\mathscr{F}_n, \qquad \tilde{\mathsf{P}}_n = \tilde{\mathsf{P}}\,|\,\mathscr{F}_n$$

for the restrictions of these measures to $\mathscr{F}_n$, i.e., let P_n and $\tilde{\mathsf{P}}_n$ be measures on $(\Omega, \mathscr{F}_n)$ and for $B \in \mathscr{F}_n$ let

$$\mathsf{P}_n(B) = \mathsf{P}(B), \qquad \tilde{\mathsf{P}}_n(B) = \tilde{\mathsf{P}}(B).$$

Definition 1. The probability measure $\tilde{\mathsf{P}}$ is *absolutely continuous* with respect to P (notation, $\tilde{\mathsf{P}} \ll \mathsf{P}$) if $\tilde{\mathsf{P}}(A) = 0$ whenever $\mathsf{P}(A) = 0$, $A \in \mathscr{F}$.

When $\tilde{\mathsf{P}} \ll \mathsf{P}$ and $\mathsf{P} \ll \tilde{\mathsf{P}}$ the measures $\tilde{\mathsf{P}}$ and P are *equivalent* (notation, $\tilde{\mathsf{P}} \sim \mathsf{P}$).

The measures $\tilde{\mathsf{P}}$ and P are *singular* (or *orthogonal*) if there is a set $A \in \mathscr{F}$ such that $\tilde{\mathsf{P}}(A) = 1$ and $\mathsf{P}(\bar{A}) = 1$ (notation, $\tilde{\mathsf{P}} \perp \mathsf{P}$).

Definition 2. We say that $\tilde{\mathsf{P}}$ is *locally absolutely continuous* with respect to P (notation, $\tilde{\mathsf{P}} \overset{\text{loc}}{\ll} \mathsf{P}$) if

$$\tilde{\mathsf{P}}_n \ll \mathsf{P}_n \tag{2}$$

for every $n \geq 1$.

The fundamental question that we shall consider in this section is the determination of conditions under which local absolute continuity $\tilde{\mathsf{P}} \overset{\text{loc}}{\ll} \mathsf{P}$ implies one of the properties $\tilde{\mathsf{P}} \ll \mathsf{P}$, $\tilde{\mathsf{P}} \sim \mathsf{P}$, $\tilde{\mathsf{P}} \perp \mathsf{P}$. It will become clear that martingale theory is the mathematical apparatus that lets us give definitive answers to these questions.

Let us then suppose that $\tilde{\mathsf{P}} \overset{\text{loc}}{\ll} \mathsf{P}$. We write

$$z_n = \frac{d\tilde{\mathsf{P}}_n}{d\mathsf{P}_n}$$

the Radon–Nikodým derivative of $\tilde{\mathsf{P}}_n$ with respect to P_n. It is clear that z_n is $\mathscr{F}_n$-measurable; and if $A \in \mathscr{F}_n$ then

$$\begin{aligned}\int_A z_{n+1}\, d\mathsf{P} &= \int_A \frac{d\tilde{\mathsf{P}}_{n+1}}{d\mathsf{P}_{n+1}}\, d\mathsf{P} = \tilde{\mathsf{P}}_{n+1}(A) = \tilde{\mathsf{P}}_n(A) \\ &= \int_A \frac{d\tilde{\mathsf{P}}_n}{d\mathsf{P}_n}\, d\mathsf{P} = \int_A z_n\, d\mathsf{P}.\end{aligned}$$

It follows that, *with respect to* P, *the stochastic sequence* $Z = (z_n, \mathscr{F}_n)_{n\geq 1}$ *is a martingale.*

Write

$$z_\infty = \overline{\lim}\, z_n.$$

Since $\mathsf{E}z_n = 1$, it follows from Theorem 1, §4, that $\lim z_n$ exists P-a.s. and therefore $\mathsf{P}(z_\infty = \lim z_n) = 1$. (In the course of the proof of Theorem 1 it will be established that $\lim z_n$ exists also for $\tilde{\mathsf{P}}$, so that $\tilde{\mathsf{P}}(z_\infty = \lim z_n) = 1$.)

The key to problems on absolute continuity and singularity is *Lebesgue's decomposition.*

Theorem 1. *Let* $\tilde{\mathsf{P}} \overset{\text{loc}}{\ll} \mathsf{P}$. *Then for every* $A \in \mathscr{F}$,

$$\tilde{\mathsf{P}}(A) = \int_A z_\infty \, d\mathsf{P} + \tilde{\mathsf{P}}\{A \cap (z_\infty = \infty)\}, \tag{3}$$

and the measures $\mu(A) = \tilde{\mathsf{P}}\{A \cap (z_\infty = \infty)\}$ *and* $\mathsf{P}(A)$, $A \in \mathscr{F}$, *are singular.*

Proof. Let us notice first that the classical *Lebesgue decomposition* shows that if P and $\tilde{\mathsf{P}}$ are two measures, there are unique measures λ and μ such that $\tilde{\mathsf{P}} = \lambda + \mu$, where $\lambda \ll \mathsf{P}$ and $\mu \perp \mathsf{P}$. Conclusion (3) can be thought of as a specialization of this decomposition under the assumption that $\tilde{\mathsf{P}}_n \ll \mathsf{P}_n$, $n \geq 1$.

Let us introduce the probability measures

$$\mathsf{Q} = \tfrac{1}{2}(\mathsf{P} + \tilde{\mathsf{P}}), \qquad \mathsf{Q}_n = \tfrac{1}{2}(\mathsf{P}_n + \tilde{\mathsf{P}}_n), \qquad n \geq 1,$$

and the notation

$$\tilde{\mathfrak{z}} = \frac{d\tilde{\mathsf{P}}}{d\mathsf{Q}}, \qquad \mathfrak{z} = \frac{d\mathsf{P}}{d\mathsf{Q}}, \qquad \tilde{\mathfrak{z}}_n = \frac{d\tilde{\mathsf{P}}_n}{d\mathsf{Q}_n}, \qquad \mathfrak{z}_n = \frac{d\mathsf{P}_n}{d\mathsf{Q}_n}.$$

Since $\tilde{\mathsf{P}}(\tilde{\mathfrak{z}} = 0) = \mathsf{P}(\mathfrak{z} = 0) = 0$, we have $\mathsf{Q}(\tilde{\mathfrak{z}} = 0, \mathfrak{z} = 0) = 0$. Consequently the product $\tilde{\mathfrak{z}} \cdot \mathfrak{z}^{-1}$ can be defined consistently on the set $\Omega \backslash \{\tilde{\mathfrak{z}} = 0, \mathfrak{z} = 0\}$; we define it to be zero on the set $\{\tilde{\mathfrak{z}} = 0, \mathfrak{z} = 0\}$.

Since $\tilde{\mathsf{P}}_n \ll \mathsf{P}_n \ll \mathsf{Q}_n$, we have (see (II.7.36))

$$\frac{d\tilde{\mathsf{P}}_n}{d\mathsf{Q}_n} = \frac{d\tilde{\mathsf{P}}_n}{d\mathsf{P}_n} \cdot \frac{d\mathsf{P}_n}{d\mathsf{Q}_n} \quad (\mathsf{Q}\text{-a.s.}) \tag{4}$$

i.e.,

$$\tilde{\mathfrak{z}}_n = z_n \mathfrak{z}_n \quad (\mathsf{Q}\text{-a.s.}) \tag{5}$$

whence

$$z_n = \tilde{\mathfrak{z}}_n \cdot \mathfrak{z}_n^{-1} \quad (\mathsf{Q}\text{-a.s.})$$

where, as before, we take $\tilde{\mathfrak{z}}_n \cdot \mathfrak{z}_n^{-1} = 0$ on the set $\{\tilde{\mathfrak{z}}_n = 0, \mathfrak{z}_n = 0\}$, which is of Q-measure zero.

Each of the sequences $(\tilde{\mathfrak{z}}_n, \mathscr{F}_n)$ and $(\mathfrak{z}_n, \mathscr{F}_n)$ is (with respect to Q) a uniformly integrable martingale and consequently the limits $\lim \tilde{\mathfrak{z}}_n$ and $\lim \mathfrak{z}_n$ exist. Moreover (Q-a.s.)

$$\lim \tilde{\mathfrak{z}}_n = \tilde{\mathfrak{z}}, \qquad \lim \mathfrak{z}_n = \mathfrak{z}. \tag{6}$$

From this and the equations $z_n = \tilde{\mathfrak{z}}_n \mathfrak{z}_n^{-1}$ (Q-a.s.) and $\mathsf{Q}(\tilde{\mathfrak{z}} = 0, \mathfrak{z} = 0) = 0$, it follows that (Q-a.s.) the limit $\lim z_n = z_\infty$ exists and is equal to $\tilde{\mathfrak{z}} \cdot \mathfrak{z}^{-1}$.

It is clear that $\mathsf{P} \ll \mathsf{Q}$ and $\tilde{\mathsf{P}} \ll \mathsf{Q}$. Therefore $\lim z_n$ exists both with respect to P and with respect to $\tilde{\mathsf{P}}$.

Now let

$$\lambda(A) = \int_A z_\infty \, d\mathsf{P}, \qquad \mu(A) = \tilde{\mathsf{P}}\{A \cap (z_\infty = \infty)\}.$$

To establish (3), we must show that

$$\tilde{\mathsf{P}}(A) = \lambda(A) + \mu(A), \qquad \lambda \ll \mathsf{P}, \quad \mu \perp \mathsf{P}.$$

We have

$$\begin{aligned}\tilde{\mathsf{P}}(A) &= \int_A \tilde{\mathfrak{z}} \, d\mathsf{Q} = \int_A \tilde{\mathfrak{z}} \mathfrak{z} \overset{\oplus}{\mathfrak{z}} \, d\mathsf{Q} + \int_A \tilde{\mathfrak{z}} [1 - \mathfrak{z} \overset{\oplus}{\mathfrak{z}}] \, d\mathsf{Q} \\ &= \int_A \tilde{\mathfrak{z}} \overset{\oplus}{\mathfrak{z}} \, d\mathsf{P} + \int_A [1 - \mathfrak{z} \overset{\oplus}{\mathfrak{z}}] \, d\tilde{\mathsf{P}} = \int_A z_\infty \, d\mathsf{P} + \tilde{\mathsf{P}}\{A \cap (\mathfrak{z} = 0)\}, \end{aligned} \tag{7}$$

where the last equation follows from

$$\mathsf{P}\left\{\overset{\oplus}{\mathfrak{z}} = \mathfrak{z}^{-1}\right\} = 1, \qquad \tilde{\mathsf{P}}\{z_\infty = \tilde{\mathfrak{z}} \cdot \mathfrak{z}^{-1}\} = 1.$$

Furthermore,

$$\begin{aligned}\tilde{\mathsf{P}}\{A \cap (\mathfrak{z} = 0)\} &= \tilde{\mathsf{P}}\{A \cap (\mathfrak{z} = 0) \cap (\tilde{\mathfrak{z}} > 0)\} \\ &= \tilde{\mathsf{P}}\{A \cap (\tilde{\mathfrak{z}} \cdot \mathfrak{z}^{-1} = \infty)\} = \tilde{\mathsf{P}}\{A \cap (z_\infty = \infty)\},\end{aligned}$$

which, together with (7), establishes (3).

It is clear from the construction of λ that $\lambda \ll \mathsf{P}$ and that $\mathsf{P}(z_\infty < \infty) = 1$. But we also have

$$\mu(z_\infty < \infty) = \tilde{\mathsf{P}}\{(z_\infty < \infty) \cap (z_\infty = \infty)\} = 0.$$

Consequently the theorem is proved.

The Lebesgue decomposition (3) implies the following useful tests for absolute continuity or singularity for locally absolutely continuous probability measures.

Theorem 2. *Let* $\tilde{\mathsf{P}} \overset{\text{loc}}{\ll} \mathsf{P}$, *i.e.* $\tilde{\mathsf{P}}_n \ll \mathsf{P}_n$, $n \geq 1$. *Then*

$$\tilde{\mathsf{P}} \ll \mathsf{P} \Leftrightarrow \mathsf{E}z_\infty = 1 \Leftrightarrow \tilde{\mathsf{P}}(z_\infty < \infty) = 1, \tag{8}$$

$$\tilde{\mathsf{P}} \perp \mathsf{P} \Leftrightarrow \mathsf{E}z_\infty = 0 \Leftrightarrow \tilde{\mathsf{P}}(z_\infty = \infty) = 1, \tag{9}$$

where E *denotes averaging with respect to* P.

PROOF. Putting $A = \Omega$ in (3), we find that

$$\mathsf{E}z_\infty = 1 \Leftrightarrow \tilde{\mathsf{P}}(z_\infty = \infty) = 0, \tag{10}$$

$$\mathsf{E}z_\infty = 0 \Leftrightarrow \tilde{\mathsf{P}}(z_\infty = \infty) = 1. \tag{11}$$

If $\tilde{\mathsf{P}}(z_\infty = \infty) = 0$, it again follows from (3) that $\tilde{\mathsf{P}} \ll \mathsf{P}$.

Conversely, let $\tilde{\mathsf{P}} \ll \mathsf{P}$. Then since $\mathsf{P}(z_\infty = \infty) = 0$, we have $\tilde{\mathsf{P}}(z_\infty = \infty) = 0$.

In addition, if $\tilde{\mathsf{P}} \perp \mathsf{P}$ there is a set $B \in \mathscr{F}$ with $\tilde{\mathsf{P}}(B) = 1$ and $\mathsf{P}(B) = 0$. Then $\tilde{\mathsf{P}}(B \cap (z_\infty = \infty)) = 1$ by (3), and therefore $\tilde{\mathsf{P}}(z_\infty = \infty) = 1$. If, on the other hand, $\tilde{\mathsf{P}}(z_\infty = \infty) = 1$ the property $\tilde{\mathsf{P}} \perp \mathsf{P}$ is evident, since $\mathsf{P}(z_\infty = \infty) = 0$.

This completes the proof of the theorem.

2. It is clear from Theorem 2 that the tests for absolute continuity or singularity can be expressed either in terms of P (verify the equation $\mathsf{E}z_\infty = 1$ or $\mathsf{E}z_\infty = 0$), or in terms of $\tilde{\mathsf{P}}$ (verify that $\tilde{\mathsf{P}}(z_\infty < \infty) = 1$ or that $\tilde{\mathsf{P}}(z_\infty = \infty) = 1$).

By Theorem 5 of §6, Chapter II, the condition $\mathsf{E}z_\infty = 1$ is equivalent to the uniform integrability (with respect to P) of the family $\{z_n\}_{n \geq 1}$. This allows us to give simple *sufficient conditions for the absolute continuity* $\tilde{\mathsf{P}} \ll \mathsf{P}$. For example, if

$$\sup_n \mathsf{E}[z_n \ln^+ z_n] < \infty \tag{12}$$

or if

$$\sup_n \mathsf{E}z_n^{1+\varepsilon} < \infty, \qquad \varepsilon > 0, \tag{13}$$

then, by Lemma 3 of §6, Chapter II, the family of random variables $\{z_n\}_{n \geq 1}$ is uniformly integrable and therefore $\tilde{\mathsf{P}} \ll \mathsf{P}$.

In many cases it is preferable to verify the property of absolute continuity or of singularity by using a test in terms of $\tilde{\mathsf{P}}$, since then the question is reduced to the investigation of the probability of the "tail" event $\{z_\infty < \infty\}$, where one can use propositions like the "zero-one" law.

Let us show, by way of illustration, that the "Kakutani dichotomy" can be deduced from Theorem 2.

Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space, let $(R^\infty, \mathscr{B}_\infty)$ be a measurable space of sequences $x = (x_1, x_2, \ldots)$ of numbers with $\mathscr{B}_\infty = \mathscr{B}(R^\infty)$, and let $\mathscr{B}_n = \sigma\{x: (x_1, \ldots, x_n)\}$. Let $\xi = (\xi_1, \xi_2, \ldots)$ and $\tilde{\xi} = (\tilde{\xi}_1, \tilde{\xi}_2, \ldots)$ be sequences of independent random variables.

Let P and $\tilde{\mathsf{P}}$ be the probability distributions on $(R^\infty, \mathscr{B}_\infty)$ for ξ and $\tilde{\xi}$, respectively, i.e.

$$P(B) = \mathsf{P}\{\xi \in B\}, \qquad \tilde{P}(B) = \mathsf{P}\{\tilde{\xi} \in B\}, \qquad B \in \mathscr{B}_\infty.$$

Also let

$$P_n = P|\mathscr{B}_n, \qquad \tilde{P}_n = \tilde{P}|\mathscr{B}_n$$

be the restrictions of P and $\tilde{P}$ to $\mathscr{B}_n$ and let

$$P_{\xi_n}(A) = \mathsf{P}(\xi_n \in A), \qquad P_{\tilde{\xi}_n}(A) = \mathsf{P}(\tilde{\xi}_n \in A), \qquad A \in \mathscr{B}(R^1).$$

Theorem 3 (Kakutani Dichotomy). *Let* $\xi = (\xi_1, \xi_2, \ldots)$ *and* $\tilde{\xi} = (\tilde{\xi}_1, \tilde{\xi}_2, \ldots)$ *be sequences of independent random variables for which*

$$P_{\tilde{\xi}_n} \ll P_{\xi_n}, \qquad n \geq 1. \tag{14}$$

Then either $\tilde{\mathsf{P}} \ll \mathsf{P}$ *or* $\tilde{\mathsf{P}} \perp \mathsf{P}$.

PROOF. Condition (14) is evidently equivalent to $\tilde{\mathsf{P}}_n \ll \mathsf{P}_n$, $n \geq 1$, i.e. $\tilde{\mathsf{P}} \overset{\text{loc}}{\ll} \mathsf{P}$. It is clear that

$$z_n = \frac{d\tilde{P}_n}{dP_n} = q_1(x_1) \cdots q_n(x_n),$$

where

$$q_i(x_i) = \frac{dP_{\tilde{\xi}_i}}{dP_{\xi_i}}(x_i). \tag{15}$$

Consequently

$$\{x: z_\infty < \infty\} = \{x: \ln z_\infty < \infty\} = \left\{x: \sum_{i=1}^{\infty} \ln q_i(x_i) < \infty\right\}.$$

The event $\{x: \sum_{i=1}^\infty \ln q_i(x_i) < \infty\}$ is a tail event. Therefore, by the Kolmogorov zero-one law (Theorem 1 of §1, Chapter IV) the probability $\tilde{\mathsf{P}}\{x: z_\infty < \infty\}$ has only two values (0 or 1), and therefore by Theorem 2 either $\tilde{\mathsf{P}} \perp \mathsf{P}$ or $\tilde{\mathsf{P}} \ll \mathsf{P}$.

This completes the proof of the theorem.

3. The following theorem provides, in "predictable" terms, a test for absolute continuity or singularity.

Theorem 4. *Let* $\tilde{\mathsf{P}} \overset{\text{loc}}{\ll} \mathsf{P}$ *and let*

$$\alpha_n = z_n z_{n-1}^{\oplus}, \qquad n \geq 1,$$

with $z_0 = 1$. *Then (with* $\mathscr{F}_0 = \{\varnothing, \Omega\}$*)*

$$\tilde{\mathsf{P}} \ll \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\left\{\sum_{n=1}^{\infty} [1 - \mathsf{E}(\sqrt{\alpha_n}|\mathscr{F}_{n-1})] < \infty\right\} = 1, \tag{16}$$

$$\tilde{\mathsf{P}} \perp \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\left\{\sum_{n=1}^{\infty} [1 - \mathsf{E}(\sqrt{\alpha_n}|\mathscr{F}_{n-1})] = \infty\right\} = 1. \tag{17}$$

PROOF. Since

$$\tilde{\mathsf{P}}_n\{z_n = 0\} = \int_{\{z_n=0\}} z_n \, d\mathsf{P} = 0,$$

we have (P-a.s.)

$$z_n = \prod_{k=1}^{n} \alpha_k = \exp\left\{\sum_{k=1}^{n} \ln \alpha_k\right\}. \tag{18}$$

Putting $A = \{z_\infty = 0\}$ in (3), we find that $\tilde{\mathsf{P}}\{z_\infty = 0\} = 0$. Therefore, by (18), we have ($\tilde{\mathsf{P}}$-a.s.)

$$\begin{aligned}\{z_\infty < \infty\} = \{0 < z_\infty < \infty\} &= \{0 < \lim z_n < \infty\}\\ &= \left\{-\infty < \lim \sum_{k=1}^{n} \ln \alpha_k < \infty\right\}.\end{aligned} \tag{19}$$

Let us introduce the function

$$u(x) = \begin{cases} x, & |x| \le 1, \\ \operatorname{sign} x, & |x| > 1. \end{cases}$$

Then

$$\left\{-\infty < \lim \sum_{k=1}^{n} \ln \alpha_k < \infty\right\} = \left\{-\infty < \lim \sum_{k=1}^{n} u(\ln \alpha_k) < \infty\right\}. \tag{20}$$

Let $\tilde{\mathsf{E}}$ denote averaging with respect to $\tilde{\mathsf{P}}$ and let η be an $\mathscr{F}_n$-measurable integrable random variable. It follows from the properties of conditional expectations (Problem 4) that

$$z_{n-1}\tilde{\mathsf{E}}(\eta|\mathscr{F}_{n-1}) = \mathsf{E}(\eta z_n|\mathscr{F}_{n-1}) \quad (\mathsf{P}\text{- and } \tilde{\mathsf{P}}\text{-a.s.}), \tag{21}$$

$$\tilde{\mathsf{E}}(\eta|\mathscr{F}_{n-1}) = z_{n-1}^{\oplus}\mathsf{E}(\eta z_n|\mathscr{F}_{n-1}) \quad (\tilde{\mathsf{P}}\text{-a.s.}). \tag{22}$$

Recalling that $\alpha_n = z_{n-1}^{\oplus} z_n$, we obtain the following useful formula from (22):

$$\tilde{\mathsf{E}}(\eta|\mathscr{F}_{n-1}) = \mathsf{E}(\alpha_n \eta|\mathscr{F}_{n-1}) \quad (\tilde{\mathsf{P}}\text{-a.s.}). \tag{23}$$

From this it follows, in particular, that

$$\mathsf{E}(\alpha_n|\mathscr{F}_{n-1}) = 1 \quad (\tilde{\mathsf{P}}\text{-a.s.}). \tag{24}$$

By (23),

$$\tilde{\mathsf{E}}[u(\ln \alpha_n)|\mathscr{F}_{n-1}] = \mathsf{E}[\alpha_n u(\ln \alpha_n)|\mathscr{F}_{n-1}] \quad (\tilde{\mathsf{P}}\text{-a.s.}).$$

Since $xu(\ln x) \ge x - 1$ for $x \ge 0$, we have, by (24),

$$\tilde{\mathsf{E}}[u(\ln \alpha_n)|\mathscr{F}_{n-1}] \ge 0 \quad (\tilde{\mathsf{P}}\text{-a.s.}).$$

It follows that the stochastic sequence $X = (X_n, \mathscr{F}_n)$ with

$$X_n = \sum_{k=1}^{n} u(\ln \alpha_k)$$

is a submartingale with respect to $\tilde{\mathsf{P}}$; and $|\Delta X_n| = |u(\ln \alpha_n)| \leq 1$.

Then, by Theorem 5 of §5, we have ($\tilde{\mathsf{P}}$-a.s.)

$$\left\{-\infty < \lim \sum_{k=1}^{n} u(\ln \alpha_k) < \infty\right\} = \left\{\sum_{k=1}^{\infty} \tilde{\mathsf{E}}[u(\ln \alpha_k) + u^2(\ln \alpha_k)|\mathscr{F}_{k-1}] < \infty\right\}. \tag{25}$$

Hence we find, by combining (19), (20), (22), and (25), that (P-a.s.)

$$\begin{aligned}\{z_\infty < \infty\} &= \left\{\sum_{k=1}^{\infty} \tilde{\mathsf{E}}[u(\ln \alpha_k) + u^2(\ln \alpha_k)|\mathscr{F}_{k-1}] < \infty\right\} \\ &= \left\{\sum_{k=1}^{\infty} \mathsf{E}[\alpha_k u(\ln \alpha_k) + \alpha_k u^2(\ln \alpha_k)|\mathscr{F}_{k-1}] < \infty\right\}\end{aligned}$$

and consequently, by Theorem 2,

$$\tilde{\mathsf{P}} \ll \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\left\{\sum_{k=1}^{\infty} \mathsf{E}[\alpha_k u(\ln \alpha_k) + \alpha_k u^2(\ln \alpha_k)|\mathscr{F}_{k-1}] < \infty\right\} = 1, \tag{26}$$

$$\tilde{\mathsf{P}} \perp \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\left\{\sum_{k=1}^{\infty} \mathsf{E}[\alpha_k u(\ln \alpha_k) + \alpha_k u^2(\ln \alpha_k)|\mathscr{F}_{k-1}] = \infty\right\} = 1. \tag{27}$$

We now observe that by (24),

$$\mathsf{E}[(1 - \sqrt{\alpha_n})^2|\mathscr{F}_{n-1}] = 2\mathsf{E}[1 - \sqrt{\alpha_n}|\mathscr{F}_{n-1}] \quad (\tilde{\mathsf{P}}\text{-a.s.})$$

and for $x \geq 0$ there are constants A and B $(0 < A < B < \infty)$ such that

$$A(1 - \sqrt{x})^2 \leq xu(\ln x) + xu^2(\ln x) + 1 - x \leq B(1 - \sqrt{x})^2. \tag{28}$$

Hence (16) and (17) follow from (26), (27) and (24), (28).

This completes the proof of the theorem.

Corollary 1. *If, for all $n \geq 1$, the σ-algebras $\sigma(\alpha_n)$ and $\mathscr{F}_{n-1}$ are independent with respect to* P(*or* $\tilde{\mathsf{P}}$), *and* $\tilde{\mathsf{P}} \overset{\text{loc}}{\ll} \mathsf{P}$, *then we have the dichotomy*: either $\tilde{\mathsf{P}} \ll \mathsf{P}$ or $\tilde{\mathsf{P}} \perp \mathsf{P}$. Correspondingly,

$$\tilde{\mathsf{P}} \ll \mathsf{P} \Leftrightarrow \sum_{n=1}^{\infty} [1 - \mathsf{E}\sqrt{\alpha_n}] < \infty,$$

$$\tilde{\mathsf{P}} \perp \mathsf{P} \Leftrightarrow \sum_{n=1}^{\infty} [1 - \mathsf{E}\sqrt{\alpha_n}] = \infty.$$

In particular, in the Kakutani situation (see Theorem 3) $\alpha_n = q_n$ and

$$\tilde{\mathsf{P}} \ll \mathsf{P} \Leftrightarrow \sum_{n=1}^{\infty} [1 - \mathsf{E}\sqrt{q_n(x_n)}] < \infty,$$

$$\tilde{\mathsf{P}} \perp \mathsf{P} \Leftrightarrow \sum_{n=1}^{\infty} [1 - \mathsf{E}\sqrt{q_n(x_n)}] = \infty.$$

Corollary 2. *Let* $\tilde{\mathsf{P}} \overset{\text{loc}}{\ll} \mathsf{P}$. *Then*

$$\tilde{\mathsf{P}}\left\{\sum_{n=1}^{\infty} \mathsf{E}(\alpha_n \ln \alpha_n | \mathscr{F}_{n-1}) < \infty\right\} = 1 \Rightarrow \tilde{\mathsf{P}} \ll \mathsf{P}.$$

For the proof, it is enough to notice that

$$x \ln x + \tfrac{3}{2}(1 - x) \geq 1 - x^{1/2}, \tag{29}$$

for all $x \geq 0$, and apply (16) and (24).

Corollary 3. *Since the series* $\sum_{n=1}^{\infty} [1 - \mathsf{E}(\sqrt{\alpha_n} | \mathscr{F}_{n-1})]$, *which has nonnegative* ($\tilde{\mathsf{P}}$-*a.s.*) *terms, converges or diverges with the series* $\sum |\ln \mathsf{E}(\sqrt{\alpha_n} | \mathscr{F}_{n-1})|$, *conclusions* (16) *and* (17) *of Theorem* 4 *can be put in the form*

$$\tilde{\mathsf{P}} \ll \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\left\{\sum_{n=1}^{\infty} |\ln \mathsf{E}(\sqrt{\alpha_n} | \mathscr{F}_{n-1})| < \infty\right\} = 1, \tag{30}$$

$$\tilde{\mathsf{P}} \perp \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\left\{\sum_{n=1}^{\infty} |\ln \mathsf{E}(\sqrt{\alpha_n} | \mathscr{F}_{n-1})| = \infty\right\} = 1. \tag{31}$$

Corollary 4. *Let there exist constants A and B such that* $0 \leq A < 1, B \geq 0$ *and*

$$\mathsf{P}\{1 - A \leq \alpha_n \leq 1 + B\} = 1, \qquad n \geq 1.$$

Then if $\tilde{\mathsf{P}} \overset{\text{loc}}{\ll} \mathsf{P}$ *we have*

$$\tilde{\mathsf{P}} \ll \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\left\{\sum_{n=1}^{\infty} \mathsf{E}[(1 - \alpha_n)^2 | \mathscr{F}_{n-1}] < \infty\right\} = 1,$$

$$\tilde{\mathsf{P}} \perp \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\left\{\sum_{n=1}^{\infty} \mathsf{E}[(1 - \alpha_n)^2 | \mathscr{F}_{n-1}] = \infty\right\} = 1.$$

For the proof it is enough to notice that when $x \in [1 - A, 1 + B]$, where $0 \leq A < 1, B \geq 0$, there are constants c and C $(0 < c < C < \infty)$ such that

$$c(1 - x)^2 \leq (1 - \sqrt{x})^2 \leq C(1 - x)^2. \tag{32}$$

4. With the notation of Subsection 2, let us suppose that $\xi = (\xi_1, \xi_2, \ldots)$ and $\tilde{\xi} = (\tilde{\xi}_1, \tilde{\xi}_2, \ldots)$ are Gaussian sequences, $\tilde{\mathsf{P}}_n \sim \mathsf{P}_n$, $n \geq 1$. Let us show that, for such sequences, the "Hájek–Feldman dichotomy," *either* $\tilde{\mathsf{P}} \sim \mathsf{P}$ *or* $\tilde{\mathsf{P}} \perp \mathsf{P}$, follows from the "predictable" test given above.

By the theorem on normal correlation (Theorem 2 of §13, Chapter II) the conditional expectations $M(x_n | \mathscr{B}_{n-1})$ and $\tilde{M}(x_n | \mathscr{B}_{n-1})$, where M and $\tilde{M}$

are averages with respect to P and $\tilde{P}$, respectively, are linear functions of $x_1, \ldots, x_{n-1}$. We denote these linear functions by $a_{n-1}(x)$ and $\tilde{a}_{n-1}(x)$ and put

$$b_{n-1} = (M[x_n - a_{n-1}(x)]^2)^{1/2},$$
$$\tilde{b}_{n-1} = (\tilde{M}[x_n - \tilde{a}_{n-1}(x)]^2)^{1/2}.$$

Again by the theorem on normal correlation, there are sequences $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots)$ and $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \ldots)$ of independent Gaussian random variables with zero means and unit variances, such that

$$\begin{aligned} x_n &= a_{n-1}(x) + b_{n-1}\varepsilon_n, \quad (\mathsf{P}\text{-a.s.}), \\ x_n &= \tilde{a}_{n-1}(x) + \tilde{b}_{n-1}\tilde{\varepsilon}_n \quad (\tilde{\mathsf{P}}\text{-a.s.}). \end{aligned} \tag{33}$$

Notice that if $b_{n-1} = 0$, or $\tilde{b}_{n-1} = 0$, it is generally necessary to extend the probability space in order to construct ε_n or $\tilde{\varepsilon}_n$. However, if $b_{n-1} = 0$ the extended vector $(x_1, \ldots, x_n)$ will be contained (P-a.s.) in the linear manifold $x_n = a_{n-1}(x)$, and since by hypothesis $\tilde{\mathsf{P}}_n \sim \mathsf{P}_n$, we have $b_{n-1} = 0$, $a_{n-1} = \tilde{a}_{n-1}(x)$, and $\alpha_n(x) = 1$ (P- or $\tilde{\mathsf{P}}$-a.s.). Hence we may suppose without loss of generality that $b_n^2 > 0$, $\tilde{b}_n^2 > 0$ for all $n \geq 1$, since otherwise the contribution of the corresponding terms of the sum $\sum_{n=1}^{\infty}[1 - M\sqrt{\alpha_n}|B_{n-1}]$ (see (16) and (17)) is zero.

Using the Gaussian hypothesis, we find from (33) that, for $n \geq 1$,

$$\alpha_n = d_{n-1}^{-1} \exp\left\{-\frac{(x_n - a_{n-1}(x))^2}{2b_{n-1}^2} + \frac{(x_n - \tilde{a}_{n-1}(x))^2}{2\tilde{b}_{n-1}^2}\right\}, \tag{34}$$

where $d_n = |\tilde{b}_n \cdot \tilde{b}_n^{-1}|$ and

$$a_0(x) = \mathsf{E}\xi_1, \qquad \tilde{a}_0(x) = \mathsf{E}\tilde{\xi}_1,$$
$$b_0^2 = \mathsf{V}\xi_1, \qquad \tilde{b}_0^2 = \mathsf{V}\tilde{\xi}_1.$$

From (34),

$$\ln M(\alpha_n^{1/2}|\mathscr{B}_{n-1}) = \tfrac{1}{2}\ln\frac{2d_{n-1}}{1 + d_{n-1}^2} - \frac{d_{n-1}^2}{1 + d_{n-1}^2}\left(\frac{a_{n-1}(x) - \tilde{a}_{n-1}(x)}{b_{n-1}}\right)^2.$$

Since $\ln[2d_{n-1}/(1 + d_{n-1}^2)] \leq 0$, statement (30) can be written in the form

$$\tilde{P} \ll P \Leftrightarrow \tilde{P}\left\{\sum_{n=1}^{\infty}\left[\tfrac{1}{2}\ln\frac{1 + d_{n-1}^2}{2d_{n-1}} + \frac{d_{n-1}^2}{1 + d_{n-1}^2}\cdot\left(\frac{a_{n-1}(x) - \tilde{a}_{n-1}(x)}{b_{n-1}}\right)^2\right] < \infty\right\} = 1. \tag{35}$$

The series

$$\sum_{n=1}^{\infty}\ln\frac{1 + d_{n-1}^2}{2d_{n-1}} \quad \text{and} \quad \sum_{n=1}^{\infty}(d_{n-1}^2 - 1)$$

converge or diverge together; hence it follows from (35) that

$$\tilde{P} \ll P \Leftrightarrow \tilde{P}\left\{\sum_{n=0}^{\infty}\left[\left(\frac{\tilde{b}_n^2}{b_n^2}-1\right)^2+\frac{\Delta_n^2(x)}{b_n^2}\right]<\infty\right\}=1, \tag{36}$$

where $\Delta_n(x) = a_n(x) - \tilde{a}_n(x)$.

Since $a_n(x)$ and $\tilde{a}_n(x)$ are linear, the sequence of random variables $\{\Delta_n(x)/b_n\}_{n\geq 0}$ is a Gaussian system (with respect to both $\tilde{\mathsf{P}}$ and P). As follows from a lemma that will be proved below, such sequences satisfy an analog of the zero-one law:

$$\tilde{P}\left\{\sum\left(\frac{\Delta_n(x)}{b_n}\right)^2<\infty\right\}=1 \Leftrightarrow \sum \tilde{M}\left(\frac{\Delta_n(x)}{b_n}\right)^2<\infty. \tag{37}$$

Hence it follows from (36) that

$$\tilde{P} \ll P \Leftrightarrow \sum_{n=0}^{\infty}\left[\tilde{M}\left(\frac{\Delta_n(x)}{b_n}\right)^2+\left(\frac{\tilde{b}_n^2}{b_n^2}-1\right)^2\right]<\infty$$

and in a similar way

$$\tilde{P} \perp P \Leftrightarrow \sum_{n=0}^{\infty}\left[\tilde{M}\left(\frac{\Delta_n(x)}{b_n}\right)^2+\left(\frac{\tilde{b}_n^2}{b_n^2}-1\right)^2\right]=\infty.$$

Then it is clear that if $\tilde{\mathsf{P}}$ and P are not singular measures, we have $\tilde{\mathsf{P}} \ll \mathsf{P}$. But by hypothesis, $\tilde{\mathsf{P}}_n \sim \mathsf{P}_n$, $n \geq 1$; hence by symmetry we have $\mathsf{P} \ll \tilde{\mathsf{P}}$. Therefore we have the following theorem.

Theorem 5 (Hájek–Feldman Dichotomy). *Let $\xi = (\xi_1, \xi_2, \ldots)$ and $\tilde{\xi} = (\tilde{\xi}_1, \tilde{\xi}_2, \ldots)$ be Gaussian sequences whose finite-dimensional distributions are equivalent: $\tilde{P}_n \sim P_n$, $n \geq 1$. Then either $\tilde{P} \sim P$ or $\tilde{P} \perp P$. Moreover,*

$$\begin{aligned}\tilde{P} \sim P &\Leftrightarrow \sum_{n=0}^{\infty}\left[\tilde{M}\left(\frac{\Delta_n(x)}{b_n}\right)^2+\left(\frac{\tilde{b}_n^2}{b_n^2}-1\right)^2\right]<\infty,\\ \tilde{P} \perp P &\Leftrightarrow \sum_{n=0}^{\infty}\left[\tilde{M}\left(\frac{\Delta_n(x)}{b_n}\right)^2+\left(\frac{\tilde{b}_n^2}{b_n^2}-1\right)^2\right]=\infty.\end{aligned} \tag{38}$$

Let us now prove the zero-one law for Gaussian sequences that we need for the proof of Theorem 5.

Lemma. *Let $\beta = (\beta_n)_{n\geq 1}$ be a Gaussian sequence defined on $(\Omega, \mathscr{F}; \mathsf{P})$. Then*

$$\mathsf{P}\left\{\sum_{n=1}^{\infty}\beta_n^2<\infty\right\}=1 \Leftrightarrow \sum_{n=1}^{\infty}\mathsf{E}\beta_n^2<\infty. \tag{39}$$

Proof. The implication $\Leftarrow$ follows from Fubini's theorem. To establish the opposite proposition, we first suppose that $\mathsf{E}\beta_n = 0$, $n \geq 1$. Here it is enough to show that

$$\mathsf{E}\sum_{n=1}^{\infty}\beta_n^2 \leq \left[\mathsf{E}\exp\left(-\sum_{n=1}^{\infty}\beta_n^2\right)\right]^{-2}, \tag{40}$$

since then the condition $\mathsf{P}\{\sum \beta_n^2 < \infty\} = 1$ will imply that the right-hand side of (40) is finite.

Select an $n \geq 1$. Then it follows from §§11 and 13, Chapter II, that there are independent Gaussian random variables $\beta_{k,n}$, $k = 1, \ldots, r \leq n$, with $\mathsf{E}\beta_{k,n} = 0$, such that

$$\sum_{k=1}^{n} \beta_k^2 = \sum_{k=1}^{r} \beta_{k,n}^2.$$

If we write $\mathsf{E}\beta_{k,n}^2 = \lambda_{k,n}$, we easily see that

$$\mathsf{E} \sum_{k=1}^{r} \beta_{k,n}^2 = \sum_{k=1}^{r} \lambda_{k,n} \tag{41}$$

and

$$\mathsf{E} \exp\left(-\sum_{k=1}^{r} \beta_{k,n}^2\right) = \prod_{k=1}^{r} (1 + 2\lambda_{k,n})^{-1/2}. \tag{42}$$

Comparing the right-hand sides of (41) and (42), we obtain

$$\mathsf{E} \sum_{k=1}^{n} \beta_k^2 = \mathsf{E} \sum_{k=1}^{r} \beta_{k,n}^2 \leq \left[\mathsf{E} \exp\left(-\sum_{k=1}^{r} \beta_{k,n}^2\right)\right]^{-2} = \left[\mathsf{E} \exp\left(-\sum_{k=1}^{n} \beta_k^2\right)\right]^{-2},$$

from which, by letting $n \to \infty$, we obtain the required inequality (40).

Now suppose that $\mathsf{E}\beta_n \not\equiv 0$.

Let us consider again the sequence $\tilde{\beta} = (\tilde{\beta}_n)_{n \geq 1}$ with the same distribution as $\beta = (\beta_n)_{n \geq 1}$ but independent of it (if necessary, extending the original probability space). If $\mathsf{P}\{\sum_{n=1}^{\infty} \beta_n^2 < \infty\} = 1$, then $\mathsf{P}\{\sum_{n=1}^{\infty} (\beta_n - \tilde{\beta}_n)^2 < \infty\} = 1$, and by what we have proved,

$$2 \sum_{n=1}^{\infty} \mathsf{E}(\beta_n - \mathsf{E}\beta_n)^2 = \sum_{n=1}^{\infty} \mathsf{E}(\beta_n - \tilde{\beta}_n)^2 < \infty.$$

Since

$$(\mathsf{E}\beta_n)^2 \leq 2\beta_n^2 + 2(\beta_n - \mathsf{E}\beta_n)^2,$$

we have $\sum_{n=1}^{\infty} (\mathsf{E}\beta_n)^2 < \infty$ and therefore

$$\sum_{n=1}^{\infty} \mathsf{E}\beta_n^2 = \sum_{n=1}^{\infty} (\mathsf{E}\beta_n)^2 + \sum_{n=1}^{\infty} \mathsf{E}(\beta_n - M\beta_n)^2 < \infty.$$

This completes the proof of the lemma.

5. We continue the discussion of the example in Subsection 3 of the preceding section, assuming that $\xi_0, \xi_1, \ldots$ are independent Gaussian random variables with $\mathsf{E}\xi_i = 0$, $\mathsf{V}\xi_i = V_i > 0$.

Again we let

$$X_{n+1} = \theta X_n + \xi_{n+1}$$

for $n \geq 1$, where $X_0 = \xi_0$, and the unknown parameter θ that is to be estimated has values in R. Let $\hat{\theta}_n$ be the least-squares estimator (see (5.22)).

Theorem 6. *A necessary and sufficient condition for the estimator $\hat{\theta}_n$, $n \geq 1$, to be strongly consistent is that*

$$\sum_{n=0}^{\infty} \frac{V_n}{V_{n+1}} = \infty. \tag{43}$$

PROOF. *Sufficiency.* Let P_θ denote the probability distribution on $(R^\infty, \mathscr{B}_\infty)$ corresponding to the sequence $(X_0, X_1, \ldots)$ when the true value of the unknown parameter is θ. Let E_θ denote an average with respect to P_θ.

We have already seen that

$$\hat{\theta}_n = \theta + \frac{M_n}{\langle M \rangle_n},$$

where

$$\langle M \rangle_n = \sum_{k=0}^{n-1} \frac{X_k^2}{V_{k+1}}.$$

According to the lemma from the preceding subsection,

$$P_\theta(\langle M \rangle_\infty = \infty) = 1 \Leftrightarrow E_\theta \langle M \rangle_\infty = \infty,$$

i.e., $\langle M \rangle_\infty = \infty$ (P_θ-a.s.) if and only if

$$\sum_{k=0}^{\infty} \frac{E_\theta X_k^2}{V_{k+1}} = \infty. \tag{44}$$

But

$$E_\theta X_k^2 = \sum_{i=0}^{k} \theta^{2i} V_{k-i}$$

and

$$\begin{aligned} \sum_{k=0}^{\infty} \frac{E_\theta X_k^2}{V_{k+1}} &= \sum_{k=0}^{\infty} \frac{1}{V_{k+1}} \left(\sum_{i=0}^{k} \theta^{2i} V_{k-i} \right) \\ &= \sum_{k=0}^{\infty} \theta^{2k} \sum_{i=k}^{\infty} \frac{V_{i-k}}{V_{i+1}} = \sum_{i=0}^{\infty} \frac{V_i}{V_{i+1}} + \sum_{k=1}^{\infty} \theta^{2k} \left(\sum_{i=k}^{\infty} \frac{V_{i-k}}{V_{i+1}} \right). \end{aligned} \tag{45}$$

Hence (44) follows from (43) and therefore, by Theorem 4, the estimator $\hat{\theta}_n$, $n \geq 1$, is strongly consistent for every θ.

Necessity. For all $\theta \in R$, let $P_\theta(\hat{\theta}_n \to \theta) = 1$. It follows that if $\theta_1 \neq \theta_2$, the measures P_{θ_1} and P_{θ_2} are singular ($P_{\theta_1} \perp P_{\theta_2}$). In fact, since the sequence $(X_0, X_1, \ldots)$ is Gaussian, by Theorem 5 of §5 the measures P_{θ_1} and P_{θ_2} are either singular or equivalent. But they cannot be equivalent, since if $P_{\theta_1} \sim P_{\theta_2}$, but $P_{\theta_1}(\hat{\theta}_n \to \theta_1) = 1$, then also $P_{\theta_2}(\hat{\theta}_n \to \theta_1) = 1$. However, by hypothesis, $P_{\theta_2}(\hat{\theta}_n \to \theta_2) = 1$ and $\theta_2 \neq \theta_1$. Therefore $P_{\theta_1} \perp P_{\theta_2}$ for $\theta_1 \neq \theta_2$.

According to (5.38),

$$P_{\theta_1} \perp P_{\theta_2} \Leftrightarrow (\theta_1 - \theta_2)^2 \sum_{k=0}^{\infty} E_{\theta_1}\left[\frac{X_k^2}{V_{k+1}}\right] = \infty$$

for $\theta_1 \neq \theta_2$. Taking $\theta_1 = 0$ and $\theta_2 \neq 0$, we obtain from (45) that

$$P_0 \perp P_{\theta_2} \Leftrightarrow \sum_{i=0}^{\infty} \frac{V_i}{V_{i+1}} = \infty,$$

which establishes the necessity of (43).

This completes the proof of the theorem.

6. Problems

1. Prove (6).

2. Let $\tilde{\mathsf{P}}_n \sim \mathsf{P}_n$, $n \geq 1$. Show that

$$\tilde{\mathsf{P}} \sim \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\{z_\infty < \infty\} = \mathsf{P}\{z_\infty > 0\} = 1,$$
$$\tilde{\mathsf{P}} \perp \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\{z_\infty = \infty\} = 1 \quad \text{or} \quad \mathsf{P}\{z_\infty = 0\} = 1.$$

3. Let $\tilde{\mathsf{P}}_n \ll \mathsf{P}_n$, $n \geq 1$, let τ be a stopping time (with respect to $(\mathscr{F}_n)$), and let $\tilde{\mathsf{P}}_\tau = \tilde{\mathsf{P}} | \mathscr{F}_\tau$ and $\mathsf{P}_\tau = \mathsf{P} | \mathscr{F}_\tau$ be the restrictions of $\tilde{\mathsf{P}}$ and P to the σ-algebra $\mathscr{F}_\tau$. Show that $\tilde{\mathsf{P}}_\tau \ll \mathsf{P}_\tau$ if and only if $\{\tau = \infty\} = \{z_\infty < \infty\}$ ($\tilde{\mathsf{P}}$-a.s.). (In particular, if $\tilde{\mathsf{P}}\{\tau < \infty\} = 1$ then $\tilde{\mathsf{P}}_\tau \ll \mathsf{P}_\tau$.)

4. Prove (21) and (22).

5. Verify (28), (29), and (32).

6. Prove (34).

7. In Subsection 2 let the sequences $\xi = (\xi_1, \xi_2, \ldots)$ and $\tilde{\xi} = (\tilde{\xi}_1, \tilde{\xi}_2, \ldots)$ consist of independent identically distributed random variables. Show that if $P_{\tilde{\xi}_1} \ll P_{\xi_1}$, then $\tilde{P} \ll P$ if and only if the measures $P_{\tilde{\xi}_1}$ and P_{ξ_1} coincide. If, however, $P_{\tilde{\xi}_1} \ll P_{\xi_1}$ and $P_{\tilde{\xi}_1} \neq P_{\xi_1}$ then $\tilde{P} \perp P$.

§7. Asymptotics of the Probability of the Outcome of a Random Walk with Curvilinear Boundary

1. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables. Let $S_n = \xi_1 + \cdots + \xi_n$, let $g = g(n)$ be a "boundary," $n \geq 1$, and let

$$\tau = \inf\{n \geq 1 : S_n < g(n)\}$$

be the first time at which the random walk (S_n) is found below the boundary $g = g(n)$. (As usual, $\tau = \infty$ if $\{\cdot\} = \varnothing$.)

It is difficult to discover the exact form of the distribution of the time τ. In the present section we find the asymptotic form of the probability $\mathsf{P}(\tau > n)$ as $n \to \infty$, for a wide class of boundaries $g = g(n)$ and assuming that the ξ_i are normally distributed. The method of proof is based on the idea of an absolutely continuous change of measure together with a number of the properties of martingales and Markov times that were presented earlier.

Theorem 1. *Let $\xi_1, \xi_2, \ldots$ be independent identically distributed random variables, with $\xi_i \sim \mathcal{N}(0, 1)$. Suppose that $g = g(n)$ is such that $g(1) < 0$ and, for $n \geq 2$,*

$$0 \leq \Delta g(n + 1) \leq \Delta g(n), \tag{1}$$

where $\Delta g(n) = g(n) - g(n - 1)$ and

$$\ln n = o\left(\sum_{k=2}^{n} [\Delta g(k)]^2\right), \qquad n \to \infty. \tag{2}$$

Then

$$\mathsf{P}(\tau > n) = \exp\left\{-\frac{1}{2}\sum_{k=2}^{n} [\Delta g(k)]^2(1 + o(1))\right\}, \qquad n \to \infty. \tag{3}$$

Before starting the proof, let us observe that (1) and (2) are satisfied if, for example,

$$g(n) = an^{\nu} + b, \qquad \tfrac{1}{2} < \nu \leq 1, \quad a + b < 0,$$

or (for sufficiently large n)

$$g(n) = n^{\nu}L(n), \qquad \tfrac{1}{2} \leq \nu \leq 1,$$

where $L(n)$ is a slowly varying function (for example, $L(n) = C(\ln n)^{\beta}$ with arbitrary β for $\frac{1}{2} < \nu < 1$ or with $\beta > 0$ for $\nu = \frac{1}{2}$).

2. We shall need the following two auxiliary propositions for the proof of Theorem 1.

Let us suppose that $\xi_1, \xi_2, \ldots$ is a sequence of independent identically distributed random variables, $\xi_i \sim \mathcal{N}(0, 1)$. Let $\mathscr{F}_0 = \{\varnothing, \Omega\}$, $\mathscr{F}_n = \sigma\{\omega: \xi_1, \ldots, \xi_n\}$, and let $\alpha = (\alpha_n, \mathscr{F}_{n-1})$ be a predictable sequence with $\mathsf{P}(|\alpha_n| \leq C) = 1$, $n \geq 1$, where C is a constant. Form the sequence $z = (z_n, \mathscr{F}_n)$ with

$$z_n = \exp\left\{\sum_{k=1}^{n} \alpha_k \xi_k - \frac{1}{2}\sum_{k=1}^{n} \alpha_k^2\right\}, \qquad n \geq 1. \tag{4}$$

It is easily verified that (with respect to P) the sequence $z = (z_n, \mathscr{F}_n)$ is a martingale with $\mathsf{E}z_n = 1$, $n \geq 1$.

Choose a value $n \geq 1$ and introduce a probability measure $\tilde{\mathsf{P}}_n$ on the measurable space $(\Omega, \mathscr{F}_n)$ by putting

$$\tilde{\mathsf{P}}_n(A) = \mathsf{E}I(A)z_n, \qquad A \in \mathscr{F}_n. \tag{5}$$

Lemma 1. *With respect to* $\tilde{\mathsf{P}}_n$, *the random variables* $\tilde{\xi}_k = \xi_k - \alpha_k$, $1 \le k \le n$, *are independent and normally distributed,* $\tilde{\xi}_k \sim \mathcal{N}(0, 1)$.

PROOF. Let $\tilde{\mathsf{E}}_n$ denote averaging with respect to $\tilde{\mathsf{P}}_n$. Then for $\lambda_k \in R$, $1 \le k \le n$,

$$\tilde{\mathsf{E}}_n \exp\left\{i \sum_{k=1}^{n} \lambda_k \tilde{\xi}_k\right\} = \mathsf{E} \exp\left\{i \sum_{k=1}^{n} \lambda_k \tilde{\xi}_k\right\} z_n$$
$$= \mathsf{E}\left[\exp\left\{i \sum_{k=1}^{n-1} \lambda_k \tilde{\xi}_k\right\} z_{n-1} \cdot \mathsf{E}\left\{\exp\left(i\lambda_n(\xi_n - \alpha_n) + \alpha_n \xi_n - \frac{\alpha_n^2}{2}\right)\bigg|\mathscr{F}_{n-1}\right\}\right]$$
$$= \mathsf{E}\left[\exp\left\{i \sum_{k=1}^{n-1} \lambda_k \tilde{\xi}_k\right\} z_{n-1}\right] \exp\{-\tfrac{1}{2}\lambda_n^2\} = \cdots = \exp\left\{-\frac{1}{2}\sum_{k=1}^{n} \lambda_k^2\right\}.$$

Now the desired conclusion follows from Theorem 4 of §12, Chapter II.

Lemma 2. *Let* $X = (X_n, \mathscr{F}_n)_{n\ge 1}$ *be a square-integrable martingale with mean zero and*

$$\sigma = \inf\{n \ge 1 : X_n \le -b\},$$

where b is a constant, $b > 0$. *Suppose that*

$$\mathsf{P}(X_1 < -b) > 0.$$

Then there is a constant $C > 0$ *such that, for all* $n \ge 1$,

$$\mathsf{P}(\sigma > n) \ge \frac{C}{\mathsf{E}X_n^2}. \tag{6}$$

PROOF. By Corollary 1 to Theorem VII.2.1 we have $\mathsf{E}X_{\sigma\wedge n} = 0$, whence

$$-\mathsf{E}I(\sigma \le n)X_\sigma = \mathsf{E}I(\sigma > n)X_n. \tag{7}$$

On the set $\{\sigma \le n\}$

$$-X_\sigma \ge b > 0.$$

Therefore, for $n \ge 1$,

$$-\mathsf{E}I(\sigma \le n)X_\sigma \ge b\mathsf{P}(\sigma \le n) \ge b\mathsf{P}(\sigma = 1) = b\mathsf{P}(X_1 < -b) > 0. \tag{8}$$

On the other hand, by the Cauchy–Schwarz inequality,

$$\mathsf{E}I(\sigma > n)X_n \le [\mathsf{P}(\sigma > n)\cdot \mathsf{E}X_n^2]^{1/2}, \tag{9}$$

which, with (7) and (8), leads to the required inequality with

$$C = (b\mathsf{P}(X_1 < -b))^2.$$

PROOF OF THEOREM 1. It is enough to show that

$$\varliminf_{n\to\infty} \ln \mathsf{P}(\tau > n) \bigg/ \sum_{k=2}^{n} [\Delta g(k)]^2 \ge -\tfrac{1}{2} \tag{10}$$

and

$$\overline{\lim_{n\to\infty}} \ln \mathsf{P}(\tau > n) \Big/ \sum_{k=2}^{n} [\Delta g(k)]^2 \leq -\tfrac{1}{2}. \tag{11}$$

For this purpose we consider the (nonrandom) sequence $(\alpha_n)_{n\geq 1}$ with

$$\alpha_1 = 0, \quad \alpha_n = \Delta g(n), \qquad n \geq 2,$$

and the probability measure $(\tilde{\mathsf{P}}_n)_{n\geq 1}$ defined by (5). Then by Hölder's inequality

$$\tilde{\mathsf{P}}_n(\tau > n) = \mathsf{E} I(\tau > n)\, z_n \leq (\mathsf{P}(\tau > n))^{1/q}(\mathsf{E} z_n^p)^{1/p}, \tag{12}$$

where $p > 1$ and $q = p/(p-1)$.

The last factor is easily calculated explicitly:

$$(\mathsf{E} z_n^p)^{1/p} = \exp\left\{\frac{p-1}{2} \sum_{k=2}^{n} [\Delta g(k)]^2\right\}. \tag{13}$$

Now let us estimate the probability $\tilde{\mathsf{P}}_n(\tau > n)$ that appears on the left-hand side of (12). We have

$$\tilde{\mathsf{P}}_n(\tau > n) = \tilde{\mathsf{P}}_n(S_k \geq g(k), 1 \leq k \leq n) = \tilde{\mathsf{P}}_n(\tilde{S}_k \geq g(1), 1 \leq k \leq n),$$

where $\tilde{S}_k = \sum_{i=1}^{k} \tilde{\xi}_i, \tilde{\xi}_i = \xi_i - \alpha_i$. By Lemma 1, the variables are independent and normally distributed, $\tilde{\xi}_i \sim \mathcal{N}(0, 1)$, with respect to the measure $\tilde{\mathsf{P}}_n$. Then by Lemma 2 (applied to $b = -g(1)$, $\mathsf{P} = \tilde{\mathsf{P}}_n$, $X_n = \tilde{S}_n$) we find that

$$\tilde{\mathsf{P}}(\tau > n) \geq \frac{c}{n}, \tag{14}$$

where c is a constant.

Then it follows from (12)–(14) that, for every $p > 1$,

$$\mathsf{P}(\tau > n) \geq C_p \exp\left\{-\frac{p}{2} \sum_{k=2}^{n} [\Delta g(k)]^2 - \frac{p}{p-1} \ln n\right\}, \tag{15}$$

where C_p is a constant. Then (15) implies the lower bound (10) by the hypotheses of the theorem, since $p > 1$ is arbitrary.

To obtain the upper bound (11), we first observe that since $z_n > 0$ (P- and $\tilde{\mathsf{P}}$-a.s.), we have by (5)

$$\mathsf{P}(\tau > n) = \tilde{\mathsf{E}}_n I(\tau > n) z_n^{-1}, \tag{16}$$

where $\tilde{\mathsf{E}}_n$ denotes an average with respect to $\tilde{\mathsf{P}}_n$.

In the case under consideration, $\alpha_1 = 0$, $\alpha_n = \Delta g(n)$, $n \geq 2$, and therefore for $n \geq 2$

$$z_n^{-1} = \exp\left\{-\sum_{k=2}^{n} \Delta g(k) \cdot \xi_k + \frac{1}{2} \sum_{k=2}^{n} [\Delta g(k)]^2\right\}.$$

By the formula for summation by parts (see the proof of Lemma 2 of §3, Chapter IV)

$$\sum_{k=2}^{n} \Delta g(k) \cdot \xi_k = \Delta g(n) \cdot S_n - \sum_{k=2}^{n} S_{k-1} \Delta(\Delta g(k)).$$

Hence if we recall that by hypothesis $\Delta g(k) \geq 0$ and $\Delta(\Delta g(k)) \leq 0$, we find that, on the set $\{\tau > n\} = \{S_k \geq g(k), 1 \leq k \leq n\}$,

$$\begin{aligned}\sum_{k=2}^{n} \Delta g(k) \cdot \xi_k &\geq \Delta g(n) \cdot g(n) - \sum_{k=3}^{n} g(k-1)\Delta(\Delta g(k)) - \xi_1 \Delta g(2) \\ &= \sum_{k=2}^{n} [\Delta g(k)]^2 + g(1)\Delta g(2) - \xi_1 \Delta g(2).\end{aligned}$$

Thus, by (16),

$$\begin{aligned}\mathsf{P}(\tau > n) &\leq \exp\left\{-\frac{1}{2}\sum_{k=2}^{n}[\Delta g(k)]^2 - g(1)\Delta g(2)\right\} \tilde{\mathsf{E}}_n I(\tau > n) e^{-\xi_1 \Delta g(2)} \\ &\leq \exp\left\{-\frac{1}{2}\sum_{k=2}^{n}[\Delta g(k)]^2\right\} \tilde{\mathsf{E}}_n I(\tau > n) e^{-\xi_1 \Delta g(2)},\end{aligned}$$

where

$$\tilde{\mathsf{E}}_n I(\tau > n) e^{-\xi_1 \Delta g(2)} \leq \mathsf{E} z_n e^{-\xi_1 \Delta g(2)} = \mathsf{E} e^{-\xi_1 \Delta g(2)} < \infty.$$

Therefore

$$\mathsf{P}(\tau > n) \leq C \exp\left\{-\frac{1}{2}\sum_{k=2}^{n}[\Delta g(k)]^2\right\},$$

where C is a positive constant; this establishes the upper bound (11).

This completes the proof of the theorem.

3. The idea of an absolutely continuous change of measure can be used to study similar problems, including the case of a two-sided boundary. We present (without proof) a result in this direction.

Theorem 2. *Let $\xi_1, \xi_2, \ldots$ be independent identically distributed random variables with $\xi_i \sim \mathcal{N}(0, 1)$. Suppose that $f = f(n)$ is a positive function such that*

$$f(n) \to \infty, \qquad n \to \infty,$$

and

$$\sum_{k=2}^{n} [\Delta f(k)]^2 = o\left(\sum_{k=1}^{n} f^{-2}(k)\right), \qquad n \to \infty.$$

Then if

$$\sigma = \inf\{n \geq 1 : |S_n| \geq f(n)\},$$

we have

$$\mathsf{P}(\sigma > n) = \exp\left\{-\frac{\pi^2}{8}\sum_{k=1}^{n} f^{-2}(k)(1 + o(1))\right\}, \qquad n \to \infty. \tag{17}$$

4. Problems

1. Show that the sequence defined in (4) is a martingale.

2. Establish (13).

3. Prove (17).

§8. Central Limit Theorem for Sums of Dependent Random Variables

1. In §4, Chapter III, the central limit theorem for sums $S_n = \xi_{n1} + \cdots + \xi_{nn}$, $n \geq 1$, of random variables $\xi_{n1}, \ldots, \xi_{nn}$ was established under the assumptions of their independence, finiteness of second moments, and negligibility in the limit of their terms. In the present section, we give up both the assumption of independence and even that of the finiteness of the absolute values of the first-order moments. However, the negligibility in the limit of the terms will be retained.

Thus, we suppose that on the probability space $(\Omega, \mathscr{F}, \mathsf{P})$ there are given stochastic sequences

$$\xi^n = (\xi_{nk}, \mathscr{F}_k^n), \qquad 0 \leq k \leq n, \qquad n \geq 1,$$

with $\xi_{n0} = 0$, $\mathscr{F}_0^n = \{\varnothing, \Omega\}$, $\mathscr{F}_k^n \subseteq \mathscr{F}_{k+1}^n \subseteq \mathscr{F}$. We set

$$X_t^n = \sum_{k=0}^{[nt]} \xi_{nk}, \qquad 0 \leq t \leq 1.$$

Theorem 1. *For a given t, $0 < t \leq 1$, let the following conditions be satisfied: for each $\varepsilon \in (0, 1)$, as $n \to \infty$,*

(A) $\displaystyle\sum_{k=1}^{[nt]} \mathsf{P}(|\xi_{nk}| > \varepsilon \,|\, \mathscr{F}_{k-1}^n) \xrightarrow{\mathsf{P}} 0,$

(B) $\displaystyle\sum_{k=1}^{[nt]} \mathsf{E}\big[\xi_{nk} I(|\xi_{nk}| \leq 1) \,\big|\, \mathscr{F}_{k-1}^n\big] \xrightarrow{\mathsf{P}} 0,$

(C) $\displaystyle\sum_{k=1}^{[nt]} \mathsf{V}\big[\xi_{nk} I(|\xi_{nk}| \leq \varepsilon) \,\big|\, \mathscr{F}_{k-1}^n\big] \xrightarrow{\mathsf{P}} \sigma_t^2 \geq 0.$

Then

$$X_t^n \xrightarrow{d} \mathscr{N}(0, \sigma_t^2).$$

Remark 1. Hypotheses (A) and (B) guarantee that X_t^n can be represented in the form $X_t^n = Y_t^n + Z_t^n$ with $Z_t^n \xrightarrow{P} 0$ and $Y_t^n = \sum_{k=0}^{[nt]} \eta_{nk}$, where the sequence $\eta^n = (\eta_{nk}, \mathscr{F}_k^n)$ is a martingale-difference, and $\mathsf{E}(\eta_{nk}|\mathscr{F}_{k-1}^n) = 0$ with $|\eta_{nk}| \le c$, uniformly for $1 \le k \le n$ and $n \ge 1$. Consequently, in the cases under consideration, the proof reduces to proving the central limit theorem for martingale-differences.

In the case when the variables $\xi_{n1}, \ldots, \xi_{nn}$ are *independent*, conditions (A), (B), and (C), with $t = 1$, and $\sigma^2 = \sigma_1^2$, become

(a) $\sum_{k=1}^{n} \mathsf{P}(|\xi_{nk}| > \varepsilon) \to 0,$

(b) $\sum_{k=1}^{n} \mathsf{E}[\xi_{nk} I(|\xi_{nk}| \le 1)] \to 0,$

(c) $\sum_{k=1}^{n} \mathsf{V}[\xi_{nk} I(|\xi_{nk}| \le \varepsilon)] \to \sigma^2.$

These are well known; see the book by Gnedenko and Kolmogorov [G5]. Hence we have the following corollary to Theorem 1.

Corollary. *If* $\xi_{n1}, \ldots, \xi_{nn}$ *are independent random variables,* $n \ge 1$, *then*

$$(a), (b), (c) \Rightarrow X_1^n \xrightarrow{d} \mathscr{N}(0, \sigma^2).$$

Remark 2. In hypothesis (C), the case $\sigma_t^2 = 0$ is *not excluded.* Hence, in particular, Theorem 1 yields a convergence condition for degenerate distributions ($X_t^n \xrightarrow{d} 0$).

Remark 3. The method used to prove Theorem 1 lets us state and prove the following more general proposition.

Let $0 < t_1 < t_2 < \cdots < t_j \le 1$, $\sigma_{t_1}^2 \le \sigma_{t_2}^2 \le \cdots \le \sigma_{t_j}^2$, $\sigma_0^2 = 0$, and let $\varepsilon_1, \ldots, \varepsilon_j$ be independent Gaussian random variables with zero means and $\mathsf{E}\varepsilon_k^2 = \sigma_{t_k}^2 - \sigma_{t_{k-1}}^2$. Form the (Gaussian) vectors $(W_{t_1}, \ldots, W_{t_j})$ with $W_{t_k} = \varepsilon_1 + \cdots + \varepsilon_k$.

Let conditions (A), (B), and (C) be satisfied for $t = t_1, \ldots, t_j$. Then the joint distribution $(\mathsf{P}_{t_1,\ldots,t_j}^n)$ of the random variables $(X_{t_1}^n, \ldots, X_{t_j}^n)$ converges weakly to the Gaussian distribution $\mathsf{P}(t_1, \ldots, t_j)$ of the variables $(W_{t_1}, \ldots, W_{t_j})$:

$$\mathsf{P}_{t_1,\ldots,t_j}^n \xrightarrow{w} \mathsf{P}_{t_1,\ldots,t_j}.$$

2. The first assertion of the following theorem shows that condition (A) is equivalent to the condition of negligibility in the limit already introduced in §4, chapter III:

$$\text{(A*)} \qquad \max_{1 \le k \le [nt]} |\xi_{nk}| \xrightarrow{P} 0.$$

Theorem 2.
(1) *Condition* (A) *is equivalent to* (A*).
(2) *Assuming* (A) *or* (A*), *condition* (C) *is equivalent to*

$$(\mathrm{C}^*)\ \sum_{k=0}^{[nt]} [\xi_{nk} - \mathsf{E}(\xi_{nk} I(|\xi_{nk}| \le 1) \,|\, \mathscr{F}_{k-1}^n)]^2 \xrightarrow{\mathsf{P}} \sigma_t^2.$$

Theorem 3. *For each $n \ge 1$ let the sequence*

$$\xi^n = (\xi_{nk}, \mathscr{F}_k^n), \qquad 1 \le k \le n,$$

be a square-integrable martingale-difference:

$$\mathsf{E}\xi_{nk}^2 < \infty, \qquad \mathsf{E}(\xi_{nk} \,|\, \mathscr{F}_{k-1}^n) = 0.$$

Suppose that the Lindeberg condition is satisfied: for $\varepsilon > 0$,

$$(\mathrm{L})\ \sum_{k=0}^{[nt]} \mathsf{E}[\xi_{nk}^2 I(|\xi_{nk}| > \varepsilon) \,|\, \mathscr{F}_{k-1}^n] \xrightarrow{\mathsf{P}} 0.$$

Then (C) *is equivalent to*

$$\langle X^n \rangle_t \xrightarrow{\mathsf{P}} \sigma_t^2, \tag{1}$$

where (quadratic variation)

$$\langle X^n \rangle_t = \sum_{k=0}^{[nt]} \mathsf{E}(\xi_{nk}^2 \,|\, \mathscr{F}_{k-1}^n), \tag{2}$$

and (C*) *is equivalent to*

$$[X^n]_t \xrightarrow{\mathsf{P}} \sigma_t^2, \tag{3}$$

where (quadratic variation)

$$[X^n]_t = \sum_{k=0}^{[nt]} \xi_{nk}^2. \tag{4}$$

The next theorem is a corollary of Theorems 1–3.

Theorem 4. *Let the square-integrable martingale-differences $\xi^n = (\xi_{nk}, \mathscr{F}_k^n)$, $n \ge 1$, satisfy (for a given t, $0 < t \le 1$) the Lindeberg condition* (L). *Then*

$$\sum_{k=0}^{[nt]} \mathsf{E}(\xi_{nk}^2 | \mathscr{F}_{k-1}^n) \xrightarrow{\mathsf{P}} \sigma_t^2 \Rightarrow X_t^n \xrightarrow{d} \mathscr{N}(0, \sigma_t^2), \tag{5}$$

$$\sum_{k=0}^{[nt]} \xi_{nk}^2 \xrightarrow{\mathsf{P}} \sigma_t^2 \Rightarrow X_t^n \xrightarrow{d} \mathscr{N}(0, \sigma_t^2). \tag{6}$$

3. Proof of Theorem 1. Let us represent X_t^n in the form

$$X_t^n = \sum_{k=0}^{[nt]} \xi_{nk} I(|\xi_{nk}| \le 1) + \sum_{k=0}^{[nt]} \xi_{nk} I(|\xi_{nk}| > 1)$$

$$= \sum_{k=0}^{[nt]} \mathsf{E}[\xi_{nk} I(|\xi_{nk}| \le 1) | \mathscr{F}_{k-1}^n] + \sum_{k=0}^{[nt]} \xi_{nk} I(|\xi_{nk}| > 1)$$

$$+ \sum_{k=0}^{[nt]} \{\xi_{nk} I(|\xi_{nk}| \le 1) - \mathsf{E}[\xi_{nk} I(|\xi_{nk}| \le 1) | \mathscr{F}_{k-1}^n]\}. \tag{7}$$

We define

$$\begin{aligned} B_t^n &= \sum_{k=0}^{[nt]} \mathsf{E}[\xi_{nk} I(|\xi_{nk}| \le 1) | \mathscr{F}_{k-1}^n], \\ \mu_k^n(\Gamma) &= I(\xi_{nk} \in \Gamma), \\ \nu_k^n(\Gamma) &= \mathsf{P}(\xi_{nk} \in \Gamma \mid \mathscr{F}_{k-1}^n), \end{aligned} \tag{8}$$

where Γ is a set from the smallest σ-algebra $\sigma(R \setminus \{0\})$ and $\mathsf{P}(\xi_{nk} \in \Gamma | \mathscr{F}_{k-1}^n)$ is the regular conditional distribution of ξ_{nk} with respect to $\mathscr{F}_{k-1}^n$.

Then (7) can be rewritten in the following form:

$$X_t^n = B_t^n + \sum_{k=0}^{[nt]} \int_{|x|>1} x \, d\mu_k^n + \sum_{k=0}^{[nt]} \int_{|x| \le 1} x \, d(\mu_k^n - \nu_k^n), \tag{9}$$

which is known as the *canonical* decomposition of $(X_t^n, \mathscr{F}_t^n)$. (The integrals are to be understood as Lebesgue–Stieltjes integrals, defined for every sample point.)

According to (B) we have $B_t^n \xrightarrow{\mathsf{P}} 0$. Let us show that (A) implies

$$\sum_{k=0}^{[nt]} \int_{|x|>1} |x| \, d\mu_k^n \xrightarrow{\mathsf{P}} 0. \tag{10}$$

We have

$$\sum_{k=0}^{[nt]} \int_{|x|>1} |x| \, d\mu_k^n = \sum_{k=0}^{[nt]} |\xi_{nk}| I(|\xi_{nk}| > 1). \tag{11}$$

For every $\delta \in (0, 1)$,

$$\left\{\sum_{k=0}^{[nt]} |\xi_{nk}| I(|\xi_{nk}| > 1) > \delta\right\} \subseteq \left\{\sum_{k=0}^{[nt]} I(|\xi_{nk}| > 1) > \delta\right\}. \tag{12}$$

It is clear that

$$\sum_{k=0}^{[nt]} I(|\xi_{nk}| > 1) = \sum_{k=0}^{[nt]} \int_{|x|>1} d\mu_k^n (\equiv U_{[nt]}^n).$$

By (A),

$$V^n_{[nt]} \equiv \sum_{k=0}^{[nt]} \int_{|x|>1} d\nu^n_k \xrightarrow{\mathrm{P}} 0, \tag{13}$$

and V^n_k is $\mathscr{F}^n_{k-1}$-measurable.

Then by the corollary to Theorem 2 of §3, Chapter VII,

$$V^n_{[nt]} \xrightarrow{\mathrm{P}} 0 \Rightarrow U^n_{[nt]} \xrightarrow{\mathrm{P}} 0. \tag{14}$$

(By the same corollary and the inequality $\Delta U^n_{[nt]} \le 1$, we also have the converse implication

$$U^n_{[nt]} \xrightarrow{\mathrm{P}} 0 \Rightarrow V^n_{[nt]} \xrightarrow{\mathrm{P}} 0, \tag{15}$$

which will be needed in the proof of Theorem 2.)

The required proposition (10) now follows from (11)–(14).

Thus

$$X^n_t = Y^n_t + Z^n_t, \tag{16}$$

where

$$Y^n_t = \sum_{k=0}^{[nt]} \int_{|x|\le 1} x\, d(\mu^n_k - \nu^n_k), \tag{17}$$

and

$$Z^n_t = B^n_t + \sum_{k=0}^{[nt]} \int_{|x|>1} x\, d\mu^n_k \xrightarrow{\mathrm{P}} 0. \tag{18}$$

It then follows by Problem 1 that to establish that

$$X^n_t \xrightarrow{d} \mathscr{N}(0, \sigma^2_t)$$

we need only show that

$$Y^n_t \xrightarrow{d} \mathscr{N}(0, \sigma^2_t). \tag{19}$$

Let us represent Y^n_t in the form

$$Y^n_t = Y^n_{[nt]}(\varepsilon) + \Delta^n_{[nt]}(\varepsilon), \qquad \varepsilon \in (0, 1],$$

where

$$Y^n_{[nt]}(\varepsilon) = \sum_{k=0}^{[nt]} \int_{\varepsilon<|x|\le 1} x\, d(\mu^n_k - \nu^n_k), \tag{20}$$

$$\Delta^n_{[nt]}(\varepsilon) = \sum_{k=0}^{[nt]} \int_{|x|\le\varepsilon} x\, d(\mu^n_k - \nu^n_k). \tag{21}$$

As in the proof of (14), it is easily verified that, because of (A), we have $Y^n_{[nt]}(\varepsilon) \xrightarrow{\mathrm{P}} 0$, $n \to \infty$.

The sequence $\Delta^n(\varepsilon) = (\Delta_k^n(\varepsilon), \mathscr{F}_k^n)$, $1 \le k \le n$, is a square-integrable martingale with quadratic variation

$$\langle \Delta^n(\varepsilon)\rangle_k = \sum_{i=0}^k \left[\int_{|x|\le\varepsilon} x^2\, d\nu_i^n - \left(\int_{|x|\le\varepsilon} x\, d\nu_i^n\right)^2\right]$$

$$= \sum_{i=0}^k \mathsf{V}[\xi_{ni} I(|\xi_{ni}| \le \varepsilon) \,|\, \mathscr{F}_{i-1}^n].$$

Because of (C),

$$\langle \Delta^n(\varepsilon)\rangle_{[nt]} \xrightarrow{\mathsf{P}} \sigma_t^2.$$

Hence, for every $\varepsilon \in (0, 1]$,

$$\max\{\gamma_{[nt]}^n(\varepsilon), |\langle \Delta^n(\varepsilon)\rangle_{[nt]} - \sigma_t^2|\} \xrightarrow{\mathsf{P}} 0.$$

By Problem 2 there is then a sequence of numbers $\varepsilon_n \downarrow 0$ such that

$$\gamma_{[nt]}^n(\varepsilon_n) \xrightarrow{\mathsf{P}} 0, \qquad \langle \Delta^n(\varepsilon_n)\rangle_{[nt]} \xrightarrow{\mathsf{P}} \sigma_t^2.$$

Therefore, again by Problem 1, it is enough to prove only that

$$M_{[nt]}^n \xrightarrow{d} \mathscr{N}(0, \sigma_t^2), \tag{22}$$

where

$$M_k^n \equiv \Delta_k^n(\varepsilon_n) = \sum_{i=0}^k \int_{|x|\le\varepsilon_n} x(\mu_i^n - \nu_i^n). \tag{23}$$

For $\Gamma \in \sigma(R\backslash\{0\})$, let

$$\tilde{\mu}_k^n(\Gamma) = I(\Delta M_k^n \in \Gamma), \qquad \tilde{\nu}_k^n(\Gamma) = \mathsf{P}(\Delta M_k^n \in \Gamma \,|\, \mathscr{F}_{k-1}^n)$$

be a regular conditional probability, $\Delta M_k^n = M_k^n - M_{k-1}^n$, $k \ge 1$, $M_0^n = 0$. Then the square-integrable martingale $M^n = (M_k^n, \mathscr{F}_k^n)$, $1 \le k \le n$, can evidently be written in the form

$$M_k^n = \sum_{i=1}^k \Delta M_i^n = \sum_{i=1}^k \int_{|x|\le 2\varepsilon_n} x\, d\tilde{\mu}_k^n.$$

(Notice that $|\Delta M_i^n| \le 2\varepsilon_n$ by (23).)

To establish (22) we have to show that, for every real λ,

$$\mathsf{E}\exp\{i\lambda M_{[nt]}^n\} \to \exp(-\tfrac{1}{2}\lambda^2\sigma_t^2). \tag{24}$$

Put

$$G_k^n = \sum_{j=1}^k \int_{|x|\le 2\varepsilon_n} (e^{i\lambda x} - 1)\, d\tilde{\nu}_j^n$$

and

$$\mathscr{E}_k^n(G^n) = \prod_{j=1}^k (1 + \Delta G_k^n).$$

Observe that

$$1 + \Delta G_k^n = 1 + \int_{|x|\le 2\varepsilon_n} (e^{i\lambda x} - 1)\, d\tilde{\nu}_k^n = \int_{|x|\le 2\varepsilon_n} e^{i\lambda x}\, d\tilde{\nu}_k^n$$
$$= \mathsf{E}[\exp(i\lambda \Delta M_k^n) | \mathscr{F}_{k-1}^n]$$

and consequently

$$\mathscr{E}_k^n(G^n) = \prod_{j=1}^k (1 + \Delta G_k^n).$$

On the basis of a lemma that will be proved in Subsection 4, (24) will follow if for every real λ

$$|\mathscr{E}_{[nt]}^n(G^n)| = \left| \prod_{j=1}^{[nt]} \mathsf{E}\left[\exp(i\lambda \Delta M_j^n) \middle| \mathscr{F}_{j-1}^n\right] \right| \ge C(\lambda) > 0 \tag{25}$$

and

$$\mathscr{E}_{[nt]}^n(G^n) \xrightarrow{\mathsf{P}} \exp(-\tfrac{1}{2}\lambda^2\sigma_t^2). \tag{26}$$

To see this we represent $\mathscr{E}_k^n(G^n)$ in the form

$$\mathscr{E}_k^n(G^n) = \exp(G_k^n) \cdot \prod_{j=1}^k (1 + \Delta G_j^n) \exp(-\Delta G_j^n).$$

(Compare the function $E_t(A)$ defined by (76) in §6, Chapter II.)

Since

$$\int_{|x|\le 2\varepsilon_n} x\, d\tilde{\nu}_j^n = \mathsf{E}(\Delta M_j^n | \mathscr{F}_{j-1}^n) = 0,$$

we have

$$G_k^n = \sum_{j=1}^k \int_{|x|\le 2\varepsilon_n} (e^{i\lambda x} - 1 - i\lambda x)\, d\tilde{\nu}_j^n. \tag{27}$$

Therefore

$$|\Delta G_k^n| \le \int_{|x|\le 2\varepsilon_n} |e^{i\lambda x} - 1 - i\lambda x|\, d\tilde{\nu}_k^n \le \tfrac{1}{2}\lambda^2 \int_{|x|\le 2\varepsilon_n} x^2\, d\tilde{\nu}_k^n$$
$$\le \tfrac{1}{2}\lambda^2 (2\varepsilon_n)^2 \to 0 \tag{28}$$

and

$$\sum_{j=1}^k |\Delta G_j^n| \le \tfrac{1}{2}\lambda^2 \sum_{j=1}^k \int_{|x|\le 2\varepsilon_n} x^2\, d\tilde{\nu}_j^n = \tfrac{1}{2}\lambda^2 \langle M^n \rangle_k. \tag{29}$$

By (C),

$$\langle M^n \rangle_{[nt]} \xrightarrow{\mathsf{P}} \sigma_t^2. \tag{30}$$

Suppose first that $\langle M^n\rangle_k \le a$ (P-a.s.), $k \le [nt]$, where $a \ge \sigma_t^2 + 1$. Then by (28), (29), and Problem 3,

$$\prod_{k=1}^{[nt]} (1 + \Delta G_k^n)\exp(-\Delta G_k^n) \xrightarrow{\mathsf{P}} 1, \qquad n \to \infty,$$

and therefore to establish (26) we only have to show that

$$G_{[nt]}^n \to -\tfrac{1}{2}\lambda^2\sigma_t^2, \tag{31}$$

i.e., after (27), (29), and (30), that

$$\sum_{k=1}^{[nt]} \int_{|x|\le 2\varepsilon_n} (e^{i\lambda x} - 1 - i\lambda x + \tfrac{1}{2}\lambda^2 x^2)\, d\tilde{\nu}_k^n \xrightarrow{\mathsf{P}} 0. \tag{32}$$

But

$$|e^{i\lambda x} - 1 - i\lambda x + \tfrac{1}{2}\lambda^2 x^2| \le \tfrac{1}{6}|\lambda x|^3$$

and therefore

$$\sum_{k=1}^{[nt]} \int_{|x|\le 2\varepsilon_n} |e^{i\lambda x} - 1 - i\lambda x + \tfrac{1}{2}\lambda^2 x^2|\, d\tilde{\nu}_k^n \le \tfrac{1}{6}|\lambda|^3(2\varepsilon_n) \sum_{k=1}^{[nt]} \int_{|x|\le 2\varepsilon_n} x^2\, d\tilde{\nu}_k^n$$
$$= \tfrac{1}{3}\varepsilon_n|\lambda|^3\langle M_n\rangle_{[nt]} \le \tfrac{1}{3}\varepsilon_n|\lambda|^3 a \to 0, \qquad n \to \infty.$$

Therefore if $\langle M^n\rangle_{[nt]} \le a$ (P-a.s.), (31) is established and consequently (26) is established also.

Let us now verify (25). Since $|e^{i\lambda x} - 1 - i\lambda x| \le \frac{1}{2}(\lambda x)^2$, we find from (28) that, for sufficiently large n,

$$|\mathscr{E}_k^n(G^n)| = \Big|\prod_{j=1}^{k}(1 + \Delta G_j^n)\Big| \ge \prod_{j=1}^{k}(1 - \tfrac{1}{2}\lambda^2\Delta\langle M^n\rangle_j)$$
$$= \exp\Big\{\sum_{j=1}^{k} \ln(1 - \tfrac{1}{2}\lambda^2\Delta\langle M^n\rangle_j)\Big\}.$$

But

$$\ln(1 - \tfrac{1}{2}\lambda^2\Delta\langle M^n\rangle_j) \ge -\frac{\tfrac{1}{2}\lambda^2\Delta\langle M^n\rangle_j}{1 - \tfrac{1}{2}\lambda^2\Delta\langle M^n\rangle_j}$$

and $\Delta\langle M^n\rangle_j \le (2\varepsilon_n)^2 \downarrow 0$, $n \to \infty$. Therefore there is an $n_0 = n_0(\lambda)$ such that for all $n \ge n_0(\lambda)$,

$$|\mathscr{E}_k^n(G^n)| \ge \exp\{-\lambda^2\langle M^n\rangle_k\}$$

and therefore

$$|\mathscr{E}_{[nt]}^n(G^n)| \ge \exp\{-\lambda^2\langle M^n\rangle_{[nt]}\} \ge e^{-\lambda^2 a}.$$

Hence the theorem is proved under the assumption that $\langle M^n\rangle_{[nt]} \le a$ (P-a.s.). To remove this assumption, we proceed as follows.

Let

$$\tau^n = \min\{k \le [nt]: \langle M^n \rangle_k \ge \sigma_t^2 + 1\},$$

taking $\tau^n = \infty$ if $\langle M^n \rangle_{[nt]} \le \sigma_t^2 + 1$.

Then for $\overline{M}_k^n = M_{k \wedge \tau^n}^n$ we have

$$\langle \overline{M}^n \rangle_{[nt]} = \langle M^n \rangle_{[nt] \wedge \tau^n} \le 1 + \sigma_t^2 + 2\varepsilon_n^2 \le 1 + \sigma_t^2 + 2\varepsilon_1^2 \ (=a),$$

and by what has been proved,

$$\mathsf{E} \exp\{i\lambda \overline{M}_{[nt]}^n\} \to \exp(-\tfrac{1}{2}\lambda^2 \sigma_t^2).$$

But

$$\lim_n |\mathsf{E}\{\exp(i\lambda M_{[nt]}^n) - \exp(i\lambda \overline{M}_{[nt]}^n)\}| \le 2 \lim_n \mathsf{P}(\tau^n < \infty) = 0.$$

Consequently

$$\lim_n \mathsf{E} \exp(i\lambda M_{[nt]}^n) = \lim_n \mathsf{E}\{\exp(i\lambda M_{[nt]}^n) - \exp(i\lambda \overline{M}_{[nt]}^n)\} + \lim_n \mathsf{E} \exp(i\lambda \overline{M}_{[nt]}^n) = \exp(-\tfrac{1}{2}\lambda^2 \sigma_t^2).$$

This completes the proof of Theorem 1.

Remark. To prove the statement made in Remark 2 to Theorem 1, we need (using the Cramér–Wold method [B3]) to show that for all real numbers $\lambda_1, \dots, \lambda_j$

$$\mathsf{E} \exp i\left[\lambda_1 M_{[nt_1]}^n + \sum_{k=2}^{j} \lambda_k (M_{[nt_k]}^n - M_{[nt_{k-1}]}^n)\right] \to \exp(-\tfrac{1}{2}\lambda_1^2 \sigma_{t_1}^2) - \sum_{k=2}^{j} \tfrac{1}{2}\lambda_k^2 (\sigma_{t_k}^2 - \sigma_{t_{k-1}}^2).$$

The proof of this is similar to the proof of (24), replacing $(M_k^n, \mathscr{F}_k^n)$ by the square-integrable martingales $(\hat{M}_k^n, \mathscr{F}_k^n)$,

$$\hat{M}_k^n = \sum_{i=1}^{k} v_i \Delta M_i^n,$$

where $v_i = \lambda_1$ for $i \le [nt_1]$ and $v_i = \lambda_j$ for $[nt_{j-1}] < i \le [nt_j]$.

4. In this subsection we prove a simple lemma which lets us reduce the verification of (24) to the verification of (25) and (26).

Let $\eta^n = (\eta_{nk}, \mathscr{F}_k^n)$, $1 \le k \le n$, $n \ge 1$, be stochastic sequences, let

$$Y^n = \sum_{k=1}^{n} \eta_{nk},$$

let

$$\mathscr{E}^n(\lambda) = \prod_{k=1}^{n} \mathsf{E}[\exp(i\lambda \eta_{nk}) | \mathscr{F}_{k-1}^n], \qquad \lambda \in R,$$

and let Y be a random variable with

$$\mathscr{E}(\lambda) = \mathsf{E}e^{i\lambda Y}, \qquad \lambda \in R.$$

Lemma. *If (for a given λ) $|\mathscr{E}^n(\lambda)| \geq c(\lambda) > 0$, $n \geq 1$, a sufficient condition for the limit relation*

$$\mathsf{E}e^{i\lambda Y^n} \to \mathsf{E}e^{i\lambda Y} \tag{33}$$

is that

$$\mathscr{E}^n(\lambda) \xrightarrow{\mathsf{P}} \mathscr{E}(\lambda). \tag{34}$$

PROOF. Let

$$m^n(\lambda) = \frac{e^{i\lambda Y^n}}{\mathscr{E}^n(\lambda)}.$$

Then $|m^n(\lambda)| \leq c^{-1}(\lambda) < \infty$, and it is easily verified that

$$\mathsf{E}m^n(\lambda) = 1.$$

Hence by (34) and the Lebesgue dominated convergence theorem,

$$\begin{aligned}|\mathsf{E}e^{i\lambda Y^n} - \mathsf{E}e^{i\lambda Y}| &= |\mathsf{E}(e^{i\lambda Y^n} - \mathscr{E}(\lambda))| \\ &= |\mathsf{E}(m^n(\lambda)[\mathscr{E}^n(\lambda) - \mathscr{E}(\lambda)])| \leq c^{-1}(\lambda)\mathsf{E}\,|\mathscr{E}^n(\lambda) - \mathscr{E}(\lambda)| \to 0, \qquad n \to \infty.\end{aligned}$$

Remark. It follows, from (33) and the hypothesis that $\mathscr{E}^n(\lambda) \geq c(\lambda) > 0$, that $\mathscr{E}(\lambda) \neq 0$. In fact, the conclusion of the lemma remains valid without the assumption that $|\mathscr{E}^n(\lambda)| \geq c(\lambda) > 0$, if restated in the form: if $\mathscr{E}^n(\lambda) \xrightarrow{\mathsf{P}} \mathscr{E}(\lambda)$ and $\mathscr{E}(\lambda) \neq 0$, then (33) holds (Problem 5).

5. PROOF OF THEOREM 2. (1) Let $\varepsilon > 0$, $\delta \in (0, \varepsilon)$, and for simplicity let $t = 1$. Since

$$\max_{1 \leq k \leq n} |\xi_{nk}| \leq \varepsilon + \sum_{k=1}^{n} |\xi_{nk}| I(|\xi_{nk}| > \varepsilon)$$

and

$$\left\{\sum_{k=1}^{n} |\xi_{nk}| I(|\xi_{nk}| > \varepsilon) > \delta\right\} \subseteq \left\{\sum_{k=1}^{n} I(|\xi_{nk}| > \varepsilon) > \delta\right\},$$

we have

$$\begin{aligned}\mathsf{P}\left\{\max_{1 \leq k \leq n} |\xi_{nk}| > \varepsilon + \delta\right\} &\leq \mathsf{P}\left\{\sum_{k=1}^{n} I(|\xi_{nk}| > \varepsilon) > \delta\right\} \\ &= \mathsf{P}\left\{\sum_{k=1}^{n} \int_{|x| > \varepsilon} d\mu_k^n > \delta\right\}.\end{aligned}$$

If (A) is satisfied, i.e.,

$$\mathsf{P}\left\{\sum_{k=1}^{n} \int_{|x| > \varepsilon} d\nu_k^n > \delta\right\} \to 0$$

then (compare (14)) we also have

$$\mathsf{P}\left\{\sum_{k=1}^{n} \int_{|x|>\varepsilon} d\mu_k^n > \delta\right\} \to 0.$$

Therefore (A) $\Rightarrow$ (A*).

Conversely, let

$$\sigma_n = \min\{k \le n : |\xi_{nk}| \ge \varepsilon/2\},$$

supposing that $\sigma_n = \infty$ if $\max_{1\le k\le n}|\xi_{nk}| < \varepsilon/2$. By (A*), $\lim_n \mathsf{P}(\sigma_n < \infty) = 0$.

Now observe that, for every $\delta \in (0, 1)$, the sets

$$\left\{\sum_{k=1}^{n\wedge\sigma_n} I(|\xi_{nk}| \ge \varepsilon/2) > \delta\right\} \quad \text{and} \quad \left\{\max_{1\le k\le n\wedge\sigma_n} |\xi_{nk}| \ge \tfrac{1}{2}\varepsilon\right\}$$

coincide, and by (A*)

$$\sum_{k=1}^{n\wedge\sigma_n} I(|\xi_{nk}| \ge \varepsilon/2) = \sum_{k=1}^{n\wedge\sigma_n} \int_{|x|\ge\varepsilon/2} d\mu_k^n \xrightarrow{\mathsf{P}} 0.$$

Therefore by (15)

$$\sum_{k=1}^{n\wedge\sigma_n} \int_{|x|\ge\varepsilon} d\nu_k^n \le \sum_{k=1}^{n\wedge\sigma_n} \int_{|x|\ge\varepsilon/2} d\nu_k^n \xrightarrow{\mathsf{P}} 0,$$

which, together with the property $\lim_n \mathsf{P}(\sigma_n < \infty) = 0$, prove that (A*) $\Rightarrow$ (A).

(2) Again suppose that $t = 1$. Choose an $\varepsilon \in (0, 1]$ and consider the square-integrable martingales

$$\Delta^n(\delta) = (\Delta_k^n(\delta), \mathscr{F}_k^n) \qquad (1 \le k \le n),$$

with $\delta \in (0, \varepsilon]$. For the given $\varepsilon \in (0, 1]$, we have, according to (C),

$$\langle \Delta^n(\varepsilon)\rangle_n \xrightarrow{\mathsf{P}} \sigma_1^2.$$

It is then easily deduced from (A) that for every $\delta \in (0, \varepsilon]$

$$\langle \Delta^n(\delta)\rangle_n \xrightarrow{\mathsf{P}} \sigma_1^2. \tag{35}$$

Let us show that it follows from (C*), (A), and (A*) that, for every $\delta \in (0, \varepsilon]$,

$$[\Delta^n(\delta)]_n \xrightarrow{\mathsf{P}} \sigma_1^2, \tag{36}$$

where

$$[\Delta^n(\delta)]_n = \sum_{k=1}^{n} \left[\xi_{nk} I(|\xi_{nk}| \le \delta) - \int_{|x|\le\delta} x\, d\nu_k^n\right]^2.$$

In fact, it is easily verified that by (A)

$$[\Delta^n(\delta)]_n - [\Delta^n(1)]_n \xrightarrow{\mathsf{P}} 0. \tag{37}$$

But

$$\left|\sum_{k=1}^{n}\left[\xi_{nk} - \int_{|x|\le 1} x\, d\nu_k^n\right]^2 - \sum_{k=1}^{n}\left[\xi_{nk} I(|\xi_{nk}| \le 1) - \int_{|x|\le 1} x\, d\nu_k^n\right]^2\right|$$

$$\le \sum_{k=1}^{n} I(|\xi_{nk}| > 1)\left[(\xi_{nk})^2 + 2|\xi_{nk}|\left|\int_{|x|\le 1} x\, d(\mu_k^n - \nu_k^n)\right|\right]$$

$$\le 5\sum_{k=1}^{n} I(|\xi_{nk}| > 1)|\xi_{nk}|^2$$

$$\le 5 \max_{1\le k\le n} |\xi_{nk}|^2 \sum_{k=1}^{n}\int_{|x|>1} d\mu_k^n \to 0. \tag{38}$$

Hence (36) follows from (37) and (38).

Consequently to establish the equivalence of (C) and (C*) it is enough to establish that when (C) is satisfied (for a given $\varepsilon \in (0, 1]$), then (C*) is also satisfied for every $a > 0$:

$$\lim_{\delta\to 0} \overline{\lim_n}\, \mathsf{P}\{|[\Delta^n(\sigma)]_n - \langle\Delta^n(\delta)\rangle_n| > a\} = 0. \tag{39}$$

Let

$$m_k^n(\delta) = [\Delta^n(\delta)]_k - \langle\Delta^n(\delta)\rangle_k, \qquad 1 \le k \le n.$$

The sequence $m^n(\delta) = (m_k^n(\delta), \mathscr{F}_k^n)$ is a square-integrable martingale, and $(m^n(\delta))^2$ is dominated (in the sense of the definition of §3 on p. 467) by the sequences $[m^n(\delta)]$ and $\langle m^n(\delta)\rangle$.

It is clear that

$$[m^n(\delta)]_n = \sum_{k=1}^{n} (\Delta m_k^n(\delta))^2 \le \max_{1\le k\le n} |\Delta m_k^n(\delta)|\{[\Delta^n(\delta)]_n + \langle\Delta^n(\delta)\rangle_n\}$$

$$\le 3\delta^2\{[\Delta^n(\delta)]_n + \langle\Delta^n(\delta)\rangle_n\}. \tag{40}$$

Since $[\Delta^n(\delta)]$ and $\langle\Delta^n(\delta)\rangle$ dominate each other, it follows from (40) that $(m^n(\delta))^2$ is dominated by the sequences $6\delta^2[\Delta^n(\delta)]$ and $6\delta^2\langle\Delta^n(\delta)\rangle$.

Hence if (C) is satisfied, then for sufficiently small δ (for example, for $\delta < \frac{1}{6}b(\sigma_1^2 + 1)$)

$$\overline{\lim_n}\, \mathsf{P}(6\delta^2\langle\Delta^n(\delta)\rangle_n > b) = 0,$$

and hence, by the corollary to Theorem 2 of §3, we have (39). If (C*) is also satisfied, then for the same values of δ,

$$\overline{\lim_n}\, \mathsf{P}(6\delta^2[\Delta^n(\delta)]_k > b) = 0. \tag{41}$$

Since $|\Delta[\Delta^n(\delta)]_k| \le (2\delta)^2$, the validity of (39) follows from (41) and another appeal to Theorem 2 of §3.

This completes the proof of Theorem 2.

6. Proof of Theorem 3. On account of the Lindeberg condition (L), the equivalence of (C) and (1), and of (C*) and (3), can be established by direct calculation (Problem 6).

7. Proof of Theorem 4. Condition (A) follows from the Lindeberg condition (L). As for condition (B), it is sufficient to observe that when ξ^n is a martingale-difference, the variables B_t^n that appear in the canonical decomposition (9) can be represented in the form

$$B_t^n = -\sum_{k=0}^{[nt]} \int_{|x|>1} x\,d\nu_n^k.$$

Therefore $B_t^n \xrightarrow{P} 0$ by the Lindeberg condition (L).

8. The fundamental theorem of the present section, namely Theorem 1, was proved under the hypothesis that the terms that are summed are uniformly asymptotically infinitesimal. It is natural to ask for conditions for the central limit theorem without such a hypothesis. For independent random variables, examples of such theorems are given by Theorem 1 (assuming finite second moments) or Theorem 5 (assuming finite first moments) from §4, Chapter III.

We quote (without proof) an analog of the first of these theorems, applicable only to sequences $\xi^n = (\xi_{nk}, \mathscr{F}_k^n)$ that are square-integrable martingale differences.

Let $\mathscr{F}_{nk}(x) = \mathsf{P}(\xi_{nk} \le x \mid \mathscr{F}_{k-1}^n)$ be a regular distribution function of ξ_{nk} with respect to $\mathscr{F}_{k-1}^n$, and let $\Delta_{nk} = \mathsf{E}(\xi_{nk}^2 \mid \mathscr{F}_{k-1}^n)$.

Theorem 5. *If a square-integrable martingale-difference* $\xi_n = (\xi_{nk}, \mathscr{F}_k^n)$, $0 \le k \le n$, $n \ge 1$, $\xi_{n0} = 0$, *satisfies the condition*

$$\sum_{k=0}^{[nt]} \Delta_{nk} \xrightarrow{P} \sigma_t^2, \qquad 0 \le \sigma_t^2 < \infty,$$

and for every $\varepsilon > 0$

$$\sum_{k=0}^{[nt]} \int_{|x|>\varepsilon} |x| \left| \mathscr{F}_{nk}(x) - \Phi\left(\frac{x}{\sqrt{\Delta_{nk}}}\right) \right| dx \xrightarrow{P} 0,$$

then

$$X_t^n \xrightarrow{d} \mathscr{N}(0, \sigma_t^2).$$

9. Problems

1. Let $\xi_n = \eta_n + \zeta_n$, $n \ge 1$, where $\eta_n \xrightarrow{d} \eta$ and $\zeta_n \xrightarrow{d} 0$. Prove that $\xi_n \xrightarrow{d} \eta$.

2. Let $(\xi_n(\varepsilon))$, $n \ge 1$, $\varepsilon > 0$, be a family of random variables such that $\xi_n(\varepsilon) \xrightarrow{P} 0$ for each $\varepsilon > 0$ as $n \to \infty$. Using, for example, Problem 11 of §10, Chapter II, prove that there is a sequence $\varepsilon_n \downarrow 0$ such that $\xi_n(\varepsilon_n) \xrightarrow{P} 0$.

3. Let (α_k^n), $1 \le k \le n$, $n \ge 1$, be a complex-valued random variable such that (P-a.s.)

$$\sum_{k=1}^{n} |\alpha_k^n| \le C, \qquad |\alpha_k^n| \le a_n \downarrow 0.$$

Show that then (P-a.s.)

$$\lim_n \prod_{k=1}^{n} (1 + \alpha_k^n) \exp(-\alpha_k^n) = 1.$$

4. Prove the statement made in Remark 2 to Theorem 1.

5. Prove the statement made in Remark 1 to the lemma.

6. Prove Theorem 3.

7. Prove Theorem 5.

§9. Discrete Version of Itô's Formula

1. In the stochastic analysis of Brownian motion and other related processes (for example, martingales, local martingales, semi-martingales) with continuous time *Itô's change-of-variables formula* plays a key role (see, for example, [J1], [L12]).

This section may be viewed as a prelude to Itô's formula for Brownian motion. In it, we present a discrete (in time) version of Itô's formula and show briefly how a corresponding formula for continuous time could be obtained using a limiting procedure.

2. Let $X = (X_n)_{0 \le n \le N}$ and $Y = (Y_n)_{0 \le n \le N}$ be two sequences of random variables on the probability space $(\Omega, \mathscr{F}, P)$, $X_0 = Y_0 = 0$ and

$$[X, Y] = ([X, Y]_n)_{0 \le n \le N},$$

where

$$[X, Y]_n = \sum_{i=1}^{n} \Delta X_i \Delta Y_i \tag{1}$$

is the *square covariance* of $(X_0, X_1, \dots, X_n)$ and $(Y_0, Y_1, \dots, Y_n)$ (see §1). Also, suppose that $F = F(x)$ is an absolutely continuous function,

$$F(x) = F(0) + \int_0^x f(y)\, dy, \tag{2}$$

where $f = f(y)$, $y \in \mathbb{R}$ is a Borel function such that

$$\int_{|y| \le c} |f(y)|\, dy < \infty, \qquad c > 0.$$

The change-of-variables formula in which we are interested concerns the possibility of representing the sequence

$$F(X) = (F(X_n))_{0 \le n \le N} \tag{3}$$

in terms of 'natural' functional from the sequence X. Of course, this requires explanation and will be explained here.

Given the function $f = f(x)$, we form the square covariance $[X, f(X)]$ as follows:

$$\begin{aligned}[X, f(X)]_n &= \sum_{k=1}^{n} \Delta f(X_k)\Delta X_k \\ &= \sum_{k=1}^{n} (f(X_k) - f(X_{k-1}))(X_k - X_{k-1}).\end{aligned} \tag{4}$$

We introduce two 'discrete integrals' (cf. Definition 5 of §1):

$$I_n(X, f(X)) = \sum_{k=1}^{n} f(X_{k-1})\Delta X_k, \tag{5}$$

$$\tilde{I}_n(X, f(X)) = \sum_{k=1}^{n} f(X_k)\Delta X_k. \tag{6}$$

Then

$$[X, f(X)]_n = \tilde{I}_n(X, f(X)) - I_n(X, f(X)). \tag{7}$$

For fixed N, we introduce a new (reversed) sequence $\tilde{X} = (\tilde{X}_n)_{0 \le n \le N}$ with

$$\tilde{X}_n = X_{N-n}. \tag{8}$$

Then clearly,

$$\tilde{I}_N(X, f(X)) = -I_N(\tilde{X}, f(\tilde{X}))$$

and analogously,

$$\tilde{I}_n(X, f(X)) = -\{I_N(\tilde{X}, f(\tilde{X})) - I_{N-n}(\tilde{X}, f(\tilde{X}))\}$$

(we set $I_0 = \tilde{I}_0 = 0$).

Thus,

$$[X, f(X)]_N = -\{I_N(\tilde{X}, f(\tilde{X})) + I_N(\tilde{X}, f(X))\}$$

and for $0 < n < N$ we have:

$$\begin{aligned}[X, f(X)]_n &= -\{I_N(\tilde{X}, f(\tilde{X})) - I_{N-n}(\tilde{X}, f(\tilde{X}))\} - I_n(X, f(X)) \\ &= -\left\{\sum_{k=N-n+1}^{N} f(\tilde{X}_k)\Delta\tilde{X}_k + \sum_{k=1}^{n} f(X_k)\Delta X_k\right\}.\end{aligned} \tag{9}$$

Remark 1. We note that the structures of the right-hand sides of (7) and (9) are different. Equation (7) contains two different forms of "discrete integral." The integral $I_n(X, f(X))$ is a "forward integral," while $\tilde{I}_n(X, f(X))$ is a "backward integral." In (9), both integrals are "forward integrals," over two *differ-*

ent sequences X and $\tilde{X}$.

3. Since for any function $g = g(x)$

$$g(X_{k-1}) + \tfrac{1}{2}[g(X_k) - g(X_{k-1})] - \tfrac{1}{2}[g(X_k) + g(X_{k-1})] = 0,$$

it is clear that

$$F(X_n) = F(X_0) + \sum_{k=1}^{n} g(X_{k-1})\Delta X_k + \tfrac{1}{2}[X, g(X)]_n + \sum_{k=1}^{n} \left\{(F(X_k) - F(X_{k-1})) - \frac{g(X_{k-1}) + g(X_k)}{2}\Delta X_k\right\}. \tag{10}$$

In particular, if $g(x) = f(x)$, where $f(x)$ is the function of (2), then

$$F(X_n) = F(X_0) + I_n(X, f(X)) + \tfrac{1}{2}[X, f(X)]_n + R_n(X, f(X)), \tag{11}$$

where

$$R_n(X, f(X)) = \sum_{k=1}^{n} \int_{X_{k-1}}^{X_k} \left[f(x) - \frac{f(X_{k-1}) + f(X_k)}{2}\right] dx. \tag{12}$$

From analysis, it is well known that if the function $f''(x)$ is continuous, then the following formula ("trapezoidal rule") holds:

$$\begin{aligned}\int_a^b \left[f(x) - \frac{f(a) + f(b)}{2}\right] dx &= \int_a^b (x - a)(x - b)\frac{f''(\xi(x))}{2!}\, dx \\ &= \frac{(b-a)^3}{2}\int_0^1 x(x-1)f''(\xi(a + x(b-a)))\, dx \\ &= \frac{(b-a)^3}{2} f''(\xi(a + \bar{x}(b-a)))\int_0^1 x(x-1)\, dx \\ &= -\frac{(b-a)^3}{12} f''(\eta),\end{aligned}$$

where $\xi(x)$, $\bar{x}$ and η are "intermediate" points in the interval $[a, b]$.

Thus, in (11)

$$R_n(X, f(X)) = -\frac{1}{12}\sum_1^n f''(\eta_k)(\Delta X_k)^3$$

where $X_{k-1} \le \eta_k \le X_k$, whence

$$|R_n(X, f(X))| \le \frac{1}{12} \sup f''(\eta) \sum_{k=1}^{n} |\Delta X_k|^3$$

where the supremum is taken over all η such that

$$\min(X_0, X_1, \ldots, X_n) \le \eta \le \max(X_0, X_1, \ldots, X_n).$$

We shall refer to formula (11) as the *discrete analogue of Itô's formula*. We note that the right-hand side of this formula contains the following three 'natural' ingredients: 'the discrete integral' $I_n(X, f(X))$, the square covariance $[X, f(X)]$, and the 'residual' term $R_n(X, f(X))$.

4. EXAMPLE 1. If $f(x) = a + bx$, then $R_n(X, f(X)) = 0$ and formula (11) takes the following form:

$$F(X_n) = F(X_0) + I_n(X, f(X)) + \tfrac{1}{2}[X, f(X)]_n. \tag{13}$$

EXAMPLE 2. Let

$$f(x) = \operatorname{sign} x = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0. \end{cases}$$

Then $F(x) = |x|$.

Let $X_k = S_k$, where

$$S_k = \xi_1 + \xi_2 + \cdots + \xi_k,$$

where $\xi_1, \xi_2, \ldots$ are independent Bernoulli random variables taking values ± 1 with probability 1/2. If we also set $S_0 = 0$, we obtain

$$|S_n| = \sum_{k=1}^{n} \operatorname{sign} S_{k-1} \Delta S_k + N_n, \tag{14}$$

where $S_0 = 0$ and

$$N_n = \{0 \le k < n: S_k = 0\}.$$

We note that the sequence of discrete integrals $(\sum_{k=1}^{n} \operatorname{sign} S_{k-1} \Delta S_k)_{n \ge 1}$ forms a martingale and therefore,

$$E|S_n| = EN_n.$$

It is not difficult to show that as $n \to \infty$,

$$E|S_n| \sim \sqrt{\frac{n}{2\pi}}, \tag{15}$$

thus, application of Itô's discrete formula gives the asymptotic behavior of the average number of changes of sign on the path of the random walk $(S_k)_{k<n}$: $EN_n \sim \sqrt{\frac{n}{2\pi}}$.

5. **Remarks.** Let $B = (Bt)_{0 \le t \le 1}$ be a Brownian movement and $X_k = B_{k/n}$, $k = 0, 1, \ldots, n$. Then application of formula (11) leads to the following result:

$$F(B_1) = F(B_0) + \sum_{k=1}^{n} f(B_{(k-1)/n}) \Delta B_{k/n} + \frac{1}{2}[f(B_{\cdot/n}), B_{\cdot/n}]_n + R_n(B_{\cdot/n}, f(B_{\cdot/n})). \tag{16}$$

It is known from the stochastic calculus of Brownian motion that

$$\sum_{k=1}^{n} |B_{k/n} - B_{(k-1)/n}|^3 \xrightarrow{P} 0 \tag{17}$$

and if $f = f(x) \in L^2_{\text{loc}}$ (i.e., $\int_{|x| \le k} f^2(x)\,dx < \infty$ for any $k > 0$), then the limit

$$L^2\text{-}\lim_n \sum_{k=0}^{n} f(B_{(k-1)/n})\Delta B_{k/n}, \tag{18}$$

exists and is denoted by

$$\int_0^1 f(B_s)\,dB_s$$

and is called Itô's *stochastic integral* of $f(B_s)$ with respect to Brownian motion ([J1], [L12]). In addition, if the function $f(x)$ has a second derivative and $|f''(x)| \le C$, then from (9), we obtain that $P\text{-}\lim R_n(B_{\cdot/n}, f(B_{\cdot/n})) = 0$. Thus, from (8) and (10), it follows that the limit

$$P\text{-}\lim[f(B_{\cdot/n}), B_{\cdot/n}]_n$$

exists and is denoted by

$$[f(B), B]_1,$$

and that the following formula holds (P-a.s.)

$$F(B_1) = F(B_0) + \int_0^1 f(B_s)\,dB_s + \frac{1}{2}[f(B), B]_1. \tag{19}$$

It can be shown that, for the given smoothness assumptions

$$[f(B), B]_1 = \int_0^1 f'(B_s)\,ds, \tag{20}$$

which leads to the known *formula of Itô for Brownian motion*:

$$F(B_1) = F(B_0) + \int_0^1 f(B_s)\,dB_s + \frac{1}{2}\int_0^1 f'(B_s)\,ds. \tag{21}$$

6. Problems

1. Show that formula (14) is true.

2. Establish that the property (16) is true.

3. Prove formula (15).

§10. Applications to Calculations of the Probability of Ruin in Insurance

1. The material studied in the present section is a good illustration of the fact that the theory of martingales provides a quick and simple way of calculating the *risk* of an insurance company.

We shall assume that the evolution of the capital $X = (X_t)_{t\geq 0}$ of a certain insurance company takes place in a probability space $(\Omega, \mathscr{F}, \mathrm{P})$ as follows.

The initial capital is $X_0 = u > 0$. Insurance payments arrive continuously at a constant rate $c > 0$ (in time Δt the amount arriving is $c\Delta t$) and claims are received at random times $T_1, T_2, \dots$ $(0 < T_1 < T_2 < \cdots)$ where the amounts to be paid out at these times are described by a nonnegative random variables $\xi_1, \xi_2, \dots$.

Thus, taking into account receipts and claims, the capital X_t at time $t > 0$ is determined by the formula

$$X_t = u + ct - S_t, \tag{1}$$

where

$$S_t = \sum_{i\geq 1} \xi_i I(T_i \leq t). \tag{2}$$

We denote

$$T = \inf\{t \geq 0 : X_t \leq 0\}$$

the first time at which the insurance company's capital becomes less than or equal to zero ('time of ruin'). Of course, if $X_t > 0$ for *all* $t \geq 0$, then the time T is given to be equal to $+\infty$.

One of the main questions relating to the operation of an insurance company is the calculation of the *probability of ruin*, $P(T < \infty)$, and the *probability of ruin before time t*, $P(T \leq t)$ (inclusively).

2. To calculate these probabilities we assume we are in the framework of the classical Cramer–Lundeberg model characterized by the following assumptions:

A The times $T_1, T_2, \dots$ at which claims are received are such that the variables $(T_0 \equiv 0)$

$$\sigma_i = T_i - T_{i-1}, \qquad i \geq 1$$

are independent, identically-distributed random variables having an exponential distribution with density $\lambda e^{-\lambda t}$, $t \geq 0$ (see Table 2, §3, Chapter II).

B The random variables $\xi_1, \xi_2, \dots$ are independent and identically distributed with distribution function $F(x) = P(\xi_1 \leq x)$ such that $F(0) = 0$, $\mu = \int_0^\infty x\, dF(x) < \infty$.

C The sequences $(T_1, T_2, \dots)$ and $(\xi_1, \xi_2, \dots)$ are independent sequences (in the sense of Definition 6, §5, Chapter II).

We denote the process of the number of claims by $N = (N_t)_{t\geq 0}$, i.e., set

$$N_t = \sum_{i\geq 1} I(T_i \leq t). \tag{3}$$

It is clear that this process has a piecewise-constant trajectory with jumps by a unit value at times $T_1, T_2, \dots$ and with value $N_0 = 0$.

Since

$$\{T_k > t\} = \{\sigma_1 + \cdots + \sigma_k > t\} = \{N_t < k\},$$

under the assumption A we find that, according to Problem 6, §2, Chapter II,

$$P(N_t < k) = P(\sigma_1 + \cdots + \sigma_k > t) = \sum_{i=0}^{k-1} e^{-\lambda t}\frac{(\lambda t)^i}{i!}.$$

Whence

$$P(N_t = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}, \qquad k = 0, 1, \ldots, \tag{4}$$

i.e., the random variable N_t has a Poisson distribution (see Table 1. §3, Chapter II) with parameter λt. Here, $\mathsf{E}N_t = \lambda t$.

The so-called *Poisson process* $N = (N_t)_{t\geq 0}$ constructed in this way is (together with the *Brownian motion*; §13, Chapter II) an example of another classical random (stochastic) process with continuous time. Like the Brownian motion, the Poisson process is a process with independent increments (see §13, Chapter II), where, for $s < t$, the increments $N_t - N_s$ have a Poisson distribution with parameter $\lambda(t - s)$ (these properties are not difficult to derive from assumption A and the explicit construction (3) of the process N).

3. From assumption C we find that

$$\begin{aligned}\mathsf{E}(X_t - X_\cdot) &= ct - \mathsf{E}S_t = ct - \mathsf{E}\sum_i \xi_i I(T_i \leq t)\\ &= ct - \sum_i \mathsf{E}\xi_i \mathsf{E}I(T_i \leq t) = ct - \mu \sum_i P(T_i \leq t)\\ &= ct - \mu \sum_i P(N_t \geq i) = ct - \mu \mathsf{E}N_t = t(c - \lambda\mu).\end{aligned}$$

Thus, we see that, in the case under consideration, a natural requirement for an insurance company to operate with a clear profit (i.e. $\mathsf{E}(X_t - X_0) > 0$, $t > 0$) is that

$$c > \lambda\mu. \tag{5}$$

In the following analysis, an important role is played by the function

$$h(z) = \int_0^\infty (e^{zx} - 1)\, dF(x), \qquad z \geq 0, \tag{6}$$

which is equal to $\hat{F}(-z) - 1$, where

$$\hat{F}(s) = \int_0^\infty e^{-sx}\, dF(x)$$

is the Laplace–Stieltjes transformation (s is a complex number).

Denoting

$$g(z) = \lambda h(z) - cz \qquad \xi_0 \equiv 0,$$

we find that for $r > 0$ with $h(r) < \infty$,

$$\begin{aligned}\mathsf{E}e^{-r(X_t - X_0)} &= \mathsf{E}e^{-r(X_t - u)} = e^{-rct} \cdot \mathsf{E}e^{r\sum_{i=0}^{N_t}\xi_i} \\ &= e^{-rct} \sum_{n=0}^{\infty} \mathsf{E}e^{r\sum_{i=0}^{N_t}\xi_i} \mathsf{P}(N_t = n) \\ &= e^{-rct} \sum_{n=0}^{\infty} (1 + h(r))^n \frac{e^{-\lambda t}(\lambda t)^n}{n!} \\ &= e^{-rct} \cdot e^{\lambda t h(r)} = e^{t[\lambda h(r) - cr]} = e^{tg(r)}.\end{aligned}$$

Analogously, it can be shown that for any $s < t$

$$\mathsf{E}e^{-r(X_t - X_s)} = e^{(t-s)g(r)}. \tag{7}$$

Let $\mathscr{G}_t = \sigma(X_s, s \le t)$. Since the process $X = (X_t)_{t \ge 0}$ is a process with independent increments (Problem 2) (P-a.s.)

$$\mathsf{E}(e^{-r(X_t - X_s)} | \mathscr{G}_s) = \mathsf{E}e^{-r(X_t - X_s)} = e^{(t-s)g(r)},$$

then (P-a.s.)

$$\mathsf{E}(e^{-rX_t - tg(r)} | \mathscr{G}_s) = e^{-rX_s - sg(r)}. \tag{8}$$

Denoting

$$Z = e^{-rX_t - tg(r)}, \qquad t \ge 0 \tag{9}$$

we see that property (8) rewritten in the form

$$\mathsf{E}(Z_t | \mathscr{G}_s) = Z_s, \quad s \le t \tag{10}$$

is a continuous analogue of the *martingale property* (2) of Definition 1 of §1.

By analogy with Definition 3 of §1, we shall say that the random variable $\tau = \tau(\omega)$ with values in $[0, +\infty]$ is a *Markov time* relative to the system of σ-algebras $(\mathscr{G}_t)_{t \ge 0}$ if for each $t \ge 0$ the set

$$\{\tau(\omega) \le t\} \in \mathscr{G}_t.$$

The process $Z = (Z_t)_{t \ge 0}$ is nonnegative with $\mathsf{E}Z_t = e^{-ru} < \infty$. Thus, by analogy with Definition 1 of §1, the process $Z = (Z_t)_{t \ge 0}$ with *continuous time* is a martingale.

It turns out that for martingales with continuous time, Theorem 1 of §2 remains valid (with self-evident changes to the notation). In particular,

$$\mathsf{E}Z_{t \wedge \tau} = \mathsf{E}Z_0 \tag{11}$$

for any Markov time τ (see for example [L5], §2, Chapter 3).

By virtue of (9) we find from (11) that for time $\tau = T$

$$\begin{aligned}e^{-ru} &= \mathsf{E}e^{-rX_{t \wedge T} - (t \wedge T)g(r)} \\ &\ge \mathsf{E}[e^{-rX_{t \wedge T} - (t \wedge T)g(r)} | T \le t]\mathsf{P}(T \le t) \\ &= \mathsf{E}[e^{-rX_T - Tg(r)} | T \le t]P(T \le t) \\ &\ge \mathsf{E}[e^{-Tg(r)} | T \le t]P(T \le t) \ge \min_{0 \le s \le t} e^{-sg(r)} P(T \le t).\end{aligned}$$

Moreover,

$$P(T \le t) \le \frac{e^{-ru}}{\min_{0 \le s \le t} e^{-sg(r)}} = e^{-ru} \max_{0 \le s \le t} e^{sg(r)}. \tag{12}$$

Let us consider the function

$$g(r) = \lambda h(r) - cr$$

in more detail. Clearly, $g(0) = 0$, $g'(0) = \lambda\mu - c < 0$ (by virtue of (5)) and $g''(r) = \lambda h''(r) \ge 0$. Thus, there exists a unique positive value $r = R$ with $g(R) = 0$.

Noting that for $r > 0$

$$\begin{aligned}\int_0^\infty e^{rx}(1 - F(x))\,dx &= \int_0^\infty \int_x^\infty e^{rx}\,dF(y)\,dx \\ &= \int_0^\infty \left(\int_0^y e^{rx}\,dx\right) dF(y) \\ &= \frac{1}{r}\int_0^\infty (e^{ry} - 1)\,dF(y) = \frac{1}{r}h(r),\end{aligned}$$

R may be asserted to be the (unique) root of the equation

$$\frac{\lambda}{c}\int_0^\infty e^{rx}(1 - F(x))\,dx = 1. \tag{13}$$

Let us set $r = R$ in (12). Then we obtain, for any $t > 0$,

$$P(T \le t) \le e^{-Ru} \tag{14}$$

whence

$$P(T < \infty) \le e^{-Ru}. \tag{15}$$

Moreover, we prove the following

Theorem. *Suppose that in the Cramer–Lundeberg model assumptions* A, B, C *and property* (5) *are satisfied (i.e.,* $\lambda\mu < c$*). Then the bound of* (15) *holds for the probability of ruin* $P(T < \infty)$*, where* R *is the positive root of equation* (13).

4. **Remark.** The above discussion could be greatly simplified from a stochastic point of view if we assumed a geometric distribution for the σ_i ($P(\sigma_i = k) = q^{k-1}p$, $k = 1, 2, \ldots$) instead of an exponential distribution. In this case, all the time variables (T_i, T) would take discrete values, it would not be necessary to call upon the results of the theory of martingales for *continuous* time and the whole study could strictly be carried out based *only* on the 'discrete' (in time) methods from the theory of martingales studied in the present chapter.

However, we have turned our attention to the "continuous" (in time) scheme to illustrate both the method and the usefulness of the general theory of martingales for the case of *continuous time*, based on the given example.

5. Problems

1. Prove that the process $N = (N_t)_{t \geq 0}$ (under assumption A) is a process with independent increments, where $N_t - N_s$ has a Poisson distribution with parameter $\lambda(t - s)$.

2. Prove that the process $X = (X_t)_{t \geq 0}$ is a process with independent increments.

3. Consider the problem of determining the probability of ruin $P(T < \infty)$ assuming that the variables σ_i have a geometric (rather than exponential) distribution ($P(\sigma_i = k) = q^{k-1}p$, $k = 1, 2, \ldots$).

CHAPTER VIII

Sequences of Random Variables That Form Markov Chains

§1. Definitions and Basic Properties

1. In Chapter I (§12), for finite probability spaces, we took the basic idea to be that of *Markov dependence* between random variables. We also presented a variety of examples and considered the simplest regularities that are possessed by random variables that are connected by a Markov chain.

In the present chapter we give a general definition of a stochastic sequence of random variables that are connected by Markov dependence, and devote our main attention to the asymptotic properties of Markov chains with countable state spaces.

2. Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space with a distinguished nondecreasing family $(\mathscr{F}_n)$ of σ-algebras, $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \cdots \subseteq \mathscr{F}$.

Definition. A stochastic sequence $X = (X_n, \mathscr{F}_n)$ is called a *Markov chain* (with respect to the measure P) if

$$\mathsf{P}\{X_n \in B | \mathscr{F}_m\} = \mathsf{P}\{X_n \in B | X_m\} \quad (\mathsf{P}\text{-a.s.}) \tag{1}$$

for all $n \geq m \geq 0$ and all $B \in \mathscr{B}(R)$.

Property (1), the *Markov property*, can be stated in a number of ways. For example, it is equivalent to saying that

$$\mathsf{E}[g(X_n) | \mathscr{F}_m] = \mathsf{E}[g(X_n) | X_m] \quad (\mathsf{P}\text{-a.s.}) \tag{2}$$

for every bounded Borel function $g = g(x)$.

Property (1) is also equivalent to the statement that, for a given "present" X_m, the "future" F and the "past" P are independent, i.e.

$$\mathsf{P}(FP \mid X_m) = \mathsf{P}(F \mid X_m)\mathsf{P}(P \mid X_m), \tag{3}$$

where $F \in \sigma\{\omega: X_i, i \geq m\}$, and $B \in \mathscr{F}_n$, $n \leq m$.

In the special case when

$$\mathscr{F}_n = \mathscr{F}_n^X = \sigma\{\omega: X_0, \ldots, X_n\}$$

and the stochastic sequence $X = (X_n, \mathscr{F}_n^X)$ is a Markov chain, we say that *the sequence* $\{X_n\}$ *itself is a Markov chain.* It is useful to notice that if $X = \{X_n, \mathscr{F}_n\}$ is a Markov chain, then (X_n) is also a Markov chain.

Remark. It was assumed in the definition that the variables X_m are real-valued. In a similar way, we can also define Markov chains for the case when X_n takes values in some measurable space $(E, \mathscr{E})$. In this case, if all singletons are measurable, the space is called a *phase space*, and we say that $X = (X_n, \mathscr{F}_n)$ is a Markov chain with values in the phase space $(E, \mathscr{E})$. When E is finite or countably infinite (and $\mathscr{E}$ is the σ-algebra of all its subsets) we say that the Markov chain is *discrete*. In turn, a discrete chain with a finite phase space is called a *finite* chain.

The theory of finite Markov chains, as presented in §12, Chapter I, shows that a fundamental role is played by the one-step transition probabilities $\mathsf{P}(X_{n+1} \in B \mid X_n)$. By Theorem 3, §7, Chapter II, there are functions $P_{n+1}(x; B)$, the *regular conditional probabilities*, which (for given x) are measures on $(R, \mathscr{B}(R))$, and (for given B) are measurable functions of x, such that

$$\mathsf{P}(X_{n+1} \in B \mid X_n) = P_{n+1}(X_n; B) \quad (\mathsf{P}\text{-a.s.}). \tag{4}$$

The functions $P_n = P_n(x, B)$, $n \geq 0$, are called *transition functions*, and in the case when they coincide ($P_1 = P_2 = \cdots$), the corresponding Markov chain is said to be *homogeneous* (in time).

From now on we shall consider only homogeneous Markov chains, and the transition function $P_1 = P_1(x, B)$ will be denoted simply by $P = P(x, B)$.

Besides the transition function, an important probabilistic property of a Markov chain is the *initial distribution* $\pi = \pi(B)$, that is, the probability distribution defined by $\pi(B) = \mathsf{P}(X_0 \in B)$.

The set of pairs (π, P), where π is an initial distribution and P is a transition function, completely determines the probabilistic properties of X, since every finite-dimensional distribution can be expressed (Problem 2) in terms of π and P: for every $n \geq 0$ and $A \in \mathscr{B}(R^{n+1})$

$$\mathsf{P}\{(X_0, \ldots, X_n) \in A\} = \int_R \pi(dx_0) \int_R P(x_0; dx_1) \cdots \int_R I_A(x_0, \ldots, x_n) P(x_{n-1}; dx_n). \tag{5}$$

We deduce, by a standard limiting process, that for any $\mathscr{B}(R^{n+1})$-measurable function $g(x_0, \ldots, x_n)$, either of constant sign or bounded,

$$\mathsf{E}g(X_0, \ldots, X_n) = \int_R \pi(dx_0) \int_R P(x_0; dx_1) \cdots \int_R g(x_0, \ldots, x_n) P(x_{n-1}; dx_n). \quad (6)$$

3. Let $P^{(n)} = P^{(n)}(x; B)$ denote a regular variant of the *n-step transition probability*:

$$\mathsf{P}(X_n \in B \mid X_0) = P^{(n)}(X_0; B) \quad (\mathsf{P}\text{-a.s.}). \quad (7)$$

It follows at once from the Markov property that for all k and l, $(k, l \geq 1)$,

$$P^{(k+l)}(X_0; B) = \int_R P^{(k)}(X_0; dy) P^{(l)}(y; B) \quad (\mathsf{P}\text{-a.s.}). \quad (8)$$

It does *not* follow, of course, that for *all* $x \in R$

$$P^{(k+l)}(x; B) = \int_R P^{(k)}(x; dy) P^{(l)}(y; B). \quad (9)$$

It turns out, however, that regular variants of the transition probabilities *can be chosen* so that (9) will be satisfied for *all* $x \in R$ (see the discussion in the historical and bibliographical notes, p. 559).

Equation (9) is the *Kolmogorov–Chapman* equation (compare (I.12.13)) and is the starting point for the study of the probabilistic properties of Markov chains.

4. It follows from our discussion that with every Markov chain $X = (X_n, \mathscr{F}_n)$, defined on $(\Omega, \mathscr{F}, \mathsf{P})$ there is associated a set (π, P). It is natural to ask what properties a set (π, P) must have in order for $\pi = \pi(B)$ to be a probability distribution on $(R, \mathscr{B}(R))$ and for $P = P(x; B)$ to be a function that is measurable in x for given B, and a probability measure on B for every x, so that π will be the initial distribution, and P the transition function, for some Markov chain. As we shall now show, no additional hypotheses are required.

In fact, let us take $(\Omega, \mathscr{F})$ to be the measurable space $(R^\infty, \mathscr{B}(R^\infty))$. On the sets $A \in \mathscr{B}(R^{n+1})$ we define a probability measure by the right-hand side of formula (5). It follows from §9, Chapter II, that a probability measure P exists on $(R^\infty, \mathscr{B}(R^\infty))$ for which

$$\mathsf{P}\{\omega: (x_0, \ldots, x_n) \in A\} = \int_R \pi(dx_0) \int_R P(x_0; dx_1) \cdots \int_R I_A(x_0, \ldots, x_n) P(x_{n-1}; dx_n). \quad (10)$$

Let us show that if we put $X_n(\omega) = x_n$ for $\omega = (x_0, x_1, \ldots)$, the sequence $X = (X_n)_{n \geq 0}$ will constitute a Markov chain (with respect to the measure P just constructed).

In fact, if $B \in \mathscr{B}(R)$ and $C \in \mathscr{B}(R^{n+1})$, then

$$\begin{aligned}
\mathsf{P}\{X_{n+1} \in B, (X_0, \ldots, X_n) \in C\} \\
&= \int_R \pi(dx_0) \int_R P(x_0; dx_1) \cdots \int_R I_B(x_{n+1}) I_C(x_0, \ldots, x_n) P(x_n; dx_{n+1}) \\
&= \int_R \pi(dx_0) \int_R P(x_0; dx_1) \cdots \int_R P(x_n; B) I_C(x_0, \ldots, x_n) P(x_{n-1}; dx_n) \\
&= \int_{\{\omega: (X_0, \ldots, X_n) \in C\}} P(X_n; B)\, dP,
\end{aligned}$$

whence (P-a.s.)

$$\mathsf{P}\{X_{n+1} \in B \mid X_0, \ldots, X_n\} = P(X_n; B). \tag{11}$$

Similarly we can verify that (P-a.s.)

$$\mathsf{P}\{X_{n+1} \in B \mid X_n\} = P(X_n; B). \tag{12}$$

Equation (1) now follows from (11) and (12). It can be shown in the same way that for every $k \geq 1$ and $n \geq 0$,

$$\mathsf{P}\{X_{n+k} \in B \mid X_0, \ldots, X_n\} = \mathsf{P}\{X_{n+k} \in B \mid X_n\} \quad (\mathsf{P}\text{-a.s.}).$$

This implies the homogeneity of Markov chains.

The Markov chain $X = (X_n)$ that we have constructed is known as the Markov chain generated by (π, P). To emphasize that the measure P on $(R^\infty, \mathscr{B}(R^\infty))$ has precisely the initial distribution π, it is often denoted by P_π.

If π is concentrated at the single point x, we write P_x instead of P_π, and the corresponding Markov chain is called the chain *generated by the point* x (since $\mathsf{P}_x\{X_0 = x\} = 1$).

Consequently, each transition function $P = P(x, B)$ is in fact connected with the *whole family of probability measures* $\{\mathsf{P}_x, x \in R\}$, and therefore with the whole family of Markov chains that arise when the sequence $(X_n)_{n \geq 0}$ is considered with respect to the measures P_x, $x \in R$. From now on, we shall use the phrase "Markov chain with given transition function" to mean the family of Markov chains in the sense just described.

We observe that the measures P_π and P_x constructed from the transition function $P = P(x, B)$ are *consistent* in the sense that, when $A \in \mathscr{B}(R^\infty)$,

$$\mathsf{P}_\pi\{(X_0, X_1, \ldots) \in A \mid X_0 = x\} = \mathsf{P}_x\{(X_0, X_1, \ldots) \in A\} \quad (\pi\text{-a.s.}) \tag{13}$$

and

$$\mathsf{P}_\pi\{(X_0, X_1, \ldots) \in A\} = \int_R \mathsf{P}_x\{(X_0, X_1, \ldots) \in A\} \pi(dx). \tag{14}$$

5. Let us suppose that $(\Omega, \mathscr{F}) = (R^\infty, \mathscr{B}(R^\infty))$ and that we are considering a sequence $X = (X_n)$ that is defined coordinate-wise, that is, $X_n(\omega) = x_n$ for $\omega = (x_0, x_1, \ldots)$. Also let $\mathscr{F}_n = \sigma\{\omega: X_0, \ldots, X_n\}$, $n \geq 0$.

Let us define the *shifting operators* θ_n, $n \geq 0$, on Ω by the equation

$$\theta_n(x_0, x_1, \ldots) = (x_n, x_{n+1}, \ldots),$$

and let us define, for every random variable $\eta = \eta(\omega)$, the random variables $\theta_n \eta$ by putting

$$(\theta_n \eta)(\omega) = \eta(\theta_n \omega).$$

In this notation, the Markov property of homogeneous chains can (Problem 1) be given the following form: For every $\mathscr{F}$-measurable $\eta = \eta(\omega)$, every $n \geq 0$, and $B \in \mathscr{B}(R)$,

$$\mathsf{P}\{\theta_n \eta \in B \mid \mathscr{F}_n\} = \mathsf{P}_{X_n}\{\eta \in B\} \quad (\mathsf{P}\text{-a.s.}). \tag{15}$$

This form of the Markov property allows us to give the following important generalization: (15) remains valid if we replace n by stopping times τ.

Theorem. *Let $X = (X_n)$ be a homogeneous Markov chain defined on $(R^\infty, \mathscr{B}(R^\infty), \mathsf{P})$ and let τ be a stopping time. Then the following strong Markov property is valid:*

$$\mathsf{P}\{\theta_\tau \eta \in B \mid \mathscr{F}_\tau\} = \mathsf{P}_{X_\tau}\{\eta \in B\} \quad (\mathsf{P}\text{-a.s.}). \tag{16}$$

PROOF. If $A \in \mathscr{F}_\tau$ then

$$\begin{aligned} \mathsf{P}\{\theta_\tau \eta \in B, A\} &= \sum_{n=0}^{\infty} \mathsf{P}\{\theta_\tau \eta \in B, A, \tau = n\} \\ &= \sum_{n=0}^{\infty} \mathsf{P}\{\theta_n \eta \in B, A, \tau = n\}. \end{aligned} \tag{17}$$

The events $A \cap \{\tau = n\} \in \mathscr{F}_n$, and therefore

$$\begin{aligned} \mathsf{P}\{\theta_n \eta \in B, A \cap \{\tau = n\}\} &= \int_{A \cap \{\tau = n\}} \mathsf{P}\{\theta_n \eta \in B \mid \mathscr{F}_n\}\, d\mathsf{P} \\ &= \int_{A \cap \{\tau = n\}} \mathsf{P}_{X_n}\{\eta \in B\}\, d\mathsf{P} = \int_{A \cap \{\tau = n\}} \mathsf{P}_{X_\tau}\{\eta \in B\}\, d\mathsf{P}, \end{aligned}$$

which, with (17), establishes (16).

Corollary. *If σ is a stopping time such that $\mathsf{P}(\sigma \geq \tau) = 1$ and σ is $\mathscr{F}_\tau$-measurable, then*

$$\mathsf{P}\{X_\sigma \in B, \sigma < \infty \mid \mathscr{F}_\tau\} = \mathsf{P}_{X_\tau}(B) \qquad (\{\sigma < \infty\}; \mathsf{P}\text{-}a.s.). \tag{18}$$

6. As we said above, we are going to consider only discrete Markov chains (with phase space $E = \{\ldots, i, j, k, \ldots\}$). To simplify the notation, we shall now denote the transition functions $P(i; \{j\})$ by p_{ij} and call them transition

probabilities; an n-step transition probability from i to j will be denoted by $p_{ij}^{(n)}$.

Let $E = \{1, 2, \ldots\}$. The principal questions that we study in §§2–4 are intended to clarify the conditions under which:

(A) The *limits* $\pi_j = \lim p_{ij}^{(n)}$ *exist* and are independent of i;
(B) The limits $(\pi_1, \pi_2, \ldots)$ form a *probability distribution*, that is, $\pi_i \geq 0$, $\pi_i = 1$;
(C) The chain is *ergodic*, that is, the limits $(\pi_1, \pi_2, \ldots)$ have the properties $\pi_i > 0$, $\sum_{i=1}^{\infty} \pi_i = 1$;
(D) There is one and only one *stationary probability distribution* $\mathbb{Q} = (q_1, q_2, \ldots)$, that is, one such that $q_i \geq 0$, $\sum_{i=1}^{\infty} q_i = 1$, and $q_j = \sum_i q_i p_{ij}$, $j \in E$.

In the course of answering these questions we shall develop a classification of the states of a Markov chain as they depend on the arithmetic and asymptotic properties of $p_{ij}^{(n)}$ and $p_{ii}^{(n)}$.

7. Problems

1. Prove the equivalence of definitions (1), (2), (3) and (15) of the Markov property.

2. Prove formula (5).

3. Prove equation (18).

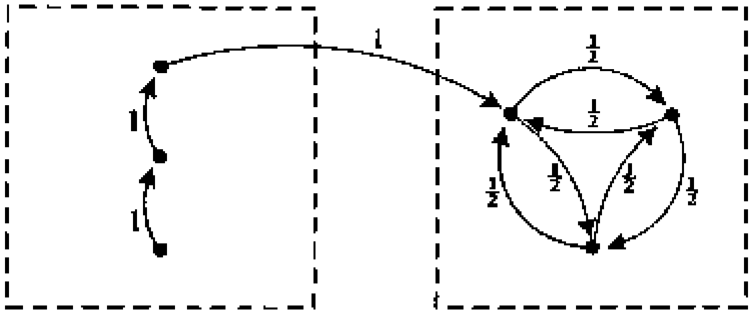
4. Let $(X_n)_{n\geq 0}$ be a Markov chain. Show that the reversed sequence $(\ldots, X_n, X_{n-1}, \ldots, X_0)$ is also a Markov chain.

§2. Classification of the States of a Markov Chain in Terms of Arithmetic Properties of the Transition Probabilities $p_{ij}^{(n)}$

1. We say that a state $i \in E = \{1, 2, \ldots\}$ is *inessential* if, with positive probability, it is possible to escape from it after a finite number of steps, without ever returning to it; that is, there exist m and j such that $p_{ij}^{(m)} > 0$, but $p_{ji}^{(n)} = 0$ for all n and j.

Let us delete all the inessential states from E. Then the remaining set of *essential* states has the property that a wandering particle that encounters it can never leave it (Figure 36). As will become clear later, it is essential states that are the most interesting.

Let us now consider the set of essential states. We say that state j is *accessible* from the point i $(i \to j)$ if there is an $m \geq 0$ such that $p_{ij}^{(m)} > 0$



Inessential state Essential state

Figure 36

($p_{ij}^{(0)} = 1$ if $i = j$, and 0 if $i \neq j$). States i and j *communicate* ($i \leftrightarrow j$) if j is accessible from i and i is accessible from j.

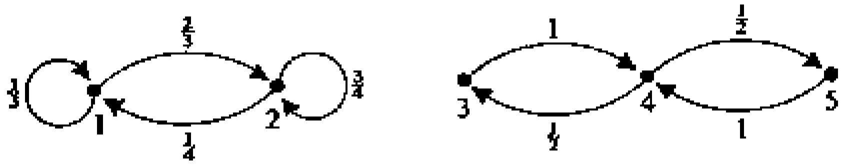
By the definition, the relation "$\leftrightarrow$" is symmetric and reflexive. It is easy to verify that it is also transitive ($i \leftrightarrow j$, $j \leftrightarrow k \Rightarrow i \leftrightarrow k$). Consequently the set of essential states separates into a finite or countable number of disjoint sets $E_1, E_2, \ldots$, each of which consists of communicating sets but with the property that passage between different sets is impossible.

By way of abbreviation, we call the sets $E_1, E_2, \ldots$ *classes* or *indecomposable classes* (of essential communicating sets), and we call a Markov chain *indecomposable* if its states form a single indecomposable class.

As an illustration we consider the chain with matrix

$$\mathbb{P} = \left\| \begin{array}{cc:ccc} \frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 \end{array} \right\| = \begin{pmatrix} \mathbb{P}_1 & 0 \\ 0 & \mathbb{P}_2 \end{pmatrix}.$$

The graph of this chain, with set of states $E = \{1, 2, 3, 4, 5\}$ has the form



It is clear that this chain has two indecomposable classes $E_1 = \{1, 2\}$, $E_2 = \{3, 4, 5\}$, and the investigation of their properties reduces to the investigation of the two separate chains whose states are the sets E_1 and E_2, and whose transition matrices are $\mathbb{P}_1$ and $\mathbb{P}_2$.

Now let us consider any indecomposable class E, for example the one sketched in Figure 37.

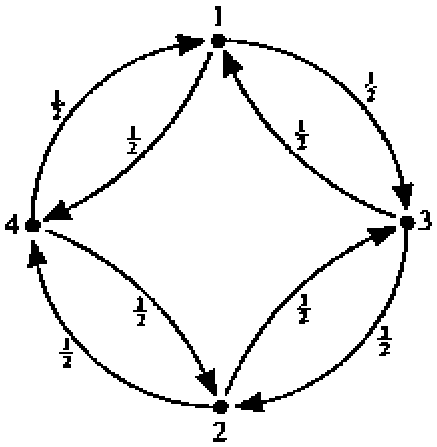


Figure 37. Example of a Markov chain with period $d = 2$.

Observe that in this case a return to each state is possible only after an even number of steps; a transition to an adjacent state, after an odd number; the transition matrix has block structure,

$$\mathbb{P} = \left(\begin{array}{cc|cc} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \hline \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{array}\right).$$

Therefore it is clear that the class $E = \{1, 2, 3, 4\}$ separates into two subclasses $C_0 = \{1, 2\}$ and $C_1 = \{3, 4\}$ with the following *cyclic* property: after one step from C_0 the particle necessarily enters C_1, and from C_1 it returns to C_0.

This example suggests a classification of indecomposable classes into *cyclic subclasses*.

2. Let us say that state j has period $d = d(j)$ if the following two conditions are satisfied:

(1) $p_{jj}^{(n)} > 0$ only for values of n of the form dm;
(2) d is the largest number satisfying (1).

In other words, d is the *greatest common divisor* of the numbers n for which $p_{jj}^{(n)} > 0$. (If $p_{jj}^{(n)} = 0$ for all $n \geq 1$, we put $d(j) = 0$.)

Let us show that all states of a single indecomposable class E have the same period d, which is therefore naturally called the *period* of the class, $d = d(E)$.

Let i and $j \in E$. Then there are numbers k and l such that $p_{ij}^{(k)} > 0$ and $p_{ji}^{(l)} > 0$. Consequently $p_{ii}^{(k+l)} \geq p_{ij}^{(k)} p_{ji}^{(l)} > 0$, and therefore $k + l$ is divisible by $d(i)$. Suppose that $n > 0$ and n is not divisible by $d(i)$. Then $n + k + l$ is also not divisible by $d(i)$ and consequently $p_{ii}^{(n+k+l)} = 0$. But

$$p_{ii}^{(n+k+l)} \geq p_{ij}^{(k)} p_{jj}^{(n)} p_{ji}^{(l)}$$

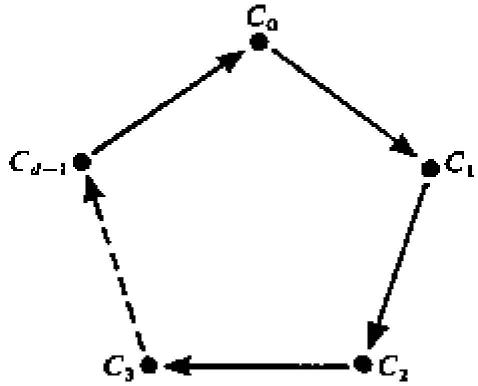


Figure 38. Motion among cyclic subclasses.

and therefore $p_{jj}^{(n)} = 0$. It follows that if $p_{jj}^{(n)} > 0$ we have n divisible by $d(i)$, and therefore $d(i) \leq d(j)$. By symmetry, $d(j) \leq d(i)$. Consequently $d(i) = d(j)$.

If $d(j) = 1$ ($d(E) = 1$), the state j (or class E) is said to be *aperiodic*.

Let $d = d(E)$ be the period of an indecomposable class E. The transitions within such a class may be quite freakish, but (as in the preceding example) there is a cyclic character to the transitions from one group of states to another. To show this, let us select a state i_0 and introduce (for $d \geq 1$) the following subclasses:

$$C_0 = \{j \in E: p_{i_0 j}^{(n)} > 0 \Rightarrow n \equiv 0 (\text{mod } d)\};$$

$$C_1 = \{j \in E: p_{i_0 j}^{(n)} > 0 \Rightarrow n \equiv 1 (\text{mod } d)\};$$

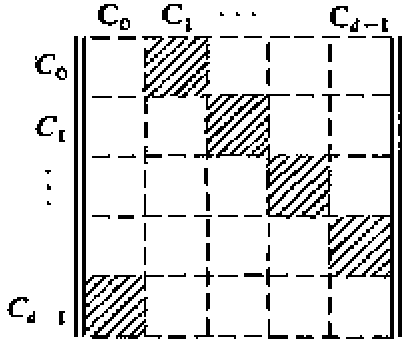
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$C_{d-1} = \{j \in E: p_{i_0 j}^{(n)} > 0 \Rightarrow n \equiv d - 1 (\text{mod } d)\}.$$

Clearly $E = C_0 + C_1 + \cdots + C_{d-1}$. Let us show that the motion from subclass to subclass is as indicated in Figure 38.

In fact, let state $i \in C_p$ and $p_{ij} > 0$. Let us show that necessarily $j \in C_{p+1 (\text{mod } d)}$. Let n be such that $p_{i_0 i}^{(n)} > 0$. Then $n = ad + p$ and therefore $n \equiv p (\text{mod } d)$ and $n + 1 \equiv p + 1 (\text{mod } d)$. Hence $p_{i_0 j}^{(n+1)} > 0$ and $j \in C_{p+1 (\text{mod } d)}$.

Let us observe that it now follows that the transition matrix $\mathbb{P}$ of an indecomposable chain has the following block structure:



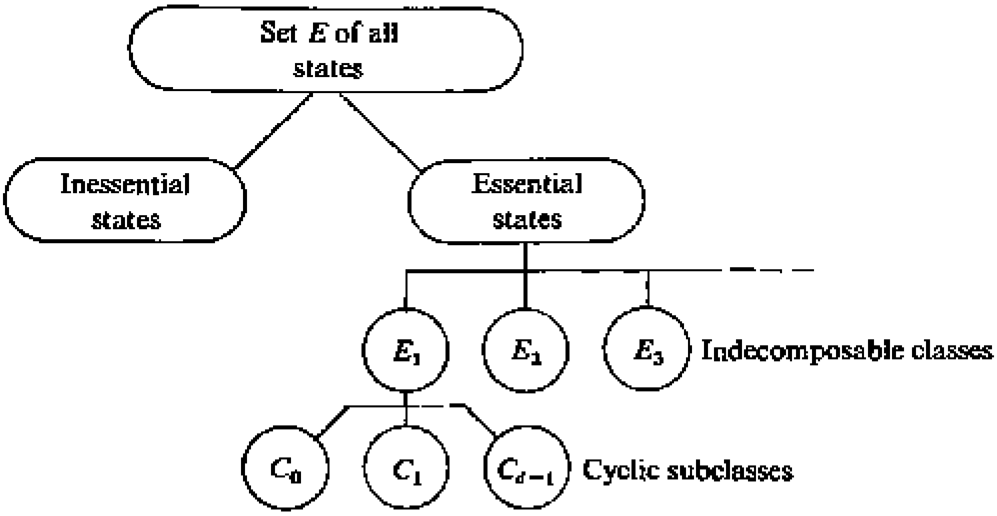


Figure 39. Classification of states of a Markov chain in terms of arithmetic properties of the probabilities p_{ij}^n.

Consider a subclass C_p. If we suppose that a particle is in the set C_0 at the initial time, then at time $s = p + dt$, $t = 0, 1, \ldots$, it will be in the subclass C_p. Consequently, with each subclass C_p we can connect a new Markov chain with transition matrix $(p_{ij}^d)_{i,j \in C_p}$, which is indecomposable and aperiodic. Hence if we take account of the classification that we have outlined (see the summary in Figure 39) we infer that in studying problems on limits of probabilities $p_{ij}^{(n)}$ we can restrict our attention to *aperiodic indecomposable chains*.

3. Problems

1. Show that the relation "$\leftrightarrow$" is transitive.

2. For Example 1, §5, show that when $0 < p < 1$, all states belong to a single class with period $d = 2$.

3. Show that the Markov chains discussed in Examples 4 and 5 of §5 are aperiodic.

§3. Classification of the States of a Markov Chain in Terms of Asymptotic Properties of the Probabilities $p_{ii}^{(n)}$

1. Let $\mathbb{P} = \|p_{ij}\|$ be the transition matrix of a Markov chain,

$$f_{ii}^{(k)} = \mathsf{P}_i\{X_k = i, X_l \neq i, 1 \leq l \leq k-1\} \tag{1}$$

and for $i \neq j$

$$f_{ij}^{(k)} = \mathsf{P}_i\{X_k = j, X_l \neq j, 1 \leq l \leq k-1\}. \tag{2}$$

For $X_0 = i$, these are respectively the *probability of first return to state i at time k*, and the *probability of first arrival at state j at time k*.

Using the strong Markov property (1.16), we can show as in (I.12.38) that

$$p_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)}. \tag{3}$$

For each $i \in E$ we introduce

$$f_{ii} = \sum_{n=1}^{\infty} f_{ii}^{(n)}, \tag{4}$$

which is the probability that a particle that leaves state i will sooner or later return to that state. In other words, $f_{ii} = \mathsf{P}_i\{\sigma_i < \infty\}$, where $\sigma_i = \inf\{n \geq 1: X_n = i\}$ with $\sigma_i = \infty$ when $\{\cdot\} = \varnothing$.

We say that a state i is *recurrent* if

$$f_{ii} = 1,$$

and *nonrecurrent* if

$$f_{ii} < 1.$$

Every recurrent state can, in turn, be classified according to whether the *average time of return* is finite or infinite.

Let us say that a recurrent state i is *positive* if

$$\mu_i^{-1} \equiv \left(\sum_{n=1}^{\infty} n f_{ii}^{(n)}\right)^{-1} > 0,$$

and *null* if

$$\mu_i^{-1} \equiv \left(\sum_{n=1}^{\infty} n f_{ii}^{(n)}\right)^{-1} = 0.$$

Thus we obtain the classification of the states of the chain, as displayed in Figure 40.

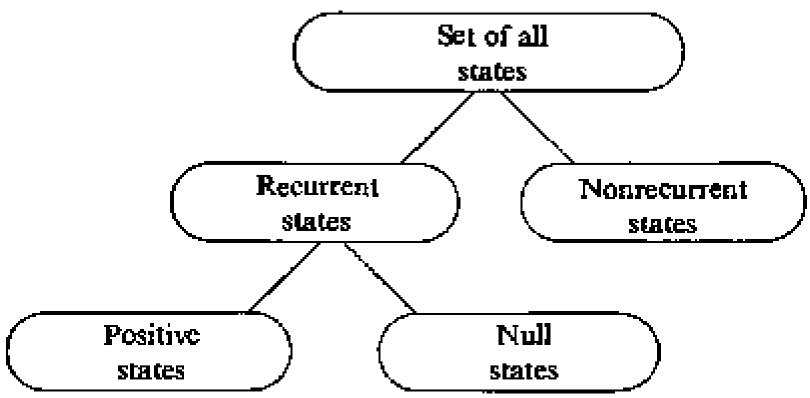


Figure 40. Classification of the states of a Markov chain in terms of the asymptotic properties of the probabilities $p_{ii}^{(n)}$.

2. Since the calculation of the functions $f_{ii}^{(n)}$ can be quite complicated, it is useful to have the following tests for whether a state i is recurrent or not.

Lemma 1

(a) *The state i is recurrent if and only if*

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty. \tag{5}$$

(b) *If state j is recurrent and $i \leftrightarrow j$ then state i is also recurrent.*

PROOF. (a) By (3),

$$p_{ii}^{(n)} = \sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)},$$

and therefore (with $p_{ii}^{(0)} = 1$)

$$\begin{aligned}\sum_{n=1}^{\infty} p_{ii}^{(n)} &= \sum_{n=1}^{\infty} \sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)} = \sum_{k=1}^{\infty} f_{ii}^{(k)} \sum_{n=k}^{\infty} p_{ii}^{(n-k)} \\ &= f_{ii} \sum_{n=0}^{\infty} p_{ii}^{(n)} = f_{ii}\left(1 + \sum_{n=1}^{\infty} p_{ii}^{(n)}\right).\end{aligned}$$

Therefore if $\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty$, we have $f_{ii} < 1$ and therefore state i is nonrecurrent. Furthermore, let $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$. Then

$$\sum_{n=1}^{N} p_{ii}^{(n)} = \sum_{n=1}^{N} \sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)} = \sum_{k=1}^{N} f_{ii}^{(k)} \sum_{n=k}^{N} p_{ii}^{(n-k)} \le \sum_{k=1}^{N} f_{ii}^{(k)} \sum_{l=0}^{N} p_{ii}^{(l)},$$

and therefore

$$f_{ii} = \sum_{k=1}^{\infty} f_{ii}^{(k)} \ge \sum_{k=1}^{N} f_{ii}^{(k)} \ge \frac{\sum_{n=1}^{N} p_{ii}^{(n)}}{\sum_{l=0}^{N} p_{ii}^{(l)}} \to 1, \qquad N \to \infty.$$

Thus if $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$ then $f_{ii} = 1$, that is, the state i is recurrent.

(b) Let $p_{ij}^{(s)} > 0$ and $p_{ji}^{(t)} > 0$. Then

$$p_{ii}^{(n+s+t)} \ge p_{ij}^{(s)} p_{jj}^{(n)} p_{ji}^{(t)},$$

and if $\sum_{n=1}^{\infty} p_{jj}^{(n)} = \infty$, then also $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$, that is, the state i is recurrent.

3. From (5) it is easy to deduce a first result on the asymptotic behavior of $p_{ij}^{(n)}$.

Lemma 2. *If state j is nonrecurrent then*

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} < \infty \tag{6}$$

for every i, and therefore

$$p_{ij}^{(n)} \to 0, \qquad n \to \infty. \tag{7}$$

PROOF. By (3) and Lemma 1,

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} = \sum_{n=1}^{\infty} \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)} = \sum_{k=1}^{\infty} f_{ij}^{(k)} \sum_{n=0}^{\infty} p_{jj}^{(n)}$$

$$= f_{ij} \sum_{n=0}^{\infty} p_{jj}^{(n)} \le \sum_{n=0}^{\infty} p_{jj}^{(n)} < \infty.$$

Here we used the inequality $f_{ij} = \sum_{k=1}^{\infty} f_{ij}^{(k)} \le 1$, which holds because the series represents the probability that a particle starting at i eventually arrives at j. This establishes (6) and therefore (7).

Let us now consider recurrent states.

Lemma 3. *Let j be a recurrent state with $d(j) = 1$.*

(a) *If i communicates with j, then*

$$p_{ij}^{(n)} \to \frac{1}{\mu_j}, \qquad n \to \infty. \tag{8}$$

If in addition j is a positive state then

$$p_{ij}^{(n)} \to \frac{1}{\mu_j} > 0, \qquad n \to \infty. \tag{9}$$

If, however, j is a null state, then

$$p_{ij}^{(n)} \to 0, \qquad n \to \infty. \tag{10}$$

(b) *If i and j belong to different classes of communicating states, then*

$$p_{ij}^{(n)} \to \frac{f_{ij}}{\mu_j}, \qquad n \to \infty. \tag{11}$$

The proof of the lemma depends on the following theorem from analysis.

Let $f_1, f_2, \ldots$ be a sequence of nonnegative numbers with $\sum_{i=1}^{\infty} f_i = 1$, such that the greatest common divisor of the indices j for which $f_j > 0$ is 1. Let $u_0 = 1$, $u_n = \sum_{k=1}^{n} f_k u_{n-k}$, $n = 1, 2, \ldots$, and let $\mu = \sum_{n=1}^{\infty} n f_n$. Then $u_n \to 1/\mu$ as $n \to \infty$. (For a proof, see [F1], §10 of Chapter XIII.)

Taking account of (3), we apply this to $u_n = p_{jj}^{(n)}$, $f_k = f_{jj}^{(k)}$. Then we immediately find that

$$p_{jj}^{(n)} \to \frac{1}{\mu_j},$$

where $\mu_j = \sum_{n=1}^{\infty} n f_{jj}^{(n)}$.

Taking $p_{jj}^{(s)} = 0$ for $s < 0$, we can rewrite (3) in the form

$$p_{ij}^{(n)} = \sum_{k=1}^{\infty} f_{ij}^{(k)} p_{jj}^{(n-k)}. \tag{12}$$

By what has been proved, we have $p_{jj}^{(n-k)} \to \mu_j^{-1}$, $n \to \infty$, for each given k. Therefore if we suppose that

$$\lim_n \sum_{k=1}^{\infty} f_{ij}^{(k)} p_{jj}^{(n-k)} = \sum_{k=1}^{\infty} f_{ij}^{(k)} \lim_n p_{jj}^{(n-k)}, \tag{13}$$

we immediately obtain

$$p_{ij}^{(n)} \to \frac{1}{\mu_j}\left(\sum_{k=1}^{\infty} f_{ij}^{(k)}\right) = \frac{1}{\mu_j} f_{ij}, \tag{14}$$

which establishes (11).

Recall that f_{ij} is the probability that a particle starting from state i arrives, sooner or later, at state j. State j is recurrent, and if i communicates with j, it is natural to suppose that $f_{ij} = 1$. Let us show that this is indeed the case.

Let f'_{ij} be the probability that a particle, starting from state i, visits state j infinitely often. Clearly $f_{ij} \geq f'_{ij}$. Therefore if we show that, for a recurrent state j and a state i that communicates with it, the probability $f'_{ij} = 1$, we will have established that $f_{ij} = 1$.

According to part (b) of Lemma 1, the state i is also recurrent, and therefore

$$f_{ii} = \sum f_{ii}^{(n)} = 1. \tag{15}$$

Let

$$\sigma_i = \inf\{n \geq 1 : X_n = i\}$$

be the first time (for times $n \geq 1$) at which the particle reaches state i; take $\sigma_i = \infty$ if no such time exists.

Then

$$1 = f_{ii} = \sum_{n=1}^{\infty} f_{ii}^{(n)} = \sum_{n=1}^{\infty} \mathsf{P}_i(\sigma_i = n) = \mathsf{P}_i(\sigma_i < \infty), \tag{16}$$

and consequently to say that state i is recurrent means that a particle starting at i will eventually return to the same state (at a random time σ_i). But after returning to this state the "life" of the particle starts over, so to speak (because of the strong Markov property). Hence it appears that if state i is recurrent the particle must return to it infinitely often:

$$\mathsf{P}_i\{X_n = i \text{ for infinitely many } n\} = 1. \tag{17}$$

Let us now give a formal proof.

Let i be a state (recurrent or nonrecurrent). Let us show that the probability of return to that state at least r times is $(f_{ii})^r$.

For $r = 1$ this follows from the definition of f_{ii}. Suppose that the proposition has been proved for $r = m - 1$. Then by using the strong Markov property and (16), we have

$$\begin{aligned}
&\mathsf{P}_i\,(\text{number of returns to } i \text{ is greater than or equal to } m)\\
&= \sum_{k=1}^{\infty} \mathsf{P}_i\begin{pmatrix}\sigma_i = k, \text{ and the number of returns to } i \text{ after time } k\\ \text{is at least } m-1\end{pmatrix}\\
&= \sum_{k=1}^{\infty} \mathsf{P}_i(\sigma_i = k)\mathsf{P}_i\left(\begin{matrix}\text{at least } m-1 \text{ values}\\ \text{of } X_{\sigma_i+1}, X_{\sigma_i+2}, \ldots \text{ equal } i\end{matrix}\,\middle|\,\sigma_i = k\right)\\
&= \sum_{k=1}^{\infty} \mathsf{P}_i(\sigma_i = k)\mathsf{P}_i(\text{at least } m-1 \text{ values of } X_1, X_2, \ldots \text{ equal } i)\\
&= \sum_{k=1}^{\infty} f_{ii}^{(k)}(f_{ii})^{m-1} = f_{ii}^{m}.
\end{aligned}$$

Hence it follows in particular that formula (17) holds for a recurrent state i. If the state is nonrecurrent, then

$$\mathsf{P}_i\{X_n = i \text{ for infinitely many } n\} = 0. \tag{18}$$

We now turn to the proof that $f'_{ij} = 1$. Since the state i is recurrent, we have by (17) and the strong Markov property

$$\begin{aligned}
1 &= \sum_{k=1}^{\infty} \mathsf{P}_i(\sigma_j = k) + \mathsf{P}_i(\sigma_j = \infty)\\
&= \sum_{k=1}^{\infty} \mathsf{P}_i\begin{pmatrix}\sigma_j = k, & \begin{matrix}\text{the number of returns to } i\\ \text{after time } k \text{ is infinite}\end{matrix}\end{pmatrix} + \mathsf{P}_i(\sigma_j = \infty)\\
&= \sum_{k=1}^{\infty} \mathsf{P}_i\begin{pmatrix}\sigma_j = k, & \begin{matrix}\text{infinitely many values of}\\ X_{\sigma_j+1}, X_{\sigma_j+2}, \ldots \text{ equal } i\end{matrix}\end{pmatrix} + \mathsf{P}_i(\sigma_j = \infty)\\
&= \sum_{k=1}^{\infty} \mathsf{P}_i(\sigma_j = k)\cdot \mathsf{P}_i\left(\begin{matrix}\text{infinitely many}\\ \text{values of } X_{\sigma_j+1}, X_{\sigma_j+2},\\ \ldots, \text{ equal } i\end{matrix}\,\middle|\,\begin{matrix}\sigma_j = k\\ X_{\sigma_j} = j\end{matrix}\right) + \mathsf{P}_i(\sigma_j = \infty)\\
&= \sum_{k=1}^{\infty} f_{ij}^{(k)}\cdot \mathsf{P}_j\begin{pmatrix}\text{infinitely many values}\\ \text{of } X_1, X_2, \ldots \text{ equal } i\end{pmatrix} + (1 - f_{ij})\\
&= \sum_{k=1}^{\infty} f_{ij}^{(k)} f'_{ij} + (1 - f_{ij}) = f'_{ij} f_{ij} + (1 - f_{ij}).
\end{aligned}$$

Thus

$$1 = f'_{ij} f_{ij} + 1 - f_{ij}$$

and therefore

$$f_{ij} = f'_{ij}\cdot f_{ij}.$$

Since $i \leftrightarrow j$, we have $f_{ij} > 0$, and consequently $f'_{ij} = 1$ and $f_{ij} = 1$.

Therefore if we assume (13), it follows from (14) and the equation $f_{ij} = 1$ that, for communicating states i and j,

$$p_{ij}^{(n)} \to \frac{1}{\mu_j}, \qquad n \to \infty.$$

As for (13), its validity follows from the theorem on dominated convergence together with the remark that

$$p_{jj}^{(n-k)} \to \frac{1}{\mu_j}, \quad n \to \infty, \qquad \sum_{k=1}^{\infty} f_{ij}^{(k)} = f_{ij} \leq 1.$$

This completes the proof of the lemma.

Next we consider periodic states.

Lemma 4. *Let j be a recurrent state and let $d(j) > 1$.*

(a) *If i and j belong to the same class (of states), and if i belongs to the cyclic subclass C_r and j to C_{r+a}, then*

$$p_{ij}^{(nd+a)} \to \frac{d}{\mu_j}. \tag{19}$$

(b) *With an arbitrary i,*

$$p_{ij}^{(nd+a)} \to \left[\sum_{r=0}^{\infty} f_{ij}^{(rd+a)}\right] \cdot \frac{d}{\mu_j}, \qquad a = 0, 1, \ldots, d-1. \tag{20}$$

Proof. (a) First let $a = 0$. With respect to the transition matrix $\mathbb{P}^d$ the state j is recurrent and aperiodic. Consequently, by (8),

$$p_{ij}^{(nd)} \to \frac{1}{\sum_{k=1}^{\infty} k f_{jj}^{(kd)}} = \frac{d}{\sum_{k=1}^{\infty} kd f_{jj}^{(kd)}} = \frac{d}{\mu_j}.$$

Suppose that (19) has been proved for $a = r$. Then

$$p_{ij}^{(nd+r+1)} = \sum_{k=1}^{\infty} p_{ik} p_{kj}^{(nd+r)} \to \sum_{k=1}^{\infty} p_{ik} \cdot \frac{d}{\mu_j} = \frac{d}{\mu_j}.$$

(b) Clearly

$$p_{ij}^{(nd+a)} = \sum_{k=1}^{nd+a} f_{ij}^{(k)} p_{jj}^{(nd+a+k)}, \qquad a = 0, 1, \ldots, d-1.$$

State j has period d, and therefore $p_{jj}^{(nd+a-k)} = 0$, except when $k - a$ has the form $r \cdot d$. Therefore

$$p_{ij}^{(nd+a)} = \sum_{r=0}^{n} f_{ij}^{(rd+a)} p_{jj}^{((n-r)d)}$$

and the required result (20) follows from (19).

This completes the proof of the lemma.

Lemmas 2–4 imply, in particular, the following result about limits of $p_{ij}^{(n)}$.

Theorem 1. *Let a Markov chain be indecomposable (that is, its states form a single class of essential communicating states) and aperiodic.*

Then:

(a) *If all states are either null or nonrecurrent, then, for all i and j,*

$$p_{ij}^{(n)} \to 0, \qquad n \to \infty; \tag{21}$$

(b) *if all states j are positive, then, for all i,*

$$p_{ij}^{(n)} \to \frac{1}{\mu_j} > 0, \qquad n \to \infty; \tag{22}$$

4. Let us discuss the conclusion of this theorem in the case of a Markov chain with a finite number of states, $E = \{1, 2, \dots, r\}$. Let us suppose that the chain is indecomposable and aperiodic. It turns out that then it is automatically increasing and positive:

$$\begin{pmatrix} \text{indecomposability} \\ d = 1 \end{pmatrix} \Rightarrow \begin{pmatrix} \text{indecomposability} \\ \text{recurrence} \\ \text{positivity} \\ d = 1 \end{pmatrix} \tag{23}$$

For the proof, we suppose that all states are nonrecurrent. Then by (21) and the finiteness of the set of states of the chain,

$$1 = \lim_n \sum_{j=1}^r p_{ij}^{(n)} = \sum_{j=1}^r \lim_n p_{ij}^{(n)} = 0. \tag{24}$$

The resulting contradiction shows that not all states can be nonrecurrent. Let i_0 be a recurrent state and j an arbitrary state. Since $i_0 \leftrightarrow j$, Lemma 1 shows that j is also recurrent.

Thus all states of an aperiodic indecomposable chain are recurrent.

Let us now show that all recurrent states are positive.

If we suppose that they are all null states, we again obtain a contradiction with (24). Consequently there is at least one positive state, say i_0. Let i be any other state. Since $i \leftrightarrow i_0$, there are s and t such that $p_{i_0 i}^{(s)} > 0$ and $p_{i i_0}^{(t)} > 0$, and therefore

$$p_{ii}^{(n+s+t)} \geq p_{ii_0}^{(s)} p_{i_0 i_0}^{(n)} p_{i_0 i}^{(t)} \to p_{ii_0}^{(s)} \frac{1}{\mu_{i_0}} \cdot p_{i_0 i}^{(t)} > 0. \tag{25}$$

Hence there is a positive ε such that $p_{ii}^{(n)} \geq \varepsilon > 0$ for all sufficiently large n. But $p_{ii}^{(n)} \to 1/\mu_i$ and therefore $\mu_i > 0$. Consequently (23) is established.

Let $\pi_j = 1/\mu_j$. Then $\pi_j > 0$ by (22) and since

$$1 = \lim_n \sum_{j=1}^r p_{ij}^{(n)} = \sum_{j=1}^r \pi_j,$$

the (aperiodic indecomposable) chain is ergodic. Clearly, for all ergodic finite chains,

$$\textit{there is an } n_0 \textit{ such that } \min_{i,j} p_{ij}^{(n)} > 0 \textit{ for all } n \geq n_0. \tag{26}$$

It was shown in §12 of Chapter I that the converse is also valid: (26) implies ergodicity.

Consequently we have the following implications:

$$\begin{pmatrix}\text{indecomposability}\\ d = 1\end{pmatrix} \Leftrightarrow \begin{pmatrix}\text{indecomposability}\\ \text{recurrence}\\ \text{positivity}\\ d = 1\end{pmatrix} \Rightarrow \text{ergodicity} \Leftrightarrow (26).$$

However, we can prove more.

Theorem 2. *For a finite Markov chain*

$$\begin{pmatrix}\textit{indecomposability}\\ d = 1\end{pmatrix} \Leftrightarrow \begin{pmatrix}\textit{indecomposability}\\ \textit{recurrence}\\ \textit{positivity}\\ d = 1\end{pmatrix} \Leftrightarrow (\textit{ergodicity}) \Leftrightarrow (26).$$

PROOF. We have only to establish

$$(\text{ergodicity}) \Rightarrow \begin{pmatrix}\text{indecomposability}\\ \text{recurrence}\\ \text{positivity}\\ d = 1\end{pmatrix}.$$

Indecomposability follows from (26). As for aperiodicity, increasingness, and positivity, they are valid in more general situations (the existence of a limiting distribution is sufficient), as will be shown in Theorem 2, §4.

5. PROBLEMS

1. Consider an indecomposable chain with states 0, 1, 2, ... A necessary and sufficient condition for it to be nonrecurrent is that the system of equations $u_j = \sum_i u_i p_{ij}$, $j = 0, 1, \ldots$, has a bounded solution such that $u_i \not\equiv c$, $i = 0, 1, \ldots$.

2. A sufficient condition for an indecomposable chain with states 0, 1, ... to be recurrent is that there is a sequence $(u_0, u_1, \ldots)$ with $u_i \to \infty$, $i \to \infty$, such that $u_j \geq \sum_i u_i p_{ij}$ for all $j \neq 0$.

3. A necessary and sufficient condition for an indecomposable chain with states 0, 1, ... to be recurrent and positive is that the system of equations $u_j = \sum_i u_i p_{ij}$, $j = 0, 1, \ldots$, has a solution, not identically zero, such that $\sum_i |u_i| < \infty$.

4. Consider a Markov chain with states 0, 1, ... and transition probabilities

$$p_{00} = r_0, \qquad p_{01} = p_0 > 0,$$

$$p_{ij} = \begin{cases} p_i > 0, & j = i + 1, \\ r_i \geq 0, & j = i, \\ q_i > 0, & j = i - 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\rho_0 = 1$, $\rho_m = (q_1 \ldots q_m)/(p_1 \ldots p_m)$. Prove the following propositions.

$$\text{Chain is recurrent} \Leftrightarrow \sum \rho_m = \infty,$$

$$\text{Chain is nonrecurrent} \Leftrightarrow \sum \rho_m < \infty,$$

$$\text{Chain is positive} \Leftrightarrow \sum \frac{1}{p_m \rho_m} < \infty,$$

$$\text{Chain is null} \Leftrightarrow \sum \rho_m = \infty, \sum \frac{1}{p_m \rho_m} = \infty.$$

5. Show that $f_{ik} \geq f_{ij} f_{jk}$ and $\sup_n p_{ij}^{(n)} \leq f_{ij} \leq \sum_{n=1}^{\infty} p_{ij}^{(n)}$.

6. Show that for every Markov chain with countably many states, the limit of $p_{ij}^{(n)}$ always exists in the *Cesàro sense*:

$$\lim_n \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} = \frac{f_{ij}}{\mu_j}.$$

7. Consider a Markov chain $\xi_0, \xi_1, \ldots$ with $\xi_{k+1} = (\xi_k^+) + \eta_{k+1}$, where $\eta_1, \eta_2, \ldots$ is a sequence of independent identically distributed random variables with $\mathsf{P}(\eta_k = j) = p_j$, $j = 0, 1, \ldots$. Write the transition matrix and show that if $p_0 > 0$, $p_0 + p_1 < 1$, the chain is recurrent if and only if $\sum_k k p_k \leq 1$.

§4. On the Existence of Limits and of Stationary Distributions

1. We begin with some necessary conditions for the existence of stationary distributions.

Theorem 1. *Let a Markov chain with countably many states $E = \{1, 2, \ldots\}$ and transition matrix $\mathbb{P} = \|p_{ij}\|$ be such that the limits*

$$\lim_n p_{ij}^{(n)} = \pi_j,$$

exist for all i and j and do not depend on i.

Then

(a) $\sum_i \pi_i \le 1$, $\sum_i \pi_i p_{ij} = \pi_j$;
(b) *either all* $\pi_j = 0$ *or* $\sum_j \pi_j = 1$;
(c) *if all* $\pi_j = 0$, *there is no stationary distribution*; *if* $\sum_i \pi_j = 1$, *then* $\Pi = (\pi_1, \pi_2, \ldots)$ *is the unique stationary distribution.*

PROOF. By Fatou's lemma,

$$\sum_j \pi_j = \sum_j \lim_n p_{ij}^{(n)} \le \underline{\lim}_n \sum_j p_{ij}^{(n)} = 1.$$

Moreover,

$$\sum_i \pi_i p_{ij} = \sum_i \left(\lim_n p_{ki}^{(n)}\right) p_{ij} \le \underline{\lim}_n \sum_i p_{ki}^{(n)} p_{ij} = \underline{\lim}_n p_{kj}^{(n+1)} = \pi_j,$$

that is, for each j,

$$\sum_i \pi_i p_{ij} \le \pi_j.$$

Suppose that

$$\sum_i \pi_i p_{ij_0} < \pi_{j_0}$$

for some j_0. Then

$$\sum_j \pi_j > \sum_j \left(\sum_i \pi_i p_{ij}\right) = \sum_i \pi_i \sum_j p_{ij} = \sum_i \pi_i.$$

This contradiction shows that

$$\sum_i \pi_i p_{ij} = \pi_j \tag{1}$$

for all j.

It follows from (1) that

$$\sum_i \pi_i p_{ij}^{(n)} = \pi_j.$$

Therefore

$$\pi_j = \lim_n \sum_i \pi_i p_{ij}^{(n)} = \sum_i \pi_i \lim_n p_{ij}^{(n)} = \left(\sum_i \pi_i\right)\pi_j,$$

that is, for all j,

$$\pi_j\left(1 - \sum_i \pi_i\right) = 0,$$

from which (b) follows.

Now let $\mathbb{Q} = (q_1, q_2, \ldots)$ be a stationary distribution. Since $\sum_i q_i p_{ij}^{(n)} = q_j$ and therefore $\sum_i q_i \pi_j = q_j$, that is, $\pi_j = q_j$ for all j, this stationary distribution must coincide with $\Pi = (\pi_1, \pi_2, \ldots)$. Therefore if all $\pi_j = 0$, there is no stationary distribution. If, however, $\sum_j \pi_j = 1$, then $\Pi = (\pi_1, \pi_2, \ldots)$ is the unique stationary distribution.

This completes the proof of the theorem.

Let us state and prove a fundamental result on the existence of a unique stationary distribution.

Theorem 2. *For Markov chains with countably many states, there is a unique stationary distribution if and only if the set of states contains precisely one positive recurrent class (of essential communicating states).*

Proof. Let N be the number of positive recurrent classes.

Suppose $N = 0$. Then all states are either nonrecurrent or are recurrent null states, and by (3.10) and (3.20), $\lim_n p_{ij}^{(n)} = 0$ for all i and j. Consequently, by Theorem 1, there is no stationary distribution.

Let $N = 1$ and let C be the unique positive recurrent class. If $d(C) = 1$ we have, by (3.8),

$$p_{ij}^{(n)} \to \frac{1}{\mu_j} > 0, \qquad i, j \in C.$$

If $j \notin C$, then j is nonrecurrent, and $p_{ij}^{(n)} \to 0$ for all i as $n \to \infty$, by (3.7).

Put

$$q_j = \begin{cases} \dfrac{1}{\mu_j} > 0, & j \in C, \\ 0, & j \notin C. \end{cases}$$

Then, by Theorem 1, the set $\mathbb{Q} = (q_1, q_2, \ldots)$ is the unique stationary distribution.

Now let $d = d(C) > 1$. Let $C_0, \ldots, C_{d-1}$ be the cyclic subclasses. With respect to $\mathbb{P}^d$, each subclass C_k is a recurrent aperiodic class. Then if i and $j \in C_k$ we have

$$p_{ij}^{(nd)} \to \frac{d}{\mu_j} > 0$$

by (3.19). Therefore on each set C_k, the set d/μ_j, $j \in C_k$, forms (with respect to $\mathbb{P}^d$) the unique stationary distribution. Hence it follows, in particular, that $\sum_{j \in C_k} (d/\mu_j) = 1$, that is, $\sum_{j \in C_k} (1/\mu_j) = 1/d$.

Let us put

$$q_j = \begin{cases} \dfrac{1}{\mu_j}, & j \in C = C_0 + \cdots + C_{d-1}, \\ 0, & j \notin C, \end{cases}$$

and show that, for the original chain, the set $\mathbb{Q} = (q_1, q_2, \ldots)$ is the unique stationary distribution.

In fact, for $i \in C$,

$$p_{ii}^{(nd)} = \sum_{j \in C} p_{ij}^{(nd-1)} p_{ji}.$$

Then by Fatou's lemma,

$$\frac{d}{\mu_i} = \lim_n p_{ii}^{(nd)} \geq \sum_{j \in C} \underline{\lim}_n \, p_{ij}^{(nd-1)} p_{ji} = \sum_{j \in C} \frac{1}{\mu_j} p_{ji}$$

and therefore

$$\frac{1}{\mu_i} \geq \sum_{j \in C} \frac{1}{\mu_j} p_{ji}.$$

But

$$\sum_{i \in C} \frac{1}{\mu_i} = \sum_{k=0}^{d-1} \left(\sum_{i \in C_k} \frac{1}{\mu_i} \right) = \sum_{k=0}^{d-1} \frac{1}{d} = 1.$$

As in Theorem 1, it can now be shown that in fact

$$\frac{1}{\mu_i} = \sum_{j \in C} \frac{1}{\mu_j} p_{ji}.$$

This shows that the set $\mathbb{Q} = (q_1, q_2, \ldots)$ is a stationary distribution, which is unique by Theorem 1.

Now let there be $N \geq 2$ positive recurrent classes. Denote them by $C^1, \ldots, C^N$, and let $\mathbb{Q}^i = (q_1^i, q_2^i, \ldots)$ be the stationary distribution corresponding to the class C^i and constructed according to the formula

$$q_j^i = \begin{cases} \dfrac{1}{\mu_j} > 0, & j \in C^i, \\ 0, & j \notin C^i. \end{cases}$$

Then, for all nonnegative numbers $a_1, \ldots, a_N$ such that $a_1 + \cdots + a_N = 1$, the set $a_1\mathbb{Q}^1 + \cdots + a_N\mathbb{Q}^N$ will also form a stationary distribution, since

$$(a_1\mathbb{Q}^1 + \cdots + a_N\mathbb{Q}^N)\mathbb{P} = a_1\mathbb{Q}^1\mathbb{P} + \cdots + a_N\mathbb{Q}^N\mathbb{P} = a_1\mathbb{Q}^1 + \cdots + a_N\mathbb{Q}^N.$$

Hence it follows that when $N \geq 2$ there is a continuum of stationary distributions. Therefore there is a unique stationary distribution only in the case $N = 1$.

This completes the proof of the theorem.

2. The following theorem answers the question of when there is a limit distribution for a Markov chain with a countable set of states E.

Theorem 3. *A necessary and sufficient condition for the existence of a limit distribution is that there is, in the set E of states of the chain, exactly one aperiodic positive recurrent class C such that $f_{ij} = 1$ for all $j \in C$ and $i \in E$.*

Proof. *Necessity.* Let $q_j = \lim p_{ij}^{(n)}$ and let $\mathbb{Q} = (q_1, q_2, \ldots)$ be a distribution ($q_i \geq 0$, $\sum_i q_i = 1$). Then by Theorem 1 this limit distribution is the unique stationary distribution, and therefore by Theorem 2 there is one and only one recurrent positive class C. Let us show that this class has period $d = 1$. Suppose the contrary, that is, let $d > 1$. Let $C_0, C_1, \ldots, C_{d-1}$ be the cyclic subclasses. If $i \in C_0$ and $j \in C_1$, then by (19), $p_{ij}^{(nd+1)} \to d/\mu_i$ and $p_{ij}^{(nd)} = 0$ for all n. But $d/\mu_j > 0$, and therefore $p_{ij}^{(n)}$ does not have a limit as $n \to \infty$; this contradicts the hypothesis that $\lim_n p_{ij}^{(n)}$ exists. Now let $j \in C$ and $i \in E$. Then, by (3.11), $p_{ij}^{(n)} \to f_{ij}/\mu_j$. Consequently $\pi_j = f_{ij}/\mu_j$. But π_j is independent of i. Therefore $f_{ij} = f_{ji} = 1$.

Sufficiency. By (3.11), (3.10) and (3.7),

$$p_{ij}^{(n)} \to \begin{cases} \dfrac{f_{ij}}{\mu_j}, & j \in C, \quad i \in E, \\ 0, & j \notin C, \quad i \in E. \end{cases}$$

Therefore if $f_{ij} = 1$ for all $j \in C$ and $i \in E$, then $q_j = \lim_n p_{ij}^{(n)}$ is independent of i. Class C is positive and therefore $q_j > 0$ for $j \in C$. Then, by Theorem 1, we have $\sum_j q_j = 1$ and the set $\mathbb{Q} = (q_1, q_2, \ldots)$ is a limit distribution.

3. Let us summarize the results obtained above on the existence of a limit distribution, the uniqueness of a stationary distribution and ergodicity, for the case of finite chains.

Theorem 4. *We have the following implications for finite Markov chains:*

$$\begin{array}{ccc}
(\textit{ergodicity}) & \overset{\{1\}}{\Leftrightarrow} & \begin{pmatrix}\textit{chain indecomposable,} \\ \textit{recurrent, positive,} \\ \textit{with } d = 1\end{pmatrix} \\
\Downarrow & & \Downarrow \\
\begin{pmatrix}\textit{limit distribution} \\ \textit{exists}\end{pmatrix} & \overset{\{2\}}{\Leftrightarrow} & \begin{pmatrix}\textit{there exists exactly one} \\ \textit{recurrent positive class} \\ \textit{with } d = 1\end{pmatrix} \\
\Downarrow & & \Downarrow \\
\begin{pmatrix}\textit{unique stationary} \\ \textit{distribution}\end{pmatrix} & \overset{\{3\}}{\Leftrightarrow} & \begin{pmatrix}\textit{there exists exactly one} \\ \textit{recurrent positive class}\end{pmatrix}
\end{array}$$

Proof. The "vertical" implications are evident. {1} is established in Theorem 2, §3; {2} in Theorem 3; {3} in Theorem 2.

4. Problems

1. Show that, in Example 1 of §5, neither stationary nor a limit distribution occurs.

2. Discuss the question of stationarity and limit distribution for the Markov chain with transition matrix

$$\mathbb{P} = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}.$$

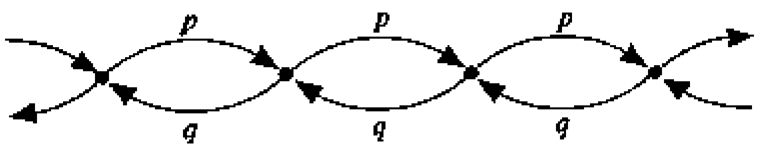
3. Let $\mathbb{P} = \|p_{ij}\|$ be a finite doubly stochastic matrix, that is, $\sum_{i=1}^{m} p_{ij} = 1, j = 1, \dots, m$. Show that the stationary distribution of the corresponding Markov chain is the vector $\mathbb{Q} = (1/m, \dots, 1/m)$.

§5. Examples

1. We present a number of examples to illustrate the concepts introduced above, and the results on the classification and limit behavior of transition probabilities.

Example 1. A *simple random walk* is a Markov chain such that a particle remains in each state with a certain probability, and goes to the next state with a certain probability.

The simple random walk corresponding to the graph



describes the motion of a particle among the states $E = \{0, \pm 1, \dots\}$ with transitions one unit to the right with probability p and to the left with probability q. It is clear that the transition probabilities are

$$p_{ij} = \begin{cases} p, & j = i + 1, \\ q, & j = i - 1, p + q = 1, \\ 0 & \text{otherwise}. \end{cases}$$

If $p = 0$, the particle moves deterministically to the left; if $p = 1$, to the right. These cases are of little interest since all states are inessential. We therefore assume that $0 < p < 1$.

With this assumption, the states of the chain form a single class (of essential communicating states). A particle can return to each state after $2, 4, 6, \dots$ steps. Hence the chain has period $d = 2$.

Since, for each $i \in E$,

$$p_{ii}^{(2n)} = C_{2n}^n (pq)^n = \frac{(2n)!}{(n!)^2}(pq)^n,$$

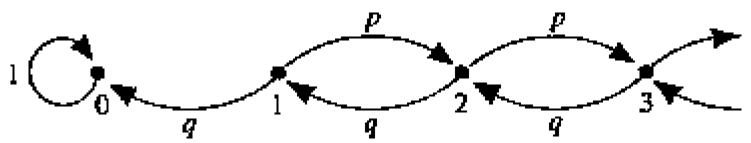
then by Stirling's formula (which says $n! \sim \sqrt{2\pi n}\, n^n e^{-n}$) we have

$$p_{ii}^{(2n)} \sim \frac{(4pq)^n}{\sqrt{\pi n}}.$$

Therefore $\sum_n p_{ii}^{(2n)} = \infty$ if $p = q$, and $\sum_n p_{ii}^{(2n)} < \infty$ if $p \neq q$. In other words, the chain is recurrent if $p = q$, but if $p \neq q$ it is nonrecurrent. It was shown in §10, Chapter I, that $f_{ii}^{(2n)} \sim 1/(2\sqrt{\pi}n^{3/2})$, $n \to \infty$, if $p = q = \frac{1}{2}$. Therefore $\mu_i = \sum_n (2n) f_{ii}^{(2n)} = \infty$, that is, all recurrent states are null states. Hence by Theorem 1 of §3, $p_{ij}^{(n)} \to 0$ as $n \to \infty$ for all i and j.

There are no stationary, limit, or ergodic distributions.

EXAMPLE 2. Consider a simple random walk with $E = \{0, 1, 2, \ldots\}$, where 0 is an absorbing barrier:



$0 < p < 1$

State 0 forms a unique positive recurrent class with $d = 1$. All other states are nonrecurrent. Therefore, by Theorem 2 of §4, there is a unique stationary distribution

$$\Pi = (\pi_0, \pi_1, \pi_2, \ldots)$$

$$\text{with } \pi_0 = 1 \text{ and } \pi_i = 0,\ i \geq 1.$$

Let us now consider the question of limit distributions. Clearly $p_{00}^{(n)} = 1$, $p_{ij}^{(n)} \to 0$, $j \geq 1$, $i \geq 0$. Let us now show that for $i \geq 1$ the numbers $\alpha(i) = \lim_n p_{i0}^{(n)}$ are given by the formulas

$$\alpha(i) = \begin{cases} \left(\dfrac{q}{p}\right)^i, & p > q, \\ 1, & p \leq q. \end{cases} \tag{1}$$

We begin by observing that since state 0 is absorbing we have $p_{i0}^{(n)} = \sum_{k \leq n} f_{i0}^{(k)}$ and consequently $\alpha(i) = f_{i0}$, that is, the probability $\alpha(i)$ is the probability that a particle starting from state i sooner or later reaches the null state. By the method of §12, Chapter I (see also §2 of Chapter VII) we can obtain the recursion relation

$$\alpha(i) = p\alpha(i + 1) + q\alpha(i - 1), \tag{2}$$

with $\alpha(0) = 1$. The general solution of this equation has the form

$$\alpha(i) = a + b(q/p)^i, \tag{3}$$

and the condition $\alpha(0) = 1$ imposes the condition $a + b = 1$.

If we suppose that $q > p$, then since $\alpha(i)$ is bounded we see at once that $b = 0$, and therefore $\alpha(i) = 1$. This is quite natural, since when $q > p$ the particle tends to move toward the null state.

If, on the other hand, $p > q$ the opposite is true: the particle tends to move to the right, and so it is natural to expect that

$$\alpha(i) \to 0, \qquad i \to \infty, \tag{4}$$

and consequently $a = 0$ and

$$\alpha(i) = \left(\frac{q}{p}\right)^i. \tag{5}$$

To establish this equation, we shall not start from (4), but proceed differently.

In addition to the absorbing barrier at 0 we introduce an absorbing barrier at the integral point N. Let us denote by $\alpha_N(i)$ the probability that a particle that starts at i reaches the zero state before reaching N. Then $\alpha_N(i)$ satisfies (2) with the boundary conditions

$$\alpha_N(0) = 1, \qquad \alpha_N(N) = 0,$$

and, as we have already shown in §9, Chapter I,

$$\alpha_N(i) = \frac{\left(\frac{q}{p}\right)^i - \left(\frac{q}{p}\right)^N}{1 - \left(\frac{q}{p}\right)^N}, \qquad 0 \le i \le N. \tag{6}$$

Hence

$$\lim_N \alpha_N(i) = \left(\frac{q}{p}\right)^i$$

and consequently to prove (5) we have only to show that

$$\alpha(i) = \lim_N \alpha_N(i). \tag{7}$$

This is intuitively clear. A formal proof can be given as follows.

Let us suppose that the particle starts from a given state i. Then

$$\alpha(i) = \mathsf{P}_i(A), \tag{8}$$

where A is the event in which there is an N such that a particle starting from i reaches the zero state before reaching state N. If

$$A_N = \{\text{particle reaches 0 before } N\},$$

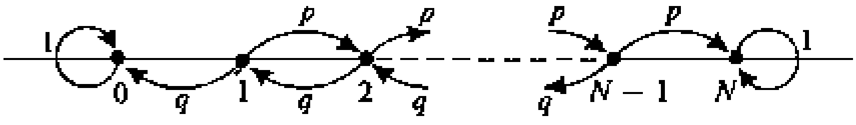
then $A = \bigcup_{N=i+1}^{\infty} A_N$. It is clear that $A_N \subseteq A_{N+1}$ and

$$\mathsf{P}_i\left(\bigcup_{N=i+1}^{\infty} A_N\right) = \lim_{N\to\infty} \mathsf{P}_i(A_N). \tag{9}$$

But $\alpha_N(i) = \mathsf{P}_i(A_N)$, so that (7) follows directly from (8) and (9).

Thus if $p > q$ the limit $\lim p_{i0}^{(1)}$ depends on i, and consequently there is no limit distribution in this case. If, however, $p \le q$, then in all cases $\lim p_{i0}^{(n)} = 1$ and $\lim p_{ij}^{(n)} = 0$, $j \ge 1$. Therefore in this case the limit distribution has the form $\Pi = (1, 0, 0, \ldots)$.

EXAMPLE 3. Consider a simple random walk with absorbing barriers at 0 and N:

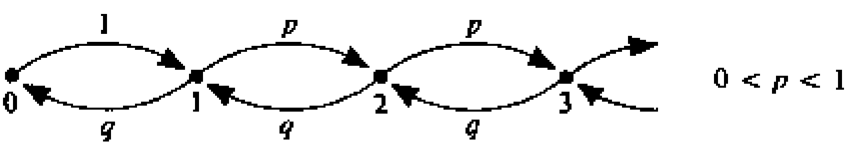


Here there are two positive recurrent classes $\{0\}$ and $\{N\}$. All other states $\{1, \ldots, N-1\}$ are nonrecurrent. It follows from Theorem 1, §3, that there are infinitely many stationary distributions $\Pi = (\pi_0, \pi_1, \ldots, \pi_N)$ with $\pi_0 = a$, $\pi_N = b$, $\pi_1 = \cdots = \pi_{N-1} = 0$, where $a \ge 0$, $b \ge 0$, $a + b = 1$. From Theorem 4, §4 it also follows that there is no limit distribution. This is also a consequence of the equations (Subsection 2, §9, Chapter I)

$$\lim_{n\to\infty} p_{i0}^{(n)} = \begin{cases} \dfrac{\left(\dfrac{q}{p}\right)^i - \left(\dfrac{q}{p}\right)^N}{1 - \left(\dfrac{q}{p}\right)^N}, & p \ne q, \\[2ex] 1 - \dfrac{i}{N}, & p = q, \end{cases} \tag{10}$$

$$\lim_n p_{iN}^{(n)} = 1 - \lim_n p_{i0}^{(n)} \quad \text{and} \quad \lim_n p_{ij}^{(n)} = 0,\ 1 \le j \le N-1.$$

EXAMPLE 4. Consider a simple random walk with $E = \{0, 1, \ldots\}$ and a *reflecting barrier at* 0:



It is easy to see that the chain is periodic with period $d = 2$. Suppose that $p > q$ (the moving particle tends to move to the right). Let $i > 1$; to determine the probability f_{i1} we may use formula (1), from which it follows that

$$f_{i1} = \left(\frac{q}{p}\right)^{i-1} < 1, \qquad i > 1.$$

All states of this chain communicate with each other. Therefore if state i is recurrent, state 1 will also be recurrent. But (see the proof of Lemma 3 in §3) in that case f_{i1} must be 1. Consequently when $p > q$ all the states of the chain are nonrecurrent. Therefore $p_{ij}^{(n)} \to 0$, $n \to \infty$ for i and $j \in E$, and there is neither a limit distribution nor a stationary distribution.

Now let $p \le q$. Then, by (1), $f_{i1} = 1$ for $i > 1$ and $f_{11} = q + pf_{21} = 1$. Hence the chain is recurrent.

Consider the system of equations determining the stationary distribution $\Pi = (\pi_0, \pi_1, \dots)$:

$$\begin{aligned} \pi_0 &= \pi_1 q, \\ \pi_1 &= \pi_0 + \pi_2 q, \\ \pi_2 &= \pi_1 p + \pi_3 q, \end{aligned}$$

that is,

$$\begin{aligned} \pi_1 &= \pi_1 q + \pi_2 q, \\ \pi_2 &= \pi_2 q + \pi_3 q, \\ &\dots\dots\dots \end{aligned}$$

whence

$$\pi_j = \left(\frac{p}{q}\right)\pi_{j-1}, \qquad j = 2, 3, \dots.$$

If $p = q$ we have $\pi_1 = \pi_2 = \dots$, and consequently

$$\pi_0 = \pi_1 = \pi_2 = \cdots = 0.$$

In other words, if $p = q$, there is no stationary distribution, and therefore no limit distribution. From this and Theorem 3, §4, it follows, in particular, that in this case all states of the chain are null states.

It remains to consider the case $p < q$. From the condition $\sum_{j=0}^{\infty} \pi_j = 1$ we find that

$$\pi_1\left[q + 1 + \left(\frac{p}{q}\right) + \left(\frac{p}{q}\right)^2 + \cdots\right] = 1,$$

that is

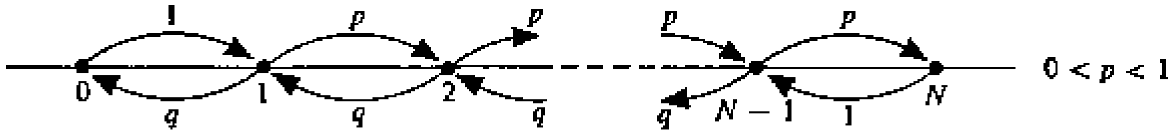
$$\pi_1 = \frac{q - p}{2q}$$

and

$$\pi_j = \frac{q-p}{2q} \cdot \left(\frac{p}{q}\right)^{j-1}; \qquad j \geq 2.$$

Therefore the distribution Π is the unique stationary distribution. Hence when $p < q$ the chain is recurrent and positive (Theorem 2, §4). The distribution Π is also a limit distribution and is ergodic.

EXAMPLE 5. Again consider a simple random walk with *reflecting barriers at* 0 *and* N:



All the states of the chain are periodic with period $d = 2$, recurrent, and positive. According to Theorem 4 of §4, the chain is ergodic. Solving the system $\pi_j = \sum_{i=0}^{N} \pi_i p_{ij}$ subject to $\sum_{j=0}^{N} \pi_j = 1$, we obtain the ergodic distribution

$$\pi_i = \frac{\left(\frac{p}{q}\right)^{i-1}}{1 + \sum_{j=1}^{N-1} \left(\frac{p}{q}\right)^{j-1}}, \qquad 2 \leq j \leq N-1,$$

and

$$\pi_0 = \pi_1 q, \qquad \pi_N = \pi_{N-1} p.$$

2. EXAMPLE 6. It follows from Example 1 that the simple random walk considered there on the integral points of the line is recurrent if $p = q$, but nonrecurrent if $p \neq q$. Now let us consider simple random walks in the plane and in space, from the point of view of recurrence or nonrecurrence.

For the plane, we suppose that a particle in any state (i, j) moves up, down, to the right or to the left with probability $\frac{1}{4}$ (Figure 41).

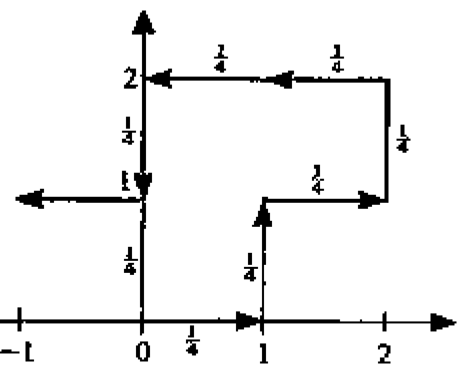


Figure 41. A walk in the plane.

For definiteness, consider the state (0, 0). Then the probability $P_k = p^{(k)}_{(0,0),(0,0)}$ of going from (0, 0) to (0, 0) in k steps is given by

$$P_{2n+1} = 0, \qquad n = 0, 1, 2, \ldots,$$
$$P_{2n} = \sum_{\{(i,j):\, i+j=n,\, 0 \le i \le n\}} \frac{(2n)!}{i!\,i!\,j!\,j!} (\tfrac{1}{4})^{2n}, \quad n = 1, 2, \ldots.$$

Multiplying numerators and denominators by $(n!)^2$, we obtain

$$P_{2n} = (\tfrac{1}{4})^{2n} C_{2n}^n \sum_{i=0}^{n} C_n^i C_n^{n-i} = (\tfrac{1}{4})^{2n} (C_{2n}^n)^2,$$

since

$$\sum_{i=0}^{n} C_n^i C_n^{n-i} = C_{2n}^n.$$

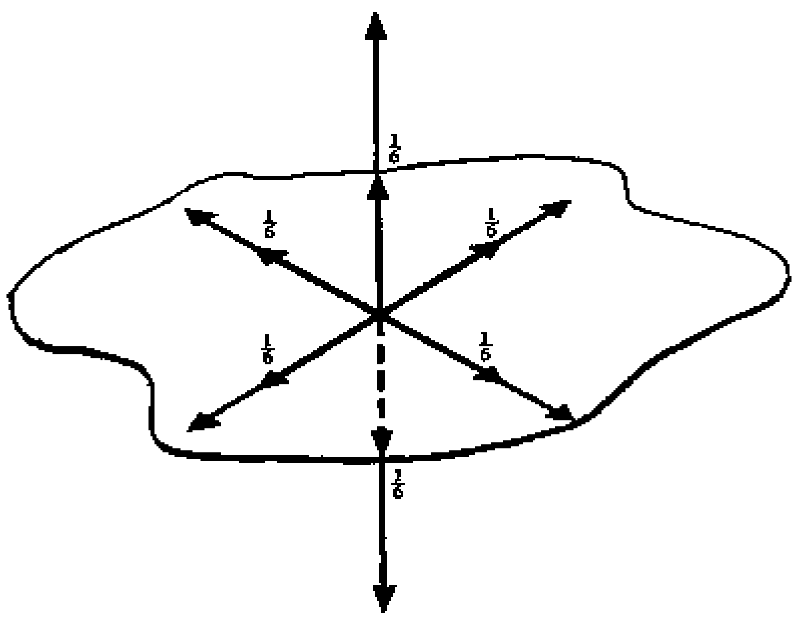
Applying Stirling's formula, we find that

$$P_{2n} \sim \frac{1}{\pi n},$$

and therefore $\sum P_{2n} = \infty$. Consequently the state (0, 0) (likewise any other) is *recurrent*.

It turns out, however, that in *three or more dimensions* the symmetric random walk is *nonrecurrent*. Let us prove this for walks on the integral points (i, j, k) in space.

Let us suppose that a particle moves from (i, j, k) by one unit along a coordinate direction, with probability $\frac{1}{6}$ for each.



Then if P_k is the probability of going from (0, 0, 0) to (0, 0, 0) in k steps, we have

$$P_{2n+1} = 0, \qquad n = 0, 1, \ldots,$$

$$\begin{aligned} P_{2n} &= \sum_{\{(i,j):0\le i+j\le n,\, 0\le i\le n,\, 0\le j\le n\}} \frac{(2n)!}{(i!)^2\,(j!)^2((n-i-j)!)^2}\left(\tfrac{1}{6}\right)^{2n} \\ &= \frac{1}{2^{2n}} C_{2n}^n \sum_{\{(i,j):0\le i+j\le n,\, 0\le i\le n,\, 0\le j\le n\}} \left[\frac{n!}{i!\,j!\,(n-i-j)!}\right]^2 \left(\tfrac{1}{3}\right)^{2n} \\ &\le C_n \frac{1}{2^{2n}} C_{2n}^n \frac{1}{3^n} \sum_{\{(i,j):0\le i+j\le n,\, 0\le i\le n,\, 0\le j\le n\}} \frac{n!}{i!\,j!\,(n-i-j)!}\left(\tfrac{1}{3}\right)^n \\ &= C_n \frac{1}{2^{2n}} C_{2n}^n \frac{1}{3^n}, \qquad n = 1, 2, \ldots, \end{aligned} \tag{11}$$

where

$$C_n = \max_{\{(i,j):0\le i+j\le n,\, 0\le i\le n,\, 0\le j\le n\}} \left[\frac{n!}{i!\,j!\,(n-i-j)!}\right]. \tag{12}$$

Let us show that when n is large, the max in (12) is attained for $i \sim n/3$, $j \sim n/3$. Let i_0 and j_0 be the values at which the max is attained. Then the following inequalities are evident:

$$\frac{n!}{j_0!\,(i_0-1)!\,(n-j_0-i_0+1)!} \le \frac{n!}{j_0!\,i_0!\,(n-j_0-i_0)!},$$

$$\begin{aligned} \frac{n!}{j_0!\,(i_0+1)!\,(n-j_0-i_0-1)!} &\le \frac{n!}{(j_0-1)!\,i_0!\,(n-j_0-i_0+1)!} \\ &\le \frac{n!}{(j_0+1)!\,i_0!\,(n-j_0-i_0-1)!}, \end{aligned}$$

whence

$$n - i_0 - 1 \le 2j_0 \le n - i_0 + 1,$$
$$n - j_0 - 1 \le 2i_0 \le n - j_0 + 1,$$

and therefore we have, for large n, $i_0 \sim n/3$, $j_0 \sim n/3$, and

$$C_n \sim \frac{n!}{\left[\left(\frac{n}{3}\right)!\right]^3}.$$

By Stirling's formula,

$$C_n \frac{1}{2^{2n}} C_{2n}^n \frac{1}{3^n} \sim \frac{3\sqrt{3}}{2\pi^{3/2} n^{3/2}},$$

and since

$$\sum_{n=1}^{\infty} \frac{3\sqrt{3}}{2\pi^{3/2} n^{3/2}} < \infty,$$

we have $\sum_n P_{2n} < \infty$. Consequently the state (0, 0, 0), and likewise any other state, is nonrecurrent. A similar result holds for dimensions greater than 3.

Thus we have the following result (Pólya):

Theorem. *For R^1 and R^2, the symmetric random walk is recurrent; for R^n, $n \geq 3$, it is nonrecurrent.*

3. Problems

1. Derive the recursion relation (1).
2. Establish (4).
3. Show that in Example 5 all states are aperiodic, recurrent, and positive.
4. Classify the states of a Markov chain with transition matrix

$$\mathbb{P} = \begin{pmatrix} p & q & 0 & 0 \\ 0 & 0 & p & q \\ p & q & 0 & 0 \\ 0 & 0 & p & q \end{pmatrix},$$

where $p + q = 1$, $p \geq 0$, $q \geq 0$.

Historical and Bibliographical Notes

Introduction

The history of probability theory up to the time of Laplace is described by Todhunter [T1]. The period from Laplace to the end of the nineteenth century is covered by Gnedenko and Sheinin in [K10]. Maistrov [M1] discusses the history of probability theory from the beginning to the thirties of the present century. There is a brief survey in Gnedenko [G4]. For the origin of much of the terminology of the subject see Aleksandrova [A3].

For the basic concepts see Kolmogorov [K8], Gnedenko [G4], Borovkov [B4], Gnedenko and Khinchin [G6], A. M. and I. M. Yaglom [Y1], Prokhorov and Rozanov [P5], Feller [F1, F2], Neyman [N3], Loève [L7], and Doob [D3]. We also mention [M3] which contains a large number of problems on probability theory.

In putting this text together, the author has consulted a wide range of sources. We mention particularly the books by Breiman [B5], Ash [A4, A5], and Ash and Gardner [A6], which (in the author's opinion) contain an excellent selection and presentation of material.

For current work in the field see, for example, *Annals of Probability* (formerly *Annals of Mathematical Statistics*) and *Theory of Probability and its Applications* (translation of *Teoriya Veroyatnostei i ee Primeneniya*).

Mathematical Reviews and *Zentralblatt für Mathematik* contain abstracts of current papers on probability and mathematical statistics from all over the world.

For tables for use in computations, sec [A1].

Chapter I

§1. Concerning the construction of probabilistic models see Kolmogorov [K7] and Gnedenko [G4]. For further material on problems of distributing objects among boxes see, e.g., Kolchin, Sevastyanov and Chistyakov [K3].

§2. For other probabilistic models (in particular, the one-dimensional Ising model) that are used in statistical physics, see Isihara [I2].

§3. Bayes's formula and theorem form the basis for the "Bayesian approach" to mathematical statistics. See, for example, De Groot [D1] and Zacks [Z1].

§4. A variety of problems about random variables and their probabilistic description can be found in Meshalkin [M3].

§5. A combinatorial proof of the law of large numbers (originating with James Bernoulli) is given in, for example, Feller [F1]. For the empirical meaning of the law of large numbers see Kolmogorov [K7].

§6. For sharper forms of the local and integrated theorems, and of Poisson's theorem, see Borovkov [B4] and Prokhorov [P3].

§7. The examples of Bernoulli schemes illustrate some of the basic concepts and methods of mathematical statistics. For more detailed discussions see, for example, Cramér [C5] and van der Waerden [W1].

§8. Conditional probability and conditional expectation with respect to a partition will help the reader understand the concepts of conditional probability and conditional expectation with respect to σ-algebras, which will be introduced later.

§9. The ruin problem was considered in essentially the present form by Laplace. See Gnedenko and Sheinin [K10]. Feller [F1] contains extensive material from the same circle of ideas.

§10. Our presentation essentially follows Feller [F1]. The method for proving (10) and (11) is taken from Doherty [D2].

§11. Martingale theory is throughly covered in Doob [D3]. A different proof of the ballot theorem is given, for instance, in Feller [F1].

§12. There is extensive material on Markov chains in the books by Feller [F1], Dynkin [D4], Kemeny and Snell [K2], Sarymsakov [S1], and Sirazhdinov [S8]. The theory of branching processes is discussed by Sevastyanov [S3].

Chapter II

§1. Kolmogorov's axioms are presented in his book [K8].

§2. Further material on algebras and σ-algebras can be found in, for example, Kolmogorov and Fomin [K8], Neveu [N1], Breiman [B5], and Ash [A5].

§3. For a proof of Carathéodory's theorem see Loève [I7] or Halmos [H1].

§§4–5. More material on measurable functions is available in Halmos [H1].

§6. See also Kolmogorov and Fomin [K8], Halmos [H1], and Ash and Gardner [A6]. The Radon–Nikodým theorem is proved in these books. The inequality

$$\mathsf{P}(|\xi| \geq \varepsilon) \leq \frac{\mathsf{E}\xi^2}{\varepsilon^2}$$

is sometimes called Chebyshev's inequality, and the inequality

$$\mathsf{P}(|\xi| \geq \varepsilon) \leq \frac{\mathsf{E}|\xi|^r}{\varepsilon^r}, \qquad r > 0,$$

is called *Markov's inequality.*

For Pratt's lemma see [P2].

§7. The definitions of conditional probability and conditional expectation with respect to a σ-algebra were given by Kolmogorov [K8]. For additional material see Breiman [B5] and Ash [A5]. The result quoted in the Corollary to Theorem 5 can be found in [M5].

§8. See also Borovkov [B4], Ash [A5], Cramér [C5], and Gnedenko [G4].

§9. Kolmogorov's theorem on the existence of a process with given finite-dimensional distribution is in his book [K8]. For Ionescu–Tulcea's theorem see also Neveu [N1] and Ash [A5]. The proof in the text follows [A5].

§§10–11. See also Kolmogorov and Fomin [K9], Ash [A5], Doob [D3], and Loève [L7].

§12. The theory of characteristic functions is presented in many books. See, for example, Gnedenko [G4], Gnedenko and Kolmogorov [G5], Ramachandran [R1], Lukacs [L8], and Lukacs and Laha [L9]. Our presentation of the connection between moments and semi-invariants follows Leonov and Shiryaev [L4].

§13. See also Ibragimov and Rozanov [I1], Breiman [B5], and Liptser and Shiryayev [L5].

Chapter III

§1. Detailed investigations of problems on weak convergence of probability measures are given in Gnedenko and Kolmogorov [G5] and Billingsley [B3].

§2. Prokhorov's theorem appears in his paper [P4].

§3. The monograph [G5] by Gnedenko and Kolmogorov studies the limit theorems of probability theory by the method of characteristic functions. See also Billingsley [B3]. Problem 2 includes both Bernoulli's law of large numbers and Poisson's law of large numbers (which assumes that $\xi_1, \xi_2, \ldots$ are independent and take only two values (1 and 0), but in general are differently distributed: $\mathsf{P}(\xi_i = 1) = p_i$, $\mathsf{P}(\xi_i = 0) = 1 - p_i$, $i \geq 1$).

§4. Here we give the standard proof of the central limit theorem for sums of independent random variables under the Lindeberg condition. Compare [G5] and [P6].

In the first edition, we gave the proof of Theorem 3 here.

§5. Questions of the validity of the central limit theorem without the hypothesis of negligibility in the limit have already attracted the attention of P. Lévy. A detailed account of the current state of the theory of limit theorems in the *nonclassical setting* is contained in Zolotarev [Z4]. The statement and proof of Theorem 1 were given by Rotar [R5].

§6. The presentation uses material from Gnedenko and Kolmogorov [G5], Ash [A5], and Petrov [P1], [P6].

§7. The Lévy–Prokhorov metric was introduced in a well-known work by Prokhorov [P4], to whom the results on metrizability of weak convergence of measures given on metric spaces are also due. Concerning the metric $\|P - \tilde{P}\|^*_{BL}$, see Dudley [D6] and Pollard [P7].

§8. Theorem 1 is due to Skorokhod. Useful material on the method of a single probability space may be found in Borovkov [B4] and in Pollard [P7].

§§9–10. A number of books contain a great deal of material touching on these questions: Jacod and Shiryaev [J1], LeCam [L10], Greenwood and Shiryaev [G7], and Liese and Vajda [L11].

§11. Petrov [P6] contains a lot of material on estimates of the rate of convergence in the central limit theorem. The proof given of the theorem of Berry and Esseen is contained in Gnedenko and Kolmogorov [G5].

§12. The proof follows Presman [P8].

Chapter IV

§1. Kolmogorov's zero-or-one law appears in his book [K8]. For the Hewitt–Savage zero-or-one law see also Borovkov [B4], Breiman [B5], and Ash [A5].

§§2–4. Here the fundamental results were obtained by Kolmogorov and Khinchin (see [K8] and the references given there). See also Petrov [P1] and Stout [S9]. For probabilistic methods in number theory see Kubilius [K11].

§5. Regarding these questions, see Petrov [P6], Borovkov [B4], and Dacunha-Castelle and Duflo [D5].

Chapter V

§§1–3. Our exposition of the theory of (strict sense) stationary random processes is based on Breiman [B5], Sinai [S7], and Lamperti [L2]. The simple proof of the maximal ergodic theorem was given by Garsia [G1].

Chapter VI

§1. The books by Rozanov [R4], and Gihman and Skorohod [G2, G3] are devoted to the theory of (wide sense) stationary random processes. Example 6 was frequently presented in Kolmogorov's lectures.

§2. For orthogonal stochastic measures and stochastic integrals see also Doob [D3], Gihman and Skorohod [G3], Rozanov [R4], and Ash and Gardner [A6].

§3. The spectral representation (2) was obtained by Cramér and Loève (see, for example, [I7]). The same representation (in different language) is contained in Kolmogorov [K5]. Also see Doob [D3], Rozanov [R4], and Ash and Gardner [A6].

§4. There is a detailed exposition of problems of statistical estimation of the covariance function and spectral density in Hannan [H2, H3].

§§5–6. See also Rozanov [R4], Lamperti [L2], and Gihman and Skorohod [G2, G3].

§7. The presentation follows Lipster and Shiryaev [L5].

Chapter VII

§1. Most of the fundamental results of the theory of martingales were obtained by Doob [D3]. Theorem 1 is taken from Meyer [M4]. Also see Meyer and Dellacherie [M5], Lipster and Shiryaev [15], and Gihman and Skorohod [G3].

§2. Theorem 1 is often called the theorem "on transformation under a system of optional stopping" (Doob [D3]). For the identities (14) and (15) and Wald's fundamental identity see Wald [W2].

§3. Chow and Teicher [C3] contains an illuminating study of the results presented here, including proofs of the inequalities of Khinchin, Marcinkiewicz and Zygmund, Burkholder, and Davis. Theorem 2 was given by Lenglart [L3].

§4. See Doob [D3].

§5. Here we follow Kabanov, Liptser and Shiryaev [K1], Engelbert and Shiryaev [E1], and Neveu [N2]. Theorem 4 and the example were given by Liptser.

§6. This approach to problems of absolute continuity and singularity, and the results given here, can be found in Kabanov, Liptser and Shiryaev [K1]. Theorem 6 was obtained by Kabanov.

§7. Theorems 1 and 2 were given by Novikov [N4]. Lemma 1 is a discrete analog of Girsanov's lemma (see [K1]).

§8. See also Liptser and Shiryaev [L12] and Jacod and Shiryaev [J1], which discuss limit theorems for random processes of a rather general nature (for example, martingales, semi-martingales).

Chapter VIII

§1. For the basic definitions see Dynkin [D4], Ventzel [V2], Doob [D3], and Gihman and Skorohod [G3]. The existence of regular transition probabilities such that the Kolmogorov–Chapman equation (9) is satisfied for all $x \in R$ is proved in [N1] (corollary to Proposition V.2.1) and in [G3] (Volume I, Chapter II, §4). Kuznetsov (see Abstracts of the Twelfth European Meeting of Statisticians, Varna, 1979) has established the validity (which is far from trivial) of a similar result for Markov processes with continuous times and values in universal measurable spaces.

§§2–5. Here the presentation follows Kolmogorov [K4], Borovkov [B4], and Ash [A4].

References†

[A1] M. Abramovitz and I. A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables.* National Bureau of Standards, Washington, D.C., 1964.

[A2] P. S. Aleksandrov. *Einführung in die Mengenlehre und die Theorie der reellen Funktionen.* DVW, Berlin, 1956.

[A3] N. V. Aleksandrova. *Mathematical Terms* [*Matematicheskie terminy*]. Vysshaia Shkola, Moscow, 1978.

[A4] R. B. Ash. *Basic Probability Theory.* Wiley, New York, 1970.

[A5] R. B. Ash. *Real Analysis and Probability.* Academic Press, New York, 1972.

[A6] R. B. Ash and M. F. Gardner. *Topics in Stochastic Processes.* Academic Press, New York, 1975.

[B1] S. N. Bernshtein. Chebyshev's work on the theory of probability (in Russian), in *The Scientific Legacy of P. L. Chebyshev* [*Nauchnoe nasledie P. L. Chebysheva*], pp. 43–68, Akademiya Nauk SSSR, Moscow–Leningrad, 1945.

[B2] S. N. Bernshtein. *Theory of Probability* [*Teoriya veroyatnosteĭ*], 4th ed. Gostehizdat, Moscow, 1946.

[B3] P. Billingsley. *Convergence of Probability Measures.* Wiley, New York, 1968.

[B4] A. A. Borovkov. *Wahrscheinlichkeitstheorie: eine Einführung,* first edition Birkhäuser, Basel–Stuttgart, 1976; *Theory of Probability,* second edition [*Teoriya veroyatnosteĭ*]. "Nauka," Moscow, 1986.

[B5] L. Breiman. *Probability.* Addison-Wesley, Reading, MA, 1968.

[C1] P. L. Chebyshev. *Theory of Probability: Lectures Given in 1879 and 1880* [*Teoriya veroyatnosteĭ: Lektsii akad. P. L. Chebysheva chitannye v 1879, 1880 gg.*]. Edited by A. N. Krylov from notes by A. N. Lyapunov. Moscow–Leningrad, 1936.

[C2] Y. S. Chow, H. Robbins, and D. Siegmund. *The Theory of Optimal Stopping.* Dover, New York, 1991.

† Translator's note: References to translations into Russian have been replaced by references to the original works. Russian references have been replaced by their translations whenever I could locate translations; otherwise they are reproduced (with translated titles). Names of journals are abbreviated according to the forms used in *Mathematical Reviews* (1982 versions).

[C3] Y. S. Chow and H. Teicher. *Probability Theory: Independence, Interchangability, Martingales.* Springer-Verlag, New York, 1978.
[C4] Kai-Lai Chung. *Markov Chains with Stationary Transition Probabilities.* Springer-Verlag, New York, 1967.
[C5] H. Cramér, *Mathematical Methods of Statistics.* Princeton University Press, Princeton, NJ, 1957.
[D1] M. H. De Groot. *Optimal Statistical Decisions.* McGraw-Hill, New York, 1970.
[D2] M. Doherty. An amusing proof in fluctuation theory. *Lecture Notes in Mathematics*, no. 452, 101–104, Springer-Verlag, Berlin, 1975.
[D3] J. L. Doob. *Stochastic Processes.* Wiley, New York, 1953.
[D4] E. B. Dynkin. *Markov Processes.* Plenum, New York, 1963.
[E1] H. J. Engelbert and A. N. Shiryaev. On the sets of convergence and generalized submartingales. *Stochastics* 2 (1979), 155–166.
[F1] W. Feller. *An Introduction to Probability Theory and Its Applications*, vol. 1, 3rd ed. Wiley, New York, 1968.
[F2] W. Feller. *An Introduction to Probability Theory and Its Applications*, vol. 2, 2nd ed. Wiley, New York, 1966.
[G1] A. Garcia. A simple proof of E. Hopf's maximal ergodic theorem. *J. Math. Mech.* 14 (1965), 381–382.
[G2] I. I. Gihman [Gikhman] and A. V. Skorohod [Skorokhod]. *Introduction to the theory of random processes*, first edition. Saunders, Philadelphia, 1969; second edition [*Vvedenie v teoriyu sluchaĭnyh protsessov*]. "Nauka", Moscow, 1977.
[G3] I. I. Gihman and A. V. Skorohod. *Theory of Stochastic Processes*, 3 vols. Springer-Verlag, New York–Berlin, 1974–1979.
[G4] B. V. Gnedenko. *The theory of probability.* Mir, Moscow, 1988.
[G5] B. V. Gnedenko and A. N. Kolmogorov. *Limit Distributions for Sums of Independent Random Variables*, revised edition. Addison-Wesley, Reading, 1968.
[G6] B. V. Gnedenko and A. Ya. Khinchin. *An Elementary Introduction to the Theory of Probability.* Freeman, San Francisco, 1961; ninth edition [*Elementarnoe vvedenie v teoriyu veroyatnosteĭ*]. "Nauka", Moscow, 1982.
[H1] P. R. Halmos. *Measure Theory.* Van Nostrand, New York, 1950.
[H2] E. J. Hannan. *Time Series Analysis.* Methuen, London, 1960.
[H3] E. J. Hannan. *Multiple Time Series.* New York, Wiley, 1970.
[I1] I. A. Ibragimov and Yu. V. Linnik. *Independent and Stationary Sequences of Random Variables.* Walters-Noordhoff, Groningen, 1971.
[I2] I. A. Ibragimov and Yu. A. Rozanov. *Gaussian Random Processes.* Springer-Verlag, New York, 1978.
[I3] A. Isihara. *Statistical Physics.* Academic Press, New York, 1971.
[K1] Yu. M. Kabanov, R. Sh. Liptser, and A. N. Shiryaev. On the question of the absolute continuity and singularity of probability measures. *Math. USSR-Sb.* 33 (1977), 203–221.
[K2] J. Kemeny and L. J. Snell. *Finite Markov Chains.* Van Nostrand, Princeton, 1960.
[K3] V. F. Kolchin, B. A. Sevastyanov, and V. P. Chistyakov. *Random Allocations.* Halsted, New York, 1978.
[K4] A. N. Kolmogorov, Markov chains with countably many states. *Byull. Moskov. Univ.* 1 (1937), 1–16 (in Russian).
[K5] A. N. Kolmogorov. Stationary sequences in Hilbert space. *Byull. Moskov. Univ. Mat.* 2 (1941), 1–40 (in Russian).
[K6] A. N. Kolmogorov, The contribution of Russian science to the development

of probability theory. *Uchen. Zap. Moskov. Univ.* 1947, no. 91, 56ff. (in Russian).

[K7] A. N. Kolmogorov, Probability theory (in Russian), in *Mathematics: Its Contents, Methods, and Value* [*Matematika, ee soderzhanie, metody i znachenie*]. Akad. Nauk SSSR, vol. 2, 1956.

[K8] A. N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea, New York, 1956; second edition [*Osnovnye poniatiya teorii veroyatnosteĭ*]. "Nauka", Moscow, 1974.

[K9] A. N. Kolmogorov and S. V. Fomin. *Elements of the Theory of Functions and Functionals Analysis*. Graylok, Rochester, 1957 (vol. 1), 1961 (vol. 2); sixth edition [*Elementy teorii funktsiĭ i funktsional'nogo analiza*]. "Nauka", Moscow, 1989.

[K10] A. N. Kolmogorov and A. P. Yushkevich, editors. *Mathematics of the Nineteenth Century* [*Matematika XIX veka*]. Nauka, Moscow, 1978.

[K11] J. Kubilius. *Probabilistic Methods in the Theory of Numbers*. American Mathematical Society, Providence, 1964.

[L1] J. Lamperti. *Probability*. Benjamin, New York, 1966.

[L2] J. Lamperti. *Stochastic Processes*. Springer-Verlag, New York, 1977.

[L3] E. Lenglart. Relation de domination entre deux processus. *Ann. Inst. H. Poincaré*. Sect. B (N.S.) 13 (1977), 171–179.

[L4] V. P. Leonov and A. N. Shiryaev. On a method of calculation of semi-invariants. *Theory Probab. Appl.* 4 (1959), 319–329.

[L5] R. S. Liptser and A. N. Shiryaev. *Statistics of Random Processes*. Springer-Verlag, New York, 1977.

[L6] R. Sh. Liptser and A. N. Shiryaev. A functional central limit theorem for semimartingales. *Theory Probab. Appl.* **25** (1980), 667–688.

[L7] M. Loève. *Probability Theory*. Springer-Verlag, New York, 1977–78.

[L8] E. Lukacs. *Characteristic Functions*. Hafner, New York, 1960.

[L9] E. Lukacs and R. G. Laha. *Applications of Characteristic Functions*. Hafner, New York, 1964.

[M1] D. E. Maistrov. *Probability Theory: A Historical Sketch*. Academic Press, New York, 1974.

[M2] A. A. Markov. *Calculus of Probabilities* [*Ischislenie veroyatnosteĭ*], 3rd ed. St. Petersburg, 1913.

[M3] L. D. Meshalkin. *Collection of Problems on Probability Theory* [*Sbornik zadach po teorii veroyatnosteĭ*]. Moscow University Press, 1963.

[M4] P.-A. Meyer, Martingales and stochastic integrals. I. *Lecture Notes in Mathematics*, no. 284. Springer-Verlag, Berlin, 1972.

[M5] P.-A. Meyer and C. Dellacherie. Probabilities and potential. *North-Holland Mathematical Studies*, no. 29. Hermann, Paris; North-Holland, Amsterdam, 1978.

[N1] J. Neveu. *Mathematical Foundations of the Calculus of Probability*. Holden-Day, San Francisco, 1965.

[N2] J. Neveu. *Discrete Parameter Martingales*. North-Holland, Amsterdam, 1975.

[N3] J. Neyman. *First Course in Probability and Statistics*. Holt, New York, 1950.

[N4] A. A. Novikov. On estimates and the asymptotic behavior of the probability of nonintersection of moving boundaries by sums of independent random variables. *Math. USSR-Izv.* **17** (1980), 129–145.

[P1] V. V. Petrov. *Sums of Independent Random Variables*. Springer-Verlag, Berlin, 1975.

[P2] J. W. Pratt. On interchanging limits and integrals. *Ann. Math. Stat.* 31 (1960), 74–77.

[P3] Yu. V. Prohorov [Prokhorov]. Asymptotic behavior of the binomial distribution. *Uspekhi Mat. Nauk* 8, no. 3(55) (1953), 135–142 (in Russian).
[P4] Yu. V. Prohorov. Convergence of random processes and limit theorems in probability theory. *Theory Probab. Appl.* 1 (1956), 157–214.
[P5] Yu. V. Prokhorov and Yu. A. Rozanov. *Probability theory*. Springer-Verlag, Berlin–New York, 1969; second edition [*Teoriia veroiatnostei*]. "Nauka", Moscow, 1973.
[R1] B. Ramachandran. *Advanced Theory of Characteristic Functions*. Statistical Publishing Society, Calcutta, 1967.
[R2] A. Rényi. *Probability Theory*, North-Holland, Amsterdam, 1970.
[R3] V. I. Rotar. An extension of the Lindeberg–Feller theorem. *Math. Notes* 18 (1975), 660–663.
[R4] Yu. A. Rozanov. *Stationary Random Processes*. Holden-Day, San Francisco, 1967.
[S1] T. A. Sarymsakov. *Foundations of the Theory of Markov Processes* [*Osnovy teorii protsessov Markova*]. GITTL, Moscow, 1954.
[S2] V. V. Sazonov. Normal approximation: Some recent advances. *Lecture Notes in Mathematics*, no. 879. Springer-Verlag, Berlin–New York, 1981.
[S3] B. A. Sevyastyanov [Sewastjanow]. *Verzweigungsprozesse*. Oldenbourg, Munich–Vienna, 1975.
[S4] A. N. Shiryaev. *Random Processes* [*Sluchainye processy*]. Moscow State University Press, 1972.
[S5] A. N. Shiryaev. *Probability, Statistics, Random Processes* [*Veroyatnost, statistika, sluchainye protsessy*], vols. I and II. Moscow State University Press, 1973, 1974.
[S6] A. N. Shiryaev. *Statistical Sequential Analysis* [*Statisticheskiĭ posledovatelnyi analiz*]. Nauka, Moscow, 1976.
[S7] Ya. G. Sinai. *Introduction to Ergodic Theory*. Princeton Univ. Press, Princeton, 1976.
[S8] S. H. Sirazhdinov. *Limit Theorems for Stationary Markov Chains* [*Predelnye teoremy dlya odnorodnyh tsepeĭ Markova*]. Akad. Nauk Uzbek. SSR, Tashkent, 1955.
[S9] W. F. Stout. *Almost Sure Convergence*. Academic Press, New York, 1974.
[T1] I. Todhunter. *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace*. Macmillan, London, 1865.
[V1] E. Valkeila. A general Poisson approximation theorem. *Stochastics* 7 (1982), 159–171.
[V2] A. D. Venttsel. *A Course in the Theory of Stochastic Processes*. McGraw-Hill, New York, 1981.
[W1] B. L. van der Waerden. *Mathematical Statistics*. Springer-Verlag, Berlin–New York, 1969.
[W2] A. Wald. *Sequential Analysis*. Wiley, New York, 1947.
[Y1] A. M. Yaglom and I. M. Yaglom. *Probability and Information*. Reidel, Dordrecht, 1983.
[Z1] S. Zacks. *The Theory of Statistical Inference*. Wiley, New York, 1971.
[Z2] V. M. Zolotarev. A generalization of the Lindeberg–Feller theorem. *Theory Probab. Appl.* 12 (1967), 606–618.
[Z3] V. M. Zolotarev. Théorèmes limites pour les sommes de variables aléatoires indépendantes qui ne sont pas infinitésimales. *C.R. Acad. Sci. Paris*. Ser. A–B 264 (1967), A799–A800.
[D5] D. Dacunha-Castelle and M. Duflo. *Probabilités et statistiques. 1. Problèmes à temps fixe. 2. Problèmes à temps mobile*. Masson, Paris, 1982; *Probability and Statistics*. Springer-Verlag, New York, 1986 (English translation).

[D6] R. M. Dudley. Distances of probability measures and random variables. *Ann. Math. Statist.* 39 (1968), 1563–1572.

[G7] P. E. Greenwood and A. N. Shiryaev. *Contiguity and the Statistical Invariance Principle*. Gordon and Breach, New York, 1985.

[J1] J. Jacod and A. N. Shiryaev. *Limit Theorems for Stochastic Processes*. Springer-Verlag, Berlin Heidelberg, 1987.

[K12] V. S. Korolyuk. *Aide-mémoire de théorie des probabilités et de statistique mathématique*. Mir, Moscow, 1983; second edition [*Spravochnik po teorii veroyatnosteĭ i matematicheskoĭ statistike*]. "Nauka", Moscow, 1985.

[L10] L. LeCam. *Asymptotic Methods in Statistical Theory*. Springer-Verlag, New York, 1986.

[L11] F. Liese and I. Vajda. *Convex Statistical Distances*. Teubner, Leipzig, 1987.

[L12] R. Sh. Liptser and A. N. Shiryaev. *Theory of Martingales*. Kluwer, Dordrecht, Boston, 1989.

[P6] V. V. Petrov. *Limit Theorems for Sums of Independent Random Variables* [*Predel'nye teoremy dlya summ nezavisimyh sluchaĭnyh velichin*]. Nauka, Moscow, 1987.

[P7] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

[P8] E. L. Presman. Approximation in variation of the distribution of a sum of independent Bernoulli variables with a Poisson law. *Theory of Probability and Its Applications* 30 (1985), no. 2, 417–422.

[R6] Yu. A. Rozanov. *The Theory of Probability, Stochastic Processes, and Mathematical Statistics* [*Teoriya veroyatnosteĭ, sluchaĭnye protsessy i matematicheskaya statistika*]. Nauka, Moscow, 1985.

[S10] B. A. Sevastyanov. *A Course in the Theory of Probability and Mathematical Statistics* [*Kurs teorii veroyatnosteĭ i matematicheskoĭ statistiki*]. Nauka, Moscow, 1982.

[S11] A. N. Shiryaev. *Probability*. Springer-Verlag, Berlin Heidelberg, 1984 (English translation); *Wahrscheinlichkeit*, 1988 (German translation).

[Z4] V. M. Zolotarev. *Modern theory of summation of random variables* [*Sovremennaya teoriya summirovaniya nezavisimyh sluchaĭnyh velichin*]. Nauka, Moscow, 1986.

Index of Symbols

Index

Also see the Historical and Bibliographical Notes (pp. 597–602), and the Index of Symbols.

Graduate Texts in Mathematics

continued from page ii

61 WHITEHEAD. Elements of Homotopy Theory.
62 KARGAPOLOV/MERLZIAKOV. Fundamentals of the Theory of Groups.
63 BOLLOBAS. Graph Theory.
64 EDWARDS. Fourier Series. Vol. I 2nd ed.
65 WELLS. Differential Analysis on Complex Manifolds. 2nd ed.
66 WATERHOUSE. Introduction to Affine Group Schemes.
67 SERRE. Local Fields.
68 WEIDMANN. Linear Operators in Hilbert Spaces.
69 LANG. Cyclotomic Fields II.
70 MASSEY. Singular Homology Theory.
71 FARKAS/KRA. Riemann Surfaces. 2nd ed.
72 STILLWELL. Classical Topology and Combinatorial Group Theory. 2nd ed.
73 HUNGERFORD. Algebra.
74 DAVENPORT. Multiplicative Number Theory. 2nd ed.
75 HOCHSCHILD. Basic Theory of Algebraic Groups and Lie Algebras.
76 IITAKA. Algebraic Geometry.
77 HECKE. Lectures on the Theory of Algebraic Numbers.
78 BURRIS/SANKAPPANAVAR. A Course in Universal Algebra.
79 WALTERS. An Introduction to Ergodic Theory.
80 ROBINSON. A Course in the Theory of Groups. 2nd ed.
81 FORSTER. Lectures on Riemann Surfaces.
82 BOTT/TU. Differential Forms in Algebraic Topology.
83 WASHINGTON. Introduction to Cyclotomic Fields. 2nd ed.
84 IRELAND/ROSEN. A Classical Introduction to Modern Number Theory. 2nd ed.
85 EDWARDS. Fourier Series. Vol. II. 2nd ed.
86 VAN LINT. Introduction to Coding Theory. 2nd ed.
87 BROWN. Cohomology of Groups.
88 PIERCE. Associative Algebras.
89 LANG. Introduction to Algebraic and Abelian Functions. 2nd ed.
90 BRØNDSTED. An Introduction to Convex Polytopes.
91 BEARDON. On the Geometry of Discrete Groups.
92 DIESTEL. Sequences and Series in Banach Spaces.
93 DUBROVIN/FOMENKO/NOVIKOV. Modern Geometry—Methods and Applications. Part I. 2nd ed.
94 WARNER. Foundations of Differentiable Manifolds and Lie Groups.
95 SHIRYAEV. Probability. 2nd ed.
96 CONWAY. A Course in Functional Analysis. 2nd ed.
97 KOBLITZ. Introduction to Elliptic Curves and Modular Forms. 2nd ed.
98 BRÖCKER/TOM DIECK. Representations of Compact Lie Groups.
99 GROVE/BENSON. Finite Reflection Groups. 2nd ed.
100 BERG/CHRISTENSEN/RESSEL. Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions.
101 EDWARDS. Galois Theory.
102 VARADARAJAN. Lie Groups, Lie Algebras and Their Representations.
103 LANG. Complex Analysis. 3rd ed.
104 DUBROVIN/FOMENKO/NOVIKOV. Modern Geometry—Methods and Applications. Part II.
105 LANG. $SL_2(\mathbf{R})$.
106 SILVERMAN. The Arithmetic of Elliptic Curves.
107 OLVER. Applications of Lie Groups to Differential Equations. 2nd ed.
108 RANGE. Holomorphic Functions and Integral Representations in Several Complex Variables.
109 LEHTO. Univalent Functions and Teichmüller Spaces.
110 LANG. Algebraic Number Theory.
111 HUSEMÖLLER. Elliptic Curves.
112 LANG. Elliptic Functions.
113 KARATZAS/SHREVE. Brownian Motion and Stochastic Calculus. 2nd ed.
114 KOBLITZ. A Course in Number Theory and Cryptography. 2nd ed.
115 BERGER/GOSTIAUX. Differential Geometry: Manifolds, Curves, and Surfaces.
116 KELLEY/SRINIVASAN. Measure and Integral. Vol. I.
117 SERRE. Algebraic Groups and Class Fields.
118 PEDERSEN. Analysis Now.
119 ROTMAN. An Introduction to Algebraic Topology.
120 ZIEMER. Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation.
121 LANG. Cyclotomic Fields I and II. Combined 2nd ed.
122 REMMERT. Theory of Complex Functions. *Readings in Mathematics*

123 EBBINGHAUS/HERMES et al. Numbers. *Readings in Mathematics*
124 DUBROVIN/FOMENKO/NOVIKOV. Modern Geometry—Methods and Applications. Part III.
125 BERENSTEIN/GAY. Complex Variables: An Introduction.
126 BOREL. Linear Algebraic Groups. 2nd ed.
127 MASSEY. A Basic Course in Algebraic Topology.
128 RAUCH. Partial Differential Equations.
129 FULTON/HARRIS. Representation Theory: A First Course. *Readings in Mathematics*
130 DODSON/POSTON. Tensor Geometry.
131 LAM. A First Course in Noncommutative Rings.
132 BEARDON. Iteration of Rational Functions.
133 HARRIS. Algebraic Geometry: A First Course.
134 ROMAN. Coding and Information Theory.
135 ROMAN. Advanced Linear Algebra.
136 ADKINS/WEINTRAUB. Algebra: An Approach via Module Theory.
137 AXLER/BOURDON/RAMEY. Harmonic Function Theory.
138 COHEN. A Course in Computational Algebraic Number Theory.
139 BREDON. Topology and Geometry.
140 AUBIN. Optima and Equilibria. An Introduction to Nonlinear Analysis.
141 BECKER/WEISPFENNING/KREDEL. Gröbner Bases. A Computational Approach to Commutative Algebra.
142 LANG. Real and Functional Analysis. 3rd ed.
143 DOOB. Measure Theory.
144 DENNIS/FARB. Noncommutative Algebra.
145 VICK. Homology Theory. An Introduction to Algebraic Topology. 2nd ed.
146 BRIDGES. Computability: A Mathematical Sketchbook.
147 ROSENBERG. Algebraic K-Theory and Its Applications.
148 ROTMAN. An Introduction to the Theory of Groups. 4th ed.
149 RATCLIFFE. Foundations of Hyperbolic Manifolds.
150 EISENBUD. Commutative Algebra with a View Toward Algebraic Geometry.
151 SILVERMAN. Advanced Topics in the Arithmetic of Elliptic Curves.
152 ZIEGLER. Lectures on Polytopes.
153 FULTON. Algebraic Topology: A First Course.
154 BROWN/PEARCY. An Introduction to Analysis.
155 KASSEL. Quantum Groups.
156 KECHRIS. Classical Descriptive Set Theory.
157 MALLIAVIN. Integration and Probability.
158 ROMAN. Field Theory.
159 CONWAY. Functions of One Complex Variable II.
160 LANG. Differential and Riemannian Manifolds.
161 BORWEIN/ERDÉLYI. Polynomials and Polynomial Inequalities.
162 ALPERIN/BELL. Groups and Representations.
163 DIXON/MORTIMER. Permutation Groups.
164 NATHANSON. Additive Number Theory: The Classical Bases.
165 NATHANSON. Additive Number Theory: Inverse Problems and the Geometry of Sumsets.
166 SHARPE. Differential Geometry: Cartan's Generalization of Klein's Erlangen Program.
167 MORANDI. Field and Galois Theory.
168 EWALD. Combinatorial Convexity and Algebraic Geometry.
169 BHATIA. Matrix Analysis.
170 BREDON. Sheaf Theory. 2nd ed.
171 PETERSEN. Riemannian Geometry.
172 REMMERT. Classical Topics in Complex Function Theory.
173 DIESTEL. Graph Theory.
174 BRIDGES. Foundations of Real and Abstract Analysis.
175 LICKORISH. An Introduction to Knot Theory.
176 LEE. Riemannian Manifolds.
177 NEWMAN. Analytic Number Theory.
178 CLARKE/LEDYAEV/STERN/WOLENSKI. Nonsmooth Analysis and Control Theory.
179 DOUGLAS. Banach Algebra Techniques in Operator Theory. 2nd ed.
180 SRIVASTAVA. A Course on Borel Sets.
181 KRESS. Numerical Analysis.
182 WALTER. Ordinary Differential Equations.
183 MEGGINSON. An Introduction to Banach Space Theory.
184 BOLLOBAS. Modern Graph Theory.
185 COX/LITTLE/O'SHEA. Using Algebraic Geometry.
186 RAMAKRISHNAN/VALENZA. Fourier Analysis on Number Fields.

This book contains a systematic treatment of probability from the ground up, starting with intuitive ideas and gradually developing more sophisticated subjects, such as random walks, martingales, Markov chains, ergodic theory, weak convergence of probability measures, stationary stochastic processes, and the Kalman-Bucy filter. Many examples are discussed in detail, and there are a large number of exercises. The book is accessible to advanced undergraduates and can be used as a text for self-study.

This new edition contains substantial revisions and updated references. The reader will find a deeper study of topics such as the distance between probability measures, metrization of weak convergence, and contiguity of probability measures. Proofs for a number of some important results which were merely stated in the first edition have been added. The author has included new material on the probability of large deviations, on the central limit theorem for sums of dependent random variables, and on a discrete version of Ito's formula.

ISBN 0-387-94549-0

EAN

9 780387 945491 >

献给 A. H. 柯尔莫戈洛夫诞辰一百周年 (1903 — 1987)

目 录

第一章 初等概率论

我们把只涉及有限个事件概率的那一部分概率论称做初等概率论.

A. H. 柯尔莫戈洛夫,《概率论的基本概念》, 32

§1. 有限种结局试验的概率模型

1. 基本事件空间① 考虑某项试验, 其结果在某一组条件 (“条件复形”) 下由有限种不同的结局 (现象) $\omega_1, \cdots, \omega_N$ 描绘. 关于这些结局的实际本性并不重要, 重要的是不同结局的个数 N 是有限的. 我们把这些结局 $\omega_1, \cdots, \omega_N$ 称做基本事件, 而把一切结局的全体

$$\Omega = \{\omega_1, \cdots, \omega_N\}$$

称做 (有限) *基本事件空间*, 或*样本空间*.

引进基本事件空间的概念, 是形成不同试验的*概率模型* (概率的理论) 的第一步. 下面是描绘基本事件空间构造的几个例子.

例 1 对于 “掷一枚硬币”, 基本事件空间由两个点组成:

$$\Omega = \{Z, F\},$$

其中 Z 表示出现 “正面”, 而 F 表示出现 “反面”. (这时假设, 诸如 “硬币在棱上立着”, “硬币丢失” …… 的情况不会出现. 假设不出现 “正面” 就出现 “反面”.)

例 2 将一枚硬币重复掷 n 次, 基本事件空间为

$$\Omega = \{\omega : \omega = (a_1, \cdots, a_n)\}, \quad a_i = \text{Z或 F},$$

且基本事件的总数 $N(\Omega) = 2^n$.

例 3 首先掷一枚硬币. 如果硬币掷出正面 Z, 则掷六面分别刻有 1,2,3,4,5,6 个点的色子; 假如硬币掷出反面 F, 则再掷一次硬币. 那么, 该试验的基本事件空间为

$$\Omega = \{Z1, Z2, Z3, Z4, Z5, Z6, FZ, FF\},$$

其中 Z 表示硬币掷出 “正面”, F 表示硬币掷出 “反面”, 而 1,2,3,4,5,6 表示色子掷出的点数

① 原书个别 “小节” 有 “标题”. 为便于读者使用, 在没有标题的 “小节”, 我们特别加上了标题. ——译者

注 在讲概率论时, 一般不提 “试验在一组条件下进行”, 而默认这样的 “条件” 存在. 然而, 这样的 “条件” 即便对于初等概率论也很重要, 不同的 “条件” 对同一试验, 也可导致完全不同的概率模型和概率的理论. (参见下文的第 3 小节 “事件的关系和运算” 开头关于这个问题的说明.)

2. 随机抽样与随机分配 现在研究更复杂一些的例子, 这些例子讨论自含 M 个不同球的箱子中, 抽取 n 个球的不同方法.

例 4 *放回抽样*. 如果每次将抽到的球在下一次抽球前放回箱中, 则试验称做放回抽样. 这时, 由 n 个球形成的每一个样本可以表示为向量 $(a_1, \cdots, a_n)$, 其中 $a_i (i = 1, 2, \cdots, n)$ 是第 i 次抽到的球的编号. 易见, 对于放回抽样, 每个 a_i 可以是 $1, 2, \cdots, M$ 等 M 个数中任何一个数. 基本事件空间的描述, 本质上与如下情形有关: 诸如 (4,1,2,1) 和 (1,4,2,1) 是认为是不同的基本事件, 还是同一基本事件. 因此, 习惯上区分两种情形: *有序抽样*和*无序抽样*. 在前一种情形下, 由同一些元素组成的两个样本, 只要其中元素的前后顺序有所不同, 就视为*不同*的样本. 在后一种情形下, 不管元素的顺序, 只要由同一些元素组成的样本, 都视为*同一个*样本. 为强调具体的样本究竟是哪一种, 对于*有序样本*, 我们将使用记号 $(a_1, \cdots, a_n)$, 而*无序样本*则记作 $[a_1, \cdots, a_n]$.

于是, *放回有序样本*的基本事件空间 Ω 具有如下构造:

$$\Omega = \{\omega : \omega = (a_1, \cdots, a_n), a_i = 1, 2, \cdots, M\},$$

且不同结局 (样本) 的个数, 等于

$$N(\Omega) = M^n, \tag{1}$$

即组合分析中的*自 M 个元素中选 n 个且允许重复的置换的总数*.

放回无序样本. 即排列组合中的 “*自 M 个元素中选 n 个且允许重复的组合的总数*”, 其基本事件空间为:

$$\Omega = \{\omega : \omega = [a_1, \cdots, a_n], a_i = 1, 2, \cdots, M\}.$$

易见, 不同的无序样本总数 $N(\Omega)$, 小于有序样本总数. 现在证明, 不同无序样本总数:

$$N(\Omega) = C_{M+n-1}^n, \tag{2}$$

其中 C_k^l *是自 k 个元素中取 l 个的组合数*:

$$C_k^l = \frac{k!}{l!(k-l)!}.$$

我们用数学归纳法证明. 以 $N(M, n)$ 表示我们感兴趣的结局的个数. 易见, 对于一切 $k \leqslant M$,

$$N(k, 1) = k = C_k^1.$$

现在假设 $N(k,n)=C_{k+n-1}^{n}(k\leqslant M)$，并且证明，如果将 n 换成 $n+1$ 此式仍然成立。对于无序样本 $[a_1,\cdots,a_n]$，可以认为其顺序按不减顺序排列：$a_1\leqslant a_2\leqslant\cdots\leqslant a_{n+1}$。显然，无序样本的个数，当 $a_1=1$ 时为 $N(M,n)$；当 $a_1=2$ 时为 $N(M-1,n)$；依次类推，得

$$\begin{aligned}N(M,n+1)&=N(M,n)+N(M-1,n)+\cdots+N(1,n)\\&=C_{M+n-1}^{n}+C_{M-1+n-1}^{n}+\cdots+C_{n}^{n}\\&=(C_{M+n}^{n+1}-C_{M+n-1}^{n+1})+(C_{M-1+n}^{n+1}-C_{M-1+n-1}^{n+1})\\&\quad+\cdots+(C_{n+1}^{n+1}-C_{n}^{n})+C_{n}^{n}=C_{M+n}^{n+1}.\end{aligned}$$

其中用到二项式系数的如下性质：

$$C_{k}^{l-1}+C_{k}^{l}=C_{k+1}^{l}.$$

这一等式，是由帕斯卡 (B. Pascal) 三角形计算二项式系数的基础，很容易证明。

例 5 *不放回抽样*。假设 $n\leqslant M$，且凡是抽到的球都不再放回。这里，也考虑两种可能的抽法：有序抽样和无序抽样。

对于*不放回有序抽样* (组合分析中称做自 M 个元素中选 n 个的有序无重复排列)，其基本事件空间为

$$\Omega=\{\omega:\omega=(a_1,\cdots,a_n),a_k\neq a_l,k\neq l,a_i=1,2,\cdots,M\},$$

而 Ω 中的元素的总数等于 $M(M-1)\cdots(M-n+1)$。将此数记作 $(M)_n$ 或 A_M^n，称做"*自 M 个元素中选 n 个的排列数*"。

对于*不放回无序抽样* (组合分析中称做自 M 个元素中选 n 个*无重复的组合*)，其基本事件空间为

$$\Omega=\{\omega:\omega=[a_1,\cdots,a_n],a_k\neq a_l,k\neq l,a_i=1,2,\cdots,M\},$$

共含

$$N(\Omega)=C_M^n \tag{3}$$

个元素。事实上，由 n 个不同元素形成的无序数组 $[a_1,\cdots,a_n]$，可以得到恰好 $n!$ 个有序数组，从而

$$N(\Omega)\times n!=(M)_n,\quad N(\Omega)=\frac{(M)_n}{n!}=C_M^n.$$

如果自含 M 个球的箱中选 n 个球，则可以将不同结局的个数归纳为表 1-1：

表 1－1

	抽样方式		不同抽法的总数
自含 M 个球箱子 Ω 的 n 次抽样	放回	有序	M^n
		无序	C_{M+n-1}^{n}
	不放回	有序	$(M)_n$
		无序	C_M^n

对于 $M=3,n=2$，相应基本事件空间的结构列为表 1－2：

表 1－2

	抽样方式		基本事件
自含 M 个球箱子 Ω 的 n 次抽样	放回	有序	(1, 1)(1, 2)(1, 3)(2, 1)(2, 2)(2, 3)(3, 1)(3, 2)(3, 3)
		无序	[1, 1][2, 2][3, 3][1, 2][1, 3][2, 3]
	不放回	有序	(1, 2)(1, 3)(2, 1)(2, 3)(3, 1)(3, 2)
		无序	[1, 2][1, 3][2, 3]

例 6 *按箱分配质点*。考虑按箱分配质点问题：将 n 个质点 (小球) 分配到 M 个箱中去，并研究质点分配问题中基本事件空间的构造。问题产生于统计物理，例如，在研究 n 个粒子 (如质子、电子、……) 按 M 种状态的分布。

假设 M 个箱子一一编号为 $1,2,\cdots,M$，而 n 个质点可以辨别 (两两不同)，且分别编号 $1,2,\cdots,n$。那么，n 个质点分配到 M 个箱子中，完全可以由数组 $(a_1,\cdots,a_n)$ 描绘，其中 a_i 表示第 i 个质点编号为 a_i 的箱中；假如 n 个质点不可辨别 (完全相同)，则它们分配到 M 个箱中完全可以由数组 $[a_1,\cdots,a_n]$ 描绘，其中 a_i 表示第 i 个质点分到编号为 a_i 的箱中。

比较例 4 和例 5 所讨论的情形，可以得到如下对应关系：

$$(\text{有序抽样})\Leftrightarrow(\text{质点可以辨}),$$
$$(\text{无序抽样})\Leftrightarrow(\text{质点不可辨}).$$

这种对应关系表示，在"自盛有 M 个球的箱子中抽选 n 个球"的问题中，有序 (无序) 抽样相当于：可辨质点 (不可辨质点) 分配到 M 个箱中的分配问题。

下面的对应关系具有类似的含义：

$$(\text{放回抽样})\Leftrightarrow(\text{箱中可容纳任意多质点}),$$
$$(\text{不放回抽样})\Leftrightarrow(\text{箱中最多可容纳一个质点}).$$

由这些对应关系，可以建立如下类型的对应关系：

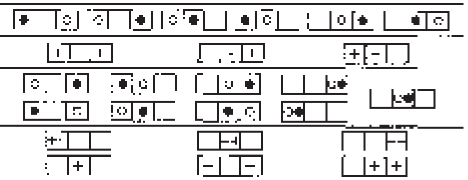
$$(\text{无序不放回抽样})\Leftrightarrow(\text{不可辨质点按箱分配，且每箱最多容纳一个质点})$$

等等. 由于这些对应关系, 就可以在可辨和不可辨质点按箱分配问题中, 利用例 4 和例 5 描绘基本事件空间, 既包括 "有抑制" 的箱子 (最多容纳一个质点的箱子) 和 "无抑制" 的箱子 (可容纳任意个质点的箱子).

表 1-3 演示按两个箱子分配 3 个小球的情形, 其中 "○" 和 "●" 分别表示白球和黑球. 如果分到箱中的球不可辨, 则用 "+" 表示.

表 1-3

自含 M 个球箱子 Ω 中的 n 次抽样	分配方式		基本事件
	有抑制	可辨	(见下图)
		不可辨	(见下图)
	无抑制	可辨	(见下图)
		不可辨	(见下图)



对于质点按箱分配问题中, 利用上面指出的 "抽样" 与 "分配" 之间的对偶性, 就可以明显地求出分配结局的个数. 相应的结果归纳为表 1-4, 其中也包含表 1-1 的结果.

表 1-4

自含 M 个球箱子 Ω 中的 n 次抽样	抽样方式		不同抽法的总数	分配方式		自含 M 个球箱子 Ω 中的 n 次抽样
	放回	有序	M^n ①	可辨	有抑制	
		无序	C^n_{M+n-1} ②	不可辨		
	不放回	有序	$(M)_n$ ③	可辨	无抑制	
		无序	C^n_M ④	不可辨		
	抽样方式		不同抽法的总数	分配方式		

统计物理学中, 把不满足泡利 (W. Pauli) 抑制原理 ("每箱中最多容纳一个质点") 的不同 (不可辨) 质点, 称做麦克斯韦–波尔茨曼 (J. C. Maxwell-L. Boltzmann) (物理) 统计 (见表 1-4 ①); 有抑制不可辨质点的情形称做玻色–爱因斯坦 [S. N. Bose-A. Einstein] 统计 (见表 1-4 ②). 如果质点不可辨但满足抑制原理, 则称做费米–迪拉克 (E. Fermi-P. A. M. Dirac) 统计 (见表 1-4 ④). 熟知, 例如电子、质子和中子都服从费米–迪拉克统计, 光子和 π 介子服从玻色–爱因斯坦统计 (见表 1-4 ②). 此外, 在物理学中再没有发现可辨粒子不服从抑制原理的情形.

3. 事件及其关系和运算 除基本事件空间的概念外, 现在引进重要概念事件. 事件的概念, 是建立所考察试验的各种概率模型 ("理论") 的基础. 在试验的结果中, 试验者一般并不关心究竟出现了哪种具体的结局, 而关心出现的结局属于

一切结局集合的哪个子集. 满足试验条件的一切子集 $A \subseteq \Omega$, 分为两种类型: "结局 $\omega \in A$" 或 "结局 $\omega \notin A$". 我们称这样的子集 A 为事件.

例如, 将一枚硬币重复掷三次, 一切可能结局的空间 Ω, 由 8 个点构成:

$$\Omega = \{000, 001, 010, 011, 100, 101, 110, 111\},$$

其中 0 和 1 分别表示掷出 "正面" 和 "反面". 如果由 "一组条件" 可以记录 (确定、测量等) 所有 3 次掷硬币的结果, 则例如

$$A = \{000, 001, 010, 100\}$$

就是事件: "将一枚硬币重复掷三次" 正面至少出现两次. 假如由 "一组条件" 只能确定第一次掷出的结果, 则 A 已经不能称为事件, 因为关于 "试验的具体结局 ω 是否属于 A", 既不能肯定也不能否定.

从给定的作为事件的某一组集合出发, 经过某些变换可以形成一些新事件. 这些变换, 可以用逻辑的语言表述为 "或"、"与"、"非", 而用集合论的语言表述为 "并"、"交" 和 "补".

并 对于两个集合 A 与 B, 称

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ 或 } \omega \in B\}$$

为集合 A 与 B 的并, 表示由属于集合 A 或 B 的点形成的集合. 用概率论的语言, $A \bigcup B$ 表示事件 {事件 A 与 B 至少出现一个}.

交 称 AB 或

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ 且 } \omega \in B\}$$

为两个集合 A 与 B 的交, 表示由既属于集合 A 同时又属于集合 B 的点形成的集合. 用概率论的语言, $A \bigcap B$ 表示事件 {事件 A 与 B 同时出现}.

例如, 如果 $A = \{00, 01, 10\}$, $B = \{11, 10, 01\}$, 则

$$A \cup B = \{00, 01, 10, 11\} = \Omega, \quad A \cap B = \{10, 01\}.$$

补 如果 A 是 Ω 的子集, 则称 $\overline{A} = \{\omega \in \Omega : \omega \notin A\}$ 为集合 A 补, 表示 Ω 中不属于 A 的点的集合.

差 属于 B 但不属于 A 的点的集合称做 B 与 A 的差, 记作 $B \setminus A$. 那么, $\overline{A} = \Omega \setminus A$. 用概率的语言, $\overline{A}$ 表示事件 "A 不出现". 例如, 事件 $A = \{00, 01, 10\}$, 则事件 $\overline{A} = \{11\}$ 表示接连两次掷出反面.

不可能事件和必然事件 一般用 $\varnothing$ 表示空集. 概率论中空集 $\varnothing$ 称做不可能事件, 集合 Ω 自然称做必然事件. 集合 A 和 $\overline{A}$ 没有共同点, 因而 $A \bigcap \overline{A}$ 是空集.①

① 对于事件 A 和 B, 若 $A \bigcap B = \varnothing$, 则称 A 和 B 不相容, 否则称 A 和 B 相容. ——译者

和 当 A 与 B 不相交时 ($AB=\varnothing$), 集合 A 与 B 的并称做集合 A 与 B 的和, 记作 $A+B$.

事件代数 考虑集合 $A\subseteq\Omega$ 的某个集系 $\mathscr{A}_0$, 则利用集合运算 $\cup,\cap$ 与 $\setminus$ 可以由 $\mathscr{A}_0$ 构造新集系, 其中元素也是事件. 给这些事件补充上必然事件 Ω 和不可能事件 $\varnothing$, 得集系 $\mathscr{A}$, 则 $\mathscr{A}$ 是代数. 所谓 "代数" 即 Ω 的这样的集系, 满足

1) $\Omega\in\mathscr{A}$.

2) 若 $A\in\mathscr{A}, B\in\mathscr{A}$, 则集合 $A\cup B, A\cap B, A\setminus B$ 也都属于 $\mathscr{A}$.

由以上的叙述, 可见作为事件系最好考虑本身是代数的集系. 以后, 我们正是考虑这样的集系.

我们列举几个事件代数的例子.

a) $\mathscr{A}=\{\Omega,\varnothing\}$ —— 集系由 Ω 和空集 $\varnothing$ 构成, 称做平凡代数;

b) $\mathscr{A}=\{A,\overline{A},\Omega,\varnothing\}$ —— 事件 A 产生的集系;

c) $\mathscr{A}=\{A:A\subseteq\Omega\}$ —— Ω 全部子集的集系 (包括空集 $\varnothing$).

易见, 所有这些事件代数是按下面的原则得到的.

分割 我们称集系

$$\mathscr{D}=\{D_1,\cdots,D_n\}$$

构成集合 Ω 的一个分割, 而 $D_1,\cdots,D_n$ 是该分割的原子, 如果 $D_1,\cdots,D_n$ 非空且两两不相容, 而它们的和等于 Ω:

$$D_1+\cdots+D_n=\Omega.$$

例如, 假定集合 Ω 由 3 个点构成: $\Omega=\{1,2,3\}$, 则存在 5 个不同的分割:

$$\begin{aligned}
&\mathscr{D}_1=\{D_1\} && D_1=\{1,2,3\};\\
&\mathscr{D}_2=\{D_1,D_2\} && D_1=\{1,2\}, D_2=\{3\};\\
&\mathscr{D}_3=\{D_1,D_2\} && D_1=\{1,3\}, D_2=\{2\};\\
&\mathscr{D}_4=\{D_1,D_2\} && D_1=\{2,3\}, D_2=\{1\};\\
&\mathscr{D}_5=\{D_1,D_2,D_3\} && D_1=\{1\}, D_2=\{2\}, D_3=\{3\}.
\end{aligned}$$

(关于有限集合的分割的总数参见本节的习题 2.)

如果考虑 $\mathscr{D}$ 中一切集合的并连同空集 $\varnothing$, 则得到的集系是代数, 称做 $\mathscr{D}$ 产生的代数, 记作 $\sigma(\mathscr{D})$. 于是, 代数 $\sigma(\mathscr{D})$ 的元素由空集 $\varnothing$ 与分割 $\mathscr{D}$ 之原子中集合的和组成.

这样, 如果 $\mathscr{D}$ 是 Ω 的某一分割, 则它与代数 $\mathscr{B}=\sigma(\mathscr{D})$ 一一对应.

逆命题也正确. 设 $\mathscr{B}$ 是有限空间 Ω 的子集的代数, 则存在唯一分割 $\mathscr{D}$, 其原子是代数 $\mathscr{B}$ 的元素, 并且 $\mathscr{B}=\sigma(\mathscr{D})$. 事实上, 假设集合 $\mathscr{D}\in\mathscr{B}$ 并且具有性质: 对于任意 $B\in\mathscr{B}$, 集合 $D\cap B$ 要么与 D 重合, 要么是空集. 那么, 这样集合 D 的全体组成分割 $\mathscr{D}$ 并且具有所要求的性质 $\mathscr{B}=\sigma(\mathscr{D})$. 对于例 a), 取只含一个集合的 $D_1=\Omega$ 平凡分割; 对于例 b), $\mathscr{D}=\{A,\overline{A}\}$. 对于例 c), $\mathscr{D}$ 是只含一个点的集合 $\{\omega_i\}, \omega_i\in\Omega$ 的最细小分割, 即 $\mathscr{D}$ 产生的代数是 Ω 的一切子集的代数.

对两个分割 $\mathscr{D}_1$ 和 $\mathscr{D}_2$, 如果 $\sigma(\mathscr{D}_1)\subseteq\sigma(\mathscr{D}_2)$, 则称分割 $\mathscr{D}_2$ 比分割 $\mathscr{D}_1$ "细小", 记作 $\mathscr{D}_1\preccurlyeq\mathscr{D}_2$.

像前面一样, 假设空间 Ω 有有限个点 $\omega_1,\cdots,\omega_N$ 构成. 记 $N(\mathscr{A})$ 为例 c) 中组成体系 $\mathscr{A}$ 的集合的总数. 我们证明 $N(\mathscr{A})=2^N$. 事实上, 每一个非空集合 $A\in\mathscr{A}$ 可以表示为 $A=\{\omega_{i_1},\cdots,\omega_{i_k}\}(1\leqslant k\leqslant N)$, 其中 $\omega_{i_j}\in\Omega$. 将该集合与由 0 或 1 形成的序列

$$(0,\cdots,0,1,0,\cdots,0,1,\cdots),$$

其中在编号为 $i_1,\cdots,i_k$ 的位置上为 1, 而在其余位置上为 0. 那么, 对于固定的 k, 形如 $A=\{\omega_{i_1},\cdots,\omega_{i_k}\}$ 的不同集合 A 的总数等于 k 个 1 (k 个不可辨质点) 分配 N 个位置 (N 个箱子) 不同分法的总数. 根据表 1-1 的情形④, 这样分法的总数等于 C_N^k. 由此可见

$$N(\mathscr{A})=1+C_N^1+\cdots+C_N^N=2^N,$$

其中包括空集 $\varnothing$.

4. 概率空间 为建立只有有限种可能结局的随机试验的概率模型 (理论), 我们暂时完成了最初的两步: 引进了基本事件空间 Ω, 并建立了 Ω 子集的某种体系 $\mathscr{A}$ —— 代数, 其中的子集称做事件. 有时把 $\mathscr{E}=(\Omega,\mathscr{A})$ 等同于试验. 现在进行下一步: 赋予每一个基本事件 (每一种结局或现象) $\omega_i\in\Omega(i=1,\cdots,N)$ 某种 "权", 记作 $p(\omega_i)$ 或 p_i, 称做基本事件 (结局) ω_i 的概率. 假设 $p(\omega_i)$ 满足条件:

a) $0\leqslant p(\omega_i)\leqslant 1$ (非负性);

b) $p(\omega_1)+\cdots+p(\omega_N)=1$ (规范性).

从给定的基本事件 ω_i 的概率 $p(\omega_i)$ 出发, 按公式

$$\mathbf{P}(A)=\sum_{\{i:\omega_i\in A\}}p(\omega_i) \tag{4}$$

定义任意事件 $A\in\mathscr{A}$ 的概率.

定义 1 通常称

$$(\Omega,\mathscr{A},\mathbf{P})$$

为 "概率空间", 其中 $\Omega=\{\omega_1,\cdots,\omega_N\}$, $\mathscr{A}$ 是 Ω 的子集的代数, 而 $\mathbf{P}=\{\mathbf{P}(A):A\in\mathscr{A}\}$. 概率空间决定 (定义) 只有有限种可能结局 (基本事件) 的, 随机试验的概率模型 (理论). 显然, $\mathbf{P}(\{\omega_i\})=p(\omega_i)(i=1,\cdots,N)$

$$\mathbf{P}(\varnothing)=0, \tag{5}$$

$$\mathbf{P}(\Omega)=1, \tag{6}$$

$$\mathbf{P}(A\cup B)=\mathbf{P}(A)+\mathbf{P}(B)-\mathbf{P}(A\cap B). \tag{7}$$

特别, 若 $A\cap B=\varnothing$, 则

$$\mathrm{P}(A\cup B)=\mathrm{P}(A)+\mathrm{P}(B); \tag{8}$$

而

$$\mathrm{P}(\bar{A})=1-\mathrm{P}(A). \tag{9}$$

5. 古典型概率 在一些具体的情形下建立概率模型时, 给出基本事件空间 Ω 和代数 $\mathscr{A}$, 一般并不复杂. 这时, 在初等概率论里, 一般用 Ω 的全部子集的代数当做代数 $\mathscr{A}$. 较困难的问题, 是定义基本事件的概率. 实际上, 对这个问题的回答已经超出了概率论的范围, 我们不在此过多地讨论这个问题. 我们的基本任务, 并不是如何赋予某一个试验基本事件的概率, 而是根据基本事件的概率, 计算复合事件 ($\mathscr{A}$ 中的事件) 的概率.

从数学观点来看十分清楚, 对于有限基本事件空间, 只要赋予基本事件 $\omega_1,\cdots,\omega_N$ 以满足 $p_1+\cdots+p_N=1$ 的非负实数 $p_1,\cdots,p_N$, 就可以得到一切可以想象的 (有限的) 概率空间.

对于具体的情形, 所确定的数值的正确性, 可以一定程度地利用以后将要介绍的大数定律来验证. 在上述情形下, 根据大数定律, 对于给定的在相同条件下进行的较长"独立"试验系列, 基本事件出现的频率"十分接近"它们相应的概率.

鉴于赋予试验基本事件以概率值的困难, 我们指出, 存在许多实际情形, 在这些情形下由于对称性或均衡性的直观, 把一切可能出现的基本事件视为等可能的是合理的. 因此, 假如基本事件空间 Ω 由点 $\omega_1,\cdots,\omega_N$ 构成, 其中 $N<\infty$, 则

$$p(\omega_1)=\cdots=p(\omega_N)=\frac{1}{N},$$

从而对于任何事件 $A\in\mathscr{A}$,

$$\mathrm{P}(A)=\frac{N(A)}{N}, \tag{10}$$

其中 $N(A)$ 是事件 A 所含基本事件的个数.

这样求概率的方法称做古典型方法. 显然, 这时求概率 $\mathrm{P}(A)$ 归结为计算导致事件 A 的基本事件的个数. 这一般用排列组合的方法来实现. 因此, 在涉及有限集合的概率的计算中, 排列组合的方法占有重要的地位.

例 7 *重合问题*. 假设箱中有 M 个球, 分别编号为 $1,2,\cdots,M$. 进行 n 次放回抽样, 并且认为样本是有序的. 显然, 这时

$$\Omega=\{\omega:\omega=(a_1,\cdots,a_n),a_i=1,\cdots,M\},$$

而 $N(\Omega)=M^n$. 根据古典型概率的求法, 认为所有 M^n 个结局是等可能的. 提出下面的问题: 事件

$$A=\{\omega:a_i\neq a_j,i\neq j\}$$

的概率如何? 即事件"抽到的元素无重复"的概率如何? 显然

$$N(A)=M(M-1)\cdots(M-n+1)=(M)_n.$$

因此

$$\mathrm{P}(A)=\frac{(M)_n}{M^n}=\left(1-\frac{1}{M}\right)\left(1-\frac{2}{M}\right)\cdots\left(1-\frac{n-1}{M}\right). \tag{11}$$

该例题有如下有趣的解释. 假设某一班中有 n 名学生, 每个学生的生日在一年 365 天中每一天是等可能的. 问"n 个学生中至少有两人的生日在同一天"的概率 $P_{365}(n)$ 如何? 可以把"选择"生日, 视为在 365 个球中任选一个球, 则根据 (11) 式

$$P_{365}(n)=1-\frac{(365)_n}{365^n}.$$

对于某些 n, 下表列出了概率 $P_{365}(n)$ 的值:

n	4	16	22	23	40	64	70	100
$P_{365}(n)$	0.016 36	0.283 60	0.475 69	0.507 30	0.891 23	0.997 11	0.999 16	0.999 999 69

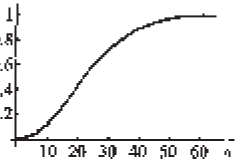
对于充分大的 M,

$$\ln\frac{(M)_n}{M^n}=\sum_{k=1}^{n-1}\ln\left(1-\frac{k}{M}\right)\sim-\frac{1}{M}\sum_{k=1}^{n-1}k=-\frac{n(n-1)}{2M},$$

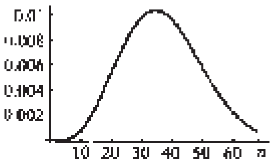
亦即

$$P_M(n)\equiv1-\frac{(M)_n}{M^n}\sim1-e^{-\frac{n(n-1)}{2M}}\quad[\equiv\tilde{P}_M(n)],\quad M\to\infty.$$

下面是函数 $P_{365}(n)$ 的图形. 在相同的尺度下, $\tilde{P}_{365}(n)$ 逼近的图形实际上与函数 $P_{365}(n)$ 的图形相重合. 在区间 $[0,60]$ 的范围内二者的最大差别近似等于 0.01 (在 $n=30$ 的邻域内).



$P_{365}(n)$ 和 $\tilde{P}_{365}(n)$ 图形



$P_{365}(n)-\tilde{P}_{365}(n)$ 图形

有趣的是, (和期望的相反!) 当一个班的人数并不是太多, 只有 23 名学生时, "n 个学生中至少有两人的生日在同一天"的概率就已达 1/2.

例 8 *抽彩*. 考虑按如下规则的抽彩. 假设总共有 M 张编号为 $1,2,\cdots,M$ 彩票, 其中编号为 $1,2,\cdots,n\ (M\geqslant 2n)$ 的中彩. 一人买了 n 张彩票, 问至少有一张中奖的概率 (记作 P) 如何?

由于抽取彩票的顺序，对于所购买的彩票是否中奖无意义，所以可以认为基本事件空间有如下结构：

$$\Omega = \{\omega : \omega = [a_1, \cdots, a_n], a_k \neq a_l, k \neq l, a_i = 1, \cdots, M\}.$$

由表 1－1 知 $N(\Omega) = C_M^n$. 现在假设

$$A_0 = \{\omega : \omega = [a_1, \cdots, a_n], a_k \neq a_l, k \neq l, a_i = n+1, \cdots, M\}$$

是事件"所买的彩票中没有中奖的"，仍然由表 1－1 知 $N(A_0) = C_{M-n}^n$. 因此，

$$\mathbf{P}(A_0) = \frac{C_{M-n}^n}{C_M^n} = \frac{(M-n)_n}{(M)_n} = \left(1 - \frac{n}{M}\right)\left(1 - \frac{n}{M-1}\right)\cdots\left(1 - \frac{n}{M-n+1}\right)$$

即

$$P = 1 - \mathbf{P}(A_0) = 1 - \left(1 - \frac{n}{M}\right)\left(1 - \frac{n}{M-1}\right)\cdots\left(1 - \frac{n}{M-n+1}\right).$$

如果 $M = n^2$ 且 $n \to \infty$，则 $\mathbf{P}(A_0) \to e^{-1}$，

$$P \to 1 - e^{-1} \approx 0.632,$$

其中收敛的速度相当快：当 $n = 10$ 时，概率已经为 $P = 0.670$.

6. 练习题

1. 验证运算 $\bigcap$ 和 $\bigcup$ 的下列性质的正确性：

交换律：$A \bigcup B = B \bigcup A, A \bigcap B = B \bigcap A$;

结合律：$A \bigcup (B \bigcup C) = (A \bigcup B) \bigcup C, A \bigcap (B \bigcap C) = (A \bigcap B) \bigcap C$;

分配律：$A \bigcap (B \bigcup C) = (A \bigcap B) \bigcup (A \bigcap C), A \bigcup (B \bigcap C) = (A \bigcup B) \bigcap (A \bigcup C)$;

幂等性：$A \bigcup A = A, A \bigcap A = A$;

对偶律：$\overline{A \bigcup B} = \overline{B} \bigcap \overline{A}, \overline{A \bigcap B} = \overline{B} \bigcup \overline{A}$.

2. 设集合 Ω 含 N 个元素. 证明不同分割的贝尔 (A.G.Bell) 数 B_N 决定于

$$B_N = e^{-1} \sum_{k=1}^{\infty} \frac{k^N}{k!}. \tag{12}$$

提示 证明

$$B_N = \sum_{k=0}^{N-1} C_{N-1}^k B_k, \quad B_0 = 1,$$

然后证明 (12) 式满足上面的递推公式.

3. 证明，对于任何有限集合系 $A_1, \cdots, A_n$，有（概率的半可加性）

$$\mathbf{P}(A_1 \cup \cdots \cup A_n) \leqslant \mathbf{P}(A_1) + \cdots + \mathbf{P}(A_n).$$

4. 设 A 和 B 是任意二事件. 证明，$A\overline{B} \bigcup B\overline{A}$ 表示事件"A 和 B 中必出现一个且只出现一个"，并且

$$\mathbf{P}(A\overline{B} \cup B\overline{A}) = \mathbf{P}(A) + \mathbf{P}(B) - 2\mathbf{P}(AB).$$

5. 设有 $A_1, \cdots, A_n$，而量 $S_0, S_1, \cdots, S_n$ 决定于如下关系式：$S_0 = 1$,

$$S_r = \sum_{J_r} \mathbf{P}(A_{k_1} \cap \cdots \cap A_{k_r}), \quad 1 \leqslant r \leqslant n,$$

其中对集合 $\{1, 2, \cdots, n\}$ 的无序子集 $J_r = [k_1, \cdots, k_r] (k_i \neq k_j, i \neq j)$ 求和.

(1) 设 B_m 表示事件"在 $A_1, \cdots, A_n$ 中必出现一个且只出现一个". 证明

$$\mathbf{P}(B_m) = \sum_{r=m}^{n} (-1)^{r-m} C_r^m S_r.$$

特别，对于 $m = 0$,

$$\mathbf{P}(B_0) = 1 - S_1 + S_2 - \cdots \pm S_n.$$

(2) 证明，"$A_1, \cdots, A_n$ 中有 m 个同时出现"的概率等于

$$\mathbf{P}(B_m) + \cdots + \mathbf{P}(B_n) = \sum_{r=m}^{n} (-1)^{r-m} C_{r-1}^{m-1} S_r.$$

特别，"$A_1, \cdots, A_n$ 中至少出现一个"的概率等于

$$\mathbf{P}(B_1) + \cdots + \mathbf{P}(B_n) = S_1 - S_2 + \cdots \mp S_n.$$

(3) 证明下列性质：

(a) 邦弗尔罗尼 (Bonferroni) 不等式①：对于任意 $k = 1, 2, \cdots, (2k \leqslant n)$，有.

$$S_1 - S_2 + \cdots - S_{2k} \leqslant \mathbf{P}\left(\bigcup_{i=1}^{n} A_i\right) \leqslant S_1 - S_2 + \cdots + S_{2k-1};$$

(b) 庞加莱 (J. H. Poincare) 恒等式：

$$\mathbf{P}\left(\bigcup_{r=1}^{n} A_r\right) = \sum_{r=1}^{n} (-1)^{r-1} S_r;$$

(c) 弗雷歇 (M. Fréshet) 不等式：

$$\frac{S_{r+1}}{C_n^{r+1}} \leqslant \frac{S_r}{C_n^r}, \quad (r = 0, 1 \cdots, n-1);$$

① 见 W. Feller, An Introduction to probability theory, V. 1, ch. IV. —— 译者

(d) 冈贝尔 (E. J. Gumbel) 不等式:

$$\frac{C_n^{r+1}-S_{r+1}}{C_{n-1}^{r}} \leqslant \frac{C_n^{r}-S_r}{C_{n-1}^{r-1}},\quad (r=1,2,\cdots,n-1);$$

6. 证明 $\mathbf{P}(A\bigcap B\bigcap C)\geqslant \mathbf{P}(A)+\mathbf{P}(B)+\mathbf{P}(C)-2$, 并用归纳法证明

$$\mathbf{P}\left(\bigcap_{i=1}^{n}A_i\right)\geqslant\sum_{i=1}^{n}\mathbf{P}(A_i)-(n-1).$$

7. 在关于 n 和 M 不同类型 (类型: $n=xM, M\to\infty$ 或 $n=x\sqrt{M}, M\to\infty$, 其中 x 固定) 的假设下, 讨论例 7 中概率 $P_M(n)$ 的渐近性. 将结果与 §6 中的局部极限定理进行比较.

§2.　某些经典模型和分布

1. 二项分布　假设将一枚硬币接连掷 n 次, 观测结果用有序数组 $(a_1,\cdots,a_n)$ 表示, 其中, 当第 i 次掷出现正面时 $a_i=1$, 当第 i 次掷出现反面时 $a_i=0$. 基本事件空间具有如下形式:

$$\Omega=\{\omega:\omega=(a_1,\cdots,a_n), a_i=0 \text{ 或 } 1\}.$$

赋予每一个基本事件 $\omega=(a_1,\cdots,a_n)$ 概率 (权重)

$$p(\omega)=p^{\sum a_i}q^{n-\sum a_i},$$

其中 p 和 q 非负且 $p+q=1$. 首先证明, 这样定义概率 (权重) $p(\omega)$ 是合理的. 为此, 只需验证

$$\sum_{\omega\in\Omega}p(\omega)=1.$$

考虑所有满足

$$\sum_i a_i=k,\quad (k=0,1,\cdots,n)$$

的基本事件 $\omega=(a_1,\cdots,a_n)$. 根据表 1-4 (k 个不可辨的 "1" 分配到 n 个位置上), 这样的基本事件个数等于 C_n^k. 因此

$$\sum_{\omega\in\Omega}p(\omega)=\sum_{k=0}^{n}C_n^k p^k q^{n-k}=(p+q)^n=1.$$

设 $\mathscr{A}$ 是空间 Ω 的一切子集的代数, 在 $\mathscr{A}$ 上定义了概率:

$$\mathbf{P}(A)=\sum_{\omega\in A}p(\omega),\quad A\in\mathscr{A},$$

其中 $\mathbf{P}(\{\omega\})=p(\omega), \omega\in\Omega$, 从而定义了一个概率模型. 自然称做描绘 n 次掷硬币的概率模型.

当 $n=1$ 时, 基本事件空间仅含两个点 $\omega=1$ ("成功") 和 $\omega=0$ ("失败"), 而概率 $p(1)=p$, 自然称做 "成功的概率". 以后, 我们将看到, 作为 n 次 "独立" 试验的结果, 可以得到由上面讨论的描绘 n 次掷硬币的概率模型, 其中 p 是每次试验成功的概率.

事件

$$A_k=\{\omega:\omega=(a_1,\cdots,a_n), a_1+\cdots+a_n=k\},\quad k=0,1,\cdots,n,$$

表示恰好 k 次 "成功". 由以上的讨论, 可见

$$\mathbf{P}(A_k)=C_n^k p^k q^{n-k},\ \text{且}\ \sum_{k=0}^{n}\mathbf{P}(A_k)=1.\tag{1}$$

概率 $(\mathbf{P}(A_0),\mathbf{P}(A_1),\cdots,\mathbf{P}(A_n))$ 组称做二项分布 (在容量为 n 的样本中 "成功" 的次数服从二项分布).

二项分布在概率论中起着十分重要的作用, 出现在多种不同的概率模型中. 记

$$P_n(k)=\mathbf{P}(A_k),\quad k=0,1,\cdots,n.$$

对于 $p=1/2$ (掷 "对称" 硬币) 的情形, 当 $n=5,10,20$ 时, 图 1 是二项分布的示意图.

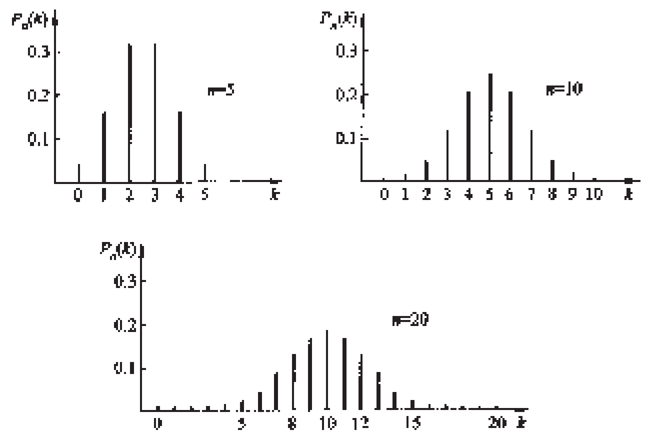


图 1　二项概率的示意图 ($n=5,10,20$)

我们再引进一个 (实际上与上一个等价的) 模型, 它描绘某一质点的随机游动. 假设质点从 0 出发, 经过单位时间向上或向下移动一步 (图 2).

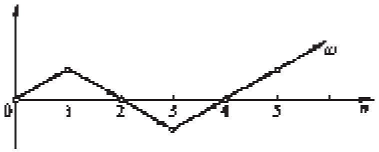


图 2

于是, 经过 n 步质点最多向上移动 n 步或向下移动 n 步. 显然, 质点运动的每一条 "轨道" ω, 可以完全由数组 $(a_1,\cdots,a_n)$ 描绘, 其中当质点在第 i 步向上移动时 $a_i=+1$, 而当质点在第 i 步向下移动时 $a_i=-1$. 赋予每条 "轨道" ω 以概率,

$$p(\omega)=p^{\nu(\omega)}q^{n-\nu(\omega)},$$

其中 $\nu(\omega)$ 是数组 $\omega=(a_1,\cdots,a_n)$ 中 "+1" 的个数, 即

$$\nu(\omega)=\frac{(a_1+\cdots+a_n)+n}{2},$$

而 p 和 q 非负且 $p+q=1$. 由于

$$\sum_{\omega\in\Omega}p(\omega)=1,$$

可见概率组 $\{p(\omega)\}$, 连同轨道 $\omega=(a_1,\cdots,a_n)$ 的空间 Ω 及其子集, 确实决定质点 n 步运动的概率模型.

现在提出另一个问题: 事件 A_k "质点经 n 步到达纵坐标为 k 的点" 的概率如何? 一切满足 $\nu(\omega)-[n-\nu(\omega)]=k$, 即满足

$$\nu(\omega)=\frac{n+k}{2}$$

的轨道都满足上面提出的条件. 这样轨道的条数等于 $C_n^{(n+k)/2}$ (见表 1 · 4), 因此

$$\mathrm{P}(A_k)=C_n^{(n+k)/2}p^{(n+k)/2}q^{(n-k)/2}.$$

于是, 可以认为二项分布 $\{\mathrm{P}(A_{-n}),\cdots,\mathrm{P}(A_0),\cdots,\mathrm{P}(A_n)\}$, 描绘质点移动 n 步后位置的概率分布. 特别, 对于 "对称" (即 $p=q=1/2$) 的情形, 每条轨道的概率等于 2^{-n}, 故

$$\mathrm{P}(A_k)=C_n^{(n+k)/2}2^{-n}.$$

下面讨论当 $n\to\infty$ 时这些概率的渐进性质.

如果移动步数等于 $2n$, 则由二项式系数的性质可见, 在概率 $\mathrm{P}(A_k)(|k|\leqslant 2n)$ 中最大概率为

$$\mathrm{P}(A_0)=C_{2n}^{n}2^{-2n}.$$

由斯特林 (J.Stirling) 公式[*] (见 (6) 式)

$$n!\sim\sqrt{2\pi n}\,e^{-n}n^n.$$

所以

$$C_{2n}^{n}=\frac{(2n)!}{(n!)^2}\sim\frac{2^{2n}}{\sqrt{\pi n}},$$

即对充分大的 n,

$$\mathrm{P}(A_0)\sim\frac{1}{\sqrt{\pi n}}.$$

对于质点移动 $2n$ 步的情形, 图 3 是产生二项分布的示意图 (注意, 与图 2 不同, 时轴是铅直的, 并且正向朝上).

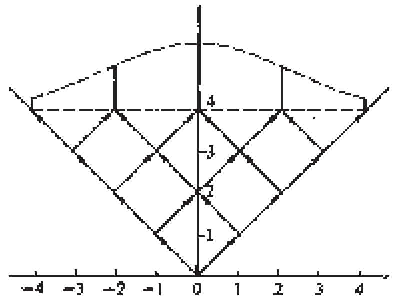


图 3 二项分布的产生

2. 多项分布 推广上述模型, 假设基本事件空间有如下构造

$$\Omega=\{\omega:\omega=(a_1,\cdots,a_n),a_i=b_1,\cdots,b_r\},$$

其中 $b_1,\cdots,b_r$ 是给定的数. 设 $\nu_i(\omega)$ 是序列 $\omega=(a_1,\cdots,a_n)$ 中等于 $b_i(i=1,\cdots,r)$ 元素的个数, 而基本事件的 ω 概率等于

$$p(\omega)=p_1^{\nu_1(\omega)}\cdots p_r^{\nu_r(\omega)},$$

其中 $p_i\geqslant 0, p_1+\cdots+p_r=1$. 注意,

$$\sum_{\omega\in\Omega}p(\omega)=\sum_{\substack{n_1\geqslant 0,\cdots,n_r\geqslant 0\\ n_1+\cdots+n_r=n}}C_n(n_1,\cdots,n_r)p_1^{n_1}\cdots p_r^{n_r},$$

[*] 关系式 $f(n)\sim g(n)$ 表示 $f(n)/g(n)\to 1\ (n\to\infty)$.

和式中以 $C_n(n_1,\cdots,n_r)$ 表示有序数组 $(a_1,\cdots,a_n)$ 的个数，其中元素 b_1 重复 n_1 次，$\cdots$ 元素 b_r 重复 n_r 次. 因为，有 $C_n^{n_1}$ 种方法将 n_1 个元素 b_1 安排在 n 个位置上，有 $C_{n-n_1}^{n_2}$ 种方法将 n_2 个元素 b_2 安排在 $n-n_1$ 个位置上，等等. 所以

$$C_n(n_1,\cdots,n_r)=C_n^{n_1}C_{n-n_1}^{n_2}\cdots C_{n-(n_1+\cdots+n_{r-1})}^{n_r}$$
$$=\frac{n!}{n_1!(n-n_1)!}\times\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}\times\cdots\times 1=\frac{n!}{n_1!\cdots n_r!},$$

因此

$$\sum_{\omega\in\Omega}p(\omega)=\sum_{\left\{\substack{n_1\geqslant 0,\cdots,n_r\geqslant 0\\ n_1+\cdots+n_r=n}\right\}}\frac{n!}{n_1!\cdots n_r!}p_1^{n_1}\cdots p_r^{n_r}=(p_1+\cdots+p_r)^n=1,$$

于是，多项分布确实是概率分布.

设

$$A_{n_1,\cdots,n_r}=\{\omega:\nu_1(\omega)=n_1,\cdots,\nu_r(\omega)=n_r\},$$

则

$$\mathbf{P}(A_{n_1,\cdots,n_r})=C_n(n_1,\cdots,n_r)p_1^{n_1}\cdots p_r^{n_r}. \tag{2}$$

概率组 $\{\mathbf{P}(A_{n_1,\cdots,n_r})\}$ 称做*多项分布*.

注意，多项分布及其特例 —— 二项分布的产生与放回抽样相联系.

3. 多元超几何分布 出现在不放回抽样中.

作为例子，假设一箱子中有编号为 $1,2,\cdots,M$ 的 M 个不同的球，其中 M_1 个球具有颜色 $b_1,\cdots,M_r$ 个球具有颜色 b_r，且 $M_1+\cdots+M_r=M$. 现在从箱中进行 $n(n<M)$ 次不放回抽样. 基本事件空间为

$$\Omega=\{\omega:\omega=(a_1,\cdots,a_n),a_k\neq a_l,k\neq l;a_i=1,\cdots,M\},$$

而 $N(\Omega)=(M)_n$. 假设基本事件是等可能的，而 $B_{n_1,\cdots,n_r}$ 表示事件："n_1 个球具有颜色 $b_1,\cdots,n_r$ 个球具有颜色 b_r"，且 $n_1+\cdots+n_r=n$. 求事件 $B_{n_1,\cdots,n_r}$ 的概率. 容易证明，

$$N(B_{n_1,\cdots,n_r})=C_n(n_1,\cdots,n_r)(M_1)_{n_1}\cdots(M_r)_{n_r},$$

因此

$$\mathbf{P}(B_{n_1,\cdots,n_r})=\frac{N(B_{n_1,\cdots,n_r})}{N(\Omega)}=\frac{C_{M_1}^{n_1}\cdots C_{M_r}^{n_r}}{C_M^n}. \tag{3}$$

概率组 $\{\mathbf{P}(B_{n_1,\cdots,n_r})\}$ 称做*多元超几何分布*. 当 $r=2$ 时此分布简称为*超几何分布*，因为其母函数是超几何函数.

多元超几何分布的构造相当复杂，因为概率

$$\mathbf{P}(B_{n_1,n_2})=\frac{C_{M_1}^{n_1}C_{M_2}^{n_2}}{C_M^n},\ n_1+n_2=n,\ M_1+M_2=M \tag{4}$$

包含 9 个阶乘数. 易见，如果当 $M\to\infty,M_1\to\infty$ 时，且 $M_1/M\to p$，从而 $M_2/M\to 1-p$，则

$$\mathbf{P}(B_{n_1,n_2})\to C_{n_1+n_2}^{n_1}p^{n_1}(1-p)^{n_2}. \tag{5}$$

换句话说，在上述条件下，超几何分布逼近二项分布. 这直观上是明显的，因为当 M 和 M_1 (有限) 较大时，由不放回抽样得到几乎与放回抽样一样的结果.

例 利用公式 (4)，求体育抽彩中猜中 6 个 "幸运" 号码. 其实际意义如下.

假设有编号为 $1,2,\cdots,49$ 的 49 个球，其中 6 个球是 "幸运的" (例如，6 个红球，其余是白球). 随意从中不放回抽出 6 个球. 问抽出的 6 个球全部都是 "幸运的" 概率如何？设 $M=49,M_1=6,n_1=6,n_2=0$，则我们感兴趣的事件为

$$B_{6,0}=\{6\text{ 个球都是 “幸运的”}\},$$

而根据公式 (4)，其概率为

$$\mathbf{P}(B_{6,0})=\frac{1}{C_{49}^6}\approx 7.2\times 10^{-8}.$$

4. 斯特林公式 随 n 的增大，阶乘数 $n!$ 增长的非常快. 例如，

$$10!=3\ 628\ 800,$$
$$15!=1\ 307\ 674\ 368\ 000,$$

而 100! 包含 158 位数字. 因此，无论是从理论上还是实际计算的角度，斯特林公式

$$n!=\sqrt{2\pi n}\left(\frac{n}{e}\right)^n\exp\left\{\frac{\theta_n}{12n}\right\},\ 0<\theta_n<1$$

都很重要. 数学分析的大量文献中，都有斯特林公式证明 (亦可参见第八章 §8 练习题 1).

5. 练习题

1. 证明 (5) 式的命题.

2. 证明，对于多项分布的概率，在满足 $np_i-1<k_i\leqslant(n+r-1)p_i(i=1,2,\cdots,r)$ 的点 $(k_1,\cdots,k_r)$，达到最大值.

3. 一维伊金格 (Ezenger) 模型. 有 n 个质点，分别位于编号 $1,2,\cdots,n$ 的点. 假设质点分为两种类型：类型 I 有 n_1 个质点，类型 II 有 n_2 个质点，$n_1+n_2=n$，并且全部 n 个质点的 $n!$ 排列是等可能的.

建立相应的概率模型，并求事件

$$A_n(m_{11},m_{12},m_{21},m_{22})=\{\nu_{11}=m_{11},\nu_{12}=m_{12},\nu_{21}=m_{21},\nu_{22}=m_{22}\}$$

的概率，其中 ν_{ij} —— 排在 II 型质点 j 之后的 I 型质点 i 的个数 $(i,j=1,2)$.

4. 利用概率和组合的方法, 证明下列恒等式:

$$\sum_{k=0}^{n} C_n^k = 2^n, \qquad \sum_{k=0}^{n} (C_n^k)^2 = C_{2n}^n,$$

$$\sum_{k=0}^{n} (-1)^{n-k} C_m^k = C_{m-1}^n, \; m \geqslant n+1,$$

$$\sum_{k=0}^{m} k(k-1) C_m^k = m(m-1)2^{m-2}, \; m \geqslant 2,$$

$$kC_n^k = nC_{n-1}^{k-1},$$

$$C_n^m = \sum_{j=0}^{m} C_k^j C_{n-k}^{m-j},$$

$(0 \leqslant m \leqslant n, 0 \leqslant k \leqslant n$, 对于 $j < 0$ 和 $j > l$, 设 $C_l^j = 0)$.

5. 假设 N 是某个总体的容量, 要求在对总体的全部元素没有简单重复计数的情况下, 以 "最少的费用" 估计 N. 例如, 在估计某个国家、城市 …… 人口时, 对类似的问题感兴趣.

拉普拉斯在 1786 年法国人口为 N 时, 曾经提出下面的方法.

从总体中选择若干个元素 (例如选择 M 个元素), 并且作上标记. 然后将这 M 个元素放回原总体, 并且与无标记的元素 "很好地混合". 然后从 "混合好了的" 总体中再抽取 n 个元素, 以 X 表示其中有标记的元素个数.

(1) 证明, 由超几何分布的公式 (4), 相应的概率 $\mathbf{P}_{N,M,n}\{X = m\}$ 可以表示为:

$$\mathbf{P}_{N,M,n}\{X = m\} = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}.$$

(2) 假设 M, n 和 m 固定, 对 N 求上面概率的最大值, 即求总体的 "最大似然" 容量 N, 使 (对于给定的 M, n) 有标记的元素的个数 X 等于 m.

(3) 证明, 总体容量的最大似然估计值 (记作 $\widehat{N}$), 由公式表示:

$$\widehat{N} = [Mnm^{-1}],$$

其中 $[\cdot]$ 表示整数部分.

这样得到的估计量 $\widehat{N}$, 称做 N 的最大似然估计量. (在 §7 练习题 4 将继续讨论该问题).

6. (对照 §1 练习题 2.) 设 Ω 含 N 个元素, 而 $\tilde{d}(N)$ 表示具有如下性质的不同分割的个数: 分割的每一个子集有奇数个元素. 证明

$$\tilde{d}(1) = 1, \quad \tilde{d}(2) = 1, \quad \tilde{d}(3) = 2,$$
$$\tilde{d}(4) = 5, \quad \tilde{d}(6) = 12, \quad \tilde{d}(7) = 37,$$

而一般, 有

$$\sum_{n=1}^{\infty} \frac{\tilde{d}(n)x^n}{n!} = e^{\sinh x} - 1, \quad |x| < 1.$$

§3. 条件概率, 独立性

1. 事件的条件概率 事件的概率的概念, 可以回答下面一类问题: 假如箱中有 M 个球, 其中 M_1 个白球和 M_2 个黑球; 以 A 表示事件: 随意抽出一个球是白球. 问事件 A 的概率 $\mathbf{P}(A)$ 如何? 按古典型概率 $\mathbf{P}(A) = M_1/M$.

下面引进的条件概率的概念, 可以回答这样一类问题: 在第一次抽到白球 (事件 A) 的条件下, 在第二次也抽到白球 (事件 B) 的概率如何? (这里讨论的是不放回抽样).

这里自然应该这样讨论: 假如第一次抽到的球是白色的, 那么第二次抽球前箱中还有 $M-1$ 个球, 其中 $M_1 - 1$ 个白球, 另外还有 M_2 个黑球; 直观上显然, 我们感兴趣的 (条件) 概率等于 $(M_1 - 1)/(M-1)$.

现在给出与直观概念一致的, 条件概率的定义.

设 $(\Omega, \mathscr{A}, \mathbf{P})$ 是 (有限) 概率空间, 而 A 是某个事件 (即 $A \in \mathscr{A}$).

定义 1 设 $\mathbf{P}(A) > 0$, 称

$$\frac{\mathbf{P}(AB)}{\mathbf{P}(A)} \tag{1}$$

为在事件 A 的条件下, 事件 B 的条件概率 (记作 $\mathbf{P}(B|A)$).

对于古典方法 (§1 的第 4 小节), 概率定义为

$$\mathbf{P}(A) = \frac{N(A)}{N(\Omega)}, \quad \mathbf{P}(AB) = \frac{N(AB)}{N(\Omega)},$$

则

$$\mathbf{P}(B|A) = \frac{N(AB)}{N(A)}. \tag{2}$$

由定义 1 直接得出, 条件概率如下的性质:

$$\mathbf{P}(A|A) = 1,$$
$$\mathbf{P}(\varnothing|A) = 0,$$
$$\mathbf{P}(B|A) = 1, \quad B \supseteq A,$$
$$\mathbf{P}(B_1 + B_2|A) = \mathbf{P}(B_1|A) + \mathbf{P}(B_2|A).$$

由这些性质可见, 对于固定的事件 A, 在概率空间 $(\Omega \bigcap A, \mathscr{A} \bigcap A)$ 上的条件概率 $\mathbf{P}(\cdot|A)$, 以及在空间 $(\Omega, \mathscr{A})$ 上的概率 $\mathbf{P}(\cdot)$ 具有同样的性质, 其中

$$\mathscr{A} \bigcap A = \{B \bigcap A : B \in \mathscr{A}\}.$$

显然,

$$\mathbf{P}(B|A)+\mathbf{P}(\overline{B}|A)=1,$$

然而, 一般

$$\mathbf{P}(B|A)+\mathbf{P}(B|\overline{A})\neq 1,$$

$$\mathbf{P}(B|A)+\mathbf{P}(\overline{B}|\overline{A})\neq 1.$$

例 1　考虑有两个孩子的家庭. 问一家中有两个男孩的概率如何? 假设

a) 岁数较大的是男孩;

b) 至少有一个是男孩.

显然, 基本事件空间为 (b——男孩, g——女孩):

$$\Omega=\{\mathrm{bb},\mathrm{bg},\mathrm{gb},\mathrm{gg}\},$$

其中 bg 表示 "岁数较大的是男孩, 岁数较小的是女孩", 等等.

假设 4 个基本事件都是等可能的, 那么,

$$\mathbf{P}(\mathrm{bb})=\mathbf{P}(\mathrm{bg})=\mathbf{P}(\mathrm{gb})=\mathbf{P}(\mathrm{gg})=\frac{1}{4}.$$

设事件 A"年长的是男孩", B"年幼的是男孩". 因此, $A\bigcup B$ 表示事件 "至少一个是男孩", 而 $A\bigcap B$ 表示事件 "两个全部是男孩". 我们感兴趣的问题是条件概率: a) $\mathbf{P}(AB|A)$, b) $\mathbf{P}(AB|A\bigcup B)$:

$$\mathbf{P}(AB|A)=\frac{\mathbf{P}(AB)}{\mathbf{P}(A)}=\frac{1/4}{1/2}=\frac{1}{2},$$

$$\mathbf{P}(AB|A\bigcup B)=\frac{\mathbf{P}(AB)}{\mathbf{P}(A\bigcup B)}=\frac{1/4}{3/4}=\frac{1}{3}.$$

2. 全概率公式和乘法公式　另一个简单而重要的公式 (3), 称做全概率公式, 是利用条件概率计算复合事件的基本工具.

考虑基本事件空间 Ω 的某个分割 $\mathscr{D}=\{A_1,\cdots,A_n\}$, 且 $\mathbf{P}(A_i)>0(i=1,2,\cdots,n)$. (这样的分割又称做不相容事件的完全事件组.) 显然,

$$B=BA_1+\cdots+BA_n,$$

因此

$$\mathbf{P}(B)=\sum_{i=1}^{n}\mathbf{P}(BA_i),$$

其中

$$\mathbf{P}(BA_i)=\mathbf{P}(B|A_i)\mathbf{P}(A_i).$$

从而, 有全概率公式:

$$\mathbf{P}(B)=\sum_{i=1}^{n}\mathbf{P}(B|A_i)\mathbf{P}(A_i).\qquad(3)$$

特别, 如果 $0<\mathbf{P}(A)<1$, 则

$$\mathbf{P}(B)=\mathbf{P}(B|A)\mathbf{P}(A)+\mathbf{P}(B|\overline{A})\mathbf{P}(\overline{A}).\qquad(4)$$

例 2　箱中有 M 个球, 其中 m 个是 "幸运的". 现在不放回地从中先后抽取两个球, 求后抽到的球是 "幸运的" 概率. 假设所有结局是等可能的, 而且关于先抽到的第一个球的情况未知. 设事件 $A=\{$ 先抽到的球是 "幸运的"$\}$, $B=\{$ 后抽到的球是 "幸运的"$\}$. 那么,

$$\mathbf{P}(B|A)=\frac{\mathbf{P}(AB)}{\mathbf{P}(A)}=\frac{\dfrac{m(m-1)}{M(M-1)}}{\dfrac{m}{M}}=\frac{m-1}{M-1},$$

$$\mathbf{P}(B|\overline{A})=\frac{\mathbf{P}(\overline{A}B)}{\mathbf{P}(\overline{A})}=\frac{\dfrac{m(M-m)}{M(M-1)}}{\dfrac{M-m}{M}}=\frac{m}{M-1},$$

从而, 由全概率公式, 有

$$\mathbf{P}(B)=\mathbf{P}(B|A)\mathbf{P}(A)+\mathbf{P}(B|\overline{A})\mathbf{P}(\overline{A})$$
$$=\frac{m-1}{M-1}\times\frac{m}{M}+\frac{m}{M-1}\times\frac{M-m}{M}=\frac{m}{M}.$$

有趣的是概率 $\mathbf{P}(A)$ 也等于 m/M. 这样, 尽管不知道先抽出的球的情况, 但是后抽出的是 "幸运球" 的概率并没有改变.

由条件概率的定义知, 当 $\mathbf{P}(A)>0$ 时, 有

$$\mathbf{P}(AB)=\mathbf{P}(B|A)\mathbf{P}(A),\qquad(5)$$

该式称做乘法公式. 用数学归纳法可以将其推广: 假设事件组 $A_1,\cdots,A_n$ 满足条件: $\mathbf{P}(A_1\cdots A_n)>0$, 则

$$\mathbf{P}(A_1\cdots A_n)=\mathbf{P}(A_1)\mathbf{P}(A_2|A_1)\cdots\mathbf{P}(A_n|A_1\cdots A_{n-1}),\qquad(6)$$

其中 $A_1A_2\cdots A_n=A_1\bigcap A_2\bigcap\cdots\bigcap A_n$.

3. 贝叶斯公式　设事件 A 和 B 的概率大于 0: $\mathbf{P}(A)>0,\mathbf{P}(B)>0$. 则与 (5) 式同样, 有

$$\mathbf{P}(AB)\quad\mathbf{P}(A|B)\mathbf{P}(B).\qquad(7)$$

由 (5) 和 (7) 式得所谓贝叶斯 (Bayes) 公式:

$$\mathbf{P}(A|B)=\frac{\mathbf{P}(A)\mathbf{P}(B|A)}{\mathbf{P}(B)}. \tag{8}$$

假如事件组 $A_1,\cdots,A_n$ 是 Ω 的一个分割, 则由 (3) 和 (8) 式得

$$\mathbf{P}(A_i|B)=\frac{\mathbf{P}(A_i)\mathbf{P}(B|A_i)}{\sum_{j=1}^{n}\mathbf{P}(A_j)\mathbf{P}(B|A_j)}, \tag{9}$$

称做贝叶斯定理.

在统计应用中, 事件 $A_1,\cdots,A_n$ 组成事件组 ($A_1+\cdots+A_n=\Omega$), 常称做"假设"或"假说", 而 $\mathbf{P}(A_i)$ 称做假设 A_i 的验前概率*). 条件概率 $\mathbf{P}(A_i|B)$ 称做假设 A_i 在事件 B 出现后的验后概率.

例 3 假设匣中有两枚硬币: A_1 —— 对称的硬币, "正面" Z 出现的概率等于 1/2, 而 A_2 —— 不对称的硬币, "正面" Z 出现的概率等于 1/3. 随意选出一枚硬币并将其投掷, 结果掷出正面. 问抽到硬币为对称硬币的概率如何?

建立相应的概率模型. 这里自然取集合 $\Omega=\{A_1\mathrm{Z},A_1\mathrm{F},A_2\mathrm{Z},A_2\mathrm{F}\}$, 可以描绘选取和投掷的结局, 其中 $A_1\mathrm{Z}$ 表示"选中硬币" A_1, 结果掷出正面 Z, 等等, 而 F 表示硬币掷出反面. 根据条件, 所考虑结局的概率应该是:

$$\mathbf{P}(A_1)=\mathbf{P}(A_2)=\frac{1}{2}$$

和

$$\mathbf{P}(\mathrm{Z}|A_1)=\frac{1}{2},\quad \mathbf{P}(\mathrm{Z}|A_2)=\frac{1}{3}.$$

这些条件唯一决定各结局的概率:

$$\mathbf{P}(A_1\mathrm{Z})=\frac{1}{4},\mathbf{P}(A_1\mathrm{F})=\frac{1}{4},\mathbf{P}(A_2\mathrm{Z})=\frac{1}{6},\mathbf{P}(A_2\mathrm{F})=\frac{1}{3}.$$

那么, 根据贝叶斯公式, 所求的概率为

$$\mathbf{P}(A_1|\mathrm{Z})=\frac{\mathbf{P}(A_1)\mathbf{P}(\mathrm{Z}|A_1)}{\mathbf{P}(A_1)\mathbf{P}(\mathrm{Z}|A_1)+\mathbf{P}(A_2)\mathbf{P}(\mathrm{Z}|A_2)}=\frac{3}{5},$$

从而

$$\mathbf{P}(A_2|\mathrm{Z})=\frac{2}{5}.$$

*) a priori —— 验前, a posteriori —— 验后.

4. 独立性 在这一小节将引进的独立性的概念, 它对于概率论在一定意义上有核心作用: 正是独立性的概念决定了概率论的特色, 使它从研究有测度的可测空间的一般理论中分离出来.

对于两个事件 A 和 B, 如果事件 A 的出现, 对事件 B 出现的概率不产生丝毫影响, 自然应该认为事件 B 不依赖于事件 A. 换句话说, 称"事件 B 对事件 A 独立", 如果

$$\mathbf{P}(B|A)=\mathbf{P}(B), \tag{10}$$

其中假设 $\mathbf{P}(A)>0$.

因为

$$\mathbf{P}(B|A)=\frac{\mathbf{P}(AB)}{\mathbf{P}(A)},$$

所以由 (10) 可见

$$\mathbf{P}(AB)=\mathbf{P}(A)\mathbf{P}(B). \tag{11}$$

同样, 设 $\mathbf{P}(B)>0$, 则自然称"事件 A 对事件 B 独立", 如果

$$\mathbf{P}(A|B)=\mathbf{P}(A).$$

由此仍然得到关系式 (11), 于是该式关于事件 A 和 B 是对称的, 并且当事件 A 和 B 的概率 (一个或两个) 可能为 0 时 (11) 式仍然成立. 由此导出独立性的定义.

定义 2 称事件 A 和 B (关于概率 $\mathbf{P}$) 为独立的或统计独立的, 如果

$$\mathbf{P}(AB)=\mathbf{P}(A)\mathbf{P}(B).$$

在概率论中, 往往不但需要考虑事件 (集合) 的独立性, 而且需要研究事件 (集合) 组的独立性. 下面引进相应的定义.

定义 3 称 Ω 子集系的代数 $\mathscr{A}_1$ 和 $\mathscr{A}_2$ (关于概率 $\mathbf{P}$) 为独立的或统计独立的, 如果对于相应地属于 $\mathscr{A}_1$ 和 $\mathscr{A}_2$ 的两个任意子集 A_1 和 A_2 独立.

作为例子, 考虑两个代数:

$$\mathscr{A}_1=\{A_1,\overline{A}_1,\varnothing,\Omega\} \text{ 和 } \mathscr{A}_2=\{A_2,\overline{A}_2,\varnothing,\Omega\},$$

其中 A_1 和 A_2 分别是 $\mathscr{A}_1$ 和 $\mathscr{A}_2$ 中的集合. 不难证明, 两个集合代数 $\mathscr{A}_1$ 和 $\mathscr{A}_2$ 独立当且仅当对任意事件 A_1 和 A_2 独立. 事实上, $\mathscr{A}_1$ 和 $\mathscr{A}_2$ 独立说明, 16 个事件偶 A_1 和 A_2, A_1 和 $\overline{A}_2,\cdots,\Omega$ 和 Ω 独立. 从而 A_1 和 A_2 独立. 相反, 若 A_1 和 A_2 独立, 则需要证明其余 15 个事件偶也独立. 例如, 现在验证 A_1 和 $\overline{A}_2$ 独立. 有

$$\begin{aligned}\mathbf{P}(A_1\overline{A}_2)&=\mathbf{P}(A_1)-\mathbf{P}(A_1A_2)=\mathbf{P}(A_1)-\mathbf{P}(A_1)\mathbf{P}(A_2)\\&=\mathbf{P}(A_1)[1-\mathbf{P}(A_2)]=\mathbf{P}(A_1)\mathbf{P}(\overline{A}_2).\end{aligned}$$

同样可以证明其余事件偶独立.

5. 全体独立和两两独立 两个集合以及两个集合代数独立的概念, 可以推广到任意有限个集合以及集合代数的情形.

定义 4 称集合 $A_1, \cdots, A_n$ (关于概率 $\mathbf{P}$) 全体独立或全体统计独立, 如果对于任何 $k=1,2,\cdots,n$ 和 $1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant n$,

$$\mathbf{P}(A_{i_1} \cdots A_{i_k}) = \mathbf{P}(A_{i_1}) \cdots \mathbf{P}(A_{i_k}). \tag{12}$$

定义 5 称集合 (关于概率 $\mathbf{P}$) 代数 $\mathscr{A}_1, \cdots, \mathscr{A}_n$ 全体独立或全体统计独立, 如果相应地属于 $\mathscr{A}_1, \cdots, \mathscr{A}_n$ 的任何集合 $A_1, \cdots, A_n$ 独立.

需要指出, 由事件两两独立, 一般它们未必全体独立. 事实上, 例如, 若 $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ 且所有基本事件都等可能, 则事件

$$A = \{\omega_1, \omega_2\}, \quad B = \{\omega_1, \omega_3\}, \quad C = \{\omega_1, \omega_4\}$$

两两独立, 但

$$\mathbf{P}(ABC) = \frac{1}{4} \neq \left(\frac{1}{2}\right)^3 = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C).$$

还需要指出, 对于某些事件 A, B, C, 由

$$\mathbf{P}(ABC) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C),$$

一般并不能得出这些事件两两独立. 事实上, 设空间 Ω 由有序数偶组成 (i,j), 其中 $i,j = 1,2,\cdots,6$, 且这些数偶都等可能. 那么, 对于事件

$$A = \{(i,j) : j = 1, 2 \text{ 或 } 5\}, \quad B = \{(i,j) : j = 4, 5 \text{ 或 } 6\}, \quad C = \{(i,j) : i + j = 9\},$$

则

$$\mathbf{P}(AB) = \frac{1}{6} \neq \frac{1}{4} = \mathbf{P}(A)\mathbf{P}(B),$$
$$\mathbf{P}(AC) = \frac{1}{36} \neq \frac{1}{18} = \mathbf{P}(A)\mathbf{P}(C),$$
$$\mathbf{P}(BC) = \frac{1}{12} \neq \frac{1}{18} = \mathbf{P}(B)\mathbf{P}(C),$$

然而

$$\mathbf{P}(ABC) = \frac{1}{36} = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C).$$

6. 伯努利模型 从独立性的概念出发, 详细讨论 §2 中引进的导出二项分布的经典模型 $(\Omega, \mathscr{A}, \mathbf{P})$. 在此模型中,

$$\Omega = \{\omega : \omega = (a_1, \cdots, a_n), a_i = 0, 1\},$$
$$\mathscr{A} = \{A : A \subseteq \Omega\},$$

而 $\mathbf{P}(\{\omega\}) = p(\omega)$, 其中

$$p(\omega) = p^{\sum a_i}(1-p)^{n-\sum a_i}. \tag{13}$$

考虑事件 $A \subseteq \Omega$. 如果事件 A 只决定于 a_k 的值, 则称该事件依赖于 k 时的试验. 事件 A_k 和 $\overline{A}_k$ 就是这样事件的例子:

$$A_k = \{\omega : a_k = 1\}, \quad \overline{A}_k = \{\omega : a_k = 0\}.$$

考虑代数序列: $\mathscr{A}_1, \cdots, \mathscr{A}_n$, 其中 $\mathscr{A}_k = \{A_k, \overline{A}_k, \varnothing, \Omega\}$, 并证明对于 (13) 的情形, 这些代数独立.

显然,

$$\mathbf{P}(A_k) = \sum_{\{\omega : a_k = 1\}} p(\omega) = \sum_{\{\omega : a_k = 1\}} p^{\sum a_i}(1-p)^{n-\sum a_i}$$
$$= p \sum_{\{a_i, i \neq k\}} p^{\sum_{i \neq k} a_i} q^{(n-1) - \sum_{i \neq k} a_i} = p \sum_{l=0}^{n-1} C_{n-1}^{l} p^l q^{(n-1)-l} = p,$$

而类似的计算, 可得 $\mathbf{P}(\overline{A}_k) = q$, 且当 $k \neq l$ 时, 有

$$\mathbf{P}(A_k A_l) = p^2, \quad \mathbf{P}(A_k \overline{A}_l) = pq, \quad \mathbf{P}(\overline{A}_k \overline{A}_l) = q^2.$$

由此容易证明, 代数 $\mathscr{A}_k$ 和 $\mathscr{A}_l$ $(k \neq l)$ 独立.

类似地可以证明代数 $\mathscr{A}_1, \mathscr{A}_2, \cdots, \mathscr{A}_n$ 独立. 因此可以把所讨论的模型 $(\Omega, \mathscr{A}, \mathbf{P})$, 称做适合具有两种结局且成功的概率为 p 的 n 次独立试验的模型. 伯努利 (J. Bernoulli) 最早研究了该模型, 并且对这种模型证明了大数定律 (§5). 因此, 该模型又称做 (具有 "成功" 与 "失败" 两种结局且 "成功" 概率为 p 的) 伯努利模型.

深入研究伯努利模型的概率空间表明, 它具有下面将叙述的 "概率空间直积" 的构造.

假设有 n 个有限概率空间 $(\Omega_1, \mathscr{B}_1, \mathbf{P}_1), \cdots, (\Omega_n, \mathscr{B}_n, \mathbf{P}_n)$. 组成点 $\omega = (a_1, \cdots, a_n)$ 的空间 $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$, 其中 $a_i \in \Omega_i$. 记

$$\mathscr{A} = \mathscr{B}_1 \otimes \mathscr{B}_2 \otimes \cdots \otimes \mathscr{B}_n$$

是 Ω 的子集的代数, 由形如 $A = B_1 \times B_2 \times \cdots \times B_n$ $(B_i \in \mathscr{B}_i)$ 的集合构成. 最后, 对于 $\omega = (a_1, \cdots, a_n)$, 令 $p(\omega) = p_1(a_1) \cdots p_n(a_n)$, 且把集合 $A = B_1 \times B_2 \times \cdots \times B_n$ 的概率定义为:

$$\mathbf{P}(A) = \sum_{\{a_1 \in B_1, \cdots, a_n \in B_n\}} p_1(a_1) \cdots p_n(a_n).$$

不难验证 $\mathbf{P}(\Omega) = 1$, 因此 $(\Omega, \mathscr{A}, \mathbf{P})$ 决定某个概率空间, 称做概率空间

$$(\Omega_1, \mathscr{B}_1, \mathbf{P}_1), \cdots, (\Omega_n, \mathscr{B}_n, \mathbf{P}_n)$$

的直积.

概率空间直积一条容易验证的性质是: 事件

$$A_1=\{\omega:a_1\in B_1\},\cdots,A_n=\{\omega:a_n\in B_n\}$$

关于概率 $\mathbf{P}$ 独立, 其中 $B_i\in\mathscr{B}_i$. 同样, 空间 Ω 的子集代数

$$\mathscr{A}_1=\{A_1:A_1=\{\omega:a_1\in B_1\},B_1\in\mathscr{B}_1\},$$

$$\cdots\cdots\cdots\cdots$$

$$\mathscr{A}_n=\{A_n:A_n=\{\omega:a_n\in B_n\},B_n\in\mathscr{B}_n\}$$

独立.

由上面引进的构造, 可见伯努利概型:

$$(\Omega,\mathscr{A},\mathbf{P}),\ \Omega=\{\omega:\omega=(a_1,\cdots,a_n),a_i=0,1\},$$

$$\mathscr{A}=\{A:A\subseteq\Omega\}\ \text{和}\ \mathbf{P}(\{\omega\})=p^{\sum a_i}(1-p)^{n-\sum a_i}$$

可以由概率空间 $(\Omega_1,\mathscr{B}_1,\mathbf{P}_1),\cdots,(\Omega_n,\mathscr{B}_n,\mathbf{P}_n)$ 的直积得到, 其中

$$\Omega_i=\{0,1\},\quad \mathscr{B}_i=\{\{0\},\{1\},\varnothing,\Omega_i\},$$

$$\mathbf{P}_i(\{1\})=p,\quad \mathbf{P}_i(\{0\})=q.$$

7. 练习题

1. 举例说明等式

$$\mathbf{P}(B|A)+\mathbf{P}(B|\overline{A})=1,$$

$$\mathbf{P}(B|A)+\mathbf{P}(\overline{B}|\overline{A})=1$$

一般不正确.

2. 箱中有 M 球, 其中 M_1 个白球. 考虑不放回抽取的容量为 n 的样本. 以 B_j 表示事件 "第 j 次抽到的是白球", 以 A_k 表示事件 "在容量为 n 的样本中恰好 k 个白球". 证明, 无论对于不放回抽样, 还是放回抽样, 都有

$$\mathbf{P}(B_j|A_k)=\frac{k}{n}.$$

3. 对于独立事件 $A_1,\cdots,A_n$, 证明

$$\mathbf{P}\left(\bigcup_{k=1}^{n}A_k\right)=1-\prod_{k=1}^{n}\mathbf{P}(\overline{A}_k).$$

4. 对于独立事件 $A_1,\cdots,A_n$, 其中 $\mathbf{P}(A_i)=p_i$. 证明这些事件一个都不出现的概率

$$p_0=\prod_{i=1}^{n}(1-p_i).$$

5. 设 A 和 B 是独立事件. 通过 $\mathbf{P}(A)$ 和 $\mathbf{P}(B)$ 表示下列事件的概率: 在 A 和 B 之中, 恰好出现 k 个事件, 至少出现 k 个事件, 以及最多出现 k 个事件.

6. 设事件 A 与它自己独立, 即 A 与 A 独立, 证明 $\mathbf{P}(A)$ 等于 0 或 1.

7. 设事件 A 概率 $\mathbf{P}(A)$ 等于 0 或 1, 证明 A 与任何事件 B 独立.

8. 考虑图 4 的电路图. 在 A,B,C,D 和 E 等 5 个继电器中, 各继电器独立的工作, 每个以概率 p 和 q 分别有 "断" (无信号通过) 和 "通" (有信号通过) 两种状态. 问 "在入口发送的信号, 出口可以收到" 信号的概率如何? 在 E "处于 '通' 的状态下, 出口可以收到" 信号的概率如何?

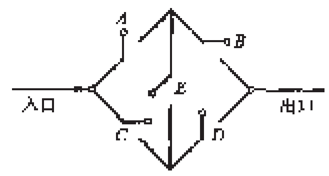


图 4

9. 设 $\mathbf{P}(A+B)>0$, 证明

$$\mathbf{P}(A|A+B)=\frac{\mathbf{P}(A)}{\mathbf{P}(A)+\mathbf{P}(B)}.$$

10. 设事件 A 对事件 $B_n(n\geqslant 1)$ 独立, 并且 $B_i\cap B_j=\varnothing, i\neq j$. 证明事件

$$A\quad\text{与}\quad\bigcup_{n=1}^{\infty}B_n$$

独立.

11. 证明, 如果 $\mathbf{P}(A|C)>\mathbf{P}(B|C)$ 和 $\mathbf{P}(A|\overline{C})>\mathbf{P}(B|\overline{C})$, 则 $\mathbf{P}(A)>\mathbf{P}(B)$.

12. 证明

$$\mathbf{P}(A|B)=\mathbf{P}(A|BC)\mathbf{P}(C|B)+\mathbf{P}(A|B\overline{C})\mathbf{P}(\overline{C}|B).$$

13. 设 X 和 Y 独立, 服从参数为 n 和 p 的二项分布, 证明

$$\mathbf{P}\{X=k|X+Y=m\}=\frac{C_n^kC_n^{m-k}}{C_{2n}^m},\quad k=0,1,\cdots,\min(m,n).$$

14. 设 A,B,C 是两两独立的等概率事件, 且 $A\cap B\cap C=\varnothing$. 求概率 $\mathbf{P}(A)$ 的最大值.

15. 在箱中原来有一个白球. 以相同的概率将一个白球或黑球放入箱中. 然后随意取出一个球, 结果是白球. 问箱中剩下的球也是白球的概率如何?

§4. 随机变量及其特征

1. 随机变量及其概率分布和分布函数 设 $(\Omega,\mathscr{A},\mathbf{P})$ 是某具有有限个结局试验的概率模型, $N(\Omega)$ 是 Ω 中基本事件的个数, 而 $\mathscr{A}$ 是 Ω 中所有子集的代数. 能够理解, 以上讨论的各种事件 $A\in\mathscr{A}$ 概率的计算, 其实基本事件空间的自然本性并不重要. 重要的只是某种数字特征, 其值依赖于基本事件. 我们想知道的是, 在一系

列 n 次试验中, 出现一定次数成功的概率如何, 分到各箱中质点个数服从哪种概率分布, 等等.

现在要引进的随机变量的概念, 可以表征在随机试验中 "测量" 结果的量. 下面, 将再引进随机变量更一般的形式.

定义 1 称任一定义在 (有限) 基本事件空间 Ω 上的数值函数 $\xi = \xi(\omega)$ 为随机变量. (第二章 §4, 引进随机变量的一般概念之后, 就可以清楚知道随机变量的 "简单" 术语的来源.)

例 1 对于接连两次掷硬币模型, 其基本事件空间为 $\Omega = \{ZZ, ZF, FZ, FF\}$. 其中 Z —— 正面, F —— 反面. 我们利用下面的表格定义随机变量 $\xi = \xi(\omega)$:

ω	ZZ	ZF	FZ	FF
$\xi(\omega)$	2	1	1	0

这里, 按实际含义, $\xi(\omega)$ 是对应于 ω 的 "正面" 出现的次数.

随机变量 ξ 的另一简单的例子是某集合 $A \in \mathscr{A}$ 的示性函数 (亦称特征函数):

$$\xi = I_A(\omega),$$

其中*)

$$I_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}$$

当实验者遇到描绘某些记载或读数的随机变量时, 则他关心的基本问题是: 该随机变量取各个数值的概率如何. 从这种观点出发, 关心的不是概率 P 在 $(\Omega, \mathscr{A})$ 上的分布, 而是概率在随机变量之可能值的集合上的分布. 由于对于所研究的情形, Ω 由有限个点构成, 则随机变量 ξ 的值域 X 也是有限的. 设 $X = \{x_1, \cdots, x_m\}$, 其中 $x_1, \cdots, x_m$ 是 ξ 的全部可能值.

记 $\mathscr{X}$ 为值域 X 上一切子集的全体, 并设 $B \in \mathscr{X}$. 当 X 是随机变量 ξ 的值域时, 集合 B 也可以视为某个事件.

在 $(X, \mathscr{X})$ 上考虑由随机变量 ξ 按公式,

$$P_\xi(B) = \mathbf{P}\{\omega : \xi(\omega) \in B\}, \quad B \in \mathscr{X}$$

产生的概率 $P_\xi(\cdot)$. 显然, 这些概率的值完全决定于:

$$P_\xi(x_i) = P\{\omega : \xi(\omega) = x_i\}, \quad x_i \in X.$$

组数 $\{P_\xi(x_1), \cdots, P_\xi(x_m)\}$ 称做随机变量 ξ 的概率分布.

*) 对于示性函数 $I_A(\omega)$, 还使用记号 $I(A)$, I_A. 关于以后常用到的一些性质参见练习题 1.

例 2 设随机变量 ξ 分别以概率 p 和 q 取 1 和 0 两个值, 其中 p 称做 "成功" 的概率, 而 q 称做 "失败" 的概率, 则称 ξ 为伯努利随机变量*). 显然, 对于随机变量 ξ, 有

$$P_\xi(x) = p^x q^{1-x}, \quad x = 0, 1. \tag{1}$$

设 ξ 是以概率

$$P_\xi(x) = C_n^x p^x q^{n-x}, \quad x = 0, 1, \cdots, n \tag{2}$$

取 $0, 1, \cdots, n$ 等 $n+1$ 个可能值, 则称 ξ 为二项随机变量, 或称 ξ 为服从二项分布的随机变量.

注意, 在这些及以后举的许多例子中, 我们不具体说明基本概率空间 $(\Omega, \mathscr{A}, \mathbf{P})$ 的构造, 而只关心随机变量的值及其概率分布.

随机变量 ξ 的构造完全由概率分布 $\{P_\xi(x_i), i = 1, 2, \cdots, m\}$ 描述. 下面引进的分布函数的概念, 提供随机变量构造的等价描述.

定义 2 设 $x \in \mathbf{R}^1$. 函数

$$F_\xi(x) = \mathbf{P}\{\omega : \xi(\omega) \leqslant x\}$$

称做随机变量 ξ 的分布函数.

显然,

$$F_\xi(x) = \sum_{\{i : x_i \leqslant x\}} P_\xi(x_i)$$

并且

$$P_\xi(x_i) = F_\xi(x_i) - F_\xi(x_i-),$$

其中

$$F_\xi(x-) = \lim_{y \uparrow x} F_\xi(y).$$

如果假设 $x_1 < x_2 < \cdots < x_m$, 而 $F_\xi(x_0) = 0$, 则

$$P_\xi(x_i) = F_\xi(x_i) - F_\xi(x_{i-1}), \quad i = 1, \cdots, m.$$

下面的图 5 是二项随机变量 ξ 的 $P_\xi(x)$ 和 $F_\xi(x)$ 的示意图.

直接由定义 2 可见, 分布函数具有下列性质:

(1) $F_\xi(-\infty) = 0$, $F_\xi(+\infty) = 1$.

(2) $F_\xi(x)$ 右连续: $F_\xi(x+) = F_\xi(x)$, 并且是阶梯函数.

*) 在概率论的文献里, 所说的 "伯努利"、"二项"、"泊松"、"高斯" …… 随机变量, 通常称做随机变量服从伯努利、二项、泊松、高斯 …… 分布.

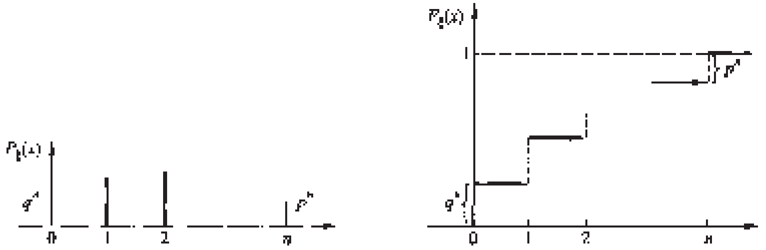


图 5

除随机变量之外, 还常研究随机向量 $\xi = (\xi_1, \cdots, \xi_r)$, 其各分量都是随机变量. 例如, 在研究多项分布时, 对象就是随机向量 $\nu = (\nu_1, \cdots, \nu_r)$, 其中 $\nu_i = \nu_i(\omega)$ 是序列 $\omega = (a_1, \cdots, a_n)$ 中等于 $b_i (i = 1, \cdots, r)$ 的分量的个数.

对于 $x_i \in X_i$ ($X_i (i = 1, \cdots, r)$ 是 ξ_i 的一切可能值的集合), 概率

$$P_\xi(x_1, \cdots, x_r) = \mathbf{P}\{\omega : \xi_1(\omega) = x_1, \cdots, \xi_r(\omega) = x_r\}$$

的全体, 称做随机向量 $\xi = (\xi_1, \cdots, \xi_r)$ 的概率分布, 而函数

$$F_\xi(x_1, \cdots, x_r) = \mathbf{P}\{\omega : \xi_1(\omega) \leqslant x_1, \cdots, \xi_r(\omega) \leqslant x_r\}$$

称做随机向量 $\xi = (\xi_1, \cdots, \xi_r)$ 的分布函数, 其中 $x_i \in \mathbb{R}^1 (i = 1, \cdots, r)$.

对于上面提到的向量 $\nu = (\nu_1, \cdots, \nu_r)$,

$$P_\nu(n_1, \cdots, n_r) = C_n(n_1, \cdots, n_r) p_1^{n_1} \cdots p_r^{n_r}$$

(见 §2 中的 (2) 式).

2. 独立随机变量 设是一组 $\xi_1, \cdots, \xi_r$ 在 $\mathbb{R}^1$ 中 (有限) 集合 X 上取值的随机变量. 记 $\mathscr{X}$ 是 X 中所有子集的代数.

定义 3 称随机变量 $\xi_1, \cdots, \xi_r$ 为 (全体) 独立的, 如果对于任意 $x_1, \cdots, x_r \in X$,

$$\mathbf{P}\{\xi_1 = x_1, \cdots, \xi_r = x_r\} = \mathbf{P}\{\xi_1 = x_1\} \cdots \mathbf{P}\{\xi_r = x_r\},$$

或等价地: 对于任意 $B_1, \cdots, B_r \in \mathscr{X}$,

$$\mathbf{P}\{\xi_1 \in B_1, \cdots, \xi_r \in B_r\} = \mathbf{P}\{\xi_1 \in B_1\} \cdots \mathbf{P}\{\xi_r \in B_r\}.$$

上面讨论的伯努利概型, 就是独立随机变量的一个最简单例子. 具体地, 设

$$\Omega = \{\omega : \omega = (a_1, \cdots, a_n), a_i = 0, 1\}, \quad p(\omega) = p^{\sum a_i}(1 - p)^{n - \sum a_i},$$

并且对于 $\omega = (a_1, \cdots, a_n)$, $\xi_i(\omega) = a_i (i = 1, \cdots, n)$. 那么, 由 §3 的证明, 事件

$$A_1 = \{\omega : a_1 = 1\}, \cdots, A_n = \{\omega : a_n = 1\}$$

独立, 从而 $\xi_1, \cdots, \xi_n$ 独立.

3. 随机变量之和的概率分布 我们以后不止一次地遇到, 作为随机变量 $\xi_1, \cdots, \xi_r$ 的函数的随机变量 $f(\xi_1, \cdots, \xi_r)$ 的概率分布问题. 现在只限于考虑求随机变量之和 $\xi + \eta$ 的分布的情形.

如果 ξ 的值域为 $X = \{x_1, \cdots, x_k\}$, η 的值域为 $Y = \{y_1, \cdots, y_l\}$, 则 $\zeta = \xi + \eta$ 的值域为 $Z = \{z : z = x_i + y_j, i = 1, \cdots, k; j = 1, \cdots, l\}$, 且显然

$$\begin{aligned} P_\zeta(z) = \mathbf{P}\{\zeta = z\} &= \mathbf{P}\{\xi + \eta = z\} \\ &= \sum_{\{(i,j): x_i + y_j = z\}} \mathbf{P}\{\xi = x_i, \eta = y_j\}. \end{aligned}$$

随机变量 ξ 和 η 独立的情形特别重要, 即 $\mathbf{P}\{\xi = x_i, \eta = y_j\} = \mathbf{P}\{\xi = x_i\}\mathbf{P}\{\eta = y_j\}$. 这时, 对于任意 $z \in Z$

$$P_\zeta(z) = \sum_{\{(i,j): x_i + y_j = z\}} P_\xi(x_i) P_\eta(y_j) = \sum_{i=1}^{k} P_\xi(x_i) P_\eta(z - x_i). \tag{3}$$

例如, 若 ξ 和 η 是分别以概率 p 和 q 取 1 和 0 为值的独立伯努利随机变量, 则 $Z = \{0, 1, 2\}$, 而

$$\begin{aligned} P_\zeta(0) &= P_\xi(0)P_\eta(0) = q^2, \\ P_\zeta(1) &= P_\xi(0)P_\eta(1) + P_\xi(1)P_\eta(0) = 2pq, \\ P_\zeta(2) &= P_\xi(1)P_\eta(1) = p^2. \end{aligned}$$

用归纳法容易证明, 若 $\xi_1, \cdots, \xi_n$ 是独立伯努利随机变量, 且 $\mathbf{P}\{\xi_i = 1\} = p$, $\mathbf{P}\{\xi_i = 0\} = q$, 则随机变量 $\zeta = \xi_1 + \cdots + \xi_n$ 服从二项分布

$$P_\zeta(k) = C_n^k p^k q^{n-k} \quad (k = 0, 1, \cdots, n). \tag{4}$$

4. 数学期望 现在研究重要概念, 随机变量的数学期望或均值.

设 $(\Omega, \mathscr{A}, \mathbf{P})$ 是 (有限) 概率空间, 而 $\xi = \xi(\omega)$ 是某一随机变量, 其值域为 $X = \{x_1, \cdots, x_k\}$. 如果设 $A_i = \{\omega : \xi(\omega) = x_i\}$, 则显然 ξ 可以表示为

$$\xi(\omega) = \sum_{i=1}^{k} x_i I(A_i), \tag{5}$$

其中集合 $A_1, \cdots, A_k$ 构成 Ω 的分割 (即 $A_1, \cdots, A_k$ 两两不相容, 且其和是 Ω; 见 §1 第 3 小节).

记 $p_i = \mathbf{P}\{\xi = x_i\}$. 直观上显然, 如果在 n 次独立重复试验中观测随机变量 ξ 的取值, 则取 x_i 的值大致应该出现 $np_i (i = 1, \cdots, k)$ 次. (最好将这一段话与大数定

律节论点对比, 关于大数定律见 §5 和 §12.) 因此, 根据 n 次试验的结果, 计算的该随机变量的 "平均值" 大致为

$$\frac{1}{n}[np_1x_1+\cdots+np_kx_k]=\sum_{i=1}^{k}p_ix_i.$$

这一事实引出下面的定义.

定义 4 实数

$$\mathrm{E}\xi=\sum_{i=1}^{k}x_i\mathrm{P}(A_i) \tag{6}$$

称做随机变量

$$\xi=\sum_{i=1}^{k}x_iI(A_i)$$

的数学期望或平均值. 由于 $A_i=\{\omega:\xi(\omega)=x_i\}$, 而 $P_\xi(x_i)=\mathrm{P}(A_i)$, 则

$$\mathrm{E}\xi=\sum_{i=1}^{k}x_iP_\xi(x_i). \tag{7}$$

注意到分布函数的定义 $F_\xi=F_\xi(x)$, 并记

$$\Delta F_\xi(x)=F_\xi(x)-F_\xi(x-),$$

得 $P_\xi(x_i)=\Delta F_\xi(x_i)$, 从而

$$\mathrm{E}\xi=\sum_{i=1}^{k}x_i\Delta F_\xi(x_i). \tag{8}$$

在研究数学期望的性质之前, 常需要随机变量 ξ 的各种不同表达形式, 例如

$$\xi(\omega)=\sum_{j=1}^{l}x_j'I(B_j),$$

其中 $B_1+\cdots+B_l=\Omega$, 但是一般在 x_j' 之中可能有相同的值. 这时, 可以按上面的公式计算数学期望, 而不需要首先变换成所有的 x_i 值两两不等的 (5) 式. 事实上,

$$\sum_{\{j:x_j'=x_i\}}x_j'\mathrm{P}(B_j)=x_i\sum_{\{j:x_j'=x_i\}}\mathrm{P}(B_j)=x_i\mathrm{P}(A_i),$$

于是

$$\sum_{j=1}^{l}x_j'P(B_j)=\sum_{i=1}^{k}x_iP(A_i).$$

5. 数学期望的基本性质 现在列举数学期望的基本性质:

1) 若 $\xi\geqslant 0$, 则 $\mathrm{E}\xi\geqslant 0$.

2) $\mathrm{E}(a\xi+b\eta)=a\mathrm{E}\xi+b\mathrm{E}\eta$, 其中 a,b 是常数.

3) 若 $\xi\geqslant\eta$, 则 $\mathrm{E}\xi\geqslant\mathrm{E}\eta$.

4) $|\mathrm{E}\xi|\leqslant\mathrm{E}|\xi|$.

5) 若 ξ 和 η 独立, 则 $\mathrm{E}\xi\eta=\mathrm{E}\xi\cdot\mathrm{E}\eta$.

6) $(\mathrm{E}|\xi\eta|)^2\leqslant\mathrm{E}\xi^2\cdot\mathrm{E}\eta^2$ (柯西－布尼亚科夫斯基不等式, 亦称柯西－施瓦茨不等式, 或施瓦茨不等式).

7) 若 $\xi=I(A)$, 则 $\mathrm{E}\xi=\mathrm{P}(A)$.

性质 1) 和 7) 显然. 为证明性质 2), 设

$$\xi=\sum x_iI(A_i),\quad \eta=\sum y_jI(B_j),$$

则

$$\begin{aligned}
a\xi+b\eta&=a\sum_{i,j}x_iI(A_i\cap B_j)+b\sum_{i,j}y_jI(A_i\cap B_j)\\
&=\sum_{i,j}(ax_i+by_j)I(A_i\cap B_j);\\
\mathrm{E}(a\xi+b\eta)&=\sum_{i,j}(ax_i+by_j)\mathrm{P}(A_i\cap B_j)\\
&=\sum_i ax_i\mathrm{P}(A_i)+\sum_j by_j\mathrm{P}(B_j)\\
&=a\sum_i x_i\mathrm{P}(A_i)+b\sum_j y_j\mathrm{P}(B_j)=a\mathrm{E}\xi+b\mathrm{E}\eta.
\end{aligned}$$

由性质 1) 和 2) 可以证明 3). 因为

$$|\mathrm{E}\xi|=\left|\sum_i x_i\mathrm{P}(A_i)\right|\leqslant\sum_i|x_i|\mathrm{P}(A_i)=\mathrm{E}|\xi|,$$

所以性质 4) 显然. 为证明性质 5), 只需注意到

$$\begin{aligned}
\mathrm{E}\xi\eta&=\mathrm{E}\left(\sum_i x_iI(A_i)\right)\left(\sum_j y_jI(B_j)\right)=\mathrm{E}\sum_{i,j}x_iy_jI(A_i\cap B_j)\\
&=\sum_{i,j}x_iy_j\mathrm{P}(A_i\cap B_j)=\sum_{i,j}x_iy_j\mathrm{P}(A_i)\mathrm{P}(B_j)\\
&=\left(\sum_i x_i\mathrm{P}(A_i)\right)\left(\sum_j y_j\mathrm{P}(B_j)\right)=\mathrm{E}\xi\cdot\mathrm{E}\eta.
\end{aligned}$$

其中在证明过程中用到: 对于独立随机变量 ξ 和 η, 事件

$$A_i = \{\omega : \xi(\omega) = x_i\} \text{ 和 } B_j = \{\omega : \eta(\omega) = y_j\}$$

独立: $\mathbf{P}(A_i \bigcap B_j) = \mathbf{P}(A_i)\mathbf{P}(B_j)$.

为证明性质 6), 注意到

$$\xi^2 = \sum_{i=1}^{l} x_i^2 I(A_i), \quad \eta^2 = \sum_{j=1}^{k} y_j^2 I(B_j)$$

和

$$\mathbf{E}\xi^2 = \sum_{i=1}^{l} x_i^2 \mathbf{P}(A_i), \quad \mathbf{E}\eta^2 = \sum_{j=1}^{k} y_j^2 \mathbf{P}(B_j).$$

设 $\mathbf{E}\xi^2 > 0, \mathbf{E}\eta^2 > 0$. 记

$$\tilde{\xi} = \frac{\xi}{\sqrt{\mathbf{E}\xi^2}}, \quad \tilde{\eta} = \frac{\eta}{\sqrt{\mathbf{E}\eta^2}}.$$

由 $2|\tilde{\xi}\tilde{\eta}| \leqslant \tilde{\xi}^2 + \tilde{\eta}^2$, 可见 $2\mathbf{E}|\tilde{\xi}\tilde{\eta}| \leqslant \mathbf{E}\tilde{\xi}^2 + \mathbf{E}\tilde{\eta}^2 = 2$. 因此

$$\mathbf{E}|\tilde{\xi}\tilde{\eta}| \leqslant 1, \quad (\mathbf{E}|\xi\eta|)^2 \leqslant \mathbf{E}\xi^2 \times \mathbf{E}\eta^2.$$

假如 $\mathbf{E}\xi^2 = 0$, 则

$$\sum_i x_i^2 \mathbf{P}\{A_i\} = 0,$$

从而 0 是 ξ 的可能值, 并且

$$\mathbf{P}\{\omega : \xi(\omega) = 0\} = 1.$$

因此, 如果 $\mathbf{E}\xi^2$ 或 $\mathbf{E}\eta^2$ 之一等于 0, 则显然 $\mathbf{E}|\xi\eta| = 0$. 于是柯西—布尼亚科夫斯基不等式仍然成立.

注 性质 5) 明显地可以推广到任意有穷个随机变量: 若 $\xi_1, \cdots, \xi_r$ 独立, 则

$$\mathbf{E}\xi_1 \cdots \xi_r = \mathbf{E}\xi_1 \cdots \mathbf{E}\xi_r.$$

这里可以仿照 $n = 2$ 的情形证明, 亦可用归纳法证明.

例 3 设 ξ 是伯努利随机变量, 以概率 p 和 q 取 1 和 0 为值, 则

$$\mathbf{E}\xi = 1 \times \mathbf{P}\{\xi = 1\} + 0 \times \mathbf{P}\{\xi = 0\} = p.$$

例 4 设 $\xi_1, \cdots, \xi_n$ 是 n 个伯努利随机变量, 以概率 $\mathbf{P}\{\xi_i = 1\} = p$ 和 $\mathbf{P}\{\xi_i = 0\} = q, p + q = 1$ 取 1 和 0 为值, 则对于

$$S_n = \xi_1 + \cdots + \xi_n,$$

其数学期望为

$$\mathbf{E}S_n = np.$$

这里有另一种求法. 易见, 如果假设 $\xi_1, \cdots, \xi_n$ 独立伯努利随机变量, 则 $\mathbf{E}S_n$ 不变. 在此条件下, 根据 (4) 式, 有

$$\mathbf{P}\{S_n = k\} = C_n^k p^k q^{n-k} \quad (k = 0, 1, \cdots, n).$$

因此,

$$\begin{aligned}
\mathbf{E}S_n &= \sum_{k=0}^{n} k\mathbf{P}\{S_n = k\} = \sum_{k=0}^{n} k C_n^k p^k q^{n-k} = \sum_{k=0}^{n} \frac{k\,n!}{k!(n-k)!} p^k q^{n-k} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)![(n-1)-(k-1)]!} p^{k-1} q^{(n-1)-(k-1)} \\
&= np \sum_{l=0}^{n-1} \frac{(n-1)!}{l![(n-1)-l]!} p^l q^{(n-1)-l} = np.
\end{aligned}$$

其实, 用前一种方法比用后一种方法得到结果更快一些.

6. 随机变量函数的数学期望 设 $\xi = \sum x_i I(A_i)$, 其中 $A_i = \{\omega : \xi(\omega) = x_i\}$, 而 $\varphi = \varphi(\xi(\omega))$ 是 $\xi(\omega)$ 的某一函数. 如果 $B_j = \{\omega : \varphi(\xi(\omega)) = y_j\}$, 则 (设 $I_{B_j}(\omega) = I(B_j)$)

$$\varphi(\xi(\omega)) = \sum_j y_j I_{B_j}(\omega),$$

从而

$$\mathbf{E}\varphi = \sum_j y_j \mathbf{P}(B_j) = \sum_j y_j P_\varphi(y_j). \tag{9}$$

同样显然 (设 $I_{A_i}(\omega) = I(A_i)$)

$$\varphi(\xi(\omega)) = \sum_i \varphi(x_i) I_{A_i}(\omega).$$

于是, 为求 $\varphi = \varphi(\xi(\omega))$ 的数学期望, 既可以利用 (9) 式, 也可以利用下面的公式

$$\mathbf{E}\varphi(\xi(\omega)) = \sum_i \varphi(x_i) P_\xi(x_i).$$

7. 方差和标准差 随机变量 ξ 的方差和标准差, 表征 ξ 取值的散布程度, 是十分重要的概念.

定义 5 称

$$\mathbf{D}\xi = \mathbf{E}(\xi - \mathbf{E}\xi)^2$$

为随机变量的方差 (记作 $\mathbf{D}\xi$). 称 $\sigma = \sqrt{\mathbf{D}\xi}$ 为标准差.

由于

$$\mathbf{E}(\xi-\mathbf{E}\xi)^2=\mathbf{E}[\xi^2-2\xi\mathbf{E}\xi+(\mathbf{E}\xi)^2]=\mathbf{E}\xi^2-(\mathbf{E}\xi)^2,$$

可见

$$\mathbf{D}\xi=\mathbf{E}\xi^2-(\mathbf{E}\xi)^2.$$

显然 $\mathbf{D}\xi\geqslant 0$. 由方差的定义, 可见对于任意常数 a, b,

$$\mathbf{D}(a+b\xi)=b^2\mathbf{D}\xi;$$

特别 $\mathbf{D}a=0, \mathbf{D}(b\xi)=b^2\mathbf{D}\xi$.

对于二随机变量 ξ 和 η,

$$\begin{aligned}\mathbf{D}(\xi+\eta)&=\mathbf{E}[(\xi-\mathbf{E}\xi)+(\eta-\mathbf{E}\eta)]^2\\&=\mathbf{D}\xi+\mathbf{D}\eta+2\mathbf{E}(\xi-\mathbf{E}\xi)(\eta-\mathbf{E}\eta).\end{aligned}$$

记

$$\operatorname{cov}(\xi,\eta)=\mathbf{E}(\xi-\mathbf{E}\xi)(\eta-\mathbf{E}\eta),$$

称做随机变量 ξ 和 η 的协方差. 如果 $\mathbf{D}\xi>0, \mathbf{D}\eta>0$, 则

$$\rho(\xi,\eta)=\frac{\operatorname{cov}(\xi,\eta)}{\sqrt{\mathbf{D}\xi\times\mathbf{D}\eta}}$$

称做随机变量 ξ 和 η 的相关系数. 不难证明 (见练习题 7), 若 $\rho(\xi,\eta)=\pm1$, 则随机变量 ξ 和 η 线性相关:

$$\eta=a\xi+b,$$

其中当 $\rho(\xi,\eta)=1$ 时 $a>0$; 当 $\rho(\xi,\eta)=-1$ 时 $a<0$.

立即可以指出, 若 ξ 和 η 独立, 则 $\xi-\mathbf{E}\xi$ 和 $\eta-\mathbf{E}\eta$ 独立, 因此根据数学期望的性质 5), 有

$$\operatorname{cov}(\xi,\eta)=\mathbf{E}(\xi-\mathbf{E}\xi)\times\mathbf{E}(\eta-\mathbf{E}\eta)=0.$$

由协方差的定义, 可见

$$\mathbf{D}(\xi+\eta)=\mathbf{D}\xi+\mathbf{D}\eta+2\operatorname{cov}(\xi,\eta). \tag{10}$$

如果 ξ 与 η 独立, 则和 $\xi+\eta$ 的方差等于方差之和:

$$\mathbf{D}(\xi+\eta)=\mathbf{D}\xi+\mathbf{D}\eta. \tag{11}$$

虽然式 (11) 可以由式 (10) 导出的, 然而 (11) 式在比 "ξ 和 η 独立" 较弱的条件下仍然成立. 具体地说, 只要假设 "ξ 和 η 不相关", 即假设 $\operatorname{cov}(\xi,\eta)=0$ 就可以了.

注 由 ξ 和 η 不相关, 一般得不出 ξ 和 η 独立. 看下面的例子. 假设随机变量 α 以概率 1/3 分别取 $0, \pi/2, \pi$ 为值, 则 $\xi=\sin\alpha$ 和 $\eta=\cos\alpha$ 不相关, 然而 ξ 和 η 不但 (关于 $\mathbf{P}$) 不独立:

$$\mathbf{P}\{\xi=1,\eta=1\}=0\neq 1/9=\mathbf{P}\{\xi=1\}\mathbf{P}\{\eta=1\},$$

而且 ξ 和 η 之间有函数关系: $\xi^2+\eta^2=1$.

性质 (10) 和 (11) 明显可以推广到任意个随机变量 $\xi_1,\cdots,\xi_n$ 的情形:

$$\mathbf{D}\left(\sum_{i=1}^n\xi_i\right)=\sum_{i=1}^n\mathbf{D}\xi_i+2\sum_{i>j}\operatorname{cov}(\xi_i,\xi_j). \tag{12}$$

特别, 若随机变量 $\xi_1,\cdots,\xi_n$ 两两独立 (实际上, 只需要求它们两两不相关), 则

$$\mathbf{D}\left(\sum_{i=1}^n\xi_i\right)=\sum_{i=1}^n\mathbf{D}\xi_i. \tag{13}$$

例 5 设 ξ 是伯努利随机变量, 以概率 p 和 q 取 1 和 0 为值, 则

$$\mathbf{D}\xi=\mathbf{E}(\xi-\mathbf{E}\xi)^2=\mathbf{E}(\xi-p)^2=(1-p)^2p+p^2q=pq.$$

由此可见, 如果 $\xi_1,\cdots,\xi_n$ 是独立同分布的伯努利随机变量序列, 且 $S_n=\xi_1+\cdots+\xi_n$, 则

$$\mathbf{D}S_n=npq. \tag{14}$$

8. 最优线性估计 考虑两个随机变量 ξ 和 η. 假设只对随机变量 ξ 进行观测. 如果随机变量 ξ 和 η 相关, 则可以预期, 已知 ξ 的值可以对未观测随机变量 η 的值, 作出某种判断.

我们把 ξ 的任何一个函数 $f=f(\xi)$ 称做 η 的一个估计量. 称估计量 $f^*=f^*(\xi)$ 为在均方意义下最优的, 如果

$$\mathbf{E}(\eta-f^*(\xi))^2=\inf_f\mathbf{E}(\eta-f(\xi))^2.$$

现在讨论, 如何在线性估计 $\lambda(\xi)=a+b\xi$ 类中求最优估计. 为此考虑函数

$$g(a,b)=\mathbf{E}[\eta-(a+b\xi)]^2.$$

将 $g(a,b)$ 分别对 a 和 b 求偏导数, 得

$$\frac{\partial g(a,b)}{\partial a}=-2\mathbf{E}[\eta-(a+b\xi)],$$

$$\frac{\partial g(a,b)}{\partial b}=-2\mathbf{E}[\eta-(a+b\xi)\xi].$$

令所得偏导数等于 0, 可以求出 λ 均方线性估计: $\lambda^* = a^* + b^*\xi$, 其中

$$a^* = \mathbf{E}\eta - b^*\mathbf{E}\xi,\quad b^* = \frac{\operatorname{cov}(\xi,\eta)}{\mathbf{D}\xi}. \tag{15}$$

即

$$\lambda^*(\xi) = \mathbf{E}\eta + \frac{\operatorname{cov}(\xi,\eta)}{\mathbf{D}\xi}(\xi - \mathbf{E}\xi). \tag{16}$$

称 $\mathbf{E}[\eta - \lambda^*(\xi)]^2$ 为估计值的均方误差. 经过简单的计算可见, 该误差等于

$$\Delta^* = \mathbf{E}[\eta - \lambda^*(\xi)]^2 = \mathbf{D}\eta - \frac{\operatorname{cov}^2(\xi,\eta)}{\mathbf{D}\xi} = \mathbf{D}\eta \times [1 - \rho^2(\xi,\eta)]. \tag{17}$$

这样, ξ 和 η 之间的相关系数 (绝对值) 越大, 均方误差的估计值 Δ^* 就越小. 特别, 如果 $|\rho(\xi,\eta)| = 1$, 则 $\Delta^* = 0$ (与练习题 7 的结果比较). 如果随机变量 ξ 和 η 不相关 (即 $\rho(\xi,\eta) = 0$), 则 $\lambda^*(\xi) = \mathbf{E}\eta$. 于是, 在随机变量 ξ 和 η 不相关的情形下, 根据 ξ 对 η 的估计就是 $\mathbf{E}\eta$ (与练习题 4 的结果比较).

9. 练习题

1. 验证示性函数 $I_A = I_A(\omega)$ 的下列性质:

$$I_\varnothing = 0,\quad I_\Omega = 1,\quad I_{\overline{A}} = 1 - I_A,$$
$$I_{AB} = I_A \times I_B,\quad I_{A\cup B} = I_A + I_B - I_{AB},$$
$$I_{A\setminus B} = I_A(1 - I_B),\quad I_{A\triangle B} = (I_A - I_B)^2 = I_A + I_B \pmod 2,$$
$$I_{E_1} = 1 - \prod_{i=1}^{n}(1 - I_{A_i}),\quad I_{E_2} = \prod_{i=1}^{n}(1 - I_{A_i}),\quad I_{E_3} = \sum_{i=1}^{n} I_{A_i},$$

其中 $A\triangle B$ 称做集合 A 与 B 的对称差: $A\triangle B = (A\setminus B)\bigcup(B\setminus A)$, 而

$$E_1 = \bigcup_{i=1}^{n} A_i,\quad E_2 = \overline{\bigcup_{i=1}^{n} A_i},\quad E_3 = \sum_{i=1}^{n} A_i.$$

2. 设 $\xi_1, \cdots, \xi_n$ 是独立随机变量, 且

$$\xi_{\min} = \min\{\xi_1, \cdots, \xi_n\},\quad \xi_{\max} = \max\{\xi_1, \cdots, \xi_n\}.$$

证明:

$$\mathbf{P}\{\xi_{\min} \geqslant x\} = \prod_{i=1}^{n}\mathbf{P}\{\xi_i \geqslant x\},\quad \mathbf{P}\{\xi_{\max} < x\} = \prod_{i=1}^{n}\mathbf{P}\{\xi_i < x\}.$$

3. 设 $\xi_1, \cdots, \xi_n$ 是独立伯努利随机变量, 且

$$\mathbf{P}\{\xi_i = 0\} = 1 - \lambda_i\Delta,$$
$$\mathbf{P}\{\xi_i = 1\} = \lambda_i\Delta,$$

其中 $\Delta > 0, \lambda_i > 0$, 而 Δ 是较小的数. 证明:

$$\mathbf{P}\{\xi_1 + \cdots + \xi_n = 1\} = \left(\sum_{i=1}^{n}\lambda_i\right)\Delta + O(\Delta^2),$$
$$\mathbf{P}\{\xi_1 + \cdots + \xi_n > 1\} = O(\Delta^2).$$

4. 证明, 当 $a = \mathbf{E}\xi$ 时 $\mathbf{E}(\xi - a)^2$ 达到下确界

$$\inf_{-\infty<a<\infty}\mathbf{E}(\xi - a)^2,\quad 即\quad \inf_{-\infty<a<\infty}\mathbf{E}(\xi - a)^2 = \mathbf{D}\xi.$$

5. 设 $F_\xi(x)$ 是随机变量 ξ 的分布函数, 而 m_e 是 $F_\xi(x)$ 的中位数, 即下列条件的点:

$$F_\xi(m_e-) \leqslant \frac{1}{2} \leqslant F_\xi(m_e).$$

证明

$$\inf_{-\infty<a<\infty}\mathbf{E}|\xi - a| = \mathbf{E}|\xi - m_e|.$$

6. 设 $P_\xi(x) = \mathbf{P}\{\xi = x\}, F_\xi(x) = \mathbf{P}\{\xi \leqslant x\}$. 证明:

(1) 对于 $a > 0$, $-\infty < b < \infty$, 有

$$P_{a\xi+b}(x) = P_\xi\left(\frac{x-b}{a}\right),$$
$$F_{a\xi+b}(x) = F_\xi\left(\frac{x-b}{a}\right).$$

(2) 如果 $y \geqslant 0$, 则

$$F_{\xi^2}(y) = F_\xi(+\sqrt{y}) - F_\xi(-\sqrt{y}) + P_\xi(-\sqrt{y}).$$

(3) 设 $\xi^+ = \max\{\xi, 0\}$, 则

$$F_{\xi^+}(x) = \begin{cases} 0, & 若\ x < 0, \\ F_\xi(x), & 若\ x = 0, \\ F_\xi(x), & 若\ x > 0. \end{cases}$$

7. 设随机变量 ξ 和 η 的方差 $\mathbf{D}\xi > 0, \mathbf{D}\eta > 0$, 它们的相关系数为 $\rho = \rho(\xi,\eta)$. 证明:

(1) $|\rho| \leqslant 1$.

(2) 若 $|\rho| = 1$, 则存在常数 a 和 b, 使 $\eta = a\xi + b$, 并且当 $\rho = 1$ 时

$$\frac{\eta - \mathbf{E}\eta}{\sqrt{\mathbf{D}\eta}} = \frac{\xi - \mathbf{E}\xi}{\sqrt{\mathbf{D}\xi}},$$

(即 $a > 0$), 当 $\rho = -1$ 时

$$\frac{\eta - \mathbf{E}\eta}{\sqrt{\mathbf{D}\eta}} = -\frac{\xi - \mathbf{E}\xi}{\sqrt{\mathbf{D}\xi}}.$$

(即 $a<0$).

8. 假设对于随机变量 ξ 和 η, $\mathbf{E}\xi=\mathbf{E}\eta=0$, $\mathbf{D}\xi=\mathbf{D}\eta=1$, 而 ξ 和 η 的相关系数为 $\rho=\rho(\xi,\eta)$. 证明:

$$\mathbf{E}\max\{\xi^2,\eta^2\}\leqslant 1+\sqrt{1-\rho^2}.$$

9. 利用等式

$$I_{B_0}=\prod_{i=1}^{n}(1-I_{A_i}),\quad 其中\quad B_0=\overline{\bigcup_{i=1}^{n}A_i},$$

证明 §1 中练习题 5 的公式:

$$\mathbf{P}(B_0)=1-S_1+S_2-\cdots+S_n.$$

10. 设 $\xi_1,\cdots,\xi_n$ 是独立随机变量, $\varphi_1=\varphi_1(\xi_1,\cdots,\xi_k)$ 和 $\varphi_2=\varphi_2(\xi_{k+1},\cdots,\xi_n)$ 分别是 $(\xi_1,\cdots,\xi_k)$ 和 $(\xi_{k+1},\cdots,\xi_n)$ 的函数. 证明 φ_1 和 φ_2 独立.

11. 证明随机变量 $\xi_1,\cdots,\xi_n$ 独立, 当且仅当对于一切 $x_1,\cdots,x_n$,

$$F_{\xi_1,\cdots,\xi_n}(x_1,\cdots,x_n)=F_{\xi_1}(x_1)\cdots F_{\xi_n}(x_n),$$

其中 $F_{\xi_1,\cdots,\xi_n}(x_1,\cdots,x_n)=\mathbf{P}\{\xi_1\leqslant x_1,\cdots,\xi_n\leqslant x_n\}$.

12. 证明随机变量 ξ 与自己独立 (即 ξ 与 ξ 独立), 当且仅当 $\xi\equiv\text{const.}$ (常数).

13. 问随机变量 ξ 满足何条件时, ξ 与 $\sin\xi$ 独立?

14. 设 ξ 和 η 是独立随机变量, 且 $\eta\neq 0$. 通过概率 $P_\xi(x)$ 和 $P_\eta(y)$ 表示概率:

$$\mathbf{P}\{\xi\eta\leqslant z\}\quad 和\quad \mathbf{P}\left\{\frac{\xi}{\eta}\leqslant z\right\}.$$

15. 设随机变量 ξ,η,ζ 满足条件: $|\xi|\leqslant 1$, $|\eta|\leqslant 1$, $|\zeta|\leqslant 1$, 证明贝尔 (A. C. Bell) 不等式:

$$|\mathbf{E}\xi\zeta-\mathbf{E}\eta\zeta|\leqslant 1-\mathbf{E}\xi\eta.$$

(例如, 见 [136])

16. 向 n 个箱子中独立地掷 k 个球. 假设每个球落入各箱的概率都等于 $1/n$, 求非空箱子个数的数学期望.

§5. 伯努利概型 I. 大数定律

1. 伯努利概型 以上的定义 "三对象"

$$(\Omega,\mathscr{A},\mathbf{P}),\ 其中\ \Omega=\{\omega:\omega=(a_1,\cdots,a_n),a_i=0,1\},$$
$$\mathscr{A}=\{A:A\subseteq\Omega\},\quad P(\{\omega\})=p^{\sum a_i}(1-p)^{n-\sum a_i}(=p(\omega)),$$

称做伯努利概型, 全称为 "有两种结局的 n 次独立试验的概率模型". 在这一节和下一节, 我们将研究 (在下面所指的意义下) 伯努利概型的某些性质. 用与之相联系的, 随机变量和事件概率的术语, 引进这些性质较为适宜.

引进随机变量 $\xi_1,\cdots,\xi_n$, 其中 $\xi_i=\xi_i(\omega)=a_i$, $i=1,\cdots,n$, 而 $\omega=(a_1,\cdots,a_n)$. 已经熟知, 伯努利随机变量 $\xi_i=\xi_i(\omega)=a_i$, $i=1,\cdots,n$ 独立而且同分布:

$$\mathbf{P}\{\xi_i=1\}=p,\ \mathbf{P}\{\xi_i=0\}=1-p=q,\quad i=1,\cdots,n.$$

随机变量 ξ_i 表示在第 i 步 (或时刻 i) 的试验结果.

设 $S_0(\omega)=0$,

$$S_k=\xi_1+\cdots+\xi_k,\quad k=1,\cdots,n.$$

我们已经 (在 §4 例 4) 证明, $\mathbf{E}S_n=np$. 从而

$$\mathbf{E}\frac{S_n}{n}=p. \tag{1}$$

即 "成功" 频率 S_n/n 的平均值等于成功的概率 p. 由此自然产生一个问题: "成功" 频率 S_n/n 对成功概率 p 的 (绝对) 偏差的大小如何?

我们首先指出, 对于充分小的 $\varepsilon>0$ 和甚至很大的 n, 也不能指望对于任意 ω, 频率 S_n/n 对成功概率 p 的 (绝对) 偏差都小于 ε, 即不能指望对于任意 ω, 不等式

$$\left|\frac{S_n(\omega)}{n}-p\right|\leqslant\varepsilon,\quad \omega\in\Omega \tag{2}$$

都成立.

事实上, 对于 $0<p<1$, 由

$$\mathbf{P}\left\{\frac{S_n}{n}=1\right\}=\{\xi_1=1,\cdots,\xi_n=1\}=p^n,$$
$$\mathbf{P}\left\{\frac{S_n}{n}=0\right\}=\{\xi_1=0,\cdots,\xi_n=0\}=q^n,$$

可见, 对于充分小的 $\varepsilon>0$ 不等式 (2) 并不成立.

不过, 我们指出当 n 很大时, 事件

$$\left\{\frac{S_n}{n}=1\right\}\quad 和\quad \left\{\frac{S_n}{n}=0\right\}$$

的概率都较小. 因此, 自然想到, 当 n 充分大时使

$$\left|\frac{S_n}{n}-p\right|>\varepsilon$$

成立的 "结局 ω 的全体" 的概率也较小. 因此, 设法估计事件

$$\left\{\omega:\left|\frac{S_n(\omega)}{n}-p\right|>\varepsilon\right\}$$

的概率. 为此我们运用如下切比雪夫不等式.①

切比雪夫 (П. Л. Чебышёв) 不等式　*设 $(\Omega, \mathscr{A}, \mathbf{P})$ 是某一概率空间, $\xi = \xi(\omega)$ 是非负随机变量. 那么, 对任意 $\varepsilon > 0$,*

$$\mathbf{P}\{\xi \geqslant \varepsilon\} \leqslant \frac{\mathbf{E}\xi}{\varepsilon}. \tag{3}$$

证明　注意到,

$$\xi = \xi I(\xi \geqslant \varepsilon) + \xi I(\xi < \varepsilon) \geqslant \xi I(\xi \geqslant \varepsilon) \geqslant \varepsilon I(\xi \geqslant \varepsilon),$$

其中 $I(A)$ 是集合 A 的示性函数.

于是, 根据数学期望的性质

$$\mathbf{E}\xi \geqslant \varepsilon \mathbf{E} I(\xi \geqslant \varepsilon) = \varepsilon \mathbf{P}\{\xi \geqslant \varepsilon\},$$

从而 (3) 式得证.

系　*设 ξ 是任意随机变量, 则对任意 $\varepsilon > 0$,*

$$\mathbf{P}\{|\xi| \geqslant \varepsilon\} \leqslant \frac{\mathbf{E}|\xi|}{\varepsilon},$$

$$\mathbf{P}\{|\xi| \geqslant \varepsilon\} = \mathbf{P}\{\xi^2 \geqslant \varepsilon^2\} \leqslant \frac{\mathbf{E}\xi^2}{\varepsilon^2}, \tag{4}$$

$$\mathbf{P}\{|\xi - \mathbf{E}\xi| \geqslant \varepsilon\} \leqslant \frac{\mathbf{D}\xi}{\varepsilon^2}.$$

利用最后一个不等式, 设 $\xi = S_n/n$, 则由 §4 (14) 式, 有

$$\mathbf{P}\left\{\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right\} \leqslant \frac{\mathbf{D}\left(\dfrac{S_n}{n}\right)}{\varepsilon^2} = \frac{\mathbf{D}S_n}{n^2\varepsilon^2} = \frac{npq}{n^2\varepsilon^2} = \frac{pq}{n\varepsilon^2}.$$

于是

$$\mathbf{P}\left\{\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right\} \leqslant \frac{pq}{n\varepsilon^2} \leqslant \frac{1}{4n\varepsilon^2}. \tag{5}$$

由此可见, 当 n 充分大时, "成功" 频率 S_n/n 对 "成功" 概率 p 的 (绝对) 偏差大于 ε 的概率充分小.

对于一切 n 和 $1 \leqslant k \leqslant n$, 记

$$P_n(k) = C_n^k p^k q^{n-k},$$

则

$$\mathbf{P}\left\{\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right\} = \sum_{\{k: |\frac{k}{n} - p| \geqslant \varepsilon\}} P_n(k).$$

① П. Л. Чебышёв 是俄罗斯数学家. 按俄语语音应译为 "切贝绍夫" (见《俄语姓名译名手册》, 商务印书馆, 1982;《新俄汉数学词汇》, 科学出版社, 1958). 英语文献一般译为 P. L. Chebyshev. 我国有的文献按英语译音, 译为 "切比雪夫" 或 "契比雪夫".——译者

实际上证明了 (5) 式:

$$\sum_{\{k: |\frac{k}{n} - p| \geqslant \varepsilon\}} P_n(k) \leqslant \frac{pq}{n\varepsilon^2} \leqslant \frac{1}{4n\varepsilon^2}. \tag{6}$$

即我们用概率的方法证明了不等式 (5). 注意, 假如不用概率的方法, 而用分析的方法, 也可以证明此不等式.

由 (6) 可见

$$\sum_{\{k: |\frac{k}{n} - p| \geqslant \varepsilon\}} P_n(k) \to 0 \, (n \to \infty). \tag{7}$$

该命题可以用图形作如下解释. 图 6 是二项分布 $\{P_n(k); 0 \leqslant k \leqslant n\} (p = 1/2)$ 的示意图.

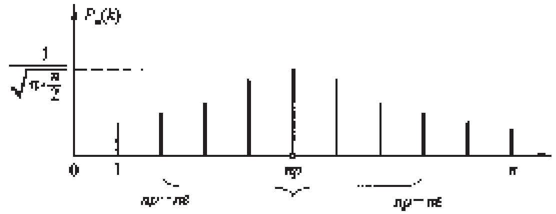


图 6　($m_1 = np - n\varepsilon$, $m_2 = np + n\varepsilon$)

由图 6 可见: 概率 $P_n(k)$ 在 $k = np$ 处达到最大值 P_m, 其中

$$P_m = \frac{1}{\sqrt{\pi \cdot \frac{n}{2}}}.$$

图 6 显示: 若将概率 $P_n(k)$ 对 k 求和, 则对于 $np - n\varepsilon \leqslant k \leqslant np + n\varepsilon$, 概率接近 1.

我们把一系列随机变量 $S_0, S_1, \cdots, S_n$ 视为某游动的质点的轨道. 那么, 对 (7) 式可以作如下解释.

过原点引 3 条直线: $k(p - \varepsilon)$, kp, $k(p + \varepsilon)$. 那么, 质点轨道总的趋势是沿直线运动. 对于任意 $\varepsilon > 0$, 可以断定, 对于充分大的 n, 表示质点在时刻 n 位置的点 S_n 位于区间 $[n(p - \varepsilon), n(p + \varepsilon)]$ 上 (见图 7).

命题 (7) 可以表示为:

$$\mathbf{P}\left\{\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right\} \to 0, \quad n \to \infty. \tag{8}$$

不过需要指出, 这里存在一定的细节. 问题在于, 假如概率 $\mathbf{P}$ 在某空间 $(\Omega, \mathscr{A})$ 上, 空间 $(\Omega, \mathscr{A})$ 上定义了独立无穷伯努利随机变量序列 $\xi_1, \xi_2, \cdots$, 则上面的写法是

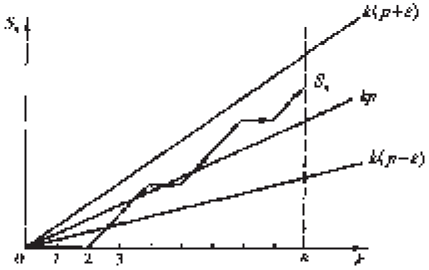


图 7

完全正确的: 确实可以建立这样的对象. 从而给 (8) 式赋予严格的概率意义 (见第二章 §9 定理 1 系 1). 现在, 如果希望赋予 (7) 式分析命题的含义, 则用初等概率的语言, 可以说明下述事实.

设 $(\Omega^{(n)}, \mathscr{A}^{(n)}, \mathbf{P}^{(n)}), n \geqslant 1$, 是伯努利概型序列:

$$\Omega^{(n)} = \{\omega^{(n)} : \omega^{(n)} = (a_1^{(n)}, \cdots, a_n^{(n)}), a_i^{(n)} = 0, 1\},$$
$$\mathscr{A}^{(n)} = \{A : A \subseteq \Omega^{(n)}\},$$
$$\mathbf{P}^{(n)}(\{\omega^{(n)}\}) = p^{\sum a_i^{(n)}} q^{n - \sum a_i^{(n)}},$$

而

$$S_k^{(n)}(\omega^{(n)}) = \xi_1^{(n)}(\omega^{(n)}) + \cdots + \xi_k^{(n)}(\omega^{(n)}),$$

其中, 对于 $n \geqslant 1$, $\xi_1^{(n)}, \cdots, \xi_n^{(n)}$ 是独立同分布伯努利随机变量序列. 那么

$$\mathbf{P}^{(n)}\left\{\omega^{(n)} : \left|\frac{S_n^{(n)}(\omega^{(n)})}{n} - p\right| \geqslant \varepsilon\right\} = \sum_{\{k : |\frac{k}{n} - p| \geqslant \varepsilon\}} P_n(k) \to 0, \quad (n \to \infty). \tag{9}$$

式 (7) ~ (9) 的命题称做

“伯努利大数定律”

应该指出, J. 伯努利证明的恰好是命题 (7). 其证明非常严格, 利用了二项分布 “尾部” 概率的估计, 即对于满足 $|k/n - p| \geqslant \varepsilon$ 的 k 估计概率 $P_n(k)$. 对于充分大的 n, 二项分布 “尾部” 概率

$$\sum_{\{k : |\frac{k}{n} - p| \geqslant \varepsilon\}} P_n(k)$$

的直接计算问题相当繁杂, 况且所得 “频率 S_n/n 对概率 p 绝对偏差小于 ε” 的概率估计式很难实际应用. 因此, 对于任意 p, 棣莫弗和拉普拉斯所创造的概率 $P_n(k)$ 的

渐近公式特别重要. 不仅重新证明了大数定律, 而且得到了更精确的所谓局部及积分极限定理. 该定理的实质在于, 对于充分大 n 的和至少满足 $k \sim np$ 的 k, 有

$$P_n(k) \sim \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}},$$

而

$$\sum_{\{k : |\frac{k}{n} - p| \leqslant \varepsilon\}} P_n(k) \sim \frac{1}{\sqrt{2\pi}} \int_{-\varepsilon\sqrt{n/pq}}^{\varepsilon\sqrt{n/pq}} e^{-\frac{x^2}{2}} dx.$$

2. 大数定律的意义　下一节将给出上述结果的确切表述和证明. 现在, 我们讨论大数定律的现实意义, 及其经验解释.

假设进行大量, 例如 N 系列试验, 而每一系列试验包括 “n 次独立试验, 而每次试验以概率 p 出现某事件 C”. 设 S_n^i/n 是事件 C 在第 i 系列试验中出现的频率, N_ε 是 “频率对概率的绝对偏差不大于 ε” 系列数, 即

$$N_\varepsilon \text{等于使} \left|\frac{S_n^i}{n} - p\right| \leqslant \varepsilon \text{的} i \text{个数}.$$

那么, 由大数定律可见

$$\frac{N_\varepsilon}{N} \sim P_\varepsilon, \tag{10}$$

其中

$$P_\varepsilon = \mathbf{P}\left\{\left|\frac{S_n^1}{n} - p\right| \leqslant \varepsilon\right\}.$$

这里, 重要的是强调, 将 (10) 式精确化的尝试无疑将必须利用某一概率测度, 像估计频率 S_n/n 对概率 p 的偏差一样. 这种估计只有在引进概率测度 $\mathbf{P}$ 后才有可能.

3. 观测次数　考虑上面得到的估计

$$\mathbf{P}\left\{\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right\} = \sum_{\{k : |\frac{k}{n} - p| \geqslant \varepsilon\}} P_n(k) \leqslant \frac{1}{4n\varepsilon^2}. \tag{11}$$

为回答下面数理统计的典型问题: 对任意 $0 < p < 1$, 保证不等式

$$\mathbf{P}\left\{\left|\frac{S_n}{n} - p\right| \leqslant \varepsilon\right\} \geqslant 1 - \alpha \tag{12}$$

成立的最小观测次数 n 如何? 其中 α 是给定的通常较小的数.

由 (11) 式可见, 满足 (12) 式的最小观测次数, 是满足

$$n \geqslant \frac{1}{4\varepsilon^2 \alpha} \tag{13}$$

的最小整数 n.

例如，若取 $\alpha = 0.05, \varepsilon = 0.02$，则观测次数为 12 500 就可以满足 (12) 式，而且不依赖于参数 p.

我们在下面 (§6，第 5 小节) 将看到，观测次数可以大为减小，因为切比雪夫不等式作为概率

$$\mathbf{P}\left\{\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right\}$$

的上侧估计太粗略.

4. 熵 记

$$C(n,\varepsilon) = \left\{\omega : \left|\frac{S_n(\omega)}{n} - p\right| \leqslant \varepsilon\right\}.$$

由大数定律可见，对于任何 $\varepsilon > 0$，当 n 充分大时，集合 $C(n,\varepsilon)$ 的概率接近 1. 这时，$C(n,\varepsilon)$ 中的轨道 (实现) ω 自然称做典型的 [或 $C(n,\varepsilon)$−典型的].

提出下面的问题：典型实现的条数 $N(C(n,\varepsilon))$ 以及每一条典型实现的权 $p(\omega)$ 如何?

为此，首先注意到，基本事件空间 Ω 中点的总数 $N(\Omega) = 2^n$，而对于 $p = 0$ 或 1，典型轨道只有一条：$(1,1,\cdots,1)$ 或 $(0,0,\cdots,0)$，即 $C(n,\varepsilon) = 1$. 但是，假如 $p = 1/2$，则直观上显然，"几乎一切" 轨道 [只有 $(1,1,\cdots,1)$ 或 $(0,0,\cdots,0)$ 除外] 都是典型的，因而轨道的条数接近 2^n.

结果表明，对于 $0 < p < 1$，所提出问题有完全确定的答案：无论是典型轨道数，还是权重 $p(\omega)$，都决定于 p 的某一专门函数——"熵".

为更深入地揭示相应结果的内容，我们考虑 §2 第 2 小节中比伯努利概型更加一般的模型.

设 $(p_1, p_2, \cdots, p_r)$ 是一有限概率分布，即满足条件 $p_1 + p_2 + \cdots + p_r = 1$ 的非负实数. 称

$$H = -\sum_{i=1}^{r} p_i \ln p_i \tag{14}$$

为概率分布 $(p_1, p_2, \cdots, p_r)$ 的熵，其中 ln 是自然对数，且 $0 \ln 0 = 0$. 显然，$H \geqslant 0$，而且 $H = 0$ 当且仅当在 $p_1, p_2, \cdots, p_r$ 中除某一个为 1 之外都等于 0. 函数 $f(x) = -x \ln x (0 \leqslant x \leqslant 1)$ 是 (向上) 凸函数. 熟知，由凸函数的性质，有

$$\frac{f(x_1) + \cdots + f(x_r)}{r} \leqslant f\left(\frac{x_1 + \cdots + x_r}{r}\right).$$

从而

$$H = -\sum_{i=1}^{r} p_i \ln p_i \leqslant -r \times \frac{p_1 + \cdots + p_r}{r} \times \ln\left(\frac{p_1 + \cdots + p_r}{r}\right) = \ln r.$$

换句话说，当 $p_1 = p_2 = \cdots = p_r = 1/r$ 时，熵达到其最大值 (对于 $r = 2$，函数 $H = H(p)$ 的图形见图 8).

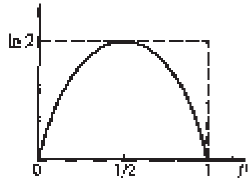


图 8 函数 $H(p)$

如果把 $p_1, p_2, \cdots, p_r$ 看成某些事件，如 $A_1, A_2, \cdots, A_r$ 出现的概率，则完全清楚：出现某个事件 "不确定性的程度"，对于不同的分布是不同的. 例如，$p_1 = 1, p_2 = \cdots = p_r = 0$，则该分布不具有任何不确定性：可以满怀信心地说，试验结果必然出现事件 A_1. 不过，如果 $p_1 = p_2 = \cdots = p_r = 1/r$，则这样的分布具有最大的不确定性，因为甚至不能说哪个事件出现的可能性更大些.

为比较不同分布的不确定性，需要有不同分布的不确定性之度量的数字特征. 熵 H 正是不确定性度量的恰当的数字特征. 由下面的讨论可见，熵在统计力学、编码理论和通讯理论中起着重要作用.

假设

$$\Omega = \{\omega : \omega = (a_1, \cdots, a_n), a_i = 1, \cdots, r\}$$

是基本事件空间，其中 $p(\omega) = p_1^{\nu_1(\omega)} \cdots p_r^{\nu_r(\omega)}$，其中 $\nu_i(\omega)$ 是序列 ω 中第 i 个元素 a_i 的个数，而 $(p_1, \cdots, p_r)$ 是某一概率分布.

对于任意 $\varepsilon > 0$，设

$$C(n,\varepsilon) = \left\{\omega : \left|\frac{\nu_i(\omega)}{n} - p_i\right| < \varepsilon, i = 1, \cdots, r\right\}.$$

显然，

$$\mathbf{P}(C(n,\varepsilon)) \geqslant 1 - \sum_{i=1}^{r} \mathbf{P}\left\{\left|\frac{\nu_i(\omega)}{n} - p_i\right| \geqslant \varepsilon\right\},$$

并且，由于大数定律知，该式也适用于随机变量

$$\xi_k(\omega) = \begin{cases} 1, & 若\ a_k = i, \\ 0, & 若\ a_k \neq i, \end{cases} \quad k = 1, \cdots, n,$$

概率

$$\mathbf{P}\left\{\left|\frac{\nu_i(\omega)}{n} - p_i\right| \geqslant \varepsilon\right\}$$

充分地小. 因此对于充分大的 n，$C(n,\varepsilon)$ 的概率接近 1. 同 $r = 2$ 的情形一样，进入 $C(n,\varepsilon)$ 的轨道称为典型的.

如果所有 $p_i > 0 (i = 1, \cdots, r)$，则对于任何 $\omega \in \Omega$，权重

$$p(\omega) = \exp\left\{-n \sum_{k=1}^{r}\left[-\frac{\nu_k(\omega)}{n} \ln p_k\right]\right\}.$$

因此，如果 ω 是典型轨道，则由 (14) 式，有

$$\left|\sum_{i=1}^{r}\left[-\frac{\nu_i(\omega)}{n} \ln p_i\right] - H\right| \leqslant -\sum_{i=1}^{r}\left|\frac{\nu_i(\omega)}{n} - p_i\right| \times \ln p_i \leqslant -\varepsilon \sum_{i=1}^{r} \ln p_i.$$

由此可见, 典型轨道的概率接近 e^{-nH}, 因为由于大数定律当 n 充分大时, 典型轨道的条数 "几乎" 穷尽 Ω 中的所有点, 而 Ω 中轨道的条数应该为量级 e^{nH}. 将以上的讨论归纳为下面的定理.

定理 (麦克米兰 [B. McMillan]) 设 $p_i > 0 (i = 1, \cdots, r), 0 < \varepsilon < 1$, 则存在一自然数 $n_0 = n_0(\varepsilon; p_1, \cdots, p_r)$, 使对于一切 $n > n_0$:

a) $e^{n(H-\varepsilon)} \leqslant N(C(n, \varepsilon_1)) \leqslant e^{n(H+\varepsilon)}$;

b) $e^{-n(H+\varepsilon)} \leqslant p(\omega) \leqslant e^{-n(H-\varepsilon)}, \omega \in C(n, \varepsilon_1)$;

c) $\mathbf{P}(C(n, \varepsilon_1)) = \sum\limits_{\omega \in C(n,\varepsilon_1)} p(\omega) \to 1 (n \to \infty)$,

其中

$$\varepsilon_1 = \min\left[\varepsilon, \varepsilon\left(-2\sum_{k=1}^{r} \ln p_k\right)^{-1}\right].$$

证明 命题 c) 由大数定律得出. 对于其余命题注意到, 如果 $\omega \in C(n, \varepsilon_1)$, 则

$$np_k - \varepsilon_1 n < \nu_k(\omega) < np_k + \varepsilon_1 n \quad (k = 1, \cdots, r),$$

因此

$$\begin{aligned} p(\omega) = \exp\left\{-\sum \nu_k \ln p_k\right\} &< \exp\left\{-n\sum p_k \ln p_k - \varepsilon_1 n \sum \ln p_k\right\} \\ &\leqslant \exp\left\{-n\left(H - \frac{\varepsilon}{2}\right)\right\}. \end{aligned}$$

类似地, 有

$$p(\omega) > \exp\left\{-n\left(H + \frac{\varepsilon}{2}\right)\right\}.$$

从而, 命题 b) 得证.

最后, 由于

$$\mathbf{P}(C(n, \varepsilon_1)) \geqslant N(C(n, \varepsilon_1)) \times \min_{\omega \in C(n,\varepsilon_1)} p(\omega),$$

则

$$N(C(n, \varepsilon_1)) \leqslant \frac{\mathbf{P}(C(n, \varepsilon_1))}{\min\limits_{\omega \in C(n,\varepsilon_1)} p(\omega)} < \frac{1}{e^{-n(H+\frac{\varepsilon}{2})}} = e^{n(H+\frac{\varepsilon}{2})};$$

类似地, 有

$$N(C(n, \varepsilon_1)) \geqslant \frac{\mathbf{P}(C(n, \varepsilon_1))}{\max\limits_{\omega \in C(n,\varepsilon_1)} p(\omega)} > \mathbf{P}(C(n, \varepsilon_1)) e^{n(H-\frac{\varepsilon}{2})}.$$

由于 $\mathbf{P}(C(n, \varepsilon_1)) \to 1 (n \to \infty)$, 可见存在 n_1, 使对于 $n > n_1$, 有 $\mathbf{P}(C(n, \varepsilon_1)) > 1 - \varepsilon$, 故

$$N(C(n, \varepsilon_1)) \geqslant (1-\varepsilon)\exp\left\{n\left(H - \frac{\varepsilon}{2}\right)\right\} = \exp\left\{n(H - \varepsilon) + \left[\frac{n\varepsilon}{2} + \ln(1-\varepsilon)\right]\right\}.$$

假设 n_2 满足: 对于 $n > n_2$, 有

$$\frac{n\varepsilon}{2} + \ln(1 - \varepsilon) > 0.$$

于是, 对于 $n \geqslant n_0 = \max\{n_1, n_2\}$, 有

$$N(C(n, \varepsilon_1)) \geqslant e^{n(H-\varepsilon)}. \tag{14}$$

5. 用概率方法证明维尔斯特拉斯定理 利用伯努利概型的大数定律, 可以给著名的维尔斯特拉斯 (K. T. W. Wierstrass) 定理 "用多项式逼近连续函数" 以简单而雅致的证明.

设 $f(p)$ 是线段 $[0, 1]$ 上的连续函数. 引进多项式

$$B_n(p) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) \mathrm{C}_n^k p^k (1-p)^{n-k}, \quad 0 \leqslant p \leqslant 1, n \geqslant 0. \tag{15}$$

该多项式称做伯恩斯坦 (С. Н. Бернштейн) 多项式[①], 用提供维尔斯特拉斯定理的此证明的作者命名.

如果 $\xi_1, \cdots, \xi_n$ 是独立伯努利随机变量序列, 且 $\mathbf{P}\{\xi_i = 1\} = p, \mathbf{P}\{\xi_i = 0\} = q$, 设 $S_n = \xi_1 + \cdots + \xi_n$, 则

$$\mathbf{E} f(S_n/n) = B_n(p).$$

由于在闭区间 $[0, 1]$ 上的连续函数 $f = f(p)$ 一致连续, 可见对于任意 $\varepsilon > 0$ 存在 $\delta > 0$, 使当 $|x - y| \leqslant \delta$ 时 $|f(x) - f(y)| \leqslant \varepsilon$. 显然, 这样的函数有界: $|f(x)| \leqslant M < \infty$.

因此由不等式 (5), 可见

$$\begin{aligned} |f(p) - B_n(p)| &= \left|\sum_{k=0}^{n}\left[f(p) - f\left(\frac{k}{n}\right)\right] \mathrm{C}_n^k p^k q^{n-k}\right| \\ &\leqslant \sum_{\{k:|\frac{k}{n}-p|\leqslant\delta\}} \left|f(p) - f\left(\frac{k}{n}\right)\right| \mathrm{C}_n^k p^k q^{n-k} \\ &\qquad + \sum_{\{k:|\frac{k}{n}-p|>\delta\}} \left|f(p) - f\left(\frac{k}{n}\right)\right| \mathrm{C}_n^k p^k q^{n-k} \\ &\leqslant \varepsilon + 2M \sum_{\{k:|\frac{k}{n}-p|>\delta\}} \mathrm{C}_n^k p^k q^{n-k} \leqslant \varepsilon + \frac{2M}{4n\delta^2} = \varepsilon + \frac{M}{2n\delta^2}. \end{aligned}$$

由此可见, 对于伯恩斯坦多项式 (15),

$$\lim_{n\to\infty} \max_{0\leqslant p\leqslant 1} |f(p) - B_n(p)| = 0.$$

这正是维尔斯特拉斯定理的结论.

① С. Н. 伯恩斯坦 (С. Н. Бернштейн, 1880 — 1968), 乌兹别克斯坦统计学家. —— 译者

6. 练习题

1. 设随机变量 ξ 和 η 的相关系数为 ρ. 证明切比雪夫不等式二维类似:

$$\mathbf{P}\{|\xi-\mathbf{E}\xi|\geqslant\varepsilon\sqrt{\mathbf{D}\xi}\ \text{或}\ |\eta-\mathbf{E}\eta|\geqslant\varepsilon\sqrt{\mathbf{D}\eta}\}\leqslant\frac{1+\sqrt{1-\rho^2}}{\varepsilon^2}.$$

(提示. 利用 §4 中练习题 8 的结果.)

2. 设 $f=f(x)$ 为非负偶函数, 且当 $x>0$ 时非减. 设 $\xi=\xi(\omega)$ 是非负随机变量, 且 $|\xi(\omega)|\leqslant C$, 证明对于任意 $\varepsilon>0$, 有

$$\mathbf{P}\{|\xi|\geqslant\varepsilon\}\geqslant\frac{\mathbf{E}f(\xi)-f(\varepsilon)}{f(C)}.$$

特别, 对于 $f(x)=x^2$,

$$\frac{\mathbf{E}\xi^2-\varepsilon^2}{C^2}\leqslant\mathbf{P}\{|\xi-\mathbf{E}\xi|\geqslant\varepsilon\}\leqslant\frac{\mathbf{D}\xi}{\varepsilon^2}.$$

3. 设 $\xi_1,\cdots,\xi_n$ 是独立随机变量序列, $\mathbf{D}\xi_i\leqslant C$, 证明

$$\mathbf{P}\left\{\left|\frac{\xi_1+\cdots+\xi_n}{n}-\frac{\mathbf{E}(\xi_1+\cdots+\xi_n)}{n}\right|\geqslant\varepsilon\right\}\leqslant\frac{C}{n\varepsilon^2}. \tag{16}$$

(与关系式 (8) 有同样的补充说明, 由不等式 (16), 可见在比伯努利概型更一般的情形下, 大数定律依然成立.)

4. 设 $\xi_1,\cdots,\xi_n$ 是独立伯努利随机变量, 且 $\mathbf{P}\{\xi_i=1\}=p>0, \mathbf{P}\{\xi_i=-1\}=q(p+q=1)$. 证明有如下伯恩斯坦估计: 对于任意 $a>0$, 有

$$\mathbf{P}\left\{\left|\frac{S_n}{n}-(2p-1)\right|\geqslant\varepsilon\right\}\leqslant 2e^{-a\varepsilon^2 n},$$

其中 $S_n=\xi_1+\cdots+\xi_n, \varepsilon>0$.

5. 设 ξ 是非负随机变量, 而 $a>0$. 求概率 $\mathbf{P}\{\xi\geqslant a\}$ 的上确界, 假如已知

(1) $\mathbf{E}\xi=20$;

(2) $\mathbf{E}\xi=20, \mathbf{D}\xi=25$;

(3) $\mathbf{E}\xi=20, \mathbf{D}\xi=25$, 且 ξ 关于数学期望对称.

§6. 伯努利概型 II. 极限定理 (棣莫弗–拉普拉斯局部定理、泊松定理)

1. 棣莫弗–拉普拉斯局部定理 像上一节一样, 考虑

$$S_n=\xi_1+\cdots+\xi_n.$$

那么

$$\mathbf{E}\frac{S_n}{n}=p, \tag{1}$$

而由 §4, (14) 式

$$\mathbf{E}\left(\frac{S_n}{n}-p\right)^2=\frac{pq}{n}. \tag{2}$$

由 (1) 式可见, $S_n/n\sim p$, 其中关于等价符号 "$\sim$" 曾以概率

$$\mathbf{P}\left\{\left|\frac{S_n}{n}-p\right|\geqslant\varepsilon\right\}$$

的估计的形式, 在大数定律中得到确切的解释. 自然想到, 由 "关系式" (2) 亦可给

$$\left|\frac{S_n}{n}-p\right|\sim\sqrt{\frac{pq}{n}} \tag{3}$$

以确切的概率意义. 例如, 考虑形如

$$\mathbf{P}\left\{\left|\frac{S_n}{n}-p\right|\leqslant x\sqrt{\frac{pq}{n}}\right\},\quad x\in\mathbb{R}^1$$

的概率, 或 (由于 $\mathbf{E}S_n=np, \mathbf{D}S_n=npq$) 考虑概率

$$\mathbf{P}\left\{\left|\frac{S_n-\mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}}\right|\leqslant x\right\}.$$

如果对 $n\geqslant 1$, 仍记

$$P_n(k)=C_n^k p^k q^{n-k},\quad 0\leqslant k\leqslant n,$$

则概率

$$\mathbf{P}\left\{\left|\frac{S_n-\mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}}\right|\leqslant x\right\}=\sum_{\left\{k:\left|\frac{k-np}{\sqrt{npq}}\right|\leqslant x\right\}}P_n(k). \tag{4}$$

现在提出问题: 当 $n\to\infty$ 时求概率 $P_n(k)$ 及其和

$$\sum_{\left\{k:\left|\frac{k-np}{\sqrt{npq}}\right|\leqslant x\right\}}P_n(k)$$

满足的便于应用的渐近公式.

下面的定理, 对于既满足 $|k-np|=O(\sqrt{npq})$ 又满足 $|k-np|=o(npq)^{2/3}$ 的 k 值, 给出了答案.

局部极限定理 设 $0<p<1$, 则对满足 $|k-np|=o(npq)^{2/3}$ 的所有 k, 一致有

$$P_n(k)\sim\frac{1}{\sqrt{2\pi npq}}e^{-\frac{(k-np)^2}{2npq}}, \tag{5}$$

即当 $n\to\infty$ 时

$$\sup_{\{k:|k-np|\leqslant\varphi(n)\}}\left|\frac{P_n(k)}{\frac{1}{\sqrt{2\pi npq}}e^{-\frac{(k-np)^2}{2npq}}}-1\right|\to 0, \tag{6}$$

其中 $\varphi(n)$ 是任意满足 $\varphi(n)=o(npq)^{2/3}$ 的非负函数.

证明 对于证明, 主要用到斯特林 (J. Stering) 公式 (§6, (2) 式):

$$n!=\sqrt{2\pi n}e^{-n}n^n(1+R(n)),$$

其中当 $n\to\infty$ 时 $R(n)\to 0$.

根据斯特林公式, 若 $n\to\infty, k\to\infty, n-k\to\infty$, 则

$$\begin{aligned}C_n^k&=\frac{n!}{k!(n-k)!}\\&=\frac{\sqrt{2\pi n}e^{-n}n^n(1+R(n))}{\sqrt{2\pi k}\cdot\sqrt{2\pi(n-k)}e^{-k}k^k(n-k)^{n-k}e^{-(n-k)}(1+R(k))(1+R(n-k))}\\&=\frac{1}{\sqrt{2\pi n\frac{k}{n}\left(1-\frac{k}{n}\right)}}\times\frac{1+\varepsilon(n,k,n-k)}{\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}},\end{aligned}$$

其中当 $n\to\infty, k\to\infty, n-k\to\infty$ 时, 显然 $\varepsilon(n,k,n-k)\to 0$.

因此

$$P_n(k)=C_n^kp^kq^{n-k}=\frac{1}{\sqrt{2\pi n\frac{k}{n}\left(1-\frac{k}{n}\right)}}\times\frac{p^k(1-p)^{n-k}(1+\varepsilon)}{\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}}.$$

记 $\hat{p}=k/n$, 则

$$\begin{aligned}P_n(k)&=\frac{1}{\sqrt{2\pi n\hat{p}(1-\hat{p})}}\left(\frac{p}{\hat{p}}\right)^k\left(\frac{1-p}{1-\hat{p}}\right)^{n-k}(1+\varepsilon)\\&=\frac{1}{\sqrt{2\pi n\hat{p}(1-\hat{p})}}\exp\left\{k\ln\frac{p}{\hat{p}}+(n-k)\ln\frac{1-p}{1-\hat{p}}\right\}\times(1+\varepsilon)\\&=\frac{1}{\sqrt{2\pi n\hat{p}(1-\hat{p})}}\exp\left\{n\left[\frac{k}{n}\ln\frac{p}{\hat{p}}+\left(1-\frac{k}{n}\right)\ln\frac{1-p}{1-\hat{p}}\right]\right\}\times(1+\varepsilon)\\&=\frac{1}{\sqrt{2\pi n\hat{p}(1-\hat{p})}}\exp\{-nH(\hat{p})\}\times(1+\varepsilon),\end{aligned}$$

其中

$$H(x)=x\ln\frac{x}{p}+(1-x)\ln\frac{1-x}{1-p}.$$

所考虑的 k 满足 $|k-np|=o(npq)^{2/3}$, 故当 $n\to\infty$ 时 $p-\hat{p}\to 0$.

由于对 $0<x<1$

$$H'(x)=\ln\frac{x}{p}-\ln\frac{1-x}{1-p},$$

$$H''(x)=\frac{1}{x}+\frac{1}{1-x},$$

$$H'''(x)=-\frac{1}{x^2}+\frac{1}{(1-x)^2}.$$

因此, 若将 $H(\hat{p})$ 表示为 $H(p+(\hat{p}-p))$, 并利用泰勒 (B. Taylor) 公式, 则当 n 充分大时, 有

$$\begin{aligned}H(\hat{p})&=H(p)+H'(p)(\hat{p}-p)+\frac{1}{2}H''(p)(\hat{p}-p)^2+O(|\hat{p}-p|^3)\\&=\frac{1}{2}\left(\frac{1}{p}+\frac{1}{q}\right)(\hat{p}-p)^2+O(|\hat{p}-p|^3).\end{aligned}$$

从而

$$P_n(k)=\frac{1}{\sqrt{2\pi n\hat{p}(1-\hat{p})}}\exp\left\{-\frac{n}{2pq}(\hat{p}-p)^2+nO(|\hat{p}-p|^3)\right\}\times(1+\varepsilon).$$

注意到

$$\frac{n}{2pq}(\hat{p}-p)^2=\frac{n}{2pq}\left(\frac{k}{n}-p\right)^2=\frac{(k-np)^2}{2npq}.$$

因此

$$P_n(k)=\frac{1}{\sqrt{2\pi npq}}e^{-\frac{(k-np)^2}{2npq}}(1+\varepsilon'(n,k,n-k)),$$

其中

$$1+\varepsilon'(n,k,n-k)=(1+\varepsilon(n,k,n-k))e^{nO(|\hat{p}-p|^3)}\sqrt{\frac{p(1-p)}{\hat{p}(1-\hat{p})}}.$$

易见

$$\sup|\varepsilon'(n,k,n-k)|\to 0,\quad n\to\infty,$$

其中 sup 对满足

$$|k-np|\leqslant\varphi(n),\quad\varphi(n)=o(npq)^{2/3}$$

的 k 来求. □

系 局部极限定理的结论, 可以表述为如下等价的形式: 对于一切 $x\in\mathbb{R}^1$, 若 $x=o(npq)^{1/6}$, 而 $np+x\sqrt{npq}$ 是集合 $\{0,1,\cdots,n\}$ 中的整数, 则

$$P_n(np+x\sqrt{npq})\sim\frac{1}{\sqrt{2\pi npq}}e^{-\frac{x^2}{2}},\tag{7}$$

即当 $n\to\infty$ 时, 有

$$\sup_{\{x:|x|\leqslant\psi(n)\}}\left|\frac{P_n(np+x\sqrt{npq})}{\frac{1}{\sqrt{2\pi npq}}e^{-\frac{x^2}{2}}}-1\right|\to 0,\tag{8}$$

其中 $\psi(n)=o(npq)^{1/6}$.

注意到关于 §5 中 (8) 式的说明, 可以用概率的语言将上面得到的结果表述为:

$$\mathbf{P}\{S_n=k\}\sim\frac{1}{\sqrt{2\pi npq}}e^{-\frac{(k-np)^2}{2npq}},\ |k-np|=o(npq)^{2/3},\tag{9}$$

$$\mathbf{P}\left\{\frac{S_n-np}{\sqrt{npq}}=x\right\}\sim\frac{1}{\sqrt{2\pi npq}}e^{-\frac{x^2}{2}},\ x=o(npq)^{1/6}.\tag{10}$$

(假设在 (10) 式中 $np+x\sqrt{npq}$ 取 $0,1,\cdots,n$ 为值.)

假如设

$$t_k=\frac{k-np}{\sqrt{npq}},\quad \Delta t_k=t_{k+1}-t_k=\frac{1}{\sqrt{npq}},$$

则 (10) 式具有如下形式:

$$\mathbf{P}\left\{\frac{S_n-\mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}}=t_k\right\}\sim\frac{\Delta t_k}{\sqrt{2\pi}}e^{-\frac{t_k^2}{2}},\quad t_k=o(npq)^{1/6}. \tag{11}$$

显然当 $n\to\infty$ 时

$$\Delta t_k=\frac{1}{\sqrt{npq}}\to 0,$$

而点 t_k 的集合 $\{t_k\}$ "充满" 整个数轴. 因此, 自然想到由 (11) 式可以得到积分公式:

$$\mathbf{P}\left\{a<\frac{S_n-np}{\sqrt{npq}}\leqslant b\right\}\sim\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}\mathrm{d}x,\quad -\infty<a\leqslant b<\infty.$$

下面将给出确切的表述.

2. 棣莫弗 – 拉普拉斯积分定理 对于 $-\infty<a\leqslant b<\infty$, 设

$$P_n(a,b]=\sum_{a<x\leqslant b}P_n(np+x\sqrt{npq}),$$

其中对一切使 $np+x\sqrt{npq}$ 为整数的 x 求和.

由局部定理可见 (亦见 (11) 式), 对于由 $k=np+t_k\sqrt{npq}$ 决定且满足条件 $|t_k|\leqslant T<\infty$ 的 t_k, 有

$$P_n(np+t_k\sqrt{npq})=\frac{\Delta t_k}{\sqrt{2\pi}}e^{-\frac{t_k^2}{2}}[1+\varepsilon(t_k,n)], \tag{12}$$

其中

$$\sup_{|t_k|\leqslant T}|\varepsilon(t_k,n)|\to 0,\ n\to\infty. \tag{13}$$

从而, 对于固定的 a,b ($-T\leqslant a\leqslant b\leqslant T$, 而 $T<\infty$),

$$\sum_{a<t_k\leqslant b}P_n(np+t_k\sqrt{npq})=\sum_{a<t_k\leqslant b}\frac{\Delta t_k}{\sqrt{2\pi}}e^{-\frac{t_k^2}{2}}+\sum_{a<t_k\leqslant b}\varepsilon(t_k,n)\frac{\Delta t_k}{\sqrt{2\pi}}e^{-\frac{t_k^2}{2}}$$

$$=\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}\mathrm{d}x+R_n^{(1)}(a,b)+R_n^{(2)}(a,b), \tag{14}$$

其中

$$R_n^{(1)}(a,b)=\sum_{a<t_k\leqslant b}\frac{\Delta t_k}{\sqrt{2\pi}}e^{-\frac{t_k^2}{2}}-\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}\mathrm{d}x,$$

$$R_n^{(2)}(a,b)=\sum_{a<t_k\leqslant b}\varepsilon(t_k,n)\frac{\Delta t_k}{\sqrt{2\pi}}e^{-\frac{t_k^2}{2}}.$$

由熟知的积分和的性质,

$$\sup_{-T\leqslant a\leqslant b\leqslant T}|R_n^{(1)}(a,b)|\to 0,\quad n\to\infty. \tag{15}$$

同样易见

$$\sup_{-T\leqslant a\leqslant b\leqslant T}|R_n^{(2)}(a,b)|\leqslant\sup_{|t_k|\leqslant T}|\varepsilon(t_k,n)|\sum_{|t_k|\leqslant T}\frac{\Delta t_k}{\sqrt{2\pi}}e^{-\frac{t_k^2}{2}}$$

$$\leqslant\sup_{|t_k|\leqslant T}|\varepsilon(t_k,n)|\times\left[\frac{1}{\sqrt{2\pi}}\int_{-T}^{T}e^{-\frac{x^2}{2}}\mathrm{d}x+\sup_{-T\leqslant a\leqslant b\leqslant T}|R_n^{(1)}(a,b)|\right]\to 0, \tag{16}$$

其中右侧收敛于 0, 是因为 (15) 式以及数学分析中熟知的事实

$$\frac{1}{\sqrt{2\pi}}\int_{-T}^{T}e^{-\frac{x^2}{2}}\mathrm{d}x\leqslant\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{x^2}{2}}\mathrm{d}x=1. \tag{17}$$

记

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}\mathrm{d}t,$$

则由 (14) ~ (16) 式可见

$$\sup_{-T\leqslant a\leqslant b\leqslant T}|P_n(a,b]-[\Phi(b)-\Phi(a)]|\to 0,\ n\to\infty. \tag{18}$$

现在证明, 该式不仅对于有限 T 成立, 而且对于 $T=\infty$ 也成立. 由于 (17) 式, 对于任意给定的 $\varepsilon>0$, 存在有限 $T=T(\varepsilon)$, 使

$$\frac{1}{\sqrt{2\pi}}\int_{-T}^{T}e^{-\frac{x^2}{2}}\mathrm{d}x>1-\frac{\varepsilon}{4}. \tag{19}$$

根据 (18) 式, 对于任意 $\varepsilon>0$, 存在 N, 使对于一切 $n>N$ 和 $T=T(\varepsilon)$, 有

$$\sup_{-T\leqslant a\leqslant b\leqslant T}|P_n(a,b]-[\Phi(b)-\Phi(a)]|<\frac{\varepsilon}{4}. \tag{20}$$

由此和 (19) 式, 可见

$$P_n(-T,T]>1-\frac{\varepsilon}{2},$$

因此

$$P_n(-\infty,-T]+P_n(T,\infty)\leqslant\frac{\varepsilon}{2},$$

其中

$$P_n(-\infty,T]=\lim_{S\downarrow-\infty}P_n(S,T],\ P_n(T,\infty)=\lim_{S\uparrow\infty}P_n(T,S].$$

这样，对于任意 $-\infty \leqslant -T \leqslant a \leqslant b \leqslant T \leqslant \infty$，有

$$\left|P_n(a,b]-\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}\mathrm{d}x\right| \leqslant \left|P_n(-T,T]-\frac{1}{\sqrt{2\pi}}\int_{-T}^T e^{-\frac{x^2}{2}}\mathrm{d}x\right|$$

$$+\left|P_n(a,-T]-\frac{1}{\sqrt{2\pi}}\int_a^{-T} e^{-\frac{x^2}{2}}\mathrm{d}x\right|+\left|P_n(T,b]-\frac{1}{\sqrt{2\pi}}\int_T^b e^{-\frac{x^2}{2}}\mathrm{d}x\right|$$

$$\leqslant \frac{\varepsilon}{4}+P_n(-\infty,-T]+\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-T} e^{-\frac{x^2}{2}}\mathrm{d}x+P_n(T,\infty)+\frac{1}{\sqrt{2\pi}}\int_T^{\infty} e^{-\frac{x^2}{2}}\mathrm{d}x$$

$$\leqslant \frac{\varepsilon}{4}+\frac{\varepsilon}{2}+\frac{\varepsilon}{8}+\frac{\varepsilon}{8}=\varepsilon.$$

注意到 (18) 式，由此容易证明 $P_n(a,b]$ 关于 $-\infty\leqslant a<b\leqslant\infty$ 一致趋向 $\Phi(b)-\Phi(a)$。于是，证明了下面的定理。

棣莫弗－拉普拉斯积分定理 设 $0<p<1$，

$$P_n(k)=\mathrm{C}_n^k p^k q^{n-k},\quad P_n(a,b]=\sum_{a<x\leqslant b}P_n(np+x\sqrt{npq}).$$

那么

$$\sup_{-\infty\leqslant a<b\leqslant\infty}\left|P_n(a,b]-\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}\mathrm{d}x\right|\to 0,\quad n\to\infty. \tag{21}$$

精确到 §5 中 (8) 式所指，可以将 (21) 式的结果用概率论语言表示为：

$$\sup_{-\infty\leqslant a<b\leqslant\infty}\left|\mathbf{P}\left\{a<\frac{S_n-\mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}}\leqslant b\right\}-\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}\mathrm{d}x\right|\to 0,\quad n\to\infty.$$

由此可见，对于任意 $-\infty\leqslant A<B\leqslant\infty$，有

$$\mathbf{P}\{A<S_n\leqslant B\}-\left[\Phi\left(\frac{B-np}{\sqrt{npq}}\right)-\Phi\left(\frac{A-np}{\sqrt{npq}}\right)\right]\to 0,\quad n\to\infty. \tag{22}$$

例 将规则的色子掷 12 000 次. 问 "6 点" 出现的次数属于区间 (1800, 2100] 的概率 P 如何？

所求概率等于

$$P=\sum_{1800<k\leqslant 2100}\mathrm{C}_{12\,000}^k\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{n-k}.$$

显然，用 "手算" 精确地算出该和的值是相当困难的. 假如利用积分定理，则求得此概率 P 大致等于 ($n=12\,000, p=1/6, A=1800, B=2100$)：

$$\Phi\left(\frac{2100-2000}{\sqrt{12\,000\times\frac{1}{6}\times\frac{5}{6}}}\right)-\Phi\left(\frac{1800-2000}{\sqrt{12\,000\times\frac{1}{6}\times\frac{5}{6}}}\right)$$

$$=\Phi(\sqrt{6})-\Phi(-2\sqrt{6})\approx\Phi(2.449)-\Phi(-4.898)\approx 0.993,$$

其中 $\Phi(2.449)$ 和 $\Phi(-4.898)$ 的值，由正态分布函数的 $\Phi(x)$ 数值表查出 (参见第 6 小节).

3. 二项概率的正态逼近 把二项概率 $P_n(np+x\sqrt{npq})$ (假设只考虑使 $np+x\sqrt{npq}$ 为整数的 x) 标在图上 (图 9).

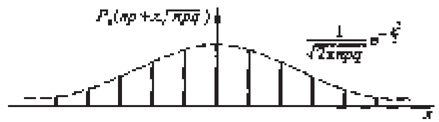


图 9

那么，局部定理表明，对于 $x=o(npq)^{1/6}$，概率 $P_n(np+x\sqrt{npq})$ 的值较好地 "位于" 正态密度曲线上：

$$\frac{1}{\sqrt{2\pi npq}}e^{-\frac{x^2}{2}}.$$

由积分定理知，概率

$$P_n(a,b]=\mathbf{P}\{a\sqrt{npq}<S_n-np\leqslant b\sqrt{npq}\}$$

$$=\mathbf{P}\{np+a\sqrt{npq}<S_n\leqslant np+b\sqrt{npq}\}$$

的值可以较好地由积分

$$\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}\mathrm{d}x$$

逼近.

记

$$F_n(x)=P_n(-\infty,x]\left(=\mathbf{P}\left\{\frac{S_n-np}{\sqrt{npq}}\leqslant x\right\}\right).$$

那么，由 (21) 式可见

$$\sup_{-\infty\leqslant x\leqslant\infty}|F_n(x)-\Phi(x)|\to 0,\quad n\to\infty. \tag{23}$$

重要的是在 (21) 和 (23) 式中，随着 n 的增长趋向 0 的速度如何. 这里引用的结果是贝里－埃森 (A. C. Berry-C. G. Esseen) 定理的特殊情形 (第三章 §11)：

$$\sup_{-\infty<x<\infty}|F_n(x)-\Phi(x)|\leqslant\frac{p^2+q^2}{\sqrt{npq}}. \tag{24}$$

要特别强调，估计 $1/\sqrt{npq}$ 的量级不能再提高了. 这指的是，当 p 的值接近 0 或 1 时，甚至对于充分大的 n，用函数 $\Phi(x)$ 逼近 $F_n(x)$ 的效果可能不佳. 因此产生一个问题，当 p 或 q 的值较小时，能否为我们关心的概率，找到比局部和积分定理给出的正态分布更好的逼近. 为此我们指出，当 $p=1/2$ 时二项分布 $\{P_n(k)\}$ 具有对称的形状 (图 10 左边的图). 不过，当 p 较小时，二项分布的形状是非对称的 (图 10 的右边的图)，不能指望用正态逼近有好结果.

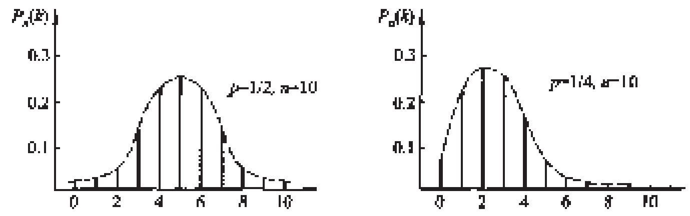


图 10

4. 泊松定理 结果表明, 对于较小的 p 值, 概率的所谓泊松分布可以很好地逼近二项分布的概率 $P_n(k)$.

设

$$P_n(k)=\begin{cases} C_n^k p^k q^{n-k}, & 若\ k=0,1,\cdots,n;\\ 0, & 若\ k=n+1,n+2,\cdots, \end{cases}$$

且假设 p 是 n 的函数 $p=p(n)$.

泊松定理 设 $p(n)\to 0, n\to\infty$, 且 $np(n)\to\lambda$, 其中 $\lambda>0$. 那么, 对于任意 $k=0,1,\cdots$, 有

$$P_n(k)\to\pi_k,\quad n\to\infty, \tag{25}$$

其中

$$\pi_k=\frac{\lambda^k e^{-\lambda}}{k!},\quad k=0,1,\cdots. \tag{26}$$

证明 证明十分简单. 由于根据条件

$$p(n)=\frac{\lambda}{n}+o\left(\frac{1}{n}\right),$$

可见对于固定的 $k=0,1,\cdots$ 和充分大的 n,

$$P_n(k)=C_n^k p^k q^{n-k}=\frac{n(n-1)\cdots(n-k+1)}{k!}\left[\frac{\lambda}{n}+o\left(\frac{1}{n}\right)\right]^k\times\left[1-\frac{\lambda}{n}+o\left(\frac{1}{n}\right)\right]^{n-k}.$$

由

$$n(n-1)\cdots(n-k+1)\left[\frac{\lambda}{n}+o\left(\frac{1}{n}\right)\right]^k=\frac{n(n-1)\cdots(n-k+1)}{n^k}[\lambda+o(1)]^k\to\lambda^k, n\to\infty.$$

和

$$\left[1-\frac{\lambda}{n}+o\left(\frac{1}{n}\right)\right]^{n-k}\to e^{-\lambda}, n\to\infty,$$

立即得 (25) 式. □

数组 $\{\pi_k, k=0,1\cdots\}$, 满足

$$\pi_k\geqslant 0, \sum_{k=0}^{\infty}\pi_k=1,$$

因此可以作为概率分布, 称做泊松分布. 注意, 以上所讨论的概率分布都集中在有限个点上. 泊松分布, 是我们第一次遇到的集中在可数个点上的 (离散型) 分布的例子.

下面引进的 (Ю. В. 普罗霍洛夫的) 结果, 给出了当 $n\to\infty$ 时概率 $P_n(k)$ 收敛于 π_k 的速度: 如果 $np(n)=\lambda$, 则[*]

$$\sum_{k=0}^{\infty}|P_n(k)-\pi_k|\leqslant\frac{2\lambda}{n}\times\min(2,\lambda). \tag{27}$$

5. 棣莫弗–拉普拉斯定理与大数定律 我们再回到棣莫弗–拉普拉斯极限定理. (在 §5 对 (8) 式说明的前提下) 说明如何由棣莫弗–拉普拉斯极限定理, 得出大数定律. 因为

$$\mathbf{P}\left\{\left|\frac{S_n}{n}-p\right|\leqslant\varepsilon\right\}=\mathbf{P}\left\{\left|\frac{S_n-np}{\sqrt{npq}}\right|\leqslant\varepsilon\sqrt{\frac{n}{pq}}\right\},$$

所以由 (21) 式可见, 对于 $\varepsilon>0$, 当 $n\to\infty$ 时

$$\mathbf{P}\left\{\left|\frac{S_n}{n}-p\right|\leqslant\varepsilon\right\}-\frac{1}{\sqrt{2\pi}}\int_{-\varepsilon\sqrt{n/pq}}^{\varepsilon\sqrt{n/pq}}e^{-x^2/2}dx\to 0, \tag{28}$$

因此

$$\mathbf{P}\left\{\left|\frac{S_n}{n}-p\right|\leqslant\varepsilon\right\}\to 1,\quad n\to\infty,$$

此即大数定律的结论.

由 (28) 式, 可见

$$\mathbf{P}\left\{\left|\frac{S_n}{n}-p\right|\leqslant\varepsilon\right\}\sim\frac{1}{\sqrt{2\pi}}\int_{-\varepsilon\sqrt{n/pq}}^{\varepsilon\sqrt{n/pq}}e^{-x^2/2}dx,\quad n\to\infty, \tag{29}$$

然而, 切比雪夫不等式只能给出下面的估计

$$\mathbf{P}\left\{\left|\frac{S_n}{n}-p\right|\leqslant\varepsilon\right\}\geqslant 1-\frac{pq}{n\varepsilon^2}.$$

在 §5 第 3 小节关于为使不等式

$$\mathbf{P}\left\{\left|\frac{S_n}{n}-p\right|\leqslant\varepsilon\right\}\geqslant 1-\alpha$$

[*] 较细结果的证明, 将在第三章 §12 介绍.

成立所需要的观测次数, 由切比雪夫不等式得到下面的估计

$$n \geqslant \left[\frac{1}{4\varepsilon^2\alpha}\right] (= n_1(\alpha)),$$

其中 $[x]$ 是 x 的整数部分. 例如, 对于 $\varepsilon=0.02, \alpha=0.05$, 需要 12 500 次观测. 现在, 利用等价关系式 (29) 解决了同一问题.

我们由关系式

$$\frac{1}{\sqrt{2\pi}}\int_{-k(\alpha)}^{k(\alpha)} e^{-\frac{x^2}{2}}\mathrm{d}x = 1-\alpha$$

求 $k(\alpha)$. 由于

$$\varepsilon\sqrt{\frac{n}{pq}} \geqslant 2\varepsilon\sqrt{n},$$

并由不等式

$$2\varepsilon\sqrt{n}\geqslant k(\alpha) \tag{30}$$

求出 (最小整数), 得

$$\mathrm{P}\left\{\left|\frac{S_n}{n}-p\right| \leqslant \varepsilon\right\} \geqslant 1-\alpha. \tag{31}$$

由 (30) 式, 可见 $n \geqslant n_2(\alpha)$, 其中

$$n_2(\alpha) = \left[\frac{k^2(\alpha)}{4\varepsilon^2}\right]$$

可以保证 (31) 式成立, 其逼近精度容易由 (24) 式得到.

取 $\varepsilon=0.02, \alpha=0.05$, 可见只需要 2 500 次观测, 而不是切比雪夫不等式要求的 12 500 次. 下面对于一些 α 值, 列举相应的 $k(\alpha)$ 值:

α:	0.50	0.317 3	0.10	0.05	0.045 4	0.01	0.002 7
$k(\alpha)$:	0.675	1.000	1.645	1.960	2.000	2.576	3.000

6. 正态分布 前面在棣莫弗-拉普拉斯积分定理里, 引进的函数

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{y^2}{2}}\mathrm{d}y, \tag{32}$$

在概率论里起非常重要的作用, 称做*正态分布函数*或*高斯分布函数*. 函数

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}, \quad x\in\mathbb{R}^1,$$

称做*正态密度*或*高斯密度*.

我们已经见到过在有限或可数点集上的 (离散型) 分布. 正态分布属于概率论中另外一种非常重要类型的分布——连续型分布. 正态分布之所以非常重要, 首先是

因为在相当一般的条件下, 大量独立随机变量 (未必是伯努利变量) 之和的分布, 可以很好地用正态分布来逼近 (第三章 §4). 我们现在讨论函数 $\varphi(x)$ 和 $\Phi(x)$ 的一些简单性质, 而图 11 和图 12 分别是 $\varphi(x)$ 和 $\Phi(x)$ 的图形.

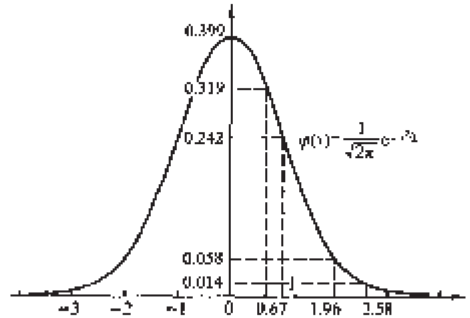


图 11 正态分布密度 $\varphi(x)$ 图

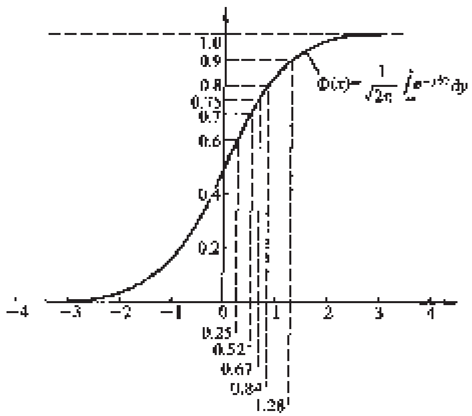


图 12 正态分布函数 $\Phi(x)$ 图

函数 $\varphi(x)$ 的图形是关于纵轴对称的钟形曲线, 随着 $|x|$ 的增长下降得非常快: $\varphi(1)=0.241\ 97, \varphi(2)=0.053\ 991, \varphi(3)=0.004\ 432, \varphi(4)=0.000\ 134, \varphi(5)=0.000\ 016$. 该曲线在 $x=0$ 达到最大值 $(2\pi)^{-1/2}\approx 0.399$.

随着 x 的增长, 函数

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\mathrm{d}t$$

曲线趋向 1 极快: $\Phi(1)=0.841\ 345, \Phi(2)=0.977\ 250, \Phi(3)=0.998\ 650, \Phi(4)$

0.999 968, $\Phi(5)=0.999\ 997$.

关于函数 $\varphi(x)$ 和 $\Phi(x)$, 以及概率论和数理统计中其他一些基本函数的数值表, 参见 [6].

需要指出, 在进行计算时, 除函数 $\Phi(x)$ 外, 还常使用与之相近的误差函数

$$\operatorname{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt, \quad x>0.$$

显然, 对于 $x>0$, 有

$$\Phi(x)=\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right], \quad \operatorname{erf}(x)=2\Phi(\sqrt{2}x)-1.$$

7. 成功频率对概率的偏差满足一定要求的试验次数 在 §5 第 3 小节的最后曾经指出, 由切比雪夫不等式给出的事件

$$\left\{\omega:\left|\frac{S_n}{n}-p\right|\geqslant\varepsilon\right\}$$

概率的估计是相当粗略的. 这一估计对于非负随机变量 X, 是由切比雪夫不等式

$$\mathrm{P}\{X\geqslant\varepsilon\}\leqslant\frac{\mathrm{E}X^2}{\varepsilon^2}$$

得到的. 不过可以利用切比雪夫不等式的另一种形式

$$\mathrm{P}\{X\geqslant\varepsilon\}=\mathrm{P}\{X^{2k}\geqslant\varepsilon^{2k}\}\leqslant\frac{\mathrm{E}X^{2k}}{\varepsilon^{2k}}. \tag{33}$$

然而, 还可以更进一步, 利用切比雪夫不等式的 "指数" 形式: 对于 $X\geqslant0$ 和 $\lambda>0$,

$$\mathrm{P}\{X\geqslant\varepsilon\}=\mathrm{P}\{e^{\lambda X}\geqslant e^{\lambda\varepsilon}\}\leqslant\mathrm{E}e^{\lambda(X-\varepsilon)}. \tag{34}$$

由于 $\lambda>0$ 的任意性, 可见

$$\mathrm{P}\{X\geqslant\varepsilon\}\leqslant\inf_{\lambda>0}\mathrm{E}e^{\lambda(X-\varepsilon)}. \tag{35}$$

我们讨论, 当 $X=S_n/n$, $S_n=\xi_1+\cdots+\xi_n$, $\mathrm{P}\{\xi_i=1\}=p$, $\mathrm{P}\{\xi_i=0\}=q$, $i\geqslant1$ 时, 沿此路径将导致何种结果.

记 $\varphi(\lambda)=\mathrm{E}e^{\lambda\xi_1}$, 则

$$\varphi(\lambda)=1-p+pe^{\lambda},$$

且在假设 $\xi_1,\cdots,\xi_n$ 独立的条件下, 有

$$\mathrm{E}e^{\lambda S_n}=[\varphi(\lambda)]^n.$$

因此, 对于 $0<a<1$,

$$\begin{aligned}\mathrm{P}\left\{\frac{S_n}{n}\geqslant a\right\}&\leqslant\inf_{\lambda>0}\mathrm{E}e^{\lambda\left(\frac{S_n}{n}-a\right)}=\inf_{\lambda>0}e^{-n\left[\frac{\lambda}{n}a-\ln\varphi\left(\frac{\lambda}{n}\right)\right]}\\&=\inf_{s>0}e^{-n[as-\ln\varphi(s)]}=e^{-n\sup_{s>0}[as-\ln\varphi(s)]}.\end{aligned} \tag{36}$$

类似地

$$\mathrm{P}\left\{\frac{S_n}{n}\leqslant a\right\}\leqslant e^{-n\sup_{s<0}[as-\ln\varphi(s)]}. \tag{37}$$

当 $p\leqslant a\leqslant1$ 时, 函数 $f(s)=as-\ln(1-p+pe^s)$ 在点 $s_0\,(f'(s_0)=0)$ 达到最大值, 其中点 s_0 决定于

$$e^{s_0}=\frac{a(1-p)}{p(1-a)}.$$

因此

$$\sup_{s>0}f(s)=H(a),$$

其中

$$H(a)=a\ln\frac{a}{p}+(1-a)\ln\frac{1-a}{1-p},$$

是前面证明局部定理时引进的函数 (第 1 小节).

这样, 当 $p\leqslant a\leqslant1$ 时

$$\mathrm{P}\left\{\frac{S_n}{n}\geqslant a\right\}\leqslant e^{-nH(a)}, \tag{38}$$

而由于 $H(p+x)\geqslant2x^2$, $0\leqslant p+x\leqslant1$, 则对于 $\varepsilon>0$, $0\leqslant p\leqslant1$, 有

$$\mathrm{P}\left\{\frac{S_n}{n}-p\geqslant\varepsilon\right\}\leqslant e^{-2n\varepsilon^2}. \tag{39}$$

类似可得, 当 $a\leqslant p\leqslant1$ 时

$$\mathrm{P}\left\{\frac{S_n}{n}\leqslant a\right\}\leqslant e^{-nH(a)}, \tag{40}$$

从而, 对任何 $\varepsilon>0$, $0\leqslant p\leqslant1$, 有

$$\mathrm{P}\left\{\frac{S_n}{n}-p\leqslant-\varepsilon\right\}\leqslant e^{-2n\varepsilon^2}. \tag{41}$$

于是

$$\mathrm{P}\left\{\left|\frac{S_n}{n}-p\right|\geqslant\varepsilon\right\}\leqslant2e^{-2n\varepsilon^2}. \tag{42}$$

由此可见, 对于任意 $0\leqslant p\leqslant1$, 保证不等式

$$\mathrm{P}\left\{\left|\frac{S_n}{n}-p\right|\leqslant\varepsilon\right\}\geqslant1-\alpha \tag{43}$$

成立的观测次数 $n_3(\alpha)$ 决定于下面的公式:

$$n_3(\alpha) = \left[\frac{\ln(2/\alpha)}{2\varepsilon^2}\right], \tag{44}$$

其中 $[x]$ 是 x 的整数部分. 不取"整数部分", 直接将 $n_3(\alpha)$ 与 $n_1(\alpha) = [(4\alpha\varepsilon^2)^{-1}]$ 比较, 可见

$$\frac{n_1(\alpha)}{n_3(\alpha)} = \frac{1}{4\alpha\varepsilon^2}\Big/\frac{\ln(2/\alpha)}{2\varepsilon^2} = \frac{1}{2\alpha\ln\frac{2}{\alpha}} \uparrow \infty, \quad \alpha \downarrow 0.$$

由此可见, 当 $\alpha \downarrow 0$ 时, 由指数型切比雪夫不等式 (34) 估计的最小必须要的观测次数, 比用一般切比雪夫不等式估计的次数更为准确, 特别是对于较小的 α. 利用将证明的关系式

$$\frac{1}{\sqrt{2\pi}}\int_x^{\infty} e^{-\frac{y^2}{2}}dy \sim \frac{1}{\sqrt{2\pi}x}e^{-\frac{x^2}{2}}, \quad x \to \infty,$$

可以证明, 当 $\alpha \downarrow 0$ 时 $k^2(\alpha) \sim 2\ln(2/\alpha)$. 于是

$$\frac{n_2(\alpha)}{n_3(\alpha)} \to 1, \alpha \downarrow 0.$$

像 (38) ~ (42) 式类型的关系式, 在概率论中称做*大偏差概率的不等式*. 下面是对这一名称的解释.

利用棣莫弗-拉普拉斯定理, 可以简单地估计事件 $\{|S_n - np| \leqslant x\sqrt{n}\}$ 的概率, 此事件表示 S_n 对 np (数量级 $\sqrt{n}$) 的"标准"离差. 而不等式 (39),(41) 和 (42) 对

$$\{|S_n - np| \leqslant x\sqrt{n}\}$$

给出的估计描绘量级大于 $\sqrt{n}$ 的离差, 其数量级为 n.

我们将在第四章 §5 中, 在更一般的情形下研究关于大偏差概率的问题.

6. 练习题

1. 设 $n = 100, p = 1/10, p = 2/10, p = 3/10, p = 4/10, p = 5/10$. 利用 (例如文献 [6] 中的) 二项分布以及泊松分布的数值表①, 将概率

$$\mathbf{P}\{10 < S_{100} \leqslant 12\}, \quad \mathbf{P}\{20 < S_{100} \leqslant 22\},$$
$$\mathbf{P}\{33 < S_{100} \leqslant 35\}, \quad \mathbf{P}\{40 < S_{100} \leqslant 42\},$$
$$\mathbf{P}\{50 < S_{100} \leqslant 52\}$$

与正态逼近和泊松逼近的相应数值进行比较.

2. 设 $p = 1/2$, 而 $Z_n = 2S_n - n$ (n 组试验中 1 比 0 多出的个数). 证明

$$\sup_j \left|\sqrt{\pi n}\mathbf{P}\{Z_{2n} = j\} - e^{-j^2/4n}\right| \to 0, \quad n \to \infty.$$

① 亦可利用: 中国科学院数学研究所编的《常用数理统计表》, 科学出版社, 1974 年. ——译者

3. 证明泊松定理 (对于 $p = \lambda/n$) 的收敛速度为

$$\sup_k \left|P_n(k) - \frac{\lambda^k e^{-\lambda}}{k!}\right| \leqslant \frac{\lambda^2}{n}.$$

(证明可以参见第三章 §12.)

§7. 伯努利概型中"成功"概率的估计

1. "成功"概率估计的概念和性质 (相合性, 无偏性和有效性). 以上讨论的伯努利概型

$$(\Omega, \mathscr{A}, \mathbf{P}), \Omega = \{\omega : \omega = (x_1, \cdots, x_n), x_i = 0, 1\},$$
$$\mathscr{A} = \{A : A \subseteq \Omega\}, \mathbf{P}(\{\omega\}) = p(\omega),$$
$$p(\omega) = p^{\sum x_i}(1-p)^{n-\sum x_i},$$

假设 p ("成功"的概率) 的数值已知.

现在假设 p 事先未知, 并希望根据对试验结局的观测结果, 或由对随机变量 $\xi_1, \cdots, \xi_n$ 的观测结果来确定 p, 其中 $\xi_i(\omega) = x_i$. 这是*数理统计*的典型问题之一, 有不同的提法. 我们下面讨论问题的两种提法: *估计问题*和*建立置信区间问题*.

沿用数理统计中普遍采用的记号, 未知参数 p 记作 θ, 并认为是验前的 (a priori), 且 θ 的值属于集合 $\Theta = [0, 1]$. 通常称"由

$$\mathscr{E} = (\Omega, \mathscr{A}, \mathbf{P}_\theta, \theta \in \Theta), \quad \mathbf{P}_\theta(\{\omega\}) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

确定一个 (对应于'n 次独立试验'且'成功'概率为 $\theta \in \Theta$ 的) *概率 – 统计模型*", 而任何在 Θ 中取值的函数 $T_n = T_n(\omega)$ 称做*估计量*.

如果设

$$S_n = \xi_1 + \cdots + \xi_n, \quad T_n^* = \frac{S_n}{n},$$

则由大数定律可见, 估计量 T_n^* 称做*相合的*, 若对于任意 $\varepsilon > 0$, 有

$$\mathbf{P}_\theta\{|T_n^* - \theta| \geqslant \varepsilon\} \to 0, \quad n \to \infty. \tag{1}$$

此外, 估计量 T_n^* 称做*无偏的*, 如果对于任意 $\theta \in \Theta$,

$$\mathbf{E}_\theta T_n^* = \theta, \tag{2}$$

其中 $\mathbf{E}_\theta$ 是对应于概率 $\mathbf{P}_\theta$ 的数学期望.

估计量的无偏性是一条很自然的性质, 它反映如下事实: 由任何合理的估计量, 至少在平均意义下都应当得到所期待的结果. 不过, 估计量 T_n^* 并非唯一无偏估计量. 例如, 对于 $b_1 + \cdots + b_n = n$, 任何一个估计量

$$T_n = \frac{b_1\xi_1 + \cdots + b_n\xi_n}{n}$$

都是无偏的. 这些估计量都服从大数定律 (1) (至少对于 $|t_i| \leqslant K < \infty$). 从而, 这些估计量 T_n 也和 T_n^* 一样是 "好" 估计量.

于是, 产生一个问题: 如何比较不同的无偏估计量, 它们之中哪一个应称做最好的、最优的.

按估计量的本身的含义, 自然应当认为估计量对被估计的参数的偏差越小越好. 基于这样的考虑, 称估计量 $\tilde{T}_n$ (在无偏估计 T_n 类中) 为有效的, 如果

$$\mathbf{D}_\theta \tilde{T}_n = \inf_{T_n} \mathbf{D}_\theta T_n, \quad \theta \in \Theta, \tag{3}$$

其中 $\mathbf{D}_\theta T_n$ 是估计量 T_n 的方差, 即 $\mathbf{E}_\theta(T_n-\theta)^2$.

现在证明上面所考虑的估计量 T_n^* 是有效估计量. 事实上, 有

$$\mathbf{D}_\theta T_n^* = \mathbf{D}_\theta\left(\frac{S_n}{n}\right) = \frac{\mathbf{D}_\theta S_n}{n^2} = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}. \tag{4}$$

因此, 为证明估计量 T_n^* 有效, 只需证明

$$\inf_{T_n} \mathbf{D}_\theta T_n \geqslant \frac{\theta(1-\theta)}{n}. \tag{5}$$

对于 $\theta=0$ 和 1, 不等式显然. 现在设 $\theta \in (0,1)$ 且

$$p_\theta(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}.$$

显然 $\mathbf{P}_\theta(\{\omega\}) = p_\theta(\omega)$, 其中

$$p_\theta(\omega) = \prod_{i=1}^n p_\theta(x_i).$$

记

$$L_\theta(\omega) = \ln p_\theta(\omega).$$

那么

$$L_\theta(\omega) = \ln\theta \sum_{i=1}^n x_i + \ln(1-\theta)\sum_{i=1}^n (1-x_i),$$

$$\frac{\partial L_\theta(\omega)}{\partial\theta} = \frac{\sum_{i=1}^n (x_i-\theta)}{\theta(1-\theta)}.$$

因为

$$1 = \mathbf{E}_\theta 1 = \sum_\omega p_\theta(\omega)$$

且由估计量 T_n 的无偏性

$$\theta = \mathbf{E}_\theta T_n = \sum_\omega T_n(\omega) p_\theta(\omega),$$

所以经对 θ 求导, 得

$$0 = \sum_\omega \frac{\partial p_\theta(\omega)}{\partial\theta} = \sum_\omega \frac{\frac{\partial p_\theta(\omega)}{\partial\theta}}{p_\theta(\omega)} p_\theta(\omega) = \mathbf{E}_\theta\left[\frac{\partial L_\theta(\omega)}{\partial\theta}\right],$$

$$1 = \sum_\omega T_n \frac{\frac{\partial p_\theta(\omega)}{\partial\theta}}{p_\theta(\omega)} p_\theta(\omega) = \mathbf{E}_\theta\left[T_n \frac{\partial L_\theta(\omega)}{\partial\theta}\right].$$

因此,

$$1 = \mathbf{E}_\theta\left[(T_n-\theta)\frac{\partial L_\theta(\omega)}{\partial\theta}\right].$$

而根据柯西—布尼亚科夫斯基不等式

$$1 \leqslant \mathbf{E}_\theta(T_n-\theta)^2 \times \mathbf{E}_\theta\left[\frac{\partial L_\theta(\omega)}{\partial\theta}\right]^2.$$

于是,

$$\mathbf{E}_\theta(T_n-\theta)^2 \geqslant \frac{1}{I_n(\theta)}, \tag{6}$$

其中

$$I_n(\theta) = \mathbf{E}_\theta\left[\frac{\partial L_\theta(\omega)}{\partial\theta}\right]^2$$

称做费希尔信息量.

由 (6) 式得无偏估计量的所谓拉奥—克拉默 (C. R. Rao—G. Cramér) 不等式的特殊情形:

$$\inf_{T_n} \mathbf{D}_\theta T_n \geqslant \frac{1}{I_n(\theta)}. \tag{7}$$

对于所讨论的情形, 有

$$I_n(\theta) = \mathbf{E}_\theta\left[\frac{\partial L_\theta(\omega)}{\partial\theta}\right]^2 = \mathbf{E}_\theta\left[\frac{\sum_{i=1}^n(\xi_i-\theta)}{\theta(1-\theta)}\right]^2 = \frac{n\theta(1-\theta)}{[\theta(1-\theta)]^2} = \frac{n}{\theta(1-\theta)},$$

从而不等式 (5) 得证. 如我们曾提到的那样, 由此可见 $T_n^* = S_n/n$ 是未知参数 θ 的有效估计量.

2. "成功" 概率的置信区间 显然, 把 T_n^* 当作 θ 的 "点" 估计量, 我们就犯了某种错误. 甚至有可能出现这样的情形, 由观测数据 $x_1,\cdots,x_n$ 计算的数值 T_n^*, 对 θ 的真值有相当大的偏差. 因此最好再指出误差的大小.

不能指望对所有基本事件 ω, $T_n^* = T_n^*(\omega)$ 都能与未知参数 θ 的真值差异甚小, 这也是毫无意义的. 不过, 由大数定律, 知对于充分大的 n 和任意 $\delta > 0$, 事件 $\{|\theta - T_n^*| > \delta\}$ 的概率都充分小.

根据切比雪夫不等式, 有

$$\mathbf{P}_\theta\{|\theta-T_n^*|>\delta\}\leqslant\frac{\mathrm{D}_\theta T_n^*}{\delta^2}=\frac{\theta(1-\theta)}{n\delta^2},$$

因此, 对于任意 $\lambda>0$, 有

$$\mathbf{P}_\theta\left\{|\theta-T_n^*|\leqslant\lambda\sqrt{\frac{\theta(1-\theta)}{n}}\right\}\geqslant 1-\frac{1}{\lambda^2}.$$

例如, 取 $\lambda=3$, 则事件

$$\left\{|\theta-T_n^*|\leqslant 3\sqrt{\frac{\theta(1-\theta)}{n}}\right\}$$

出现的概率 $\mathbf{P}_\theta$ 大于 0.8888 $(1-1/3^2=8/9\approx 0.8888)$. 特别, 因为 $\theta(1-\theta)\leqslant 1/4$, 故事件

$$\left\{|\theta-T_n^*|\leqslant\frac{3}{2\sqrt{n}}\right\}$$

出现的概率 $\mathbf{P}_\theta$ 大于 0.8888.

于是,

$$\mathbf{P}_\theta\left\{|\theta-T_n^*|\leqslant\frac{3}{2\sqrt{n}}\right\}=\mathbf{P}_\theta\left\{T_n^*-\frac{3}{2\sqrt{n}}\leqslant\theta\leqslant T_n^*+\frac{3}{2\sqrt{n}}\right\}\geqslant 0.8888.$$

换句话说, 可以断定 "未知参数 θ 的真值" 以大于 0.8888 的概率属于区间

$$\left[T_n^*-\frac{3}{2\sqrt{n}},T_n^*+\frac{3}{2\sqrt{n}}\right].$$

有时将此命题简单地表示为

$$\theta\simeq T_n^*\pm\frac{3}{2\sqrt{n}}\ (\geqslant 88\%),$$

其中 "$\geqslant 88\%$" 表示 "在概率不小于 88% 的情形下".

区间

$$\left[T_n^*-\frac{3}{2\sqrt{n}},T_n^*+\frac{3}{2\sqrt{n}}\right]$$

就是 "未知参数的置信区间" 的一个例子.

定义 称形如

$$[\psi_1(\omega),\psi_2(\omega)]$$

的集合为置信度 $1-\delta$ 的置信区间 (或显著性水平为 δ 的置信区间), 如果对于一切 $\theta\in\Theta$, 有

$$\mathbf{P}_\theta\{\psi_1(\theta)\leqslant\theta\leqslant\psi_2(\theta)\}\geqslant 1-\delta,$$

其中 $\psi_1(\omega)$ 和 $\psi_2(\omega)$ 是基本事件的两个函数.

上面的讨论表明, 区间

$$\left[T_n^*-\frac{\lambda}{2\sqrt{n}},T_n^*+\frac{\lambda}{2\sqrt{n}}\right]$$

的置信度为 $1-1/\lambda^2$. 实际上, 置信度要远远高些, 因为由切比雪夫不等式只能得到事件概率的粗略估计. 为得到更精确的结果, 注意到

$$\left\{\omega:|\theta-T_n^*|\leqslant\lambda\sqrt{\frac{\theta(1-\theta)}{n}}\right\}=\{\omega:\psi_1(T_n^*,n)\leqslant\theta\leqslant\psi_2(T_n^*,n)\},$$

其中 $\psi_1=\psi_1(T_n^*,n)$ 和 $\psi_2=\psi_2(T_n^*,n)$ 是如下二次方程

$$(\theta-T_n^*)^2=\frac{\lambda^2}{n}\theta(1-\theta)$$

的根, 该方程描绘图 13 所示的椭圆. 现在记

$$F_\theta^n(x)=\mathbf{P}_\theta\left\{\frac{S_n-n\theta}{\sqrt{n\theta(1-\theta)}}\leqslant x\right\}.$$

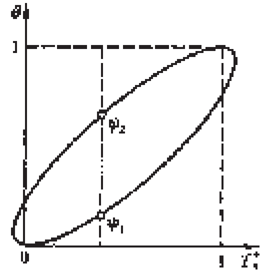


图 13

那么, 由 §6(24) 式, 可见

$$\sup_x|F_\theta^n(x)-\Phi(x)|\leqslant\frac{1}{\sqrt{n\theta(1-\theta)}}.$$

因此, 假如事先已知

$$0<\Delta\leqslant\theta\leqslant 1-\Delta<1,$$

其中 Δ 是某一常数, 则

$$\sup_x|F_\theta^n(x)-\Phi(x)|\leqslant\frac{1}{\Delta\sqrt{n}}.$$

从而

$$\mathbf{P}_\theta\{\psi_1(T_n^*,n)\leqslant\theta\leqslant\psi_2(T_n^*,n)\}=\mathbf{P}_\theta\left\{|\theta-T_n^*|\leqslant\lambda\sqrt{\frac{\theta(1-\theta)}{n}}\right\}$$

$$=\mathbf{P}_\theta\left\{\frac{|S_n-n\theta|}{\sqrt{n\theta(1-\theta)}}\leqslant\lambda\right\}\geqslant[2\Phi(\lambda)-1]-\frac{2}{\Delta\sqrt{n}}.$$

设 λ^* 是满足

$$[2\Phi(\lambda)-1]-\frac{2}{\Delta\sqrt{n}}\leqslant 1-\delta^*$$

的最小 λ 值, 其中 δ^* 是给定的显著性水平. 记 $\delta=\delta^*-2/(\Delta\sqrt{n})$, 则 λ^* 是如下方程的根:

$$\Phi(\lambda)=1-\frac{\delta}{2}.$$

当 n 较大时, 可以忽略项 $2/(\Delta\sqrt{n})$, 认为 λ^* 满足关系式

$$\Phi(\lambda^*) = 1 - \frac{\delta^*}{2}.$$

例如, 若 $\lambda^* = 3$, 则 $1-\delta^* = 0.9973\cdots$ 因此大致以概率 0.9973, 有

$$T_n^* - 3\sqrt{\frac{\theta(1-\theta)}{n}} \leqslant \theta \leqslant T_n^* + 3\sqrt{\frac{\theta(1-\theta)}{n}}, \tag{8}$$

而经迭代和忽略量级为 $O(n^{-3/4})$ 的项, 得

$$T_n^* - 3\sqrt{\frac{T_n^*(1-T_n^*)}{n}} \leqslant \theta \leqslant T_n^* + 3\sqrt{\frac{T_n^*(1-T_n^*)}{n}}. \tag{9}$$

由此可见, 对于充分大 n 的, 置信区间

$$\left[T_n^* - \frac{3}{2\sqrt{n}}, T_n^* + \frac{3}{2\sqrt{n}}\right]$$

的置信度为 0.9973, 然而由切比雪夫不等式得到的置信度只有 0.8889.

由此可以得到如下实用的结果. 假如进行大数量 N 系列试验, 每系列试验根据 n 次观测的结果估计参数 θ, 那么, 平均在 99.73% 的情形下, 每一系列 n 次试验, 估计量与参数真值的差别不大于 $3/(2\sqrt{n})$. (关于这一结果, 亦可参见 §5 的末尾.)

3. 练习题

1. 假设事先已知参数 θ 在 $\Theta_0 \subseteq [0,1]$ 中取值. 说明何时对于只在 Θ_0 中取值的参数 θ 存在无偏估计.

2. 在上题的条件下求拉奥-克拉默不等式, 并讨论估计量的有效性.

3. 在第 1 题的条件下, 讨论建立 θ 的置信区间的问题.

4. 在 §2 第 5 题的条件下, 假设 $N \gg M$, $N \gg n$, 讨论估计量 $\hat{N}$ 的无偏性和有效性. 建立 θ 的置信区间 (见 (8) 式和 (9) 式), N 的置信区间 $[\hat{N} - a(\hat{N}), \hat{N} - b(\hat{N})]$, 使

$$\mathbf{P}_{N,M,n}\{\hat{N} - a(\hat{N}) \leqslant N \leqslant \hat{N} + b(\hat{N})\} \approx 1 - \alpha,$$

其中 α 是某一较小的数.

§8. 关于分割的条件概率与条件数学期望

1. 条件概率　设 $(\Omega, \mathscr{A}, \mathbf{P})$ 是概率空间, 而

$$\mathscr{D} = \{D_1, \cdots, D_k\}$$

是 Ω 某一分割: $D_i \in \mathscr{A}, \mathbf{P}(D_i) > 0, i = 1, \cdots, k; D_1 + \cdots + D_k = \Omega$. 其次, 设事件 $A \in \mathscr{A}$, 而 $\mathbf{P}(A|D_i) > 0$, 是事件 A 关于事件 D_i 的条件概率.

一组条件概率 $\{\mathbf{P}(A|D_i), i = 1, \cdots, k\}$ 可以与随机变量

$$\pi(\omega) = \sum_{i=1}^{k} \mathbf{P}(A|D_i) I_{D_i}(\omega) \tag{1}$$

相联系 (与 §4 (5) 式比较), 且 $\pi(\omega)$ 在原子 D_i 上取 $\mathbf{P}(A|D_i)$ 为值. 为强调 $\pi(\omega)$ 确实与此分割 $\mathscr{D}$ 相联系, 将其记作

$$\mathbf{P}(A|\mathscr{D}) \quad 或 \quad \mathbf{P}(A|\mathscr{D})(\omega),$$

并称之为事件 A 关于分割 $\mathscr{D}$ 的条件概率.

这一概念, 以及将要引进的关于 σ-代数的条件概率的概念, 在概率论中起着重要的作用, 下面将逐步展开叙述.

条件概率的如下两条明确的性质:

$$\mathbf{P}(A + B|\mathscr{D}) = \mathbf{P}(A|\mathscr{D}) + \mathbf{P}(B|\mathscr{D}); \tag{2}$$

如果 $\mathscr{D}$ 是只含 Ω 中一个集合的平凡分割, 则

$$\mathbf{P}(A|\mathscr{D}) = \mathbf{P}(A) \tag{3}$$

把条件概率 $\mathbf{P}(A|\mathscr{D})$ 定义为随机变量, 就可以考虑其数学期望, 利用数学期望可以用如下紧凑的形式将 §3 中全概率公式 (3) 写成:

$$\mathbf{E}\mathbf{P}(A|\mathscr{D}) = \mathbf{P}(A). \tag{4}$$

事实上, 由于

$$\mathbf{P}(A|\mathscr{D})(\omega) = \sum_{i=1}^{k} \mathbf{P}(A|D_i) I_{D_i}(\omega),$$

则根据数学期望的定义 (见 §4 (5) 和 (6) 式), 有

$$\mathbf{E}\mathbf{P}(A|\mathscr{D}) = \sum_{i=1}^{k} \mathbf{P}(A|D_i)\mathbf{P}(D_i) = \sum_{i=1}^{k} \mathbf{P}(AD_i) = \mathbf{P}(A).$$

现在, 设 $\eta = \eta(\omega)$ 是以大于 0 的概率取 $y_1, \cdots, y_k$ 为值的随机变量:

$$\eta(\omega) = \sum_{j=1}^{k} y_j I_{D_j}(\omega),$$

其中 $D_j = \{\omega : \eta(\omega) = y_j\}$. 分割 $\mathscr{D}_\eta = \{D_1, \cdots, D_k\}$ 称做随机变量 η 诱导的分割. 条件概率 $\mathbf{P}(A|\mathscr{D}_\eta)$ 或 $\mathbf{P}(A|\mathscr{D}_\eta)(\omega)$ 称做事件 A 关于随机变量 η 的条件概率. 我们把 $\mathbf{P}(A|\eta = y_j)$ 也理解为条件概率 $\mathbf{P}(A|D_j)$, 其中 $D_j = \{\omega : \eta(\omega) = y_j\}$.

类似地，若 $\eta_1,\eta_2,\cdots,\eta_m$ 是随机变量，而 $\mathscr{D}_{\eta_1,\eta_2,\cdots,\eta_m}$ 是诱导的以

$$D_{y_1,y_2,\cdots,y_m}=\{\omega:\eta_1=y_1,\eta_2=y_2,\cdots,\eta_m=y_m\}$$

为原子的分割，则 $\mathbf{P}(A|\mathscr{D}_{\eta_1,\eta_2,\cdots,\eta_m})$ 记作 $\mathbf{P}(A|\eta_1,\eta_2,\cdots,\eta_m)$，并称为事件 A 关于随机变量 $\eta_1,\eta_2,\cdots,\eta_m$ 的条件概率.

例 1 设 ξ 和 η 是两个独立同分布随机变量，分别以概率 p 和 q 取 1 和 0 为值. 对于 $k=0,1,2$，我们现在求事件 $A=\{\omega:\xi+\eta=k\}(k=0,1,2)$ 关于 η 的条件概率 $\mathbf{P}(A|\eta)$.

为此，首先注意到有用的一般事实：如果 ξ 和 η 是两个独立随机变量，则当 $\mathbf{P}\{\eta=y\}>0$ 时，有

$$\mathbf{P}\{\xi+\eta=z|\eta=y\}=\mathbf{P}\{\xi+y=z\}. \tag{5}$$

事实上，

$$\begin{aligned}\mathbf{P}\{\xi+\eta=z|\eta=y\}&=\frac{\mathbf{P}\{\xi+\eta=z,\eta=y\}}{\mathbf{P}\{\eta=y\}}=\frac{\mathbf{P}\{\xi+y=z,\eta=y\}}{\mathbf{P}\{\eta=y\}}\\&=\frac{\mathbf{P}\{\xi+y=z\}\mathbf{P}\{\eta=y\}}{\mathbf{P}\{\eta=y\}}=\mathbf{P}\{\xi+y=z\}.\end{aligned}$$

由此公式，可见

$$\begin{aligned}\mathbf{P}(A|\eta)(\omega)&=\mathbf{P}(\xi+\eta=k|\eta)(\omega)\\&=\mathbf{P}(\xi+\eta=k|\eta=0)I_{\{\eta=0\}}(\omega)+\mathbf{P}(\xi+\eta=k|\eta=1)I_{\{\eta=1\}}(\omega)\\&=\mathbf{P}(\xi=k)I_{\{\eta=0\}}(\omega)+\mathbf{P}(\xi=k-1)I_{\{\eta=1\}}(\omega).\end{aligned}$$

于是

$$\mathbf{P}(\xi+\eta=k|\eta)(\omega)=\begin{cases}qI_{\{\eta=0\}}(\omega), & 若\ k=0,\\ pI_{\{\eta=0\}}(\omega)+qI_{\{\eta=1\}}(\omega), & 若\ k=1,\\ pI_{\{\eta=1\}}(\omega), & 若\ k=2,\end{cases} \tag{6}$$

或同样地

$$\mathbf{P}(\xi+\eta=k|\eta)=\begin{cases}q(1-\eta), & 若\ k=0,\\ p(1-\eta)+q\eta, & 若\ k=1,\\ p\eta, & 若\ k=2.\end{cases} \tag{7}$$

2. 条件数学期望 设 $\xi=\xi(\omega)$ 是随机变量，其值域为 $X=\{x_1,\cdots,x_l\}$：

$$\xi=\sum_{j=1}^{l}x_jI_{A_j},\quad A_j=\{\omega:\xi=x_j\},$$

而 $\mathscr{D}=\{D_1,\cdots,D_k\}$ 是某个分割. 曾经定义了 ξ 关于概率 $\mathbf{P}(A_j)(j=1,\cdots,l)$ 的数学期望：

$$\mathbf{E}\xi=\sum_{j=1}^{l}x_j\mathbf{P}(A_j). \tag{8}$$

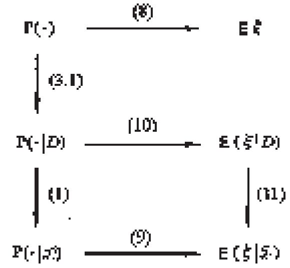


图 14

自然类似地利用概率 $\mathbf{P}(A_j|\mathscr{D})$，定义随机变量 ξ 关于分割 $\mathscr{D}$ 的条件数学期望，记作 $\mathbf{E}(\xi|\mathscr{D})$ 或 $\mathbf{E}(\xi|\mathscr{D})(\omega)$：

$$\mathbf{E}(\xi|\mathscr{D})=\sum_{j=1}^{l}x_j\mathbf{P}(A_j|\mathscr{D}). \tag{9}$$

根据这一定义，条件数学期望 $\mathbf{E}(\xi|\mathscr{D})(\omega)$ 是一随机变量，对于属于同一原子 D_i 的基本事件 ω，取同一个值

$$\sum_{j=1}^{l}x_j\mathbf{P}(A_j|D_i).$$

这一事实说明，可以由另一途径定义条件数学期望. 具体地说，首先由公式

$$\mathbf{E}(\xi|D_i)=\sum_{j=1}^{l}x_j\mathbf{P}(A_j|D_i)\left(=\frac{\mathbf{E}(\xi I_{D_i})}{\mathbf{P}(D_i)}\right) \tag{10}$$

定义随机变量 ξ 关于事件 D_i 的条件数学期望，然后按定义设

$$\mathbf{E}(\xi|\mathscr{D})(\omega)=\sum_{i=1}^{k}\mathbf{E}(\xi|D_i)I_{D_i}(\omega) \tag{11}$$

(参见流程图 14).

注意，$\mathbf{E}(\xi|D)$ 和 $\mathbf{E}(\xi|\mathscr{D})(\omega)$ 的值与随机变量 ξ 的表示方法无关.

下面指出的条件数学期望的性质，可以直接由其定义得到：

$$\mathbf{E}(a\xi+b\eta|\mathscr{D})=a\mathbf{E}(\xi|\mathscr{D})+b\mathbf{E}(\eta|\mathscr{D})\ (a,b\ 是常数); \tag{12}$$

$$\mathbf{E}(\xi|\Omega)=\mathbf{E}\xi; \tag{13}$$

$$\mathbf{E}(C|\mathscr{D})=C,\quad C\ 是常数. \tag{14}$$

如果 $\xi=I_A(\omega)$，则

$$\mathbf{E}(\xi|\mathscr{D})=\mathbf{P}(A|\mathscr{D}). \tag{15}$$

其中，最后一条性质表明，由条件数学期望的性质，可以直接得出条件概率的性质.

另一重要的性质是全概率公式 (4) 的推广：

$$\mathbf{E}\mathbf{E}(\xi|\mathscr{D})=\mathbf{E}\xi. \tag{16}$$

为证明此式, 只需注意到, 由 (4) 式, 可见

$$\mathbf{E}\mathbf{E}(\xi|\mathscr{D}) = \mathbf{E}\sum_{j=1}^{l} x_j \mathbf{P}(A_j|\mathscr{D}) = \sum_{j=1}^{l} x_j \mathbf{E}\mathbf{P}(A_j|\mathscr{D}) = \sum_{j=1}^{l} x_j \mathbf{P}(A_j) = \mathbf{E}\xi.$$

设 $\mathscr{D}=\{D_1,\cdots,D_k\}$ 是某个分割, 而 $\eta=\eta(\omega)$ 是某一随机变量. 我们称随机变量 η 关于此分割为可测的或 $\mathscr{D}$-可测的, 如果 $\mathscr{D}_\eta \preccurlyeq \mathscr{D}$, 即 $\eta=\eta(\omega)$ 可以表示为

$$\eta(\omega)=\sum_{i=1}^{k} y_i I_{D_i}(\omega),$$

其中有些 y_i 值可能相等. 换句话说, 随机变量 η 为 $\mathscr{D}$-可测的, 当且仅当它在分割 $\mathscr{D}$ 的原子上取常数值.

例 2 如果 $\mathscr{D}$ 为平凡分割: $\mathscr{D}=\{\Omega\}$, 则随机变量 η 为 $\mathscr{D}$-可测的, 当且仅当 $\eta\equiv C$, 其中 C 为常数. 任何随机变量 η 关于分割 $\mathscr{D}_\eta$-可测.

假设随机变量 η 为 $\mathscr{D}$-可测, 那么

$$\mathbf{E}(\xi\eta|\mathscr{D})=\eta\mathbf{E}(\xi|\mathscr{D}). \tag{17}$$

特别

$$\mathbf{E}(\eta|\mathscr{D})=\eta \quad (\mathbf{E}(\eta|\mathscr{D}_\eta)=\eta). \tag{18}$$

为证明 (17) 式注意到, 如果

$$\xi=\sum_{j=1}^{l} x_j I_{A_j},\quad A_j=\{\omega:\xi=x_j\},$$

则

$$\xi\eta=\sum_{j=1}^{l}\sum_{i=1}^{k} x_j y_i I_{A_j D_i},$$

即

$$\mathbf{E}(\xi\eta|\mathscr{D}) = \sum_{j=1}^{l}\sum_{i=1}^{k} x_j y_i \mathbf{P}(A_jD_i|\mathscr{D}) = \sum_{j=1}^{l}\sum_{i=1}^{k} x_j y_i \sum_{m=1}^{k} \mathbf{P}(A_jD_i|D_m)I_{D_m}(\omega)$$
$$= \sum_{j=1}^{l}\sum_{i=1}^{k} x_j y_i \mathbf{P}(A_jD_i|D_i)I_{D_i}(\omega) = \sum_{j=1}^{l}\sum_{i=1}^{k} x_j y_i \mathbf{P}(A_j|D_i)I_{D_i}(\omega). \tag{19}$$

另一方面, 考虑到 $I_{D_i}^2=I_{D_i}$ 和 $I_{D_i}I_{D_m}=0, i\neq m$, 有

$$\eta\mathbf{E}(\xi|\mathscr{D}) = \left[\sum_{i=1}^{k} y_i I_{D_i}(\omega)\right]\times\left[\sum_{j=1}^{l} x_j\mathbf{P}(A_j|\mathscr{D})\right]$$
$$= \left[\sum_{i=1}^{k} y_i I_{D_i}(\omega)\right]\times\sum_{m=1}^{k}\left[\sum_{j=1}^{l} x_j \mathbf{P}(A_j|D_m)\right]\cdot I_{D_m}(\omega)$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{l} x_j y_i \mathbf{P}(A_j|D_i)I_{D_i}(\omega).$$

于是, (19) 式与 (17) 式一同得证.

我们再证明条件数学期望一条重要性质. 设 $\mathscr{D}_1$ 和 $\mathscr{D}_2$ 是两个分割, 且 $\mathscr{D}_1\preccurlyeq\mathscr{D}_2$ (分割 $\mathscr{D}_1$ 比分割 $\mathscr{D}_2$ "细小"). 那么, "望远" 性质成立:

$$\mathbf{E}[\mathbf{E}(\xi|\mathscr{D}_2)|\mathscr{D}_1]=\mathbf{E}(\xi|\mathscr{D}_1). \tag{20}$$

为证明 (20) 式, 设 $\mathscr{D}_1=\{D_{11},\cdots,D_{1m}\}$, $\mathscr{D}_2=\{D_{21},\cdots,D_{2n}\}$. 那么, 若

$$\xi=\sum_{j=1}^{l} x_j I_{A_j},$$

则

$$\mathbf{E}(\xi|\mathscr{D}_2)=\sum_{j=1}^{l} x_j\mathbf{P}(A_j|\mathscr{D}_2),$$

且只需证明

$$\mathbf{E}[\mathbf{P}(A_j|\mathscr{D}_2)|\mathscr{D}_1]=\mathbf{P}(A_j|\mathscr{D}_1). \tag{21}$$

由于

$$\mathbf{P}(A_j|\mathscr{D}_2)=\sum_{q=1}^{n} \mathbf{P}(A_j|D_{2q})I_{D_{2q}},$$

则

$$\begin{aligned}
\mathbf{E}[\mathbf{P}(A_j|\mathscr{D}_2)|\mathscr{D}_1] &= \sum_{q=1}^{n}\mathbf{P}(A_j|D_{2q})\mathbf{P}(D_{2q}|\mathscr{D}_1)\\
&= \sum_{q=1}^{n}\mathbf{P}(A_j|D_{2q})\left[\sum_{p=1}^{m}\mathbf{P}(D_{2q}|D_{1p})I_{D_{1p}}\right]\\
&= \sum_{p=1}^{m}I_{D_{1p}}\sum_{q=1}^{n}\mathbf{P}(A_j|D_{2q})\mathbf{P}(D_{2q}|D_{1p})\\
&= \sum_{p=1}^{m}I_{D_{1p}}\sum_{\{q:D_{2q}\subseteq D_{1p}\}}\mathbf{P}(A_j|D_{2q})\mathbf{P}(D_{2q}|D_{1p})\\
&= \sum_{p=1}^{m}I_{D_{1p}}\sum_{\{q:D_{2q}\subseteq D_{1p}\}}\frac{\mathbf{P}(A_jD_{2q})}{\mathbf{P}(D_{2q})}\times\frac{\mathbf{P}(D_{2q})}{\mathbf{P}(D_{1p})}\\
&= \sum_{p=1}^{m}I_{D_{1p}}\mathbf{P}(A_j|D_{1p}) = \mathbf{P}(A_j|\mathscr{D}_1).
\end{aligned}$$

于是, (21) 式得证.

假如分割 $\mathscr{D}$ 是由随机变量 $\eta_1,\cdots,\eta_k$ 诱导的 $(\mathscr{D}=\mathscr{D}_{\eta_1,\cdots,\eta_k})$, 则条件数学期望 $\mathbf{E}(\xi|\mathscr{D}_{\eta_1,\cdots,\eta_k})$ 称做随机变量 ξ 关于 $\eta_1,\cdots,\eta_k$ 的条件数学期望, 记作 $\mathbf{E}(\xi|\eta_1,\cdots,\eta_k)$ 或 $\mathbf{E}(\xi|\eta_1,\cdots,\eta_k)(\omega)$.

直接由 $\mathbf{E}(\xi|\eta)$ 的定义可见, 如果 ξ 和 η 独立, 则

$$\mathbf{E}(\xi|\eta)=\mathbf{E}\xi. \tag{22}$$

由 (18) 式, 可见

$$\mathbf{E}(\eta|\eta)=\eta. \tag{23}$$

性质 (22) 有如下推广. 设随机变量 ξ 与分割 $\mathscr{D}$ 独立 (即对于任意 $D_i\in\mathscr{D}$, 随机变量 ξ 与 I_{D_i} 独立). 那么,

$$\mathbf{E}(\xi|\mathscr{D})=\mathbf{E}\xi. \tag{24}$$

作为特殊情形, 由 (20) 式, 可以得到如下有用的公式:

$$\mathbf{E}[\mathbf{E}(\xi|\eta_1,\eta_2)|\eta_1]=\mathbf{E}(\xi|\eta_1). \tag{25}$$

例 3 对于例 1 中的随机变量 ξ 和 η, 求 $\mathbf{E}(\xi+\eta|\eta)$. 由 (22) 式和 (23) 式, 有

$$\mathbf{E}(\xi+\eta|\eta)=\mathbf{E}\xi+\eta=p+\eta.$$

由 (8) 式亦可得到此结果:

$$\mathbf{E}(\xi+\eta|\eta)=\sum_{k=0}^{2}k\mathbf{P}(\xi+\eta=k|\eta)=p(1-\eta)+q\eta+2p\eta=p+\eta.$$

例 4 设 ξ 和 η 是独立同分布随机变量. 验证,

$$\mathbf{E}(\xi|\xi+\eta)=\mathbf{E}(\eta|\xi+\eta)=\frac{\xi+\eta}{2}. \tag{26}$$

事实上, 为简便计, 假设 ξ 和 η 取 $1,2,\cdots,m$ 为值; 对于 $1\leqslant k\leqslant m,\ 2\leqslant l\leqslant 2m$, 有

$$\begin{aligned}
\mathbf{P}(\xi=k|\xi+\eta=l) &= \frac{\mathbf{P}\{\xi=k,\xi+\eta=l\}}{\mathbf{P}\{\xi+\eta=l\}}=\frac{\mathbf{P}\{\xi=k,\eta=l-k\}}{\mathbf{P}\{\xi+\eta=l\}}\\
&= \frac{\mathbf{P}\{\xi=k\}\mathbf{P}\{\eta=l-k\}}{\mathbf{P}\{\xi+\eta=l\}}=\frac{\mathbf{P}\{\eta=k\}\mathbf{P}\{\xi=l-k\}}{\mathbf{P}\{\xi+\eta=l\}}\\
&= \mathbf{P}(\eta=k|\xi+\eta=l).
\end{aligned}$$

从而 (26) 的第一个等式得证. 为证明第二个等式, 只需注意到, 由于 ξ 和 η 同分布,

$$2\mathbf{E}(\xi|\xi+\eta)=\mathbf{E}(\xi|\xi+\eta)+\mathbf{E}(\eta|\xi+\eta)=\mathbf{E}(\xi+\eta|\xi+\eta)=\xi+\eta.$$

3. 关于 $\mathbf{E}(\xi|\mathscr{B})$ 和 $\mathbf{E}(\xi|\mathscr{D})$ 在 §1 曾指出, 有限集合 Ω 的每一个分割 $\mathscr{D}=\{D_1,\cdots,D_k\}$ 对应着 Ω 子集的一个代数 $\alpha(\mathscr{D})$. 恰好相反, 有限空间 Ω 的任何代数 $\mathscr{B}$ 可以由某一个分割 $\mathscr{D}$ 派生: $\mathscr{B}=\alpha(\mathscr{D})$. 因此在有限空间 Ω 的任何代数与分割之间存在一一对应关系. 这一事实是指, 以后将引进的关于特殊集系 (σ−代数) 的条件数学期望的概念.

对于有限空间, 代数和 $\sigma-$ 代数的概念完全相同. 在这种情形下, 如果 $\mathscr{B}$ 是代数, 则以后 (在第二章 §7) 引进的随机变量 ξ 关于代数 $\mathscr{B}$ 的条件数学期望 $\mathbf{E}(\xi|\mathscr{B})$, 千能与 $\mathbf{E}(\xi|\mathscr{D})$ 一致. 其中 $\mathbf{E}(\xi|\mathscr{D})$ 是 ξ 关于分割 $\mathscr{D}$ (即关于 $\sigma-$ 代数 $\mathscr{B}=\alpha(\mathscr{D})$) 的条件数学期望. 实际上, 对于有限空间, 我们以后不再区分 $\mathbf{E}(\xi|\mathscr{B})$ 和 $\mathbf{E}(\xi|\mathscr{D})$: 认为 $\mathbf{E}(\xi|\mathscr{B})$ 按定义就是 $\mathbf{E}(\xi|\mathscr{D})$.

4. 练习题

1. 试举一例: 随机变量 ξ 和 η 不独立, 但是

$$\mathbf{E}(\xi|\eta)=\mathbf{E}\xi.$$

(与命题 (32) 对照.)

2. 随机变量

$$\mathbf{D}(\xi|\mathscr{D})=\mathbf{E}\{(\xi-\mathbf{E}(\xi|\mathscr{D}))^2|\mathscr{D}\}$$

称做关于分割 $\mathscr{D}$ 的*条件方差*.

证明方差

$$\mathbf{D}\xi=\mathbf{E}\mathbf{D}(\xi|\mathscr{D})+\mathbf{D}\mathbf{E}(\xi|\mathscr{D}).$$

3. 基于 (17) 式, 证明: 对于任意函数 $f=f(\eta)$, 条件数学期望具有如下性质:

$$\mathbf{E}[f(\eta)\mathbf{E}(\xi|\eta)]=\mathbf{E}[\xi f(\eta)].$$

4. 设 ξ 和 η 是随机变量. 证明

$$\inf_f \mathbf{E}[\eta - f(\xi)]^2,$$

在函数 $f^*(\xi) = \mathbf{E}(\eta|\xi)$ 上达到下确界, 即条件数学期望 $\mathbf{E}(\eta|\xi)$, 在均方意义下是 η 对 ξ 的最优估计量.

5. 设 $\xi_1, \cdots, \xi_k, \tau$ 是独立随机变量, 而 $\xi_1, \cdots, \xi_k$ 同分布, 且 τ 的可能值为 $1, 2, \cdots, n$. 证明, 对于随机个随机变量之和 $S_\tau = \xi_1 + \cdots + \xi_\tau$, 有

$$\mathbf{E}(S_\tau|\tau) = \tau\mathbf{E}\xi_1, \quad \mathbf{D}(S_\tau|\tau) = \tau\mathbf{D}\xi_1$$

和

$$\mathbf{E}S_\tau = \mathbf{E}\tau \times \mathbf{E}\xi_1, \quad \mathbf{D}S_\tau = \mathbf{E}\tau \times \mathbf{D}\xi_1 + \mathbf{D}\tau \times (\mathbf{E}\xi_1)^2.$$

6. 证明等式 (24).

§9. 随机游动 I. 掷硬币博弈的破产概率和平均持续时间

1. 关于随机游动 在 §6 中证明的伯努利概型的极限定理, 远远不只是给出了计算概率 $\mathbf{P}\{S_n = k\}$ 和 $\mathbf{P}\{A < S_n \leqslant B\}$ 的方便公式. 其意义还在于, 这些定理具有适合多方面应用的特点, 也就是说, 这些定理不仅适用于只有两个可能值的伯努利随机变量 $\xi_1, \xi_2, \cdots$, 而且适用于更为一般类源的随机变量. 因此, 伯努利概型作为一种简单的模型, 可以导出很多更为一般模型所固有的许多概率规律性.

在这一节和下一节, 将研究一系列新的概率规律性, 有些甚至带有非常不可预测的特点. 全部研究仍然局限于由伯努利概型所描绘的随机游动, 不过许多结论对于更一般情形的随机游动仍然成立.

2. 二人博弈和破产概率 考虑伯努利概型: $(\Omega, \mathscr{A}, \mathbf{P})$, 其中

$$\Omega = \{\omega : \omega = (x_1, \cdots, x_n), x_i = \pm 1\}, \mathscr{A} = \{A : A \subseteq \Omega\}$$
$$\mathbf{P}(\{\omega\}) = p^{\nu(\omega)} q^{n-\nu(\omega)}, \quad \nu(\omega) = \frac{\sum x_i + n}{2},$$

其中 $\mathscr{A}$ 是 Ω 的一切子集的集系. 设 $\xi_i(\omega) = x_i (i = 1, \cdots, n)$, 则 $\xi_1, \cdots, \xi_n$ 是独立伯努利随机变量序列:

$$\mathbf{P}\{\xi_i = 1\} = p, \quad \mathbf{P}\{\xi_i = -1\} = q, \quad p + q = 1.$$

设 $S_0 = 0, S_k = \xi_1 + \cdots + \xi_k, 1 \leqslant k \leqslant n$. 序列 $(S_k)_{k \leqslant n}$ 可以视为由 0 出发的某"质点"随机游动的轨道. 这时 $S_{k+1} = S_k + \xi_{k+1}$, 即如果质点于 k 时在点 S_k, 则于 $k+1$ 时质点或者以概率 p 向上移动一步, 或者以概率 q 向下移动一步.

设 A 和 $B(A \leqslant 0 \leqslant B)$ 是两个整数. 与所考虑的随机游动相联系的有趣的问题之一就是, 随机游动的质点经 n 步越出区间 (A, B) 的概率如何? 同样有趣的问题是, 经 n 步在点 A 或点 B 越出区间 (A, B) 的概率如何?

如果用博弈来说明, 则该问题的自然性就显得格外清楚. 假设有两人 (甲和乙) 对弈, 开始他们的"赌金"相等, 分别为 $(-A)$ 和 B. 若 $\xi_i = +1$, 就认为乙付给甲一单位赌金; 若 $\xi_i = -1$, 则相反甲付给乙一单位赌金. 那么, $S_k = \xi_1 + \cdots + \xi_k$ 表示经 k 局甲赢得乙的数额 ($S_k < 0$ 实际上表示甲输给乙的数额).

在 $k \leqslant n$ 时, 即当首次 $S_k = B(S_k = A)$ 时, 乙 (甲) 的赌金成为 0, 即出现乙 (甲) 破产. (如果 $k < n$, 则应该认为在时刻 k 对弈停止, 尽管在 n 时前包括 n 时游动是确定的.)

在给出问题的确切表述之前, 首先引进一系列记号.

设 x 是区间 $[A, B]$ 上的整数, 而对于 $0 \leqslant k \leqslant n$, 设 $S_k^x = x + S_k$,

$$\tau_k^x = \min\{0 \leqslant l \leqslant k : S_l^x = A 或 B\}, \tag{1}$$

其中约定: 如果对于一切 $0 \leqslant l \leqslant k$, 有 $A < S_l^x < B$, 则认为 $\tau_k^x = k$.

对于每一个 $0 \leqslant k \leqslant n$ 和 $x \in [A, B]$, 时刻 τ_k^x 称做停止时间 (见 §11), 是一定义在基本事件空间 Ω 上的整数值随机变量 (注意, 没有明显地标出 τ_k^x 依赖于 ω).

显然, 对于一切 $l < k$, 集合 $\{\omega : \tau_k^x = l\}$ 是一事件: "于 0 时始于点 x 的随机游动 $\{S_i^x : 0 \leqslant i \leqslant k\}$, 于 l 时越出区间 (A, B)". 同样显然, 对于 $l \leqslant k$, 集合 $\{\omega : \tau_k^x = l, S_l^x = A\}$ 与 $\{\omega : \tau_k^x = l, S_l^x = B\}$ 是事件 "游动的质点于 l 时相应地在点 A 和 B 越出区间 (A, B)".

对于一切 $0 \leqslant k \leqslant n$, 设

$$\mathscr{A}_k^x = \sum_{0 \leqslant l \leqslant k} \{\omega : \tau_k^x = l, S_l^x = A\}, \tag{2}$$
$$\mathscr{B}_k^x = \sum_{0 \leqslant l \leqslant k} \{\omega : \tau_k^x = l, S_l^x = B\};$$

且设

$$\alpha_k(x) = \mathbf{P}(\mathscr{A}_k^x), \quad \beta_k(x) = \mathbf{P}(\mathscr{B}_k^x)$$

是经时间段 $[0, k]$ 分别于点 A 和 B 处越出区间 (A, B) 的概率. 对于这些概率可以得到递推关系式, 并由此依次求出 $\alpha_1(x), \cdots, \alpha_n(x)$ 和 $\beta_1(x), \cdots, \beta_n(x)$.

这样, 设 $A < x < B$. 显然 $\alpha_0(x) = \beta_0(x) = 0$. 设 $1 \leqslant k \leqslant n$, 那么由 §8.(3) 式, 有

$$\begin{aligned}\beta_k(x) &= \mathbf{P}(\mathscr{B}_k^x) \\ &= \mathbf{P}(\mathscr{B}_k^x|S_1^x = x+1)\mathbf{P}\{\xi_1 = 1\} + \mathbf{P}(\mathscr{B}_k^x|S_1^x = x-1)\mathbf{P}\{\xi_1 = -1\} \\ &= p\mathbf{P}(\mathscr{B}_k^x|S_1^x = x+1) + q\mathbf{P}(\mathscr{B}_k^x|S_1^x = x-1).\end{aligned} \tag{3}$$

现在证明

$$\mathbf{P}(\mathscr{B}_k^x|S_1^x = x+1) = \mathbf{P}(\mathscr{B}_{k-1}^{x+1}), \quad \mathbf{P}(\mathscr{B}_k^x|S_1^x = x-1) = \mathbf{P}(\mathscr{B}_{k-1}^{x-1}).$$

为此注意到, 集合 $\mathscr{B}_k^x$ 可以表示为

$$\mathscr{B}_k^x = \{\omega : (x, x+\xi_1, \cdots, x+\xi_1+\cdots+\xi_k) \in B_k^x\},$$

其中 B_k^x 是形如

$$(x, x+x_1, \cdots, x+x_1+\cdots+x_k), \quad x_i = \pm 1$$

轨道的集合, 这样的轨道是在时段 $[0, k]$ 内, 在点 B 首次越出区间 (A, B) (见图 15).

把集合 B_k^x 表示为 $B_k^{x,x+1} + B_k^{x,x-1}$, 其中 $B_k^{x,x+1}$ 和 $B_k^{x,x-1}$ 是 B_k^x 中分别对应于 $x_1 = +1$ 和 $x_1 = -1$ 轨道.

易见, $B_k^{x,x+1}$ 中的每一条轨道与轨道

$$(x+1, x+1+x_2, \cdots, x+1+x_2+\cdots+x_k),$$

一一对应. 对于 $B_k^{x,x-1}$ 的轨道, 也有同样的性质. 由于这一性质, 以及随机变量 $\xi_1, \cdots, \xi_k$ 独立同分布和 §8 式 (6), 有

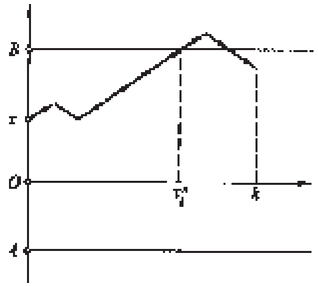


图 15 B_k^x 轨道的例

$$\begin{aligned}\mathbf{P}(\mathscr{B}_k^x|S_1^x = x+1) &= \mathbf{P}(\mathscr{B}_k^x|\xi_1 = 1)\\ &= \mathbf{P}\{(x, x+\xi_1, \cdots, x+\xi_1+\cdots+\xi_k) \in B_k^x|\xi_1 = 1\}\\ &= \mathbf{P}\{(x+1, x+1+\xi_2, \cdots, x+1+\xi_2+\cdots+\xi_k) \in B_{k-1}^{x+1}\}\\ &= \mathbf{P}\{(x+1, x+1+\xi_1, \cdots, x+1+\xi_1+\cdots+\xi_{k-1}) \in B_{k-1}^{x+1}\}\\ &= \mathbf{P}(\mathscr{B}_{k-1}^{x+1}).\end{aligned}$$

同样地

$$\mathbf{P}(\mathscr{B}_k^x|S_1^x = x-1) = \mathbf{P}(\mathscr{B}_{k-1}^{x-1}).$$

因此, 对于 $x \in (A, B)$ 和 $k \leqslant n$, 由 (3) 式, 有

$$\beta_k(x) = p\beta_{k-1}(x+1) + q\beta_{k-1}(x-1), \tag{4}$$

其中

$$\beta_l(B) = 1, \quad \beta_l(A) = 0, \quad 0 \leqslant l \leqslant n. \tag{5}$$

同理

$$\alpha_k(x) = p\alpha_{k-1}(x+1) + q\alpha_{k-1}(x-1), \tag{6}$$

其中

$$\alpha_l(A) = 1, \quad \alpha_l(B) = 0, \quad 0 \leqslant l \leqslant n.$$

由于 $\alpha_0(x) = \beta_0(x) = 0, x \in (A, B)$, 则所得递推公式可以 (至少原则上可以) 用来求概率 $\alpha_1(x), \cdots, \alpha_n(x)$ 和 $\beta_1(x), \cdots, \beta_n(x)$. 暂时不作具体计算, 而考虑当 n 充分大时这些概率的值的问题.

为此注意到, 由于 $\mathscr{B}_{k-1}^x \subset \mathscr{B}_k^x, k \leqslant n$, 可见 $\beta_{k-1}(x) \leqslant \beta_k(x) \leqslant 1$. 因此自然想到 (实际上确实是这样, 见下面第 3 小节), 当 n 充分大时概率 $\beta_n(x)$ 接近如下方程的解 $\beta(x)$:

$$\beta(x) = p\beta(x+1) + q\beta(x-1), \tag{7}$$

且满足由 (4) 和 (5) 两式求极限所得的边界条件:

$$\beta(B) = 1, \ \beta(A) = 0. \tag{8}$$

为求解问题 (7) 和 (8), 首先假设 $p \neq q$. 不难看到, 所考虑的方程有两个特解: a 和 $b(q/p)^x$, 其中 a 和 b 是常数. 因此, 我们以如下形式求通解 $\beta(x)$:

$$\beta(x) = a + b(q/p)^x. \tag{9}$$

考虑到 (8) 式, 对于一切 $A \leqslant x \leqslant B$, 有

$$\beta(x) = \frac{(q/p)^x - (q/p)^A}{(q/p)^B - (q/p)^A}. \tag{10}$$

现在证明这是该问题的唯一解. 为此, 只需证明, 问题 (7) 和 (8) 式的一切解都可以表示为 (9) 的形式.

设 $\tilde{\beta}(x)$ 是满足 $\tilde{\beta}(A) = 0, \tilde{\beta}(B) = 1$ 的某个解, 总存在常数 $\tilde{a}$ 和 $\tilde{b}$, 使

$$\tilde{a} + \tilde{b}(q/p)^A = \tilde{\beta}(A), \quad \tilde{a} + \tilde{b}(q/p)^{A+1} = \tilde{\beta}(A+1).$$

因此由方程 (7) 可见

$$\tilde{\beta}(A+2) = \tilde{a} + \tilde{b}(q/p)^{A+2},$$

且一般

$$\tilde{\beta}(x) = \tilde{a} + \tilde{b}(q/p)^x.$$

于是, 所求得的解 (10) 是所考虑问题的唯一解.

由类似的讨论可见, 方程

$$\alpha(x) = p\alpha(x+1) + q\alpha(x-1), \quad x \in (A, B) \tag{11}$$

满足边界条件

$$\alpha(A) = 1, \quad \alpha(B) = 0 \tag{12}$$

的唯一解是

$$\alpha(x) = \frac{(q/p)^B - (q/p)^x}{(q/p)^B - (q/p)^A}, \quad A \leqslant x \leqslant B. \tag{13}$$

假如 $p = q = 1/2$，则问题 (7)、(8) 和 (11)、(12) 唯一解 $\beta(x)$ 和 $\alpha(x)$ 相应为：

$$\beta(x) = \frac{x - A}{B - A} \tag{14}$$

和

$$\alpha(x) = \frac{B - x}{B - A}. \tag{15}$$

注意，对于任意 $0 \leqslant p \leqslant 1$，

$$\alpha(x) + \beta(x) = 1. \tag{16}$$

假如对局二人甲和乙的最初的"赌金"分别为 $x - A$ 和 $B - x$，对于无限多局的对弈，宜把 $\alpha(x)$ 和 $\beta(x)$ 相应地称做对局二人甲和乙的破产概率，当然应假定存在无限独立伯努利随机变量序列 $\xi_1, \xi_2, \cdots$，其中 $\xi_i = +1$ 表示甲赢，而 $\xi_i = -1$ 表示甲输。为在本小节开始考虑的概率空间 $(\Omega, \mathscr{A}, \mathbf{P})$ 上，存在上述独立伯努利随机变量序列，$(\Omega, \mathscr{A}, \mathbf{P})$ 显得太"贫乏"。以后（第二章 §9）我们会看到，确实可以建立这样的无限序列，而且 $\alpha(x)$ 和 $\beta(x)$ 确实是在无限多步破产的概率。

现在讨论由这些公式得到的一系列推论。

如果设 $A = 0, 0 \leqslant x \leqslant B$，则函数 $\beta(x)$ 按其本来含义，是自状态 x 出发的质点在到达 0 点之前到达点 B 的概率。由 (10) 和 (14) 式，可见（图 16）：

$$\beta(x) = \begin{cases} x/B, & 若\ p = q = 1/2, \\ \dfrac{(q/p)^x - 1}{(q/p)^B - 1}, & 若\ p \neq q. \end{cases} \tag{17}$$

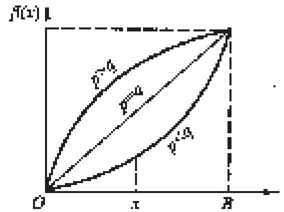


图 16 $\beta(x)$ 的图形

[$\beta(x)$ 是当质点 B 自点 x 出发早于点 0 到达点 B 的概率]

其次，假设不等式 $q > p$ 成立，即对于甲是不利博弈，其破产的极限概率 $\alpha(0)$ 为：

$$\alpha = \frac{(q/p)^B - 1}{(q/p)^B - (q/p)^A}.$$

现在假设对弈条件有如下变化：甲和乙的赌资仍然是 $(-A)$ 和 B，但是现在每人支付的金额为 1/2 单位，而不是以前的 1 单位。换句话说，现在假设 $\mathbf{P}\{\xi_i = 1/2\} = p, \mathbf{P}\{\xi_i = -1/2\} = q$。这时，若以 $\alpha_{1/2}$ 表示甲破产的概率，则

$$\alpha_{1/2} = \frac{(q/p)^{2B} - 1}{(q/p)^{2B} - (q/p)^{2A}}.$$

因此，如果 $q > p$，则

$$\alpha_{1/2} = \alpha \times \frac{(q/p)^B + 1}{(q/p)^B - (q/p)^A} > \alpha.$$

由此得出结论：*如果对弈对于甲不利（即 $q > p$），则增加一倍赌金可以减小他破产的概率。*

3. $\alpha_n(x)$ 和 $\beta_n(x)$ 收敛于 $\alpha(x)$ 和 $\beta(x)$ 的速度 现在讨论 $\alpha_n(x)$ 和 $\beta_n(x)$ 收敛于极限值 $\alpha(x)$ 和 $\beta(x)$ 的速度问题。为简便计，假设 $x = 0$，记

$$\alpha_n = \alpha_n(0), \quad \beta_n = \beta_n(0), \quad \gamma_n = 1 - (\alpha_n + \beta_n).$$

显然，

$$\gamma_n = \mathbf{P}\{A < S_k < B, 0 \leqslant k \leqslant n\},$$

其中 $\{A < S_k < B, 0 \leqslant k \leqslant n\}$ 就是事件

$$\bigcap_{0 \leqslant k \leqslant n} \{A < S_k < B\}.$$

设 $n = rm$，其中 r 和 m 是整数；设

$$\zeta_1 = \xi_1 + \cdots + \xi_m,$$
$$\zeta_2 = \xi_{m+1} + \cdots + \xi_{2m},$$
$$\cdots\cdots\cdots\cdots$$
$$\zeta_r = \xi_{m(r-1)+1} + \cdots + \xi_{rm}.$$

那么，设 $C = |A| + B$，易见

$$\{A < S_k < B, 1 \leqslant k \leqslant rm\} \subseteq \{|\zeta_1| < C, \cdots, |\zeta_r| < C\},$$

因此由随机变量 $\zeta_1, \cdots, \zeta_r$ 独立同分布，可见

$$\gamma_n \leqslant \mathbf{P}\{|\zeta_1| < C, \cdots, |\zeta_r| < C\} = \prod_{i=1}^{r} \mathbf{P}\{|\zeta_i| < C\} = (\mathbf{P}\{|\zeta_1| < C\})^r. \tag{18}$$

注意到 $\mathbf{D}\zeta_1 = m[1 - (p - q)^2]$，因此当 $0 < p < 1$ 时，对于充分大的 m，有

$$\mathbf{P}\{|\zeta_1| < C\} \leqslant \varepsilon_1, \tag{19}$$

其中 $\varepsilon_1 < 1$，因为假如 $\mathbf{P}\{|\zeta_1| < C\} = 1$，则 $\mathbf{D}\zeta_1 \leqslant C^2$。

假如 $p = 0$ 或 1，则对于充分大的 m，概率 $\mathbf{P}\{|\zeta_1| < C\} = 0$，从而对于一切 $0 \leqslant p \leqslant 1$，(19) 式成立。

由 (18) 和 (19) 两式，可见对于充分大的 n，

$$\gamma_n \leqslant \varepsilon^n, \tag{20}$$

其中 $\varepsilon = \varepsilon_1^{1/m} < 1$。

由 (16) 式，$\alpha + \beta = 1$。因此

$$(\alpha - \alpha_n) + (\beta - \beta_n) = \gamma_n$$

而因为 $\alpha \geqslant \alpha_n, \beta \geqslant \beta_n$, 所以

$$0 \leqslant \alpha - \alpha_n \leqslant \gamma_n \leqslant \varepsilon^n,$$
$$0 \leqslant \beta - \beta_n \leqslant \gamma_n \leqslant \varepsilon^n,$$

其中 $\varepsilon < 1$. 对于 $\alpha(x) - \alpha_n(x)$ 和 $\beta(x) - \beta_n(x)$ 有类似的估计.

4. 平均游动时间　现在讨论平均随机游动时间.

设 $m_k(x) = \mathbf{E}\tau_k^x$ 是停止时间 $\tau_k^x, (k \leqslant n)$ 的数学期望. 仿照推导 $\beta_n(x)$ 的递推公式的方法, 对于 $x \in (A, B)$, 得

$$\begin{aligned}
m_k(x) &= \mathbf{E}\tau_k^x = \sum_{1\leqslant l\leqslant k} l\mathbf{P}\{\tau_k^x = l\} \\
&= \sum_{1\leqslant l\leqslant k} l[p\mathbf{P}\{\tau_k^x = l|\xi_1 = 1\} + q\mathbf{P}\{\tau_k^x = l|\xi_1 = -1\}] \\
&= \sum_{1\leqslant l\leqslant k} l[p\mathbf{P}\{\tau_{k-1}^{x+1} = l-1\} + q\mathbf{P}\{\tau_{k-1}^{x-1} = l-1\}] \\
&= \sum_{0\leqslant l\leqslant k-1} (l+1)[p\mathbf{P}\{\tau_{k-1}^{x+1} = l\} + q\mathbf{P}\{\tau_{k-1}^{x-1} = l\}] \\
&= pm_{k-1}(x+1) + qm_{k-1}(x-1) \\
&\qquad + \sum_{0\leqslant l\leqslant k-1} [p\mathbf{P}\{\tau_{k-1}^{x+1} = l\} + q\mathbf{P}\{\tau_{k-1}^{x-1} = l\}] \\
&= pm_{k-1}(x+1) + qm_{k-1}(x-1) + 1.
\end{aligned}$$

于是, 对于 $x \in (A, B)$ 和 $0 \leqslant k \leqslant n$, 函数 $m_k(x)$ 满足递推关系式:

$$m_k(x) = 1 + pm_{k-1}(x+1) + qm_{k-1}(x-1), \tag{21}$$

其中 $m_0(x) = 0$. 由这些方程连同边界条件

$$m_k(A) = m_k(B) = 0, \tag{22}$$

可以依次求出 $m_1(x), \cdots, m_n(x)$.

由于 $m_k(x) \leqslant m_{k-1}(x)$, 可见存在极限

$$m(x) = \lim_{n\to\infty} m_n(x),$$

且由 (21) 式知 $m(x)$ 满足方程

$$m(x) = 1 + pm(x+1) + qm(x-1) \tag{23}$$

和边界条件

$$m(A) = m(B) = 0. \tag{24}$$

为求该方程的解, 首先假设

$$m(x) < \infty, x \in (A, B). \tag{25}$$

那么, 如果 $p \neq q$, 则特解的形式为 $x/(q-p)$, 而通解 (见 (9) 式) 可以表示为:

$$m(x) = \frac{x}{q-p} + a + b\left(\frac{q}{p}\right)^x.$$

由此并考虑到边界条件 $m(A) = m(B) = 0$, 即可求出

$$m(x) = \frac{1}{p-q}[B\beta(x) + A\alpha(x) - x], \tag{26}$$

其中 $\beta(x)$ 和 $\alpha(x)$ 由 (10) 和 (13) 两式决定. 如果 $p = q = 1/2$, 则方程 (23) 的通解形如

$$m(x) = a + bx - x^2,$$

因此, 由于边界条件 $m(A) = m(B) = 0$, 可见

$$m(x) = (B - x)(x - A). \tag{27}$$

特别, 由此可见, 如果对弈二人的初始赌金相等: $B = -A$, 则

$$m(0) = B^2.$$

假设 $B = 10$ 且每秒一局, 那么某一对手破产前的 (极限) 平均时间相当长: 等于 100 秒.

公式 (26) 和 (27) 是在假设 $m(x) < \infty, x \in (A, B)$ 的条件下得到的, 现在证明, 实际上 $m(x)$ 对于一切 $x \in (A, B)$ 有限. 我们只限于讨论 $x = 0$ 的情形, 一般情形可以类似地证明.

设 $p = q = 1/2$. 将数列 $S_0, S_1, \cdots, S_n$ 与停止时间 $\tau_n = \tau_n^0$, 由随机变量 $S_{\tau_n} = S_{\tau_n}(\omega)$ 联系在一起, 其中

$$S_{\tau_n}(\omega) = \sum_{k=0}^{n} S_k(\omega) I_{\{\tau_n = k\}}(\omega). \tag{28}$$

随机变量 S_{τ_n} 的直观意义是清楚的: 它是随机游动在停止时刻 τ_n 的值. 这时, 假如 $\tau_n < n$, 则 $S_{\tau_n} = A$ 或 B; 而如果 $\tau_n = n$, 则 $A \leqslant S_{\tau_n} \leqslant B$.

现在证明, 对于 $p = q = 1/2$, 有

$$\mathbf{E}S_{\tau_n} = 0, \tag{29}$$

$$\mathbf{E}S_{\tau_n}^2 = \mathbf{E}\tau_n. \tag{30}$$

为证明 (29) 式, 注意到

$$\begin{aligned}\mathbf{E}S_{\tau_n}^2 &= \sum_{k=0}^{n}\mathbf{E}[S_k I_{\{\tau_n=k\}}(\omega)]\\ &= \sum_{k=0}^{n}\mathbf{E}[S_n I_{\{\tau_n=k\}}(\omega)] + \sum_{k=0}^{n}\mathbf{E}[(S_k-S_n)I_{\{\tau_n=k\}}(\omega)]\\ &= \mathbf{E}S_n + \sum_{k=0}^{n}\mathbf{E}[(S_k-S_n)I_{\{\tau_n=k\}}(\omega)],\end{aligned} \tag{31}$$

其中显然 $\mathbf{E}S_n=0$. 现在证明

$$\sum_{k=0}^{n}\mathbf{E}[(S_k-S_n)I_{\{\tau_n=k\}}(\omega)]=0.$$

对于 $0\leqslant k<n$, 有 $\{\tau_n>k\}=\{A<S_1<B,\cdots,A<S_k<B\}$. 事件

$$\{\omega: A<S_1<B,\cdots,A<S_k<B\}$$

显然可以表示为

$$\{\omega:(\xi_1,\cdots,\xi_k)\in A_k\}, \tag{32}$$

其中 A_k 是集合 $\{-1,+1\}^k$ 的某一子集. 换句话说, 这是一个只决定于随机变量 $\xi_1,\cdots,\xi_k$ 的值, 而不依赖于随机变量 $\xi_{k+1},\cdots,\xi_n$ 值的集合. 由于集合

$$\{\tau_n=k\}=\{\tau_n>k-1\}\setminus\{\tau_n>k\},$$

可见它也是形如 (32) 的集合. 由于随机变量 $\xi_1,\cdots,\xi_k$ 的独立, 以及由 §4 的练习题 10 可见, *对于任意* $0\leqslant k<n$, 随机变量 S_n-S_k 和 $I_{\{\tau_n=k\}}$ 独立, 从而

$$\mathbf{E}[(S_n-S_k)I_{\{\tau_n=k\}}]=\mathbf{E}[S_n-S_k]\times\mathbf{E}I_{\{\tau_n=k\}}=0.$$

于是 (29) 式得证.

同样可以证明 (30) 式:

$$\begin{aligned}\mathbf{E}S_{\tau_n}^2 &= \sum_{k=0}^{n}\mathbf{E}S_k^2 I_{\{\tau_n=k\}} = \sum_{k=0}^{n}\mathbf{E}\{[S_n+(S_k-S_n)]^2 I_{\{\tau_n=k\}}\}\\ &= \sum_{k=0}^{n}[\mathbf{E}S_n^2 I_{\{\tau_n=k\}} + 2\mathbf{E}S_n(S_k-S_n)I_{\{\tau_n=k\}} + \mathbf{E}(S_n-S_k)^2 I_{\{\tau_n=k\}}]\\ &= \mathbf{E}S_n^2 - \sum_{k=0}^{n}\mathbf{E}(S_n-S_k)^2 I_{\{\tau_n=k\}} = n - \sum_{k=0}^{n}(n-k)\mathbf{P}\{\tau_n=k\}\\ &= \sum_{k=0}^{n}k\mathbf{P}\{\tau_n=k\} = \mathbf{E}\tau_n.\end{aligned}$$

于是, 对于 $p=q=1/2$, 公式 (29) 和 (30) 得证. 对于任意 $p,q\,(p+q=1)$ 的情形, 类似地可以证明:

$$\mathbf{E}S_{\tau_n}=(p-q)\mathbf{E}\tau_n, \tag{33}$$

$$\mathbf{E}[S_{\tau_n}-\tau_n\mathbf{E}\xi_1]^2=\mathbf{D}\xi_1\times\mathbf{E}\tau_n, \tag{34}$$

其中 $\mathbf{E}\xi_1=p-q$, $\mathbf{D}\xi_1=1-(p-q)^2$.

现在利用得到的关系式, 证明

$$\lim_{n\to\infty}m_n(0)=m(0)<\infty.$$

对于 $p=q=1/2$, 由 (30) 式, 有

$$\mathbf{E}\tau_n\leqslant\max\{A^2,B^2\}, \tag{35}$$

如果 $p\neq q$, 则由 (33) 式, 有

$$\mathbf{E}\tau_n\leqslant\frac{\max\{|A|,B\}}{|p-q|}, \tag{36}$$

由此可见 $m(0)<\infty$.

还要指出, 对于 $p=q=1/2$, 有

$$\mathbf{E}\tau_n=\mathbf{E}S_{\tau_n}^2=A^2\alpha_n+B^2\beta_n+\mathbf{E}[S_n^2 I_{\{A<S_n<B\}}I_{\{\tau_n=n\}}];$$

因此

$$A^2\alpha_n+B^2\beta_n\leqslant\mathbf{E}\tau_n\leqslant A^2\alpha_n+B^2\beta_n+\max\{A^2,B^2\}\times\gamma_n.$$

由此及不等式 (20) 可见, 当 $n\to\infty$ 时数学期望 $\mathbf{E}\tau_n$ 以指数的速度收敛于极限值:

$$m(0)=A^2\alpha+B^2\beta=A^2\times\frac{B}{B-A}-B^2\times\frac{A}{B-A}=|AB|.$$

对于 $p\neq q$, 有类似的结果 —— $\mathbf{E}\tau_n$ 以指数的速度收敛于极限值:

$$\mathbf{E}\tau_n\to m(0)=\frac{\alpha A+\beta B}{p-q}.$$

5. 练习题

1. 证明 (33) 和 (34) 式的推广: 下列公式成立:

$$\mathbf{E}S_{\tau_n^x}^x=x+(p-q)\mathbf{E}\tau_n^x,$$

$$\mathbf{E}[S_{\tau_n^x}^x-\tau_n^x\mathbf{E}\xi_1]^2=\mathbf{D}\xi_1\times\mathbf{E}\tau_n^x+x^2.$$

2. 讨论当水平 $A\to-\infty$ 时, $\alpha(x),\beta(x)$ 和 $m(x)$ 趋向何量的问题.

3. 假设伯努利概型中 $p=q=1/2$. 问当 n 充分大时 $\mathrm{E}|S_n|$ 是何数量级?

4. 甲乙二人各自独立地掷一枚对称的硬币. 证明, 他们各自掷 n 次, 二人掷出正面的次数相同的概率为

$$\frac{1}{2^{2n}}\sum_{k=0}^{n}(C_n^k)^2.$$

并由此导出恒等式 (亦见 §2 练习题 4)

$$\sum_{k=0}^{n}(C_n^k)^2=C_{2n}^n.$$

设 σ_n 表示时刻: 对手甲和乙首次各自掷出正面的次数相等 (若这样的时刻不存在, 则令 $\sigma_n=n+1$). 求数学期望 $\mathrm{E}\min\{\sigma_n,n\}$.

§10. 随机游动 II. 反射原理. 反正弦定律

1. 质点首次返回 0 的时间 像上一节一样, 假设 $\xi_1,\cdots,\xi_{2n}$ 是独立同分布伯努利随机变量序列, 且

$$\mathrm{P}\{\xi_i=1\}=p,\quad \mathrm{P}\{\xi_i=-1\}=q,$$
$$S_k=\xi_1+\cdots+\xi_k,\quad 1\leqslant k\leqslant 2n,\quad S_0=0.$$

记

$$\sigma_{2n}=\min\{1\leqslant k\leqslant 2n: S_k=0\};$$

对于一切 $1\leqslant k\leqslant 2n$, 若 $S_k\neq 0$, 则设 $\sigma_{2n}=\infty$.

量 σ_{2n} 的直观含义十分清楚: 它是首次返回 0 的时刻. 这一节将要研究首次返回 0 的时刻 σ_{2n} 的性质, 这时假设所考虑的随机游动对称, 即 $p=q=1/2$.

对于 $0\leqslant k\leqslant n$, 记

$$u_{2k}=\mathrm{P}\{S_{2k}=0\},\quad f_{2k}=\mathrm{P}\{\sigma_{2n}=2k\}. \tag{1}$$

显然, $u_0=1$ 而

$$u_{2k}=C_{2k}^k\cdot\frac{1}{2^{2k}}.$$

我们最近的目标是证明: 对于 $1\leqslant k\leqslant n$, 概率决定于公式:

$$f_{2k}=\frac{1}{2k}u_{2(k-1)}. \tag{2}$$

易见, 对于 $1\leqslant k\leqslant n$, 有

$$\{\sigma_{2n}=2k\}=\{S_1\neq 0,\cdots,S_{2k-1}\neq 0,S_{2k}=0\},$$

由对称性, 可见

$$\begin{aligned}f_{2k}&=\mathrm{P}\{S_1\neq 0,\cdots,S_{2k-1}\neq 0,S_{2k}=0\}\\&=2\mathrm{P}\{S_1>0,\cdots,S_{2k-1}>0,S_{2k}=0\}.\end{aligned} \tag{3}$$

我们把序列 $(S_0,\cdots,S_k)$ 称做长为 k 的路径, 以 $L_k(A)$ 表示具有性质 A 的长为 k 的路径条数. 那么,

$$\begin{aligned}f_{2k}&=\frac{1}{2^{2n-1}}\sum_{(a_{2k+1},\cdots,a_{2n})}[L_{2n}(S_1>0,\cdots,S_{2k-1}>0,S_{2k}=0,\\&\qquad S_{2k+1}=a_{2k+1},\cdots,S_{2n}=a_{2k+1}+\cdots+a_{2n})]\\&=\frac{1}{2^{2k-1}}L_{2k}(S_1>0,\cdots,S_{2k-1}>0,S_{2k}=0),\end{aligned} \tag{4}$$

其中对一切 $\{a_{2k+1},\cdots,a_{2n}\}$, $a_i=\pm 1$ 的组合求和.

于是, 求概率 f_{2k} 归结为求路径的条数 $L_{2k}\{S_1>0,\cdots,S_{2k-1}>0,S_{2k}=0\}$.

引理 1 设 a,b 为非负整数, $a-b>0$, $k=a+b$, 则

$$L_k(S_1>0,\cdots,S_{k-1}>0,S_k=a-b)=\frac{a-b}{k}C_k^a. \tag{5}$$

证明 事实上

$$\begin{aligned}&L_k(S_1>0,\cdots,S_{k-1}>0,S_k=a-b)\\&=L_k(S_1=1,S_2>0,\cdots,S_{k-1}>0,S_k=a-b)\\&=L_k(S_1=1,S_k=a-b)\\&\quad-L_k(S_1=1,S_k=a-b;\exists i,2\leqslant i\leqslant k-1,\text{ 使 }S_i\leqslant 0).\end{aligned} \tag{6}$$

换句话说, 始自点 $(1,1)$ 止于点 $(k,a-b)$ 的正路径 $(S_1,S_2,\cdots,S_k)$ 的条数, 等于由点 $(1,1)$ 到点 $(k,a-b)$ 的全部路径条数, 减去与时轴相切或相交的路径条数*).

注意到,

$$L_k(S_1=1,S_k=a-b;\exists i,2\leqslant i\leqslant k-1,\text{ 使 }S_i\leqslant 0)=L_k(S_1=-1,S_k=a-b), \tag{7}$$

即自点 $\alpha=(1,1)$ 到点 $\beta=(k,a-b)$ 且与时轴相切或相交的所有路径条数, 等于由点 $\alpha^*=(1,-1)$ 到点 $\beta=(k,a-b)$ 的所有路径的条数. 这一事实称做反射原理. 容易证明, 在连接 α 和 β 两点的路径 $A=(S_1,\cdots,S_a,S_{a+1},\cdots,S_k)$, 与在连接 α^* 和 β 两点的路径 $B=(-S_1,\cdots,-S_a,S_{a+1},\cdots,S_k)$, 之间存在一一对应关系 (图 17). 其中 a 是路径 A 和路径 B 为 0 的第一个点; 其证明可以由上述两条路径的一一对应关系导出.

*) 路径 $(S_1,S_2,\cdots,S_k)$ 称做正的 (非负的), 如果全部 $S_i>0$ ($S_i\geqslant 0$); 路径称做与时轴相切的, 如果对于所有 $1\leqslant j\leqslant k$, $S_j\geqslant 0$ ($S_j\leqslant 0$), 且存在 $1\leqslant i\leqslant k$, $S_i=0$; 路径称做与时轴相交的, 如果存在两个时点 i 和 j, 使 $S_i>0$ ($S_j<0$).

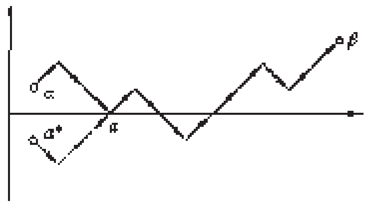


图 17 反射原理

由 (6) 式和 (7) 式可见

$$\begin{aligned}&L_k(S_1>0,\cdots,S_{k-1}>0,S_k=a-b)\\&=L_k(S_1=1,S_k=a-b)-L_k(S_1=-1,S_k=a-b)\\&=C_{k-1}^{a-1}-C_{k-1}^{a}=\frac{a-b}{k}C_k^a,\end{aligned}$$

于是, 命题 (6) 得证.

现在, 回到概率 f_{2k} 的计算. 由 (4) 式和 (5) 式 ($a=k, b=k-1$), 可见

$$\begin{aligned}f_{2k}&=2^{-2k+1}L_{2k}(S_1>0,\cdots,S_{2k-1}>0,S_{2k}=0)\\&=2^{-2k+1}L_{2k-1}(S_1>0,\cdots,S_{2k-1}=1)\\&=2^{-2k+1}\times\frac{1}{2k-1}\times C_{2k-1}^{k}=\frac{1}{2k}u_{2(k-1)}.\end{aligned}$$

于是, 公式 (2) 得证.

我们再给出公式 (2) 的一个证明. 证明基于如下事实:

$$\frac{1}{2k}u_{2(k-1)}=u_{2(k-1)}-u_{2k}. \tag{8}$$

显然,

$$\{\sigma_{2n}=2k\}=\{\sigma_{2n}>2(k-1)\}\setminus\{\sigma_{2n}>2k\},$$

$$\{\sigma_{2n}>2l\}=\{S_1\neq 0,\cdots,S_{2l}\neq 0\},$$

即

$$\{\sigma_{2n}=2k\}=\{S_1\neq 0,\cdots,S_{2(k-1)}\neq 0\}\setminus\{S_1\neq 0,\cdots,S_{2k}\neq 0\}.$$

因此

$$f_{2k}=\mathbf{P}\{S_1\neq 0,\cdots,S_{2(k-1)}\neq 0\}-\mathbf{P}\{S_1\neq 0,\cdots,S_{2k}\neq 0\},$$

从而, 由于 (8) 式, 为证明等式

$$f_{2k}=\frac{1}{2k}u_{2(k-1)},$$

只需证明

$$L_{2k}(S_1\neq 0,\cdots,S_{2k}\neq 0)=L_{2k}(S_{2k}=0). \tag{9}$$

为此, 注意到, 显然

$$L_{2k}(S_1\neq 0,\cdots,S_{2k}\neq 0)=2L_{2k}(S_1>0,\cdots,S_{2k}>0).$$

因此为验证 (9) 式, 只需证明

$$2L_{2k}(S_1>0,\cdots,S_{2k}>0)=L_{2k}(S_1\geqslant 0,\cdots,S_{2k}\geqslant 0), \tag{10}$$

且

$$L_{2k}(S_1\geqslant 0,\cdots,S_{2k}\geqslant 0)=L_{2k}(S_{2k}=0). \tag{11}$$

如果证明在至少一个 $S_i=0$ 的路径 $A=(S_1,\cdots,S_{2k})$, 与正路径 $B=(S_1,\cdots,S_{2k})$ 之间存在一一对应关系, 则就可以证明等式 (10).

设 $A=(S_1,\cdots,S_{2k})$ 是在点为首次 0 (即 $S_a=0$) 的非负路径. 自点 $(a,2)$ 引一轨道 $(S_a+2,S_{a+1}+2,\cdots,S_{2k}+2)$ (图 18 中虚线), 那么路径 $B=(S_1,\cdots,S_{a-1},S_a+2,\cdots,S_{2k}+2)$ 是正的. 相反, 设 $B=(S_1,\cdots,S_{2k})$ 是某一正路径, 而 b 是 $S_b=1$ 的最后时刻 (图 19), 那么路径 $A=(S_1,\cdots,S_b,S_{b+1}-2,\cdots,S_k-2)$ 是非负的. 由上面的构造可见, 在正路径和 (至少有一个 $S_i=0$) 的非负路径之间, 存在一一对应关系. 于是等式 (10) 得证.

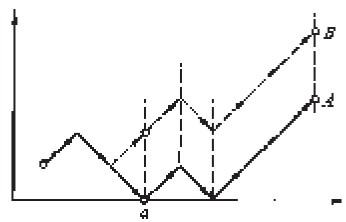


图 18

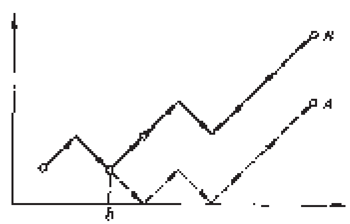


图 19

现在证明等式 (11). 由于对称性和 (10) 式, 只需要证明

$$\begin{aligned}&L_{2k}(S_1>0,\cdots,S_{2k}>0)+L_{2k}(S_1\geqslant 0,\cdots,S_{2k}\geqslant 0,\\&\qquad\text{且 }\exists i,0\leqslant i\leqslant 2k,S_i=0)=L_{2k}(S_{2k}=0).\end{aligned}$$

路径的集合 $(S_{2k}=0)$ 可以表示为两个集合 $\mathscr{C}_1$ 与 $\mathscr{C}_2$ 之和, 其中 $\mathscr{C}_1$ 是只有一个极小值的路径 $(S_0,\cdots,S_{2k})$, 而 $\mathscr{C}_2$ 是至少在两个点上达到极小值的路径.

设 $C_1\in\mathscr{C}_1$ (图 20), 而 γ 是极小值点. 将路径 $C_1=(S_0,S_1,\cdots,S_{2k})$ 对应于一正路径 C_1^* (图 21). 路径 C_1^* 是由如下方式得到的: 将轨道 $(S_0,S_1,\cdots,S_l)$ 关于过点

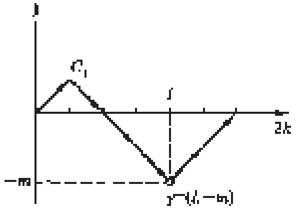


图 20

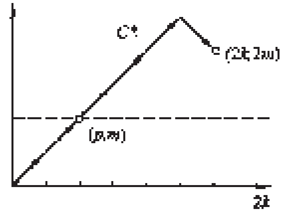


图 21

l 铅直线进行反射, 而把所得轨道自点 $(2k,0)$ 延伸后, 将其置于右、上方. 然后将坐标原点置于点 $(l,-m)$. 所得到的轨道是正路径.

同样, 如果路径 $C_2\in\mathscr{C}_2$, 则用同样的方式可以将其对应于某一非负路径 C_2^*.

相反, 设 $C_1^*=(S_1>0,\cdots,S_{2k}>0)$ 是某一正路径, 其中 $S_{2k}=2m$ (图 21). 将路径 C_1^* 对应于一正路径 C_1 (图 20). 路径 C_1 是由如下方式得到的: 设 p 是使 $S_p=m$ 的最后一点; 将轨道 $(S_p,\cdots,S_{2m})$ 关于铅直线 $x=p$ 反射, 将其置于右、上方, 使其第一个端点与点 $(0,0)$ 重合. 然后将坐标原点置于所得轨道的左端点 (这恰好是图 20 上轨道). 所得路径有唯一极小值, 且 $S_{2k}=0$. 对路径

$$(S_1\geqslant 0,\cdots,S_{2k}\geqslant 0,\ 且\ \exists i, 0\leqslant i\leqslant 2k, S_i=0)$$

施用类似的构造可以导致至少有两个极小值且 $S_{2k}=0$ 的路径. 从而建立了一一对应关系, 于是证明了 (11) 式.

这样, 证明了等式 (9), 以及下面的公式:

$$f_{2k}=u_{2(k-1)}-u_{2k}=\frac{1}{2k}u_{2(k-1)}.$$

由斯特林 (J. Stirling) 公式, 有

$$u_{2k}=C_{2k}^k\times 2^{-2k}\sim\frac{1}{\sqrt{\pi k}},\quad k\to\infty.$$

因此

$$f_{2k}\sim\frac{1}{2\sqrt{\pi}k^{3/2}},\quad k\to\infty.$$

由此可见, 首返 0 的时间的数学期望

$$\mathbf{E}\min(\sigma_{2n},2n)=\sum_{k=1}^{n}2k\mathbf{P}\{\sigma_{2n}=2k\}+2nu_{2n}=\sum_{k=1}^{n}u_{2(k-1)}=2nu_{2n}$$

是相当大的.

此外

$$\sum_{k=1}^{\infty}u_{2(k-1)}=\infty,$$

从而, 对于步数无限的情形, 游动趋于 0 的平均时间的极限等于 ∞.

这一事实可以解释所研究的、对称随机游动许多意想不到的性质. 例如, 两个实力相当的对手 (即 $p=q=1/2$) 对弈时, 自然地期望经 $2n$ 局对弈, 平局的平均次数 (即使 $S_i=0$ 的时刻 i 的次数) 与 $2n$ 成正比. 然而, 平均平局次数却是数量级为 $\sqrt{2n}$ 的量 (参见第一章 §9, (17) 式). 特别, 由此可以得出与期望相反的、"典型的"的实现: 游动 $(S_0,S_1,\cdots,S_n)$ 事实上并不具有正弦的特点 (质点一半时间在正侧, 另一半时间在负侧), 而具有长而缓慢的波动的特点. 所谓反正弦定律说明了命题的这种提法. 我们现在就开始介绍反正弦定律.

2. 反正弦定律 记 $P_{2k,2n}$ 为质点在时段上 $[0,2n]$ 的正方度过 $2k$ 个单位时间的概率*).

引理 2 设 $u_0=1$ 和 $0\leqslant k\leqslant n$, 则

$$P_{2k,2n}=u_{2k}u_{2n-2k}. \tag{12}$$

证明 上面曾证明 $f_{2k}=u_{2(k-1)}-u_{2k}$. 现在证明,

$$u_{2k}=\sum_{r=1}^{k}f_{2r}u_{2(k-r)}. \tag{13}$$

由于 $\{S_{2k}=0\}\subseteq\{\sigma_{2n}\leqslant 2k\}$, 可见

$$\{S_{2k}=0\}=\{S_{2k}=0\}\cap\{\sigma_{2n}\leqslant 2k\}=\sum_{1\leqslant l\leqslant k}\{S_{2k}=0\}\cap\{\sigma_{2n}=2l\}.$$

从而

$$u_{2k}=\mathbf{P}\{S_{2k}=0\}=\sum_{1\leqslant l\leqslant k}\mathbf{P}\{S_{2k}=0,\sigma_{2n}=2l\}$$
$$=\sum_{1\leqslant l\leqslant k}\mathbf{P}\{S_{2k}=0|\sigma_{2n}=2l\}\mathbf{P}\{\sigma_{2n}=2l\}.$$

而由

$$\mathbf{P}\{S_{2k}=0|\sigma_{2n}=2l\}=\mathbf{P}\{S_{2k}=0|S_1\neq 0,\cdots,S_{2l-1}\neq 0,S_{2l}=0\}$$
$$=\mathbf{P}\{S_{2l}+(\xi_{2l+1}+\cdots+\xi_{2k})=0|S_1\neq 0,\cdots,S_{2l-1}\neq 0,S_{2l}=0\}$$
$$=\mathbf{P}\{S_{2l}+(\xi_{2l+1}+\cdots+\xi_{2k})=0|S_{2l}=0\}$$
$$=\mathbf{P}\{\xi_{2l+1}+\cdots+\xi_{2k}=0\}=\mathbf{P}\{S_{2(k-l)}=0\},$$

可见

$$u_{2k}=\sum_{1\leqslant l\leqslant k}\mathbf{P}\{S_{2(k-l)}=0\}\mathbf{P}\{\sigma_{2n}=2l\},$$

*) 称质点在时间段 $[m-1,m]$ 位于正方, 如果 S_{m-1} 或 S_m 至少有一个值是正的.

于是,(13) 式得证.

现在证明 (12) 式. 当 $k=0$ 和 $k=n$ 时 (12) 式显然成立. 现在设 $1 \leqslant k \leqslant n-1$. 假如质点在正方度过 $2k$ 个时刻, 则它一定通过 0. 设 $2r$ 是首次返回 0 的时刻. 有两种可能的情形: 当 $S_i \geqslant 0$ 时 $i \leqslant 2r$, 当 $S_i \leqslant 0$ 时 $i \leqslant 2r$.

易见, 对于第一种情形, 路径条数等于

$$\left(\frac{1}{2} \times 2^{2r} f_{2r}\right) 2^{2(n-r)} P_{2(k-r),2(n-r)} = \frac{1}{2} \times 2^{2n} f_{2r} P_{2(k-r),2(n-r)}.$$

对于第二种情形, 相应的路径条数等于

$$\frac{1}{2} \times 2^{2n} f_{2r} P_{2k,2(n-r)}.$$

从而, 对于 $1 \leqslant k \leqslant n-1$, 有

$$P_{2k,2n} = \frac{1}{2} \sum_{r=1}^{k} f_{2r} P_{2(k-r),2(n-r)} + \frac{1}{2} \sum_{r=1}^{k} f_{2r} P_{2k,2(n-r)}. \tag{14}$$

假设对于 $m=1,\cdots,n-1$, 公式 $P_{2k,2m}=u_{2m}u_{2m-2k}$ 成立. 那么, 由 (13) 和 (14) 两式, 可见

$$\begin{aligned} P_{2k,2n} &= \frac{1}{2} u_{2n-2k} \sum_{r=1}^{k} f_{2r} u_{2k-2r} + \frac{1}{2} u_{2k} \sum_{r=1}^{k} f_{2r} u_{2n-2r-2k} \\ &= \frac{1}{2} u_{2n-2k} u_{2k} + \frac{1}{2} u_{2k} u_{2n-2k} = u_{2k} u_{2n-2k}. \end{aligned} \qquad \square$$

现在, 设 $\gamma(2n)$ 是质点在区间 $[0,2n]$ 内位于正方的时间单位数, 则对于 $x<1$, 有

$$\mathbf{P}\left\{\frac{1}{2} < \frac{\gamma(2n)}{2n} \leqslant x\right\} = \sum_{\{k:\frac{1}{2}<\frac{2k}{2n}\leqslant x\}} P_{2k,2n}.$$

由于当 $k \sim \infty$ 时, 有

$$u_{2k} \sim \frac{1}{\sqrt{\pi k}},$$

则当 $k \to \infty, n-k \to \infty$ 时

$$P_{2k,2n} = u_{2k} u_{2(n-k)} \sim \frac{1}{\pi\sqrt{k(n-k)}}.$$

因此

$$\sum_{\{k:\frac{1}{2}<\frac{2k}{2n}\leqslant x\}} P_{2k,2n} - \sum_{\{k:\frac{1}{2}<\frac{2k}{2n}\leqslant x\}} \frac{1}{\pi n}\left[\frac{k}{n}\left(1-\frac{k}{n}\right)\right]^{-1/2} \to 0, n \to \infty,$$

故

$$\sum_{\{k:\frac{1}{2}<\frac{2k}{2n}\leqslant x\}} P_{2k,2n} - \frac{1}{\pi}\int_{1/2}^{x} \frac{dt}{\sqrt{t(1-t)}} \to 0, n \to \infty.$$

而由对称性, 可见

$$\sum_{\{k:\frac{k}{n}\leqslant\frac{1}{2}\}} P_{2k,2n} \to \frac{1}{2}, n \to \infty$$

且

$$\frac{1}{\pi}\int_{1/2}^{x} \frac{dt}{\sqrt{t(1-t)}} = \frac{2}{\pi}\arcsin\sqrt{x} - \frac{1}{2}.$$

于是, 证明了下面的定理.

定理 (反正弦定律) *质点位于正方的时间所占的份额不大于 x 的概率, 趋于 $2\pi^{-1}\arcsin\sqrt{x}$:*

$$\sum_{\{k:\frac{k}{n}\leqslant x\}} P_{2k,2n} \to \frac{2}{\pi}\arcsin\sqrt{x}. \tag{15}$$

注意, 积分中

$$\int_0^x \frac{dt}{\sqrt{t(1-t)}}$$

的被积函数 $\rho(t)=[t(1-t)]^{-1/2}$ 是在点 $t=0$ 和 $t=1$ 趋向无穷的一条 U 形曲线.

由此可见, 对于充分大的 n, 有

$$\mathbf{P}\left\{0 < \frac{\gamma(2n)}{2n} \leqslant \Delta\right\} > \mathbf{P}\left\{\frac{1}{2} < \frac{\gamma(2n)}{2n} \leqslant \frac{1}{2}+\Delta\right\},$$

即形象地说, 质点位于正方接近 0 (或 1), 比自然期望的值 1/2, 更有可能.

利用反正弦函数表, 以及在 (15) 式中的收敛速度实际上是很快的情况, 可以求得下面的概率:

$$\begin{aligned} \mathbf{P}\left\{\frac{\gamma(2n)}{2n} \leqslant 0.024\right\} &\approx 0.1, \\ \mathbf{P}\left\{\frac{\gamma(2n)}{2n} \leqslant 0.1\right\} &\approx 0.2, \\ \mathbf{P}\left\{\frac{\gamma(2n)}{2n} \leqslant 0.2\right\} &\approx 0.3, \\ \mathbf{P}\left\{\frac{\gamma(2n)}{2n} \leqslant 0.65\right\} &\approx 0.6. \end{aligned}$$

这样, 如果考虑时段 $[0,1000]$, 则大致在十分之一的情形下, 质点总共有 24 个时间单位在正方度过, 但多数时间 (976 时间单位) 在负方.

3. 练习题

1. 当 $n\to\infty$ 时, $\mathbf{E}\min(\sigma_{2n}, 2n)\to\infty$ 的速度如何?

2. 设 $\tau_n=\min\{1\leqslant k\leqslant n: S_k=1\}$, 如果对于一切 $1\leqslant k\leqslant n, S_k<1$, 设 $\tau_n=\infty$. 问对于对称 $(p=q=1/2)$ 和非对称 $(p\neq q)$ 游动, 当 $n\to\infty$ 时, $\mathbf{E}\min(\tau_n, n)$ 的极限如何?

3. 基于 §10 的思路和方法, 对于伯努利对称 $(p=q=1/2)$ 随机游动 $\{S_k, k\leqslant n\}$, 其中 $S_0=0, S_k=\xi_1+\cdots+\xi_k$, 证明满足下列关系式 ($N$ 是正整数):

$$\mathbf{P}\{\max_{1\leqslant k\leqslant n} S_k\geqslant N, S_n<N\}=\mathbf{P}\{S_n>N\};$$
$$\mathbf{P}\{\max_{1\leqslant k\leqslant n} S_k\geqslant N\}=2\mathbf{P}\{S_n\geqslant N\}-\mathbf{P}\{S_n=N\};$$
$$\mathbf{P}\{\max_{1\leqslant k\leqslant n} S_k=N\}=\mathbf{P}\{S_n=N\}+\mathbf{P}\{S_n=N+1\}.$$

§11. 鞅. 鞅对随机游动的某些应用

1. 引言 前面研究的伯努利变量 $\xi_1,\cdots,\xi_n$ 形成独立随机变量序列. 在这一节和下一节将引进形成鞅和马尔可夫 A. A. Марков 链的两类非独立随机变量.

鞅论将在第七章中详细介绍. 而现在仅给出定义, 证明关于对于停止时间保持鞅性的一个定理, 给出鞅用于证明所谓关于"表决"的定理. 同样地, 该定理将用于 §10 中 (5) 式命题的另一个证明, 该命题曾经利用反射定理证明.

2. 鞅的定义和例 设 $(\Omega,\mathscr{A},\mathbf{P})$ 是有限概率空间, $\mathscr{D}_1\preccurlyeq\mathscr{D}_2\preccurlyeq\cdots\preccurlyeq\mathscr{D}_n$ 是某一分割序列.

定义 1 随机变量序列 $\xi=(\xi_k)_{1\leqslant k\leqslant n}$ (关于分割 $\mathscr{D}_1\preccurlyeq\mathscr{D}_2\preccurlyeq\cdots\preccurlyeq\mathscr{D}_n$) 称做鞅, 如果:

1) ξ_k 是 $\mathscr{D}_k-$ 可测的,

2) $\mathbf{E}(\xi_{k+1}|\mathscr{D}_k)=\xi_k, 1\leqslant k\leqslant n-1$.

为强调随机变量 $\xi=(\xi_1,\cdots,\xi_n)$ 的哪个分割系是鞅, 我们还将使用记号

$$\xi=(\xi_k,\mathscr{D}_k)_{1\leqslant k\leqslant n}, \tag{1}$$

不过为简便计常省略下标 $1\leqslant k\leqslant n$.

如果分割 $\mathscr{D}_k$ 是由随机变量 $\xi_1,\cdots,\xi_k$ 诱导的, 即

$$\mathscr{D}_k=\mathscr{D}_{\xi_1,\cdots,\xi_k},$$

我们往往不提 $\xi=(\xi_k,\mathscr{D}_k)$ 是鞅, 干脆称序列本身 $\xi=(\xi_k)$ 是鞅.

下面举几个鞅的例子.

例 1 设 $\eta_1,\cdots,\eta_n$ 是独立伯努利随机变量:

$$\mathbf{P}\{\eta_k=1\}=\mathbf{P}\{\eta_k=-1\}=\frac{1}{2},$$
$$S_k=\eta_1+\cdots+\eta_k, \mathscr{D}_k=\mathscr{D}_{\eta_1,\cdots,\eta_k}.$$

注意, 分割 $\mathscr{D}_k$ 的构造十分简单:

$$\mathscr{D}_1=\{D^+,D^-\},$$

其中 $D^+=\{\omega:\eta_1=+1\}, D^-=\{\omega:\eta_1=-1\}$;

$$\mathscr{D}_2=\{D^{++},D^{+-},D^{-+},D^{--}\},$$

其中 $D^{-+}=\{\omega:\eta_1=-1,\eta_2=+1\},\cdots,D^{--}=\{\omega:\eta_1=-1,\eta_2=-1\}$; 依此类推.

易见, $\mathscr{D}_{\eta_1,\cdots,\eta_k}=\mathscr{D}_{S_1,\cdots,S_k}$.

现在证明, 序列 $(S_k,\mathscr{D}_k)_{1\leqslant k\leqslant n}$ 构成鞅.

随机变量 S_k 是 $\mathscr{D}_k-$ 可测的, 而由 §8 的 (12) 式, (18) 式和 (24) 式, 可见

$$\mathbf{E}(S_{k+1}|\mathscr{D}_k)=\mathbf{E}(S_k+\eta_{k+1}|\mathscr{D}_k)=\mathbf{E}(S_k|\mathscr{D}_k)+\mathbf{E}(\eta_{k+1}|\mathscr{D}_k)=S_k+\mathbf{E}\eta_{k+1}=S_k,$$

而这正是鞅的性质所要求的.

假如设 $S_0=0$, 取平凡分割: $\mathscr{D}_0=\{\Omega\}$, 则序列 $(S_k,\mathscr{D}_k)_{0\leqslant k\leqslant n}$ 仍然是鞅.

例 2 设 $\eta_1,\cdots,\eta_n$ 是独立伯努利随机变量. $\mathbf{P}\{\eta_k=1\}=p, \mathbf{P}\{\eta_k=-1\}=q$. 如果 $p\neq q$, 则每一个满足

$$\xi_k=\left(\frac{q}{p}\right)^{S_k}, \xi_k=S_k-k(p-q), \quad 其中\ S_k=\eta_1+\cdots+\eta_k$$

的序列 $\xi=(\xi_k)$ 都是鞅.

例 3 设 η 是独立伯努利随机变量, $\mathscr{D}_1\preccurlyeq\mathscr{D}_2\preccurlyeq\cdots\preccurlyeq\mathscr{D}_n$, 且

$$\xi_k=\mathbf{E}(\eta|\mathscr{D}_k). \tag{2}$$

那么, 序列 $\xi=(\xi_k,\mathscr{D}_k)_{1\leqslant k\leqslant n}$ 是鞅. 事实上, $\mathbf{E}(\eta|\mathscr{D}_k)$ 显然 $\mathscr{D}_k-$ 可测, 且由 §8 的 (20) 式, 可见

$$\mathbf{E}(\xi_{k+1}|\mathscr{D}_k)=\mathbf{E}[\mathbf{E}(\eta|\mathscr{D}_{k+1})|\mathscr{D}_k]=\mathbf{E}(\eta|\mathscr{D}_k)=\xi_k.$$

因此, 我们指出, 如果 $\xi=(\xi_k,\mathscr{D}_k)_{1\leqslant k\leqslant n}$ 是任意鞅, 则由 §8 的 (20) 式, 可见

$$\xi_k=\mathbf{E}(\xi_{k+1}|\mathscr{D}_k)=\mathbf{E}[\mathbf{E}(\xi_{k+2}|\mathscr{D}_{k+1})|\mathscr{D}_k]=\mathbf{E}(\xi_{k+2}|\mathscr{D}_k)=\cdots=\mathbf{E}(\xi_n|\mathscr{D}_k). \tag{3}$$

于是, 一切鞅 $\xi=(\xi_k,\mathscr{D}_k)_{1\leqslant k\leqslant n}$ 的集合, 包含所有形如 (2) 式的鞅. (注意, 对于无穷序列 $\xi=(\xi_k,\mathscr{D}_k)_{k\geqslant 1}$, 一般并非如此; 参见第七章 §1 练习题 6.)

例 4 设 $\eta_1, \cdots, \eta_n$ 是独立同分布随机变量序列, $S_k = \eta_1 + \cdots + \eta_k$;

$$\mathscr{D}_1 = \mathscr{D}_{S_n}, \mathscr{D}_2 = \mathscr{D}_{S_n, S_{n-1}}, \cdots, \mathscr{D}_n = \mathscr{D}_{S_n, \cdots, S_1}.$$

证明序列 $\xi = (\xi_k, \mathscr{D}_k)_{1 \leqslant k \leqslant n}$, 其中

$$\xi_1 = \frac{S_n}{n}, \xi_2 = \frac{S_{n-1}}{n-1}, \cdots, \xi_k = \frac{S_{n+1-k}}{n+1-k}, \cdots, \xi_n = S_1,$$

是鞅. 首先, 显然 $\mathscr{D}_k \preccurlyeq \mathscr{D}_{k+1}$ 且 ξ_k 为 $\mathscr{D}_k-$ 可测; 其次, 由于对称性, 对于 $j \leqslant n-k+1$, 有

$$\mathbf{E}(\eta_j|\mathscr{D}_k) = \mathbf{E}(\eta_1|\mathscr{D}_k) \tag{4}$$

(与 §8 中 (20) 式比较). 因此

$$(n-k+1)\mathbf{E}(\eta_1|\mathscr{D}_k) = \sum_{j=1}^{n-k+1} \mathbf{E}(\eta_j|\mathscr{D}_k) = \mathbf{E}(S_{n-k+1}|\mathscr{D}_k) = S_{n-k+1}.$$

于是

$$\xi_k = \frac{S_{n-k+1}}{n-k+1} = \mathbf{E}(\eta_1|\mathscr{D}_k),$$

故由例 3 可见序列 $\xi = (\xi_k, \mathscr{D}_k)_{1 \leqslant k \leqslant n}$ 是鞅.

注 由已证明的序列 $\xi = (\xi_k, \mathscr{D}_k)_{1 \leqslant k \leqslant n}$ 的鞅性可知, 为何有时称 $(S_k/k)_{1 \leqslant k \leqslant n}$ 构成逆鞅或"反向鞅"的原因 (对照第七章 §1 练习题 5).

例 5 设 $\eta_1, \cdots, \eta_n$ 是独立伯努利随机变量:

$$\mathbf{P}\{\eta_i = +1\} = \mathbf{P}\{\eta_i = -1\} = \frac{1}{2}, S_k = \eta_1 + \cdots + \eta_k.$$

设 A 和 B 是两个整数, $A < 0 < B$. 那么, 对于任意 $0 < \lambda < \pi/2$, 序列 $\xi = (\xi_k, \mathscr{D}_k)$, 其中 $\mathscr{D}_k = \mathscr{D}_{S_1, \cdots, S_k}$, 且

$$\xi_k = (\cos \lambda)^{-k} \exp\left\{ i\lambda \left(S_k - \frac{A+B}{2} \right) \right\} \tag{5}$$

构成复鞅 (ξ_k 的实部和虚部都是鞅).

3. 停止时间 由鞅的定义, 可见对于一切 k, 数学期望 $\mathbf{E}\xi_k$ 相同:

$$\mathbf{E}\xi_k = \mathbf{E}\xi_1.$$

事实上, 如果将 (确定性) 时间 k 换成所谓停止时间, 鞅的这一性质仍然成立. 对于相应的提法, 我们引进下面的定义.

定义 2 随机变量 $\tau = \tau(\omega)$ (关于分割 $(\mathscr{D}_k)_{1 \leqslant k \leqslant n}, \mathscr{D}_1 \preccurlyeq \mathscr{D}_2 \preccurlyeq \cdots \preccurlyeq \mathscr{D}_n$) 称做停止时间, 如果对于 $k = 1, \cdots, n$, 随机变量 $I_{\{\tau = k\}}(\omega)$ 是 $\mathscr{D}_k-$ 可测的.

假如把分割 $\mathscr{D}_k$ 视为 k 步观测结果诱导的 (例如, 由随机变量 $\eta_1, \cdots, \eta_k$ 诱导的分割 $\mathscr{D}_k = \mathscr{D}_{\eta_1, \cdots, \eta_k}$), 则变量 $I_{\{\tau = k\}}(\omega)$ 的 $\mathscr{D}_k-$ 可测性表示: 事件 $\{\tau = k\}$ 存在与否仅取决于 k 步的观测结果 (而不依赖于"将来").

如果 $\mathscr{B}_k = \alpha(\mathscr{D}_k)$, 则变量 $I_{\{\tau = k\}}(\omega)$ 的 $\mathscr{D}_k-$ 可测性等价于假设:

$$\{\tau = k\} \in \mathscr{B}_k. \tag{6}$$

我们已经遇到过停止时间的例子: §9 和 §10 引进的时间 τ_k^x, σ_{2n}, 就是停止时间的例. 上面列举的停止时间是形如

$$\begin{aligned} \tau^A &= \min\{0 < k \leqslant n : \xi_k \in A\}, \\ \sigma^A &= \min\{0 \leqslant k \leqslant n : \xi_k \in A\} \end{aligned} \tag{7}$$

的停止时间的特殊情形. (相应为在 0 之后首次到达 和首次) 到达某一序列 $\xi_0, \xi_1, \cdots, \xi_n$ 的集合 A 的时间.

4. 停止时间的应用

定理 1 设 $\xi = (\xi_k, \mathscr{D}_k)_{1 \leqslant k \leqslant n}$ 是鞅, 而 τ 是关于分割 $(\mathscr{D}_k)_{1 \leqslant k \leqslant n}$ 的某一停止时间, 则

$$\mathbf{E}(\xi_\tau|\mathscr{D}_1) = \xi_1, \tag{8}$$

其中

$$\xi_\tau = \sum_{k=1}^{n} \xi_k I_{\{\tau = k\}} \tag{9}$$

而

$$\mathbf{E}\xi_\tau = \mathbf{E}\xi_1. \tag{10}$$

证明 (仿照 §9 中 (29) 式的证明). 设 $D \in \mathscr{D}_1$, 则利用条件数学期望的性质, 并注意到 (3) 的鞅性, 有

$$\begin{aligned} \mathbf{E}(\xi_\tau|D) &= \frac{\mathbf{E}(\xi_\tau I_D)}{\mathbf{P}(D)} = \frac{1}{\mathbf{P}(D)} \sum_{l=1}^{n} \mathbf{E}(\xi_l I_{\{\tau = l\}} I_D) \\ &= \frac{1}{\mathbf{P}(D)} \sum_{l=1}^{n} \mathbf{E}[\mathbf{E}(\xi_n|\mathscr{D}_l) I_{\{\tau = l\}} I_D] = \frac{1}{\mathbf{P}(D)} \sum_{l=1}^{n} \mathbf{E}[\mathbf{E}(\xi_n I_{\{\tau = l\}} I_D|\mathscr{D}_l)] \\ &= \frac{1}{\mathbf{P}(D)} \sum_{l=1}^{n} \mathbf{E}[\xi_n I_{\{\tau = l\}} I_D] = \frac{1}{\mathbf{P}(D)} \mathbf{E}(\xi_n I_D) = \mathbf{E}(\xi_n|D), \end{aligned}$$

从而

$$\mathbf{E}(\xi_\tau|\mathscr{D}_1) = \mathbf{E}(\xi_n|\mathscr{D}_1) = \xi_1.$$

由此显然地得等式 $\mathbf{E}\xi_\tau = \mathbf{E}\xi_1$. □

系 对于例 1 的鞅 $(S_k,\mathscr{F}_k)_{1\leqslant k\leqslant n}$ 和任意 (关于 $\mathscr{F}_k$ 的) 停止时间 τ, 有等式

$$\mathbf{E}S_\tau=0,\quad \mathbf{E}S_\tau^2=\mathbf{E}\tau. \tag{11}$$

等式 (11) 称做瓦尔德 (A. Wald) 恒等式 (对照 §10, (29) 式和 (30) 式; 以及第七章 §2 的练习题 1 和定理 3).

利用定理 1 证明如下命题.

定理 2 ("表决" 定理) 设 $\eta_1,\cdots,\eta_n$ 是独立同分布随机变量序列, 其中 η_i 在集合上 $\{0,1,\cdots\}$ 只取有限个值, $S_k=\eta_1+\cdots+\eta_k$, $1\leqslant k\leqslant n$, 则

$$\mathbf{P}\{S_k<k,k\in[1,n]|S_n\}=\left(1-\frac{S_n}{n}\right)^+. \tag{12}$$

其中 $a^+=\max(a,0)$.

证明 (12) 式在集合 $\{\omega:S_n\geqslant n\}$ 上显然. 因此, 只需证明对于使 $S_n<n$ 的基本事件证明 (12).

考虑在例 4 中引进的鞅 $\xi=(\xi_k,\mathscr{G}_k)_{1\leqslant k\leqslant n}$, 其中

$$\xi_k=\frac{S_{n+1-k}}{n+1-k},\quad \mathscr{G}_k=\mathscr{G}_{S_{n+1-k},\cdots,S_n}.$$

记

$$\tau=\min\{1\leqslant k\leqslant n:\xi_k\geqslant 1\}.$$

在集合

$$\{\xi_k<1,k\in[1,n]\}=\left\{\max_{1\leqslant l\leqslant n}\frac{S_l}{l}<1\right\}$$

上设 $\tau=n$. 显然, 在此集合上 $\xi_\tau=\xi_n=S_1=0$, 因而

$$\left\{\max_{1\leqslant l\leqslant n}\frac{S_l}{l}<1\right\}\supseteq\left\{\max_{1\leqslant l\leqslant n}\frac{S_l}{l}<1,S_n<n\right\}\subseteq\{\xi_\tau=0\}. \tag{13}$$

现在考虑同时使

$$S_n<n,\ \max_{1\leqslant l\leqslant n}\frac{S_l}{l}\geqslant 1$$

的基本事件. 记 $\sigma=n+1-\tau$. 易见,

$$\sigma=\min\{1\leqslant k\leqslant n:S_k\geqslant k\},$$

因此 (由于 $S_n<n$) $\sigma<n$, $S_\sigma\geqslant\sigma$, 而 $S_{\sigma+1}<\sigma+1$. 可见 $\eta_{\sigma+1}=S_{\sigma+1}-S_\sigma<(\sigma+1)-\sigma=1$, 即 $\eta_{\sigma+1}=0$. 所以 $\sigma\leqslant S_\sigma=S_{\sigma+1}<\sigma+1$, 因而 $S_\sigma=\sigma$ 且

$$\xi_\tau=\frac{S_{n+1-\tau}}{n+1-\tau}=\frac{S_\sigma}{\sigma}=1.$$

从而

$$\left\{\max_{1\leqslant l\leqslant n}\frac{S_l}{l}\geqslant 1,S_n<n\right\}\subseteq\{\xi_\tau=1\}. \tag{14}$$

由 (13) 式和 (14) 式, 可见

$$\left\{\max_{1\leqslant l\leqslant n}\frac{S_l}{l}\geqslant 1,S_n<n\right\}=\{\xi_\tau=1\}\cap\{S_n<n\}.$$

因此, 在集合 $S_n<n$ 上

$$\mathbf{P}\left\{\max_{1\leqslant l\leqslant n}\frac{S_l}{l}\geqslant 1\Big|S_n\right\}=\mathbf{P}\{\xi_\tau=1|S_n\}=\mathbf{E}(\xi_\tau|S_n),$$

其中因为 ξ_τ 只有 0 和 1 两个可能值, 所以最后一个等式成立.

最后, 注意到, $\mathbf{E}(\xi_\tau|S_n)=\mathbf{E}(\xi_\tau|\mathscr{G}_1)$, 且由定理 1 知 $\mathbf{E}(\xi_\tau|\mathscr{G}_1)=\xi_1=S_n/n$. 于是, 在集合 $\{S_n<n\}$ 上, 有

$$\mathbf{P}\{S_k<k,k\in[1,n]|S_n\}=1-\frac{S_n}{n}. \qquad \square$$

我们用定理 2 给出 §10 中引理 1 的另一证明, 并说明表决定理名称的来历.

设 $\xi_1,\cdots,\xi_n$ 是独立伯努利随机变量序列, 且

$$\mathbf{P}\{\xi_i=1\}=\mathbf{P}\{\xi_i=-1\}=\frac{1}{2};$$

$S_k=\xi_1+\cdots+\xi_k$, 而 a 和 b 非负正数: $a-b>0$, $a+b=n$. 现在证明,

$$\mathbf{P}\{S_1>0,\cdots,S_n>0|S_n=a-b\}=\frac{a-b}{a+b}. \tag{15}$$

事实上, 由对称性, 可见

$$\begin{aligned}
&\mathbf{P}\{S_1>0,\cdots,S_n>0|S_n=a-b\}\\
&=\mathbf{P}\{S_1<0,\cdots,S_n<0|S_n=-(a-b)\}\\
&=\mathbf{P}\{S_1+1<1,\cdots,S_n+n<n|S_n+n=n-(a-b)\}\\
&=\mathbf{P}\{\eta_1<1,\cdots,\eta_1+\cdots+\eta_n<n|\eta_1+\cdots+\eta_n=n-(a-b)\}\\
&=\left[1-\frac{n-(a-b)}{n}\right]^+=\frac{a-b}{n}=\frac{a-b}{a+b},
\end{aligned}$$

其中设 $\eta_k=\xi_k+1$, 并用到了 (12) 式.

§10 中的 (5) 式曾经借助反射原理证明的, 现在显然可以由式 (15) 导出 §10 中的 (5) 式.

假设有两个候选人 A 和 B, 我们可以把 $\xi_i=1$ 视为投 A 一票, 而把 $\xi_i=-1$ 视为投 B 一票. 假如有 k 个人参加投票, 而 S_k 是 A 和 B 得票的差额; 假设 A 最后共得 a 票, 而 B 最后共得 b 票, 且 $a-b>0$, $a+b=n$, 则

$$\mathbf{P}\{S_1>0,\cdots,S_n>0|S_n=a-b\}$$

表示 A 得票始终领先于 B 的概率. 根据 (15) 此概率等于 $(a-b)/n$.

5. 练习题

1. 设 $\mathscr{D}_0 \preccurlyeq \mathscr{D}_1 \preccurlyeq \cdots \preccurlyeq \mathscr{D}_n$ 是分割序列, $\mathscr{D}_0 = \{\Omega\}$; η_k 是 $\mathscr{D}_k$ – 可测随机变量, $1 \leqslant k \leqslant n$. 证明序列 $\xi = (\xi_k, \mathscr{D}_k)_{1\leqslant k\leqslant n}$, 其中

$$\xi_k = \sum_{l=1}^{k} [\eta_l - \mathbf{E}(\eta_l|\mathscr{D}_{l-1})]$$

是鞅.

2. 设对于随机变量 $\eta_1, \cdots, \eta_k$, $\mathbf{E}(\eta_k|\eta_1, \cdots, \eta_{k-1}) = 0$. 证明序列 $\xi = (\xi_k)_{1\leqslant k\leqslant n}$, 其中 $\xi_1 = \eta_1$ 且

$$\xi_{k+1} = \sum_{i=1}^{k} \eta_{i+1} f_i(\eta_1, \cdots, \eta_i),$$

其中 f_i 是构成鞅的某一函数.

3. 证明, 任何鞅 $\xi = (\xi_k, \mathscr{D}_i)_{1\leqslant i\leqslant n}$ 都有不相关增量: 对于 $a < b < c < d$, 有

$$\operatorname{cov}(\xi_d - \xi_c, \xi_b - \xi_a) = 0.$$

4. 设对于随机变量序列 $\xi = (\xi_1, \cdots, \xi_n)$, 其中 ξ_k 为 $\mathscr{D}_k$ – 可测 ($\mathscr{D}_1 \preccurlyeq \mathscr{D}_2 \preccurlyeq \cdots \preccurlyeq \mathscr{D}_n$). 证明, 为使序列 $\xi = (\xi_1, \cdots, \xi_n)$ (关于分割系 $(\mathscr{D}_k)$) 为鞅的充分和必要条件是, 对于任意 (关于 $(\mathscr{D}_k)$ 的) 停止时间 τ, 有 $\mathbf{E}\xi_\tau = \mathbf{E}\xi_1$. (条件 "对于任意停止时间", 可以换成条件 "对于只有两个可能值的停止时间".)

5. 证明, 如果 $\xi = (\xi_k, \mathscr{D}_k)_{1\leqslant k\leqslant n}$ 是鞅, 而 τ 是停止时间, 则对于任意 k,

$$\mathbf{E}[\xi_n I_{\{\tau=k\}}] = \mathbf{E}[\xi_k I_{\{\tau=k\}}].$$

6. 设 $\xi = (\xi_k, \mathscr{D}_k)$ 和 $\eta = (\eta_k, \mathscr{D}_k)$ 是两个鞅, 且 $\xi_1 = \eta_1 = 0$. 证明

$$\mathbf{E}\xi_n\eta_n = \sum_{k=2}^{n} \mathbf{E}(\xi_k - \xi_{k-1})(\eta_k - \eta_{k-1}),$$

特别,

$$\mathbf{E}\xi_n^2 = \sum_{k=2}^{n} \mathbf{E}(\xi_k - \xi_{k-1})^2.$$

7. 设 $\eta_1, \cdots, \eta_n$ 是独立同分布机变量序列, 其中 $\mathbf{E}\eta_i = 0$. 证明序列 $\xi = (\xi_k)$, 其中

$$\xi_k = \left(\sum_{i=1}^{k} \eta_i\right)^2 - k\mathbf{E}\eta_1^2,$$

$$\xi_k = \frac{\exp\{\lambda(\eta_1 + \cdots + \eta_k)\}}{[\mathbf{E}\exp(\lambda\eta_1)]^k}$$

都是鞅.

8. 设 $\eta_1, \cdots, \eta_n$ 是独立同分布随机变量序列, 其值域为有限集合 Y. 设

$$f_0(y) = \mathrm{P}\{\eta_1 = y\} > 0, y \in Y; f_1(y) \geqslant 0, \sum_{y\in Y} f_1(y) = 1.$$

证明

$$\xi = (\xi_k, \mathscr{D}_k^\eta), \quad \xi_k = \frac{f_1(\eta_1)\cdots f_1(\eta_k)}{f_0(\eta_1)\cdots f_0(\eta_k)}, \quad \mathscr{D}_k^\eta = \mathscr{D}_{\eta_1, \cdots, \eta_k}$$

是鞅. (关系式 ξ_k 称做似然比, 在数理统计中有十分重要的应用.)

§12. 马尔可夫链. 遍历性定理. 强马尔可夫性

1. **马尔可夫链** 在以上研究的伯努利概型中, $\Omega = \{\omega : \omega = (x_1, \cdots, x_n), x_i = 0, 1\}$, 每个基本事件 ω 的概率为式 $\mathrm{P}(\{\omega\}) = p(\omega)$, 其中

$$p(\omega) = p(x_1)\cdots p(x_n), \quad p(x) = p^x q^{1-x}. \tag{1}$$

在此情况下随机变量 $\xi_1, \cdots, \xi_n$ 独立同分布, 其中 $\xi_i(\omega) = x_i$, 且

$$\mathbf{P}\{\xi_1 = x\} = \cdots = \mathbf{P}\{\xi_n = x\} = p(x), \quad x = 0, 1.$$

如果将 (1) 式换成

$$p(\omega) = p_1(x_1)\cdots p_n(x_n),$$

其中 $p_i(x) = p_i^x(1-p_i)^{1-x}$, $0 \leqslant p_i \leqslant 1$, 则这时随机变量 $\xi_1, \cdots, \xi_n$ 不再是独立的, 且一般也不是同分布的:

$$\mathbf{P}\{\xi_1 = x\} = p_1(x), \cdots, \mathbf{P}\{\xi_n = x\} = p_n(x).$$

我们考虑这类概型的一种推广, 将引出由非独立随机变量形成的所谓的马尔可夫链.

假设

$$\Omega = \{\omega : \omega = (x_0, x_1, \cdots, x_n), x_i \in X\},$$

其中 X 是某一有穷集合. 假设给定非负函数: $p_0(x), p_1(x, y), \cdots, p_n(x, y)$, 满足

$$\sum_{x\in X} p_0(x) = 1,$$
$$\sum_{y\in X} p_k(x, y) = 1, \quad k = 1, \cdots, n; \ x \in X. \tag{2}$$

对于每一 $\omega = (x_0, x_1, \cdots, x_n)$, 设 $\mathbf{P}(\{\omega\}) = p(\omega)$, 其中

$$p(\omega) = p_0(x_0)p_1(x_0, x_1)\cdots p_n(x_{n-1}, x_n). \tag{3}$$

不难验证

$$\sum_{\omega\in\Omega} p(\omega) = 1.$$

全体 $p(\omega)$, 以及空间 Ω 连同其一切子集的集系决定空间 $(\Omega, \mathscr{A}, \mathbf{P})$, 通常称之为形成马尔可夫链的试验模型.

引进随机变量 $\xi_0, \xi_1, \cdots, \xi_n$, 其中对于 $\omega = (x_1, \cdots, x_n)$, $\xi_i(\omega) = x_i$. 经简单计算, 可得

$$\begin{aligned} &\mathbf{P}\{\xi_0 = a\} = p_0(a), \\ &\mathbf{P}\{\xi_0 = a_0, \xi_1 = a_1, \cdots, \xi_k = a_k\} = p_0(a_0)p_1(a_0, a_1)\cdots p_k(a_{k-1}, a_k). \end{aligned} \tag{4}$$

现在对于上述概率模型 $(\Omega, \mathscr{A}, \mathbf{P})$, 证明条件概率的一条重要性质: 假设 $\mathbf{P}\{\xi_k = a_k, \cdots, \xi_0 = a_0\} > 0$.

$$\mathbf{P}\{\xi_{k+1} = a_{k+1}|\xi_k = a_k, \cdots, \xi_0 = a_0\} = \mathbf{P}\{\xi_{k+1} = a_{k+1}|\xi_k = a_k\}. \tag{5}$$

由于 (4) 式和条件概率的定义 (§3), 可见

$$\begin{aligned} \mathbf{P}\{\xi_{k+1} = a_{k+1}|\xi_k = a_k, \cdots, \xi_0 = a_0\} &= \frac{\mathbf{P}\{\xi_{k+1} = a_{k+1}, \cdots, \xi_0 = a_0\}}{\mathbf{P}\{\xi_k = a_k, \cdots, \xi_0 = a_0\}} \\ &= \frac{p_0(a_0)p_1(a_0, a_1)\cdots p_{k+1}(a_k, a_{k+1})}{p_0(a_0)p_1(a_0, a_1)\cdots p_k(a_{k-1}, a_k)} = p_{k+1}(a_k, a_{k+1}). \end{aligned}$$

类似可得

$$\mathbf{P}\{\xi_{k+1} = a_{k+1}|\xi_k = a_k\} = p_{k+1}(a_k, a_{k+1}), \tag{6}$$

因而性质 (5) 得证.

设 $\mathscr{D}_k^\xi = \mathscr{D}_{\xi_0, \cdots, \xi_k}$ 是随机变量 $\xi_0, \cdots, \xi_k$ 诱导的分割, 而 $\mathscr{B}_k^\xi = \alpha(\mathscr{D}_k^\xi)$.

那么, 根据 §8 引进的记号, 由 (5) 式可见

$$\mathbf{P}\{\xi_{k+1} = a_{k+1}|\mathscr{B}_k^\xi\} = \mathbf{P}\{\xi_{k+1} = a_{k+1}|\xi_k\} \tag{7}$$

或

$$\mathbf{P}\{\xi_{k+1} = a_{k+1}|\xi_0, \cdots, \xi_k\} = \mathbf{P}\{\xi_{k+1} = a_{k+1}|\xi_k\}.$$

注 鉴于 (5) 式和 (7) 式以及 0 概率事件, 暂时中断我们的叙述, 作对整个以后的全部内容至关重要的说明.

在推导 (5) 式时曾经假设 $\mathbf{P}\{\xi_k = a_k, \cdots, \xi_0 = a_0\} > 0$ (即 $\mathbf{P}\{\xi_k = a_k\} > 0$). 之所以需要这样, 是因为条件概率 $\mathbf{P}(A|B)$ (暂时!) 只有在 $\mathbf{P}(B) > 0$ 的条件下才有定义.

但是, 需要指出, 如果 $B = \{\xi_k = a_k, \cdots, \xi_0 = a_0\}$ 而且 $\mathbf{P}(B) = 0$ (因此, 对于 $C = \{\xi_k = a_k\}$, $\mathbf{P}(C) = 0$), 则 "路径" 应视为不可实现的. 那么, 事件 $\{\xi_{k+1} = a_k\}$ 关于这条不可实现的 "路径" 的条件概率的问题, 也就毫无实际意义.

因此, 为明确起见, 我们以后把条件概率定义为:

$$\mathbf{P}(A|B) = \begin{cases} \dfrac{\mathbf{P}(AB)}{\mathbf{P}(B)}, & 若\ \mathbf{P}(B) > 0, \\ 0, & 若\ \mathbf{P}(B) = 0. \end{cases}$$

在这样的定义下, 无需诸如 $\mathbf{P}\{\xi_k = a_k, \cdots, \xi_0 = a_0\} > 0$ 的任何前提, (5) 式和 (7) 式仍然是正确的.

应强调指出, 所指出的与零概率事件有关的困难, 在概率论的论述中相当典型. 在第二章 §7 中将引进 (关于任意分割、σ− 代数 …… 的) 条件概率的一般定义, 可以自然地 "用于零概率的" 情形.

如果利用明显的等式

$$\mathbf{P}(AB|C) = \mathbf{P}(A|BC)\mathbf{P}(B|C),$$

则由 (7) 式, 有

$$\mathbf{P}\{\xi_n = a_n, \cdots, \xi_{k+1} = a_{k+1}|\mathscr{B}_k^\xi\} = \mathbf{P}\{\xi_n = a_n, \cdots, \xi_{k+1} = a_{k+1}|\xi_k\}, \tag{8}$$

或

$$\mathbf{P}\{\xi_n = a_n, \cdots, \xi_{k+1} = a_{k+1}|\xi_0, \cdots, \xi_k\} = \mathbf{P}\{\xi_n = a_n, \cdots, \xi_{k+1} = a_{k+1}|\xi_k\}. \tag{9}$$

此等式有如下直观的解释. 假设 ξ_k 表示质点在 "现在", $(\xi_0, \cdots, \xi_{k-1})$ 表示质点在 "过去", $(\xi_{k+1}, \cdots, \xi_n)$ 表示质点在 "将来". 那么, (9) 式表示: 在 "过去" $(\xi_0, \cdots, \xi_{k-1})$ 和 "现在" ξ_k 固定的情况下, "将来" $(\xi_{k+1}, \cdots, \xi_n)$ 仅依赖于 "现在" ξ_k, 而与质点如何到达点 ξ_k 无关, 即 "将来" 不依赖于 "过去".

设 $\mathrm{J} = \{\xi_n = a_n, \cdots, \xi_{k+1} = a_{k+1}\}$, $\mathrm{X} = \{\xi_k = a_k\}$, $\mathrm{G} = \{\xi_{k-1} = a_{k-1}, \cdots, \xi_0 = a_0\}$[①], 则由 (9) 式可见

$$\mathbf{P}(\mathrm{J}|\mathrm{XG}) = \mathbf{P}(\mathrm{J}|\mathrm{X}),$$

由此容易求出

$$\mathbf{P}(\mathrm{JG}|\mathrm{X}) = \mathbf{P}(\mathrm{J}|\mathrm{X})\mathbf{P}(\mathrm{G}|\mathrm{X}). \tag{10}$$

换句话说, 由 (7) 式可见, 在固定 "现在" X 的情况下, "将来" J 和 "过去" G 相互独立. 不难证明逆命题: 如果对于任意 $k = 0, 1, \cdots, n-1$, (10) 式成立, 则对于任意 $k = 0, 1, \cdots, n-1$, 性质 (7) 也成立.

"将来" 和 "过去" 的独立性, 或者说, 在固定 "现在" 的情况下, "将来" 和 "过去" 相互独立, 通常称做马尔可夫性, 而相应的随机变量序列 $\xi_1, \cdots, \xi_n$ 称做马尔可夫链.

① J 表示 "将来", X 表示 "现在", G 表示 "过去". —— 译者

这样, 如果由 (3) 式决定基本事件的 "权重" $p(\omega)$, 则序列 $\xi=(\xi_0,\cdots,\xi_n)$ 构成马尔可夫链, 式中 $\xi_i(\omega)=x_i$.

于是, 引出如下定义.

定义 设 $(\Omega,\mathscr{A},\mathbf{P})$ 是某一 (有限) 概率空间, $\xi=(\xi_0,\cdots,\xi_n)$ 是 (有限) 随机变量序列, 而 (有限) 集合 X 是其值域. 如果满足条件 (7), 则序列 $\xi=(\xi_0,\cdots,\xi_n)$ 称做 (有限) *马尔可夫链*.

集合 X 称做马尔可夫链的*相空间*或*状态空间*. 概率的全体

$$\{p_0(x)\},p_0(x)=\mathbf{P}\{\xi_0=x\},x\in X$$

称为*初始分布*; 而矩阵

$$(p_k(x,y)),x,y\in X;p_k(x,y)=\mathbf{P}\{\xi_k=y|\xi_{k-1}=x\}$$

称为在时刻 $k=1,\cdots,n$ (自状态 x 到状态 y) 的*转移概率矩阵*.

如果转移概率不依赖于 k: $p_k(x,y)=p(x,y)$, 则序列 $\xi=(\xi_0,\cdots,\xi_n)$ 称为转移概率矩阵是 $(p(x,y))$ 的*齐次马尔可夫链*.

注意, $(p(x,y))$ 是随机矩阵: 其元素非负, 任何一行元素之和等于 1,

$$\sum_y p(x,y)=1,\quad x\in X.$$

假设相空间 X 是有限个整数点的集合 (例如, $X=\{0,1,\cdots,N\}$, $X=\{0,\pm1,\cdots,\pm N\}$ 等). 此外, 根据习惯记 $p_i=p_0(i)$, $p_{ij}=p(i,j)$.

显然, 齐次马尔可夫链的性质, 完全决定于初始分布 p_i 和转移概率 p_{ij}. 在有些情形下, 为描绘马尔可夫链的演变, 不明显写出矩阵 (p_{ij}), 而是运用有向网络, 其结点表示 X 中的状态.

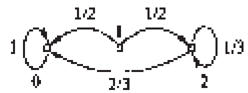
而箭头表示自状态 i 指向状态 j; 箭线上的 p_{ij} 是质点可能自点 i 转移到点 j 的概率.

例 1 设 $X=\{0,1,2\}$ 和

$$(p_{ij})=\begin{pmatrix}1&0&0\\1/2&0&1/2\\2/3&0&1/3\end{pmatrix}.$$

该矩阵对应下面的网络图:

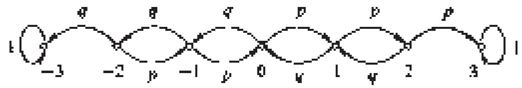
注意, 这里状态 0 是 "吸收状态": 由于 $p_{00}=1$, 可见一旦质点进入状态 0, 则它就停留在此状态. 质点自状态 1 以相同的概率进入相邻的状态 0 或 2; 对于状态 2, 质点留在其中的概率为 1/3, 而由状态 2 转移到状态 0 的概率为 2/3.



例 2 设 $X=\{0,\pm1,\cdots,\pm N\}$, $p_{00}=1$, $p_{NN}=p_{-N,-N}=1$, 而对于 $|i|<N$,

$$p_{ij}=\begin{cases}p, & 若\ j=i+1,\\ q, & 若\ j=i-1,\\ 0, & 其他情形.\end{cases}\tag{11}$$

可以用图形表示对应于这样链的转移情况 ($N=3$):



该链对应于上面研究的甲、乙两人的赌博: 每人的赌金为 N, 甲每一步以概率 p 赢乙得 +1, 以概率 q 输给乙得 −1. 假如把状态 i 视为甲赢得乙的金额, 则到达状态 N 和 $-N$, 分别表示乙破产和甲破产.

事实上, 如果 $\eta_1,\eta_2,\cdots,\eta_n$ 是相互独立的伯努利随机变量, 其中 $\mathbf{P}\{\eta_i=+1\}=p$, $\mathbf{P}\{\eta_i=-1\}=q$, 而 $S_k=\eta_1+\eta_2+\cdots+\eta_k$ 是甲赢得乙的金额, 则序列 $S_0,S_1,\cdots,S_n$ $(S_0=0)$ 形成马尔可夫链且 $p_0=1$, 而转移概率矩阵为 (11). 因为

$$\begin{aligned}&\mathbf{P}\{S_{k+1}=j|S_k=i_k,S_{k-1}=i_{k-1},\cdots,S_1=i_1\}\\&=\mathbf{P}\{S_k+\eta_{k+1}=j|S_k=i_k,S_{k-1}=i_{k-1},\cdots,S_1=i_1\}\\&=\mathbf{P}\{S_k+\eta_{k+1}=j|S_k=i_k\}=\mathbf{P}\{\eta_{k+1}=j-i_k\}.\end{aligned}$$

马尔可夫链 $S_0,S_1,\cdots,S_n$ 具有十分简单的构造:

$$S_{k+1}=S_k+\eta_{k+1},\quad 0\leqslant k\leqslant n-1,$$

其中 $\eta_1,\eta_2,\cdots,\eta_n$ 是独立随机变量序列.

同理可得, 如果 $\xi_0,\eta_1,\cdots,\eta_n$ 是独立随机变量序列, 则序列 $\xi_0,\xi_1,\cdots,\xi_n$ 也是马尔可夫链, 其中

$$\xi_{k+1}=f_k(\xi_k,\eta_{k+1}),\quad 0\leqslant k\leqslant n-1.\tag{12}$$

因此应该指出, 自然把这样构造的马尔可夫链, 视为由递推关系式

$$x_{k+1}=f_k(x_k)$$

决定的, 如 (确定性) 序列 $(x_0,x_1,\cdots,x_n)$ 的概率的类似.

现在再举一个 "排队论" 中, 形如 (12) 式的马尔可夫链的例子.

例 3 假设在出租车站每单位时间有一辆车到达. 假如在停车站无乘客等候, 则车立即离去. 以 η_k 表示于 k 时到达车站等车的乘客数量, 并且假定 $\eta_1, \eta_2, \cdots, \eta_n$ 是相互独立的随机变量. 设 ξ_k 为在时刻 k 乘客的队长. 其中 $\xi_0 = 0$. 那么, 如果 $\xi_k = i$, 则在下一时刻 $k+1$ 的队长等于

$$j = \begin{cases} \eta_{k+1}, & \text{若 } i = 0, \\ i - 1 + \eta_{k+1}, & \text{若 } i \geqslant 1. \end{cases}$$

换句话说,

$$\xi_{k+1} = (\xi_k - 1)^+ + \eta_{k+1}, \quad 0 \leqslant k \leqslant n-1,$$

其中 $a^+ = \max(a, 0)$, 从而序列 $\xi = (\xi_0, \cdots, \xi_n)$ 是马尔可夫链.

例 4 此例涉及分支过程理论. 所谓离散时间分支过程, 是这样一个随机变量序列 $\xi_0, \xi_1, \cdots, \xi_n$, 其中 ξ_k 表示在时刻 k 质点的数量, 而且质点的生与灭的情况是按如下方式进行的: 每个质点以概率 $p_j (j = 0, 1, \cdots, M)$ 裂变为 j 个质点, 裂变既不依赖于其他质点, 也与 "过去历史" 无关.

假设在初始时刻总共只有一个质点, $\xi_0 = 1$. 假如在时刻 k 有编号为 $1, 2, \cdots, \xi_k$ 的 ξ_k 个质点, 那么根据上面的描述, ξ_{k+1} 等于随机多个随机变量之和:

$$\xi_{k+1} = \eta_1^{(k)} + \cdots + \eta_{\xi_k}^{(k)},$$

其中 $\eta_i^{(k)}$ 是第 i 号质点裂变的质点的个数. 显然, 若 $\xi_k = 0$, 则 $\xi_{k+1} = 0$. 假设所有随机变量 $\eta_j^{(k)}, k \geqslant 0, j \geqslant 1$ 相互独立, 则

$$\begin{aligned} &\mathbf{P}\{\xi_{k+1} = i_{k+1} | \xi_k = i_k, \xi_{k-1} = i_{k-1}, \cdots\} \\ &= \mathbf{P}\{\xi_{k+1} = i_{k+1} | \xi_k = i_k\} = \mathbf{P}\{\eta_1^{(k)} + \cdots + \eta_{i_k}^{(k)} = i_{k+1}\}. \end{aligned}$$

由此可见, 序列 $\xi_0, \xi_1, \cdots, \xi_n$ 是马尔可夫链.

特别重要的是下面的情形: 每一个质点, 或者是以概率 q 灭绝, 或者以概率 p 裂变为两个质点, 其中 $p + q = 1$. 对于这种情形, 容易经计算得到: 转移概率

$$p_{ij} = \mathbf{P}\{\xi_{k+1} = j | \xi_k = i\}$$

由下面的公式给出.

$$p_{ij} = \begin{cases} C_i^{j/2} p^{j/2} q^{i-j/2}, & \text{若 } j = 0, \cdots, 2i, \\ 0, & \text{其他}. \end{cases}$$

2. 柯尔莫戈洛夫 – 查普曼 (Kolmogorov-Chapman) 方程 记 $\xi = (\xi_k, \Pi, \mathbb{P})$ 为齐次马尔可夫链, 其初始概率行向量为 $\Pi = (p_i)$, 而转移概率矩阵为 $\mathbb{P} = (p_{ij})$. 显然,

$$p_{ij} = \mathbf{P}\{\xi_1 = j | \xi_0 = i\} = \cdots = \mathbf{P}\{\xi_n = j | \xi_{n-1} = i\}.$$

记

$$p_{ij}^{(k)} = \mathbf{P}\{\xi_k = j | \xi_0 = i\} \quad (= \mathbf{P}\{\xi_{k+l} = j | \xi_l = i\}, l = 1, 2, \cdots)$$

为经 k 步由状态 i 到状态 j 的转移概率, 而

$$p_j^{(k)} = \mathbf{P}\{\xi_k = j\}$$

是质点在 k 时位于 j 点的概率. 亦设

$$\Pi^{(k)} = (p_1^{(k)}), \quad \mathbb{P}^{(k)} = (p_{ij}^{(k)}).$$

我们证明, 转移概率满足 "柯尔莫戈洛夫 – 查普曼方程":

$$p_{ij}^{(k+l)} = \sum_\alpha p_{i\alpha}^{(k)} p_{\alpha j}^{(l)}, \tag{13}$$

其矩阵形式为

$$\mathbb{P}^{(k+l)} = \mathbb{P}^{(k)} \mathbb{P}^{(l)}. \tag{14}$$

基于全概率公式和马尔可夫性, 关系式 (13) 很容易证明:

$$\begin{aligned} p_{ij}^{(k+l)} &= \mathbf{P}\{\xi_{k+l} = j | \xi_0 = i\} = \sum_\alpha \mathbf{P}\{\xi_{k+l} = j, \xi_k = \alpha | \xi_0 = i\} \\ &= \sum_\alpha \mathbf{P}\{\xi_{k+l} = j | \xi_k = \alpha\} \mathbf{P}\{\xi_k = \alpha | \xi_0 = i\} = \sum_\alpha p_{\alpha j}^{(l)} p_{i\alpha}^{(k)}. \end{aligned}$$

方程 (13) 的两种特殊情形最重要: 后向方程

$$p_{ij}^{(l+1)} = \sum_\alpha p_{i\alpha} p_{\alpha j}^{(l)} \tag{15}$$

和前向方程

$$p_{ij}^{(k+1)} = \sum_\alpha p_{i\alpha}^{(k)} p_{\alpha j} \tag{16}$$

(见图 22 和图 23).

前向方程和后向方程的矩阵形式相应为:

$$\mathbb{P}^{(k+1)} = \mathbb{P}^{(k)} \mathbb{P}, \tag{17}$$

$$\mathbb{P}^{(k+1)} = \mathbb{P} \mathbb{P}^{(k)}. \tag{18}$$

对于 (无条件) 概率 $p_j^{(k)}$, 类似地可以得到

$$p_j^{(k+l)} = \sum_\alpha p_\alpha^{(k)} p_{\alpha j}^{(l)}, \tag{19}$$

而其矩阵形式为

$$\Pi^{(k+l)} = \Pi^{(k)} \mathbb{P}^{(l)}.$$

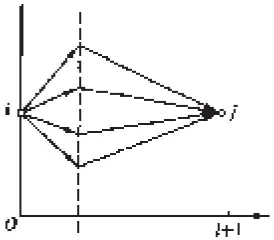


图 22 后向方程

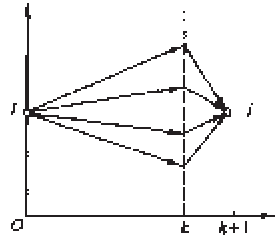


图 23 前向方程

特别, 前向方程为

$$\Pi^{(k+1)} = \Pi^{(k)}\mathbb{P},$$

而后向方程为

$$\Pi^{(k+1)} = \Pi^{(1)}\mathbb{P}^{(k)}.$$

由于 $\mathbb{P}^{(1)} = \mathbb{P}, \Pi^{(0)} = \Pi$, 可见

$$\mathbb{P}^{(k)} = \mathbb{P}^k, \quad \Pi^{(k)} = \Pi\mathbb{P}^k.$$

因此, 对于齐次马尔可夫链, k 步的转移概率 $p_{ij}^{(k)}$ 是矩阵 $\mathbb{P}$ 的 k 次幂. 从而, 齐次马尔可夫链的许多性质, 可以用矩阵方法进行研究.

例 考虑具有 0 和 1 两个状态, 而转移矩阵为

$$\mathbb{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$$

的齐次马尔可夫链. 不难计算

$$\mathbb{P}^2 = \begin{pmatrix} p_{00}^2 + p_{01}p_{10} & p_{01}(p_{00}+p_{11}) \\ p_{10}(p_{00}+p_{11}) & p_{11}^2 + p_{01}p_{10} \end{pmatrix},$$

而由归纳法, 有

$$\mathbb{P}^n = \frac{1}{2-p_{00}-p_{11}} \begin{pmatrix} 1-p_{11} & 1-p_{00} \\ 1-p_{11} & 1-p_{00} \end{pmatrix} + \frac{(p_{00}+p_{11}-1)^n}{2-p_{00}-p_{11}} \begin{pmatrix} 1-p_{00} & -(1-p_{00}) \\ -(1-p_{11}) & 1-p_{11} \end{pmatrix}$$

(假设 $|p_{00}+p_{11}-1|<1$).

由此可见, 如果矩阵 $\mathbb{P}$ 的元素使 $|p_{00}+p_{11}-1|<1$ (特别, 如所有转移概率 p_{ij} 都大于 0), 则当 $n\to\infty$ 时, 有

$$\mathbb{P}^n \to \frac{1}{2-p_{00}-p_{11}} \begin{pmatrix} 1-p_{11} & 1-p_{00} \\ 1-p_{11} & 1-p_{00} \end{pmatrix}, \tag{20}$$

从而

$$\lim_n p_{i0}^{(n)} = \frac{1-p_{11}}{2-p_{00}-p_{11}}, \quad \lim_n p_{i1}^{(n)} = \frac{1-p_{00}}{2-p_{00}-p_{11}}.$$

于是, 如果 $|p_{00}+p_{11}-1|<1$, 则所研究的马尔可夫链的行为有如下规律性: 随着时间的推移, 初始状态对"质点处于不同状态的概率"的影响逐渐消失 ($p_{ij}^{(n)}$ 收敛于不依赖于 i 的极限值 π_j, 并形成概率分布: $\pi_0\geqslant 0, \pi_1\geqslant 0, \pi_0+\pi_1=1$); 假如除此之外假设 $p_{ij}>0$, 则极限值 $\pi_0>0, \pi_1>0$ (对照下面的定理 1).

3. 遍历性 下面的定理描绘的广泛的一类马尔可夫链, 具有所谓遍历性: 极限

$$\pi_j = \lim_n p_{ij}^{(n)}$$

不但存在且不依赖于 i 构成概率分布

$$\pi_j \geqslant 0, \quad \sum_j \pi_j = 1,$$

而且对于所有 j, $\pi_j>0$ (这样的分布称做遍历分布, 详见第八章 §3).

定理 1 (遍历性定理) 设 $\mathbb{P} = \|p_{ij}\|$ 是马尔可夫链的转移概率矩阵, 且状态 $X = \{1,2,\cdots,N\}$ 有穷.

a) 如果对于某个 n_0, 有

$$\min_{i,j} p_{ij}^{(n_0)} > 0, \tag{21}$$

则存在 $\pi_1,\cdots,\pi_N$, 使

$$\pi_j > 0, \sum_j \pi_j = 1, \tag{22}$$

而且对每个 $j\in X$ 和任意 $i\in X$, 有

$$p_{ij}^{(n)} \to \pi_j, n\to\infty. \tag{23}$$

b) 相反, 如果存在满足条件 (22) 和 (23) 的数 $\pi_1,\cdots,\pi_N$, 则存在满足条件 (21) 的 n_0.

c) 式 (22) 中的数 $(\pi_1,\cdots,\pi_N)$ 满足方程组

$$\pi_j = \sum_\alpha \pi_\alpha p_{\alpha j}, \quad j=1,\cdots,N. \tag{24}$$

证明 a) 记

$$m_j^{(n)} = \min_i p_{ij}^{(n)}, \quad M_j^{(n)} = \max_i p_{ij}^{(n)}.$$

由于

$$p_{ij}^{(n+1)} = \sum_\alpha p_{i\alpha} p_{\alpha j}^{(n)}, \tag{25}$$

可见

$$m_j^{(n+1)} = \min_i p_{ij}^{(n+1)} = \min_i \sum_\alpha p_{i\alpha} p_{\alpha j}^{(n)} \geqslant \min_i \sum_\alpha p_{i\alpha} \min_\alpha p_{\alpha j}^{(n)} = m_j^{(n)},$$

因此 $m_j^{(n)} \leqslant m_j^{(n+1)}$, 且类似地 $M_j^{(n)} \geqslant M_j^{(n+1)}$. 从而, 为证明命题 (23), 只需证明:

$$M_j^{(n)} - m_j^{(n)} \to 0, n \to \infty, j = 1, \cdots, N.$$

设 $\varepsilon = \min_{i,j} p_{ij}^{(n_0)} > 0$. 则

$$\begin{aligned} p_{ij}^{(n_0+n)} &= \sum_\alpha p_{i\alpha}^{(n_0)} p_{\alpha j}^{(n)} = \sum_\alpha [p_{i\alpha}^{(n_0)} - \varepsilon p_{j\alpha}^{(n)}] p_{\alpha j}^{(n)} + \varepsilon \sum_\alpha p_{j\alpha}^{(n)} p_{\alpha j}^{(n)} \\ &= \sum_\alpha [p_{i\alpha}^{(n_0)} - \varepsilon p_{j\alpha}^{(n)}] p_{\alpha j}^{(n)} + \varepsilon p_{jj}^{(2n)}. \end{aligned}$$

而由于 $p_{i\alpha}^{(n_0)} - \varepsilon p_{j\alpha}^{(n)} \geqslant 0$, 可见

$$p_{ij}^{(n_0+n)} \geqslant m_j^{(n)} \sum_\alpha [p_{i\alpha}^{(n_0)} - \varepsilon p_{j\alpha}^{(n)}] + \varepsilon p_{jj}^{(2n)} = m_j^{(n)}(1-\varepsilon) + \varepsilon p_{jj}^{(2n)}.$$

于是

$$m_j^{(n_0+n)} \geqslant m_j^{(n)}(1-\varepsilon) + \varepsilon p_{jj}^{(2n)}.$$

类似可得

$$M_j^{(n_0+n)} \leqslant M_j^{(n)}(1-\varepsilon) + \varepsilon p_{jj}^{(2n)}.$$

由上面两个不等式, 得

$$M_j^{(n_0+n)} - m_j^{(n_0+n)} \leqslant (M_j^{(n)} - m_j^{(n)})(1-\varepsilon).$$

从而

$$M_j^{(kn_0+n)} - m_j^{(kn_0+n)} \leqslant (M_j^{(n)} - m_j^{(n)})(1-\varepsilon)^k \downarrow 0, \quad k \to \infty.$$

于是, 当 $n \to \infty$ 时, 对于某序列 $\{n_p\}$, $M_j^{(n_p)} - m_j^{(n_p)} \to 0$. 由于差 $M_j^{(n)} - m_j^{(n)}$ 关于 n 的单调性可见, 当 $n \to \infty$ 时, $M_j^{(n)} - m_j^{(n)} \to 0$.

若记 $\pi_j = \lim_n m_j^{(n)}$, 则由所得到的估计可见, 对于 $n \geqslant n_0$, 有

$$|p_{ij}^{(n)} - \pi_j| \leqslant M_j^{(n)} - m_j^{(n)} \leqslant (1-\varepsilon)^{[n/n_0]-1},$$

即 $p_{ij}^{(n)}$ 以几何速度收敛于极限值 π_j.

同样显然, $m_j^{(n)} \geqslant m_j^{(n_0)} \geqslant \varepsilon > 0, n \geqslant n_0$, 即 $\pi_j > 0$.

b) 因为状态的个数有限, 且 $\pi_j > 0$, 故由 (23) 式可以直接得到 (21) 式.

c) 由 (23) 式和 (25) 式, 得方程 (24). □

4. 马尔可夫链的大数定律 在马尔可夫链的理论中, 方程组

$$\pi_j = \sum_\alpha \pi_\alpha p_{\alpha j}, \quad j = 1, \cdots, N \tag{24*}$$

有重要作用 (对照 (24) 式). 其任何非负解 $Q = (q_1, \cdots, q_N)$ 都满足条件 $\sum_\alpha q_\alpha = 1$, 习惯上称做转移矩阵为 (p_{ij}) 的, 马尔可夫链的平稳分布或不变分布. 现在对这一名称作如下说明.

设初始分布为 $Q = (q_1, \cdots, q_N)$, 即假设 $p_j = q_j, j = 1, \cdots, N$. 那么,

$$p_j^{(1)} = \sum_\alpha q_\alpha p_{\alpha j} = q_j$$

且一般 $p_j^{(n)} = q_j$. 换句话说, 假如取 $Q = (q_1, \cdots, q_N)$ 做初始分布, 则此分布随时间的改变无变化, 即对于任意 k,

$$\mathrm{P}\{\xi_k = j\} = \mathrm{P}\{\xi_0 = j\}, \quad j = 1, \cdots, N.$$

不但如此, 具有这样初始分布 $Q = (q_1, \cdots, q_N)$ 的马尔可夫链 $\xi = (\xi_k, Q, \mathbb{P})$ 是平稳的: 对于任意 k, 随机变量 $(\xi_k, \xi_{k+1}, \cdots, \xi_{k+l})$ 的联合概率分布不依赖于 k (假设 $k + l \leqslant n$).

条件 (21) 保证了存在不依赖于 i 的极限

$$\pi_j = \lim_n p_{ij}^{(n)},$$

而且遍历分布存在, 即存在 $(\pi_1, \cdots, \pi_N), \pi_j > 0$. 这时分布 $(\pi_1, \cdots, \pi_N)$ 也是平稳分布. 现在证明分布 $(\pi_1, \cdots, \pi_N)$ 是唯一平稳分布.

事实上, 假设 $(\tilde\pi_1, \cdots, \tilde\pi_N)$ 是另一个平稳分布, 那么

$$\tilde\pi_j = \sum_\alpha \tilde\pi_\alpha p_{\alpha j} = \cdots = \sum_\alpha \tilde\pi_\alpha p_{\alpha j}^{(n)}.$$

由于 $p_{\alpha j}^{(n)} \to \pi_j$, 可见

$$\tilde\pi_j = \sum_\alpha (\tilde\pi_\alpha \pi_j) = \pi_j.$$

注意, 对于非遍历链也可能存在平稳概率分布 (并且唯一). 事实上, 若

$$\mathbb{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

则

$$\mathbb{P}^{2n}=\begin{pmatrix}1&0\\0&1\end{pmatrix},\quad \mathbb{P}^{2n+1}=\begin{pmatrix}0&1\\1&0\end{pmatrix},$$

从而极限 $\lim\limits_{n} p_{ij}^{(n)}$ 不存在. 然而, 方程组

$$q_j=\sum_{\alpha} q_\alpha p_{\alpha j},\quad j=1,2$$

变为方程组

$$\begin{cases}q_1=q_2,\\ q_2=q_1.\end{cases}$$

$(1/2,1/2)$ 是其满足条件 $q_1+q_2=1$ 的唯一解 (q_1,q_2).

还应指出, 对于前面在第 2 小节讨论的例子, 方程组 (24*) 为 (其中 $\pi_j=q_j$):

$$\begin{cases}q_0=q_0p_{00}+q_1p_{10},\\ q_1=q_0p_{01}+q_1p_{11}.\end{cases}$$

所以, 注意到条件 $q_1+q_2=1$, 由此可见, 唯一平稳分布 (q_1,q_2) 就是已求得的分布:

$$q_0=\frac{1-p_{11}}{2-p_{00}-p_{11}},\quad q_1=\frac{1-p_{00}}{2-p_{00}-p_{11}}.$$

现在, 讨论由遍历性定理得出的若干推论.

设 A 是一状态组, $A\subseteq X$, 而

$$I_A(x)=\begin{cases}1, & 若\ x\in A,\\ 0, & 若\ x\notin A.\end{cases}$$

引进随机变量

$$\nu_A(n)=\frac{I_A(\xi_0)+\cdots+I_A(\xi_n)}{n+1},$$

即质点在集合 A 中度过的时间的比率. 由于

$$\mathbf{E}[I_A(\xi_k)|\xi_0=i]=\mathbf{P}[\xi_k\in A|\xi_0=i]=\sum_{j\in A}p_{ij}^{(k)}\quad (=p_i^{(k)}(A)),$$

可见

$$\mathbf{E}[\nu_A(n)|\xi_0=i]=\frac{1}{n+1}\sum_{k=0}^{n}p_i^{(k)}(A),$$

特别

$$\mathbf{E}[\nu_{\{j\}}(n)|\xi_0=i]=\frac{1}{n+1}\sum_{k=0}^{n}p_{ij}^{(k)}.$$

由数学分析 (亦见第四章 §3 的引理 1) 知, 如果序列 $a_n\to a$, 则

$$\frac{a_0+\cdots+a_n}{n+1}\to a,\quad n\to\infty.$$

因此, 如果当 $k\to\infty$ 时 $p_{ij}^{(k)}\to\pi_j$, 则

$$\mathbf{E}\nu_{\{j\}}(n)\to\pi_j,\ \mathbf{E}\nu_A(n)\to\pi_A,\quad 其中\ \pi_A=\sum_{j\in A}\pi_j.$$

实际上, 对于遍历链可以证明得更多, 即可以证明随机变量 $I_A(\xi_0),\cdots,I_A(\xi_n),\cdots$ 服从大数定律.

大数定律. *如果 $\xi_0,\xi_1,\cdots$ 是有限遍历马尔可夫链, 则对于任意 $\varepsilon>0$, 任何 $A\subseteq X$, 以及任意初始分布,*

$$\mathbf{P}\{|\nu_A(n)-\pi_A|>\varepsilon\}\to 0,\quad n\to\infty. \tag{26}$$

在证明之前需要指出, 不能直接利用 §5 中关于伯努利随机变量序列 $I_A(\xi_0),\cdots,I_A(\xi_n),\cdots$ 的结果, 因为一般说来这些变量是不独立的. 然而, 如果仍然利用切比雪夫不等式, 以及如下事实: 对于有限状态遍历链, 存在 $0<\rho<1$ 和 $C>0$, 使

$$|p_{ij}^{(n)}-\pi_j|\leqslant C\rho^n, \tag{27}$$

则可以用与独立变量同样的方法证明 (26) 式.

考虑状态 i 和 j (有可能重合), 并证明对于 $\varepsilon>0$, 有

$$\mathbf{P}\{|\nu_{\{j\}}(n)-\pi_j|>\varepsilon|\xi_0=i\}\to 0,\quad n\to\infty. \tag{28}$$

由切比雪夫不等式, 有

$$\mathbf{P}\{|\nu_{\{j\}}(n)-\pi_j|>\varepsilon|\xi_0=i\}\leqslant\frac{\mathbf{E}\{|\nu_{\{j\}}(n)-\pi_j|^2|\xi_0=i\}}{\varepsilon^2}.$$

因此, 只需证明

$$\mathbf{E}\{|\nu_{\{j\}}(n)-\pi_j|^2|\xi_0=i\}\to 0,\quad n\to\infty.$$

通过简单的计算, 可得

$$\begin{aligned}&\mathbf{E}\{|\nu_{\{j\}}(n)-\pi_j|^2|\xi_0=i\}\\ &=\frac{1}{(n+1)^2}\mathbf{E}\left\{\left[\sum_{k=0}^{n}(I_{\{j\}}(\xi_k)-\pi_j)\right]^2\middle|\xi_0=i\right\}=\frac{1}{(n+1)^2}\sum_{k=0}^{n}\sum_{l=0}^{n}m_{ij}^{(k,l)},\end{aligned}$$

其中

$$\begin{aligned}m_{ij}^{(k,l)}&=\mathbf{E}\{[I_{\{j\}}(\xi_k)I_{\{j\}}(\xi_l)]|\xi_0=i\}-\pi_j\mathbf{E}[I_{\{j\}}(\xi_k)|\xi_0=i]\\ &\quad-\pi_j\mathbf{E}[I_{\{j\}}(\xi_l)|\xi_0=i]+\pi_j^2\\ &=p_{ij}^{(s)}p_{jj}^{(t)}-\pi_jp_{ij}^{(k)}-\pi_jp_{ij}^{(l)}+\pi_j^2,\\ &\qquad s=\min(k,l),\ t=|k-l|.\end{aligned}$$

由 (27) 式
$$p_{ij}^{(n)} = \pi_j + \varepsilon_{ij}^{(n)}, \quad |\varepsilon_{ij}^{(n)}| \leqslant C\rho^n.$$
所以
$$|m_{ij}^{(k,l)}| \leqslant C_1[\rho^k + \rho^l + \rho^k + \rho^l],$$
其中 C_1 是某一常数. 从而
$$\frac{1}{(n+1)^2}\sum_{k=0}^{n}\sum_{l=0}^{n} m_{ij}^{(k,l)} \leqslant \frac{C_1}{(n+1)^2}\sum_{k=0}^{n}\sum_{l=0}^{n}[\rho^k+\rho^l+\rho^k+\rho^l]$$
$$\leqslant \frac{4C_1}{(n+1)^2}\times\frac{2(n+1)}{1-\rho} = \frac{8C_1}{(n+1)(1-\rho)} \to 0, \quad n\to\infty.$$
由此就可以证明关系式 (28). 而由 (28) 式可以明显地得到命题 (26).

5. 游动时间的概率和期望的递推方程 对于由伯努利模型诱导的随机游动 $S_0, S_1, \cdots$, 在 §9 对于越出某一边界时间的概率和数学期望, 曾经得到递推方程. 现在对于马尔可夫链导出类似的方程.

假设 $\xi = (\xi_0,\xi_1,\cdots)$ 是转移概率矩阵为 (p_{ij})、相空间为 $X = \{0,\pm1,\cdots,\pm N\}$ 的马尔可夫链. 设 A 和 B 是两个整数, $-N \leqslant A \leqslant 0 \leqslant B \leqslant N$ 且 $x \in X$. 以 $\mathscr{B}_{k+1}$ 表示经右端点首次越出区间 (A,B) 的轨道 $(x_0,x_1,\cdots,x_k)$, $x_i\in X$ 的集合, 即越出集合 (A,B) 进入集合 $(B,B+1,\cdots,N)$ 的轨道 $(x_0,x_1,\cdots,x_k)$, $x_i \in X$ 的全体.

对于 $A\leqslant x\leqslant B$, 设
$$\beta_k(x) = \mathbf{P}\{(\xi_0,\cdots,\xi_k)\in\mathscr{B}_{k+1}|\xi_0=x\}.$$

为求这些概率 (即马尔可夫链经右端点首次越出区间 (A,B) 的概率), 可以利用推导后向方程的方法. 有
$$\begin{aligned}\beta_k(x) &= \mathbf{P}\{(\xi_0,\cdots,\xi_k)\in\mathscr{B}_{k+1}|\xi_0=x\}\\ &= \sum_y p_{xy}\mathbf{P}\{(\xi_0,\cdots,\xi_k)\in\mathscr{B}_{k+1}|\xi_0=x,\xi_1=y\},\end{aligned}$$
其中, 利用马尔可夫性和链的齐性, 不难证明
$$\begin{aligned}&\mathbf{P}\{(\xi_0,\cdots,\xi_k)\in\mathscr{B}_{k+1}|\xi_0=x,\xi_1=y\}\\ &= \mathbf{P}\{(x,y,\xi_2,\cdots,\xi_k)\in\mathscr{B}_{k+1}|\xi_0=x,\xi_1=y\}\\ &= \mathbf{P}\{(y,\xi_2,\cdots,\xi_k)\in\mathscr{B}_k|\xi_1=y\}\\ &= \mathbf{P}\{(y,\xi_1,\cdots,\xi_{k-1})\in\mathscr{B}_k|\xi_0=y\} = \beta_{k-1}(y).\end{aligned}$$
因此, 对于 $A<x<B$, $1\leqslant k\leqslant n$,
$$\beta_k(x) = \sum_y p_{xy}\beta_{k-1}(y).$$

这时, 显然
$$\beta_k(x) = 1, \quad x = B, B+1, \cdots, N,$$
和
$$\beta_k(x) = 0, \quad x = -N, \cdots, A.$$

对于经左端点首次越出区间 (A,B) 的概率 $\alpha_k(x)$, 可以类似地导出方程.

设 $\tau_k = \min\{0\leqslant l\leqslant k: \xi_l \notin (A,B)\}$, 并且当集合 $\{\cdot\}=\varnothing$ 时 $\tau_k = k$. 那么, 用对 $m_k(x) = \mathbf{E}(\tau_k|\xi_0 = x)$ 用的同样方法, 可以导出如下方程:
$$m_k(x) = 1 + \sum_y m_{k-1}(y)p_{xy}$$
(其中 $1\leqslant k\leqslant n$, $A<x<B$). 这时
$$m_k(x) = 0, \quad x\notin (A,B).$$

显然, 如果转移概率矩阵由 (11) 式给出, 则 $\alpha_k(x)$, $\beta_k(x)$ 和 $m_k(x)$ 的方程就是 §9 中相应的方程, 这些方程的推导方法实质上与这里方法一样.

最重要的是, 在游动时间无限的情况下, 将这些方程用于极限情形. 像 §9 中一样, 当 $k\to\infty$ 时, 相应的方程通过形式地极限过度, 可以由以上得到的方程导出.

作为例子, 考虑状态为 $\{0,1,\cdots,B\}$ 的马尔可夫链, 假设其转移概率为
$$p_{00} = 1, \quad p_{BB} = 1,$$
而对于 $1\leqslant i\leqslant B-1$
$$p_{ij} = \begin{cases} p_i > 0, & 若\ j = i+1,\\ r_i, & 若\ j = i,\\ q_i > 0, & 若\ j = i-1,\end{cases}$$
其中 $p_i + r_i + q_i = 1$.

该链可以用如下图形表示:

由此可见, 状态 0 和 B 是 "吸收状态"; 对于任何其他状态 i, 质点以概率 r_i 滞留, 以概率 p_i 向右移动一步, 以概率 q_i 向左移动一步.

现在求①
$$\alpha(x) = \lim_{k\to\infty}\alpha_k(x),$$

① $\alpha_k(x)$ 是轨道在时间段 $[0,k]$ 上经点 A 处越出区间 (A,B) 的概率 (见 §9). ——译者

即质点自点 x 出发早于到达状态 B 而到达状态 0 的极限概率. 当 $k\to\infty$ 时, 对于 $0<j<B$, 在 $\alpha_k(x)$ 的方程中求极限, 得

$$\alpha(j)=q_j\alpha(j-1)+r_j\alpha(j)+p_j\alpha(j+1),$$

且边界条件为

$$\alpha(0)=1,\quad \alpha(B)=0.$$

由于 $r_j=1-q_j-p_j$, 则

$$p_j[\alpha(j+1)-\alpha(j)]=q_j[\alpha(j)-\alpha(j-1)].$$

从而

$$\alpha(j+1)-\alpha(j)=\rho_j[\alpha(1)-1],$$

其中

$$\rho_j=\frac{q_1\cdots q_j}{p_1\cdots p_j},\quad \rho_0=1.$$

因为

$$\alpha(j+1)-1=\sum_{i=0}^{j}[\alpha(i+1)-\alpha(i)],$$

所以

$$\alpha(j+1)-1=[\alpha(1)-1]\sum_{i=0}^{j}\rho_i.$$

如果 $j=B-1$, 则 $\alpha(j+1)=\alpha(B)=0$, 即

$$\alpha(1)-1=-\frac{1}{\sum\limits_{i=0}^{B-1}\rho_i}.$$

于是,

$$\alpha(1)=\frac{\sum\limits_{i=1}^{B-1}\rho_i}{\sum\limits_{i=0}^{B-1}\rho_i},\quad \alpha(j)=\frac{\sum\limits_{i=j}^{B-1}\rho_i}{\sum\limits_{i=0}^{B-1}\rho_i},\quad j=1,\cdots,B.$$

(与 §9 中相应的结果比较.)

现在设

$$m(x)=\lim_{k\to\infty}m_k(x)$$

是游动的质点在到达状态 0 或 B 之一前, 所度过时间的极限值. 那么, $m(0)=m(B)=0$,

$$m(x)=1+\sum_y m(y)p_{xy},$$

从而, 对于所研究的例子, 对于一切 $j=1,\cdots,B-1$, 有

$$m(j)=1+q_jm(j-1)+r_jm(j)+p_jm(j+1).$$

为求 $m(j)$, 记

$$M(j)=m(j)-m(j-1),\quad j=1,\cdots,B.$$

那么,

$$p_jM(j+1)=q_jM(j)-1,\quad j=1,\cdots,B-1,$$

从而求得

$$M(j+1)=\rho_jM(1)-R_j,$$

其中

$$\rho_j=\frac{q_1\cdots q_j}{p_1\cdots p_j},\quad R_j=\frac{1}{p_j}\left[1+\frac{q_j}{p_{j-1}}+\cdots+\frac{q_j\cdots q_2}{p_{j-1}\cdots p_1}\right].$$

因此

$$\begin{aligned}m(j)&=m(j)-m(0)=\sum_{i=0}^{j-1}M(i+1)\\&=\sum_{i=0}^{j-1}[\rho_im(1)-R_i]=m(1)\sum_{i=0}^{j-1}\rho_i-\sum_{i=0}^{j-1}R_i.\end{aligned}$$

最后, 只剩下求 $m(1)$. 由于 $m(B)=0$, 可见

$$m(1)=\frac{\sum\limits_{i=0}^{B-1}R_i}{\sum\limits_{i=0}^{B-1}\rho_i},$$

而对于 $1<j\leqslant B$,

$$m(j)=\sum_{i=0}^{j-1}\rho_i\times\frac{\sum\limits_{i=0}^{B-1}R_i}{\sum\limits_{i=0}^{B-1}\rho_i}-\sum_{i=0}^{j-1}R_i.$$

(与 §9 相应的结果比较, 其中 $r_i=0, p_i=p, q_i=q$).

6. **强马尔可夫性** 这一小节研究马尔可夫性 (8) 的一种强化: 将普通时间 k 换成随机时间后马尔可夫性仍然成立 (见下面定理 2). 这一称为强马尔可夫性的概念的重要性, [例如在推导递推公式 (38) 的过程中] 将得到证实. 而公式 (38) 的递推公式对于马尔可夫链状态的分类至关重要 (第八章).

设 $\xi=(\xi_0,\cdots,\xi_n)$ 是齐次马尔可夫链, (p_{ij}) 是转移概率矩阵, $\mathscr{D}^\xi=(\mathscr{D}_k^\xi)_{0\leqslant k\leqslant n}$ 是分割系, 其中 $\mathscr{D}_k^\xi=\mathscr{D}_{\xi_0,\cdots,\xi_k}$. 以 $\mathscr{B}_k^\xi$ 表示由分割 $\mathscr{D}_k^\xi$ 生成的代数 $\alpha(\mathscr{D}_k^\xi)$.

我们首先给予马尔可夫性 (8) 另一种表达形式. 设 $B\in\mathscr{B}_k^\xi$, 并证明

$$\begin{aligned}&\mathbf{P}(\xi_n=a_n,\cdots,\xi_{k+1}=a_{k+1}|B\cap\{\xi_k=a_k\})\\&=\mathbf{P}(\xi_n=a_n,\cdots,\xi_{k+1}=a_{k+1}|\xi_k=a_k)\end{aligned}\tag{29}$$

(假设 $\mathbf{P}(B\cap\{\xi_k=a_k\})>0$). 实际上, 集合 B 可以表示为

$$B=\sum\nolimits^*\{\xi_0=a_0^*,\cdots,\xi_k=a_k^*\},$$

其中 $\sum^*$ 表示对某个数组 $(a_0^*,\cdots,a_k^*)$ 求和. 因此

$$\begin{aligned}&\mathbf{P}(\xi_n=a_n,\cdots,\xi_{k+1}=a_{k+1}|B\cap\{\xi_k=a_k\})\\&=\frac{\mathbf{P}(\{\xi_n=a_n,\cdots,\xi_k=a_k\}\cap B)}{\mathbf{P}(\{\xi_k=a_k\}\cap B)}\\&=\frac{\sum^*\mathbf{P}(\{\xi_n=a_n,\cdots,\xi_k=a_k\}\cap\{\xi_0=a_0^*,\cdots,\xi_k=a_k^*\})}{\mathbf{P}(\{\xi_k=a_k\}\cap B)}.\end{aligned}\tag{30}$$

而由于马尔可夫性, 可见

$$\begin{aligned}&\mathbf{P}(\{\xi_n=a_n,\cdots,\xi_k=a_k\}\cap\{\xi_0=a_0^*,\cdots,\xi_k=a_k^*\})\\&=\begin{cases}\mathbf{P}(\xi_n=a_n,\cdots,\xi_{k+1}=a_{k+1}|\xi_0=a_0^*,\cdots,\xi_k=a_k^*)\\ \qquad\times\mathbf{P}(\xi_0=a_0^*,\cdots,\xi_k=a_k^*), & 若\ a_k=a_k^*,\\ 0, & 若\ a_k\neq a_k^*,\end{cases}\\&=\begin{cases}\mathbf{P}(\xi_n=a_n,\cdots,\xi_{k+1}=a_{k+1}|\xi_k=a_k)\\ \qquad\times\mathbf{P}(\xi_0=a_0^*,\cdots,\xi_k=a_k^*), & 若\ a_k=a_k^*,\\ 0, & 若\ a_k\neq a_k^*,\end{cases}\\&=\begin{cases}\mathbf{P}(\xi_n=a_n,\cdots,\xi_{k+1}=a_{k+1}|\xi_k=a_k)\\ \qquad\times\mathbf{P}(\{\xi_k=a_k\}\cap B), & 若\ a_k=a_k^*,\\ 0, & 若\ a_k\neq a_k^*.\end{cases}\end{aligned}$$

从而 (30) 式中的和 $\sum^*$ 等于

$$\mathbf{P}(\xi_n=a_n,\cdots,\xi_{k+1}=a_{k+1}|\xi_k=a_k)\mathbf{P}(\{\xi_k=a_k\}\cap B),$$

于是, (29) 得证.

设 τ (关于分割系 $\mathscr{D}^\xi=(\mathscr{D}_k^\xi)_{0\leqslant k\leqslant n}$) 是停止时间 (见 §11 定义 2).

定义 称代数 $\mathscr{B}_n^\xi$ 中的集合 B 属于集合系 $\mathscr{B}_\tau^\xi$, 如果对于每一个 $1\leqslant k\leqslant n$,

$$B\cap\{\tau=k\}\in\mathscr{B}_k^\xi.\tag{31}$$

不难验证, 这样集合 B 的全体 $\mathscr{B}_\tau^\xi$ 构成代数 (称之为到时刻 τ 之前观测到的事件的代数).

定理 2 *设 $\xi=(\xi_0,\cdots,\xi_n)$ 是齐次马尔可夫链, (p_{ij}) 是转移概率矩阵, τ (关于分割系 $\mathscr{D}^\xi$) 是停止时间, $B\in\mathscr{B}_\tau^\xi$, 而 $A=\{\omega:\tau+l\leqslant n\}$. 那么, 如果*

$$\mathbf{P}(A\cap B\cap\{\xi_\tau=a_0\})>0,$$

则下面的强马尔可夫性成立:

$$\begin{aligned}&\mathbf{P}(\xi_{\tau+l}=a_l,\cdots,\xi_{\tau+1}=a_1|A\cap B\cap\{\xi_\tau=a_0\})\\&=\mathbf{P}(\xi_{\tau+l}=a_l,\cdots,\xi_{\tau+1}=a_1|A\cap\{\xi_\tau=a_0\}),\end{aligned}\tag{32}$$

且 (当 $\mathbf{P}(A\cap\{\xi_\tau=a_0\})>0$ 时)

$$\mathbf{P}(\xi_{\tau+l}=a_l,\cdots,\xi_{\tau+1}=a_1|A\cap\{\xi_\tau=a_0\})=p_{a_0a_1}\cdots p_{a_{l-1}a_l}.\tag{33}$$

证明 为简便计, 我们只证明 $l=1$ 的情形. 由于 $B\cap\{\tau=k\}\in\mathscr{B}_k^\xi$, 可见根据 (29) 式, 有

$$\begin{aligned}&\mathbf{P}(\xi_{\tau+1}=a_1,A\cap B\cap\{\xi_\tau=a_0\})\\&=\sum_{k\leqslant n-1}\mathbf{P}\{\xi_{k+1}=a_1,\xi_k=a_0,\tau=k,B\}\\&=\sum_{k\leqslant n-1}\mathbf{P}\{\xi_{k+1}=a_1|\xi_k=a_0,\tau=k,B\}\mathbf{P}\{\xi_k=a_0,\tau=k,B\}\\&=\sum_{k\leqslant n-1}\mathbf{P}\{\xi_{k+1}=a_1|\xi_k=a_0\}\mathbf{P}\{\xi_k=a_0,\tau=k,B\}\\&=p_{a_0a_1}\sum_{k\leqslant n-1}\mathbf{P}\{\xi_k=a_0,\tau=k,B\}=p_{a_0a_1}\mathbf{P}(A\cap B\cap\{\xi_\tau=a_0\}),\end{aligned}$$

于是, (32) 和 (33) 式得证 (对于 (33) 式的情形, 应设 $B=\Omega$). □

注 1 显然, 对于 $l=1$ 的情形, 强马尔可夫性 (32) 和 (33) 式等价于: 对于任意 $C\subseteq X$,

$$\mathbf{P}(\xi_{\tau+1}\in C|A\cap B\cap\{\xi_\tau=a_0\})=\mathbf{P}_{a_0}(C),\tag{34}$$

其中

$$\mathbf{P}_{a_0}(C)=\sum_{a_1\in C}p_{a_0a_1}.$$

而 (34) 式可以表述为: 在集合 $A=\{\tau\leqslant n-1\}$ 上,

$$\mathbf{P}\{\xi_{\tau+1}\in C|\mathscr{F}_\tau^\xi\}=\mathbf{P}_{\xi_\tau}(C). \tag{35}$$

这在一般齐次马尔可夫过程论中, 是强马尔可夫性常用的形式之一.

注 2　即使没有事件 $A\cap B\cap\{\xi_\tau=a_0\}$ 和 $A\cap\{\xi_\tau=a_0\}$ 的概率大于 0 的要求, 性质 (32) 和 (33) (只要利用注 1 所描绘的约定) 仍然成立.

7. n 步的转移概率 $p_{ij}^{(n)}$　设 $\xi=(\xi_0,\cdots,\xi_n)$ 是齐次马尔可夫链, (p_{ij}) 是转移概率矩阵,

$$f_{ii}^{(k)}=\mathbf{P}(\xi_k=i;\xi_l\neq i,1\leqslant l\leqslant k-1|\xi_0=i), \tag{36}$$

而对于 $i\neq j$,

$$f_{ij}^{(k)}=\mathbf{P}(\xi_k=j;\xi_l\neq j,1\leqslant l\leqslant k-1|\xi_0=i). \tag{37}$$

分别是在时刻 k 首达状态 i 的概率和在时刻 k 首达状态 j 的概率.

现在证明

$$p_{ij}^{(n)}=\sum_{k=1}^n f_{ij}^{(k)}p_{jj}^{(n-k)},\ 其中\ p_{jj}^{(0)}=1 \tag{38}$$

此式的直观意义很明显: 为经 n 步自状态 i 到达状态 j, 需要经 $k(1\leqslant k\leqslant n)$ 步首达状态 j, 然后经 $n-k$ 步由 j 返回 j. 现在给出 (38) 式的严格证明.

设 j 固定且

$$\tau_j=\min\{1\leqslant k\leqslant n:\xi_k=j\},$$

若 $\{\cdot\}=\varnothing$, 则设 $\tau_j=n+1$. 那么 $f_{ij}^{(k)}=\mathbf{P}(\tau_j=k|\xi_0=i)$, 而

$$\begin{aligned}p_{ij}^{(n)}&=\mathbf{P}(\xi_n=j|\xi_0=i)=\sum_{k=1}^n\mathbf{P}(\xi_n=j,\tau_j=k|\xi_0=i)\\&=\sum_{k=1}^n\mathbf{P}(\xi_{\tau_j+n-k}=j,\tau_j=k|\xi_0=i),\end{aligned} \tag{39}$$

其中最后一个等式成立, 因为在集合 $\{\tau_j=k\}$ 上 $\xi_{\tau_j+n-k}=\xi_n$. 而且, 对于任意 $1\leqslant k\leqslant n$, 集合 $\{\tau_j=k\}=\{\tau_j=k,\xi_{\tau_j}=j\}$. 因此, 如果 $\mathbf{P}\{\xi_0=i,\tau_j=k\}>0$, 则由定理 2, 有

$$\begin{aligned}\mathbf{P}(\xi_{\tau_j+n-k}=j|\xi_0=i,\tau_j=k)&=\mathbf{P}(\xi_{\tau_j+n-k}=j|\xi_0=i,\tau_j=k,\xi_{\tau_j}=j)\\&=\mathbf{P}(\xi_{\tau_j+n-k}=j|\xi_{\tau_j}=j)=p_{jj}^{(n-k)},\end{aligned}$$

而根据 (37) 式, 有

$$p_{ij}^{(n)}=\sum_{k=1}^n\mathbf{P}(\xi_{\tau_j+n-k}=j|\xi_0=i,\tau_j=k)\mathbf{P}(\tau_j=k|\xi_0=i)=\sum_{k=1}^n p_{jj}^{(n-k)}f_{ij}^{(k)},$$

于是, 关系 (38) 式得证.

8. 练习题

1. 设 $\xi=(\xi_0,\cdots,\xi_n)$ 是马尔可夫链, X 是其值域, $f=f(x),x\in X$ 是某一函数. 问序列 $(f(\xi_0),\cdots,f(\xi_n))$ 是不是马尔可夫链? 序列 $(\xi_n,\xi_{n-1},\cdots,\xi_0)$ 是不是 "后向" 马尔可夫链?

2. 设 $\mathbb{P}=\|p_{ij}\|,1\leqslant i,j\leqslant r$ 是随机矩阵, 而 λ 是其特征值, 即特征方程 $\det(\mathbb{P}-\lambda E)=0$ 的根. 证明 $\lambda_1=1$ 是特征值, 而一切其他特征值 $\lambda_2,\cdots,\lambda_r$ 的绝对值都不大于 1. 如果所有特征值 $\lambda_1,\cdots,\lambda_r$ 两两不同, 则 $p_{ij}^{(k)}$ 有如下表示式:

$$p_{ij}^{(k)}=\pi_j+a_{ij}(2)\lambda_2^k+\cdots+a_{ij}(r)\lambda_r^k,$$

其中 $\pi_j,a_{ij}(2),\cdots,a_{ij}(r)$ 可以通过矩阵 $\mathbb{P}$ 的元素表示. (特别, 用这一代数方法分析马尔可夫链的性质, 可见当 $|\lambda_2|<1,\cdots,|\lambda_r|<1$ 时, 对于每一个 j, 存在不依赖于 i 的极限 $\lim_k p_{ij}^{(k)}$.)

3. 设 $\xi=(\xi_0,\cdots,\xi_n)$ 是齐次马尔可夫链, X 是其状态集, (p_{xy}) 是转移概率矩阵, 记

$$T\varphi(x)=\mathbf{E}[\varphi(\xi_1)|\xi_0=x]\left(=\sum_y\varphi(y)p_{xy}\right),$$

假设非负函数 $\varphi=\varphi(x)$ 满足方程

$$T\varphi(x)=\varphi(x),\quad x\in X.$$

证明随机变量序列

$$\zeta=(\zeta_k,\mathscr{F}_k^\xi),\quad 其中\ \zeta_k=\varphi(\xi_k)$$

是鞅.

4. 设 $\xi=(\xi_n,\Pi,\mathbb{P})$ 和 $\tilde\xi=(\tilde\xi_n,\tilde\Pi,\mathbb{P})$ 是两个马尔可夫链, 具有不同的初始分布, 相应为 $\Pi=(p_1,\cdots,p_r)$ 和 $\tilde\Pi=(\tilde p_1,\cdots,\tilde p_r)$. 设 $\Pi^{(n)}=(p_1^{(n)},\cdots,p_r^{(n)})$ 和 $\tilde\Pi^{(n)}=(\tilde p_1^{(n)},\cdots,\tilde p_r^{(n)})$, 证明

$$\sum_{i=1}^r|\tilde p_i^{(n)}-p_i^{(n)}|\leqslant 2(1-r\varepsilon)^n.$$

5. 设 P 和 Q 为随机矩阵. 证明 PQ 和 $\alpha P+(1-\alpha)Q,0\leqslant\alpha\leqslant 1$ 也是随机矩阵.

6. 考虑齐次马尔可夫链 $(\xi_0,\cdots,\xi_n)$, $X=\{0,1\}$ 是其状态集, 其转移概率矩阵为

$$\begin{pmatrix}1-p & p\\ q & 1-q\end{pmatrix},$$

其中 $0<p<1,0<q<1$. 设 $S_n=\xi_0+\cdots+\xi_n$. 作为 (§6) 棣莫弗—拉普拉斯定理

的推广, 证明

$$\mathbf{P}\left\{\frac{S_n-\dfrac{p}{p-q}n}{\sqrt{\dfrac{npq(2-p-q)}{(p+q)^3}}}\leqslant x\right\}\to\Phi(x),\qquad n\to\infty.$$

说明, 当 $p+q=1$ 时, 随机变量 $\xi_1,\cdots,\xi_n$ 独立, 而上述命题归结为:

$$\mathbf{P}\left\{\frac{S_n-pn}{\sqrt{npq}}\leqslant x\right\}\to\Phi(x),\qquad n\to\infty.$$

第二章　概率论的数学基础

像几何学和代数学一样, 概率论作为数学学科, 可以并且应该完全公理化. 这意味着, 在给出研究对象及其关系之后, 还应给出这些关系应服从的公理. 在此之后全部的叙述应仅仅以这些公理为基础, 而不依赖于这些对象及其关系的一般的具体值.

A. H. 柯尔莫戈洛夫,《概率论的基本概念》[32]

§1. 有无限种结局试验的概率模型. 柯尔莫戈洛夫公理化体系

1. 代数和有限 · 可加测度 上一章介绍的模型, 提供了具有有限种结局的试验的概率 · 统计描述. 例如, "三对象"

$$(\Omega, \mathscr{A}, \mathbf{P}), \text{ 其中 } \Omega = \{\omega : \omega = (a_1, \cdots, a_n), a_i = 0, 1\},$$
$$\mathscr{A} = \{A : A \subseteq \Omega\}, \mathbf{P}(\{\omega\}) = p^{\sum a_i}(1-p)^{n-\sum a_i} (= p(\omega)),$$

是 n 次 "独立" 地掷硬币模型, 其中 "正面" 出现的概率为 p. 在此模型中, 一切结局的个数 $N(\Omega)$, 即集合中点的个数是有限的, 并且等于 2^n.

现在的问题是, 建立无限次 "独立地" 掷硬币模型, 其中每次掷 "正面" 出现的概率为 p.

作为结局的全体, 自然取集合

$$\Omega = \{\omega : \omega = (a_1, a_2, \cdots), a_i = 0, 1\},$$

即一切序列 $\omega = (a_1, a_2, \cdots)$ 的空间, 集合 ω 的元素有 0 和 1 两个可能值.

问集合 Ω 的势 (基数) $N(\Omega)$ 如何? 熟知, 任意实数 $a \in [0, 1)$ 可以唯一地分解为 (可能含无限个 0 的) 二进制小数:

$$a = \frac{a_1}{2} + \frac{a_2}{2^2} + \cdots (a_i = 0, 1).$$

由此可见, 在 ω 的集合 Ω 与 a 的集合 $[0, 1)$ 之间存在一一对应关系, 说明集合 Ω 的势等于连续统的势.

于是, 如果希望建立描绘 "无限次 '独立地' 掷硬币类型" 的试验的概率模型, 则不得不考虑相当复杂来源的空间 Ω.

我们现在尝试弄清楚, 如何在 "无限次 '独立地' 掷 (对称 [$p = q = 1/2$] 的) 硬币" 的模型中, 合理地确定 (赋予) 概率.

由于作为 Ω 可以取集合 $[0, 1)$, 所以把上面提出的问题, 可以视为 "在集合 $[0, 1)$ 上随机选点" 的模型中求概率的问题. 由于直观的对称性, 可以认为一切结局都是

"等可能的". 由于集合 $[0,1)$ 是不可数的, 假如考虑到点属于集合 $[0,1)$ 的概率应该等于 1, 则每个 ω 的概率必为 0. 但是如此定义概率 $(p(\omega)=0, \omega\in[0,1))$ 没有什么意义. 然而, 我们通常关心的并不是个别结局出现的概率, 而是试验结局属于某个给定集合 (事件) A 的概率. 在初等概率论中, 根据 "权重" $p(\omega)$ 即可求事件 A 的概率.

$$\mathbf{P}(A)=\sum_{\omega\in A}p(\omega).$$

对于现在考虑的情况. 例如, 当 $p(\omega)=0, \omega\in[0,1)$ 时, 我们无法求 "自 $[0,1)$ 上随机选的点" 属于集合 $[0,1/2)$ 的概率. 然而, 直观上很明显这一概率应该等于 $1/2$.

这种情况提示我们, 当空间 Ω 不可数时, 不应求个别结局的概率, 而应求 Ω 中某些集合的概率. 第一章的论证说明, 在其上确定概率的集合的全体, 关于并、交与补应该是封闭的. 鉴于这种情况, 我们引进如下定义.

定义 1 设 Ω 是点 ω 的某一集合, Ω 的子集系 $\mathscr{A}$ 称做代数, 如果

a) $\Omega\in\mathscr{A}$,

b) $A,B\in\mathscr{A}\Rightarrow A\bigcup B\in\mathscr{A}, A\bigcap B\in\mathscr{A}$,

c) $A\in\mathscr{A}\Rightarrow\overline{A}\in\mathscr{A}$.

注意, 在条件 b) 中, 只需要求 $A\bigcup B\in\mathscr{A}$ 或 $A\bigcap B\in\mathscr{A}$ 之一成立, 因为

$$A\bigcup B=\overline{\overline{A}\bigcap\overline{B}},\quad A\bigcap B=\overline{\overline{A}\bigcup\overline{B}}$$

为表述概率模型的概念, 需要下面的概念.

定义 2 设 $\mathscr{A}$ 是 Ω 的子集的代数. 在 $[0,\infty]$ 上取值的集函数 $\mu=\mu(A), A\in\mathscr{A}$, 称做定义在 $\mathscr{A}$ 上的有限 − 可加测度, 如果 $\mathscr{A}$ 中的任意两个不相交集合 A 和 B, 有

$$\mu(A+B)=\mu(A)+\mu(B).\tag{1}$$

具有 $\mu(\Omega)<\infty$ 的有限 − 可加测度 $\mu=\mu(A)$ 称做有限的, 而当 $\mu(\Omega)=1$ 时, 称做有限 − 可加概率测度或有限 − 可加概率.

2. 概率模型 现在给出试验的概率模型, 假设其 (广义) 结局 (现象) 属于集合 Ω.

定义 3 三元对象

$$(\Omega,\mathscr{A},\mathbf{P}),$$

的总体, 其中

a) Ω 是点 ω 的集合,

b) $\mathscr{A}$ 是 Ω 的子集的代数,

c) $\mathbf{P}$ 是 $\mathscr{A}$ 上的有限 − 可加概率,

称做概率模型、广义 (试验的) 概率 "理论".

不过, 事实表明, 为建立有成效的数学理论, 这一概率模型显得过于广泛. 因此, 不但对所考虑的 Ω 的子集类, 而且对容许的概率测度类, 不得不加以限制.

定义 4 集合 Ω 的子集系 $\mathscr{F}$ 称做 σ − 代数, 如果 $\mathscr{F}$ 是代数, 而且满足下列性质 (定义 1 之性质 b) 的加强):

b*) 如果 $A_n\in\mathscr{F}, n=1,2,\cdots$, 则

$$\bigcup A_n\in\mathscr{F},\ \bigcap A_n\in\mathscr{F}$$

(这时, 只需要求二式之一成立).

定义 5 空间 Ω 连同其子集的 σ − 代数 $\mathscr{F}$, 称做可测空间, 记作 $(\Omega,\mathscr{F})$.

定义 6 定义在集合 Ω 的子集的代数 $\mathscr{A}$ 上的有限 − 可加测度 μ, 称做完全可加或 σ − 可加测度, 亦简称测度, 如果对于两两不相交的集合 $A_1, A_2, \cdots\in\mathscr{A}$, 且

$$\sum_{n=1}^{\infty}A_n\in\mathscr{A},$$

有

$$\mu\left(\sum_{n=1}^{\infty}A_n\right)=\sum_{n=1}^{\infty}\mu(A_n).$$

测度 μ 称做 σ − 有限的, 如果空间 Ω 可以表示为

$$\Omega=\sum_{n=1}^{\infty}\Omega_n,\ \Omega_n\in\mathscr{A},$$

且 $\mu(\Omega_n)<\infty, n=1,2,\cdots$.

代数 $\mathscr{A}$ 上的测度 (注意, 指 σ − 可加测度) $\mathbf{P}$, 如果满足条件 $\mathbf{P}(\Omega)=1$, 则 $\mathbf{P}$ 称做 (定义在代数 $\mathscr{A}$ 的集合上的) 概率测度或简称概率.

现在指出概率测度的某些性质.

1° 若 $\varnothing$ 是空集, 则

$$\mathbf{P}(\varnothing)=0.$$

2° 若 $A,B\in\mathscr{A}$, 则

$$\mathbf{P}(A\cup B)=\mathbf{P}(A)+\mathbf{P}(B)-\mathbf{P}(A\cap B).$$

3° 若 $A,B\in\mathscr{A}$ 且 $B\subseteq A$, 则

$$\mathbf{P}(B)\leqslant\mathbf{P}(A).$$

4° 若 $A_n\in\mathscr{A}, n=1,2,\cdots$, 且 $\bigcup A_n\in\mathscr{A}$, 则

$$\mathbf{P}(A_1\cup A_2\cup\cdots)\leqslant\mathbf{P}(A_1)+\mathbf{P}(A_2)+\cdots.$$

前三条性质显然. 为证明性质 4, 只需注意到, 对于 $n \geqslant 2$, 有

$$\bigcup_{n=1}^{\infty} A_n = \sum_{n=1}^{\infty} B_n, B_1 = A_1, B_n = \left(\bigcap_{k=1}^{n-1} A_k\right) \cap A_n, B_i \cap B_j = \varnothing,$$

因而

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbf{P}\left(\sum_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mathbf{P}(B_n) \leqslant \sum_{n=1}^{\infty} \mathbf{P}(A_n).$$

下面的定理证明, 集合的有限 - 可加函数, 同时也是 σ - 可加的, 该定理有许多应用.

定理 设 $\mathbf{P}$ 是定义在代数 $\mathscr{A}$ 上的 σ- 可加集合函数, 且 $\mathbf{P}(\Omega) = 1$. 那么, 如下 4 个条件等价.

1) $\mathbf{P}$ 为 σ- 可加 ($\mathbf{P}$ 是概率);

2) $\mathbf{P}$ 上连续, 即对于任意集合 $A_1, A_2, \cdots \in \mathscr{A}$, 若

$$A_n \subseteq A_{n+1}, \bigcup_{n=1}^{\infty} A_n \in \mathscr{A},$$

则

$$\lim_n \mathbf{P}(A_n) = \mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right);$$

3) $\mathbf{P}$ 下连续, 即对于任意集合 $A_1, A_2, \cdots \in \mathscr{A}$, 若

$$A_n \supseteq A_{n+1}, \bigcap_{n=1}^{\infty} A_n \in \mathscr{A},$$

则

$$\lim_n \mathbf{P}(A_n) = \mathbf{P}\left(\bigcap_{n=1}^{\infty} A_n\right);$$

4) $\mathbf{P}$ 在 "$\varnothing$" 连续, 即对于任意集合 $A_1, A_2, \cdots \in \mathscr{A}$, 若

$$A_{n+1} \subseteq A_n, \bigcap_{n=1}^{\infty} A_n = \varnothing,$$

则

$$\lim_n \mathbf{P}(A_n) = 0.$$

证明 1)$\Rightarrow$2). 由于

$$\bigcup_{n=1}^{\infty} A_n = A_1 + (A_2 \backslash A_1) + (A_3 \backslash A_2) + \cdots,$$

则

$$\begin{aligned}\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) &= \mathbf{P}(A_1) + \mathbf{P}(A_2 \backslash A_1) + \mathbf{P}(A_3 \backslash A_2) + \cdots \\ &= \mathbf{P}(A_1) + \mathbf{P}(A_2) - \mathbf{P}(A_1) + \mathbf{P}(A_3) - \mathbf{P}(A_2) + \cdots \\ &= \lim_n \mathbf{P}(A_n).\end{aligned}$$

2)$\Rightarrow$3). 设 $n \geqslant 1$. 则

$$\mathbf{P}(A_n) = \mathbf{P}(A_1 \backslash (A_1 \backslash A_n)) = \mathbf{P}(A_1) - \mathbf{P}(A_1 \backslash A_n).$$

集合序列 $\{A_1 \backslash A_n\}_{n \geqslant 1}$ 是非降的 (见 §1 表 1), 而

$$\bigcup_{n=1}^{\infty} (A_1 \backslash A_n) = A_1 \backslash \bigcap_{n=1}^{\infty} A_n.$$

那么, 由 2) 可见

$$\lim_n \mathbf{P}(A_1 \backslash A_n) = \mathbf{P}\left(\bigcup_{n=1}^{\infty} (A_1 \backslash A_n)\right).$$

因此

$$\begin{aligned}\lim_n \mathbf{P}(A_n) &= \mathbf{P}(A_1) - \lim_n \mathbf{P}(A_1 \backslash A_n) \\ &= \mathbf{P}(A_1) - \mathbf{P}\left(\bigcup_{n=1}^{\infty} (A_1 \backslash A_n)\right) = \mathbf{P}(A_1) - \mathbf{P}\left(A_1 \backslash \bigcap_{n=1}^{\infty} A_n\right) \\ &= \mathbf{P}(A_1) - \mathbf{P}(A_1) + \mathbf{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \mathbf{P}\left(\bigcap_{n=1}^{\infty} A_n\right).\end{aligned}$$

3)$\Rightarrow$4). 显然.

4)$\Rightarrow$1). 设两两不相交的集合 $A_1, A_2, \cdots \in \mathscr{A}$, 且 $\sum_{n=1}^{\infty} A_n \in \mathscr{A}$, 则

$$\mathbf{P}\left(\sum_{i=1}^{\infty} A_i\right) = \mathbf{P}\left(\sum_{i=1}^{n} A_i\right) + \mathbf{P}\left(\sum_{i=n+1}^{\infty} A_i\right),$$

而由于

$$\sum_{i=n+1}^{\infty} A_i \downarrow \varnothing, n \to \infty,$$

因此,

$$\begin{aligned}\sum_{i=1}^{\infty} \mathbf{P}(A_i) &= \lim_n \sum_{i=1}^{n} \mathbf{P}(A_i) = \lim_n \mathbf{P}\left(\sum_{i=1}^{n} A_i\right) \\ &= \lim_n \left[\mathbf{P}\left(\sum_{i=1}^{\infty} A_i\right) - \mathbf{P}\left(\sum_{i=n+1}^{\infty} A_i\right)\right] \\ &= \mathbf{P}\left(\sum_{i=1}^{\infty} A_i\right) - \lim_n \mathbf{P}\left(\sum_{i=n+1}^{\infty} A_i\right) = \mathbf{P}\left(\sum_{i=1}^{\infty} A_i\right).\end{aligned}$$

□

3. 柯尔莫戈洛夫公理化体系 设 Ω 是试验 E 的结局 (现象) 的集合. 现在可以表述柯尔莫戈洛夫公理化体系, 该公理化体系已经被普遍地接受, 并且是建立试验 E 的概率模型的基础.

基本定义 考虑三个对象的全体

$$(\Omega, \mathscr{F}, \mathbf{P}),$$

其中

a) Ω 是点 ω 的集合;

b) $\mathscr{F}$ 是 Ω 子集的 σ - 代数;

c) $\mathbf{P}$ 是 $\mathscr{F}$ 上的概率.

称 $(\Omega, \mathscr{F}, \mathbf{P})$ 为 (试验的) 概率模型或概率空间. 这时, 结局的空间 Ω 称为基本事件空间, $\mathscr{F}$ 中的集合 A 称做事件, 而 $\mathbf{P}(A)$ 称做事件 A 的概率.

由这一定义可见, 概率的公理本质上依赖于集合论与测度论的工具. 因此, 下面的表 2－1 很重要, 该表将集合论与概率论相应的概念进行对照和比较. 在下面的两节将陆续给出概率论中最重要的可测空间的例子, 以及其中确定概率方法的例子.

表 2－1

记号	集合论	概率论
ω	元素, 点	结局, 基本事件
Ω	点集	结局的空间, 基本事件空间, 必然事件
$\mathscr{F}$	子集的 σ - 代数	事件的 σ - 代数
$A \in \mathscr{F}$	点集	事件 (若 $\omega \in A$, 则称事件 A 出现了)
$\overline{A} = \Omega \setminus A$	集合 A 的补集, 不属于 A 的点 ω 的集合	事件 "事件 A 不出现"
$A \cup B$	集合 A 与 B 的并, 属于 A 或 B 的点 ω 的集合	事件 "事件 A 或 B 至少一个出现"
$A \cap B$ 或 AB	集合 A 与 B 的交, 同属于 A 和 B 的点 ω 的集合	事件 "事件 A 與 B 同时出现"
$\varnothing$	空集	不可能事件
$A \cap B = \varnothing$	集合 A 与 B 不相交	事件 A 与 B 不相容 (不能同时出现)
$A + B$	集合 A 与 B 之和, 即不相交集合 A 与 B 的并	事件 "出现了二不相容事件之一"
$A \setminus B$	集合 A 与 B 之差, 属于 A 不属于 B 的点 ω 的集合	事件 "A 出现但 B 不出现"
$A \triangle B$	集合的对称差: 集合 $(A \setminus B) \cup (B \setminus A)$	事件 "A 或 B 仅出现一个"
$\bigcup_{n=1}^{\infty} A_n$	集合 $A_1, A_2, \cdots$ 的并: 至少属于 $A_1, A_2, \cdots$ 之一的点 ω 的集合	事件 "$A_1, A_2, \cdots$ 的并": 事件 "$A_1, A_2, \cdots$ 至少出现一个"
$\sum_{n=1}^{\infty} A_n$	集合 $A_1, A_2, \cdots$ 的和: 两两不交集合 $A_1, A_2, \cdots$ 的并	事件 "两两不相容事件 $A_1, A_2, \cdots$ 中至少出现一个"
$\bigcap_{n=1}^{\infty} A_n$	集合 $A_1, A_2, \cdots$ 的交, 同属于 $A_1, A_2, \cdots$ 的点 ω 的集合	事件 "$A_1, A_2, \cdots$ 的交": 事件 "$A_1, A_2, \cdots$ 同时出现"
$A_n \uparrow A$ 或 $A = \lim_n \uparrow A_n$	递增集合序列 (A_n) 收敛于集合 A: $A_1 \subseteq A_2 \subseteq \cdots, A = \bigcup_{n=1}^{\infty} A_n$	递增事件序列 $A_1, A_2, \cdots$ 收敛于事件 A

续表

记号	集合论	概率论
$A_n \downarrow A$ 或 $A = \lim_n \downarrow A_n$	递减集合序列 (A_n) 收敛于集合 A: $A_1 \supseteq A_2 \supseteq \cdots, A = \bigcap_{n=1}^{\infty} A_n$	递减事件序列 $A_1, A_2, \cdots$ 收敛于事件 A
$\overline{\lim}_n A_n$ 或 $\limsup A_n$ 或 {无限多个 A_n}	集合 $A = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$	在事件序列 $A_1, A_2, \cdots$ 中出现无限多次的 ω 的集合
$\underline{\lim}_n A_n$ 或 $\liminf A_n$	集合 $A = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$	事件 "在 $A_1, A_2, \cdots$ 中出现无限多个事件, 其中仅有限个事件可能除外"

注 在建立用于描述"现象与条件"的概率模型的试验模型时, 应 (根据第一章 §? (? 的注释)) 事先说明, 是"在何条件组合下"考虑有关试验的. 通常我们并不作这种说明, 不过都读者应当清楚, 这些"条件"是什么.

4. 练习题

1. 设 $\Omega = \{r : r \in [0,1]\}$ 是区间 $[0,1]$ 上有理点的集合, $\mathscr{A}$ 是集合代数: 其中每一个都是有限个形如 $\{r : a < r < b\}, \{r : a \leqslant r < b\}, \{r : a < r \leqslant b\}, \{r : a \leqslant r \leqslant b\}$ 的不相交集合 A 之并, 而 $\mathbf{P}(A) = b - a$. 证明, $\mathbf{P}(A), A \in \mathscr{A}$, 是有限 - 可加的集函数, 但不是 σ - 可加的集函数.

2. 设 Ω 是可数集合, 而 $\mathscr{F}$ 是一切子集的全体. 设 $\mu(A) = 0$, 如果 A 有限, 而 $\mu(A) = \infty$, 如果 A 无限. 证明集函数 μ 有限 - 可加, 但非 σ - 可加.

3. 设 μ 是 σ - 代数 $\mathscr{F}$ 上的有限测度, $A_n \in \mathscr{F}, n = 1, 2, \cdots$, 而

$$A = \lim_n A_n \left(\text{即 } A = \overline{\lim_n} A_n = \underline{\lim_n} A_n\right).$$

证明, $\mu(A) = \lim_n \mu(A_n)$.

4. 证明, $\mathbf{P}(A \triangle B) = \mathbf{P}(A) + \mathbf{P}(B) - 2\mathbf{P}(A \cap B)$. (对照第一章 §1 练习题 4.)

5. 设

$$\rho_1(A, B) = \mathbf{P}(A \triangle B),$$

$$\rho_2(A, B) = \begin{cases} \dfrac{\mathbf{P}(A \triangle B)}{\mathbf{P}(A \cup B)}, & \text{若 } \mathbf{P}(A \cup B) \neq 0, \\ 0, & \text{若 } \mathbf{P}(A \cup B) = 0. \end{cases}$$

证明, $\rho_1(A, B)$ 和 $\rho_2(A, B)$ 可以作为"距离", 即满足"三角形不等式".

6. 设 μ 是代数 $\mathscr{A}$ 上的有限 - 可加测度, $A_1, A_2, \cdots \in \mathscr{A}$ 是两两不相交集合,

$$A = \sum_{i=1}^{\infty} A_i \in \mathscr{A}.$$

证明

$$\mu(A) \geqslant \sum_{i=1}^{\infty} \mu(A_i).$$

7. 证明

$$\limsup A_n = \liminf A_n;\quad \liminf A_n = \limsup \overline{A}_n;$$

$$\liminf A_n \subseteq \limsup A_n;\quad \limsup(A_n \cup B_n) = \limsup A_n \cup \limsup B_n;$$

$$\liminf(A_n \cap B_n) = \liminf A_n \cap \liminf B_n;$$

$$\limsup A_n \cap \liminf B_n \subseteq \limsup(A_n \cap B_n) \subseteq \limsup A_n \cap \limsup B_n.$$

如果 $A_n \uparrow A$ 或 $A_n \downarrow A$, 则

$$\liminf A_n = \limsup A_n.$$

8. 设 $\{x_n\}$ 是一数列, 而 $A_n = (-\infty, x_n)$. 证明: $x = \limsup x_n$ 与 $A = \limsup A_n$ 有如下联系: $(-\infty, x) \subseteq A \subseteq (-\infty, x]$. 换句话说, A 等于 $(-\infty, x)$ 或 $(-\infty, x]$.

9. 举例说明, 对于取 $+\infty$ 为值的测度, 由完全可加性一般推不出在 "零" $\varnothing$ 的连续性.

10. 证明布尔 (G. Boole) 不等式: $\mathrm{P}(A \bigcap B) \geqslant 1 - \mathrm{P}(\overline{A}) - \mathrm{P}(\overline{B})$.

11. 设 $A_1, \cdots, A_n$ 是 $\mathscr{F}$ 中的事件. 称这一事件组为可重置的或可交换的 (exchangeable 或 interchangeable), 如果对于一切 $1 \leqslant k \leqslant n$ 和任意下标 $1 \leqslant i_1 < \cdots < i_k \leqslant n$, 概率 $\mathrm{P}(A_{i_1}, \cdots, A_{i_k})$ 等于同一个值 $(=p_k)$. 对这样的事件, 证明下面的公式:

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) = np_1 - C_n^2 p_2 + C_n^3 p_3 - \cdots + (-1)^{n-1} p_n.$$

12. 设 $(A_k)_{k\geqslant 1}$ 是无限可重置事件序列, 即对于一切 $n \geqslant 1$ 和任意下标 $1 \leqslant i_1 < \cdots < i_n$, 概率 $\mathrm{P}(A_{i_1}, \cdots, A_{i_n})$ 等于同一个值 $(= p_n)$. 证明

$$\mathrm{P}\left(\varliminf_n A_n\right) = \mathrm{P}\left(\bigcap_{k=1}^{\infty} A_k\right) = \lim_{j\to\infty} p_j,$$

$$\mathrm{P}\left(\varlimsup_n A_n\right) = \mathrm{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = 1 - \lim_{j\to\infty} (-1)^j \Delta^j(p_0),$$

其中 $p_0 = 1$, $\Delta^1(p_n) = p_{n+1} - p_n$, $\Delta^j(p_n) = \Delta^1(\Delta^{j-1}(p_n))$, $j \geqslant 2$.

13. 设 $(A_n)_{n\geqslant 1}$ 是一集合序列, $I(A_n)$ 是集合 A_n, $n \geqslant 1$ 的示性函数. 证明

$$I\left(\varliminf_n A_n\right) = \varliminf_n I(A_n),\ I\left(\varlimsup_n A_n\right) = \varlimsup_n I(A_n),$$

$$I\left(\bigcup_{n=1}^{\infty} A_n\right) \leqslant \sum_{n=1}^{\infty} I(A_n).$$

14. 证明

$$I\left(\bigcup_{n=1}^{\infty} A_n\right) = \max_{n\geqslant 1} I(A_n),\ I\left(\bigcap_{n=1}^{\infty} A_n\right) = \min_{n\geqslant 1} I(A_n).$$

15. 证明

$$\mathrm{P}\left(\varlimsup_n A_n\right) \geqslant \varlimsup_n \mathrm{P}(A_n),\ \mathrm{P}\left(\varliminf_n A_n\right) \leqslant \varliminf_n \mathrm{P}(A_n).$$

16. 设 $A^* = \varlimsup_n A_n$ 和 $A_* = \varliminf_n A_n$. 证明 $\mathrm{P}(A_n - A_*) \to 0$, $\mathrm{P}(A^* - A_n) \to 0$.

17. 在上一题的记号下, 记 $A = A^* = A_*$. 证明, 若 $A_n \to A$, 则 $\mathrm{P}(A \triangle A_n) \to 0$.

18. 设集合 A_n 在 $\mathrm{P}(A \triangle A^*) = \mathrm{P}(A \triangle A_*) = 0$ 意义下收敛于集合 A. 证明 $\mathrm{P}(A \triangle A_n) \to 0$.

19. 证明, 集合 A 与 B 的对称差 $A \triangle B$, 具有下面的性质.

$$I(A \triangle B) = I(A) + I(B) (\mathrm{mod}\,2).$$

(由此可以导出 $\mathrm{P}(A \triangle B) = \mathrm{P}(A) + \mathrm{P}(B) - 2\mathrm{P}(A \bigcap B)$; 见练习题 4.) 证明, 对称差的下列性质:

$$(A \triangle B) \triangle C = A \triangle (B \triangle C),\ (A \triangle B) \triangle (B \triangle C) = A \triangle C,$$

$$A \triangle B = C \Leftrightarrow A = B \triangle C.$$

§2. 代数和 σ－ 代数. 可测空间.

1. 代数和 $\sigma-$ 代数 在建立 (试验的) 概率空间时, 代数和 $\sigma-$ 代数是组成元素. 关于代数和 $\sigma-$ 代数, 将举一些例子并介绍一系列性质.

设 Ω 是一基本事件空间. 集合系

$$\mathscr{F}_* = \{\varnothing, \Omega\},\ \mathscr{F}^* = \{A : A \subseteq \Omega\}$$

显然是代数, 同时也是 $\sigma-$ 代数. 这时, $\mathscr{F}_*$ 是平凡的、最 "贫乏" 的 $\sigma-$ 代数, 而 $\mathscr{F}^*$ 是由 Ω 的一切子集构成的、最 "广泛" 的 $\sigma-$ 代数.

对于有限空间 Ω, $\sigma-$ 代数 $\mathscr{F}^*$ 非常直观, 一般这正是在初等概率论里所考虑的 "事件" 系. 对于不可数空间 Ω, 集系 $\mathscr{F}^*$ 显得过于广泛, 致使在这样的集系上不是总能 "合适地" 定义概率.

如果 $A \subseteq \Omega$, 则集系

$$\mathscr{F}_A = \{A, \overline{A}, \varnothing, \Omega\}$$

也是代数 ($\sigma-$ 代数) 的例, 并称做由集合 A 诱导的或由集合 A 生成的代数 ($\sigma-$ 代数).

这一集系是由分割生成的集系的特殊情形. 确切地说, 设

$$\mathscr{D} = \{D_1, D_2, \cdots\}$$

是某一把 Ω 分为非空集合的有限分割:

$$\Omega = D_1 + D_2 + \cdots,\ D_i \cap D_j = \varnothing,\ i \neq j.$$

那么, 由分割的有限或可数个元素之并形成的集系 $\mathscr{A}=\alpha(\mathscr{D})$ (连同空集) 是代数 (σ- 代数).

下面的引理有重要意义, 因为它证明"建立含给定集系的最小代数 (σ- 代数)"原则上是可能的.

引理 1 *设 $\mathscr{E}$ 是 Ω 中的一个集系. 那么, 存在含 $\mathscr{E}$ 中所有集合的最小代数 $\alpha(\mathscr{E})$ 和最小 σ- 代数 $\sigma(\mathscr{E})$.*

证明 Ω 中一切子集的集系 $\mathscr{F}^*$ 是 σ- 代数. 因此, 至少存在一个包含 $\mathscr{E}$ 的代数和一个包含 $\mathscr{E}$ 的 σ- 代数. 现在, 由属于包含的 $\mathscr{E}$ 的任意代数 (σ- 代数) 的一切集合, 组成集系 $\alpha(\mathscr{E})$ ($\sigma(\mathscr{E})$). 不难验证, 代数 $\alpha(\mathscr{E})$ (σ- 代数 $\sigma(\mathscr{E})$) 就是引理所要求的最小代数 (σ- 代数).

注 1 常把集系 $\alpha(\mathscr{E})$ (相应的 $\sigma(\mathscr{E})$) 称做由集系 $\mathscr{E}$ 诱导的 (生成的) 最小代数 (相应的最小 σ- 代数).

前面已经指出 (§1 第 3 小节), σ- 代数在概率空间的定义中占重要地位. 因此, 希望得到构造由某一代数 $\mathscr{A}$ 生成的 σ- 代数的构造方法 (引理 1 仅给出了这样 σ- 代数的存在性, 但没有提供其有效的构造方法).

下面是可以想到的且很自然的、由 $\mathscr{A}$ 构造 σ- 代数 $\sigma(\mathscr{A})$ 的方法之一.

设 $\mathscr{E}$ 是 Ω 的一个子集系. 设 $\widehat{\mathscr{E}}$ 是 Ω 的一个子集系, 包含 $\mathscr{E}$ 中的一切子集及其补集, 以及 $\mathscr{E}$ 中集合的有限或可数个集合的并. 令 $\mathscr{A}_0=\mathscr{A}$, $\mathscr{A}_1=\widehat{\mathscr{A}}_0$, $\mathscr{A}_2=\widehat{\mathscr{A}}_1,\cdots$. 显然对于每个 n, 集合 $\mathscr{A}_n$ 属于 $\sigma(\mathscr{A})$, 而且似乎可以指望, 对于某个 n 有 $\mathscr{A}_n=\sigma(\mathscr{A})$, 或者至少 $\bigcup\mathscr{A}_n=\sigma(\mathscr{A})$.

然而, 事情一般并非如此. 事实上, 取 $\Omega=(0,1]$, 而作为代数 $\mathscr{A}$, 考虑由空集 ∅ 以及形如 $(a,b]$ 的区间之有限和诱导的 Ω 的子集系, 其中 a,b 是有理数. 不难证明, 这时集系 $\bigcup_{n=1}^{\infty}\mathscr{A}_n$ 严格小于 σ- 代数 $\sigma(\mathscr{A})$.

我们以后主要关心的, 不是如何由代数 $\mathscr{A}$ 构造最小 σ- 代数 $\sigma(\mathscr{A})$ 的问题, 而是确定给定的不同集系是否 σ- 代数 $\sigma(\mathscr{A})$ 的问题.

为得到上述问题的答案, 要用到如下重要的概念——"单调类".

定义 1 Ω 的子集系 $\mathscr{M}$, 称做单调类, 如果由 $A_n\in\mathscr{M}, n=1,2,\cdots, A_n\uparrow A$ 或 $A_n\downarrow A$, 可得 $A\in\mathscr{M}$.

设 $\mathscr{E}$ 是某一集系. 记 $\mu(\mathscr{E})$ 为包含 $\mathscr{E}$ 的最小单调类. (仿照引理 1 的证明, 同样可以证明最小单调类存在性.)

引理 2 *代数 $\mathscr{A}$ 同时又是最小 σ- 代数 $\sigma(\mathscr{A})$ 的充分必要条件是, $\mathscr{A}$ 是最小单调类.*

证明 每一个 σ- 代数显然是单调类. 现在设 $\mathscr{A}$ 是单调类, 而 $A_n\in\mathscr{A}, n=1,2,\cdots$. 显然,

$$B_n=\bigcup_{i=1}^{n}A_i\in\mathscr{A} \text{ 且 } B_n\subseteq B_{n+1}, n=1,2,\cdots.$$

从而根据单调类的定义,

$$B_n\uparrow\bigcup_{i=1}^{\infty}A_i\in\mathscr{A}.$$

类似可得 $\bigcap_{i=1}^{\infty}A_i\in\mathscr{A}$. □

我们利用该引理证明, 说明如下"σ- 代数"与"单调类"两概念之间联系的定理.

定理 1 *设 $\mathscr{A}$ 是代数, 则*

$$\mu(\mathscr{A})=\sigma(\mathscr{A}). \qquad (1)$$

证明 由引理 2, 知 $\mu(\mathscr{A})\subseteq\sigma(\mathscr{A})$. 因此, 只需证明 $\mu(\mathscr{A})$ 是 σ- 代数. 但是, 由于 $\mathscr{M}=\mu(\mathscr{A})$ 是单调类, 故仍由引理 2 知, 只需证明 $\mu(\mathscr{A})$ 是代数.

对于任意 $A\in\mathscr{M}$, 证明 $\bar{A}\in\mathscr{M}$. 为此运用如下以后常用的"适当集合原理".

以

$$\widetilde{\mathscr{M}}=\{B: B\in\mathscr{M}, \bar{B}\in\mathscr{M}\}$$

表示"使 $\mu(\mathscr{A})$ 是代数"所有集合的全体. 显然 $\mathscr{A}\subseteq\widetilde{\mathscr{M}}\subseteq\mathscr{M}$. 现在证明 $\widetilde{\mathscr{M}}$ 是单调类.

设 $B_n\in\widetilde{\mathscr{M}}$, 则 $B_n\in\mathscr{M}$, $\bar{B}_n\in\mathscr{M}$, 因此

$$\lim\uparrow B_n\in\mathscr{M}, \lim\uparrow\bar{B}_n\in\mathscr{M}, \lim\downarrow B_n\in\mathscr{M}, \lim\downarrow\bar{B}_n\in\mathscr{M}.$$

从而

$$\overline{\lim\uparrow B_n}=\lim\downarrow\bar{B}_n\in\mathscr{M}, \overline{\lim\downarrow B_n}=\lim\uparrow\bar{B}_n\in\mathscr{M},$$
$$\overline{\lim\uparrow\bar{B}_n}=\lim\downarrow B_n\in\mathscr{M}, \overline{\lim\downarrow\bar{B}_n}=\lim\uparrow B_n\in\mathscr{M}.$$

即 $\widetilde{\mathscr{M}}$ 是单调类. 由于 $\widetilde{\mathscr{M}}\subseteq\mathscr{M}$, 可见 $\mathscr{M}$ 是最小单调类, 所以 $\widetilde{\mathscr{M}}=\mathscr{M}$. 并且如果 $A\in\mathscr{M}=\mu(\mathscr{A})$, 则 $\bar{A}\in\mathscr{M}$, 即类 $\mathscr{M}$ 关于补的运算封闭.

现在证明, 类 $\mathscr{M}$ 关于交的运算也封闭.

设 $A\in\mathscr{M}$, 且

$$\mathscr{M}_A=\{B: B\in\mathscr{M}, A\cap B\in\mathscr{M}\}.$$

由等式

$$\lim\downarrow(A\cap B_n)=A\cap\lim\downarrow B_n, \lim\uparrow(A\cap B_n)=A\cap\lim\uparrow B_n,$$

可见 $\mathscr{M}_A$ 是单调类.

其次, 容易验证

$$(A\in\mathscr{M}_B)\Leftrightarrow(B\in\mathscr{M}_A). \qquad (2)$$

现在设 $A\in\mathscr{A}$, 则由于 $\mathscr{A}$ 是代数, 可见对于任意 $B\in\mathscr{A}$, 集合 $A\cap B\in\mathscr{A}$, 因此

$$\mathscr{A}\subseteq\mathscr{M}_A\subseteq\mathscr{M}.$$

由于 $\mathscr{M}_A$ 是单调类, 而 $\mathscr{M}$ 是最小单调类, 可见对于任意 $A\in\mathscr{A}$, $\mathscr{M}_A=\mathscr{M}$. 由 (2) 式可见, 对于 $A\in\mathscr{A}$, $B\in\mathscr{M}$, 有

$$(A\in\mathscr{M}_B)\Leftrightarrow(B\in\mathscr{M}_A=\mathscr{M}).$$

从而, 如果 $A\in\mathscr{A}$, 则对于任意 $B\in\mathscr{M}$, 有

$$A\in\mathscr{M}_B.$$

因为 $A\in\mathscr{A}$ 是任意的, 所以由此得

$$\mathscr{A}\subseteq\mathscr{M}_B\subseteq\mathscr{M}.$$

因此, 对于任意 $B\in\mathscr{M}$,

$$\mathscr{M}_B=\mathscr{M},$$

即对于任意 $B\in\mathscr{M}$, $C\in\mathscr{M}$, 有 $C\bigcap B\in\mathscr{M}$.

从而, $\mathscr{M}$ 关于补和交 (从而关于并) 的运算是封闭类, 即 $\mathscr{M}=\mu(\mathscr{A})$ 是代数. 于是定理得证. □

由对以上所作证明的分析, 可见在考虑按 "适当集合原理" 形成的集系时, 这些集系关于某些理论上的集合运算的封闭性是重要的.

从这种观点出发, 在整个有关 "单调类" 的问题中, 引进所谓 "$\pi-$ 系" 和 "$\lambda-$ 系" 的概念是有益的. 其实, 在证明定理 1 时已经用到这些概念. 这些概念, 还可以用来证明涉及所研究问题的一系列命题 (例如, 定理 2), 而且往往比直接验证有关集系是 "单调类" 更加方便.

定义 2 ("$\pi-\lambda-$ 系") 设 Ω 是某个空间. 空间 Ω 子集的集系 $\mathscr{P}$, 称做 $\pi-$ 系, 如果它关于有限次的交运算封闭: 若 $A_1,\cdots,A_n\in\mathscr{P}$, 则 $\bigcap_{k=1}^n A_k\in\mathscr{P}$, $n\geqslant 1$.

空间 Ω 子集的集系 $\mathscr{L}$, 称做 $\lambda-$ 系, 如果

(λ_a) $\Omega\in\mathscr{L}$,

(λ_b) $(A,B\in\mathscr{L}$ 且 $A\subseteq B)\Rightarrow(B\setminus A\in\mathscr{L})$,

(λ_c) $(A_n\in\mathscr{L}, n\geqslant 1$, 且 $A_n\uparrow A)\Rightarrow(A\in\mathscr{L})$.

同时是 $\pi-$ 系和 $\lambda-$ 系的空间 Ω 中的子集的集系 $\mathscr{D}$, 称做邓肯 (Е. Б. Дынкин) $\pi-\lambda-$ 系或 $d-$ 系.

注 2 指出如下事实是有益的: $\lambda-$ 系的定义中的条件 (λ_a), (λ_b) 和 (λ_c), 等价于条件 (λ_a), (λ'_b) 和 (λ'_c), 其中

(λ'_b) 若 $A\in\mathscr{L}$, 则 $\overline{A}\in\mathscr{L}$;

(λ'_c) 若 $A_n\in\mathscr{L}$, $n\geqslant 1$, $A_n\bigcap A_m=\varnothing\,(m\neq n)$, 则 $\bigcup A_n\in\mathscr{L}$.

还应指出, 显然任何代数都是 $\pi-$ 系.

设 $\mathscr{E}$ 是某一集系, 则以 $\pi(\mathscr{E})$, $\lambda(\mathscr{E})$ 和 $d(\mathscr{E})$ 分别表示包含 $\mathscr{E}$ 的 $\pi-$ 系, $\lambda-$ 系和 $d-$ 系.

由下面的定理, 可见 $\pi-\lambda-$ 系的作用. 为说明该定理的含义, 我们指出, 每一 $\sigma-$ 代数都是 $\lambda-$ 系. 然而, 逆命题一般不成立. 例如, 若 $\Omega=\{1,2,3,4\}$, 则集系

$$\mathscr{D}=\{\varnothing,\Omega,\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\}\}$$

是 $\lambda-$ 系, 但不是 $\sigma-$ 代数.

不过, 若再补充要求 $\lambda-$ 系同时又是 $\pi-$ 系, 则所得 $\pi-\lambda-$ 系就是 $\sigma-$ 代数.

定理 2 (关于 $\pi-\lambda-$ 系) a) 任何 $\pi-\lambda-$ 系 $\mathscr{E}$ 都是 $\sigma-$ 代数.

b) 设 $\mathscr{E}$ 是集合的 $\pi-$ 系, 则 $\lambda(\mathscr{E})=d(\mathscr{E})=\sigma(\mathscr{E})$.

c) 设 $\mathscr{E}$ 是集合的 $\pi-$ 系, 而 $\mathscr{L}$ 是某一 $\lambda-$ 系且 $\mathscr{E}\subseteq\mathscr{L}$, 则 $\sigma(\mathscr{E})\subseteq\mathscr{L}$.

证明 a) (由于 (λ_a)) 系 $\mathscr{E}$ 包含 Ω, 且 (由于 (λ'_b)) 关于求补及有限次交封闭 (根据 (λ'_b) 及关于 $\mathscr{E}$ 是 $\pi-$ 系的假设). 而根据假设 $\mathscr{E}$ 是 $\pi-$ 系. 因此, (根据 §1 的定义 1) 集系 $\mathscr{E}$ 是代数. 为证明集系 $\mathscr{E}$ 也是 $\sigma-$ 代数, (根据 §1 的定义 4) 需要证明: 如果集合 $B_1,B_2,\cdots\in\mathscr{E}$, 则其并 $\bigcup_n B_n$ 也属于 $\mathscr{E}$.

设 $A_1=B_1$, $A_n=B_n\cap\overline{A}_1\bigcap\cdots\bigcap\overline{A}_{n-1}$, $n\geqslant 1$, 则根据 (λ'_c), $\bigcup A_n\in\mathscr{E}$. 由于 $\bigcup B_n=\bigcup A_n$, 则 $\bigcup B_n\in\mathscr{E}$.

于是 $\pi-\lambda-$ 系是 $\sigma-$ 代数.

b) 考虑 $\lambda-$ 系 $\lambda(\mathscr{E})$ 和 $\sigma-$ 代数 $\sigma(\mathscr{E})$. 如前面已经指出, 任何 $\sigma-$ 代数都是 $\lambda-$ 系. 那么, 由于 $\sigma(\mathscr{E})\supseteq\mathscr{E}$, 则 $\sigma(\mathscr{E})=\lambda(\sigma(\mathscr{E}))\supseteq\lambda(\mathscr{E})$. 因此, $\lambda(\mathscr{E})\subseteq\sigma(\mathscr{E})$.

假如现在能证明, $\lambda-$ 系 $\lambda(\mathscr{E})$ 也是 $\pi-$ 系, 那么由命题 a), 可见 $\lambda(\mathscr{E})$ 是包含 $\mathscr{E}$ 的 $\sigma-$ 代数. 由于 $\sigma(\mathscr{E})$ 是包含 $\mathscr{E}$ 的最小 $\sigma-$ 代数, 故由已经证明的 $\lambda(\mathscr{E})\subseteq\sigma(\mathscr{E})$, 可见 $\lambda(\mathscr{E})=\sigma(\mathscr{E})$.

因此, 只需要证明 $\lambda(\mathscr{E})$ 是 $\pi-$ 系.

仿照定理 1 的证明, 并且利用 "适当集合原理".

设

$$\mathscr{E}_1=\{B\in\lambda(\mathscr{E}):\ 对于一切\ A\in\mathscr{E},\ B\bigcap A\in\lambda(\mathscr{E})\}.$$

若 $B\in\mathscr{E}$, 则 $B\bigcap A\in\mathscr{E}$ (因为 $\mathscr{E}$ 是 $\pi-$ 系). 因而 $\mathscr{E}\subseteq\mathscr{E}_1$. 因为 (按 $\mathscr{E}_1$ 的定义) $\mathscr{E}_1$ 是 $\lambda-$ 系, 所以 $\lambda(\mathscr{E})\subseteq\lambda(\mathscr{E}_1)=\mathscr{E}_1$. 另一方面, 按 $\mathscr{E}_1$ 的定义, 有包含关系 $\mathscr{E}_1\subseteq\lambda(\mathscr{E})$.

从而 $\mathscr{E}_1=\lambda(\mathscr{E})$.

现在设

$$\mathscr{E}_2=\{B\in\lambda(\mathscr{E}):\ 对于一切\ A\in\lambda(\mathscr{E}),\ B\bigcap A\in\lambda(\mathscr{E})\}.$$

同 $\mathscr{E}_1$ 一样, $\mathscr{E}_2$ 也是 $\lambda-$ 系.

取集合 $B\in\mathscr{E}$, 那么根据 $\mathscr{E}_1$ 的定义, 对于 $A\in\mathscr{E}_1=\lambda(\mathscr{E})$, 有 $B\bigcap A\in\lambda(\mathscr{E})$. 从而, 根据 $\mathscr{E}_2$ 的定义, 可见 $\mathscr{E}\subseteq\mathscr{E}_2$, 且 $\lambda(\mathscr{E})\subseteq\lambda(\mathscr{E}_2)=\mathscr{E}_2$. 因此, $\lambda(\mathscr{E})\supseteq\mathscr{E}_2$, 所以 $\lambda(\mathscr{E})=\mathscr{E}_2$. 故对于一切 $A,B\in\lambda(\mathscr{E})$, 集合 $B\bigcap A\in\lambda(\mathscr{E})$, 即 $\lambda(\mathscr{E})$ 是 $\pi-$ 系. 这样, $\lambda(\mathscr{E})$ 是 $\pi-\lambda-$ 系 (或 $\lambda(\mathscr{E})=\sigma(\mathscr{E})$), 如前面所指出的那样. 由此可见 $\lambda(\mathscr{E})=\sigma(\mathscr{E})$.

于是, 命题 b) 得证.

c) 由于 $\mathscr{E}\subseteq\mathscr{D}$ 且 $\mathscr{D}$ 是 $\lambda-$ 系, 可见 $\lambda(\mathscr{E})\subseteq\lambda(\mathscr{D})=\mathscr{D}$. 由 b) 知 $\lambda(\mathscr{E})=\sigma(\mathscr{E})$. 于是, $\sigma(\mathscr{E})\subseteq\mathscr{D}$. □

注 3 定理 2 的结果可以直接由定理 1 得到 (练习题 10).

下面两个命题的证明, 是 "适当集合原理" 和关于 $\pi-\lambda-$ 系的定理 2 之很好的演示.

引理 3 *设 $\mathbf{P}$ 和 $\mathbf{Q}$ 是可测空间 $(\Omega,\mathscr{F})$ 上的两个概率测度; $\mathscr{E}$ 是 $\mathscr{F}$ 中集合的 $\pi-$ 系, 且测度 $\mathbf{P}$ 和 $\mathbf{Q}$ 在 $\mathscr{E}$ 中的集合上相等. 那么, 这两个测度 $\mathbf{P}$ 和 $\mathbf{Q}$ 在 $\sigma-$ 代数 $\sigma(\mathscr{E})$ 上也相等. 特别, 若 $\mathscr{A}$ 是代数, 则测度 $\mathbf{P}$ 和 $\mathbf{Q}$ 在 $\sigma-$ 代数 $\sigma(\mathscr{A})$ 上也相等.*

证明 利用 "适当集合原理", 作为这样的集合, 取

$$\mathscr{D}=\{A\in\sigma(\mathscr{E}):\mathbf{P}(A)=\mathbf{Q}(A)\}.$$

显然 $\Omega\in\mathscr{D}$; 如果 $A\in\mathscr{D}$, 则由于 $\mathbf{P}(\overline{A})=1-\mathbf{P}(A)=1-\mathbf{Q}(A)=\mathbf{Q}(\overline{A})$, 显然 $\overline{A}\in\mathscr{D}$. 如果 $A_1,A_2,\cdots\in\mathscr{D}$ 且两两不相交, 则由测度 $\mathbf{P}$ 和 $\mathbf{Q}$ 的可数可加性, 有

$$\mathbf{P}\left(\bigcup_n A_n\right)=\sum_n\mathbf{P}(A_n)=\sum_n\mathbf{Q}(A_n)=\mathbf{Q}\left(\bigcup_n A_n\right).$$

从而, 性质 (λ_a), (λ_b') 和 (λ_c') 成立, 因此 $\mathscr{D}$ 是 $\lambda-$ 系.

根据引理的条件 $\mathscr{E}\subseteq\mathscr{D}$, 按 $\mathscr{E}$ 是 $\pi-$ 系. 那么, 由定理 2 的命题 c), 可见 $\sigma(\mathscr{E})\subseteq\mathscr{D}$. 由于适当集合的定义, 这一性质恰好表示测度 $\mathbf{P}$ 和 $\mathbf{Q}$ 在 $\sigma-$ 代数 $\sigma(\mathscr{E})$ 上相等. □

引理 4 *设 $\mathscr{A}_1,\mathscr{A}_2,\cdots,\mathscr{A}_n$ 是 (关于测度 $\mathbf{P}$) 相互独立的事件代数. 那么, $\sigma-$ 代数 $\sigma(\mathscr{A}_1),\sigma(\mathscr{A}_2),\cdots,\sigma(\mathscr{A}_n)$ (关于测度 $\mathbf{P}$) 也相互独立.*

证明 我们首先指出, 在概率的一般理论中, 集合以及集合系 (代数, $\sigma-$ 代数, $\cdots$) 的独立性的定义, 和初等概率论中独立性的定义完全一致 (见第一章 §4, 定义 2 ~ 5).

设 $A_2,\cdots,A_n$ 相应为 $\mathscr{A}_2,\cdots,\mathscr{A}_n$ 中的集合, 记

$$\mathscr{D}_1=\left\{A\in\sigma(\mathscr{A}_1):\mathbf{P}(A\cap A_2\cap\cdots\cap A_n)=\mathbf{P}(A)\prod_{k=2}^n\mathbf{P}(A_k)\right\}. \tag{3}$$

现在证明 $\mathscr{D}_1$ 是 $\lambda-$ 系.

显然 $\Omega\in\mathscr{D}_1$, 即性质 (λ_a) 成立. 设 $A,B\in\mathscr{D}_1$, 而 $A\subseteq B$. 那么, 由

$$\mathbf{P}(A\cap A_2\cap\cdots\cap A_n)=\mathbf{P}(A)\prod_{k=2}^n\mathbf{P}(A_k)$$

和

$$\mathbf{P}(B\cap A_2\cap\cdots\cap A_n)=\mathbf{P}(B)\prod_{k=2}^n\mathbf{P}(A_k),$$

因此由第二式减去第一式, 得

$$\mathbf{P}((B\setminus A)\cap A_2\cap\cdots\cap A_n)=\mathbf{P}(B\setminus A)\prod_{k=2}^n\mathbf{P}(A_k).$$

因而条件 (λ_b) 成立. 最后, 如果集合 $B_k\in\sigma(\mathscr{A}_1)$, $k\geqslant1$, 且 $B_k\uparrow B$, 则

$$B_k\cap A_2\cap\cdots\cap A_n\uparrow B\cap A_2\cap\cdots\cap A_n.$$

因此, 因为概率 $\mathbf{P}$ 的上连续性 (见 §1 的定理), 当 $k\to\infty$ 时对

$$\mathbf{P}(B_k\cap A_2\cap\cdots\cap A_n)=\mathbf{P}(B_k)\prod_{i=2}^n\mathbf{P}(A_i)$$

求极限, 得

$$\mathbf{P}(B\cap A_2\cap\cdots\cap A_n)=\mathbf{P}(B)\prod_{i=2}^n\mathbf{P}(A_i),$$

即条件 (λ_c) 成立.

因此, $\mathscr{D}_1$ 是 $\lambda-$ 系, 且 $\mathscr{D}_1\supseteq\mathscr{A}_1$. 根据定理 2 的命题 c), 可得 $\mathscr{D}_1\supseteq\sigma(\mathscr{A}_1)$.

于是, 证明了集系 $\sigma(\mathscr{A}_1),\mathscr{A}_2,\cdots,\mathscr{A}_n$ 相互独立.

对于集系 $\sigma(\mathscr{A}_2),\mathscr{A}_3,\cdots,\mathscr{A}_n,\sigma(\mathscr{A}_1)$ 进行类似的论证, 可见集系 $\sigma(\mathscr{A}_2),\mathscr{A}_3,\cdots,\mathscr{A}_n,\sigma(\mathscr{A}_1)$ 相互独立, 或等价于集系 $\mathscr{A}_3,\cdots,\mathscr{A}_n,\sigma(\mathscr{A}_1),\sigma(\mathscr{A}_2)$.

继续这一过程, 可得由 $\sigma-$ 代数 $\sigma(\mathscr{A}_1),\sigma(\mathscr{A}_2),\cdots,\sigma(\mathscr{A}_n)$ 构成的相互独立的集系. □

注 4 我们再次分析, 为使集系形成 $\sigma-$ 代数, 需要满足什么条件.

为此, 如果集系关于可数次交运算

$$A_1,A_2,\cdots\in\mathscr{E}\Rightarrow\bigcap_{n=1}^\infty A_n\in\mathscr{E}$$

封闭, 则称之为 π^*- 系.

那么, 由 $\sigma-$ 代数的定义可见, 如果某一代数 $\mathscr{E}$ 同时又是 π^*- 系, 则它就是 $\sigma-$ 代数.

基于"π–λ 系"概念的处理方法略有不同，这里不是从"代数"的概念出发，而是从"λ– 系"概念出发. 并且，如由定理 2 的命题 a) 可见，若此 λ– 系同时又是 π– 系，则它也是 σ– 代数.

显然，这恰好是上两种处理方法的区别所在.

当验证某集系是否 σ– 代数时，我们从检验该集系是否代数开始，就是说，进行此检验只需考虑集合的有限和 (或交). "可数性" (这正式 "全部要做的") 出现在下列情形：需要检验一个集系是否 π– 系.

对于 "λ–π 方法"，验证性质 "有关集系是 σ– 代数"，我们首先从 "确定该集系是否 λ– 系" 开始，λ– 系的性质 (λ_c) 或 (λ'_c) 已经涉及"可数" 运算. 但是，在第二阶段当验证该集系是否 π– 系时，我们只进行集合的有限交或和的运算.

在结束叙述有关 "单调类" 的有关结果时，我们给出其一种函数形式. (第八章 §2 定理 1 之引理的证明，可以作为引用的例子.)

定理 3 *设 $\mathscr{E}$ 是 $\mathscr{F}$ 中集合的 π– 系，而 $\mathscr{H}$ 是实数值 $\mathscr{F}$– 可测函数的全体，具有下列性质：*

(h_1) 若 $A\in\mathscr{E}$，则函数 $I_A\in\mathscr{H}$；

(h_2) 若 $f\in\mathscr{H}, h\in\mathscr{H}$，则对于任何实数 c, $f+h\in\mathscr{H}$, $cf\in\mathscr{H}$；

(h_3) 若函数 $h_n\in\mathscr{H}, n\geqslant 1, 0\leqslant h_n\uparrow h$，则 $h\in\mathscr{H}$.

那么，集系 $\mathscr{H}$ 包含一切关于 σ– 代数 $\sigma(\mathscr{E})$ 可测的有界函数.

证明 设 $\mathscr{L}=\{A\in\mathscr{F}: I_A\in\mathscr{H}\}$. 由 (h_1)，可见 $\mathscr{E}\subseteq\mathscr{L}$. 由于 (h_2) 和 (h_3)，可见集系 $\mathscr{L}$ 是 λ– 系 (练习题 11). 因此由定理 2 的命题 c)，可见 $\sigma(\mathscr{E})\subseteq\mathscr{L}$. 从而，由 $A\in\sigma(\mathscr{E})$，可见 $I_A\in\mathscr{H}$. 根据性质 (h_2)，由此可见，所有简单函数也属于集系 $\mathscr{H}$. (所谓简单函数，即形如 I_{A_i}, $A_i\in\sigma(\mathscr{E})$ 的函数的有限线性组合). 最后，由性质知，一切关于 σ– 代数 $\sigma(\mathscr{E})$ 可测的有界函数属于集系 $\mathscr{H}$. □

注 5 设 $X_1,\cdots,X_n$ 是可测空间 $(\Omega,\mathscr{F})$ 上的随机变量，$\mathscr{F}^X=\sigma(X_1,\cdots,X_n)$，而 $f=f(\omega)$ 是 $\mathscr{F}^X$– 可测函数. 那么，存在博雷尔函数 $F=F(x_1,\cdots,x_n)$，使 $f(\omega)=F(X_1(\omega),\cdots,X_n(\omega))$.

为证明这一命题，只需利用定理 3. 在定理 3 中，作为 "适当集合原理" 中函数的适当集合，取非负博雷尔函数 $F=F(x_1,\cdots,x_n)$ 的集合 $\mathscr{H}$，而作为集合 $\mathscr{E}$，取

$$\mathscr{E}=\{\omega: X_1(\omega)\leqslant x_1,\cdots,X_n(\omega)\leqslant x_n; x_i\in\mathbb{R}, i=1,\cdots,n\}.$$

由定理 3 知，任意非负 $\mathscr{F}^X$– 可测的函数 $f=f(\omega)$，都可以通过非负博雷尔函数 $F=F(x_1,\cdots,x_n)$ 表示为 $f(\omega)=F(X_1(\omega),\cdots,X_n(\omega))$. 一般 (未必非负函数)，函数 f 通过极限过度可以表示为 $f=f^+-f^-$.

下面研究对于概率论最重要的各种可测空间 $(\Omega,\mathscr{F})$.

2. 可测空间 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ 设 $\mathbb{R}=(-\infty,\infty)$ 是实数轴，对于任意 $-\infty\leqslant a<b<\infty$，设

$$(a,b]=\{x\in\mathbb{R}: a<x\leqslant b\}.$$

假定把区间 $(a,\infty]$ 理解为区间 (a,∞). (这样的假定之所以必要，为使区间 $(-\infty,b]$ 的补集具有同样的形式，即左开而右闭.)

以 $\mathscr{A}$ 表示集合 $\mathbb{R}$ 上，有限个形如 $(a,b]$ 的不相交区间之和构成的集系：

$$A\in\mathscr{A},\quad 其中\ A=\sum_{i=1}^{n}(a_i,b_i],\ n<\infty.$$

不难验证，如果给 $\mathscr{A}$ 补充上空集 $\varnothing$，则所得集系形成代数，但不是 σ– 代数，因为如果 $A_n=(0,1-1/n]\in\mathscr{A}$，但是

$$\bigcup_n A_n=(0,1)\notin\mathscr{A}.$$

设 $\mathscr{B}(\mathbb{R})$ 是包含集系 $\mathscr{A}$ 的最小 σ– 代数 $\sigma(\mathscr{A})$. 在数学分析中，这一 σ– 代数有重要应用，称做数轴上集合的博雷尔代数，其中的元素称做博雷尔集.

若以 $\mathscr{I}$ 表示形如 $I=(a,b]$ 的区间 I 的集系，而以 $\sigma(\mathscr{I})$ 表示包含 $\mathscr{I}$ 的最小 σ– 代数，则不难验证 $\sigma(\mathscr{I})$ 是博雷尔代数. 换句话说，由于 $\sigma(\mathscr{I})=\sigma(\alpha(\mathscr{I}))$，博雷尔代数，可以不通过代数 $\mathscr{A}$，由集系 $\mathscr{I}$ 得到.

由于对于 $a<b$，

$$(a,b)=\bigcup_{n=1}^{\infty}\left(a,b-\frac{1}{n}\right],\quad [a,b]=\bigcap_{n=1}^{\infty}\left(a-\frac{1}{n},b\right].$$

$$\{a\}=\bigcap_{n=1}^{\infty}\left(a-\frac{1}{n},a\right],$$

可见，博雷尔代数，除包含形如 $(a,b]$ 的区间外，还包含单点集 $\{a\}$，以及如下 6 种集合：

$$(a,b),\ [a,b],\ [a,b),\ (-\infty,b),\ (-\infty,b],\ (a,-\infty). \tag{4}$$

还应指出，由 $\mathbb{R}$ 中的集系生成的一切最小 σ– 代数，都由 (4) 式中同一种不相交区间之有限和构成，并且等于 $\mathscr{B}(\mathbb{R})$. 因此，构造博雷尔代数 $\mathscr{B}(\mathbb{R})$，可以不从形如 $(a,b]$ 的区间出发，而从上述 6 种区间的任何一种出发.

有时，不得不涉及扩展数轴上 $\overline{\mathbb{R}}=[-\infty,\infty]$ 的集合的代数 $\mathscr{B}(\overline{\mathbb{R}})$，称做由 $\overline{\mathbb{R}}$ 上的集系诱导的、由形如

$$(a,b]=\{x\in\overline{\mathbb{R}}: a<x\leqslant b\},\quad -\infty\leqslant a<b\leqslant\infty$$

的不相交区间之有限和构成的最小 σ– 代数，其中 $(-\infty,b]$ 即集合 $\{x\in\overline{\mathbb{R}}: -\infty\leqslant x\leqslant b\}$.

注 1 对于可测空间 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$，又常使用记号 $(\mathbb{R},\mathscr{B})$, $(\mathbb{R}^1,\mathscr{B}_1)$.

注 2 在数轴 R 上引进 (与通常的欧几里得度量 $|x-y|$ 等价的) 度量:

$$\rho_1(x,y)=\frac{|x-y|}{1+|x-y|},$$

而以 $\mathscr{B}_0(R)$ 表示,由有限个 "形如 $S_\rho(x^0)$ 的不相交开集之和" 诱导的、最小 σ− 代数,其中对于 $\rho>0, x^0\in R$,

$$S_\rho(x^0)=\{x\in R:\rho_1(x,x^0)<\rho\}.$$

那么,$\mathscr{B}_0(R)=\mathscr{B}(R)$ (见练习题 7).

3. 可测空间 $(R^n,\mathscr{B}(R^n))$ 设 $R^n=R\times\cdots\times R$ 是 n 条 (份) 数轴的直积 (或欧几里得直积),亦即有序数组 $x=(x_1,\cdots,x_n)$ 的集合,其中 $-\infty<x_k<\infty, k=1,\cdots,n$. 集合 $I=I_1\times\cdots\times I_n$,其中 $I_k=(a_k,b_k]$,即集合 $\{x\in R^n: x_k\in I_k, k=1,\cdots,n\}$ 称做*矩形*,而 I_k 称做此矩形的*边*. 以 $\mathscr{I}$ 表示有限不相交矩形 I 之和构成的集合的全体. 由矩形系 $\mathscr{I}$ 生成的最小 σ− 代数 $\sigma(\mathscr{I})$,称做中集合的博雷尔代数,记作 $\mathscr{B}(R^n)$. 现在说明,可以用另一种方式得到这一博雷尔代数.

与矩形 $I=I_1\times\cdots\times I_n$ 同时,考虑具有博雷尔边的矩形 $B=B_1\times\cdots\times B_n$ (B_k 是数轴上的博雷尔集,且在直积 $R^n=R\times\cdots\times R$ 中占据第 k 个位置). 包含具有博雷尔边的一切矩形的最小 σ− 代数,记作

$$\mathscr{B}(R)\otimes\cdots\otimes\mathscr{B}(R).$$

并且称做 σ− 代数 $\mathscr{B}(R)$ 的直积. 我们证明,实际上,

$$\mathscr{B}(R^n)=\mathscr{B}(R)\otimes\cdots\otimes\mathscr{B}(R).$$

换句话说,"由有限个不相交矩形 $B=B_1\times\cdots\times B_n$ 之和" 形成的集系生成的最小 σ− 代数,与 "具有博雷尔边的、有限个不相交矩形 $I=I_1\times\cdots\times I_n$ 之和" 形成的集系,二者重合.

证明本质上依赖于下面的引理.

引理 5 *设 $\mathscr{E}$ 是 Ω 的某一集系,集合 $B\subseteq\Omega$;假设根据定义,*

$$\mathscr{E}\cap B=\{A\cap B: A\in\mathscr{E}\}, \tag{5}$$

而 $\sigma(\mathscr{E}\cap B)$ 是由集系 $\mathscr{E}\cap B$ 生成的 B 的子集最小 σ− 代数. 那么

$$\sigma(\mathscr{E}\cap B)=\sigma(\mathscr{E})\cap B. \tag{6}$$

证明 由于 $\mathscr{E}\subseteq\sigma(\mathscr{E})$,可见

$$\mathscr{E}\cap B\subseteq\sigma(\mathscr{E})\cap B. \tag{7}$$

由于 $\sigma(\mathscr{E})\cap B$ 是 B 中的 σ- 代数,因此由 (7) 式,可见

$$\sigma(\mathscr{E}\cap B)\subseteq\sigma(\mathscr{E})\cap B.$$

为证明相反的包含关系,仍利用适当集合原理. 记

$$\mathscr{C}_B=\{A\in\sigma(\mathscr{E}): A\cap B\in\sigma(\mathscr{E}\cap B)\}.$$

由于 $\sigma(\mathscr{E})$ 和 $\sigma(\mathscr{E}\cap B)$ 是 σ- 代数,则 $\mathscr{C}_B$ 也是 σ- 代数,而且显然

$$\mathscr{E}\subseteq\mathscr{C}_B\subseteq\sigma(\mathscr{E}).$$

由 $\sigma(\mathscr{E})\subseteq\sigma(\mathscr{C}_B)=\mathscr{C}_B\subseteq\sigma(\mathscr{E})$,可见 $\sigma(\mathscr{E})=\mathscr{C}_B$. 从而,对于每一集合 $A\in\sigma(\mathscr{E})$,有

$$A\cap B\in\sigma(\mathscr{E}\cap B).$$

于是 $\sigma(\mathscr{E})\cap B\subseteq\sigma(\mathscr{E}\cap B)$. □

现在证明 σ- 代数 $\mathscr{B}(R^n)$ 与 $\mathscr{B}(R)\otimes\cdots\otimes\mathscr{B}(R)$ 重合 (相等). 对于 $n=1$,二者显然重合. 现在证明,对于 $n=2$,二者重合.

因为 $\mathscr{B}(R^2)\subseteq\mathscr{B}(R)\otimes\mathscr{B}(R)$,只需证明,任意 $B_1\times B_2$ 属于 $\mathscr{B}(R^2)$.

设 $R^2=R_1\times R_2$,其中 R_1 和 R_2 分别为 "第一条" 和 "第二条" 实数轴; $\tilde{\mathscr{B}}_1=\mathscr{B}_1\times R_2$, $\tilde{\mathscr{B}}_2=R_1\times\mathscr{B}_2$,其中 $\mathscr{B}_1\times R_2(R_1\times\mathscr{B}_2)$ 是形如 $B_1\times R_2(R_1\times B_2)$ 集合的全体,其中 $B_1\in\mathscr{B}_1(B_2\in\mathscr{B}_2)$. 假设 $\mathscr{I}_1$ 和 $\mathscr{I}_2$ 是 R_1 和 R_2 中的区间的全体,而 $\tilde{\mathscr{I}}_1=\mathscr{I}_1\times R_2$, $\tilde{\mathscr{I}}_2=R_1\times\mathscr{I}_2$. 那么,如果 $\tilde{B}_1=B_1\times R_2$, $\tilde{B}_2=R_1\times B_2$,则由 (6) 式,有

$$\begin{aligned}B_1\times B_2=\tilde{B}_1\cap\tilde{B}_2&\in\tilde{\mathscr{B}}_1\cap\tilde{B}_2=\sigma(\tilde{\mathscr{I}}_1)\cap\tilde{B}_2\\&=\sigma(\tilde{\mathscr{I}}_1\cap\tilde{B}_2)\subseteq\sigma(\tilde{\mathscr{I}}_1\cap\tilde{\mathscr{I}}_2)=\sigma(\mathscr{I}_1\times\mathscr{I}_2).\end{aligned}$$

而这正是要证明的.

任意 $n>2$ 的情形用类似的方法证明. □

注 设 $\mathscr{B}_0(R^n)$ 是由不相交开 "球" 集有限个 "形如 $S_\rho(x^0)$ 的之和" 诱导的最小 σ− 代数,其中对于 $\rho>0, x^0\in R^n$,

$$S_\rho(x^0)=\{x\in R^n:\rho_n(x,x^0)<\rho\},$$

而度量为

$$\rho_n(x,x^0)=\sum_{k=1}^{n}2^{-k}\rho_1(x_k,x_k^0),$$

其中 $x=(x_1,\cdots,x_n)$, $x^0=(x_1^0,\cdots,x_n^0)$.

于是,$\mathscr{B}_0(R^n)=\mathscr{B}(R^n)$ (练习题 7).

4. 可测空间 $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty))$ 空间 $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty))$ 是建立具有无限步概率模型的基础, 因此它在概率论中有重要意义.

空间 $\mathbb{R}^\infty$ 是有序数列

$$x=(x_1, x_2, \cdots), \quad -\infty < x_k < \infty, k=1,2,\cdots$$

集合. 记 I_k 和 B_k 相应为区间 $(a_k, b_k]$ 和 (坐标为 x_k 的) 第 k 条直线的博雷尔集. 考虑柱集:

$$\mathscr{I}(I_1\times\cdots\times I_n)=\{x: x=(x_1, x_2, \cdots), x_1\in I_1, \cdots, x_n\in I_n\}, \tag{8}$$

$$\mathscr{I}(B_1\times\cdots\times B_n)=\{x: x=(x_1, x_2, \cdots), x_1\in B_1, \cdots, x_n\in B_n\}, \tag{9}$$

$$\mathscr{I}(B^n)=\{x: x=(x_1, x_2, \cdots, x_n)\in B^n\}, \tag{10}$$

其中 B^n 是 $\mathscr{B}(\mathbb{R}^n)$ 中的博雷尔集. 柱集中的每一个"柱" $\mathscr{I}(B_1\times\cdots\times B_n)$ 或 $\mathscr{I}(B_n)$, 可以视为以 $\mathbb{R}^{n+1}, \mathbb{R}^{n+2}, \cdots$, 为底的柱, 因为

$$\mathscr{I}(B_1\times\cdots\times B_n)=\mathscr{I}(B_1\times\cdots\times B_n\times\mathbb{R}),$$
$$\mathscr{I}(B^n)=\mathscr{I}(B^{n+1}),$$

其中 $B^{n+1}=B^n\times\mathbb{R}$.

不相交柱集 $\mathscr{I}(I_1\times\cdots\times I_n)$ 的有限和构成的集合是代数. 同样, 不相交柱集 $\mathscr{I}(B_1\times\cdots\times B_n)$ 的并构成的集合也是代数. 柱集系 $\mathscr{I}(B^n)$ 也是代数. 记 $\mathscr{B}(\mathbb{R}^\infty)$, $\mathscr{B}_1(\mathbb{R}^\infty)$ 和 $\mathscr{B}_2(\mathbb{R}^\infty)$ 相应为包含 (8), (9) 和 (10) 式中一切集合的最小 σ- 代数 (σ 代数 $\mathscr{B}_1(\mathbb{R}^\infty)$ 常表示为 $\mathscr{B}(\mathbb{R})\otimes\mathscr{B}(\mathbb{R})\otimes\cdots$). 显然

$$\mathscr{B}(\mathbb{R}^\infty)\subseteq\mathscr{B}_1(\mathbb{R}^\infty)\subseteq\mathscr{B}_2(\mathbb{R}^\infty).$$

实际上, 这三个 σ- 代数重合.

为证明对于每一个 $n=1,2,\cdots$ 记

$$\mathscr{C}_n=\{A\subseteq\mathbb{R}^n: \{x=(x_1, \cdots, x_n)\in A\}\in\mathscr{B}(\mathbb{R}^\infty)\}.$$

设 $B^n\in\mathscr{B}(\mathbb{R}^n)$, 那么

$$B^n\in\mathscr{C}_n.$$

由于 $\mathscr{C}_n$ 是 σ- 代数, 故

$$\mathscr{B}(\mathbb{R}^n)\subseteq\sigma(\mathscr{C}_n)=\mathscr{C}_n.$$

由此可见 $\mathscr{B}_2(\mathbb{R}^\infty)\subset\mathscr{B}(\mathbb{R}^\infty)$.

于是, $\mathscr{B}(\mathbb{R}^\infty)=\mathscr{B}_1(\mathbb{R}^\infty)=\mathscr{B}_2(\mathbb{R}^\infty)$.

以后, 我们称 $\mathscr{B}(\mathbb{R}^\infty)$ 中的集合为 $(\mathbb{R}^\infty)$ 中的博雷尔集.

注 设 $\mathscr{B}_0(\mathbb{R}^\infty)$ 是由不相交开"球"集有限个"形如 $S_\rho(x^0)$ 的之和"诱导的最小 σ- 代数, 其中对于 $\rho>0, x^0\in\mathbb{R}^\infty$,

$$S_\rho(x^0)=\{x\in\mathbb{R}^\infty: \rho_\infty(x, x^0)<\rho\},$$

而度量为

$$\rho_\infty(x, x^0)=\sum_{k=1}^{\infty}2^{-k}\rho_1(x_k, x_k^0),$$

其中 $x=(x_1, x_2, \cdots)$, $x^0=(x_1^0, x_2^0, \cdots)$. 于是, $\mathscr{B}(\mathbb{R}^\infty)=\mathscr{B}_0(\mathbb{R}^\infty)$ (练习题 7).

举几个 $\mathbb{R}^\infty$ 中的博雷尔集的例子.

(a) $\{x\in\mathbb{R}^\infty: \sup x_n > a\}$, $\{x\in\mathbb{R}^\infty: \inf x_n < a\}$;

(b) $\{x\in\mathbb{R}^\infty: \overline{\lim}\, x_n\leqslant a\}$, $\{x\in\mathbb{R}^\infty: \underline{\lim}\, x_n > a\}$, 其中

$$\overline{\lim}\, x_n=\inf_n\sup_{m\geqslant n}x_m, \quad \underline{\lim}\, x_n=\sup_n\inf_{m\geqslant n}x_m;$$

(c) $\{x\in\mathbb{R}^\infty: x_n\to\}$ —— 使极限 $\lim x_n$ 存在并有限的 $x\in\mathbb{R}^\infty$ 的集合;

(d) $\{x\in\mathbb{R}^\infty: \lim x_n > a\}$;

(e) $\left\{x\in\mathbb{R}^\infty: \sum\limits_{n=1}^{\infty}|x_n|>a\right\}$;

(f) $\left\{x\in\mathbb{R}^\infty: 至少对于一个\ n\geqslant 1, \sum\limits_{k=1}^{n}x_k=0\right\}$.

例如, 为验证 (a) 中的集合属于 $\mathscr{B}(\mathbb{R}^\infty)$, 只需注意到

$$\{x: \sup x_n > a\}=\bigcup_n\{x: x_n > a\}\in\mathscr{B}(\mathbb{R}^\infty),$$

$$\{x: \inf x_n < a\}=\bigcup_n\{x: x_n < a\}\in\mathscr{B}(\mathbb{R}^\infty).$$

5. 可测空间 $(\mathbb{R}^T, \mathscr{B}(\mathbb{R}^T))$ 其中 T 是任意集合, 而空间 $\mathbb{R}^T$ 是实函数 $x=(x_t)$, $t\in T$, 的全体*). 我们主要考虑 T 是数轴上不可数子集的情形. 为简便和确定计, 现在可以假设 $T=[0,\infty)$.

考虑如下三种类型的柱集.

$$\mathscr{I}_{t_1,\cdots,t_n}(I_1\times\cdots\times I_n)=\{x: x_{t_1}\in I_1, \cdots, x_{t_n}\in I_n\}, \tag{11}$$

$$\mathscr{I}_{t_1,\cdots,t_n}(B_1\times\cdots\times B_n)=\{x: x_{t_1}\in B_1, \cdots, x_{t_n}\in B_n\}, \tag{12}$$

$$\mathscr{I}_{t_1,\cdots,t_n}(B^n)=\{x: (x_{t_1}, \cdots, x_{t_n})\in B^n\}, \tag{13}$$

其中 I_k 是形如 $(a_k, b_k]$ 的集合, B_k 是数轴上的博雷尔集, 而 B^n 是 $\mathbb{R}^n$ 中的博雷尔集.

*) 对于 $\mathbb{R}^T$ 中的函数, 以后还使用下面的记号: $x=(x_t)_{t\in T}$, $x=\{x_t\}$, $t\in T$.

集合 $\mathscr{I}_{t_1,\cdots,t_n}(I_1\times\cdots\times I_n)$ 是如下函数的全体: 在时刻 $t_1,\cdots,t_n$ 函数 "经过窗口" $I_1,\cdots,I_n$, 而在其余时刻取任意值 (图 24).

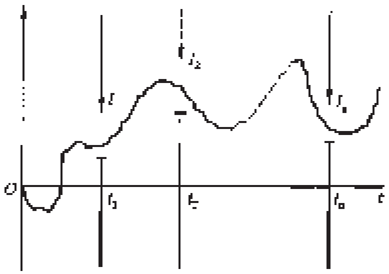


图 24

记 $\mathscr{B}(\mathbb{R}^T)$, $\mathscr{B}_1(\mathbb{R}^T)$ 和 $\mathscr{B}_2(\mathbb{R}^T)$ 为相应地包含柱集 (11), (12) 和 (13) 式的最小 $\sigma-$ 代数. 显然

$$\mathscr{B}(\mathbb{R}^T)\subseteq\mathscr{B}_1(\mathbb{R}^T)\subseteq\mathscr{B}_2(\mathbb{R}^T). \tag{14}$$

然而, 实际上三个 $\sigma-$ 代数重合, 而且可以完全地描绘这些集合的结构.

定理 4　*设 T 是不可数集合, 则 $\mathscr{B}(\mathbb{R}^T)=\mathscr{B}_1(\mathbb{R}^T)=\mathscr{B}_2(\mathbb{R}^T)$, 且任意集合 $A\in\mathscr{B}(\mathbb{R}^T)$ 有如下结构: 在 T 中存在最多可数个点 $t_1,t_2,\cdots$, 和博雷尔集 $B\in\mathscr{B}(\mathbb{R}^\infty)$, 使*

$$A=\{x: x=(x_{t_1},x_{t_2},\cdots)\in B\}. \tag{15}$$

证明　设 $\mathscr{E}$ 是形如 (15) 式的集合的全体 (对于不同的数组 $(t_1,t_2,\cdots)$ 和 $B\in\mathscr{B}(\mathbb{R}^\infty)$). 若对应于 $A_1,A_2,\cdots\in\mathscr{E}$ 的数组为 $T^{(1)}=(t_1^{(1)},t_2^{(1)},\cdots)$, $T^{(2)}=(t_1^{(2)},t_2^{(2)},\cdots)$, 则集合 $T^{(\infty)}=\bigcup_k T^{(k)}$ 可以取作一个统一集系: 一切 A_i 表示为

$$A_i=\{x: x=(x_{\tau_1},x_{\tau_2},\cdots)\in B_i\},$$

其中 B_i 是 (同一) $\sigma-$ 代数 $\mathscr{B}(\mathbb{R}^\infty)$ 的集合, 而 $\tau_i\in T^{(\infty)}$.

由此可见, 集系 $\mathscr{E}$ 是 $\sigma-$ 代数. 显然, 此 $\sigma-$ 代数包含一切形如 (13) 式的柱集, 而且由于 $\mathscr{B}_2(\mathbb{R}^T)$ 是包含这些集合的最小 $\sigma-$ 代数, 则由此并连同 (14) 式, 得

$$\mathscr{B}(\mathbb{R}^T)\subseteq\mathscr{B}_1(\mathbb{R}^T)\subseteq\mathscr{B}_2(\mathbb{R}^T)\subseteq\mathscr{E}. \tag{16}$$

考虑 $\mathscr{E}$ 中表示为 (15) 式的集合 A. 如果固定 $(t_1,t_2,\cdots)$, 则用与空间 $(\mathbb{R}^\infty,\mathscr{B}(\mathbb{R}^\infty))$ 情形同样的论述, 可见集合 A 是由 (11) 式的柱集生成的 $\sigma-$ 代数的元素. 然而, 这一 $\sigma-$ 代数显然属于 $\sigma-$ 代数 $\mathscr{B}(\mathbb{R}^T)$. 于是连同 (16) 式就可以证明定理 2 的两个命题.

这样, $\sigma-$ 代数 $\mathscr{B}(\mathbb{R}^T)$ 中任何博雷尔集 A, 决定于 (至少在可数个点 $t_1,t_2,\cdots$ 上) 加在函数上 $x(t),t\in T$, 的约束. 特别, 由此可见, 依赖于函数在不可数点上的 "性质" 的集合

$$A_1=\{x: \text{对于一切 } t\in[0,1],\ x_t<C\},$$

$$A_2=\{x: \text{至少对于一个 } t\in[0,1],\ x_t=0\},$$

$$A_3=\{x: \text{在固定的点 } t_0\in[0,1],\ x_t \text{ 连续}\},$$

未必是博雷尔集. 而所提到的三个集合确实不属于 $\mathscr{B}(\mathbb{R}^{[0,1]})$.

现在证明这对于 A_1 成立. 假如 $A_1\in\mathscr{B}(\mathbb{R}^{[0,1]})$, 那么根据定理 4, 存在这样的点 $(t_1^0,t_2^0,\cdots)$ 和集合 $B^0\in\mathscr{B}(\mathbb{R}^\infty)$, 使

$$\left\{x: \sup_t x_t<C, t\in[0,1]\right\}=\{x: x=(x_{t_1^0},x_{t_2^0},\cdots)\in B^0\}.$$

显然, 函数 $y_t=C-1$ 属于 A_1, 从而 $(y_{t_1^0},y_{t_2^0},\cdots)\in B^0$. 定义函数

$$z_t=\begin{cases}C-1, & \text{若 } t\in(t_1^0,t_2^0,\cdots),\\ C+1, & \text{若 } t\notin(t_1^0,t_2^0,\cdots).\end{cases}$$

显然,

$$(y_{t_1^0},y_{t_2^0},\cdots)=(z_{t_1^0},z_{t_2^0},\cdots),$$

从而, 函数 $z=(z_t)$ 属于集合

$$\{x: (x_{t_1^0},x_{t_2^0},\cdots)\in B^0\}.$$

但是 $z=(z_t)$ 明显不属于集合 $\{x: \sup x_t<C\}$. 所得矛盾说明 $A_1\notin\mathscr{B}(\mathbb{R}^{[0,1]})$.

由于关于一切函数 $x=(x_t),t\in[0,1]$ 空间中的 $\sigma-$ 代数 $\mathscr{B}(\mathbb{R}^{[0,1]})$, 集合 A_1,A_2 和 A_3 的不可测性, 自然地应考虑较窄的函数类, 以便使这些集合成为可测的. 直观上明显, 作为这样的空间, 例如可以考虑连续函数的空间.

6. 可测空间 $(C,\mathscr{B}(C))$.　设 $T=[0,1]$, 而 C 是连续函数 $x=(x_t),0\leqslant t\leqslant 1$ 的空间. 该空间关于均匀度量

$$\rho(x,y)=\sup_{t\in T}|x_t-y_t|$$

是可测空间. 在空间 C 中可以引进两个 $\sigma-$ 代数: $\mathscr{B}(C)$ 是由柱集生成的 $\sigma-$ 代数, 而 $\mathscr{B}_0(C)$ 是由 (关于度量 $\rho(x,y)$ 的) 开集生成的 $\sigma-$ 代数. 现在证明, 两个 $\sigma-$ 代数实际上重合 $\mathscr{B}(C)=\mathscr{B}_0(C)$.

设 $B=\{x: x_{t_0}<b\}$ 是某一柱集. 易见, 此集合是开集. 由此可见

$$\{x: x_{t_1}<b_1,\cdots,x_{t_n}<b_n\}\in\mathscr{B}_0(C), \text{ 即 } \mathscr{B}(C)\subseteq\mathscr{B}_0(C).$$

相反，考虑集合 $B_\rho = \{y: y \in S_\rho(x^0)\}$，其中 x^0 是 C 中的某一函数，而

$$S_\rho(x^0) = \{x \in C: \sup_{t \in T} |x_t - x_t^0| < \rho\}$$

是以 x^0 为球心的开球。由 C 中函数的连续性，可见

$$\begin{aligned} B_\rho &= \{y \in C: y \in S_\rho(x^0)\} = \{y \in C: \max_t |y_t - x_t^0| < \rho\} \\ &= \bigcap_{t_k} \{y \in C: |y_{t_k} - x_{t_k}^0| < \rho\} \in \mathscr{B}(C), \end{aligned} \tag{17}$$

其中 t_k 是区间 $[0,1]$ 上的有理点。于是 $\mathscr{B}_0(C) \subseteq \mathscr{B}(C)$。

空间 $(C, \mathscr{B}_0(C), \rho)$ 是完备的而且是可分的；[5]，[57]。

7. 可测空间 $(D, \mathscr{B}(D))$ 设 D 是一函数 $x = (x_t), 0 \leqslant t \leqslant 1$，的空间，其中每个函数右连续 $x_t = (x_{t+}), t < 1$，而且对于任意 $t > 0$，有左极限。

像空间 C 一样，在空间 D 上可以定义度量 $d = d(x,y)$，使 $\mathscr{B}_0(D) = \mathscr{B}(D)$，其中 $\mathscr{B}_0(D)$ 是由开集生成的 $\sigma-$ 代数，$\mathscr{B}(D)$ 是由柱集生成的 $\sigma-$ 代数。这时 $(D, \mathscr{B}(D), d)$ 是可分空间；[5]，[57]。关于均匀度量是斯科罗霍德 (A. B. Скороход) 引进的，定义为：

$$d(x,y) = \inf\left\{\varepsilon > 0: \exists \lambda \in \Lambda: \sup_t |x_t - y_{\lambda(t)}| + \sup_t |t - \lambda(t)| \leqslant \varepsilon\right\}, \tag{18}$$

其中 Λ 是 $[0,1]$ 上的严格递增连续函数 $\lambda(t)$，且 $\lambda(0) = 0, \lambda(1) = 1$ 的集合。

8. 可测空间 $\left(\prod_{t \in T} \Omega_t, \boxtimes_{t \in T} \mathscr{F}_t\right)$ 设空间 $(\mathbb{R}^T, \mathscr{B}(\mathbb{R}^T))$ 是 T "份" 数轴，连同其博雷尔集系。在概率论中，除 $(\mathbb{R}^T, \mathscr{B}(\mathbb{R}^T))$ 外，还考虑用如下方式构成的可测空间

$$\left(\prod_{t \in T} \Omega_t, \boxtimes_{t \in T} \mathscr{F}_t\right).$$

设 T 是下标的集合，$(\Omega_t, \mathscr{F}_t)$ 是可测空间 $t \in T$。记

$$\Omega = \prod_{t \in T} \Omega_t$$

为一切函数 $\omega = (\omega_t), t \in T$，的集合；对于每个 $t \in T, \omega_t \in \Omega_t$。

不难验证，一切有限个不相交柱集的并

$$\mathscr{F}_{t_1, \cdots, t_n}(B_1 \times \cdots \times B_n) = \{\omega: \omega_{t_1} \in B_1, \cdots, \omega_{t_n} \in B_n\}, B_{t_i} \in \mathscr{F}_{t_i}$$

构成代数。以 $\boxtimes_{t \in T} \mathscr{F}_t$ 表示包含一切柱集的最小 $\sigma-$ 代数，而可测空间 $\left(\prod \Omega_t, \boxtimes \mathscr{F}_t\right)$ 是可测空间 $(\Omega_t, \mathscr{F}_t)$ 的直积。

9. 练习题

1. 设 $\mathscr{B}_1$ 和 $\mathscr{B}_2$ 是空间 Ω 的子集的 $\sigma-$ 代数。问集系

$$\mathscr{B}_1 \cap \mathscr{B}_2 = \{A: A \in \mathscr{B}_1 \text{ 和 } A \in \mathscr{B}_2\},$$
$$\mathscr{B}_1 \cup \mathscr{B}_2 = \{A: A \in \mathscr{B}_1 \text{ 或 } A \in \mathscr{B}_2\}$$

是否 $\sigma-$ 代数？

2. 设 $\mathscr{D} = \{D_1, D_2, \cdots\}$ 是 Ω 的某一可数分割，而 $\mathscr{B} = \sigma(\mathscr{D})$。问 $\sigma-$ 代数 $\mathscr{B}$ 的势如何？

3. 证明

$$\mathscr{B}(\mathbb{R}^n) \otimes \mathscr{B}(\mathbb{R}) = \mathscr{B}(\mathbb{R}^{n+1}).$$

4. 证明第 4 小节的集合 (b)~(f) 属于 $\mathscr{B}(\mathbb{R}^\infty)$。

5. 证明第 5 小节的集合 A_2 和 A_3 不属于 $\mathscr{B}(\mathbb{R}^{[0,1]})$。

6. 证明函数 (18) 确实是度量。

7. 证明 $\mathscr{B}_0(\mathbb{R}^n) = \mathscr{B}(\mathbb{R}^n), n \geqslant 1$，和 $\mathscr{B}_0(\mathbb{R}^\infty) = \mathscr{B}(\mathbb{R}^\infty)$。

8. 设 $C = C[0, \infty)$ 是定义连续函数 $x = (x_t), t \geqslant 0$ 的空间。证明，像 $C = C[0,1]$ 的情形一样，该空间关于度量

$$\rho(x,y) = \sum_{n=1}^{\infty} 2^{-n} \min\left[\sup_{0 \leqslant t \leqslant n} |x_t - y_t|, 1\right], x, y \in C,$$

是完备可分度量空间，而 $\mathscr{B}_0(C) = \mathscr{B}(C)$，其中 $\mathscr{B}_0(C)$ 是由开集生成的 $\sigma-$ 代数，$\mathscr{B}(C)$ 是由柱集生成的 $\sigma-$ 代数。

9. 证明，(定义 2 的) 条件 $(\lambda_a), (\lambda_b), (\lambda_c)$，与 (注 2 中的) 条件 $(\lambda_a), (\lambda_b'), (\lambda_c')$ 等价。

10. 由定理 1 证明定理 2。

11. 证明，定理 3 中的 $\mathscr{D}$ 系是 $\lambda-$ 系。

12. 称 $\sigma-$ 代数为可数 − 生成的或可分的，如果它有某个可数集类生成。

(a) 证明，空间 $\Omega = (0,1]$ 中博雷尔子集的 $\sigma-$ 代数 $\mathscr{B}$ 是可数 − 生成的。

(b) 证明，(例如) 如下的情形是可能的：两个 $\sigma-$ 代数 $\mathscr{F}_1$ 和 $\mathscr{F}_2$，且 $\mathscr{F}_1 \subset \mathscr{F}_2$，但 $\mathscr{F}_2$ 可数生成，而 $\mathscr{F}_1$ 不是可数生成的。

13. 证明，$\sigma-$ 代数 $\mathscr{G}$ 是可数 − 生成的充分和必要条件是：对于某一随机变量 $X, \mathscr{G} = \sigma(X)$ ($\sigma(X)$ 的定义见 §4 第 4 小节)。

14. 举例说明可分 $\sigma-$ 代数，其相应的子 $\sigma-$ 代数是不可分的。

15. 证明 $X_1, X_2, \cdots$ 是独立随机变量序列 §4, §5，如果对于任意 $n \geqslant 1, \sigma(X_n)$ 和 $\sigma(X_1, \cdots, X_{n-1})$ 相互独立。

16. 举例说明两个 $\sigma-$ 代数之并，不是 $\sigma-$ 代数。

17. 设 $\mathscr{A}_1$ 和 $\mathscr{A}_2$ 是两个独立的集系, 且每一个都是 $\pi-$ 系. 证明 $\sigma(\mathscr{A}_1)$ 和 $\sigma(\mathscr{A}_2)$ 也相互独立. 举例说明, 虽然两个不是 $\pi-$ 系的集系 $\mathscr{A}_1$ 和 $\mathscr{A}_2$ 独立, 但是 $\sigma(\mathscr{A}_1)$ 和 $\sigma(\mathscr{A}_2)$ 不独立.

18. 设 $\mathscr{L}$ 是 $\lambda-$ 系, 则

$$\{A, B \in \mathscr{L}, A \cap B = \varnothing\} \Rightarrow \{A \cup B \in \mathscr{L}\}.$$

19. 设 $\mathscr{F}_1$ 和 $\mathscr{F}_2$ 是在 Ω 上的子集的两个 $\sigma-$ 代数. 记

$$d(\mathscr{F}_1, \mathscr{F}_2) = 4 \sup_{\substack{A_1 \in \mathscr{F}_1 \\ A_2 \in \mathscr{F}_2}} |\mathbf{P}(A_1A_2) - \mathbf{P}(A_1)\mathbf{P}(A_2)|.$$

证明, $d(\mathscr{F}_1, \mathscr{F}_2)$ 表征 $\mathscr{F}_1$ 和 $\mathscr{F}_2$ 的相依程度, 且具有如下性质:

(a) $0 \leqslant d(\mathscr{F}_1, \mathscr{F}_2) \leqslant 1$;

(b) 若 $\mathscr{F}_1$ 和 $\mathscr{F}_2$ 独立, 则 $d(\mathscr{F}_1, \mathscr{F}_2) = 0$;

(c) $d(\mathscr{F}_1, \mathscr{F}_2) = 1$, 当且仅当 $\mathscr{F}_1$ 和 $\mathscr{F}_2$ 的交含概率为 1/2 的事件.

20. 利用引理 1 的证明方法, 证明存在唯一含 $\mathscr{E}$ 集系 $\lambda-$ 系 $\lambda(\mathscr{E})$ 和 $\pi-$ 系 $\pi(\mathscr{E})$.

21. 设 $\mathscr{A}$ 是一集合代数: 对于任意不相交的集合序列 $(A_n)_{n\geqslant 1}$, $A_n \in \mathscr{A}$, 且

$$\bigcup_{n=1}^{\infty} A_n \in \mathscr{A},$$

证明 $\mathscr{A}$ 是 $\sigma-$ 代数.

22. 设 $(\mathscr{F}_n)_{n\geqslant 1}$ 是递增 $\sigma-$ 代数序列: $\mathscr{F}_n \subseteq \mathscr{F}_{n+1}$, $n \geqslant 1$. 证明 $\bigcup_{n=1}^{\infty} \mathscr{F}_n$ 是 (一般刚好是) 代数.

23. 设 $\mathscr{F}$ 是代数 (或 $\sigma-$ 代数), 而 C 是 $\mathscr{F}$ 中某一集合. 考虑由 $\mathscr{F} \bigcup \{C\}$ 生成的最小代数 (相应地 $\sigma-$ 代数). 证明这一代数 (相应地 $\sigma-$ 代数) 的一切元素具有集合 $(A \bigcap C) \bigcup (B \bigcap \bar{C})$ 的形式, 其中 $A, B \in \mathscr{F}$.

24. 设 $\bar{\mathbf{R}} = R \bigcup \{-\infty\} \bigcup \{\infty\}$ 是扩展的数轴. 博雷尔 $\sigma-$ 代数 $\mathscr{B}(\bar{\mathbf{R}})$ 定义为 (见第 2 小节): 由集合 $[-\infty, x]$, $x \in \mathbf{R}$, 生成的 $\sigma-$ 代数, 其中 $[-\infty, x] = \{-\infty\} \bigcup (-\infty, x]$. 证明这一 $\sigma-$ 代数 $\mathscr{B}(\bar{\mathbf{R}})$ 与由集合生成的任何 $\sigma-$ 代数重合:

(a) $[-\infty, x)$, $x \in \mathbf{R}$, 或

(b) $(x, \infty]$, $x \in \mathbf{R}$, 或

(c) 一切有限区间 $\{-\infty\}$ 和与 $\{\infty\}$.

§3. 在可测空间上建立概率测度的方法

1. 可测空间 $(\mathbf{R}, \mathscr{B}(\mathbf{R}))$　设 $\mathbf{P}(A)$ 是定义在数轴的博雷尔集上的概率测度. 取 $A = (-\infty, x]$ 并设

$$F(x) = \mathbf{P}(-\infty, x], \quad x \in \mathbf{R}. \tag{1}$$

这样定义的函数具有下列性质:

1) $F(x)$ 是非减函数;

2) $F(-\infty) = 0$, $F(+\infty) = 1$, 其中

$$F(-\infty) = \lim_{x\downarrow -\infty} F(x), \quad F(+\infty) = \lim_{x\uparrow\infty} F(x);$$

3) $F(x)$ 在每一点 $x \in \mathbf{R}$ 右连续且有左极限.

第一条性质显然, 而另两条性质可以由概率测度的连续性得到.

定义 1　满足上述性质 1)~3) 的任意函数 $F(x)$, 称做 (数轴 $\mathbf{R}$ 上) 的*分布函数*.

于是, 与 $(\mathbf{R}, \mathscr{B}(\mathbf{R}))$ 上的每一个概率测度 $\mathbf{P}$, (由于 (1) 式) 有一个分布函数与之相对应. 逆命题仍然成立.

定理 1　*设 $F = F(x)$ 是数轴 $\mathbf{R}$ 上一分布函数. 那么, 在 $(\mathbf{R}, \mathscr{B}(\mathbf{R}))$ 上存在唯一一个概率测度 $\mathbf{P}$, 使对于任意 $-\infty \leqslant a < b < \infty$, 有*

$$\mathbf{P}(a, b] = F(b) - F(a). \tag{2}$$

证明　设 $\mathscr{A}$ 是集合 $A \subset \mathbf{R}$ 的代数, 其中每一个集合是形如 $(a, b]$ 的有限个不相交区间的和:

$$A = \sum_{k=1}^{n} (a_k, b_k].$$

在这些集合上定义一集函数 $\mathbf{P}_0$:

$$\mathbf{P}_0(A) = \sum_{k=1}^{n} [F(b_k) - F(a_k)], \quad A \in \mathscr{A}. \tag{3}$$

此式在代数 $\mathscr{A}$ 上, 由 F 唯一确定一有限 − 可加集函数. 因此, 如果可以证明在代数 $\mathscr{A}$ 上函数 F 也是可数 − 可加的, 则概率测度 $\mathbf{P}$ 在 $\mathscr{B}(\mathbf{R})$ 上存在性和唯一性, 可以直接由测度论的一般结果得到 (这里不加证明而直接引用, 例如, 其证明可以参见 [42], [70]).

卡拉泰奥多里 (C. Carathéodory) 定理　*设 Ω 是某一空间, $\mathscr{A}$ 是其子集的代数, 而 $\mathscr{B} = \sigma(\mathscr{A})$ 是含 $\mathscr{A}$ 的最小 $\sigma-$ 代数. 记 μ_0 为 $(\Omega, \mathscr{A})$ 上的 $\sigma-$ 有限测度 (即 $\sigma-$ 可加集函数). 那么, 在 $(\Omega, \mathscr{B})$ 上存在且唯一测度 μ, 是 μ_0 的开拓, 即*

$$\mu(A) = \mu_0(A), \quad A \in \mathscr{A}.$$

这样, 我们现在证明, 函数 $\mathbf{P}_0$ 在 $\mathscr{A}$ 上可数 − 可加 (即是概率测度). 根据 §1 的定理 1, 为此只需验证 $\mathbf{P}_0$ 在 $\varnothing$ 连续, 即

$$\mathbf{P}(A_n) \downarrow 0, \quad A_n \downarrow \varnothing, \quad A_n \in \mathscr{A}.$$

设 $A_1, A_2, \cdots$ 是自 $\mathscr{A}$ 选出的一集合序列, 且 $A_n \downarrow \varnothing$. 首先, 假设 A_n 属于某一闭区间 $[-N, N]$, $N < \infty$. 因为 A_n 由形如 $(a, b]$ 的有限个区间的和构成, 且由于函数 $F(x)$ 的右连续性: 当 $a' \downarrow a$ 时

$$\mathbf{P}_0(a', b] = F(b) - F(a') \to F(b) - F(a) = \mathbf{P}_0(a, b],$$

可见对于每一个 A_n 存在一个集合 $B_n \in \mathscr{A}$, 使其闭包 $[B_n] \subseteq A_n$ 且

$$\mathbf{P}_0(A_n) - \mathbf{P}_0(B_n) \leqslant \varepsilon 2^{-n},$$

其中 ε 是某一事先给定的正数.

根据假设 $\bigcap A_n = \varnothing$, 因而 $\bigcap [B_n] = \varnothing$. 因为 $[B_n]$ 是闭集, 所以存在 $n_0 = n_0(\varepsilon)$, 使

$$\bigcap_{n=1}^{n_0} [B_n] = \varnothing. \tag{4}$$

实际上, $[-N, N]$ 是紧统, 而集系 $\{[-N, N] \setminus [B_n]\}_{n \geqslant 1}$ 是该紧统的开覆盖. 那么, 根据海涅–博雷尔 [H. E. Heine-E. Borel] 引理 (例如, 参见 [1], [33]), 存在有限子覆盖:

$$\bigcup_{n=1}^{n_0} ([-N, N] \setminus [B_n]) = [-N, N].$$

即

$$\bigcap_{n=1}^{n_0} [B_n] = \varnothing.$$

考虑到 (4) 式, 以及 $A_{n_0} \subseteq A_{n_0-1} \subseteq \cdots \subseteq A_1$, 有

$$\mathbf{P}_0(A_{n_0}) = \mathbf{P}_0\left(A_{n_0} \setminus \bigcap_{k=1}^{n_0} B_k\right) + \mathbf{P}_0\left(\bigcap_{k=1}^{n_0} B_k\right) = \mathbf{P}_0\left(A_{n_0} \setminus \bigcap_{k=1}^{n_0} B_k\right)$$
$$\leqslant \mathbf{P}_0\left(\bigcup_{k=1}^{n_0} (A_k \setminus B_k)\right) \leqslant \sum_{k=1}^{n_0} \mathbf{P}_0(A_k \setminus B_k) \leqslant \sum_{k=1}^{n_0} \varepsilon 2^{-k} \leqslant \varepsilon.$$

因此 $\mathbf{P}_0(A_n) \downarrow 0, n \to \infty$.

现在去掉条件: 对于某个 $N, A_n \subset [-N, N]$. 对于给定的 $\varepsilon > 0$, 选择 N, 使 $\mathbf{P}_0[-N, N] > 1 - \varepsilon/2$. 那么, 由于

$$A_n = A_n \cap [-N, N] + A_n \cap \overline{[-N, N]},$$

可见

$$\mathbf{P}_0(A_n) = \mathbf{P}_0(A_n \cap [-N, N]) + \mathbf{P}_0(A_n \cap \overline{[-N, N]})$$
$$\leqslant \mathbf{P}_0(A_n \cap [-N, N]) + \varepsilon/2,$$

由上一段的结果 (只是将其中的 A_n 换成 $A_n \bigcap [-N, N]$) 可得, 对于充分大的 n, 有 $\mathbf{P}_0(A_n \bigcap [-N, N]) \leqslant \varepsilon/2$. 于是, 仍然得到 $\mathbf{P}_0(A_n) \downarrow 0, n \to \infty$. □

这样, 在 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上的概率测度 $\mathbf{P}$ 与数轴上 $\mathbb{R}$ 的函数 F 之间, 存在一一对应关系. 习惯上, 把由函数 F 建立的测度 $\mathbf{P}$, 称做对应于分布函数 F 的勒贝格–斯蒂尔切斯 (H. L. Lebesgue-T. J. Stieltjes) 概率测度.

特别重要的情形是, 如果

$$F(x) = \begin{cases} 0, & 若\ x < 0, \\ x, & 若\ 0 \leqslant x \leqslant 1, \\ 1, & 若\ x > 1. \end{cases}$$

这时, 相应的概率测度 (记作 λ), 称做区间 $[0, 1]$ 上的勒贝格测度. 显然, $\lambda(a, b] = b - a$, 其中 $0 \leqslant a \leqslant b \leqslant 1$. 换句话说, 区间 $(a, b]$ (以及区间 $(a, b), [a, b], [a, b)$) 上的勒贝格测度就等于区间的长度 $b - a$.

以

$$\mathscr{B}([0, 1]) = \{A \cap [0, 1] : A \in \mathscr{B}(\mathbb{R})\}$$

表示区间 $[0, 1]$ 上博雷尔集合的全体. 除博雷尔集合外, 往往需要考虑区间 $[0, 1]$ 上所谓勒贝格集合. 称集合 $\Lambda \subseteq [0, 1]$ 属于集系 $\overline{\mathscr{B}}([0, 1])$, 如果存在博雷尔集合 A, B, 使 $A \subseteq \Lambda \subseteq B$ 且 $\lambda(B \setminus A) = 0$. 不难验证, $\overline{\mathscr{B}}([0, 1])$ 是 $\sigma-$ 代数. 正是把 $\overline{\mathscr{B}}([0, 1])$ 称做区间 $[0, 1]$ 上勒贝格集合的集系. 显然 $\mathscr{B}([0, 1]) \subseteq \overline{\mathscr{B}}([0, 1])$.

暂时仅定义在 $\mathscr{B}([0, 1])$ 中集合上的测度 λ, 可以自然地开拓到勒贝格集系 $\overline{\mathscr{B}}([0, 1])$ 上. 具体地说, 如果 $\Lambda \in \overline{\mathscr{B}}([0, 1])$ 且 $A \subseteq \Lambda \subseteq B$, 其中集系 $A, B \in \mathscr{B}([0, 1])$ 且 $\lambda(B \setminus A) = 0$, 则设 $\overline{\lambda}(\Lambda) = \lambda(A)$. 不难验证, 这样定义的集函数 $\overline{\lambda} = \overline{\lambda}(\Lambda), \Lambda \in \overline{\mathscr{B}}([0, 1])$, 是 $([0, 1], \overline{\mathscr{B}}([0, 1]))$ 上的概率测度, 称做 (勒贝格集系上的) 勒贝格测度.

注 1 所采用的测度完备化 (开拓) 方法, 不只适用于所考虑的情形. 例如, 假设 $(\Omega, \mathscr{F}, \mathbf{P})$ 是某一概率空间. 以 $\overline{\mathscr{F}}^{\mathbf{P}}$ 表示 Ω 的一切子集 Λ 的全体: 对于其中每一个子集 Λ, 存在 $A, B \in \mathscr{F}$, 使 $A \subseteq \Lambda \subseteq B$ 且 $\mathbf{P}(B \setminus A) = 0$. 自然, (利用等式 $\mathbf{P}(\Lambda) = \mathbf{P}(B)$) 也可以为 $\Lambda \in \overline{\mathscr{F}}^{\mathbf{P}}$ 定义概率测度. 用这种方法得到的概率空间 $(\Omega, \overline{\mathscr{F}}^{\mathbf{P}}, \mathbf{P})$, 称做空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 关于测度 $\mathbf{P}$ 的完备化.

如果对于概率测度 $\mathbf{P}$, 有 $\overline{\mathscr{F}}^{\mathbf{P}} = \mathscr{F}$, 则测度 $\mathbf{P}$ 称做完备的, 而相应的空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 称做完备概率空间.

注 2 现在简单地阐述卡拉泰奥多里定理的证明思路, 假设其中 $\mu_0(\Omega) = 1$.

设 A 是 Ω 中的集合, $A_1, A_2, \cdots$ 是 $\mathscr{A}$ 中的集合, 并且覆盖集合 A : $A \subseteq \bigcup_{n=1}^{\infty} A_n$. 定义集合 A 的外测度 $\mu^*(A)$ 如下.

$$\mu^*(A) = \inf \sum_{n=1}^{\infty} \mu_0(A_n),$$

其中 inf 对集合 A 的一切上述覆盖 $(A_1, A_2, \cdots)$ 来求; 而

$$\mu_*(A) = 1 - \mu^*(\overline{A})$$

称做集合 A 的内测度 $\mu_*(A)$.

设 $\overline{\mathscr{A}}$ 是 Ω 中满足 $\mu_*(A) = \mu^*(A)$ 的集合 A 的全体. 不难证明, 集系 $\overline{\mathscr{A}}$ 是 σ-代数 (练习题 12); 因而 $\mathscr{A} \subseteq \sigma(\mathscr{A}) \subseteq \overline{\mathscr{A}}$. 赋予 $\overline{\mathscr{A}}$ 中集合 A 以 "测度" $\mu(A)$, 令其等于 $\mu^*(A)(= \mu_*(A))$. 函数 $\mu(A)$ 确实是测度 (练习题 13). 即 $\mu(A)$ 确实是可数-可加集函数 (并且是概率测度, 因为 $\mu(\Omega) = \mu_0(\Omega) = 1$).

由等式 $\mathbf{P}(a, b] = F(b) - F(a)$ 建立的, 概率测度 $\mathbf{P}$ 与分布函数 F 之间的一一对应关系, 使得可以由相应的分布函数构造各种概率测度.

离散型测度 若分布函数 $F = F(x)$ 是阶梯函数, 则对应的测度称做离散型测度. 对于离散型测度, 分布函数仅在 (有限或可数个) 点 $x_1, x_2, \cdots$ 改变其数值: $\Delta F(x_i) > 0, i = 1, 2, \cdots$, 其中 $\Delta F(x) = F(x) - F(x-)$(图 25). 在这种情况下测度集中在点 $x_1, x_2, \cdots$ 上:

$$\mathbf{P}(\{x_k\}) = \Delta F(x_k) > 0, \quad \sum_k \mathbf{P}(\{x_k\}) = 1.$$

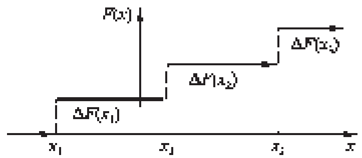


图 25

数组 $(p_1, p_2, \cdots)$, 其中 $p_k = \mathbf{P}(\{x_k\})$, 称做离散型概率分布, 而相应的分布函数 $F = F(x)$ 称做离散型的.

下表是最常见的离散概率分布类型及其名称.

表 2-2

分布名称	概率 p_k	参数
离散均匀	$\frac{1}{N}, k = 1, 2, \cdots, N$	$N = 1, 2, \cdots$
伯努利 (J. Bernoulli)	$p_1 = p, p_0 = q$	$0 \leqslant p \leqslant 1, q = 1 - p$
二项	$C_n^k p^k q^{n-k} (k = 0, 1, \cdots, n)$	$0 \leqslant p \leqslant 1, q = 1 - p \ (n = 1, 2, \cdots)$
泊松 (S. D. Poisson)	$\frac{\lambda^k}{k!} e^{-\lambda} (k = 0, 1, \cdots)$	$\lambda > 0$
几何	$pq^{k-1} (k = 1, 2, \cdots)$	$0 < p \leqslant 1, q = 1 - p$
负二项 (帕斯卡) (B. Pascal)	$C_{k-1}^{r-1} p^r q^{k-r} (k = r, r+1, \cdots)$	$0 < p \leqslant 1, q = 1 - p \ (r = 1, 2, \cdots)$

绝对连续测度 称测度为绝对连续的, 如果存在非负博雷尔函数 $f(t), t \in \mathbb{R}$, 使其分布函数 $F = F(x)$ 可以表示为:

$$F(x) = \int_{-\infty}^{x} f(t) dt, \tag{5}$$

其中的积分理解为黎曼 (G. F. B. Riemann) 积分 (而在一般情形下是勒贝格积分 (§6)).

函数 $f(t), t \in \mathbb{R}$, 称做分布函数的密度 (概率分布密度, 或简称为密度), 而分布函数 $F = F(x)$ 本身称为绝对连续的.

表 2-3

分布名称	概率 p_k	参数
$[a, b]$ 上的均匀	$\frac{1}{b-a} (a \leqslant x \leqslant b)$	$a, b \in \mathbb{R}; a < b$
正态或高斯 (C. F. Gauss)	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}, x \in \mathbb{R}$	$m \in \mathbb{R}, \sigma > 0$
伽马 (Γ)	$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, x \geqslant 0$	$\alpha > 0, \beta > 0$
贝塔 (B)	$\frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 \leqslant x \leqslant 1$	$\alpha > 0, \beta > 0$
指数 (参数为 $\alpha = 1$, $\beta = \lambda^{-1}$ 的 Γ 分布)	$\lambda e^{-\lambda x}, x \geqslant 0$	$\lambda > 0$
双侧指数	$\frac{\lambda}{2} e^{-\lambda\lvert x - \alpha\rvert}, x \in \mathbb{R}$	$\lambda > 0, \alpha \in \mathbb{R}$
χ^2 (卡方) (Γ 分布: $\alpha = 1/2, \beta = 2$)	$\frac{1}{2^{n/2}\Gamma(n/2)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, x \geqslant 0$	(自由度) $n = 1, 2, \cdots$
t (学生)	$\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, x \in \mathbb{R}$	(自由度) $n = 1, 2, \cdots$
F	$\frac{\left(\frac{m}{n}\right)^{\frac{m}{2}}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} x^{\frac{m}{2}-1} \left(1 + \frac{mx}{n}\right)^{-\frac{m+n}{2}}, x \geqslant 0$	(第一自由度) $m = 1, 2, \cdots$ (第二自由度) $n = 1, 2, \cdots$
柯西 (A. L. Cauchy)	$\frac{\theta}{\pi(x^2 + \theta^2)}, x \in \mathbb{R}$	$\theta > 0$

显然, 任何黎曼可积且在数轴上的积分为 1 即 $\int_{-\infty}^{\infty} f(x) dx = 1$ 的、非负博雷尔函数 $f = f(x), x \in \mathbb{R}$, 由 (5) 式决定 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上某一概率测度的分布函数. 表 2-3

列出了概率论与数理统计中特别重要的, 不同类型的概率密度 $f = f(x)$ 的例子, 并且指出了其名称和参数. (在表中未指明 x 值的, 认为 $f(x) = 0$.)

奇异测度 称点 x 为分布函数 $F(x)$ 的增长点, 如果对于任意 $\varepsilon > 0$, 有 $F(x+\varepsilon) - F(x-\varepsilon) > 0$. 测度称为*奇异的*, 如果其分布函数 $F(x)$ 连续, 但是在其增长点集合的勒贝格测度等于 0. 回避有关这样函数构造的细节 (例如, 可以参见 [70]), 我们仅限于举一个 "传统的" 例子.

用下面的康托尔 (G. Cantor) 方法, 构造区间 $[0,1]$ 上的分布函数 $F(x)$.

将区间 $[0,1]$ 分成三等份, 并且设 (图 25)

$$F_1(x) = \begin{cases} 1/2, & 若\ x \in (1/3, 2/3), \\ 0, & 若\ x = 0, \\ 1, & 若\ x = 1, \end{cases}$$

用线性内插法再给函数 $F_1(x)$ 补充上定义.

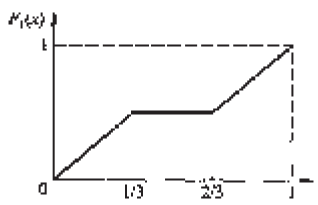


图 26

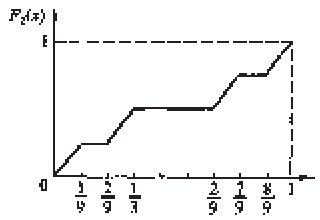


图 27

其次, 将区间 $[0, 1/3]$ 和 $[2/3, 1]$ 中每一个仍然分成三等份, 并且建立函数 (图 27):

$$F_2(x) = \begin{cases} 1/2, & 若\ x \in (1/3, 2/3), \\ 1/4, & 若\ x \in (1/9, 2/9), \\ 3/4, & 若\ x \in (7/9, 8/9), \\ 0, & 若\ x = 0, \\ 1, & 若\ x = 1, \end{cases}$$

也用线性内插法再给函数 $F_2(x)$ 补充上定义.

重复这一过程, 将建成函数序列 $F_n(x), n = 1, 2, \cdots$. 函数序列 $\{F_n(x)\}$ 收敛于某非减连续函数 $F(x)$ (称为*康托尔函数*), 而且其增长点的集合的勒贝格测度为 0. 事实上, 由函数 $F(x)$ 的构造过程可见, $F(x)$ 为常数的区间 $(1/3, 2/3), (1/9, 2/9), (7/9, 8/9), \cdots$ 的总长度为

$$\frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \cdots = \frac{1}{3}\sum_{n=0}^{\infty}\left(\frac{2}{3}\right)^n = 1. \tag{6}$$

以 $\mathscr{N}$ 表示康托尔 (G. Cantor) 函数 $F(x)$ 的增长点的集合, 由 (6) 式可见 $\lambda(\mathscr{N}) = 0$. 同时, 假如 μ 是对应于康托尔函数 $F(x)$ 的测度, 则 $\mu(\mathscr{N}) = 1$. (在这种情形下, 称测度 μ 关于勒贝格测度 λ 为奇异的.)

我们不准备过多的讨论关于分布函数的可能类型的问题, 只限于指出, 实际上上面指出的三种类型包含所有的分布函数. 确切地说, 任何分布函数 $F(x)$ 都可以表示为:

$$F(x) = \alpha_1 F_1(x) + \alpha_2 F_2(x) + \alpha_3 F_3(x),$$

其中 $F_1(x)$ 是离散型分布函数, $F_2(x)$ 是连续型分布函数, $F_3(x)$ 是奇异型分布函数, 而 $\alpha_i (i = 1, 2, 3)$ 是非负实数, 而且 $\alpha_1 + \alpha_2 + \alpha_3 = 1$ (练习题 18).

2. 勒贝格测度到数轴上的开拓 定理 1 在 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上的概率测度与 $\mathbb{R}$ 上的分布函数之间, 建立了一一对应关系. 分析该定理的证明可见, 实际上有更为一般的结果, 其中包括可以在整个数轴上引进所谓勒贝格测度.

设 μ 是 $(\Omega, \mathscr{A})$ 上某一 σ- 有限测度, 其中 $\mathscr{A}$ 是 Ω 子集的代数. 结果表明, 关于测度 μ 自代数 $\mathscr{A}$ 开拓到最小 σ- 代数 $\sigma(\mathscr{A})$ 的卡拉泰奥多里定理结论, 对于 σ- 有限测度仍然成立, 这为定理 1 的推广提供了可能性.

使对于有限区间 I 的测度 $\mu(I) < \infty$ 的任何 σ- 有限测度 μ, 称为 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上的*勒贝格 - 斯蒂尔切斯测度*. 我们把定义在数轴 $\mathbb{R}$ 上, 值域为 $(-\infty, \infty)$ 的任何不减右连续函数 $G = G(x)$, 称为*广义分布函数*.

定理 1 可以这样推广, 使公式

$$\mu(a, b] = G(b) - G(a), a < b,$$

仍然可以建立勒贝格 - 斯蒂尔切斯测度 μ 与广义分布函数 G 之间的一一对应关系.

实际上, 如果 $G(+\infty) - G(-\infty) < \infty$, 则定理 1 的证明完全适用, 并且无需作任何改变, 因为可以将这种情形归结到 $G(+\infty) - G(-\infty) = 1$ 和 $G(-\infty) = 0$ 的情形.

现在假设 $G(+\infty) - G(-\infty) = \infty$. 设

$$G_n(x) = \begin{cases} G(x), & 若\ |x| \leqslant n, \\ G(n), & 若\ x > n, \\ G(-n), & 若\ x < -n. \end{cases}$$

在代数 $\mathscr{A}$ 上定义有限 - 可加测度 μ_0, 使其在 $(a, b]$ 的值 $\mu_0(a, b] = G(b) - G(a)$, 而设 μ_n 已经是 (按定理 1) 建立的、对应于函数的 $G_n(x)$ 的可数 - 可加测度.

显然, 在 $\mathscr{A}$ 上 $\mu_n \uparrow \mu_0$. 设 $A_1, A_2, \cdots$ 是 $\mathscr{A}$ 上的两两不相交集合, 且 $A \equiv \sum A_n \in \mathscr{A}$. 那么, (§1 练习题 6)

$$\mu_0(A) \geqslant \sum_{n=1}^{\infty} \mu_0(A_n).$$

而且, 如果 $\sum_{n=1}^{\infty}\mu_0(A_n)=\infty$, 则

$$\mu_0(A)=\sum_{n=1}^{\infty}\mu_0(A_n).$$

现在假设 $\sum\mu_0(A_n)<\infty$. 那么

$$\mu_0(A)=\lim_n\mu_n(A)=\lim_n\sum_{k=1}^{\infty}\mu_n(A_k).$$

根据所作的假设 $\sum\mu_0(A_n)<\infty$. 因此, 由 $\mu_n\leqslant\mu_0$ 可见

$$0\leqslant\mu_0(A)-\sum_{k=1}^{\infty}\mu_0(A_k)=\lim_n\left[\sum_{k=1}^{\infty}\mu_n(A_k)-\mu_0(A_k)\right]\leqslant 0.$$

这样, σ-有限可加测度 μ_0 在 $\mathscr{A}$ 上是有限可加的, 因此, (根据卡拉泰奥多里定理) 它可以开拓到 $\sigma(\mathscr{A})$ 上的 σ-有限测度 μ.

$G(x)=x$ 的情形特别重要, 对应于这一广义分布函数的测度, 称为 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ 上的勒贝格测度. 像区间 $[0,1]$ 的情形一样, 在数轴上可以引进勒贝格集合系 $\overline{\mathscr{B}}(\mathbb{R})$ $(\Lambda\in\overline{\mathscr{B}}(\mathbb{R})$, 如果存在勒贝格集合 A 和 B, 使 $A\subseteq\Lambda\subseteq B$, $\lambda(B\setminus A)=0)$. 对于 A 和 B 也可以定义勒贝格测度 λ (若 $A\subseteq\Lambda\subseteq B$, $\Lambda\in\overline{\mathscr{B}}(\mathbb{R})$ 且 $\lambda(B\setminus A)=0$, 则 $\lambda(\Lambda)=\lambda(A)$).

3. 可测空间 $(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n))$ 如同数轴的情形, 假设 $\mathbf{P}$ 是 $(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n))$ 上的某一测度.

记

$$F_n(x_1,\cdots,x_n)=\mathbf{P}((-\infty,x_1]\times\cdots\times(-\infty,x_n]),$$

其更紧凑的形式为

$$F_n(x)=\mathbf{P}(-\infty,x],$$

其中 $x=(x_1,\cdots,x_n)$, $(-\infty,x]=(-\infty,x_1]\times\cdots\times(-\infty,x_n]$.

引进 $\mathbb{R}^n\to\mathbb{R}$ 的差分算子 Δ_{a_i,b_i}, 按如下公式运作 $(a_i\leqslant b_i)$:

$$\Delta_{a_ib_i}F_n(x_1,\cdots,x_n)=F_n(x_1,\cdots,x_{i-1},b_i,x_{i+1},\cdots,x_n)$$
$$-F_n(x_1,\cdots,x_{i-1},a_i,x_{i+1},\cdots,x_n).$$

通过简单的运算, 可得

$$\Delta_{a_1b_1}\cdots\Delta_{a_nb_n}F_n(x_1,\cdots,x_n)=\mathbf{P}(a,b],\tag{7}$$

其中 $(a,b]=(a_1,b_1]\times\cdots\times(a_n,b_n]$. 特别, 一个不同于一维情形之处是, 一般

$$\mathbf{P}(a,b]\neq F_n(b)-F_n(a).$$

由于 $\mathbf{P}(a,b]\geqslant 0$, 故由 (7) 式可见, 对于任意 $a=(a_1,\cdots,a_n)$, $b=(b_1,\cdots,b_n)$

$$\Delta_{a_1b_1}\cdots\Delta_{a_nb_n}F_n(x_1,\cdots,x_n)\geqslant 0.\tag{8}$$

由 $\mathbf{P}$ 的连续性, 可见 $F_n(x_1,\cdots,x_n)$ 对于变量的全体右连续, 即如果 $x=(x_1,\cdots,x_n)$, $x^{(k)}=(x_1^{(k)},\cdots,x_n^{(k)})$, 则当 $x^{(k)}\downarrow x$ 时

$$F_n(x^{(k)})\downarrow F_n(x),\ k\to\infty.\tag{9}$$

同样明显,

$$F_n(+\infty,\cdots,+\infty)=1\tag{10}$$

和

$$\lim_{x\downarrow y}F_n(x_1,\cdots,x_n)=0,\tag{11}$$

假如至少有一个 y 的坐标为 $-\infty$.

定义 2 满足条件 (8)~(11) 的任意函数 $F=F_n(x_1,\cdots,x_n)$, 称做 (空间 $\mathbb{R}^n$ 上的) n 维分布函数.

运用与定理 1 同样的论述, 可以证明下面的定理.

定理 2 设 $F=F_n(x_1,\cdots,x_n)$ 是 $\mathbb{R}^n$ 中某一分布函数, 则在 $(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n))$ 上存在唯一概率测度 $\mathbf{P}$, 使

$$\mathbf{P}(a,b]=\Delta_{a_1b_1}\cdots\Delta_{a_nb_n}F_n(x_1,\cdots,x_n).\tag{12}$$

举几个 n 维分布函数的例子.

设 $F^1,\cdots,F^n$ 是 $\mathbb{R}$ 上的一维分布函数, 而

$$F_n(x_1,\cdots,x_n)=F^1(x_1)\cdots F^n(x_n).$$

显然, 该函数右连续, 且满足条件 (10) 和 (11). 不难验证

$$\Delta_{a_1b_1}\cdots\Delta_{a_nb_n}F_n(x_1,\cdots,x_n)=\prod[F^k(b_k)-F^k(a_k)]\geqslant 0.$$

因而 $F_n(x_1,\cdots,x_n)$ 是一分布函数.

函数

$$F^k(x_k)=\begin{cases}0, & 若\ x_k<0,\\ x_k, & 若\ 0\leqslant x_k\leqslant 1,\\ 1, & 若\ x_k>1\end{cases}$$

的情形特别重要. 这时, 对于一切 $0\leqslant x_k\leqslant 1$, $k=1,\cdots,n$, 有

$$F_n(x_1,\cdots,x_n)=x_1\cdots x_n.$$

对应于这一 n 维分布函数的概率测度，称做 $(0,1]^n$ 上的 n 维勒贝格测度.

多数 n 维分布函数具有如下形式

$$F_n(x_1,\cdots,x_n)=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n} f_n(t_1,\cdots,t_n)\mathrm{d}t_1\cdots\mathrm{d}t_n,$$

其中 $f_n(t_1,\cdots,t_n)$ 非负函数，且满足

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_n(t_1,\cdots,t_n)\mathrm{d}t_1\cdots\mathrm{d}t_n=1,$$

而积分可以理解为黎曼积分 (在更一般的情形下应理解为勒贝格积分). 函数 $f=f_n(t_1,\cdots,t_n)$ 称做 n 维概率分布函数的密度，n 维概率分布密度，或简称为 n 维密度.

当 $n=1$ 时，函数

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}},\ x\in\mathbb{R},$$

其中 $\sigma>0$，是 (非退化) 高斯分布密度或正态分布密度. 当 $n>1$ 时，存在这一密度的自然类似情形.

设 $\mathbb{B}=(r_{ij})$ 是一 $n\times n$ 阶非负定对称矩阵：

$$\sum_{i,j=1}^{n} r_{ij}\lambda_i\lambda_j\geqslant 0,\ \lambda_i\in\mathbb{R},\ i=1,\cdots,n,$$

$$r_{ij}=r_{ji}.$$

当 $\mathbb{B}$ 是正定矩阵时，其行列式 $|\mathbb{B}|=\det\mathbb{B}>0$，从而有逆矩阵 $A=(a_{ij})=\mathbb{B}^{-1}$. 那么，函数

$$f_n(x_1,\cdots,x_n)=\frac{|A|^{1/2}}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}\sum a_{ij}(x_i-m_i)(x_j-m_j)\right\},\tag{13}$$

其中 $m_i\in\mathbb{R}$，$i=1,\cdots,n$. $f_n(x_1,\cdots,x_n)$ 具有下列性质：$f_n(x_1,\cdots,x_n)>0$，且在整个空间上的 (n 重黎曼) 积分等于 1 (这将在 §13 证明)，且由于它是正的，故是概率密度.

这一函数称做 n 维 (非退化) 高斯分布密度或正态分布密度. 其均值向量为 $m=(m_1,\cdots,m_n)$，而协方差矩阵为 $\mathbb{B}=A^{-1}$.

当 $n=2$ 时，密度 $f_2(x_1,x_2)$ 为：

$$\begin{aligned}f_2(x_1,x_2)=&\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\right.\\&\left.\times\left[\frac{(x_1-m_1)^2}{\sigma_1^2}-2\rho\frac{(x_1-m_1)(x_2-m_2)}{\sigma_1\sigma_2}+\frac{(x_2-m_2)^2}{\sigma_2^2}\right]\right\},\end{aligned}\tag{14}$$

其中 $\sigma_i>0$，$|\rho|<1$ (关于参数 m_i,σ_i 和 ρ 的含义将在 §8 中说明.) 图 28 是二维正态分布的示意图.

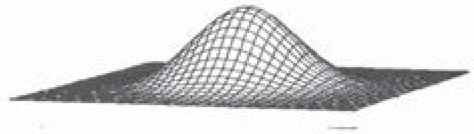


图 28 二维正态密度的图形

注 像 $n=1$ 的情形一样，定理 2 可以推广到 (与定义类似的) $(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n))$ 上的勒贝格–斯蒂尔切斯测度，以及 $\mathbb{R}^n$ 上的广义分布函数. 当广义分布函数 $G_n(x_1,\cdots,x_n)$ 等于 $x_1\cdots x_n$ 时，相应的测度称为空间 $\mathbb{R}^n$ 的博雷尔集合上的勒贝格测度. 显然，对于博雷尔集合上的勒贝格测度，

$$\lambda(a,b]=\prod_{i=1}^{n}(b_i-a_i),$$

即 "矩形"

$$(a,b]=(a_1,b_1]\times\cdots\times(a_n,b_n]$$

的勒贝格测度等于其 "体积".

4. 可测空间 $(\mathbb{R}^\infty,\mathscr{B}(\mathbb{R}^\infty))$ 对于空间 $\mathbb{R}^n$，$n\geqslant 1$，的情形，概率测度是按如下模式建立的：首先从基本集合 —— 形如 $(a,b]$ 的 "矩形" 出发，然后将其自然地扩展到形如集合 $A=\sum(a_i,b_i]$，最后根据卡拉泰奥多里定理将其开拓到 $\mathscr{B}(\mathbb{R}^n)$ 中的集合.

类似的建立概率模型的模式，也 "适用于" 空间 $(\mathbb{R}^\infty,\mathscr{B}(\mathbb{R}^\infty))$ 的情形. 以

$$\mathscr{I}_n(B)=\{x\in\mathbb{R}^\infty:(x_1,\cdots,x_n)\in\mathscr{B}\},B\in\mathscr{B}(\mathbb{R}^n)$$

表示空间 $\mathbb{R}^\infty$ 中 $B\in\mathscr{B}(\mathbb{R}^\infty)$ 为 "底" 的柱集的集合. 我们将看到，正是把柱集自然地视为 $\mathbb{R}^\infty$ 中的基本集合，并根据其概率的值定义 $\mathscr{B}(\mathbb{R}^\infty)$ 中集合上的概率测度.

假设 $\mathbf{P}$ 是空间 $(\mathbb{R}^\infty,\mathscr{B}(\mathbb{R}^\infty))$ 上的某一概率测度. 对于 $n=1,2,\cdots$，记

$$P_n(B)=\mathbf{P}(\mathscr{I}_n(B)),B\in\mathscr{B}(\mathbb{R}^n).\tag{15}$$

相应地定义在 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$，$(\mathbb{R}^2,\mathscr{B}(\mathbb{R}^2))$，$\cdots$ 上的概率测度 $P_1,P_2,\cdots$ 序列，具有如下明显的一致性：对于 $n=1,2,\cdots$ 和 $B\in\mathscr{B}(\mathbb{R}^n)$，

$$P_{n+1}(B\times\mathbb{R})=P_n(B).\tag{16}$$

特别值得注意的是，相反的结果也成立.

定理 3 (柯尔莫戈洛夫关于 $(\mathbb{R}^\infty,\mathscr{B}(\mathbb{R}^\infty))$ 中的测度) 假设 $P_1,P_2,\cdots$ 是 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$, $(\mathbb{R}^2,\mathscr{B}(\mathbb{R}^2)),\cdots$ 上, 具有一致性 (16) 的概率测度序列. 那么, 在 $(\mathbb{R}^\infty,\mathscr{B}(\mathbb{R}^\infty))$ 上存在唯一的概率测度 $\mathbf{P}$, 使对于每个 $n=1,2,\cdots$, 有

$$\mathbf{P}(\mathscr{I}_n(B))=P_n(B),\ B\in\mathscr{B}(\mathbb{R}^n). \tag{17}$$

证明 假设 $B^n\in\mathscr{B}(\mathbb{R}^n)$ 而 $\mathscr{I}_n(B^n)$ 是以 B^n 为 "底" 的柱集. 赋予该柱集以测度 $\mathbf{P}(\mathscr{I}_n(B^n))=P_n(B^n)$.

现在证明, 由于一致性条件, 这样的定义是适定的, 即 $\mathbf{P}(\mathscr{I}_n(B^n))$ 的值与柱集的表现方法无关. 事实上, 假设同一柱集有两种表示方法:

$$\mathscr{I}_n(B^n)=\mathscr{I}_{n+k}(B^{n+k}).$$

那么, 由此可见, 若 $(x_1,\cdots,x_{n+k})\in\mathbb{R}^{n+k}$, 则

$$(x_1,\cdots,x_n)\in B^n\Leftrightarrow(x_1,\cdots,x_{n+k})\in B^{n+k}. \tag{18}$$

从而, 由 (16) 式, 有

$$\begin{aligned}P_n(B^n)&=P_{n+1}(\{(x_1,\cdots,x_{n+1}):(x_1,\cdots,x_n)\in B^n\})=\cdots\\&=P_{n+k}(\{(x_1,\cdots,x_{n+k}):(x_1,\cdots,x_n)\in B^n\})=P_{n+k}(B^{n+k}).\end{aligned}$$

设 $\mathscr{A}(\mathbb{R}^\infty)$ 是一切柱集 $\hat{B}^n=\mathscr{I}_n(B^n)$, $B^n\in\mathscr{B}(\mathbb{R}^n)$, $n=1,2,\cdots$ 的全体. 易见, $\mathscr{A}(\mathbb{R}^\infty)$ 是代数.

现在设 $\hat{B}^1,\cdots,\hat{B}^k$ 是 $\mathscr{A}(\mathbb{R}^\infty)$ 中的不相交集合. 不失普遍性, 可以假设: 对于某个 n, 有 $\hat{B}_i=\mathscr{I}_n(B_i^n)$, $i=1,2,\cdots,k$, 其中 $B_1^n,\cdots,B_k^n$ 是 $\mathscr{B}(\mathbb{R}^n)$ 中的两两不相交集合. 那么,

$$\mathbf{P}\left(\sum_{i=1}^k\hat{B}_i\right)=\mathbf{P}\left(\sum_{i=1}^k\mathscr{I}_n(B_i^n)\right)=P_n\left(\sum_{i=1}^k B_i^n\right)=\sum_{i=1}^k P_n(B_i^n)=\sum_{i=1}^k\mathbf{P}(\hat{B}_i),$$

即集函数 $\mathbf{P}$ 在代数 $\mathscr{A}(\mathbb{R}^\infty)$ 上有限 − 可加.

现在证明, $\mathbf{P}$ 在 "零" 连续 (从而在 $\mathscr{A}(\mathbb{R}^\infty)$ 上 $\sigma-$ 可加; 见 §1 的定理), 也就是说, 若当 $n\to\infty$ 时 $\hat{B}_n\downarrow\varnothing$, 则当 $n\to\infty$ 时 $\mathbf{P}(\hat{B}_n)\to0$.

假设相反, 即设 $\lim\mathbf{P}(\hat{B}_n)=\delta>0$. 不失普遍性, 可以认为:

$$\hat{B}_n=\{x:(x_1,\cdots,x_n)\in B_n\},\ B_n\in\mathscr{B}(\mathbb{R}^n).$$

利用概率测度 P_n 在 $(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n))$ 中的如下性质 (见例 9): 如果 $B_n\in\mathscr{B}(\mathbb{R}^n)$, 则对于任意给定的 $\delta>0$, 存在紧统 $A_n\in\mathscr{B}(\mathbb{R}^n)$, 使 $A_n\subseteq B_n$, $P_n(B_n\setminus A_n)\leqslant\delta/2^{n+1}$. 故若 $\hat{A}_n=\{x:(x_1,\cdots,x_n)\in A_n\}$, 则

$$\mathbf{P}(\hat{B}_n\setminus\hat{A}_n)=P_n(B_n\setminus A_n)\leqslant\delta/2^{n+1}.$$

考虑集合 $\hat{C}_n=\bigcap_{k=1}^n\hat{A}_k$, 假设

$$\hat{C}_n=\{x:x=(x_1,\cdots,x_n)\in C_n\}.$$

考虑到集合序列 $\hat{B}_n$ 递降, 得

$$\mathbf{P}(\hat{B}_n\setminus\hat{C}_n)\leqslant\sum_{k=1}^n\mathbf{P}(\hat{B}_n\setminus\hat{A}_k)\leqslant\sum_{k=1}^n\mathbf{P}(\hat{B}_k\setminus\hat{A}_k)\leqslant\delta/2.$$

但是按假设的条件 $\mathbf{P}(\hat{B}_n)\to\delta>0$, 由此应得

$$\lim_n\mathbf{P}(\hat{C}_n)\geqslant\frac{\delta}{2}>0.$$

现在证明这与 $\hat{C}_n\downarrow\varnothing$ 矛盾.

事实上, 在集合 $\hat{C}_n$ 中各选一点 $\hat{x}^{(n)}=(x_1^{(n)},x_2^{(n)},\cdots)$, 那么, 对于任意 $n\geqslant1$, 有 $\hat{x}^{(n)}=(x_1^{(n)},\cdots,x_n^{(n)})\in C_n$.

设 (n_1) 是序列 (n) 的这样一个子列, 使 $x_1^{(n_1)}\to x_1^0$, 其中 x_1^0 是 C_1 中某一点. (这样的子列存在, 因为一切 $x_1^{(n)}\in C_1$, 而 C_1 是紧统.) 由子列 (n_1) 选这样一个子列 (n_2), 使得 $(x_1^{(n_2)},x_2^{(n_2)})\to(x_1^0,x_2^0)\in C_2$. 同样, 设 $(x_1^{(n_k)},\cdots,x_k^{(n_k)})\to(x_1^0,\cdots,x_k^0)\in C_k$. 最后, 得对角序列 (m_k), 其中 m_k 是子列 (n_k) 的第 k 项. 那么, 对于任意 $i=1,2,\cdots$, 当 $m_k\to\infty$ 时, 有 $x_i^{(m_k)}\to x_i^0$, 并且对于任意 $n=1,2,\cdots$, 点 $(x_1^0,x_2^0,\cdots)\in\hat{C}_n$. 显然, 这与假设 $\hat{C}_n\downarrow\varnothing$, $n\to\infty$ 矛盾.

于是, 集函数 $\mathbf{P}$ 在代数 $\mathscr{A}(\mathbb{R}^\infty)$ 上 $\sigma-$ 可加. 说明根据卡拉泰奥多里定理, 可以将 $\mathbf{P}$ 开拓为 $(\mathbb{R}^\infty,\mathscr{B}(\mathbb{R}^\infty))$ 上的 (概率) 测度. □

注 现在考虑的情形, 空间 $\mathbb{R}^\infty$ 是 $\mathbb{R}$ 的直积 $\mathbb{R}^\infty=\mathbb{R}\times\mathbb{R}\times\cdots$ 自然产生一个问题, 假如把 $(\mathbb{R}^\infty,\mathscr{B}(\mathbb{R}^\infty))$ 换成可测空间 $(\Omega_i,\mathscr{F}_i)$ 的直积, 定理 3 是否仍然成立.

只要仔细分析一下该定理的证明就会发现, 用到的直线具有拓扑特点唯一性质, 就是本质上用到: "在 $\mathscr{B}(\mathbb{R}^n)$ 的任意集合 A 中存在紧统, 使其概率测度可以任意地接近集合 A 概率测度". 不过熟知, 不仅空间 $(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n))$ 具有该性质, 而且任何完备可分的、具有由开集生成的 $\sigma-$ 代数的可测空间都具有这一性质.

这样, 定理 3 对于如下情形仍然成立: $P_1,P_2,\cdots$ 是空间

$$(\Omega_1,\mathscr{F}_1),(\Omega_1\times\Omega_2,\mathscr{F}_1\otimes\mathscr{F}_2),\cdots$$

上的满足一致性的概率测度序列, 其中 $(\Omega_i,\mathscr{F}_i)$ 是完备可分度量空间, $\sigma-$ 代数 $\mathscr{F}_i$ 由开集生成, 而用 $(\Omega_1\times\Omega_2\times\cdots,\mathscr{F}_1\otimes\mathscr{F}_2\otimes\cdots)$ 代替 $(\mathbb{R}^\infty,\mathscr{B}(\mathbb{R}^\infty))$.

在 §9 (定理 2) 将证明, 对于任意可测空间 $(\Omega_n,\mathscr{F}_n)$, 如果测度 P_n, $n\geqslant1$ 是以某种特别方式构造的情形, 定理 3 的结果仍然成立. 在 (关于所述可测空间, 或测度 $\{P_n\}$ 族结构的拓扑特点无任何假设的) 一般情形下, 定理 3 可能不正确. 下面的例子就说明这一点.

考虑空间 $\Omega=(0,1]$. 它显然不完备. 在此空间上按如下模式建立 $\sigma-$ 代数序列 $\mathscr{F}_1\subseteq\mathscr{F}_2\subseteq\cdots$. 对于一切 $n=1,2,\cdots$, 设

$$\varphi_n(\omega)=\begin{cases}1, & 若\ 0<\omega<1/n,\\ 0, & 若\ 1/n\leqslant\omega\leqslant 1,\end{cases}$$

$$\mathscr{C}_n=\{A\in\Omega: A=[\omega:\varphi_n(\omega)\in B], B\in\mathscr{B}(\mathbb{R})\}$$

而 $\mathscr{F}_n=\sigma\{\mathscr{C}_1,\cdots,\mathscr{C}_n\}$ 是包含集系 $\mathscr{C}_1,\cdots,\mathscr{C}_n$ 的最小 $\sigma-$ 代数. 显然 $\mathscr{F}_1\subseteq\mathscr{F}_2\subseteq\cdots$. 设 $\mathscr{F}=\sigma(\bigcup\mathscr{F}_n)$ 是包含集系一切 $\mathscr{F}_n$ 的最小 $\sigma-$ 代数. 考虑可测空间 $(\Omega,\mathscr{F}_n)$, 并在其上以如下方式定义概率测度 P_n:

$$P_n\{\omega:(\varphi_1(\omega),\cdots,\varphi_n(\omega))\in B^n\}=\begin{cases}1, & 若\ (1,\cdots,1)\in B^n,\\ 0, & 若不然,\end{cases}$$

其中 $B^n\in\mathscr{B}(\mathbb{R}^n)$. 不难验证, 测度族是一致的: 如果 $A\in\mathscr{F}_{n-1}$, 则 $P_{n-1}(A)=P_n(A)$. 然而可以证明, 在 $(\Omega,\mathscr{F})$ 上不存在概率测度 $\mathbf{P}$, 使其收缩 $\mathbf{P}|\mathscr{F}_n$(即将测度 $\mathbf{P}$ 仅局限在 $\mathscr{F}_n$ 中的集合上) 与 P_n, $n=1,2,\cdots$ 重合. 事实上, 假如这样的测度 $\mathbf{P}$ 存在, 那么, 对于任意 $n=1,2,\cdots$, 有

$$\mathbf{P}\{\omega:\varphi_1(\omega)=\cdots=\varphi_n(\omega)=1\}=P_n\{\omega:\varphi_1(\omega)=\cdots=\varphi_n(\omega)=1\}=1. \tag{19}$$

但是,

$$\{\omega:\varphi_1(\omega)=\cdots=\varphi_n(\omega)=1\}=(0,1/n)\downarrow\varnothing,$$

与 (19) 式矛盾, 从而也与函数 $\mathbf{P}$ 的可数可加性 (与在"零" $\varnothing$ 的连续性) 矛盾.

现在举一个 $(\mathbb{R}^\infty,\mathscr{B}(\mathbb{R}^\infty))$ 上概率测度的例. 设 $F_1(x),F_2(x),\cdots$ 是一维分布函数序列. 定义函数 $G_1(x)=F_1(x)$, $G_2(x_1,x_2)=F_1(x_1)F_2(x_2),\cdots$. 记 $P_1,P_2,\cdots$ 为与这些函数相对应空间 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$, $(\mathbb{R}^2,\mathscr{B}(\mathbb{R}^2)),\cdots$ 上的概率测度. 那么, 由定理 3 可见, 在 $(\mathbb{R}^\infty,\mathscr{B}(\mathbb{R}^\infty))$ 上存在一测度, 使

$$\mathbf{P}\{x\in\mathbb{R}^\infty:(x_1,\cdots,x_n)\in B\}=P_n(B), B\in\mathscr{B}(\mathbb{R}^n),$$

特别,

$$\mathbf{P}\{x\in\mathbb{R}^\infty:x_1\leqslant a_1,\cdots,x_n\leqslant a_n\}=F_1(a_1)\cdots F_n(a_n).$$

设 $F_i(x)$ 是伯努利分布函数:

$$F_i(x)=\begin{cases}0, & 若\ x<0,\\ q, & 若\ 0\leqslant x<1,\\ 1, & 若\ x\geqslant 1.\end{cases}$$

设 Ω 是一切数值序列 $x=(x_1,x_2,\cdots)$, $x_i=0,1$, 的空间. $\mathscr{B}(\mathbb{R}^\infty)\bigcap\Omega$ 是 Ω 中博雷尔子集的 $\sigma-$ 代数. 可以证明, 在 $(\Omega,\mathscr{B}(\mathbb{R}^\infty)\bigcap\Omega)$ 上存在概率测度 $\mathbf{P}$, 使对于任意 $n=1,2,\cdots$, 有

$$\mathbf{P}\{x:x_1=a_1,\cdots,x_n=a_n\}=p^{\sum a_i}q^{n-\sum a_i}.$$

需要指出, 正是因为缺少这一结果, 我们在第一章 §5 中不能以 (8) 式的形式表述大数定律.

5. 可测空间 $(\mathbb{R}^T,\mathscr{B}(\mathbb{R}^T))$　设 T 是下标 $t\in T$ 的任意集合, $\mathbb{R}_t$ 而是下标为 t 的数轴. 考虑不同下标 $t_i\in T$ 任意无序数组 $\tau=[t_1,\cdots,t_n]$, $n\geqslant 1$; 设 P_τ 是 $(\mathbb{R}^\tau,\mathscr{B}(\mathbb{R}^\tau))$ 上的概率测度, 其中 $\mathbb{R}^\tau=\mathbb{R}_{t_1}\times\cdots\times\mathbb{R}_{t_n}$.

设 τ 在一切有限无序数组的集合取值, 则称概率测度族 $\{P_\tau\}$ 为一致的, 如果对于任意两个数组 $\tau=[t_1,\cdots,t_n]$ 和 $\sigma=[s_1,\cdots,s_k]$ 且 $\sigma\subseteq\tau$, 对于任何 $B\in\mathscr{B}(\mathbb{R}^\sigma)$, 有

$$\begin{aligned}&P_\sigma\{(x_{s_1},\cdots,x_{s_k}):(x_{s_1},\cdots,x_{s_k})\in B\}\\&=P_\tau\{(x_{t_1},\cdots,x_{t_n}):(x_{s_1},\cdots,x_{s_k})\in B\}.\end{aligned} \tag{20}$$

定理 4 ($(\mathbb{R}^T,\mathscr{B}(\mathbb{R}^T))$ 上的柯尔莫戈洛夫测度开拓定理)　*设 $\{P_\tau\}$ 是 $(\mathbb{R}^\tau,\mathscr{B}(\mathbb{R}^\tau))$ 上的一致概率测度族, 则在 $(\mathbb{R}^T,\mathscr{B}(\mathbb{R}^T))$ 上存在唯一概率测度 $\mathbf{P}$, 使对于一切不同下标 $t_i\in T$ 的无序数组 $\tau=[t_1,\cdots,t_n]$, $B\in\mathscr{B}(\mathbb{R}^\tau)$ 和 $\mathscr{I}_\tau=\{x\in\mathbf{R}^T:(x_{t_1},\cdots,x_{t_n})\in B\}$, 有*

$$\mathbf{P}\{\mathscr{I}_\tau(B)\}=P_\tau(B). \tag{21}$$

证明　设集合 $\hat{B}\in\mathscr{B}(\mathbb{R}^T)$. 根据 §2 定理 3 存在有限或可数集合 $S=\{s_1,s_2,\cdots\}\subseteq T$, 使 $\hat{B}=\{x:(x_{s_1},x_{s_2},\cdots)\in B\}$, 其中 $B\in\mathscr{B}(\mathbb{R}^S)$, $\mathbb{R}^S=\mathbb{R}_{s_1}\times\mathbb{R}_{s_2}\times\cdots$. 换句话说, $\hat{B}=\mathscr{I}_S(B)$ 是"底"为 $B\in\mathscr{B}(\mathbb{R}^S)$ 的柱集.

在这样的柱集 $\hat{B}=\mathscr{I}_S(B)$ 上定义一集函数 $\mathbf{P}$:

$$\mathbf{P}(\mathscr{I}_S(B))=P_S(B), \tag{22}$$

其中由定理 3 知, 概率测度 P_S 存在.

我们现在证明, $\mathbf{P}$ 就是定理所断定存在的测度. 为此需要证明两点. 第一, 验证定义 (22) 适定, 即在用不同方式表示 $\hat{B}$ 的情形下, $\mathbf{P}(\hat{B})$ 的值不变; 第二, 集函数 $\mathbf{P}$ 可数可加.

这样, 设 $\hat{B}=\mathscr{I}_{S_1}(B_1)$ 和 $\hat{B}=\mathscr{I}_{S_2}(B_2)$. 那么, 显然 $\hat{B}=\mathscr{I}_{S_1\bigcup S_2}(B_3)$, 其中 $B_3\in\mathscr{B}(\mathbb{R}^{S_1\bigcup S_2})$. 因此只需证明, 若 $S\subseteq S'$ 且 $B\in\mathscr{B}(\mathbb{R}^S)$, 则 $P_{S'}(B')=P_S(B)$, 其中

$$B'=\{(x_{s'_1},x_{s'_2},\cdots):(x_{s_1},x_{s_2},\cdots)\in B\},$$

而 $S'=\{s'_1,s'_2,\cdots\}$, $S=\{s_1,s_2,\cdots\}$.

由于一致性条件 (20), 等式 (22) 可以直接由定理 3 得出, 从而证明 $\mathbf{P}(\hat{B})$ 的值与集合 $\hat{B}$ 的表示方法无关.

其次, 为验证集函数 $\mathbf{P}$ 的可数可加性, 假设 $\{\hat{B}_n\}$ 是 $\mathscr{B}(\mathbb{R}^T)$ 中某一两两不交的集合序列. 则存在有限或可数集合 $S \subseteq T$, 使对于任何 $n \geqslant 1$, $\hat{B}_n = \mathscr{T}_S(B_n)$, 其中 $B_n \in \mathscr{B}(\mathbb{R}^S)$. 由于 P_S 是概率测度, 则

$$\mathbf{P}\left(\sum \hat{B}_n\right) = \mathbf{P}\left(\sum \mathscr{T}_S(B_n)\right) = P_S\left(\sum B_n\right)$$
$$= \sum P_S(B_n) = \sum \mathbf{P}(\mathscr{T}_S(B_n)) = \sum \mathbf{P}(\hat{B}_n).$$

最后, 直接由测度 $\mathbf{P}$ 的构造可得性质 (21). □

注 1 需要强调 T 是下标的任意集合. 这时由于定理 3 的注可见, 如果用任意完备可分度量空间 Ω_t (连同由开集生成的 σ- 代数) 代替数轴 $\mathbb{R}_t$, 则定理仍然成立.

注 2 曾假定所讨论的概率测度族 $\{P_\tau\}$, 对于一切不同下标的无序数组 $\tau = [t_1, \cdots, t_n]$ 是给定的. 对此应着重强调, 作为 $\tau = [t_1, \cdots, t_n]$ 的函数, 这些测度 P_τ 实质上是由不同点 $\{t_1\}, \cdots, \{t_n\}$ 构成的集合的函数. (例如, 数组 $[a, b]$ 和数组 $[b, a]$ 应视为同一数组, 因为二者都是由点 $\{a\}$ 和 $\{b\}$ 构成的同一集合.) 有时作为开始的取概率测度族 $\{P_\tau\}$, 其中 τ 在一切不同下标的有序数组 $\tau = (t_1, \cdots, t_n)$ 的集合上取值. (这时, 数组 (a, b) 和数组 (b, a) 应视为不同的数组, 因为对于有序数组元素的先后顺序至关重要.) 在这种情形下, 为使定理 4 仍然成立, 除条件 (20) 外需要再增加一个一致性条件:

$$P_{(t_1, \cdots, t_n)}(A_{t_1} \times \cdots \times A_{t_n}) = P_{(t_{i_1}, \cdots, t_{i_n})}(A_{t_{i_1}} \times \cdots \times A_{t_{i_n}}), \tag{23}$$

其中 $(i_1, \cdots, i_n)$ 是数 $(1, \cdots, n)$ 的任意排列, $A_{t_i} \in \mathscr{B}(\mathbb{R}_{t_i})$; 条件 (23) 作为概率测度 $\mathbf{P}$ 存在的必要条件, 由 (21) 式 (将 $P_{[t_1, \cdots, t_n]}(B)$ 换成 $P_{(t_1, \cdots, t_n)}(B)$) 可见是显然的.

我们下面总假设所考虑的数组 τ 是无序的. 如果 T 是数轴上的集合 (或为某一完全有序集合), 则不失普遍性, 可以认为对于所考虑的数组 $\tau = [t_1, \cdots, t_n]$, 有 $t_1 < t_2 < \cdots < t_n$. (例如, 集合 τ 由数值点 $\{a_1\}, \{a_2\}, \cdots, \{a_n\}$ 组成, 则 τ 总可以表示为 $\tau = [t_1, t_2, \cdots, t_n]$, 其中 $t_1 < t_2 < \cdots < t_n$.) 于是, 这里所有 "有限维" 概率只需给出这样的数组 $\tau = [t_1, t_2, \cdots, t_n]$, 其中 $t_1 < t_2 < \cdots < t_n$, 其中 $t_1 = \min(a_1, \cdots, a_n)$, $t_n = \max(a_1, \cdots, a_n)$. 这样, 在这种情形下所有 "有限维" 概率, 只需对于这样的数组 $\tau = [t_1, t_2, \cdots, t_n]$ 给出就可以了, 其中 $t_1 < t_2 < \cdots < t_n$.

现在考虑 $T = [0, \infty)$ 的情形. 这种情形下 $\mathbb{R}^T$ 是一切实函数 $x = (x_t)_{t \geqslant 0}$ 的空间. $(\mathbb{R}^{[0,\infty)}, \mathscr{B}(\mathbb{R}^{[0,\infty)}))$ 上概率测度的一个重要的例子是所谓维纳 (N. Weiner) 测度, 其构造如下.

考虑高斯概率密度族 $\{\varphi_t(y|x)\}_{t \geqslant 0}$ (是 y 的函数, 而 x 固定):

$$\varphi_t(y|x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}}, \quad y \in \mathbb{R};$$

对于每一数组 $\tau = [t_1, t_2, \cdots, t_n]$, $t_1 < t_2 < \cdots < t_n$, 集合 $B = I_1 \times \cdots \times I_n$, 其中 $I_k = (a_k, b_k)$, 以及由公式

$$\begin{aligned} P_\tau(B) &= P_\tau(I_1 \times \cdots \times I_n) \\ &= \int_{I_1} \cdots \int_{I_n} \varphi_{t_1}(a_1|0)\varphi_{t_2-t_1}(a_2|a_1) \cdots \varphi_{t_n-t_{n-1}}(a_n|a_{n-1})\mathrm{d}a_1 \cdots \mathrm{d}a_n \end{aligned} \tag{24}$$

定义的测度 $P_\tau(B)$ (积分理解为黎曼积分). 然后, 对于每一柱集

$$\mathscr{T}_{t_1 \cdots t_n}(I_1 \times \cdots \times I_n) = \{x \in \mathbb{R}^T : x_{t_1} \in I_1, \cdots, x_{t_n} \in I_n\}$$

定义集函数 $\mathbf{P}$, 设

$$\mathbf{P}(\mathscr{T}_{t_1 \cdots t_n}(I_1 \times \cdots \times I_n)) = P_{[t_1, \cdots, t_n]}(I_1 \times \cdots \times I_n).$$

这样赋予柱集 $\mathscr{T}_{t_1, \cdots, t_n}(I_1 \times \cdots \times I_n)$ 测度的方法, 其直观含义如下.

集合 $\mathscr{T}_{t_1, \cdots, t_n}(I_1 \times \cdots \times I_n)$ 是于时间 $t_1, \cdots, t_n$ 通过 "窗口" $I_1, \cdots, I_n$ (§2 图 24) 的, 所有函数的集合. 我们把 $\varphi_{t_k - t_{k-1}}(a_k|a_{k-1})\mathrm{d}a_k$ 看作 "质点自点 a_{k-1} 出发, 经时间 $t_k - t_{k-1}$ 到达点 a_k 的 $\mathrm{d}a_k$ 一邻域" 的概率. 那么, 在 (24) 式中所考虑的密度的积, 表示运动的 "质点" 在时间区间 $[0, t_1], [t_1, t_2], \cdots, [t_{n-1}, t_n]$ 平移增量一定的独立性.

不难看到, 这样建立的测度族 $\{P_\tau\}$ 是一致的, 从而可以开拓为 $(\mathbb{R}^{[0,\infty)}, \mathscr{B}(\mathbb{R}^{[0,\infty)}))$ 上的测度. 这样得到的测度, 在概率论中起重要作用. 该测度是维纳引进的, 称为维纳测度.

6. 练习题

1. 设 $F(x) = \mathbf{P}(-\infty, x]$. 证明下列公式:

$$\mathbf{P}(a, b] = F(b) - F(a), \mathbf{P}(a, b) = F(b-) - F(a),$$
$$\mathbf{P}[a, b] = F(b) - F(a-), \mathbf{P}[a, b) = F(b-) - F(a-),$$
$$\mathbf{P}(\{x\}) = F(x) - F(x-),$$

其中 $F(x-) = \lim\limits_{y \uparrow x} F(y)$.

2. 证明公式 (7).

3. 证明定理 2.

4. 证明, 分布函数 $F = F(x)$ 在 $\mathbb{R}$ 上最多有可数个间断点. 问对于 $\mathbb{R}^n$ 上的分布函数有何相应的结果?

5. 考虑函数

$$G(x, y) = \begin{cases} 1, & 若\ x + y \geqslant 0, \\ 0, & 若\ x + y < 0, \end{cases}$$

$G(x, y) = [x + y]$ 是 $x + y$ 的整数部分.

证明，两个函数都具有性质：对于每一个自变量不连续、递增，但不是 R^2 上的（广义）分布函数.

6. 设 μ 是对应于连续广义分布函数的勒贝格 – 斯蒂尔切斯测度. 证明，若集合 A 最多是可数的，则 $\mu(A)=0$.

7. 设 c 是连续统的势. 证明博雷尔集 R^n 的势等于 c，而勒贝格 σ– 代数的势等于 2^c.

8. 设 $(\Omega,\mathscr{F},\mathbf{P})$ 是某一概率空间，$\mathscr{A}$ 是 Ω 子集的代数，而 $\sigma(\mathscr{A})=\mathscr{F}$. 利用适当集合原理，证明对于任意 $\varepsilon>0$ 和 $B\in\mathscr{F}$，存在集合 $A\in\mathscr{A}$，使

$$\mathbf{P}(A\triangle B)\leqslant\varepsilon.$$

9. 设 $\mathbf{P}$ 是 $(R^n,\mathscr{B}(R^n))$ 上的概率测度. 证明，对于任意 $\varepsilon>0$ 和 $B\in\mathscr{B}(R^n)$，存在紧集 A_1 和开集 A_2，使 $A_1\subseteq B\subseteq A_2$，且 $\mathbf{P}(A_2\setminus A_1)\leqslant\varepsilon$.（该结果曾用于定理 3 的证明.）

10. 设 $\mathbf{P}$ 是给定概率测度. 验证由 $\mathbf{P}_\tau(B)=\mathbf{P}\{\mathscr{I}_\tau(B)\}$ 建立的测度族的一致性（对照 (21) 式）.

11. 验证表 2 – 2 和表 2 – 3 中的 "分布" 确实是概率分布.

12. 证明第 1 小节的注 2 中的 $\overline{\mathscr{B}}$ 是 σ– 代数.

13. 证明第 1 小节的注 2 中的集函数 $\overline{\mu}(A)$，$A\in\overline{\mathscr{B}}$，是测度.

14. 举一例说明，若测度 μ_0 是 $\mathscr{A}$ 上的有限 – 可加测度（但不可数 – 可加），则不可能开拓为 $\sigma(\mathscr{A})$ 上的可数 – 可加测度.

15. 证明，Ω 子集的 σ – 代数 $\mathscr{A}$ 上的任何有限 – 可加概率测度，都可以开拓为 Ω 的一切子集上的有限 – 可加概率测度.

16. 设 $\mathbf{P}$ 是 Ω 子集的 σ – 代数 $\mathscr{F}$ 上给定的概率测度. 假设 $C\subseteq\Omega$，但是 $C\notin\mathscr{F}$. 证明测度 $\mathbf{P}$ 可以开拓到 σ – 代数 $\sigma(\mathscr{F}\cup\{C\})$，并且保持可数可加性.

17. 证明连续型分布函数 F 的承载子是一完全集合（即 $\operatorname{supp}F$ 的承载子是闭集，且具有如下性质：若 $x\in\operatorname{supp}F$，$\varepsilon>0$，则存在 $y\in\operatorname{supp}F$，使 $0<|x-y|<\varepsilon$）. 证明（任意）分布函数的承载子都是闭集.

18. 证明关于每一分布函数 F 结构的如下基本结果（见第 1 小节末尾）：每一个分布函数都是凸组合

$$F=\alpha_1F_{\mathrm{d}}+\alpha_2F_{\mathrm{abc}}+\alpha_3F_{\mathrm{sc}},$$

其中 F_{d} 是离散型分布函数，F_{abc} 是绝对连续分布函数，F_{sc} 是奇异连续分布函数；而 $\alpha_i\geqslant0$，$\alpha_1+\alpha_2+\alpha_3=1$.

19. 证明，对于康托尔函数和对于康托尔增长点的集合 $\mathscr{N}$（与 $\operatorname{supp}F$ 的承载子重合）中的每一个点 x，有如下表现：

$$x=\sum_{k=1}^{\infty}\frac{\alpha_k(x)}{3^k},$$

其中 $\alpha_k(x)=0$ 或 2，而且对于这样的点，有

$$F(x)=\sum_{k=1}^{\infty}\frac{\alpha_k(x)}{2^{k+1}}.$$

20. 设 C 是 R 上一闭集. 建立分布函数 F，使其承载子为 $\operatorname{supp}F=C$.

21. 证明，二项分布（见第 1 小节）的分布函数

$$B_n(m;p)=\mathbf{P}_n\{\nu\leqslant m\}=\sum_{k=0}^{m}C_n^kp^kq^{n-k}$$

可以通过（不完全）B 函数 $B_n(m;p)$ 表示：

$$B_n(m;p)=\frac{1}{\mathrm{B}(m+1,n-m)}\int_p^1x^m(1-x)^{n-m-1}\mathrm{d}x.$$

22. 证明，泊松分布的分布函数

$$F(n;\lambda)=\sum_{k=0}^{n}\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$$

可以通过（不完全）Γ 函数表示：

$$F(n;\lambda)=\frac{1}{n!}\int_\lambda^\infty x^n\mathrm{e}^{-x}\mathrm{d}x.$$

23. 在描绘分布密度 $f=f(x)$ 的形状时，除均值和方差外，标准特征是参数 "偏度" (skewness)

$$\alpha_3=\frac{\mu_3}{\sigma^3}$$

和参数 "峰度" (peakedness 或 kurtosis)

$$\alpha_4=\frac{\mu_4}{\sigma^4},$$

其中

$$\mu_k=\int_{-\infty}^{\infty}(x-\mu)^kf(x)\mathrm{d}x,\ \mu=\int_{-\infty}^{\infty}xf(x)\mathrm{d}x,\quad\sigma^2=\mu_2.$$

对于表 2 – 3 中所引的分布，讨论关于参数 α_3 和 α_4 的值的问题.

24. 对于服从参数 $\beta=1$ 的 Γ 分布的随机变量 X（见表 2 – 3），证明

$$\mathbf{E}X^k=\frac{\Gamma(k+\alpha)}{\Gamma(\alpha)}.$$

特别，$\mathbf{E}X=\alpha$，$\mathbf{E}X^2=\alpha(\alpha+1)$，因此 $\mathbf{D}X=\alpha$.

当 $\beta\neq1$ 时求类似的公式.

25. 对于服从 B 分布的随机变量 X (见表 2－3), 证明

$$\mathrm{E}X^k = \frac{\mathrm{B}(r+k,s)}{\mathrm{B}(r,s)}.$$

26. 对于二项分布, 试验次数 n 固定, 考虑在 n 次试验中 "成功" 次数 ν 恰好等于 r 的概率 $\mathbf{P}_n\{\nu=r\}$. 此概率 $\mathbf{P}_n\{\nu=r\}=\mathrm{C}_n^r p^r q^{n-r}, 0\leqslant r\leqslant n$, 其中 p 是每次试验 "成功" 的概率. 这些概率形成二项分布 (n 给定). 如果考虑问题 "最早出现 r 次 '成功' (随机地) 发生在第 $\tau=k(k\geqslant r)$ 次试验" 的概率, 将导出负二项分布 (或逆二项分布). 证明, 事件 $\{\tau=k\}$ 的概率为

$$\mathbf{P}^r\{\tau=k\}=\mathrm{C}_{k-1}^{r-1}p^r q^{k-r},\quad k=r,r+1,\cdots,$$

其中 $r=1,2,\cdots$ (p 是每次试验 "成功" 的概率). 这些概率 $\mathbf{P}^r\{\tau=k\}, k=r,r+1,\cdots$, 的全体形成负二项分布. 证明, 对于给定的 r, $\mathbf{E}^r\tau=rq/p$.

§4.　随机变量 I

1. 随机变量及其概率分布和分布函数　设 $(\Omega,\mathscr{F})$ 是某一可测空间, $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ 是数轴及其博雷尔集系 $\mathscr{B}(\mathbb{R})$ 的可测空间.

定义 1　定义在 $(\Omega,\mathscr{F})$ 上的实函数 $\xi=\xi(\omega)$, 称做 $\mathscr{F}$－可测函数或随机变量, 如果对于任意 $B\in\mathscr{B}(\mathbb{R})$, 有

$$\{\omega:\xi(\omega)\in B\}\in\mathscr{F}, \tag{1}$$

或者, 如果 $\xi^{-1}(B)\equiv\{\omega:\xi(\omega)\in B\}$ 是 Ω 中的可测集.

当 $(\Omega,\mathscr{F})=(\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n))$ 时, $\mathscr{B}(\mathbb{R}^n)$－可测函数称做博雷尔可测函数.

任意 (可测) 集合 $A\in\mathscr{F}$ 的示性函数 $I_A(\omega)$, 是随机变量最简单的例子.

表示为

$$\xi(\omega)=\sum_{i=1}^{\infty}x_i I_{A_i}(\omega) \tag{2}$$

的随机变量, 其中 $\sum A_i=\Omega$, $A_i\in\mathscr{F}$, 称做离散型随机变量. 假如 (2) 式中的和含有限项, 则相应的随机变量称做简单的.

按照第一章 §4 的解释, 可以说随机变量试验的某一数量特征, 其取值依赖于 "偶然" ω. 这时, 可测性 (1) 的要求之所以重要, 其原因是: 如果在 $(\Omega,\mathscr{F})$ 上给定了概率, 则事件 $\{\omega:\xi(\omega)\in B\}$ 的概率才有意义, 其中 $\{\omega:\xi(\omega)\in B\}$ 表示 "随机变量 ξ 的值属于某一给定的博雷尔集合 B".

于是, 我们给出如下定义.

定义 2　$(\mathbb{R},\mathscr{B}(\mathbb{R}))$ 上的概率测度 P_ξ 连同

$$P_\xi(B)=\mathbf{P}\{\omega:\xi(\omega)\in B\}, B\in\mathscr{B}(\mathbb{R}),$$

称做随机变量 ξ 在 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ 上的概率分布.

定义 3　函数

$$F_\xi(x)=\mathbf{P}\{\omega:\xi(\omega)\leqslant x\}, x\in\mathbb{R},$$

称做随机变量 ξ 的分布函数.

对于离散型随机变量 ξ, 测度 P_ξ 集中在有限或可数个点上, 并且可以表示为

$$P_\xi(B)=\sum_{\{k:x_k\in B\}}p(x_k), \tag{3}$$

其中 $p(x_k)=\mathbf{P}\{\xi=x_k\}=\Delta F_\xi(x_k)$.

显然, 也有相反的结论: 如果 P_ξ 可以表示为 (3) 式, 则 ξ 是离散型随机变量.

随机变量 ξ 称做连续的, 如果其分布函数 $F_\xi(x)$ 对 $x\in\mathbb{R}$ 连续.

随机变量 ξ 称做绝对连续的[①], 如果存在非负函数 $f=f_\xi(x)$, 使

$$F_\xi(x)=\int_{-\infty}^{x}f_\xi(y)\mathrm{d}y, x\in\mathbb{R} \tag{4}$$

(积分可以理解为黎曼积分, 不过在更一般的情形下, 应理解为勒贝格积分; 参见下面 §6), 其中函数 $f=f_\xi(x)$ 称做密度.

2. 函数 $\xi=\xi(\omega)$ 为 $\mathscr{F}$－可测的充分和必要条件　判断函数 $\xi=\xi(\omega)$ 是否随机变量, 需要对于所有集合 $B\in\mathscr{F}$ 验证性质 (1) 是否成立. 下面的引理表明, 对于这样集合类的 "测试" 可以简化.

引理 1　*设 $\mathscr{E}$ 是某一集系, 且 $\sigma(\mathscr{E})=\mathscr{B}(\mathbb{R})$. 函数 $\xi=\xi(\omega)$ 为 $\mathscr{F}$－可测的充分和必要条件是: 对于一切 $E\in\mathscr{E}$, 有*

$$\{\omega:\xi(\omega)\in E\}\in\mathscr{F}. \tag{5}$$

证明　必要性显然. 为证明充分性仍利用适当集合原理 (§2).

设 $\mathscr{D}$ 是满足 $\xi^{-1}(D)\in\mathscr{F}, D\in\mathscr{B}(\mathbb{R})$ 的博雷尔集系. 不难验证 "取逆像" 运算, 保持并、交与补的集合运算不变:

$$\begin{aligned}\xi^{-1}\left(\bigcup_\alpha B_\alpha\right)&=\bigcup_\alpha\xi^{-1}(B_\alpha),\\ \xi^{-1}\left(\bigcap_\alpha B_\alpha\right)&=\bigcap_\alpha\xi^{-1}(B_\alpha),\\ \xi^{-1}(\overline{B}_\alpha)&=\overline{\xi^{-1}(B_\alpha)}.\end{aligned} \tag{6}$$

① 在我国文献中, 亦常称绝对随机变量为连续型随机变量.——译者

由此可见，集系 $\mathscr{D}$ 是 σ-代数. 因此

$$\mathscr{E}\subseteq\mathscr{D}\subseteq\mathscr{B}(\mathbb{R}),$$

故

$$\sigma(\mathscr{E})\subseteq\sigma(\mathscr{D})=\mathscr{D}\subseteq\mathscr{B}(\mathbb{R}).$$

由于 $\sigma(\mathscr{E})=\mathscr{B}(\mathbb{R})$，从而 $\mathscr{D}=\mathscr{B}(\mathbb{R})$.

系 函数 $\xi=\xi(\omega)$ 为随机变量的充分和必要条件是，对于任何 $x\in\mathbb{R}$，有

$$\{\omega:\xi(\omega)<x\}\in\mathscr{F}$$

或

$$\{\omega:\xi(\omega)\leqslant x\}\in\mathscr{F}.$$

由如下事实可以立即得到系的证明：每一个集系

$$\mathscr{E}_1=\{x:x<c,c\in\mathbb{R}\},$$
$$\mathscr{E}_2=\{x:x\leqslant c,c\in\mathbb{R}\}$$

都产生 σ-代数 $\mathscr{B}(\mathbb{R})$，即 $\sigma(\mathscr{E}_1)=\sigma(\mathscr{E}_2)=\mathscr{B}(\mathbb{R})$ (参见 §2).

基于下面的引理，可以用其他随机变量的函数来构造新的随机变量.

引理 2 设 $\varphi=\varphi(x)$ 是博雷尔函数，而 $\xi=\xi(\omega)$ 是一随机变量. 那么，复合函数 $\eta=\varphi\circ\xi$，即函数 $\eta(\omega)=\varphi(\xi(\omega))$ 也是随机变量.

证明 引理的结论由如下事实可得：对于 $B\in\mathscr{B}(\mathbb{R})$，

$$\{\omega:\eta(\omega)\in B\}=\{\omega:\varphi(\xi(\omega))\in B\}=\{\omega:\xi(\omega)\in\varphi^{-1}(B)\}\in\mathscr{F},\tag{7}$$

因为 $\varphi^{-1}(B)\in\mathscr{B}(\mathbb{R})$. □

这样，由于 $x^n,x^+,x^-,|x|$ 是博雷尔函数，可见如果 ξ 是随机变量，则 $\xi^n,\xi^+=\max\{\xi,0\},\xi^-=-\min\{\xi,0\},|\xi|$ 也都是随机变量 (练习题 3).

3. 广义随机变量[①] 从给定的随机变量组 $\{\xi_n\}$ 出发，可以建立随机变量新的函数，例如，

$$\sum_{n=1}^{\infty}|\xi_n|,\ \overline{\lim_n}\,\xi_n,\ \underline{\lim_n}\,\xi_n,\ \cdots$$

注意，这些函数一般已经在扩充数轴 $[-\infty,\infty]$ 上取值. 因此，最好将可测函数类加以扩充，使之也可以取 $\pm\infty$ 为值.

① Расширенная случайная величина (extended random variable)，亦可译为"扩充随机变量"——译者

定义 4 称定义在 $(\Omega,\mathscr{F})$ 上，取值于 $\overline{\mathbb{R}}=[-\infty,\infty]$ 上的函数 $\xi=\xi(\omega)$ 为广义随机变量，如果对于任何博雷尔集 $B\in\mathscr{B}(\overline{\mathbb{R}})$ (σ-代数 $\mathscr{B}(\overline{\mathbb{R}})=\sigma(\mathscr{B}(\mathbb{R}),\pm\infty)$) 满足条件 (1).

下面的定理虽然简单，但是在建立勒贝格积分时是关键 (§6).

定理 1 a) 对于任意 (包括广义) 随机变量 $\xi=\xi(\omega)$，存在简单随机变量列 $\xi_1,\xi_2,\cdots$，使 $|\xi_n|\leqslant|\xi|$，且对于一切 $\omega\in\Omega$，当 $n\to\infty$ 时 $\xi_n(\omega)\to\xi(\omega)$.

b) 在上述条件下，如果 $\xi(\omega)\geqslant 0,\omega\in\Omega$，则存在简单随机变量序列 $\xi_1,\xi_2,\cdots$，使对于一切 $\omega\in\Omega$，当 $n\to\infty$ 时 $\xi_n(\omega)\uparrow\xi(\omega)$.

证明 首先证明第 2 个命题. 对于 $n=1,2,\cdots$，设

$$\xi_n(\omega)=\sum_{k=1}^{n2^n}\frac{k-1}{2^n}I_{\{\frac{k-1}{2^n}\leqslant\xi(\omega)<\frac{k}{2^n}\}}(\omega)+nI_{\{\xi(\omega)\geqslant n\}}(\omega).$$

可以直接验证，对于一切 $\omega\in\Omega$，所建立的序列 $\xi_n(\omega)\uparrow\xi(\omega)$. 如果注意到，$\xi$ 可以表示为 $\xi=\xi^+-\xi^-$，则由已经证明的命题 b)，可以立即得到命题 a). □

现在证明，广义随机变量类关于逐项收敛封闭. 为此，首先指出，如果 $\xi_1,\xi_2,\cdots$ 是广义随机变量序列，则 $\sup\xi_n,\inf\xi_n,\overline{\lim}\,\xi_n$ 和 $\underline{\lim}\,\xi_n$ 也是随机变量 (有可能是广义的). 由

$$\{\omega:\sup\xi_n>x\}=\bigcup_n\{\omega:\xi_n>x\}\in\mathscr{F},$$
$$\{\omega:\inf\xi_n<x\}=\bigcup_n\{\omega:\xi_n<x\}\in\mathscr{F};$$
$$\overline{\lim}\,\xi_n=\inf_n\sup_{m\geqslant n}\xi_m,\quad \underline{\lim}\,\xi_n=\sup_n\inf_{m\geqslant n}\xi_m,$$

可以直接得到上面所指出的性质.

定理 2 如果 $\xi_1,\xi_2,\cdots$ 是扩充随机变量序列，而 $\xi(\omega)=\lim\xi_n(\omega),\omega\in\Omega$，则 $\xi=\xi(\omega)$ 也是广义随机变量.

证明 定理的证明，可以立即由上面的结果和下面的关系式得到：

$$\begin{aligned}\{\omega:\xi_n<x\}&=\{\omega:\lim\xi_n(\omega)<x\}\\&=\{\omega:\overline{\lim}\,\xi_n(\omega)=\underline{\lim}\,\xi_n(\omega)\}\cap\{\overline{\lim}\,\xi_n(\omega)<x\}\\&=\Omega\cap\{\overline{\lim}\,\xi_n(\omega)<x\}=\{\overline{\lim}\,\xi_n(\omega)<x\}\in\mathscr{F}.\end{aligned}$$

4. 随机变量序列之和、差、积、商的极限 我们现在讨论随机变量的简单函数的一些性质，这些变量定义在可测空间 $(\Omega,\mathscr{F})$ 上，且有可能取值于扩充数轴 $\overline{\mathbb{R}}=[-\infty,\infty]$[*)].

*) 以下关于 $\overline{\mathbb{R}}$ 中的算术运算，作如下通常的约定：若 $a\in\mathbb{R}$，则 $a\pm\infty=\pm\infty$，$a/\pm\infty=0$；若 $a<0$，则 $a\times\infty=-\infty$，此外，$0\times(\pm\infty)=0$，$\infty+\infty=\infty$，$-\infty-\infty=-\infty$.

如果 ξ 和 η 是两个随机变量, 则 $\xi+\eta, \xi-\eta, \xi\eta$ 和 ξ/η 也是随机变量 (假设有关运算有意义, 即不出现形如 $\infty-\infty, \infty/\infty, 0/0$ 等不定式).

设 $\{\xi_n\}$ 和 $\{\eta_n\}$ 是两个随机变量序列, 分别收敛于 ξ 和 η (见定理 1). 那么

$$\xi_n \pm \eta_n \to \xi \pm \eta,$$

$$\xi_n\eta_n \to \xi\eta,$$

$$\frac{\xi_n}{\eta_n + \frac{1}{n} I_{\{\eta_n=0\}}(\omega)} \to \frac{\xi}{\eta}.$$

这些关系式左侧的各个函数都是简单随机变量. 因此, 由定理 2 知 $\xi\pm\eta, \xi\eta, \xi/\eta$ 都是随机变量.

5. 随机变量的函数 设 $\xi=\xi(\omega)$ 是随机变量. 考虑 $\mathscr{F}$ 中形如 $\{\omega:\xi(\omega)\in B\}, B\in\mathscr{B}(\mathbb{R})$ 的集合, 它们构成 $\sigma-$ 代数, 称为由随机变量 ξ 生成的 $\sigma-$ 代数, 记作 $\mathscr{F}_\xi$ 或 $\sigma(\xi)$.

如果 φ 是某一博雷尔函数, 则由引理 2 知函数 $\eta=\varphi\circ\xi$ 也是随机变量, 并且 $\mathscr{F}_\xi-$ 可测: $\{\omega:\eta(\omega)\in B\}\in\mathscr{F}_\xi, B\in\mathscr{B}(\mathbb{R})$ (见 (7) 式). 实际上, 相反的结果也成立.

定理 3 *设 $\eta=\eta(\omega)$ 是 $\mathscr{F}_\xi-$ 可测随机变量. 那么, 存在某一博雷尔函数 φ, 使对于每一个 $\omega\in\Omega$, 有 $\eta=\varphi\circ\xi$, 即 $\eta(\omega)=\varphi(\xi(\omega))$.*

证明 设 Φ_ξ 是一切 $\mathscr{F}_\xi-$ 可测函数 $\eta=\eta(\omega)$ 类, 而 $\tilde{\Phi}_\xi$ 是可以表示为 $\varphi\circ\eta$ 的 $\mathscr{F}_\xi-$ 可测函数类, 其中 φ 是某一博雷尔函数. 显然, $\tilde{\Phi}_\xi\subseteq\Phi_\xi$. 定理的结论表明, 实际上是 $\tilde{\Phi}_\xi=\Phi_\xi$.

设 $A\in\mathscr{F}_\xi$ 和 $\eta(\omega)=I_A(\omega)$. 现在证明 $\eta\in\tilde{\Phi}_\xi$. 事实上, 若 $A\in\mathscr{F}_\xi$, 则存在 $B\in\mathscr{B}(\mathbb{R})$, 使 $A=\{\omega:\xi(\omega)\in B\}$. 记

$$\chi_B(x)=\begin{cases}1, & 若\ x\in B,\\ 0, & 若\ x\notin B.\end{cases}$$

那么 $I_A(\omega)=\chi_B(\xi(\omega))\in\tilde{\Phi}_\xi$. 由此可见, 任何 $\mathscr{F}_\xi-$ 可测简单函数

$$\sum_{i=1}^{n} c_i I_{A_i}(\omega), A_i\in\mathscr{F}_\xi$$

仍然属于 $\tilde{\Phi}_\xi$.

现在设 η 是任意 $F_\xi-$ 可测简单函数. 根据定理 1, 存在 $\mathscr{F}_\xi-$ 可测的简单函数序列 $\{\eta_n\}$: $\eta_n=\eta_n(\omega)$, 使得当 $n\to\infty$ 时 $\eta_n(\omega)\to\eta(\omega), \omega\in\Omega$. 由上面证明的结果知, 存在博雷尔函数 $\varphi_n=\varphi_n(x)$, 使 $\eta_n=\varphi_n(\xi(\omega))$. 在这种情况下当 $n\to\infty$ 时 $\varphi_n(\xi(\omega))\to\eta(\omega), \omega\in\Omega$.

记 $B=\{x\in\mathbb{R}:\lim_n\varphi_n(x)存在\}$. 此集合是博雷尔集, 所以函数

$$\varphi(x)=\begin{cases}\lim\limits_n\varphi_n(x), & 若\ x\in B,\\ 0, & 若\ x\notin B\end{cases}$$

仍然是博雷尔函数 (见练习题 6).

那么, 显然对于所有 $\omega\in\Omega$, 有

$$\eta(\omega)=\lim_n\varphi_n(\xi(\omega))=\varphi(\xi(\omega)).$$

于是, $\tilde{\Phi}_\xi=\Phi_\xi$. □

6. 阶梯随机变量 考虑可测空间 $(\Omega,\mathscr{F})$, 和空间 Ω 的有限或可数分割: $\mathscr{D}=\{D_1,D_2,\cdots\}, D_i\in\mathscr{F}, \sum_i D_i=\Omega$. 组成含空集 $\varnothing$ 和 $\sum_\alpha D_\alpha$ 的代数 $\mathscr{A}$, 其中和式含有限或可数项. 显然, 集系 $\mathscr{A}$ 是单调类, 因此根据 §2 引理 2, 代数 $\mathscr{A}$ 同时也是 $\sigma-$ 代数, 记作 $\sigma(\mathscr{D})$, 称为由 $\mathscr{D}$ 生成的 $\sigma-$ 代数. 显然 $\sigma(\mathscr{D})\subseteq F$.

引理 3 *设 $\xi=\xi(\omega)$ 是 $\sigma(\mathscr{D})-$ 可测随机变量, 则 ξ 可以表示为:*

$$\xi(\omega)=\sum_{k=1}^{\infty}x_k I_{D_k}(\omega), \tag{8}$$

其中 $x_k\in\mathbb{R}, k\geqslant 1$, 即 $\xi(\omega)$ 在分割的元素 $D_k, k\geqslant 1$ 上为常数.

证明 任取分割的一个集合 D_k, 证明 $\sigma(\mathscr{D})-$ 可测函数 ξ 在 D_k 上为常数.

为此, 设

$$x_k=\sup\{c: D_k\cap\{\omega:\xi(\omega)<c\}=\varnothing\}.$$

由于

$$\{\omega:\xi(\omega)<x_k\}=\bigcup_{r<x_k}\{\omega:\xi(\omega)<r\},$$

而 $r<x_k$ 是有理数, 则 $D_k\bigcap\{\omega:\xi(\omega)<x_k\}=\varnothing$.

现在设 $c>x_k$, 则 $D_k\bigcap\{\omega:\xi(\omega)<c\}\neq\varnothing$. 由于集合 $\{\omega:\xi(\omega)<c\}$ 具有 $\sum_\alpha D_\alpha$ 的形式, 其中对于有限或可数个下标求和, 故

$$D_k\cap\{\omega:\xi(\omega)<c\}=D_k.$$

由此可见, 对于一切 $c>x_k$,

$$D_k\cap\{\omega:\xi(\omega)\geqslant c\}=\varnothing;$$

由于

$$\{\omega:\xi(\omega)>x_k\}=\bigcup_{r>x_k}\{\omega:\xi(\omega)\geqslant r\},$$

其中对于有限或可数个下标求和. 故

$$D_k \cap \{\omega : \xi(\omega) = x_k\} = \varnothing.$$

于是, $D_k \cap \{\omega : \xi(\omega) \neq x_k\} = \varnothing$. 从而

$$D_k \subseteq \{\omega : \xi(\omega) = x_k\}.$$

而这正是需要证明的. □

7. 练习题

1. 证明随机变量 ξ 连续的充分和必要条件是: 对于一切 $x \in \mathbb{R}$, $\mathbf{P}\{\xi = x\} = 0$.

2. 设 $|\xi|$ 为 $\mathscr{F}-$ 可测, 问 ξ 是否也 $\mathscr{F}-$ 可测?

3. 证明函数 $x^+, x^+ = \max(x, 0)$, $x^- = -\min(x, 0)$, $|x| = x^+ + x^-$ 是博雷尔函数.

4. 如果 ξ 和 η 都 $\mathscr{F}-$ 可测, 则 $\{\omega : \xi(\omega) = \eta(\omega)\} \in \mathscr{F}$.

5. 设 ξ 和 η 是 $(\Omega, \mathscr{F})$ 上的两个随机变量, 而集合 $A \in \mathscr{F}$. 证明, 函数

$$\zeta(\omega) = \xi(\omega) I_A + \eta(\omega) I_{\bar{A}}$$

也是随机变量.

6. 设 $\xi_1, \cdots, \xi_n$ 是随机变量, 而 $\varphi(x_1, \cdots, x_n)$ 是博雷尔函数. 证明 $\varphi(\xi_1(\omega), \cdots, \xi_n(\omega))$ 也是随机变量.

7. 设 ξ 和 η 是以 $1, 2, \cdots, N$ 为值的两个随机变量, 且 $\mathscr{F}_\xi = \mathscr{F}_\eta$. 证明存在数列的 $(1, 2, \cdots, N)$ 的排列 $(i_1, i_2, \cdots, i_N)$, 使对于每一个 $j = 1, 2, \cdots, N$, 集合 $\{\omega : \xi = j\}$ 与 $\{\omega : \eta = i_j\}$ 相等.

8. 举一随机变量 ξ 的例, 使其分布函数有密度 $f(x)$, 但是当 $x \to \infty$ 时 $f(x)$ 的极限 $\lim_{x\to\infty} f(x)$ 不存在, 从而 $f(x)$ 在无穷大不为 0.

9. 设 ξ 和 η 是有界随机变量 $(|\xi| \leqslant c_1, |\eta| \leqslant c_2)$. 证明, 若对于一切 $m, n \geqslant 1$, 有

$$\mathbf{E}\xi^m\eta^n = \mathbf{E}\xi^m \times \mathbf{E}\eta^n,$$

则 ξ 和 η 独立.

10. 设 ξ 和 η 是随机变量, 其分布函数 $F_\xi = F_\eta$. 证明, 若对 $x \in \mathbb{R}$, $\{\omega : \xi(\omega) = x\} \neq \varnothing$, 则存在 $y \in \mathbb{R}$, 使 $\{\omega : \xi(\omega) = x\} = \{\omega : \eta(\omega) = y\}$.

11. 设 E 是 $\mathbb{R}$ 上的有限或可数集合 ξ 是映射: $\Omega \to E$. 证明 ξ 是 $(\Omega, \mathscr{F})$ 上是随机变量, 当且仅当对于每一个 $x \in E$, $\{\omega : \xi(\omega) = x\} \in \mathscr{F}$.

§5. 随机元

1. 随机函数、随机向量和随机过程 除随机变量外, 在概率论及其应用中, 还研究更一般性质的随机对象. 例如, 随机点、随机向量、随机函数、随机过程、随机

场、随机集合、随机测度等等. 因此, 希望有一个关于任意性质的随机对象的一般概念.

定义 1 设 $(\Omega, \mathscr{F})$ 和 $(E, \mathscr{E})$ 是两个可测空间. 定义在 Ω 上取值于 E 的函数 $X = X(\omega)$, 称做 (取值于 E 的) $\mathscr{F}/\mathscr{E}$ 可测函数或随机元. 如果对于任何 $B \in \mathscr{E}$, 有

$$\{\omega : X(\omega) \in B\} \in \mathscr{F}. \tag{1}$$

取值于 E 的随机元, 亦常称为 E- 值随机变量.

下面讨论这一定义的一些特殊情形.

假如 $(E, \mathscr{E}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, 则随机元的定义与随机变量的定义一致 (§4).

假如 $(E, \mathscr{E}) = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$, 则随机元 $X(\omega)$ 是 $\mathbb{R}^n$ 中的 "随机点". 如果 π_k 是 $\mathbb{R}^n$ 在第 k 坐标轴上的射影, 则 $X(\omega)$ 可以表示为

$$X(\omega) = (\xi_1(\omega), \cdots, \xi_n(\omega)), \tag{2}$$

其中 $\xi_k = \pi_k \circ X$.

由条件 (1) 可见, ξ_k 是普通的随机变量. 事实上, 对于任意 $B \in \mathscr{B}(\mathbb{R})$, 由于 $\mathbb{R} \times \cdots \times \mathbb{R} \times B \times \mathbb{R} \times \cdots \times \mathbb{R} \in \mathscr{B}(\mathbb{R}^n)$, 有

$$\begin{aligned}
&\{\omega : \xi_k(\omega) \in B\} \\
&= \{\omega : \xi_1(\omega) \in \mathbb{R}, \cdots, \xi_{k-1}(\omega) \in \mathbb{R}, \xi_k(\omega) \in B, \xi_{k+1}(\omega) \in \mathbb{R}, \cdots, \xi_n(\omega) \in \mathbb{R}\} \\
&= \{\omega : X(\omega) \in (\mathbb{R} \times \cdots \times \mathbb{R} \times B \times \mathbb{R} \times \cdots \times \mathbb{R})\} \in \mathscr{F}.
\end{aligned}$$

定义 2 我们把任意有序随机变量组 $(\eta_1(\omega), \cdots, \eta_n(\omega))$, 称做 *n* *– 维随机向量*.

按照这一定义, 任意在 $\mathbb{R}^n$ 取值的随机元 $X(\omega)$ 是 $n-$ 维随机向量. 反过来说, 任何 $n-$ 维随机向量 $X(\omega) = (\xi_1(\omega), \cdots, \xi_n(\omega))$, 是 $\mathbb{R}^n$ 中的随机元. 事实上, 如果 $B_k \in \mathscr{B}(\mathbb{R})$, $k = 1, \cdots, n$, 则

$$\{\omega : X(\omega) \in (B_1 \times \cdots \times B_n)\} = \prod_{k=1}^{n} \{\omega : \xi_k(\omega) \in B_k\} \in \mathscr{F}.$$

含 $B_1 \times \cdots \times B_n$ 的最小 $\sigma-$ 代数等于 $\mathscr{B}(\mathbb{R}^n)$. 那么, 由 §4 中引理 1 的明显推广立即可得, 对于任意 $B \in \mathscr{B}(\mathbb{R}^n)$, 集合 $\{\omega : X(\omega) \in B\} \in \mathscr{F}$.

设 $(E, \mathscr{E}) = (\mathbf{Z}, \mathscr{B}(\mathbf{Z}))$, 其中 $\mathbf{Z}$ 是复数 $z = x + iy$, $x, y \in \mathbb{R}$ 的集合, 而 $\mathscr{B}(\mathbf{Z})$ 含形如 $\{z : z = x + iy, a_1 < x \leqslant b_1, a_2 < y \leqslant b_2\}$ 集合的最小 $\sigma-$ 代数. 由以上的讨论, 可见复值随机变量 $Z(\omega)$ 可以表示为 $Z(\omega) = X(\omega) + iY(\omega)$, 其中 $X(\omega)$ 和 $Y(\omega)$ 是随机变量. 因此, $Z(\omega)$ 亦称为*复随机变量*.

设 $(E, \mathscr{E}) = (\mathbb{R}^T, \mathscr{B}(\mathbb{R}^T))$, 其中 T 是数轴的某一子集. 这时, 任何可以表示为 $X = (\xi_t)_{t \in T}$, $\xi_t = \pi_t \circ X$ 的随机元 $X = X(\omega)$, 称做定义在时间区间 T 上的*随机函数*.

像随机向量一样, 任何随机函数同时也是随机过程. 见下面的定义.

定义 3 设 T 是数轴的某一子集. 随机变量组 $X = (\xi_t)_{t\in T}$ 称做时间区间 T 上的随机过程.

对于 $T = \{1, 2, \cdots\}$ 的情形, $X = (\xi_1, \xi_2, \cdots)$ 称为离散时间随机过程或随机序列.

对于 $T = [0, 1], (-\infty, \infty), [0, \infty), \cdots$ 的情形, $X = (\xi_t)_{t\in T}$ 称为连续时间随机过程.

利用 $\sigma-$ 代数 $\mathscr{B}(\mathbb{R}^T)$ 的构造 (§2), 不难证明, (在定义 3 的意义下的) 任何随机过程 $X = (\xi_t)_{t\in T}$, 同时也是随机函数 (值域为 $\mathbb{R}^T$ 的随机元).

定义 4 设 $X = (\xi_t)_{t\in T}$ 是随机过程. 对于每个固定的 $\omega \in \Omega$, 函数 $(\xi_t(\omega))_{t\in T}$ 称做过程对应于结局 ω 的实现或轨道.

与 §4 的定义 3 类似, 自然地引出下面的定义.

定义 5 设 $X = (\xi_t)_{t\in T}$ 是随机过程.

1) $(\mathbb{R}^T, \mathscr{B}(\mathbb{R}^T))$ 上的概率测度 P_X, 其中

$$P_X(B) = \mathbf{P}\{\omega : X(\omega) \in B\}, \quad B \in \mathscr{B}(\mathbb{R}^T),$$

称做过程 X 的概率分布.

2) 对于 $t_1 < t_2 \cdots < t_n, t_i \in T$, 概率

$$P_{t_1, \cdots, t_n}(B) = \mathbf{P}\{\omega : (\xi_{t_1}, \cdots, \xi_{t_n}) \in B\}, B \in \mathscr{B}(\mathbb{R}^n),$$

称做过程 X 有限维概率 (或有限维概率分布).

3) 对于 $t_1 < t_2 \cdots < t_n, t_i \in T$ 函数

$$F_{t_1, \cdots, t_n}(x_1, \cdots, x_n) = \mathbf{P}\{\omega : \xi_{t_1} \leqslant x_1, \cdots, \xi_{t_n} \leqslant x_n\}$$

称做过程 $X = (\xi_t)_{t\in T}$ 的有限维概率分布函数.

设 $(E, \mathscr{E}) = (C, \mathscr{B}_0(C))$, 其中 C 是区间 $T = [0, 1]$ 上连续函数 $x = (x_t)_{t\in T}$ 的空间, 其中 $\sigma-$ 代数 $\mathscr{B}_0(C)$ 是由开集生成的 (§2). 下面证明, 空间 $(C, \mathscr{B}_0(C))$ 的任意随机元 X, 同时也是在定义 3 意义上的随机过程 (且具有连续轨道).

实际上根据 §2, 集合 $A = \{x \in C : x_t < a\}$ 是 $\mathscr{B}_0(C)$ 中的开集. 因此

$$\{\omega : \xi_t(\omega) < a\} = \{\omega : X(\omega) \in A\} \in \mathscr{F}.$$

另一方面, 设 $X = (\xi_t(\omega))_{t\in T}$ 是 (在定义 3 意义上的) 随机过程, 对于每个 $\omega \in \Omega$, 其轨道是连续函数. 根据 §2 的 (17) 式

$$\{x \in C : x \in S_\rho(X^0(\omega))\} = \bigcap_{t_k}\{x \in C : |x_{t_k} - x^0_{t_k}| < \rho\},$$

其中 t_k 是线段 $[0, 1]$ 上的有理点. 因此,

$$\{\omega : X(\omega) \in S_\rho(X^0(\omega))\} = \bigcap_{t_k}\{\omega : |\xi_{t_k}(\omega) - \xi^0_{t_k}(\omega)| < \rho\} \in \mathscr{F}.$$

于是, 对于任何 $\{\omega : X(\omega) \in B\} \in \mathscr{F}$, 有 $B \in \mathscr{B}_0(C)$.

通过类似的讨论也可以证明, 在 §2 (第 7 小节) 中引进的空间 $(D, \mathscr{B}_0(D))$ 上的随机元, 可以视为随机过程, 其轨道属于无第一类间断点的函数空间, 并且逆命题成立.

2. 随机元的独立性 设 $(\Omega, \mathscr{F}, \mathbf{P})$ 是概率空间, 而 $(E_\alpha, \mathscr{E}_\alpha)$ 是可测空间, 其中 α 的值域是某 (任意) 集合 $\mathfrak{A}$.

定义 6 称 $\mathscr{F}/\mathscr{E}_\alpha$ 可测函数 $(X_\alpha(\omega)), \alpha \in \mathfrak{A}$, 独立 (或全体独立), 如果对于任何有限下标数组 $\alpha_1, \cdots, \alpha_n$, 随机元 $X_{\alpha_1}, \cdots, X_{\alpha_n}$ 独立, 即

$$\mathbf{P}\{X_{\alpha_1} \in B_{\alpha_1}, \cdots, X_{\alpha_n} \in B_{\alpha_n}\} = \mathbf{P}\{X_{\alpha_1} \in B_{\alpha_1}\} \cdots \mathbf{P}\{X_{\alpha_n} \in B_{\alpha_n}\}, \tag{3}$$

其中 $B_\alpha \in \mathscr{E}_\alpha$.

设 $\mathfrak{A} = \{1, 2, \cdots, n\}$, ξ_α 是随机变量, $\alpha \in \mathfrak{A}$, 而

$$F_\xi(x_1, \cdots, x_n) = \mathbf{P}\{\xi_1 \leqslant x_1, \cdots, \xi_n \leqslant x_n\}$$

是向量 $(\xi_1, \cdots, \xi_n)$ 的 $n-$ 维分布函数. 设 $F_{\xi_i}(x_i)$ 是随机变量 $\xi_i, i = 1, \cdots, n$, 的分布函数.

定理 随机变量 $\xi_1, \cdots, \xi_n$ 独立的充分和必要条件是, 对于所有向量 $(x_1, \cdots, x_n) \in \mathbb{R}^n$, 有

$$F_\xi(x_1, \cdots, x_n) = F_{\xi_1}(x_1) \cdots F_{\xi_n}(x_n). \tag{4}$$

证明 必要性显然. 为证明充分性, 设 $a = (a_1, \cdots, a_n), b = (b_1, \cdots, b_n)$ 是任意向量, 则

$$\mathbf{P}_\xi(a, b] = \mathbf{P}\{a_1 < \xi_1 \leqslant b_1, \cdots, a_n < \xi_n \leqslant b_n\},$$
$$\mathbf{P}_{\xi_i}(a_i, b_i] = \mathbf{P}\{a_i < \xi_i \leqslant b_i\}.$$

那么, 由 §3 的 (7) 式, 以及 (4) 式, 得

$$\mathbf{P}_\xi(a, b] = \prod_{i=1}^n [F_{\xi_i}(b_i) - F_{\xi_i}(a_i)] = \prod_{i=1}^n P_{\xi_i}(a_i, b_i].$$

从而,

$$\mathbf{P}\{\xi_1 \in I_1, \cdots, \xi_n \in I_n\} = \prod_{i=1}^n \mathbf{P}\{\xi_i \in I_i\}, \tag{5}$$

其中 $I_i = (a_i, b_i]$.

固定 $I_2, \cdots, I_n$, 对于任意 $B_1 \in \mathscr{B}(R)$, 有

$$\mathbf{P}\{\xi_1 \in B_1, \xi_2 \in I_2, \cdots, \xi_n \in I_n\} = \mathbf{P}\{\xi_1 \in B_1\} \prod_{i=2}^{n} \mathbf{P}\{\xi_i \in I_i\}. \tag{6}$$

设 $\mathscr{M}$ 是 $\mathscr{B}(R)$ 中满足 (6) 式的集合的全体 (§2, "适当集合原理"). 由不相交的形如 $I_1 = (a_1, b_1]$ 的区间形成的代数 $\mathscr{A}$ 显然属于 $\mathscr{M}$. 因此 $\mathscr{A} \subseteq \mathscr{M} \subseteq \mathscr{B}(\mathbf{R})$. 由概率测度的可数可加性 (从而也由其连续性), 可见集系 $\mathscr{M}$ 是单调类. 因此 (见 §2, 第 1 小节)

$$\mu(\mathscr{A}) \subseteq \mathscr{M} \subseteq \mathscr{B}(R).$$

然而, 根据 §2 的定理 1, $\mu(\mathscr{A}) = \sigma(\mathscr{A}) = \mathscr{B}(\mathbf{R})$, 故 $\mathscr{M} = \mathscr{B}(\mathbf{R})$.

于是 (6) 式得证. 现在固定 $B_1, I_3, \cdots, I_n$, 并将 I_2 换成 B_2, 运用与证明 (6) 式同样的方法继续这一过程, 则显然可以得到需要的等式:

$$\mathbf{P}\{\xi_1 \in B_1, \cdots, \xi_n \in B_n\} = \mathbf{P}\{\xi_1 \in B_1\} \cdots \mathbf{P}\{\xi_n \in B_n\},$$

其中 $B_i \in \mathscr{B}(R)$. □

3. 练习题

1. 设 $\xi_1, \cdots, \xi_n$ 是离散型随机变量, 证明它们独立的充分和必要条件是, 对于任意实数 $x_1, \cdots, x_n$, 有

$$\mathbf{P}\{\xi_1 = x_1, \cdots, \xi_n = x_n\} = \prod_{i=1}^{n} \mathbf{P}\{\xi_i = x_i\}.$$

2. 证明任何随机函数 $X(\omega) = (\xi_t(\omega))_{t \in T}$ (在定义 3 的意义上) 是随机过程, 而且反之亦然.

3. 设 $X_1, \cdots, X_n$ 是分别在 $(E_1, \mathscr{E}_1), \cdots, (E_n, \mathscr{E}_n)$ 取值的随机元. 其次, 设 $(E_1', \mathscr{E}_1'), \cdots, (E_n', \mathscr{E}_n')$ 是可测空间, 而 $g_1, \cdots, g_n$ 相应为 $\mathscr{E}_1/\mathscr{E}_1', \cdots, \mathscr{E}_n/\mathscr{E}_n'$ — 可测函数. 证明, 若 $X_1, \cdots, X_n$ 独立, 则 $g_1 \circ X_1, \cdots, g_n \circ X_n$ 也独立.

4. 设 $X_1, X_2, \cdots$ 是可交换无限随机变量序列 (即这样的序列, 每一组随机变量, 由具有不同下标的 k 个元素构成, 例如, $X_{i_1}, \cdots, X_{i_k}$ 只与 k 有关, 与 $i_1, \cdots, i_k$ 具体值无关, 其中 $i_1, \cdots, i_k$ 两两不等; 对照 §1 的练习题 11). 证明, 若 $\mathbf{E}X_n^2 < \infty, n \geqslant 1$, 则协方差 $\operatorname{cov}(X_1, X_2) \geqslant 0$.

5. 设 ξ, η, ζ 是独立随机变量, 证明随机变量 $\xi + \eta$ 与 ζ^2 独立.

6. 设由随机变量 $\xi_1, \cdots, \xi_m, \eta_1, \cdots, \eta_n$ 组成随机向量 $X = (\xi_1, \cdots, \xi_m)$ 和 $Y = (\eta_1, \cdots, \eta_n)$, 假设满足下列条件:

(i) 随机变量 $\xi_1, \cdots, \xi_m$ 独立;

(ii) 随机变量 $\eta_1, \cdots, \eta_n$ 独立;

(iii) 随机向量 X 和 Y, 作为分别取值于 $\mathbf{R}^m$ 和 $\mathbf{R}^n$ 的随机元独立.

证明随机变量 $\xi_1, \cdots, \xi_m, \eta_1, \cdots, \eta_n$ 独立.

7. 假设有两个随机向量 $X = (\xi_1, \cdots, \xi_m)$ 和 $Y = (\eta_1, \cdots, \eta_n)$. 已知随机变量 $\xi_1, \cdots, \xi_m, \eta_1, \cdots, \eta_n$ 独立.

(i) 证明, 随机向量 X 和 Y, 作为随机元独立 (对照练习题 6).

(ii) 设 $f: \mathbf{R}^m \to \mathbf{R}, g: \mathbf{R}^n \to \mathbf{R}$ 是博雷尔函数, 证明随机变量 $f(\xi_1, \cdots, \xi_m)$ 和 $g(\eta_1, \cdots, \eta_n)$ 独立.

§6. 勒贝格积分. 数学期望

1. 引言与记号 在第一章 §4, 对于 $(\Omega, \mathscr{F}, \mathbf{P})$ 是有限概率空间, $\xi = \xi(\omega)$ 是简单随机变量

$$\xi(\omega) = \sum_{k=1}^{n} x_k I_{A_k}(\omega) \tag{1}$$

的情形, 曾经定义了数学期望 $\mathbf{E}\xi$ 的概念. 对于任意概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$, 使用与简单随机变量的数学期望同样的概念. 具体地说, 根据定义, 设

$$\mathbf{E}\xi = \sum_{k=1}^{n} x_k \mathbf{P}(A_k). \tag{2}$$

这一定义, 像有限概率空间一样, (在 $\mathbf{E}\xi$ 的值与形如 (1) 的表现无关的意义上) 是适定的, 类似地可以证明数学期望的简单性质 (见第一章 §4 第 5 小节).

这一节的目的是: 给出任意随机变量数学期望 $\mathbf{E}\xi$ 的定义, 研究其性质. 按分析的观点, 数学期望 $\mathbf{E}\xi$ 是 $\mathscr{F}$ — 可测函数 $\xi = \xi(\omega)$ 对测度 $\mathbf{P}$ 的勒贝格积分. 对于数学期望, 除 $\mathbf{E}\xi$ 之外还使用如下记号

$$\int_\Omega \xi(\omega) \mathbf{P}(d\omega), \quad \int_\Omega \xi d\mathbf{P}.$$

2. 数学期望的定义 设 $\xi = \xi(\omega)$ 是非负随机变量. 构造一非负简单随机变量序列 $\{\xi_n\}_{n \geqslant 1}$, 使对于每个 $\omega \in \Omega$, 当 $n \to \infty$ 时 $\xi_n(\omega) \uparrow \xi(\omega)$ (见 §4, 定理 1).

由于 $\mathbf{E}\xi_n \leqslant \mathbf{E}\xi_{n+1}$ (对照第一章 §4 第 5 小节的性质 3)), 则存在 $\lim_n \mathbf{E}\xi_n$, 其中的极限有可能以 $+\infty$ 为值.

定义 1 称

$$\mathbf{E}\xi = \lim_n \mathbf{E}\xi_n \tag{3}$$

为非负随机变量 $\xi = \xi(\omega)$ 的勒贝格积分, 或 $\xi = \xi(\omega)$ 的数学期望.

为使该定义是适定的, 需要证明其中的极限与逼近序列 $\{\xi_n\}$ 的选择无关. 换句话说, 需要证明, 若 $\xi_n \uparrow \xi, \eta_m \uparrow \xi$, 其中 $\{\eta_m\}$ 也是非负简单函数序列, 则

$$\lim_n \mathbf{E}\xi_n = \lim_m \mathbf{E}\eta_m. \tag{4}$$

引理 1 设 η 和 $\xi_n(n \geqslant 1)$ 是非负简单随机变量, 且

$$\xi_n \uparrow \xi \geqslant \eta.$$

那么,

$$\lim_n \mathbf{E}\xi_n \geqslant \mathbf{E}\eta. \tag{5}$$

证明 设 $\varepsilon > 0$ 且

$$A_n = \{\omega : \xi_n \geqslant \eta - \varepsilon\}.$$

显然,$A_n \uparrow \Omega$, 而

$$\xi_n = \xi_n I_{A_n} + \xi_n I_{\bar{A}_n} \geqslant \xi_n I_{A_n} \geqslant (\eta - \varepsilon) I_{A_n}.$$

因此, 由简单随机变量的数学期望的性质, 可见

$$\begin{aligned}\mathbf{E}\xi_n &\geqslant \mathbf{E}(\eta - \varepsilon)I_{A_n} = \mathbf{E}\eta I_{A_n} - \varepsilon \mathbf{P}(A_n)\\ &= \mathbf{E}\eta - \mathbf{E}\eta I_{\bar{A}_n} - \varepsilon \mathbf{P}(A_n) \geqslant \mathbf{E}\eta - C\mathbf{P}(\bar{A}_n) - \varepsilon,\end{aligned}$$

其中 $C = \max_\omega \eta(\omega)$. 由于 $\varepsilon > 0$ 是任意的, 得所要证明的不等式 (5). □

由这一引理可见

$$\lim_n \mathbf{E}\xi_n \geqslant \lim_m \mathbf{E}\eta_m;$$

而由对称性, 有

$$\lim_n \mathbf{E}\xi_n \leqslant \lim_m \mathbf{E}\eta_m;$$

于是, (4) 式得证.

下面的注释往往是有益的.

注 1 非负随机变量的数学期望可以表示为:

$$\mathbf{E}\xi = \sup_{\{s \in S : s \leqslant \xi\}} \mathbf{E}s, \tag{6}$$

其中 $S = \{s\}$ 是非负简单随机变量的集合 (练习题 1).

这样, 对于非负简单随机变量, 定义了数学期望. 下面考虑一般情形.

设 ξ 是随机变量, $\xi^+ = \max(\xi, 0), \xi^- = -\min(\xi, 0)$.

定义 2 称随机变量 ξ 的数学期望 $\mathbf{E}\xi$ 存在或有定义, 如果 $\mathbf{E}\xi^+$ 和 $\mathbf{E}\xi^-$ 中至少一个有限:

$$\min\{\mathbf{E}\xi^+, \mathbf{E}\xi^-\} < \infty.$$

这时, 随机变量 ξ 的数学期望定义为:

$$\mathbf{E}\xi = \mathbf{E}\xi^+ - \mathbf{E}\xi^-.$$

数学期望 $\mathbf{E}\xi$ 又称为函数 ξ 对概率测度的勒贝格积分. (关于勒贝格积分的其他定义方法, 见第 11 小节.)

定义 3 称随机变量 ξ 的数学期望有限 (或者 ξ 可积), 如果 $\mathbf{E}\xi^+ < \infty$ 和 $\mathbf{E}\xi^- < \infty$.

因为 $|\xi| = \xi^+ + \xi^-$, 所以 $\mathbf{E}\xi$ 有限等价于 $\mathbf{E}|\xi| < \infty$. (在此意义上按勒贝格可积具有 "绝对的" 特点.)

注 2 除数学期望 $\mathbf{E}\xi$ 外, 随机变量 ξ 的重要数字特征还有 $\mathbf{E}\xi^r$ (如果它存在) 和 $\mathbf{E}|\xi|^r, r > 0$, 并相应地称为随机变量 ξ 的 r 阶矩和 r 阶绝对矩.

注 3 在上面给出的勒贝格积分 $\int_\Omega \xi d\mathbf{P}$ 的定义, 曾假设: 测度 $\mathbf{P}$ 是概率测度 ($\mathbf{P}(\Omega) = 1$), 而 $\mathscr{F}$ – 可测函数 (随机变量) ξ 在 $\mathbb{R} = (-\infty, \infty)$ 中取值. 现在假设测度 μ 是定义在可测空间 $(\Omega, \mathscr{F})$ 上, 且可能取 $+\infty$ 为值的任意测度, 而 $\xi = \xi(\omega)$ 是在 $\bar{\mathbb{R}} = [-\infty, \infty]$ 上取值的 $\mathscr{F}$ – 可测函数 (广义随机变量). 这时, 勒贝格积分

$$\int_\Omega \xi(\omega)\mu(d\omega)$$

用同样的方法定义: 首先, 对非负简单函数 ξ (按 (2) 式将 $\mathbf{P}$ 换成 μ), 然后, 对任意非负简单函数 ξ, 一般按公式

$$\int_\Omega \xi(\omega)\mu(d\omega) = \int_\Omega \xi^+\mu(d\omega) - \int_\Omega \xi^-\mu(d\omega)$$

定义, 只要不出现 $\infty - \infty$ 的不定式.

对于数学分析, $(\Omega, \mathscr{F}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 且 μ 是勒贝格测度 λ 的情形特别重要. 这时, 将积分

$$\int_\mathbb{R} \xi(x)\lambda(dx)$$

记作

$$\int_\mathbb{R} \xi(x)dx, \text{ 或 } (\mathrm{L})\int_{-\infty}^{\infty} \xi(x)dx,$$

以强调它与黎曼积分

$$(\mathrm{R})\int_{-\infty}^{\infty} \xi(x)dx$$

的区别. 假如 (勒贝格 – 斯蒂尔切斯) 测度对应于某广义分布函数 $G = G(x)$, 则积分

$$\int_\mathbb{R} \xi(x)\mu(dx)$$

也称做勒贝格 – 斯蒂尔切斯积分, 并记为

$$(\mathrm{L-S})\int_\mathbb{R} \xi(x)G(dx),$$

以示其与相应的黎曼 – 斯蒂尔切斯积分 (见下面第 11 小节)

$$(\mathrm{R-S})\int_{-\infty}^{\infty} \xi(x)G(dx)$$

为区别.

由下面性质 D, 可见假如定义了 $\mathbf{E}\xi$, 则对于任何 $A\in\mathscr{F}$, 也就定义了数学期望 $\mathbf{E}(\xi I_A)$, 亦常使用下列记号:

$$\mathbf{E}(\xi;A) \text{ 或 } \int_\Omega \xi I_A d\mathbf{P},\ \int_A \xi d\mathbf{P}.$$

其中最后一个积分习惯上称做"ξ 对测度 $\mathbf{P}$ 在集合 A 上的勒贝格积分".

类似, 对于任意测度 μ, 下面两种表达式等价:

$$\int_\Omega \xi I_A d\mu,\quad \int_A \xi d\mu.$$

特别, 如果 μ 是 n-维勒贝格—斯蒂尔切斯测度, $A=(a_1,b_1]\times\cdots\times(a_n,b_n]$, 则可以将 $\int_A \xi d\mu$ 写成

$$\int_{a_1}^{b_1}\cdots\int_{a_n}^{b_n}\xi(x_1,\cdots,x_n)\mu(dx_1\cdots dx_n).$$

如果 μ 是勒贝格测度, 则将 $\mu(dx_1\cdots dx_n)$ 简单写成 $dx_1\cdots dx_n$.

3. 随机变量 ξ 的数学期望 $\mathbf{E}\xi$ 的性质

A. *若 c 是常数, 且 $\mathbf{E}\xi$ 存在, 则 $\mathbf{E}(c\xi)$ 也存在, 并且*

$$\mathbf{E}(c\xi)=c\mathbf{E}\xi.$$

B. *若 $\xi\leqslant\eta$, 则*

$$\mathbf{E}\xi\leqslant\mathbf{E}\eta,$$

即, 若 $-\infty<\mathbf{E}\xi$, 则

$$-\infty<\mathbf{E}\eta \text{ 且 } \mathbf{E}\xi\leqslant\mathbf{E}\eta,$$

或若 $\mathbf{E}\eta<\infty$, 则

$$\mathbf{E}\xi<\infty \text{ 且 } \mathbf{E}\xi\leqslant\mathbf{E}\eta.$$

C. *若 $\mathbf{E}\xi$ 存在, 则*

$$|\mathbf{E}\xi|\leqslant\mathbf{E}|\xi|.$$

D. *若 $\mathbf{E}\xi$ 存在, 则对于每个 $A\in\mathscr{F}$, 数学期望 $\mathbf{E}(\xi I_A)$ 存在; 若 $\mathbf{E}\xi$ 有限, 则 $\mathbf{E}(\xi I_A)$ 也有限.*

E. *若 ξ 和 η 是非负随机变量, 或满足 $\mathbf{E}|\xi|<\infty, \mathbf{E}|\eta|<\infty$, 则*

$$\mathbf{E}(\xi+\eta)=\mathbf{E}\xi+\mathbf{E}\eta.$$

(关于这一性质的推广见练习题 2)

证明 性质 A ~ E 的证明.

(a) 对于简单随机变量, 命题显然. 设 $c\geqslant 0;\ \xi\geqslant 0, \xi_n\uparrow\xi$, 其中 ξ_n 是简单随机变量. 那么, 由于 $c\xi_n\uparrow c\xi$, 可见

$$\mathbf{E}(c\xi)=\lim_n\mathbf{E}(c\xi_n)=c\lim_n\mathbf{E}\xi_n=c\mathbf{E}\xi.$$

在一般情形下, 需要利用表示 $\xi=\xi^+-\xi^-$, 并注意到, 对于 $c\geqslant 0$, $(c\xi)^+=c\xi^+$, $(c\xi)^-=c\xi^-$, 而对于 $c<0$, $(c\xi)^+=-c\xi^-$, $(c\xi)^-=-c\xi^+$.

(b) 若 $0\leqslant\xi\leqslant\eta$, 则 $\mathbf{E}\xi$ 和 $\mathbf{E}\eta$ 存在且 $\mathbf{E}\xi\leqslant\mathbf{E}\eta$, 由此 (6) 式得证. 现在设 $\mathbf{E}\xi>-\infty$, 则 $\mathbf{E}\xi^-<\infty$. 若 $\xi\leqslant\eta$, 则 $\xi^+\leqslant\eta^+$ 和 $\xi^-\geqslant\eta^-$. 因此 $\mathbf{E}\eta^-\leqslant\mathbf{E}\xi^-<\infty$, 从而 $\mathbf{E}\eta$ 存在, 且

$$\mathbf{E}\xi=\mathbf{E}\xi^+-\mathbf{E}\xi^-\leqslant\mathbf{E}\eta^+-\mathbf{E}\eta^-=\mathbf{E}\eta.$$

类似地考虑 $\mathbf{E}\eta<\infty$ 的情形.

(c) 由于 $-|\xi|\leqslant\xi\leqslant|\xi|$, 则由性质 A 和 B, 可见

$$-\mathbf{E}|\xi|\leqslant\mathbf{E}\xi\leqslant\mathbf{E}|\xi|,$$

即 $|\mathbf{E}\xi|\leqslant\mathbf{E}|\xi|$.

(d) 由性质 B, 以及

$$(\xi I_A)^+=\xi^+I_A\leqslant\xi^+,\ (\xi I_A)^-=\xi^-I_A\leqslant\xi^-,$$

可见 $\mathbf{E}(\xi I_A)$ 也有限.

(e) 设 $\xi\geqslant 0, \eta\geqslant 0$, 而 $\{\xi_n\}$ 和 $\{\eta_n\}$ 是简单函数序列, 且 $\xi_n\uparrow\xi$ 和 $\eta_n\uparrow\eta$. 那么,

$$\mathbf{E}(\xi_n+\eta_n)=\mathbf{E}\xi_n+\mathbf{E}\eta_n, \text{ 且}$$
$$\mathbf{E}(\xi_n+\eta_n)\uparrow\mathbf{E}(\xi+\eta), \mathbf{E}\xi_n\uparrow\mathbf{E}\xi, \mathbf{E}\eta_n\uparrow\mathbf{E}\eta,$$

从而 $\mathbf{E}(\xi+\eta)=\mathbf{E}\xi+\mathbf{E}\eta$. 对于 $\mathbf{E}|\xi|<\infty, \mathbf{E}|\eta|<\infty$ 的情形, 如果利用

$$\xi=\xi^+-\xi^-, \eta=\eta^+-\eta^-,$$
$$\xi^+\leqslant|\xi|, \xi^-\leqslant|\xi|, \eta^+\leqslant|\eta|, \eta^-\leqslant|\eta|,$$

则可以将其归结为讨论过的情形.

数学期望的进一步性质 数学期望的下一组性质 F ~ J, 涉及"$\mathbf{P}$-几乎必然"的概念. *称某一性质"$\mathbf{P}$-几乎必然"成立, 如果存在概率为 0 的集合 $\mathscr{N}\in\mathscr{F}$, $\mathbf{P}(\mathscr{N})=0$, 使该性质对于每一点 $\omega\in\Omega-\mathscr{N}$ 成立.* 与"$\mathbf{P}$-几乎必然"($\mathbf{P}$-a.c.) 的概念同时, 常使用"$\mathbf{P}$-几乎处处"($\mathbf{P}$-a.e.) 的概念. 或者简称"几乎必然"(a.c.), 相应地简称"几乎处处"(a.e.).

F. *若 $\xi=0$ (a.c.), 则 $\mathbf{E}\xi=0$.*

事实上, 如果 ξ 是简单随机变量, $\xi = \sum_{k=1}^{m} x_k I_{A_k}(\omega)$, $x_k \neq 0$, 则根据条件 $\mathbf{P}(A_k) = 0$, 因此 $\mathbf{E}\xi = 0$. 如果 $\xi \geqslant 0$ 且 $0 \leqslant s \leqslant \xi$, 其中 s 是简单随机变量, 故 $s = 0$ (a. e.), 从而 $\mathbf{E}s = 0$ 且

$$\mathbf{E}\xi = \sup_{\{s \in S: s \leqslant \xi\}} \mathbf{E}s = 0.$$

由于 $\xi = \xi^{+} - \xi^{-}$, 且 $\xi^{+} \leqslant |\xi|$, $\xi^{-} \leqslant |\xi|$, 而 $|\xi| = 0$ (a.e.), 所以, 通过极限过度, 可以将一般情形归结为 ξ 是简单随机变量的情形.

G. 若 $\xi = \eta$ (a.e.), 且 $\mathbf{E}|\xi| < \infty$, 则 $\mathbf{E}|\eta| < \infty$ 且 $\mathbf{E}\xi = \mathbf{E}\eta$ (亦见练习题 3).

实际上, 设 $\mathscr{N} = \{\omega : \xi \neq \eta\}$, 则 $\mathbf{P}(\mathscr{N}) = 0$ 且

$$\xi = \xi I_{\mathscr{N}} + \xi I_{\overline{\mathscr{N}}}, \quad \eta = \eta I_{\mathscr{N}} + \eta I_{\overline{\mathscr{N}}}.$$

由性质 **E** 和 **F**, 得

$$\mathbf{E}\xi = \mathbf{E}\xi I_{\mathscr{N}} + \mathbf{E}\xi I_{\overline{\mathscr{N}}} = \mathbf{E}\xi I_{\overline{\mathscr{N}}} = \mathbf{E}\eta I_{\overline{\mathscr{N}}}.$$

由于 $\mathbf{E}\eta I_{\mathscr{N}} = 0$, 故根据性质 **E**, 有 $\mathbf{E}\xi = \mathbf{E}\eta I_{\mathscr{N}} + \mathbf{E}\eta I_{\overline{\mathscr{N}}} = \mathbf{E}\eta$.

H. 若 $\xi \geqslant 0$ 且 $\mathbf{E}\xi = 0$, 则 $\xi = 0$ (a.e.).

为证明, 记 $A = \{\omega : \xi(\omega) > 0\}$, $A_n = \{\omega : \xi(\omega) \geqslant 1/n\}$. 显然, $A_n \uparrow A$, $0 \leqslant \xi I_{A_n} \leqslant \xi I_A$. 因此由性质 **B** 有

$$0 \leqslant \mathbf{E}\xi I_{A_n} \leqslant \mathbf{E}\xi = 0.$$

从而

$$0 = \mathbf{E}\xi I_{A_n} \geqslant \frac{1}{n}\mathbf{P}(A_n),$$

即对于一切 $n \geqslant 1$, $\mathbf{P}(A_n) = 0$. 由于 $\mathbf{P}(A) = \lim \mathbf{P}(A_n)$, 故 $\mathbf{P}(A) = 0$.

I. 设 ξ 和 η 满足 $\mathbf{E}|\xi| < \infty$, $\mathbf{E}|\eta| < \infty$ 且对于一切 $A \in \mathscr{F}$, 有 $\mathbf{E}(\xi I_A) \leqslant \mathbf{E}(\eta I_A)$. 证明 $\xi \leqslant \eta$ (a.e.).

实际上, 设 $B = \{\omega : \xi(\omega) > \eta(\omega)\}$, 则 $\mathbf{E}(\eta I_B) \leqslant \mathbf{E}(\xi I_B) \leqslant \mathbf{E}(\eta I_B)$, 即 $\mathbf{E}(\xi I_B) = \mathbf{E}(\eta I_B)$. 由性质 **E**, 有 $\mathbf{E}((\xi - \eta) I_B) = 0$; 而由性质 **H**, 有 $(\xi - \eta) I_B = 0$ (a.e.), 因此 $\mathbf{P}(B) = 0$.

J. 设 ξ 是广义随机变量, 且 $\mathbf{E}|\xi| < \infty$. 证明 $|\xi| < \infty$ (a.e.).

事实上, 假设 $A = \{\omega : |\xi(\omega)| = \infty\}$, 且 $\mathbf{P}(A) > 0$. 则 $\mathbf{E}|\xi| \geqslant \mathbf{E}(|\xi| I_A) = \infty \times \mathbf{P}(A) = \infty$, 而这与假设 $\mathbf{E}|\xi| < \infty$ 矛盾 (亦见练习题 4).

4. 数学期望的极限定理 在这一小节讨论关于在数学期望 (勒贝格积分) 号下的极限过程.

定理 1 (单调收敛性) 设 $\eta, \xi, \xi_1, \xi_2, \cdots$ 是随机变量序列.

a) 如果对于一切 $n \geqslant 1$, $\xi_n \geqslant \eta$, 有 $\mathbf{E}\eta > -\infty$, 且 $\xi_n \uparrow \xi$, 则

$$\mathbf{E}\xi_n \uparrow \mathbf{E}\xi.$$

b) 如果对于一切 $n \geqslant 1$, $\xi_n \leqslant \eta$, 有 $\mathbf{E}\eta < \infty$, 且 $\xi_n \downarrow \xi$, 则

$$\mathbf{E}\xi_n \downarrow \mathbf{E}\xi.$$

证明 a) 首先假设 $\eta \geqslant 0$. 假设对于每一个 $k \geqslant 1$, 简单函数序列 $\{\xi_k^{(n)}\}_{n \geqslant 1}$ 当 $n \to \infty$ 时 $\xi_k^{(n)} \uparrow \xi_k$. 记 $\zeta^{(n)} = \max_{1 \leqslant k \leqslant n} \xi_k^{(n)}$. 那么

$$\zeta^{(n-1)} \leqslant \zeta^{(n)} = \max_{1 \leqslant k \leqslant n} \xi_k^{(n)} \leqslant \max_{1 \leqslant k \leqslant n} \xi_k = \xi_n.$$

设 $\zeta = \lim_n \zeta^{(n)}$. 因为对于 $1 \leqslant k \leqslant n$,

$$\xi_k^{(n)} \leqslant \zeta^{(n)} \leqslant \xi_n,$$

那么, 若当 $n \to \infty$ 时取极限, 则对于任意 $k \geqslant 1$, 有

$$\xi_k \leqslant \zeta \leqslant \xi,$$

因此 $\xi = \zeta$.

由于 $\zeta^{(n)}$ 是简单随机变量且 $\zeta^{(n)} \uparrow \zeta$, 则

$$\mathbf{E}\xi = \mathbf{E}\zeta = \lim \mathbf{E}\zeta^{(n)} \leqslant \lim \mathbf{E}\xi_n.$$

另一方面, 显然, 由于 $\xi_n \leqslant \xi_{n+1} \leqslant \xi$, 则

$$\lim \mathbf{E}\xi_n \leqslant \mathbf{E}\xi.$$

从而 $\lim \mathbf{E}\xi_n = \mathbf{E}\xi$.

现在假设 η 是任意随机变量, 且 $\mathbf{E}\eta > -\infty$.

如果 $\mathbf{E}\eta = \infty$, 则由于性质 **B**, 可见 $\mathbf{E}\xi_n = \mathbf{E}\xi = \infty$, 从而命题得证. 设 $\mathbf{E}\eta < \infty$, 则考虑到关于的假设 $\mathbf{E}\eta > -\infty$, 得 $\mathbf{E}|\eta| < \infty$. 显然, 对于一切 $\omega \in \Omega$, $0 \leqslant \xi_n - \eta \uparrow \xi - \eta$. 从而, 根据上面已证明的结果, 有 $\mathbf{E}(\xi_n - \eta) \uparrow \mathbf{E}(\xi - \eta)$. 因此 (由性质 **E**) 和练习题 2)

$$\mathbf{E}\xi_n - \mathbf{E}\eta \uparrow \mathbf{E}\xi - \mathbf{E}\eta.$$

因为 $\mathbf{E}|\eta| < \infty$, 所以当 $n \to \infty$ 时 $\mathbf{E}\xi_n \uparrow \mathbf{E}\xi$.

如果将原变量加上负号, 则可以由命题 a) 得到命题 b). □

系 设 $\{\eta_n\}_{n \geqslant 1}$ 是非负随机变量序列, 那么

$$\mathbf{E}\sum_{n=1}^{\infty} \eta_n = \sum_{n=1}^{\infty} \mathbf{E}\eta_n.$$

证明 由数学期望的性质 **E** (亦见练习题 2), 单调收敛定理, 以及当 $k \to \infty$ 时

$$\sum_{n=1}^{k} \eta_n \uparrow \sum_{n=1}^{\infty} \eta_n$$

即可得到. □

定理 2 (法图 [P. Fatou] 引理)　设 $\eta, \xi_1, \xi_2, \cdots$ 是随机变量.

a) 如果对于一切 $n \geqslant 1$, $\xi_n \geqslant \eta$, 且 $\mathbf{E}\eta > -\infty$, 则

$$\mathbf{E}\varliminf \xi_n \leqslant \varliminf \mathbf{E}\xi_n.$$

b) 如果对于一切 $n \geqslant 1$, $\xi_n \leqslant \eta$, 且 $\mathbf{E}\eta < \infty$, 则

$$\varlimsup \mathbf{E}\xi_n \leqslant \mathbf{E}\varlimsup \xi_n.$$

c) 如果对于一切 $n \geqslant 1$, $|\xi_n| \leqslant \eta$, 且 $\mathbf{E}\eta < \infty$, 则

$$\mathbf{E}\varliminf \xi_n \leqslant \varliminf \mathbf{E}\xi_n \leqslant \varlimsup \mathbf{E}\xi_n \leqslant \mathbf{E}\varlimsup \xi_n. \tag{7}$$

证明　设 $\zeta_n = \inf_{m \geqslant n} \xi_m$, 则

$$\varliminf \xi_n = \lim_n \inf_{m \geqslant n} \xi_m = \lim_n \zeta_n.$$

明显地 $\zeta_n \uparrow \varliminf \xi_n$, 且对于一切 $n \geqslant 1$, 有 $\zeta_n \geqslant \eta$. 那么, 有定理 1, 有

$$\mathbf{E}\varliminf_n \xi_n = \mathbf{E}\lim_n \zeta_n = \lim_n \mathbf{E}\zeta_n = \varliminf_n \mathbf{E}\zeta_n \leqslant \varliminf_n \mathbf{E}\xi_n.$$

从而命题 a) 得证. 由 a) 可得 b). 由 a) 和 b) 可得 c). □

定理 3 (勒贝格控制收敛定理)　设随机变量 $\eta, \xi, \xi_1, \xi_2, \cdots$ 满足条件: $|\xi_n| \leqslant \eta$, $\mathbf{E}\eta < \infty$, 且 $\xi_n \to \xi$ (a.c.). 那么, $\mathbf{E}|\xi| < \infty$,

$$\mathbf{E}|\xi_n| \to \mathbf{E}\xi, n \to \infty, \tag{8}$$

且

$$\mathbf{E}|\xi_n - \xi| \to 0, n \to \infty. \tag{9}$$

证明　根据假设 $\varliminf \xi_n = \varlimsup \xi_n = \xi$ (a.c.). 因此, 由性质 G 和法图引理 (命题 c))

$$\mathbf{E}\xi = \mathbf{E}\varliminf_n \xi_n \leqslant \varliminf_n \mathbf{E}\xi_n = \varlimsup_n \mathbf{E}\xi_n = \mathbf{E}\varlimsup_n \xi_n = \mathbf{E}\xi,$$

因而 (8) 式得证. 显然, $|\xi| \leqslant \eta$. 所以 $\mathbf{E}|\xi| < \infty$.

只需注意到 $|\xi_n - \xi| \leqslant 2\eta$, 即可证明命题 (9).

系　设随机变量 $\eta, \xi, \xi_1, \xi_2, \cdots$ 满足条件: $|\xi_n| \leqslant \eta$, $\xi_n \to \xi$ (a.c.), 且对于某个 $p > 0$, $\mathbf{E}\eta^p < \infty$, 则 $\mathbf{E}|\xi|^p < \infty$, 且当 $n \to \infty$ 时 $\mathbf{E}|\xi_n - \xi|^p \to 0$.

为证明此系, 只需注意到, $|\xi| \leqslant \eta$ 和

$$|\xi_n - \xi|^p \leqslant (|\xi_n| + |\xi|)^p \leqslant (2\eta)^p.$$

在法图引理和控制收敛定理中, 保障公式 (7) ~ (9) 成立的条件 "$|\xi_n| \leqslant \eta$, $\mathbf{E}\eta < \infty$", 可以略为减弱. 为表述相应的结果 (定理 4), 现在给出如下定义.

定义 4　随机变量族 $\{\xi_n\}_{n \geqslant 1}$ 称做 (对测度 $\mathbf{P}$) 一致可积的, 如果

$$\sup_n \int_{\{|\xi_n| > c\}} |\xi_n| \mathbf{P}(d\omega) \to 0, c \to \infty, \tag{10}$$

或 (在其他记号下)

$$\sup_n \mathbf{E}[|\xi_n| I_{\{|\xi_n| > c\}}] \to 0, c \to \infty. \tag{11}$$

显然, 如果随机变量 $\xi_n, n \geqslant 1$ 满足条件 $|\xi_n| \leqslant \eta$, $\mathbf{E}\eta < \infty$, 则随机变量族 $\{\xi_n\}_{n \geqslant 1}$ 是一致可积的.

定理 4　设随机变量族 $\{\xi_n\}_{n \geqslant 1}$ 是一致可积的.

a) 那么

$$\mathbf{E}\varliminf_n \xi_n \leqslant \varliminf_n \mathbf{E}\xi_n \leqslant \varlimsup_n \mathbf{E}\xi_n \leqslant \mathbf{E}\varlimsup_n \xi_n.$$

b) 此外, 如果 $\xi_n \to \xi$ (a.c.), 则随机变量 ξ 可积, 且

$$\mathbf{E}\xi_n \to \mathbf{E}\xi, \quad n \to \infty,$$

$$\mathbf{E}|\xi_n - \xi| \to 0, \quad n \to \infty.$$

证明　a) 对于任意 $c > 0$, 有

$$\mathbf{E}\xi_n = \mathbf{E}\xi_n I_{\{\xi_n < -c\}} + \mathbf{E}\xi_n I_{\{\xi_n \geqslant -c\}}. \tag{12}$$

由于一致可积性, 对于任意 $\varepsilon > 0$, 可以选择充分大的 c, 使

$$\sup_n |\mathbf{E}\xi_n I_{\{\xi_n < -c\}}| < \varepsilon. \tag{13}$$

由法图引理, 可见

$$\varliminf \mathbf{E}\xi_n I_{\{\xi_n \geqslant -c\}} \geqslant \mathbf{E}\varliminf \xi_n I_{\{\xi_n \geqslant -c\}}.$$

因为 $\xi_n I_{\{\xi_n \geqslant -c\}} \geqslant \xi_n$, 所以

$$\varliminf \mathbf{E}\xi_n I_{\{\xi_n \geqslant -c\}} \geqslant \mathbf{E}\varliminf \xi_n. \tag{14}$$

由 (12)~(14) 式可得

$$\varliminf \mathbf{E}\xi_n \geqslant \mathbf{E}\varliminf \xi_n - \varepsilon.$$

由于 $\varepsilon > 0$ 是任意的, 可见 $\varliminf \mathbf{E}\xi_n \geqslant \mathbf{E}\varliminf \xi_n$.

类似地, 可以证明: $\varlimsup \mathbf{E}\xi_n \leqslant \mathbf{E}\varlimsup \xi_n$.

至于命题 b), 则其证明与定理 3 相应命题的证明相同. □

下面的定理, 更完整地揭示了控制收敛定理的意义, 并给出了在数学期望号下取极限的充分和必要条件.

定理 5 设 $0 \leqslant \xi_n \to \xi$ (P-a.c.), 而 $\mathbf{E}\xi_n < \infty$. 那么, $\mathbf{E}\xi_n \to \mathbf{E}\xi < \infty$ 的充分和必要条件是: 随机变量族 $\{\xi_n\}_{n\geqslant 1}$ 一致可积.

证明 由定理 4 的命题 b) 可见充分性成立. 为证明必要性, 考虑 (有限或可数) 集合 $A = \{a : \mathbf{P}\{\xi = a\} > 0\}$. 那么, 对于每一个 $a \notin A$, $\xi_n I_{\{\xi_n < a\}} \to \xi I_{\{\xi < a\}}$, 并且随机变量族 $\{\xi_n I_{\{\xi_n < a\}}\}_{n\geqslant 1}$ 一致可积. 因此, 由于 "充分性" 知, 对于每一个 $a \notin A$, $\mathbf{E}\xi_n I_{\{\xi_n < a\}} \to \mathbf{E}\xi I_{\{\xi < a\}}$, 因而当 $n \to \infty$ 时

$$\mathbf{E}\xi_n I_{\{\xi_n \geqslant a\}} \to \mathbf{E}\xi I_{\{\xi \geqslant a\}}, \quad a \notin A. \tag{15}$$

固定 $\varepsilon > 0$, 选取充分大的 $a_0 \notin A$, 使 $\mathbf{E}\xi I_{\{\xi \geqslant a_0\}} < \varepsilon/2$; 然后选取 N_0, 使对于一切 $n \geqslant N_0$, 有

$$\mathbf{E}\xi_n I_{\{\xi_n \geqslant a_0\}} \leqslant \mathbf{E}\xi I_{\{\xi \geqslant a_0\}} + \frac{\varepsilon}{2},$$

因而 $\mathbf{E}\xi_n I_{\{\xi_n \geqslant a_0\}} \leqslant \varepsilon$. 最后, 选取充分大的 $a_1 \geqslant a_0$, 使对于一切 $n \leqslant N_0$ $\mathbf{E}\xi_n I_{\{\xi_n \geqslant a_1\}} \leqslant \varepsilon$. 那么,

$$\sup_n \mathbf{E}\xi_n I_{\{\xi_n \geqslant a_1\}} \leqslant \varepsilon,$$

于是, $\{\xi_n\}_{n\geqslant 1}$ 的一致可积性得证. $\Box$

5. 一致可积的准则 首先指出, 如果随机变量族 $\{\xi_n\}_{n\geqslant 1}$ 一致可积, 则

$$\sup_n \mathbf{E}|\xi_n| < \infty. \tag{16}$$

实际上, 对于固定的 $\varepsilon > 0$ 和充分大的 $c > 0$,

$$\begin{aligned}\sup_n \mathbf{E}|\xi_n| &= \sup_n [\mathbf{E}|\xi_n| I_{\{|\xi_n| \geqslant c\}} + \mathbf{E}|\xi_n| I_{\{|\xi_n| < c\}}] \\ &\leqslant \sup_n \mathbf{E}|\xi_n| I_{\{|\xi_n| \geqslant c\}} + \sup_n \mathbf{E}|\xi_n| I_{\{|\xi_n| < c\}} \leqslant \varepsilon + c.\end{aligned}$$

因此 (16) 式得证.

结果表明, 条件 (16) 连同所谓 "一致连续性" 条件, 是一致可积性的充分和必要条件.

引理 2 随机变量族 $\{\xi_n\}_{n\geqslant 1}$ 一致可积的充分和必要条件, 是 $\mathbf{E}|\xi_n|$ $(n \geqslant 1)$ 一致有界 (即 (16) 式成立) 和 $\mathbf{E}\{|\xi_n| I_A\}$ $(n \geqslant 1)$ 一致连续 (即当 $\mathbf{P}(A) \to 0$ 时 $\sup_n \mathbf{E}\{|\xi_n| I_A\} \to 0$).

证明 必要性. 上面已经验证过条件 (16). 其次, 有

$$\begin{aligned}\mathbf{E}\{|\xi_n| I_A\} &= \mathbf{E}\{|\xi_n| I_{A\cap\{|\xi_n| \geqslant c\}}\} + \mathbf{E}\{|\xi_n| I_{A\cap\{|\xi_n| < c\}}\} \\ &\leqslant \mathbf{E}\{|\xi_n| I_{\{|\xi_n| \geqslant c\}}\} + c\mathbf{P}(A).\end{aligned} \tag{17}$$

选择 c 充分大, 使

$$\sup_n \mathbf{E}\{|\xi_n| I_{\{|\xi_n| \geqslant c\}}\} \leqslant \frac{\varepsilon}{2},$$

那么, 如果 $\mathbf{P}(A) \leqslant \varepsilon/2c$, 则由 (17) 式, 可见

$$\sup_n \mathbf{E}\{|\xi_n| I_A\} \leqslant \varepsilon.$$

因此必要性得证.

充分性. 设 $\varepsilon > 0$, $\delta > 0$, 由条件 $\mathbf{P}(A) < \delta$, 关于 n 一致, 有 $\mathbf{E}\{|\xi_n| I_A\} \leqslant \varepsilon$. 因为对于任意 $c > 0$,

$$\mathbf{E}|\xi_n| \geqslant \mathbf{E}\{|\xi_n| I_{\{|\xi_n| \geqslant c\}}\} \geqslant c\mathbf{P}\{|\xi_n| \geqslant c\}$$

(对照切比雪夫不等式), 所以

$$\sup_n \mathbf{P}\{|\xi_n| \geqslant c\} \leqslant \frac{1}{c} \sup_n \mathbf{E}|\xi_n| \to 0, \quad c \to \infty.$$

因此, 对于充分大的 c, 作为集合 A 可以取集合 $\{|\xi_n| \geqslant c\}$ $(n \geqslant 1)$ 中的任意一个. 于是, 一致可积性得证. $\Box$

下面的引理给出了便于应用的一致可积性的充分条件.

引理 3 设 $\xi_1, \xi_2, \cdots$ 是可积随机变量序列, 而 $G = G(t)$ 是定义在 $[0, \infty)$ 上的非负增函数, 满足

$$\lim_{t\to\infty} \frac{G(t)}{t} = \infty, \tag{18}$$

$$\sup_n \mathbf{E}G(|\xi_n|) < \infty. \tag{19}$$

那么, 随机变量族 $\{\xi_n\}_{n\geqslant 1}$ 一致可积.

证明 设 $\varepsilon > 0$,

$$M = \sup_n \mathbf{E}G(|\xi_n|), \quad a = \frac{M}{\varepsilon}.$$

取 c 充分大, 使对于 $t \geqslant c$, 有 $G(t)/t \geqslant a$. 那么, 关于 $n \geqslant 1$ 一致有

$$\mathbf{E}[|\xi_n| I_{\{|\xi_n| \geqslant c\}}] \leqslant \frac{1}{a}\mathbf{E}[G(|\xi_n|) I_{\{|\xi_n| \geqslant c\}}] \leqslant \frac{M}{a} = \varepsilon. \quad \Box$$

6. 独立随机变量之积的数学期望 如同第一章 §4 第 5 小节, 对于 ξ 和 η 是独立简单随机变量的情形, 证明了 $\mathbf{E}\xi\eta = \mathbf{E}\xi \times \mathbf{E}\eta$. 现在证明类似命题的一般情形 (亦见练习题 6).

定理 6 假设 ξ 和 η 是独立随机变量, 且 $\mathbf{E}|\xi| < \infty$, $\mathbf{E}|\eta| < \infty$. 那么, $\mathbf{E}|\xi\eta| < \infty$, 且

$$\mathbf{E}\xi\eta = \mathbf{E}\xi \times \mathbf{E}\eta. \tag{20}$$

证明 首先, 假设 $\xi \geqslant 0$, $\eta \geqslant 0$. 记

$$\xi_n = \sum_{k=0}^{\infty} \frac{k}{n} I_{\{\frac{k}{n} \leqslant \xi < \frac{k+1}{n}\}}, \quad \eta_n = \sum_{k=0}^{\infty} \frac{k}{n} I_{\{\frac{k}{n} \leqslant \eta < \frac{k+1}{n}\}},$$

则 $\xi_n \leqslant \xi$, $|\xi_n - \xi| \leqslant 1/n$ 和 $\eta_n \leqslant \eta$, $|\eta_n - \eta| \leqslant 1/n$. 由于 $\mathbf{E}\xi < \infty$, $\mathbf{E}\eta < \infty$, 则根据勒贝格控制收敛定理,

$$\lim_n \mathbf{E}\xi_n = \mathbf{E}\xi, \quad \lim_n \mathbf{E}\eta_n = \mathbf{E}\eta.$$

其次, 由于 ξ 和 η 独立性, 当 $n \to \infty$ 时, 有

$$\begin{aligned}\mathbf{E}\xi_n\eta_n &= \sum_{k,l \geqslant 0} \frac{kl}{n^2}\mathbf{E}I_{\{\frac{k}{n} \leqslant \xi < \frac{k+1}{n}\}} I_{\{\frac{l}{n} \leqslant \eta < \frac{l+1}{n}\}} \\ &= \sum_{k,l \geqslant 0} \frac{kl}{n^2}\mathbf{E}I_{\{\frac{k}{n} \leqslant \xi < \frac{k+1}{n}\}} \times \mathbf{E}I_{\{\frac{l}{n} \leqslant \eta < \frac{l+1}{n}\}} = \mathbf{E}\xi_n \cdot \mathbf{E}\eta_n.\end{aligned}$$

因此, 当 $n \to \infty$ 时, 有

$$\begin{aligned}|\mathbf{E}\xi\eta - \mathbf{E}\xi_n\eta_n| &\leqslant \mathbf{E}|\xi\eta - \xi_n\eta_n| \leqslant \mathbf{E}[|\xi||\eta - \eta_n|] + \mathbf{E}[|\eta_n||\xi - \xi_n|] \\ &\leqslant \frac{1}{n}\mathbf{E}\xi + \frac{1}{n}\mathbf{E}\left(\eta + \frac{1}{n}\right) \to 0.\end{aligned}$$

于是,

$$\mathbf{E}\xi\eta = \lim_n \mathbf{E}\xi_n\eta_n = \lim_n \mathbf{E}\xi_n \times \lim_n \mathbf{E}\eta_n = \mathbf{E}\xi \times \mathbf{E}\eta,$$

并且 $\mathbf{E}\xi\eta < \infty$.

如果利用关系式

$$\xi = \xi^+ - \xi^-, \eta = \eta^+ - \eta^-, \xi\eta = \xi^+\eta^+ - \xi^-\eta^+ - \xi^+\eta^- + \xi^-\eta^-,$$

则一般情形可以归结为上面讨论的情形. □

7. 与数学期望有关的不等式 现在证明与数学期望有关的若干不等式, 这些不等式在概率论以及数学分析中都有系统的应用, 其中许多已经在初等概率论中介绍过 (见第一章 §4 和 §5).

切比雪夫不等式. 设 ξ 是非负随机变量, 则对于任意 $\varepsilon > 0$, 有

$$\mathbf{P}\{\xi \geqslant \varepsilon\} \leqslant \frac{\mathbf{E}\xi}{\varepsilon}. \tag{21}$$

证明 易见

$$\mathbf{E}\xi \geqslant \mathbf{E}[\xi I_{\{\xi \geqslant \varepsilon\}}] \geqslant \varepsilon \mathbf{E}I_{\{\xi \geqslant \varepsilon\}} = \varepsilon\mathbf{P}\{\xi \geqslant \varepsilon\},$$

由此立即得不等式 (21). 由不等式 (21), 可得切比雪夫不等式的如下等价形式: 如果是任意随机变量 ξ, 则

$$\mathbf{P}\{\xi \geqslant \varepsilon\} \leqslant \frac{\mathbf{E}\xi^2}{\varepsilon^2} \tag{22}$$

和

$$\mathbf{P}\{|\xi - \mathbf{E}\xi| \geqslant \varepsilon\} \leqslant \frac{\mathbf{D}\xi}{\varepsilon^2}, \tag{23}$$

其中 $\mathbf{D}\xi = \mathbf{E}(\xi - \mathbf{E}\xi)^2$ 是随机变量 ξ 的方差.

柯西–布尼亚科夫斯基不等式 假设对于随机变量 ξ 和 η, $\mathbf{E}\xi^2 < \infty$, $\mathbf{E}\eta^2 < \infty$. 那么, $\mathbf{E}|\xi\eta| < \infty$ 且

$$(\mathbf{E}|\xi\eta|)^2 \leqslant \mathbf{E}\xi^2 \times \mathbf{E}\eta^2. \tag{24}$$

证明 假设 $\mathbf{E}\xi^2 > 0$, $\mathbf{E}\eta^2 > 0$. 那么, 记

$$\tilde{\xi} = \frac{\xi}{\sqrt{\mathbf{E}\xi^2}}, \quad \tilde{\eta} = \frac{\eta}{\sqrt{\mathbf{E}\eta^2}}.$$

则由 $2|\tilde{\xi}\tilde{\eta}| \leqslant \tilde{\xi}^2 + \tilde{\eta}^2$, 可见,

$$2\mathbf{E}|\tilde{\xi}\tilde{\eta}| \leqslant \mathbf{E}\tilde{\xi}^2 + \mathbf{E}\tilde{\eta}^2 = 2,$$

即 $\mathbf{E}|\tilde{\xi}\tilde{\eta}| \leqslant 1$, 由此得不等式 (24).

假如 $\mathbf{E}\xi^2 = 0$, 则根据性质 I, $\xi = 0$ (a.c.), 而根据性质 F, $\mathbf{E}\xi\eta = 0$, 于是, 不等式 (24) 仍然成立.

延森 (I. L. Jencen) 不等式 设对于 $x \in \mathbf{R}$, $g = g(x)$ 是凹 (亦称下凸) 博雷尔函数, 且 $\mathbf{E}|\xi| < \infty$. 那么,

$$g(\mathbf{E}\xi) \leqslant \mathbf{E}g(\xi). \tag{25}$$

证明 由于 $g = g(x)$ 是凹函数, 故对于每一点 $x_0 \in \mathbf{R}$, 存在 $\lambda(x_0)$, 使对于一切 $x \in R$,

$$g(x) \geqslant g(x_0) + (x - x_0)\lambda(x_0). \tag{26}$$

设 $x = \xi$ 和 $x_0 = \mathbf{E}\xi$, 则由 (26) 式可见

$$g(\xi) \geqslant g(\mathbf{E}\xi) + (x - \mathbf{E}\xi)\lambda(\mathbf{E}\xi),$$

于是, 在上式两侧同求数学期望, 得 $\mathbf{E}g(\xi) \geqslant g(\mathbf{E}\xi)$. □

由延森不等式可以导出一系列有用的不等式. 下面举几个例子.

李亚普诺夫不等式 对于 $0 < s < t$, 有

$$(\mathbf{E}|\xi|^s)^{1/s} \leqslant (\mathbf{E}|\xi|^t)^{1/t}. \tag{27}$$

易见, 若记 $r = t/s$, $\eta = |\xi|^s$, 并在延森不等式中设 $g(x) = |x|^r$, 则得 $|\mathbf{E}\eta|^r \leqslant \mathbf{E}|\eta|^r$, 即

$$(\mathbf{E}|\xi|^s)^{t/s} \leqslant \mathbf{E}|\xi|^t,$$

由此立即得不等式 (27).

由李亚普诺夫不等式, 得如下绝对矩之间的一系列不等式:

$$\mathbf{E}|\xi| \leqslant (\mathbf{E}|\xi|^2)^{1/2} \leqslant \cdots \leqslant (\mathbf{E}|\xi|^n)^{1/n}. \tag{28}$$

赫尔德 (O. L. Hölder) 不等式 设 $1 < p < \infty, 1 < q < \infty$ 且 $1/p + 1/q = 1$. 如果 $\mathbf{E}|\xi|^p < \infty, \mathbf{E}|\eta|^q < \infty$, 则 $\mathbf{E}|\xi\eta| < \infty$ 且

$$\mathbf{E}|\xi\eta| \leqslant (\mathbf{E}|\xi|^p)^{1/p}(\mathbf{E}|\eta|^q)^{1/q}. \tag{29}$$

证明 如果 $\mathbf{E}|\xi|^p = 0$ 或 $\mathbf{E}|\eta|^q = 0$, 则立即得不等式 (29), 就像柯西 – 李亚普诺夫不等式的情形一样 (当 $p = q = 2$ 时, 赫尔德不等式就是柯西 – 李亚普诺夫不等式).

现在设 $\mathbf{E}|\xi|^p > 0, \mathbf{E}|\eta|^q > 0$. 记

$$\tilde{\xi} = \frac{|\xi|}{(\mathbf{E}|\xi|^p)^{1/p}}, \quad \tilde{\eta} = \frac{|\eta|}{(\mathbf{E}|\eta|^q)^{1/q}}.$$

不难证明不等式

$$x^a y^b \leqslant ax + by, \tag{30}$$

其中 $x > 0, y > 0, a > 0, b > 0, a + b = 1$. 实际上, 由对数函数的凸性, 可见

$$\ln(ax + by) \geqslant a \ln x + b \ln y = \ln x^a y^b;$$

经对数还原, 立即得 (30) 式.

在 (30) 式中设 $x = \tilde{\xi}^p, y = \tilde{\eta}^q, a = 1/p, b = 1/q$, 得

$$\tilde{\xi}\tilde{\eta} \leqslant \frac{1}{p}\tilde{\xi}^p + \frac{1}{q}\tilde{\eta}^q;$$

$$\mathbf{E}\tilde{\xi}\tilde{\eta} \leqslant \frac{1}{p}\mathbf{E}\tilde{\xi}^p + \frac{1}{q}\mathbf{E}\tilde{\eta}^q = \frac{1}{p} + \frac{1}{q} = 1.$$

于是, (29) 式得证. □

闵可夫斯基 (Г. Минковский) 不等式 如果 $\mathbf{E}|\xi|^p < \infty, \mathbf{E}|\eta|^p < \infty, 1 \leqslant p < \infty$, 则 $\mathbf{E}|\xi + \eta|^p < \infty$, 且

$$(\mathbf{E}|\xi + \eta|^p)^{1/p} \leqslant (\mathbf{E}|\xi|^p)^{1/p} + (\mathbf{E}|\eta|^p)^{1/p}. \tag{31}$$

证明 首先证明如下代数不等式: 如果 $a, b > 0$ 和 $p \geqslant 1$, 则

$$(a + b)^p \leqslant 2^{p-1}(a^p + b^p). \tag{32}$$

实际上, 考虑函数 $F(x) = (a + x)^p - 2^{p-1}(a^p + x^p)$. 那么

$$F'(x) = p(a + x)^{p-1} - 2^{p-1}px^{p-1};$$

而由于 $p \geqslant 1$, 可见对于 $x < a, F'(a) = 0, F'(x) > 0$, 而对于 $x > a, F'(x) < 0$. 因此,

$$F(b) \leqslant \max F(x) = F(a) = 0,$$

从而不等式 (32) 得证.

利用不等式 (32), 得

$$|\xi + \eta|^p \leqslant (|\xi| + |\eta|)^p \leqslant 2^{p-1}(|\xi|^p + |\eta|^p). \tag{33}$$

因此, 如果 $\mathbf{E}|\xi|^p < \infty, \mathbf{E}|\eta|^p < \infty$, 则 $\mathbf{E}|\xi + \eta|^p < \infty$.

如果 $p = 1$, 则由 (33) 式可得 (31) 式.

现在假设 $p > 1$. 取 $q > 1$, 使 $1/p + 1/q = 1$, 则

$$|\xi + \eta|^p = |\xi + \eta||\xi + \eta|^{p-1} \leqslant |\xi||\xi + \eta|^{p-1} + |\eta||\xi + \eta|^{p-1}. \tag{34}$$

因为 $(p - 1)q = p$, 所以

$$\mathbf{E}(|\xi + \eta|^{p-1})^q = \mathbf{E}|\xi + \eta|^p < \infty$$

因此, 由赫尔德不等式, 有

$$\mathbf{E}(|\xi||\xi + \eta|^{p-1}) \leqslant (\mathbf{E}|\xi|^p)^{1/p}(\mathbf{E}|\xi + \eta|^{(p-1)q})^{1/q} = (\mathbf{E}|\xi|^p)^{1/p}(\mathbf{E}|\xi + \eta|^p)^{1/q}.$$

同样可得

$$\mathbf{E}(|\eta||\xi + \eta|^{p-1}) \leqslant (\mathbf{E}|\eta|^p)^{1/p}(\mathbf{E}|\xi + \eta|^p)^{1/q}.$$

从而, 由 (34) 式有

$$\mathbf{E}|\xi + \eta|^p \leqslant (\mathbf{E}|\xi + \eta|^p)^{1/q}\left[(\mathbf{E}|\xi|^p)^{1/p} + (\mathbf{E}|\eta|^p)^{1/p}\right]. \tag{35}$$

如果 $\mathbf{E}|\xi + \eta|^p = 0$, 则欲证明的 (31) 式显然. 现在假设 $\mathbf{E}|\xi + \eta|^p > 0$, 则由 (35) 式可见

$$(\mathbf{E}|\xi + \eta|^p)^{1-1/q} \leqslant \left[(\mathbf{E}|\xi|^p)^{1/p} + (\mathbf{E}|\eta|^p)^{1/p}\right].$$

于是, 由于 $1 - 1/q = 1/p$, 得所要证明的 (31) 式. □

8. 拉东 – 尼科迪姆 (J. Radon-O. M. Nikodým) 定理 设随机变量 ξ 有数学期望 $\mathbf{E}\xi$. 那么, 由性质 D, 决定一集函数

$$\mathbf{Q}(A) = \int_A \xi \mathrm{d}\mathbf{P}, A \in \mathscr{F}. \tag{36}$$

现在证明, 该函数是可数 – 可加的.

首先假设 ξ 是非负随机变量. 若 $A_1, A_2, \cdots$ 是 $\mathscr{F}$ 中的两两不相交集合, 记 $A = \sum A_n$, 则根据定理 1 的系, 有

$$\mathbf{Q}(A) = \mathbf{E}(\xi I_A) = \mathbf{E}\left(\xi I_{\sum A_n}\right) = \mathbf{E}\left(\sum \xi I_{A_n}\right) = \sum \mathbf{E}(\xi I_{A_n}) = \sum \mathbf{Q}(A_n).$$

如果 ξ 是有数学期望 $\mathbf{E}\xi$ 的任意随机变量, 则由下列关系式 (a)～(e), 可见函数 $\mathbf{Q}(A)$ 的可数 – 可加:

(a) 表现

$$\mathbf{Q}(A) = \mathbf{Q}^{+}(A) - \mathbf{Q}^{-}(A), \tag{37}$$

其中

$$\mathbf{Q}^{+}(A) = \int_A \xi^{+} d\mathbf{P},\ \mathbf{Q}^{-}(A) = \int_A \xi^{-} d\mathbf{P};$$

(b) (上面已证明的) 非负随机变量的可数—可加性;

(c) $\min\{\mathbf{Q}^{+}(\Omega), \mathbf{Q}^{-}(\Omega)\} < \infty$.

这样, 如果 $\mathbf{E}\xi$ 存在, 则函数 $\mathbf{Q} = \mathbf{Q}(A)$ 是带符号的测度, 即可表示为 $\mathbf{Q} = \mathbf{Q}_1 - \mathbf{Q}_2$ 的可数—可加集函数, 其中测度 Q_1 或 Q_2 至少有一个有限.

现在证明, 函数 $\mathbf{Q} = \mathbf{Q}(A)$ 具有如下重要性质: 关于测度 $\mathbf{P}$ 绝对连续:

$$如果\ \mathbf{P}(A) = 0,\ 则\ \mathbf{Q}(A) = 0,\quad (A \in \mathscr{F}).$$

(该性质简记为 $\mathbf{Q} \ll \mathbf{P}$).

为证明这一事实, 只需考虑非负随机变量的情形. 如果

$$\xi = \sum_{k=1}^{n} x_k I_{A_k}$$

是简单非负随机变量, 且 $\mathbf{P}(A) = 0$, 则

$$\mathbf{Q}(A) = \mathbf{E}(\xi I_A) = \sum_{k=1}^{n} x_k \mathbf{P}(A_k \cap A) = 0.$$

如果 $\{\xi_n\}_{n\geqslant 1}$ 是非负简单函数序列, 且 $\xi_n \uparrow \xi \geqslant 0$, 则根据控制收敛定理, 有

$$\mathbf{Q}(A) = \mathbf{E}(\xi I_A) = \lim \mathbf{E}(\xi_n I_{A_n}) = 0.$$

因为对于任何 $n \geqslant 1$ 及使 $\mathbf{P}(A) = 0$ 的 A, $\mathbf{E}(\xi_n I_{A_n}) = 0$.

于是, 勒贝格积分

$$\mathbf{Q}(A) = \int_A \xi d\mathbf{P},$$

作为集合 $A \in \mathscr{F}$ 的函数, 是关于测度 $\mathbf{P}$ ($\mathbf{Q} \ll \mathbf{P}$) 绝对连续的带符号的测度. 非常值得注意的是逆命题也成立.

拉东—尼科迪姆定理　设 $(\Omega, \mathscr{F})$ 是可测空间, μ 是 σ-有限测度, 而 λ 是关于 μ 绝对连续的带符号的测度 (即 $\lambda = \lambda_1 - \lambda_2$, 其中测度 λ_1 或 λ_2 至少有一个有限). 那么, 存在在 $\bar{\mathbb{R}} = [-\infty, \infty]$ 上取值的 $\mathscr{F}$—可测函数 $f = f(\omega)$, 使

$$\lambda(A) = \int_A f(\omega)\mu(d\omega),\quad A \in \mathscr{F}. \tag{38}$$

函数 $f = f(\omega)$ 精确到 μ—测度 0 唯一. 倘若 $h = h(\omega)$ 是另一 $\mathscr{F}$—可测函数, 使

$$\lambda(A) = \int_A h(\omega)\mu(d\omega),\quad A \in \mathscr{F},$$

则 $\mu\{\omega : f(\omega) \neq h(\omega)\} = 0$.

如果 μ 是测度, 则 $f = f(\omega)$ 的值域为 $\bar{\mathbb{R}}_{+} = [0, \infty]$.

式 (38) 中的函数 $f = f(\omega)$, 称做拉东—尼科迪姆导数, 或测度 λ 关于测度 μ 的密度, 并且记作

$$\frac{d\lambda}{d\mu}\quad 或\quad \frac{d\lambda}{d\mu}(\omega).$$

这些导数的一系列性质, 将在下面的 §7 中第 8 小节介绍. 特别指出 §7 中的公式 (35), 在置换测度的情况下重新计算数学期望时常要用到.

具体地说, 设 $\mathbf{P}$ 和 $\widetilde{\mathbf{P}}$ 是两个概率测度, $\mathbf{E}$ 和 $\widetilde{\mathbf{E}}$ 是相应的数学期望. 假设测度 $\widetilde{\mathbf{P}}$ 关于测度 $\mathbf{P}$ 绝对连续 ($\widetilde{\mathbf{P}} \ll \mathbf{P}$). 那么, 对于一切非负随机变量 $\xi = \xi(\omega)$, 有如下 "数学期望换算公式":

$$\widetilde{\mathbf{E}}\xi = \mathbf{E}\left(\xi\frac{d\widetilde{\mathbf{P}}}{d\mathbf{P}}\right). \tag{39}$$

在不假定随机变量 ξ 非负的情况下, 数学期望换算公式仍然成立: 在随机变量 ξ 关于测度 $\widetilde{\mathbf{P}}$ 可积, 当且仅当随机变量 $\xi\dfrac{d\widetilde{\mathbf{P}}}{d\mathbf{P}}$ 关于测度 $\mathbf{P}$ 可积; 这时 (39) 式成立.

公式 (39) 的证明并不复杂: 对于简单函数 ξ 它可以由导数 $d\widetilde{\mathbf{P}}/d\mathbf{P}$ 的定义得到; 对于非负函数 ξ, 首先, 由 §4 定理 1 的 b) 可见, 存在简单函数 $\xi_n \uparrow \xi\ (n \to \infty)$; 其次, 利用由 §4 中单调函数收敛的定理 1 的 a), 即可证明 (39) 式. 如果 ξ 是任意随机变量, 则根据 (39) 式, 有

$$\widetilde{\mathbf{E}}|\xi| = \mathbf{E}|\xi|\frac{d\widetilde{\mathbf{P}}}{d\mathbf{P}}.$$

"ξ 对测度 $\widetilde{\mathbf{P}}$ 的可积性" 与 "$\xi(d\widetilde{\mathbf{P}}/d\mathbf{P})$ 对测度 $\mathbf{P}$ 的可积性" 等价. 至于公式 (39) 本身的证明, 只需要考虑表示 $\xi = \xi^{+} - \xi^{-}$ 即可.

这里不加证明直接引用的拉东—尼科迪姆定理 (例如, 其证明可以参见 [70]), 将在 (§7) 建立条件数学期望时起关键作用.

9. 勒贝格积分中的变量替换　如果

$$\xi = \sum_{i=1}^{n} x_i I_{A_i}$$

是简单随机变量, $A_i = \{\omega : \xi = x_i\}$, 则

$$\mathbf{E}g(\xi) = \sum g(x_i)\mathbf{P}(A_i) = \sum g(x_i)\Delta F_\xi(x_i).$$

换句话说, 为计算 (简单) 随机变量的函数的数学期望, 没有必要直接利用全部概率测度 $\mathbf{P}$, 只需知道概率分布 P_ξ, 或等价地只需知道随机变量 ξ 的分布函数 F_ξ.

下面的定理归纳了这一性质.

定理 7 (勒贝格积分中的变量替换) 设 $(\Omega,\mathscr{F})$ 和 $(E,\mathscr{E})$ 是两个可测空间. $X=X(\omega)$ 是以 E 为值域的 $\mathscr{F}/\mathscr{E}$ 可测函数, $\mathbf{P}$ 是 $(\Omega,\mathscr{F})$ 上的概率测度, P_X 是由 $X=X(\omega)$ 诱导的 $(E,\mathscr{E})$ 上的概率测度:

$$P_X(A)=\mathbf{P}\{\omega:X(\omega)\in A\},A\in\mathscr{E}. \tag{40}$$

那么, 对于任意 $\mathscr{E}$- 可测函数 $g=g(x),x\in E$,

$$\int_A g(x)P_X(\mathrm{d}x)=\int_{X^{-1}(A)}g(X(\omega))\mathbf{P}(\mathrm{d}\omega),\quad A\in\mathscr{E} \tag{41}$$

(其含义是: 如果有一个积分存在, 则另一个也存在, 并且二者相等).

证明 设集合 $A\in\mathscr{E}$ 和 $g(x)=I_B(x)$, 其中 $B\in\mathscr{E}$. 那么, 欲证的关系式 (41) 变为等式

$$P_X(AB)=\mathbf{P}(X^{-1}(A)\cap X^{-1}(B)), \tag{42}$$

其正确性由 (40) 式和 $X^{-1}(A)\cap X^{-1}(B)=X^{-1}(A\cap B)$ 可见.

由 (42) 式可见, (41) 式对于非负简单函数 $g=g(x)$ 成立, 而由单调收敛定理知, (41) 式对于任意非负 $\mathscr{E}$- 可测函数成立.

而对于一般情形, 需要将函数 g 表示为 g^+-g^-. 注意到, 因为 (41) 式对于函数 g^+ 和 g^- 成立, 而且例如, 若

$$\int_A g^+(x)P_X(\mathrm{d}x)<\infty,$$

则

$$\int_{X^{-1}(A)}g^+(X(\omega))\mathbf{P}(\mathrm{d}\omega)<\infty,$$

说明若

$$\int_A g(x)P_X(\mathrm{d}x)$$

存在, 则积分存在

$$\int_{X^{-1}(A)}g(X(\omega))\mathbf{P}(\mathrm{d}\omega). \qquad \square$$

设 $(E,\mathscr{E})=(\mathbb{R},\mathscr{B}(\mathbb{R}))$, 而 $\xi=\xi(\omega)$ 是概率分布为 P_ξ 的随机变量. 那么, 如果 $g=g(x)$ 是博雷尔函数, 而积分

$$\int_A g(x)P_\xi(\mathrm{d}x)\quad 或 \quad \int_{\xi^{-1}(A)}g(\xi(\omega))\mathbf{P}(\mathrm{d}\omega)$$

存在, 则

$$\int_A g(x)P_\xi(\mathrm{d}x)=\int_{\xi^{-1}(A)}g(\xi(\omega))\mathbf{P}(\mathrm{d}\omega).$$

特别, 当 $A=\mathbb{R}$ 时得

$$\mathbf{E}g(\xi(\omega))=\int_\Omega g(\xi(\omega))\mathbf{P}(\mathrm{d}\omega)=\int_{\mathbb{R}}g(x)P_\xi(\mathrm{d}x). \tag{43}$$

测度 P_ξ 可以唯一地由分布函数还原 (§3 定理 1). 因此, 勒贝格积分

$$\int_{\mathbb{R}}g(x)P_\xi(\mathrm{d}x)$$

常记作

$$\int_{\mathbb{R}}g(x)F_\xi(\mathrm{d}x)\quad 或 \quad \int_{\mathbb{R}}g\,\mathrm{d}F_\xi,$$

并称做 (关于分布函数 $F_\xi(x)$ 所对应的测度的) 勒贝格 – 斯蒂尔切斯积分.

考虑分布函数 $F_\xi(x)$ 有密度 $f_\xi(x)$ 的情形, 即

$$F_\xi(x)=\int_{-\infty}^{x}f_\xi(y)\mathrm{d}y, \tag{44}$$

其中 $f_\xi=f_\xi(x)$ 非负博雷尔函数, 而积分是集合 $(-\infty,x]$ 上关于勒贝格测度的勒贝格积分 (见第 2 小节注 3). 在 (44) 式的记号下, (43) 式有如下形式:

$$\mathbf{E}g(\xi(\omega))=\int_{-\infty}^{\infty}g(x)f_\xi(x)\mathrm{d}x, \tag{45}$$

其中的积分是函数 $g(x)f_\xi(x)$ 按勒贝格测度的勒贝格积分. 事实上, 如果 $g(x)=I_B(x)$, $B\in\mathscr{B}(\mathbb{R})$, 则 (45) 式有如下形式

$$P_\xi(B)=\int_B f_\xi(x)\mathrm{d}x,\ B\in\mathscr{B}(\mathbb{R}), \tag{46}$$

其正确性由 §3 的定理 1 和如下公式可见:

$$F_\xi(b)-F_\xi(a)=\int_a^b f_\xi(x)\mathrm{d}x.$$

一般情形的证明, 与定理 7 的证明相同.

10. 傅比尼 (G. Fubini) 定理 考虑带测度 μ 的可测空间 $(\Omega,\mathscr{F})$, 其中 $\Omega=\Omega_1\times\Omega_2$, $\mathscr{F}=\mathscr{F}_1\otimes\mathscr{F}_2$, 而测度 $\mu=\mu_1\times\mu_2$ 是有限测度 μ_1 和 μ_2 的直积, 即在 $\mathscr{F}$ 上的测度:

$$\mu_1\times\mu_2(A\times B)=\mu_1(A)\mu_2(B),\quad A\in\mathscr{F}_1,B\in\mathscr{F}_2$$

(由下面将要证明的定理 8, 可见这样测度的存在性).

下面的定理有重要意义, 就像数学分析中关于 "二重黎曼积分化为二次积分" 的著名定理一样.

定理 8 (傅比尼定理)　设 $\xi=\xi(\omega_1,\omega_2)$ 是 $\mathscr{F}_1\otimes\mathscr{F}_2$ 可测函数, 并且关于测度 $\mu_1\times\mu_2$ 可积:

$$\int_{\Omega_1\times\Omega_2}|\xi(\omega_1,\omega_2)|\mathrm{d}(\mu_1\times\mu_2)<\infty. \tag{47}$$

那么, 积分

$$\int_{\Omega_1}\xi(\omega_1,\omega_2)\mu_1(\mathrm{d}\omega_1)\quad 和\quad \int_{\Omega_2}\xi(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)$$

1) 对于 μ_2– 几乎一切 ω_2 和 μ_1– 几乎一切 ω_1 有定义;

2) 是 $\mathscr{F}_1$– 和 $\mathscr{F}_2$ 可测函数, 并且相应地

$$\mu_2\left\{\omega_2:\int_{\Omega_1}|\xi(\omega_1,\omega_2)|\mu_1(\mathrm{d}\omega_1)=\infty\right\}=0,$$
$$\mu_1\left\{\omega_1:\int_{\Omega_2}|\xi(\omega_1,\omega_2)|\mu_2(\mathrm{d}\omega_2)=\infty\right\}=0; \tag{48}$$

3)

$$\int_{\Omega_1\times\Omega_2}\xi(\omega_1,\omega_2)\mathrm{d}(\mu_1\times\mu_2)=\int_{\Omega_1}\left[\int_{\Omega_2}\xi(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)\right]\mu_1(\mathrm{d}\omega_1)$$
$$=\int_{\Omega_2}\left[\int_{\Omega_1}\xi(\omega_1,\omega_2)\mu_1(\mathrm{d}\omega_1)\right]\mu_2(\mathrm{d}\omega_2). \tag{49}$$

证明　(i) 首先, 证明对于任意固定 $\omega_1\in\Omega_1$, ω_2 的函数 $\xi_{\omega_1}(\omega_2)=\xi(\omega_1,\omega_2)$ 是 $\mathscr{F}_2$– 可测的.

设 $F\in\mathscr{F}_1\otimes\mathscr{F}_2$ 和 $\xi(\omega_1,\omega_2)=I_F(\omega_1,\omega_2)$. 记 $F_{\omega_1}=\{\omega_2\in\Omega_2:(\omega_1,\omega_2)\in F\}$ 为 F 在点 ω_1 处的截线, 并设 $\mathscr{C}_{\omega_1}=\{F\in\mathscr{F}:F_{\omega_1}\in\mathscr{F}_2\}$. 需要证明对于任意 ω_1, $\mathscr{C}_{\omega_1}=\mathscr{F}$.

如果 $F=A\times B$, $A\in\mathscr{F}_1$, $B\in\mathscr{F}_2$, 则

$$(A\times B)_{\omega_1}=\begin{cases}B, & \omega_1\in A,\\ \varnothing, & \omega_1\notin A.\end{cases}$$

因此, 具有可测边的矩形属于 $\mathscr{C}_{\omega_1}$. 其次, 如果 $F\in\mathscr{F}$, 则 $(\bar F)_{\omega_1}=\overline{F_{\omega_1}}$, 而若 $\{F^n\}_{n\geqslant1}$ 是 $\mathscr{F}$ 中的集合, 则 $(\bigcup F^n)_{\omega_1}=\bigcup F^n_{\omega_1}$. 由此可见 $\mathscr{C}_{\omega_1}=\mathscr{F}$.

现在假设 $\xi(\omega_1,\omega_2)\geqslant0$. 那么, 由于对于每一 ω_1, 函数 $\xi_{\omega_1}(\omega_2)=\xi(\omega_1,\omega_2)$ 是 $\mathscr{F}_2$– 可测的, 则积分

$$\int_{\Omega_2}\xi(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)$$

有定义. 现在证明该积分是 $\mathscr{F}_1$– 可测函数, 且

$$\int_{\Omega_1}\left[\int_{\Omega_2}\xi(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)\right]\mu_1(\mathrm{d}\omega_1)=\int_{\Omega_1\times\Omega_2}\xi(\omega_1,\omega_2)\mathrm{d}(\mu_1\times\mu_2). \tag{50}$$

假设 $\xi(\omega_1,\omega_2)=I_{A\times B}(\omega_1,\omega_2)$, $A\in\mathscr{F}_1$, $B\in\mathscr{F}_2$. 则由于 $I_{A\times B}(\omega_1,\omega_2)=I_A(\omega_1)I_B(\omega_2)$, 可见

$$\int_{\Omega_2}I_{A\times B}(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)=I_A(\omega_1)\int_{\Omega_2}I_B(\omega_2)\mu_2(\mathrm{d}\omega_2). \tag{51}$$

因而, (51) 式左侧的积分是 $\mathscr{F}_1$ 可测函数.

设 $\xi(\omega_1,\omega_2)=I_F(\omega_1,\omega_2)$, $F\in\mathscr{F}=\mathscr{F}_1\otimes\mathscr{F}_2$. 证明积分

$$f(\omega_1)=\int_{\Omega_2}I_F(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)$$

$\mathscr{F}_1$– 可测. 为此, 记 $\mathscr{C}=\{F\in\mathscr{F}:f(\omega_1)$ 为 $\mathscr{F}_1$– 可测$\}$. 由于已经证明集合 $A\times B$ 属于 $\mathscr{C}$, $(A\in\mathscr{F}_1, B\in\mathscr{F}_2)$, 说明由这样集合组成的代数 $\mathscr{A}$ 也属于 $\mathscr{C}$. 由单调收敛定理可见, 集系 $\mathscr{C}$ 是单调类, 故 $\mathscr{C}=\mu(\mathscr{C})$ 是含 $\mathscr{C}$ 的最小单调类. 所以, 由于包含关系 $\mathscr{A}\subseteq\mathscr{C}\subseteq\mathscr{F}$, 以及 §2 定理 1, 可见 $\mathscr{F}=\sigma(\mathscr{A})=\mu(\mathscr{A})\subseteq\mu(\mathscr{C})=\mathscr{C}\subseteq\mathscr{F}$, 即 $\mathscr{C}=\mathscr{F}$.

最后, 如果 $\xi(\omega_1,\omega_2)$ 是任意非负 $\mathscr{F}$– 可测函数, 则由单调收敛定理和 §4 定理 2, 可见下面积分的 $\mathscr{F}_1$ 可测性:

$$\int_{\Omega_2}\xi(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2).$$

现在证明, 定义在 $\mathscr{F}=\mathscr{F}_1\otimes\mathscr{F}_2$ 上, 且具有性质: $\mu=\mu_1\times\mu_2(A\times B)=\mu_1(A)\mu_2(B)$, $A\in\mathscr{F}_1$, $B\in\mathscr{F}_2$ 的测度 $\mu=\mu_1\times\mu_2$, 测度确实存在而且唯一. 对于 $F\in\mathscr{F}$,

$$\mu(F)=\int_{\Omega_1}\left[\int_{\Omega_2}I_{F_{\omega_1}}(\omega_2)\mu_2(\mathrm{d}\omega_2)\right]\mu_1(\mathrm{d}\omega_1).$$

如已经证明的, 上面的二次积分中里边的积分是 $\mathscr{F}_1$– 可测函数, 因而集函数 $\mu(F)$ 确实对 $F\in\mathscr{F}$ 有定义. 显然, 若 $F=A\times B$, 则 $\mu(A\times B)=\mu_1(A)\mu_2(B)$. 设 $\{F^n\}$ 是 $\mathscr{F}$ 中两两不相交的集合, 则

$$\mu\left(\sum_n F^n\right)=\int_{\Omega_1}\left[\int_{\Omega_2}I_{(\sum F^n)_{\omega_1}}(\omega_2)\mu_2(\mathrm{d}\omega_2)\right]\mu_1(\mathrm{d}\omega_1)$$
$$=\int_{\Omega_1}\sum_n\left[\int_{\Omega_2}I_{F^n_{\omega_1}}(\omega_2)\mu_2(\mathrm{d}\omega_2)\right]\mu_1(\mathrm{d}\omega_1)$$
$$=\sum_n\int_{\Omega_1}\left[\int_{\Omega_2}I_{F^n_{\omega_1}}(\omega_2)\mu_2(\mathrm{d}\omega_2)\right]\mu_1(\mathrm{d}\omega_1)=\sum_n\mu(F^n).$$

于是, μ 是 $\mathscr{F}$ 上的 (σ– 有限) 测度.

由卡拉泰奥多里定理, 可见该测度是唯一具有性质 $\mu(A\times B)=\mu_1(A)\mu_2(B)$ 的测度.

现在证明 (50) 式. 如果 $\xi(\omega_1,\omega_2)=I_{A\times B}(\omega_1,\omega_2)$, $A\in\mathscr{F}_1$, $B\in\mathscr{F}_2$, 则

$$\int_{\Omega_1\times\Omega_2} I_{A\times B}(\omega_1,\omega_2)\mathrm{d}(\mu_1\times\mu_2)=\mu_1\times\mu_2(A\times B). \tag{52}$$

而由于 $I_{A\times B}(\omega_1,\omega_2)=I_A(\omega_1)I_B(\omega_2)$, 可见

$$\begin{aligned}&\int_{\Omega_1}\left[\int_{\Omega_2} I_{A\times B}(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)\right]\mu_1(\mathrm{d}\omega_1)\\ &=\int_{\Omega_1}\left[I_A(\omega_1)\int_{\Omega_2}I_B(\omega_2)\mu_2(\mathrm{d}\omega_2)\right]\mu_1(\mathrm{d}\omega_1)=\mu_1(A)\mu_2(B).\end{aligned} \tag{53}$$

根据测度 $\mu_1\times\mu_2$ 的定义

$$\mu_1\times\mu_2(A\times B)=\mu_1(A)\mu_2(B),$$

由 (52) 和 (53) 式, 可见对于 $\xi(\omega_1,\omega_2)=I_{A\times B}(\omega_1,\omega_2)$, 式 (50) 成立.

现在设 $\xi(\omega_1,\omega_2)=I_F(\omega_1,\omega_2)$, $F\in\mathscr{F}$. 集合的函数

$$\lambda(F)=\int_{\Omega_1\times\Omega_2} I_F(\omega_1,\omega_2)\mathrm{d}(\mu_1\times\mu_2),\ F\in\mathscr{F},$$

显然是 σ- 有限测度. 不难验证, 集函数

$$\nu(F)=\int_{\Omega_1}\left[\int_{\Omega_2}I_F(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)\right]\mu_1(\mathrm{d}\omega_1)$$

也是这样的测度. 上面已经证明, 测度 λ 和 ν 在形如 $F=A\times B$ 的集合上重合, 因此也在代数 $\mathscr{A}$ 上重合. 由此根据卡拉泰奥多里定理, 可见测度 λ 和 ν 对于一切 $F\in\mathscr{F}$ 重合.

(ii) 现在证明傅比尼定理本身的命题. 由 (47) 式, 有

$$\int_{\Omega_1\times\Omega_2}\xi^+(\omega_1,\omega_2)\mathrm{d}(\mu_1\times\mu_2)<\infty,\ \int_{\Omega_1\times\Omega_2}\xi^-(\omega_1,\omega_2)\mathrm{d}(\mu_1\times\mu_2)<\infty.$$

由已证明的知, 积分

$$\int_{\Omega_2}\xi^-(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)$$

是 ω_1 的 $\mathscr{F}_1$- 可测函数, 且

$$\int_{\Omega_1}\left[\int_{\Omega_2}\xi^+(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)\right]\mu_1(\mathrm{d}\omega_1)=\int_{\Omega_1\times\Omega_2}\xi^+(\omega_1,\omega_2)\mathrm{d}(\mu_1\times\mu_2)<\infty.$$

因此, 由练习题 4 (亦见第 3 小节的性质 J)

$$\int_{\Omega_2}\xi^+(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)<\infty\ (\mu_1-\text{a.e.}).$$

同样,

$$\int_{\Omega_2}\xi^-(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)<\infty\ (\mu_1-\text{a.e.}).$$

从而,

$$\int_{\Omega_2}|\xi(\omega_1,\omega_2)|\mu_2(\mathrm{d}\omega_2)<\infty\ (\mu_1-\text{a.e.}).$$

显然, 除测度 μ_1 为 0 的集合 $\mathscr{N}$ 外, 有

$$\int_{\Omega_2}\xi(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)=\int_{\Omega_2}\xi^+(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)-\int_{\Omega_2}\xi^-(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2). \tag{54}$$

假设其中对于 $\omega_1\in\mathscr{N}$, 积分等于 0, 则可以认为 (54) 式对于一切 $\omega_1\in\Omega_1$ 成立. 那么, 考虑到 (50) 式, 将 (54) 式对测度 μ_1 积分, 得

$$\begin{aligned}&\int_{\Omega_1}\left[\int_{\Omega_2}\xi(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)\right]\mu_1(\mathrm{d}\omega_1)\\ &=\int_{\Omega_1}\left[\int_{\Omega_2}\xi^+(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)\right]\mu_1(\mathrm{d}\omega_1)-\int_{\Omega_1}\left[\int_{\Omega_2}\xi^-(\omega_1,\omega_2)\mu_2(\mathrm{d}\omega_2)\right]\mu_1(\mathrm{d}\omega_1)\\ &=\int_{\Omega_1\times\Omega_2}\xi^+(\omega_1,\omega_2)\mathrm{d}(\mu_1\times\mu_2)-\int_{\Omega_1\times\Omega_2}\xi^-(\omega_1,\omega_2)\mathrm{d}(\mu_1\times\mu_2)\\ &=\int_{\Omega_1\times\Omega_2}\xi(\omega_1,\omega_2)\mathrm{d}(\mu_1\times\mu_2).\end{aligned}$$

类似地证明 (48) 式第一个关系式和下面的等式:

$$\int_{\Omega_1\times\Omega_2}\xi(\omega_1,\omega_2)\mathrm{d}(\mu_1\times\mu_2)=\int_{\Omega_2}\left[\int_{\Omega_1}\xi(\omega_1,\omega_2)\mu_1(\mathrm{d}\omega_1)\right]\mu_2(\mathrm{d}\omega_2).$$ □

系 如果

$$\int_{\Omega_1}\left[\int_{\Omega_2}|\xi(\omega_1,\omega_2)|\mu_2(\mathrm{d}\omega_2)\right]\mu_1(\mathrm{d}\omega_1)<\infty,$$

则傅比尼定理的结论仍然成立.

事实上, 由条件 (50) 可见 (47) 式, 因此傅比尼定理的一切结论仍然成立.

例 设随机向量 (ξ,η) 有二维概率密度 $f_{\xi,\eta}(x,y)$, 即

$$\mathbf{P}\{(\xi,\eta)\in B\}=\int_B f_{\xi,\eta}(x,y)\mathrm{d}x\mathrm{d}y,\ B\in\mathscr{B}(\mathbf{R}^2),$$

其中 $f_{\xi,\eta}(x,y)$ 是非负 $\mathscr{B}(\mathbf{R}^2)$- 可测函数, 而积分是对二维勒贝格测度的积分.

我们证明, ξ 和 η 的一维概率分布也有密度 $f_\xi(x)$ 和 $f_\eta(y)$, 并且

$$f_\xi(x)=\int_{-\infty}^{\infty}f_{\xi,\eta}(x,y)\mathrm{d}y,\ f_\eta(y)=\int_{-\infty}^{\infty}f_{\xi,\eta}(x,y)\mathrm{d}x. \tag{55}$$

事实上, 如果 $A\in\mathscr{B}(\mathbf{R})$, 则根据傅比尼定理

$$\mathbf{P}\{\xi\in A\}=\mathbf{P}\{(\xi,\eta)\in A\times\mathbf{R}\}=\int_{A\times\mathbf{R}}f_{\xi,\eta}(x,y)\mathrm{d}x\mathrm{d}y=\int_A\left[\int_{\mathbf{R}}f_{\xi,\eta}(x,y)\mathrm{d}y\right]\mathrm{d}x,$$

因此, 说明 ξ 有概率分布密度, 而 (55) 式的第一个公式得证. 类似地可以证明 (55) 式的第二个公式.

根据 §6 的定理, 随机变量 ξ 和 η 独立的充分和必要条件是

$$F_{\xi,\eta}(x,y) = F_\xi(x)F_\eta(y),\ (x,y) \in \mathbb{R}^2.$$

现在证明, 当随机变量 ξ 和 η 有联合密度 $f_{\xi,\eta}(x,y)$ 时, 它们独立的充分必要条件是:

$$f_{\xi,\eta}(x,y) = f_\xi(x)f_\eta(y) \tag{56}$$

(等式应理解为关于二维勒贝格测度几乎必然成立).

实际上, 如果 (56) 式成立, 则根据傅比尼定理, 有

$$\begin{aligned} F_{\xi,\eta}(x,y) &= \int_{(-\infty,x]\times(-\infty,y]} f_{\xi,\eta}(s,t)\,\mathrm{d}s\mathrm{d}t = \int_{(-\infty,x]\times(-\infty,y]} f_\xi(s)f_\eta(t)\,\mathrm{d}s\mathrm{d}t \\ &= \int_{(-\infty,x]} f_\xi(s)\mathrm{d}s\left[\int_{(-\infty,y]} f_\eta(t)\mathrm{d}t\right] = F_\xi(x)F_\eta(y), \end{aligned}$$

从而 ξ 和 η 独立.

相反, 如果 ξ 和 η 独立, 且有联合密度 $f_{\xi,\eta}(x,y)$, 则仍根据傅比尼定理, 有

$$\begin{aligned} \int_{(-\infty,x]\times(-\infty,y]} f_{\xi,\eta}(s,t)\,\mathrm{d}s\mathrm{d}t &= \left[\int_{(-\infty,x]} f_\xi(s)\mathrm{d}s\right] \times \left[\int_{(-\infty,y]} f_\eta(t)\mathrm{d}t\right] \\ &= \int_{(-\infty,x]\times(-\infty,y]} f_\xi(s)f_\eta(t)\,\mathrm{d}s\mathrm{d}t. \end{aligned}$$

由此可见, 对于任意 $B \in \mathscr{B}(\mathbb{R}^2)$, 有

$$\int_B f_{\xi,\eta}(x,y)\mathrm{d}x\mathrm{d}y = \int_B f_\xi(x)f_\eta(y)\mathrm{d}x\mathrm{d}y.$$

则有性质 I, 易见条件 (56) 成立.

11. 勒贝格积分和黎曼积分的各种定义及其关系 这一小节, 将要讨论勒贝格积分和黎曼积分各种定义的问题, 以及它们之间的关系.

首先需要指出, 勒贝格积分的构造, 与需要积分的函数定义在什么可测空间 $(\Omega,\mathscr{F})$ 无关. 然而黎曼积分, 对于抽象函数根本没有定义, 对于空间 $\Omega = \mathbb{R}^n$ 的情形, 则依次定义: 首先对 $\mathbb{R}^1$ 定义, 然后随着相应的变化, 依次对于 $n > 1$ 的情形进行移植.

应该强调, 黎曼积分和勒贝格积分的构造, 基于不同的思想基础. 构造黎曼积分的第一步, 将 $x \in \mathbb{R}^1$ 点按其在 x 轴上临近程度的特征进行分组. 而在构造勒贝格积分时, 则将 $x \in \mathbb{R}^1$ 点按另一种特征分组: 按被积函数值的临近程度进行分组. 这两种不同处理方法的结果, 使仅对于不 "十分" 间断的函数, 黎曼积分的积分和才有极限; 然而, 勒贝格积分的积分和, 却对更广泛的函数类收敛于极限值

回忆黎曼–斯蒂尔切斯积分的定义. 假设 $G(x)$ 是 $\mathbb{R}^1$ 上的某一广义分布函数 (见 §3 第 2 小节), μ 是与其对应的勒贝格–斯蒂尔切斯测度. 而 $g = g(x)$ 是在区间 $[a,b]$ 之外为 0 的有界函数.

考虑区间 $[a,b]$ 的分割 $\mathscr{P} = \{x_0,\cdots,x_n\}$, 其中

$$a = x_0 < x_1 < \cdots < x_n = b.$$

建立上积分和与下积分和:

$$\overline{\sum_{\mathscr{P}}} = \sum_{i=1}^n \overline{g}_i[G(x_{i+1}) - G(x_i)],\quad \underline{\sum_{\mathscr{P}}} = \sum_{i=1}^n \underline{g}_i[G(x_{i+1}) - G(x_i)],$$

其中

$$\overline{g}_i = \sup_{x_{i-1}<y\leqslant x_i} g(y),\quad \underline{g}_i = \inf_{x_{i-1}<y\leqslant x_i} g(y).$$

定义简单函数 $\overline{g}_{\mathscr{P}}(x)$ 和 $\underline{g}_{\mathscr{P}}(x)$, 并对于 $x_{i-1} < x \leqslant x_i$, 设

$$\overline{g}_{\mathscr{P}}(x) = \overline{g}_i,\ \underline{g}_{\mathscr{P}}(x) = \underline{g}_i,$$

并设 $\overline{g}_{\mathscr{P}}(a) = \underline{g}_{\mathscr{P}}(a) = g(a)$. 那么显然, 与根据勒贝格–斯蒂尔切斯积分的构造 (见第 2 小节的注 3), 有

$$\overline{\sum_{\mathscr{P}}} = (\mathrm{L-S})\int_a^b \overline{g}_{\mathscr{P}}(x)G(\mathrm{d}x)$$

和

$$\underline{\sum_{\mathscr{P}}} = (\mathrm{L-S})\int_a^b \underline{g}_{\mathscr{P}}(x)G(\mathrm{d}x).$$

现在设 $\{\mathscr{P}_k\}$ 是分割序列, 满足 $\mathscr{P}_k \subseteq \mathscr{P}_{k+1}$, 且对于 $\mathscr{P}_k = \{x_0^{(k)},\cdots,x_{n_k}^{(k)}\}$, 当 $k \to \infty$ 时 $\max\limits_{0\leqslant i\leqslant n_k} |x_{i+1}^{(k)} - x_i^{(k)}| \to 0$. 那么, $\overline{g}_{\mathscr{P}_1} \geqslant \overline{g}_{\mathscr{P}_2} \geqslant \cdots \geqslant g \geqslant \cdots \geqslant \underline{g}_{\mathscr{P}_2} \geqslant \underline{g}_{\mathscr{P}_1}$, 且如果 $|g(x)| \leqslant C$, 则根据控制收敛定理, 有

$$\begin{aligned} \lim_{k\to\infty} \overline{\sum_{\mathscr{P}_k}} &= (\mathrm{L-S})\int_a^b \overline{g}(x)G(\mathrm{d}x), \\ \lim_{k\to\infty} \underline{\sum_{\mathscr{P}_k}} &= (\mathrm{L-S})\int_a^b \underline{g}(x)G(\mathrm{d}x), \end{aligned} \tag{57}$$

其中 $\overline{g}(x) = \lim_k \overline{g}_{\mathscr{P}_k}(x)$, $\underline{g}(x) = \lim_k \underline{g}_{\mathscr{P}_k}(x)$.

如果极限 $\lim_k \overline{\sum}_{\mathscr{P}_k}$ 和 $\lim_k \underline{\sum}_{\mathscr{P}_k}$ 有限, 相等, 其共同值与分割序列 $\{\mathscr{P}_k\}$ 的选择无关, 则称函数 $g = g(x)$ 黎曼-斯蒂尔切斯可积, 而它们极限的共同值记作

$$(\mathrm{R-S})\int_a^b g(x)G(\mathrm{d}x),\quad 或\quad (\mathrm{R-S})\int_a^b g(x)\mathrm{d}G(x), \tag{58}$$

当 $G(x)=x$ 时, 该积分称做黎曼积分, 记作

$$(\mathrm{R})\int_a^b g(x)\mathrm{d}x.$$

现在假设

$$(\mathrm{L-S})\int_a^b g(x)G(\mathrm{d}x)$$

是相应的勒贝格–斯蒂尔切斯积分 (见第 2 小节, 注 3).

定理 9　*如果函数 $g=g(x)$ 在 $[a,b]$ 上连续, 则它黎曼–斯蒂尔切斯可积, 且*

$$(\mathrm{R-S})\int_a^b g(x)G(\mathrm{d}x)=(\mathrm{L-S})\int_a^b g(x)G(\mathrm{d}x). \tag{59}$$

证明　由于函数 $g=g(x)$ 连续, 可见 $\overline{g}(x)=g(x)=\underline{g}(x)$. 因此, 由 (57) 式, 有

$$\lim_{k\to\infty}\overline{\sum_{\mathscr{P}_k}}=\lim_{k\to\infty}\underline{\sum_{\mathscr{P}_k}}.$$

于是, 连续函数 $g=g(x)$ 黎曼–斯蒂尔切斯可积, 并且 (仍然由于 (57) 式) 其积分等于勒贝格–斯蒂尔切斯积分. □

对于直线 R 上勒贝格测度的情形, 我们比较详细地讨论黎曼积分和勒贝格积分之间的关系问题.

定理 10　*假设 $g=g(x)$ 在 $[a,b]$ 上是有界函数.*

a) 函数 $g=g(x)$ 在 $[a,b]$ 上黎曼可积, 当且仅当它 (关于 $\overline{\mathscr{B}}([a,b])$ 上的勒贝格 $\overline{\lambda}$ 测度) 几乎处处连续.

b) 假如函数 $g=g(x)$ 在 $[a,b]$ 上黎曼可积, 则它也勒贝格可积, 且

$$(\mathrm{R})\int_a^b g(x)\mathrm{d}x=(\mathrm{L})\int_a^b g(x)\overline{\lambda}(\mathrm{d}x). \tag{60}$$

证明　a) 假设函数 $g=g(x)$ 黎曼可积. 那么, 由 (57) 式, 可见

$$(\mathrm{L})\int_a^b \overline{g}(x)\lambda(\mathrm{d}x)=(\mathrm{L})\int_a^b \underline{g}(x)\lambda(\mathrm{d}x).$$

因为一般 $\underline{g}(x)\leqslant g(x)\leqslant \overline{g}(x)$, 所以由性质 H, 有

$$\underline{g}(x)=g(x)=\overline{g}(x)\quad(\lambda-\text{a.e.}). \tag{61}$$

由此不难验证, 函数 $g=g(x)$ (关于测度 λ) 几乎处处连续.

相反, 假设函数 $g=g(x)$ (关于测度 λ) 几乎处处连续. 这时 (61) 式成立, 因此 $g(x)$ 仅在 $\overline{\lambda}(\mathscr{N})=0$ 的集合 $\mathscr{N}$ 上不同于函数 $\overline{g}(x)$. 那么

$$\begin{aligned}\{x:g(x)\leqslant c\}&=\{x:g(x)\leqslant c\}\cap\overline{\mathscr{N}}+\{x:g(x)\leqslant c\}\cap\mathscr{N}\\&=\{x:\overline{g}(x)\leqslant c\}\cap\overline{\mathscr{N}}+\{x:g(x)\leqslant c\}\cap\mathscr{N}.\end{aligned}$$

显然, $\{x:\overline{g}(x)\leqslant c\}\cap\overline{\mathscr{N}}\in\overline{\mathscr{B}}([a,b])$, 而由于 $\{x:g(x)\leqslant c\}\cap\mathscr{N}$ 是 $\mathscr{N}$ 的子集, 而 $\mathscr{N}$ 是关于勒贝格测度 $\overline{\lambda}$ 的 0 测集, 故 $\mathscr{N}\in\overline{\mathscr{B}}([a,b])$. 从而, 函数 $g=g(x)$ 为 $\overline{\mathscr{B}}([a,b])$- 可测的, 并且作为有界函数勒贝格可积. 所以根据性质 G, 有

$$(\mathrm{L})\int_a^b g(x)\overline{\lambda}(\mathrm{d}x)=(\mathrm{L})\int_a^b \overline{g}(x)\lambda(\mathrm{d}x)=(\mathrm{L})\int_a^b \underline{g}(x)\lambda(\mathrm{d}x).$$

于是, 命题 a) 得证.

b) 如果函数 $g=g(x)$ 黎曼可积, 则由 a) 知它连续 ($\overline{\lambda}$-a.e.). 由定理 9 知, 对于连续函数, $g=g(x)$ 也勒贝格可积, 并且黎曼积分与勒贝格积分相等. □

注 1　设 μ 是 $\mathscr{B}([a,b])$ 上的某一勒贝格–斯蒂尔切斯测度. 对于子集 $\Lambda\subseteq[a,b]$ 的集系 $\overline{\mathscr{B}}_\mu([a,b])$, 存在集合 $A,B\in\mathscr{B}([a,b])$, 使 $A\subseteq\Lambda\subseteq B$, $\mu(B\setminus A)=0$. 设 $\overline{\mu}$ 是测度 μ 到 $\overline{\mathscr{B}}([a,b])$ 上的开拓 (对于使 $A\subseteq\Lambda\subseteq B$, $\mu(B\setminus A)=0$ 的 Λ, $\overline{\mu}(\Lambda)=\mu(A)$). 那么, 如果用测度 $\overline{\mu}$ 代替勒贝格测度 $\overline{\lambda}$, 而相应地用黎曼–斯蒂尔切斯积分和勒贝格–斯蒂尔切斯积分, 分别代替黎曼积分和勒贝格积分, 则定理仍然成立.

注 2　勒贝格积分定义 (见第 1 小节定义 1 和 2, 以及公式 (3) 和 (6)) 既从概念上, 又 "纯粹表面地" 区别于黎曼和黎曼–斯蒂尔切斯积分, 而后两种积分用到上、下积分和 (见 (57) 式).

现在较详细地比较这些定义.

设 $(\Omega,\mathscr{F},\mu)$ 是某一带测度 μ 的可测空间.

对于每一 $\mathscr{F}$– 可测非负函数 $f=f(\omega)$, 定义两个 (下和上) 积分 L_*f 和 L^*f (记作 $\int_* f\mathrm{d}\mu$ 和 $\int^* f\mathrm{d}\mu$).

$$L_*f=\sup\sum_i\Big(\inf_{\omega\in A_i}f(\omega)\Big)\mu(A_i),$$

$$L^*f=\inf\sum_i\Big(\sup_{\omega\in A_i}f(\omega)\Big)\mu(A_i),$$

其中 sup 和 inf 对空间 Ω 的一切 $\mathscr{F}$– 可测有限分割 $(A_1,A_2,\cdots,A_n)$; $A_1,A_2,\cdots,A_n$ 是 $\mathscr{F}$– 可测集合, $(A_1+A_2+\cdots+A_n)=\Omega$, $n\geqslant1$.

可以证明 $L_*f\leqslant L^*f$, 而如果函数 f 有界且测度 μ 有限, 则 $L_*f=L^*f$ (练习题 20).

定义函数 f 对测度 μ 的积分 Lf 的方法之一 (达布–杨 [J. G. Darboux-W. H. Young]): *称函数 f 对测度 μ 是可积的, 如果 $L_*f=L^*f$, 并在这种情况下记作 $Lf=L_*f(=L^*f)$.*

如果现在转向在第 1 小节给出的勒贝格积分 $\mathbf{E}f$ 的定义 1, 则可以看到 (练习题 21)

$$\mathbf{E}f=L_*f.$$

于是, 可以认为, 对于有界非负函数 $f = f(x)$, 勒贝格方法和达布—杨方法导致同一结果 ($\mathrm{E}f = L f = L_* f = L^* f$).

这些积分方法的差异出现在如下情形中: 被积函数无界, 或者测度 μ 无限.

例如, 按勒贝格积分的观点, 相应地对于 $f(x) = x^{-1/2}I_{(0,1]}$ 和 $f(x) = x^{-2}I_{(1,\infty)}$ 积分

$$\int_{(0,1]} \frac{\mathrm{d}x}{x^{1/2}} \quad 和 \quad \int_{(1,\infty)} \frac{\mathrm{d}x}{x^2}$$

相等, 且都等于 $L_* f$. 然而, $L^* f = \infty$.

这样, $L_* f < L^* f$, 说明所考虑的函数在达布—杨意义上不可积, 但在勒贝格意义上可积.

在所阐述的、运用下积分 $L_* f$ 及上积分 $L^* f$ 方法的范围内, 我们转向黎曼意义上的积分法.

假设 $\Omega = (0,1]$, $\mathscr{F} = \mathscr{B}$ (博雷尔 σ-代数), 而 $\mu = \lambda$ 是勒贝格测度. 设 $f = f(x)$, $x \in \Omega$ 是某一有界函数 (关于其可测性暂时不作假设).

仿照 $L_* f$ 和 $L^* f$, 引进下黎曼积分 $R_* f$ 和上黎曼积分 $R^* f$, 设

$$R_* f = \sup \sum_i \left(\inf_{\omega \in B_i} f(\omega) \right) \lambda(B_i),$$

$$R^* f = \inf \sum_i \left(\sup_{\omega \in B_i} f(\omega) \right) \lambda(B_i),$$

其中 $(B_1, B_2, \cdots, B_n)$ 是 $\Omega = (0,1]$ 的有限分割, 而 B_i 是形如 $(a_i, b_i]$ 的集合 (B_i 与 $L_* f$ 和 $L^* f$ 定义中 A_i 的不同, A_i 是任意 $\mathscr{F}$-可测集合).

由上面引进的定义, 显然

$$R_* f \leqslant L_* f \leqslant L^* f \leqslant R^* f.$$

在定理 9 和 10 中引进的黎曼可积性的性质, 可以重新表述, 并给出下列条件:

(a) $R^* f = R_* f$;

(b) 函数 f 之间断点集合 D_f 的勒贝格测度等于 0: $\lambda(D_f) = 0$;

(c) 存在一常数 $R(f)$, 使对于任意 $\varepsilon > 0$, 存在 $\delta > 0$, 使得

$$\left| R(f) - \sum_i f(\omega_i)\lambda((a_i, b_i]) \right| < \varepsilon$$

且对于两两不相交区间 $(a_i, b_i]$, $\sum (a_i, b_i] = (0, 1]$ 的任何有限族, 有 $\lambda((a_i, b_i]) < \delta$, $\omega_i \in (a_i, b_i]$.

利用定理 9 和 10 的论证, 可以证明 (练习题 22), 如果函数 f 有界, 则

(A) 条件 (a), (b), (c) 等价, 并且

(B) 在 (a), (b), (c) 中的任何条件下, 有

$$R(f) = R_* f = R^* f.$$

12. 勒贝格 – 斯蒂尔切斯积分的分部积分法 在这一小节, 我们引进关于"勒贝格—斯蒂尔切斯积分中分部积分"的有用的定理.

设在 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上有两个广义分布函数 $F = F(x)$ 和 $G = G(x)$.

定理 11 对于任意实数 a 和 b, $a < b$, 如下分部积分公式成立:

$$F(b)G(b) - F(a)G(a) = \int_a^b F(s-)\mathrm{d}G(s) + \int_a^b G(s)\mathrm{d}F(s), \tag{62}$$

或等价地,

$$\begin{aligned} &F(b)G(b) - F(a)G(a) \\ &= \int_a^b F(s-)\mathrm{d}G(s) + \int_a^b G(s-)\mathrm{d}F(s) + \sum_{a<s\leqslant b} \Delta F(s)\Delta G(s), \end{aligned} \tag{63}$$

其中

$$F(s-) = \lim_{t \uparrow s} F(t), \ \Delta F(s) = F(s) - F(s-).$$

注 1 积分公式 (62) 可以写成如下"微分"形式:

$$\mathrm{d}(FG) = F_- \mathrm{d}G + G\mathrm{d}F. \tag{64}$$

注 2 对于 ($(a,b]$ 上的) 有界变差函数 F 和 G, 定理的结论仍然成立. (每一个这样的右连续且有左极限的函数, 可以表示为两个单调不减函数之差.)

证明 首先回忆, 根据第 2 小节的假设, 将 $\int_a^b$ 积分理解为 $\int_{(a,b]}$. 因此 (见 §3 中公式 (2)):

$$[F(b) - F(a)][G(b) - G(a)] = \int_a^b \mathrm{d}F(s) \times \int_a^b \mathrm{d}G(t).$$

由此根据富比尼定理 ($F \times G$ 是对应于 F 和 G 的测度的直积), 可见

$$\begin{aligned} &[F(b) - F(a)][G(b) - G(a)] = \int_{(a,b]\times(a,b]} \mathrm{d}(F \times G)(s,t) \\ &= \int_{(a,b]\times(a,b]} I_{\{s\geqslant t\}}(s,t)\mathrm{d}(F \times G)(s,t) + \int_{(a,b]\times(a,b]} I_{\{s<t\}}(s,t)\mathrm{d}(F \times G)(s,t) \\ &= \int_{(a,b]} [G(s) - G(a)]\mathrm{d}F(s) + \int_{(a,b]} [F(t-) - F(a)]\mathrm{d}G(t) \\ &= \int_a^b G(s)\mathrm{d}F(s) + \int_a^b F(s-)\mathrm{d}G(s) - G(a)[F(b) - F(a)] \\ &\quad - F(a)[G(b) - G(a)], \end{aligned} \tag{65}$$

其中 I_A 是集合 A 的示性函数.

由 (65) 式直接可得 (62) 式. 同样, 由 (62) 式得 (63) 式, 为此只需注意到

$$\int_a^b [G(s)-G(s-)]\,\mathrm{d}F(s) = \sum_{a<s\leqslant b} \Delta G(s)\Delta F(s). \tag{66}$$

□

系 1 如果 $F(x)$ 和 $G(x)$ 是分布函数, 则

$$F(x)G(x) = \int_{-\infty}^x F(s-)\,\mathrm{d}G(s) + \int_{-\infty}^x G(s)\,\mathrm{d}F(s). \tag{67}$$

如果分布函数

$$F(x) = \int_{-\infty}^x f(s)\,\mathrm{d}s,$$

则

$$F(x)G(x) = \int_{-\infty}^x F(s)\,\mathrm{d}G(s) + \int_{-\infty}^x G(s)f(s)\,\mathrm{d}s. \tag{68}$$

系 2 设随机变量 ξ 的分布函数为 $F=F(x)$, 而且 $\mathbf{E}|\xi|^n<\infty$, 那么,

$$\int_0^\infty x^n\,\mathrm{d}F(x) = n\int_0^\infty x^{n-1}[1-F(x)]\,\mathrm{d}x, \tag{69}$$

$$\int_{-\infty}^0 |x|^n\,\mathrm{d}F(x) = n\int_0^\infty x^{n-1}F(-x)\,\mathrm{d}x, \tag{70}$$

$$\mathbf{E}|\xi|^n = \int_{-\infty}^\infty |x|^n\,\mathrm{d}F(x) = n\int_0^\infty x^{n-1}[1-F(x)+F(-x)]\,\mathrm{d}x. \tag{71}$$

为证明 (69) 式, 注意到

$$\begin{aligned}\int_0^b x^n\,\mathrm{d}F(x) &= -\int_0^b x^n\,\mathrm{d}[1-F(x)] \\ &= -b^n[1-F(b)] + n\int_0^b x^{n-1}[1-F(x)]\,\mathrm{d}x.\end{aligned} \tag{72}$$

现在, 在 $\mathbf{E}|\xi|^n<\infty$ 的条件下, 证明

$$b^n[1-F(b)+F(-b)] \leqslant b^n\mathbf{P}(|\xi|\geqslant b) \to 0,\ b\to\infty. \tag{73}$$

事实上,

$$\mathbf{E}|\xi|^n = \sum_{k=1}^\infty \int_{k-1}^k |x|^n\,\mathrm{d}F(x) < \infty,$$

因而

$$\sum_{k\geqslant b+1}\int_{k-1}^k |x|^n\,\mathrm{d}F(x) \to 0,\ b\to\infty.$$

由于

$$\sum_{k\geqslant b+1}\int_{k-1}^k |x|^n\,\mathrm{d}F(x) \geqslant b^n\mathbf{P}(|\xi|\geqslant b),$$

可见 (73) 式成立.

在 (72) 式中令 $b\to\infty$ 求极限, 得欲证明的 (69) 式. 类似可以证明 (70) 式. 由 (69) 和 (70) 式得公式 (71).

13. 有界变差函数 设 $A=A(t), t\geqslant 0$ 是右连续、有左极限且局部有界变差函数 (即在每一个有限区间 $[a,b]$ 上有有界变差). 考虑方程

$$Z_t = 1+\int_0^t Z_{s-}\,\mathrm{d}A(s), \tag{74}$$

其微分形式为

$$\mathrm{d}Z_t = Z_{t-}\,\mathrm{d}A(t),\quad Z_0 = 1. \tag{75}$$

利用上一小节证明的分部积分公式, 可以 (在局部有界变差函数类中) 求出方程 (74) 之解的明显形式.

引进函数 (称做随机分量; §7):

$$\mathscr{E}_t(A) = e^{A(t)-A(0)}\prod_{0<s\leqslant t}(1+\Delta A(s))e^{-\Delta A(s)}, \tag{76}$$

其中 $\Delta A(s) = A(s)-A(s-), s>0$, 而 $\Delta A(0)=0$.

函数 $A(s), 0\leqslant s\leqslant t$ 具有有界变差, 故最多有可数个间断点, 而且级数 $\sum_{0<s\leqslant t}|\Delta A(s)|$ 收敛. 由此可见函数

$$\prod_{0<s\leqslant t}(1+\Delta A(s))e^{-\Delta A(s)},\quad t\geqslant 0,$$

是局部有界变差函数.

如果记

$$A^c(t) = A(t) - \sum_{0<s\leqslant t}\Delta A(s)$$

为函数 $A(t)$ 的连续分量, 则 (76) 式可以改写成如下形式:

$$\mathscr{E}_t(A) = e^{A^c(t)-A^c(0)}\prod_{0<s\leqslant t}(1+\Delta A(s)). \tag{77}$$

记

$$F(t) = e^{A^c(t)-A^c(0)},\ G(t) = \prod_{0<s\leqslant t}(1+\Delta A(s)),\quad G(0)=1,$$

那么，由 (62) 式，有

$$\begin{aligned}\mathscr{E}_t(A) = F(t)G(t) &= 1 - \int_0^t F(s)\mathrm{d}G(s) - \int_0^t G(s-)\mathrm{d}F(s) \\ &= 1 + \sum_{0<s\leqslant t} F(s)G(s-)\Delta A(s) + \int_0^t G(s-)F(s)\mathrm{d}A^c(s) \\ &= 1 + \int_0^t \mathscr{E}_{s-}(A)\mathrm{d}A(s).\end{aligned}$$

于是，$\mathscr{E}_t(A), t \geqslant 0$ 是方程 (74) 的 (局部有界) 解.

现在证明，此解在局部有界类中唯一.

假设有两个局部有界解；而 $Y = Y(t), t \geqslant 0$ 是二者之差，那么

$$Y(t) = \int_0^t Y(s-)\mathrm{d}A(s).$$

设

$$T = \inf\{t \geqslant 0 : Y(t) \neq 0\}.$$

如果对于一切 $t \geqslant 0, Y(t) = 0$，则认为 $T = \infty$.

由于 $A(t), t \geqslant 0$ 是局部有界变差函数，则存在两个广义分布函数 $A_1(t)$ 和 $A_2(t)$，使 $A(t) = A_1(t) - A_2(t)$. 若设 $T < \infty$，则存在有限 $T' > T$，使

$$[A_1(T') + A_2(T')] - [A_1(T) + A_2(T)] \leqslant \frac{1}{2}.$$

那么，由方程

$$Y(t) = \int_T^t Y(s-)\mathrm{d}A(s), t \geqslant T,$$

可见

$$\sup_{t\leqslant T'} |Y(t)| \leqslant \frac{1}{2} \sup_{t\leqslant T'} |Y(t)|.$$

因为 $\sup_{t\leqslant T'} |Y(t)| < \infty$，所以对于 $T < t \leqslant T', Y(t) = 0$，而这与假设矛盾.

于是，证明了下面的定理.

定理 12 *在局部有界函数类中，方程 (74) 有由 (76) 式给出的解，而且是唯一解.*

14. 练习题

1. 证明 (6) 式.

2. 证明性质 E 的如下推广. 设对于随机变量 ξ 和 η 数学期望 $\mathbf{E}\xi$ 和 $\mathbf{E}\eta$ 存在，且 $\mathbf{E}\xi + \mathbf{E}\eta$ 有意义 (不是 $\infty - \infty$ 或 $-\infty + \infty$)，则

$$\mathbf{E}(\xi + \eta) = \mathbf{E}\xi + \mathbf{E}\eta.$$

3. 推广性质 G: 证明，如果 $\xi = \eta$ (a.c.)，且 $\mathbf{E}\xi$ 存在，则 $\mathbf{E}\eta$ 也存在，并且 $\mathbf{E}\xi = \mathbf{E}\eta$.

4. 假设 ξ 是广义随机变量，μ 是 σ- 有限测度，且 $\int_\Omega |\xi|\mathrm{d}\mu < \infty$. 证明 $|\xi| < \infty$ (a. c.). (与性质 J 比较.)

5. 设 μ 是 σ- 有限测度，ξ 和 η 是广义随机变量，且 $\int \xi\mathrm{d}\mu$ 和 $\int \eta\mathrm{d}\mu$ 存在. 证明，如果对于一切 $A \in \mathscr{F}$，有

$$\int_A \xi\mathrm{d}\mu \leqslant \int_A \eta\mathrm{d}\mu,$$

则 $\xi \leqslant \eta$ (μ-a.c.). (与性质 I 比较).

6. 设 ξ 和 η 是独立非负随机变量，证明 $\mathbf{E}\xi\eta = \mathbf{E}\xi \times \mathbf{E}\eta$.

7. 利用法图引理，证明

$$\mathbf{P}(\underline{\lim} A_n) \leqslant \underline{\lim}\mathbf{P}(A_n), \quad \mathbf{P}(\overline{\lim} A_n) \geqslant \overline{\lim}\mathbf{P}(A_n).$$

8. 举例说明，控制收敛定理的条件 “$|\xi_n| \leqslant \eta, \mathbf{E}\eta < \infty$” 一般不能减弱.

9. 举例说明，法图引理的条件 “$\xi_n \leqslant \eta, \mathbf{E}\eta > -\infty$” 一般不能去掉.

10. 证明法图引理的如下变形：若随机变量序列 $\{\xi_n^+\}_{n\geqslant 1}$ 一致可积，则

$$\overline{\lim}\mathbf{E}\xi_n \leqslant \mathbf{E}\overline{\lim}\xi_n.$$

11. 定义在 [0, 1] 上的狄利克雷 (P. G. L. Dirichlet) 函数

$$d(x) = \begin{cases} 1, & 若\ x\ 是无理数, \\ 0, & 若\ x\ 是有理数, \end{cases}$$

勒贝格可积，但是黎曼不可积. 为什么?

12. 举例说明，定义在 [0,1] 上且黎曼可积的函数序列 $\{f_n\}_{n\geqslant 1}, |f_n| \leqslant 1$，且关于勒贝格测度几乎处处有 $f_n \to f$，但是 f 黎曼不可积.

13. 设 $\{a_{ij}, i, j \geqslant 1\}$ 是实数序列，且 $\sum_{i,j} |a_{ij}| < \infty$. 由傅比尼定理导出

$$\sum_{(i,j)} a_{ij} = \sum_i \left(\sum_j a_{ij}\right) = \sum_j \left(\sum_i a_{ij}\right). \tag{78}$$

14. 举一数列 $\{a_{ij}, i, j \geqslant 1\}$ 的例子，使 $\sum_{i,j} |a_{ij}| = \infty$ 而等式 (78) 不成立.

15. 从简单函数出发，利用在积分号下取极限的勒贝格定理，证明，关于利用换元积分法的如下结果.

设 $h = h(x)$ 是区间 $[a, b]$ 上的非减连续可微函数，而 $f(x)$ 在 $[h(a), h(b)]$ 上的 (关于勒贝格测度) 可积. 那么，函数 $f(h(y))h'(y)$ 在区间 $[a, b]$ 上可积，且

$$\int_{h(a)}^{h(b)} f(x)\mathrm{d}x = \int_a^b f(h(y))h'(y)\mathrm{d}y.$$

16. 证明 (70) 式.

17. 设 $\xi, \xi_1, \xi_2, \cdots$ 是非负可积随机变量, 满足 $\mathbf{E}\xi_n \to \mathbf{E}\xi$, 且对于任意 $\varepsilon > 0$, 概率 $\mathbf{P}\{|\xi - \xi_n| > \varepsilon\} \to 0$. 证明 $\mathbf{E}|\xi_n - \xi| \to 0, n \to \infty$.

18. 设 ξ 是可积随机变量 ($\mathbf{E}|\xi| < \infty$). 证明, 对于任意 $\varepsilon > 0$, 存在 $\delta > 0$, 使对于任何 $A \in \mathscr{F}$, $\mathbf{P}(A) < \delta$, 有 $\mathbf{E}I_A|\xi| < \varepsilon$ ("勒贝格积分的绝对连续性").

19. 设对于随机变量 ξ, η, ζ 和 $\xi_n, \eta_n, \zeta_n, n \geqslant 1$, 有[*)]

$$\xi_n \xrightarrow{\mathbf{P}} \xi, \eta_n \xrightarrow{\mathbf{P}} \eta, \zeta_n \xrightarrow{\mathbf{P}} \zeta, \eta_n \leqslant \xi_n \leqslant \zeta_n, n \geqslant 1,$$

$$\mathbf{E}\zeta_n \to \mathbf{E}\zeta, \mathbf{E}\eta_n \to \mathbf{E}\eta,$$

且数学期望 $\mathbf{E}\xi, \mathbf{E}\eta, \mathbf{E}\zeta$ 有限. 证明普拉特 (Pratt) 引理: $\mathbf{E}\xi_n \to \mathbf{E}\xi$. 而如果满足 $\eta_n \leqslant 0 \leqslant \zeta_n$, 则 $\mathbf{E}|\xi_n - \xi| \to 0$.

由此证明, 如果 $\xi_n \xrightarrow{\mathbf{P}} \xi, \mathbf{E}|\xi_n| \to \mathbf{E}|\xi|$ 和 $\mathbf{E}|\xi| < \infty$, 则 $\mathbf{E}|\xi_n - \xi| \to 0$.

举例说明, 在普拉特引理的条件下, 一般 $\mathbf{E}|\xi_n - \xi| \not\to 0$.

20. 证明 $L_*f \leqslant L^*f$; 而如果函数 f 有界, 且测度 μ 有限, 则 $L_*f = L^*f$ (见第 11 小节之注 2).

21. 证明, 对于有界函数 f, 数学期望 $\mathbf{E}f = L_*f$ (见第 11 小节之注 2).

22. 证明第 11 小节之注 2 的最后一个命题.

23. 设 $F(x)$ 是随机变量 X 的分布函数. 证明

$$\mathbf{E}X^+ < \infty \Leftrightarrow \int_a^\infty \ln \frac{1}{F(x)} \mathrm{d}x < \infty, \text{对于某个 } a.$$

24. 证明, 如果对于 $p > 0$,

$$\lim_{x \to \infty} x^p \mathbf{P}\{|\xi| > x\} = 0,$$

则对于一切 $r < p$, $\mathbf{E}|\xi|^r < \infty$. 举例说明, 当 $r = p$ 时有可能 $\mathbf{E}|\xi|^p = \infty$.

25. 举一密度 $f(x)$ 的例子, $f(x)$ 是偶函数, 然而所有非奇数阶矩

$$\int_{-\infty}^\infty x^k f(x)\mathrm{d}x = 0, \quad k = 1, 3, \cdots$$

26. 举一随机变量序列 $\xi_n, n \geqslant 1$ 的例子, 使

$$\mathbf{E}\sum_{n=1}^\infty \xi_n \neq \sum_{n=1}^\infty \mathbf{E}\xi_n.$$

27. 设对于随机变量 X, 当 $a > 1$ 时, 有

$$\frac{\mathbf{P}\{|X| > an\}}{\mathbf{P}\{|X| > n\}} \to 0, n \to \infty.$$

[*)] 收敛 $\xi_n \xrightarrow{\mathbf{P}} \xi$ 称做依概率收敛, 表示对于任意 $\varepsilon > 0$, 概率 $\mathbf{P}\{|\xi_n - \xi| > \varepsilon\} \to 0, n \to \infty$. 详见 §10.

证明 X 一切阶矩存在. 提示: 利用公式:

$$\mathbf{E}|X|^N = N\int_0^\infty x^{N-1}\mathbf{P}\{|X| > x\}\mathrm{d}x.$$

28. 设对于随机变量 X 以概率 p_k 取 $k = 0, 1, 2, \cdots$ 为值. 函数

$$F(s) = \sum_{k=0}^\infty p_k s^k, \quad |s| \leqslant 1,$$

称做随机变量 X 的母函数. 证明下列公式:

(i) 若 X 是泊松随机变量, 即 $p_k = e^{-\lambda}\lambda^k/k!$, 其中 $\lambda > 0, k = 0, 1, 2, \cdots$, 则

$$F(s) = e^{-\lambda(1-s)}, |s| \leqslant 1;$$

(ii) 若随机变量 X 有几何分布, 即 $p_k = pq^k$, 其中 $0 < p < 1, q = 1 - p, k = 0, 1, 2, \cdots$, 则

$$F(s) = \frac{p}{1 - sq}, \quad |s| \leqslant 1.$$

29. 除母函数 $F(s)$ 外尚使用矩母函数: $M(s) = \mathbf{E}e^{sX}$ (假设 s 使 $\mathbf{E}e^{sX} < \infty$).

(a) 证明, 若矩母函数 $M(s)$ 对于 0 的某一邻域的一切 $s (s \in [-a, a], a > 0)$ 有定义, 则当 $s = 0$ 时对于所有 $k = 1, 2, \cdots$, 导数 $M^{(k)}(s)$ 存在, 而且

$$M^{(k)}(0) = \mathbf{E}X^k$$

(这一性质决定了 $M(s)$ 的名称)

(b) 举一随机变量的例子, 对于一切 $s > 0$, 其矩母函数 $M(s) = \infty$.

(c) 证明, 对于一切 $s \in \mathbb{R}$, 参数 $\lambda > 0$ 的泊松随机变量的矩母函数为 $M(s) = e^{-\lambda(1-e^s)}$.

30. 设 $0 < r < \infty, X_n \in L^r, X_n \xrightarrow{\mathbf{P}} X$. 证明下列条件等价:

(i) 随机变量族 $\{|X_n|^r, n \geqslant 1\}$ 一致可积;

(ii) 在 L^r 中 $X_n \to X$;

(iii) $\mathbf{E}|X_n|^r \to \mathbf{E}|X|^r < \infty$.

31. 斯皮策 (F. Spitzer) 恒等式. 设 $X_1, X_2, \cdots$ 是独立同分布随机变量序列, $\mathbf{P}\{X_1 \leqslant 1\} = 1$; $S_n = X_1 + \cdots + X_n$. 则对于 $|u|, |t| < 1$, 有

$$\sum_{n=0}^\infty t^n \mathbf{E}e^{uM_n} = \exp\left\{\sum_{n=1}^\infty \frac{t^n}{n}\mathbf{E}(uS_n^+)\right\},$$

其中 $M_n = \max\{0, X_1, X_2, \cdots, X_n\}, S_n^+ = \max\{0, S_n\}$.

32. 设 $S_0 = 0, S_n = X_1 + X_2 + \cdots + X_n, n \geqslant 1$ 是简单对称随机游动, 而 $\tau = \min\{n > 0 : S_n \geqslant 0\}$. 证明

$$\mathbf{E}\min(\tau, 2m) = 2\mathbf{E}|S_{2m}| = 4m\mathbf{P}\{S_{2m} = 0\}, m \geqslant 0.$$

33. 设 ξ 是标准正态随机变量: $\xi\sim\mathcal{N}(0,1)$. 利用分部积分法证明 $\mathbf{E}\xi^k=(k-1)\mathbf{E}\xi^{k-2}$, 并由此导出公式:

$$\mathbf{E}\xi^{2k-1}=0 \text{ 和 } \mathbf{E}\xi^{2k}=1\times 3\times\cdots\times(2k-3)(2k-1)\quad (=(2k-1)!!).$$

34. 证明, 函数 $x^{-1}\sin x, x\in\mathbb{R}$ 按黎曼可积, 但是按 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ 上的勒贝格测度不可积.

35. 证明函数

$$\xi(\omega_1,\omega_2)=e^{-\omega_1\omega_2}-2e^{-2\omega_1\omega_2},\quad \omega_1\in\Omega_1=[1,\infty),\quad \omega_2\in\Omega_2=(0,1],$$

关于勒贝格测度,

(a) 对于每一个 ω_2, 对 $\omega_1\in\Omega_1$ 可积;

(b) 对于每一个 ω_1, 对 $\omega_2\in\Omega_2$ 可积, 但是傅比尼定理不成立.

36. 证明莱维 (P. P. Lévy) 定理: 设随机变量 $\xi_1,\xi_2,\cdots$ 可积 (对于一切 $n\geqslant 1, \mathbf{E}|\xi_n|<\infty$), $\sup_n \mathbf{E}\xi_n<\infty$ 且 $\xi_n\uparrow\xi$, 则随机变量 ξ 可积, 且 $\mathbf{E}\xi_n\uparrow\mathbf{E}\xi$ (对照定理 1 之 a)).

37. 证明法图引理的如下变式: 如果 $0\leqslant\xi_n\to\xi$ (P-a.e.), 且 $\mathbf{E}\xi_n\leqslant A<\infty$, 则 ξ 可积, 且 $\mathbf{E}\xi\leqslant A$.

38. (黎曼积分法与勒贝格积分法的联系.) 设博雷尔函数 $f=f(x)$ 对勒贝格测度可积: $\int_{\mathbb{R}}|f(x)|\mathrm{d}x<\infty$. 证明对于任意 $\varepsilon>0$, 存在

(a) 具有有界区间 A_i 的阶梯函数

$$f_\varepsilon(x)=\sum_{i=1}^n f_i I_{A_i}(x),\quad 使\quad \int_{\mathbb{R}}|f(x)-f_\varepsilon(x)|\mathrm{d}x<\varepsilon;$$

(b) 具有有界承载子的连续可积函数 $g_\varepsilon(x)$, 使

$$\int_{\mathbb{R}}|f(x)-g_\varepsilon(x)|\mathrm{d}x<\varepsilon.$$

39. 证明, 如果 ξ 是可积随机变量, 则

$$\mathbf{E}\xi=\int_0^\infty \mathbf{P}\{\xi>x\}\mathrm{d}x-\int_{-\infty}^0\mathbf{P}\{\xi<x\}\mathrm{d}x.$$

40. 证明, 如果 ξ 和 η 是可积随机变量, 则

$$\mathbf{E}\xi-\mathbf{E}\eta=\int_{-\infty}^{\infty}[\mathbf{P}\{\eta<x\leqslant\xi\}-\mathbf{P}\{\xi<x\leqslant\eta\}]\mathrm{d}x.$$

41. 设 ξ 是非负随机变量 $(\xi\geqslant 0)$, 其拉普拉斯变换为 $\varphi_\xi(\lambda)=\mathbf{E}e^{-\lambda\xi}, \lambda\geqslant 0$.

(a) 证明, 对于任意 $0<r<1$,

$$\mathbf{E}\xi^r=\frac{r}{\Gamma(1-r)}\int_0^\infty\frac{1-\varphi_\xi(\lambda)}{\lambda^{r+1}}\mathrm{d}\lambda.$$

提示 对于 $s\geqslant 0, 0<r<1$,

$$\frac{1}{r}\Gamma(1-r)s^r=\int_0^\infty\frac{1-e^{-s\lambda}}{\lambda^{r+1}}\mathrm{d}\lambda.$$

(b) 证明, 如果 $\xi>0$, 则对于任意 $r>0$,

$$\mathbf{E}\xi^{-r}=\frac{1}{r\Gamma(r)}\int_0^\infty\varphi_\xi(\lambda^{1/r})\mathrm{d}\lambda.$$

提示 对于 $s\geqslant 0, r>0$,

$$s=\frac{r}{\Gamma(1/r)}=\int_0^\infty\exp\{-(\lambda/s)^r\}\mathrm{d}\lambda.$$

§7. 关于 $\sigma-$ 代数的条件概率和条件数学期望

1. **条件数学期望和条件概率一般定义的必要性** 设 $(\Omega,\mathscr{F},\mathbf{P})$ 是概率空间, 而事件 $A\in\mathscr{F}$ 的概率 $\mathbf{P}(A)>0$; $\mathscr{D}=\{D_1,D_2,\cdots\}$ 是空间 Ω 的有限或可数分割, 其中 $\mathbf{P}(D_i)>0, i\geqslant 1$. 像有限概率空间一样, 称

$$\mathbf{P}(B|A)\equiv\frac{\mathbf{P}(AB)}{\mathbf{P}(A)}$$

为事件 B 关于事件 A 的条件概率, 而事件 B 关于分割 $\mathscr{D}$ 的条件概率, 记作 $\mathbf{P}(B|\mathscr{D})$ 或 $\mathbf{P}(B|\mathscr{D})(\omega)$, 定义为: 对于 $\omega\in D_i, i\geqslant 1$, 等于 $\mathbf{P}(B|D_i)$ 的随机变量

$$\mathbf{P}(B|\mathscr{D})(\omega)=\sum_{i\geqslant 1}\mathbf{P}(B|D_i)I_{D_i}(\omega).$$

类似地, 如果 ξ 是随机变量, 假设其数学期望 $\mathbf{E}\xi$ 存在, 则 ξ 关于事件 $A(\mathbf{P}(A)>0)$ 的条件数学期望定义为

$$\mathbf{E}(\xi|A)\equiv\frac{\mathbf{E}(\xi I_A)}{\mathbf{P}(A)}$$

(对照第一章 §8 的 (10) 式).

随机变量 $\mathbf{P}(B|\mathscr{D})$ 显然关于 $\sigma-$ 代数 $\mathscr{G}=\sigma(\mathscr{D})$ 可测, 因此 $\mathbf{P}(B|\mathscr{D})$ 亦记作 $\mathbf{P}(B|\mathscr{G})$ (见第一章 §8).

在概率论中往往有必要考虑关于零概率事件的条件概率.

例如, 考虑如下试验. 设 ξ 是在 $[0,1]$ 上均匀分布的随机变量. 如果 $\xi=x$, 则掷一次硬币正面出现的概率为 x, 反面出现的概率为 $1-x$. 以 ν 表示将这样的硬币掷 n 次, 正面出现的次数. 问 "条件概率 $\mathbf{P}\{\nu=k|\xi=x\}$ 如何?" 尽管直观上清楚 "这一概率似乎应该等于 $C_n^k x^k(1-x)^{n-k}$", 然而, 由于当 $\mathbf{P}\{\xi=x\}=0$ 时, 我们所关心的 "条件概率 $\mathbf{P}\{\nu=k|\xi=x\}$" 暂时没有定义.

下面将给出关于 $\sigma-$ 代数 $\mathscr{G}, \mathscr{G}\subseteq\mathscr{F}$, 条件数学期望 (其中包括条件概率) 的一般定义, 并且与第一章 §8 对于有限概率空间的相应定义作比较.

2. 条件数学期望和条件概率的一般定义 设 $(\Omega,\mathscr{F},\mathbf{P})$ 是概率空间，$\mathscr{G}$ 是 σ-代数，$\mathscr{G}\subseteq\mathscr{F}$（$\mathscr{G}$ 是 $\mathscr{F}$ 的子 σ-代数），而 $\xi=\xi(\omega)$ 是随机变量. 回忆在 §6 中分两个阶段引进了数学期望 $\mathbf{E}\xi$ 的定义：首先对非负随机变量 ξ，然后对一般随机变量 ξ，并利用等式

$$\mathbf{E}\xi=\mathbf{E}\xi^{+}-\mathbf{E}\xi^{-},$$

其中（为避免形如 $\infty-\infty$ 的不定式）仅假设

$$\min(\mathbf{E}\xi^{-},\mathbf{E}\xi^{+})<\infty.$$

在定义条件数学期望 $\mathbf{E}(\xi|\mathscr{G})$ 时，也采用类似的两阶段构造过程.

定义 1 1) 非负随机变量 ξ 关于 σ-代数 $\mathscr{G}$ 的条件数学期望，定义为非负（广义）随机变量 $\mathbf{E}(\xi|\mathscr{G})$ 或 $\mathbf{E}(\xi|\mathscr{G})(\omega)$，如果

a) $\mathbf{E}(\xi|\mathscr{G})$ 为 $\mathscr{G}$-可测；

b) 对于任意 $A\in\mathscr{G}$，有

$$\int_A\xi\,\mathrm{d}\mathbf{P}=\int_A\mathbf{E}(\xi|\mathscr{G})\,\mathrm{d}\mathbf{P}.\tag{1}$$

2) 假设任意随机变量 ξ 关于 σ-代数 $\mathscr{G}$ 的条件数学期望 $\mathbf{E}(\xi|\mathscr{G})$ 或 $\mathbf{E}(\xi|\mathscr{G})(\omega)$ 有定义，如果 $\mathbf{P}$-a.c.

$$\min\{\mathbf{E}(\xi^{+}|\mathscr{G}),\ \mathbf{E}(\xi^{-}|\mathscr{G})\}<\infty,$$

并且通过如下公式表示

$$\mathbf{E}(\xi|\mathscr{G})=\mathbf{E}(\xi^{+}|\mathscr{G})-\mathbf{E}(\xi^{-}|\mathscr{G}),$$

且在使 $\mathbf{E}(\xi^{+}|\mathscr{G})=\mathbf{E}(\xi^{-}|\mathscr{G})=\infty$ 的一切基本事件的（零概率）集合上，任意规定差 $\mathbf{E}(\xi^{+}|\mathscr{G})-\mathbf{E}(\xi^{-}|\mathscr{G})$，例如令其等于 0.

首先证明，对于非负随机变量 ξ，$\mathbf{E}(\xi|\mathscr{G})$ 确实存在. 根据 §6 第 8 小节，集函数

$$\mathbf{Q}(A)=\int_A\xi\,\mathrm{d}\mathbf{P},\quad A\in\mathscr{G}\tag{2}$$

是在 $(\Omega,\mathscr{G})$ 上关于测度 $\mathbf{P}$ 绝对连续的测度，其中 $\mathbf{P}$ 是在 $(\Omega,\mathscr{F})$，$\mathscr{G}\subseteq\mathscr{F}$ 上的测度. 因此（根据拉东–尼科迪姆定理）存在这样的一个 $\mathscr{G}$-可测广义随机变量 $\mathbf{E}(\xi|\mathscr{G})$，使

$$\mathbf{Q}(A)=\int_A\mathbf{E}(\xi|\mathscr{G})\,\mathrm{d}\mathbf{P}.\tag{3}$$

由 (2) 和 (3) 式得 (1) 式.

注 1 根据拉东–尼科迪姆定理，条件数学期望 $\mathbf{E}(\xi|\mathscr{G})$ 仅精确到 $\mathbf{P}$-零测集唯一. 换句话说，作为 $\mathbf{E}(\xi|\mathscr{G})(\omega)$ 可以取任何 $\mathscr{G}$-可测函数 $f(\omega)$，其中 $f(\omega)$ 满足

$$\mathbf{Q}(A)=\int_A f(\omega)\,\mathrm{d}\mathbf{P},\ A\in\mathscr{G},$$

称做条件数学期望的变式.

还应指出，根据对拉东–尼科迪姆定理的注释，

$$\mathbf{E}(\xi|\mathscr{G})=\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\mathbf{P}}(\omega),\tag{4}$$

即条件数学期望是测度 $\mathbf{Q}$ 对测度 $\mathbf{P}$ 的拉东–尼科迪姆导数，其中测度 $\mathbf{Q}$ 和测度 $\mathbf{P}$ 都是空间 $(\Omega,\mathscr{G})$ 上的测度.

有益地指出，由 (1) 式可见，如果对于非负随机变量 ξ，$\mathbf{E}\xi<\infty$，则 $\mathbf{E}(\xi|\mathscr{G})<\infty$ ($\mathbf{P}$-a.c.). 类似地，如果 $\xi\leqslant0$，$\mathbf{E}\xi>-\infty$，则 $\mathbf{E}(\xi|\mathscr{G})>-\infty$ ($\mathbf{P}$-a.c.).

注 2 关于 (1) 需要指出，一般不能设 $\mathbf{E}(\xi|\mathscr{G})=\xi$，因为随机变量 ξ 未必 $\mathscr{G}$-可测.

注 3 假设对于随机变量 ξ，$\mathbf{E}\xi$ 存在，则 $\mathbf{E}(\xi|\mathscr{G})$ 也许可以定义为使 (1) 式成立的 $\mathscr{G}$-可测函数. 通常正是这样做的. 我们引进的定义 $\mathbf{E}(\xi|\mathscr{G})=\mathbf{E}(\xi^{+}|\mathscr{G})-\mathbf{E}(\xi^{-}|\mathscr{G})$ 具有如下优点：在 $\mathscr{G}$ 是平凡 σ-代数的情形下，即在 $\mathscr{G}=\{\varnothing,\Omega\}$ 的情形下，$\mathbf{E}(\xi|\mathscr{G})=\mathbf{E}\xi$，这时并不假定 $\mathbf{E}\xi$ 存在.（例如，对于随机变量 ξ，假如 $\mathbf{E}\xi^{+}=\infty$，$\mathbf{E}\xi^{-}=\infty$，$\mathscr{G}=\mathscr{F}$，则 $\mathbf{E}\xi$ 没有定义，但是在按定义 1，$\mathbf{E}(\xi|\mathscr{G})$ 存在，并且 $\xi=\xi^{+}-\xi^{-}$.）

注 4 假设条件数学期望 $\mathbf{E}(\xi|\mathscr{G})$ 有定义. 随机变量

$$\mathbf{D}(\xi|\mathscr{G})=\mathbf{E}\{[\xi-\mathbf{E}(\xi|\mathscr{G})]^2|\mathscr{G}\}$$

称做随机变量 ξ 关于 σ-代数 $\mathscr{G}$ 的条件方差.（与第一章 §8 练习题 2 中 $\mathbf{D}(\xi|\mathscr{D})$ 关于 $\mathscr{D}$ 的定义，以及 §4 中方差的定义比较.）

定义 2 设 $B\in\mathscr{F}$. 条件数学期望 $\mathbf{E}(I_B|\mathscr{G})$ 记作 $\mathbf{P}(B|\mathscr{G})$ 或 $\mathbf{P}(B|\mathscr{G})(\omega)$，并且称做事件 B 关于 σ-代数 $\mathscr{G}(\mathscr{G}\subseteq\mathscr{F})$ 的条件概率.

由定义 1 和 2，对于每一个固定的 $B\in\mathscr{F}$，条件概率 $\mathbf{P}(B|\mathscr{G})$ 是一随机变量，满足：

a) $\mathbf{P}(B|\mathscr{G})$ 为 $\mathscr{G}$-可测；

b) 对于任意 $A\in\mathscr{G}$，

$$\mathbf{P}(A\cap B)=\int_A\mathbf{P}(B|\mathscr{G})\,\mathrm{d}\mathbf{P}.\tag{5}$$

定义 3 设 ξ 是随机变量，$\mathscr{G}_\eta$ 是某随机元素 η 诱导的 σ-代数. 那么，如果 $\mathbf{E}(\xi|\mathscr{G}_\eta)$ 有定义，则记作 $\mathbf{E}(\xi|\eta)$ 或 $\mathbf{E}(\xi|\eta)(\omega)$，并称做 ξ 关于 η 的条件数学期望.

条件概率 $\mathbf{P}(B|\mathscr{G}_\eta)$，记作 $\mathbf{P}(B|\eta)$ 或 $\mathbf{P}(B|\eta)(\omega)$，称做事件 B 关于 η 的条件概率.

3. 关于分割和关于 σ-代数的条件数学期望的关系 现在证明，这里给出的 $\mathbf{E}(\xi|\mathscr{G})$ 的定义与第一章 §8 中条件数学期望的定义一致.

设 $\mathscr{D}=\{D_1,D_2,\cdots\}$ 是以 $D_i\left(\sum_i D_i=\Omega\right)$ 为原子的有限或可数分割，其中 $\mathbf{P}(D_i)>0$，$i\geqslant1$.

定理 1 如果 $\mathscr{G}=\sigma(\mathscr{D})$，而 ξ 是随机变量，且其数学期望 $\mathbf{E}\xi$ 有定义，则

$$\mathbf{E}(\xi|\mathscr{G})=\mathbf{E}(\xi|D_i)\quad (\text{在 } D_i \text{ 上 } \mathbf{P}-\text{a.c.}) \tag{6}$$

或

$$\mathbf{E}(\xi|\mathscr{G})=\frac{\mathbf{E}(\xi I_{D_i})}{\mathbf{P}(D_i)}\quad (\text{在 } D_i \text{ 上 } \mathbf{P}-\text{a.c.}).$$

(其中 "$\xi=\eta$ (在 A 上 $\mathbf{P}$—a.c.)" 或 "$\xi=\eta$ (A;$\mathbf{P}$—a.c.)" 表示 $\mathbf{P}(A\cap\{\xi\neq\eta\})=0$.)

证明 根据 §4 引理 3，在 D_i 上 $\mathbf{E}(\xi|\mathscr{G})=K_i$，其中 K_i 是常数。由

$$\int_{D_i}\xi\mathrm{d}\mathbf{P}=\int_{D_i}\mathbf{E}(\xi|\mathscr{G})\mathrm{d}\mathbf{P}=K_i\mathbf{P}(D_i),$$

可见

$$K_i=\frac{1}{\mathbf{P}(D_i)}\int_{D_i}\xi\mathrm{d}\mathbf{P}=\frac{\mathbf{E}(\xi I_{D_i})}{\mathbf{P}(D_i)}=\mathbf{E}(\xi|D_i). \qquad \square$$

于是，第一章引进的、关于分割 $\mathscr{D}=\{D_1,D_2,\cdots\}$ 的条件数学期望 $\mathbf{E}(\xi|\mathscr{D})$ 的概念，是关于 $\sigma-$代数 $\mathscr{G}=\sigma(\mathscr{D})$ 的条件数学期望概念的特殊情形。

4. 条件数学期望的性质 假设下面所考虑的随机变量 ξ,η 的条件数学期望都存在，而 $\sigma-$代数 $\mathscr{G}\subseteq\mathscr{F}$.

A*. 若 C 是常数，而 $\xi=C$ (a.c.)，则 $\mathbf{E}(\xi|\mathscr{G})=C$ (a.c.).

B*. 若 $\xi\leqslant\eta$ (a.c.)，则 $\mathbf{E}(\xi|\mathscr{G})\leqslant\mathbf{E}(\eta|\mathscr{G})$ (a.c.).

C*. $|\mathbf{E}(\xi|\mathscr{G})|\leqslant\mathbf{E}(|\xi||\mathscr{G})$ (a.c.).

D*. 若 a,b 为常数，且 $a\mathbf{E}\xi+b\mathbf{E}\eta$ 存在，则

$$\mathbf{E}(a\xi+b\eta|\mathscr{G})=a\mathbf{E}(\xi|\mathscr{G})+b\mathbf{E}(\eta|\mathscr{G})\ (\text{a.c.}).$$

E*. 若 $\mathscr{F}_*=\{\varnothing,\Omega\}$ 是平凡 $\sigma-$代数，则

$$\mathbf{E}(\xi|\mathscr{F}_*)=\mathbf{E}\xi\ (\text{a.c.}).$$

F*. $\mathbf{E}(\xi|\mathscr{F})=\xi$ (a.c.).

G*. $\mathbf{E}[\mathbf{E}(\xi|\mathscr{G})]=\mathbf{E}\xi$.

H*. 若 $\mathscr{G}_1\subseteq\mathscr{G}_2$，则有 (第一) "望远性"：

$$\mathbf{E}[\mathbf{E}(\xi|\mathscr{G}_2)|\mathscr{G}_1]=\mathbf{E}(\xi|\mathscr{G}_1)\ (\text{a.c.}).$$

I*. 若 $\mathscr{G}_1\supseteq\mathscr{G}_2$，则有 (第二) "望远性"：

$$\mathbf{E}[\mathbf{E}(\xi|\mathscr{G}_2)|\mathscr{G}_1]=\mathbf{E}(\xi|\mathscr{G}_2)\ (\text{a.c.}).$$

J*. 若随机变量 ξ 的数学期望 $\mathbf{E}\xi$ 存在，且与 $\sigma-$代数 $\mathscr{G}$ 无关 (即不依赖于 $I_B, B\in\mathscr{G}$)，则

$$\mathbf{E}(\xi|\mathscr{G})=\mathbf{E}\xi\ (\text{a.c.}).$$

K*. 若 η 是 $\mathscr{G}-$可测随机变量，$\mathbf{E}|\xi|<\infty$，$\mathbf{E}|\xi\eta|<\infty$，则

$$\mathbf{E}(\xi\eta|\mathscr{G})=\eta\mathbf{E}(\xi|\mathscr{G})\ (\text{a.c.}).$$

数学期望性质的证明

a*. 等于常数的函数关于 $\mathscr{G}$ 可测，因此只需验证等式①：

$$\int_A\xi\mathrm{d}\mathbf{P}=\int_A C\mathrm{d}\mathbf{P},\ A\in\mathscr{G}.$$

由于假设 $\xi=C$ (a.c.)，则由 §6 第 3 小节的性质 G，知上面的等式显然.

b*. 若 $\xi\leqslant\eta$ (a.c.)，则由 §6 第 3 小节的性质 B，有

$$\int_A\xi\mathrm{d}\mathbf{P}\leqslant\int_A\eta\mathrm{d}\mathbf{P},\ A\in\mathscr{G}.$$

从而，有

$$\int_A\mathbf{E}(\xi|\mathscr{G})\mathrm{d}\mathbf{P}\leqslant\int_A\mathbf{E}(\eta|\mathscr{G})\mathrm{d}\mathbf{P},\ A\in\mathscr{G}.$$

于是，由 §6 第 3 小节的性质 I，得所要求的不等式.

c*. 若考虑到明显的不等式 $-|\xi|\leqslant\xi\leqslant|\xi|$，则由 B* 可得所要求的不等式.

d*. 若集合 $A\in\mathscr{G}$，则由 §6 第 14 小节的练习题 2，有

$$\int_A(a\xi+b\eta)\mathrm{d}\mathbf{P}=\int_A a\xi\mathrm{d}\mathbf{P}+\int_A b\eta\mathrm{d}\mathbf{P}=\int_A a\mathbf{E}(\xi|\mathscr{G})\mathrm{d}\mathbf{P}+\int_A b\mathbf{E}(\eta|\mathscr{G})\mathrm{d}\mathbf{P}$$
$$=\int_A[a\mathbf{E}(\xi|\mathscr{G})+b\mathbf{E}(\eta|\mathscr{G})]\mathrm{d}\mathbf{P}.$$

于是性质 D* 得证.

e*. 该性质可以由下列性质得出：第一，$\mathbf{E}\xi$ 是 $\mathscr{F}_*-$可测函数；第二，如果 $A=\Omega$ 或 $A=\varnothing$，则显然

$$\int_A\xi\mathrm{d}\mathbf{P}=\int_A\mathbf{E}\xi\mathrm{d}\mathbf{P}.$$

f*. 由于 ξ 为 $\mathscr{F}-$可测，且

$$\int_A\xi\mathrm{d}\mathbf{P}=\int_A\mathbf{E}(\xi|\mathscr{F})\mathrm{d}\mathbf{P},\ A\in\mathscr{F},$$

则 $\mathbf{E}(\xi|\mathscr{F})=\xi$ (a.c.).

g*. 如果设 $\mathscr{G}_1=\{\varnothing,\Omega\}$，$\mathscr{G}_2=\mathscr{G}$，则性质 G* 可以由性质 E* 和 H* 得出.

h*. 设 $A\in\mathscr{G}_1$，则

$$\int_A\mathbf{E}(\xi|\mathscr{G}_1)\mathrm{d}\mathbf{P}=\int_A\xi\mathrm{d}\mathbf{P}.$$

①小写题号 a* ~ k* 下相应为命题 A* ~ K* 的证明. 译者

因为 $\mathscr{G}_1 \subseteq \mathscr{G}_2$, 则 $A \in \mathscr{G}_2$. 因此

$$\int_A \mathbf{E}[\mathbf{E}(\xi|\mathscr{G}_2)|\mathscr{G}_1]\mathrm{d}\mathbf{P} = \int_A \mathbf{E}(\xi|\mathscr{G}_2)\mathrm{d}\mathbf{P} = \int_A \xi \mathrm{d}\mathbf{P}.$$

从而, 对于 $A \in \mathscr{G}_1$, 有

$$\int_A \mathbf{E}(\xi|\mathscr{G}_1)\mathrm{d}\mathbf{P} = \int_A \mathbf{E}[\mathbf{E}(\xi|\mathscr{G}_2)|\mathscr{G}_1]\mathrm{d}\mathbf{P}.$$

仿照 §6 第 3 小节的性质 I (亦见 §6 的练习题 5), 可得

$$\mathbf{E}(\xi|\mathscr{G}_1) = \mathbf{E}[\mathbf{E}(\xi|\mathscr{G}_2)|\mathscr{G}_1] \quad \text{(a.c.)}.$$

i*. 设 $A \in \mathscr{G}_1$, 则根据 $\mathbf{E}[\mathbf{E}(\xi|\mathscr{G}_2)|\mathscr{G}_1]$ 的定义

$$\int_A \mathbf{E}[\mathbf{E}(\xi|\mathscr{G}_2)|\mathscr{G}_1]\mathrm{d}\mathbf{P} = \int_A \mathbf{E}(\xi|\mathscr{G}_2)\mathrm{d}\mathbf{P}.$$

因为函数 $\mathbf{E}(\xi|\mathscr{G}_2)$ 为 $\mathscr{G}_2$- 可测, 并且由于 $\mathscr{G}_2 \subseteq \mathscr{G}_1$, 则 $\mathbf{E}(\xi|\mathscr{G}_2)$ 也 $\mathscr{G}_1$- 可测. 可见 $\mathbf{E}(\xi|\mathscr{G}_2)$ 是条件数学期望 $\mathbf{E}[\mathbf{E}(\xi|\mathscr{G}_2)|\mathscr{G}_1]$ 的变式之一. 于是, 性质 **I*** 得证.

j*. 由于 $\mathbf{E}\xi$ 为 $\mathscr{G}$- 可测函数, 故只需验证对于任意 $B \in \mathscr{G}$, 有

$$\int_B \xi \mathrm{d}\mathbf{P} = \int_B \mathbf{E}\xi \mathrm{d}\mathbf{P},$$

即 $\mathbf{E}(\xi I_B) = \mathbf{E}\xi \times \mathbf{E}I_B$. 如果 $\mathbf{E}|\xi| < \infty$, 则这由 §6 的定理 6 立刻得到. 对于一般情形, 不是用 §6 的定理 6, 而是利用 §6 的练习题 6 证明上面的积分等式.

k*. 为证明性质 **K***, 需要利用下面定理 2 的命题 a), 因此下面再证明 (见 232 页).

定理 2 (数学期望符号下的收敛性) 设 $\{\xi_n\}_{n\geqslant 1}$ 是 (广义) 随机变量序列.

a) 若 $|\xi_n| \leqslant \eta$, $\mathbf{E}\eta < \infty$ 且 $\xi_n \to \xi$ (a.c.), 则

$$\mathbf{E}(\xi_n|\mathscr{G}) \to \mathbf{E}(\xi|\mathscr{G}) \text{ (a.c.)},$$

且

$$\mathbf{E}(|\xi_n - \xi||\mathscr{G}) \to 0 \text{ (a.c.)}.$$

b) 若 $\xi_n \geqslant \eta$, $\mathbf{E}\eta > -\infty$ 且 $\xi_n \uparrow \xi$ (a.c.), 则

$$\mathbf{E}(\xi_n|\mathscr{G}) \uparrow \mathbf{E}(\xi|\mathscr{G}) \text{ (a.c.)}.$$

c) 若 $\xi_n \leqslant \eta$, $\mathbf{E}\eta < \infty$ 且 $\xi_n \downarrow \xi$ (a.c.), 则

$$\mathbf{E}(\xi_n|\mathscr{G}) \downarrow \mathbf{E}(\xi|\mathscr{G}) \text{ (a.c.)}.$$

d) 若 $\xi_n \geqslant \eta$, $\mathbf{E}\eta > -\infty$, 则

$$\mathbf{E}(\underline{\lim}\,\xi_n|\mathscr{G}) \leqslant \underline{\lim}\,\mathbf{E}(\xi_n|\mathscr{G}) \text{ (a.c.)}.$$

e) 若 $\xi_n \leqslant \eta$, $\mathbf{E}\eta < \infty$, 则

$$\overline{\lim}\,\mathbf{E}(\xi_n|\mathscr{G}) \leqslant \mathbf{E}(\overline{\lim}\,\xi_n|\mathscr{G}) \text{ (a.c.)}.$$

f) 若 $\xi_n \geqslant 0$, 则

$$\mathbf{E}\left(\sum \xi_n|\mathscr{G}\right) = \sum \mathbf{E}(\xi_n|\mathscr{G}) \text{ (a.c.)}.$$

证明 a) 设

$$\zeta_n = \sup_{m\geqslant n} |\xi_m - \xi|.$$

由于 $\xi_n \to \xi$ (a.c.), 可见 $\zeta_n \downarrow 0$ (a.c.). 数学期望 $\mathbf{E}\xi_n$ 和 $\mathbf{E}\xi$ 有限, 由于性质 **D*** 和 **C*** (a.c.),

$$|\mathbf{E}(\xi_n|\mathscr{G}) - \mathbf{E}(\xi|\mathscr{G})| = |\mathbf{E}(\xi_n - \xi|\mathscr{G})| \leqslant \mathbf{E}(|\xi_n - \xi||\mathscr{G}) \leqslant \mathbf{E}(\zeta_n|\mathscr{G}).$$

由于 $\mathbf{E}(\zeta_{n+1}|\mathscr{G}) \leqslant \mathbf{E}(\zeta_n|\mathscr{G})$ (a.c.), 则存在极限 $h = \lim_n \mathbf{E}(\zeta_n|\mathscr{G})$ (a.c.). 那么,

$$0 \leqslant \int_\Omega h\mathrm{d}\mathbf{P} \leqslant \int_\Omega \mathbf{E}(\zeta_n|\mathscr{G})\mathrm{d}\mathbf{P} = \int_\Omega \zeta_n \mathrm{d}\mathbf{P} \to 0, \quad n \to \infty,$$

其中由于 $0 \leqslant \zeta_n \leqslant 2\eta$, $\mathbf{E}\eta < \infty$, 可见最后的结果可以由控制收敛定理得出. 从而,

$$\int_\Omega h\mathrm{d}\mathbf{P} = 0,$$

只由 §6 第 3 小节的性质 **H**, 知 $h = 0$ (a.c.).

b) 设 $\eta \equiv 0$. 由于 $\mathbf{E}(\xi_n|\mathscr{G}) \leqslant \mathbf{E}(\xi_{n+1}|\mathscr{G})$ (a.c.), 则存在极限 $\zeta(\omega) = \lim_n \mathbf{E}(\xi_n|\mathscr{G})$ (a.c.). 那么, 由等式

$$\int_A \xi_n \mathrm{d}\mathbf{P} = \int_A \mathbf{E}(\xi_n|\mathscr{G})\mathrm{d}\mathbf{P}, \quad A \in \mathscr{G},$$

和控制收敛定理, 有

$$\int_A \xi \mathrm{d}\mathbf{P} = \int_A \mathbf{E}(\xi|\mathscr{G})\mathrm{d}\mathbf{P} = \int_A \zeta \mathrm{d}\mathbf{P}, \quad A \in \mathscr{G}.$$

从而, 根据与 §6 第 3 小节的性质 I 类似的性质和 §6 的练习题 5, 可见 $\xi = \zeta$ (a.c.).

为证明一般情形, 注意到, $0 \leqslant \xi_n^+ \uparrow \xi^+$, 且根据已经证明的

$$\mathbf{E}(\xi_n^+|\mathscr{G}) \uparrow \mathbf{E}(\xi^+|\mathscr{G}) \text{ (a.c.)}. \tag{7}$$

由于 $0 \leqslant \xi_n^- \leqslant \xi^-$，$\mathbf{E}\xi^- < \infty$，故由 a) 可见

$$\mathbf{E}(\xi_n^-|\mathscr{G}) \to \mathbf{E}(\xi^-|\mathscr{G}).$$

于是，由此连同 (7) 式，命题 b) 得证.

由 b) 得命题 c).

d) 设 $\zeta_n = \inf_{m \geqslant n} \xi_m$，则 $\zeta_n \uparrow \zeta$，其中 $\zeta = \underline{\lim}\, \xi_n$. 根据 b)，$\mathbf{E}(\zeta_n|\mathscr{G}) \uparrow \mathbf{E}(\zeta|\mathscr{G})$ (a.c.)，因此

$$\mathbf{E}(\underline{\lim}\, \xi_n|\mathscr{G}) = \mathbf{E}(\zeta|\mathscr{G}) = \lim_n \mathbf{E}(\zeta_n|\mathscr{G}) = \underline{\lim}_n \mathbf{E}(\zeta_n|\mathscr{G}) \leqslant \underline{\lim}\, \mathbf{E}(\xi_n|\mathscr{G}) \quad (\text{a.c.}).$$

由 d) 得命题 e).

f) 如果 $\xi_n \geqslant 0$，则由命题 $\mathbf{D}^*$，可见

$$\mathbf{E}\left(\sum_{k=1}^{n} \xi_k \Big| \mathscr{G}\right) = \sum_{k=1}^{n} \mathbf{E}(\xi_k|\mathscr{G}) \quad (\text{a.c.}),$$

由此连同 b) 即可证明所要求的结果. □

现在证明命题 $\mathbf{K}^*$. 设 $\eta = I_B$，$B \in \mathscr{G}$. 那么，对于任意 $A \in \mathscr{G}$，有

$$\int_A \xi\eta \,d\mathbf{P} = \int_{A\cap B} \xi \,d\mathbf{P} = \int_{A\cap B} \mathbf{E}(\xi|\mathscr{G})\,d\mathbf{P} = \int_A I_B \mathbf{E}(\xi|\mathscr{G})\,d\mathbf{P} = \int_A \eta \mathbf{E}(\xi|\mathscr{G})\,d\mathbf{P}.$$

由于勒贝格积分的可加性，等式

$$\int_A \xi\eta \,d\mathbf{P} = \int_A \eta \mathbf{E}(\xi|\mathscr{G})\,d\mathbf{P},\ A \in \mathscr{G}, \tag{8}$$

对于简单随机变量

$$\eta = \sum_{k=1}^{n} y_k I_{B_k},\ B_k \in \mathscr{G},$$

仍然成立. 所以由 §6 第 3 小节的性质 I，对于简单随机变量，有

$$\mathbf{E}(\xi\eta|\mathscr{G}) = \eta \mathbf{E}(\xi|\mathscr{G}) \quad (\text{a.c.}). \tag{9}$$

现在假设 η 是任意 $\mathscr{G}$ - 可测随机变量，而 $\{\eta_n\}_{n\geqslant 1}$ 是 $\mathscr{G}$ - 可测简单随机变量序列，且 $|\eta_n| \leqslant |\eta|$，$\eta_n \to \eta$. 那么，由 (9) 式，可见

$$\mathbf{E}(\xi\eta_n|\mathscr{G}) = \eta_n \mathbf{E}(\xi|\mathscr{G}) \quad (\text{a.c.}).$$

显然 $|\xi\eta_n| \leqslant |\xi\eta|$，其中 $|\xi\eta| < \infty$. 因此，根据定理 2 的命题 a)，

$$\mathbf{E}(\xi\eta_n|\mathscr{G}) \to \mathbf{E}(\xi\eta|\mathscr{G}) \quad (\text{a.c.}).$$

其次，因为 $\mathbf{E}|\xi| < \infty$，则 $\mathbf{E}(\xi|\mathscr{G})$ 有限 (a.c.) (见性质 $\mathbf{C}^*$，§6 第 3 小节的性质 $\mathbf{J}$). 于是 $\eta_n \mathbf{E}(\xi|\mathscr{G}) \to \eta \mathbf{E}(\xi|\mathscr{G})$ (a.c.). (关于 $\mathbf{E}(\xi|\mathscr{G})$ 几乎必然有限的假设是重要的，因为在 §4 第 4 小节曾约定：$0 \times \infty = 0$，但是如果 $\eta_n = 1/n$，$\eta \equiv 0$ 就会出现 $(1/n) \times \infty$ 不趋向 $0 \times \infty = 0$ 的情形).

注 如果仅满足条件：η 为 $\mathscr{G}$ - 可测，且 $\mathbf{E}(\xi|\mathscr{G})$ 有定义，则命题 $\mathbf{K}^*$ 仍然成立.

5. 条件数学期望 $\mathbf{E}(\xi|\mathscr{G}_\eta)$ 的结构 现在比较详细地讨论条件数学期望 $\mathbf{E}(\xi|\mathscr{G}_\eta)$ 的结构，其中上面曾约定 $\mathbf{E}(\xi|\mathscr{G}_\eta)$ 也通过 $\mathbf{E}(\xi|\eta)$ 表示.

由于 $\mathbf{E}(\xi|\eta)$ 是 $\mathscr{G}_\eta$ - 可测函数，根据 §4 的定理 3(比较确切地说，该定理对于广义随机变量明显的变式)，存在定义在 $\overline{\mathbb{R}}$ 上取值于 $\overline{\mathbb{R}}$ 的博雷尔函数 $m = m(x)$，使对于一切 $\omega \in \Omega$，有

$$m(\eta(\omega)) = \mathbf{E}(\xi|\eta)(\omega). \tag{10}$$

我们用 $\mathbf{E}(\xi|\eta = y)$ 表示函数 $m(y)$，并称做 ξ 关于事件 $\{\eta = y\}$ 的条件数学期望，或 ξ 在 $\eta = y$ 条件下的条件数学期望.

根据定义

$$\int_A \xi \,d\mathbf{P} = \int_A \mathbf{E}(\xi|\eta)\,d\mathbf{P} = \int_A m(\eta)\,d\mathbf{P},\ A \in \mathscr{G}_\eta. \tag{11}$$

因此，根据 §6 (关于勒贝格积分变量替换) 的定理 7，有

$$\int_{\{\omega:\eta\in B\}} m(\eta)\,d\mathbf{P} = \int_B m(y) P_\eta(dy),\ B \in \mathscr{B}(\overline{\mathbb{R}}), \tag{12}$$

其中 P_η 是 η 的概率分布. 从而 $m = m(y)$ 是博雷尔函数，且对于一切 $B \in \mathscr{B}(\overline{\mathbb{R}})$，有

$$\int_{\{\omega:\eta\in B\}} \xi \,d\mathbf{P} = \int_B m(y) P_\eta(dy). \tag{13}$$

这说明可以经另一途径定义条件数学期望 $\mathbf{E}(\xi|\eta = y)$.

定义 4 设 ξ 和 η 是随机变量 (也可能是广义的)，且存在 $\mathbf{E}\xi$. 我们称任何满足

$$\int_{\{\omega:\eta\in B\}} \xi \,d\mathbf{P} = \int_B m(y) P_\eta(dy),\ B \in \mathscr{B}(\overline{\mathbb{R}}), \tag{14}$$

的 $\mathscr{B}(\overline{\mathbb{R}})$ - 可测函数 $m = m(x)$，为随机变量 ξ 在 $\eta = y$ 条件下的条件数学期望.

如果注意到集函数

$$\mathbf{Q}(B) = \int_{\{\omega:\eta\in B\}} \xi \,d\mathbf{P}$$

是带符号的测度，并且关于测度 P_η 绝对连续，则仍然由拉东 – 尼科迪姆定理可见函数 $m = m(x)$ 存在.

现在假设 $m(y)$ 是按定义 4 的条件数学期望. 那么，仍利用关于勒贝格积分变量替换的定理，可得

$$\int_{\{\omega:\eta\in B\}} \xi \,d\mathbf{P} = \int_B m(y) P_\eta(dy) = \int_{\{\omega:\eta\in B\}} m(\eta)\,d\mathbf{P},\ B \in \mathscr{B}(\overline{\mathbb{R}}).$$

函数 $m(\eta)$ 为 $\mathscr{G}_\eta$ - 可测，且集合 $\{\omega : \eta \in B\}$，$B \in \mathscr{B}(\overline{\mathbb{R}})$ 包含 $\mathscr{G}_\eta$ 中所有集合.

由此可见，$m(\eta)$ 是数学期望 $\mathbf{E}(\xi|\eta)$. 于是，由 $\mathbf{E}(\xi|\eta = y)$ 可以得到 $\mathbf{E}(\xi|\eta)$；相反，由 $\mathbf{E}(\xi|\eta)$ 可以得到 $\mathbf{E}(\xi|\eta = y)$.

直观上，条件数学期望 $m(y)=\mathbf{E}(\xi|\eta=y)$ 是比 $\mathbf{E}(\xi|\eta)$ 更简单明了的对象. 不过，条件数学期望 $\mathbf{E}(\xi|\eta)$ 作为 $\mathscr{G}_\eta$ - 可测随机变量更便于使用.

应该指出，前面引进的性质 $\mathrm{A}^*\sim\mathrm{K}^*$，以及定理 2 论断，容易移植到条件数学期望 $\mathbf{E}(\xi|\eta=y)$(为此，只需将"几乎必然"换成"P_η－几乎必然"). 例如，性质 K^* 可以转述为：如果，$\mathbf{E}|\xi|<\infty$，$\mathbf{E}|\xi f(\eta)|<\infty$，其中 $f=f(y)$ 是 $\mathscr{B}(\overline{\mathbf{R}})$－可测函数，则

$$\mathbf{E}(\xi f(\eta)|\eta=y)=f(y)\mathbf{E}(\xi|\eta=y)\ (P_\eta-\text{a.c.}). \tag{15}$$

其次 (对照性质 J^*). 设 ξ 和 η 独立，则

$$\mathbf{E}(\xi|\eta=y)=\mathbf{E}\xi\quad(P_\eta-\text{a.c.}).$$

还要指出，如果 $B\in\mathscr{B}(\mathbf{R}^2)$ 且 ξ 和 η 独立，则

$$\mathbf{E}[I_B(\xi,\eta)|\eta=y]=\mathbf{E}I_B(\xi,y)\quad(P_\eta-\text{a.c.}). \tag{16}$$

如果 $\varphi=\varphi(x,y)$ 是 $\mathscr{B}(\mathbf{R}^2)$－可测函数，使 $\mathbf{E}|\varphi(\xi,\eta)|<\infty$，则

$$\mathbf{E}[\varphi(\xi,\eta)|\eta=y]=\mathbf{E}\varphi(\xi,y)\quad(P_\eta-\text{a.c.}).$$

为证明 (16) 式，我们指出以下事实. 假如 $B=B_1\times B_2$，则为证明 (16) 式，只需验证

$$\int_{\{\omega:\eta\in A\}}I_{B_1\times B_2}(\xi,\eta)\mathbf{P}(\mathrm{d}\omega)=\int_{\{y\in A\}}\mathbf{E}I_{B_1\times B_2}(\xi,y)P_\eta(\mathrm{d}y).$$

等式左侧是 $\mathbf{P}\{\xi\in B_1,\eta\in A\cap B_2\}$，右侧是 $\mathbf{P}\{\xi\in B_1\}\mathbf{P}\{\eta\in A\cap B_2\}$，而由于 ξ 和 η 独立，可见左、右两侧概率相等. 一般情形的证明，要运用 §2 中关于单调类的定理 1 (对照傅比尼定理证明中的相应方法).

定义 5 *在 $\eta=y$ 的条件下，$\mathbf{E}(I_A|\eta=y)$ 称做事件 $A\in\mathscr{F}$ 的条件概率，记作 $\mathbf{P}(A|\eta=y)$.*

显然，$\mathbf{P}(A|\eta=y)$ 也可以定义为满足

$$\mathbf{P}(A\cap\{\eta\in B\})=\int_B\mathbf{P}(A|\eta=y)P_\eta(\mathrm{d}y),\quad B\in\mathscr{B}(\overline{\mathbf{R}}), \tag{17}$$

的可测函数.

6. 计算条件概率和条件数学期望的例

例 1 设 η 是离散型随机变量：

$$\mathbf{P}\{\eta=y_k\}>0,\quad\sum_{k=1}^\infty\mathbf{P}\{\eta=y_k\}=1.$$

那么

$$\mathbf{P}(A|\eta=y_k)=\frac{\mathbf{P}(A\cap\{\eta=y_k\})}{\mathbf{P}\{\eta=y_k\}},\quad k\geqslant1.$$

对于 $y\notin\{y_1,y_2,\cdots\}$，条件概率 $\mathbf{P}(A|\eta=y)$ 可以任意地定义，例如，令其等于 0.

假设随机变量 ξ 的数学期望 $\mathbf{E}\xi$ 存在，则

$$\mathbf{E}(\xi|\eta=y_k)=\frac{1}{\mathbf{P}\{\eta=y_k\}}\int_{\{\omega:\eta=y_k\}}\xi\mathrm{d}\mathbf{P}.$$

对于 $y\notin\{y_1,y_2,\cdots\}$，条件数学期望 $\mathbf{E}(\xi|\eta=y)$ 可以任意地定义，例如，令其等于 0.

例 2 设 (ξ,η) 是随机向量，其分布具有密度 $f_{\xi,\eta}(x,y)$：

$$\mathbf{P}\{(\xi,\eta)\in B\}=\int_B f_{\xi,\eta}(x,y)\mathrm{d}x\mathrm{d}y,\quad B\in\mathscr{B}(\mathbf{R}^2).$$

设 $f_\xi(x)$ 和 $f_\eta(y)$ 是随机变量 ξ 和 η 的概率分布密度 (见 §6 (46),(55),(56)). 记

$$f_{\xi|\eta}(x|y)=\frac{f_{\xi,\eta}(x,y)}{f_\eta(y)}, \tag{18}$$

其中，如果 $f_\eta(y)=0$，则设 $f_{\xi|\eta}(x|y)=0$.

那么，

$$\mathbf{P}\{\xi\in C|\eta=y\}=\int_C f_{\xi|\eta}(x|y)\mathrm{d}x,\quad C\in\mathscr{B}(\mathbf{R}), \tag{19}$$

即 $f_{\xi|\eta}(x|y)$ 是条件概率分布密度.

事实上，为证明 (19) 式，只需对于 $B\in\mathscr{B}(\mathbf{R})$，$A=\{\xi\in C\}$，证明 (17) 式. 由 §6 的 (43) 式和 (45) 式以及傅比尼定理，可见

$$\begin{aligned}\int_B\left[\int_C f_{\xi|\eta}(x|y)\mathrm{d}x\right]P_\eta(\mathrm{d}y)&=\int_B\left[\int_C f_{\xi|\eta}(x|y)\mathrm{d}x\right]f_\eta(y)\mathrm{d}y\\&=\int_{C\times B}f_{\xi|\eta}(x|y)f_\eta(y)\mathrm{d}x\mathrm{d}y=\int_{C\times B}f_{\xi,\eta}(x,y)\mathrm{d}x\mathrm{d}y\\&=\mathbf{P}\{(\xi,\eta)\in C\times B\}=\mathbf{P}(\{\xi\in C\}\cap\{\eta\in B\}),\end{aligned}$$

于是 (17) 式得证.

如果 $\mathbf{E}\xi$ 存在，则类似可得

$$\mathbf{E}(\xi|\eta=y)=\int_{-\infty}^\infty xf_{\xi|\eta}(x|y)\mathrm{d}x. \tag{20}$$

例 3 假设某设备工作的时间是一非负随机变量 $\eta=\eta(\omega)$，其分布函数 $F_\eta(y)$ 有密度 $f_\eta(y)$(对于 $y<0$，自然 $F_\eta(y)=f_\eta(y)=0$). 求条件数学期望 $\mathbf{E}(\eta-a|\eta\geqslant a)$，即在设备已经工作了时间 a 的条件下，它还能再继续工作的平均时间.

设 $\mathbf{P}\{\eta\geqslant a\}>0$. 那么，根据定义 (见第 1 小节) 和 §6 的 (45) 式，可见

$$\mathbf{E}(\eta-a|\eta\geqslant a)=\frac{\mathbf{E}[(\eta-a)I_{\{\eta\geqslant a\}}]}{\mathbf{P}\{\eta\geqslant a\}}=\frac{\int_\Omega(\eta-a)I_{\{\eta\geqslant a\}}\mathbf{P}(\mathrm{d}\omega)}{\mathbf{P}\{\eta\geqslant a\}}=\frac{\int_a^\infty(y-a)f_\eta(y)\mathrm{d}y}{\int_a^\infty f_\eta(y)\mathrm{d}y}.$$

有趣的是，如果随机变量 η 服从指数分布，即

$$f_\eta(y)=\begin{cases}\lambda e^{-\lambda y}, & 若\ y\geqslant 0,\\ 0, & 若\ y<0,\end{cases} \tag{21}$$

则 $\mathbf{E}\eta=\mathbf{E}(\eta|\eta\geqslant 0)=1/\lambda$，且对于任意 $a>0$，$\mathbf{E}(\eta-a|\eta\geqslant a)=1/\lambda$. 换句话说，在设备已经工作了时间 a 的条件下，它还能再继续工作的平均时间，与 a 的值无关，就等于 $\mathbf{E}\eta$.

在 (21) 式的条件下，求条件分布 $\mathbf{P}(\eta-a\leqslant x|\eta\geqslant a)$. 有

$$\mathbf{P}(\eta-a\leqslant x|\eta\geqslant a)=\frac{\mathbf{P}\{a\leqslant\eta\leqslant a+x\}}{\mathbf{P}\{\eta\geqslant a\}}=\frac{F_\eta(a+x)-F_\eta(a)+\mathbf{P}\{\eta=a\}}{1-F_\eta(a)+\mathbf{P}\{\eta=a\}}$$
$$=\frac{[1-e^{-\lambda(a+x)}]-[1-e^{-\lambda a}]}{1-[1-e^{-\lambda a}]}=\frac{e^{-\lambda a}[1-e^{-\lambda x}]}{e^{-\lambda a}}=1-e^{-\lambda x}.$$

于是，条件分布 $\mathbf{P}(\eta-a\leqslant x|\eta\geqslant a)$ 等于无条件分布 $\mathbf{P}\{\eta\leqslant x\}$. 指数分布这一特别好的性质是特有的：不存在其他分布，使其密度具有同样性质 $\mathbf{P}(\eta-a\leqslant x|\eta\geqslant a)=\mathbf{P}\{\eta\leqslant x\}, a\geqslant 0, 0\leqslant x<\infty$.

例 4 (蒲丰 [G. de Buffon] 掷针问题) 假设向平面上一无限长和单位宽的 "走廊" (两条直线所夹的区域) 上 (图 29)，"随机" 地投掷一单位长度的针. 问针 (至少) 与 "走廊" 的一侧相交的概率如何?

为解决所提 (几何概率的) 问题，首先说清何谓 "随机" 地投掷一单位长度的针. 以 ξ 表示针落地后其中点到 "走廊" 左侧的距离. 假设在区间 [0,1] 上均匀分布，而角 θ 在 $[-\pi/2,\pi/2]$ 上均匀 ($F_\theta(da)=da$) 分布 (图 29). 此外，假设 ξ 和 θ 独立.

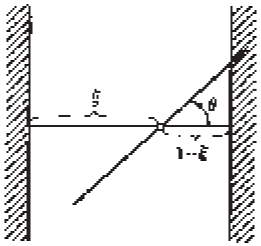


图 29

设 A 是表示事件针与 "走廊" 的某一侧相交. 易见，如果

$$B=\left\{(a,x):|a|\leqslant\frac{\pi}{2}, x\in\left[0,\frac{1}{2}\cos a\right]\cup\left[1-\frac{1}{2}\cos a,1\right]\right\},$$

则 $A=\{\omega:(\theta,\xi)\in B\}$. 因而所求概率为

$$\mathbf{P}(A)=\mathbf{E}I_A(\omega)=\mathbf{E}I_B(\theta(\omega),\xi(\omega)).$$

由于性质 G' 和 (16) 式，有

$$\begin{aligned}\mathbf{E}I_B(\theta(\omega),\xi(\omega))&=\mathbf{E}[\mathbf{E}(I_B(\theta(\omega),\xi(\omega))|\theta(\omega))]\\&=\int_\Omega\mathbf{E}[I_B(\theta(\omega),\xi(\omega))|\theta(\omega)]\mathbf{P}(d\omega)\\&=\int_{-\pi/2}^{\pi/2}\mathbf{E}[I_B(\theta(\omega),\xi(\omega))|\theta(\omega)=a]F_\theta(da)\\&=\frac{1}{\pi}\int_{-\pi/2}^{\pi/2}\mathbf{E}I_B(a,\xi(\omega))da=\frac{1}{\pi}\int_{-\pi/2}^{\pi/2}\cos a\,da=\frac{2}{\pi},\end{aligned}$$

其中用到

$$\mathbf{E}I_B(a,\xi(\omega))=\mathbf{P}\left\{\xi\in\left[0,\frac{1}{2}\cos a\right]\cup\left[1-\frac{1}{2}\cos a,1\right]\right\}=\cos a.$$

这样，"随机地" 向 "走廊" 中投掷的针落地后，至少与 "走廊" 的一侧相交的概率等于 $2/\pi$. 这一结果，可以作为通过试验确定 π 值的基础. 实际上，假设将针独立地掷 N 次，并定义随机变量

$$\eta_i=\begin{cases}1, & 若第\ i\ 次掷针与走廊侧边相交,\\ 0, & 若第\ i\ 次掷针与走廊侧边不交.\end{cases}$$

那么，由大数定律 (例如，见第一章 §5 (6) 式)，可见对于任意 $\varepsilon>0$，

$$\mathbf{P}\left\{\left|\frac{\eta_1+\cdots+\eta_N}{N}-\mathbf{P}(A)\right|>\varepsilon\right\}\to 0,\quad N\to\infty.$$

在此意义上，频率

$$\frac{\eta_1+\cdots+\eta_N}{N}\approx\mathbf{P}(A)=\frac{2}{\pi},$$

于是

$$\frac{2N}{\eta_1+\cdots+\eta_N}\approx\pi.$$

正是这一公式曾经是用统计方法确定 π 值的基础. 在 1850 年 R. 沃尔弗 (R. Wolf) 掷针 5000 次，结果得的值 3.1596. 看来，这种方法，是在数值分析中利用概率统计方法最早的用法之一，也就是现代熟知的 "蒙特卡罗法".

注 所讨论的例 4 (蒲丰掷针问题) 是几何概率的典型问题. 在这类问题中，相当常见的是，由诸如 "对称性" 之类的简单几何直观，可见如何确定 "一组基本事件" 的概率. (对照第一章 §1 第 3 和 4 小节，以及这一章的 §3.) 下面练习题的第 9~12 题是几何概率的题

7. 条件概率的正则性 设 $\{\xi_n\}_{n\geqslant 1}$ 是非负随机变量序列，则可见定理 2 的命题 f)，

$$\mathbf{E}\left(\sum\xi_n|\mathscr{G}\right)=\sum\mathbf{E}(\xi_n|\mathscr{G})\quad(\text{a.c.}).$$

特别, 若 $B_1, B_2, \cdots$ 是两两不相交的集合, 则

$$\mathbf{P}\left(\sum B_n|\mathscr{G}\right) = \sum \mathbf{P}(B_n|\mathscr{G}) \quad (\text{a.c.}). \tag{22}$$

重要的是强调这一等式只是几乎必然成立. 因而, 对于固定的 ω 不能将条件概率 $\mathbf{P}(B|\mathscr{G})(\omega)$ 看成 B 的测度. 也许除去一个零测集 $\mathscr{N}$ 之后, 就可以对于每一个 $\omega \notin \mathscr{N}$, 使得 $\mathbf{P}(\cdot|\mathscr{G})(\omega)$ 是概率测度. 然而, 一般并非如此, 看下面的例子. 对于给定的 $B_1, B_2, \cdots$, 以 $\mathscr{N}(B_1, B_2, \cdots)$ 表示不满足可数可加性 (22) 式的结局 ω 的集合. 那么, 特殊集合 $\mathscr{N}$ 为:

$$\mathscr{N} = \bigcup \mathscr{N}(B_1, B_2, \cdots), \tag{23}$$

其中对 $\mathscr{F}$ 中一切两两不相交的集合 $B_1, B_2, \cdots$ 求并. 虽然每一个集合的 $\mathbf{P}$-测度等于 0, 然而集合 $\mathscr{N}$ 的 $\mathbf{P}$-测度可能不等于 0 (因为 (23) 式的并含不可数项). 例如, 个别点的勒贝格测度为 0, 而集合 $\mathscr{N} = [0,1)$ 的测度等于 1, 因为 $\mathscr{N} = [0,1)$ 是单点集 $\{x\}, 0 \leqslant x < 1$, 的不可数的和.

假如对于每一个 $\omega \in \Omega$, 当条件概率 $\mathbf{P}(\cdot|\mathscr{G})(\omega)$ 是概率测度时, 将更为方便. 例如, 因为这时计算条件数学期望 $\mathbf{E}(\xi|\mathscr{G})(\omega)$, 可以利用对测度 $\mathbf{P}(\cdot|\mathscr{G})$ 求平均

$$\mathbf{E}(\xi|\mathscr{G})(\omega) = \int_\Omega \xi(\tilde\omega)\mathbf{P}(d\tilde\omega|\mathscr{G})(\omega) \quad (\text{a.c.})$$

简单地实现 (对照第一章 §8 (10) 式).

现在引进正则条件概率的概念.

定义 6 对于一切 $\omega \in \Omega, B \in \mathscr{F}$, 定义的函数 $P(\omega; B)$, 称做关于 σ-代数 $\mathscr{G} \subseteq \mathscr{F}$ 为正则条件概率, 如果:

a) 对于每一个 $\omega \in \Omega$, $P(\omega; \cdot)$ 是 $\mathscr{F}$ 上的概率测度;

b) 对于每一个 $B \in \mathscr{F}$, $P(\omega; B)$ 作为 ω 的函数是条件概率 $\mathbf{P}(B|\mathscr{G})(\omega)$ 的变式之一, 即

$$P(\omega; B) = \mathbf{P}(B|\mathscr{G})(\omega) \quad (\text{a.c.}).$$

定理 3 设 $P(\omega; B)$ 关于 $\mathscr{G}$ 是正则条件概率, ξ 是可积随机变量, 则

$$\mathbf{E}(\xi|\mathscr{G})(\omega) = \int_\Omega \xi(\tilde\omega)P(\omega; d\tilde\omega) \quad (\text{a.c.}). \tag{24}$$

证明 如果 $\xi = I_B, B \in \mathscr{F}$, 则要证明的 (24) 式变为

$$\mathbf{P}(B|\mathscr{G})(\omega) = P(\omega; B) \quad (\text{a.c.}),$$

并且由定义 6 的 b), 知此等式成立. 从而, (24) 式对于简单函数成立.

现在设 $\xi \geqslant 0$ 和 $\xi_n \uparrow \xi$, 其中 ξ_n 是简单函数. 那么, 由定理 2 的性质 h), 有

$$\mathbf{E}(\xi|\mathscr{G})(\omega) = \lim_n \mathbf{E}(\xi_n|\mathscr{G})(\omega) \quad (\text{a.c.}).$$

因为对于每一个 $\omega \in \Omega$, $P(\omega; \cdot)$ 是测度, 则根据单调收敛定理, 有

$$\lim_n \mathbf{E}(\xi_n|\mathscr{G})(\omega) = \lim_n \int_\Omega \xi_n(\tilde\omega)P(\omega; d\tilde\omega) = \int_\Omega \xi(\tilde\omega)P(\omega; d\tilde\omega).$$

由于 $\xi = \xi^+ - \xi^-$, 故一般情形可以归结为上面已经证明的情形. □

系 设 $\mathscr{G} = \mathscr{G}_\eta$, 其中 η 是随机变量, 而 (ξ, η) 是随机向量, 且具有概率分布密度 $f_{\xi,\eta}(x, y)$. 假设 $\mathbf{E}|g(\xi)| < \infty$, 那么

$$\mathbf{E}[g(\xi)|\eta = y] = \int_{-\infty}^{\infty} g(x) f_{\xi|\eta}(x|y) dx,$$

其中 $f_{\xi|\eta}(x|y)$ 是条件分布密度 (见 (18) 式).

为表述正则条件概率存在的基本结果, 需要下面的定义.

定义 7 设 $(E, \mathscr{E})$ 是可测空间, $X = X(\omega)$ 是值空间为 E 的随机元, $\mathscr{G}$ 是 $\mathscr{F}$ 的 σ-子代数. 对于 $\omega \in \Omega$ 和 $B \in \mathscr{E}$ 定义的函数 $Q(\omega; B)$, 称做 X 关于 σ-代数 $\mathscr{G}$ 的正则条件分布, 如果:

a) 对于任意 $\omega \in \Omega$, $Q(\omega; B)$ 是 $(E, \mathscr{E})$ 上的概率测度;

b) 对于每一个集 $B \in \mathscr{E}$, $Q(\omega; B)$ 作为 ω 的函数, 是条件概率 $\mathbf{P}(X \in B|\mathscr{G})(\omega)$ 的变式之一, 即

$$Q(\omega; B) = \mathbf{P}(X \in B|\mathscr{G})(\omega) \quad (\text{a.c.}).$$

定义 8 设 ξ 是随机变量. 函数 $F(\omega; x), \omega \in \Omega, x \in \mathbb{R}$ 称做 ξ 关于 σ-代数 $\mathscr{G}$ 的正则分布函数, 如果:

a) 对于每一 $\omega \in \Omega$, $F(\omega; x)$ 是 $\mathbb{R}$ 上的分布函数;

b) 对于每一 $x \in \mathbb{R}$, $F(\omega; x) = \mathbf{P}(\xi \leqslant x|\mathscr{G})(\omega) \quad (\text{a.c.})$.

定理 4 随机变量 ξ 关于 σ-代数 $\mathscr{G} \subseteq \mathscr{F}$ 的正则分布函数和正则条件分布永远存在.

证明 对于每一个有理数 $r \in \mathbb{R}$, 设

$$F_r(\omega) = \mathbf{P}(\xi \leqslant r|\mathscr{G})(\omega), \text{ 其中 } \mathbf{P}(\xi \leqslant r|\mathscr{G})(\omega) = \mathbf{E}(I_{\{\xi \leqslant r\}}|\mathscr{G})(\omega)$$

是事件 $\{\xi \leqslant r\}$ 关于 $\mathscr{G}$ 的条件概率的一种变式. 设 $\{r_i\}$ 是 $\mathbb{R}$ 上有理数集. 如果 $r_i < r_j$, 则由于性质 B*, 有 $\mathbf{P}(\xi \leqslant r_i|\mathscr{G}) \leqslant \mathbf{P}(\xi \leqslant r_j|\mathscr{G})$ (a.c.). 这说明, 若

$$A_{ij} = \{\omega : F_{r_j}(\omega) < F_{r_i}(\omega)\}, \quad A = \bigcup A_{ij},$$

则 $\mathbf{P}(A) = 0$. 即使分布函数 $F_r(\omega), r \in \{r_i\}$, 不单调的集合的测度等于 0.

现在, 设

$$B_i = \{\omega : \lim_{n\to\infty} F_{r_i + 1/n}(\omega) \neq F_{r_i}(\omega)\}, \quad B = \bigcup_{i=1}^{\infty} B_i.$$

显然 $I_{\{\xi\le r_i+1/n\}}\downarrow I_{\{\xi\le r_i\}}, n\to\infty$. 因此，根据定理 2 的命题 a)，有 $F_{r_i+1/n}(\omega)\downarrow F_{r_i}(\omega)$ (a.e.). 说明（在有理点上）不右连续点的集合 B 的测度也是 0：$\mathbf{P}(B)=0$.

其次，设

$$C=\{\omega:\lim_{n\to\infty}F_n(\omega)\neq 1\}\cup\{\omega:\lim_{n\to-\infty}F_n(\omega)\neq 0\}.$$

那么，由于 $\{\xi\leqslant n\}\uparrow\Omega, n\to\infty$，而 $\{\xi\leqslant n\}\downarrow\varnothing, n\to-\infty$，则 $\mathbf{P}(C)=0$.

现在，设

$$F(\omega;x)=\begin{cases}\lim\limits_{r\downarrow x}F_r(\omega), & 若\ \omega\notin A\cup B\cup C,\\ G(x), & 若\ \omega\in A\cup B\cup C,\end{cases}$$

其中 $G(x)$ 是 $\mathbf{R}$ 上的任意分布函数. 下面证明函数 $F(\omega;x)$ 满足定义 8.

设 $\omega\notin A\bigcup B\bigcup C$，那么 $F(\omega;x)$ 显然是 x 的不减函数. 对于 $x<x'\leqslant r$，当 $r\downarrow x$ 时，则

$$F(\omega;x)\leqslant F(\omega;x')\leqslant F(\omega;r)=F_r(\omega)\downarrow F(\omega;x).$$

因此 $F(\omega;x)$ 右连续. 类似，有

$$\lim_{x\to\infty}F(\omega;x)=1,\quad \lim_{x\to-\infty}F(\omega;x)=0.$$

由于当 $\omega\in A\bigcup B\bigcup C$ 时，$F(\omega;x)=G(x)$，则对于每个 $\omega\in\Omega$，$F(\omega;x)$ 是 $\mathbf{R}$ 上的分布函数，即满足定义 8 中的条件 a).

根据构造，$\mathbf{P}(\xi\leqslant r|\mathscr{G})(\omega)=F_r(\omega)=F(\omega;r)$. 如果 $r\downarrow x$，则由于已经证明的 $F(\omega;x)$ 的右连续性，可见对于一切 $\omega\in\Omega$，有 $F(\omega;r)\downarrow F(\omega;x)$. 由定理 2 的命题 a) 知，几乎处处有 $\mathbf{P}(\xi\leqslant r|\mathscr{G})(\omega)\to\mathbf{P}(\xi\leqslant x|\mathscr{G})(\omega)$. 因此，$F(\omega;x)=\mathbf{P}(\xi\leqslant x|\mathscr{G})(\omega)$ (a.e.). 于是，定义 8 的性质 b) 得证.

现在证明变量 ξ 关于 $\mathscr{G}$ 的正则条件分布存在.

假设 $F(\omega;x)$ 是上面建立的函数. 设

$$Q(\omega;B)=\int_B F(\omega;\mathrm{d}x),$$

其中的积分是勒贝格－斯蒂尔切斯积分. 由勒贝格－斯蒂尔切斯积分性质可见（见 §6 第 8 小节），对于每个固定的 $\omega\in\Omega$，$Q(\omega;B)$ 是 B 的测度. 我们利用"适当集合原理"（见 §2 第 1 小节），证明 $Q(\omega;B)$ 是条件概率 $\mathbf{P}(B|\mathscr{G})(\omega)$ 的变式.

设 $\mathscr{E}$ 是 $\mathscr{B}(\mathbf{R})$ 中使 $Q(\omega;B)=\mathbf{P}(\xi\in B|\mathscr{G})(\omega)$ (a.e.) 集合 B 的全体. 由于 $F(\omega;x)$ 几乎必然等于 $\mathbf{P}(\xi\leqslant x|\mathscr{G})(\omega)$，可见形如 $B=(-\infty,x], x\in\mathbf{R}$，的集合 B 属于集系 $\mathscr{E}$. 说明集系 $\mathscr{E}$ 也包括一切形如 $(a,b]$ 的区间，以及由有限个不相交形如 $(a,b]$ 的区间的和组成的代数 $\mathscr{A}$. 那么，由测度 $Q(\omega;B)$（其中 ω 固定）的连续性，以及定理 2 的命题 b)，可见 $\mathscr{E}$ 是单调类. 因为 $\mathscr{A}\subseteq\mathscr{E}\subseteq\mathscr{B}(\mathbf{R})$，所以由 §2 的定理 1，有

$$\mathscr{B}(\mathbf{R})=\sigma(\mathscr{A})\subseteq\sigma(\mathscr{E})=\mu(\mathscr{E})=\mathscr{E}\subseteq\mathscr{B}(\mathbf{R}).$$

于是 $\mathscr{E}=\mathscr{B}(\mathbf{R})$. □

利用不复杂的拓扑分析，可以将定理 4 关于存在正则条件分布的论断推广到在博雷尔空间取值的随机元. 现在引进如下定义.

定义 9 可测空间 $(E,\mathscr{E})$ 称做博雷尔空间，如果它按博雷尔等价于数轴的某一博雷尔子集，即存在一一映射 $\varphi=\varphi(e):(E,\mathscr{E})\to(\mathbf{R},\mathscr{B}(\mathbf{R}))$，使得

1) $\varphi(E)\equiv\{\varphi(e):e\in E\}$ 是 $\mathscr{B}(\mathbf{R})$ 的某一集合；

2) φ 为 $\mathscr{E}$－可测：$[\varphi^{-1}(A)\in\mathscr{E}, A\in\varphi(E)\bigcap\mathscr{B}(\mathbf{R})]$；

3) φ^{-1} 为 $\mathscr{B}(\mathbf{R})/\mathscr{E}$－可测：$\{\varphi(B)\in\varphi(E)\bigcap\mathscr{B}(\mathbf{R}), B\in\mathscr{E}\}$.

定理 5 *设 $X=X(\omega)$ 是在博雷尔空间 $(E,\mathscr{E})$ 取值的随机元. 那么，存在 X 的关于 σ－代数 $\mathscr{G}\subseteq\mathscr{F}$ 的正则条件分布.*

证明 设 $\varphi=\varphi(e)$ 是定义 9 中的函数. 由于定义 9 中的 2)，$\varphi(X(\omega))$ 是随机变量. 故由定理 4，随机变量 $\varphi(X(\omega))$ 关于 $\mathscr{G}$ 的条件分布 $Q(\omega;A), A\in\varphi(E)\bigcap\mathscr{B}(\mathbf{R})$ 有定义.

引进函数 $\tilde{Q}(\omega;B)=Q(\omega;\varphi(B)), B\in\mathscr{E}$. 由定义 9 的 3) $\varphi(B)\in\varphi(E)\bigcap\mathscr{B}(\mathbf{R})$，因而 $\tilde{Q}(\omega;B)$ 有定义. 显然，对于每一个 ω，$\tilde{Q}(\omega;B)$ 关于 $B\ (B\in\mathscr{E})$ 是测度. 现在固定 $B\in\mathscr{E}$. 由于 $\varphi=\varphi(e)$ 是一一映射，有

$$\tilde{Q}(\omega;B)=Q(\omega;\varphi(B))=\mathbf{P}\{\varphi(X)\in\varphi(B)|\mathscr{G}\}(\omega)=\mathbf{P}\{X\in B|\mathscr{G}\}(\omega)\quad(\text{a.e.}).$$

于是，$\tilde{Q}(\omega;B)$ 是 X 的关于 $\mathscr{G}$ 的条件正则分布. □

系 *设 $X=X(\omega)$ 是在完全可分度量空间 $(E,\mathscr{E})$ 取值的随机元. 那么，存在 X 的关于 $\mathscr{G}$ 的正则条件分布. 特别，对于空间 $(\mathbf{R}^n,\mathscr{B}(\mathbf{R}^n))$ 和 $(\mathbf{R}^\infty,\mathscr{B}(\mathbf{R}^\infty))$ 存在这样的分布.*

其证明可以从两方面得到：第一，定理 5；第二，由拓扑中著名结果知，这样的空间是博雷尔空间.

8. 贝叶斯定理的推广 由上面阐述的条件数学期望的理论，可以得到在统计学中用到的贝叶斯定理的推广.

回忆，$\mathscr{D}=\{A_1,\cdots,A_n\}$ $(\mathbf{P}(A_i)>0)$ 是空间 Ω 的分割，则由第一章 §3 (9) 式的贝叶斯定理知，对于任意 B，$\mathbf{P}(B)>0$，有

$$\mathbf{P}(A_i|B)=\frac{\mathbf{P}(A_i)\mathbf{P}(B|A_i)}{\sum\limits_{j=1}^{n}\mathbf{P}(A_j)\mathbf{P}(B|A_j)}.\tag{25}$$

因此，如果 $\theta=\sum_{i=1}^{n}a_iI_{A_i}$ 是离散型随机变量，则根据第一章 §8 的 (10) 式，有

$$\mathbf{E}[g(\theta)|B]=\frac{\sum\limits_{i=1}^{n}g(a_i)\mathbf{P}(A_i)\mathbf{P}(B|A_i)}{\sum\limits_{j=1}^{n}\mathbf{P}(A_j)\mathbf{P}(B|A_j)},\tag{26}$$

或

$$\mathbf{E}[g(\theta)|B]=\frac{\int_{-\infty}^{\infty} g(a)\mathbf{P}(B|\theta=a)P_\theta(\mathrm{d}a)}{\int_{-\infty}^{\infty}\mathbf{P}(B|\theta=a)P_\theta(\mathrm{d}a)}, \tag{27}$$

其中 $P_\theta(A)=\mathbf{P}\{\theta\in A\}$.

由刚引过的 $\mathbf{E}[g(\theta)|B]$ 的定义, 不难证明, (27) 式对于任何事件 B, $\mathbf{P}(B)>0$, 以及随机变量 $\theta=\theta(\omega)$ 和函数 $g=g(a)$ 且 $\mathbf{E}|g(\theta)|<\infty$, 仍然成立.

现在, 对于某 σ-代数 $\mathscr{G}$ $(\mathscr{G}\subseteq\mathscr{F})$ 的条件数学期望 $\mathbf{E}[g(\theta)|\mathscr{G}]$, 考虑 (27) 式的类似.

设

$$\mathbf{Q}(B)=\int_B g(\theta(\omega))\mathbf{P}(\mathrm{d}\omega),\quad B\in\mathscr{G}, \tag{28}$$

则由 (4) 式可见

$$\mathbf{E}[g(\theta)|\mathscr{G}]=\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\mathbf{P}}(\omega). \tag{29}$$

与 σ-代数 $\mathscr{G}$ 同时, 再考虑 σ-代数 $\mathscr{G}_\theta$, 则由 (5) 式可见

$$\mathbf{P}(B)=\int_\Omega \mathbf{P}(B|\mathscr{G}_\theta)\mathrm{d}\mathbf{P}, \tag{30}$$

或根据在勒贝格积分号下的变量替换公式, 有

$$\mathbf{P}(B)=\int_{-\infty}^{\infty}\mathbf{P}(B|\theta=a)P_\theta(\mathrm{d}a). \tag{31}$$

由于

$$\mathbf{Q}(B)=\mathbf{E}[g(\theta)I_B]=\mathbf{E}[g(\theta)\mathbf{E}(I_B|\mathscr{G}_\theta)],$$

可见

$$\mathbf{Q}(B)=\int_{-\infty}^{\infty} g(a)\mathbf{P}(B|\theta=a)P_\theta(\mathrm{d}a). \tag{32}$$

现在假设这些条件概率 $\mathbf{P}(B|\theta=a)$ 都是正则的, 并且可以表示为

$$\mathbf{P}(B|\theta=a)=\int_B\rho(\omega;a)\lambda(\mathrm{d}\omega), \tag{33}$$

其中 $\rho=\rho(\omega;a)$ 是非负 (对两个变量) 可测的函数, 而 λ 是 $(\Omega,\mathscr{G})$ 上的某 σ-有限测度. 设 $\mathbf{E}|g(\theta)|<\infty$. 现在证明, 依概率 1, 有 (广义贝叶斯定理)

$$\mathbf{E}[g(\theta)|\mathscr{G}](\omega)=\frac{\int_{-\infty}^{\infty} g(a)\rho(\omega;a)P_\theta(\mathrm{d}a)}{\int_{-\infty}^{\infty}\rho(\omega;a)P_\theta(\mathrm{d}a)}. \tag{34}$$

为证明 (34) 式, 需要下面的引理.

引理 *设 $(\Omega,\mathscr{F})$ 是某个可测空间.*

a) 设 μ 和 λ $(\mu\ll\lambda)$ 是 σ-有限测度, $f=f(\omega)$ 是 $\mathscr{F}$-可测函数, 则

$$\int_\Omega f\,\mathrm{d}\mu=\int_\Omega f\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\mathrm{d}\lambda \tag{35}$$

(即如果其中一个积分存在, 则另一个也存在, 并且二者相等).

b) 如果 ν 是带符号的测度, 而 μ 和 λ $(\mu\ll\lambda)$ 是 σ-有限测度, 且 $\nu\ll\mu\ll\lambda$, 则

$$\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}=\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\times\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\quad(\lambda-\text{a.e.}) \tag{36}$$

和

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu}=\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\Big/\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\quad(\mu-\text{a.e.}). \tag{37}$$

证明 a) 由于

$$\mu(A)=\int_A\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)\mathrm{d}\lambda,\ A\in\mathscr{F},$$

可见对于任意简单函数 $f=\sum f_i I_{A_i}$, (35) 式自然成立. 一般情形, 由表示 $f=f^+-f^-$ 和单调收敛定理可以得到 (35) 式 (对照 §6 (39) 式的证明).

b) 由命题 a), 其中设 $f=\mathrm{d}\nu/\mathrm{d}\mu$, 可得

$$\nu(A)=\int_A\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)\mathrm{d}\mu=\int_A\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)\mathrm{d}\lambda.$$

因此 $\nu\ll\lambda$, 故

$$\nu(A)=\int_A\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\mathrm{d}\lambda.$$

由于集合 A 的任意性和 §6 的性质 I, 可得 (36) 式.

由 (36) 式, 以及如下等式可得 (37) 式:

$$\mu\left\{\omega:\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}=0\right\}=\int_{\{\omega:\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}=0\}}\left(\frac{\mathrm{d}\mu}{\mathrm{d}\lambda}\right)\mathrm{d}\lambda=0$$

(在集合 $\{\omega:\mathrm{d}\mu/\mathrm{d}\lambda=0\}$ 上 (37) 式的右侧可以任意决定, 例如, 令其等于 0). □

为证明 (34) 式注意到, 由傅比尼定理和假设 (33), 有

$$\mathbf{Q}(B)=\int_B\left[\int_{-\infty}^{\infty}g(a)\rho(\omega;a)P_\theta(\mathrm{d}a)\right]\lambda(\mathrm{d}\omega), \tag{38}$$

$$\mathbf{P}(B)=\int_B\left[\int_{-\infty}^{\infty}\rho(\omega;a)P_\theta(\mathrm{d}a)\right]\lambda(\mathrm{d}\omega). \tag{39}$$

那么, 由引理, 有

$$\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\mathbf{P}}=\frac{\mathrm{d}\mathbf{Q}/\mathrm{d}\lambda}{\mathrm{d}\mathbf{P}/\mathrm{d}\lambda}\quad(\mathbf{P}-\text{a.e.}).$$

由此并考虑到 (38), (39) 和 (29) 式, 得 (34) 式.

注 如果代替随机变量 θ, 考虑在某一可测空间 $(E,\mathscr{E})$ 取值的随机元, 则 (34) 式仍然成立 (只要把在 $\mathbf{R}$ 上积分换成在 E 上积分).

下面讨论 (34) 式的某些特殊情形.

假设 σ- 代数 $\mathscr{G}$ 由随机变量 ξ 生成: $\mathscr{G}=\sigma(\xi)$.

假设

$$\mathbf{P}(\xi\in A\mid\theta=a)=\int_A q(x;a)\lambda(dx),\quad A\in\mathscr{B}(\mathbf{R}), \tag{40}$$

其中 $q=q(x;a)$ 是关于两个变量非负可测的函数, λ 是 $(\mathbf{R},\mathscr{B}(\mathbf{R}))$ 上的某一 σ- 有限测度. 那么, 由勒贝格积分在符号下的变量替换和 (34) 式, 可得

$$\mathbf{E}[g(\theta)\mid\xi=x]=\frac{\int_{-\infty}^{\infty}g(a)q(x;a)P_\theta(da)}{\int_{-\infty}^{\infty}q(x;a)P_\theta(da)}. \tag{41}$$

特别, 设 (θ,ξ), $\theta=\sum a_iI_{A_i}$, $\xi=\sum x_jI_{B_j}$ 是两个离散型随机变量. 那么, 作为 λ 选计数测度 $(\lambda(\{x_i\})=1, i=1,2,\cdots)$, 由 (40) 式, 得 (对照 (26) 式)

$$\mathbf{E}[g(\theta)\mid\xi=x_j]=\frac{\sum_i g(a_i)\mathbf{P}\{\xi=x_j\mid\theta=a_i\}\mathbf{P}\{\theta=a_i\}}{\sum_i \mathbf{P}\{\xi=x_j\mid\theta=a_i\}\mathbf{P}\{\theta=a_i\}}. \tag{42}$$

现在设 (θ,ξ) 是密度为 $f_{\theta,\xi}(a,x)$ 的两个绝对连续随机变量, 则由于 (19) 式, 可见 (40) 式成立, 其中 $q(x;a)=f_{\xi|\theta}(x|a)$, 而 λ 是勒贝格测度. 因此

$$\mathbf{E}[g(\theta)\mid\xi=x]=\frac{\int_{-\infty}^{\infty}g(a)f_{\xi|\theta}(x|a)f_\theta(a)da}{\int_{-\infty}^{\infty}f_{\xi|\theta}(x|a)f_\theta(a)da}. \tag{43}$$

9. 条件数学期望的换算公式 我们再引进广义贝叶斯定理 (见 (34) 式) 的一种形式, 其 (下面的) 提法在与概率测度变换有关的问题中特别适宜.

定理 6 *假设 $\mathbf{P}$ 和 $\tilde{\mathbf{P}}$ 是可测空间 $(\Omega,\mathscr{F})$ 上的两个概率测度, 且测度 $\tilde{\mathbf{P}}$ 关于测度 $\mathbf{P}$ 绝对连续 (记作 $\tilde{\mathbf{P}}\ll\mathbf{P}$), 而 $d\tilde{\mathbf{P}}/d\mathbf{P}$ 是测度 $\tilde{\mathbf{P}}$ 关于测度 $\mathbf{P}$ 的拉东–尼科迪姆导数. 设 $\mathscr{G}$ 是 $\mathscr{F}(\mathscr{G}\subseteq\mathscr{F})$ 的子 σ- 代数; $\mathbf{E}(\cdot\mid\mathscr{G})$ 和 $\tilde{\mathbf{E}}(\cdot\mid\mathscr{G})$ 相应为按测度 $\mathbf{P}$ 和测度 $\tilde{\mathbf{P}}$ 关于 $\mathscr{G}$ 的条件数学期望; ξ 是非负 ($\mathscr{F}$- 可测的) 随机变量. 则下列"条件数学期望的换算公式"成立:*

$$\tilde{\mathbf{E}}(\xi\mid\mathscr{G})=\frac{\mathbf{E}\left(\xi\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)}{\mathbf{E}\left(\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)}\quad(\tilde{\mathbf{P}}\text{-a.c.}). \tag{44}$$

对于使条件数学期望 $\tilde{\mathbf{E}}(\xi\mid\mathscr{G})$ 存在的一切随机变量 ξ, 公式 (44) 成立.

证明 首先注意到, 事件

$$\left\{\omega:\mathbf{E}\left(\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)=0\right\}$$

的 $\tilde{\mathbf{P}}$ 测度 (以及 $\mathbf{P}$ 测度) 等于 0. 事实上, 如果 $A\in\mathscr{G}$, 则

$$\int_A\mathbf{E}\left(\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)d\mathbf{P}=\int_A\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}d\mathbf{P}=\int_A d\tilde{\mathbf{P}}=\tilde{\mathbf{P}}(A).$$

因而

$$\tilde{\mathbf{P}}(A)=\tilde{\mathbf{P}}\left\{\omega:\mathbf{E}\left(\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)=0\right\}=0.$$

设 $\xi\geqslant 0$. 根据条件数学期望的定义, $\tilde{\mathbf{E}}(\xi\mid\mathscr{G})$ 是 $\mathscr{G}$- 可测随机变量, 且对于任意 $A\in\mathscr{G}$, 满足

$$\tilde{\mathbf{E}}[I_A\tilde{\mathbf{E}}(\xi\mid\mathscr{G})]=\tilde{\mathbf{E}}[I_A\xi]. \tag{45}$$

因此, 为证明 (44) 式, 只需证明 (44) 式右侧的 ($\mathscr{G}$- 可测) 随机变量满足下面的不等式:

$$\tilde{\mathbf{E}}\left[I_A\times\frac{\mathbf{E}\left(\xi\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)}{\mathbf{E}\left(\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)}\right]=\tilde{\mathbf{E}}[I_A\xi]. \tag{46}$$

利用条件数学期望的性质, 以及 §6 的 (39) 式, 可得

$$\tilde{\mathbf{E}}\left[I_A\times\frac{\mathbf{E}\left(\xi\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)}{\mathbf{E}\left(\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)}\right]=\mathbf{E}\left[I_A\times\frac{\mathbf{E}\left(\xi\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)}{\mathbf{E}\left(\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)}\times\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\right]$$

$$=\mathbf{E}\left[I_A\times\frac{\mathbf{E}\left(\xi\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)}{\mathbf{E}\left(\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)}\times\mathbf{E}\left(\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)\right]=\mathbf{E}\left[I_A\times\mathbf{E}\left(\xi\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\middle|\mathscr{G}\right)\right]$$

$$=\mathbf{E}\left[I_A\xi\frac{d\tilde{\mathbf{P}}}{d\mathbf{P}}\right]=\tilde{\mathbf{E}}[I_A\xi].$$

从而, 对于非负随机变量 ξ, 证明了 (45) 式. 对于任意可积随机变量 ξ, 一般情形的证明, 完全类似于 §6 的 (39) 式的证明. □

10. 充分统计量和因子分解定理　上面引进的广义贝叶斯定理 ((41),(42) 和 (43) 式), 是数理统计中 "贝叶斯方法" 的基本工具之一. 它回答如下问题: 根据对与随机参数 θ 相联系的随机变量 ξ 的观测结果, 如何重新分配关于随机变量 θ 的知识.

下面将讨论条件数学期望概念在由观测结果估计未知参数问题中的另外一种应用. (需要强调: 与上面讨论的 θ 是随机变量的情形不同, 现在 θ 是事先给定的参数集 Θ 中的某一普通参数 (对照第一章 §7)).

实质上, 这里说的是关于数理统计中的一个重要概念 —— 充分子 σ - 代数 (或 σ - 子代数) 的概念.

设 $(\Omega, \mathscr{F})$ 是某一可测空间, $\mathscr{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ 是依赖于参数的概率测度 P_θ 族, 其中 θ 属于参数的集合 Θ. 简称组合 $\mathscr{E} = (\Omega, \mathscr{F}, \mathscr{P})$ 为概率 - 统计模型或概率 - 统计试验 (对照第一章 §7).

为说明下面引进的定义 10, 假设有结局 ω 的某一 $\mathscr{F}$ - 可测函数 (统计量), 和由该函数 $T = T(\omega)$ 诱导的 σ - 代数 $\mathscr{G} = \sigma(T(\omega))$. 显然, $\mathscr{G} \subseteq \mathscr{F}$, 且一般在 $\mathscr{F}$ 中可能有不属于 $\mathscr{G}$ 的元素 (即 $\mathscr{F}$ 比 $\mathscr{G}$ "丰富"). 但是, 完全可能出现这样的情形: 从定义的角度对什么样的参数 θ "作出" 使得除了关于 $T = T(\omega)$ 的知识之外, 任何关于参数 θ 的其他信息, 没有也不需要有. 在这种意义上, 统计量 T 自然应称为充分的.

注 1　在 $T(\omega) = \omega$ 的情形下, 即已知的是结局本身 (而不是其函数), 可以指出如下两种极端的情形.

一种是对于一切 $\theta \in \Theta$, 所有概率 P_θ 都相等. 显然, 在这种情形下, 任何结局 ω 都不能提供关于参数 θ 的值的任何信息.

另一种情形是, 所有概率 P_θ 的承载子, "处于" $\mathscr{F}$ 的不同子集中 (即, 对于任何两个参数 θ_1 和 θ_2, 测度 P_{θ_1} 和 P_{θ_2} 是奇异的: 存在两个集合 (承载子) Ω_{θ_1} 和 Ω_{θ_2}, 使

$$\mathrm{P}_{\theta_1}(\Omega \setminus \Omega_{\theta_1}) = 0, \quad \mathrm{P}_{\theta_2}(\Omega \setminus \Omega_{\theta_2}) = 0, \quad \Omega_{\theta_1} \cap \Omega_{\theta_2} = \varnothing$$

成立). 在这种情形下, 参数 θ 的值由所得结局 ω "唯一" 确定.

对于这两种极端情形没什么兴趣. 人们感兴趣的是, 所有测度 P_θ 都等价的情形 (而这时其承载子不可区分).

现在, 我们把 σ - 代数 $\mathscr{G} = \sigma(T(\omega))$, 以及更一般地把任何子 σ - 代数 $\mathscr{G} \subseteq \mathscr{F}$, 视为在试验结局为 $\omega \in \Omega$ 时, 所获得的 "信息".

有了这样的信息 $\mathscr{G}$, (像贝叶斯定理的情形一样) 自然关心概率 P_θ 如何变化, 即相应的条件概率 $\mathrm{P}_\theta(\cdot|\mathscr{G})(\omega)$ 如何. 这时 (与注 1 的第一种情形类似) 明显, 假如这些概率对于所有 ω 不依赖于 θ, 与由 "信息" $\mathscr{G}$ 可以得到的相比, 则由 "更多的信息" $\widetilde{\mathscr{G}} \supseteq \mathscr{G}$ 得不到关于参数 θ 的任何更新的内容.

这时, "信息" $\mathscr{G}$ 可以称为详尽无遗的, 或者习惯上, 称做充分的.

下面的定义表达了上面的论断.

定义 10　设 $\mathscr{E} = (\Omega, \mathscr{F}, \mathscr{P})$ 为概率 - 统计模型, $\mathscr{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ 和 $\mathscr{G}$ 是 $\mathscr{F}$ ($\mathscr{G} \subseteq \mathscr{F}$) 的子 σ - 代数. 称子 σ - 代数 $\mathscr{G}$ 对于 $\mathscr{P}$ 族是充分的, 如果存在不依赖于条件概率 $\mathrm{P}_\theta(\cdot|\mathscr{G})(\omega)$ ($\theta \in \Theta, \omega \in \Omega$) 的变式, 即存在函数 $P(A;\omega)$, $A \in \mathscr{F}$, $\omega \in \Omega$, 使对于一切 $A \in \mathscr{F}$ 和 $\theta \in \Theta$, 有

$$\mathrm{P}_\theta(A|\mathscr{G})(\omega) = P(A;\omega), \ (\mathrm{P}_\theta\text{ - a.c.}), \tag{47}$$

即对于一切 $\theta \in \Theta$, $P(A;\omega)$ 是 $\mathrm{P}_\theta(A|\mathscr{G})(\omega)$ 的变式.

如果 $\mathscr{G} = \sigma(T(\omega))$, 则统计量 $T = T(\omega)$ 称做 $\mathscr{P}$ 族的充分统计量.

注 2　在统计研究中, 求充分统计量的重要性, 在于力图寻找结局 ω 的这样函数 $T = T(\omega)$, 使之在保留 (关于参数 θ 值的) 信息的情况下压缩数据. 例如, 设想对于很大的 n, 有 $\omega = (x_1, x_2, \cdots, x_n)$, $x_i \in \mathbb{R}$. 那么, 由于现有数据 $x_1, x_2, \cdots, x_n$ 的维数 n 很大, 使建立参数 θ 的 "好" 估计量 (例如, 第一章 §7) 可能成为很困难的问题. 不过, 完全可能出现这样的情形 (这在第一章 §7 我们已经见到), 为建立参数的 "好" 估计量, 完全不需要知道原始数据 $x_1, x_2, \cdots, x_n$, 只需构造统计量的数据的和: $T(\omega) = x_1 + x_2 + \cdots + x_n$.

显然, 这样的统计量确实可以本质地压缩了数据 (与计算量). 同时对于建立参数 θ 的 "好" 估计量, 也是充分的.

下面的因子分解定理提供了, 保障 σ - 代数 $\mathscr{G}$ 对于概率测度 P_θ 族 $\mathscr{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ 的充分性的条件.

定理 7　设 $\mathscr{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ 是被控制测度族, 即存在 $(\Omega, \mathscr{F})$ 上的 σ - 有限测度 λ, 使对于 $\theta \in \Theta$, 测度 P_θ 关于 λ 绝对连续 ($\mathrm{P}_\theta \ll \lambda$). 设

$$g_\theta^{(\lambda)}(\omega) = \frac{\mathrm{d}\mathrm{P}_\theta}{\mathrm{d}\lambda}(\omega)$$

是测度 P_θ 关于测度 λ 的拉东 - 尼科迪姆导数.

子 σ - 代数 $\mathscr{G}$ 对于概率测度族 $\mathscr{P}$ 是充分的, 当且仅当函数 $g_\theta^{(\lambda)}(\omega)$ 有因子分解: 存在非负函数 $\hat{g}_\theta^{(\lambda)}(\omega)$ 和 $h(\omega)$, 其中 $\hat{g}_\theta^{(\lambda)}(\omega)$ 是 $\mathscr{G}$ - 可测函数, $h(\omega)$ 是 $\mathscr{F}$ - 可测函数, 且对于一切 $\theta \in \Theta$, 有

$$g_\theta^{(\lambda)}(\omega) = \hat{g}_\theta^{(\lambda)}(\omega)h(\omega) \quad (\lambda\text{ - a.e.}). \tag{48}$$

假如把测度 λ 取为测度 P_{θ_0}, 其中 θ_0 是 Θ 的某个参数, 则对于 σ - 代数 $\mathscr{G}$ 的充分性, 充分和必要条件是: 导数

$$g_\theta^{(\theta_0)}(\omega) = \frac{\mathrm{d}\mathrm{P}_\theta}{\mathrm{d}\mathrm{P}_{\theta_0}}(\omega)$$

本身为 $\mathscr{G}$ - 可测的.

证明 1) 充分性. 根据假设作为控制测度 λ 是 σ– 有限测度. 表示存在 $\mathscr{F}$– 可测不相交集合 $\Omega_k, k \geqslant 1$, 使 $\Omega = \sum_{k\geqslant 1} \Omega_k$ 且 $0 < \lambda(\Omega_k) < \infty, k \geqslant 1$.

建立测度

$$\tilde{\lambda}(\cdot) = \sum_{k\geqslant 1} \frac{1}{2^k} \frac{\lambda(\Omega_k \cap \cdot)}{1 + \lambda(\Omega_k)}.$$

该测度有限: $\tilde{\lambda}(\Omega) < \infty$ 且 $\tilde{\lambda}(\Omega) > 0$. 不失普遍性, 可以认为 λ 是概率测度: $\tilde{\lambda}(\Omega) = 1$.

那么, 由条件数学期望的换算公式 (44), 对于任意 $\mathscr{F}$– 可测有限随机变量 $X = X(\omega)$, 有

$$\mathbf{E}_\theta(X|\mathscr{G}) = \frac{\mathbf{E}_{\tilde{\lambda}}\left(X \dfrac{d\mathbf{P}_\theta}{d\tilde{\lambda}} \,\Big|\, \mathscr{G}\right)}{\mathbf{E}_{\tilde{\lambda}}\left(\dfrac{d\mathbf{P}_\theta}{d\tilde{\lambda}} \,\Big|\, \mathscr{G}\right)} \quad (\mathbf{P}_\theta\text{ - a.c.}). \tag{49}$$

根据 (48) 式, 有

$$g_\theta^{(\tilde{\lambda})} = \frac{d\mathbf{P}_\theta}{d\tilde{\lambda}} = \frac{d\mathbf{P}_\theta}{d\lambda} \times \frac{d\lambda}{d\tilde{\lambda}} = g_\theta^{(\lambda)} \frac{d\lambda}{d\tilde{\lambda}} = \tilde{g}_\theta^{(\lambda)} h \frac{d\lambda}{d\tilde{\lambda}}. \tag{50}$$

因此 (49) 式有如下形式:

$$\mathbf{E}_\theta(X|\mathscr{G}) = \frac{\mathbf{E}_{\tilde{\lambda}}\left(X \tilde{g}_\theta^{(\lambda)} h \dfrac{d\lambda}{d\tilde{\lambda}} \,\Big|\, \mathscr{G}\right)}{\mathbf{E}_{\tilde{\lambda}}\left(\tilde{g}_\theta^{(\lambda)} h \dfrac{d\lambda}{d\tilde{\lambda}} \,\Big|\, \mathscr{G}\right)} \quad (\mathbf{P}_\theta\text{ - a.c.}). \tag{51}$$

但是 $\tilde{g}_\theta^{(\lambda)}$ 为 $\mathscr{G}$– 可测, 且

$$\mathbf{E}_\theta(X|\mathscr{G}) = \frac{\mathbf{E}_{\tilde{\lambda}}\left(X h \dfrac{d\lambda}{d\tilde{\lambda}} \,\Big|\, \mathscr{G}\right)}{\mathbf{E}_{\tilde{\lambda}}\left(h \dfrac{d\lambda}{d\tilde{\lambda}} \,\Big|\, \mathscr{G}\right)} \quad (\mathbf{P}_\theta - \text{a.c.}). \tag{52}$$

这里的右侧与 θ 无关, 从而性质 (47) 成立. 于是, 根据定义 10, σ– 代数 $\mathscr{G}$ 是充分的.

2) 必要性. 为证明必要性需要有补充假设: 测度族 $\mathscr{P} = \{\mathbf{P}_\theta, \theta \in \Theta\}$ 不但被某 σ– 有限测度 λ 控制, 而且存在某个 $\theta_0 \in \Theta$, 使一切测度 $\mathbf{P}_\theta \ll \mathbf{P}_{\theta_0}$, 即对于任意 $\theta \in \Theta$, 测度 $\mathbf{P}_\theta$ 关于测度 $\mathbf{P}_{\theta_0}$ 绝对连续. (一般情形的证明变得比较复杂, 参见 [106] 的定理 34.6.)

这样, $\mathscr{G}$ 是充分 σ– 代数, 即满足性质 (47). 假设 $\mathbf{P}_\theta \ll \mathbf{P}_{\theta_0}, \theta \in \Theta$, 证明, 对于每一个 $\theta \in \Theta$, $g_\theta^{(\theta_0)} = d\mathbf{P}_\theta/d\mathbf{P}_{\theta_0}$ 是 $\mathscr{G}$– 可测函数.

设 $A \in \mathscr{F}$. 那么, 对于每一个 $\theta \in \Theta$, 利用条件数学期望的性质, 可得 (其中 $g_\theta^{(\theta_0)} = d\mathbf{P}_\theta/d\mathbf{P}_{\theta_0}$):

$$\begin{aligned}\mathbf{P}_\theta(A) &= \mathbf{E}_\theta I_A = \mathbf{E}_\theta\mathbf{E}_\theta(I_A|\mathscr{G}) = \mathbf{E}_\theta\mathbf{E}_{\theta_0}(I_A|\mathscr{G}) = \mathbf{E}_{\theta_0}[g_\theta^{(\theta_0)}\mathbf{E}_{\theta_0}(I_A|\mathscr{G})] \\ &= \mathbf{E}_{\theta_0}\mathbf{E}_{\theta_0}[g_\theta^{(\theta_0)}\mathbf{E}_{\theta_0}(I_A|\mathscr{G})|\mathscr{G}] = \mathbf{E}_{\theta_0}\left([\mathbf{E}_{\theta_0}(g_\theta^{(\theta_0)}|\mathscr{G})] \times [\mathbf{E}_{\theta_0}(I_A|\mathscr{G})]\right) \\ &= \mathbf{E}_{\theta_0}\mathbf{E}_{\theta_0}[I_A\mathbf{E}_{\theta_0}(g_\theta^{(\theta_0)}|\mathscr{G})|\mathscr{G}] = \mathbf{E}_{\theta_0} I_A\mathbf{E}_{\theta_0}(g_\theta^{(\theta_0)}|\mathscr{G}) = \int_A \mathbf{E}_{\theta_0}(g_\theta^{(\theta_0)}|\mathscr{G})d\mathbf{P}_{\theta_0}.\end{aligned}$$

从而, 导数的变式 $g_\theta^{(\theta_0)} = d\mathbf{P}_\theta/d\mathbf{P}_{\theta_0}$ 是 $\mathscr{G}$– 可测函数 $\mathbf{E}_{\theta_0}(g_\theta^{(\theta_0)}|\mathscr{G})$.

这样, 当 $\lambda = \mathbf{P}_{\theta_0}$ 时, 由 σ– 代数 $\mathscr{G}$ 的充分性, 可得因子分解的性质 (48), 其中 $\tilde{g}_\theta^{(\theta_0)}(\omega) = g_\theta^{(\theta_0)}, h \equiv 1$.

对于一般情形 (仍然需要补充假设 $\mathbf{P}_\theta \ll \mathbf{P}_{\theta_0}, \theta \in \Theta$), 有

$$g_\theta^{(\lambda)} = \frac{d\mathbf{P}_\theta}{d\lambda} = \frac{d\mathbf{P}_\theta}{d\mathbf{P}_{\theta_0}} \times \frac{d\mathbf{P}_{\theta_0}}{d\lambda} = g_\theta^{(\theta_0)}\frac{d\mathbf{P}_{\theta_0}}{d\lambda}.$$

记

$$\tilde{g}_\theta^{(\theta_0)} = g_\theta^{(\theta_0)}, \quad h = \frac{d\mathbf{P}_{\theta_0}}{d\lambda},$$

得所要证明的因子分解表达式 (48). □

注 3 有必要强调, 对于任意测度族 $\mathscr{P} = \{\mathbf{P}_\theta, \theta \in \Theta\}$, (在没有任何控制型假设的情况下) 充分 σ– 代数显然存在. 作为这样的 σ– 代数, 可以取最 "丰富" 的 σ– 代数 $\mathscr{F}$.

事实上, 这时对于每一个可积随机变量 X, $\mathbf{E}_\theta(X|\mathscr{F}) = X\ (\mathbf{P}_\theta - \text{a.c.})$, 从而性质 (47) 成立.

显然, 这样的充分 σ– 代数没有什么意义, 因为这时不会有任何 "数据的压缩". 寻找最小 σ– 代数 $\mathscr{G}_{\min}$, 即一切充分 σ– 代数的交才有真正的意义 (参见 §2 的引理 1 的证明, 可见最小 σ– 代数存在). 然而明显地建立这样的 σ– 代数并不容易 (不过可以参见 [107] 第二章 §14 ~ §15).

注 4 假设 $\mathscr{P} = \{\mathbf{P}_\theta, \theta \in \Theta\}$ 是被控制族 ($\mathbf{P}_\theta \ll \lambda, \theta \in \Theta, \lambda$ 是 σ– 有限测度), 而对于一切 $\theta \in \Theta$, 密度 $g_\theta^{(\lambda)} = (d\mathbf{P}_\theta/d\lambda)(\omega)$ 可以表示为

$$g_\theta^{(\lambda)}(\omega) = G_\theta^{(\lambda)}(T(\omega))h(\omega), \quad (\lambda\text{ - a.c.}), \tag{53}$$

其中 $T = T(\omega)$ 是取值于集合 E 的 (及在 E 上分出的 σ– 代数 $\mathscr{E}$) 某 $\mathscr{F}/\mathscr{E}$– 可测函数 (随机元, 见 §5). 设函数 $G_\theta^{(\lambda)}(t), t \in E$, 和 $h(\omega), \omega \in \Omega$, 是非负的且相应为 $\mathscr{E}$– 和 $\mathscr{F}$– 可测的.

比较 (48) 式和 (53) 式, 可见 σ– 代数 $\mathscr{G} = \sigma(T(\omega))$ 是充分的, 而函数 $T = T(\omega)$ (在定义 10 的意义下) 是充分统计量.

注意, 在被控制的情形下, 通常正式以满足因子分解式 (53) 的函数 $T = T(\omega)$ 当作充分统计量的定义.

例 5 (指数族)　假设 $\Omega = \mathbb{R}^n, \mathscr{F} = \mathscr{B}(\mathbb{R}^n)$, 测度 $\mathbf{P}_\theta$ 是按如下方式构造的: 对于 $\omega = (x_1, x_2, \cdots, x_n), x_i \in \mathbf{R}$, 则

$$\mathbf{P}_\theta(\mathrm{d}\omega) = \mathbf{P}_\theta(\mathrm{d}x_1)\cdots\mathbf{P}_\theta(\mathrm{d}x_n), \tag{54}$$

其中测度 $\mathbf{P}_\theta(\mathrm{d}x), x \in \mathbf{R}$, 有如下构造:

$$\mathbf{P}_\theta(\mathrm{d}x) = \alpha(\theta)e^{\beta(\theta)s(x)}\gamma(x)\lambda(\mathrm{d}x). \tag{55}$$

这里, $x \to s(x)$ 是某一 $\mathscr{B}$ 可测函数, 而 $\alpha(\theta), \beta(\theta), \gamma(x), \lambda(\mathrm{d}x)$ 含义明显. (测度族 $\mathbf{P}_\theta, \theta \in \Theta$, 是称为指数族的最简单例子.) 由 (54) 式和 (55) 式, 可见

$$\mathbf{P}_\theta(\mathrm{d}\omega) = \alpha^n(\theta)e^{\beta(\theta)[s(x_1)+\cdots+s(x_n)]}\gamma(x_1)\cdots\gamma(x_n)\mathrm{d}x_1\cdots\mathrm{d}x_n. \tag{56}$$

由 (56) 式与 (53) 式比较, 可见统计量 $T(\omega) = s(x_1)+\cdots+s(x_n), \omega = (x_1, \cdots, x_n)$, (对于所考虑的指数族) 是充分统计量.

如果对于 $\omega = (x_1, \cdots, x_n)$, 记 $X_1(\omega) = x_1, \cdots, X_n(\omega) = x_n$, 则 (按测度的直积原则组成的) 测度 $\mathbf{P}_\theta$ 的构造, 使 $X_1, \cdots, X_n$ 关于这些测度是独立同分布随机变量序列.

于是, 根据序列 $X_1(\omega), \cdots, X_n(\omega)$ 建立的统计量 $T(\omega) = s(X_1(\omega)) + \cdots + s(X_n(\omega))$, 是充分统计量. (在练习题 20 中要求回答, 这一统计量是否最小充分统计量).

例 6　设 $\Omega = \mathbb{R}^n, \mathscr{F} = \mathscr{B}(\mathbb{R}^n)$, 参数 $\theta > 0$ 且对于 $\omega = (x_1, \cdots, x_n)$, (关于勒贝格测度 λ 的) 密度为

$$\frac{\mathrm{d}\mathbf{P}_\theta}{\mathrm{d}\lambda}(\omega) = \begin{cases} \theta^{-n}, & \text{若 } 0 \leqslant x_i \leqslant \theta, i = 1, \cdots, n, \\ 0, & \text{若不然}. \end{cases}$$

如果

$$T(\omega) = \max_{1\leqslant i\leqslant n} x_i,$$

$$h(\omega) = \begin{cases} 1, & \text{若 } x_i \geqslant 0, i = 1, \cdots, n, \\ 0, & \text{若不然}, \end{cases}$$

$$G_\theta^{(\lambda)}(t) = \begin{cases} \theta^{-n}, & \text{若 } 0 \leqslant t \leqslant \theta, \\ 0, & \text{若不然}, \end{cases}$$

则

$$\frac{\mathrm{d}\mathbf{P}_\theta}{\mathrm{d}\lambda}(\omega) = G_\theta^{(\lambda)}(T(\omega))h(\omega). \tag{57}$$

于是, $T(\omega) = \max\limits_{1\leqslant i\leqslant n} x_i$ 是充分统计量得证.

11. 无偏估计量, 拉奥–布莱克韦尔 (C. R. Rao-D. H. Blackwell) 定理　假设 Θ 是直线上的一子集, $\mathscr{E} = (\Omega, \mathscr{F}, \mathscr{P} = \{\mathbf{P}_\theta : \theta \in \Theta\})$ 是概率统计模型. 现在考虑的问题是: 建立参数 θ 的 "好" 估计量.

任何随机变量 $\tilde{\theta} = \tilde{\theta}(\omega)$ 都将视为参数 θ 的估计量 (对照第一章 §7).

下面的定理表明, 充分 σ 代数的概念可以改进估计量的 "质量". 估计量 $\tilde{\theta}$ 的质量, 是用其对参数 θ 的真值的均方偏差来衡量的. 更确切地说, 称参数 θ 的估计量 $\tilde{\theta}$ 是无偏的, 如果 $\mathbf{E}_\theta|\tilde{\theta}| < \infty$ 且 $\mathbf{E}_\theta\tilde{\theta} = \theta$ (对照第一章 §7 性质 2).

定理 8 (拉奥和布莱克韦尔)　*设 $\mathscr{G}$ 对于概率测度族 $\mathscr{P}$ 是充分 σ 代数, 而 $\tilde{\theta} = \tilde{\theta}(\omega)$ 是参数 θ 的某一估计量.*

a) 如果 $\tilde{\theta}$ 是无偏估计量, 则估计量

$$T = \mathbf{E}_\theta(\tilde{\theta}|\mathscr{G}) \tag{58}$$

也是无偏的.

b) 估计量 T 在如下意义上优于估计量 $\tilde{\theta}$:

$$\mathbf{E}_\theta(T-\theta)^2 \leqslant \mathbf{E}_\theta(\tilde{\theta}-\theta)^2, \quad \theta \in \Theta. \tag{59}$$

证明　性质 a) 由下列等式可见:

$$\mathbf{E}_\theta T = \mathbf{E}_\theta\mathbf{E}_\theta(\tilde{\theta}|\mathscr{G}) = \mathbf{E}_\theta\tilde{\theta} = \theta.$$

为证明性质 b) 只需注意到: 由延森不等式 (只需在练习题 5 中设 $g(x) = (x-\theta)^2$), 有

$$[\mathbf{E}_\theta(\tilde{\theta}|\mathscr{G}) - \theta]^2 \leqslant \mathbf{E}_\theta[(\tilde{\theta}-\theta)^2|\mathscr{G}].$$

在不等式两侧同时求数学期望 $\mathbf{E}_\theta(\cdot)$, 得要证明的 (59) 式. □

12. 练习题

1. 设 ξ 和 η 是独立同分布随机变量, $\mathbf{E}\xi$ 有定义. 证明

$$\mathbf{E}(\xi|\xi+\eta) = \mathbf{E}(\eta|\xi+\eta) = \frac{\xi+\eta}{2} \text{ (a.c.)}.$$

2. 设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量, 且 $\mathbf{E}|\xi_1| < \infty$. 证明

$$\mathbf{E}(\xi_1|S_n, S_{n+1}, \cdots) = \frac{S_n}{n} \text{ (a.c.)},$$

其中 $S_n = \xi_1 + \cdots + \xi_n$.

3. 假设对于随机元 (X, Y), 存在正则分布 $P_x(B) = \mathbf{P}(Y \in B|X = x)$. 证明, 如果 $\mathbf{E}|g(X,Y)| < \infty$, 则 P_x–a.c., 有

$$\mathbf{E}[g(X,Y)|X = x] = \int g(x,y)P_x(\mathrm{d}y).$$

4. 设随机变量 ξ 的分布函数为 $F_\xi(x)$, 证明

$$\mathbf{E}(\xi|a<\xi\leqslant b)=\frac{\int_a^b x\mathrm{d}F_\xi(x)}{F_\xi(b)-F_\xi(a)}$$

(假设 $F_\xi(b)-F_\xi(a)>0$).

5. 设 $g=g(x)$ 是凹 (向上凸) 博雷尔函数, $\mathbf{E}|g(\xi)|<\infty$. 证明, 对于条件数学期望, (a.c.) 有琴森不等式:

$$g(\mathbf{E}(\xi|\mathscr{G}))\leqslant\mathbf{E}[g(\xi)|\mathscr{G}].$$

6. 证明随机变量 ξ 和 σ – 代数 $\mathscr{G}$ 独立 (即对于任意 $B\in\mathscr{G}$, 随机变量 ξ 与 I_B 独立), 当且仅当对于每一博雷尔函数 $g(x)$, $\mathbf{E}|g(\xi)|<\infty$, 有 $\mathbf{E}[g(\xi)|\mathscr{G}]=\mathbf{E}g(\xi)$.

7. 设 ξ 是非负随机变量, $\mathscr{G}$ 是 σ – 代数 $\mathscr{G}\subseteq\mathscr{F}$, $\mathbf{Q}$ 是由等式 $\mathbf{Q}(A)=\int_A\xi\mathrm{d}\mathbf{P}$ 定义在集合 $A\in\mathscr{G}$ 上的测度. 证明, $\mathbf{E}(\xi|\mathscr{G})<\infty$ (a.c.) 的充分和必要条件为 $\mathbf{Q}$ 是有限测度.

8. 证明条件概率 $\mathbf{P}(A|B)$ 在如下意义上 "连续": 如果 $\lim_n A_n=A$, $\lim_n B_n=B$ 且 $\mathbf{P}(B_n)>0$, $\mathbf{P}(B)>0$, 则 $\lim_n\mathbf{P}(A_n|B_n)=\mathbf{P}(A|B)$.

9. 设 $\Omega=(0,1)$, $\mathscr{F}=\mathscr{B}((0,1))$, $\mathbf{P}$ 是勒贝格测度; $X(\omega)$ 和 $Y(\omega)$ 是在 $(0,1)$ 上均匀分布的两个独立随机变量. 考虑第三个随机变量 $Z(\omega)=|X(\omega)-Y(\omega)|$ 是 "点" $X(\omega)$ 和 $Y(\omega)$ 之间的距离. 证明 $Z(\omega)$ 的分布函数 $F_Z(z)$ 有密度:

$$f_Z(z)=\begin{cases}2(1-z), & 若\ 0\leqslant z\leqslant 1,\\ 0, & 若不然.\end{cases}$$

(由此可见 $\mathbf{E}Z=1/3$.)

10. 向半径为 R 的圆 $\{(x,y):x^2+y^2\leqslant R^2\}$ 上 "随机地" 掷两点 A_1 和 A_2, 即独立地选两点且选到点 $A_i(\rho_i,\theta_i)$, $i=1,2$, 的概率 (在极坐标系中) 为:

$$\mathbf{P}\{\rho_i\in\mathrm{d}r,\ \theta_i\in\mathrm{d}\theta\}=\frac{r\mathrm{d}r\mathrm{d}\theta}{\pi R^2},\quad i=1,2.$$

证明点 A_1 和 A_2 之间的距离 ρ 的分布密度为:

$$f_\rho(r)=\frac{2r}{\pi R^2}\left[2\arccos\left(\frac{r}{2R}\right)-\frac{r}{R}\sqrt{1-\left(\frac{r}{2R}\right)^2}\right],$$

其中 $0<r<2R$.

11. 在单位正方形中 (其顶点为 (0,0),(0,1),(1,1),(1,0)) "随机地" (解释何意!) 选一点 $P=(x,y)$. 证明该点 P 离点 (1,1) 比离点 (−1/2,1/2) 更近.

12. A 和 B 两人约好在 7 点到 8 点之间会面. 但是两人忘记了确切的会面时间, 只好在 7 点到 8 点之间 "随机地" 到达会面地点, 并且最多等 10 分钟. 证明他们能见面的概率等于 11/36.

13. 设 $X_1,X_2,\cdots$ 是独立随机变量序列, $S_n=\sum_{i=1}^n X_i$, 而 $\sigma(S_2)$ 是由 S_2 生成的 σ – 代数. 证明 S_1 和 S_3 关于 $\sigma(S_2)$ 条件独立.

14. 称 σ – 代数 $\mathscr{F}_1$ 和 $\mathscr{F}_2$ 关于 σ – 代数 $\mathscr{F}_3$ 条件独立, 如果对于任意 $A_i\in\mathscr{F}_i$, $i=1,2$,

$$\mathbf{P}(A_1A_2|\mathscr{F}_3)=\mathbf{P}(A_1|\mathscr{F}_3)\mathbf{P}(A_2|\mathscr{F}_3).$$

证明 $\mathscr{F}_1$ 和 $\mathscr{F}_2$ 关于 $\mathscr{F}_3$ 条件独立, 与下列任何条件之一等价 ($\mathbf{P}$ – a.c.):

(a) 对于一切 $A_1\in\mathscr{F}_1$, 有 $\mathbf{P}(A_1|\sigma(\mathscr{F}_2\bigcup\mathscr{F}_3))=\mathbf{P}(A_1|\mathscr{F}_3)$;

(b) 对于一切集合 $B\in\mathscr{P}_1$, 其中集系 $\mathscr{P}_1$ 形成 π – 系 $\mathscr{F}_1=\sigma(\mathscr{P}_1)$, 有 $\mathbf{P}(B|\sigma(\mathscr{F}_2\bigcup\mathscr{F}_3))=\mathbf{P}(B|\mathscr{F}_3)$;

(c) 对于一切分别属于 π – 系 $\mathscr{P}_1$ 和 $\mathscr{P}_2$ 的集合 B_1 和 B_2, 而且其中 $\mathscr{F}_1=\sigma(\mathscr{P}_1)$ 和 $\mathscr{F}_2=\sigma(\mathscr{P}_2)$, 有 $\mathbf{P}(B_1B_2|\sigma(\mathscr{F}_2\bigcup\mathscr{F}_3))=\mathbf{P}(B_1|\mathscr{F}_3)\mathbf{P}(B_2|\mathscr{F}_3)$;

(d) 对于一切 $\sigma(\mathscr{F}_2\bigcup\mathscr{F}_3)$ – 可测且有数学期望 $\mathbf{E}X$ 的随机变量 X (见 §6 定义 2), 有 $\mathbf{E}(X|\sigma(\mathscr{F}_2\bigcup\mathscr{F}_3))=\mathbf{E}(X|\mathscr{F}_3)$.

15. 证明关于条件数学期望的法图引理的如下扩展形式 (对照定理 2 之 (d)).

设概率空间 $(\Omega,\mathscr{F},\mathbf{P})$, 而 $(\xi_n)_{n\geqslant 1}$ 是随机变量序列, 且数学期望 $\mathbf{E}\xi_n$, $n\geqslant 1$ 而 $\mathbf{E}\varliminf\xi_n$ 存在 (亦可取 $\pm\infty$ 为值; 见 §6 定义 2).

假设 $\mathscr{G}$ 是 $\mathscr{F}$ 中事件的子 σ – 代数, 而且当 $a\to\infty$ 时,

$$\sup_{n\geqslant 1}\mathbf{E}(\xi_n^- I_{\{\xi_n^-\geqslant a\}}|\mathscr{G})\to 0\quad(\mathbf{P}-\text{a.c.}).$$

证明

$$\mathbf{E}(\varliminf\xi_n|\mathscr{G})\leqslant\varliminf\mathbf{E}(\xi_n|\mathscr{G})\quad(\mathbf{P}-\text{a.c.}).$$

16. 假设像上题一样, 对于随机变量序列 $(\xi_n)_{n\geqslant 1}$, 数学期望 $\mathbf{E}\xi_n$, $n\geqslant 1$, 存在, 而 $\mathscr{G}$ 是 $\mathscr{F}$ 中事件的子 σ – 代数, 且

$$\sup_n\lim_{k\to\infty}\mathbf{E}(|\xi_n|I_{\{|\xi_n|\geqslant k\}}|\mathscr{G})=0\quad(\mathbf{P}-\text{a.c.}).\tag{60}$$

证明, 如果 $\xi_n\to\xi$ ($\mathbf{P}$ – a.c.), 且数学期望 $\mathbf{E}\xi_n$, $n\geqslant 1$, 存在, 则

$$\mathbf{E}(\xi_n|\mathscr{G})\to\mathbf{E}(\xi|\mathscr{G})\quad(\mathbf{P}-\text{a.c.}).$$

17. 假设在上题的条件下, 式 (60) 换成满足条件: 对于某一 $\alpha>1$,

$$\sup_n\mathbf{E}(|\xi_n|^\alpha|\mathscr{G})<\infty\quad(\mathbf{P}-\text{a.c.}).$$

那么,

$$\mathbf{E}(\xi_n|\mathscr{G})\to\mathbf{E}(\xi|\mathscr{G})\quad(\mathbf{P}-\text{a.c.}).$$

18. 设对于某个 $p\geqslant 1$, $\xi_n\xrightarrow{L^p}\xi$. 证明 $\mathbf{E}(\xi_n|\mathscr{G})\xrightarrow{L^p}\mathbf{E}(\xi|\mathscr{G})$.

19. (a) 设 $\mathbf{D}(X|Y)=\mathbf{E}[(X-\mathbf{E}(X|Y))^2|Y]$. 则 $\mathbf{D}X=\mathbf{E}\mathbf{D}(X|Y)+\mathbf{D}\mathbf{E}(X|Y)$.

(b) 证明 $\operatorname{cov}(X,Y)=\operatorname{cov}(X,\mathbf{E}(Y|X))$.

20. 说明例 5 中的充分统计量 $T(\omega)=s(X_1(\omega))+\cdots+s(X_n(\omega))$ 是否最小的.

21. 证明因子分解公式 (57).

22. 对于第 10 小节的例 5, 证明

$$\mathbf{E}_\theta(X_i|T)=\frac{n-1}{2n}T,$$

其中对于 $\omega=(x_1,\cdots,x_n)$, $X_i(\omega)=x_i$, $i=1,\cdots,n$.

§8. 随机变量 II

1. 方差, 协方差和相关系数 在第一章引进了简单随机变量的一些简单数字特征: 如方差、协方差和相关系数. 在一般情形下也相应地引进这些概念. 具体地说, 假设 $(\Omega,\mathscr{F},\mathbf{P})$ 是概率空间, 而随机变量 $\xi=\xi(\omega)$ 有数学期望 $\mathbf{E}\xi$.

称

$$\mathbf{D}\xi=\mathbf{E}(\xi-\mathbf{E}\xi)^2$$

为随机变量 ξ 的*方差*.

称 $\sigma=+\sqrt{\mathbf{D}\xi}$ 为随机变量 ξ 的*标准差*.

如果随机变量 ξ 有高斯 (正态) 密度:

$$f_\xi(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}},\quad \sigma>0,\quad -\infty<x<+\infty,\tag{1}$$

则 (1) 式中参数的含义非常简单:

$$m=\mathbf{E}\xi,\quad \sigma^2=\mathbf{D}\xi.$$

这样, 称为*高斯分布*或*正态分布*的随机变量 ξ 的密度, 完全决定于其均值 m 和方差 σ^2. (因此, 对于参数为 m 和 σ^2 的正态分布, 常使用记号: $\xi\sim N(m,\sigma^2)$.)

现在假设 (ξ,η) 是随机向量, 称

$$\operatorname{cov}(\xi,\eta)=\mathbf{E}(\xi-\mathbf{E}\xi)(\eta-\mathbf{E}\eta)\tag{2}$$

为 ξ 和 η 的*协方差* (假设其中的各个数学期望存在).

如果 $\operatorname{cov}(\xi,\eta)=0$, 则称随机变量 ξ 和 η *不相关*.

如果 $0<\mathbf{D}\xi<\infty$, $0<\mathbf{D}\eta<\infty$, 则称

$$\rho(\xi,\eta)=\frac{\operatorname{cov}(\xi,\eta)}{\sqrt{\mathbf{D}\xi\times\mathbf{D}\eta}}\tag{3}$$

为随机变量 ξ 和 η 的*相关系数*.

在第一章 §4 曾讲述了简单随机变量的方差、协方差和相关系数的性质. 在一般情形下, 这些性质的提法完全类似.

设随机向量 $\xi=(\xi_1,\cdots,\xi_n)$ 的分量有有限二阶矩. 称 $n\times n$ 阶矩阵 $\boldsymbol{R}=(R_{ij})_{n\times n}$ 为随机向量 (ξ,η) 的*协方差矩阵*. 其中 $R_{ij}=\operatorname{cov}(\xi_i,\xi_j)$. 显然, $\boldsymbol{R}$ 是对称矩阵. 此外, 由于对于任意 $\lambda_i\in\mathbb{R}$, $i=1,\cdots,n$,

$$\sum_{i,j=1}^n R_{ij}\lambda_i\lambda_j=\mathbf{E}\left[\sum_{i=1}^n(\xi_i-\mathbf{E}\xi_i)\lambda_i\right]^2\geqslant 0,$$

可见 $\boldsymbol{R}$ 是*非负定矩阵*, 即

$$\sum_{i,j=1}^n R_{ij}\lambda_i\lambda_j\geqslant 0.$$

下面的引理说明逆命题亦成立.

引理 *$n\times n$ 阶矩阵 $\boldsymbol{R}=(R_{ij})_{n\times n}$ 为随机向量 (ξ,η) 的协方差矩阵的充分和必要条件是, $\boldsymbol{R}$ 为对称和非负定矩阵, 或者等价地: 存在 $n\times k(1\leqslant k\leqslant n)$ 阶矩阵 $\boldsymbol{A}$, 使*

$$\boldsymbol{R}=\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}},$$

其中上标 "T" 是矩阵的 "转置" 符号: $\boldsymbol{A}^{\mathrm{T}}$ 是矩阵 $\boldsymbol{A}$ 的转置.

证明 上面已经证明, 任何协方差矩阵都是对称和非负定矩阵.

反之, 设 $\boldsymbol{R}$ 是对称的非负定矩阵. 由矩阵理论知, 对于任意对称非负定矩阵, 存在正交矩阵 $\boldsymbol{Z}$ (即 $\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}=\boldsymbol{E}$ 是单位矩阵), 使

$$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z}=\boldsymbol{D},$$

其中 $\boldsymbol{D}$ 是对角线上元素为非负 $(d_i\geqslant 0, i=1,\cdots,n)$ 的对角矩阵:

$$\boldsymbol{D}=\begin{pmatrix}d_1&0&\cdots&0\\0&d_2&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&d_n\end{pmatrix}.$$

由此可见,

$$\boldsymbol{R}=\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^{\mathrm{T}}=(\boldsymbol{Z}\boldsymbol{B})(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}}),$$

其中 $\boldsymbol{B}$ 是对角线上元素为 $b_i=+\sqrt{d_i}(i=1,\cdots,n)$ 的对角矩阵. 因此, 如果设 $\boldsymbol{A}=\boldsymbol{Z}\boldsymbol{B}$, 则对于 $\boldsymbol{R}$, 得所要求的表现 $\boldsymbol{R}=\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$.

显然, $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ 就是所要求的对称非负定矩阵. 因此, 只需证明 $\boldsymbol{R}$ 是某随机向量的协方差矩阵.

设 $\eta_1, \eta_2, \cdots, \eta_n$ 是独立同正态分布 $N(0,1)$ 的随机变量序列. (由 §9 定理 1 之系 1, 可见这样的随机变量序列存在; 实际上, 这容易由 §1 定理 2 得到) 那么, 随机向量 $\xi = \boldsymbol{A}\eta$ 具有所要求的性质 (注意, $\xi = \boldsymbol{A}\eta$ 应视为列向量). 事实上,

$$\mathrm{E}\xi\xi^{\mathrm{T}} = \mathrm{E}(\boldsymbol{A}\eta)(\boldsymbol{A}\eta)^{\mathrm{T}} = \boldsymbol{A}(\mathrm{E}\eta\eta^{\mathrm{T}})\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{AEA}^{\mathrm{T}} = \boldsymbol{AA}^{\mathrm{T}}.$$

(如果 $\zeta = (\zeta_{ij})$ 是以随机变量为元素的矩阵, 则 $\mathbf{E}\zeta$ 表示 $(\mathbf{E}\zeta_{ij})$.) ∎

现在考虑二维高斯 (正态) 密度:

$$f_{\xi,\eta}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_1)^2}{\sigma_1^2} - 2\rho\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right]\right\}, \tag{4}$$

其中含 5 个参数: $m_1, m_2, \sigma_1, \sigma_2, \rho$ (对照 §3 之 (14) 式), 其中 $|m_1| < \infty$, $|m_2| < \infty$, $\sigma_1 > 0$, $\sigma_2 > 0$, $|\rho| < 1$. (见 §3 图 28.) 经过简单的计算, 可以说明这些参数的概率意义:

$$m_1 = \mathbf{E}\xi, \quad \sigma_1^2 = \mathbf{D}\xi,$$
$$m_2 = \mathbf{E}\eta, \quad \sigma_2^2 = \mathbf{D}\eta,$$
$$\rho = \rho(\xi,\eta).$$

在第一章 §4 曾经说明, 如果随机变量 ξ 和 η 不相关, 则还不能说明它们独立. 但是, 假如 (ξ,η) 是正态的 (高斯的), 则由 ξ 和 η 不相关, 可见它们独立.

事实上, 如果在 (4) 式中 $\rho = 0$, 则

$$f_{\xi,\eta}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{(x-m_1)^2}{2\sigma_1^2}\right\}\exp\left\{-\frac{(y-m_2)^2}{2\sigma_2^2}\right\}.$$

但是, 由于 §6 的 (55) 式和本小节 (4) 式, 有

$$f_\xi(x) = \int_{-\infty}^{\infty} f_{\xi,\eta}(x,y)\mathrm{d}y = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{(x-m_1)^2}{2\sigma_1^2}\right\},$$
$$f_\eta(y) = \int_{-\infty}^{\infty} f_{\xi,\eta}(x,y)\mathrm{d}x = \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left\{-\frac{(y-m_2)^2}{2\sigma_2^2}\right\}.$$

于是,

$$f_{\xi,\eta}(x,y) = f_\xi(x)f_\eta(y),$$

由此可见, 随机变量 ξ 和 η 独立 (见 §6 的第 9 小节末尾).

2. 最优估计量 在前面 §7 引进的条件数学期望概念, 具有说服力的重要性在于它在解决关于参数估计理论的如下问题中的应用 (对照第一章 §4 的第 8 小节).

设 (ξ,η) 是随机向量, 其中 ξ 可以观测, 而 η 不可观测. 问如何根据对 ξ 的观测值, "估计" 不可观测的分量 η?

为使问题更加明确, 我们引进估计量的概念. 设 $\varphi = \varphi(x)$ 是博雷尔函数. 称随机变量 $\varphi(\xi)$ 为由 ξ 对 η 的估计量, 而称 $\mathbf{E}[\eta - \varphi(\xi)]^2$ 为该估计量的 (均根方) 误差. 估计量 $\varphi^*(\xi)$ 称做 (在根均方意义上) 最优的, 如果

$$\Delta = \mathbf{E}[\eta - \varphi^*(\xi)]^2 = \inf_{\varphi}\mathbf{E}[\eta - \varphi(\xi)]^2, \tag{5}$$

其中 inf 在一切博雷尔函数 $\varphi = \varphi(x)$ 类中求求.

定理 1 设 $\mathbf{E}\eta^2 < \infty$. 那么, 最优估计量 $\varphi^*(x)$ 存在, 并且作为 $\varphi^*(\xi)$ 可以选任意函数:

$$\varphi^*(x) = \mathbf{E}(\eta|\xi = x). \tag{6}$$

证明 不失普遍性, 可以只考虑满足 $\mathbf{E}\varphi^2(\xi) < \infty$ 的 $\varphi(\xi)$. 那么, 如果 $\varphi(\xi)$ 是所欲求的估计量, 而 $\varphi^*(\xi) = \mathbf{E}(\eta|\xi)$, 则

$$\begin{aligned}\mathbf{E}[\eta - \varphi(\xi)]^2 &= \mathbf{E}\{[\eta - \varphi^*(\xi)] - [\varphi^*(\xi) - \varphi(\xi)]\}^2 \\ &= \mathbf{E}[\eta - \varphi^*(\xi)]^2 + \mathbf{E}[\varphi^*(\xi) - \varphi(\xi)]^2 + 2\mathbf{E}[\eta - \varphi^*(\xi)][\varphi^*(\xi) - \varphi(\xi)] \\ &\geqslant \mathbf{E}[\eta - \varphi^*(\xi)]^2,\end{aligned}$$

由于 $\mathbf{E}[\varphi^*(\xi) - \varphi(\xi)]^2 \geqslant 0$, 并且由条件数学期望的性质, 可见

$$\begin{aligned}\mathbf{E}[\eta - \varphi^*(\xi)][\varphi^*(\xi) - \varphi(\xi)] &= \mathbf{E}\{\mathbf{E}[\eta - \varphi^*(\xi)][\varphi^*(\xi) - \varphi(\xi)]|\xi\} \\ &= \mathbf{E}\{[\varphi^*(\xi) - \varphi(\xi)]\mathbf{E}[\eta - \varphi^*(\xi)|\xi]\} = 0.\end{aligned}$$ ∎

注 1 由定理的证明, 可见如果 ξ 不是随机变量, 而是在某可测空间 $(E,\mathscr{E})$ 取值的随机元, 则其结论仍然成立. 在后一种情形下, 估计量 $\varphi = \varphi(x)$ 应理解为 $\mathscr{E}/\mathscr{B}(\mathbb{R})$ 可测函数.

现在, 假设 (ξ,η) 是服从高斯分布 (正态分布) 的随机向量, 并讨论 $\varphi^*(x)$ 的结构, 其中 (4) 式为二维正态分布的密度.

由 §7 中 (1), (4) 和 (18) 各式知, 条件概率分布密度 $f_{\eta|\xi}(y|x)$ 为

$$f_{\eta|\xi}(y|x) = \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_2^2}}\exp\left\{-\frac{(y-m(x))^2}{2\sigma_2^2(1-\rho^2)}\right\}, \tag{7}$$

其中

$$m(x) = m_2 + \frac{\sigma_2}{\sigma_1}\rho(x - m_1). \tag{8}$$

那么, 由 §7 定理 3 的系, 可见

$$\mathbf{E}(\eta|\xi = x) = \int_{-\infty}^{\infty} y f_{\eta|\xi}(y|x)\mathrm{d}y = m(x), \tag{9}$$

且

$$\begin{aligned}\mathbf{D}(\eta|\xi=x)&=\mathbf{E}\{[\eta-\mathbf{E}(\eta|\xi=x)]^2|\xi=x\}\\&=\int_{-\infty}^{\infty}[y-m(x)]^2 f_{\eta|\xi}(y|x)\mathrm{d}y=\sigma_2^2(1-\rho^2).\end{aligned}\tag{10}$$

注意, 条件方差 $\mathbf{D}(\eta|\xi=x)$ 不依赖于 x, 因此,

$$\Delta=\mathbf{E}[\eta-\mathbf{E}(\eta|\xi=x)]^2=\sigma_2^2(1-\rho^2).\tag{11}$$

公式 (9) 和 (11) 是在假设 $\mathbf{D}\xi>0, \mathbf{D}\eta>0$ 条件下得到的. 假如 $\mathbf{D}\xi>0$, 而 $\mathbf{D}\eta=0$, 则 (9) 和 (11) 两式显然成立.

于是, 得到下面的结果 (对照第一章 §4 的 (16) 式和 (17) 式).

定理 2 (正态相关定理) *设 (ξ,η) 是 $\mathbf{D}\xi>0$ 的高斯 (正态) 随机向量. 由 ξ 对 η 的最优估计量是*

$$\mathbf{E}(\eta|\xi)=\mathbf{E}\eta+\frac{\operatorname{cov}(\xi,\eta)}{\mathbf{D}\xi}(\xi-\mathbf{E}\xi),\tag{12}$$

而其误差为

$$\Delta\equiv\mathbf{E}[\eta-\mathbf{E}(\eta|\xi)]^2=\mathbf{D}\eta-\frac{\operatorname{cov}^2(\xi,\eta)}{\mathbf{D}\xi}.\tag{13}$$

注 2 曲线 $y(x)=\mathbf{E}(\eta|\xi=x)$ 称做 η 在 ξ 上的或 η 关于 ξ 的回归曲线. 在正态情形下 $\mathbf{E}(\eta|\xi=x)=ax+b$, 因而 η 关于 ξ 的回归是线性的. 因此第一章 §4 中, 对于最优线性估计及其误差, (12) 和 (13) 式的右侧与 (16) 和 (17) 式的右侧对应相等, 就毫不意外.

系 *设 ε_1 和 ε_2 是均值为 0 和方差为 1 的独立正态随机变量, 而*

$$\xi=a_1\varepsilon_1+a_2\varepsilon_2,\quad \eta=b_1\varepsilon_1+b_2\varepsilon_2.$$

那么, $\mathbf{E}\xi=\mathbf{E}\eta=0$, $\mathbf{D}\xi=a_1^2+a_2^2$, $\mathbf{D}\eta=b_1^2+b_2^2$, $\operatorname{cov}(\xi,\eta)=a_1b_1+a_2b_2$, 而且若 $a_1^2+a_2^2>0$, 则

$$\mathbf{E}(\eta|\xi)=\frac{a_1b_1+a_2b_2}{a_1^2+a_2^2}\xi,\tag{14}$$

$$\Delta=\frac{(a_1b_2-a_2b_1)^2}{a_1^2+a_2^2}.\tag{15}$$

3. 随机变量的函数的分布 假设一个随机变量是另一个随机变量的函数. 现在, 讨论求随机变量函数的概率分布的问题.

设随机变量 ξ 的分布函数为 $F_\xi(x)$ (如果它有密度, 则记作 $f_\xi(x)$), $\varphi=\varphi(x)$ 是某一博雷尔函数, 而 $\eta=\varphi(\xi)$. 记 $I_y=(-\infty,y]$. 则

$$F_\eta(y)=\mathbf{P}\{\eta\leqslant y\}=\mathbf{P}\{\varphi(\xi)\in I_y\}=\mathbf{P}\{\xi\in\varphi^{-1}(I_y)\}=\int_{\varphi^{-1}(I_y)}F_\xi(\mathrm{d}x).\tag{16}$$

于是, 随机变量 $\eta=\varphi(\xi)$ 的分布函数 $F_\eta(y)$, 可以通过随机变量 ξ 的分布函数 $F_\xi(x)$ 和函数 φ 来表示.

例如, 若 $\eta=a\xi+b, a>0$, 则

$$F_\eta(y)=\mathbf{P}\left\{\xi\leqslant\frac{y-b}{a}\right\}=F_\xi\left(\frac{y-b}{a}\right).\tag{17}$$

设 $\eta=\xi^2$, 则对于 $y<0$, 显然 $F_\eta(y)=0$, 而对于 $y\geqslant0$, 有

$$\begin{aligned}F_\eta(y)&=\mathbf{P}\{\xi^2\leqslant y\}=\mathbf{P}\{-\sqrt{y}\leqslant\xi\leqslant\sqrt{y}\}\\&=F_\xi(\sqrt{y})-F_\xi(-\sqrt{y})+\mathbf{P}\{\xi=-\sqrt{y}\}.\end{aligned}\tag{18}$$

现在考虑求密度 $f_\eta(y)$ 的问题.

假设随机变量 ξ 的值域是 (有限或无限) 开区间 $I=(a,b)$, 而函数 $\varphi=\varphi(x)$ 对于 $x\in I$ 有定义、连续可微, 并且是单调的: 或严格单调增加, 或严格单调减小. 此外, 假设 $\varphi'(x)\neq0, x\in I$.

记 $h(y)=\varphi^{-1}(y)$, 并且不失普遍性假设 $\varphi(x)$ 严格增加. 那么, 对于 $y\in\{\varphi(x): x\in I\}$, 有

$$\begin{aligned}F_\eta(y)&=\mathbf{P}\{\eta\leqslant y\}=\mathbf{P}\{\varphi(\xi)\leqslant y\}=\mathbf{P}\{\xi\leqslant\varphi^{-1}(y)\}\\&=\mathbf{P}\{\xi\leqslant h(y)\}=\int_{-\infty}^{h(y)}f_\xi(x)\mathrm{d}x.\end{aligned}\tag{19}$$

根据 §6 练习题 15,

$$\int_{-\infty}^{h(y)}f_\xi(x)\mathrm{d}x=\int_{-\infty}^{y}f_\xi(h(z))h'(z)\mathrm{d}z.\tag{20}$$

因此

$$f_\eta(y)=f_\xi(h(y))h'(y).\tag{21}$$

类似地, 如果 $\varphi(x)$ 严格减小, 则

$$f_\eta(y)=f_\xi(h(y))(-h'(y)).$$

于是, 在两种情形下都有

$$f_\eta(y)=f_\xi(h(y))|h'(y)|.\tag{22}$$

例如, 若 $\eta=a\xi+b, a\neq0$, 则

$$h(y)=\frac{y-b}{a},\quad f_\eta(y)=\frac{1}{|a|}f_\xi\left(\frac{y-b}{a}\right).$$

如果 $\xi\sim\mathscr{N}(m,\sigma^2)$, 而 $\eta=e^\xi$, 则由 (22) 式, 有

$$f_\eta(y)=\begin{cases}\dfrac{1}{\sqrt{2\pi}\sigma y}\exp\left\{-\dfrac{\ln(y/M)^2}{2\sigma^2}\right\}, & 若\ y>0,\\ 0, & 若\ y\leqslant0,\end{cases}\tag{23}$$

其中 $M = e^m$.

以 (23) 式为密度的概率分布称做对数正态分布.

如果函数 $\varphi = \varphi(x)$ 不是严格增加或严格减小的, 则 (22) 式对于 $\eta = \varphi(\xi)$ 不能用. 不过, 对于许多应用, 其如下推广就完全够用.

设函数 $\varphi = \varphi(x)$ 定义在集合 $\sum_{k=1}^{n}[a_k, b_k]$ 上, 并且在每一个开区间 $I_k = (a_k, b_k)$ 上连续可微, 以及或者严格增加, 或者严格减小, 并且 $\varphi'(x) \neq 0, x \in I_k$. 此外, 假设 $h_k = h_k(y)$ 是 $\varphi(x), x \in I_k$ 的反函数. 那么, 有 (22) 式如下推广:

$$f_\eta(y) = \sum_{k=1}^{n} f_\xi(h_k(y))\,|h_k'(y)|\,I_{D_k}(y), \tag{24}$$

其中 D_k 是函数 $h_k(y)$ 的定义域.

例如, 若 $\eta = \xi^2$, 则设 $I_1 = (-\infty, 0)$, $I_2 = (0, \infty)$, 得 $h_1(y) = -\sqrt{y}$, $h_2(y) = \sqrt{y}$, 因而

$$f_\eta(y) = \begin{cases} \dfrac{1}{2\sqrt{y}}[f_\xi(\sqrt{y}) + f_\xi(-\sqrt{y})], & 若\ y > 0, \\ 0, & 若\ y \leqslant 0. \end{cases} \tag{25}$$

注意, 由于 $\mathrm{P}\{\xi = -\sqrt{y}\} = 0$, 这一结果亦可由 (18) 式得到. 特别, 如果 $\xi \sim \mathcal{N}(0, 1)$, 则

$$f_{\xi^2}(y) = \begin{cases} \dfrac{1}{\sqrt{2\pi y}}e^{-\frac{y}{2}}, & 若\ y > 0, \\ 0, & 若\ y \leqslant 0. \end{cases} \tag{26}$$

经不复杂的计算亦可得到:

$$f_{|\xi|}(y) = \begin{cases} f_\xi(y) + f_\xi(-y), & 若\ y > 0, \\ 0, & 若\ y \leqslant 0. \end{cases} \tag{27}$$

$$f_{+\sqrt{|\xi|}}(y) = \begin{cases} 2y(f_\xi(y^2) + f_\xi(-y^2)), & 若\ y > 0, \\ 0, & 若\ y \leqslant 0. \end{cases} \tag{28}$$

4. 随机变量多元函数的分布　现在考虑多个随机变量的函数.

如果 ξ 和 η 是联合分布函数为 $F_{\xi,\eta}(x, y)$ 的随机变量, 而 $\varphi = \varphi(x, y)$ 是某一博雷尔函数, 则立即可以得到 $\zeta = \varphi(\xi, \eta)$ 的分布函数:

$$F_\zeta(z) = \int_{\{(x,y):\, \varphi(x,y) \leqslant z\}} \mathrm{d}F_{\xi,\eta}(x, y). \tag{29}$$

例如, 如果 $\varphi(x, y) = x + y$, 而 ξ 和 η 独立 (即 $F_{\xi,\eta}(x, y) = F_\xi(x)F_\eta(y)$), 则利用傅比尼定理, 有

$$\begin{aligned} F_\zeta(z) &= \int_{\{(x,y):\, x+y \leqslant z\}} \mathrm{d}F_\xi(x)\mathrm{d}F_\eta(y) = \int_{\mathbb{R}^2} I_{\{x+y \leqslant z\}}(x, y)\mathrm{d}F_\xi(x)\mathrm{d}F_\eta(y) \\ &= \int_{-\infty}^{\infty} \mathrm{d}F_\xi(x)\left[\int_{-\infty}^{\infty} I_{\{x+y \leqslant z\}}(x, y)\mathrm{d}F_\eta(y)\right] = \int_{-\infty}^{\infty} F_\eta(z - x)\mathrm{d}F_\xi(x). \end{aligned} \tag{30}$$

类似地可得

$$F_\zeta(z) = \int_{-\infty}^{\infty} F_\xi(z - y)\mathrm{d}F_\eta(y). \tag{31}$$

对于两个分布函数 F 和 G, 函数

$$H(z) = \int_{-\infty}^{\infty} F(z - x)\mathrm{d}G(x)$$

通常记作 $F * G$, 并称做 F 和 G *卷积*.

这样, 两个随机变量 ξ 与 η 之和 $\zeta = \xi + \eta$ 的分布函数 F_ζ, 是它们分布函数 F_ξ 与 F_η 的卷积:

$$F_\zeta = F_\xi * F_\eta.$$

显然, 这时 $F_\xi * F_\eta = F_\eta * F_\xi$.

现在假设独立随机变量 ξ 和 η 各自有密度 f_ξ 和 f_η. 那么, 仍然由傅比尼定理及 (31) 式, 有

$$\begin{aligned} F_\zeta(z) &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{z-y} f_\xi(u)\mathrm{d}u\right] f_\eta(y)\mathrm{d}y = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{z} f_\xi(u - y)\mathrm{d}u\right] f_\eta(y)\mathrm{d}y \\ &= \int_{-\infty}^{z}\left[\int_{-\infty}^{\infty} f_\xi(u - y)f_\eta(y)\mathrm{d}y\right]\mathrm{d}u. \end{aligned}$$

由此, 得

$$f_\zeta(z) = \int_{-\infty}^{\infty} f_\xi(z - y)f_\eta(y)\mathrm{d}y, \tag{32}$$

同理, 有

$$f_\zeta(z) = \int_{-\infty}^{\infty} f_\eta(z - x)f_\xi(x)\mathrm{d}x. \tag{33}$$

下面举几个应用这些公式的例子.

设 $\xi_1, \xi_2, \cdots, \xi_n$ 是独立同分布随机变量序列, 其共同的密度是在 $[-1, 1]$ 均匀分布密度:

$$f(x) = \begin{cases} \dfrac{1}{2}, & 若\ |x| \leqslant 1, \\ 0, & 若\ |x| > 1. \end{cases}$$

那么, 由 (32) 式, 可见

$$f_{\xi_1+\xi_2}(x) = \begin{cases} \dfrac{2 - |x|}{4}, & 若\ |x| \leqslant 2, \\ 0, & 若\ |x| > 2, \end{cases}$$

$$f_{\xi_1+\xi_2+\xi_3}(x) = \begin{cases} \dfrac{(3 - |x|)^2}{16}, & 若\ 1 \leqslant |x| \leqslant 3, \\ \dfrac{3 - x^2}{8}, & 若\ 0 \leqslant |x| \leqslant 1, \\ 0, & 若\ |x| > 3, \end{cases}$$

而由归纳法, 可见对于一般情形, 有

$$f_{\xi_1+\cdots+\xi_n}(x)=\begin{cases}\dfrac{1}{2^n(n-1)!}\displaystyle\sum_{k=0}^{\left[\frac{n+x}{2}\right]}(-1)^k C_n^k(n+x-2k)^{n-1}, & 若\ |x|\leqslant n,\\ 0, & 若\ |x|>n.\end{cases}$$

现在设 $\xi\sim N(m_1,\sigma_1^2)$, $\eta\sim N(m_2,\sigma_2^2)$. 若记

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2},$$

则

$$f_\xi(x)=\frac{1}{\sigma_1}\varphi\left(\frac{x-m_1}{\sigma_1}\right),\ f_\eta(x)=\frac{1}{\sigma_2}\varphi\left(\frac{x-m_2}{\sigma_2}\right),$$

由 (32) 式, 得

$$f_{\xi+\eta}(x)=\frac{1}{\sqrt{\sigma_1^2+\sigma_2^2}}\varphi\left(\frac{x-(m_1+m_2)}{\sqrt{\sigma_1^2+\sigma_2^2}}\right).$$

于是, 两个独立正态随机变量之和, 仍然是正态随机变量, 其均值为 m_1+m_2, 而方差为 $\sigma_1^2+\sigma_2^2$.

设 $\xi_1,\xi_2,\cdots,\xi_n$ 是独立同分布随机变量, 都服从均值为 0、方差为 1 的正态分布. 那么, 由 (26) 式 (用归纳法) 不难求出

$$f_{\xi_1^2+\cdots+\xi_n^2}(x)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma(n/2)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, & 若\ x>0,\\ 0, & 若\ x\leqslant 0.\end{cases}\tag{34}$$

随机变量 $\xi_1^2+\cdots+\xi_n^2$ 通常记作 χ_n^2, 而其概率分布称做自由度为 n 的 χ^2 分布 ("卡方" 分布) (对照 §3 表 2－3).

如果记 $\chi_n=+\sqrt{\chi_n^2}$, 则由 (28) 和 (34) 式, 可见

$$f_{\chi_n}(x)=\begin{cases}\dfrac{2}{2^{n/2}\Gamma(n/2)}x^{n-1}e^{-\frac{x^2}{2}}, & 若\ x>0,\\ 0, & 若\ x\leqslant 0.\end{cases}\tag{35}$$

具有此密度的概率分布, 通常称做自由度为 n 的 χ 分布 ("卡" 分布).

仍设独立随机变量 ξ 和 η 各自的密度为 f_ξ 和 f_η. 那么,

$$F_{\xi\eta}(z)=\iint\limits_{\{(x,y):xy\leqslant z\}}f_\xi(x)f_\eta(y)\,dx\,dy,$$

$$F_{\xi/\eta}(z)=\iint\limits_{\{(x,y):x/y\leqslant z\}}f_\xi(x)f_\eta(y)\,dx\,dy.$$

由此, 不难得到

$$f_{\xi\eta}(z)=\int_{-\infty}^{\infty}f_\xi\left(\frac{z}{y}\right)f_\eta(y)\frac{dy}{|y|}=\int_{-\infty}^{\infty}f_\eta\left(\frac{z}{x}\right)f_\xi(x)\frac{dx}{|x|},\tag{36}$$

和

$$f_{\xi/\eta}(z)=\int_{-\infty}^{\infty}f_\xi(zy)f_\eta(y)|y|\,dy.\tag{37}$$

在 (37) 式中, 设

$$\xi=\xi_0,\quad \eta=\sqrt{\frac{\xi_1^2+\cdots+\xi_n^2}{n}},$$

其中 $\xi_0,\xi_1,\cdots,\xi_n$ 是均值为 0、方差为 $\sigma^2>0$ 的独立正态随机变量. 由 (35) 式, 可得

$$f_t(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}.\tag{38}$$

其中随机变量

$$t=\frac{\xi_0}{\sqrt{\frac{1}{n}(\xi_1^2+\cdots+\xi_n^2)}}$$

习惯上用 t 表示, 相应的分布称做 t 分布或 "学生" (Student) 分布[①] (对照 §3 表 2－3). 注意, 该分布不依赖于 σ.

5. 练习题

1. 验证公式 (9), (10), (24), (27), (28), (34) ～ (38) 的正确性.

2. 设 $\xi_1,\xi_2,\cdots,\xi_n$, $n\geqslant 2$ 是独立同分布随机变量, 其共同的分布函数为 $F(x)$ (假如密度存在, 则记作 $f(x)$). 记 $\bar{\xi}=\max\{\xi_1,\cdots,\xi_n\}$, $\underline{\xi}=\min\{\xi_1,\cdots,\xi_n\}$, $\rho=\bar{\xi}-\underline{\xi}$. 证明

$$F_{\bar{\xi},\underline{\xi}}(y,x)=\begin{cases}(F(y))^n-[F(y)-F(x)]^n, & 若\ y>x,\\ (F(y))^n, & 若\ y\leqslant x,\end{cases}$$

$$f_{\bar{\xi},\underline{\xi}}(y,x)=\begin{cases}n(n-1)[F(y)-F(x)]^{n-2}f(x)f(y), & 若\ y>x,\\ 0, & 若\ y\leqslant x,\end{cases}$$

$$F_\rho(x)=\begin{cases}n\displaystyle\int_{-\infty}^{\infty}[F(y)-F(y-x)]^{n-1}f(y)\,dy, & 若\ x\geqslant 0,\\ 0, & 若\ x<0,\end{cases}$$

$$f_\rho(x)=\begin{cases}n(n-1)\displaystyle\int_{-\infty}^{\infty}[F(y)-F(y-x)]^{n-2}f(y-x)f(y)\,dy, & 若\ x>0,\\ 0, & 若\ x\leqslant 0.\end{cases}$$

---

① 戈塞特 (W. S. Gosset, 1876－1937) 是最早创立 t 分布的英国统计学家、化学家, 推断统计学派的先驱者, 发表有关论文时使用的笔名是 Student, 因此 t 分布亦称 "学生分布". 戈塞特于 1908 年《生物统计学》(Biometrika, 亦译《生物计量学》) 上发表的论文 "平均值的可能误差" 中首先提出了 t 分布, 并且描绘了其性质. —— 译者

3. 设 ξ_1 和 ξ_2 是独立的泊松随机变量, 其参数相应为 λ_1 和 λ_2. 证明 $\xi_1+\xi_2$ 仍然服从泊松分布, 参数为 $\lambda_1+\lambda_2$.

4. 在 (4) 式中设 $m_1=m_2=0$. 证明

$$f_{\xi/\eta}(z)=\frac{\sigma_1\sigma_2\sqrt{1-\rho^2}}{\pi(\sigma_2^2z^2-2\rho\sigma_1\sigma_2z+\sigma_1^2)}.$$

5. 量 $\rho^*(\xi,\eta)=\sup_{u,v}\rho(u(\xi),v(\eta))$, 其中对使相关系数 $\rho(u(\xi),v(\eta))$ 有定义的一切博雷尔函数 $u=u(x),v=v(x)$ 求上确界 (sup). $\rho^*(\xi,\eta)$ 称做 ξ 和 η 的最大相关系数. 证明随机变量 ξ 和 η 独立, 当且仅当 $\rho^*(\xi,\eta)=0$.

6. 设 $\tau_1,\tau_2,\cdots,\tau_k$ 是独立同分布非负随机变量, 都服从参数为 λ 的指数分布, 其共同的密度为

$$f(t)=\lambda e^{-\lambda t},\quad t\geqslant 0.$$

证明, 随机变量 $\tau_1+\tau_2+\cdots+\tau_k$ 的分布有密度:

$$\frac{\lambda^k t^{k-1}e^{-\lambda t}}{(k-1)!},\quad t\geqslant 0,$$

并且

$$\mathrm{P}\{\tau_1+\tau_2+\cdots+\tau_k>t\}=\sum_{i=0}^{k-1}e^{-\lambda t}\frac{(\lambda t)^i}{i!}.$$

7. 设 $\xi\sim N(0,\sigma^2)$. 证明对于任意 $p\geqslant 1$,

$$\mathrm{E}|\xi|^p=\frac{2^{p/2}}{\sqrt{\pi}}\Gamma\left(\frac{p+1}{2}\right)\sigma^p,$$

而

$$\Gamma(s)=\int_0^{\infty}x^{s-1}e^{-x}dx$$

是欧拉 (L. Euler) Γ 函数. 特别, 对于任意 $n\geqslant 1$,

$$\mathrm{E}\xi^{2n}=(2n-1)!!\sigma^{2n}.$$

8. 假设 ξ 和 η 是独立随机变量, 并且 $\xi+\eta$ 的分布与 ξ 的分布是同一分布. 证明 $\eta=0$ (a.e.).

9. 假设 (X,Y) 在单位圆 $\{(x,y):x^2+y^2\leqslant 1\}$ 上有均匀分布, 记 $W=X^2+Y^2$,

$$U=X\sqrt{-\frac{2\ln W}{W}},\quad V=Y\sqrt{-\frac{2\ln W}{W}}.$$

证明, U 和 V 是独立 $N(0,1)$– 分布随机变量.

10. 设 U 和 V 是独立同在 $(0,1)$ 均匀分布的随机变量, 定义

$$X=\sqrt{-2\ln V}\cos(2\pi U),\quad Y=\sqrt{-2\ln V}\sin(2\pi U).$$

证明, X 和 Y 是独立 $N(0,1)$– 分布随机变量.

11. 举一例子: ξ 和 η 是正态随机变量, 但是 $\xi+\eta$ 不服从正态分布.

12. 设 $X_1,\cdots,X_n$ 是独立同分布随机变量, 其共同的密度和分布函数相应为 $f=f(x)$ 和 $F=F(x)$, 而

$$\mathscr{R}_n=\max\{X_1,\cdots,X_n\}-\min\{X_1,\cdots,X_n\}$$

是样本 $(X_1,\cdots,X_n)$ 的 "极差". 证明随机变量 $\mathscr{R}_n$ 的密度 $f_{\mathscr{R}_n}(x),x>0$, 为

$$f_{\mathscr{R}_n}(x)=n(n-1)\int_{-\infty}^{\infty}[F(y)-F(y-x)]^{n-2}f(y)f(y-x)dy.$$

特别, 对于在 $[0,1]$ 上均匀分布的随机变量 $X_1,\cdots,X_n$,

$$f_{\mathscr{R}_n}(x)=\begin{cases}n(n-1)x^{n-2}(1-x), & 若\ 0\leqslant x\leqslant 1,\\ 0, & 若\ x<0\ 或\ x>1.\end{cases}$$

13. 设 $F(x)$ 是分布函数. 证明, 对于任意 $a>0$, 下列函数也是分布函数:

$$G_1(x)=\frac{1}{a}\int_x^{x+a}F(u)du,\quad G_2(x)=\frac{1}{2a}\int_{x-a}^{x+a}F(u)du.$$

14. 设随机变量 X 服从参数为 $\lambda>0$ 的指数分布 $(f_X(x)=\lambda e^{-\lambda x},x\geqslant 0)$, 求随机变量 $Y=X^{1/\alpha},\alpha>0$ 的概率密度 (相应的分布称做韦布尔 [W. Weibull] 分布).

设 $\lambda=1$. 求随机变量 $Y=-\ln X$ 的分布密度 (相应的分布称做双指数分布).

15. 设随机变量 X 和 Y 的联合分布密度 $f(x,y)$ 形如 $f(x,y)=g(\sqrt{x^2+y^2})$.

(a) 求随机变量 $\rho=\sqrt{X^2+Y^2}$ 和 $\theta=\tan^{-1}(Y/X)$ 的联合分布密度, 证明 ρ 和 θ 独立.

(b) 设 $U=X\cos\alpha+Y\sin\alpha$ 和 $V=-X\sin\alpha+Y\cos\alpha$. 证明随机变量 U 和 V 的联合分布密度与 $f(x,y)$ 相同. (这反映 "随机变量 X 和 Y 的联合分布关于 '旋转' 的不变性" 这一事实.)

16. 设 $X_1,\cdots,X_n$ 是独立同分布随机变量, 其共同的密度和分布函数相应为 $f=f(x)$ 和 $F=F(x)$, 记 (对照练习题 12) $X_{(1)}=\min\{X_1,\cdots,X_n\}$ 是 $X_1,\cdots,X_n$ 的最小值, $X_{(2)}$ 是第二小值, 等等, $X_{(n)}=\max\{X_1,\cdots,X_n\}$ 是 $X_1,\cdots,X_n$ 的最大值 (这样定义的随机变量 $X_{(1)},\cdots,X_{(n)}$ 称做随机变量 $X_1,\cdots,X_n$ 的顺序统计量). 证明:

(a) 随机变量 $X_{(k)}$ 的概率分布密度为

$$nf(x)C_{n-1}^{k-1}[F(x)]^{k-1}[1-F(x)]^{n-k};$$

(b) 随机变量 $X_{(1)},\cdots,X_{(n)}$ 的联合密度为

$$f(x_1,\cdots,x_n)=\begin{cases}n!f(x_1)\cdots f(x_n), & 若\ x_1<\cdots<x_n,\\ 0, & 若不然.\end{cases}$$

17. 设 $X_1, \cdots, X_n$ 是独立同分布随机变量, 且都服从正态分布 $N(\mu, \sigma^2)$. 设

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \quad \text{其中 } n > 1, \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

则随机变量 $\overline{X}$ 和 S^2 分别称做样本均值和样本方差. 证明:

(a) $\mathbf{E}S^2 = \sigma^2$;

(b) $\overline{X}$ 和 S^2 独立;

(c) $\overline{X} \sim N(\mu, \sigma^2/n)$, 而 $(n-1)S^2/\sigma^2$ 服从自由度为 $n-1$ 的 χ^2 分布.

18. 设 $X_1, \cdots, X_n$ 是独立同分布随机变量, $N(N = 1, 2, \cdots)$ 是不依赖于 $X_1, \cdots, X_n$ 的随机变量, 且 $\mathbf{E}N < \infty, \mathbf{D}N < \infty$; $S_N = X_1 + \cdots + X_N$. 证明

$$\mathbf{D}S_N = \mathbf{D}X_1\mathbf{E}N + (\mathbf{E}X_1)^2\mathbf{D}N, \quad \frac{\mathbf{D}S_N}{\mathbf{E}S_N} = \frac{\mathbf{D}X_1}{\mathbf{E}X_1} + \mathbf{E}X_1\frac{\mathbf{D}N}{\mathbf{E}N}.$$

19. 设 $M(t) = \mathbf{E}e^{tX}$ 是随机变量 X 的母函数. 证明对于任意 $t > 0$, 有 $\mathbf{P}\{X \geqslant 0\} \leqslant M(t)$.

20. 设 $X, X_1, \cdots, X_n$ 是独立同分布随机变量, $S_n = X_1 + \cdots + X_n, S_0 = 0$, 而

$$M_n = \max_{0 \leqslant j \leqslant n} S_j, \quad M = \sup_{n \geqslant 0} S_n.$$

证明

(a) 对于 $n \geqslant 1$, $\overline{M}_n$ 和 $(M_{n-1} + X)^+$ 同分布;

(b) 如果 $S_n \to -\infty$ ($\mathbf{P}$-a.s.), 则 M 和 $(M + X)^+$ 同分布;

(c) 如果 $-\infty < \mathbf{E}X < 0$ 且 $\mathbf{E}X^2 < \infty$, 则

$$\mathbf{E}M = \frac{\mathbf{D}X - \mathbf{D}(S - X)^+}{-2\mathbf{E}X}.$$

21. 在上题的条件下, 对于任意 $\varepsilon > 0$, 设

$$\overline{M}(\varepsilon) = \sup_{n \geqslant 0}(S_n - n\varepsilon).$$

证明

$$\lim_{\varepsilon \downarrow 0} \varepsilon \overline{M}(\varepsilon) = \frac{\mathbf{D}X}{2}.$$

§9. 建立具有给定有限维分布的过程

1. 具有给定分布函数的随机变量存在性 设 $\xi = \xi(\omega)$ 是概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 上的随机变量, 而

$$F_\xi(x) = \mathbf{P}\{\omega : \xi(\omega) \leqslant x\}$$

是其分布函数. 自然, 在 §3 定义 1 的意义上, $F_\xi(x)$ 是数轴上的分布函数.

现在提出如下问题. 假设 $F = F(x)$ 是 $\mathbf{R}$ 上某一分布函数, 问是否存在以 $F(x)$ 为分布函数的随机变量?

证实问题的这一提法的一个理由如下. 许多概率论的命题从如下的句子开始: "设 ξ 是分布函数为 $F(x)$ 的随机变量, 则 ……". 因此, 为使类似的论断有意义, 应该确信所考虑的对象确实存在. 由于为给出随机变量, 应先给出其定义域 $(\Omega, \mathscr{F})$; 而为了提起其概率分布, 应该有 $(\Omega, \mathscr{F})$ 上的概率测度 $\mathbf{P}$. 于是, 关于具有给定分布函数 $F(x)$ 的随机变量存在性的正确提法是:

是否存在概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 和此概率空间上的随机变量 $\xi = \xi(\omega)$, 使

$$\mathbf{P}\{\omega : \xi(\omega) \leqslant x\} = F(x)?$$

现在证明, 对于所提问题的答案是肯定的, 其实答案已经包含在 §3 的定理 1 中. 事实上, 设

$$\Omega = \mathbf{R}, \quad \mathscr{F} = \mathscr{B}(\mathbf{R}).$$

那么, 由 §3 的定理 1 可见, 在 $(\mathbf{R}, \mathscr{B}(\mathbf{R}))$ 上存在 (并且唯一) 概率测度 $\mathbf{P}$, 使对于任意 $a < b$, $\mathbf{P}(a, b] = F(b) - F(a)$.

设 $\xi(\omega) = \omega$. 那么,

$$\mathbf{P}\{\omega : \xi(\omega) \leqslant x\} = \mathbf{P}\{\omega : \omega \leqslant x\} = \mathbf{P}(-\infty, x] = F(x).$$

于是, 建立了所要求的概率空间和随机变量.

2. 具有给定有限维分布的随机过程的存在性 现在对于随机过程提出类似的问题.

对于 $t \in T \subseteq \mathbf{R}$, 设 $X = (\xi_t)_{t \in T}$ 是概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 上的随机过程 (见 §5 定义 3).

按照物理的观点, 随机过程最重要的概率特征是其有限维分布函数族 $\{F_{t_1, \cdots, t_n}(x_1, \cdots, x_n)\}$:

$$F_{t_1, \cdots, t_n}(x_1, \cdots, x_n) = \mathbf{P}\{\omega : \xi_{t_1} \leqslant x_1, \cdots, \xi_{t_n} \leqslant x_n\}, \tag{1}$$

其中对于一切数组 $t_1, \cdots, t_n, t_1 < \cdots < t_n$, $F_{t_1, \cdots, t_n}(x_1, \cdots, x_n)$ 是给定的.

由 (1) 式可见, 对于一切数组 $t_1, \cdots, t_n (t_1 < \cdots < t_n)$, $F_{t_1, \cdots, t_n}(x_1, \cdots, x_n)$ 是 n 元分布函数 (见 §3 定义 2). 而且分布函数族 $\{F_{t_1, \cdots, t_n}(x_1, \cdots, x_n)\}$ 满足如下一致性条件 (见 §3 (20) 式):

$$\begin{aligned} &F_{t_1, \cdots, t_k, \cdots, t_n}(x_1, \cdots, \infty, \cdots, x_n) \\ &= F_{t_1, \cdots, t_{k-1}, t_{k+1}, \cdots, t_n}(x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_n). \end{aligned} \tag{2}$$

现在, 自然地提出这样的问题: 分布函数 $F_{t_1, \cdots, t_n}(x_1, \cdots, x_n)$ 族 $\{F_{t_1, \cdots, t_n}(x_1, \cdots, x_n)\}$ (见 §3 定义 2), 在什么条件下可以做某一随机过程的有限维分布函数族? 非常出色的是, 一致性条件 (2) 可以穷尽一切附加条件.

定理 1 (柯尔莫戈洛夫关于过程存在定理性) 设 $\{F_{t_1,\cdots,t_n}(x_1,\cdots,x_n)\}$ 是给定的满足一致性条件 (2) 的有限维分布函数族，其中 $t_i \in T \subseteq \mathbf{R}, n \geqslant 1, t_1 < t_2 < \cdots < t_n$. 那么，在概率空间 $(\Omega,\mathscr{F},\mathbf{P})$ 上存在随机过程 $X=(\xi_t)_{t\in T}$，使

$$\mathbf{P}\{\omega:\xi_{t_1}\leqslant x_1,\cdots,\xi_{t_n}\leqslant x_n\}=F_{t_1,\cdots,t_n}(x_1,\cdots,x_n). \tag{3}$$

证明 设

$$\Omega=\mathbf{R}^T,\quad \mathscr{F}=\mathscr{B}(\mathbf{R}^T).$$

即作为空间 Ω 取实函数的空间 $\omega=(\omega_t)_{t\in T}$，而 $\mathscr{F}$ 取由柱集生成的 σ 代数.

设 $\tau=[t_1,t_2,\cdots,t_n], t_1<t_2<\cdots<t_n$. 那么，根据 §3 定理 2，在空间 $(\mathbf{R}^n,\mathscr{B}(\mathbf{R}^n))$ 中可以建立 (并且是唯一的) 概率测度 P_τ，使

$$P_\tau\{(\omega_{t_1},\cdots,\omega_{t_n}):\omega_{t_1}\leqslant x_1,\cdots,\omega_{t_n}\leqslant x_n\}=F_{t_1,\cdots,t_n}(x_1,\cdots,x_n). \tag{4}$$

由一致性条件 (2) 可见，概率测度 $\{P_\tau\}$ 族也满足一致性条件 (见 §3 的 (20) 式). 根据 §3 的定理 4，在空间 $(\mathbf{R}^T,\mathscr{B}(\mathbf{R}^T))$ 上存在概率测度 $\mathbf{P}$，使对于 $\tau=[t_1,\cdots,t_n], t_1<\cdots<t_n$,

$$\mathbf{P}\{\omega:(\omega_{t_1},\cdots,\omega_{t_n})\in B\}=P_\tau(B).$$

由此亦可见，条件 (4) 成立. 因此，作为欲求的随机过程 $X=(\xi_t(\omega))_{t\in T}$，可以取如下的随机过程:

$$\xi_t(\omega)=\omega_t,\quad t\in T. \tag{5}$$

□

注 1 所建立的概率空间 $(\mathbf{R}^T,\mathscr{B}(\mathbf{R}^T),\mathbf{P})$ 常称做标准的，而由 (5) 式表示随机过程方法常称为建立过程的坐标方法.

注 2 设 $(E_\alpha,\mathscr{E}_\alpha)$ 是完全可分度量空间，而 α 属于任意下标的集合 $\mathfrak{A}$. 设 $\{P_\tau\}$ 是 $(E_{\alpha_1}\times\cdots\times E_{\alpha_n},\mathscr{E}_{\alpha_1}\otimes\cdots\otimes\mathscr{E}_{\alpha_n})$ 上的有限维分布函数族: $P_\tau, \tau=[\alpha_1,\cdots,\alpha_n]$. 那么，存在概率空间 $(\Omega,\mathscr{F},\mathbf{P})$ 和 $\mathscr{F}/\mathscr{E}_\alpha$- 可测函数族 $(X_\alpha(\omega))_{\alpha\in\mathfrak{A}}$，使对于任意 $\tau=[\alpha_1,\cdots,\alpha_n]$ 和 $B\in\mathscr{E}_{\alpha_1}\otimes\cdots\otimes\mathscr{E}_{\alpha_n}$，有

$$\mathbf{P}\{(X_{\alpha_1},\cdots,X_{\alpha_n})\in B\}=P_\tau(B).$$

如果对于每一个 $\omega=(\omega_\alpha), \alpha\in\mathfrak{A}$，设 $\Omega=\prod_\alpha E_\alpha, \mathscr{F}=\prod_\alpha\mathscr{E}_\alpha$ 和 $X_\alpha(\omega)=\omega_\alpha$，则这一结果由 §3 的定理 4 得到，而它推广了定理 1 的命题.

系 1 设 $F_1(x), F_2(x),\cdots$ 是一元分布函数序列. 那么，存在概率空间 $(\Omega,\mathscr{F},\mathbf{P})$ 和独立随机变量序列 $\xi_1,\xi_2,\cdots$，使

$$\mathbf{P}\{\omega:\xi_i(\omega)\leqslant x\}=F_i(x). \tag{6}$$

特别，存在概率空间 $(\Omega,\mathscr{F},\mathbf{P})$，在 $(\Omega,\mathscr{F},\mathbf{P})$ 上定义了无限伯努利随机变量序列 (参见第一章 §5 第 2 小节). 这里，作为 Ω 可以取空间:

$$\Omega=\{\omega:\omega=(a_1,a_2,\cdots), a_i=0,1\}$$

(参见定理 2).

为证明该系，只需设 $F_{1,\cdots,n}(x_1,\cdots,x_n)=F_1(x_1)\cdots F_n(x_n)$，并且应用定理 1.

系 2 设 $T=[0,\infty)$，而对于 $s,t\in T, t>s, x\in\mathbf{R}, B\in\mathscr{B}(\mathbf{R}), \{P(s,x;t,B)\}$ 是非负函数族，并且满足条件:

a) 对于固定的 s,x,t, $P(s,x;t,B)$ 是 B 的概率测度;

b) 对于固定的 s,t 和 B, $P(s,x;t,B)$ 是 x 的博雷尔函数;

c) 对于一切 $0\leqslant s<t<\tau$ 和 $B\in\mathscr{B}(\mathbf{R})$，满足柯尔莫戈洛夫–查普曼方程:

$$P(s,x;\tau,B)=\int_{\mathbf{R}}P(s,x;t,\mathrm{d}y)P(t,y;\tau,B). \tag{7}$$

此外，假设 $\pi=\pi(\cdot)$ 是 $(\mathbf{R},\mathscr{B}(\mathbf{R}))$ 上的概率测度.

那么，存在概率空间 $(\Omega,\mathscr{F},\mathbf{P})$ 和随机过程 $X=(\xi_t)_{t\geqslant0}$，使对于 $0=t_0<t_1<\cdots<t_n$，有

$$\begin{aligned}&\mathbf{P}\{\xi_{t_0}\leqslant x_0,\xi_{t_1}\leqslant x_1,\cdots,\xi_{t_n}\leqslant x_n\}\\&=\int_{-\infty}^{x_0}\pi(\mathrm{d}y_0)\int_{-\infty}^{x_1}P(0,y_0;t_1,\mathrm{d}y_1)\cdots\int_{-\infty}^{x_n}P(t_{n-1},y_{n-1};t_n,\mathrm{d}y_n).\end{aligned} \tag{8}$$

这样建立的过程 X 称做马尔可夫过程，π 称做其初始分布，而 $\{P(s,x;t,B)\}$ 称做过程的转移概率族.

系 3 设 $T=\{0,1,2,\cdots\}$，而对于 $k\geqslant1, x\in\mathbf{R}, B\in\mathscr{B}(\mathbf{R})$, $\{P_k(x;B)\}$ 是非负函数族，并且 (对于固定的 k 和 x) $P_k(x;B)$ 是 B 的概率测度，而 (对于固定的 k 和 B) $P_k(x;B)$ 是 x 的可测函数. 此外，假设 $\pi=\pi(\cdot)$ 是 $(\mathbf{R},\mathscr{B}(\mathbf{R}))$ 上的概率测度.

那么，可以建立概率空间 $(\Omega,\mathscr{F},\mathbf{P})$，使对于 $(\Omega,\mathscr{F},\mathbf{P})$ 上的随机变量族 $X=(\xi_0,\xi_1,\cdots)$，有

$$\begin{aligned}&\mathbf{P}\{\xi_0\leqslant x_0,\xi_1\leqslant x_1,\cdots,\xi_n\leqslant x_n\}\\&=\int_{-\infty}^{x_0}\pi(\mathrm{d}y_0)\int_{-\infty}^{x_1}P_1(y_0;\mathrm{d}y_1)\cdots\int_{-\infty}^{x_n}P_n(y_{n-1};\mathrm{d}y_n).\end{aligned}$$

3. 测度的开拓和随机序列的存在性 由系 1 知，存在分布函数相应为 $F_1,F_2,\cdots$ 的独立随机变量 $\xi_1,\xi_2,\cdots$ 序列.

现在设 $(E_1,\mathscr{E}_1),(E_2,\mathscr{E}_2),\cdots$ 是完全可分度量空间，而 $P_1,P_2,\cdots$ 是在这些空间上的概率测度. 那么，由注 2 可见，存在概率空间 $(\Omega,\mathscr{F},\mathbf{P})$ 和独立随机元序列 $X_1,X_2,\cdots$，使 X_n 为 $\mathscr{F}/\mathscr{E}_n$- 可测，且 $\mathbf{P}\{X_n\in B\}=P_n(B), B\in\mathscr{E}_n$. 实际上，当 $(E_n,\mathscr{E}_n)$ 是任意可测空间时，这一结果仍然成立.

定理 2 (图尔恰 [I. Tulcea] 关于测度的开拓和随机序列的存在性)　设 $(\Omega_n, \mathscr{F}_n)$, $n=1,2,\cdots$ 是任意可测空间, $\Omega=\prod \Omega_n$, $\mathscr{F}=\prod \mathscr{F}_n$. 假设在 $(\Omega_1,\mathscr{F}_1)$ 上给定概率测度 P_1, 而对于 $(\omega_1,\cdots,\omega_n)\in\Omega_1\times\cdots\times\Omega_n$, $n\geqslant 1$, 在 $(\Omega_{n+1},\mathscr{F}_{n+1})$ 上给定概率测度 $P(\omega_1,\cdots,\omega_n;\cdot)$; 并且对于每一个 $B\in\mathscr{F}_{n+1}$, $P(\omega_1,\cdots,\omega_n;B)$ 是 $(\omega_1,\cdots,\omega_n)$ 的可测函数. 对于 $A_i\in\mathscr{F}_i$, $n\geqslant 1$, 设

$$P_n(A_1\times\cdots\times A_n)=\int_{A_1}P_1(d\omega_1)\int_{A_2}P(\omega_1;d\omega_2)\cdots\int_{A_n}P(\omega_1,\cdots,\omega_{n-1};d\omega_n). \tag{9}$$

那么, 在 $(\Omega,\mathscr{F})$ 上存在唯一概率测度 $\mathbf{P}$, 使对于任意 $n\geqslant 1$,

$$\mathbf{P}\{\omega:\omega_1\in A_1,\cdots,\omega_n\in A_n\}=P_n(A_1\times\cdots\times A_n), \tag{10}$$

并存在随机变量序列 $X=(X_1(\omega),X_2(\omega),\cdots)$, 使对于 $A_i\in\mathscr{F}_i$,

$$\mathbf{P}\{\omega:X_1(\omega)\in A_1,\cdots,X_n(\omega)\in A_n\}=P_n(A_1\times\cdots\times A_n). \tag{11}$$

证明　1) 证明的第一步: 对于每一个 $n>1$, 证明由 (9) 式在 "矩形" $A_1\times\cdots\times A_n$ 上给定的集函数 P_n, 可以开拓到 σ-代数 $\mathscr{F}_1\otimes\cdots\otimes\mathscr{F}_n$ 上.

为此, 对于每一个 $n\geqslant 2$ 和 $B\in\mathscr{F}_1\otimes\cdots\otimes\mathscr{F}_n$, 设

$$\begin{aligned}P_n(B)=&\int_{\Omega_1}P_1(d\omega_1)\int_{\Omega_2}P(\omega_1;d\omega_2)\cdots\int_{\Omega_{n-1}}P(\omega_1,\cdots,\omega_{n-2};d\omega_{n-1})\\&\times\int_{\Omega_n}I_B(\omega_1,\cdots,\omega_{n-1},\omega_n)P(\omega_1,\cdots,\omega_{n-1};d\omega_n).\end{aligned} \tag{12}$$

易见, 对于 $B=A_1\times\cdots\times A_n$, (12) 式的右侧与 (9) 式的右侧相同. 此外, 对于 $n=2$, 如同 §6 的定理 8, P_2 是测度. 于是, 利用归纳法容易证明, 对于每一个 $n\geqslant 2$, P_n 是测度.

2) 证明的第二步: 与关于在 $(R^\infty,\mathscr{B}(R^\infty))$ 中测度开拓的柯尔莫戈洛夫定理 (§3 定理 3) 相同. 具体地说, 对于任意柱集 $J_n(B)=\{\omega\in\Omega:(\omega_1,\cdots,\omega_n)\in B\}$, $B\in\mathscr{F}_1\otimes\cdots\otimes\mathscr{F}_n$, 利用等式

$$\mathbf{P}(J_n(B))=P_n(B) \tag{13}$$

定义集函数 $\mathbf{P}$. 由 (12) 式, 以及 $P(\omega_1,\cdots,\omega_k;\cdot)$ 是测度, 容易证明定义 (13) 具有一致性: $\mathbf{P}(J_n(B))$ 的值不依赖于 (13) 式中柱集的表示方法.

由 (13) 式在柱集上定义的集函数 $\mathbf{P}$, 显然也是定义在包含所有柱集的代数上的集函数 $\mathbf{P}$. 由此可见, 集函数 $\mathbf{P}$ 在此代数上是有限可加测度. 只剩下验证集函数 $\mathbf{P}$ 在此代数上的可数可加性, 然后应用卡拉泰奥多里定理.

在 §3 的定理 3 进行过上面提到的证明: 证明基于空间 $(R^n,\mathscr{B}(R^n))$ 的如下性质: 对于每一个博雷尔集 B, 存在测度任意接近 B 的测度的紧统 $A\subseteq B$. 对于现在的情形, 这一方法有如下形式变化.

像 §3 的定理 3 一样, 设 $\{\hat{B}_n\}_{n\geqslant 1}$ 是递减为空集 $\varnothing$ 的柱集序列,

$$\hat{B}_n=\{\omega:(\omega_1,\cdots,\omega_n)\in B_n\}$$

但是

$$\lim_{n\to\infty}\mathbf{P}(\hat{B}_n)>0. \tag{14}$$

由 (12) 式可见, 对于 $n>1$, 有

$$\mathbf{P}(\hat{B}_n)=\int_{\Omega_1}f_n^{(1)}(\omega_1)P_1(d\omega_1),$$

其中

$$f_n^{(1)}(\omega_1)=\int_{\Omega_2}P(\omega_1;d\omega_2)\cdots\int_{\Omega_n}I_{B_n}(\omega_1,\cdots,\omega_{n-1},\omega_n)P(\omega_1,\cdots,\omega_{n-1};d\omega_n).$$

由于 $\hat{B}_{n+1}\subseteq\hat{B}_n$, 则 $B_{n+1}\subseteq B_n\times\Omega_{n+1}$. 因而 $I_{B_{n+1}}(\omega_1,\cdots,\omega_{n+1})\leqslant I_{B_n}(\omega_1,\cdots,\omega_n)I_{\Omega_{n+1}}(\omega_{n+1})$. 因此, 函数序列 $\{f_n^{(1)}(\omega_1)\}_{n\geqslant 1}$ 是减小的. 设

$$f^{(1)}(\omega_1)=\lim_n f_n^{(1)}(\omega_1).$$

那么, 根据控制收敛定理

$$\lim_n\mathbf{P}(\hat{B}_n)=\lim_n\int_{\Omega_1}f_n^{(1)}(\omega_1)P_1(d\omega_1)=\int_{\Omega_1}f^{(1)}(\omega_1)P_1(d\omega_1).$$

根据假设, $\lim_n\mathbf{P}(\hat{B}_n)>0$. 由此可见, 由于若 $\omega_1\notin B_1$, 则对于所有 $n\geqslant 1$, $f_n^{(1)}(\omega_1)=0$, 故存在 $\omega_1^0\in B_1$, 使 $f^{(1)}(\omega_1^0)>0$.

其次, 对于 $n>2$,

$$f_n^{(1)}(\omega_1^0)=\int_{\Omega_2}f_n^{(2)}(\omega_2)P(\omega_1^0;d\omega_2), \tag{15}$$

其中

$$\begin{aligned}f_n^{(2)}(\omega_2)=&\int_{\Omega_3}P(\omega_1^0,\omega_2;d\omega_3)\cdots\\&\cdots\int_{\Omega_n}I_{B_n}(\omega_1^0,\omega_2,\cdots,\omega_n)P(\omega_1^0,\omega_2,\cdots,\omega_{n-1};d\omega_n).\end{aligned}$$

同序列的 $\{f_n^{(1)}(\omega_1)\}$ 的情形一样, 证明序列 $\{f_n^{(2)}(\omega_2)\}$ 是减少的. 设

$$f^{(2)}(\omega_2)=\lim_{n\to\infty}f_n^{(2)}(\omega_2).$$

那么, 由 (15) 式, 可见

$$0<f^{(1)}(\omega_1^0)=\int_{\Omega_2}f^{(2)}(\omega_2)P(\omega_1^0;d\omega_2),$$

且存在点 $\omega_2^0 \in \Omega_2$, 使 $f^{(2)}(\omega_2^0) > 0$. 这时 $(\omega_1^0, \omega_2^0) \in B_2$. 继续这一过程, 对于任意 $n \geqslant 1$, 存在点 $(\omega_1^0, \cdots, \omega_n^0) \in B_n$. 从而, 点 $(\omega_1^0, \cdots, \omega_n^0, \cdots) \in \bigcap \hat{B}_n$. 但是, 根据假设 $\bigcap \hat{B}_n = \varnothing$. 出现的矛盾说明 $\lim_n \mathbf{P}(\hat{B}_n) = 0$.

于是, 定理涉及概率测度 $\mathbf{P}$ 存在性的部分得证.

3) 定理的结尾部分显然可以由上面的结果得到. 为此只需要设 $X_n(\omega) = \omega_n$, $n \geqslant 1$. □

系 4　*设 $(E_n, \mathscr{E}_n)_{n \geqslant 1}$ 是任意可测空间, $(P_n)_{n \geqslant 1}$ 是该空间上的概率测度. 那么, 存在概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 和相应地取值于可测空间 $(E_1, \mathscr{E}_1), (E_2, \mathscr{E}_2), \cdots$ 的独立随机元族 $X_1, X_2, \cdots$, 并且*

$$\mathbf{P}\{\omega : X_n(\omega) \in B\} = P_n(B), \quad B \in \mathscr{E}_n, \ n \geqslant 1.$$

系 5　*设 $E = \{1, 2, \cdots\}$, $\{p_k(x, y)\}$, $k \geqslant 1$, $x, y \in E$ 是非负函数族, 并且满足*

$$\sum_{y \in E} p_k(x, y) = 1, x \in E, k \geqslant 1.$$

此外, 设 $\pi = \pi(\cdot)$ 是 E 上的概率分布 $(\pi(x) \geqslant 0, \sum_{x \in E} \pi(x) = 1)$.

那么, 存在概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$, 且在此空间上存在随机变量族 $X = \{\xi_0, \xi_1, \cdots\}$, 使对于一切 $x_i \in E$, $n \geqslant 1$, 有

$$\mathbf{P}\{\xi_0 = x_0, \xi_1 = x_1, \cdots, \xi_n = x_n\} = \pi(x_0) p_1(x_0, x_1) \cdots p_n(x_{n-1}, x_n) \qquad (16)$$

(对照第一章 §12 之 (4) 式). 作为空间 Ω 可以取

$$\Omega = \{\omega : \omega = (x_0, x_1, \cdots), x_i \in E\}.$$

满足 (16) 式的随机变量序列 $X = \{\xi_0, \xi_1, \cdots\}$, 称做具有可数状态集 E 的马尔可夫链, $\{p_k(x, y)\}$ 称为其转移概率矩阵. 而称为其初始概率分布 π. (对照第一章 §12 的定义, 以及第八章 §1 的定义.)

4. **更新过程**　柯尔莫戈洛夫定理 (定理 1) 指出了, 具有给定有限维分布函数的过程的存在性. 这时, 定理的证明用到典则概率空间, 而构造过程使用坐标式方法. 这本身就说明了过程构造的复杂性.

按这种观点, 在尽量少地使用 "概率结构" 的情况下, 构建具有给定性质的随机过程的情形会引起人们很大兴趣.

为演示这种可能性, 我们现在考虑所谓*更新过程* (其特殊情形是泊松过程; 见第二章 §10).

设 $(\sigma_1, \sigma_2, \cdots)$ 是独立同分布随机变量序列, 其共同的分布函数为 $F = F(x)$. (定理 1 的系 1 保障这样随机变量序列的存在.)

由 $(\sigma_1, \sigma_2, \cdots)$ 序列形成新的序列 $(T_0, T_1, \cdots)$, 其中 $T_0 = 0$, 且

$$T_n = \sigma_1 + \cdots + \sigma_n, n \geqslant 1.$$

为直观计, 我们把 T 视为时间; T_n 表示第 n 个质点出现的时间 (例如, 电话的第 n 次呼唤). 那么, σ_n 描绘第 $(n-1)$ 到第 n 个质点 (第 n 次呼唤) 持续的时间.

习惯上把随机过程 $N = (N_t)_{t \geqslant 0}$, 其中 N_t 是 (构造性地给定的) 集

$$N_t = \sum_{n=1}^{\infty} I(T_n \leqslant t), \qquad (17)$$

称做*更新过程*.

显然, N_t 亦可定义为

$$N_t = \max\{n : T_n \leqslant t\}, \qquad (18)$$

即 N_t 是出现在区间 $(0, t]$ 上的质点 (呼唤) 数. 这时显然

$$\{N_t \geqslant n\} = \{T_n \leqslant t\}. \qquad (19)$$

这一简单的公式十分有用, 因为它可以把对随机过程 $N = (N_t)_{t \geqslant 0}$ 概率性质的研究, 归结为对独立随机变量 $\sigma_1, \cdots, \sigma_n (n \geqslant 1)$ 之和 $T_n = \sigma_1 + \cdots + \sigma_n$ 的性质的研究 (参见第四章 §3 第 4 小节, 以及第七章 §2 第 4 小节).

由 (17) 式立即可见, *更新函数* $m(t) = \mathbf{E}N_t (t \geqslant 0)$ 决定于分布函数 $F_n(t) = \mathbf{P}\{T_n \leqslant t\}$ 如下:

$$m(t) = \sum_{n=1}^{\infty} F_n(t). \qquad (20)$$

5. **练习题**

1. 设 $\Omega = [0, 1]$, $\mathscr{F}$ 是 $[0, 1]$ 上的博雷尔集类, $\mathbf{P}$ 是 $[0, 1]$ 上的勒贝格测度. 我们称空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 为 "通用的", 如果对于 $(\Omega, \mathscr{F}, \mathbf{P})$ 上的任意分布函数 $F(x)$, 可以定义一随机变量 $\xi = \xi(\omega)$, 使其分布函数 $F_\xi(x) = \mathbf{P}\{\xi \leqslant x\}$ 恰好是 $F(x)$. 证明空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 为 "通用的". (提示: $\xi(\omega) = F^{-1}(\omega)$, 其中 $F^{-1}(\omega) = \sup\{x : F(x) < \omega\}$, $0 < \omega < 1$, 而 $\xi(0), \xi(1)$ 可以是任意的.)

2. 验证定理 1 和 2 的系中分布族的一致性.

3. 由定理 1 导出定理 2 的系 2 的命题.

4. 设 F_n 是随机变量 $T_n (n \geqslant 1)$ 的分布函数 (第 4 小节). 证明

$$F_{n+1}(t) = \int_0^t F_n(t - s) dF(s), n \geqslant 1, \text{其中 } F_1 = F.$$

5. 证明 $\mathbf{P}\{N_t = n\} = F_n(t) - F_{n+1}(t)$ (见 (17) 式).

6. 证明, 在上面第 4 小节引进的更新函数 $m(t)$ 满足*更新方程*:

$$m(t) = F(t) + \int_0^t m(t - x) dF(x). \qquad (21)$$

7. 证明，在有限区间上有界的函数类中，由 (20) 式定义的函数是方程 (21) 的唯一解.

8. 设 T 是任意集合.

(i) 假设对于每个 $t \in T$，给定一概率空间 $(\Omega_t, \mathscr{F}_t, \mathbf{P}_t)$. 记

$$\Omega = \prod_{t \in T} \Omega_t, \quad \mathscr{F} = \prod_{t \in T} \mathscr{F}_t.$$

证明，在 $(\Omega, \mathscr{F})$ 上存在唯一概率测度 $\mathbf{P}$，使

$$\mathbf{P}\left(\prod_{t \in T} B_t\right) = \prod_{t \in T} \mathbf{P}(B_t),$$

其中对于除有限个之外的一切下标 t，$B_t \in \mathscr{F}_t, t \in T, B_t = \Omega_t$. (提示：这里 $\mathbf{P}$ 在相应的代数上，并利用定理 2 的证明方法.)

(ii) 假设对于每个 $t \in T$，给定可测空间 $(E_t, \mathscr{E}_t)$ 和该空间上的概率测度 $\mathbf{P}_t$. 证明存在概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 及独立随机元素 $(X_t)_{t \in T}$，使 X_t 为 $\mathscr{F}/\mathscr{E}_t$- 可测的，且 $\mathbf{P}\{X_t \in B\} = \mathbf{P}_t(B), B \in \mathscr{E}_t$.

§10.　随机变量序列收敛的各种形式

1. 收敛性的基本类型　同数学分析一样，在概率论中也需要考虑随机变量不同形式的收敛性. 现在，讨论如下形式的收敛性：*依概率收敛，依概率 1 收敛，p 阶平均收敛，依分布收敛.*

从定义开始叙述. 设 $\xi, \xi_1, \xi_2, \cdots$ 是某一概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 上的随机变量.

定义 1　随机变量序列 $\xi_1, \xi_2, \cdots$ (亦记作 (ξ_n)，$(\xi_n)_{n \geqslant 1}$ 或 $\{\xi_n\}$，$\{\xi_n\}_{n \geqslant 1}$)，称做*依概率收敛于随机变量 ξ 的* (记作 $\xi_n \xrightarrow{\mathbf{P}} \xi$)，如果对于任意 $\varepsilon > 0$，有

$$\mathbf{P}\{|\xi_n - \xi| > \varepsilon\} \to 0, \quad n \to \infty. \tag{1}$$

我们已经遇到过这种形式的收敛性，即伯努利大数定律：

$$\mathbf{P}\left\{\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right\} \to 0, \quad n \to \infty$$

(记号见第一章 §5). 在数学分析中这种形式的收敛性，称做*依测度收敛*.

定义 2　随机变量序列 $\xi_1, \xi_2, \cdots$，称做*依概率 1 (几乎必然或几乎处处) 收敛于随机变量 ξ*，如果

$$\mathbf{P}\{\omega : \xi_n \nrightarrow \xi\} = 0, \quad n \to \infty, \tag{2}$$

即如果使 $\xi_n(\omega)$ 收敛于 $\xi(\omega)$ 的结局 ω 的集合的概率等于 1.

这种形式 (依概率 1) 的收敛性，有多种记号：

$$\xi_n \to \xi(\mathbf{P}\text{-a.c.}), \text{ 或 } \xi_n \to \xi(\text{a.c.}), \text{ 或 } \xi_n \xrightarrow{\text{a.c.}} \xi, \text{ 或 } \xi_n \xrightarrow{\text{a.s.}} \xi.$$

定义 3　随机变量序列 $\xi_1, \xi_2, \cdots$，称做 *p 阶平均收敛于随机变量 ξ*，如果

$$\mathbf{E}|\xi_n - \xi|^p \to 0, \quad n \to \infty. \tag{3}$$

在数学分析中这种形式的收敛性称做 L^p- 收敛. 因此，(3) 式可写为 $\xi_n \xrightarrow{L^p} \xi$. 特别，若 $p = 2$，则这种收敛称做*平方平均收敛 (或均方收敛)*，记作 $\xi = \text{l.i.m.}\,\xi_n$ (l.i.m. 是 "limit in mean"——平均极限的缩写).

定义 4　随机变量序列 $\xi_1, \xi_2, \cdots$，称*按分布 (或依分布) 收敛于随机变量 ξ*，如果对于任意有界连续函数 $f = f(x)$，有

$$\mathbf{E}f(\xi_n) \to \mathbf{E}f(\xi), \quad n \to \infty. \tag{4}$$

记作：$\xi_n \xrightarrow{d} \xi$，$\xi_n \xrightarrow{\text{law}} \xi$ 其中 d 是 distribution (分布) 的字头，law 是 (分布) 律.

这一收敛性名称的来历，像将由第三章 §1 表明的那样，条件 (4) 等价于：分布函数 $F_{\xi_n}(x)$ 在分布函数 $F_\xi(x)$ 的每一个连续点上收敛于 $F_\xi(x)$ (称做 $F_{\xi_n}(x)$ *基本收敛于* $F_\xi(x)$，记作 $F_{\xi_n} \Rightarrow F_\xi$).

应该强调，随机的按分布收敛性，只有通过其分布函数的收敛性定义. 因此，关于这种形式的收敛性，当且仅当随机变量定义在不同的概率空间上才有意义. 这种形式的收敛性，在第三章将详细研究，到时要特别说明，为什么在收敛性 $F_{\xi_n} \Rightarrow F_\xi$ 的定义中，仅要求在 $F_\xi(x)$ 的连续点上收敛，而不是在一切 x 点上.

2. 基本随机变量序列的概念　在数学分析中，在解决给定函数序列 (在一定意义下) 收敛性的问题时，基本序列 (或柯西序列) 的概念是非常重要的. 对于现在所研究的随机变量序列收敛性的情形，也引进类似的概念.

称随机变量序列 $\{\xi_n\}_{n \geqslant 1}$ (或简记为 $\{\xi_n\}$) *为依概率为基本的*，如果满足条件：对于任意 $\varepsilon > 0$，有 $\mathbf{P}\{|\xi_n - \xi_m| \geqslant \varepsilon\} \to 0, n, m \to \infty$；称 $\{\xi_n\}_{n \geqslant 1}$ *为依概率 1 为基本的*，若对于几乎一切 $\omega \in \Omega$，$\{\xi_n(\omega)\}_{n \geqslant 1}$ 为基本序列；称 $\{\xi_n(\omega)\}_{n \geqslant 1}$ *在 L^p (或 $p(0 < p < \infty)$ 阶平均) 的意义上为基本的*，若 $\mathbf{E}|\xi_n - \xi_m|^p \to 0, n, m \to \infty, 0 < p < \infty$.

3. 随机变量序列的依概率 1 收敛

定理 1　a) *使 $\xi_n \to \xi(\mathbf{P}\text{-a.c.})$ 的充分和必要条件是，对于任意 $\varepsilon > 0$，*

$$\mathbf{P}\left\{\sup_{k \geqslant n} |\xi_k - \xi| \geqslant \varepsilon\right\} \to 0, \quad n \to \infty. \tag{5}$$

b) *随机变量序列 $\{\xi_n\}_{n \geqslant 1}$ 依概率 1 为基本的，当且仅当对于任意 $\varepsilon > 0$，*

$$\mathbf{P}\left\{\sup_{k \geqslant n, l \geqslant n} |\xi_k - \xi_l| \geqslant \varepsilon\right\} \to 0, \quad n \to \infty, \tag{6}$$

或等价地

$$\mathbf{P}\left\{\sup_{k\ge 0}|\xi_{n+k}-\xi_n|\ge\varepsilon\right\}\to 0,\quad n\to\infty. \tag{7}$$

证明 a) 设 $A_k^\varepsilon=\{\omega:|\xi_k-\xi|\ge\varepsilon\}$, $A^\varepsilon=\overline{\lim}A_n^\varepsilon\equiv\bigcap_{n=1}^\infty\bigcup_{k\ge n}A_k^\varepsilon$. 那么,

$$\{\xi_n\nrightarrow\xi\}=\bigcup_{\varepsilon\ge 0}A^\varepsilon=\bigcup_{m=1}^\infty A^{1/m}.$$

因为

$$\mathbf{P}(A^\varepsilon)=\lim_n\mathbf{P}\left(\bigcup_{k\ge n}A_k^\varepsilon\right),$$

所以命题 a) 的结论由下面一系列等价关系得出:

$$\mathbf P\{\omega:\xi_n\nrightarrow\xi\}=0\Leftrightarrow \mathbf P\left(\bigcup_{\varepsilon>0}A^\varepsilon\right)=0\Leftrightarrow\mathbf P\left(\bigcup_{m=1}^\infty A^{1/m}\right)=0\Leftrightarrow$$

$$\Leftrightarrow \mathbf P(A^{1/m})=0,\ m\ge1\Leftrightarrow \mathbf P(A^\varepsilon)=0,\ \varepsilon>0\Leftrightarrow$$

$$\Leftrightarrow\mathbf P\left(\bigcup_{k\ge n}A_k^\varepsilon\right)\to0,\ n\to\infty,\varepsilon>0\Leftrightarrow\mathbf P\left\{\sup_{k\ge n}|\xi_k-\xi|\ge\varepsilon\right\}\to0,\ n\to\infty,\varepsilon>0.$$

b) 记

$$B_{k,l}^\varepsilon=\{\omega:|\xi_k-\xi_l|\ge\varepsilon\},\quad B^\varepsilon=\bigcap_{n=1}^\infty\bigcup_{\substack{k\ge n\\ l\ge n}}B_{k,l}^\varepsilon.$$

那么, $\{\omega:\{\xi_n(\omega)\}$ 不是基本的$\}=\bigcup_{\varepsilon>0}B^\varepsilon$, 像 a) 一样可以证明 $\mathbf P\{\omega:\{\xi_n(\omega)\}$ 不是基本的$\}=0$ 与 (6) 式等价. 而由下面明显的不等式, 可见 (6) 式与 (7) 式等价:

$$\sup_{k\ge0}|\xi_{n+k}-\xi_n|\le\sup_{k\ge0,\,l\ge0}|\xi_{n+k}-\xi_{n+l}|\le2\sup_{k\ge0}|\xi_{n+k}-\xi_n|.$$ □

系 由于

$$\mathbf P\left\{\sup_{k\ge n}|\xi_k-\xi|\ge\varepsilon\right\}=\mathbf P\left(\bigcup_{k\ge n}\{|\xi_k-\xi|\ge\varepsilon\}\right)\le\sum_{k\ge n}\mathbf P\{|\xi_k-\xi|\ge\varepsilon\},$$

可见 $\xi_n\to\xi\ (\mathbf P\text{-a.c.})$ 的充分条件是: 对于任意 $\varepsilon>0$, 有

$$\sum_{k=1}^\infty\mathbf P\{|\xi_k-\xi|\ge\varepsilon\}<\infty. \tag{8}$$

现在关于条件 (8) 应当指出, 推导该式的论断可以用来建立如下简单但十分重要的结果, 在研究依概率 1 成立的性质时是基本工具.

设 $A_1,A_2,\cdots$ 是 $\mathscr F$ 中的某一事件序列. 我们 (见 §1 表 2-1) 曾使用记号 {无限多个 A_n} 表示事件 $\overline{\lim}A_n$ 为 "在 $A_1,A_2,\cdots$ 中有无限多个出现".

引理 1 (博雷尔 - 坎泰利 [E. Borel-F. P. Cantelli]) a) 如果 $\sum\mathbf P(A_n)<\infty$, 则概率 $\mathbf P\{\text{无限多个}\ A_n\}=0$.

b) 如果 $\sum\mathbf P(A_n)=\infty$ 且事件 $A_1,A_2,\cdots$ 独立, 则概率 $\mathbf P\{\text{无限多个}\ A_n\}=1$.

证明 a) 根据定义

$$\{\text{无限多个}\ A_n\}=\overline{\lim}A_n=\bigcap_{n=1}^\infty\bigcup_{k\ge n}A_k.$$

因此,

$$\mathbf P\{\text{无限多个}\ A_n\}=\mathbf P\left(\bigcap_{n=1}^\infty\bigcup_{k\ge n}A_k\right)=\lim\mathbf P\left(\bigcup_{k\ge n}A_k\right)\le\sum_{k\ge n}\mathbf P(A_k),$$

由此命题 a) 得证.

b) 如果 $A_1,A_2,\cdots$ 独立, 则 $\overline A_1,\overline A_2,\cdots$ 也独立. 那么, 对于任意 $N\ge n$, 有

$$\mathbf P\left(\bigcap_{k=n}^N\overline A_k\right)=\prod_{k=n}^N\mathbf P(\overline A_k).$$

由此易见

$$\mathbf P\left(\bigcap_{k=n}^\infty\overline A_k\right)=\prod_{k=n}^\infty\mathbf P(\overline A_k). \tag{9}$$

由不等式 $\ln(1-x)\le -x,\ 0\le x<1$, 可得,

$$\ln\prod_{k=n}^\infty[1-\mathbf P(A_k)]=\sum_{k=n}^\infty\ln[1-\mathbf P(A_k)]\le-\sum_{k=n}^\infty\mathbf P(A_k)=-\infty.$$

从而, 对于任意 n,

$$\mathbf P\left(\bigcap_{k=n}^\infty\overline A_k\right)=0.$$

即 $\mathbf P\{\text{无限多个}\ A_n\}=1$. □

系 1 如果 $A_n^\varepsilon=\{\omega:|\xi_n-\xi|\ge\varepsilon\}$, 则条件 (8) 表示

$$\sum_{n=1}^\infty\mathbf P(A_n^\varepsilon)<\infty,\quad\varepsilon>0,$$

而根据博雷尔 - 坎泰利引理 $\mathbf P(A^\varepsilon)=0,\ \varepsilon>0$, 其中 $A^\varepsilon=\overline{\lim}A_n^\varepsilon(=\{\text{无限多个}\ A_n^\varepsilon\})$, 于是, 有

$$\sum_{k=1}^\infty\mathbf P\{|\xi_k-\xi|\ge\varepsilon\}<\infty,\ \varepsilon>0\Rightarrow\mathbf P(A^\varepsilon)=0,\ \varepsilon>0\Leftrightarrow\mathbf P\{\omega:\xi_n\nrightarrow\xi\}=0$$

而关于这一点上面已经指出.

系 2 设 $\{\varepsilon_n\}_{n\geqslant 1}$ 是正数序列, $\varepsilon_n \downarrow 0, n\to\infty$. 那么, 如果

$$\sum_{n=1}^{\infty} \mathbf{P}\{|\xi_n-\xi|\geqslant \varepsilon_n\}<\infty, \tag{10}$$

则 $\xi_n \xrightarrow{\text{a.s.}} \xi$.

事实上, 设 $A_n=\{|\xi_n-\xi|\geqslant\varepsilon_n\}$. 那么, 根据博雷尔 – 坎泰利引理 $\mathbf{P}\{$无限多个 $A_n\}=0$. 对于几乎每一个结局 $\omega\in\Omega$, 存在 $N=N(\omega)$, 使对于 $n\geqslant N(\omega)$, 有 $|\xi_n(\omega)-\xi(\omega)|<\varepsilon_n$. 由于 $\varepsilon_n\downarrow 0$, 可见对于几乎一切 $\omega\in\Omega$, $\xi_n(\omega)\to\xi(\omega)$.

4. 各种收敛性的蕴涵关系

定理 2 有如下蕴涵关系:

$$\xi_n \xrightarrow{\text{a.s.}} \xi \Rightarrow \xi_n \xrightarrow{\mathbf{P}} \xi, \tag{11}$$

$$\xi_n \xrightarrow{L^p} \xi \Rightarrow \xi_n \xrightarrow{\mathbf{P}} \xi, \quad p>0, \tag{12}$$

$$\xi_n \xrightarrow{\mathbf{P}} \xi \Rightarrow \xi_n \xrightarrow{d} \xi. \tag{13}$$

证明 由 (5) 式可见命题 (11) 成立; 由切比雪夫不等式, 可见蕴涵关系 (12) 成立.

现在证明蕴涵关系 (13). 设 $f(x)$ 是连续函数, $|f(x)|\leqslant c$, $\varepsilon>0$, 而 N 满足 $\mathbf{P}\{|\xi|>N\}\leqslant\varepsilon/(4c)$.

选择 $\delta>0$, 使得对于一切 $|x|\leqslant N$ 和 $|x-y|\leqslant\delta$, 不等式

$$|f(x)-f(y)|\leqslant\frac{\varepsilon}{2}$$

成立.

那么, (对照第一章 §5 第 5 小节, 维尔斯特拉斯定理的"概率的"证明):

$$\begin{aligned}\mathbf{E}|f(\xi_n)-f(\xi)| &= \mathbf{E}(|f(\xi_n)-f(\xi)|;|\xi_n-\xi|\leqslant\delta,|\xi|\leqslant N)\\ &\quad+\mathbf{E}(|f(\xi_n)-f(\xi)|;|\xi_n-\xi|\leqslant\delta,|\xi|>N)\\ &\quad+\mathbf{E}(|f(\xi_n)-f(\xi)|;|\xi_n-\xi|>\delta)\\ &\leqslant\varepsilon/2+\varepsilon/2+2c\mathbf{P}\{|\xi_n-\xi|>\delta\}=\varepsilon+2c\mathbf{P}\{|\xi_n-\xi|>\delta\}.\end{aligned}$$

由于 $\mathbf{P}\{|\xi_n-\xi|>\delta\}\to 0$, 可见对于充分大的 n, $\mathbf{E}|f(\xi_n)-f(\xi)|\leqslant 2\varepsilon$. 由于 $\varepsilon>0$ 是任意的, 可见蕴涵关系 (13) 得证. □

现在举一些例子, 其中有的说明 (11), (12) 式中相反的蕴涵关系一般不成立.

例 1 ($\xi_n \xrightarrow{\mathbf{P}} \xi \nRightarrow \xi_n \xrightarrow{\text{a.s.}} \xi$; $\xi_n \xrightarrow{L^p} \xi \nRightarrow \xi_n \xrightarrow{\text{a.s.}} \xi$). 设 $\Omega=[0,1]$, $\mathscr{F}=\mathscr{B}[0,1]$, $\mathbf{P}$ 是 $[0,1]$ 上的勒贝格测度. 设

$$A_n^i=\left[\frac{i-1}{n},\frac{i}{n}\right],\ \xi_n^i=I_{A_n^i}(\omega),\ i=1,2,\cdots,n;\ n\geqslant 1.$$

那么, 随机变量序列

$$\{\xi_1^1;\xi_2^1,\xi_2^2;\xi_3^1,\xi_3^2,\xi_3^3;\cdots\}$$

依概率收敛, 也对 $p>0$ 阶平均收敛, 但是它在任何点 $\omega\in[0,1]$ 都不收敛.

例 2 ($\xi_n \xrightarrow{\mathbf{P}} \xi \Leftarrow \xi_n \xrightarrow{\text{a.s.}} \xi \nRightarrow \xi_n \xrightarrow{L^p} \xi\ (p>0)$). 仍设 $\Omega=[0,1]$, $\mathscr{F}=\mathscr{B}[0,1]$, $\mathbf{P}$ 是 $[0,1]$ 上的勒贝格测度, 而

$$\xi_n(\omega)=\begin{cases} e^n, & 若\ 0\leqslant\omega\leqslant 1/n,\\ 0, & 若\ \omega>1/n.\end{cases}$$

那么, 序列 $\{\xi_n\}$ 依概率 1 (故依概率) 收敛于 0, 但是对于任何 $p>0$,

$$\mathbf{E}|\xi_n|^p=\frac{e^{pn}}{n}\to\infty,\quad n\to\infty.$$

例 3 ($\xi_n \xrightarrow{L^p} \xi \nRightarrow \xi_n \xrightarrow{\text{a.s.}} \xi$). 设 $\{\xi_n\}$ 是独立随机变量序列, 且

$$\mathbf{P}\{\xi_n=1\}=p_n,\quad \mathbf{P}\{\xi_n=0\}=1-p_n.$$

那么, 不难证明:

$$\xi_n \xrightarrow{\mathbf{P}} 0 \Leftrightarrow p_n\to 0,\quad n\to\infty; \tag{14}$$

$$\xi_n \xrightarrow{L^p} 0 \Leftrightarrow p_n\to 0,\quad n\to\infty; \tag{15}$$

$$\xi_n \xrightarrow{\text{a.s.}} 0 \Leftrightarrow \sum_{n=1}^{\infty}p_n<\infty. \tag{16}$$

特别, 当 $p_n=1/n$ 时, 对于任何 $p>0$, $\xi_n \xrightarrow{L^p} 0$, 然而 ξ_n 并不几乎处处收敛于 0.

下面的定理涉及一个重要的情形: 由几乎处处收敛可以导出 L^1 平均收敛.

定理 3 设 $\{\xi_n\}_{n\geqslant 1}$ 是非负随机变量序列, 且 $\xi_n \xrightarrow{\text{a.s.}} \xi$, $\mathbf{E}\xi_n\to\mathbf{E}\xi<\infty$. 那么,

$$\mathbf{E}|\xi_n-\xi|\to 0,\quad n\to\infty. \tag{17}$$

证明 对于充分大的 n, 有 $\mathbf{E}\xi_n<\infty$. 因此

$$\begin{aligned}\mathbf{E}|\xi-\xi_n| &= \mathbf{E}(\xi-\xi_n)I_{\{\xi\geqslant\xi_n\}}+\mathbf{E}(\xi_n-\xi)I_{\{\xi_n>\xi\}}\\ &= 2\mathbf{E}(\xi-\xi_n)I_{\{\xi\geqslant\xi_n\}}+\mathbf{E}(\xi_n-\xi).\end{aligned}$$

由于 $0\leqslant(\xi-\xi_n)I_{\{\xi\geqslant\xi_n\}}\leqslant\xi$, 则由控制收敛定理知 $\lim_n\mathbf{E}(\xi-\xi_n)I_{\{\xi\geqslant\xi_n\}}=0$, 于是由假设 $\mathbf{E}\xi_n\to\mathbf{E}\xi$ 证得 (17) 式. □

注 如果将依概率 1 收敛换成依概率收敛, 则控制收敛定理 (§6 定理 3) 仍然成立 (见练习题 1). 因此, 在定理 3 中收敛"$\xi_n \xrightarrow{\text{a.s.}} \xi$"可以换成收敛"$\xi_n \xrightarrow{\mathbf{P}} \xi$".

5. 柯西收敛准则 由数学分析知，任意基本数列 $\{x_n\}$, $x_n \in \mathbb{R}$，都是收敛的 (柯西准则). 我们对于随机变量序列引进类似的结果.

定理 4 (柯西几乎必然收敛准则) *随机变量序列 $\{\xi_n\}$ 依概率 1 收敛 (于某一随机变量 ξ) 的充分必要条件是 $\{\xi_n\}$ 依概率 1 为基本序列.*

证明 若 $\xi_n \xrightarrow{\text{a.s.}} \xi$，则

$$\sup_{k\geqslant n,\, l\geqslant n} |\xi_k - \xi_l| \leqslant \sup_{k\geqslant n} |\xi_k - \xi| + \sup_{l\geqslant n} |\xi_l - \xi|,$$

由此 (见定理 1) 可得定理的必要性.

现在假设 $\{\xi_n\}$ 依概率 1 为基本序列. 记 $\mathscr{N} = \{\omega : \{\xi_n(\omega)\}$ 不是基本序列$\}$. 那么，对于 $\omega \in \Omega\setminus\mathscr{N}$，数列 $\{\xi_n(\omega)\}$ 是基本的，故根据数列的柯西定理，数列 $\{\xi_n(\omega)\}$ 有极限 $\lim \xi_n(\omega)$. 设

$$\xi(\omega) = \begin{cases} \lim \xi_n(\omega), & 若\ \omega \in \Omega\setminus\mathscr{N}, \\ 0, & 若\ \omega \in \mathscr{N}. \end{cases} \tag{18}$$

于是，所定义的函数 $\xi(\omega)$ 是随机变量，因而显然 $\xi_n \xrightarrow{\text{a.s.}} \xi$. □

定理 5 *设 $\{\xi_n\}$ 对于依概率收敛是基本随机变量序列，则从中可以分离出依概率 1 收敛的子序列 $\{\xi_{n_k}\}$.*

证明 设 $\{\xi_n\}$ 对于依概率收敛是基本随机变量序列. 由于定理 4，只需证明从中可以分出几乎处处收敛的子序列.

设 $n_1 = 1$，由归纳法确定一个 n_k：使之对于任意 $s \geqslant n$, $t \geqslant n$，是满足

$$\mathrm{P}\{|\xi_t - \xi_s| > 2^{-k}\} < 2^{-k}$$

的最小 $n > n_{k-1}$. 那么，

$$\sum_k \mathrm{P}\{|\xi_{n_{k+1}} - \xi_{n_k}| > 2^{-k}\} < \sum_k 2^{-k} < \infty,$$

而由博雷尔–坎泰利引理

$$\mathrm{P}\{|\xi_{n_{k+1}} - \xi_{n_k}| > 2^{-k}，无限多个\ k\} = 0.$$

因此，依概率 1，有

$$\sum_k |\xi_{n_{k+1}} - \xi_{n_k}| < \infty.$$

设 $\mathscr{N} = \{\omega : \sum |\xi_{n_{k+1}} - \xi_{n_k}| = \infty\}$. 那么，如果设

$$\xi(\omega) = \begin{cases} \xi_{n_1}(\omega) + \displaystyle\sum_{k=1}^{\infty} [\xi_{n_{k+1}}(\omega) - \xi_{n_k}(\omega)], & 若\ \omega \in \Omega\setminus\mathscr{N}, \\ 0, & 若\ \omega \in \mathscr{N}, \end{cases}$$

则 $\xi_{n_k} \xrightarrow{\text{a.s.}} \xi$.

如果原序列依概率收敛，则它按依概率为基本的 (见下面的 (19) 式)，从而这正是所讨论的情形. □

定理 6 (柯西依概率收敛准则) *随机变量序列 $\{\xi_n\}$ 依概率收敛的充分必要条件是，$\{\xi_n\}$ 按依概率收敛是基本随机变量序列.*

证明 如果 $\xi_n \xrightarrow{\mathrm{P}} \xi$，则

$$\mathrm{P}\{|\xi_n - \xi_m| \geqslant \varepsilon\} \leqslant \mathrm{P}\{|\xi_n - \xi| \geqslant \varepsilon/2\} + \mathrm{P}\{|\xi_m - \xi| \geqslant \varepsilon/2\}, \tag{19}$$

即 $\{\xi_n\}$ 按依概率收敛是基本随机变量序列.

相反，若 $\{\xi_n\}$ 按依概率收敛是基本随机变量序列，则根据定理 5，存在随机变量子序列 $\{\xi_{n_k}\}$ 和随机变量 ξ，使 $\xi_{n_k} \xrightarrow{\text{a.s.}} \xi$. 那么，

$$\mathrm{P}\{|\xi_n - \xi| \geqslant \varepsilon\} \leqslant \mathrm{P}\{|\xi_n - \xi_{n_k}| \geqslant \varepsilon/2\} + \mathrm{P}\{|\xi_{n_k} - \xi| \geqslant \varepsilon/2\},$$

即 $\xi_n \xrightarrow{\mathrm{P}} \xi$. □

关于 $p(>0)$ 阶平均收敛，我们首先对空间 L^p 作若干说明.

以 $L^p = L^p(\Omega, \mathscr{F}, \mathrm{P})$ 表示满足

$$\mathrm{E}|\xi|^p \equiv \int_\Omega |\xi|^p \,\mathrm{dP} < \infty$$

的随机变量 $\xi = \xi(\omega)$ 的空间. 假设 $p \geqslant 1$，并设

$$\|\xi\|_p = (\mathrm{E}|\xi|^p)^{1/p}.$$

显然

$$\|\xi\|_p \geqslant 0, \tag{20}$$

$$\|c\xi\|_p = |c|\,\|\xi\|_p，\ c\ 是常数， \tag{21}$$

由闵可夫斯基不等式 (§6 (31) 式)，有

$$\|\xi + \eta\|_p \leqslant \|\xi\|_p + \|\eta\|_p. \tag{22}$$

这样，根据泛函分析中熟知的定义，定义在 L^p 上并满足条件 (20)~(22) 式的函数 $\|\cdot\|_p$ (对于 $p \geqslant 1$) 称做*半范数*.

为使之成为范数尚需具有如下性质

$$\|\xi\|_p = 0 \Rightarrow \xi = 0. \tag{23}$$

当然这一般并不成立，因为根据性质 **H**(§6)，只是可以断定 ξ 几乎必然为 0，而不是恒等于 0.

鉴于这种情况，需要用略有不同的观点对待空间 L^p. 具体地说，将每一个随机变量 $\xi \in L^p$ 和 L^p 中与之等价的随机变量类 $[\xi]$ 相联系 (称 ξ 和 η 等价，如果 $\mathrm{P}\{\xi = \eta\} = 1$). 不难证明，等价性具有自反性、对称性和传递性，说明线性空间 L^p

可以分割为两两不相交的等价随机变量类. 如果以 $[L^p]$ 表示 L^p 中等价的随机变量类 $[\xi]$ 的全体, 并且定义:

$$[\xi]+[\eta]=[\xi+\eta],$$
$$a[\xi]=[a\xi],\ a \text{ 是常数},$$
$$\|[\xi]\|_p=\|\xi\|_p,$$

则 $[L^p]$ 就是线性赋范空间.

泛函分析中关于空间 $[L^p]$ 的元素, 通常称做 "函数", 而不是 "等价函数类". 因此, 我们以后不再使用记号 $[L^p]$, 而正是把 L^p 理解为等价函数类的集合, 像以前一样简单地称元素, 函数, 随机变量 …… .

泛函分析重要的结果之一, 就是要证明空间 $L^p, p\geqslant 1$, 是完备的, 即任何基本序列都是收敛的. 下面用概率的语言表述并证明这一结果.

定理 7 (柯西 $p(p\geqslant 1)$ 阶平均收敛准则) *空间 L^p 中的随机变量序列 $\{\xi_n\}$, $p(p\geqslant 1)$ 阶平均收敛于取值于 L^p 的随机变量的充分必要条件是, $\{\xi_n\}$ 是 p 阶平均收敛的基本随机变量序列.*

证明 由闵可夫斯基不等式, 可见必要性成立. 设 $\{\xi_n\}$ 是基本随机变量序列 $(\|\xi_n-\xi_m\|_p\to 0, n,m\to\infty)$. 仿照定理 5 的证明, 选择一子序列 $\{\xi_{n_k}\}$ 使 $\xi_{n_k}\xrightarrow{\text{a.s.}}\xi$, 其中 ξ 是某一 $\|\xi\|_p<\infty$ 的随机变量.

设 $n_1=1$. 由归纳法确定一个 n_k, 使之对于任意 $s\geqslant n, t\geqslant n$, 是满足

$$\|\xi_t-\xi_s\|_p<2^{-2k}$$

的最小 $n>n_{k-1}$. 记

$$A_k=\{\omega:|\xi_{n_{k+1}}-\xi_{n_k}|\geqslant 2^{-k}\}.$$

那么, 由切比雪夫不等式, 有

$$\mathbf{P}(A_k)\leqslant\frac{\mathbf{E}|\xi_{n_{k+1}}-\xi_{n_k}|^p}{2^{-kp}}\leqslant\frac{2^{-2kp}}{2^{-kp}}=2^{-kp}\leqslant 2^{-k}.$$

如同定理 5, 由此可见存在随机变量 ξ, 使 $\xi_{n_k}\xrightarrow{\text{a.s.}}\xi$.

由此可见当 $n\to\infty$ 时 $\|\xi_n-\xi\|_p\to 0$. 为此固定 $\varepsilon>0$ 并选择 $N=N(\varepsilon)$, 使对于 $n\geqslant N, m\geqslant N$, 有 $\|\xi_n-\xi_m\|_p^p<\varepsilon$. 那么, 对于任意固定的 $n\geqslant N$, 由法图引理 (§6), 有

$$\begin{aligned}\mathbf{E}|\xi_n-\xi|^p&=\mathbf{E}\left\{\lim_{n_k\to\infty}|\xi_n-\xi_{n_k}|^p\right\}=\mathbf{E}\left\{\varliminf_{n_k\to\infty}|\xi_n-\xi_{n_k}|^p\right\}\\&\leqslant\varliminf_{n_k\to\infty}\mathbf{E}|\xi_n-\xi_{n_k}|^p=\varliminf_{n_k\to\infty}\|\xi_n-\xi_{n_k}\|_p^p\leqslant\varepsilon.\end{aligned}$$

从而, 当 $n\to\infty$ 时 $\mathbf{E}|\xi_n-\xi|^p\to 0$. 由于 $\xi=(\xi-\xi_n)+\xi_n$, 故由闵可夫斯基不等式可见 $\mathbf{E}|\xi|^p<\infty$. □

注 1 按照泛函分析的术语, 完备赋范线性空间称为巴拿赫 (S. Banach) 空间. 这样, 空间 $L^p, p\geqslant 1$, 是巴拿赫空间.

注 2 如果 $0<p<1$, 则 $\|\xi\|_p=(\mathbf{E}|\xi|^p)^{1/p}$ 不满足三角形不等式 (22), 从而不是范数. 然而, 空间 (等价类) $L^p, 0<p<1$, 关于度量 $d(\xi,\eta)=\mathbf{E}|\xi-\eta|^p$ 是完备的.

注 3 设随机变量 $\xi=\xi(\omega)$ 满足条件: $\|\xi\|_\infty<\infty$, 其中 $\|\xi\|_\infty$ 决定于

$$\|\xi\|_\infty=\operatorname{ess\,sup}|\xi|=\inf\{0\leqslant c\leqslant\infty:\mathbf{P}(|\xi|>c)=0\},$$

称做 ξ 的本质上确界. 记 $L^\infty=L^\infty(\Omega,\mathscr{F},\mathbf{P})$ 是随机变量 $\xi=\xi(\omega)$ 的 (等价类) 空间.

函数 $\|\cdot\|_\infty$ 是范数, 且空间 L^∞ 关于此范数是完备的.

6. 练习题

1. 利用定理 5, 证明 §6 定理 3 和定理 4 的命题: "依概率 1 收敛" 可以换成 "依概率收敛".

2. 证明 L^∞ 是完备空间.

3. 证明, 若 $\xi_n\xrightarrow{\mathbf{P}}\xi$ 且 $\eta_n\xrightarrow{\mathbf{P}}\eta$, 则 ξ 和 η 等价 (即 $\mathbf{P}\{\xi\neq\eta\}=0$).

4. 设 $\xi_n\xrightarrow{\mathbf{P}}\xi$ 和 $\eta_n\xrightarrow{\mathbf{P}}\eta$, 且 ξ 和 η 等价. 证明对于任意 $\varepsilon>0$, 有

$$\mathbf{P}\{|\xi_n-\eta_n|\geqslant\varepsilon\}\to 0,\ n\to\infty.$$

5. 设 $\xi_n\xrightarrow{\mathbf{P}}\xi$ 和 $\eta_n\xrightarrow{\mathbf{P}}\eta$. 证明, 如果 $\varphi=\varphi(x,y)$ 是连续函数, 则 $\varphi(\xi_n,\eta_n)\xrightarrow{\mathbf{P}}\varphi(\xi,\eta)$ (斯鲁茨基 [E. E. Слуцкий] 引理).

6. 设 $(\xi_n-\xi)^2\xrightarrow{\mathbf{P}}0$. 证明 $\xi_n^2\xrightarrow{\mathbf{P}}\xi^2$.

7. 证明, 若 $\xi_n\xrightarrow{d}C$, 其中 C 为常数, 则有依概率收敛:

$$\xi_n\xrightarrow{d}C\ \Rightarrow\ \xi_n\xrightarrow{\mathbf{P}}C.$$

8. 设序列 $\{\xi_n\}_{n\geqslant 1}$ 满足条件: 对于某个 $p>0$, $\sum_{n=1}^{\infty}\mathbf{E}|\xi_n|^p<\infty$. 证明 $\xi_n\xrightarrow{\mathbf{P}\text{-a.c.}}0$.

9. 设 $\{\xi_n\}_{n\geqslant 1}$ 是同分布随机变量序列, 证明

$$\mathbf{E}|\xi_1|<\infty\Leftrightarrow\sum_{n=1}^{\infty}\mathbf{P}\{|\xi_1|>\varepsilon n\}<\infty,\varepsilon>0$$
$$\Leftrightarrow\sum_{n=1}^{\infty}\mathbf{P}\left\{\left|\frac{\xi_n}{n}\right|>\varepsilon\right\}<\infty,\varepsilon>0\Rightarrow\frac{\xi_n}{n}\xrightarrow{\mathbf{P}\text{-a.c.}}0.$$

10. 设 $\{\xi_n\}_{n\geqslant 1}$ 是一随机变量序列. 假设存在随机变量 ξ 和子序列 $\{n_k\}$ 使 $\xi_{n_k}\xrightarrow{\mathbf{P}\text{-a.c.}}\xi$ 和 $\max\limits_{n_{k-1}<l\leqslant n_k}|\xi_l-\xi_{n_{k-1}}|\xrightarrow{\mathbf{P}\text{-a.c.}}0$, $k\to\infty$. 证明 $\xi_n\to\xi\ (\mathbf{P}\text{-a.c.})$.

11. 在随机变量的集合中引进“度量 d”如下:

$$d(\xi,\eta)=\mathbf{E}\frac{|\xi-\eta|}{1+|\xi-\eta|},$$

并且认为几乎处处相等的随机变量相同. 证明 $d=d(\xi,\eta)$ 确实是度量, 并且依概率收敛等价于在此度量下的收敛.

12. 证明在随机变量的集合中, 不存在与几乎处处收敛等价的度量.

13. 设 $X_1 \leqslant X_2 \leqslant \cdots$, 且 $X_n \xrightarrow{\mathbf{P}} X$, 证明 $X_n \xrightarrow{\mathbf{P}\text{-a.c.}} X$.

14. 设 $X_n \xrightarrow{\mathbf{P}\text{-a.c.}} X$, 则 $\overline{X}_n \xrightarrow{\mathbf{P}\text{-a.c.}} X$ (切萨罗 [E. Cesaro] 求和法) 其中

$$\overline{X}_n=\frac{1}{n}\sum_{k=1}^{n}X_k.$$

举例说明 $\mathbf{P}$– 几乎处处收敛不能换成依概率收敛.

15. 设 $(\Omega,\mathscr{F},\mathbf{P})$ 是概率空间, $X_n \xrightarrow{\mathbf{P}} X$ 证明, 如果 $\mathbf{P}$ 是原子测度, 则也依概率 1 有 $X_n \to X$ (若对于任意 $B \in \mathscr{F}$, 要么 $\mathbf{P}(B\bigcap A)=\mathbf{P}(A)$, 要么 $\mathbf{P}(B\bigcap A)=0$, 则集合 $A\in\mathscr{F}$ 称做 $\mathbf{P}$– 原子. 测度 $\mathbf{P}$ 称做 $\mathbf{P}$– 原子的, 如果存在可数个不相交的 $\mathbf{P}$– 原子族 $\{A_n\}$, 使 $\mathbf{P}(\bigcup_{n=1}^{\infty}A_n)=1$.)

16. 根据空间 (第一) 博雷尔 – 坎泰利引理, 如果对于任意 $\varepsilon>0$,

$$\sum_{n=1}^{\infty}\mathbf{P}\{|\xi_n|>\varepsilon\}<\infty,$$

则序列 $\xi_n \xrightarrow{\mathbf{P}\text{-a.c.}} 0$. 举例说明, 在 $\sum_{n=1}^{\infty}\mathbf{P}(|\xi_n|>\varepsilon)=\infty(\varepsilon>0)$ 的条件下, $\xi_n \xrightarrow{\mathbf{P}\text{-a.c.}} 0$ 也可能成立.

17. (第二博雷尔 – 坎泰利引理). 设 $\Omega=(0,1)$, $\mathscr{B}=\mathscr{B}((0,1))$, $\mathbf{P}$ 是勒贝格测度. 考虑事件 $A_n=(0,1/n)$. 证明 $\sum\mathbf{P}(A_n)=\infty$, 但是 $(0,1)$ 上的每一个 ω 最多只可能属于有限个集合 $A_1,\cdots,A_{[1/\omega]}$, 即 $\mathbf{P}\{无限多个 A_n\}=0$.

18. 举一随机变量序列的例子 $\{\xi_n\}$, 使之依概率 1 有 $\limsup\xi_n=\infty$, $\liminf\xi_n=-\infty$, 然而存在随机变量 η: $\xi_n \xrightarrow{\mathbf{P}} \eta$.

19. 设 Ω 是有限或可数集合, 证明, 若 $\xi_n \xrightarrow{\mathbf{P}} \xi$, 则 $\xi_n \xrightarrow{\mathbf{P}\text{-a.c.}} \xi$.

20. 设 $A_1,A_2,\cdots$ 是独立事件序列且 $\sum_{n=1}^{\infty}\mathbf{P}(A_n)<\infty$. 证明, 对于 $S_n=\sum_{k=1}^{n}I(A_k)$, 满足“第二博雷尔 – 坎泰利引理”的条件:

$$\lim_n\frac{S_n}{\mathbf{E}S_n}=1\quad(\mathbf{P}\text{-a.c.}).$$

21. 设 $(X_n)_{n\geqslant1}$ 和 $(Y_n)_{n\geqslant1}$ 是两个随机变量序列, 且其一切有限维分布相同 $(F_{X_1,\cdots,X_n}=F_{Y_1,\cdots,Y_n}, n\geqslant1)$. 证明, 若 $X_n \xrightarrow{\mathbf{P}} X$, 则 $Y_n \xrightarrow{\mathbf{P}} Y$, 其中 Y 是与 X 同分布的随机变量.

22. 设 $(X_n)_{n\geqslant1}$ 是独立随机变量序列, 且 $X_n \xrightarrow{\mathbf{P}} X$, 其中 X 是某一随机变量. 证明 X 是退化随机变量.

23. 证明, 对于每一个随机变量序列 $\xi_1,\xi_2,\cdots$, 存在这样的数列 $a_1,a_2,\cdots$, 使 $\xi_n/a_n \xrightarrow{\mathbf{P}\text{-a.c.}} 0$.

24. 设 $\xi_1,\xi_2,\cdots$ 是随机变量序列, $S_n=\xi_1+\cdots+\xi_n$, $n\geqslant1$. 证明 $\{S_n\to\}$ —— 使级数 $\sum_{k\geqslant1}\xi_k(\omega)$ 收敛的 $\omega\in\Omega$ 的集合, 可以表示为如下形式:

$$\{S_n\to\}=\bigcap_{N\geqslant1}\bigcup_{m\geqslant1}\bigcap_{k\geqslant m}\left\{\sup_{l\geqslant k}|S_l-S_k|\leqslant N^{-1}\right\}.$$

相应地, 若级数 $\sum_{k\geqslant1}\xi_k(\omega)$ 发散, 则

$$\{S_n\nrightarrow\}=\bigcup_{N\geqslant1}\bigcap_{m\geqslant1}\bigcup_{k\geqslant m}\left\{\sup_{l\geqslant k}|S_l-S_k|>N^{-1}\right\}.$$

25. 证明第二博雷尔 – 坎泰利引理的如下类型 (引理 1 的命题 b)): 假设事件 $A_1,A_2,\cdots$ (未必独立) 满足条件

$$\sum_{n=1}^{\infty}\mathbf{P}(A_n)=\infty \quad 和 \quad \liminf_n\frac{\displaystyle\sum_{1\leqslant i,k\leqslant n}\mathbf{P}(A_i\cap A_k)}{\left[\displaystyle\sum_{1\leqslant k\leqslant n}\mathbf{P}(A_k)\right]^2}\leqslant1,$$

则 $\mathbf{P}\{无限多个 A_n\}=1$.

26. 证明在第二博雷尔 – 坎泰利引理中, 代替 $A_1,A_2,\cdots$ 独立, 只需要求它们两两独立.

27. 证明 0 – 1 律的如下形式 (对照第四章 §1 的 0 – 1 律): 如果 $A_1,A_2,\cdots$ 两两独立, 则

$$\mathbf{P}\{无限多个 A_n\}=\begin{cases}0, & 若 \sum\mathbf{P}(A_n)<\infty,\\ 1, & 若 \sum\mathbf{P}(A_n)=\infty.\end{cases}$$

28. 设对于任意事件序列 $A_1,A_2,\cdots$, 满足 $\lim_n\mathbf{P}(A_n)=0$ 和 $\sum_n\mathbf{P}(A_n\bigcap\overline{A}_{n+1})<\infty$. 证明 $\mathbf{P}\{无限多个 A_n\}=0$.

29. 证明, 若 $\sum_n\mathbf{P}\{|\xi_n|>n\}<\infty$, 则 $\limsup_n(|\xi_n|/n)\leqslant1(\mathbf{P}\text{-a.c.})$.

30. 设 $\xi_n\downarrow\xi(\mathbf{P}\text{-a.c.})$, $\mathbf{E}|\xi_n|<\infty$, $n\geqslant1$ 且 $\inf_n\mathbf{E}\xi_n>-\infty$. 证明 $\xi_n \xrightarrow{L^1} \xi$, 即 $\mathbf{E}|\xi_n-\xi|\to0$.

31. 利用博雷尔 – 坎泰利引理证明, 如果 $\mathbf{P}\{无限多个 A_n\}=1$, 当且仅当对于每一个集合 A, $\mathbf{P}(A)>0$, $\sum_n\mathbf{P}(A\bigcap A_n)=\infty$.

32. 设事件 $A_1,A_2,\cdots$ 独立, 且对于一切 $n\geqslant1$, $\mathbf{P}(A_n)<1$, 证明 $\mathbf{P}\{无限多个 A_n\}=1$, 当且仅当 $\mathbf{P}(\bigcup A_n)=1$.

33. 设 $X_1, X_2, \cdots$ 是独立随机变量, 且 $\mathbf{P}\{X_n=0\}=1/n$, $\mathbf{P}\{X_n=1\}=1-1/n$; 记 $E_n=\{X_n=0\}$. 证明

$$\sum_{n=1}^{\infty}\mathbf{P}(E_n)=\infty,\quad \sum_{n=1}^{\infty}\mathbf{P}(\overline{E}_n)=\infty.$$

并由此得出结论: $\lim_n X_n$ ($\mathbf{P}$-a.e.) 不存在.

34. 对于随机变量序列 $X_1, X_2, \cdots$, 证明 $X_n \xrightarrow{\mathbf{P}} 0$, 当且仅当对于某个 $r>0$, 有

$$\mathbf{E}\frac{|X_n|^r}{1+|X_n|^r}\to 0.$$

特别, 如果 $S_n=X_1+\cdots+X_n$, 则

$$\frac{S_n-\mathbf{E}S_n}{n}\xrightarrow{\mathbf{P}}0 \Leftrightarrow \mathbf{E}\frac{(S_n-\mathbf{E}S_n)^2}{n^2+(S_n-\mathbf{E}S_n)^2}\to 0.$$

证明, 对于任意随机变量序列 $X_1, X_2, \cdots$, 有

$$\max_{1\leqslant k\leqslant n}|X_k|\xrightarrow{\mathbf{P}}0 \Rightarrow \frac{S_n}{n}\xrightarrow{\mathbf{P}}0.$$

35. 假设对于独立同分布伯努利随机变量序列 $X_1, X_2, \cdots$: $\mathbf{P}\{X_k=\pm 1\}=1/2$;

$$U_n=\sum_{k=1}^{n}\frac{X_k}{2^k},\quad n\geqslant 1.$$

证明, $U_n \xrightarrow{\mathbf{P}\text{-a.e.}} U$, 其中 U 是在 $(-1,+1)$ 上均匀分布的随机变量.

§11. 具有有限二阶矩的随机变量的希尔伯特空间

1. 随机变量的希尔伯特 (D. Hilbert) 空间　在上面讨论的巴拿赫空间 $L^p, p\geqslant 1$ 中, 起重要作用的空间 $L^2=L^2(\Omega,\mathscr{F},\mathbf{P})$ 是具有有限二阶矩的等价随机变量类.

对于 $\xi,\eta\in L^2$, 记

$$(\xi,\eta)\equiv\mathbf{E}\xi\eta. \tag{1}$$

显然, 对于 $\xi,\eta,\zeta\in L^2$, 有

$$(a\xi+b\eta,\zeta)=a(\xi,\zeta)+b(\eta,\zeta),\ (\xi,\xi)\geqslant 0,$$

且

$$(\xi,\xi)=0\Leftrightarrow\xi=0.$$

这样, (ξ,η) 是数量积. 关于由此数量积诱导的范数

$$\|\xi\|=(\xi,\xi)^{1/2}, \tag{2}$$

空间 L^2 是完备的 (证明见 §10). 因此, 按照泛函分析的术语, 引进数量积 (1) 的空间是 (具有有限二阶矩的) 随机变量的希尔伯特空间.

概率论中, 在研究仅由随机变量的前两阶矩决定的性质时, 广泛使用希尔伯特空间方法 ("L^2-理论"). 这里, 我们仅限于介绍 (第六章) 叙述 L^2-理论所必须的基本概念和事实.

2. 正交随机变量系　空间 L^2 中的两个随机变量 ξ 和 η 称做正交的 $(\xi\perp\eta)$, 如果其数量积 $(\xi,\eta)\equiv\mathbf{E}\xi\eta=0$. 根据 §8, 随机变量 ξ 和 η 称做不相关的, 如果 $\mathrm{cov}(\xi,\eta)=0$, 即如果

$$\mathbf{E}\xi\eta=\mathbf{E}\xi\times\mathbf{E}\eta.$$

由此可见, 对于均值为 0 的随机变量, 正交性与不相关性相同.

如果对于任何随机变量 $\xi,\eta\in M(\xi\neq\eta)$, $\xi\perp\eta$, 则系 $M\subseteq L^2$ 称做*正交随机变量系*.

此外, 如果对于一切 $\xi\in M$, 其范数 $\|\xi\|=1$, 则 M 称做随机变量的*规范正交系*.

3. 最优线性估计量　设 $M=\{\eta_1,\cdots,\eta_n\}$ 是规范正交系, 而 ξ 是 L^2 中的某一随机变量. 在形如 $\sum_{i=1}^{n}a_i\eta_i$ 的线性估计类中, 求随机变量 ξ (在均方意义上) 的最优估计量 (对照 §8 的第 2 小节).

通过简单的计算, 可得

$$\begin{aligned}\mathbf{E}\left|\xi-\sum_{i=1}^{n}a_i\eta_i\right|^2&\equiv\left\|\xi-\sum_{i=1}^{n}a_i\eta_i\right\|^2=\left(\xi-\sum_{i=1}^{n}a_i\eta_i,\ \xi-\sum_{i=1}^{n}a_i\eta_i\right)\\&=\|\xi\|^2-2\sum_{i=1}^{n}a_i(\xi,\eta_i)+\left(\sum_{i=1}^{n}a_i\eta_i,\sum_{i=1}^{n}a_i\eta_i\right)\\&=\|\xi\|^2-2\sum_{i=1}^{n}a_i(\xi,\eta_i)+\sum_{i=1}^{n}a_i^2\\&=\|\xi\|^2-\sum_{i=1}^{n}|(\xi,\eta_i)|^2+\sum_{i=1}^{n}|a_i-(\xi,\eta_i)|^2\\&\geqslant\|\xi\|^2-\sum_{i=1}^{n}|(\xi,\eta_i)|^2,\end{aligned}\tag{3}$$

其中用到了等式

$$a_i^2-2a_i(\xi,\eta_i)=|a_i-(\xi,\eta_i)|^2-|(\xi,\eta_i)|^2.$$

由此可见, 对于一切实数 $a_1,\cdots,a_n$, 当 $a_i=(\xi,\eta_i)$, $i=1,2,\cdots,n$ 时,

$$\mathbf{E}\left|\xi-\sum_{i=1}^{n}a_i\eta_i\right|^2$$

达到下确界.

这样，由 $\eta_1, \cdots, \eta_n$ 对 ξ (在均方意义上) 的最优估计量为

$$\hat{\xi} = \sum_{i=1}^{n} (\xi, \eta_i)\eta_i. \tag{4}$$

这时

$$\Delta = \inf \mathbf{E}\left|\xi - \sum_{i=1}^{n} a_i\eta_i\right|^2 = \mathbf{E}|\xi - \hat{\xi}|^2 = \|\xi\|^2 - \sum_{i=1}^{n} |(\xi, \eta_i)|^2 \tag{5}$$

(对照第一章 §4 (17) 式和 §8 (13) 式).

由 (5) 式亦可得如下贝塞尔 (F. W. Bessel) 不等式：如果 $M = \{\eta_1, \eta_2, \cdots\}$ 是规范正交系，且 $\xi \in L^2$，则

$$\sum_{i=1}^{\infty} |(\xi, \eta_i)|^2 \leqslant \|\xi\|^2, \tag{6}$$

其中达到等式当且仅当

$$\xi = \underset{n}{\text{l.i.m.}} \sum_{i=1}^{n} (\xi, \eta_i)\eta_i. \tag{7}$$

最优线性估计量 $\hat{\xi}$ 常记作 $\hat{\mathbf{E}}(\xi|\eta_1, \cdots, \eta_n)$，并称为 ξ 关于 $\eta_1, \cdots, \eta_n$ 的广义条件数学期望.

关于这一术语有如下解释. 假如考虑随机变量 ξ 关于 $\eta_1, \cdots, \eta_n$ 的一切可能的估计量 $\varphi = \varphi(\eta_1, \cdots, \eta_n)$ (φ 是博雷尔函数)，则 $\varphi^* = \mathbf{E}(\xi|\eta_1, \cdots, \eta_n)$，即 ξ 关于 $\eta_1, \cdots, \eta_n$ 的条件数学期望是最优估计量 (对照 §8 的定理 1). 因此，最优线性估计量类似地记作 $\hat{\mathbf{E}}(\xi|\eta_1, \cdots, \eta_n)$，并称为广义条件数学期望. 为此我们指出，如果随机变量 $\eta_1, \cdots, \eta_n$ 构成高斯系 (见下面 §13)，则 $\mathbf{E}(\xi|\eta_1, \cdots, \eta_n)$ 与 $\hat{\mathbf{E}}(\xi|\eta_1, \cdots, \eta_n)$ 相等.

下面说明估计量 $\hat{\xi} = \hat{\mathbf{E}}(\xi|\eta_1, \cdots, \eta_n)$ 的几何意义.

以 $\mathscr{L} = \mathscr{L}(\eta_1, \cdots, \eta_n)$ 表示规范正交随机变量 $\eta_1, \cdots, \eta_n$ 系生成的线性流形 (即形如 $\sum_{i=1}^{n} a_i\eta_i (a_i \in \mathbb{R})$ 的随机变量系).

那么，由以上的叙述可见，ξ 有如下 "正交分解"：

$$\xi = \hat{\xi} + (\xi - \hat{\xi}), \tag{8}$$

其中 $\hat{\xi} \in \mathscr{L}$，而且在对于任意 $\lambda \in \mathscr{L}$，$\xi - \hat{\xi} \perp \lambda$ 意义上 $\xi - \hat{\xi} \perp \mathscr{L}$. 因此自然称 $\hat{\xi}$ 为 ξ 在 $\mathscr{L}$ 上的投影 (即 $\mathscr{L}$ 中 "最接近" ξ 的元素)，$\xi - \hat{\xi}$ 称做 $\mathscr{L}$ 的垂线.

4. 线性无关性 由随机变量 $\eta_1, \cdots, \eta_n$ 的规范正交性假设，可以简单地找到由 $\eta_1, \cdots, \eta_n$ 对 ξ 的最优估计 $\hat{\xi}$ (投影). 假如不要求规范正交性假设，则情况比较复杂. 不过，下面将要证明，在有随机变量 $\eta_1, \cdots, \eta_n$ 的情形，在一定意义上可以归为已经研究的规范正交变量的情形. 为简便计，在以后的叙述中，将假设所考虑的一切随机变量的均值为 0.

称随机变量 $\eta_1, \cdots, \eta_n$ 为线性无关的，如果等式

$$\sum_{i=1}^{n} a_i\eta_i = 0 \quad (\mathbf{P}\text{-a.c.})$$

只有当所有 a_i 都等于零时才成立.

考虑协方差矩阵

$$\boldsymbol{R} = \mathbf{E}\eta\eta^{\mathrm{T}},$$

其中 $\eta = (\eta_1, \cdots, \eta_n)$ 视为列向量. $\boldsymbol{R} = \mathbf{E}\eta\eta^{\mathrm{T}}$ 是非负定对称矩阵，如同在 §8 已经指出的，存在正交矩阵 $\boldsymbol{Z}$，可将 $\boldsymbol{R}$ 化为对角矩阵：

$$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z} = \boldsymbol{D},$$

其中

$$\boldsymbol{D} = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{pmatrix}$$

的对角线上的非负元素 $d_i \geqslant 0 (i = 1, \cdots, n)$，是矩阵 $\boldsymbol{R}$ 的特征数，即特征方程 $\det(\boldsymbol{R} - \lambda\boldsymbol{E}) = 0$ 的根 λ ($\boldsymbol{E}$ 是单位矩阵).

如果随机变量 $\eta_1, \cdots, \eta_n$ 线性无关，则克拉默 (J. P. Gram) 行列式 (即 $\det \boldsymbol{R}$) 不等于 0，即全部 $d_i > 0 (i = 1, \cdots, n)$. 设

$$\boldsymbol{B} = \begin{pmatrix} \sqrt{d_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{d_n} \end{pmatrix}$$

和

$$\beta = \boldsymbol{B}^{-1}\boldsymbol{Z}^{\mathrm{T}}\eta. \tag{9}$$

那么，向量 β 的协方差矩阵

$$\mathbf{E}\beta\beta^{\mathrm{T}} = \boldsymbol{B}^{-1}\boldsymbol{Z}^{\mathrm{T}}\mathbf{E}\eta\eta^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{B}^{-1} = \boldsymbol{B}^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z}\boldsymbol{B}^{-1} = \boldsymbol{E},$$

从而向量 $\beta = (\beta_1, \cdots, \beta_n)$ 的分量是两两不相关的随机变量. 同样明显

$$\eta = (\boldsymbol{Z}\boldsymbol{B})\beta. \tag{10}$$

于是，如果随机变量 $\eta_1, \cdots, \eta_n$ 线性无关，则存在这样一个规范正交随机变量系 $\beta_1, \cdots, \beta_n$，使 (9) 和 (10) 式成立. 这时

$$\mathscr{L}(\eta_1, \cdots, \eta_n) = \mathscr{L}(\beta_1, \cdots, \beta_n).$$

所介绍的得到规范正交随机变量系 $\beta_1,\cdots,\beta_n$ 的方法，在一系列问题中显得不是很有用. 问题在于，假如把 η_k 视为随机变量 $(\eta_1,\cdots,\eta_k)$ 序列在时刻 k 的值，那么上面建立的正交随机变量 $\beta_1,\cdots,\beta_k$ 系的值 β_k 不仅依赖于"过去" $(\eta_1,\cdots,\eta_k)$，而且依赖于"将来" $(\eta_{k+1},\cdots,\eta_n)$. 下面将要介绍的*克拉默－施密特* (E. Schmidt) *正交化过程*不但没有这一缺陷，而且它具有如下优点：它可以用于线性无关的无限随机变量序列 (即任何有限个随机变量序列都线性无关).

设 $\eta_1,\eta_2,\cdots$ 是 L^2 中线性无关的随机变量序列. 用归纳法以下面介绍的方式建立序列 $\varepsilon_1,\varepsilon_2,\cdots$. 设 $\varepsilon_1=\eta_1/\|\eta_1\|$. 假设 $\varepsilon_1,\cdots,\varepsilon_{n-1}$ 已经选定，而且它们正交，则设

$$\varepsilon_n=\frac{\eta_n-\hat{\eta}_n}{\|\eta_n-\hat{\eta}_n\|},\tag{11}$$

其中 $\hat{\eta}_n$ 是 η_n 在由正交随机变量 $\varepsilon_1,\cdots,\varepsilon_{n-1}$ 生成的线性流形 $\mathscr{L}=\mathscr{L}(\varepsilon_1,\cdots,\varepsilon_{n-1})$ 上的投影.

$$\hat{\eta}_n=\sum_{k=1}^{n-1}(\eta_n,\varepsilon_k)\varepsilon_k.\tag{12}$$

由于 $\eta_1,\cdots,\eta_n$ 线性无关，$\mathscr{L}(\eta_1,\cdots,\eta_{n-1})=\mathscr{L}(\varepsilon_1,\cdots,\varepsilon_{n-1})$ 可见 $\|\eta_n-\hat{\eta}_n\|>0$，从而 ε_n 有定义.

根据构造 $\|\varepsilon_n\|=1,n\geqslant 1$，而且显然 $(\varepsilon_n,\varepsilon_k)=0,k<n$. 因此 $\varepsilon_1,\varepsilon_2,\cdots$ 是正交序列. 这时根据 (11) 式有

$$\eta_n=\hat{\eta}_n+b_n\varepsilon_n,$$

其中 $b_n=\|\eta_n-\hat{\eta}_n\|$，而 $\hat{\eta}_n$ 决定于 (12) 式.

现在假设 $\eta_1,\cdots,\eta_n$ 是任意随机变量系 (未必是线性无关的). 设 $\det R=0$，其中 $R=(r_{ij})$ 是向量 $(\eta_1,\cdots,\eta_n)$ 的协方差矩阵；设

$$\operatorname{rang} R=r<n.$$

那么，由代数熟知，对于二次型

$$Q(a)=\sum_{i,j=1}^{n}r_{ij}a_ia_j,\quad a=(a_1,\cdots,a_n),$$

恰好存在 $n-r$ 个线性无关向量 $a^{(1)},\cdots,a^{(n-r)}$，使 $Q(a^{(i)})=0,i=1,\cdots,n-r$.

因为

$$Q(a)=\mathrm{E}\left(\sum_{k=1}^{n}a_k\eta_k\right)^2,$$

从而，依概率 1，有

$$\sum_{k=1}^{n}a_k^{(i)}\eta_k=0,\ i=1,\cdots,n-r.$$

换句话说，在随机变量 $\eta_1,\cdots,\eta_n$ 之间，恰好存在 $n-r$ 个线性关系式. 因此，例如 $\eta_1,\cdots,\eta_r$ 线性无关，则其余一切随机变量 $\eta_{r+1},\cdots,\eta_n$ 可以通过 $\eta_1,\cdots,\eta_r$ 线性表示，故 $\mathscr{L}(\eta_1,\cdots,\eta_n)=\mathscr{L}(\eta_1,\cdots,\eta_r)$. 由此可见，利用正交化过程，可以得到 r 个正交随机变量 $\varepsilon_1,\cdots,\varepsilon_r$，使之一切通过 $\eta_1,\cdots,\eta_n$ 线性表示，并且 $\mathscr{L}(\eta_1,\cdots,\eta_n)=\mathscr{L}(\varepsilon_1,\cdots,\varepsilon_r)$.

5. 正交基底和正交化 假设 $\eta_1,\eta_2,\cdots$ 是 L^2 中的随机变量序列. 以 $\mathscr{L}(\eta_1,\cdots,\eta_n)$ 表示随机变量 $\eta_1,\eta_2,\cdots$ 生成的*线性流形*，即形如 $\sum_{i=1}^{n}a_i\eta_i,n\geqslant 1,a_i\in\mathbb{R}$ 的随机变量的全体. 以 $\overline{\mathscr{L}}=\overline{\mathscr{L}}(\eta_1,\cdots,\eta_n)$ 表示随机变量 $\eta_1,\eta_2,\cdots$ 生成的*闭线性流形*，即 $\mathscr{L}$ 中的随机变量 $\eta_1,\eta_2,\cdots$ 及其均方极限的全体.

称随机变量系 $\eta_1,\eta_2,\cdots$ 是空间 L^2 中*可数正交基* (或*完全正交系*)，如果：

a) $\eta_1,\eta_2,\cdots$ 是正交系，

b) $\overline{\mathscr{L}}(\eta_1,\eta_2,\cdots)=L^2$.

具有可数正交基的希尔伯特空间，称做*可分的*.

由于条件 b)，对于任意 $\xi\in L^2$ 和 $\varepsilon>0$，存在实数 $a_1,\cdots,a_n$，使

$$\left\|\xi-\sum_{i=1}^{n}a_i\eta_i\right\|\leqslant\varepsilon.$$

那么，根据 (3) 式，

$$\left\|\xi-\sum_{i=1}^{n}(\xi,\eta_i)\eta_i\right\|\leqslant\varepsilon.$$

从而，对于可分正交基的希尔伯特空间 L^2，任意随机元素 ξ 可以表示为

$$\xi=\sum_{i=1}^{\infty}(\xi,\eta_i)\eta_i,\tag{13}$$

确切地说，

$$\xi=\operatorname*{l.i.m.}_{n}\sum_{i=1}^{n}(\xi,\eta_i)\eta_i.$$

由此及 (3) 式，可见有如下帕塞瓦尔 (M. A. Parseval) *等式*：

$$\|\xi\|^2=\sum_{i=1}^{\infty}|(\xi,\eta_i)|^2,\ \xi\in L^2.\tag{14}$$

不难证明其逆命题也成立：若 $\eta_1,\eta_2,\cdots$ 是某一正交系，且 (13) 式或 (14) 式中任何条件成立，则此正交系是基底.

举几个可分希尔伯特空间及其基底的例子.

例 1 设 $\Omega=\mathbb{R},\mathscr{F}=\mathscr{B}(\mathbb{R})$，而 $\mathbf{P}$ 是高斯测度：

$$\mathbf{P}(-\infty,a]=\int_{-\infty}^{a}\varphi(x)\mathrm{d}x,\quad\varphi(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}.$$

记 $D = \mathrm{d}/\mathrm{d}x$ 并引进函数

$$H_n(x) = \frac{(-1)^n D^n \varphi(x)}{\varphi(x)}, \quad n \geqslant 0. \tag{15}$$

不难求出:

$$\begin{aligned} D\varphi(x) &= -x\varphi(x), \\ D^2\varphi(x) &= (x^2-1)\varphi(x), \\ D^3\varphi(x) &= (3x-x^3)\varphi(x), \\ &\cdots\cdots \end{aligned} \tag{16}$$

由此可见, $H_n(x)$ 是多项式 (称做埃尔米特 (G. Hermite) 多项式). 由 (15) 和 (16) 式, 有

$$\begin{aligned} H_0(x) &= 1, \\ H_1(x) &= x, \\ H_2(x) &= x^2 - 1, \\ H_3(x) &= x^3 - 3x, \\ &\cdots\cdots \end{aligned}$$

由简单的计算, 可得

$$(H_m, H_n) = \int_{-\infty}^{\infty} H_m(x)H_n(x)\mathbf{P}(\mathrm{d}x) = \int_{-\infty}^{\infty} H_m(x)H_n(x)\varphi(x)\mathrm{d}x = n!\delta_{nm},$$

其中 δ_{nm} 是克罗内克 (L. Kronecker) 记号 ($\delta_{nm} = 0$, 若 $m \neq n$; $\delta_{nm} = 1$, 若 $m = n$). 因此, 如果设

$$h_n(x) = \frac{H_n(x)}{\sqrt{n!}},$$

则这些赋范埃尔米特多项式 $\{h_n(x)\}_{n\geqslant 0}$ 为规范正交基. 由泛函分析 [33, 第七章 §3] 知, 如果

$$\lim_{\varepsilon \downarrow 0} \int_{-\infty}^{\infty} e^{\varepsilon|x|}\mathbf{P}(\mathrm{d}x) < \infty, \tag{17}$$

则函数系 $\{1, x, x^2, \cdots\}$ 在 L^2 中是完备的, 即 L^2 中的任意函数 $\xi = \xi(x)$, 要么可以表示为 $\sum_{i=1}^{n} a_i\eta_i(x)$, 其中 $\eta_i(x) = x^i$, 要么可以表示为这些和 (在均方意义上) 的极限. 如果对于序列 $\eta_1(x), \eta_2(x), \cdots (\eta_i(x) = x^i)$ 运用克拉默 – 施密特正交化过程, 则所得规范化正交系恰好与赋范埃尔米特多项式系相同. 条件 (17) 对于现在考虑的情形成立. 从而, 多项式 $\{h_n(x)\}_{n\geqslant 0}$ 是基底, 说明在所考虑的概率空间中, 任意随机变量 $\xi = \xi(x)$ 都可以表示为

$$\xi(x) = \underset{n}{\mathrm{l.i.m.}} \sum_{i=1}^{n} (\xi, h_i)h_i(x). \tag{18}$$

例 2 设 $\Omega = \{0, 1, 2, \cdots\}$, $\mathscr{P} = \{P_1, P_2, \cdots\}$ 是泊松分布:

$$P_x = \frac{\lambda^x}{x!}e^{-\lambda}, x = 0, 1, \cdots; \lambda > 0.$$

记 $\Delta f(x) = f(x) - f(x-1)$ $(f(x) = 0, x < 0)$, 并仿照 (15) 式定义泊松 – 沙利耶 (M. Charlier) 多项式

$$\Pi_n(x) = \frac{(-1)^n \Delta^n P_x}{P_x}, \quad n \geqslant 1, \Pi_0 = 1. \tag{19}$$

由于

$$(\Pi_m, \Pi_n) = \sum_{x=0}^{\infty} \Pi_m(x)\Pi_n(x)P_x = c_n\delta_{mn},$$

其中 c_n 是正常数, 则赋范泊松 – 沙利耶多项式 $\{\pi_n(x)\}_{n\geqslant 0}$, 其中 $\pi_n(x) = \Pi_n(x)/\sqrt{c_n}$ 为规范正交系, 并且因为满足条件 (17) 式, $\{\pi_n(x)\}_{n\geqslant 0}$ 也是规范化正交基.

例 3 在本例中将要引进的拉德马赫 – 哈尔 (H. Rademacher—A. Haar) 规范正交函数系, 不但在函数论中, 而且在概率论中也很重要.

设 $\Omega = [0, 1)$, $\mathscr{F} = \mathscr{B}([0, 1))$, 而 $\mathbf{P}$ 是勒贝格测度. 在 §1 曾经指出, 每一个数 $x \in [0, 1)$ 都可以表示为二进制小数:

$$x = \frac{x_1}{2} + \frac{x_2}{2^2} + \cdots,$$

其中 $x_i = 0$ 或 1. (为了唯一性, 我们约定: 只考虑在二进制小数中含无限个 0 的数. 例如, 对于

$$\frac{1}{2} = \frac{1}{2} + \frac{0}{2^2} + \frac{0}{2^3} + \cdots = \frac{0}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots,$$

我们只取第一个.)

引进随机变量 $\xi_1(x), \xi_2(x), \cdots$, 设

$$\xi_n(x) = x_n.$$

那么, 对于只有 0 和 1 两个可能值的任何 a_i, 有

$$\begin{aligned} &\mathbf{P}\{x : \xi_1 = a_1, \cdots, \xi_n = a_n\} \\ &= \mathbf{P}\left\{x : \frac{a_1}{2} + \frac{a_2}{2^2} + \cdots + \frac{a_n}{2^n} \leqslant x < \frac{a_1}{2} + \frac{a_2}{2^2} + \cdots + \frac{a_n}{2^n} + \frac{1}{2^n}\right\} \\ &= \mathbf{P}\left\{x : x \in \left[\frac{a_1}{2} + \frac{a_2}{2^2} + \cdots + \frac{a_n}{2^n}, \frac{a_1}{2} + \frac{a_2}{2^2} + \cdots + \frac{a_n}{2^n} + \frac{1}{2^n}\right)\right\} = \frac{1}{2^n}. \end{aligned}$$

由此可见, $\xi_1, \xi_2, \cdots$ 是独立伯努利随机变量序列 (图 30 表示 $\xi_1 = \xi_1(x)$ 和 $\xi_2 = \xi_2(x)$ 的构造).

现在如果设 $R_n(x) = 1 - 2\xi_n(x), n \geqslant 1$, 则不难验证函数系 $\{R_n\}$ (图 31 是拉德马赫函数) 是正交的:

$$\mathbf{E}R_nR_m = \int_0^1 R_n(x)R_m(x)\mathrm{d}x = \delta_{nm}.$$

注意 $(1,R_n)\equiv \mathbf{E}R_n=0$. 由此可见该函数系不是完备的.

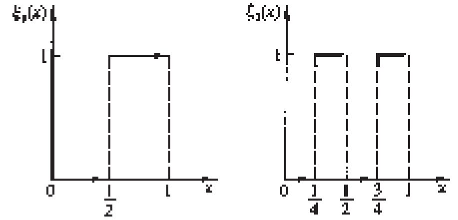


图 30 伯努利变量

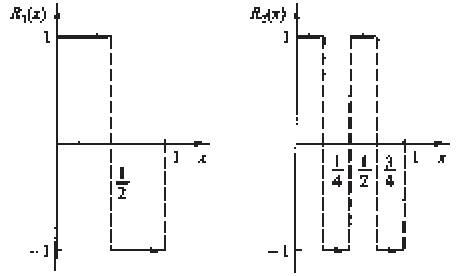


图 31 拉德马赫函数

然而, 拉德马赫系可以用来构造所谓哈尔系, 而哈尔系便于操作, 且不仅是规范正交的, 并且是完备的.

仍然假设 $\Omega=[0,1)$, $\mathscr{F}=\mathscr{B}([0,1))$. 设

$H_1(x)=1$,

$H_2(x)=R_1(x)$,

. . . .

$$H_n(x)=\begin{cases}2^{j/2}R_{j+1}(x), & 若\ \dfrac{k-1}{2^j}\leqslant x<\dfrac{k}{2^j},\ n=2^j+k,\ 1\leqslant k\leqslant 2^j,\ j\geqslant 1,\\ 0, & 若不然.\end{cases}$$

不难验证, $H_n(x)$, $n\geqslant 3$, 可以表示为:

$$H_{2^m+1}(x)=\begin{cases}2^{m/2}, & 若\ 0\leqslant x<2^{-(m+1)},\\ -2^{m/2}, & 若\ 2^{-(m+1)}\leqslant x<2^{-m},\\ 0, & 其他.\end{cases}$$

$$H_{2^m+j}(x)=H_{2^m+1}\left(x-\frac{j-1}{2^m}\right),\ j=1,\cdots,2^m,\ m=1,2,\cdots.$$

图 32 是前 8 个哈尔函数的示意图. 由图 32 可以得到哈尔函数组成的构造和性质的印象.

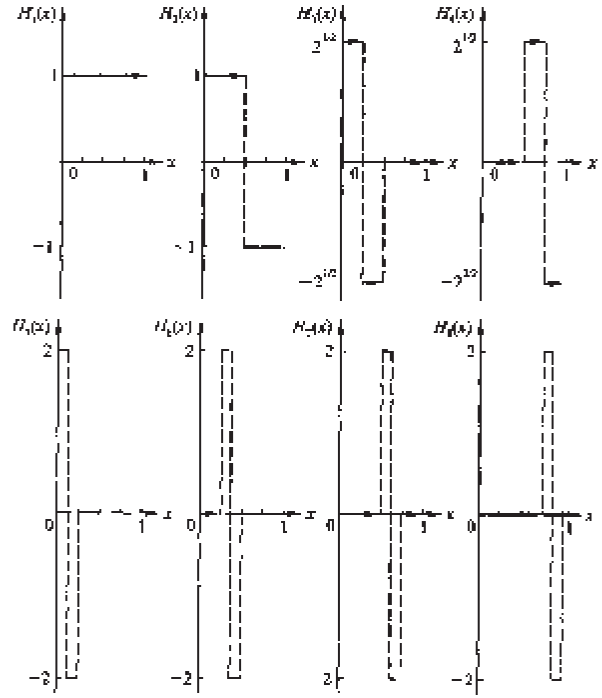


图 32 哈尔函数 $H_1(x),\cdots,H_8(x)$

不难验证, 哈尔函数系是规范化正交的, 并且在 L^1 和 L^2 中是完备的, 即如果对于 $p=1$ 或 2, 哈尔函数 $f=f(x)\in L^p$, 则

$$\int_0^1\left|f(x)-\sum_{k=1}^{n}(f,H_k)H_k(x)\right|^p\mathrm{d}x\to 0,\ n\to\infty;$$

而且, 依概率 1 (对勒贝格测度) 有

$$\sum_{k=1}^{n}(f,H_k)H_k(x)\to f(x),\ n\to\infty.$$

我们将在第七章 §4 证明这些事实. 这些结果由鞅收敛的一般理论导出, 并且是鞅方法在函数论中的运用很好的实例.

6. 最优线性估计 如果 $\eta_1, \cdots, \eta_n$ 是某一有限规范正交系，则如上面已经证明的，对于任意随机变量 $\xi \in L^2$，在线性流形 $\mathscr{L} = \mathscr{L}(\eta_1, \cdots, \eta_n)$ 中，存在随机变量 $\hat{\xi}$(ξ 在 $\mathscr{L}$ 上的投影)，使

$$\|\xi - \hat{\xi}\| = \inf_{\zeta}\{\|\xi - \zeta\| : \zeta \in \mathscr{L}(\eta_1, \cdots, \eta_n)\}.$$

这时 $\hat{\xi} = \sum_{i=1}^{n}(\xi, \eta_i)\eta_i$. 这一结果容许自然地推广到 $\eta_1, \eta_2, \cdots$ 是可数正交系 (未必是基底) 的情形. 具体地说, 有如下结果.

定理 *假设 $\eta_1, \eta_2, \cdots$ 是规范正交随机变量系, $\overline{\mathscr{L}} = \overline{\mathscr{L}}(\eta_1, \eta_2, \cdots)$ 是由 $\eta_1, \eta_2, \cdots$ 生成的封闭线性流形. 那么, 存在并且唯一随机元 $\hat{\xi} \in \overline{\mathscr{L}}$, 使*

$$\|\xi - \hat{\xi}\| = \inf_{\zeta}\{\|\xi - \zeta\| : \zeta \in \overline{\mathscr{L}}\}. \tag{20}$$

这时,

$$\hat{\xi} = \underset{n}{\text{l.i.m.}} \sum_{i=1}^{n}(\xi, \eta_i)\eta_i, \tag{21}$$

且 $\xi - \hat{\xi} \perp \zeta, \zeta \in \overline{\mathscr{L}}$.

证明 记 $d = \inf\{\|\xi - \zeta\| : \zeta \in \overline{\mathscr{L}}\}$, 并选一序列 $\zeta_1, \zeta_2, \cdots$, 使 $\|\xi - \zeta_n\| \to d$. 现在证明此序列是基本的. 简单的计算表明,

$$\|\zeta_n - \zeta_m\|^2 = 2\|\zeta_n - \xi\|^2 + 2\|\zeta_m - \xi\|^2 - 4\left\|\frac{1}{2}(\zeta_n + \zeta_m) - \xi\right\|^2.$$

由于明显 $(\zeta_n + \zeta_m)/2 \in \overline{\mathscr{L}}$, 可见

$$\left\|\frac{1}{2}(\zeta_n + \zeta_m) - \xi\right\|^2 \geqslant d^2.$$

从而 $\|\zeta_n - \zeta_m\|^2 \to 0, n, m \to \infty$.

空间 L^2 是完备的 (§10 定理 7). 因此存在随机元 $\hat{\xi}$, 使 $\|\zeta_n - \hat{\xi}\| \to 0$. 由于集合 $\overline{\mathscr{L}}$ 是封闭的, 故 $\hat{\xi} \in \overline{\mathscr{L}}$. 其次, 因为 $\|\xi - \zeta_n\| \to d$, 所以 $\|\xi - \hat{\xi}\| = d$, 从而证明了所需要元素 $\hat{\xi}$ 的存在性.

现在证明 $\hat{\xi}$ 是 $\overline{\mathscr{L}}$ 中唯一具有所要求性质的元素. 假设 $\tilde{\xi} \in \overline{\mathscr{L}}$ 且

$$\|\xi - \hat{\xi}\| = \|\xi - \tilde{\xi}\| = d.$$

那么 (由于练习题 3)

$$\|\hat{\xi} + \tilde{\xi} - 2\xi\|^2 + \|\hat{\xi} - \tilde{\xi}\|^2 = 2\|\hat{\xi} - \xi\|^2 + 2\|\tilde{\xi} - \xi\|^2 = 4d^2.$$

但是

$$\|\hat{\xi} + \tilde{\xi} - 2\xi\|^2 = 4\left\|\frac{1}{2}(\hat{\xi} + \tilde{\xi}) - \xi\right\|^2 \geqslant 4d^2.$$

从而, $\|\hat{\xi} - \tilde{\xi}\|^2 = 0$. 这就证明了在 $\overline{\mathscr{L}}$ 中 "最接近" ξ 的元素的唯一性.

现在证明 $\xi - \hat{\xi} \perp \zeta, \zeta \in \overline{\mathscr{L}}$. 由 (20) 式, 对于任意 $c \in \mathbb{R}$,

$$\|\xi - \hat{\xi} - c\zeta\| \geqslant \|\xi - \hat{\xi}\|.$$

由于

$$\|\xi - \hat{\xi} - c\zeta\|^2 = \|\xi - \hat{\xi}\|^2 + c^2\|\zeta\|^2 - 2(\xi - \hat{\xi}, c\zeta),$$

所以

$$c^2\|\zeta\|^2 \geqslant 2(\xi - \hat{\xi}, c\zeta). \tag{22}$$

取 $c = \lambda(\xi - \hat{\xi}, \zeta), \lambda \in \mathbb{R}$, 则由 (22) 式, 得

$$(\xi - \hat{\xi}, \zeta)^2[\lambda^2\|\zeta\|^2 - 2\lambda] \geqslant 0.$$

对于充分小的正数 λ, 不等式 $\lambda^2\|\zeta\|^2 - 2\lambda < 0$ 成立. 因此 $(\xi - \hat{\xi}, \zeta) = 0, \zeta \in \overline{\mathscr{L}}$.

只剩下证明 (21) 式.

集合 $\overline{\mathscr{L}} = \overline{\mathscr{L}}(\eta_1, \eta_2, \cdots)$ 是 L^2 的封闭子空间, 因此本身是希尔伯特空间 (具有与 L^2 同样的数量积). 随机变量族 $\eta_1, \eta_2, \cdots$ 是希尔伯特空间 $\overline{\mathscr{L}}$ 的基底, 从而

$$\hat{\xi} = \underset{n}{\text{l.i.m.}} \sum_{k=1}^{n}(\hat{\xi}, \eta_k)\eta_k. \tag{23}$$

由于 $\xi - \hat{\xi} \perp \eta_k, k \geqslant 1$, 可见 $(\hat{\xi}, \eta_k) = (\xi, \eta_k), k \geqslant 1$. 于是, 连同 (23) 式就证明了 (21) 式. □

注 像有限维情形一样, $\hat{\xi}$ 称为在 $\overline{\mathscr{L}} = \overline{\mathscr{L}}(\eta_1, \eta_2, \cdots)$ 上的投影, $\xi - \hat{\xi}$ 称做垂线, 而表达式

$$\xi = \hat{\xi} + (\xi - \hat{\xi})$$

称做正交分解.

随机变量 $\hat{\xi}$ 记作 $\hat{\mathbf{E}}(\xi|\eta_1, \eta_2, \cdots)$ (对照第 3 小节的 $\hat{\mathbf{E}}(\xi|\eta_1, \cdots, \eta_n)$), 并且称做 ($\xi$ 关于 $\eta_1, \eta_2, \cdots$ 的) 广义条件数学期望. 按由 $\eta_1, \eta_2, \cdots$ 估计 ξ 的观点, 随机变量 $\hat{\xi}$ 是最优线性估计量, 而由 (5) 和 (23) 式可见, 估计量的误差为

$$\Delta = \mathbf{E}|\xi - \hat{\xi}|^2 = \|\xi - \hat{\xi}\|^2 = \|\xi\|^2 - \sum_{i=1}^{\infty}|(\xi, \eta_i)|^2.$$

7. 练习题

1. 证明, 如果 $\xi = \text{l.i.m.}\,\xi_n$, 则 $\|\xi_n\| \to \|\xi\|$.
2. 证明, 如果 $\xi = \text{l.i.m.}\,\xi_n$ 和 $\eta = \text{l.i.m.}\,\eta_n$, 则 $(\xi_n, \eta_n) \to (\xi, \eta)$.
3. 证明, 范数 $\|\cdot\|$ 满足 "平行四边形" 性质:

$$\|\xi + \eta\|^2 + \|\xi - \eta\|^2 = 2(\|\xi\|^2 + \|\eta\|^2).$$

4. 证明, 正交随机变量族 $\{\xi_1,\cdots,\xi_n\}$, 满足 "毕达哥拉斯 (Pythagoras) 定理":

$$\left\|\sum_{i=1}^{n}\xi_i\right\|^2=\sum_{i=1}^{n}\|\xi_i\|^2.$$

5. 设 $\xi_1,\xi_2,\cdots$ 是正交随机变量序列, $S_n=\xi_1+\cdots+\xi_n$. 证明, 若 $\sum_{n=1}^{\infty}\mathrm{E}\xi_n^2<\infty$, 则存在随机变量 $S(\mathrm{E}S^2<\infty)$, 使 $\mathrm{l.i.m.}\,S_n=S$, 即 $\|S_n-S\|^2=\mathrm{E}|S_n-S|^2\to 0, n\to\infty$.

6. 证明拉德马赫函数 R_n 可以定义为:

$$R_n(x)=\mathrm{sign}(\sin 2^n\pi x),\quad 0\leqslant x\leqslant 1,\quad n=1,2,\cdots$$

7. 证明, 对于 $\xi\in L^2(\mathscr{F})$,

$$\|\xi\|\geqslant\|\mathrm{E}(\xi|\mathscr{G})\|,$$

并且等式成立, 当且仅当几乎必然 $\xi=\mathrm{E}(\xi|\mathscr{G})$.

8. 证明, 若 $\xi,\eta\in L^2(\mathscr{F})$, $\mathrm{E}(\xi|\eta)=\eta$, $\mathrm{E}(\eta|\xi)=\xi$, 则 $\mathrm{P}\{\xi=\eta\}=1$.

9. 有 $\mathscr{F}$ 的 3 个子 $\sigma-$ 代数序列: $(\mathscr{G}_n^{(1)})$, $(\mathscr{G}_n^{(2)})$ 和 $(\mathscr{G}_n^{(3)})$, 设 ξ 是有界随机变量. 已知对于每个 n, 有

$$\mathscr{G}_n^{(1)}\subseteq\mathscr{G}_n^{(2)}\subseteq\mathscr{G}_n^{(3)},\quad \mathrm{E}(\xi|\mathscr{G}_n^{(1)})\xrightarrow{\mathrm{P}}\eta,\quad \mathrm{E}(\xi|\mathscr{G}_n^{(3)})\xrightarrow{\mathrm{P}}\eta.$$

证明 $\mathrm{E}(\xi|\mathscr{G}_n^{(2)})\xrightarrow{\mathrm{P}}\eta$.

§12. 特征函数

1. 复数值随机变量 特征函数方法是概率论的基本的分析工具之一. 这在第三章中证明极限定理时, 特别是证明中心极限定理时表现得最明显. 中心极限定理是棣莫弗 – 拉普拉斯定理的推广. 这里, 我们将限于介绍特征函数的定义及其基本性质.

首先, 作一个一般性的说明.

除 (取实数值的) 随机变量外, 特征函数还要求考虑复数值随机变量 (见 §5 第 1 小节).

随机变量的许多有关定义和性质, 都可以很容易移植到复数情形. 例如, 若数学期望为 $\mathrm{E}\xi$ 和 $\mathrm{E}\eta$, 则认为复数值随机变量 $\zeta=\xi+i\eta$ 的数学期望定义为 $\mathrm{E}\zeta=\mathrm{E}\xi+i\mathrm{E}\eta$. 由随机元独立性的定义 6 (§5) 不难得到, 复数值随机变量 $\zeta_1=\xi_1+i\eta_1$ 和 $\zeta_2=\xi_2+i\eta_2$ 相互独立, 当且仅当随机向量 (ξ_1,η_1) 和 (ξ_2,η_2) 独立, 或等价地 $\sigma-$ 代数 $\mathscr{F}_{\xi_1,\eta_1}$ 和 $\mathscr{F}_{\xi_2,\eta_2}$ 相互独立.

与具有有限二阶矩的实数值随机变量的空间 L^2 同时, 我们引进复数值随机变量 $\zeta=\xi+i\eta$ 的希尔伯特空间, 其中 $\mathrm{E}|\zeta|^2<\infty$, $|\zeta|^2=\xi^2+\eta^2$, 并在其中引进数量积

$(\zeta_1,\zeta_2)=\mathrm{E}\zeta_1\bar{\zeta}_2$, 其中 $\bar{\zeta}_2$ 是复值共轭随机变量. 以后实数值随机变量和复数值随机变量, 一般都称为随机变量, 除非必要时才明确指出具体是哪一种.

还需要指出下列记号. 在向量 $a\in\mathbf{R}^n$ 的代数运算中, 向量视为列向量:

$$a=\begin{pmatrix}a_1\\ \vdots\\ a_n\end{pmatrix},$$

并以 $a^{\mathrm{T}}=(a_1,\cdots,a_n)$ 表示行向量. 对于 $a,b\in\mathbf{R}^n$, 把 $\sum_{i=1}^{n}a_ib_i$ 理解为 a 和 b 的数量积 (a,b). 显然, $(a,b)=a^{\mathrm{T}}b$.

如果 $a\in\mathbf{R}^n$, 而 $\boldsymbol{R}=(r_{ij})$ 是 $n\times n$ 阶矩阵, 则

$$(\boldsymbol{R}a,a)=a^{\mathrm{T}}\boldsymbol{R}a=\sum_{i,j=1}^{n}a_ir_{ij}a_j. \tag{1}$$

2. 特征函数的定义

定义 1 设 $F=F(x)$, $x=(x_1,\cdots,x_n)$ 是 $(\mathbf{R}^n,\mathscr{B}(\mathbf{R}^n))$ 上的 n 维分布函数, 则函数

$$\varphi(t)=\int_{\mathbf{R}^n}e^{i(t,x)}\mathrm{d}F(x),\quad t\in\mathbf{R}^n, \tag{2}$$

称做 $F(x)$ 的特征函数.

定义 2 设 $\xi=(\xi_1,\cdots,\xi_n)$ 是空间 $(\Omega,\mathscr{F},\mathrm{P})$ 上在 $\mathbf{R}^n$ 中取值的随机向量, 则函数

$$\varphi_\xi(t)=\int_{\mathbf{R}^n}e^{i(t,x)}\mathrm{d}F_\xi(x),\quad t\in\mathbf{R}^n, \tag{3}$$

称做 $F_\xi(x)$ 的特征函数, 其中 $F_\xi=F_\xi(x)$, $x=(x_1,\cdots,x_n)$ 是随机向量 $\xi=(\xi_1,\cdots,\xi_n)$ 的分布函数.

如果 $F(x)$ 有密度 $f=f(x)$, 则

$$\varphi(t)=\int_{\mathbf{R}^n}e^{i(t,x)}f(x)\mathrm{d}x,\quad t\in\mathbf{R}^n.$$

换句话说, 这时特征函数 $\varphi(t)$ 恰好是函数 $f=f(x)$ 的傅里叶 (J. Fourier) 变换.

由 (3) 式和 §6 (关于在勒贝格积分号下求极限的) 定理 7, 可见随机向量的特征函数也可以由如下等式定义:

$$\varphi_\xi(t)=\mathrm{E}e^{i(t,\xi)},\quad t\in\mathbf{R}^n, \tag{4}$$

现在讨论特征函数的基本性质, 但是只证明 $n=1$ 的情形. 一些涉及一般情形的相对重要的结果, 将在练习题中给出.

设 $\xi = \xi(\omega)$ 是随机变量，$F_\xi = F_\xi(x)$ 是其分布函数，而

$$\varphi_\xi(t) = \mathbf{E}e^{it\xi}$$

是其特征函数.

立即可以指出，若 $\eta = a\xi + b$，则

$$\varphi_\eta(t) = \mathbf{E}e^{it\eta} = \mathbf{E}e^{it(a\xi+b)} = e^{itb}\mathbf{E}e^{iat\xi}.$$

因此

$$\varphi_\eta(t) = e^{itb}\varphi_\xi(at). \tag{5}$$

其次，如果 $\xi_1, \xi_2, \cdots, \xi_n$ 是独立随机变量，而 $S_n = \xi_1 + \cdots + \xi_n$，则

$$\varphi_{S_n}(t) = \prod_{i=1}^{n} \varphi_{\xi_i}(t). \tag{6}$$

事实上，由于 (有界) 独立随机变量乘积的数学期望 (无论是实数值随机变量，还是复数值随机变量，可见 §6 定理 6 和练习题 1)，等于其数学期望的乘积，可见

$$\varphi_{S_n}(t) = \mathbf{E}e^{it(\xi_1+\cdots+\xi_n)} = \mathbf{E}e^{it\xi_1}\cdots\mathbf{E}e^{it\xi_n} = \prod_{i=1}^{n}\varphi_{\xi_i}(t).$$

对于用特征函数方法，证明独立随机变量之和的极限定理 (第三章 §3)，性质 (6) 是关键，这时，通过和的各项表示分布函数 F_{S_n} 是相当复杂的，具体地说，

$$F_{S_n} = F_{\xi_1} * \cdots * F_{\xi_n},$$

其中符号 "$*$" 表示分布的卷积 (见 §8 第 4 小节).

下面是特征函数的例子.

例 1 设 ξ 是伯努利随机变量，且 $\mathbf{P}\{\xi = 1\} = p$，$\mathbf{P}\{\xi = 0\} = q$，$p + q = 1$，$0 < p < 1$，则

$$\varphi_\xi(t) = pe^{it} + q.$$

如果 $\xi_1, \cdots, \xi_n$ 是独立与 ξ 同分布的随机变量，则对于 $T_n = (S_n - np)/\sqrt{npq}$，有

$$\begin{aligned}\varphi_{T_n}(t) &= \varphi_{(S_n-np)/\sqrt{npq}}(t) = \mathbf{E}e^{it(S_n-np)/\sqrt{npq}} \\ &= e^{-it\sqrt{np/q}}[pe^{it/\sqrt{npq}} + q]^n = [pe^{it\sqrt{q/(np)}} + qe^{-it\sqrt{p/(nq)}}]^n.\end{aligned} \tag{7}$$

由此可见，当 $n \to \infty$ 时，有

$$\varphi_{T_n}(t) \to e^{-\frac{t^2}{2}}. \tag{8}$$

例 2 设 $\xi \sim \mathscr{N}(m, \sigma^2)$，$|m| < \infty$，$\sigma^2 > 0$. 现在证明

$$\varphi_\xi(t) = e^{itm - \frac{t^2\sigma^2}{2}}. \tag{9}$$

设 $\eta = (\xi - m)/\sigma$，则 $\eta \sim \mathscr{N}(0, 1)$. 而由 (5) 式

$$\varphi_\xi(t) = e^{itm}\varphi_\eta(\sigma t),$$

故只需验证

$$\varphi_\eta(t) = e^{-\frac{t^2}{2}}. \tag{10}$$

下面的一连串式子就可以证明 (10) 式：

$$\begin{aligned}\varphi_\eta(t) = \mathbf{E}e^{it\eta} &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{itx}e^{-\frac{x^2}{2}}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\sum_{n=0}^{\infty}\frac{(itx)^n}{n!}e^{-\frac{x^2}{2}}\,dx \\ &= \sum_{n=0}^{\infty}\frac{(it)^n}{n!}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}x^n e^{-\frac{x^2}{2}}\,dx = \sum_{n=0}^{\infty}\frac{(it)^{2n}}{(2n)!}(2n-1)!! \\ &= \sum_{n=0}^{\infty}\frac{(it)^{2n}}{(2n)!}\frac{(2n)!}{2^n n!} = \sum_{n=0}^{\infty}\left(-\frac{t^2}{2}\right)^n\frac{1}{n!} = e^{-\frac{t^2}{2}},\end{aligned}$$

其中用到如下关系式 (见 §8 练习题 7)：

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}x^{2n}e^{-\frac{x^2}{2}}\,dx = \mathbf{E}\eta^{2n} = (2n-1)!!.$$

例 3 设 ξ 是泊松随机变量，

$$\mathbf{P}\{\xi = k\} = \frac{\lambda^k}{k!}e^{-\lambda},\quad k = 0, 1, \cdots$$

则

$$\mathbf{E}e^{it\xi} = \sum_{k=0}^{\infty}e^{itk}\frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^{it})^k}{k!} = \exp\{\lambda(e^{it} - 1)\}. \tag{11}$$

3. 特征函数的性质 我们在 §9 第 1 小节曾指出，对于 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上每一个分布函数 $F(x)$，都存在一随机变量 ξ，使 $F(x)$ 恰好是 ξ 分布函数. 因此，在叙述特征函数时 (无论是按定义 1，还是按定义 2)，可以局限于考虑随机变量 $\xi = \xi(\omega)$ 的特征函数 $\varphi(t) = \varphi_\xi(t)$.

定理 1 *设随机变量 ξ 的分布函数为 $F = F(x)$，而*

$$\varphi(t) = \mathbf{E}e^{it\xi}$$

是其特征函数.

那么，有下列性质：

1) $|\varphi(t)|\leqslant\varphi(0)=1$;

2) $\varphi(t)$ 对于 $t\in\mathbb{R}$ 一致连续;

3) $\varphi(t)=\overline{\varphi(-t)}$;

4) $\varphi(t)$ 是实数值函数的充分和必要条件是, 其分布函数 $F=F(x)$ 对称, 即

$$\int_B \mathrm{d}F(x)=\int_{-B}\mathrm{d}F(x),\quad B\in\mathscr{B}(\mathbb{R}),\quad -B=\{-x:x\in B\};$$

5) 如果对于某个 $n\geqslant1$, $\mathbf{E}|\xi|^n<\infty$, 则对于一切 $r\leqslant n$, 存在导数 $\varphi^{(r)}(t)$, 且

$$\varphi^{(r)}(t)=\int_{\mathbb{R}}(\mathrm{i}x)^r\mathrm{e}^{\mathrm{i}tx}\mathrm{d}F(x), \tag{12}$$

$$\mathbf{E}\xi^r=\frac{\varphi^{(r)}(0)}{\mathrm{i}^r}, \tag{13}$$

$$\varphi(t)=\sum_{r=0}^{n}\frac{(\mathrm{i}t)^r}{r!}\mathbf{E}\xi^r+\frac{(\mathrm{i}t)^n}{n!}\varepsilon_n(t), \tag{14}$$

其中 $|\varepsilon_n(t)|\leqslant 3\mathbf{E}|\xi|^n$ 且 $\varepsilon_n(t)\to0, t\to0$;

6) 如果存在有限导数 $\varphi^{(2n)}(0)$, 则 $\mathbf{E}\xi^{2n}<\infty$;

7) 如果对于一切 $n\geqslant1$, $\mathbf{E}|\xi|^n<\infty$ 且

$$\overline{\lim_n}\frac{(\mathbf{E}|\xi|^n)^{1/n}}{n}=\frac{1}{T}<\infty,$$

则对于一切 $|t|<T$,

$$\varphi(t)=\sum_{n=0}^{\infty}\frac{(\mathrm{i}t)^n}{n!}\mathbf{E}\xi^n. \tag{15}$$

证明 性质 1) 和 3) 显然. 性质 2) 由下面的估计式

$$|\varphi(t+h)-\varphi(t)|=|\mathbf{E}\mathrm{e}^{\mathrm{i}t\xi}(\mathrm{e}^{\mathrm{i}h\xi}-1)|\leqslant\mathbf{E}|\mathrm{e}^{\mathrm{i}h\xi}-1|$$

和控制收敛定理, 以及当 $h\to0$ 时 $\mathbf{E}|\mathrm{e}^{\mathrm{i}h\xi}-1|\to0$, 可以得到.

证明性质 4). 假设分布函数 $F=F(x)$ 对称, 则对于有界博雷尔奇函数 $g(x)$, 有

$$\int_{\mathbb{R}}g(x)\mathrm{d}F(x)=0$$

(注意, 对于简单奇函数, 这由函数 F 的对称性的定义立即可以得到). 因此

$$\int_{\mathbb{R}}\sin tx\,\mathrm{d}F(x)=0,\ 故\ \varphi(t)=\mathbf{E}\cos t\xi.$$

相反, 设 $\varphi(t)$ 是实数值函数, 则由于 3), 可见

$$\varphi_{-\xi}(t)=\varphi_{\xi}(-t)=\overline{\varphi_{\xi}(t)}=\varphi_{\xi}(t),\quad t\in\mathbb{R}.$$

由此 (这将在下面的定理 2 证明), 可见随机变量 $-\xi$ 和 ξ 的分布函数 $F_{-\xi}$ 和 F_{ξ} 相等. 因此 (根据 §3 的定理 1), 对于 $B\in\mathscr{B}(\mathbb{R})$, 有

$$\mathbf{P}\{\xi\in B\}=\mathbf{P}\{-\xi\in B\}=\mathbf{P}\{\xi\in -B\}.$$

证明性质 5). 如果 $\mathbf{E}|\xi|^n<\infty$, 则由李亚普诺夫不等式 (§6 的 (28) 式), 有 $\mathbf{E}|\xi|^r<\infty, r\leqslant n$.

考虑关系式

$$\frac{\varphi(t+h)-\varphi(t)}{h}=\mathbf{E}\mathrm{e}^{\mathrm{i}t\xi}\left(\frac{\mathrm{e}^{\mathrm{i}h\xi}-1}{h}\right).$$

由于

$$\left|\frac{\mathrm{e}^{\mathrm{i}h\xi}-1}{h}\right|\leqslant|x|\quad 和\quad \mathbf{E}|\xi|<\infty,$$

可见根据控制收敛定理, 存在

$$\lim_{h\to0}\mathbf{E}\mathrm{e}^{\mathrm{i}t\xi}\left(\frac{\mathrm{e}^{\mathrm{i}h\xi}-1}{h}\right),$$

等于

$$\mathbf{E}\mathrm{e}^{\mathrm{i}t\xi}\lim_{h\to0}\left(\frac{\mathrm{e}^{\mathrm{i}h\xi}-1}{h}\right)=\mathrm{i}\mathbf{E}(\xi\mathrm{e}^{\mathrm{i}t\xi})=\mathrm{i}\int_{-\infty}^{\infty}x\mathrm{e}^{\mathrm{i}tx}\mathrm{d}F(x). \tag{16}$$

因此, 存在导数 $\varphi'(t)$ 且

$$\varphi'(t)=\mathrm{i}\mathbf{E}(\xi\mathrm{e}^{\mathrm{i}t\xi})=\mathrm{i}\int_{-\infty}^{\infty}x\mathrm{e}^{\mathrm{i}tx}\mathrm{d}F(x).$$

用归纳法可证明, 导数 $\varphi^{(r)}(t)(1<r\leqslant n)$ 的存在性, 和 (12) 式的正确性.

直接由 (12) 式得出 (13) 式. 现在证明表达式 (14).

由于对于实数 y, 有

$$\mathrm{e}^{\mathrm{i}y}=\cos y+\mathrm{i}\sin y=\sum_{k=0}^{n-1}\frac{(\mathrm{i}y)^k}{k!}+\frac{(\mathrm{i}y)^n}{n!}[\cos\theta_1 y+\mathrm{i}\sin\theta_2 y],$$

其中 $|\theta_1|\leqslant1, |\theta_2|\leqslant1$, 可见

$$\mathrm{e}^{\mathrm{i}t\xi}=\sum_{k=0}^{n-1}\frac{(\mathrm{i}t\xi)^k}{k!}+\frac{(\mathrm{i}t\xi)^n}{n!}[\cos\theta_1(\omega)t\xi+\mathrm{i}\sin\theta_2(\omega)t\xi], \tag{17}$$

且

$$\mathbf{E}\mathrm{e}^{\mathrm{i}t\xi}=\sum_{k=0}^{n-1}\frac{(\mathrm{i}t)^k}{k!}\mathbf{E}\xi^k+\frac{(\mathrm{i}t)^n}{n!}[\mathbf{E}\xi^n+\varepsilon_n(t)], \tag{18}$$

其中

$$\varepsilon_n(t)=\mathbf{E}(\xi^n[\cos\theta_1(\omega)t\xi+\mathrm{i}\sin\theta_2(\omega)t\xi-1]).$$

显然, $|\varepsilon_n(t)| \leqslant 3\mathbf{E}|\xi^n|$; 而且根据控制收敛定理, 当 $t\to 0$ 时, $\varepsilon_n(t)\to 0$.

证明性质 6). 用归纳法证明. 首先假设导数 $\varphi''(0)$ 存在并且有限. 证明 $\mathbf{E}\xi^2<\infty$. 根据洛必达 (G. L'hospital) 法则和法图引理, 有

$$\begin{aligned}\varphi''(0) &= \lim_{h\to 0}\frac{1}{2}\left[\frac{\varphi'(2h)-\varphi'(0)}{2h}+\frac{\varphi'(0)-\varphi'(-2h)}{2h}\right]\\&=\lim_{h\to 0}\frac{2\varphi'(2h)-2\varphi'(-2h)}{8h}=\lim_{h\to 0}\frac{1}{4h^2}[\varphi(2h)-2\varphi(0)+\varphi(-2h)]\\&=\lim_{h\to 0}\int_{-\infty}^{\infty}\left(\frac{e^{ihx}-e^{-ihx}}{2h}\right)^2\mathrm{d}F(x)=-\lim_{h\to 0}\int_{-\infty}^{\infty}\left(\frac{\sin hx}{hx}\right)^2x^2\mathrm{d}F(x)\\&\leqslant-\int_{-\infty}^{\infty}\lim_{h\to 0}\left(\frac{\sin hx}{hx}\right)^2x^2\mathrm{d}F(x)=-\int_{-\infty}^{\infty}x^2\mathrm{d}F(x).\end{aligned}$$

从而

$$\int_{-\infty}^{\infty}x^2\mathrm{d}F(x)\leqslant-\varphi''(0)<\infty.$$

现在假设 $\varphi^{(2k+2)}(0)$ 存在而且有限

$$\int_{-\infty}^{\infty}x^{2k}\mathrm{d}F(x)<\infty.$$

如果

$$\int_{-\infty}^{\infty}x^{2k}\mathrm{d}F(x)=0,\quad 则\quad \int_{-\infty}^{\infty}x^{2k+2}\mathrm{d}F(x)=0.$$

因此, 我们将假设

$$\int_{-\infty}^{\infty}x^{2k}\mathrm{d}F(x)>0.$$

那么, 根据性质 5), 有

$$\varphi^{(2k)}(t)=\int_{-\infty}^{\infty}(ix)^{2k}e^{itx}\mathrm{d}F(x),$$

因而

$$(-1)^k\varphi^{(2k)}(t)=\int_{-\infty}^{\infty}e^{itx}\mathrm{d}G(x),$$

其中 $G(x)=\int_{-\infty}^{x}u^{2k}\mathrm{d}F(u)$.

从而, 函数 $(-1)^k\varphi^{(2k)}(t)G^{-1}(\infty)$ 是概率分布 $G(x)\times G^{-1}(\infty)$ 的特征函数, 并且根据已证明的, 有

$$G^{-1}(\infty)\int_{-\infty}^{\infty}x^2\mathrm{d}G(x)<\infty.$$

由于 $G^{-1}(\infty)>0$, 可见

$$\int_{-\infty}^{\infty}x^{2k+2}\mathrm{d}F(x)=\int_{-\infty}^{\infty}x^2\mathrm{d}G(x)<\infty.$$

证明性质 7). 设 $0<t_0<T$. 则利用斯特林公式 (第一章 §2 (6) 式), 有

$$\overline{\lim}\frac{(\mathbf{E}|\xi|^n)^{1/n}}{n}<\frac{1}{t_0}\Rightarrow\overline{\lim}\frac{(\mathbf{E}|\xi|^n t_0^n)^{1/n}}{n}<1\Rightarrow\overline{\lim}\left(\frac{\mathbf{E}|\xi|^n t_0^n}{n!}\right)^{1/n}<1.$$

从而, 根据柯西准则, 级数 $\sum\mathbf{E}|\xi|^r t_0^r/r!$ 收敛, 因此对于任意 $|t|\leqslant t_0$, 级数 $\sum_{r=0}^{\infty}\mathbf{E}\xi^r(it)^r/r!$ 收敛. 而由 (14) 式

$$\varphi(t)=\sum_{r=0}^{n}\frac{(it)^r}{r!}\mathbf{E}\xi^r+R_n(t),\ n\geqslant 1,\ 其中\ |R_n(t)|\leqslant\frac{3|t|^n}{n!}\mathbf{E}|\xi|^n.$$

于是, 对于一切 $|t|<T$, 有

$$\varphi(t)=\sum_{r=0}^{\infty}\frac{(it)^r}{r!}\mathbf{E}\xi^r. \tag{18}$$

注 1 与 (14) 式的证明类似, 可以验证, 如果对于某个 $n\geqslant 1$, $\mathbf{E}|\xi|^n<\infty$, 则

$$\varphi(t)=\sum_{k=0}^{n}\frac{i^k(t-s)^k}{k!}\int_{-\infty}^{\infty}x^k e^{isx}\mathrm{d}F(x)+\frac{i^n(t-s)^n}{n!}\varepsilon_n(t-s), \tag{19}$$

其中 $|\varepsilon_n(t-s)|\leqslant 3\mathbf{E}|\xi^n|$, 而当 $t-s\to 0$ 时 $\varepsilon_n(t-s)\to 0$.

注 2 关于性质 7) 中的条件, 参见下面的第 9 小节中关于 "矩问题的唯一性" 的内容.

4. 特征函数唯一决定分布函数 下面的定理表明, 特征函数唯一决定分布函数.

定理 2 (唯一性) *假设 F 和 G 是具有同一特征函数的分布函数, 即对于一切 $t\in\mathbb{R}$,*

$$\int_{-\infty}^{\infty}e^{itx}\mathrm{d}F(x)=\int_{-\infty}^{\infty}e^{itx}\mathrm{d}G(x). \tag{20}$$

那么 $F(x)=G(x)$.

证明 固定 $a,b\in\mathbb{R}$, $\varepsilon>0$, 并且考虑图 33 所表示的函数 $f^\varepsilon=f^\varepsilon(x)$. 现在证明:

$$\int_{-\infty}^{\infty}f^\varepsilon(x)\mathrm{d}F(x)=\int_{-\infty}^{\infty}f^\varepsilon(x)\mathrm{d}G(x). \tag{21}$$

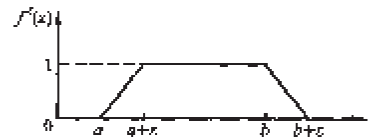


图 33

设 $n \geqslant 0$ 满足 $[a, b+\varepsilon] \subseteq [-n, n]$; 数列 $\{\delta_n\}$ 满足 $1 \geqslant \delta_n \downarrow 0, n \to \infty$. $f^\varepsilon = f^\varepsilon(x)$ 作为 $[-n, n]$ 上的连续且在端点上等值的函数, 可以用三角多项式一致逼近的 (维尔斯特拉斯—斯通 (M.Stone)) 定理, 即存在有限和

$$f_n^\varepsilon(x) = \sum_k a_k \exp\left(\mathrm{i}\pi x \frac{k}{n}\right), \tag{22}$$

使

$$\sup_{-n \leqslant x \leqslant n} |f^\varepsilon(x) - f_n^\varepsilon(x)| \leqslant \delta_n. \tag{23}$$

对于所有 $x \in \mathbb{R}$ 将周期函数 $f_n^\varepsilon(x)$ 进行开拓, 并注意到

$$\sup_x |f_n^\varepsilon(x)| \leqslant 2.$$

那么, 由于 (20) 式

$$\int_{-\infty}^{\infty} f_n^\varepsilon(x)\mathrm{d}F(x) = \int_{-\infty}^{\infty} f_n^\varepsilon(x)\mathrm{d}G(x),$$

可见

$$\begin{aligned}
&\left|\int_{-\infty}^{\infty} f^\varepsilon(x)\mathrm{d}F(x) - \int_{-\infty}^{\infty} f^\varepsilon(x)\mathrm{d}G(x)\right| = \left|\int_{-n}^{n} f^\varepsilon \mathrm{d}F - \int_{-n}^{n} f^\varepsilon \mathrm{d}G\right| \\
\leqslant &\left|\int_{-n}^{n} f_n^\varepsilon \mathrm{d}F - \int_{-n}^{n} f_n^\varepsilon \mathrm{d}G\right| + 2\delta_n \\
\leqslant &\left|\int_{-\infty}^{\infty} f_n^\varepsilon \mathrm{d}F - \int_{-\infty}^{\infty} f_n^\varepsilon \mathrm{d}G\right| + 2\delta_n + 2F(\overline{[-n, n]}) + 2G(\overline{[-n, n]}),
\end{aligned} \tag{24}$$

其中

$$F(A) = \int_A \mathrm{d}F(x), \quad G(A) = \int_A \mathrm{d}G(x).$$

当 $n \to \infty$ 时 (24) 式的右侧趋向 0, 从而 (21) 式得证.

当 $\varepsilon \to 0$ 时 $f^\varepsilon(x) \to I_{(a,b]}(x)$. 因此根据控制收敛定理, 由 (21) 式, 可见

$$\int_{-\infty}^{\infty} I_{(a,b]}(x)\mathrm{d}F(x) = \int_{-\infty}^{\infty} I_{(a,b]}(x)\mathrm{d}G(x),$$

即 $F(b) - F(a) = G(b) - G(a)$. 而由于 a 和 b 是任意的, 可见对于一切 $x \in \mathbb{R}$, $F(x) = G(x)$. 于是, 定理 2 得证. □

5. 逆转公式　上一个定理说明, 分布函数 $F = F(x)$ 唯一决定于其特征函数 $\varphi = \varphi(t)$. 下面的定理将给出函数 F 通过函数 φ 的表达式.

定理 3 (逆转公式)　*假设 $F = F(x)$ 是分布函数, 而*

$$\varphi(t) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}tx}\mathrm{d}F(x)$$

是其特征函数.

a) 对于函数 $F = F(x)$ 的任意两个连续点 a 和 $b (a < b)$, 有

$$F(b) - F(a) = \lim_{c \to \infty} \frac{1}{2\pi} \int_{-c}^{c} \frac{\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t} \varphi(t)\mathrm{d}t. \tag{25}$$

b) 如果 $\int_{-\infty}^{\infty} |\varphi(t)|\mathrm{d}t < \infty$, 则分布函数 $F(x)$ 有密度 $f(x)$:

$$F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y, \tag{26}$$

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}tx} \varphi(t)\mathrm{d}t. \tag{27}$$

证明　首先注意到, 如果函数 $F(x)$ 有密度 $f(x)$, 则

$$\varphi(t) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}tx} f(x)\mathrm{d}x, \tag{28}$$

因此 (27) 式恰好是 (可积) 函数 $\varphi(x)$ 的傅里叶变换. 在 (27) 式两侧同时积分, 并运用傅比尼定理, 得

$$\begin{aligned}
F(b) - F(a) &= \int_a^b f(x)\mathrm{d}x = \frac{1}{2\pi} \int_a^b \left[\int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}tx} \varphi(t)\mathrm{d}t\right] \mathrm{d}x \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi(t) \left[\int_a^b \mathrm{e}^{-\mathrm{i}tx}\mathrm{d}x\right] \mathrm{d}t = \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi(t) \frac{\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t}\mathrm{d}t.
\end{aligned}$$

在完成了对 (25) 式一定的说明之后, 现在开始证明 (25) 式.

a) 有

$$\begin{aligned}
\Phi_c &\equiv \frac{1}{2\pi} \int_{-c}^{c} \frac{\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t} \varphi(t)\mathrm{d}t = \frac{1}{2\pi} \int_{-c}^{c} \frac{\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t} \left[\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}tx}\mathrm{d}F(x)\right] \mathrm{d}t \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_{-c}^{c} \frac{\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t} \mathrm{e}^{\mathrm{i}tx}\mathrm{d}t\right] \mathrm{d}F(x) = \int_{-\infty}^{\infty} \Psi_c(x)\mathrm{d}F(x),
\end{aligned} \tag{29}$$

其中, 设

$$\Psi_c(x) = \frac{1}{2\pi} \int_{-c}^{c} \frac{\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t} \mathrm{e}^{\mathrm{i}tx}\mathrm{d}t,$$

并且用到傅比尼定理; 由于

$$\left|\frac{\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t} \mathrm{e}^{\mathrm{i}tx}\right| = \left|\frac{\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t}\right| = \left|\int_a^b \mathrm{e}^{-\mathrm{i}tx}\mathrm{d}x\right| \leqslant b - a,$$

且

$$\int_{-c}^{c} \int_{-\infty}^{\infty} (b - a)\mathrm{d}F(x)\mathrm{d}t \leqslant 2c(b - a) < \infty,$$

可见应用傅比尼定理是合理的. 其次,

$$\Psi_c(x)=\frac{1}{2\pi}\int_{-c}^{c}\frac{\sin t(x-a)-\sin t(x-b)}{t}\mathrm{d}t$$
$$=\frac{1}{2\pi}\int_{-c(x-a)}^{c(x-a)}\frac{\sin v}{v}\mathrm{d}v-\frac{1}{2\pi}\int_{-c(x-b)}^{c(x-b)}\frac{\sin u}{u}\mathrm{d}u. \tag{30}$$

函数

$$g(s,t)=\int_s^t\frac{\sin v}{v}\mathrm{d}v$$

关于 s 和 t 一致连续, 且当 $s\downarrow-\infty$ 和 $t\uparrow\infty$ 时

$$g(s,t)\to\pi. \tag{31}$$

因此, 存在常数 C, 使对于一切 c 和 x, 有 $|\Psi_c(x)|<C<\infty$. 此外, 由 (30) 和 (31) 式, 可见

$$\Psi_c(x)\to\Psi(x),\ c\to\infty,$$

其中

$$\Psi(x)=\begin{cases}0, & 若\ x<a\ 或\ x>b,\\ 1/2, & 若\ x=a\ 或\ x=b,\\ 1, & 若\ a<x<b.\end{cases}$$

设 μ 是 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ 上的测度, 且 $\mu(a,b]=F(b)-F(a)$. 那么, 运用控制收敛定理和 §6 练习题 1, 当 $c\to\infty$ 时, 有

$$\Phi_c=\int_{-\infty}^{\infty}\Psi_c(x)\mathrm{d}F(x)\to\int_{-\infty}^{\infty}\Psi(x)\mathrm{d}F(x)=\mu(a,b)+\frac{1}{2}\mu\{a\}+\frac{1}{2}\mu\{b\}$$
$$=F(b-)-F(a)+\frac{1}{2}[F(a)-F(a-)+F(b)-F(b-)]$$
$$=\frac{F(b)+F(b-)}{2}-\frac{F(a)+F(a-)}{2}=F(b)-F(a),$$

其中最后一个等式对于函数 $F(x)$ 的一切连续点 a 和 b 成立.

于是, (25) 式得证.

b) 设 $\int_{-\infty}^{\infty}|\varphi(t)|\mathrm{d}t<\infty$. 记

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{e}^{-\mathrm{i}tx}\varphi(t)\mathrm{d}t.$$

由控制收敛定理知, 此函数对 x 连续, 故在区间 $[a,b]$ 上可积. 所以仍可以应用傅比尼定理, 可得对于一切函数 $F(x)$ 的一切连续点 a 和 b, 有

$$\int_a^b f(x)\mathrm{d}x=\int_a^b\frac{1}{2\pi}\left[\int_{-\infty}^{\infty}\mathrm{e}^{-\mathrm{i}tx}\varphi(t)\mathrm{d}t\right]\mathrm{d}x$$
$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}\varphi(t)\left[\int_a^b\mathrm{e}^{-\mathrm{i}tx}\mathrm{d}x\right]\mathrm{d}t=\lim_{c\to\infty}\frac{1}{2\pi}\int_{-c}^{c}\varphi(t)\left[\int_a^b\mathrm{e}^{-\mathrm{i}tx}\mathrm{d}x\right]\mathrm{d}t$$
$$=\lim_{c\to+\infty}\int_{-c}^{c}\frac{1}{2\pi}\frac{\mathrm{e}^{-\mathrm{i}ta}-\mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t}\varphi(t)\mathrm{d}t=F(b)-F(a).$$

由此可见,

$$F(x)=\int_{-\infty}^{x}f(y)\mathrm{d}y,\quad x\in\mathbb{R}.$$

而且因为 $f(x)$ 连续, 且 $F(x)$ 是不减函数, 所以 $f(x)$ 是 $F(x)$ 的密度. □

注　逆转公式 (25) 给出了定理 2 的另一证明.

定理 4　*随机向量 $\xi=(\xi_1,\cdots,\xi_n)$ 的分量独立的充分和必要条件是, 其特征函数等于各分量特征函数的乘积:*

$$\mathrm{E}\mathrm{e}^{\mathrm{i}(t_1\xi_1+\cdots+t_n\xi_n)}=\prod_{k=1}^{n}\mathrm{E}\mathrm{e}^{\mathrm{i}t_k\xi_k},\ (t_1,\cdots,t_n)\in\mathbb{R}^n.$$

证明　由练习题 1, 可见必要性成立. 为证明充分性, 记 $F(x_1,\cdots,x_n)$ 为随机向量 $\xi=(\xi_1,\cdots,\xi_n)$ 的分布函数, 而 $F_k(x)$ 为 $\xi_k(1\leqslant k\leqslant n)$ 的分布函数; $G=G(x_1,\cdots,x_n)=F_1(x_1)\cdots F_n(x_n)$. 那么, 根据傅比尼定理, 对于一切 $(t_1,\cdots,t_n)\in\mathbb{R}^n$, 有

$$\int_{\mathbb{R}^n}\mathrm{e}^{\mathrm{i}(t_1x_1+\cdots+t_nx_n)}\mathrm{d}G(x_1,\cdots,x_n)=\prod_{k=1}^{n}\int_{\mathbb{R}}\mathrm{e}^{\mathrm{i}t_kx_k}\mathrm{d}F_k(x_k)$$
$$=\prod_{k=1}^{n}\mathrm{E}\mathrm{e}^{\mathrm{i}t_k\xi_k}=\mathrm{E}\mathrm{e}^{\mathrm{i}(t_1\xi_1+\cdots+t_n\xi_n)}=\int_{\mathbb{R}^n}\mathrm{e}^{\mathrm{i}(t_1x_1+\cdots+t_nx_n)}\mathrm{d}F(x_1,\cdots,x_n).$$

于是, 由定理 2 (确切地说, 由定理 2 的多元类似; 亦见练习题 3) 可见, 根据 §5 的定理, 随机变量 $\xi_1,\cdots,\xi_n$ 独立. □

6. 特征函数的必要条件　在定理 1 中包含特征函数的某些必要条件. 例如, 假如函数 $\varphi=\varphi(t)$ 不满足该定理的前 3 个条件之一, 则这一函数就不是特征函数.

验证我们所感兴趣的函数 $\varphi=\varphi(t)$ 是否特性函数, 是一项复杂的事. 下面给出这方面一些结果 (不加证明).

定理 (博赫纳 — 辛钦 [S. Bochner – А. Я. Хинчин])　*设 $\varphi(t),t\in\mathbb{R}$, 是连续函数且 $\varphi(0)=1$. $\varphi(t)$ 是特征函数的充分和必要条件是, 它是非负定函数, 即对于任意实数 $t_1,\cdots,t_n$ 和对于任意复数 $\lambda_1,\cdots,\lambda_n,n=1,2,\cdots$, 有*

$$\sum_{i,j=1}^{n}\varphi(t_i-t_j)\lambda_i\bar{\lambda}_j\geqslant 0. \tag{32}$$

条件 (32) 的必要性显然: 因为如果

$$\varphi(t)=\int_{-\infty}^{\infty} e^{itx}\,dF(x),$$

则

$$\sum_{k,j=1}^{n}\varphi(t_k-t_j)\lambda_k\overline{\lambda}_j=\int_{-\infty}^{\infty}\left|\sum_{k=1}^{n}\lambda_k e^{it_k x}\right|^2 dF(x)\geqslant 0.$$

条件 (32) 充分性的证明比较困难 (参见 [69] 的 T.2, X IX.2).

定理 (波利亚 [G. Polia]) 设 $\varphi(t)$ 是连续函数、偶函数, 在 $(-\infty,0)$ 上 (因此在 $(0,\infty)$ 上) 是凹函数 (向下凸函数), 且满足条件: $\varphi(t)\geqslant 0, \varphi(0)=1, \varphi(t)\to 0\ (t\to\infty)$, 那么, $\varphi(t)$ 是特征函数 (参见 [69] 的 T.2, X V.2).

这一定理提供了构造特征函数的相当方便的方法. 例如, 函数

$$\varphi_1(t)=e^{-|t|},$$

$$\varphi_2(t)=\begin{cases}1-|t|, & 若\ |t|\leqslant 1,\\ 0, & 若\ |t|>1\end{cases}$$

就是这样的函数.

图 34 所表示的函数 $\varphi_3(x)$ 也是特征函数. 在区间 $[-a,a]$ 上 $\varphi_3(x)=\varphi_2(x)$. 然而, 与它们对应的分布函数 F_2 和 F_3 显然不同. 这个例子表明, 在有限区间上特征函数相同, 一般分布函数未必相同.

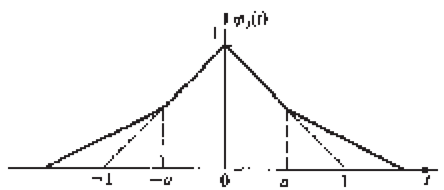


图 34

定理 (马钦凯维奇 [J. Marcinkiewicz]) 如果特征函数 $\varphi(t)$ 具有形式 $\exp\mathscr{P}(t)$, 其中 $\mathscr{P}(t)$ 是多项式, 则此多项式的次数不可能大于 2(见 [135, 7.3]).

例如, 由此定理可见, 函数 $\exp(-t^4)$ 不是特征函数.

7. 某些特殊分布的特征函数 下面的定理是用例子表明, 如何根据随机变量的特征函数, 作出关于此变量的不平凡的论断.

定理 5 设 $\varphi_\xi(t)$ 是随机变量 ξ 的特征函数.

a) 如果对于某个 $t_0\neq 0$, $|\varphi_\xi(t_0)|=1$, 则 ξ 是步长为 $h=2\pi/|t_0|$ 的格点随机变量, 即

$$\sum_{n=-\infty}^{\infty}\mathbf{P}\{\xi=a+nh\}=1, \tag{33}$$

其中 a 是常数.

b) 如果对于两个不同的点 t 和 αt, 其中 α 是无理数, 有 $|\varphi_\xi(t)|=|\varphi_\xi(\alpha t)|=1$, 则随机变量 ξ 是退化的:

$$\mathbf{P}\{\xi=a\}=1,$$

其中 a 是常数.

c) 如果 $|\varphi_\xi(t)|\equiv 1$, 则随机变量 ξ 是退化的.

证明 a) 如果 $|\varphi_\xi(t_0)|=1, t_0\neq 0$, 则存在实数 a, 使 $\varphi(t_0)=e^{it_0a}$, 那么,

$$e^{-it_0a}\int_{-\infty}^{\infty}e^{it_0x}dF(x)=1\Rightarrow\int_{-\infty}^{\infty}e^{it_0(x-a)}dF(x)=1\Rightarrow$$
$$\Rightarrow\int_{-\infty}^{\infty}\cos t_0(x-a)\,dF(x)=1\Rightarrow\int_{-\infty}^{\infty}[1-\cos t_0(x-a)]\,dF(x)=0.$$

由于 $1-\cos t_0(x-a)\geqslant 0$, 由 §6 第 2 小节性质 **H**, 可见 $(\mathbf{P}-\text{a.e.})$ 有

$$1=\cos t_0(\xi-a),$$

而这与 (33) 式等价.

b) 由假设 $|\varphi_\xi(t)|=|\varphi_\xi(\alpha t)|=1$ 和 (33) 式, 可见

$$\sum_{n=-\infty}^{\infty}\mathbf{P}\left\{\xi=a+\frac{2\pi n}{t}\right\}=\sum_{m=-\infty}^{\infty}\mathbf{P}\left\{\xi=b+\frac{2\pi m}{\alpha t}\right\}=1.$$

如果 ξ 是非退化的, 则在集合

$$\left\{a+\frac{2\pi n}{t}, n=0,\pm 1,\cdots\right\}和\left\{b+\frac{2\pi m}{\alpha t}, m=0,\pm 1,\cdots\right\}$$

上至少存在两个不同的点:

$$a+\frac{2\pi n_1}{t}=b+\frac{2\pi m_1}{\alpha t},\quad a+\frac{2\pi n_2}{t}=b+\frac{2\pi m_2}{\alpha t},$$

由此得

$$\frac{2\pi}{t}(n_1-n_2)=\frac{2\pi}{\alpha t}(m_1-m_2),$$

而这与假设 "α 是无理数" 矛盾, 故 b) 正确. 此外, 由命题 b) 可得命题 c). □

8. **随机变量的半不变量和矩的联系** 设 $\xi=(\xi_1,\cdots,\xi_k)$ 是随机向量，函数

$$\varphi_\xi(t)=\mathbf{E}e^{i(t,\xi)},\quad t=(t_1,\cdots,t_k)$$

是它的特征函数. 假设对于某个 $n\geqslant 1$, $\mathbf{E}|\xi_i|^n<\infty$, $i=1,\cdots,k$. 由 §6 的赫尔德不等式 (29) 和李雅普诺夫不等式 (27)，可见对于一切非负 $\nu_1,\cdots,\nu_k$，且 $\nu_1+\cdots+\nu_k\leqslant n$，(混合) 矩 $\mathbf{E}(\xi_1^{\nu_1}\cdots\xi_k^{\nu_k})$ 存在.

如定理 1，由此对于 $\nu_1+\cdots+\nu_k\leqslant n$，可见偏导数

$$\frac{\partial^{\nu_1+\cdots+\nu_k}}{\partial t_1^{\nu_1}\cdots\partial t_k^{\nu_k}}\varphi_\xi(t_1,\cdots,t_k)$$

的存在性和连续性. 那么，将 $\varphi_\xi(t_1,\cdots,t_k)$ 展为泰勒级数，可得

$$\varphi_\xi(t_1,\cdots,t_k)=\sum_{\nu_1+\cdots+\nu_k\leqslant n}\frac{i^{\nu_1+\cdots+\nu_k}}{\nu_1!\cdots\nu_k!}m_\xi^{(\nu_1,\cdots,\nu_k)}t_1^{\nu_1}\cdots t_k^{\nu_k}+o(|t|^n),\tag{34}$$

其中 $|t|=|t_1|+\cdots+|t_k|$，而

$$m_\xi^{(\nu_1,\cdots,\nu_k)}=\mathbf{E}\xi_1^{\nu_1}\cdots\xi_k^{\nu_k}$$

是 $\nu=(\nu_1,\cdots,\nu_k)$ 阶 (混合) 矩.

函数 $\varphi_\xi(t_1,\cdots,t_k)$ 连续，$\varphi_\xi(0,\cdots,0)=1$，因此在 $0=(0,\cdots,0)$ 的某邻域内 $(|t|<\delta)$，$\varphi_\xi(t_1,\cdots,t_k)\neq 0$. 在此邻域内存在连续偏导数

$$\frac{\partial^{\nu_1+\cdots+\nu_k}}{\partial t_1^{\nu_1}\cdots\partial t_k^{\nu_k}}\ln\varphi_\xi(t_1,\cdots,t_k),$$

其中将 $\ln z$ 理解为对数的主值 (假如 $z=re^{i\theta}$，$-\pi<\theta\leqslant\pi$，则规定 $\ln z$ 为 $\ln r+i\theta$).

因此 $\ln\varphi_\xi(t_1,\cdots,t_k)$ 可以表示为泰勒公式

$$\ln\varphi_\xi(t_1,\cdots,t_k)=\sum_{\nu_1+\cdots+\nu_k\leqslant n}\frac{i^{\nu_1+\cdots+\nu_k}}{\nu_1!\cdots\nu_k!}s_\xi^{(\nu_1,\cdots,\nu_k)}t_1^{\nu_1}\cdots t_k^{\nu_k}+o(|t|^n),\tag{35}$$

其中系数 $s_\xi^{(\nu_1,\cdots,\nu_k)}$ 称做随机向量 $\xi=(\xi_1,\cdots,\xi_k)$ 的 $\nu=(\nu_1,\cdots,\nu_k)$ 阶 (混合) 半不变量或累积量.

注意，如果 ξ 和 η 是两个独立随机向量，则

$$\ln\varphi_{\xi+\eta}(t)=\ln\varphi_\xi(t)+\ln\varphi_\eta(t),\tag{36}$$

因此

$$s_{\xi+\eta}^{(\nu_1,\cdots,\nu_k)}=s_\xi^{(\nu_1,\cdots,\nu_k)}+s_\eta^{(\nu_1,\cdots,\nu_k)}.\tag{37}$$

(此正是这一条性质说明 "半不变量" 名称的合理性.)

为简化记号，并且使 (34) 和 (35) 式具有 "一维" 的形式，我们引进如下记号. 如果 $\nu=(\nu_1,\cdots,\nu_k)$ 是具有非负整数分量的向量，则设

$$\nu!=\nu_1!\cdots\nu_k!,\quad|\nu|=\nu_1+\cdots+\nu_k,\quad t^\nu=t_1^{\nu_1}\cdots t_k^{\nu_k}.$$

亦设

$$s_\xi^{(\nu)}=s_\xi^{(\nu_1,\cdots,\nu_k)},\ m_\xi^{(\nu)}=m_\xi^{(\nu_1,\cdots,\nu_k)}.$$

那么，(34) 和 (35) 式具有如下形式:

$$\varphi_\xi(t)=\sum_{|\nu|\leqslant n}\frac{i^{|\nu|}}{\nu!}m_\xi^{(\nu)}t^\nu+o(|t|^n),\tag{38}$$

$$\ln\varphi_\xi(t)=\sum_{|\nu|\leqslant n}\frac{i^{|\nu|}}{\nu!}s_\xi^{(\nu)}t^\nu+o(|t|^n).\tag{39}$$

下面的定理及其系给出矩和半不变量的联系公式.

定理 6 设 $\xi=(\xi_1,\cdots,\xi_k)$ 是随机向量，满足 $\mathbf{E}|\xi_i|^n<\infty$, $i=1,\cdots,k$, $n\geqslant 1$. 那么，对于一切 $\nu=(\nu_1,\cdots,\nu_k)$ 且 $|\nu|\leqslant n$，有

$$m_\xi^{(\nu)}=\sum_{\lambda^{(1)}+\cdots+\lambda^{(q)}=\nu}\frac{1}{q!}\frac{\nu!}{\lambda^{(1)}!\cdots\lambda^{(q)}!}\prod_{p=1}^{q}s_\xi^{(\lambda^{(p)})},\tag{40}$$

$$s_\xi^{(\nu)}=\sum_{\lambda^{(1)}+\cdots+\lambda^{(q)}=\nu}\frac{(-1)^{q-1}}{q}\frac{\nu!}{\lambda^{(1)}!\cdots\lambda^{(q)}!}\prod_{p=1}^{q}m_\xi^{(\lambda^{(p)})},\tag{41}$$

其中 $\sum\limits_{\lambda^{(1)}+\cdots+\lambda^{(q)}=\nu}$ 表示对于一切满足 $|\lambda^{(p)}|>0$, $\lambda^{(1)}+\cdots+\lambda^{(q)}=\nu$ 的 "有序非负整数分量的向量 $\lambda^{(p)}$" 求和.

证明 由于

$$\varphi_\xi(t)=\exp\{\ln\varphi_\xi(t)\},$$

则将 $\exp\{\cdots\}$ 展成泰勒级数并考虑到 (39) 式，得

$$\varphi_\xi(t)=1+\sum_{q=1}^{n}\frac{1}{q!}\left(\sum_{1\leqslant|\lambda|\leqslant n}\frac{i^{|\lambda|}}{\lambda!}s_\xi^{(\lambda)}t^\lambda\right)^q+o(|t|^n).\tag{42}$$

比较 (38) 和 (42) 式右侧 t^λ 的项，并考虑到 $|\lambda^{(1)}|+\cdots+|\lambda^{(q)}|=|\lambda^{(1)}+\cdots+\lambda^{(q)}|$，得 (40) 式.

其次，

$$\ln\varphi_\xi(t)=\ln\left[1+\sum_{1\leqslant|\lambda|\leqslant n}\frac{i^{|\lambda|}}{\lambda!}m_\xi^{(\lambda)}t^\lambda+o(|t|^n)\right].\tag{43}$$

对较小的 z 有如下展开式

$$\ln(1+z)=\sum_{q=1}^{n}\frac{(-1)^{q-1}}{q}z^q+o(z^n).$$

将此展开式用于 (43) 式, 然后将项 t^ν 的系数与 (38) 式右侧相应项的系数比较, 得 (41) 式. □

系 1 在矩和半不变量之间有下列公式:

$$m_\xi^{(\nu)}=\sum_{\{r_1\lambda^{(1)}+\cdots+r_x\lambda^{(x)}=\nu\}}\frac{1}{r_1!\cdots r_x!}\frac{\nu!}{(\lambda^{(1)}!)^{r_1}\cdots(\lambda^{(x)}!)^{r_x}}\prod_{j=1}^{x}[s_\xi^{(\lambda^{(j)})}]^{r_j},\tag{44}$$

$$s_\xi^{(\nu)}=\sum_{\{r_1\lambda^{(1)}+\cdots+r_x\lambda^{(x)}=\nu\}}\frac{(-1)^{q-1}(q-1)!}{r_1!\cdots r_x!}\frac{\nu!}{(\lambda^{(1)}!)^{r_1}\cdots(\lambda^{(x)}!)^{r_x}}\prod_{j=1}^{x}[m_\xi^{(\lambda^{(j)})}]^{r_j},\tag{45}$$

其中 $\sum_{\{r_1\lambda^{(1)}+\cdots+r_x\lambda^{(x)}=\nu\}}$ 表示对于一切满足 $|\lambda^{(j)}|>0$, $r_1\lambda^{(1)}+\cdots+r_x\lambda^{(x)}=\nu$ 的 "有序正整数 r_j 的数组" 求和.

为证明 (44) 式, 假设 (40) 式中相应地有 r_1 个向量等于 $\lambda^{(k_1)},\cdots,r_x$ 个向量等于 $\lambda^{(k_x)}$ $(r_j>0, r_1+\cdots+r_x=q)$, 而且所有向量 $\lambda^{(k_j)}$ 两两不同. 恰好存在 $q!/(r_1!\cdots r_x!)$ 个不同的向量组 (精确到顺序) 与向量组 $(\lambda^{(1)},\cdots,\lambda^{(q)})$ 重合. 然而, 假如两个向量组, 例如, $(\lambda^{(1)},\cdots,\lambda^{(q)})$ 和 $(\tilde\lambda^{(1)},\cdots,\tilde\lambda^{(q)})$ 仅区别于分量的顺序, 则

$$\prod_{p=1}^{q}s_\xi^{(\lambda^{(p)})}=\prod_{p=1}^{q}s_\xi^{(\tilde\lambda^{(p)})}.$$

所以, 可以将仅区别于顺序的数组视为同一数组, 则由 (40) 式得 (44) 式.

类似地可以由 (41) 式得 (45) 式.

系 2 考虑 $\nu=(1,\cdots,1)$ 的特殊情形. 这时, 矩 $m_\xi^{(\nu)}=\mathbf{E}\xi_1\cdots\xi_k$ 以及相应的半不变量称做简单的.

联系简单矩和半不变量的公式, 可以由上面给出的公式得到. 不过, 最好将其写为另一种形式.

为此引进下面的记号.

设 $I_\xi=\{1,2,\cdots,k\}$ 是向量 $\xi=(\xi_1,\cdots,\xi_k)$ 之分量下标的集合. 假如 $I\subseteq I_\xi$, 则以 ξ_I 表示由下标属于 I 的向量 $\xi=(\xi_1,\cdots,\xi_k)$ 之分量构成的向量. 设 $\chi(I)$ 表示向量 $\{\chi_1,\cdots,\chi_n\}$, 其中, 若 $i\in I$, 则 $\chi_i=1$; 若 $i\notin I$, 则 $\chi_i=0$. 这些向量与集合 $I\subseteq I_\xi$ 一一对应. 因此, 记

$$m_\xi(I)=m_\xi^{(\chi(I))},\quad s_\xi(I)=s_\xi^{(\chi(I))}.$$

换句话说, $m_\xi(I)$ 和 $s_\xi(I)$ 是向量 $\xi=(\xi_1,\cdots,\xi_k)$ 之子向量 ξ_I 是简单矩和半不变量.

其次, 根据定义 (见第一章 §1 第 3 小节), 集合 I 的分割, 指不相交非空集合 I_p 组, 如果 $\sum_p I_p=I$.

鉴于这些记号, 有如下公式:

$$m_\xi(I)=\sum_{\sum_{p=1}^{q}I_p=I}\prod_{p=1}^{q}s_\xi(I_p),\tag{46}$$

$$s_\xi(I)=\sum_{\sum_{p=1}^{q}I_p=I}(-1)^{q-1}(q-1)!\prod_{p=1}^{q}m_\xi(I_p),\tag{47}$$

其中 $\sum_{\sum_{p=1}^{q}I_p=I}$ 表示对于集合 I 的一切分割求和. 其中 $1\leqslant q\leqslant N(I)$, 而 $N(I)$ 是集合 I 中的元素个数.

为证明 (46) 式, 要借助 (44) 式. 如果 $\nu=\chi(I)$ 和 $\lambda^{(1)}+\cdots+\lambda^{(q)}=\nu$, 则 $\lambda^{(p)}=\chi(I_p)$, $I_p\subseteq I$, 一切 $\lambda^{(p)}$ 各不相同, $\lambda^{(p)}!=\nu!=1$, 且每一无序组 $\{\chi(I_1),\cdots,\chi(I_q)\}$ 与分割 $I=\sum_{p=1}^{q}I_p$ 一一对应. 从而, 由 (44) 式得 (46) 式.

类似地, 可以由 (45) 式证明表达式 (47) 的正确性.

例 4 设 ξ 是随机变量 $(k=1)$, $m_n=m_\xi^{(n)}=\mathbf{E}\xi^n$, $s_n=s_\xi^{(n)}$. 那么, 由 (40) 和 (41) 式得如下公式:

$$\begin{aligned}&m_1=s_1,\\&m_2=s_2+s_1^2,\\&m_3=s_3+3s_1s_2+s_1^3,\\&m_4=s_4+3s_2^2+4s_1s_3+6s_1^2s_2+s_1^4,\\&\cdots\cdots\cdots\cdots\end{aligned}\tag{48}$$

和

$$\begin{aligned}&s_1=m_1=\mathbf{E}\xi,\\&s_2=m_2-m_1^2=\mathbf{D}\xi,\\&s_3=m_3-3m_1m_2+2m_1^3,\\&s_4=m_4-3m_2^2-4m_1m_3+12m_1^2m_2-6m_1^4,\\&\cdots\cdots\cdots\cdots\end{aligned}\tag{49}$$

例 5 设 $\xi\sim N(m,\sigma^2)$. 由 (9) 式可见

$$\ln\varphi_\xi(t)=itm-\frac{t^2\sigma^2}{2}.$$

由 (39) 式 $s_1=m$, $s_2=\sigma^2$, 而从 3 阶起一切半不变量都等于 0, 即 $s_n=0(n\geqslant 3)$.

需要指出, 由于马钦凯维奇定理, 形如 $\exp\mathscr{P}(t)$ 的函数, 其中 $\mathscr{P}(t)$ 是多项式, 则只有当该多项式的次数不大于 2 时, 才可以做特征函数. 特别, 由此可见高斯分布是具有下面性质的唯一分布: 从某个 $n\geqslant 3$ 起一切半不变量 $s_n=0$.

例 6 设 ξ 是参数为 $\lambda>0$ 的泊松随机变量, 则根据 (11) 式,

$$\ln\varphi_\xi(t)=\lambda(e^{it}-1).$$

由此可见, 对于一切 $n\geqslant 1$, 有

$$s_n=\lambda. \tag{50}$$

例 7 设 $\xi=(\xi_1,\cdots,\xi_k)$ 是随机向量, 则

$$\begin{aligned} m_\xi(1)&=s_\xi(1),\\ m_\xi(1,2)&=s_\xi(1,2)+s_\xi(1)s_\xi(2),\\ m_\xi(1,2,3)&=s_\xi(1,2,3)+s_\xi(1,2)s_\xi(3)+s_\xi(1,3)s_\xi(2)\\ &\quad+s_\xi(2,3)s_\xi(1)+s_\xi(1)s_\xi(2)s_\xi(3),\\ &\cdots\cdots \end{aligned} \tag{51}$$

这些公式表明, 简单矩可以相当对称地通过简单半不变量表示. 假如设 $\xi_1\equiv\xi_2\equiv\cdots\equiv\xi_k$, 则由 (51) 式自然地得到 (48) 式.

由 (51) 式可见公式 (48) 中系数的 "组群" 来源. 此外, 由 (51) 式可见,

$$s_\xi(1,2)=m_\xi(1,2)-m_\xi(1)m_\xi(2)=\mathrm{E}\xi_1\xi_2-\mathrm{E}\xi_1\mathrm{E}\xi_2, \tag{52}$$

即 $s_\xi(1,2)$ 恰好是随机变量 ξ_1 和 ξ_2 的协方差.

9. 矩问题的唯一性 设随机变量 ξ 的分布函数为 $F(x)$, 特征函数为 $\varphi(t)$. 假设有一切阶矩 $m_n=\mathrm{E}\xi^n$, $n\geqslant 1$.

由定理 2 知特征函数唯一决定概率分布. 现在提出另一个问题 (矩的唯一性问题): 矩 $\{m_n\}_{n\geqslant 1}$ 是否唯一决定概率分布?

更确切地说, 假设两个分布函数 F 和 G 所有的矩都相同, 即对于一切整数 $n\geqslant 0$, 有

$$\int_{-\infty}^{\infty}x^n\mathrm{d}F(x)=\int_{-\infty}^{\infty}x^n\mathrm{d}G(x). \tag{53}$$

问由此是否可以得出 F 和 G 相同的结论?

一般, 对该问题的回答是否定的. 为证实这一点, 我们考虑分布函数 F, 假设其密度为

$$f(x)=\begin{cases}ke^{-\alpha x^\lambda}, & \text{若 } x>0,\\ 0, & \text{若 } x\leqslant 0,\end{cases}$$

其中 $\alpha>0$, $0<\lambda<1/2$, 而常数 k 由规范性条件 $\displaystyle\int_0^\infty f(x)\mathrm{d}x=1$ 决定.

记 $\beta=\alpha\tan\lambda\pi$, 则

$$g(x)=\begin{cases}ke^{-\alpha x^\lambda}[1+\varepsilon\sin(\beta x^\lambda)], & |\varepsilon|<1,\ x>0,\\ 0, & \text{若 } x\leqslant 0.\end{cases}$$

显然 $g(x)\geqslant 0$. 现在验证对于所有整数 $n\geqslant 0$,

$$\int_0^\infty x^n e^{-\alpha x^\lambda}\sin\beta x^\lambda\mathrm{d}x=0. \tag{54}$$

熟知, 对于 $p>0$ 和复数 q 的实部 $\mathrm{Re}\,q>0$, 有

$$\int_0^\infty t^{p-1}e^{-qt}\mathrm{d}t=\frac{\Gamma(p)}{q^p}.$$

设 $p=(n+1)/\lambda$, $q=\alpha+i\beta$, $t=x^\lambda$, 则

$$\begin{aligned} \int_0^\infty x^{\lambda\left(\frac{n+1}{\lambda}-1\right)}e^{-(\alpha+i\beta)x^\lambda}\lambda x^{\lambda-1}\mathrm{d}x&=\lambda\int_0^\infty x^n e^{-(\alpha+i\beta)x^\lambda}\mathrm{d}x\\ &=\lambda\int_0^\infty x^n e^{-\alpha x^\lambda}\cos\beta x^\lambda\mathrm{d}x-i\lambda\int_0^\infty x^n e^{-\alpha x^\lambda}\sin\beta x^\lambda\mathrm{d}x\\ &=\frac{\Gamma\left(\frac{n+1}{\lambda}\right)}{\alpha^{\frac{n+1}{\lambda}}(1+i\tan\lambda\pi)^{\frac{n+1}{\lambda}}}. \end{aligned} \tag{55}$$

但是, 由于 $\sin(n+1)\pi=0$, 有

$$\begin{aligned} (1+i\tan\lambda\pi)^{\frac{n+1}{\lambda}}&=(\cos\lambda\pi+i\sin\lambda\pi)^{\frac{n+1}{\lambda}}(\cos\lambda\pi)^{-\frac{n+1}{\lambda}}\\ &=e^{i\pi(n+1)}(\cos\lambda\pi)^{-\frac{n+1}{\lambda}}=\cos(n+1)\pi\times(\cos\lambda\pi)^{-\frac{n+1}{\lambda}}. \end{aligned}$$

这样, (55) 式的右侧是实的, 即对于任意整数 $n\geqslant 0$, (54) 式成立. 现在 $G(x)$, 作为取以 $g(x)$ 为密度的分布函数. 那么, 由 (54) 式知分布函数 F 和 G 的一切矩对应相等, 即对于任意整数 $n\geqslant 0$, (53) 式成立.

现在引进保障矩问题唯一性的某些充分条件.

定理 7 假设 $F=F(x)$ 是分布函数, 且对于 $n\geqslant 1$,

$$\mu_n=\int_{-\infty}^{\infty}|x|^n\mathrm{d}F(x).$$

如果

$$\varlimsup_{n\to\infty}\frac{\mu_n^{1/n}}{n}<\infty, \tag{56}$$

则矩 $\{m_n\}_{n\geqslant 1}$, 其中

$$m_n=\int_{-\infty}^{\infty}x^n\mathrm{d}F(x),$$

唯一地决定分布函数 $F=F(x)$.

证明 由 (56) 式和定理 1 的命题 7), 可见存在 $t_0>0$, 使对于一切 $|t|\leqslant t_0$, 特征函数 $\varphi(t)=\displaystyle\int_{-\infty}^{\infty}e^{itx}\mathrm{d}F(x)$ 可以表示为

$$\varphi(t)=\sum_{k=0}^{\infty}\frac{(it)^k}{k!}m_k,$$

因而矩 $\{m_n\}_{n\geqslant 1}$ 对于一切 $|t|\leqslant t_0$ 唯一决定特征函数 $\varphi(t)$ 的值.

取点 s, $|s|\leqslant t_0/2$. 那么, 如同 (15) 式的证明, 由 (56) 式可以导出, 对于一切 $|t-s|\leqslant t_0$:

$$\varphi(t)=\sum_{k=0}^{\infty}\frac{(t-s)^k}{k!}\varphi^{(k)}(s),$$

其中

$$\varphi^{(k)}(s)=i^k\int_{-\infty}^{\infty}x^k e^{isx}\,dF(x)$$

唯一决定于矩 $\{m_n\}_{n\geqslant 1}$. 从而, 对于一切 $|t|\leqslant 3t_0/2$, 这些矩唯一决定特征函数 $\varphi(t)$ 的值. 继续这一过程可以断定, 对于一切 t, 矩 $\{m_n\}_{n\geqslant 1}$ 唯一决定特征函数 $\varphi(t)$, 故也唯一决定分布函数 $F(x)$.

系 1 矩唯一决定集中在有限区间上的概率分布.

系 2 矩问题唯一性的充分条件是:

$$\overline{\lim_{n\to\infty}}\frac{(m_{2n})^{1/(2n)}}{2n}<\infty. \tag{57}$$

为证明只需注意到, 对于奇数阶矩分部估计, 然后利用条件 (56).

例 设 $F(x)$ 是正态分布函数,

$$F(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{x}e^{-\frac{t^2}{2\sigma^2}}\,dt.$$

那么,

$$m_{2n-1}=0,\quad m_{2n}=\frac{(2n)!}{2^n n!}\sigma^{2n},$$

而由 (57) 式, 可见这些矩是只有正态分布才有的矩.

最后, 我们引进卡尔莱曼 (T. Carleman) 准则 (不加证明).

卡尔莱曼准则 (矩问题的唯一性) [60, 卷 2, VII.3.]

a) 设 $\{m_n\}_{n\geqslant 1}$ 是某概率分布的矩, 且

$$\sum_{n=0}^{\infty}\frac{1}{(m_{2n})^{1/(2n)}}=\infty.$$

那么, $\{m_n\}_{n\geqslant 1}$ 唯一决定概率分布.

b) 如果 $\{m_n\}_{n\geqslant 1}$ 是集中在 $[0,\infty)$ 上概率分布的矩, 则 $\{m_n\}_{n\geqslant 1}$ 唯一性的充分条件是

$$\sum_{n=0}^{\infty}\frac{1}{(m_n)^{1/(2n)}}=\infty.$$

10. 埃森不等式 $F=F(x)$ 和 $G=G(x)$ 分别是以 $f=f(t)$ 和 $g=g(t)$ 为特征函数的分布函数. 下面的定理说明 (略去证明), 根据 f 和 g 的接近程度, 如何 (在均匀尺度下) 估计 F 和 G 的接近程度. (关于该定理的应用见第三章 §11.)

定理 (埃森 [C. G. Esseen] 不等式) 设 $G(x)$ 有导数且 $\sup_x|G'(x)|\leqslant C$, 则

$$\sup_x|F(x)-G(x)|\leqslant\frac{2}{\pi}\int_0^T\left|\frac{f(t)-g(t)}{t}\right|dt+\frac{24}{\pi T}\sup_x|G'(x)|.$$

11. 常见分布的特征函数表 表 2-4 和表 2-5 是一些常见分布的特征函数表.

表 2—4

分布名称	特征函数
离散均匀	$\frac{1}{N}\frac{e^{it}}{1-e^{it}}(1-e^{itN})$
伯努利 (J. Bernoulli)	$q+pe^{it}$
二项	$(q+pe^{it})^n$
泊松 (S. D. Poisson)	$\exp\{\lambda(e^{it}-1)\}$
几何	$p(1-qe^{it})^{-1}$
负二项 (帕斯卡)	$[p(1-qe^{it})^{-1}]^r$

表 2—5

分布名称	特征函数
$[a,b]$ 上的均匀	$\frac{e^{itb}-e^{ita}}{it(b-a)}$
正态或高斯 (C. F. Gauss)	$\exp\left\{itm-\frac{\sigma^2t^2}{2}\right\}$
伽马 (Γ)	$(1-it\beta)^{-\alpha}$
贝塔 (B)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)}\sum_{k=0}^{\infty}\frac{(it)^k\Gamma(\alpha+k)}{k!\Gamma(\alpha+\beta+k)\Gamma(1+k)}$
指数	$\frac{\lambda}{\lambda-it}$
双侧指数	$\frac{\lambda^2e^{it\alpha}}{\lambda^2+t^2}$
χ^2 (卡方)	$(1-2it)^{-n/2}$
t (学生)	$\frac{\sqrt{\pi}\Gamma((n+1)/2)}{\Gamma(n/2)}\frac{e^{-\sqrt{n}\lvert t\rvert}}{2^{2(m-1)}(m-1)!}\sum_{k=0}^{m-1}(2k)!C_{m-1+k}^{2k}(2\lvert t\rvert\sqrt{n})^{m-1-k}$, 若 $m=(n+1)/2$ 是整数
柯西 (A. L. Cauchy)	$e^{-\theta\lvert t\rvert}$

12. 练习题

1. 设 ξ 和 η 是独立随机变量, $f(x) = f_1(x) + if_2(x), g(x) = g_1(x) + ig_2(x)$, 其中 $f_k(x)$ 和 $g_k(x)(k = 1, 2)$ 是博雷尔函数. 证明, 如果 $\mathbf{E}|f(\xi)| < \infty, \mathbf{E}|g(\eta)| < \infty$, 则

$$\mathbf{E}|f(\xi)g(\eta)| < \infty$$

且

$$\mathbf{E}f(\xi)g(\eta) = \mathbf{E}f(\xi) \times \mathbf{E}g(\eta).$$

2. 设 $\xi = (\xi_1, \cdots, \xi_n)$ 且 $\mathbf{E}\|\xi\|^n < \infty$, 其中 $\|\xi\| = +\sqrt{\sum \xi_i^2}$. 证明

$$\varphi_\xi(t) = \sum_{k=0}^{n} \frac{i^k}{k!}\mathbf{E}(t, \xi)^k + \varepsilon_n(t)\|t\|^n,$$

其中 $t = (t_1, \cdots, t_n)$ 且 $\varepsilon_n(t) \to 0, t \to 0$.

3. 对于 n 维分布函数 $F = F_n(x_1, \cdots, x_n)$ 和 $G = G_n(x_1, \cdots, x_n)$, 证明定理 2.

4. 设 $F = F(x_1, \cdots, x_n)$ 是 n 维分布函数, 而 $\varphi = \varphi(t_1, \cdots, t_n)$ 是其特征函数. 利用 §3(12) 式的记号, 证明逆转公式:

$$\mathbf{P}(a, b] = \lim_{c \to \infty} \frac{1}{(2\pi)^n} \int_{-c}^{c} \cdots \int_{-c}^{c} \prod_{k=1}^{n} \frac{e^{-it_k a_k} - e^{-it_k b_k}}{it_k} \varphi(t_1, \cdots, t_n) \mathrm{d}t_1 \cdots \mathrm{d}t_n.$$

(在上面引进的逆转公式中, $(a, b]$ 是函数 $\mathbf{P}(a, b]$ 的连续区间, 即对于一切 $k = 1, \cdots, n$, 点 a_k, b_k 是边缘分布函数 $F_k(x_k)$ 的连续点, 而边缘分布函数 $F_k(x_k)$ 是由联合分布函数 $F = F(x_1, \cdots, x_n)$ 将除 x_k 之外的其余变量设为 $-\infty$ 得到的.)

5. 设 $\varphi_k(t), k \geqslant 1$ 是特征函数, 而非负实数 $\lambda_k, k \geqslant 1$, 满足条件 $\sum \lambda_k = 1$. 证明 $\sum \lambda_k \varphi_k(t)$ 是特征函数.

6. 设 $\varphi(t)$ 是特征函数, 问 $\mathrm{Re}\,\varphi(t)$ 和 $\mathrm{Im}\,\varphi(t)$ 是否特征函数?

7. 设 $\varphi_1, \varphi_2, \varphi_3$ 是特征函数, 且 $\varphi_1\varphi_2 = \varphi_1\varphi_3$. 问是否由此可见 $\varphi_2 = \varphi_3$?

8. 证明表 2－4 和表 2－5 中特征函数式的正确性.

9. 设 ξ 是整数值随机变量, 而 $\varphi(t)$ 是其特征函数. 证明

$$\mathbf{P}\{\xi = k\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ikt} \varphi_\xi(t) \mathrm{d}t, \quad k = 0, \pm 1, \pm 2, \cdots.$$

10. 证明在具有勒贝格测度 μ 的空间 $L^2 = L^2([-\pi, \pi], \mathscr{B}([-\pi, \pi]))$ 中, 函数系

$$\left\{ \frac{1}{\sqrt{2\pi}} e^{i\lambda n}, n = 0, \pm 1, \cdots \right\}$$

构成规范正交系.

11. 假设在博赫纳－辛钦定理中所考虑的函数 $\varphi(t)$ 是连续的. 证明下面 (里斯 (F. Riesz)) 结果, 该结果说明在什么程度上可以去掉连续性假设.

假设 $\varphi = \varphi(t)$ 是复数值勒贝格可测函数且 $\varphi(0) = 1$. 那么, 函数 $\varphi(t)$ 是正定的, 当且仅当它 (在数轴上就勒贝格测度几乎处处) 等于某一特征函数.

12. 判断下列各函数是否特征函数:

$$\varphi(t) = e^{-|t|^k}, 0 \leqslant k \leqslant 2; \quad \varphi(t) = e^{-|t|^k}, k > 2;$$

$$\varphi(t) = (1 + |t|)^{-1}; \quad \varphi(t) = (1 + t^4)^{-1};$$

$$\varphi(t) = \begin{cases} 1 - |t|^3, & 若\ |t| \leqslant 1, \\ 0, & 若\ |t| > 1; \end{cases} \quad \varphi(t) = \begin{cases} 1 - |t|, & 若\ |t| \leqslant 1/2, \\ 1/(4|t|), & 若\ |t| > 1/2. \end{cases}$$

13. 设 $\varphi(t)$ 是分布函数 $F = F(x)$ 的特征函数, $\{x_n\}$ 是函数 $F(x)(\Delta F(x_n) \equiv F(x_n) - F(x_n-) > 0)$ 的间断点的集合. 证明

$$\lim_{T \to \infty} \frac{1}{T} \int_{-T}^{T} |\varphi(t)|^2 \mathrm{d}t = \sum_{n \geqslant 1} [\Delta F(x_n)]^2.$$

14. 称函数

$$Q(X; l) = \sup_{x \in \mathbf{R}} \mathbf{P}\{x \leqslant X \leqslant x + l\}$$

为随机变量 X 的集中函数. 证明:

(a) 若 X 和 Y 是独立随机变量, 则对于一切 $l \geqslant 0$,

$$Q(X + Y; l) \leqslant \min\{Q(X; l), Q(Y; l)\};$$

(b) 存在 x_l^*, 使随机变量 X 的 $Q(X; l) = \mathbf{P}\{x_l^* \leqslant X \leqslant x_l^* + l\}$ 和随机变量 X 的分布函数连续, 当且仅当 $Q(X; 0) = 0$.

15. 设 X 是分布函数为 $F = F(x)$ 的随机变量, $\{m_n\}_{n \geqslant 1}$ 是其矩序列, 其中 $m_k = \int_{-\infty}^{\infty} x^k \mathrm{d}F(x)$. 证明, 若对于某个 $s > 0$, 级数 $\sum\limits_{k=1}^{\infty} \frac{m_k}{k!} s^k$ 绝对收敛, 则 $\{m_n\}_{n \geqslant 1}$ 唯一决定分布函数 $F = F(x)$.

16. 设 $F = F(x)$ 分布函数, 而 $\varphi(t) = \int_{-\infty}^{\infty} e^{itx} \mathrm{d}F(x)$ 是它的特征函数. 证明

$$\lim_{c \to \infty} \frac{1}{2c} \int_{-c}^{c} e^{-itx} \varphi(t) \mathrm{d}t = F(x) - F(x-),$$

$$\lim_{c \to \infty} \frac{1}{2c} \int_{-c}^{c} |\varphi(t)|^2 \mathrm{d}t = \sum_{x \in \mathbf{R}} [F(x) - F(x-)]^2.$$

17. 证明每一个特征函数 $\varphi(t)$ 都满足不等式: $1 - \mathrm{Re}\,\varphi(2t) \leqslant 4[1 - \mathrm{Re}\,\varphi(t)]$.

18. 假设特征函数 $\varphi(t)$ 满足: $\varphi(t) = 1 + f(t) + o(t^2), t \to 0$, 其中 $f(t) = -f(-t)$. 证明 $\varphi(t) \equiv 1$.

19. 证明对于每一个 $n \geqslant 1$, 函数

$$\varphi_n(t) = \frac{e^{it} - \sum_{k=0}^{n-1} (it)^k / k!}{(it)^n / n!}$$

是特征函数.

20. 证明

$$\frac{2}{\pi} \int_{-\infty}^{\infty} \frac{1 - \mathrm{Re}\,\varphi(t)}{t^2} \mathrm{d}t = \int_{-\infty}^{\infty} |x| \mathrm{d}F(x).$$

21. 假设特征函数 $\varphi(t)$ 满足: $\varphi(t) = 1 + O(|t|^\alpha), t \to 0$, 其中 $\alpha \in (0, 2]$. 证明以 $\varphi(t)$ 为特征函数的随机变量 ξ 具有如下性质:

$$\mathbf{P}\{|\xi| > x\} = O(x^{-\alpha}), \quad x \to \infty.$$

22. 如果 $\varphi(t)$ 是特征函数, 则函数 $|\varphi(t)|^2$ 也是特征函数.

23. 设 X 和 Y 是独立同分布随机变量, 其数学期望为 0 和方差为 1. 基于特征函数的性质, 证明, 如果随机变量 $(X - Y)/\sqrt{2}$ 的分布等同于随机变量 X 和 Y 的分布 F, 则 F 是正态分布函数.

24. 如果 $\varphi = \varphi(t)$ 是特征函数, 则对于某一个 $\lambda \geqslant 0$ 函数 $e^{\lambda(\varphi - 1)}$ 也是特征函数.

25. 设 X 是分布函数为 F 的非负随机变量. 对于任意 $\lambda \geqslant 0$, 由公式

$$\widehat{F}(\lambda) = \mathbf{E}e^{-\lambda X} = \int_{[0, \infty)} e^{-\lambda x} \mathrm{d}F(x)$$

定义的函数 $\widehat{F} = \widehat{F}(\lambda)$ 称做 X 的拉普拉斯变换. 证明如下伯恩斯坦 (С. Н. Бернштейн) 准则: 函数 $f = f(\lambda), \lambda \in (0, \infty)$ 是分布函数 $F = F(x), x \in [0, \infty)$ 的拉普拉斯变换的充分和必要条件是: 函数 $f = f(\lambda)$ 是完全单调的 (即存在所有阶导数 $f^{(n)}(\lambda), n \geqslant 0$ 且 $(-1)^n f^{(n)}(\lambda) \geqslant 0$).

26. 设 $\varphi(t)$ 是特征函数. 证明下面的函数也是特征函数:

$$\int_0^1 \varphi(ut) \mathrm{d}u, \quad \int_0^\infty e^{-u} \varphi(ut) \mathrm{d}u.$$

§13. 高斯系

1. 高斯系的特点和重要性 在概率论和数理统计中, 高斯分布或正态分布、高斯随机变量、高斯过程和高斯系有特别重要的作用. 这首先因为有中心极限定理 (第三章 §4), 棣莫弗 – 拉普拉斯定理是其特殊情形 (第一章 §6). 根据这一定理, 正态分布具有通用的特点: 在并不 "拘谨" 的条件下, 大量独立随机变量或随机向量之和的分布, 可以很好地用正态分布来逼近.

正是这种情况, 从理论上说明了统计实践中普遍的 "误差律", 它表现为由大量独立的 "基本" 误差叠加形成的测量误差服从正态分布.

多维正态分布依赖于少量参数, 在建立简单的概率模型时是其毋容置疑的优点. 高斯随机变量有有限二阶矩, 故可以用希尔伯特空间方法研究其性质. 重要的情况是, 对于高斯随机变量的情形, 不相关性变为独立性, 使得有可能大为加强 "L^2– 理论" 的结果.

2. 高斯系的定义 注意到 (根据 §8) 随机变量 $\xi = \xi(\omega)$ 称做服从参数为 m 和 σ^2 的高斯分布或正态分布的 ($\xi \sim N(m, \sigma^2), |m| < \infty, \sigma^2 > 0$), 如果其密度为

$$f_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}, \tag{1}$$

其中 $\sigma = +\sqrt{\sigma^2}$.

当 $\sigma \downarrow 0$ 时, 密度 $f_\xi(x)$ "收敛于集中在点 $x = m$ 的 δ– 函数". 所以自然称随机变量 $\xi = \xi(\omega)$ 服从参数为 m 和 $\sigma^2 = 0$ 的高斯分布或正态分布 ($\xi \sim N(m, 0)$), 如果 $\mathbf{P}\{\xi = m\} = 1$.

不过可以给出另一定义, 使之既包含非退化 ($\sigma^2 > 0$) 情形又包含退化 ($\sigma^2 = 0$) 情形. 为此, 考虑特征函数 $\varphi_\xi(t) = \mathbf{E}e^{it\xi}, t \in \mathbf{R}$.

如果 $\mathbf{P}\{\xi = m\} = 1$, 则显然

$$\varphi_\xi(t) = e^{itm}. \tag{2}$$

如果 $\xi \sim N(m, \sigma^2), \sigma^2 > 0$, 则根据 §12(9) 式, 有

$$\varphi_\xi(t) = e^{itm - \frac{t^2\sigma^2}{2}}. \tag{3}$$

易见, 对于 $\sigma^2 = 0$, (3) 式的右侧与 (2) 式的右侧相同. 由此及 §12 的定理 1, 可见参数为 m 和 $\sigma^2 (|m| < \infty, \sigma^2 \geqslant 0)$ 的高斯随机变量, 可以定义为特征函数 $\varphi_\xi(t)$ 由 (3) 式定义的随机变量. 运用特征函数的方法, 在多维情形特别方便.

设 $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_n)$ 是随机向量, 而

$$\varphi_{\boldsymbol{\xi}}(\boldsymbol{t}) = \mathbf{E}e^{i(\boldsymbol{t}, \boldsymbol{\xi})}, \boldsymbol{t} = (t_1, \cdots, t_n) \in \mathbf{R}^n, \tag{4}$$

是其特征函数 (见 §12 定义 2).

定义 1 称随机向量 $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_n)$ 服从高斯分布或正态分布, 如果其特征函数 $\varphi_{\boldsymbol{\xi}}(\boldsymbol{t})$ 有如下形式:

$$\varphi_{\boldsymbol{\xi}}(\boldsymbol{t}) = e^{i(\boldsymbol{t}, \boldsymbol{m}) - \frac{1}{2}(\boldsymbol{R}\boldsymbol{t}, \boldsymbol{t})}, \tag{5}$$

其中 $\boldsymbol{m} = (m_1, \cdots, m_n), |m_k| < \infty$, 而 $\boldsymbol{R} = (r_{kl})$ 是 $n \times n$ 阶非负定矩阵 (简记为 $\boldsymbol{\xi} \sim N(\boldsymbol{m}, \boldsymbol{R})$).

鉴于所引进的定义, 首先产生一个问题, (5) 式的函数是否特征函数? 现在证明它确实是特征函数.

为此, 首先假设 $\boldsymbol{R}$ 是非退化的. 那么存在逆矩阵 $\boldsymbol{A}=\boldsymbol{R}^{-1}$, 且函数

$$f(\boldsymbol{x})=\frac{|\boldsymbol{A}|^{1/2}}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{m}),(\boldsymbol{x}-\boldsymbol{m}))\right\}, \tag{6}$$

其中 $\boldsymbol{x}=(x_1,\cdots,x_n)$, $|\boldsymbol{A}|=\det\boldsymbol{A}$. 该函数是非负的. 证明

$$\int_{\mathbf{R}^n}e^{i(\boldsymbol{t},\boldsymbol{x})}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=e^{i(\boldsymbol{t},\boldsymbol{m})-\frac{1}{2}(\boldsymbol{R}\boldsymbol{t},\boldsymbol{t})},$$

或同样地

$$I_n\equiv\int_{\mathbf{R}^n}e^{i(\boldsymbol{t},\boldsymbol{x}-\boldsymbol{m})}\frac{|\boldsymbol{A}|^{1/2}}{(2\pi)^{n/2}}e^{-\frac{1}{2}(\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{m}),(\boldsymbol{x}-\boldsymbol{m}))}\mathrm{d}\boldsymbol{x}=e^{-\frac{1}{2}(\boldsymbol{R}\boldsymbol{t},\boldsymbol{t})}. \tag{7}$$

在积分中作变量替换

$$\boldsymbol{x}-\boldsymbol{m}=\boldsymbol{Z}\boldsymbol{u},\quad \boldsymbol{t}=\boldsymbol{Z}\boldsymbol{v},$$

其中 $\boldsymbol{Z}$ 是正交矩阵, 满足

$$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z}=\boldsymbol{D},$$

而

$$\boldsymbol{D}=\begin{pmatrix}d_1 & & 0\\ & \ddots & \\ 0 & & d_n\end{pmatrix}$$

是对角矩阵, 其中 $d_i\geqslant 0$ (见 §8 引理的证明). 由于 $|\boldsymbol{R}|=\det\boldsymbol{R}\neq 0$, 可见 $d_i>0$, $i=1,\cdots,n$. 因此

$$|\boldsymbol{A}|=|\boldsymbol{R}^{-1}|=d_1^{-1}\cdots d_n^{-1}. \tag{8}$$

其次 (见 §12 第 1 小节的记号), 有

$$\begin{aligned}
&i(\boldsymbol{t},\boldsymbol{x}-\boldsymbol{m})-\frac{1}{2}(\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{m}),(\boldsymbol{x}-\boldsymbol{m}))\\
&=i(\boldsymbol{Z}\boldsymbol{v},\boldsymbol{Z}\boldsymbol{u})-\frac{1}{2}(\boldsymbol{A}\boldsymbol{Z}\boldsymbol{u},\boldsymbol{Z}\boldsymbol{u})=i(\boldsymbol{Z}\boldsymbol{v})^{\mathrm{T}}(\boldsymbol{Z}\boldsymbol{u})-\frac{1}{2}(\boldsymbol{Z}\boldsymbol{u})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{Z}\boldsymbol{u})\\
&=i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}-\frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Z}\boldsymbol{u}=i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}-\frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{u}.
\end{aligned}$$

由 (8) 式和 §12 的 (9) 式, 可得

$$\begin{aligned}
I_n&=\frac{1}{(2\pi)^{n/2}(d_1\cdots d_n)^{1/2}}\int_{\mathbf{R}^n}e^{i\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}-\frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{u}}\mathrm{d}\boldsymbol{u}\\
&=\prod_{k=1}^{n}\frac{1}{(2\pi d_k)^{1/2}}\int_{-\infty}^{\infty}e^{iv_ku_k-\frac{u_k^2}{2d_k}}\mathrm{d}u_k=\prod_{k=1}^{n}e^{-\frac{v_k^2d_k}{2}}\\
&=e^{-\frac{1}{2}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{v}}=e^{-\frac{1}{2}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z}\boldsymbol{v}}=e^{-\frac{1}{2}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{t}}=e^{-\frac{1}{2}(\boldsymbol{R}\boldsymbol{t},\boldsymbol{t})}.
\end{aligned}$$

由 (6) 式亦可见

$$\int_{\mathbf{R}^n}f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=1. \tag{9}$$

于是, 函数 (5) 是 n 维 (非退化) 高斯分布的特征函数 (§3 第 3 小节).

现在, 设 $\boldsymbol{R}$ 是退化矩阵. 设 $\varepsilon>0$, 考虑退化矩阵 $\boldsymbol{R}^{\varepsilon}=\boldsymbol{R}+\varepsilon\boldsymbol{E}$, 其中 $\boldsymbol{E}$ 是单位矩阵. 那么, 由已证明的, 可见函数

$$\varphi^{\varepsilon}(\boldsymbol{t})=\exp\left\{i(\boldsymbol{t},\boldsymbol{m})-\frac{1}{2}(\boldsymbol{R}^{\varepsilon}\boldsymbol{t},\boldsymbol{t})\right\}$$

是特征函数:

$$\varphi^{\varepsilon}(\boldsymbol{t})=\int_{\mathbf{R}^n}e^{i(\boldsymbol{t},\boldsymbol{x})}\mathrm{d}F_{\varepsilon}(\boldsymbol{x}),$$

其中 $F_{\varepsilon}(\boldsymbol{x})=F_{\varepsilon}(x_1,\cdots,x_n)$ 是 n 维分布函数.

当 $\varepsilon\to 0$ 时

$$\varphi^{\varepsilon}(\boldsymbol{t})\to\varphi(\boldsymbol{t})=\exp\left\{i(\boldsymbol{t},\boldsymbol{m})-\frac{1}{2}(\boldsymbol{R}\boldsymbol{t},\boldsymbol{t})\right\}.$$

极限函数 $\varphi(\boldsymbol{t})$ 在 0 点 $(0,\cdots,0)$ 连续. 因此, 根据定理 1 和第三章 §3 的练习题 1, $\varphi(\boldsymbol{t})$ 是特征函数.

于是, 定义 1 的适定性得证.

3. 高斯系的均值向量和协方差矩阵　现在说明特征函数的 (5) 式中, 向量 $\boldsymbol{m}$ 和矩阵 $\boldsymbol{R}=(r_{kl})$ 的含义.

由于

$$\ln\varphi_{\xi}(\boldsymbol{t})=i(\boldsymbol{t},\boldsymbol{m})-\frac{1}{2}(\boldsymbol{R}\boldsymbol{t},\boldsymbol{t})=i\sum_{k=1}^{n}t_km_k-\frac{1}{2}\sum_{k,l=1}^{n}r_{kl}t_kt_l, \tag{10}$$

由 §12 的 (35) 式, 以及矩和半不变量的换算公式, 可见

$$m_1=s_{\xi}^{(1,0,\cdots,0)}=\mathbf{E}\xi_1,\cdots,m_n=s_{\xi}^{(0,\cdots,0,1)}=\mathbf{E}\xi_n.$$

类似地, 有

$$r_{11}=s_{\xi}^{(2,0,\cdots,0)}=\mathbf{D}\xi_1,\quad r_{12}=s_{\xi}^{(1,1,\cdots,0)}=\mathrm{cov}(\xi_1,\xi_2),$$

一般

$$r_{kl}=\mathrm{cov}(\xi_k,\xi_l).$$

于是, $\boldsymbol{m}$ 是 ξ 的均值向量, 而 $\boldsymbol{R}$ 是协方差矩阵.

如果矩阵 $\boldsymbol{R}$ 非退化, 则可以通过其他途径得到这些结果. 具体地说, 这时向量有 (6) 式给出的密度 $f(\boldsymbol{x})$, 而且通过直接推算可得

$$\begin{aligned}
\mathbf{E}\xi_k&\equiv\int_{-\infty}^{\infty}x_kf(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=m_k,\\
\mathrm{cov}(\xi_k,\xi_l)&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_k-m_k)(x_l-m_l)f(x_k,x_l)\mathrm{d}x_k\mathrm{d}x_l=r_{kl}.
\end{aligned} \tag{11}$$

4. 高斯向量的性质 我们现在讨论高斯向量的若干性质.

定理 1 a) 对于高斯向量, 其分量不相关与分量独立等价.

b) 向量 $\boldsymbol{\xi}=(\xi_1,\cdots,\xi_n)$ 是高斯向量, 当且仅当对于任意向量 $\boldsymbol{\lambda}=(\lambda_1,\cdots,\lambda_n)$, $\lambda_k\in\mathbb{R}$, 随机变量 $(\boldsymbol{\xi},\boldsymbol{\lambda})=\lambda_1\xi_1+\cdots+\lambda_n\xi_n$ 服从高斯分布.

证明 a) 如果向量 $\boldsymbol{\xi}=(\xi_1,\cdots,\xi_n)$ 的分量不相关, 则由其特征函数 $\varphi_{\xi}(\boldsymbol{t})$ 的形式可见, 它等于各分量特征函数的乘积:

$$\varphi_{\xi}(\boldsymbol{t})=\prod_{k=1}^{n}\varphi_{\xi_k}(t_k).$$

所以, 由 §12 的定理 4 知, 分量 $\xi_1,\cdots,\xi_n$ 独立.

由于由独立性总是可以得出不相关性, 故逆命题显然成立.

b) 如果 $\boldsymbol{\xi}=(\xi_1,\cdots,\xi_n)$ 是高斯向量, 则对于 $t\in\mathbb{R}$ (见 (5) 式)

$$\mathbf{E}\exp[it(\lambda_1\xi_1+\cdots+\lambda_n\xi_n)]=\exp\left\{it\left(\sum\lambda_k m_k\right)-\frac{t^2}{2}\left(\sum r_{kl}\lambda_k\lambda_l\right)\right\},$$

从而

$$(\boldsymbol{\xi},\boldsymbol{\lambda})\sim N\left(\sum\lambda_k m_k,\sum r_{kl}\lambda_k\lambda_l\right).$$

相反, 随机变量 $(\boldsymbol{\xi},\boldsymbol{\lambda})=\lambda_1\xi_1+\cdots+\lambda_n\xi_n$ 服从高斯分布, 则特别

$$\mathbf{E}e^{i(\boldsymbol{\xi},\boldsymbol{\lambda})}=e^{i\mathbf{E}(\boldsymbol{\xi},\boldsymbol{\lambda})-\frac{\mathbf{D}(\boldsymbol{\xi},\boldsymbol{\lambda})}{2}}=\exp\left\{i\sum\lambda_k\mathbf{E}\xi_k-\frac{1}{2}\sum\lambda_k\lambda_l\operatorname{cov}(\xi_k,\xi_l)\right\}.$$

由于 $\lambda_1,\cdots,\lambda_n$ 的任意性, 可见 $\boldsymbol{\xi}=(\xi_1,\cdots,\xi_n)$ 是高斯向量 (见定义 1). □

注 设 $(\boldsymbol{\theta},\boldsymbol{\xi})$ 是高斯向量, 其中 $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_k)$, $\boldsymbol{\xi}=(\xi_1,\cdots,\xi_l)$. 如果向量 $\boldsymbol{\theta}$ 和 $\boldsymbol{\xi}$ 不相关, 即 $\operatorname{cov}(\theta_i,\xi_j)=0$, $i=1,\cdots,k$, $j=1,\cdots,l$, 则它们就独立.

证明方法与定理 1 的命题 a) 相同.

设 $\boldsymbol{\xi}=(\xi_1,\cdots,\xi_n)$ 是高斯向量, 为简便计, 假设其均值向量为 0. 如果 $\operatorname{rang}\boldsymbol{R}=r<n$, 则由 §11 可见, 在变量 $\xi_1,\cdots,\xi_n$ 之间存在恰好 $n-r$ 个线性关系. 这时, 例如可以认为 $\xi_1,\cdots,\xi_r$ 线性无关, 而其余一切变量都可以通过它们线性表示. 因此向量 $\boldsymbol{\xi}=(\xi_1,\cdots,\xi_n)$ 的所有基本性质都决定于前 r 个分量 $(\xi_1,\cdots,\xi_r)$, 而且它的协方差矩阵已经是退化的.

这样, 可以认为原向量的分量已经是线性无关的, 因而 $|\boldsymbol{R}|>0$.

设 $\boldsymbol{Z}$ 是把 $\boldsymbol{R}$ 变为对角矩阵的正交矩阵:

$$\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{Z}=\boldsymbol{D}.$$

如第 3 小节所指出的, 矩阵的所有对角元素都是正数, 从而有逆矩阵. 记 $\boldsymbol{B}^2=\boldsymbol{D}$, 且设

$$\boldsymbol{\beta}=\boldsymbol{B}^{-1}\boldsymbol{Z}\boldsymbol{\xi}.$$

那么, 容易证明

$$\mathbf{E}e^{i(\boldsymbol{t},\boldsymbol{\beta})}=\mathbf{E}e^{i\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{t}}=e^{-\frac{1}{2}(\boldsymbol{E}\boldsymbol{t},\boldsymbol{t})},$$

即向量 $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_n)$ 是具有不相关分量的高斯向量, 从而根据定理 1 是具有独立分量的高斯向量. 若记 $\boldsymbol{A}=\boldsymbol{Z}\boldsymbol{B}$, 则原高斯向量 $\boldsymbol{\xi}=(\xi_1,\cdots,\xi_n)$ 可以表示为

$$\boldsymbol{\xi}=\boldsymbol{A}\boldsymbol{\beta},\tag{12}$$

其中 $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_n)$ 是具有独立分量的高斯向量, 且 $\beta_k\sim N(0,1)$. 由此得如下结果: 设 $\boldsymbol{\xi}=(\xi_1,\cdots,\xi_n)$ 是具有线性不相关分量的向量, $\mathbf{E}\xi_k=0$, $k=1,\cdots,n$, 则此向量是高斯向量的充分和必要条件是, 存在独立高斯变量 $\beta_1,\cdots,\beta_n$, $\beta_k\sim N(0,1)$, 和 n 阶非退化矩阵 $\boldsymbol{A}$, 使 $\boldsymbol{\xi}=\boldsymbol{A}\boldsymbol{\beta}$. 这时 $\boldsymbol{R}=\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ 是向量 ξ 的协方差矩阵.

如果 $\boldsymbol{R}\neq 0$, 则根据格拉姆–施密特正交化方法 (见 §11), 有

$$\xi_k=\hat{\xi}_k+b_k\varepsilon_k,\quad k=1,\cdots,n,\tag{13}$$

其中由于高斯性, 向量 $\boldsymbol{\varepsilon}=(\varepsilon_1,\cdots,\varepsilon_n)\sim N(\boldsymbol{0},\boldsymbol{E})$,

$$\hat{\xi}_k=\sum_{l=1}^{k-1}(\xi_k,\varepsilon_l)\varepsilon_l,\tag{14}$$

$$b_k=\|\xi_k-\hat{\xi}_k\|,\tag{15}$$

且

$$\mathscr{L}\{\xi_1,\cdots,\xi_k\}=\mathscr{L}\{\varepsilon_1,\cdots,\varepsilon_k\}.\tag{16}$$

由正交展开式 (13), 立即得

$$\hat{\xi}_k=\mathbf{E}(\xi_k|\xi_{k-1},\cdots,\xi_1).\tag{17}$$

鉴于 (16) 和 (14) 式, 由此可见对于高斯变量, 条件数学期望 $\mathbf{E}(\xi_k|\xi_{k-1},\cdots,\xi_1)$ 是 $\xi_1,\cdots,\xi_{k-1}$ 的线性函数:

$$\mathbf{E}(\xi_k|\xi_{k-1},\cdots,\xi_1)=\sum_{i=1}^{k-1}a_i\xi_i.\tag{18}$$

(对于 $k=2$ 的情形, 在 §8 曾得到该结果.)

由于根据 §8 中对定理 1 的注知, $\mathbf{E}(\xi_k|\xi_{k-1},\cdots,\xi_1)$ (均方意义上) 是由 $\xi_1,\cdots,\xi_{k-1}$ 对 ξ_k 的最优估计量, 则由 (18) 式可见, 对于高斯分布, 最优估计量是线性的.

对于 $(\boldsymbol{\theta},\boldsymbol{\xi})$ 是高斯向量的情形, 我们利用以上的结果由向量 $\boldsymbol{\xi}=(\xi_1,\cdots,\xi_n)$, 来求向量 $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_k)$ 的最优估计. 记

$$m_{\theta}=\mathbf{E}\boldsymbol{\theta},\quad m_{\xi}=\mathbf{E}\boldsymbol{\xi}$$

为均值列向量，而

$$\mathbf{D}_{\theta\theta} \equiv \operatorname{cov}(\theta,\theta) \equiv (\operatorname{cov}(\theta_i,\theta_j)),\ 1 \leqslant i,j \leqslant k,$$
$$\mathbf{D}_{\theta\xi} = \operatorname{cov}(\theta,\xi) = (\operatorname{cov}(\theta_i,\xi_j)),\ 1 \leqslant i \leqslant k,\ 1 \leqslant j \leqslant l,$$
$$\mathbf{D}_{\xi\xi} = \operatorname{cov}(\xi,\xi) = (\operatorname{cov}(\xi_i,\xi_j)),\ 1 \leqslant i,j \leqslant l$$

是协方差矩阵. 假设矩阵 $\mathbf{D}_{\xi\xi}$ 有逆矩阵. 那么，下列定理成立 (对照 §8 定理 2.)

定理 2 (正态相关性定理，向量情形)　*对于正态随机向量 (θ,ξ)，由 ξ 对向量 θ 的最优估计量 $\mathbf{E}(\theta|\xi)$，及其误差矩阵*

$$\Delta = \mathbf{E}[\theta - \mathbf{E}(\theta|\xi)][\theta - \mathbf{E}(\theta|\xi)]^{\mathrm{T}},$$

分别由下面的公式给出：

$$\mathbf{E}(\theta|\xi) = m_\theta + \mathbf{D}_{\theta\xi}\mathbf{D}_{\xi\xi}^{-1}(\xi - m_\xi), \tag{19}$$
$$\Delta = \mathbf{D}_{\theta\theta} - \mathbf{D}_{\theta\xi}\mathbf{D}_{\xi\xi}^{-1}\mathbf{D}_{\theta\xi}^{\mathrm{T}}. \tag{20}$$

证明　记向量

$$\eta = (\theta - m_\theta) - \mathbf{D}_{\theta\xi}\mathbf{D}_{\xi\xi}^{-1}(\xi - m_\xi). \tag{21}$$

那么，可以直接验证 $\mathbf{E}\eta(\xi - m_\xi)^{\mathrm{T}} = 0$，即向量 η 与向量 $\xi - m_\xi$ 不相关. 由于向量 (θ,ξ) 的高斯性，则 (η,ξ) 也是高斯向量. 由此以及定理 1 的注，知向量 η 与 $\xi - m_\xi$ 独立. 因此，向量 η 与 ξ 独立，从而 $\mathbf{E}(\eta|\xi) = \mathbf{E}\eta = 0$. 所以

$$\mathbf{E}[\theta - m_\theta|\xi] - \mathbf{D}_{\theta\xi}\mathbf{D}_{\xi\xi}^{-1}(\xi - m_\xi) = 0.$$

于是，公式 (19) 得证.

为证明 (20) 式，考虑条件协方差

$$\operatorname{cov}(\theta,\theta|\xi) = \mathbf{E}\{[\theta - \mathbf{E}(\theta|\xi)][\theta - \mathbf{E}(\theta|\xi)]^{\mathrm{T}}|\xi\}. \tag{22}$$

由 (19) 和 (21) 式可见，$\theta - \mathbf{E}(\theta|\xi) = \eta$. 故由 η 与 ξ 独立，可见

$$\begin{aligned}\operatorname{cov}(\theta,\theta|\xi) &= \mathbf{E}(\eta\eta^{\mathrm{T}}|\xi) = \mathbf{E}\eta\eta^{\mathrm{T}}\\ &= \mathbf{D}_{\theta\theta} + \mathbf{D}_{\theta\xi}\mathbf{D}_{\xi\xi}^{-1}\mathbf{D}_{\xi\xi}\mathbf{D}_{\xi\xi}^{-1}\mathbf{D}_{\theta\xi}^{\mathrm{T}} - 2\mathbf{D}_{\theta\xi}\mathbf{D}_{\xi\xi}^{-1}\mathbf{D}_{\xi\xi}\mathbf{D}_{\xi\xi}^{-1}\mathbf{D}_{\theta\xi}^{\mathrm{T}}\\ &= \mathbf{D}_{\theta\theta} - \mathbf{D}_{\theta\xi}\mathbf{D}_{\xi\xi}^{-1}\mathbf{D}_{\theta\xi}^{\mathrm{T}}.\end{aligned}$$

因为 $\operatorname{cov}(\theta,\theta|\xi)$ 与 “偶然性” 无关，所以

$$\Delta = \mathbf{E}\operatorname{cov}(\theta,\theta|\xi) = \operatorname{cov}(\theta,\theta|\xi),$$

于是，(20) 式得证.　□

系　*设 $(\theta,\xi_1,\cdots,\xi_n)$ 是 $n+1$ 维高斯向量，而且 $\xi_1,\cdots,\xi_n$ 独立. 那么，*

$$\mathbf{E}(\theta|\xi_1,\cdots,\xi_n) = \mathbf{E}\theta + \sum_{i=1}^{n}\frac{\operatorname{cov}(\theta,\xi_i)}{\mathbf{D}\xi_i}(\xi_i - \mathbf{E}\xi_i),$$
$$\Delta = \mathbf{D}\theta - \sum_{i=1}^{n}\frac{\operatorname{cov}^2(\theta,\xi_i)}{\mathbf{D}\xi_i}$$

(对照 §8 公式 (12),(13)).

5. 高斯向量的线性流形的封闭性　设 $\xi_1,\xi_2,\cdots$ 是依概率收敛于向量 ξ 的高斯随机向量序列. 现在证明向量 ξ 仍然服从高斯分布.

由于定理 1 的命题 a)，只需对随机变量证明这一事实.

设 $m_n = \mathbf{E}\xi_n,\ \sigma_n^2 = \mathbf{D}\xi_n$. 那么，根据勒贝格控制收敛定理，有

$$\lim_{n\to\infty} e^{itm_n - \frac{1}{2}\sigma_n^2 t^2} = \lim_{n\to\infty}\mathbf{E}e^{it\xi_n} = \mathbf{E}e^{it\xi}.$$

由于左侧极限的存在性，可见存在 m_n 和 σ_n^2，使

$$m = \lim_{n\to\infty} m_n,\quad \sigma^2 = \lim_{n\to\infty}\sigma_n^2.$$

从而，

$$\mathbf{E}e^{it\xi} = e^{itm - \frac{1}{2}\sigma^2 t^2},$$

即 $\xi \sim \mathscr{N}(m,\sigma^2)$.

特别，由此可见，由高斯随机变量 $\xi_1,\xi_2,\cdots$ 生成的封闭线性流形 $\overline{\mathscr{L}}(\xi_1,\xi_2,\cdots)$ (见 §11 第 5 小节)，仍然由高斯随机变量生成.

6. 一般高斯系及其性质　现在，讨论一般高斯系的定义.

定义 2　设 $\mathfrak{A}$ 是某一下标的集合. 随机变量 $\xi = \{\xi_\alpha\},\ \alpha \in \mathfrak{A}$ 的全体称为高斯系，如果对于任意 $\alpha_1,\cdots,\alpha_n \in \mathfrak{A},\ n \geqslant 1$，随机向量 $(\xi_{\alpha_1},\cdots,\xi_{\alpha_n})$ 是高斯向量.

现在指出高斯系的一些性质.

a) 如果 $\xi = \{\xi_\alpha\},\ \alpha \in \mathfrak{A}$ 是高斯系，则任意 $\xi' = \{\xi'_{\alpha'}\},\ \alpha' \in \mathfrak{A}' \subseteq \mathfrak{A}$ 子系也是高斯系.

b) 如果 $\xi_\alpha,\ \alpha \in \mathfrak{A}$ 是独立高斯变量，则 $\xi = \{\xi_\alpha\},\ \alpha \in \mathfrak{A}$ 是高斯系.

c) 如果 $\xi = \{\xi_\alpha\},\ \alpha \in \mathfrak{A}$ 是高斯系，则包含由形如 $\sum_{i=1}^{n} c_{\alpha_i}\xi_{\alpha_i}$ 的变量及其均方极限的闭线性流形 $\overline{\mathscr{L}}(\xi)$ 是高斯系.

注意，性质 a) 的逆命题一般不成立. 例如，假设 ξ_1 和 η_1 独立，且 $\xi_1 \sim \mathscr{N}(0,1)$，$\eta_1 \sim \mathscr{N}(0,1)$. 定义系统如下

$$(\xi,\eta) = \begin{cases}(\xi_1,|\eta_1|), & \text{若 } \xi_1 \geqslant 0,\\ (\xi_1,-|\eta_1|), & \text{若 } \xi_1 < 0.\end{cases} \tag{23}$$

那么，不难验证 ξ 和 η 中每一个都是高斯变量，然而 (ξ,η) 却不是高斯向量.

设 $\xi = \{\xi_\alpha\}, \alpha \in \mathfrak{A}$ 是高斯系, 其均值 "向量" 为 $m = \{m_\alpha\}, \alpha \in \mathfrak{A}$, 而协方差 "矩阵" 为 $R = (r_{\alpha\beta})_{\alpha,\beta\in\mathfrak{A}}$, 其中 $m_\alpha = \mathbf{E}\xi_\alpha$. "矩阵" R 显然是对称的 $(r_{\alpha\beta} = r_{\beta\alpha})$, 并在如下意义上是非负定的: 对于任意在 $\mathbb{R}^{\mathfrak{A}}$ 取值的 "向量" $c = \{c_\alpha\}_{\alpha\in\mathfrak{A}}$, 只有有限个坐标 c_α 不为 0,

$$(Rc, c) = \sum_{\alpha,\beta} r_{\alpha\beta} c_\alpha c_\beta \geqslant 0 \tag{24}$$

现在提出反问题. 假设给定一参数的集合 $\mathfrak{A} = \{\alpha\}$、"向量" $m = \{m_\alpha\}_{\alpha\in\mathfrak{A}}$ 和对称非负定 "矩阵" $R = (r_{\alpha\beta})_{\alpha,\beta\in\mathfrak{A}}$. 问是否存在概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 和空间上的高斯随机变量系 $\xi = \{\xi_\alpha\}, \alpha \in \mathfrak{A}$, 使

$$\mathbf{E}\xi_\alpha = m_\alpha,$$
$$\operatorname{cov}(\xi_\alpha, \xi_\beta) = r_{\alpha\beta}, \quad \alpha, \beta \in \mathfrak{A}?$$

如果取有限组 $\alpha_1, \cdots, \alpha_n$, 则根据向量 $\overline{m} = (m_{\alpha_1}, \cdots, m_{\alpha_n})$ 和矩阵 $\overline{R} = (r_{\alpha\beta})_{\alpha,\beta\in\mathfrak{A}}$, $\alpha, \beta = \alpha_1, \cdots, \alpha_n \in \mathbb{R}^n$, 可以建立高斯分布 $F_{\alpha_1,\cdots,\alpha_n}(x_1, \cdots, x_n)$, 使其特征函数为

$$\varphi(t) = e^{i(t,\overline{m}) - \frac{1}{2}(\overline{R}t,t)}, \quad t = (t_{\alpha_1}, \cdots, t_{\alpha_n}).$$

不难验证, 分布族

$$\{F_{\alpha_1,\cdots,\alpha_n}(x_1, \cdots, x_n); \alpha_i \in \mathfrak{A}\}$$

是一致的. 从而, 根据柯尔莫戈洛夫定理 (§9 定理 1 及其注 2) 对于上面所提问题的答案是肯定的.

7. 高斯随机序列和高斯过程　假如 $\mathfrak{A} = \{1, 2, \cdots\}$, 则按 §5 所采用的术语, 随机变量系 $\xi = \{\xi_\alpha\}, \alpha \in \mathfrak{A}$ 称做随机序列, 并记作 $\xi = (\xi_1, \xi_2, \cdots)$. 高斯序列 $\xi = (\xi_1, \xi_2, \cdots)$ 完全由均值向量 $m = (m_1, m_2, \cdots)$ 和协方差矩阵 $R = (r_{ij})$, 其中 $r_{ij} = \operatorname{cov}(\xi_i, \xi_j)$ 描绘. 特别, 如果 $r_{ij} = \sigma_i^2\delta_{ij}$, 则 $\xi = (\xi_1, \xi_2, \cdots)$ 是独立随机变量的高斯序列, 其中 $\xi_i \sim N(m_i, \sigma_i^2), i \geqslant 1$.

对于 $\mathfrak{A} = [0,1], [0,\infty), (-\infty,\infty), \cdots$ 的情形, 随机变量系 $\xi = \{\xi_t\}, t \in \mathfrak{A}$ 称做连续时间随机过程.

现在举几个高斯随机过程的例子. 假如认为其均值为 0, 则这样过程的概率性质完全决定于其协方差矩阵 $R = (r_{st}), s, t \in \mathfrak{A}$ 的形式. 我们用 $r(s,t)$ 表示 r_{st}, 并称此 s 和 t 函数为*协方差函数*.

例 1　如果 $\mathfrak{A} = [0, \infty)$, 而

$$r(s,t) = \min(s,t), \tag{25}$$

则以 $r(s,t)$ 为协方差函数 (见练习题 2), 且 $B_0 = 0$ 的高斯过程 $B = \{B_t\}_{t\geqslant 0}$ 称做*布朗运动过程*或*维纳过程*.

注意, 该过程具有独立增量, 即对于任意 $t_1 < t_2 < \cdots < t_n$ 随机变量

$$B_{t_2} - B_{t_1}, \cdots, B_{t_n} - B_{t_{n-1}}$$

是独立的. 事实上, 由于高斯性只需验证它们两两不相关. 实际上, 若 $s < t < u < v$, 则

$$\begin{aligned}&\mathbf{E}(B_t - B_s)(B_v - B_u)\\&= [r(t,v) - r(t,u)] - [r(s,v) - r(s,u)]\\&= (t - t) - (s - s) = 0.\end{aligned}$$

注　§9 第 4 小节考虑的更新过程的例子 (是根据独立随机变量序列 $\sigma_1, \sigma_2, \cdots$ 的构造性给出的), 使得可能类似地建立某种类型的布朗运动的想法产生了.

以独立同分布的标准高斯随机变量序列 $\xi_1, \xi_2, \cdots$ 为基础, 其中 $\xi_i \sim N(0,1)$, 来构造布朗运动的情形确实存在.

例如, 构造随机变量

$$B_t = \frac{\sqrt{2}}{\pi} \sum_{n=1}^{\infty} \frac{\xi_n}{n + 1/2} \sin((n + 1/2)\pi t), \quad t \in [0,1]. \tag{26}$$

由下面将给出的 "两级数" 定理 (第四章 §3 定理 2), 可见对于每一个 $t \in [0,1]$, 定义 B_t 的级数 ($\mathbf{P}$ - a.c.) 收敛. 进一步更细致地讨论表明, 此级数 ($\mathbf{P}$ - a.c.) 一致收敛, 从而过程 $B = (B_t)_{0\leqslant t\leqslant 1}$ ($\mathbf{P}$ - a.c.) 有连续轨道. 该过程的正态性, 由定理 1 的 b) 和第 5 小节中关于 "高斯随机变量依概率收敛的极限, 仍然服从高斯分布" 的论断可以证明. 同样不难证明, 协方差函数为 $r(s,t) = \mathbf{E}B_sB_t = \min(s,t)$.

这样, 在 (26) 式中建立的过程 $B = (B_t)_{0\leqslant t\leqslant 1}$ 满足布朗运动过程定义中的所有要求, 此外该过程 ($\mathbf{P}$ - a.c.) 有连续轨道. 通常, (物理应用所希望的和得到证实的) 轨道的连续性, 包含在布朗运动的定义之中. 我们看到, 这样的过程确实存在.

我们再指出一个建立布朗运动过程的熟知方法. 它基于在 §11 第 5 小节中引进的哈尔函数 $H_n(x), x \in [0,1], n = 1, 2, \cdots$.

由哈尔函数 $H_n(x)$ 建立绍德尔 (A. Schauder) 函数 $S_n(t), t \in [0,1], n = 1, 2, \cdots$:

$$S_n(t) = \int_0^t H_n(x)\mathrm{d}x. \tag{27}$$

那么, 如果 $\xi_0, \xi_1, \xi_2, \cdots$ 是独立同标准高斯分布随机变量序列, 其中 $\xi_i \sim N(0,1)$, 则级数对于 $t \in [0,1]$,

$$B_t = \sum_{n=1}^{\infty} \xi_n S_n(t) \tag{28}$$

以概率 1 一致收敛. 过程 $B = (B_t)_{0\leqslant t\leqslant 1}$ 是布朗运动.

例 2 过程 $B^0=(B_t^0), t\in\mathfrak{U}=[0,1], B_0^0\equiv 0$ 和

$$r(s,t)=\min(s,t)-st, \tag{29}$$

称做条件维纳过程或布朗桥 (注意, 由于 $r(1,1)=0$, 故 $\mathbf{P}\{B_1^0=0\}=1$).

例 3 过程 $X=(X_t), t\in\mathfrak{U}=(-\infty,\infty)$ 和

$$r(s,t)=e^{-|t-s|} \tag{30}$$

称做高斯–马尔可夫过程.

8. 布朗运动的一个简单例子 我们介绍一个布朗运动的有趣性质, 其证明是 §10 中博雷尔–坎泰利引理应用的很好的演示 (确切地说该引理的系 1).

定理 3 *设 $B=(B_t)_{t\geqslant 0}$ 是标准布朗运动, 那么, 对于任意 $T>0$, 依概率 1*

$$\lim_{n\to\infty}\sum_{k=1}^{2^nT}[B_{k2^{-n}}-B_{(k-1)2^{-n}}]^2=T. \tag{31}$$

证明 不失普遍性, 可以认为 $T=1$. 设

$$A_n^\varepsilon=\left\{\omega:\left|\sum_{k=1}^{2^n}(B_{k2^{-n}}-B_{(k-1)2^{-n}})^2-1\right|\geqslant\varepsilon\right\}.$$

由于 $B_{k2^{-n}}-B_{(k-1)2^{-n}}$ 是高斯随机变量, 且均值为 0, 而方差等于 2^{-n}, 可见

$$\mathbf{D}\left(\sum_{k=1}^{2^n}[B_{k2^{-n}}-B_{(k-1)2^{-n}}]^2\right)=2^{-n+1};$$

因此, 由切比雪夫不等式, 有 $\mathbf{P}(A_n^\varepsilon)\leqslant\varepsilon^{-2}2^{-n+1}$, 从而

$$\sum_{n=1}^{\infty}\mathbf{P}(A_n^\varepsilon)\leqslant\varepsilon^{-2}\sum_{n=1}^{\infty}2^{-n+1}=2\varepsilon^{-2}<\infty. \tag{32}$$

于是, 由此估计式和博雷尔–坎泰利引理的系 1 (§10), 得所要求的关系式 (31). □

9. 练习题

1. 设 ξ_1,ξ_2,ξ_3 是独立高斯随机变量, $\xi_i\sim N(0,1)$. 证明

$$\frac{\xi_1+\xi_2\xi_3}{\sqrt{1+\xi_3^2}}\sim N(0,1).$$

(由此产生了对于有趣的研究课题: 描绘独立高斯随机变量 $\xi_1,\cdots,\xi_n$ 的非线性变换, 使变换的结果仍然是高斯随机变量.)

2. 证明, 由 (25), (29) 和 (30) 各式的函数 $r(s,t)$ 给出的 "矩阵" $R=(r(s,t))_{s,t\in\mathfrak{U}}$ 是非负定的.

3. 设 $\boldsymbol{A}$ 是 $m\times n$ 阶矩阵. 称 $\boldsymbol{A}^{\oplus}$ 为 $\boldsymbol{A}$ 的 $n\times m$ 阶伪逆矩阵, 如果存在矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{V}$, 使

$$\boldsymbol{A}\boldsymbol{A}^{\oplus}\boldsymbol{A}=\boldsymbol{A},\quad \boldsymbol{A}^{\oplus}=\boldsymbol{U}\boldsymbol{A}^{\mathrm{T}}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}.$$

证明, 由这些条件决定的矩阵 $\boldsymbol{A}^{\oplus}$ 存在而且唯一.

4. 证明在正态相关定理的公式 (19) 和 (20) 中, 对于 $\mathbf{D}_{\xi\xi}$ 是退化矩阵的情形, 如果在这些公式中用伪逆矩阵 $\mathbf{D}_{\xi\xi}^{\oplus}$ 代替 $\mathbf{D}_{\xi\xi}^{-1}$, 则正态相关定理仍然成立.

5. 设 $(\theta,\xi)=(\theta_1,\cdots,\theta_k;\xi_1,\cdots,\xi_l)$ 是具有非退化矩阵 $\Delta=\mathbf{D}_{\theta\theta}-\mathbf{D}_{\theta\xi}^{\oplus}\mathbf{D}_{\theta\xi}^{\mathrm{T}}$ 的高斯向量. 证明分布函数 $\mathbf{P}\{\theta\leqslant a|\xi\}=\mathbf{P}\{\theta_1\leqslant a_1,\cdots,\theta_k\leqslant a_k|\xi\}$ ($\mathbf{P}$-a.c.) 有密度 $p(a_1,\cdots,a_k|\xi)$, 且

$$p(a_1,\cdots,a_k|\xi)=\frac{|\Delta|^{-1/2}}{(2\pi)^{k/2}}\exp\left\{-\frac{1}{2}(a-\mathrm{E}(\theta|\xi))^{\mathrm{T}}\Delta^{-1}(a-\mathrm{E}(\theta|\xi))\right\}.$$

6. (C.H. 伯恩斯坦) 设 ξ 和 η 是具有有限方差的独立同分布随机变量. 证明, 如果 $\xi+\eta$ 和 $\xi-\eta$ 独立, 则 ξ 和 η 是高斯随机变量.

7. 默塞 (J. Mercer) 定理. 假设 $r(s,t)$ 是 $[a,b]\times[a,b](-\infty<a<b<\infty)$ 上连续的协方差函数. 证明, 对于无限多个 $\lambda_k>0$, 方程

$$\lambda\int_a^b r(s,t)u(t)\mathrm{d}t=u(s),\ a\leqslant s\leqslant b,$$

相应的连续解 $\{u_k,k\geqslant 1\}$ 的解系, 构成 $L^2(a,b)$ 中的完全规范正交系, 使

$$r(s,t)=\sum_{k=1}^{\infty}\frac{u_k(s)u_k(t)}{\lambda_k},$$

其中级数在 $[a,b]\times[a,b]$ 上绝对收敛且一致收敛.

8. 设 $X=\{X_t,t\geqslant 0\}$ 是高斯过程, 其中 $\mathbf{E}X_t=0$, 而协方差函数 $r(s,t)=e^{-|t-s|}$, $s,t\geqslant 0$. 设 $0<t_1<\cdots<t_n$, 而 $f_{t_1,\cdots,t_n}(x_1,\cdots,x_n)$ 是随机变量 $X_{t_1},\cdots,X_{t_n}$ 的密度. 证明:

$$f_{t_1,\cdots,t_n}(x_1,\cdots,x_n)=\left[(2\pi)^n\prod_{i=2}^{n}(1-e^{2(t_{i-1}-t_i)})\right]^{-1/2}$$
$$\times\exp\left\{-\frac{x_1^2}{2}-\frac{1}{2}\sum_{i=2}^{n}\frac{[x_i-e^{(t_{i-1}-t_i)}x_{i-1}]^2}{1-e^{2(t_{i-1}-t_i)}}\right\}.$$

9. 设 $f=\{f_n,n\geqslant 1\}\subset L^2(0,1)$ 是完备规范正交系, 而 $\{\xi_n\}$ 是独立同分布 $N(0,1)$ 随机变量. 证明过程

$$B_t=\sum_{n\geqslant 1}\xi_n\int_0^t f_n(u)\mathrm{d}u$$

是布朗运动.

10. 证明当 $(\xi,\eta_1,\cdots,\eta_k)$ 高斯系时, 条件数学期望 $\mathrm{E}(\xi|\eta_1,\cdots,\eta_k)$ 等于广义数学期望 $\hat{\mathrm{E}}(\xi|\eta_1,\cdots,\eta_k)$.

11. 设 $(\xi,\eta_1,\cdots,\eta_k)$ 是高斯系, 说明条件数学期望 $\mathrm{E}(\xi^n|\eta_1,\cdots,\eta_k)$, $n\geqslant 1$ (作为 $\eta_1,\cdots,\eta_k$ 的函数) 的构造.

12. 设 $X=(X_k)_{1\leqslant k\leqslant n}$ 和 $Y=(Y_k)_{1\leqslant k\leqslant n}$ 是两个高斯随机序列, 且 $\mathbf{E}X_k=\mathbf{E}Y_k$, $\mathbf{D}X_k=\mathbf{D}Y_k$, $1\leqslant k\leqslant n$, 而

$$\operatorname{cov}(X_k,X_l)\leqslant \operatorname{cov}(Y_k,Y_l),\ 1\leqslant k,l\leqslant n.$$

证明斯莱皮恩 (P. Slepian) 不等式: 对于任意 $x\in R$,

$$\mathbf{P}\left\{\sup_{1\leqslant k\leqslant n} X_k<x\right\}\leqslant \mathbf{P}\left\{\sup_{1\leqslant k\leqslant n} Y_k<x\right\}.$$

13. 证明, 如果 $B^\circ=(B^\circ_t)_{0\leqslant t\leqslant 1}$ 是布朗桥, 则过程 $B=(B_t)_{t\geqslant 0}$ 是布朗运动, 其中 $B_t=(1+t)B^\circ_{t/(1+t)}$.

14. 对于布朗运动 $B=(B_t)_{t\geqslant 0}$, 验证下列过程也是布朗运动:

$$B^{(1)}_t=-B_t;$$
$$B^{(2)}_t=tB_{1/t},\ t>0 \text{ 和 } B^{(2)}_0=0;$$
$$B^{(3)}_t=B_{t+s}-B_s,\ s>0;$$
$$B^{(4)}_t=B_T-B_{T-t},\ 0\leqslant t\leqslant T,\ T>0;$$
$$B^{(5)}_t=\frac{1}{a}B_{a^2t},\ a>0 \text{ (自模态性)}.$$

15. 设 $X=(X_k)_{1\leqslant k\leqslant n}$ 是高斯随机序列, 且

$$m=\max_{1\leqslant k\leqslant n}\mathbf{E}X_k,\quad \sigma^2=\max_{1\leqslant k\leqslant n}\mathbf{D}X_k,$$

而对于某个 a,

$$\mathbf{P}\left\{\sup_{1\leqslant k\leqslant n}(X_k-\mathbf{E}X_k)\geqslant a\right\}\leqslant\frac{1}{2}.$$

那么, 博雷尔不等式成立:

$$\mathbf{P}\left\{\sup_{1\leqslant k\leqslant n}X_k>x\right\}\leqslant 2\Psi\left(\frac{x-m-a}{\sigma}\right),$$

其中

$$\Psi(x)=\frac{1}{\sqrt{2\pi}}\int_x^{\infty}\mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y.$$

16. 设 (X,Y) 是二维高斯随机变量, $\mathbf{E}X=\mathbf{E}Y=0$, $\mathbf{E}X^2>0$, $\mathbf{E}Y^2>0$, 其相关系数为

$$\rho=\frac{\mathbf{E}XY}{\sqrt{\mathbf{E}X^2\mathbf{E}Y^2}}.$$

证明

$$\mathbf{P}\{XY<0\}=1-2\mathbf{P}\{X>0,Y>0\}=\pi^{-1}\arccos\rho.$$

17. 设 $Z=XY$, 其中 $X\sim\mathcal{N}(0,1)$, 且 $\mathbf{P}\{Y=1\}=\mathbf{P}\{Y=-1\}=1/2$. 求随机向量 (X,Z) 和 (Y,Z) 的分布, 以及 $X+Z$ 的分布. 证明, $Z\sim\mathcal{N}(0,1)$, 且 X 和 Z 不相关, 但是不独立.

18. 详细证明, 由 (26),(28) 式定义的过程 $(B_t)_{0\leqslant t\leqslant 1}$ 是布朗运动.

19. 设 $B^\mu=(B_t+\mu t)_{t\geqslant 0}$ 是带漂移的布朗运动.

(a) 求变量 $B^\mu_{t_1}+B^\mu_{t_2}$, $t_1<t_2$ 的分布;

(b) 对于 $t_0<t_1<t_2$, 求 $\mathbf{E}B^\mu_{t_0}B^\mu_{t_1}$ 和 $\mathbf{E}B^\mu_{t_0}B^\mu_{t_1}B^\mu_{t_2}$.

20. 对于上题的过程 B^μ, 对于 $t_1<t_2$ 和 $t_1>t_2$, 求条件分布:

$$\mathbf{P}\{B^\mu_{t_2}\in\cdot\,|B^\mu_{t_1}\};$$

对于 $t_0<t_1<t_2$, 求条件分布:

$$\mathbf{P}\{B^\mu_{t_1}\in\cdot\,|B^\mu_{t_0},B^\mu_{t_2}\}.$$

第三章 概率测度的接近程度和收敛性，中心极限定理

对于概率论教程形式上的结构, 极限定理是概率论初等章节的一种上层建筑, 其中所有问题都具有最终的、纯算术的特点. 不过, 概率论认识的价值, 实际上只有通过极限定理才揭示出来. 何况, 没有极限定理就无法理解我们的全部学科——概率论有关概念的真实内容.

Б. В. 格涅坚科, А. Н. 柯尔莫戈洛夫.
《独立随机变量之和的极限分布》[16]

§1. 概率测度和分布的弱收敛

1. 概述 概率论许多结果都具有极限定理的形式. 以极限定理的形式, 建立了 J. 伯努利大数定律和棣莫弗 – 拉普拉斯定理. 可以说, 从而奠定了真正概率论的基础. 特别是指出了众多研究的方向, 说明了不同形式的 "大数定律" 和 "中心极限定理" 成立的条件. 以极限定理的形式还到立了, "在稀有事件情形下, 用 '泊松分布' 逼近二项分布的" 泊松定理. 在这些定理的例子中, 以及在有关棣莫弗 - 拉普拉斯定理和泊松定理的收敛速度的结果中, 已经可以看到, 在概率论中必然涉及分布的不同形式, 而说明*收敛速度*, 需要引进分布之间各种接近程度的 "自然" 度量.

在这一章将讨论概率分布的收敛性, 及其接近程度的某些一般性问题. 这一节讨论度量空间中*概率测度弱收敛*的一般理论问题 (特别, 伯努利大数定律和棣莫弗 – 拉普拉斯定理——"中心极限定理" 的 "始祖", 正是属于这一范围.) 从 §3 开始就可以清楚地看到特征函数方法, 是证明 $\mathbf{R}^n$ 中概率分布弱收敛的极限定理最强有力的工具之一. 在 §7 中将讨论弱收敛的 "可度量化" 问题. 在 §9 中将讨论分布另一种形式的收敛性: *按变差收敛* (一种比弱收敛更强的收敛). 在中心极限定理和泊松定理中, 收敛速度最简单结果的证明将在 §11 和 §12 中给出. 在 §13 中将 §1 和 §2 中关于弱收敛的结果用于某些 (原则上重要的) 数理统计问题.

2. 伯努利模型中的收敛性 作为开始, 我们回忆伯努利模型中大数定律的提法 (第一章 §5).

设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量序列, 且 $\mathbf{P}\{\xi_i = 1\} = p, \mathbf{P}\{\xi_i = 0\} = q, p + q = 1$. 利用第二章 §10 中引进的依概率收敛的概念, J. 伯努利大数定律可以表述为:

$$\frac{S_n}{n} \xrightarrow{\mathbf{P}} p, \quad n \to \infty, \tag{1}$$

其中 $S_n = \xi_1 + \cdots + \xi_n$. (在第四章将证明, 这里实际上也依概率 1 收敛.)

记

$$F_n(x)=\mathbf{P}\left\{\frac{S_n}{n}\leqslant x\right\},$$

$$F(x)=\begin{cases}1, & 若\ x\geqslant p,\\ 0, & 若\ x<p.\end{cases} \tag{2}$$

其中 $F(x)$ 是退化随机变量 $\xi\equiv p$ 的分布函数. 设 P_n 和 P 是空间 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ 上对应于分布函数 F_n 和 F 的概率测度.

根据第二章 §10 定理 2, 依概率收敛 $S_n/n\xrightarrow{\mathbf{P}}p$ 必导致按分布收敛 $S_n/n\xrightarrow{d}p$, 而由按分布收敛可见, 对于 $\mathbb{R}$ 上的有界连续函数类 C 中的任何函数 $f=f(x)$, 有

$$\mathbf{E}f\left(\frac{S_n}{n}\right)\to\mathbf{E}f(p). \tag{3}$$

由于

$$\mathbf{E}f\left(\frac{S_n}{n}\right)=\int_{\mathbb{R}}f(x)P_n(\mathrm{d}x),\quad \mathbf{E}f(p)=\int_{\mathbb{R}}f(x)P(\mathrm{d}x),$$

则由 (3) 式, 可以将其写为

$$\int_{\mathbb{R}}f(x)P_n(\mathrm{d}x)\to\int_{\mathbb{R}}f(x)P(\mathrm{d}x),\quad f\in C, \tag{4}$$

或 (根据第二章 §6 的记号) 将其写为

$$\int_{\mathbb{R}}f(x)\mathrm{d}F_n(x)\to\int_{\mathbb{R}}f(x)\mathrm{d}F(x),\quad f\in C. \tag{5}$$

在数学分析中, 收敛性 (4) 称做 (当 $n\to\infty$ 时测度 P_n 向测度 P 的) 弱收敛, 记作 $P_n\xrightarrow{w}P$ (对照定义 2). 自然, 收敛性 (5) 亦称做分布函数 F_n 向 F 的弱收敛, 记作 $F_n\xrightarrow{w}F$.

这样, 在伯努利概型中, 可以断定

$$\frac{S_n}{n}\xrightarrow{\mathbf{P}}p\Rightarrow F_n\xrightarrow{w}F. \tag{6}$$

这样, 由 (1) 式亦不难导出, 对于由 (2) 式引进的分布函数, 对于除函数 $F(x)$ 的一个间断点 $x=p$ 之外的所有点 $x\in\mathbb{R}$, 有

$$F_n(x)\to F(x),\quad n\to\infty.$$

这一事实说明, 对于所有点 $x\in\mathbb{R}$, 当 $n\to\infty$ 时弱收敛 $F_n\xrightarrow{w}F$ 并不会导致函数 $F_n(x)$ 向 $F(x)$ 的逐点收敛. 不过, 结果表明, 无论是伯努利概型, 还是任意分布函数的一般情形, 弱收敛等价于下面定义的所谓基本收敛 (见下面定理 2).

定义 1 称定义在数轴上的分布函数序列 $\{F_n\}$, 基本收敛于分布函数 F (记作 $F_n\Rightarrow F$), 如果当 $n\to\infty$ 时,

$$F_n(x)\to F(x),\quad x\in C(F),$$

其中 $C(F)$ 是极限函数 F 的连续点的集合.

对于所考虑的伯努利概型, 函数 $F=F(x)$ 是退化的, 而由此不难导出 (参见第二章 §10 练习题 7):

$$(F_n\Rightarrow F)\Rightarrow\left(\frac{S_n}{n}\xrightarrow{\mathbf{P}}p\right).$$

这样, 考虑到下面的定理 2, 有

$$\left(\frac{S_n}{n}\xrightarrow{\mathbf{P}}p\right)\Rightarrow\left(F_n\xrightarrow{w}F\right)\Leftrightarrow(F_n\Rightarrow F)\Rightarrow\left(\frac{S_n}{n}\xrightarrow{\mathbf{P}}p\right). \tag{7}$$

因而大数定律的论断, 可以视为由 (2) 式定义的关于分布函数弱收敛的论点之一.

记

$$\begin{aligned}F_n(x)&=\mathbf{P}\left\{\frac{S_n-np}{\sqrt{npq}}\leqslant x\right\},\\ F(x)&=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-u^2/2}\mathrm{d}u.\end{aligned} \tag{8}$$

根据棣莫弗-拉普拉斯定理 (第一章 §6), 对于一切 $x\in\mathbb{R}$, $F_n(x)\to F(x)$, 从而 $F_n\Rightarrow F$. 由于上面指出的弱收敛 $F_n\xrightarrow{w}F$ 与基本收敛 $F_n\Rightarrow F$ 等价性, 则可以认为棣莫弗-拉普拉斯定理, 也是关于 (8) 式定义的分布函数的弱收敛的论断.

这两个例子证实下面定义 2 引进的概率测度弱收敛概念的合理性. 虽然对于数轴的情形, 弱收敛与分布函数的基本收敛等价, 不过作为开始还是倾向于考虑弱收敛: 第一, 因为弱收敛更便于分析; 第二, 因为弱收敛对于比数轴更一般的空间也有意义, 特别, 对于度量空间, 最重要的例子是空间 $\mathbb{R}^n$, $\mathbb{R}^\infty$, C 和 D (见第二章 §3).

3. 弱收敛和基本收敛 设 $(E,\mathscr{E},\rho)$ 是度量空间, $\rho=\rho(x,y)$ 为度量, $\mathscr{E}$ 为由开集生成的 σ-代数, 而 $\mathbf{P},\mathbf{P}_1,\mathbf{P}_2,\cdots$ 是空间 $(E,\mathscr{E},\rho)$ 上的概率测度.

定义 2 称概率测度序列 $\{\mathbf{P}_n\}$ 弱收敛于概率测度 $\mathbf{P}$ (记作 $\mathbf{P}_n\xrightarrow{w}\mathbf{P}$, 其中 w 是英文词 weak convergence (弱收敛) 的字头), 如果对于 $\mathscr{E}$ 上的连续有界函数类 $C(\mathscr{E})$ 中的任意函数 $f=f(x)$, 有

$$\int_E f(x)\mathbf{P}_n(\mathrm{d}x)\to\int_E f(x)\mathbf{P}(\mathrm{d}x). \tag{9}$$

定义 3 称概率测度序列 $\{\mathbf{P}_n\}$ 基本收敛于概率测度 $\mathbf{P}$ (记作 $\mathbf{P}_n\Rightarrow\mathbf{P}$), 如果

$$\mathbf{P}_n(A)\to\mathbf{P}(A), \tag{10}$$

其中 $A \in \mathscr{E}$ 是任意集合, 满足

$$\mathbf{P}(\partial A) = 0, \tag{11}$$

而 ∂A 表示集合 A 的边界:

$$\partial A = [A] \cap [\overline{A}],$$

其中 $[A]$ 表示集合 A 的闭包.

下面的定理表明, 概率测度弱收敛的概念与概率测度基本收敛的概念等价, 以及其他等价的提法.

定理 1 *下列各命题等价:*

(I) $\mathbf{P}_n \xrightarrow{w} \mathbf{P}$;

(II) 对于闭集 A, $\overline{\lim}\, \mathbf{P}_n(A) \leqslant \mathbf{P}(A)$;

(III) 对于开集 A, $\underline{\lim}\, \mathbf{P}_n(A) \geqslant \mathbf{P}(A)$;

(IV) $\mathbf{P}_n \Rightarrow \mathbf{P}$.

证明 (I) $\Rightarrow$ (II). 设 A 是闭集, 记

$$f_A^\varepsilon(x) = \left[1 - \frac{\rho(x, A)}{\varepsilon}\right]^+,$$

其中

$$\rho(x, A) = \inf\{\rho(x, y) : y \in A\}, \quad [x]^+ = \max[0, x].$$

再记

$$A^\varepsilon = \{x : \rho(x, A) < \varepsilon\},$$

并注意到当 $\varepsilon \downarrow 0$ 时 $A^\varepsilon \downarrow A$.

由于函数 $f_A^\varepsilon(x)$ 有界、连续并且满足

$$\mathbf{P}_n(A) = \int_E I_A(x)\mathbf{P}_n(\mathrm{d}x) \leqslant \int_E f_A^\varepsilon(x)\mathbf{P}_n(\mathrm{d}x),$$

则当 $\varepsilon \downarrow 0$ 时

$$\overline{\lim_n}\, \mathbf{P}_n(A) \leqslant \overline{\lim_n} \int_E f_A^\varepsilon(x)\mathbf{P}_n(\mathrm{d}x) = \int_E f_A^\varepsilon(x)\mathbf{P}(\mathrm{d}x) \leqslant \mathbf{P}(A^\varepsilon) \downarrow \mathbf{P}(A),$$

因此 (I) $\Rightarrow$ (II) 得证.

如果由集合转换为其相应的补集, 则蕴涵关系 (II) $\Rightarrow$ (III) 和 (III) $\Rightarrow$ (II) 显然.

(III) $\Rightarrow$ (IV). 设 $A^0 = A \backslash \partial A$ 是集合 A 的内部, 而 $[A]$ 是集合 A 的闭包. 由于 (II) 和 (III) 以及假设 $\mathbf{P}(\partial A) = 0$, 则

$$\overline{\lim_n}\, \mathbf{P}_n(A) \leqslant \overline{\lim_n}\, \mathbf{P}_n([A]) \leqslant \mathbf{P}([A]) = \mathbf{P}(A),$$

$$\underline{\lim_n}\, \mathbf{P}_n(A) \geqslant \underline{\lim_n}\, \mathbf{P}_n(A^0) \geqslant \mathbf{P}(A^0) = \mathbf{P}(A),$$

因此对于每个 A, 若 $\mathbf{P}(\partial A) = 0$, 则 $\mathbf{P}_n(A) \to \mathbf{P}(A)$.

(IV) $\Rightarrow$ (I). 设 $f = f(x)$ 是连续有界函数, $|f(x)| \leqslant M$. 记

$$D = \{t \in R : \mathbf{P}\{x : f(x) = t\} \neq 0\};$$

考虑区间 $[-M, M]$ 的分割 $T_k = (t_0, t_1, \cdots, t_k)$:

$$-M = t_0 < t_1 < \cdots < t_k = M, \quad k \geqslant 1,$$

其中 $t_i \notin D$, $i = 0, 1, \cdots, k$. (注意, 由于集合 $f^{-1}(t)$ 两两不相交, 而测度 $\mathbf{P}$ 有限, 故集合 D 有限或可数.)

设 $B_i = \{x : t_i \leqslant f(x) < t_{i+1}\}$. 由于函数 $f(x)$ 连续, 可见 $f^{-1}(t_i, t_{i+1})$ 是开集. 则

$$\partial B_i \subseteq f^{-1}\{t_i\} \cup f^{-1}\{t_{i+1}\}.$$

点 $t_i, t_{i+1} \notin D$. 因此 $\mathbf{P}(\partial B_i) = 0$. 故由 (IV), 有

$$\sum_{i=0}^{k-1} t_i \mathbf{P}_n(B_i) \to \sum_{i=0}^{k-1} t_i \mathbf{P}(B_i). \tag{12}$$

由于

$$\begin{aligned}\left|\int_E f(x)\mathbf{P}_n(\mathrm{d}x) - \int_E f(x)\mathbf{P}(\mathrm{d}x)\right| &\leqslant \left|\int_E f(x)\mathbf{P}_n(\mathrm{d}x) - \sum_{i=0}^{k-1} t_i \mathbf{P}_n(B_i)\right| \\ &+ \left|\sum_{i=0}^{k-1} t_i \mathbf{P}_n(B_i) - \sum_{i=0}^{k-1} t_i \mathbf{P}(B_i)\right| + \left|\sum_{i=0}^{k-1} t_i \mathbf{P}(B_i) - \int_E f(x)\mathbf{P}(\mathrm{d}x)\right| \\ &\leqslant 2 \max_{0 \leqslant i \leqslant k-1}(t_{i+1} - t_i) + \left|\sum_{i=0}^{k-1} t_i \mathbf{P}_n(B_i) - \sum_{i=0}^{k-1} t_i \mathbf{P}(B_i)\right|,\end{aligned}$$

而由 (12) 式及分割 T_k $(k \geqslant 1)$ 的任意性, 可见

$$\lim_n \int_E f(x)\mathbf{P}_n(\mathrm{d}x) = \int_E f(x)\mathbf{P}(\mathrm{d}x). \quad \square$$

注 1 证明中出现的函数 $f(x) = I_A(x)$ 和 $f_A^\varepsilon(x)$ 的蕴涵关系 (I) $\Rightarrow$ (II), 相应为上半连续性和一致连续性. 注意到这种情况就不难证明, 定理的每一个条件, 与下列条件之一等价:

(V) *对于一切有界一致连续函数 $f(x)$, 有*

$$\int_E f(x)\mathbf{P}_n(\mathrm{d}x) \to \int_E f(x)\mathbf{P}(\mathrm{d}x);$$

(VI) *对于一切满足利普希茨 (R. Lipschitz) 条件 (见 §7 引理 2) 的有界函数 $f(x)$, 有*

$$\int_E f(x)\mathbf{P}_n(\mathrm{d}x) \to \int_E f(x)\mathbf{P}(\mathrm{d}x);$$

(VII) 对于一切有界的上半连续函数 $f(x)$($\overline{\lim}_n f(x_n) \leqslant f(x)$, 当 $x_n \to x$ 时), 有

$$\overline{\lim_n} \int_E f(x)\mathbf{P}_n(dx) \leqslant \int_E f(x)\mathbf{P}(dx);$$

(VIII) 对于一切有界的下半连续函数 $f(x)$($\underline{\lim}_n f(x_n) \geqslant f(x)$, 当 $x_n \to x$ 时), 有

$$\underline{\lim_n} \int_E f(x)\mathbf{P}_n(dx) \geqslant \int_E f(x)\mathbf{P}(dx).$$

注 2 定理 1 可以推广到如下情形: 将空间 $(E, \mathscr{E}, \rho)$ 上的概率测度 $\mathbf{P}_n$ 和 $\mathbf{P}$, 换成任意有限测度 μ_n 和 μ (未必是概率测度). 对于这样的测度, 类似地引进弱收敛 $\mu_n \xrightarrow{w} \mu$ 的概念和基本收敛 $\mu_n \Rightarrow \mu$ 的概念, 并且像定理 1 一样证明下列条件的等价性:

(I*) $\mu_n \xrightarrow{w} \mu$;

(II*) 对于闭集 A, $\overline{\lim}_n \mu_n(A) \leqslant \mu(A)$, 且 $\mu_n(E) \to \mu(E)$;

(III*) 对于开集 A, $\underline{\lim}_n \mu_n(A) \geqslant \mu(A)$, 且 $\mu_n(E) \to \mu(E)$;

(IV*) $\mu_n \Rightarrow \mu$.

如果将条件 (V) ~ (VIII) 中的 $\mathbf{P}_n$ 和 $\mathbf{P}$, 相应地换成 μ_n 和 μ, 则 (V*) ~ (VIII*) 中的每一个条件, 都等价于 (I*) ~ (IV*) 中的任何一个条件.

4. 数轴 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 设 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 是数轴, 其中 $\mathbb{R}$ 是实数轴, 而 $\mathscr{B}(\mathbb{R})$ 是由欧几里得度量 $\rho(x, y) = |x - y|$ 生成的博雷尔 σ-代数 (见第二章 §2 第 2 小节注 2). 记 $P_n(n \geqslant 1)$ 和 P 是 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上的概率测度, 而 $F_n(n \geqslant 1)$ 和 F 是相应 P_n 和 P 的分布函数. 那么, 有下面的定理.

定理 2 下列各条件等价:

(1) $P_n \xrightarrow{w} P$;

(2) $P_n \Rightarrow P$;

(3) $F_n \xrightarrow{w} F$;

(4) $F_n \Rightarrow F$.

证明 由于 (2) ⇔ (1) ⇔ (3), 可见只需证明 (2) ⇔ (4).

如果 $P_n \Rightarrow P$, 则 (特别) 对于所有 $x \in \mathbb{R}(P\{x\} = 0)$, 有

$$P_n(-\infty, x] \to P(-\infty, x].$$

这说明 $F_n \Rightarrow F$.

现在设 $F_n \Rightarrow F$. 为证明收敛性 $P_n \Rightarrow P$, (由于定理 1) 只需证明对于任意开集 A, 有 $\underline{\lim}_n P_n(A) \geqslant P(A)$.

假如 A 是开集, 那么存在可数个 (形如 (a, b) 的) 两两不相交开区间 $I_1, I_2, \cdots$, 使 $A = \sum_{k=1}^{\infty} I_k$. 固定 $\varepsilon > 0$, 在每一个区间 $I_k = (a_k, b_k)$ 中选择一子区间 $I'_k = (a'_k, b'_k]$,

使 $a'_k, b'_k \in \mathbb{C}(F)$, 和 $P(I_k) \leqslant P(I'_k) + \varepsilon 2^{-k}$. (由于函数 $F = F(x)$ 的间断点的个数有限或可数, 这样的区间 $I'_k(k \geqslant 1)$ 确实存在.) 那么, 根据法图引理, 有

$$\underline{\lim_n} P_n(A) = \underline{\lim_n} \sum_{k=1}^{\infty} P_n(I_k) \geqslant \sum_{k=1}^{\infty} \underline{\lim_n} P_n(I_k) \geqslant \sum_{k=1}^{\infty} \underline{\lim_n} P_n(I'_k).$$

然而, 由于

$$P_n(I'_k) = F_n(b'_k) - F_n(a'_k) \to F(b'_k) - F(a'_k) = P(I'_k),$$

因此

$$\underline{\lim_n} P_n(A) \geqslant \sum_{k=1}^{\infty} P(I'_k) \geqslant \sum_{k=1}^{\infty} [P(I_k) - \varepsilon 2^{-k}] = P(A) - \varepsilon.$$

因为 $\varepsilon > 0$ 的任意性, 所以就证明了 "若 A 是开集, 则 $\underline{\lim}_n P_n(A) \geqslant P(A)$". □

5. 定义收敛类 设 $(E, \mathscr{E})$ 是可测空间. 子集系 $\mathscr{K}_0(E) \subseteq \mathscr{E}$ 称做定义类, 如果对于空间 $(E, \mathscr{E})$ 上的任意两个概率测度 $\mathbf{P}$ 和 $\mathbf{Q}$, 由等式

$$\mathbf{P}(A) = \mathbf{Q}(A), \quad 对于一切 \quad A \in \mathscr{K}_0(E)$$

可见测度 $\mathbf{P}$ 和 $\mathbf{Q}$ 等同, 即

$$\mathbf{P}(A) = \mathbf{Q}(A), \quad 对于一切 \quad A \in \mathscr{E}.$$

假如 $(E, \mathscr{E}, \rho)$ 是度量空间, 那么子集系 $\mathscr{K}_1(E) \subseteq \mathscr{E}$ 称做定义收敛类, 如果对任意测度 $\mathbf{P}, \mathbf{P}_1, \mathbf{P}_2, \cdots$, 则由

$$\mathbf{P}_n(A) \to \mathbf{P}(A), \quad 对于一切 \quad A \in \mathscr{K}_1(E) \quad 且 \quad \mathbf{P}(\partial A) = 0$$

得出

$$\mathbf{P}_n(A) \to \mathbf{P}(A), \quad 对于一切 \quad A \in \mathscr{E}, \quad 且 \quad \mathbf{P}(\partial A) = 0.$$

当 $(E, \mathscr{E}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 时, 作为定义类 $\mathscr{K}_0(\mathbb{R})$, 可以选 "初等" 集合类 $\mathscr{K} = \{(-\infty, x] : x \in \mathbb{R}\}$ (第二章 §3 定理 1). 由定理 2 的 (2) 和 (4) 式的等价性条件可见, $\mathscr{K}$ 类也是定义收敛类.

自然, 对于更一般的空间, 也产生定义收敛类的问题.

对于空间 $\mathbb{R}^n(n \geqslant 2)$, 形如 $(-\infty, x] = (-\infty, x_1] \times \cdots \times (-\infty, x_n]$, $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$ 的集合的 "初等" 集合类 $\mathscr{K}$ 是定义类 (第二章 §3 定理 2), 也是定义收敛类 (练习题 2).

对于空间 $\mathbb{R}^\infty$, 柱集是 "初等" 集合, 由柱集的概率可以唯一地确定一切博雷尔集的概率 (第二章 §3 定理 3). 结果, 这时柱集类就是定义收敛类 (练习题 3).

似乎可以指出, 对于更一般的空间, 柱集类是定义收敛类. 然而, 一般并非如此.

例如, 考虑具有均匀度量 ρ 的空间 $(C, \mathscr{B}(C), \rho)$ (第二章 §2 第 6 小节). 设 $\mathbf{P}$ 是完全集中在函数 $x(t) \equiv 0$ $(0 \leqslant t \leqslant 1)$ 上的概率测度, 而 $\mathbf{P}_n(n \geqslant 1)$ 是概率测度, 其中

对于每一个 $n \geqslant 1$, 都全部集中在函数 $x_n = x_n(t)$ 上 (见图 35). 不难验证, 对于一切 $\mathbf{P}(\partial A) = 0$ 的柱集 $A, \mathbf{P}_n(A) \to \mathbf{P}(A)$. 但是, 例如, 若取集合

$$A = \left\{\alpha \in C : |\alpha(t)| \leqslant \frac{1}{2}, 0 \leqslant t \leqslant 1\right\} \in \mathscr{B}_0(C),$$

其中 $\mathbf{P}(\partial A) = 0, \mathbf{P}_n(A) = 0, \mathbf{P}(A) = 1$, 因而 $\mathbf{P}_n \not\Rightarrow \mathbf{P}$.

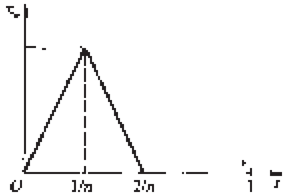


图 35

这样, 对于上面的例子, 柱集类是定义类, 但不是定义收敛类.

6. 练习题

1. 称定义在 $\mathbb{R}^m$ 上的函数 $F = F(x)$ 在点 $x \in \mathbb{R}^m$ 处连续, 如果对于任何 $\varepsilon > 0$, 存在 $\delta > 0$, 使对于一切满足不等式 $|F(x) - F(y)| < \varepsilon$ 的 $y \in \mathbb{R}^m$, 满足不等式

$$x - \delta e < y < x + \delta e,$$

其中 $e = (1, \cdots, 1) \in \mathbb{R}^m$. 此外, 亦称分布函数序列 $\{F_n\}$ 基本收敛于分布函数 F (记作 $F_n \Rightarrow F$), 如果在函数 $F = F(x)$ 的所有连续点 $x \in \mathbb{R}^m$, 有 $F_n(x) \to F(x)$.

证明, 对于 $\mathbb{R}^m (m > 1)$, 定理 2 的结论仍然成立. (见定理 1 的注.)

2. 证明, 对于空间 $\mathbb{R}^m$, "初等" 集合类 $\mathscr{K}$ 是定义收敛类.

3. 设 E 是空间 $\mathbb{R}^\infty$, C 或 D 之一, $\mathscr{E}$ 是开集生成的博雷尔集合代数, 而 $\{\mathbf{P}_n\}$ 是定义在 σ- 代数上 $\mathscr{E}$ 的概率测度序列. 称 $\{\mathbf{P}_n\}$ 在有限维分布的意义上基本收敛于概率测度 $\mathbf{P}$ (记作 $\mathbf{P}_n \stackrel{f}{\Rightarrow} \mathbf{P}$), 如果对于一切柱集 $A, \mathbf{P}(\partial A) = 0$, 当 $n \to \infty$ 时 $\mathbf{P}_n(A) \to \mathbf{P}(A)$.

证明, 对于空间为 $\mathbb{R}^\infty$ 的情形, 有

$$(\mathbf{P}_n \stackrel{f}{\Rightarrow} \mathbf{P}) \Leftrightarrow (\mathbf{P}_n \Rightarrow \mathbf{P}).$$

问对于空间 C 和 D, 上述结果是否成立?

4. 设 F 和 G 是数轴上的分布函数, 而

$$L(F, G) = \inf\{h > 0 : F(x - h) - h \leqslant G(x) \leqslant F(x + h) + h\}$$

是 (F 和 G 的) 莱维 (P. P. Lévy) 距离. 证明, "基本收敛" 与按由距离 $L(\cdot, \cdot)$ 决定的 "莱维度量的收敛" 等价:

$$(F_n \Rightarrow F) \Leftrightarrow (L(F_n, F) \to 0).$$

5. 设 $F_n \Rightarrow F$, 且分布函数 F 连续. 证明 $F_n(x)$ 收敛于 $F(x)$, 与当 $n \to \infty$ 时

$$\sup_x |F_n(x) - F(x)| \to 0$$

等价.

6. 证明定理 1 之注 1 提出的命题.

7. 证明定理 1 之注 2 中提出的条件 (I*) ~ (IV*) 等价性的正确性.

8. 证明 $\mathbf{P}_n \xrightarrow{w} \mathbf{P}$, 当且仅当序列 $\{\mathbf{P}_n\}$ 的任何子列 $\{\mathbf{P}_{n'}\}$, 都包含这样的子列 $\{\mathbf{P}_{n''}\}$, 使 $\mathbf{P}_{n''} \xrightarrow{w} \mathbf{P}$.

9. 举出 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上概率测度 $P, P_n (n \geqslant 1)$ 的例满足: $P_n \xrightarrow{w} P$, 但是对于一切博雷尔集合 $B \in \mathscr{B}(\mathbb{R})$, $P_n(B)$ 可能不收敛于 $P(B)$.

10. 举一分布函数 $F = F(x), F_n = F_n(x), n \geqslant 1$ 的例, 使 $F_n \xrightarrow{w} F$, 但是当 $n \to \infty$ 时, $\sup_x |F_n(x) - F(x)|$ 不收敛于 0.

11. 概率论方面的许多教材, 把定理 2 关于 "分布函数 F_n 收敛于分布函数 F" 的命题 "(4) ⇒ (3)" 与赫利 – 布雷 (E. Helly – J.R.Bray) 的名字相联系. 因此建议证明如下命题:

(a) 赫利 – 布雷引理. 如果 $F_n \Rightarrow F$ (见定义 1), 则

$$\lim_n \int_a^b g(x) \mathrm{d}F_n(x) = \int_a^b g(x) \mathrm{d}F(x),$$

其中 a 和 b 是分布函数 $F = F(x)$ 之连续点的集合中的两个点, 而 $g = g(x)$ 是 $[a, b]$ 上的连续函数.

(b) 赫利 – 布雷定理. 如果 $F_n \Rightarrow F$, 而 $g = g(x)$ 是 $\mathbb{R}$ 上的连续有界函数, 则

$$\lim_n \int_{-\infty}^{\infty} g(x) \mathrm{d}F_n(x) = \int_{-\infty}^{\infty} g(x) \mathrm{d}F(x).$$

12. 设 $F_n \Rightarrow F$, 而对于某个 $b > 0$, 序列

$$\left(\int |x|^b \mathrm{d}F_n(x)\right)_{n \geqslant 1}$$

有界. 证明, 对于 $0 \leqslant a \leqslant b$, 有

$$\lim_n \int |x|^a \mathrm{d}F_n(x) = \int |x|^a \mathrm{d}F(x);$$

对于任意 $k=1,2,\cdots$, 以及 $k\neq 0$, 有

$$\lim_n \int x^k dF_n(x) = \int x^k dF(x).$$

13. 设 $F_n \xrightarrow{w} F$, 而 $m=\mathrm{med}(F)$, $m_n=\mathrm{med}(F_n)$ 相应为 F 和 F_n 的中位数 (第一章 §4 练习题 5). 假设对于一切 $n\geqslant 1$, 中位数 m 和 m_n 的定义唯一, 证明 $m_n\to m$.

14. 假设函数 F 唯一决定于其各阶矩:

$$a_k=\int_{-\infty}^{\infty} x^k dF(x),\quad k=1,2,\cdots$$

设分布函数序列 $\{F_n\}_{n\geqslant 1}$ 的矩满足

$$a_{n,k}=\int_{-\infty}^{\infty} x^k dF_n(x) \to a_k=\int_{-\infty}^{\infty} x^k dF(x),\quad k=1,2,\cdots$$

证明 $F_n \xrightarrow{w} F$.

15. 证明大数定律的如下的一种形式 —— 辛钦大数定律: 设 $X_1, X_2, \cdots$ 是具有有限数学期望 $\mathbf{E}X_1=m$ 的, 两两独立且同分布的随机变量, 而 $S_n=X_1+\cdots+X_n$, 那么

$$S_n/n \xrightarrow{\mathbf{P}} m.$$

§2. 概率分布族的相对紧性和稠密性

1. 关于分布函数类的列紧性问题 如果给定概率测度序列, 则在考虑它 (弱) 收敛于某个概率测度问题之前, 当然应弄清它一般是否收敛于某个测度, 或者该序列是否至少有一个收敛的子序列.

例如, 对于序列 $\{\mathbf{P}_n\}$, 其中 $\mathbf{P}_{2n}=\mathbf{P}$, $\mathbf{P}_{2n+1}=\mathbf{Q}$, 而 $\mathbf{P}$ 和 $\mathbf{Q}$ 是不同的概率测度, 那么, $\{\mathbf{P}_n\}$ 显然不收敛, 然而它有两个收敛的子序列 $\{\mathbf{P}_{2n}\}$ 和 $\{\mathbf{P}_{2n+1}\}$.

对于非常简单构造的概率测度 $\mathbf{P}_n (n\geqslant 1)$ 的序列 $\{\mathbf{P}_n\}$, 其中 $\mathbf{P}_n$ 集中在点 $\{n\}$ 上 $(\mathbf{P}_n(\{n\})=1)$, 则 $\{\mathbf{P}_n\}$ 既不收敛, 也不含任何收敛子序列. (因为, 对于任何常数 $a<b$, $\lim_n \mathbf{P}_n(a,b]=0$, 则极限测度理应恒等于 0, 然而这与当 $n\to\infty$ 时 $1=\mathbf{P}_n(\mathbb{R})\not\to 0$ 矛盾.) 有趣的是, 在这个例子中相应的分布函数序列 $\{F_n\}$:

$$F_n(x)=\begin{cases}1, & 若\ x\geqslant n,\\ 0, & 若\ x<n,\end{cases}$$

显然是收敛的: 对于任何 $x\in\mathbb{R}$,

$$F_n(x)\to G(x)\equiv 0.$$

不过极限函数 $G=G(x)$ (按第二章 §3 定义 1) 并不是分布函数.

这个例子之所以可以借鉴, 是因为它说明分布函数类不是列紧的. 这个例子还提示, 为使分布函数序列收敛于也是分布函数的函数, 需要某些预防"质量在无穷流失"的条件.

在说明这里遇到的困难的特点之后, 我们给出基本定义.

2. 分布函数类的相对列紧性和完备性 假设所考虑的一切测度定义在度量空间 $(E,\mathscr{E},\rho)$ 上.

定义 1 称概率测度族 $\mathscr{P}=\{\mathbf{P}_\alpha:\alpha\in\mathfrak{A}\}$ 为相对列紧的, 如果 $\mathscr{P}$ 中的任何测度序列包含弱收敛于某一概率测度的子序列.

需要强调, 在这一定义中假设极限测度是概率测度, 不过它有可能不属于原概率测度族 $\mathscr{P}$. (正是因为如此, 在该定义中出现了"相对"二字.)

验证给定的概率测度族的相对列紧性相当困难. 因此, 希望有进行此检验的简单而方便的准则. 下面的概念有助于实现这一目标.

定义 2 概率测度族 $\mathscr{P}=\{\mathbf{P}_\alpha:\alpha\in\mathfrak{A}\}$ 称为完备的, 如果对于每一个 $\varepsilon>0$, 存在紧统 $K\subseteq E$, 使

$$\sup_{\alpha\in\mathfrak{A}} \mathbf{P}_\alpha(E\setminus K)\leqslant\varepsilon. \tag{1}$$

定义 3 定义在 $\mathbb{R}^n (n\geqslant 1)$ 上的分布函数族 $\mathscr{F}=\{F_\alpha:\alpha\in\mathfrak{A}\}$ 称为相对列紧的 (完备的), 如果相应的概率测度族 $\mathscr{P}=\{\mathbf{P}_\alpha:\alpha\in\mathfrak{A}\}$ 是相对列紧的 (完备的), 其中 $\mathbf{P}_\alpha$ 是按 F_α 建立的测度.

3. 概率测度族列紧性的条件 下面的结果, 在整个概率测度的弱收敛问题中有重要的作用.

定理 1 (普罗霍罗夫 [Ю. В. Прохоров] 定理) *设 $\mathscr{P}=\{\mathbf{P}_\alpha:\alpha\in\mathfrak{A}\}$ 是定义在完全可分度量空间 $(E,\mathscr{E},\rho)$ 上的概率测度族, 测度族 $\mathscr{P}$ 是相对列紧的, 当且仅当它是稠密的.*

证明 我们只对 E 是数轴的情形证明定理. (这一证明亦可用于任意欧几里得空间 $\mathbb{R}^n (n\geqslant 2)$ [[5], [5]]. 然后依次证明对于 $\mathbb{R}^\infty$ 以及对于 σ- 列紧空间定理成立; 最后, 对于一般完备可分度量空间, 通过将每一种情形归结为上一种, 并证明其成立.)

必要性. 假设 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ 上的概率测度族 $\mathscr{P}=\{\mathbf{P}_\alpha:\alpha\in\mathfrak{A}\}$ 相对列紧但不稠密. 那么, 存在这样一个 $\varepsilon>0$, 使对任意紧统 $K\subseteq\mathbb{R}$, 有

$$\sup_\alpha \mathbf{P}_\alpha(\mathbb{R}\setminus K)>\varepsilon,$$

即对于任何区间 $I=(a,b)$, 有

$$\sup_\alpha \mathbf{P}_\alpha(\mathbb{R}\setminus I)>\varepsilon.$$

由此可见, 对于任何区间 $I_n=(-n,n)$, $n\geqslant 1$, 存在这样的测度 $\mathbf{P}_{\alpha_n}$, 使

$$\mathbf{P}_{\alpha_n}(\mathbb{R}\setminus I_n)>\varepsilon.$$

既然测度族 $\mathscr{P}$ 是相对列紧的, 则在序列 $\{\mathbf{P}_{\alpha_n}\}_{n\geqslant 1}$ 中存在子序列 $\{\mathbf{P}_{\alpha_{n_k}}\}$, 使 $\mathbf{P}_{\alpha_{n_k}} \xrightarrow{w} \mathbf{Q}$, 其中 $\mathbf{Q}$ 是某一概率测度.

那么, 由于 §1 中定理 1 的等价条件 (I) 和 (II), 对于任何 $n \geqslant 1$, 有

$$\overline{\lim_{k\to\infty}}\, \mathbf{P}_{\alpha_{n_k}}(\mathbb{R}\setminus I_n) \leqslant \mathbf{Q}(\mathbb{R}\setminus I_n). \tag{2}$$

但是 $\mathbf{Q}(\mathbb{R}\setminus I_n) \downarrow 0, n\to\infty$, 而 (2) 式的左侧大于 $\varepsilon > 0$. 这一矛盾说明, 事实上由相对列紧性导致稠密性.

为证明充分性, 我们需要一个 (称为赫利定理的) 一般结果: 关于广义分布函数性质的列紧性 (第二章 §3 第 2 小节). 以 $\mathscr{J} = \{G\}$ 表示满足下列性质的函数 $G = G(x)$ (广义分布函数) 的全体:

1) $G(x)$ 不减;

2) $0 \leqslant G(-\infty), G(+\infty) \leqslant 1$;

3) $G(x)$ 右连续.

显然, $\mathscr{J}$ 包含分布函数类 $\mathscr{F} = \{F\}$, 其中 $F(-\infty) = 0, F(+\infty) = 1$.

定理 2 (赫利定理) *广义分布函数类 $\mathscr{J} = \{G\}$ 是列紧的, 即对于 $\mathscr{J}$ 中的任意函数序列 $\{G_n\}$, 在 $\mathscr{J}$ 中存在函数 $G(x) \in \mathscr{J}$ 和子序列 $\{n_k\} \subseteq \{n\}$, 使对于属于函数 $G = G(x)$ 之连续点的集合 $\mathbb{C}(G)$ 中任何点 x, 有*

$$G_{n_k}(x) \to G(x), k \to \infty.$$

证明 记 $T = \{x_1, x_2, \cdots\}$ 为 $\mathbb{R}$ 上可数且处处稠密的集合. 由于序列 $\{G_n(x_1)\}$ 有界, 故存在子序列 $N_1 = \{n_1^{(1)}, n_2^{(1)}, \cdots\}$, 使当 $i\to\infty$ 时 $G_{n_i^{(1)}}(x_1)$ 收敛于某个数 g_1; 而子序列 N_1 中存在子序列 $N_2 = \{n_1^{(2)}, n_2^{(2)}, \cdots\}$, 使当 $i\to\infty$ 时 $G_{n_i^{(2)}}(x_2)$ 收敛于某个数 g_2; 依此类推.

在集合 $T \subseteq \mathbb{R}$ 上定义函数 $G_T(x)$, 设

$$G_T(x_i) = g_i, x_i \in T,$$

并考虑 "康托尔" 对角序列 $N = \{n_1^{(1)}, n_2^{(2)}, \cdots\}$. 那么, 对于任意 $x_i \in T$, 当 $m\to\infty$ 时, 有

$$G_{n_m^{(m)}}(x_i) \to G_T(x_i).$$

最后, 对于所有 $x \in \mathbb{R}$, 定义函数 $G = G(x)$, 设

$$G(x) = \inf\{G_T(y) : y \in T, y > x\}. \tag{3}$$

可以断定, $G = G(x)$ 就是所求的函数, 且对于 $G(x)$ 的一切连续点 x, 有

$$G_{n_m^{(m)}}(x) \to G(x).$$

事实上, 由于所考虑的一切函数 G_n 都是不减的, 故对于属于集合 T 且满足不等式 $x \leqslant y$ 的一切 x, y, 有

$$G_{n_m^{(m)}}(x) \leqslant G_{n_m^{(m)}}(y).$$

因此, 对于满足上面条件的 x, y, 有 $G_T(x) \leqslant G_T(y)$. 由此和定义 (3), 可见 $G = G(x)$ 是不减函数.

现在证明 $G = G(x)$ 右连续. 设 $x_k \downarrow x$ 和 $d = \lim_k G(x_k)$. 显然 $G(x) \leqslant d$, 只需证明 $G(x) = d$. 假设相反: $G(x) < d$. 那么, 由 (3) 式可见, 存在 $y \in T, x < y$, 使 $G_T(y) < d$. 对于充分大的 $k, x < x_k < y$, 即 $G(x_k) \leqslant G_T(y) < d$, 且 $\lim_k G(x_k) < d$, 而这与 $d = \lim_k G(x_k)$ 矛盾. 于是, 所构造的函数 $G(x)$ 属于 $\mathscr{J}$.

最后, 证明收敛性: 对于任意点 $x^0 \in \mathbb{C}(G)$, 有

$$G_{n_m^{(m)}}(x^0) \to G(x^0).$$

如果 $x^0 < y \in T$, 则

$$\overline{\lim_m}\, G_{n_m^{(m)}}(x^0) \leqslant \overline{\lim_m}\, G_{n_m^{(m)}}(y) = G_T(y),$$

因此

$$\overline{\lim_m}\, G_{n_m^{(m)}}(x^0) \leqslant \inf\{G_T(y) : y \in T, y > x^0\} = G(x^0). \tag{4}$$

另一方面, 设 $x^1 < y < x^0, y \in T$, 则

$$G(x^1) \leqslant G_T(y) = \lim_m G_{n_m^{(m)}}(y) = \underline{\lim_m}\, G_{n_m^{(m)}}(y) \leqslant \underline{\lim_m}\, G_{n_m^{(m)}}(x^0).$$

因而, 当 $x^1 \uparrow x^0$ 时, 得

$$G(x^0-) \leqslant \underline{\lim_m}\, G_{n_m^{(m)}}(x^0). \tag{5}$$

而如果 $G(x^0-) = G(x^0)$, 则由 (4) 和 (5) 式可见

$$G_{n_m^{(m)}}(x^0) \to G(x^0), m\to\infty. \tag{6}$$

现在完成定理 1 的证明.

定理 1 的充分性 设概率测度族 $\mathscr{P}$ 完备, 而 $\{\mathbf{P}_n\}$ 是 $\mathscr{P}$ 中的某一概率测度序列, 以 $\{F_n\}$ 表示相应的分布函数序列.

由于赫利定理, 存在 $\{F_n\}$ 的子序列 $\{F_{n_k}\} \subseteq \{F_n\}$ 和广义分布函数 $G(x) \in \mathscr{J}$, 使对于 $x \in \mathbb{C}(G)$, $F_{n_k}(x) \to G(x)$. 现在证明, 由概率测度族 $\mathscr{P}$ 完备的假设, 可见函数 $G = G(x)$ 是 "真正的" 分布函数 $(G(-\infty) = 0, G(+\infty) = 1)$.

取 $\varepsilon > 0$, 设 $I = (a, b]$ 是使

$$\sup_n \mathbf{P}_n(\mathbb{R}\setminus I) < \varepsilon,$$

即满足

$$1-\varepsilon \leqslant \mathrm{P}_n(a,b],\ n \geqslant 1$$

的区间. 选择点 $a', b' \in \mathbb{C}(G)$, 使 $a' < a, b' > b$. 那么,

$$1-\varepsilon \leqslant \mathrm{P}_{n_k}(a,b] \leqslant \mathrm{P}_{n_k}(a',b'] = F_{n_k}(b') - F_{n_k}(a') \to G(b') - G(a').$$

由此可见 $G(+\infty) - G(-\infty) = 1$, 而由于 $0 \leqslant G(-\infty) \leqslant G(+\infty) \leqslant 1$, 可见 $G(-\infty) = 0, G(+\infty) = 1$.

于是, 极限函数 $G = G(x)$ 是分布函数, 且 $F_{n_k} \Rightarrow G$. 因此, 连同 §1 的定理 2 就证明了 $\mathrm{P}_{n_k} \xrightarrow{w} \mathrm{Q}$, 其中 Q 是根据分布函数 G 建立的概率测度. □

4. 练习题

1. 对于空间 $\mathbb{R}^n (n \geqslant 2)$, 证明定理 1 和 2.

2. 设 P_α 是直线上参数为 m_α 和 $\sigma_\alpha^2, \alpha \in \mathfrak{A}$ 的高斯测度. 证明, 测度族 $\mathscr{P} = \{\mathrm{P}_\alpha; \alpha \in \mathfrak{A}\}$ 是完备的, 当且仅当存在常数 a 和 b, 使

$$|m_\alpha| \leqslant a,\ \sigma_\alpha^2 \leqslant b,\ \alpha \in \mathfrak{A}.$$

3. 举例: 定义在 $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty))$ 上的 (1) 稠密的概率测度族 $\mathscr{P} = \{\mathrm{P}_\alpha; \alpha \in \mathfrak{A}\}$, (2) 不稠密的概率测度族 $\mathscr{P} = \{\mathrm{P}_\alpha; \alpha \in \mathfrak{A}\}$.

4. 设 P 是度量空间 $(E, \mathscr{E}, \rho)$ 上的概率测度. 称测度 P 为稠密的 (对照定义 2), 如果对于任意 $\varepsilon > 0$, 存在紧统 $K \subseteq E$, 使 $P(K) \geqslant 1-\varepsilon$. 证明下面的结果 (伍拉姆 [Woodlam] 定理): 每一个完全可分度量空间上的概率测度 P 是稠密的.

5. 设 $X = \{X_\alpha; \alpha \in \mathfrak{A}\}$ 是一随机向量族, 且对于某个 $r > 0, \sup_\alpha \mathrm{E}\|X_\alpha\|^r < \infty$. 证明分布的 $P_\alpha = \mathrm{Law}(X_\alpha)$ 分布族 $\mathscr{P} = \{P_\alpha; \alpha \in \mathfrak{A}\}$ 是稠密的.

§3. 极限定理证明的特征函数法

1. 历史情况概述 在概率论中, 最早极限定理 —— 大数定律, 棣莫弗 - 拉普拉斯定理, 以及伯努利概型中的泊松定理 的证明, 基于求极限前对分布函数 F_n 的直接分析, 而 F_n 可以十分简单地由二项概率表示. (在伯努利概型中, 随机变量只有两个可能值, 故本质上可能明显地求出函数 F_n.) 然而, 对于更复杂形式的随机变量, 类似的对函数 F_n 直接分析的方法实际上无法实现.

对于服从任意分布的、独立随机变量之和的极限定理的证明, 首先是切比雪夫实现的.

切比雪夫提出的、现在众所周知的 "切比雪夫不等式", 不但可以用初等方法证明 J. 伯努利大数定律, 而且为独立随机变量之和 $S_n = \xi_1 + \cdots + \xi_n (n \geqslant 1)$ 的大数定律, 对一切 $\varepsilon > 0$, 有

$$\mathrm{P}\left\{\left|\frac{S_n}{n} - \frac{\mathrm{E}S_n}{n}\right| \geqslant \varepsilon\right\} \to 0,\ n \to \infty, \tag{1}$$

建立了相当一般的条件. (见练习题 2.)

此外, 切比雪夫创建 (而马尔可夫完善) 了所谓 "矩法", 用这种方法可以证明棣莫弗 - 拉普拉斯定理的论断:

$$\mathrm{P}\left\{\frac{S_n - \mathrm{E}S_n}{\sqrt{\mathrm{D}S_n}} \leqslant x\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{u^2}{2}} \mathrm{d}u \,(= \Phi(x)), \tag{2}$$

并且在如下意义上具有广泛的特点: 它在关于被加随机变量的性质十分一般的假设下成立. 这正是把命题 (2) 称做概率论的中心极限定理的根据.

稍后, 李亚普诺夫提出了另一种证明中心极限定理的方法, (源于拉普拉斯的) 这种方法以概率分布的 "特征函数" 的思想为基础. 后来的发展说明, 李亚普诺夫的 "特征函数方法", 在证明最为多种多样的极限定理时, 是十分有效的, 这决定了它的发展和广泛的应用.

这一方法的实质如下.

2. 分布函数与特征函数对应的连续性 我们已经知道 (第二章 §12), 在分布函数与特征函数之间, 存在一一对应关系. 因此, 可以通过特征函数来研究分布函数的性质. 最出色的事实是, 分布函数的弱收敛 $F_n \xrightarrow{w} F$, 与相应特征函数列逐点收敛 $\varphi_n \to \varphi$, 二者等价. 并且, 下面的定理提供了 "关于数轴上分布的弱收敛定理" 证明的基本工具.

定理 1 (连续性定理) *设 $\{F_n\}$ 是分布函数序列: $F_n = F_n(x), x \in \mathbb{R}$, 而 $\{\varphi_n\}$ 是相应的特征函数序列:*

$$\varphi_n(t) = \int_{-\infty}^{\infty} e^{itx} \mathrm{d}F_n(x),\ t \in \mathbb{R}.$$

1) 如果 $F_n \xrightarrow{w} F$, 其中 $F = F(x)$ 是某一分布函数, 则 $\varphi_n(t) \to \varphi(t)$, 其中 $\varphi(t)$ 是 $F = F(x)$ 的特征函数.

2) 如果对于每个 $t \in \mathbb{R}$ 存在极限 $\lim_n \varphi_n(t)$, 而函数 $\varphi(t) = \lim_n \varphi_n(t)$ 在 $t = 0$ 连续, 则 $\varphi(t)$ 是某一概率分布 $F = F(x)$ 的特征函数, 且

$$F_n \xrightarrow{w} F.$$

证明 将弱收敛的定义分别用于函数 $\mathrm{Re}\, e^{itx}$ 和 $\mathrm{Im}\, e^{itx}$, 立即可以证明命题 1).

命题 2) 的证明, 要求事先证明几个引理.

引理 1 *设 $\{\mathrm{P}_n\}$ 是稠密概率测度族. 假设序列 $\{\mathrm{P}_n\}$ 的弱收敛子序列 $\{\mathrm{P}_{n'}\}$, 都收敛于同一概率测度 P. 那么, 整个序列 $\{\mathrm{P}_n\}$ 也弱收敛于同一概率测度 P.*

证明 假设结果相反: $\mathrm{P}_n \not\xrightarrow{w} \mathrm{P}$. 那么, 存在这样的有界连续函数 $f = f(x)$, 使

$$\int_{\mathbb{R}} f(x)\mathrm{P}_n(\mathrm{d}x) \nrightarrow \int_{\mathbb{R}} f(x)\mathrm{P}(\mathrm{d}x).$$

由此可见, 存在 $\varepsilon > 0$ 和无限数列 $\{n'\} \subseteq \{n\}$, 使

$$\left|\int_{\mathbb{R}} f(x)\mathbf{P}_{n'}(dx) - \int_{\mathbb{R}} f(x)\mathbf{P}(dx)\right| \geqslant \varepsilon > 0. \tag{3}$$

根据普罗霍罗夫定理 (§2), 由序列 $\{\mathbf{P}_{n'}\}$ 可以选出子序列 $\{\mathbf{P}_{n''}\}$, 使 $\mathbf{P}_{n''} \xrightarrow{w} \mathbf{Q}$, 其中 $\mathbf{Q}$ 是某一概率测度.

根据引理假设 $\mathbf{P} = \mathbf{Q}$, 因而应有

$$\int_{\mathbb{R}} f(x)\mathbf{P}_{n''}(dx) \to \int_{\mathbb{R}} f(x)\mathbf{P}(dx),$$

而这与 (3) 式矛盾, 从而引理得证. □

引理 2 *设 $\{\mathbf{P}_n\}$ 是 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 稠密概率测度族. 序列 $\{\mathbf{P}_n\}$ 弱收敛于某概率测度, 当且仅当于每个 $t \in \mathbb{R}$ 存在极限 $\lim_n \varphi_n(t)$, 而函数 $\varphi_n(t)$ 测度 $\mathbf{P}_n$ 的特征函数:*

$$\varphi_n(t) = \int_{\mathbb{R}} e^{itx}\mathbf{P}_n(dx).$$

证明 如果概率测度族 $\{\mathbf{P}_n\}$ 稠密, 则根据普罗霍罗夫定理, 存在数列 $\{\mathbf{P}_{n'}\}$ 和概率测度 $\mathbf{P}$, 使 $\mathbf{P}_{n'} \xrightarrow{w} \mathbf{P}$. 假设结果相反: 整个序列不收敛于 $\mathbf{P}$ ($\mathbf{P}_n \not\xrightarrow{w} \mathbf{P}$). 那么, 根据引理 1, 存在子序列 $\{\mathbf{P}_{n''}\}$ 和概率测度 $\mathbf{Q}$, 使 $\mathbf{P}_{n''} \xrightarrow{w} \mathbf{Q}$, 并且 $\mathbf{P} \neq \mathbf{Q}$.

现在利用: 对于每个 $t \in \mathbb{R}$ 存在极限 $\lim_n \varphi_n(t)$. 那么,

$$\lim_{n'} \int_{\mathbb{R}} e^{itx}\mathbf{P}_{n'}(dx) = \lim_{n''} \int_{\mathbb{R}} e^{itx}\mathbf{P}_{n''}(dx)$$

因而

$$\int_{\mathbb{R}} e^{itx}\mathbf{P}(dx) = \int_{\mathbb{R}} e^{itx}\mathbf{Q}(dx), \quad t \in \mathbb{R}.$$

由于特征函数唯一决定分布 (第二章 §12 定理 2), 故 $\mathbf{P} = \mathbf{Q}$ 与假设 $\mathbf{P}_n \not\xrightarrow{w} \mathbf{P}$ 矛盾.

最后, 由弱收敛的定义, 可见引理相反的结论成立. □

下一引理, 根据特征函数在 0 的邻域内的性质, 给出分布函数 "尾部" 的估计.

引理 3 *设 $F = F(x)$ 是数轴上的分布函数, 而 $\varphi = \varphi(t)$ 是其特征函数. 那么, 存在常数 $K > 0$, 使对于任何 $a > 0$, 有*

$$\int_{|x| \geqslant 1/a} dF(x) \leqslant \frac{K}{a}\int_0^a [1 - \mathrm{Re}\,\varphi(t)]dt. \tag{4}$$

证明 由于

$$\mathrm{Re}\,\varphi(t) = \int_{-\infty}^{\infty} \cos tx\, dF(x),$$

则傅比尼定理, 可见

$$\begin{aligned}\frac{1}{a}\int_0^a [1 - \mathrm{Re}\,\varphi(t)]dt &= \frac{1}{a}\int_0^a \left[\int_{-\infty}^{\infty}(1 - \cos tx)dF(x)\right]dt \\ &= \int_{-\infty}^{\infty}\left[\frac{1}{a}\int_0^a (1 - \cos tx)dt\right]dF(x) = \int_{-\infty}^{\infty}\left(1 - \frac{\sin ax}{ax}\right)dF(x) \\ &\geqslant \inf_{|y| \geqslant 1}\left(1 - \frac{\sin y}{y}\right) \cdot \int_{|ax| \geqslant 1} dF(x) = \frac{1}{K}\int_{|x| \geqslant 1/a} dF(x),\end{aligned}$$

其中

$$\frac{1}{K} = \inf_{|y| \geqslant 1}\left(1 - \frac{\sin y}{y}\right) = 1 - \sin 1 \geqslant \frac{1}{7},$$

于是, 当常数 $K = 7$ 时 (4) 式肯定成立. □

定理 1 的命题 2) 的证明. 设 $\varphi_n(t) \to \varphi(t)$, $n \to \infty$, 其中 $\varphi(t)$ 在 0 连续. 现在证明, 由此可见, $\{\mathbf{P}_n\}$ 是稠密概率测度族, 其中 $\mathbf{P}_n$ 是对应于分布函数 F_n 的测度.

由于 (4) 式和控制收敛定理, 当 $n \to \infty$ 时, 有

$$\begin{aligned}\mathbf{P}_n\left\{\mathbb{R} \setminus \left(-\frac{1}{a}, \frac{1}{a}\right)\right\} &= \int_{|x| \geqslant \frac{1}{a}} dF_n(x) \leqslant \frac{K}{a}\int_0^a [1 - \mathrm{Re}\,\varphi_n(t)]dt \\ &\to \frac{K}{a}\int_0^a [1 - \mathrm{Re}\,\varphi(t)]dt.\end{aligned}$$

由于根据定义 $\varphi(t)$ 在 0 连续且 $\varphi(0) = 1$, 可见对于任何 $\varepsilon > 0$, 存在 $a > 0$, 使对于一切 $n \geqslant 1$, 有

$$\mathbf{P}_n\left\{\mathbb{R} \setminus \left(-\frac{1}{a}, \frac{1}{a}\right)\right\} \leqslant \varepsilon.$$

从而测度族 $\{\mathbf{P}_n\}$ 稠密, 而引理 2 知存在概率测度 $\mathbf{P}$, 使

$$\mathbf{P}_n \xrightarrow{w} \mathbf{P}.$$

因此

$$\varphi_n(t) = \int_{-\infty}^{\infty} e^{itx}\mathbf{P}_n(dx) \to \int_{-\infty}^{\infty} e^{itx}\mathbf{P}(dx),$$

即 $\varphi_n(t) \to \varphi(t)$. 于是, $\varphi(t)$ 是概率测度 $\mathbf{P}$ 的特征函数. □

系 *设 $\{F_n\}$ 是分布函数序列, 而 $\{\varphi_n\}$ 是相应的特征函数序列. 此外, 设 F 是分布函数, 而 φ 是其特征函数. 那么, $F_n \xrightarrow{w} F$, 当且仅当对于一切 $t \in \mathbb{R}$, $\varphi_n(t) \to \varphi(t)$, $n \to \infty$.*

注 设 $\eta, \eta_1, \eta_2, \cdots$ 是随机变量, 且 $F_{\eta_n} \xrightarrow{w} F_\eta$. 那么, 根据第二章 §10 定义 4, 称随机变量 $\eta_1, \eta_2, \cdots$ 依分布收敛于 η, 并记作 $\eta_n \xrightarrow{d} \eta$. 这一记号很直观 ($d$ 是 distribution 的字头), 因此在表述极限定理时, 常认为表达式 $\eta_n \xrightarrow{d} \eta$ 比 $F_{\eta_n} \xrightarrow{w} F_\eta$ 更好.

3. 极限定理证明的特征函数方法 在下一节, 定理 1 将用于不同分布的、独立随机变量中心极限定理的证明. 证明将在所谓林德伯格 (J. W. Linderberg) 条件下进行. 然后证明, 李亚普诺夫条件可以保障林德伯格条件成立. 现在, 我们用特征函数法证明几个简单的极限定理.

定理 2 (辛钦大数定律) 设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量序列, 且 $\mathbf{E}|\xi_1| < \infty, \mathbf{E}\xi_1 = m, S_n = \xi_1 + \cdots + \xi_n$. 则 $S_n/n \xrightarrow{\mathbf{P}} m$, 即对于任何 $\varepsilon > 0$, 有

$$\mathbf{P}\left\{\left|\frac{S_n}{n} - m\right| \geqslant \varepsilon\right\} \to 0, n \to \infty.$$

证明 设 $\varphi(t) = \mathbf{E}e^{it\xi_1}$ 和 $\varphi_{S_n/n}(t) = \mathbf{E}e^{it\frac{S_n}{n}}$. 则由随机变量的独立性和第二章 §12 的公式 (6), 有

$$\varphi_{\frac{S_n}{n}}(t) = \left[\varphi\left(\frac{t}{n}\right)\right]^n.$$

而根据第二章 §12 的 (14) 式, 有

$$\varphi(t) = 1 + itm + o(t), t \to 0.$$

因此, 对于任何给定的 $t \in \mathbb{R}$, 有

$$\varphi\left(\frac{t}{n}\right) = 1 + i\frac{t}{n}m + o\left(\frac{1}{n}\right), n \to \infty,$$

从而,

$$\varphi_{\frac{S_n}{n}}(t) = \left[1 + i\frac{t}{n}m + o\left(\frac{1}{n}\right)\right]^n \to e^{itm}.$$

函数 $\varphi(t) = e^{itm}$ 在 0 连续, 并且是集中在点 m 的退化概率分布的特征函数. 于是

$$\frac{S_n}{n} \xrightarrow{d} m,$$

即 (见第二章 §10 的练习题 7)

$$\frac{S_n}{n} \xrightarrow{\mathbf{P}} m. \qquad \square$$

定理 3 (独立同分布随机变量的中心极限定理) 设 $\xi_1, \xi_2, \cdots$ 是独立同分布 (非退化) 随机变量序列, 且 $\mathbf{E}\xi_1^2 < \infty, S_n = \xi_1 + \cdots + \xi_n$. 则当 $n \to \infty$ 时, 有

$$\mathbf{P}\left\{\frac{S_n - \mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}} \leqslant x\right\} \to \Phi(x), x \in \mathbb{R}, \tag{5}$$

其中

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{u^2}{2}}du.$$

证明 设 $\mathbf{E}\xi_1 = m, \mathbf{D}\xi_1 = \sigma^2$, 而

$$\varphi(t) = \mathbf{E}e^{it(\xi_1 - m)}.$$

那么, 如果记

$$\varphi_n(t) = \mathbf{E}e^{it\frac{S_n - \mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}}},$$

则

$$\varphi_n(t) = \left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

但是, 根据第二章 §12 的 (14) 式, 有

$$\varphi(t) = 1 - \frac{\sigma^2 t^2}{2} + o(t^2), t \to 0.$$

因此, 对于任意固定的 t 和 $n \to \infty$, 有

$$\varphi_n(t) = \left[1 - \frac{\sigma^2 t^2}{2n\sigma^2} + o\left(\frac{1}{n}\right)\right]^n \to e^{-\frac{t^2}{2}}.$$

函数 $e^{-t^2/2}$ 是均值为 0 而方差为 1 的正态分布 (记作 $\mathcal{N}(0,1)$) 的特征函数, 故由定理 1 可见 (5) 式成立. 根据定理 1 的注, 可以将该结果写成

$$\frac{S_n - \mathbf{E}S_n}{\sqrt{\mathbf{D}S_n}} \xrightarrow{d} \mathcal{N}(0,1). \tag{6}$$

$\square$

上面两个定理, 涉及 "独立同分布 (规范与中心化) 随机变量之和 $S_n = \xi_1 + \cdots + \xi_n$" 的概率的渐近性质. 然而, 为表述泊松定理 (第一章 §6), 不得不考虑更一般的模型, 即所谓随机变量的系列模型.

具体地说, 假设对于每个 $n \geqslant 1$, 给定随机变量 $\xi_{n1}, \cdots, \xi_{nn}$. 换句话说, 给定随机变量的 (下) 三角矩阵:

$$\begin{pmatrix} \xi_{11} & & \\ \xi_{21} & \xi_{22} & \\ \xi_{31} & \xi_{32} & \xi_{33} \\ \cdots & \cdots & \cdots \end{pmatrix},$$

其中每一行 (第一行除外) 的随机变量相互独立. 设 $S_n = \xi_{n1} + \cdots + \xi_{nn}$.

定理 4 (泊松定理) 设对于每个 $n \geqslant 1$, 独立同分布随机变量 $\xi_{n1}, \cdots, \xi_{nn}$, 有

$$\mathbf{P}\{\xi_{nk} = 1\} = p_{nk}, \mathbf{P}\{\xi_{nk} = 0\} = q_{nk}, 1 \leqslant k \leqslant n,$$

$$p_{nk} + q_{nk} = 1, \max_{1 \leqslant k \leqslant n} p_{nk} \to 0, p_{n1} + \cdots + p_{nn} \to \lambda > 0, n \to \infty.$$

那么,

$$\mathbf{P}\{S_n = m\} \to \frac{\lambda^m}{m!}e^{-\lambda}, m = 0, 1, \cdots \tag{7}$$

证明 因为对于 $1 \leqslant k \leqslant n$,

$$\mathrm{E}e^{it\xi_{nk}} = p_{nk}e^{it} + q_{nk},$$

所以, 当 $n \to \infty$ 时, 有

$$\varphi_{S_n}(t) = \mathrm{E}e^{itS_n} = \prod_{k=1}^{n}(p_{nk}e^{it} + q_{nk})$$

$$= \prod_{k=1}^{n}[1 + p_{nk}(e^{it} - 1)] \to \exp\{\lambda(e^{it} - 1)\}.$$

函数 $\varphi(t) = \exp\{\lambda(e^{it} - 1)\}$ 是泊松分布的特征函数 (第二章 §12 第 2 小节例 3). 从而 (7) 式得证. 如果以 $\pi(\lambda)$ 表示参数为 λ 的泊松随机变量, 则仿照 (6) 式可以将 (7) 式写成

$$S_n \xrightarrow{d} \pi(\lambda). \qquad \square$$

4. 练习题

1. 对于空间 $R^n (n \geqslant 2)$ 的情形, 证明定理 1 的命题.

2. 设 $\xi_1, \xi_2, \cdots$ 是独立随机变量序列, 并且具有有限均值 $\mathrm{E}|\xi_n|$ 和方差 $\mathrm{D}\xi_n$, 而 $\mathrm{D}\xi_n \leqslant K < \infty$, 其中 K 是某一常数. 利用切比雪夫不等式证明大数定律 (1).

3. 在定理 1 的系中, 证明函数族 $\{\varphi_n\}$ 一致连续, 并且在每一个有限区间上有一致收敛 $\varphi_n \to \varphi$.

4. 设 $\varphi_{\xi_n}(t), n \geqslant 1$ 是随机变量 $\xi_n (n \geqslant 1)$ 的特征函数. 证明 $\xi_n \xrightarrow{d} 0$, 当且仅当在点 $t = 0$ 的某邻域内 $\varphi_{\xi_n}(t) \to 1, n \to \infty$.

5. 设 $X_1, X_2, \cdots$ 是 (取值于 R^k 的) 独立同分布随机向量序列, 且具有 0 均值和 (有限) 协方差矩阵 Γ. 证明

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \xrightarrow{d} N(0, \Gamma).$$

(对照定理 3.)

6. 设 $\xi_1, \xi_2, \cdots$ 和 $\eta_1, \eta_2, \cdots$ 是两个随机变量序列, 且对于每一个 n, ξ_n 与 η_n 独立. 假设当 $n \to \infty$ 时 $\xi_n \xrightarrow{d} \xi$, $\eta_n \xrightarrow{d} \eta$, 其中 ξ 与 η 独立. 证明二维随机变量序列 (ξ_n, η_n) 按分布收敛于 (ξ, η).

设 $f = f(x, y)$ 是连续函数, 证明序列 $f = f(\xi_n, \eta_n)$ 按分布收敛于 $f = f(\xi, \eta)$.

7. 举例说明, 定理 1 之命题 2 中 "极限" 特征函数 $\varphi(t) = \lim_n \varphi_n(t)$, 在 0 的连续性条件一般不能减弱. (换句话说, 假设分布 F 的特征函数 $\varphi(t)$ 在 0 不连续, 则可能出现这样的情况: $\varphi_n(t) \to \varphi(t)$, 然而 $F_n \xrightarrow{w} F$.) 举例说明: 缺少 $\varphi(t)$ 在 0 连续的条件, 可能导致破坏概率分布族 $P_n (n \geqslant 1)$ 的稠密性, 其中 $\varphi_n(t)$ 是 P_n 的特征函数.

§4. 独立随机变量之和的中心极限定理 I. 林德伯格条件

1. **林德伯格条件** 在这一节, 将在满足经典的林德伯格条件的传统假设下, 证明独立随机变量 $\xi_1, \xi_2, \cdots, \xi_n (n \geqslant 1)$ 之 (规范与中心化的) 和 S_n 的中心极限定理. 下一节将考虑更一般的情形: 第一, 中心极限定理直接为 "系列形式"; 第二, 其证明在满足所谓非经典条件下进行.

定理 1 设 $\xi_1, \xi_2, \cdots$ 是具有有限二阶矩的独立随机变量序列. 假设

$$m_k = \mathrm{E}\xi_k, \sigma_k^2 = \mathrm{D}\xi_k > 0, S_n = \xi_1 + \cdots + \xi_n, D_n^2 = \sum_{k=1}^{n}\sigma_k^2,$$

而 $F_k = F_k(x)$ 是随机变量 ξ_k 的分布函数.

假设满足 "林德伯格条件": 对于任意 $\varepsilon > 0$,

$$(\mathrm{L}) \qquad \frac{1}{D_n^2}\sum_{k=1}^{n}\int_{\{x:|x-m_k|\geqslant\varepsilon D_n\}}(x - m_k)^2\mathrm{d}F_k(x) \to 0, n \to \infty. \tag{1}$$

那么,

$$\frac{S_n - \mathrm{E}S_n}{\sqrt{\mathrm{D}S_n}} \xrightarrow{d} N(0, 1). \tag{2}$$

证明 不失普遍性可以认为 $m_k = 0 (k \geqslant 1)$. 设

$$\varphi_k(t) = \mathrm{E}e^{it\xi_k}, T_n = \frac{S_n}{\sqrt{\mathrm{D}S_n}} = \frac{S_n}{D_n}, \varphi_{S_n}(t) = \mathrm{E}e^{itS_n}, \varphi_{T_n}(t) = \mathrm{E}e^{itT_n}.$$

那么,

$$\varphi_{T_n}(t) = \mathrm{E}e^{itT_n} = \mathrm{E}e^{i\frac{t}{D_n}S_n} = \varphi_{S_n}\left(\frac{t}{D_n}\right) = \prod_{k=1}^{n}\varphi_k\left(\frac{t}{D_n}\right), \tag{3}$$

而 (由于 §3 定理 1) 为证明 (2) 式, 只需对于任意 $t \in \mathbb{R}$, 证明

$$\varphi_{T_n}(t) \to e^{-\frac{t^2}{2}}, n \to \infty. \tag{4}$$

取某个 $t \in \mathbb{R}$, 并认为在整个证明过程中 t 是固定的. 由于分解

$$e^{iy} = 1 + iy + \frac{\theta_1 y^2}{2},$$

$$e^{iy} = 1 + iy - \frac{y^2}{2} + \frac{\theta_2|y|^3}{3!},$$

对于每一个实数 y, 其中 $|\theta_1| \leqslant 1, |\theta_2| \leqslant 1 (\theta_1 = \theta_1(y), \theta_2 = \theta_2(y))$, 有

$$\varphi_k(t) = \mathrm{E}e^{it\xi_k} = \int_{-\infty}^{\infty}e^{itx}\mathrm{d}F_k(x)$$

$$= \int_{|x|\geqslant\varepsilon D_n}\left(1 + itx + \frac{\theta_1(tx)^2}{2}\right)\mathrm{d}F_k(x) + \int_{|x|<\varepsilon D_n}\left(1 + itx - \frac{t^2x^2}{2} + \frac{\theta_2|tx|^3}{6}\right)\mathrm{d}F_k(x)$$

$$= 1 + \frac{t^2}{2}\int_{|x|\geqslant\varepsilon D_n}\theta_1 x^2\mathrm{d}F_k(x) - \frac{t^2}{2}\int_{|x|<\varepsilon D_n}x^2\mathrm{d}F_k(x) + \frac{|t|^3}{6}\int_{|x|<\varepsilon D_n}\theta_2|x|^3\mathrm{d}F_k(x).$$

其中用到: 根据假设

$$m_k = \int_{-\infty}^{\infty} x\,dF_k(x) = 0.$$

从而

$$\varphi_k\left(\frac{t}{D_n}\right) = 1 - \frac{t^2}{2D_n^2}\int_{|x|<\varepsilon D_n} x^2 dF_k(x) + \frac{t^2}{2D_n^2}\int_{|x|\geqslant\varepsilon D_n} \theta_1 x^2 dF_k(x) + \frac{t^3}{6D_n^3}\int_{|x|<\varepsilon D_n} \theta_2 |x|^3 dF_k(x). \tag{5}$$

由于

$$\left|\frac{1}{2}\int_{|x|\geqslant\varepsilon D_n} \theta_1 x^2 dF_k(x)\right| \leqslant \frac{1}{2}\int_{|x|\geqslant\varepsilon D_n} x^2 dF_k(x),$$

有

$$\frac{1}{2}\int_{|x|\geqslant\varepsilon D_n} \theta_1 x^2 dF_k(x) = \tilde{\theta}_1 \int_{|x|\geqslant\varepsilon D_n} x^2 dF_k(x), \tag{6}$$

其中 $\tilde{\theta}_1 = \tilde{\theta}_1(t,k,n), |\tilde{\theta}_1| \leqslant 1/2$.

同样,

$$\left|\frac{1}{6}\int_{|x|<\varepsilon D_n} \theta_2 |x|^3 dF_k(x)\right| \leqslant \frac{1}{6}\int_{|x|<\varepsilon D_n} \frac{\varepsilon D_n}{|x|}|x|^3 dF_k(x) \leqslant \frac{1}{6}\int_{|x|<\varepsilon D_n} \varepsilon D_n x^2 dF_k(x),$$

即

$$\frac{1}{6}\int_{|x|<\varepsilon D_n} \theta_2 |x|^3 dF_k(x) = \tilde{\theta}_2 \int_{|x|<\varepsilon D_n} \varepsilon D_n x^2 dF_k(x), \tag{7}$$

其中 $\tilde{\theta}_2 = \tilde{\theta}_2(t,k,n), |\tilde{\theta}_2| \leqslant 1/6$.

现在, 设

$$A_{kn} = \frac{1}{D_n^2}\int_{|x|<\varepsilon D_n} x^2 dF_k(x), B_{kn} = \frac{1}{D_n^2}\int_{|x|\geqslant\varepsilon D_n} x^2 dF_k(x).$$

那么, 由 (5) ~ (7) 式, 可见

$$\varphi_k\left(\frac{t}{D_n}\right) - 1 = -\frac{t^2 A_{kn}}{2} + t^2\tilde{\theta}_1 B_{kn} + |t|^3\varepsilon\tilde{\theta}_2 A_{kn} (= C_{kn}). \tag{8}$$

注意到,

$$\sum_{k=1}^{n}(A_{kn} + B_{kn}) = 1, \tag{9}$$

并根据条件 (1), 有

$$\sum_{k=1}^{n} B_{kn} \to 0, n \to \infty. \tag{10}$$

因此, 对于充分大的 n, 有

$$\max_{1\leqslant k\leqslant n} |C_{kn}| \leqslant t^2\varepsilon^2 + \varepsilon|t|^3 \tag{11}$$

和

$$\sum_{k=1}^{n} |C_{kn}| \leqslant t^2 + \varepsilon|t|^3. \tag{12}$$

现在利用如下事实: 对于复数 z, 若 $|z| \leqslant 1/2$, 则

$$\ln(1+z) = z + \theta|z|^2,$$

其中 $\theta = \theta(z), |\theta| \leqslant 1$, 而 ln 表示对数的主值 ($\ln z = \ln|z| + i\arg z, -\pi < \arg z \leqslant \pi$). 那么, 对于充分大的 n, 由 (8) 式和 (11) 式, 可见对于充分小的 $\varepsilon > 0$,

$$\ln\varphi_k\left(\frac{t}{D_n}\right) = \ln(1 + C_{kn}) = C_{kn} + \theta_{kn}|C_{kn}|^2,$$

其中 $|\theta_{kn}| \leqslant 1$. 因而, 由 (3) 式可见

$$\frac{t^2}{2} + \ln\varphi_{T_n}(t) = \frac{t^2}{2} + \sum_{k=1}^{n}\ln\varphi_k\left(\frac{t}{D_n}\right) = \frac{t^2}{2} + \sum_{k=1}^{n} C_{kn} + \sum_{k=1}^{n}\theta_{kn}|C_{kn}|^2.$$

由于

$$\frac{t^2}{2} + \sum_{k=1}^{n} C_{kn} = \frac{t^2}{2}\left(1 - \sum_{k=1}^{n} A_{kn}\right) + t^2\sum_{k=1}^{n}\tilde{\theta}_1(t,k,n)B_{kn} + \varepsilon|t|^3\sum_{k=1}^{n}\tilde{\theta}_2(t,k,n)A_{kn},$$

而由 (9),(10) 两式, 对于任意 $\delta > 0$, 存在充分大的 n_0 和 $\varepsilon > 0$, 使对一切 $n \geqslant n_0$, 有

$$\left|\frac{t^2}{2} + \sum_{k=1}^{n} C_{kn}\right| \leqslant \frac{\delta}{2}.$$

此外, 由 (11),(12) 两式, 可见

$$\left|\sum_{k=1}^{n}\theta_{kn}|C_{kn}|^2\right| \leqslant \max_{1\leqslant k\leqslant n}|C_{kn}| \times \sum_{k=1}^{n}|C_{kn}| \leqslant (t^2\varepsilon^2 + \varepsilon|t|^3)(t^2 + \varepsilon|t|^3).$$

因此, 对于充分大的 n 和适当选择的 $\varepsilon > 0$, 可以使

$$\left|\sum_{k=1}^{n}\theta_{kn}|C_{kn}|^2\right| \leqslant \frac{\delta}{2}.$$

从而

$$\left|\frac{t^2}{2} + \ln\varphi_{T_n}(t)\right| \leqslant \delta.$$

这样, 对于任何实数 t, 有

$$\varphi_{T_n}(t)e^{\frac{t^2}{2}} \to 1, n \to \infty.$$

于是,

$$\varphi_{T_n}(t) \to e^{-\frac{t^2}{2}}, n \to \infty. \qquad \square$$

2. 林德伯格条件的某些特殊形式　考虑满足林德伯格条件 (1) 的某些特殊情形, 从而也就是中心极限定理成立的情形.

a) 李亚普诺夫条件. 对于某个 $\delta > 0$,

$$\frac{1}{D_n^{2+\delta}} \sum_{k=1}^{n} \mathbf{E}|\xi_k - m_k|^{2+\delta} \to 0, n \to \infty. \tag{13}$$

设 $\varepsilon > 0$, 那么

$$\mathbf{E}|\xi_k - m_k|^{2+\delta} = \int_{-\infty}^{\infty} |x - m_k|^{2+\delta} \mathrm{d}F_k(x)$$

$$\geqslant \int_{\{x: |x-m_k| \geqslant \varepsilon D_n\}} |x - m_k|^{2+\delta} \mathrm{d}F_k(x) \geqslant \varepsilon^{\delta} D_n^{\delta} \int_{\{x: |x-m_k| \geqslant \varepsilon D_n\}} (x - m_k)^2 \mathrm{d}F_k(x),$$

因此, 当 n 充分大时, 有

$$\frac{1}{D_n^2} \sum_{k=1}^{n} \int_{\{x: |x-m_k| \geqslant \varepsilon D_n\}} (x - m_k)^2 \mathrm{d}F_k(x) \leqslant \frac{1}{\varepsilon^{\delta}} \times \frac{1}{D_n^{2+\delta}} \sum_{k=1}^{n} \mathbf{E}|\xi_k - m_k|^{2+\delta}.$$

从而, 李亚普诺夫条件可以保證林德伯格条件成立.

b) 独立同分布情形. 假设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量, 且存在 $m = \mathbf{E}\xi_1$ 和方差 $0 < \sigma^2 \equiv \mathbf{D}\xi_1 < \infty$, 那么

$$\frac{1}{D_n^2} \sum_{k=1}^{n} \int_{\{x: |x-m| \geqslant \varepsilon D_n\}} (x - m)^2 \mathrm{d}F_k(x) = \frac{n}{n\sigma^2} \int_{\{x: |x-m| \geqslant \varepsilon \sigma^2 \sqrt{n}\}} (x - m)^2 \mathrm{d}F_1(x) \to 0,$$

因为当 $n \to \infty$ 时, $\{x : |x - m| \geqslant \varepsilon\sigma^2\sqrt{n}\} \downarrow \varnothing$, 而 $\sigma^2 = \mathbf{E}(\xi_1 - m)^2 < \infty$.

于是, 林德伯格条件成立, 从而由 §3 的定理 3 可得已证明的定理 1.

c) 一致有界的情形. 假设 $\xi_1, \xi_2, \cdots$ 是独立随机变量, 且对于一切 $n \geqslant 1$, 有

$$|\xi_n| \leqslant K < \infty,$$

其中 K 是某一常数, 且当 $n \to \infty$ 时, $D_n \to \infty$.

那么, 根据切比雪夫不等式

$$\int_{\{x: |x-m_k| \geqslant \varepsilon D_n\}} (x - m_k)^2 \mathrm{d}F_k(x) = \mathbf{E}[(\xi_k - m_k)^2 I(|\xi_k - m_k| \geqslant \varepsilon D_n)]$$

$$\leqslant (2K)^2 \mathbf{P}\{|\xi_k - m_k| \geqslant \varepsilon D_n\} \leqslant (2K)^2 \frac{\sigma_k^2}{\varepsilon^2 D_n^2},$$

因此, 当 $n \to \infty$ 时,

$$\frac{1}{D_n^2} \sum_{k=1}^{n} \int_{\{x: |x-m_k| \geqslant \varepsilon D_n\}} (x - m_k)^2 \mathrm{d}F_k(x) \leqslant \frac{(2K)^2}{\varepsilon^2 D_n^2} \to 0.$$

于是, 林德伯格条件仍然成立, 从而中心极限定理成立.

3. 系列形式下的林德伯格条件

注 1　设

$$T_n = \frac{S_n - \mathbf{E}S_n}{D_n}, F_{T_n}(x) = \mathbf{P}\{T_n \leqslant x\}.$$

那么, 命题 2 表示当 $n \to \infty$ 时, 对一切 $x \in \mathbf{R}$,

$$F_{T_n}(x) \to \Phi(x).$$

由于 $\Phi(x)$ 是连续函数, 则这里实际上是一致收敛 (§1 练习题 5):

$$\sup_{x \in \mathbf{R}} |F_{T_n}(x) - \Phi(x)| \to 0, n \to \infty. \tag{14}$$

特别, 由此可见

$$\mathbf{P}\{S_n \leqslant x\} - \Phi\left(\frac{x - \mathbf{E}S_n}{D_n}\right) \to 0, n \to \infty.$$

这一结论可以用语言表示为: 对于充分大的 n, 随机变量 S_n 大致服从均值为 $\mathbf{E}S_n$ 和方差为 $D_n^2 = \mathbf{D}S_n$ 的正态分布.

注 2　由于根据注 1, 当 $n \to \infty$ 时关于 x 一致收敛: $F_{T_n}(x) \to \Phi(x)$, 自然提出关于 (14) 式的收敛速度的问题. 在 $\xi_1, \xi_2, \cdots$ 独立同分布且 $\mathbf{E}|\xi_1|^3 < \infty$ 的情形下, 贝里–埃森 (A. C. Berry – C. G. Esseen) 定理 (不等式) (§11) 提供了该问题的答案:

$$\sup_x |F_{T_n}(x) - \Phi(x)| \leqslant C \frac{\mathbf{E}|\xi_1 - \mathbf{E}\xi_1|^3}{\sigma^3 \sqrt{n}}, \tag{15}$$

其中 C 是普适常数, 其具体值至今未知. (在 [90] 的第五章 §4.3 用不等式估计了该常数: $1/\sqrt{2\pi} \leqslant C \leqslant 0.7655$.)

在 §11 将给出 (15) 式的证明.

注 3　现在赋予林德伯格条件略有不同 (甚至是更为紧凑) 的形式, 特别方便用于极限定理的 "系列形式".

设 $\xi_1, \xi_2, \cdots$ 是独立随机变量序列,

$$m_k = \mathbf{E}\xi_k, \sigma_k^2 = \mathbf{D}\xi_k, D_n^2 = \sum_{k=1}^{n} \sigma_k^2 > 0 (n \geqslant 1), \xi_{nk} = \frac{\xi_k - m_k}{D_n}.$$

考虑到这些记号, 林德伯格条件 (1) 现在具有如下形式:

$$(\mathrm{L}) \qquad \sum_{k=1}^{n} \mathbf{E}\{\xi_{nk}^2 I(|\xi_{nk}| \geqslant \varepsilon)\} \to 0, n \to \infty. \tag{16}$$

如果 $S_n = \xi_{n1} + \cdots + \xi_{nn}$, 则 $\mathbf{D}S_n = 1$, 而定理 1 可以赋予这样的形式: 若满足条件 (16), 则

$$S_n \xrightarrow{d} N(0, 1).$$

在这种形式下, 中心极限定理不需要假设随机变量 ξ_{nk} 具有 $(\xi_k - m_k)/D_n$ 的特殊形式. 具体地说, 用定理 1 的证明方法, 可以证明如下结果:

定理 2 设对于每一个 $n \geqslant 1$,

$$\xi_{n1}, \xi_{n2}, \cdots, \xi_{nn}$$

是独立随机变量序列, 且 $\mathrm{E}\xi_{nk} = 0, \mathrm{D}S_n = 1$, 其中 $S_n = \xi_{n1} + \cdots + \xi_{nn}$.

那么, 林德伯格条件 (16) 是收敛 $S_n \xrightarrow{d} N(0,1)$ 的充分条件.

4. 林德伯格条件的必要性 由于

$$\max_{1 \leqslant k \leqslant n} \mathrm{E}\xi_{nk}^2 \leqslant \varepsilon^2 + \sum_{k=1}^{n} \mathrm{E}[\xi_{nk}^2 I(|\xi_{nk}| \geqslant \varepsilon)],$$

则由林德伯格条件 (16), 可见

$$\max_{1 \leqslant k \leqslant n} \mathrm{E}\xi_{nk}^2 \to 0, n \to \infty. \tag{17}$$

很出色的是, 在条件 (17) 下, 由中心极限定理成立, 就可以自然得出林德伯格条件成立①.

定理 3 设对于每一个 $n \geqslant 1$,

$$\xi_{n1}, \xi_{n2}, \cdots, \xi_{nn}$$

是独立随机变量序列, 且 $\mathrm{E}\xi_{nk} = 0, \mathrm{D}S_n = 1$, 其中 $S_n = \xi_{n1} + \cdots + \xi_{nn}$. 假设满足条件 (17). 那么, 对于中心极限定理 $S_n \xrightarrow{d} N(0,1)$ 的成立, 林德伯格条件充分而且必要.

由定理 2 立即得充分性. 为证明其必要性, 需要下面的引理 (对照 §3 的引理 3).

引理 设随机变量 ξ 的分布函数和数字特征为 $F = F(x)$, 而且 $\mathrm{E}\xi = 0, \mathrm{D}\xi = \gamma > 0$. 那么, 对于每一个 $a > 0$,

$$\int_{|x| \geqslant 1/a} x^2 \mathrm{d}F(x) \leqslant \frac{1}{a^2}[\mathrm{Re} f(\sqrt{6}a) - 1 + 3\gamma a^2], \tag{18}$$

其中 $f(t) = \mathrm{E}e^{it\xi}$ 是 ξ 的特征函数.

证明 有

$$\begin{aligned}
\mathrm{Re} f(t) - 1 + \frac{1}{2}\gamma t^2 &= \frac{1}{2}\gamma t^2 - \int_{-\infty}^{\infty} (1 - \cos tx) \mathrm{d}F(x) \\
&= \frac{1}{2}\gamma t^2 - \int_{|x|<1/a} (1 - \cos tx) \mathrm{d}F(x) - \int_{|x| \geqslant 1/a} (1 - \cos tx) \mathrm{d}F(x) \\
&\geqslant \frac{1}{2}\gamma t^2 - \frac{1}{2}t^2 \int_{|x|<1/a} x^2 \mathrm{d}F(x) - 2a^2 \int_{|x| \geqslant 1/a} x^2 \mathrm{d}F(x) \\
&= \left(\frac{1}{2}t^2 - 2a^2\right) \int_{|x| \geqslant 1/a} x^2 \mathrm{d}F(x).
\end{aligned}$$

① 条件 (17) 称做"均匀小条件". 在均匀小条件下, 林德伯格条件是独立随机变量序列服从中心极限定理的充分和必要条件. ——译者

设 $t = \sqrt{6}a$, 得所要证明的 (18) 式. □

证明定理 3 的必要性 设

$$\begin{gathered}
F_{nk}(x) = \mathrm{P}\{\xi_{nk} \leqslant x\}, f_{nk}(t) = \mathrm{E}e^{it\xi_{nk}}, \\
\mathrm{E}\xi_{nk} = 0, \mathrm{D}\xi_{nk} = \gamma_{nk} > 0, \\
\sum_{k=1}^{n} \gamma_{nk} = 1, \max_{1 \leqslant k \leqslant n} \gamma_{nk} \to 0, n \to \infty.
\end{gathered} \tag{19}$$

以 $\ln z$ 表示复数 z 之对数的主值 (即 $\ln z = \ln|z| + i\arg z, -\pi < \arg z \leqslant \pi$).

那么,

$$\ln \prod_{k=1}^{n} f_{nk}(t) = \sum_{k=1}^{n} \ln f_{nk}(t) + 2\pi i m,$$

其中 $m = m(n, t)$ 是某个整数. 从而,

$$\mathrm{Re} \ln \prod_{k=1}^{n} f_{nk}(t) = \mathrm{Re} \sum_{k=1}^{n} \ln f_{nk}(t). \tag{20}$$

由于

$$\prod_{k=1}^{n} f_{nk}(t) \to e^{-\frac{t^2}{2}},$$

故

$$\left|\prod_{k=1}^{n} f_{nk}(t)\right| \to e^{-\frac{t^2}{2}},$$

因而

$$\mathrm{Re} \ln \prod_{k=1}^{n} f_{nk}(t) = \mathrm{Re} \ln \left|\prod_{k=1}^{n} f_{nk}(t)\right| \to -\frac{1}{2}t^2. \tag{21}$$

对于 $|z| < 1$,

$$\ln(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \cdots, \tag{22}$$

而对于 $|z| \leqslant 1/2$,

$$|\ln(1+z) - z| \leqslant |z|^2. \tag{23}$$

由于 (19) 式, 对于每个固定的 t, 所有充分大的 n 和一切 $k = 1, 2, \cdots, n$, 有

$$|f_{nk}(t) - 1| \leqslant \frac{1}{2}\gamma_{nk} t^2 \leqslant \frac{1}{2}. \tag{24}$$

因此, 由 (23) 和 (24) 两式, 可见当 $n \to \infty$ 时

$$\begin{aligned}
&\left|\sum_{k=1}^{n} \{\ln[1 + (f_{nk}(t) - 1)] - (f_{nk}(t) - 1)\}\right| \leqslant \sum_{k=1}^{n} |f_{nk}(t) - 1|^2 \\
&\leqslant \frac{t^4}{4} \max_{1 \leqslant k \leqslant n} \gamma_{nk} \cdot \sum_{k=1}^{n} \gamma_{nk} = \frac{t^4}{4} \max_{1 \leqslant k \leqslant n} \gamma_{nk} \to 0,
\end{aligned}$$

从而

$$\left|\mathrm{Re}\sum_{k=1}^{n}\ln f_{nk}(t)-\mathrm{Re}\sum_{k=1}^{n}[f_{nk}(t)-1]\right|\to 0,\ n\to\infty. \tag{25}$$

由 (20),(21) 和 (25) 式, 可见

$$\mathrm{Re}\sum_{k=1}^{n}[f_{nk}(t)-1]+\frac{1}{2}t^2=\sum_{k=1}^{n}\left[\mathrm{Re}f_{nk}(t)-1+\frac{1}{2}t^2\sigma_{nk}^2\right]\to 0,\ n\to\infty.$$

设 $t=\sqrt{6}a$. 对于每一个 $a>0$, 有

$$\sum_{k=1}^{n}[\mathrm{Re}f_{nk}(\sqrt{6}a)-1+3a^2\sigma_{nk}^2]\to 0,\ n\to\infty. \tag{26}$$

最后, 由 (18) 式 (其中设 $a=1/\varepsilon$) 和 (26) 式, 得

$$\begin{aligned}\sum_{k=1}^{n}\mathrm{E}\,\xi_{nk}^2 I(|\xi_{nk}|\geqslant\varepsilon)&=\sum_{k=1}^{n}\int_{|x|\geqslant\varepsilon}x^2\mathrm{d}F_{nk}(x)\\&\leqslant\varepsilon^2\sum_{k=1}^{n}[\mathrm{Re}f_{nk}(\sqrt{6}a)-1+3a^2\sigma_{nk}^2]\to 0,\ n\to\infty.\end{aligned}$$

于是, 满足林德伯格条件得证. □

5. 练习题

1. 假设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量序列, 且 $\mathrm{E}\xi_1^2<\infty$. 证明, 当 $n\to\infty$ 时, 有

$$\max\left(\frac{|\xi_1|}{\sqrt{n}},\cdots,\frac{|\xi_n|}{\sqrt{n}}\right)\xrightarrow{d}0.$$

2. 对于伯努利概型, 直接证明, 当 $n\to\infty$ 时,

$$\sup_x|F_{T_n}(x)-\Phi(x)|,$$

的数量级为 $1/\sqrt{n}$.

3. 设 $X_1,X_2,\cdots$ 是独立可交换 (顺序的) 随机变量序列 (见第二章 §5 练习题 4), 且 $\mathrm{E}X_i=0,\mathrm{E}X_i^2=1$, 且

$$\mathrm{cov}(X_1,X_2)=\mathrm{cov}(X_1^2,X_2^2). \tag{27}$$

证明, 中心极限定理成立:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i\xrightarrow{d}N(0,1). \tag{28}$$

相反, 如果 $\mathrm{E}X_n^2<\infty$ 且 (28) 式成立, 则 (27) 式也成立.

4. 局部中心极限定理. 设 $X_1,X_2,\cdots$ 是独立同分布随机变量, 而且 $\mathrm{E}X_1=0,\mathrm{E}X_1^2=1$. 假设其特征函数 $\varphi(t)=\mathrm{E}e^{itX_1}$ 具有性质: 对于某个 $r\geqslant 1$, 有

$$\int_{-\infty}^{\infty}|\varphi(t)|^r\mathrm{d}t<\infty.$$

证明, 随机变量 $S_n/\sqrt{n}$ 的概率分布密度 $f_n(x)$ 存在, 且关于 $x\in\mathbb{R}$ 一致, 有

$$f_n(x)\to(2\pi)^{-1/2}e^{-\frac{x^2}{2}},\ n\to\infty.$$

问相应的结果对于格点随机变量如何?

5. 设 $X_1,X_2,\cdots$ 是独立同分布随机变量, 且 $\mathrm{E}X_1=0,\mathrm{E}X_1^2=1$. 假设 $d_1^2,d_2^2,\cdots$ 是非负常数满足: $d_n=o(D_n)$, 其中 $D_n^2=\sum_{k=1}^{n}d_k^2$. 证明加权变量 $d_1X_1,d_2X_2,\cdots$ 序列服从中心极限定理:

$$\frac{1}{D_n}\sum_{k=1}^{n}d_kX_k\xrightarrow{d}N(0,1),\quad n\to\infty.$$

6. 设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量, 且 $\mathrm{E}\xi_1=0,\mathrm{E}\xi_1^2=1$. 假设 $\{\tau_n\}_{n\geqslant 1}$ 是在集合 $\{1,2,\cdots\}$ 中取值的随机变量, 且 $\tau_n/n\xrightarrow{P}c$, 其中 $c>0$ 是常数. 证明 $(S_n=\xi_1+\cdots+\xi_n)$

$$\mathrm{Law}(\tau_n^{-1/2}S_{\tau_n})\to\Phi,$$

即 $\tau_n^{-1/2}S_{\tau_n}\xrightarrow{d}\xi\sim N(0,1)$.(注意, 这里并没有假设序列 $\{\tau_n\}_{n\geqslant 1}$ 以及 $\{\xi_n\}_{n\geqslant 1}$ 独立.)

7. 设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量, 且 $\mathrm{E}\xi_1=0,\mathrm{E}\xi_1^2=1$. 证明

$$\mathrm{Law}(n^{-1/2}\max_{1\leqslant m\leqslant n}S_m)\to\mathrm{Law}(|\xi|),$$

其中 $\xi\sim N(0,1)$. 换句话说, 对于 $x>0$, 证明

$$\mathrm{P}\left\{\frac{1}{\sqrt{n}}\max_{1\leqslant m\leqslant n}S_m\leqslant x\right\}\to\sqrt{\frac{2}{\pi}}\int_0^x e^{-\frac{y^2}{2}}\mathrm{d}y\left(=\frac{1}{\sqrt{2}}\mathrm{erf}(x)\right).$$

提示: 首先对于对称伯努利随机变量 $\xi_1,\xi_2,\cdots:\mathrm{P}\{\xi_n=\pm1\}=1/2$, 证明所提出的命题, 然后证明对于任意满足 $\mathrm{E}\xi_1=0,\mathrm{E}\xi_1^2=1$ 的独立同分布随机变量 $\xi_1,\xi_2,\cdots$, 有同样的极限分布. (这里极限分布的形状, 不依赖于 "具有 $\mathrm{E}\xi_n=0,\mathrm{E}\xi_n^2=1$ 的独立同分布随机变量 $\xi_1,\xi_2,\cdots$" 的具体选择, 并称做 "不变原理"; 对照 §7.)

8. 在上题的条件下, 证明

$$\mathrm{P}\{n^{-1/2}\max_{1\leqslant m\leqslant n}|S_m|\leqslant x\}\to H(x),\ x>0,$$

其中

$$H(x)=\frac{4}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}\exp\left\{-\frac{(2k+1)^2\pi^2}{8x^2}\right\}.$$

9. 设 $X_1, X_2, \cdots$ 是独立随机变量序列, 且

$$\mathbf{P}\{X_n = \pm n^{\alpha}\} = \frac{1}{2n^{\beta}}, \quad \mathbf{P}\{X_n = 0\} = 1 - \frac{1}{n^{\beta}},$$

其中 $2\alpha > \beta - 1$. 证明林德伯格条件成立当且仅当 $0 \leqslant \beta < 1$.

10. 设 $X_1, X_2, \cdots$ 是独立随机变量序列, $|X_n| \leqslant C_n$ ($\mathbf{P}$ - a.c.), 而 $C_n = o(D_n)$, 其中

$$D_n^2 = \sum_{k=1}^{n} \mathbf{E}(X_k - \mathbf{E}X_k)^2 \to \infty.$$

证明

$$\frac{S_n - \mathbf{E}S_n}{D_n} \to N(0,1), \text{ 其中 } S_n = X_1 + \cdots + X_n.$$

11. 设 $X_1, X_2, \cdots$ 是独立随机变量序列, 且 $\mathbf{E}X_n = 0, \mathbf{E}X_n^2 = \sigma_n^2$. 假设 $\{X_n\}_{n\geqslant 1}$ 服从中心极限定理, 而且对于某个 $k \geqslant 1$, 有

$$\mathbf{E}\left(D_n^{-1/2}\sum_{i=1}^{n} X_i\right)^{k} \to \frac{(2k)!}{2^k k!}.$$

证明 k 阶林德伯格条件成立, 即

$$\sum_{j=1}^{n}\int_{\{|x|>\varepsilon\}} |x|^{k}\mathrm{d}F_j(x) = o(D_n^{k}), \varepsilon > 0.$$

(普通的林德伯格条件, 对应于 $k = 2$ 的情形; 见 (1) 式.)

12. 设 $X = X(\lambda)$ 和 $Y = Y(\mu)$ 是参数相应为 $\lambda > 0$ 和 $\mu > 0$ 的独立泊松随机变量. 证明当 $\lambda \to \infty, \mu \to \infty$ 时

$$\frac{[X(\lambda) - \lambda] - [Y(\mu) - \mu]}{\sqrt{X(\lambda) + Y(\mu)}} \to N(0, 1).$$

13. 设对于 $n \geqslant 1$, $n+1$ 维随机向量 $(X_1^{(n)}, \cdots, X_{n+1}^{(n)})$ 在单位球面上均匀分布. 证明下面的庞加莱 (J. H. Poincaré) 定理成立:

$$\lim_{n\to\infty} \mathbf{P}\{\sqrt{n}X_{n+1}^{(n)} \leqslant x\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{u^2}{2}}\mathrm{d}u.$$

§5. 独立随机变量之和的中心极限定理 II. 非经典条件

1. 在非经典条件下中心极限定理的例 在 §4 曾经证明, 由林德伯格条件 (10) 引出条件

$$\max_{1\leqslant k\leqslant n} \mathbf{E}\xi_{nk}^2 \to 0$$

成立, 而由此可得所谓极限可忽略性条件 (亦称渐近小性条件), 指的是: 对于任意 $\varepsilon > 0$, 有

$$\max_{1\leqslant k\leqslant n} \mathbf{P}\{|\xi_{nk}| \geqslant \varepsilon\} \to 0, n \to \infty.$$

这样, 可以说, §4 中的定理 1 和 2 在极限可忽略性的条件下, 给出了独立同分布随机变量之和服从中心极限定理的条件. 在各被加项满足极限可忽略性条件的假设下, 极限定理习惯上称做 "经典提法" 的定理. 然而, 不难举出非退化随机变量, 在既不满足林德伯格条件, 也不满足极限可忽略性条件的情况下, 仍然服从中心极限定理的例子.

假设 $\xi_1, \xi_2, \cdots$ 是独立正态分布的随机变量序列, 且 $\mathbf{E}\xi_n = 0, \mathbf{D}\xi_1 = 1, \mathbf{D}\xi_k = 2^{k-2}, k \geqslant 2$. 设 $S_n = \xi_{n1} + \cdots + \xi_{nn}$, 其中

$$\xi_{nk} = \frac{\xi_k}{\sqrt{\sum_{i=1}^{n}\mathbf{D}\xi_i}}.$$

不难验证, 这里无论是林德伯格条件, 还是极限可忽略性条件都不满足. 然而, 由于 S_n 本身服从 $\mathbf{E}S_n = 0, \mathbf{D}S_n = 1$ 的正态分布, 所以中心极限定理显然成立.

在不假设 "经典的" 极限可忽略性条件下, 下面的定理 1 将要给出中心极限定理成立的充分 (和必要) 条件. 在此意义下, 下面提出的条件 (A) 正是 §5 的题目所反映的所谓 "非经典" 条件.

2. "非经典" 条件与林德伯格条件的联系 假设对于每个 $n \geqslant 1$, 给定 ("系列形式" 的) 独立随机变量序列:

$$\xi_{n1}, \xi_{n2}, \cdots, \xi_{nn},$$

且 $\mathbf{E}\xi_{nk} = 0, \mathbf{D}\xi_{nk} = \sigma_{nk}^2 > 0, \sum_{k=1}^{n}\sigma_{nk}^2 = 1$. 设 $S_n = \xi_{n1} + \cdots + \xi_{nn}, F_{nk}(x) = \mathbf{P}\{\xi_{nk} \leqslant x\}$,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{y^2}{2}}\mathrm{d}y, \quad \Phi_{nk}(x) = \Phi\left(\frac{x}{\sigma_{nk}}\right).$$

定理 1 随机变量和

$$S_n \xrightarrow{d} N(0,1) \tag{1}$$

的充分 (和必要) 条件是, 对于每个 $\varepsilon > 0$,

$$\text{(A)} \qquad \sum_{k=1}^{n}\int_{|x|>\varepsilon} |x||F_{nk}(x) - \Phi_{nk}(x)|\mathrm{d}x \to 0, n \to \infty. \tag{2}$$

下一个定理说明了条件 (A) 与经典林德伯格条件

$$\text{(L)} \qquad \sum_{k=1}^{n}\int_{|x|>\varepsilon} x^2\mathrm{d}F_{nk}(x) \to 0, n \to \infty \tag{3}$$

之间的联系.

定理 2 1. 林德伯格条件保障条件 (Λ) 成立:

$$(\mathrm{L}) \Rightarrow (\Lambda).$$

2. 如果当 $n\to\infty$ 时,

$$\max_{1\leqslant k\leqslant n} \mathbf{E}\xi_{nk}^2 \to 0,$$

则条件 (Λ) 保障林德伯格条件 (L) 成立:

$$(\Lambda) \Rightarrow (\mathrm{L}).$$

定理 1 的证明. 证明条件 (Λ) 的必要性相当复杂 ([88],[91],[96]). 我们在这里仅证明条件 (Λ) 的充分性.

设

$$f_{nk}(t)=\mathbf{E}e^{it\xi_{nk}},\quad f_n(t)=\mathbf{E}e^{itS_n},$$

$$\varphi_{nk}(t)=\int_{-\infty}^{\infty}e^{itx}\mathrm{d}\Phi_{nk}(x),\quad \varphi(t)=\int_{-\infty}^{\infty}e^{itx}\mathrm{d}\Phi(x).$$

由第二章 §12, 可见

$$\varphi_{nk}(t)=e^{-\frac{t^2\sigma_{nk}^2}{2}},\quad \varphi(t)=e^{-\frac{t^2}{2}}.$$

根据 §3 的定理 1, $S_n\xrightarrow{d}N(0,1)$ 当且仅当对于任意实数 t, $f_n(t)\to\varphi(t)$, $n\to\infty$.

有

$$f_n(t)-\varphi(t)=\prod_{k=1}^n f_{nk}(t)-\prod_{k=1}^n\varphi_{nk}(t).$$

由于 $|f_{nk}(t)|\leqslant 1$, $|\varphi_{nk}(t)|\leqslant 1$, 可见

$$|f_n(t)-\varphi(t)|=\left|\prod_{k=1}^n f_{nk}(t)-\prod_{k=1}^n\varphi_{nk}(t)\right|\leqslant\sum_{k=1}^n|f_{nk}(t)-\varphi_{nk}(t)|$$

$$=\sum_{k=1}^n\left|\int_{-\infty}^{\infty}e^{itx}\mathrm{d}(F_{nk}-\Phi_{nk})\right|=\sum_{k=1}^n\left|\int_{-\infty}^{\infty}\left(e^{itx}-itx+\frac{t^2x^2}{2}\right)\mathrm{d}(F_{nk}-\Phi_{nk})\right|,\quad(4)$$

其中用到

$$\int_{-\infty}^{\infty}x^k\mathrm{d}F_{nk}=\int_{-\infty}^{\infty}x^k\mathrm{d}\Phi_{nk}.$$

对于积分

$$\int_a^b\left(e^{itx}-itx+\frac{t^2x^2}{2}\right)\mathrm{d}(F_{nk}-\Phi_{nk}),$$

运用分部积分公式 (第二章 §6 定理 11), 得 $x^2[1-F_{nk}(x)+F_{nk}(-x)]\to 0$, $x\to\infty$, $x^2[1-\Phi_{nk}(x)+\Phi_{nk}(-x)]\to 0$, $x\to\infty$,

$$\int_{-\infty}^{\infty}\left(e^{itx}-itx+\frac{t^2x^2}{2}\right)\mathrm{d}(F_{nk}-\Phi_{nk})$$

$$=-it\int_{-\infty}^{\infty}(e^{itx}-1-itx)(F_{nk}(x)-\Phi_{nk}(x))\mathrm{d}x.\quad(5)$$

由 (4) 和 (5) 式, 得

$$|f_n(t)-\varphi(t)|\leqslant\sum_{k=1}^n\left|\int_{-\infty}^{\infty}\left(e^{itx}-itx+\frac{t^2x^2}{2}\right)\mathrm{d}(F_{nk}-\Phi_{nk})\right|$$

$$=\sum_{k=1}^n\left|t\int_{-\infty}^{\infty}(e^{itx}-1-itx)[F_{nk}(x)-\Phi_{nk}(x)]\mathrm{d}x\right|$$

$$\leqslant\frac{|t|^3}{2}\varepsilon\sum_{k=1}^n\int_{|x|\leqslant\varepsilon}|x||F_{nk}(x)-\Phi_{nk}(x)|\mathrm{d}x$$

$$+2t^2\sum_{k=1}^n\int_{|x|>\varepsilon}|x||F_{nk}(x)-\Phi_{nk}(x)|\mathrm{d}x$$

$$\leqslant\varepsilon|t|^3\sum_{k=1}^n\sigma_{nk}^2+2t^2\sum_{k=1}^n\int_{|x|>\varepsilon}|x||F_{nk}(x)-\Phi_{nk}(x)|\mathrm{d}x.\quad(6)$$

其中用到不等式

$$\int_{|x|\leqslant\varepsilon}|x||F_{nk}(x)-\Phi_{nk}(x)|\mathrm{d}x\leqslant 2\sigma_{nk}^2,\quad(7)$$

其正确性利用第二章 §6 的 (71) 式容易证明. 由于 $\varepsilon>0$ 是任意的, 由 (6) 式, 以及条件 (Λ) 可见 $f_n(t)\to\varphi(t)$, $n\to\infty$. □

定理 2 的证明. 1. 根据 §4, 由林德伯格条件 (L), 得 $\max_{1\leqslant k\leqslant n}\sigma_{nk}^2\to 0$. 因此, 注意到 $\sum_{k=1}^n\sigma_{nk}^2=1$, 得

$$\sum_{k=1}^n\int_{|x|>\varepsilon}x^2\mathrm{d}\Phi_{nk}(x)\leqslant\int_{|x|>\varepsilon/\sqrt{\max\limits_{1\leqslant k\leqslant n}\sigma_{nk}^2}}x^2\mathrm{d}\Phi(x)\to 0,\ n\to\infty.\quad(8)$$

由此式连同条件 (L), 可见对于任意 $\varepsilon>0$, 有

$$\sum_{k=1}^n\int_{|x|>\varepsilon}x^2\mathrm{d}[F_{nk}(x)-\Phi_{nk}(x)]\to 0,\ n\to\infty.\quad(9)$$

对于固定的 $\varepsilon>0$, 存在连续可微偶函数 $h=h(x)$: $|h(x)|\leqslant x^2$, $|h'(x)|\leqslant 4|x|$, 使

$$h(x)=\begin{cases}x^2, & 若\ |x|>2\varepsilon,\\ 0, & 若\ |x|\leqslant 2\varepsilon.\end{cases}$$

对于这样的函数 $h=h(x)$, 由 (9) 式, 可见

$$\sum_{k=1}^n\int_{|x|>\varepsilon}h(x)\mathrm{d}[F_{nk}(x)+\Phi_{nk}(x)]\to 0,\ n\to\infty.\quad(10)$$

利用分部积分法, 由 (10) 式, 得

$$\sum_{k=1}^n\int_{x\geqslant\varepsilon}h'(x)[1-F_{nk}(x)]+[1-\Phi_{nk}(x)]]\mathrm{d}x=\sum_{k=1}^n\int_{x\geqslant\varepsilon}h(x)\mathrm{d}[F_{nk}+\Phi_{nk}]\to 0,$$

$$\sum_{k=1}^n\int_{x\leqslant-\varepsilon}h'(x)[F_{nk}(x)+\Phi_{nk}(x)]\mathrm{d}x=\sum_{k=1}^n\int_{x\leqslant-\varepsilon}h(x)\mathrm{d}[F_{nk}+\Phi_{nk}]\to 0.$$

由于当 $|x| \geqslant 2\varepsilon$ 时 $h'(x) = 2x$, 可见

$$\sum_{k=1}^{n} \int_{|x| \geqslant 2\varepsilon} |x||F_{nk}(x) - \Phi_{nk}(x)|\mathrm{d}x \to 0, n \to \infty.$$

这样, 由于 $\varepsilon > 0$ 的任意性, 可见 (L) $\Rightarrow$ (Λ).

2. 由于 $\max_{1 \leqslant k \leqslant n} \sigma_{nk}^2 \to 0$ 和 (8) 式, 对于上面引进的函数 $h = h(x)$, 得

$$\sum_{k=1}^{n} \int_{|x| \geqslant \varepsilon} h(x)\mathrm{d}\Phi_{nk}(x) \leqslant \sum_{k=1}^{n} \int_{|x| \geqslant \varepsilon} x^2\mathrm{d}\Phi_{nk}(x) \to 0, n \to \infty. \tag{11}$$

其次, 由分部积分法, 可见

$$\begin{aligned}
&\left|\sum_{k=1}^{n} \int_{|x| \geqslant \varepsilon} h(x)\mathrm{d}(F_{nk} - \Phi_{nk})\right| \\
&\leqslant \left|\sum_{k=1}^{n} \int_{x \geqslant \varepsilon} h(x)\mathrm{d}[(1 - F_{nk}) - (1 - \Phi_{nk})]\right| + \left|\sum_{k=1}^{n} \int_{x \leqslant -\varepsilon} h(x)\mathrm{d}(F_{nk} - \Phi_{nk})\right| \\
&\leqslant \sum_{k=1}^{n} \int_{x \geqslant \varepsilon} |h'(x)|\,|(1 - F_{nk}) - (1 - \Phi_{nk})|\mathrm{d}x + \sum_{k=1}^{n} \int_{x \leqslant -\varepsilon} |h'(x)|\,|F_{nk} - \Phi_{nk}|\mathrm{d}x \\
&\leqslant 4\sum_{k=1}^{n} \int_{|x| \geqslant \varepsilon} |x||F_{nk}(x) - \Phi_{nk}(x)|\mathrm{d}x.
\end{aligned} \tag{12}$$

由 (11) 和 (12) 式, 可见

$$\sum_{k=1}^{n} \int_{|x| \geqslant 2\varepsilon} x^2\mathrm{d}F_{nk}(x) \leqslant \sum_{k=1}^{n} \int_{|x| \geqslant \varepsilon} h(x)\mathrm{d}F_{nk}(x) \to 0, n \to \infty,$$

即林德伯格条件 (L) 成立. □

3. 练习题

1. 证明 (5) 式.

2. 验证关系式 (10) 和 (12).

3. 假设 $N = (N_t)_{t \geqslant 0}$ 是在第二章 §9 的第 4 小节中引进的更新过程: $N_t = \sum_{n=1}^{\infty} I(T_n \leqslant t), T_n = \sigma_1 + \cdots + \sigma_n$, 其中 $\sigma_1, \sigma_2, \cdots$ 是独立同分布正随机变量序列. 假设 $\mu = \mathbf{E}\sigma_1 < \infty, 0 < \mathbf{D}\sigma_1 < \infty$, 证明中心极限定理成立:

$$\frac{N_t - t\mu^{-1}}{\sqrt{t\mu^{-3}\mathbf{D}\sigma_1}} \xrightarrow{d} \mathcal{N}(0, 1),$$

其中 $\mathcal{N}(0, 1)$ 表示均值为 0 而方差为 1 的标准正态分布随机变量.

§6. 无限可分分布和稳定分布

1. 在非经典条件下中心极限定理的例 在 §4 曾经指出, 为表述泊松分布, 不得不考虑系列形式: 对于每个 $n \geqslant 1$, 给定随机变量序列 $(\xi_{n,k}), 1 \leqslant k \leqslant n$.

设

$$T_n = \xi_{n,1} + \cdots + \xi_{n,n}, n \geqslant 1. \tag{1}$$

无限可分分布概念的产生, 与下面的问题相联系: 如何描述 "可以作为随机变量序列 $\{T_n\}_{n \geqslant 1}$ 的极限分布" 的某些分布?

一般讲, 像问题的如此这般的提法, 极限分布可以是任意的. 事实上, 假如 ξ 是某一随机变量, 而 $\xi_{n,1} = \xi, \xi_{n,k} = 0 (1 < k \leqslant n)$, 则 $T_n = \xi$, 从而极限分布就是 ξ 的分布, 即分布可以是任意的.

为使极限分布问题更富有内容, 我们在这一节将处处假设, 对于每个 $n \geqslant 1$, 随机变量 $\xi_{n,1}, \cdots, \xi_{n,n}$ 不仅独立而且同分布.

注意, 泊松定理 (§3 定理 4) 的情况正是这样. 独立同分布随机变量 $\xi_1, \xi_2, \cdots$ 之和 $S_n = \xi_1 + \cdots + \xi_n (n \geqslant 1)$ 的中心极限定理 (§3 定理 3) 就属于这种情形. 实际上, 如果设

$$\xi_{n,k} = \frac{\xi_k - \mathbf{E}\xi_k}{D_n}, D_n^2 = \mathbf{D}S_n,$$

则

$$T_n = \sum_{k=1}^{n} \xi_{n,k} = \frac{S_n - \mathbf{E}S_n}{D_n}.$$

这样, 正态分布和泊松分布, 可以视为系列形式的极限分布. 如果 $T_n \xrightarrow{d} T$, 则直观上容易理解, 由于 T_n 是独立同分布随机变量之和, 则极限随机变量 T 也应该在同样的意义上是独立同分布随机变量之和. 因此, 引出如下定义.

定义 1 随机变量 T (及其分布函数 F_T 和特征函数 φ_T) 称做无限可分的, 如果对于每个 $n \geqslant 1$, 在概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 上存在独立同分布随机变量 $\eta_1, \cdots, \eta_n$, 使*) $T \stackrel{d}{=} \eta_1 + \cdots + \eta_n$ (或同样地使, $F_T = F_{\eta_1} * \cdots * F_{\eta_n}$ 或 $\varphi_T = (\varphi_{\eta_1})^n$).

注 1 如果随机变量 T 定义的原概率空间足充分 "贫乏的", 可能出现下面的情况: 对于任意 $n \geqslant 1$, 对于某些分布函数 $F^{(n)}$ 及其特征函数 $\varphi^{(n)}$, 分布函数 F_T 及其特征函数 φ_T 可以表现为 $F_T = F^{(n)} * \cdots * F^{(n)}$ (n 次) 和 $\varphi_T = (\varphi^{(n)})^n$, 然而表达式 $T \stackrel{d}{=} \eta_1 + \cdots + \eta_n$ 不可能. 恰好有 "贫乏" 概率空间的一个例子属于杜布 (J. L. Doob; 见 [103]): 在此概率空间上有一参数为 $\lambda = 1$ 的泊松分布随机变量 (它是无限可分的: $F_T = F^{(n)} * \cdots * F^{(n)}$, 其中 $F^{(n)}$ 是对应于参数为 $\lambda = 1/n$ 的泊松分布的分布函数), 但是不存在参数为 $\lambda = 1/2$ 的泊松分布的随机变量 η_1 和 η_2.

*) 记号 "$\xi \stackrel{d}{=} \eta$" 表示随机变量 ξ 和 η 依分布相等 (相合), 即 $F_\xi(x) = F_\eta(x)$, 其中 $F_\xi(x)$ 和 $F_\eta(x)$ 是 ξ 和 η 的分布函数.

考虑到以上的结果，需要强调，上面的定义 1 实质上不明显地假设，原概率空间已经充分"丰富"。为避开此在所指出的缺陷 (练习题 11)，原概率空间已经足够"丰富"。

定理 1 随机变量 T 可以是和 $T_n=\sum_{k=1}^{n}\xi_{n,k}$ 的依分布收敛的极限，当且仅当 T 无限可分.

证明 如果 T 无限可分，则对于每个 $n\geqslant 1$，存在独立同分布随机变量 $\xi_{n,1},\cdots,\xi_{n,n}$，使 $T\overset{d}{=}\xi_{n,1}+\cdots+\xi_{n,n}$，即 $T\overset{d}{=}T_n, n\geqslant 1$.

相反，假设 $T_n\xrightarrow{d}T$. 我们现在证明 T 无限可分，即对于每个 $k\geqslant 1$，存在独立同分布随机变量 $\eta_1,\cdots,\eta_k$，使 $T\overset{d}{=}\eta_1+\cdots+\eta_k$.

固定某个 $k\geqslant 1$，并将变量 $T_{nk}=\sum_{i=1}^{nk}\xi_{nk,i}$ 表示为 $\zeta_n^{(1)}+\cdots+\zeta_n^{(k)}$，其中

$$\zeta_n^{(1)}=\xi_{nk,1}+\cdots+\xi_{nk,n},\cdots,\zeta_n^{(k)}=\xi_{nk,n(k-1)+1}+\cdots+\xi_{nk,nk}.$$

由于 $T_{nk}\xrightarrow{d}T, n\to\infty$，可见随机变量 $T_{nk}, n\geqslant 1$ 的分布函数序列相对列紧；故根据普罗霍罗夫定理知，随机变量 $T_{nk}, n\geqslant 1$ 的分布函数序列稠密 (§2). 其次，有

$$[\mathrm{P}\{\zeta_n^{(1)}>z\}]^k=\mathrm{P}\{\zeta_n^{(1)}>z,\cdots,\zeta_n^{(k)}>z\}\leqslant\mathrm{P}\{T_{nk}>kz\}$$

和

$$[\mathrm{P}\{\zeta_n^{(1)}<-z\}]^k=\mathrm{P}\{\zeta_n^{(1)}<-z,\cdots,\zeta_n^{(k)}<-z\}\leqslant\mathrm{P}\{T_{nk}<-kz\}.$$

由这两个不等式和 $T_{nk}, n\geqslant 1$ 的分布族的稠密性，可见 $\zeta_n^{(1)}, n\geqslant 1$ 的分布族的稠密性. 因此，存在子序列 $\{n_i\}\subseteq\{n\}$ 和随机向量 $(\eta_1,\cdots,\eta_k)$，而不失普遍性可以认为，定义在原 ("丰富") 概率空间上，并且

$$(\zeta_{n_i}^{(1)},\cdots,\zeta_{n_i}^{(k)})\xrightarrow{d}(\eta_1,\cdots,\eta_k),$$

或者等价地，对于任意 $\lambda_1,\cdots,\lambda_k\in\mathbb{R}$，有

$$\mathrm{E}e^{i(\lambda_1\zeta_{n_i}^{(1)}+\cdots+\lambda_k\zeta_{n_i}^{(k)})}\to\mathrm{E}e^{i(\lambda_1\eta_1+\cdots+\lambda_k\eta_k)}.$$

由于随机变量 $\zeta_{n_i}^{(1)},\cdots,\zeta_{n_i}^{(k)}$ 独立，可见

$$\mathrm{E}e^{i(\lambda_1\zeta_{n_i}^{(1)}+\cdots+\lambda_k\zeta_{n_i}^{(k)})}=\mathrm{E}e^{i\lambda_1\zeta_{n_i}^{(1)}}\cdots\mathrm{E}e^{i\lambda_k\zeta_{n_i}^{(k)}}\to\mathrm{E}e^{i\lambda_1\eta_1}\cdots\mathrm{E}e^{i\lambda_k\eta_k},$$

即

$$\mathrm{E}e^{i(\lambda_1\eta_1+\cdots+\lambda_k\eta_k)}=\mathrm{E}e^{i\lambda_1\eta_1}\cdots\mathrm{E}e^{i\lambda_k\eta_k},$$

于是由第二章 §12 定理 4 知随机变量 $\eta_1,\cdots,\eta_k$ 独立. 同样显然，它们有相同的分布. 此外，

$$T_{n_ik}=\zeta_{n_i}^{(1)}+\cdots+\zeta_{n_i}^{(k)}\xrightarrow{d}\eta_1+\cdots+\eta_k.$$

同样，有 $T_{n_ik}\xrightarrow{d}T$. 于是 (练习题 1)，得

$$T\overset{d}{=}\eta_1+\cdots+\eta_k. \qquad \square$$

注 2 如果将这一节开始的条件：对于每个 $n\geqslant 1$，随机变量 $\xi_{n,1},\cdots,\xi_{n,n}$ 同分布，换成它们满足渐近小 $\max_{1\leqslant k\leqslant n}\mathrm{P}\{|\xi_{nk}|\geqslant\varepsilon\}\to 0$ 的条件，则定理的结论仍然成立.

2. 随机变量无限可分的充分和必要条件 为验证给定的随机变量 T 是否无限可分，最简单的是考察其特征函数 $\varphi(t)$ 的形式. 如果对于每个 $n\geqslant 1$，存在这样的特征函数 $\varphi_n(t)$，使 $\varphi(t)=[\varphi_n(t)]^n$，则 T 无限可分.

对于高斯分布的情形，

$$\varphi(t)=e^{itm}e^{-\frac{t^2\sigma^2}{2}},$$

若设

$$\varphi_n(t)=e^{it\frac{m}{n}}e^{-\frac{t^2\sigma^2}{2n}},$$

则 $\varphi(t)=[\varphi_n(t)]^n$.

对于泊松分布的情形，

$$\varphi(t)=e^{\lambda(e^{it}-1)},$$

若设

$$\varphi_n(t)=e^{\frac{\lambda}{n}(e^{it}-1)},$$

则 $\varphi(t)=[\varphi_n(t)]^n$.

如果随机变量 T 服从 Γ 分布，其密度为

$$f(x)=\begin{cases}\dfrac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}, & 若\ x\geqslant 0,\\ 0, & 若\ x<0,\end{cases}$$

则 (第二章 §12 表 2－5) 其特征函数等于

$$\varphi(t)=\frac{1}{(1-i\beta t)^\alpha}.$$

从而，$\varphi(t)=[\varphi_n(t)]^n$，其中

$$\varphi_n(t)=\frac{1}{(1-i\beta t)^{\alpha/n}},$$

说明 T 无限可分.

我们 (不加证明) 引进如下关于"无限可分分布的特征函数一般形式"的一般结果.

定理 2 (柯尔莫戈洛夫 – 列维 – 辛钦) 随机变量 T 无限可分, 当且仅当其特征函数具有形式 $\varphi(t)=\exp\psi(t)$, 其中

$$\psi(t)=it\beta-\frac{t^2\sigma^2}{2}+\int_{-\infty}^{\infty}\left(e^{itx}-1-\frac{itx}{1+x^2}\right)\frac{1+x^2}{x^2}\,d\lambda(x), \tag{2}$$

而 $\beta\in\mathbb{R}, \sigma^2\geqslant 0, \lambda$ 是 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ 上的有限测度, 且 $\lambda(0)=0$.

3. 稳定随机变量 设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量序列, 且 $S_n=\xi_1+\cdots+\xi_n$. 假设存在这样的常数 $b_n, a_n>0$ 和随机变量 T, 使

$$\frac{S_n-b_n}{a_n}\xrightarrow{d}T. \tag{3}$$

试问如何描述 (3) 式中随机变量 T 的全部可能的极限分布?

如果对于独立同分布随机变量 $\xi_1,\xi_2,\cdots$, 有 $0<\sigma^2\equiv\mathrm{D}\xi_1<\infty$, 并且设 $b_n=n\mathrm{E}\xi_1$ 而 $a_n=\sigma\sqrt{n}$, 则由 §4 可见, 随机变量 T 服从标准正态分布 $\mathcal{N}(0,1)$.

如果

$$f(x)=\frac{\theta}{\pi(x^2+\theta^2)}$$

是参数为 $\theta>0$ 的柯西分布密度, 而 $\xi_1,\xi_2,\cdots$ 是密度为 $f(x)$ 的独立随机变量, 则其特征函数为 $\varphi_{\xi_1}(t)=e^{-\theta|t|}$, 而

$$\varphi_{S_n/n}(t)=\left(e^{-\frac{\theta}{n}|t|}\right)^n=e^{-\theta|t|},$$

因此随机变量 S_n/n 服从柯西分布 (分布参数仍然是 θ).

这样, 作为极限分布, 除了正态分布之外, 还可能有其他分布 (例如柯西分布). 设

$$\frac{S_n-b_n}{a_n}=\sum_{k=1}^{n}\xi_{n,k}=(T_n), \text{ 其中 } \xi_{n,k}=\frac{\xi_k}{a_n}-\frac{b_n}{na_n},\quad 1\leqslant k\leqslant n.$$

于是, 对于 T, 一切可以意料到的, 可以作为极限分布 (3) 式出现的分布, (根据定理 1) 一定是无限可分分布. 不过所考虑的随机变量 $T_n=(S_n-b_n)/a_n$ 的特点, 为这里可能出现的极限分布的结构, 可能提供了得到补充信息.

为此 (并考虑到注 1), 我们引进如下概念.

定义 2 随机变量 T (及其分布函数 $F(x)$ 和特征函数 $\varphi(t)$) 称做稳定的, 如果对于任何 $n\geqslant 1$, 存在这样的常数 $a_n>0, b_n$, 和与 T 同分布的独立随机变量 $\xi_1,\cdots,\xi_n$, 使

$$a_nT+b_n\overset{d}{=}\xi_1+\cdots+\xi_n, \tag{4}$$

或同样地

$$F\left(\frac{x-b_n}{a_n}\right)=F*\cdots*F(x)\quad\text{或}\quad[\varphi(t)]^n=\varphi(a_nt)e^{ib_nt}. \tag{5}$$

定理 3 随机变量 T 可以是随机变量

$$\frac{S_n-b_n}{a_n}, a_n>0,$$

按分布收敛的极限, 当且仅当 T 是稳定的.

证明 如果 T 是稳定的, 则根据 (4) 式

$$T\overset{d}{=}\frac{S_n-b_n}{a_n}, \text{从而 } \frac{S_n-b_n}{a_n}\xrightarrow{d}T,$$

其中 $S_n=\xi_1+\cdots+\xi_n$.

相反, 假设 $\xi_1,\xi_2,\cdots$ 是独立同分布的随机变量序列, $S_n=\xi_1+\cdots+\xi_n$, 且

$$\frac{S_n-b_n}{a_n}\xrightarrow{d}T, a_n>0.$$

现在证明 T 是稳定随机变量.

假如 T 是退化随机变量, 则它显然是稳定的. 因此, 我们将假设 T 是非退化随机变量.

固定 $k\geqslant 1$, 记

$$S_n^{(1)}=\xi_1+\cdots+\xi_n,\cdots,S_n^{(k)}=\xi_{n(k-1)+1}+\cdots+\xi_{kn};$$
$$T_n^{(1)}=\frac{S_n^{(1)}-b_n}{a_n},\cdots,T_n^{(k)}=\frac{S_n^{(k)}-b_n}{a_n}.$$

显然, 所有随机变量 $T_n^{(1)},\cdots,T_n^{(k)}$ 都同分布, 且

$$T_n^{(i)}\xrightarrow{d}T, n\to\infty, i=1,\cdots,k.$$

记

$$U_n^{(k)}=T_n^{(1)}+\cdots+T_n^{(k)}.$$

那么, 同定理 1 的证明一样, 可得

$$U_n^{(k)}\xrightarrow{d}T^{(1)}+\cdots+T^{(k)},$$

其中 $T^{(i)}(1\leqslant i\leqslant k)$ 独立, 且 $T^{(1)}\overset{d}{=}\cdots\overset{d}{=}T^{(k)}\overset{d}{=}T$.

另一方面,

$$\begin{aligned}U_n^{(k)}&=\frac{\xi_1+\cdots+\xi_{kn}-kb_n}{a_n}\\&=\frac{a_{kn}}{a_n}\left(\frac{\xi_1+\cdots+\xi_{kn}-b_{kn}}{a_{kn}}\right)+\frac{b_{kn}-kb_n}{a_n}\\&=\alpha_n^{(k)}V_{kn}+\beta_n^{(k)},\end{aligned} \tag{6}$$

其中

$$\alpha_n^{(k)}=\frac{a_{kn}}{a_n},\ \beta_n^{(k)}=\frac{b_{kn}-kb_n}{a_n},$$

而

$$V_{kn}=\frac{\xi_1+\cdots+\xi_{kn}-b_{kn}}{a_{kn}}.$$

由 (6) 式可见

$$V_{kn}=\frac{U_n^{(k)}-\beta_n^{(k)}}{\alpha_n^{(k)}},$$

其中 $V_{kn}\xrightarrow{d}T,U_n^{(k)}\xrightarrow{d}T^{(1)}+\cdots+T^{(k)},n\to\infty$.

由下面的引理知, 存在常数 $\alpha^{(k)}>0$ 和 $\beta^{(k)}$, 使 $\alpha_n^{(k)}\to\alpha^{(k)},\beta_n^{(k)}\to\beta^{(k)},n\to\infty$, 而

$$T\overset{d}{=}\frac{T^{(1)}+\cdots+T^{(k)}-\beta^{(k)}}{\alpha^{(k)}}.$$

从而, 随机变量 T 的稳定性得证. □

引理　*设 $\xi_n\xrightarrow{d}\xi$ 且存在常数 $a_n>0$ 和 b_n, 使*

$$a_n\xi_n+b_n\xrightarrow{d}\tilde{\xi},$$

而且 ξ 和 $\tilde{\xi}$ 都是非退化随机变量. 那么, 存在常数 $a>0$ 和 b, 使 $\lim a_n=a,\lim b_n=b$ 且

$$\tilde{\xi}\overset{d}{=}a\xi+b.$$

证明　设 φ_n,φ 和 $\tilde{\varphi}$ 相应为 ξ_n,ξ 和 $\tilde{\xi}$ 的特征函数. 那么, $a_n\xi_n+b_n$ 的特征函数 $\varphi_{a_n\xi_n+b_n}(t)$ 等于 $e^{itb_n}\varphi_n(a_nt)$. 根据定理 1 的系, 以及 §3 的练习题 3, 有

$$e^{itb_n}\varphi_n(a_nt)\to\tilde{\varphi}(t),\tag{7}$$

$$\varphi_n(t)\to\varphi(t),\tag{8}$$

其中收敛性为在 t 的每一变化区间上一致收敛.

设 $\{n_i\}$ 是使 $a_{n_i}\to a$ 的 $\{n\}$ 的子序列. 我们首先证明 $a<\infty$. 假设 $a=\infty$. 由 (7) 式可见, 对于任意 $c>0$,

$$\sup_{|t|\leqslant c}||\varphi_n(a_nt)|-|\tilde{\varphi}(t)||\to0,n\to\infty.$$

将 t 换成 $t_{n_i}=t_0/a_{n_i}$. 那么, 由于 $a_{n_i}\to a=\infty$, 可见

$$\left|\varphi_{n_i}\left(a_{n_i}\frac{t_0}{a_{n_i}}\right)\right|-\left|\tilde{\varphi}\left(\frac{t_0}{a_{n_i}}\right)\right|\to0,$$

因而

$$|\varphi_{n_i}(t_0)|\to|\tilde{\varphi}(0)|=1.$$

然而 $|\varphi_{n_i}(t_0)|\to|\varphi(t_0)|$. 因为对于任意 $t_0\in\mathbb{R}$, $|\varphi(t_0)|=1$, 故根据第二章 §12 定理 5, 随机变量应该是退化的, 而这与引理的假设矛盾.

这样 $a<\infty$. 现在假设存在两个子序列 $\{n_i\}$ 和 $\{n_i'\}$, 使 $a_{n_i}\to a,a_{n_i'}\to a'$, 其中 $a\neq a'$ (为确定性, 设 $0\leqslant a'<a$). 那么, 由 (7) 和 (8) 式, 可见,

$$|\varphi_{n_i}(a_{n_i}t)|\to|\varphi(at)|,|\varphi_{n_i}(a_{n_i}t)|\to|\tilde{\varphi}(t)|$$

和

$$|\varphi_{n_i'}(a_{n_i'}t)|\to|\varphi(a't)|,|\varphi_{n_i'}(a_{n_i'}t)|\to|\tilde{\varphi}(t)|,$$

故

$$|\varphi(at)|=|\varphi(a't)|.$$

从而, 对于任意 $t\in\mathbb{R}$, 有

$$|\varphi(t)|=\left|\varphi\left(\frac{a'}{a}t\right)\right|=\cdots=\left|\varphi\left(\left(\frac{a'}{a}\right)^nt\right)\right|\to1,n\to\infty.$$

因此 $|\varphi(t)|\equiv1$, 而根据第二章 §12 定理 5, 由此可见 ξ 是退化随机变量. 所产生的矛盾说明 $a=a'$, 而这表明存在有限极限 $\lim a_n=a$, 并且 $a\geqslant0$.

现在证明, 存在极限 $\lim b_n=b$, 并且 $a>0$. 由于 (8) 式在每个有限区间上一致成立, 故

$$\varphi_n(a_nt)\to\varphi(at).$$

因而, 由于 (7) 式可见, 对于所有使 $\varphi(at)\neq0$ 的 t, 存在极限 $\lim_n e^{itb_n}$. 设 $\delta>0$, 使得对于一切 $|t|<\delta$, 有 $\varphi(at)\neq0$. 那么, 对于这样的 t, 存在极限 $\lim e^{itb_n}$. 由此可见 (练习题 9) $\overline{\lim}|b_n|<\infty$.

假设存在两个子序列 $\{n_i\}$ 和 $\{n_i'\}$, 使 $\lim b_{n_i}=b$ 和 $\lim b_{n_i'}=b'$. 那么, 对于 $|t|<\delta$, 有

$$e^{itb}=e^{itb'},$$

从而 $b=b'$. 于是, 存在有限极限 $\lim b_n=b$, 而根据 (7) 式,

$$\tilde{\varphi}(t)=e^{itb}\varphi(at),$$

说明 $\tilde{\xi}\overset{d}{=}a\xi+b$. 由于 $\tilde{\xi}$ 非退化, 可见 $a>0$. □

4. 稳定分布特征函数的一般形式　我们现在 (不加证明) 给出稳定分布特征函数的一般形式.

定理 4 (列维－辛钦表现)　*随机变量 T 是稳定的, 当且仅当其特征函数 $\varphi(t)$ 具有 $\varphi(t)=\exp\psi(t)$ 的形式, 其中*

$$\psi(t)=it\beta-d|t|^\alpha\left(1-i\theta\frac{t}{|t|}G(t,\alpha)\right),\tag{9}$$

其中 $0<\alpha<2$, $\beta\in\mathbb{R}$, $d\geqslant 0$, $|\theta|\leqslant 1$, 且当 $t\to 0$ 时 $t/|t|\to 0$, 而

$$G(t,\alpha)=\begin{cases}\tan\dfrac{\pi}{2}\alpha, & 若\ \alpha\neq 1,\\ \dfrac{2}{\pi}\ln|t|, & 若\ \alpha=1.\end{cases} \tag{10}$$

我们指出, 对称稳定分布的特征函数的结构特别简单:

$$\varphi(t)=e^{-d|t|^{\alpha}}, \tag{11}$$

其中 $0<\alpha\leqslant 2$, $d\geqslant 0$.

5. 练习题

1. 证明, 如果 $\xi_n\xrightarrow{d}\xi$ 和 $\eta_n\xrightarrow{d}\eta$, 则 $\xi\overset{d}{=\!=}\eta$.

2. 证明, 如果 φ_1 和 φ_2 是两个无限可分特征函数, 则 $\varphi_1\times\varphi_2$ 也是无限可分特征函数.

3. 设 φ_n 是两个无限可分特征函数, 且对于每个 $t\in\mathbb{R}$, $\varphi_n(t)\to\varphi(t)$, 其中 $\varphi(t)$ 是一特征函数. 证明 $\varphi(t)$ 无限可分.

4. 证明, 无限可分分布的特征函数不等于 0.

5. 试举一例, 随机变量是无限可分的, 但并不是稳定的.

6. 设 ξ 是稳定随机变量, 证明, 对于一切 $r\in(0,\alpha)$, $0<\alpha<2$, 数学期望 $\mathbf{E}|\xi|^r<\infty$.

7. 证明, 如果 ξ 是参数为 $0<\alpha\leqslant 1$ 的稳定随机变量, 则其特征函数 $\varphi(t)$ 在 $t=0$ 不可微.

8. 直接证明, 函数 $\varphi(t)=e^{-d|t|^{\alpha}}$ 是特征函数, 其中 $0<\alpha\leqslant 2$, $d\geqslant 0$.

9. 设 $\{b_n\}_{n\geqslant 1}$ 是一数列, 满足条件: 对于一切 $|t|<\delta$, $\delta>0$, 极限 $\lim_n e^{itb_n}$ 存在. 证明 $\overline{\lim}_n|b_n|<\infty$.

10. 证明, 二项分布和均匀分布, 都不是无限可分分布.

11. 假设分布函数 F 及其特征函数 φ 可以表示为: $F^{(n)}*\cdots*F^{(n)}$ (n 次), $\varphi=[\varphi^{(n)}]^n$, 其中 $F^{(n)}$ 是某一分布函数, 而 $\varphi^{(n)}(n\geqslant 1)$ 是其特征函数. 证明存在 (充分"丰富" 的) 概率空间 $(\Omega,\mathscr{F},\mathbf{P})$ 和定义在它上面的随机变量 T 和 $\{\eta_k^{(n)}\}_{k\leqslant n}$ $(n\geqslant 1)$, 满足

$$T\overset{d}{=}\eta_1^{(n)}+\cdots+\eta_n^{(n)}\ (n\geqslant 1)$$

(T 服从分布 F, 而 $\eta_1,\cdots,\eta_n$ 独立同服从分布 $F^{(n)}$).

12. 举一随机变量的例子, 使之本身不是无限可分的, 然而其特征函数仍然不等于 0.

§7. 弱收敛的 "可度量性"

1. 关于弱收敛的可度量性 设 $(E,\mathscr{E},\rho)$ 是度量空间, 而 $\mathscr{P}(E)=\{P\}$ 是 $(E,\mathscr{E})$ 上的概率测度族. 自然地提出问题: 在 §1 中所考虑的弱收敛 $P_n\xrightarrow{w}P$, 是否可以 "度量", 即是否可以在 $\mathscr{P}(E)$ 中的两个测度 $\widetilde{P}$ 和 P 之间, 引进这样的距离 $\delta(P,\widetilde{P})$, 使收敛性 $\delta(P_n,P)\to 0$ 与收敛性 $P_n\xrightarrow{w}P$ 等价.

鉴于上面提出的问题, 有意conditions指出, 随机变量序列依概率收敛 $\xi_n\xrightarrow{P}\xi$, 可以度量化, 例如, 借助距离

$$d_P(\xi,\eta)=\inf\{\varepsilon>0:\mathbf{P}\{|\xi-\eta|\geqslant\varepsilon\}\leqslant\varepsilon\},$$

或者, 借助距离

$$d(\xi,\eta)=\mathbf{E}\min(1,|\xi-\eta|),\quad d(\xi,\eta)=\mathbf{E}\frac{|\xi-\eta|}{1+|\xi-\eta|}.$$

(更一般地, 可以设 $d(\xi,\eta)=\mathbf{E}g(|\xi-\eta|)$, 其中函数 $g(x)$, $x\geqslant 0$, 可以取在 0 连续的任意博雷尔非负递增函数: 当 $x>0$ 时 $g(x)>0$, $g(0)=0$, 且对于一切 $x\geqslant 0$, $y\geqslant 0$, $g(x+y)\leqslant g(x)+g(y)$.) 然而, 对于在 $(\Omega,\mathscr{F},\mathbf{P})$ 上的一切随机变量的空间不存在这样的距离 $d(\xi,\eta)$, 使得当且仅当 ξ_n 依概率 1 收敛于 ξ 时 $d(\xi_n,\xi)\to 0$. (这很容易证明: 只需考虑依概率收敛, 但是不依概率 1 收敛的随机变量序列 ξ_n, $n\geqslant 1$.) 换句话说, 依概率 1 收敛不可度量化. (见第二章 §10 中练习题 11 和 2 的命题.)

这一节的目的, 是在指出度量 $(L(P,\widetilde{P})$ 和 $\|P-\widetilde{P}\|^*_{BL})$ 的同时, 在测度 $\mathscr{P}(E)$ 的空间中, 建立弱收敛的可度量性:

$$P_n\xrightarrow{w}P\Leftrightarrow L(P_n,P)\to 0\Leftrightarrow\|P_n-P\|^*_{BL}\to 0. \tag{1}$$

2. 列维 – 普罗霍罗夫度量 $L(P,\widetilde{P})$ 设

$$\rho(x,A)=\inf\{\rho(x,y):y\in A\},$$
$$A^{\varepsilon}=\{x\in E:\rho(x,A)<\varepsilon\},\ A\in\mathscr{E}.$$

对于任意两测度 $P,\widetilde{P}\in\mathscr{P}(\mathscr{E})$, 设

$$\sigma(P,\widetilde{P})=\inf\{\varepsilon>0:P(F)\leqslant\widetilde{P}(F^{\varepsilon})+\varepsilon,对于一切闭集F\in\mathscr{E}\}, \tag{2}$$

和

$$L(P,\widetilde{P})=\max\{\sigma(P,\widetilde{P}),\sigma(\widetilde{P},P)\}. \tag{3}$$

下面的引理证明, 这样定义的所谓列维 – 普罗霍罗夫度量函数 $L(P,\widetilde{P})$, $P,\widetilde{P}\in\mathscr{P}(\mathscr{E})$, 确实是度量.

引理 1　函数 $L(P,\tilde{P})$ 具有距离的性质.

a) $L(P,\tilde{P})=L(\tilde{P},P)(=\sigma(P,\tilde{P})=\sigma(\tilde{P},P))$;

b) $L(P,\tilde{P})\leqslant L(P,\hat{P})+\sigma(\hat{P},\tilde{P})$;

c) $L(P,\tilde{P})=0$, 当且仅当 $P=\tilde{P}$.

证明　a) 只需证明, 对于 $\alpha>0,\beta>0$,

$$\text{“}P(F)\leqslant\tilde{P}(F^{\alpha})+\beta,\ \text{对于一切闭集 } F\in\mathscr{E}\text{”}\tag{4}$$

当且仅当

$$\text{“}\tilde{P}(F)\leqslant P(F^{\alpha})+\beta,\ \text{对于一切闭集 } F\in\mathscr{E}\text{”}.\tag{5}$$

设 T 是 $\mathscr{E}$ 中的闭集. 那么 T^{α} 是开集, 则不难验证 $T\subseteq E\setminus(E\setminus T^{\alpha})^{\alpha}$. 如果 (4) 式成立, 那么, 特别

$$P(E\setminus T^{\alpha})\leqslant\tilde{P}((E\setminus T^{\alpha})^{\alpha})+\beta,$$

从而

$$\tilde{P}(T)\leqslant\tilde{P}(E\setminus(E\setminus T^{\alpha})^{\alpha})\leqslant P(T^{\alpha})+\beta.$$

因此, 说明 (4) 和 (5) 两式等价. 由此可见

$$\sigma(P,\tilde{P})=\sigma(\tilde{P},P),\tag{6}$$

于是,

$$L(P,\tilde{P})=\sigma(P,\tilde{P})=\sigma(\tilde{P},P)=L(\tilde{P},P).\tag{7}$$

b) 设 $L(P,\hat{P})<\delta_1$, $L(\hat{P},\tilde{P})<\delta_2$. 那么, 对于每一个闭集 $F\in\mathscr{E}$, 有

$$\tilde{P}(F)\leqslant\hat{P}(F^{\delta_2})+\delta_2\leqslant P((F^{\delta_2})^{\delta_1})+\delta_1+\delta_2\leqslant P(F^{\delta_1+\delta_2})+\delta_1+\delta_2.$$

于是, $L(P,\tilde{P})\leqslant\delta_1+\delta_2$. 由此可见

$$L(P,\tilde{P})\leqslant L(P,\hat{P})+L(\hat{P},\tilde{P}).$$

c) 如果 $L(P,\tilde{P})=0$, 那么, 对于每一个闭集 $F\in\mathscr{E}$ 和任意 $\alpha>0$, 有

$$P(F)\leqslant\tilde{P}(F^{\alpha})+\alpha.\tag{8}$$

由于 $F^{\alpha}\downarrow F,\alpha\downarrow 0$, 则由 (8) 式当 $\alpha\downarrow 0$ 时求极限, 得 $P(F)\leqslant\tilde{P}(F)$; 由对称性, 有 $\tilde{P}(F)\leqslant P(F)$. 这样, 对于一切闭集 $F\in\mathscr{E}$, 有 $P(F)=\tilde{P}(F)$. 对于每一个博雷尔集 $A\in\mathscr{E}$ 和任意 $\varepsilon>0$, 存在这样的开集 $G_{\varepsilon}\supseteq A$ 和闭集 $F_{\varepsilon}\subseteq A$, 且 $P(G_{\varepsilon}\setminus F_{\varepsilon})\leqslant\varepsilon$. 由此可见, 度量空间 $(E,\mathscr{E},\rho)$ 上的任何概率测度 P, 完全决定于 P 在闭集上的值. 从而, 由对于一切闭集 $F\in\mathscr{E}$, 有 $\tilde{P}(F)=P(F)$. 于是, 对于一切博雷尔集 $A\in\mathscr{E}$, 有 $\tilde{P}(A)=P(A)$. □

定理 1　列维 – 普罗霍罗夫度量 $L(P,\tilde{P})$, 可以使弱收敛度量化:

$$L(P_n,P)\to 0\Leftrightarrow P_n\xrightarrow{w}P.\tag{9}$$

证明　($\Rightarrow$). 设 $L(P_n,P)\to 0,n\to\infty$. 那么, 对于任意固定的闭集 $F\in\mathscr{E}$ 和任何 $\varepsilon>0$, 根据 (2) 式和引理 1 的命题 a), 有

$$\varlimsup_n P_n(F)\leqslant P(F^{\varepsilon})+\varepsilon.\tag{10}$$

在上式中令 $\varepsilon\downarrow 0$, 得

$$\varlimsup_n P_n(F)\leqslant P(F).$$

根据 §1 的定理 1, 可见

$$P_n\xrightarrow{w}P.\tag{11}$$

($\Leftarrow$). 蕴涵关系 ($\Leftarrow$) 的证明, 基于一系列深刻而有益的事实; 它既进一步揭示弱收敛本身的内容, 又阐明其构建方法和研究收敛 "速度" 的方法.

这样, 设 $P_n\xrightarrow{w}P$. 这表明对于任意连续的有界函数 $f=f(x)$, 有

$$\int_E f(x)P_n(\mathrm{d}x)\to\int_E f(x)P(\mathrm{d}x).\tag{12}$$

现在假设 $\mathscr{G}$ 是某一等阶梯函数 $g=g(x)$ 类 (如果对于一切 $g\in\mathscr{G}$, 有 $\rho(x,y)<\delta$, 则对于任意 $\varepsilon>0$ 存在 $\delta>0$, 使 $|g(y)-g(x)|<\varepsilon$), 使得对于同一常数 $C>0$,(对于一切 $x\in E$ 和 $g\in\mathscr{G}$) $|g(x)|\leqslant C$. 根据 §8 定理 3, 对于函数类 $\mathscr{G}$, 性质 (12) 的如下条件成立:

$$P_n\xrightarrow{w}P\Rightarrow\sup_{g\in\mathscr{G}}\left|\int_E g(x)P_n(\mathrm{d}x)-\int_E g(x)P(\mathrm{d}x)\right|\to 0.\tag{13}$$

对于任意 $A\in\mathscr{E}$ 和 $\varepsilon>0$, (如同 §1 定理 1) 记

$$f_A^{\varepsilon}(x)=\left[1-\frac{\rho(x,A)}{\varepsilon}\right]^{+}.\tag{14}$$

显然

$$I_A(x)\leqslant f_A^{\varepsilon}(x)\leqslant I_{A^{\varepsilon}}(x);\tag{15}$$

且

$$|f_A^{\varepsilon}(x)-f_A^{\varepsilon}(y)|\leqslant\varepsilon^{-1}|\rho(x,A)-\rho(y,A)|\leqslant\varepsilon^{-1}\rho(x,y).$$

这样, 对于类 $\mathscr{G}^{\varepsilon}=\{f_A^{\varepsilon}(x),A\in\mathscr{E}\}$, 有 (13) 式. 因此,

$$\Delta_n=\sup_{A\in\mathscr{E}}\left|\int_E f_A^{\varepsilon}(x)P_n(\mathrm{d}x)-\int_E f_A^{\varepsilon}(x)P(\mathrm{d}x)\right|\to 0,n\to\infty.\tag{16}$$

由此及 (15) 式, 可见对于任意闭集 $A \in \mathscr{E}$ 和 $\varepsilon > 0$, 有

$$P(A^{\varepsilon}) \geqslant \int_E f_A^{\varepsilon}(x)\mathrm{d}P \geqslant \int_E f_A^{\varepsilon}(x)\mathrm{d}P_n - \Delta_n \geqslant P_n(A) - \Delta_n. \tag{17}$$

选择 $n(\varepsilon)$ 使对于一切 $n \geqslant n(\varepsilon)$, 有 $\Delta_n \leqslant \varepsilon$. 那么, 由 (17) 式, 对于 $n \geqslant n(\varepsilon)$, 有

$$P(A^{\varepsilon}) \geqslant P_n(A) - \varepsilon. \tag{18}$$

由此及定义 (2),(3) 式可见, 当 $n \geqslant n(\varepsilon)$ 时, 有 $L(P_n, P) \leqslant \varepsilon$. 于是,

$$P_n \xrightarrow{w} P \Rightarrow \Delta_n \to 0 \Rightarrow L(P_n, P) \to 0.$$

定理 (精确到 (13) 式的命题) 得证. □

3. 度量 $\|P - \tilde{P}\|^*_{BL}$ 以 BL 表示一切连续有界函数 $f = f(x), x \in E$ 的集合, 其中 $\|f\|_{\infty} = \sup_x |f(x)| < \infty$, 而且每一个函数 $f = f(x)$ 都满足利普希茨条件:

$$\|f\|_L = \sup_{x \neq y} \frac{|f(x) - f(y)|}{\rho(x, y)} < \infty.$$

设 $\|f\|_{BL} = \|f\|_{\infty} + \|f\|_L$. 以 $\|\cdot\|_{BL}$ 为范数的空间 BL 是巴拿赫空间.

定义度量 $\|P - \tilde{P}\|^*_{BL}$, 其中

$$\|P - \tilde{P}\|^*_{BL} = \sup_{f \in BL} \left\{ \left| \int f \mathrm{d}(P - \tilde{P}) \right| : \|f\|_{BL} \leqslant 1 \right\}. \tag{19}$$

(可以验证, $\|P - \tilde{P}\|^*_{BL}$ 事实上满足对于度量提出的一切要求; 练习题 2.)

定理 2 *度量 $\|P - \tilde{P}\|^*_{BL}$ 可以使弱收敛度量化:*

$$\|P_n - P\|^*_{BL} \to 0 \Leftrightarrow P_n \xrightarrow{w} P.$$

证明 由 (13) 式立即得蕴涵关系 (⇐). 为证明 (⇒) 只需验证, 在弱收敛的定义中, $P_n \xrightarrow{w} P$ 就是对于任意有界连续函数 $f = f(x)$, 满足性质 (12) 式, 故只需局限于考虑满足利普希茨条件的有界函数类. 换句话说, 如果能证明下面的结果 (引理), 则蕴涵关系 (⇒) 将得到证明.

引理 2 *弱收敛 $P_n \xrightarrow{w} P$ 成立, 当且仅当性质 (12) 对于 BL 类的任何函数 $f = f(x)$ 成立.*

证明 引理的一个方面的证明显然. 现在考虑由 (14) 式定义的函数 $f_A^{\varepsilon} = f_A^{\varepsilon}(x)$. 在证明定理 1 时已经证明, 对于任意 $\varepsilon > 0$, $\mathscr{F}^{\varepsilon} = \{f_A^{\varepsilon}(x), A \in \mathscr{E}\} \subseteq BL$. 如果现在分析 §1 的定理 1 中蕴涵关系 (I)⇒(II) 的证明, 就会发现, 在证明中实际上并非对于一切有界连续函数用到性质 (12), 而只对于类 $\mathscr{F}^{\varepsilon}$ ($\varepsilon > 0$) 中的函数用到. 由于 $\mathscr{F}^{\varepsilon} \subseteq BL$, $\varepsilon > 0$, 则显然由性质 (12) 对于类 BL 中的函数成立, 可见 §1 定理 1 中命题 II 的结论成立, 而后者 (仍然因为 §1 定理 1) 等价于弱收敛 $P_n \xrightarrow{w} P$. □

注. 定理 2 的结论可以由定理 1 推出 (反之也一样), 假如利用 (对于可分度量空间 $(E, \mathscr{E}, \rho)$ 成立的) $L(P, \tilde{P})$ 和 $\|P - \tilde{P}\|^*_{BL}$ 之间的如下不等式:

$$\|P - \tilde{P}\|^*_{BL} \leqslant 2L(P, \tilde{P}), \tag{20}$$

$$\varphi(L(P, \tilde{P})) \leqslant \|P - \tilde{P}\|^*_{BL}, \quad 其中 \quad \varphi(x) = \frac{2x^2}{2 + x}. \tag{21}$$

注意到, 对于 $x \geqslant 0$, $0 \leqslant \varphi(x) \leqslant 2/3$, 当且仅当 $x \leqslant 1$, 且对于 $0 \leqslant x \leqslant 1$, $\varphi(x) \geqslant 2x^2/3$; 由 (20) 和 (21) 式, 可见: 如果 $L(P, \tilde{P}) \leqslant 1$ 或 $\|P - \tilde{P}\|^*_{BL} \leqslant 2/3$, 则

$$\frac{2}{3}L^2(P, \tilde{P}) \leqslant \|P - \tilde{P}\|^*_{BL} \leqslant 2L(P, \tilde{P}). \tag{22}$$

4. 练习题

1. 证明, 当 $E = \mathbb{R}$ 时概率分布 P 和 $\tilde{P}$ 之间的列维 – 普罗霍罗夫度量 $L(P, \tilde{P})$, 不小于对应于 P 和 $\tilde{P}$ 的分布函数 F 和 $\tilde{F}$ 之间的列维距离 $L(F, \tilde{F})$ (见 §1 练习题 4). 举例说明这些度量间的严格不等式成立.

2. 证明公式 (19) 在决定空间 BL 中的度量.

3. 证明不等式 (20),(21) 和 (22).

4. 设 $F = F(x)$ 和 $G = G(x)$ 是两个分布函数, P_c 和 Q_c 是它们与直线 $x + y = c$ 的交点. 证明列维距离 $L(F, \tilde{F})$ (见 §1 练习题 4)

$$L(F, G) = \sup_c \frac{\overline{P_c Q_c}}{\sqrt{2}},$$

其中 $\overline{P_c Q_c}$ 表示点 P_c 和 Q_c 之间的线段的长度.

5. 证明, 一切分布函数的集合, 关于列维距离的是完备空间.

§8. 关于测度的弱收敛与随机元的几乎处处收敛的联系 ("一个概率空间的方法")

1. 随机元收敛性的定义 假设在概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 上给定随机元 $X = X(\omega)$, $X_n = X_n(\omega), n \geqslant 1$, 而 $(E, \mathscr{E}, \rho)$ 是随机元的值空间; 参见第二章 §5. 记 P 和 P_n 是 X 和 X_n 的概率分布, 即设

$$P(A) = \mathbf{P}\{\omega : X(\omega) \in A\}, \quad P_n(A) = \mathbf{P}\{\omega : X_n(\omega) \in A\}, A \in \mathscr{E}.$$

推广随机变量依分布收敛的概念 (见第二章 §10), 引进下面的定义.

定义 1 *称随机元序列 $X_n, n \geqslant 1$ 为依分布收敛或依分布律收敛的* (记号: $X_n \xrightarrow{\mathscr{D}} X$, 或 $X_n \xrightarrow{\mathscr{L}} X$, 或 $X_n \xrightarrow{law} X$), 如果 $P_n \xrightarrow{w} P$.

仿照随机变量的依概率收敛和依概率 1 收敛 (第二章 §10), 下列定义是自然的.

定义 2 称随机元序列 $X_n, n \geqslant 1$ 依概率收敛于 X ($X_n \xrightarrow{\mathbf{P}} X$), 如果

$$\mathbf{P}\{\omega : \rho(X_n(\omega), X(\omega)) \geqslant \varepsilon\} \to 0, n \to \infty. \tag{1}$$

定义 3 称随机元序列 $X_n, n \geqslant 1$, 依概率 1 (几乎必然或几乎处处) 收敛于 X ($X_n \xrightarrow{\text{a.s.}} X$, $X_n \xrightarrow{\text{a.e.}} X$), 如果 $\rho(X_n(\omega), X(\omega)) \xrightarrow{\text{a.e.}} 0, n \to \infty$.

注 1 当然, 只有在 $\rho(X_n(\omega), X(\omega))$ 作为 $\omega \in \Omega$ 的函数是随机变量时, 即当 $\rho(X_n(\omega), X(\omega))$ 为 $\mathscr{F}$ 可测时, 最后两个定义才有意义. 当空间 $(E, \mathscr{E}, \rho)$ 可分时, 这当然成立 (练习题 1).

注 2 关于定义 2 需要指出, 引进的依概率收敛性, 可以由下面的 (定义在 $(\Omega, \mathscr{F}, \mathbf{P})$ 上、取值于 E 的; 练习题 2) 随机元 X 和 Y 之间的范泰 (Fan Ky) 度量 (练习题 2)

$$d_{\mathbf{P}}(X, Y) = \inf\{\varepsilon > 0 : \mathbf{P}\{\rho(X(\omega), Y(\omega)) \geqslant \varepsilon\} \leqslant \varepsilon\} \tag{2}$$

来度量化.

注 3 如果依概率收敛和依概率 1 收敛的定义, 要求随机元定义在同一概率空间上, 则按分布收敛的定义 $X_n \xrightarrow{\mathscr{D}} X$, 只与分布的收敛性有关, 从而可以认为 $X(\omega), X_1(\omega), X_2(\omega), \cdots$ 在同一空间 E 上取值, 不过可能给定在 "各自的" 概率空间 $(\Omega, \mathscr{F}, \mathbf{P}), (\Omega_1, \mathscr{F}_1, \mathbf{P}_1), (\Omega_2, \mathscr{F}_2, \mathbf{P}_2), \cdots$ 上. 然而不失普遍性, 如果作为上述各空间的直积, 并且定义 $X(\omega, \omega_1, \omega_2, \cdots) = X(\omega), X_1(\omega, \omega_1, \omega_2, \cdots) = X_1(\omega_1), \cdots$, 则总可以认为它们定义在同一个概率空间上.

2. 按分布相等的随机元 根据定义 1 和在勒贝格积分中变量替换的定理 (第二章 §6 定理 7), 对于任意有界连续函数 $f = f(x), x \in E$, 有

$$X_n \xrightarrow{\mathscr{D}} X \Leftrightarrow \mathbf{E}f(X_n) \to f(X). \tag{3}$$

由 (3) 式可见, 根据勒贝格控制收敛定理 (第二章 §6 定理 3), 由收敛性 $X_n \xrightarrow{\text{a.e.}} X$, 立即可以得到 $X_n \xrightarrow{\mathscr{D}} X$. 因此, 显然对于 X 和 X_n 是随机变量的情形, 有同样的结果 (第二章 §10 定理 2). 更意想不到的是, 在某种意义上有相反的结果. 下面就给出其确切的提法, 然后讨论其应用.

定义 4 设随机元 $X = X(\omega')$ 和 $Y = Y(\omega'')$, 分别定义在概率空间 $(\Omega', \mathscr{F}', \mathbf{P}')$ 和 $(\Omega'', \mathscr{F}'', \mathbf{P}'')$ 上, 取值于同一空间 E. 如果 $X = X(\omega')$ 和 $Y = Y(\omega'')$ 有相同的概率分布, 则称之为按分布为相等的 (等价的或相合的; 记作 $X \stackrel{\mathscr{D}}{=} Y$).

定理 1 假设 $(E, \mathscr{E}, \rho)$ 是可分度量空间.

1. 设随机元 $X, X_n (n \geqslant 1)$, 定义在概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 上, 取值于 E 中, 且 $X_n \xrightarrow{\mathscr{D}} X$. 那么, 存在概率空间 $(\Omega^*, \mathscr{F}^*, \mathbf{P}^*)$ 和定义在该空间上且取值于 E 的随机元 $X^*, X_n^* (n \geqslant 1)$, 使

$$X_n^* \xrightarrow{\text{a.e.}} X^*,$$

且

$$X^* \stackrel{\mathscr{D}}{=} X, \quad X_n^* \stackrel{\mathscr{D}}{=} X_n, \quad n \geqslant 1.$$

2. 设 $P, P_n (n \geqslant 1)$ 是 $(E, \mathscr{E}, \rho)$ 上的概率测度, 且 $P_n \xrightarrow{w} P$. 那么, 存在概率空间 $(\Omega^*, \mathscr{F}^*, \mathbf{P}^*)$ 和定义在该空间上且取值于 E 的随机元 $X^*, X_n^* (n \geqslant 1)$, 使

$$X_n^* \xrightarrow{\text{a.e.}} X^*,$$

且

$$P^* = P, \quad P_n^* = P_n, \quad n \geqslant 1.$$

其中 P^* 和 P_n^* 是 X^* 和 X_n^* 的概率分布.

在证明定理之前, 我们指出: 第一, 只需证明命题 2; 因为, 如果把 P 和 P_n 取作 X 和 X_n 的概率分布, 则命题 1 可以由命题 2 得到. 第二, 需要指出, 这一定理最一般情形的证明技术上相当复杂. 正因为如此, 我们仅对 $E = \mathbb{R}$ 的情形证明该定理. 这一证明相当简明易懂, 并且为所求对象给出了简单描述性构造. (美中不足的是, 这一构造在一般情形下, 甚至对于 $E = \mathbb{R}^2$ 的情形, 并不 "适用".)

定理 1 的证明 ($E = \mathbb{R}$ 的情形). 设 $F = F(x)$ 和 $F_n = F_n(x)$ 是对应于 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上的测度 P 和 P_n 的分布函数; 而是 $Q = Q(u)$ 与函数 $F = F(x)$ 相联系的分位函数, 唯一地决定于如下公式

$$Q(u) = \inf\{x : F(x) \geqslant u\}, \quad 0 < u < 1. \tag{4}$$

不难验证,

$$F(x) \geqslant u \Leftrightarrow Q(u) \leqslant x. \tag{5}$$

现在考虑 $\Omega^* = (0, 1), \mathscr{F}^*(x) = \mathscr{B}(0, 1)$, 而 $\mathbf{P}^*$ 是勒贝格测度: $\mathbf{P}^*(d\omega^*) = d\omega^*$. 设 $X^*(\omega^*) = Q(\omega^*), \omega^* \in \Omega^*$. 那么,

$$\mathbf{P}^*\{\omega^* : X^*(\omega^*) \leqslant x\} = \mathbf{P}^*\{\omega^* : Q(\omega^*) \leqslant x\} = \mathbf{P}^*\{\omega^* : \omega^* \leqslant F(x)\} = F(x),$$

即所构造的随机变量 $X^*(\omega^*) = Q(\omega^*)$ 的分布恰好是 P. 类似地, 随机变量 $X_n^*(\omega^*) = Q_n(\omega^*)$ 的分布恰好是 P_n.

其次, 并不复杂地可以证明, 由在极限函数 $F = F(x)$ 的每个连续点上, $F_n(x)$ 收敛于 $F(x)$ (对于 $E = \mathbb{R}$ 的情形, 等价于 $P_n \xrightarrow{w} P$; 见 §1 的定理 1), 可见分位函数序列 $Q_n(u), n \geqslant 1$, 在极限函数 $Q = Q(u)$ 的每个连续点上, 也收敛于 $Q(u)$. 由于函数 $Q = Q(u), u \in (0, 1)$, 最多有可数个间断点, 则其勒贝格测度 $\mathbf{P}^*$ 等于 0. 因而

$$X_n^*(\omega^*) = Q_n(\omega^*) \xrightarrow{\text{a.e.}} X^*(\omega^*) = Q(\omega^*).$$

定理 1 (对于 $E = \mathbb{R}$ 的情形) 得证. □

定理 1 所描绘的, 由给定的随机元 X 和 X_n, 向定义在同一概率空间上的新随机元 X^* 和 X_n^* 的过渡, 说明了本节的标题中的名称 "一个概率空间的方法" 的含义.

现在利用这一方法, 可以比较容易地证明一系列的命题.

3. 一个概率空间方法的应用 假设 $X, X_n (n \geqslant 1)$ 是定义在同一概率空间 $(\Omega, \mathscr{F}, \mathrm{P})$ 上, 在可分度量空间 $(E, \mathscr{E}, \rho)$ 中取值的随机元, 且 $X_n \xrightarrow{\mathscr{D}} X$. 设 $h = h(x), x \in E$, 是 $(E, \mathscr{E}, \rho)$ 到另外一个可分度量空间 $(E', \mathscr{E}', \rho')$ 上的可测影射. 在概率论和数理统计中, 常遇到这样的问题: 在 $h = h(x)$ 满足什么条件时, 由收敛性 $X_n \xrightarrow{\mathscr{D}} X$ 可以得出收敛性 $h(X_n) \xrightarrow{\mathscr{D}} h(X)$.

例如, 假设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量, 且 $\mathrm{E}\xi_1 = m, \mathrm{D}\xi_1 = \sigma^2 > 0$, 而 $\overline{X}_n = (\xi_1 + \cdots + \xi_n)/n$. 由中心极限定理, 可见

$$\frac{\sqrt{n}(\overline{X}_n - m)}{\sigma} \xrightarrow{d} N(0,1).$$

问对于什么函数 $h = h(x)$ 可以保障

$$h\left(\frac{\sqrt{n}(\overline{X}_n - m)}{\sigma}\right) \xrightarrow{d} h(N(0,1))?$$

(对于连续函数 $h = h(x)$ 肯定可运用著名的曼 – 沃尔德 [H. B. Mann – A. Wald] 定理, 因此立即可得:

$$\frac{n(\overline{X}_n - m)^2}{\sigma^2} \xrightarrow{d} \chi_1^2,$$

其中 χ_1^2 是自由度为 1 的 χ^2 分布随机变量; 见第一章 §3 表 1 – 2.)

另一例子. 如果 $X = X(t, \omega), X_n = X_n(t, \omega), t \in T$, 是随机过程 (见第二章 §5), 而

$$h(X) = \sup_{t \in T} |X(t, \omega)|, \; h(X_n) = \sup_{t \in T} |X_n(t, \omega)|,$$

则上面所提问题表示: 在什么条件下由按分布过程收敛 $X_n \xrightarrow{\mathscr{D}} X$, 可以得出其上确界按分布收敛 $h(X_n) \xrightarrow{\mathscr{D}} h(X)$?

保障蕴涵关系

$$X_n \xrightarrow{\mathscr{D}} X \Rightarrow h(X_n) \xrightarrow{\mathscr{D}} h(X)$$

的条件之一是, 影射 $h = h(x)$ 连续. 事实上, 如果 $f = f(x')$ 在 E' 上是有界连续函数, 则 $f = f(h(x))$ 在 E 上也是有界连续函数. 从而

$$X_n \xrightarrow{\mathscr{D}} X \Rightarrow \mathrm{E}f(h(X_n)) \to \mathrm{E}f(h(X)).$$

下面的定理表明, 考虑到极限随机元 X 的性质, 对于函数 $h = h(x)$ 连续性的要求可以进行一定减弱.

记 $\Delta_h = \{x \in E : h(x)$ 在点 x 处不 ρ- 连续 $\}$, 换句话说, 设 Δ_h 是函数 $h = h(x)$ 的间断点的集合. 注意 $\Delta_h \in \mathscr{E}$ (练习题 4).

定理 2 1. 设 $(E, \mathscr{E}, \rho)$ 和 $(E', \mathscr{E}', \rho')$ 是可分度量空间, 且 $X_n \xrightarrow{\mathscr{D}} X$, 而影射 $h = h(x), x \in E$, 满足条件:

$$\mathrm{P}\{\omega : X \in \Delta_h\} = 0. \tag{6}$$

那么, $h(X_n) \xrightarrow{\mathscr{D}} h(X)$.

2. 设 $P, P_n (n \geqslant 1)$ 是可分度量空间 $(E, \mathscr{E}, \rho)$ 上的概率测度, 且 $P_n \xrightarrow{w} P$, 而 $h = h(x), x \in E$, 是 $(E, \mathscr{E}, \rho)$ 到可分度量空间 $(E', \mathscr{E}', \rho')$ 的可测影射. 假设

$$P\{x : x \in \Delta_h\} = 0.$$

那么, $P_n^h \xrightarrow{w} P^h$, 其中 $P_n^h(A) = P_n\{h(x) \in A\}, P^h(A) = P\{h(x) \in A\}, A \in \mathscr{E}'$.

证明 像定理 1 一样, 例如, 只需证明命题 2.

设 X^* 和 $X_n^* (n \geqslant 1)$ 是用 "一个概率空间的方法" 建立的随机元, 且满足 $X^* \stackrel{\mathscr{D}}{=} X, X_n^* \stackrel{\mathscr{D}}{=} X_n (n \geqslant 1), X_n^* \xrightarrow{\text{a.s.}} X^*$. 记

$$A^* = \{\omega^* : \rho(X_n^*, X^*) \nrightarrow 0\}, \quad B^* = \{\omega^* : X^*(\omega^*) \in \Delta_h\}.$$

那么, $\mathrm{P}^*(A^* \cup B^*) = 0$ 且对于 $\omega^* \notin A^* \cup B^*$,

$$h(X_n^*(\omega^*)) \to h(X^*(\omega^*)),$$

而这说明 $h(X_n^*) \xrightarrow{\text{a.s.}} h(X^*)$. 在第 1 小节已经指出, 由此可见 $h(X_n^*) \xrightarrow{\mathscr{D}} h(X^*)$. 由于 $h(X_n^*) \stackrel{\mathscr{D}}{=} h(X_n), h(X^*) \stackrel{\mathscr{D}}{=} h(X)$, 可见 $h(X_n) \xrightarrow{\mathscr{D}} h(X)$. □

4. 完成 §7 中 (13) 式的证明 在 §7 证明定理 1 的蕴涵关系 ($\Leftarrow$) 时, 曾经用到性质 (13). 现在进行 (13) 式的证明, 仍然借助 "一个概率空间的方法".

设 $(E, \mathscr{E}, \rho)$ 是可分度量空间, $\mathscr{G}$ 是等阶梯连续函数 $g = g(x)$ 类, 满足性质: 对于一切 $x \in E, g \in \mathscr{G}$ 和某个常数 C, 有 $|g(x)| \leqslant C$.

定理 3 设 $P, P_n (n \geqslant 1)$ 是 $(E, \mathscr{E}, \rho)$ 上的概率测度, 且 $P_n \xrightarrow{w} P$, 那么

$$\sup_{g \in \mathscr{G}} \left| \int_E g(x) P_n(\mathrm{d}x) - \int_E g(x) P(\mathrm{d}x) \right| \to 0, \quad n \to \infty. \tag{7}$$

证明 设 (7) 式成立, 则存在 $a > 0$ 和函数 $g_1, g_2, \cdots \in \mathscr{G}$, 使对于无限多个 n, 有

$$\left| \int_E g_n(x) P_n(\mathrm{d}x) - \int_E g_n(x) P(\mathrm{d}x) \right| \geqslant a > 0. \tag{8}$$

将 "一个概率空间的方法" 用于随机元 X^* 和 X_n^* (见定理 1), 可以将 (8) 式写成: 对于无限多个 n,

$$|\mathrm{E}^* g_n(X_n^*) - \mathrm{E}^* g_n(X^*)| \geqslant a > 0. \tag{9}$$

由于类 $\mathscr{G}$ 的性质, 对于任意 $\varepsilon > 0$, 存在 $\delta > 0$, 当 $\rho(x, y) < \delta$ 时, 使对于一切 $g \in \mathscr{G}$, 有 $|g(y) - g(x)| < \varepsilon$. 此外, 对于一切 $x \in E, g \in \mathscr{G}$, 有 $|g(x)| \leqslant C$. 所以

$$\begin{aligned} &|\mathbf{E}^* g_n(X_n^*) - \mathbf{E}^* g_n(X^*)| \\ &\leqslant \mathbf{E}^*\{|g_n(X_n^*) - g_n(X^*)|; \rho(X_n^*, X^*) > \delta\} \\ &\quad + \mathbf{E}^*\{|g_n(X_n^*) - g_n(X^*)|; \rho(X_n^*, X^*) \leqslant \delta\} \\ &\leqslant 2C\mathbf{P}^*\{\rho(X_n^*, X^*) > \delta\} + \varepsilon. \end{aligned}$$

由于 $X_n^* \xrightarrow{\text{a.s.}} X$, 故 $\mathbf{P}^*\{\rho(X_n^*, X^*) > \delta\} \to 0, n \to \infty$. 于是, 由于 $\varepsilon > 0$ 任意性, 可见

$$\lim_n |\mathbf{E}^* g_n(X_n^*) - \mathbf{E}^* g_n(X^*)| = 0,$$

而这与 (9) 式矛盾. □

5. 列维 – 普罗霍罗夫度量值的上估计 在这一小节把用于定理 1 的 "一个概率空间的方法" 的思想, 用来从上侧估计可分度量空间 $(E, \mathscr{E}, \rho)$ 上两个概率分布之间的列维 - 普罗霍罗夫度量的值 $L(P, \tilde{P})$.

定理 4 *对于任意两个测度 P 和 $\tilde{P}$, 存在概率空间 $(\Omega^*, \mathscr{F}^*, \mathbf{P}^*)$, 以及定义在该空间上、取值于 E 的随机元 X 和 $\tilde{X}$, 而且 X 和 $\tilde{X}$ 的分布恰好分别为 P 和 $\tilde{P}$, 并且满足:*

$$L(P, \tilde{P}) \leqslant d_{\mathbf{P}^*}(X, \tilde{X}) = \inf\{\varepsilon > 0; \mathbf{P}^*\{\rho(X, \tilde{X}) \geqslant \varepsilon\} \leqslant \varepsilon\}. \tag{10}$$

证明 根据定理 1, 确实存在概率空间 $(\Omega^*, \mathscr{F}^*, \mathbf{P}^*)$ 定义在该空间上、取值于 E 的随机元 X 和 $\tilde{X}$, 使 $\mathbf{P}^*\{X \in A\} = P(A), \mathbf{P}^*\{\tilde{X} \in A\} = \tilde{P}(A), A \in \mathscr{E}$.

设 $\varepsilon > 0$, 使

$$\mathbf{P}^*\{\rho(X, \tilde{X}) \geqslant \varepsilon\} \leqslant \varepsilon. \tag{11}$$

那么, 对于任意 $A \in \mathscr{E}$,

$$\begin{aligned} \tilde{P}(A) = \mathbf{P}^*\{\tilde{X} \in A\} &= \mathbf{P}^*\{\tilde{X} \in A, X \in A^\varepsilon\} + \mathbf{P}^*\{\tilde{X} \in A, X \notin A^\varepsilon\} \\ &\leqslant \mathbf{P}^*\{X \in A^\varepsilon\} + \mathbf{P}^*\{\rho(X, \tilde{X}) \geqslant \varepsilon\} \leqslant P(A^\varepsilon) + \varepsilon, \end{aligned}$$

其中 $A^\varepsilon = \{x \in E : \rho(x, A) < \varepsilon\}$. 因此, 根据列维 – 普罗霍罗夫度量的定义 (§7 第 3 小节):

$$L(P, \tilde{P}) \leqslant \varepsilon. \tag{12}$$

由 (11) 和 (12) 式, 并对 $\varepsilon > 0$ 求 inf, 得所需要求的 (10) 式. □

系 *设 X 和 $\tilde{X}$ 是概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 上、取值于 E 的随机元, 其概率分布为 P_X 和 $P_{\tilde{X}}$, 那么,*

$$L(P_X, P_{\tilde{X}}) \leqslant d_{\mathbf{P}}(X, \tilde{X}).$$

注 1 所作的证明显示, 事实上 (10) 式总是正确的, 只要在某一概率空间 $(\Omega^*, \mathscr{F}^*, \mathbf{P}^*)$ 上可以指出取值于 E 的随机元 X 和 $\tilde{X}$, 使其概率分布相应为 P 和 $\tilde{P}$, 而且对于 X 和 $\tilde{X}$, 有 $\{\omega^* : \rho(X(\omega^*), \tilde{X}(\omega^*)) \geqslant \varepsilon\} \in \mathscr{F}^*, \varepsilon > 0$. 因而, 估计式 (10) 的质量本质上依赖于, 根据测度 P 和 $\tilde{P}$ 建立的对象 $(\Omega^*, \mathscr{F}^*, \mathbf{P}^*)$ 和 $X, \tilde{X}$ 如何. (由 P 和 $\tilde{P}$ 建立 $\Omega^*, \mathscr{F}^*, \mathbf{P}^*$ 和 $X, \tilde{X}$ 的过程, 称做联结或匹配 —— 源于英语单词 coupling.) 例如, 可以取 "联结 – 测度" $\mathbf{P}^*$ 等于测度 P 和 $\tilde{P}$ 的直积: $\mathbf{P}^* = P \times \tilde{P}$, 然而这样的选择通常不会得到 (10) 式的好估计.

注 2 自然提出问题, (10) 式何时为等式. 为此, 我们 (不加证明) 给出如下结果: *设 P 和 $\tilde{P}$ 是可分度量空间 $(E, \mathscr{E}, \rho)$ 上的两个概率测度, 存在这样的 $(\Omega^*, \mathscr{F}^*, \mathbf{P}^*)$ 和 $X, \tilde{X}$, 使*

$$L(P, \tilde{P}) = d_{\mathbf{P}^*}(X, \tilde{X}) = \inf\{\varepsilon > 0 : \mathbf{P}^*\{\rho(X, \tilde{X}) \geqslant \varepsilon\} \leqslant \varepsilon\}.$$

6. 练习题

1. 假设概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 是某一可分度量空间, 而 $X(\omega)$ 和 $Y(\omega)$ 是定义在该空间上的任意随机元, 证明实函数 $\rho(X(\omega), Y(\omega))$ 是随机变量.

2. 证明由 (2) 式定义的函数 $d_{\mathbf{P}}(X, Y)$, 是取值于 E 的随机元空间中的度量.

3. 证明 (5) 式的正确性.

4. 证明集合 $\Delta_h = \{x \in E : h(x)$在点$x$关于距离$\rho$不连续$\} \in \mathscr{E}$.

§9. 概率测度之间的变差距离. 角谷 – 海林格距离和海林格积分. 对测度的绝对连续性和奇异性的应用

1. 概率测度间的变差距离 设 $(\Omega, \mathscr{F})$ 是可测空间, $\mathscr{P} = \{P\}$ 是空间 $(\Omega, \mathscr{F})$ 上的概率测度族.

定义 1 设 P 和 $\tilde{P}$ 为 $\mathscr{P}$ 中的两个测度, 称变量

$$\operatorname{var}(P - \tilde{P}) \equiv \sup_{\varphi} \left| \int_{\Omega} \varphi(\omega) \mathrm{d}(P - \tilde{P}) \right| \tag{1}$$

为 (带符号) 测度 $P - \tilde{P}$ 的全变差, 其中 sup 在一切满足 $|\varphi(\omega)| \leqslant 1$ 的 $\mathscr{F}$ – 可测函数 $\varphi(\omega)$ 类上来求. 称 $\operatorname{var}(P - \tilde{P})$ 为 $\mathscr{P}$ 中测度 P 和 $\tilde{P}$ 之间的*变差距离*, 记作 $\|P - \tilde{P}\|$.

引理 1 *变差距离*

$$\|P - \tilde{P}\| = 2 \sup_{A \in \mathscr{F}} |P(A) - \tilde{P}(A)|. \tag{2}$$

证明 由于对于任意 $A \in \mathscr{F}$,

$$P(A) - \tilde{P}(A) = \tilde{P}(\bar{A}) - P(\bar{A}),$$

则

$$2|P(A)-\tilde{P}(A)| = |P(A)-\tilde{P}(A)| + |P(\overline{A})-\tilde{P}(\overline{A})| \leqslant \|P-\tilde{P}\|,$$

其中由 (1) 式得最后一个不等式.

为证明相反的不等式, 借助于带符号测度 $\mu = P - \tilde{P}$ 的哈恩 (H. Hahn) 分解 (例如, 见 [34, 第六章 §5] 或 [70 第 121 页]). 根据这一分解, 测度 μ 表示为 $\mu = \mu_+ - \mu_-$, 其中 μ_+ 和 μ_- (测度 μ 的上变差 μ_+ 和下变差 μ_-) 具有如下形式:

$$\mu_+(A) = \int_{A\cap M} \mathrm{d}\mu, \quad \mu_-(A) = -\int_{A\cap \overline{M}} \mathrm{d}\mu, \quad A \in \mathscr{F},$$

其中 M 是 $\mathscr{F}$ 中的集合. 这时

$$\operatorname{var}\mu = \operatorname{var}\mu_+ + \operatorname{var}\mu_- = \mu_+(\Omega) + \mu_-(\Omega).$$

由于

$$\mu_+(\Omega) = P(M) - \tilde{P}(M), \quad \mu_-(\Omega) = \tilde{P}(\overline{M}) - P(\overline{M}),$$

可见

$$\|P - \tilde{P}\| = [P(M) - \tilde{P}(M)] + [\tilde{P}(\overline{M}) - P(\overline{M})] \leqslant 2\sup_{A\in\mathscr{F}} |P(A) - \tilde{P}(A)|.$$

定义 2 称概率测度序列 $\{P_n\}_{n\geqslant 1}$ 按变差收敛于测度 P, 记为 $P_n \xrightarrow{\text{var}} P$, 如果

$$\|P_n - P\| \to 0, \quad n \to \infty. \tag{3}$$

由这一定义和 §1 的定理 1, 不难看出, 如果定义在度量空间 $(\Omega, \mathscr{F}, \rho)$ 上的概率测度按变差收敛, 则必定也弱收敛.

分布按变差的接近程度, 看来是概率分布接近程度最强的形式. 因为, 若两个分布的变差接近, 则实际上在具体的情形下可以认为它们是无区别的. 因此, 可能造成一种印象, 似乎研究变差距离没有太大的概率意义. 然而, 例如在泊松定理中 (第一章 §6), 二项分布向泊松分布的收敛, 就是按变差收敛于 0 (下面 §12 将给出这个距离的上估计.)

下面将要举出数理统计中用到两个测度 P 和 $\tilde{P}$ 间变差距离的例子. 这种情形自然地出现在根据观测结果区分两个统计假设 H 和 $\tilde{H}$ 的问题中. 考虑两个假设, H: 真实的分布为 P; $\tilde{H}$: 真实的分布为 $\tilde{P}$. 要求判断两个概率模型: $(\Omega, \mathscr{F}, \mathrm{P})$ 或 $(\Omega, \mathscr{F}, \tilde{\mathrm{P}})$, 哪一个更符合对观测结果的 "统计". 如果把 $\omega \in \Omega$ 解释为观测结果, 则认为取值于 $[0,1]$ 的任意 $\mathscr{F}$ 可测函数 $\varphi = \varphi(\omega)$ 是 (区分假设 H 和 $\tilde{H}$ 的) 准则. 准则的统计意义是, $\varphi(\omega)$ 是"当观测结果为 ω 时, 接受假设 $\tilde{H}$ 的概率".

我们用第一类错误概率和第二类错误概率表征区分假设 H 和 $\tilde{H}$ 的准则的质量:

$\alpha(\varphi) = E\varphi(\omega)$ (="当 H 真实时, 接受 $\tilde{H}$ 的概率"),

$\beta(\varphi) = \tilde{E}[1 - \varphi(\omega)]$ (="当 $\tilde{H}$ 真实时, 接受 H 的概率"),

其中 E 和 $\tilde{E}$ 表示按测度 P 和 $\tilde{P}$ 求平均. 假如假设 H 和 $\tilde{H}$ 是对等的, 自然认为使误差概率之和 $\alpha(\varphi) + \beta(\varphi)$ 最小的准则 $\varphi^* = \varphi^*(\omega)$ 是最优的 (只要这样的准则存在).

设

$$\mathscr{E}_r(P, \tilde{P}) = \inf_{\varphi}[\alpha(\varphi) + \beta(\varphi)]. \tag{4}$$

记

$$Q = \frac{1}{2}(P + \tilde{P}) \text{ 和 } z = \frac{\mathrm{d}P}{\mathrm{d}Q}, \ \tilde{z} = \frac{\mathrm{d}\tilde{P}}{\mathrm{d}Q},$$

则

$$\mathscr{E}_r(P, \tilde{P}) = \inf_{\varphi}[E\varphi + \tilde{E}(1-\varphi)] = \inf_{\varphi} E_Q[z\varphi + \tilde{z}(1-\varphi)] = 1 + \inf_{\varphi} E_Q[\varphi(z - \tilde{z})]$$

(这里及以后用 E_Q 表示对测度 Q 的数学期望).

不难看到, 在函数

$$\varphi^*(\omega) = I\{\tilde{z} < z\}$$

上达到 inf, 且因为 $E_Q(z - \tilde{z}) = 0$, 所以

$$\mathscr{E}_r(P, \tilde{P}) = 1 - \frac{1}{2}E_Q|z - \tilde{z}| = 1 - \frac{1}{2}\|P - \tilde{P}\|, \tag{5}$$

其中由下面的引理 2, 可得最后一个等式. 于是, 由 (5) 式可见, 函数 $\mathscr{E}_r(P, \tilde{P})$ 表征区分假设的质量, 而 $\mathscr{E}_r(P, \tilde{P})$ 依赖于由变差距离表征的测度 P 和 $\tilde{P}$ 的接近程度.

引理 2 *设 Q 是某一 σ 有限测度, 满足 $P \ll Q$, $\tilde{P} \ll Q$, 而测度 P 和 $\tilde{P}$ 关于 Q 的拉东 — 尼科迪姆导数记为*

$$z = \frac{\mathrm{d}P}{\mathrm{d}Q}, \quad \tilde{z} = \frac{\mathrm{d}\tilde{P}}{\mathrm{d}Q},$$

那么,

$$\|P - \tilde{P}\| = E_Q|z - \tilde{z}|, \tag{6}$$

而且, 如果 $Q = (P + \tilde{P})/2$, 则

$$\|P - \tilde{P}\| = E_Q|z - \tilde{z}| = 2E_Q|1 - z| = 2E_Q|1 - \tilde{z}|. \tag{7}$$

证明 对于一切 $\mathscr{F}$ 可测函数 $\psi = \psi(\omega)$, $|\psi(\omega)| \leqslant 1$, 根据 z 和 $\tilde{z}$ 的定义, 有

$$|E\psi - \tilde{E}\psi| = |E_Q\psi(z - \tilde{z})| \leqslant E_Q|\psi||z - \tilde{z}| \leqslant E_Q|z - \tilde{z}|. \tag{8}$$

因此

$$\|P - \tilde{P}\| \leqslant E_Q|z - \tilde{z}|, \tag{9}$$

但是, 对于函数

$$\psi = \operatorname{sign}(\tilde{z} - z) = \begin{cases} 1, & \tilde{z} \geqslant z, \\ -1, & \tilde{z} < z, \end{cases}$$

有

$$|E\psi - \tilde{E}\psi| = E_Q|(z - \tilde{z})|. \tag{10}$$

由 (9) 和 (10) 式得所求的 (8) 式. 由于 $z + \tilde{z} = 2(Q\text{-a.e.})$, 由 (6) 式得 (7) 式. □

系 1 设 P 和 $\tilde{P}$ 是 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上的两个概率分布, 其 (关于勒贝格测度 dx 的) 概率密度为 $p(x)$ 和 $\tilde{p}(x), x \in \mathbb{R}$, 则

$$\|P - \tilde{P}\| = \int_{-\infty}^{\infty} |p(x) - \tilde{p}(x)| dx. \tag{11}$$

(作为 Q 应取 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上的勒贝格测度.)

系 2 设 P 和 $\tilde{P}$ 是两个集中在可数个点 $x_1, x_2, \cdots$ 上的离散测度: $P = \{p_1, p_2, \cdots\}$, $\tilde{P} = \{\tilde{p}_1, \tilde{p}_2, \cdots\}$, 则

$$\|P - \tilde{P}\| = \sum_{i=1}^{\infty} |p_i - \tilde{p}_i|. \tag{12}$$

(作为 Q 应取计数测度, 即 $Q(\{x_i\}) = 1, i = 1, 2, \cdots$)

2. 测度间的角谷 – 海林格 (S. Kakutani–E. Hellinger) 距离 我们现在再考虑一种概率测度接近程度的度量, 它在很大程度上与变差测度的接近程度是同种的.

设 P 和 $\tilde{P}$ 是 $(\Omega, \mathscr{F})$ 上的两个概率测度, 而 Q 是控制 P 和 $\tilde{P}$ 的另一个概率测度[①], 即 $P \ll Q, \tilde{P} \ll Q$. 仍记

$$z = \frac{dP}{dQ}, \quad \tilde{z} = \frac{d\tilde{P}}{dQ}.$$

定义 3 称非负量 $\rho(P, \tilde{P})$ 为测度 P 和 $\tilde{P}$ 之间的角谷 – 海林格距离, 如果

$$\rho^2(P, \tilde{P}) = \frac{1}{2} E_Q(\sqrt{z} - \sqrt{\tilde{z}})^2. \tag{13}$$

由于

$$E_Q(\sqrt{z} - \sqrt{\tilde{z}})^2 = \int_\Omega \left(\sqrt{\frac{dP}{dQ}} - \sqrt{\frac{d\tilde{P}}{dQ}}\right)^2 dQ, \tag{14}$$

① 一般称 Q (相对 P 和 $\tilde{P}$) 为优测度或强测度. ——译者

则量 $\rho^2(P, \tilde{P})$ 的书写形式

$$\rho^2(P, \tilde{P}) = \frac{1}{2} \int_\Omega (\sqrt{dP} - \sqrt{d\tilde{P}})^2 \tag{15}$$

就变得清晰明了.

如果设

$$H(P, \tilde{P}) = E_Q\sqrt{z\tilde{z}}, \tag{16}$$

则仿照 (15) 式, 可以形式地写为

$$H(P, \tilde{P}) = \int_\Omega \sqrt{dP\, d\tilde{P}}. \tag{17}$$

由 (13) 和 (16) 式, 以及由 (15) 和 (17) 式, 可见

$$\rho^2(P, \tilde{P}) = 1 - H(P, \tilde{P}). \tag{18}$$

量 $H(P, \tilde{P})$ 称做 1/2 阶海林格积分. 对于许多应用, 考虑 $\alpha \in (0, 1)$ 阶海林格积分 $H(\alpha; P, \tilde{P})$ 是有益的, 它由下面的公式定义:

$$H(\alpha; P, \tilde{P}) = E_Q z^\alpha \tilde{z}^{1-\alpha}, \tag{19}$$

或形式地写为

$$H(\alpha; P, \tilde{P}) = \int_\Omega (dP)^\alpha (d\tilde{P})^{1-\alpha}. \tag{20}$$

显然 $H(1/2; P, \tilde{P}) = H(P, \tilde{P})$.

为保障定义 3 是适定的, 需要证明量 $\rho^2(P, \tilde{P})$ 与控制测度的选择无关, 且 $\rho(P, \tilde{P})$ 事实上满足对 "距离" 的要求.

引理 3 1. $\alpha \in (0, 1)$ 阶海林格积分 $H(\alpha; P, \tilde{P})$ (从而 $\rho(P, \tilde{P})$) 与控制测度 Q 的选择无关.

2. 由 (13) 式定义的函数 ρ, 在概率测度的集合上是度量.

证明 1. 假如 Q' 控制 P 和 $\tilde{P}$ (即 $P \ll Q', \tilde{P} \ll Q'$), 则 Q' 也控制测度 $Q = (P + \tilde{P})/2$. 因此, 只需证明, 若 $Q \ll Q'$, 则

$$E_Q(z^\alpha \tilde{z}^{1-\alpha}) = E_{Q'}(z')^\alpha (\tilde{z}')^{1-\alpha},$$

其中

$$z' = \frac{dP}{dQ'}, \quad \tilde{z}' = \frac{d\tilde{P}}{dQ'}.$$

记 $v = dQ/dQ'$, 那么 $z' = zv, \tilde{z}' = \tilde{z}v$, 且

$$E_Q(z^\alpha \tilde{z}^{1-\alpha}) = E_{Q'}(vz^\alpha \tilde{z}^{1-\alpha}) = E_{Q'}(z')^\alpha (\tilde{z}')^{1-\alpha}.$$

于是, 第一个命题得证.

2. 如果 $\rho(P,\tilde{P})=0$, 则 $z=\tilde{z}$ (Q - a. c.), 因此 $P=\tilde{P}$. 显然, 由对称性可见 $\rho(P,\tilde{P})=\rho(\tilde{P},P)$. 最后, 设 P, P', P'' 是 3 个测度, 满足 $P\ll Q, P'\ll Q, P''\ll Q$, 且

$$z=\frac{dP}{dQ},\quad z'=\frac{dP'}{dQ},\quad z''=\frac{dP''}{dQ}.$$

利用对于 $L^2(\Omega,\mathscr{F},Q)$ 中的范数成立三角形不等式, 得

$$[E_Q(\sqrt{z}-\sqrt{z''})^2]^{1/2}\leqslant[E_Q(\sqrt{z}-\sqrt{z'})^2]^{1/2}+[E_Q(\sqrt{z'}-\sqrt{z''})^2]^{1/2},$$

即

$$\rho(P,P'')\leqslant\rho(P,P')+\rho(P',P''). \qquad \square$$

由定义 (19) 和傅比尼定理 (第二章 §6) 可以直接推出, 当测度 P 和 $\tilde{P}$ 是测度的直积时 (第二章 §6 第 10 小节): $P=P_1\times\cdots\times P_n$, $\tilde{P}=\tilde{P}_1\times\cdots\times\tilde{P}_n$, 则测度 P 和 $\tilde{P}$ 间的海林格积分等于相应积分的积:

$$H(\alpha;P,\tilde{P})=\prod_{i=1}^{n}H(\alpha;P_i,\tilde{P}_i).$$

下面的定理揭示了变差距离与角谷 – 海林格距离 (或等价地, 海林格积分) 间的联系. 特别, 它表明这些距离决定概率测度空间在 $(\Omega,\mathscr{F})$ 上的同一拓扑.

定理 1　下列不等式成立:

$$2[1-H(P,\tilde{P})]\leqslant\|P-\tilde{P}\|\leqslant\sqrt{8[1-H(P,\tilde{P})]}, \tag{21}$$

$$\|P-\tilde{P}\|\leqslant2\sqrt{1-H^2(P,\tilde{P})}. \tag{22}$$

特别,

$$2\rho^2(P,\tilde{P})\leqslant\|P-\tilde{P}\|\leqslant\sqrt{8}\rho(P,\tilde{P}). \tag{23}$$

证明　由于 $H(P,\tilde{P})\leqslant1$, 而对于 $0\leqslant x\leqslant1$, $1-x^2\leqslant2(1-x)$, 则由 (22) 式得 (21) 式的第二个不等式, 而 (22) 式由下面的一系列不等式 (其中 $Q=(P+\tilde{P})/2$) 可得:

$$\begin{aligned}\frac{1}{2}\|P-\tilde{P}\|&=E_Q|1-z|\leqslant\sqrt{E_Q|1-z|^2}=\sqrt{1-E_Qz(2-z)}=\sqrt{1-E_Qz\tilde{z}}\\&=\sqrt{1-E_Q(\sqrt{z\tilde{z}})^2}\leqslant\sqrt{1-(E_Q\sqrt{z\tilde{z}})^2}=\sqrt{1-H^2(P,\tilde{P})}.\end{aligned}$$

最后, 由不等式

$$\frac{1}{2}[\sqrt{z}-\sqrt{2-z}]^2\leqslant|z-1|,\quad z\in[0,2],$$

有 (其中 $Q=(P+\tilde{P})/2$),

$$1-H(P,\tilde{P})=\rho^2(P,\tilde{P})=\frac{1}{2}E_Q[\sqrt{z}-\sqrt{2-z}]^2\leqslant E_Q|z-1|=\frac{1}{2}\|P-\tilde{P}\|.$$

由此可得 (21) 式的第一个不等式. $\square$

注　类似地可以证明, 对于任意 $\alpha\in(0,1)$, 有

$$2[1-H(\alpha;P,\tilde{P})]\leqslant\|P-\tilde{P}\|\leqslant\sqrt{c_\alpha[1-H(\alpha;P,\tilde{P})]}, \tag{24}$$

其中 c_α 是某一常数.

系 3　设 P 和 P^n, $n\geqslant1$, 是 $(\Omega,\mathscr{F})$ 上的概率测度. 那么, 当 $n\to\infty$ 时, 有

$$\|P^n-P\|\to0\Leftrightarrow H(P^n,P)\to1\Leftrightarrow\rho(P^n,P)\to0,$$

$$\|P^n-P\|\to2\Leftrightarrow H(P^n,P)\to0\Leftrightarrow\rho(P^n,P)\to1.$$

系 4　由于根据 (5)

$$\mathscr{E}r(P,\tilde{P})=1-\frac{1}{2}\|P-\tilde{P}\|,$$

则由 (21) 式和 (22) 式, 有

$$\frac{1}{2}H^2(P,\tilde{P})\leqslant1-\sqrt{1-H^2(P,\tilde{P})}\leqslant\mathscr{E}r(P,\tilde{P})\leqslant H(P,\tilde{P}). \tag{25}$$

特别, 设

$$P^n=\underbrace{P\times\cdots\times P}_{n},\quad\tilde{P}^n=\underbrace{\tilde{P}\times\cdots\times\tilde{P}}_{n}$$

相应为 n 个同样测度的直积. 那么, 由于

$$H(P^n,\tilde{P}^n)=[H(P,\tilde{P})]^n=e^{-\lambda n},\quad\lambda=-\ln H(P,\tilde{P})\geqslant\rho^2(P,\tilde{P}),$$

则由 (25) 式, 得

$$\frac{1}{2}e^{-2\lambda n}\leqslant\mathscr{E}r(P^n,\tilde{P}^n)\leqslant e^{-\lambda n}\leqslant e^{-n\rho^2(P,\tilde{P})}. \tag{26}$$

用于上面考虑的区分两个统计假设的问题, 由这些不等式可得如下结果.

设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机元, 其概率分布为 P (假设 H) 或 $\tilde{P}$ (假设 $\tilde{H}$), 且 $P\neq\tilde{P}$, 从而 $\rho^2(P,\tilde{P})>0$. 那么, 表征根据观测结果 $\xi_1,\cdots,\xi_n$ 最优区分假设 H 和 $\tilde{H}$ 质量的函数 $\mathscr{E}r(P^n,\tilde{P}^n)$, 当 $n\to\infty$ 时以指数的速度递降为 0.

3. 海林格积分对测度的绝对连续性和奇异性的应用 利用上面引进的 α 阶海林格积分, 容易表述概率测度绝对连续性和奇异性的条件.

设 P 和 $\tilde{P}$ 是可测空间 $(\Omega, \mathscr{F})$ 上的两个概率测度. 回忆, 测度 $\tilde{P}$ 关于测度 P 绝对连续 (记作 $\tilde{P} \ll P$), 如果对于 $A \in \mathscr{F}$, 每当 $P(A) = 0$ 时必有 $\tilde{P}(A) = 0$. 如果 $\tilde{P} \ll P$ 且 $P \ll \tilde{P}$, 则称 P 和 $\tilde{P}$ 等价 ($\tilde{P} \sim P$). 测度 P 和 $\tilde{P}$ 称做奇异的或正交的 ($\tilde{P} \perp P$), 如果存在 $A \in \mathscr{F}$, 使 $P(A) = 1$ 和 $\tilde{P}(\bar{A}) = 1$ (即测度 P 和 $\tilde{P}$ "处于" 不同的集合中).

设 Q 是概率测度, 而 $P \ll Q, \tilde{P} \ll Q$,

$$z = \frac{\mathrm{d}P}{\mathrm{d}Q}, \quad \tilde{z} = \frac{\mathrm{d}\tilde{P}}{\mathrm{d}Q}.$$

定理 2 下列条件等价:

(a) $\tilde{P} \ll P$;

(b) $\tilde{P}\{z > 0\} = 1$;

(c) $H(\alpha; P, \tilde{P}) \to 1, \alpha \downarrow 0$.

定理 3 下列条件等价:

(a) $\tilde{P} \perp P$;

(b) $\tilde{P}\{z > 0\} = 0$;

(c) $H(\alpha; P, \tilde{P}) \to 0, \alpha \downarrow 0$;

(d) $H(\alpha; P, \tilde{P}) = 0$, 对于一切 $\alpha \in (0, 1)$;

(e) $H(\alpha; P, \tilde{P}) = 0$, 对于某个 $\alpha \in (0, 1)$.

证明 两个定理的证明同时进行. 根据 z 和 $\tilde{z}$ 的定义

$$P\{z = 0\} = E_Q[zI(z = 0)] = 0, \tag{27}$$

$$\begin{aligned} \tilde{P}(A \cap \{z > 0\}) &= E_Q[\tilde{z}I(A \cap \{z > 0\})] = E_Q\left[z\frac{\tilde{z}}{z}I(A \cap \{z > 0\})\right] \\ &= E\left[\frac{\tilde{z}}{z}I(A \cap \{z > 0\})\right] = E\left[\frac{\tilde{z}}{z}I(A)\right]. \end{aligned} \tag{28}$$

从而, "勒贝格分解" 成立:

$$\tilde{P}(A) = E\left[\frac{\tilde{z}}{z}I(A)\right] + \tilde{P}(A \cap \{z = 0\}), \quad A \in \mathscr{F}, \tag{29}$$

其中随机变量 $Z = \tilde{z}/z$ 称做测度 $\tilde{P}$ (的 "绝对连续分量") 关于测度 P 的勒贝格导数, 并且记作 $\mathrm{d}\tilde{P}/\mathrm{d}P$ (对照第二章 §6 拉东 - 尼科迪姆定理的注).

由此立即得出两个定理中 (a) 和 (b) 的等价性.

此外, 由于

$$z^{\alpha}\tilde{z}^{1-\alpha} \to \tilde{z}I(z > 0), \quad \alpha \downarrow 0,$$

而对于 $\alpha \in (0, 1)$

$$0 \leqslant z^{\alpha}\tilde{z}^{1-\alpha} \leqslant \alpha z + (1 - \alpha)\tilde{z} \leqslant z + \tilde{z},$$

且 $E_Q(z + \tilde{z}) = 2$. 故根据勒贝格控制收敛定理

$$\lim_{\alpha \downarrow 0} H(\alpha; P, \tilde{P}) = E_Q \tilde{z} I(z > 0) = \tilde{P}(z > 0).$$

而这说明两个定理中 (b) $\Leftrightarrow$ (c).

最后证明, 在第二个定理中 (c) $\Leftrightarrow$ (d) $\Leftrightarrow$ (e). 为此只需注意到

$$H(\alpha; P, \tilde{P}) = \tilde{E}\left(\frac{z}{\tilde{z}}\right)^{\alpha} I(\tilde{z} > 0), \quad \tilde{P}(\tilde{z} > 0) = 1.$$

这说明, 对于每一个 $\alpha \in (0, 1)$, $\tilde{P}\{z > 0\} = 0 \Leftrightarrow H(\alpha; P, \tilde{P}) = 0$, 由此可得蕴涵关系 (c) $\Leftrightarrow$ (d) $\Leftrightarrow$ (e). □

例 1 设 $P = P_1 \times P_2 \times \cdots, \tilde{P} = \tilde{P}_1 \times \tilde{P}_2 \times \cdots$, 其中 P_k 和 $\tilde{P}_k$ 是 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上的高斯测度, 其密度为

$$p_k(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x - a_k)^2}{2}}, \quad \tilde{p}_k(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x - \tilde{a}_k)^2}{2}}.$$

由于通过简单的计算, 得

$$H(\alpha; P, \tilde{P}) = \prod_{k=1}^{\infty} H(\alpha; P_k, \tilde{P}_k),$$

其中

$$H(\alpha; P_k, \tilde{P}_k) = \int_{-\infty}^{\infty} p_k^{\alpha}(x)\tilde{p}_k^{1-\alpha}(x)\mathrm{d}x = e^{-\frac{\alpha(1-\alpha)}{2}(a_k - \tilde{a}_k)^2},$$

得

$$H(\alpha; P, \tilde{P}) = e^{-\frac{\alpha(1-\alpha)}{2}\sum_{k=1}^{\infty}(a_k - \tilde{a}_k)^2}.$$

由定理 2 和 3 得

$$\tilde{P} \ll P \Leftrightarrow P \ll \tilde{P} \Leftrightarrow \tilde{P} \sim P \Leftrightarrow \sum_{k=1}^{\infty}(a_k - \tilde{a}_k)^2 < \infty,$$

$$\tilde{P} \perp P \Leftrightarrow \sum_{k=1}^{\infty}(a_k - \tilde{a}_k)^2 = \infty.$$

例 2 仍设 $P = P_1 \times P_2 \times \cdots, \tilde{P} = \tilde{P}_1 \times \tilde{P}_2 \times \cdots$, 其中 P_k 和 $\tilde{P}_k$ 是参数相应为 $\lambda_k > 0$ 和 $\tilde{\lambda}_k > 0$ 的泊松分布. 那么, 不难证明

$$\begin{aligned} \tilde{P} \ll P \Leftrightarrow P \ll \tilde{P} \Leftrightarrow \tilde{P} \sim P \Leftrightarrow \sum_{k=1}^{\infty}\left(\sqrt{\lambda_k} - \sqrt{\tilde{\lambda}_k}\right)^2 < \infty, \\ \tilde{P} \perp P \Leftrightarrow \sum_{k=1}^{\infty}\left(\sqrt{\lambda_k} - \sqrt{\tilde{\lambda}_k}\right)^2 = \infty. \end{aligned} \tag{30}$$

4. 练习题

1. 在引理 3 的记号下, 设

$$P \wedge \tilde{P} = E_Q(z \wedge \tilde{z}),$$

其中 $z \wedge \tilde{z} = \min(z, \tilde{z})$. 证明

$$\|P - \tilde{P}\| = 2(1 - P \wedge \tilde{P})$$

(从而 $\mathscr{E}_\rho(P, \tilde{P}) = P \wedge \tilde{P}$).

2. 设 $P, P_n, n \geqslant 1$, 是 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上的两个概率测度, 其 (关于勒贝格测度的) 密度为 $p(x), p_n(x), n \geqslant 1$. 假设对关于勒贝格测度几乎所有 x, $p_n(x) \to p(x)$, 证明

$$\|P - P_n\| = \int_{-\infty}^{\infty} |p(x) - p_n(x)| \mathrm{d}x \to 0, \quad n \to \infty$$

(对照第二章 §6 练习题 17).

3. 设 P 和 $\tilde{P}$ 是两个概率测度. 定义库尔贝克 (S. Kullback) 信息量 —— 相对 $\tilde{P}$ 有利于 P 的信息量 —— $K(P, \tilde{P})$, 定义为:

$$K(P, \tilde{P}) = \begin{cases} E \ln \dfrac{\mathrm{d}P}{\mathrm{d}\tilde{P}}, & 若\ P \ll \tilde{P}, \\ \infty, & 若不然. \end{cases}$$

证明

$$K(P, \tilde{P}) \geqslant -2\ln(1 - \rho^2(P, \tilde{P})) \geqslant 2\rho^2(P, \tilde{P}).$$

4. 证明公式 (11),(12).

5. 证明不等式 (24).

6. 设 $P, \tilde{P}, Q$ 是 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上的 3 个概率测度, 而 $P * Q$ 和 $\tilde{P} * Q$ 是它们的卷积 (第二章 §8 第 4 小节). 证明

$$\|P * Q - \tilde{P} * Q\| \leqslant \|P - \tilde{P}\|.$$

7. 证明 (30) 式.

8. 设 ξ 和 η 是在可测空间 $(E, \mathscr{E})$ 取值的, 概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 上的随机元, 证明

$$|\mathbf{P}\{\xi \in A\} - \mathbf{P}\{\eta \in A\}| \leqslant \mathbf{P}\{\xi \neq \eta\}, \quad A \in \mathscr{E}.$$

§10. 概率测度的临近性和完全渐近可区分性

1. 概率测度的临近性和完全渐近可区分性的概念 这些概念, 在数理统计的渐近理论中起重要作用, 对于两个测度序列的情形, 是两个测度绝对连续性和奇异性概念的自然推广.

我们从定义开始.

设 $(\Omega^n, \mathscr{F}^n)_{n \geqslant 1}$ 是某一可测空间序列, 而 $(P^n)_{n \geqslant 1}, (\tilde{P}^n)_{n \geqslant 1}$ 是概率测度序列, 其中 P^n 和 $\tilde{P}^n$ 定义在 $(\Omega^n, \mathscr{F}^n)_{n \geqslant 1}$ 上.

定义 1 称测度序列 $(\tilde{P}^n)$ 临近序列 (P^n), 记作 $(\tilde{P}^n) \lhd (P^n)$, 如果对于满足 $P^n(A^n) \to 0, n \to \infty$ 的所有 $A^n \in \mathscr{F}^n$, 有 $\tilde{P}^n(A^n) \to 0, n \to \infty$.

定义 2 称测度序列 $(\tilde{P}^n)$ 和 (P^n) 完全渐近可区分, 简称可区分, 记作 $(\tilde{P}^n) \bigtriangleup (P^n)$, 如果存在序列 $n_k \uparrow \infty, k \to \infty$ 与这样的集合 $A^{n_k} \in \mathscr{F}^{n_k}$, 使得:

$$P^{n_k}(A^{n_k}) \to 1, \quad \tilde{P}^{n_k}(A^{n_k}) \to 0, \quad k \to \infty.$$

立即可以指出, 可分性是一个对称的概念: $(\tilde{P}^n) \bigtriangleup (P^n) \Leftrightarrow (P^n) \bigtriangleup (\tilde{P}^n)$. 临近性不具备这一条性质. 假如 $(\tilde{P}^n) \lhd (P^n)$ 且 $(\tilde{P}^n) \rhd (P^n)$, 则记作 $(\tilde{P}^n) \lhd\rhd (P^n)$, 并且称该测度序列 $(\tilde{P}^n)$ 与 (P^n) 为相互临近的.

需要指出对于下面的情形: 对于一切 $n \geqslant 1$, $(\Omega^n, \mathscr{F}^n) = (\Omega, \mathscr{F})$, $P^n = P, \tilde{P}^n = \tilde{P}$, 有

$$(\tilde{P}^n) \lhd (P^n) \Leftrightarrow \tilde{P} \ll P, \tag{1}$$

$$(\tilde{P}^n) \lhd\rhd (P^n) \Leftrightarrow \tilde{P} \sim P, \tag{2}$$

$$(\tilde{P}^n) \bigtriangleup (P^n) \Leftrightarrow \tilde{P} \perp P. \tag{3}$$

这些性质以及上面给出的定义说明, 为什么临近性和完全渐近可分性, 常解释为序列 $(\tilde{P}^n)$ 和 (P^n) 的 "渐近绝对连续性" 和 "渐近奇异性".

2. 无限稠密随机变量序列的情形 下面给出的定理 1 和 2 是 §9 的定理 2 和 3 对测度序列情形的自然推广.

设 $(\Omega^n, \mathscr{F}^n)_{n \geqslant 1}$ 是可测空间序列, Q^n 是 $(\Omega^n, \mathscr{F}^n)(n \geqslant 1)$ 上的概率测度, 而 (ξ^n) 是 $(\Omega^n, \mathscr{F}^n)_{n \geqslant 1}$ 上的随机变量 (一般说来, 是广义随机变量; 见第二章 §4).

定义 3 随机变量序列 (ξ^n) 关于测度序列 Q^n 称做稠密的 (记作 $(\xi^n | Q^n)$ 稠密), 若

$$\lim_{N \uparrow \infty} \overline{\lim_{n}} Q^n(|\xi^n| > N) = 0. \tag{4}$$

(对照 §2 中概率测度族的密度的相应定义.)

下面到处将设

$$Q^n = \frac{P^n + \tilde{P}^n}{2}, \quad z^n = \frac{\mathrm{d}P^n}{\mathrm{d}Q^n}, \quad \tilde{z}^n = \frac{\mathrm{d}\tilde{P}^n}{\mathrm{d}Q^n}.$$

此外, 记

$$Z^n = \frac{\tilde{z}^n}{z^n} \tag{5}$$

为测度 $\tilde{P}^n$ 关于测度 P^n 的勒贝格导数, 并认为 $2/0=\infty$ (见 §9(29) 式). 注意, 假如 $\tilde{P}^n \ll P^n$, 则 Z^n 恰好是测度 $\tilde{P}^n$ 关于测度 P^n 的密度 $\mathrm{d}\tilde{P}^n/\mathrm{d}P^n$ 的一种变式 (第二章 §6).

对于以后, 有益地指出, 由于

$$P^n\left(z^n \leqslant \frac{1}{N}\right) = E_{Q^n}\left(z^n I\left(z^n \leqslant \frac{1}{N}\right)\right) \leqslant \frac{1}{N}, \tag{6}$$

且 $Z^n \leqslant 2/z^n$, 则

$$\left(\frac{1}{z^n}\middle| P^n\right) \text{ 稠密; } (Z^n|P^n) \text{ 稠密.} \tag{7}$$

定理 1 下列各条件等价:

(a) $(\tilde{P}^n) \lhd (P^n)$;

(b) $\left(\frac{1}{z^n}\middle|\tilde{P}^n\right)$ 稠密;

(b′) $(Z^n|\tilde{P}^n)$ 稠密;

(c) $\lim\limits_{\alpha\downarrow 0}\varliminf\limits_n H(\alpha; P^n, \tilde{P}^n) = 1$.

定理 2 下列各条件等价:

(a) $(\tilde{P}^n) \triangle (P^n)$;

(b) 对于任意 $\varepsilon > 0$, $\lim\limits_n \tilde{P}^n(z^n \geqslant \varepsilon) = 0$;

(b′) 对于任意 $N > 0$, $\varlimsup\limits_n \tilde{P}^n(Z^n \leqslant N) = 0$;

(c) $\lim\limits_{\alpha\downarrow 0}\varliminf\limits_n H(\alpha; P^n, \tilde{P}^n) = 0$;

(d) 对于任意 $\alpha \in (0,1)$, $\lim\limits_n H(\alpha; P^n, \tilde{P}^n) = 0$;

(e) 对于某个 $\alpha \in (0,1)$, $\varliminf\limits_n H(\alpha; P^n, \tilde{P}^n) = 0$.

定理 1 的证明.

(a)⇒(b). 假如 (b) 不成立, 则存在 $\varepsilon > 0$ 和这样的序列 $n_k \uparrow \infty$, 使 $\tilde{P}^{n_k}(z^{n_k} < 1/n_k) \geqslant \varepsilon$. 然而, 由 (6) 式, 有 $P^{n_k}(z^{n_k} < 1/n_k) \leqslant 1/n_k \to 0, k \to \infty$, 而这与假设 $(\tilde{P}^n) \lhd (P^n)$ 矛盾.

(b)⇔(b′). 只需注意到 $Z^n = (2/z^n) - 1$.

(b)⇒(a). 设 $A^n \in \mathscr{F}^n$ 且 $P^n(A^n) \to 0, n \to \infty$, 有

$$\begin{aligned}
\tilde{P}^n(A^n) &\leqslant \tilde{P}^n(z^n \leqslant \varepsilon) + E_{Q^n}(\tilde{z}^n I(A^n \cap \{z^n > \varepsilon\})) \\
&\leqslant \tilde{P}^n(z^n \leqslant \varepsilon) + \frac{2}{\varepsilon}E_{Q^n}(z^n I(A^n)) = \tilde{P}^n(z^n \leqslant \varepsilon) + \frac{2}{\varepsilon}P^n(A^n).
\end{aligned}$$

因此,

$$\varlimsup_n \tilde{P}^n(A^n) \leqslant \lim_n \tilde{P}^n(z^n \leqslant \varepsilon), \quad \varepsilon > 0.$$

命题 (b) 等价于

$$\lim_{\varepsilon\downarrow 0}\varlimsup_n \tilde{P}^n(z^n \leqslant \varepsilon) = 0.$$

于是, $\tilde{P}^n(A^n) \to 0$, 即 (b)⇒(a).

(b)⇒(c). 设 $\varepsilon > 0$, 则

$$\begin{aligned}
H(\alpha; P^n, \tilde{P}^n) &= E_{Q^n}[(z^n)^\alpha(\tilde{z}^n)^{1-\alpha}] \geqslant E_{Q^n}\left[\left(\frac{z^n}{\tilde{z}^n}\right)^\alpha I(z^n \geqslant \varepsilon)I(\tilde{z}^n > 0)\tilde{z}^n\right] \\
&= E_{\tilde{P}^n}\left[\left(\frac{z^n}{\tilde{z}^n}\right)^\alpha I(z^n \geqslant \varepsilon)\right] \geqslant \left(\frac{\varepsilon}{2}\right)^\alpha \tilde{P}^n(z^n \geqslant \varepsilon).
\end{aligned} \tag{8}$$

由于 $z^n + \tilde{z}^n = 2$, 可见对于 $\varepsilon > 0$,

$$\lim_{\alpha\downarrow 0}\varliminf_n H(\alpha; P^n, \tilde{P}^n) \geqslant \varliminf_{\alpha\downarrow 0}\left(\frac{\varepsilon}{2}\right)^\alpha \varliminf_n \tilde{P}^n(z^n \geqslant \varepsilon) = \varliminf_n \tilde{P}^n(z^n \geqslant \varepsilon). \tag{9}$$

由于 (b)$\varliminf\limits_{\varepsilon\downarrow 0}\varliminf\limits_n \tilde{P}^n(z^n \geqslant \varepsilon) = 1$. 因此, 由于 (9) 式和 $H(\alpha; P^n, \tilde{P}^n) \leqslant 1$, 可得命题 (c).

(c)⇒(b). 设 $\delta \in (0,1)$, 则

$$\begin{aligned}
H(\alpha; P^n, \tilde{P}^n) &= E_{Q^n}[(z^n)^\alpha(\tilde{z}^n)^{1-\alpha}I(z^n < \varepsilon)] + E_{Q^n}[(z^n)^\alpha(\tilde{z}^n)^{1-\alpha}I(z^n \geqslant \varepsilon, \tilde{z}^n \leqslant \delta)] \\
&\quad + E_{Q^n}[(z^n)^\alpha(\tilde{z}^n)^{1-\alpha}I(z^n \geqslant \varepsilon, \tilde{z}^n > \delta)] \\
&\leqslant 2\varepsilon^\alpha + 2\delta^{1-\alpha} + E_{Q^n}\left[\tilde{z}^n\left(\frac{z^n}{\tilde{z}^n}\right)^\alpha I(z^n \geqslant \varepsilon, \tilde{z}^n > \delta)\right] \\
&\leqslant 2\varepsilon^\alpha + 2\delta^{1-\alpha} + \left(\frac{2}{\delta}\right)^\alpha \tilde{P}^n(z^n \geqslant \varepsilon).
\end{aligned} \tag{10}$$

因而, 对于一切 $\alpha \in (0,1), \delta \in (0,1)$, 有

$$\lim_{\varepsilon\downarrow 0}\varliminf_n \tilde{P}^n(z^n \geqslant \varepsilon) \geqslant \left(\frac{\delta}{2}\right)^\alpha \varliminf_n H(\alpha; P^n, \tilde{P}^n) - \frac{2\delta}{2^\alpha}.$$

首先令 $\alpha \downarrow 0$ 并利用 (c), 然后令 $\delta \downarrow 0$, 得

$$\varliminf_{\varepsilon\downarrow 0}\varliminf_n \tilde{P}^n(z^n \geqslant \varepsilon) \geqslant 1.$$

于是, (b) 的正确性得证. □

定理 2 的证明.

(a)⇒(b). 设 $(\tilde{P}^n) \triangle (P^n)$, $n_k \uparrow \infty$, 而 $A^{n_k} \in \mathscr{F}^{n_k}$, 则 $P^{n_k}(A^{n_k}) \to 1, \tilde{P}^{n_k}(A^{n_k}) \to 0$. 那么, 注意到 $z^n + \tilde{z}^n = 2$, 可见

$$\begin{aligned}
\tilde{P}^{n_k}(z^{n_k} \geqslant \varepsilon) &\leqslant \tilde{P}^{n_k}(A^{n_k}) + E_{Q^{n_k}}\left[z^{n_k} \times \frac{\tilde{z}^{n_k}}{z^{n_k}}I(\overline{A}^{n_k})I(z^{n_k} \geqslant \varepsilon)\right] \\
&= \tilde{P}^{n_k}(A^{n_k}) + M_{P^{n_k}}\left[\frac{\tilde{z}^{n_k}}{z^{n_k}}I(\overline{A}^{n_k})I(z^{n_k} \geqslant \varepsilon)\right] \\
&\leqslant \tilde{P}^{n_k}(A^{n_k}) + \frac{2}{\varepsilon}P^{n_k}(\overline{A}^{n_k}).
\end{aligned}$$

从而, $\tilde{P}^{n_k}(z^{n_k} \geqslant \varepsilon) \to 0$ 故 (b) 成立.

(b)⇒(a). 如果 (b) 成立, 则存在序列 $\alpha_k \downarrow 0$, 使

$$\tilde{P}^{\alpha_k}\left(z^{\alpha_k} \geqslant \frac{1}{k}\right) \leqslant \frac{1}{k} \to 0, \quad k \to \infty.$$

由此并注意到 (见 (6) 式)

$$\tilde{P}^{\alpha_k}\left(z^{\alpha_k} \geqslant \frac{1}{k}\right) \geqslant 1 - \frac{1}{k},$$

得命题 (a).

(b)⇔(b′). 只需注意到 $Z^n = (2/z^n) - 1$.

(b)⇒(d). 由 (10) 式和 (b), 对于属于区间 (0,1) 的任意 ε 和 δ, 有

$$\varliminf_n H(\alpha; P^n, \tilde{P}^n) \leqslant 2\varepsilon^{\alpha} + 2\delta^{1-\alpha}.$$

因此,(d) 成立.

(d)⇒(c) 和 (d)⇒(e) 显然.

最后, 由 (8) 式, 有

$$\varliminf_n \tilde{P}^n(z^n \geqslant \varepsilon) \leqslant \left(\frac{2}{\varepsilon}\right)^{\alpha} \varliminf_n H(\alpha; P^n, \tilde{P}^n).$$

由于 $(2/\varepsilon)^{\alpha} \to 1, \alpha \downarrow 0$, 可见 (c)⇒(b) 和 (e)⇒(b). □

3. 独立观测模型 考虑对应于独立观测模型的一种特殊情形. 在这种情形下, 积分 $H(\alpha; P^n, \tilde{P}^n)$ 的计算和定理 1 和 2 的应用, 都没有太的困难.

假设测度 P^n 和 $\tilde{P}^n$ 是测度的直积:

$$P^n = P_1 \times \cdots \times P_n, \quad \tilde{P}^n = \tilde{P}_1 \times \cdots \times \tilde{P}_n, \quad n \geqslant 1,$$

其中 P_k 和 $\tilde{P}_k$ 是 $(\Omega_k, \mathscr{F}_k), k \geqslant 1$, 上的测度.

由于在这种情形下,

$$H(\alpha; P^n, \tilde{P}^n) = \prod_{k=1}^n H(\alpha; P_k, \tilde{P}_k) = \exp\left\{\sum_{k=1}^n \ln[1 - (1 - H(\alpha; P_k, \tilde{P}_k))]\right\},$$

则由定理 1 和 2 得如下结果:

$$(\tilde{P}^n) \vartriangleleft (P^n) \Leftrightarrow \lim_{\alpha \downarrow 0} \varlimsup_n \sum_{k=1}^n [1 - H(\alpha; P_k, \tilde{P}_k)] = 0, \tag{11}$$

$$(\tilde{P}^n) \triangle (P^n) \Leftrightarrow \lim_n \sum_{k=1}^n [1 - H(\alpha; P_k, \tilde{P}_k)] = \infty. \tag{12}$$

例 设 $(\Omega_k, \mathscr{F}_k) = (\mathbb{R}, \mathscr{B}(\mathbb{R})), \alpha_k \in (0,1)$,

$$P_k(dx) = I_{[0,1]}(x)dx, \quad \tilde{P}_k(dx) = \frac{1}{1-\alpha_k} I_{[\alpha_k, 1]}(x)dx.$$

由于这里 $H(\alpha; P_k, \tilde{P}_k) = (1-\alpha_k)^{\alpha}, \alpha \in (0,1)$, 由 (11) 式和 $H(\alpha; P_k, \tilde{P}_k) = H(1-\alpha; \tilde{P}_k, P_k)$, 可得

$$(\tilde{P}^n) \vartriangleleft (P^n) \Leftrightarrow \varlimsup_n n\alpha_n < \infty, \quad 即 \quad \alpha_n = O\left(\frac{1}{n}\right),$$

$$(P^n) \vartriangleleft (\tilde{P}^n) \Leftrightarrow \varlimsup_n n\alpha_n = 0, \quad 即 \quad \alpha_n = o\left(\frac{1}{n}\right),$$

$$(\tilde{P}^n) \triangle (P^n) \Leftrightarrow \lim_n n\alpha_n = \infty.$$

4. 练习题

1. 设 $P^n = P_1^n \times \cdots \times P_n^n, \tilde{P}^n = \tilde{P}_1^n \times \cdots \times \tilde{P}_n^n, n \geqslant 1$, 其中 P_k^n 和 $\tilde{P}_k^n$ 是参数为 $(a_k^n, 1)$ 和 $(\tilde{a}_k^n, 1)$ 的高斯测度. 求使 $(\tilde{P}^n) \vartriangleleft (P^n), (\tilde{P}^n) \triangle (P^n)$ 成立, a_k^n 和 $\tilde{a}_k^n$ 应满足的条件.

2. 设 $P^n = P_1^n \times \cdots \times P_n^n, \tilde{P}^n = \tilde{P}_1^n \times \cdots \times \tilde{P}_n^n, n \geqslant 1$, 其中 P_k^n 和 $\tilde{P}_k^n$ 是 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上的概率测度: $P_k^n(dx) = I_{[0,1]}(x)dx, \tilde{P}_k^n(dx) = I_{[a_n, 1+a_n]}(x)dx, 0 \leqslant a_n \leqslant 1$. 证明 $H(\alpha; P_k^n, \tilde{P}_k^n) = 1 - a_n$ 且

$$(\tilde{P}^n) \vartriangleleft (P^n) \Leftrightarrow (P^n) \vartriangleleft (\tilde{P}^n) \Leftrightarrow \varlimsup_n n a_n = 0,$$

$$(\tilde{P}^n) \triangle (P^n) \Leftrightarrow \varlimsup_n n a_n = \infty.$$

3. 设 $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n \geqslant 0})$ 是滤波度量空间, 即引进了 $\sigma-$ 代数流 $(\mathscr{F}_n)_{n \geqslant 0}$ 的可测空间 $(\Omega, \mathscr{F})$, 其中 $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \cdots \subseteq \mathscr{F}$. 假设 $\mathscr{F} = \sigma(\bigcup_n \mathscr{F}_n)$. 设 P 和 $\tilde{P}$ 是 $(\Omega, \mathscr{F})$ 上的两个概率测度, 而 $P_n = P|\mathscr{F}_n, \tilde{P}_n = \tilde{P}|\mathscr{F}_n$ 是 P 和 $\tilde{P}$ 在 $\mathscr{F}_n$ 上的收缩. 证明:

$$(\tilde{P}^n) \vartriangleleft (P^n) \Leftrightarrow \tilde{P} \ll P,$$

$$(\tilde{P}^n) \vartriangleleft\vartriangleright (P^n) \Leftrightarrow \tilde{P} \sim P,$$

$$(\tilde{P}^n) \triangle (P^n) \Leftrightarrow \tilde{P} \perp P.$$

§11. 中心极限定理的收敛速度

1. 中心极限定理中收敛速度的估计 设 $\xi_{n1}, \cdots, \xi_{nn}$ 是独立随机变量序列, $S_n = \xi_{n1} + \cdots + \xi_{nn}, F_n(x) = \mathbf{P}\{S_n \leqslant x\}, n \geqslant 1$. 如果 $S_n \xrightarrow{d} N(0,1)$, 则对于任意 $x \in \mathbb{R}, F_n(x) \to \Phi(x)$. 由于函数 $\Phi(x)$ 连续, 则这里实际上是一致收敛 (§1 练习题 5):

$$\sup_x |F_n(x) - \Phi(x)| \to 0, \quad n \to \infty. \tag{1}$$

自然提出关于 (1) 式的收敛速度问题. 我们考虑下面的情形:

$$S_n=\frac{\xi_1+\cdots+\xi_n}{\sigma\sqrt{n}},\quad n\geqslant 1,$$

其中 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量序列, 且 $\mathbf{E}\xi_k=0$, $\mathbf{D}\xi_k=\sigma^2>0$, 而 $\mathbf{E}|\xi_1|^3<\infty$.

定理 (贝里 - 埃森 [A. C. Berry-C. G. Esseen]) 有如下估计

$$\sup_x|F_n(x)-\Phi(x)|\leqslant\frac{C\mathbf{E}|\xi_1|^3}{\sigma^3\sqrt{n}},\tag{2}$$

其中 C 是绝对常数 $((2\pi)^{-1/2}\leqslant C<0.7655)$.

证明 为简单计, 设 $\sigma^2=1$ 和 $\beta_3=\mathbf{E}|\xi_1|^3<\infty$. 由埃森不等式 (第二章 §12 第 10 小节), 有

$$\sup_x|F_n(x)-\Phi(x)|\leqslant\frac{2}{\pi}\int_0^T\left|\frac{f_n(t)-\varphi(t)}{t}\right|\mathrm{d}t+\frac{24}{\pi T}\times\frac{1}{\sqrt{2\pi}},\tag{3}$$

其中

$$\varphi(t)=\mathrm{e}^{-\frac{t^2}{2}},\quad f_n(t)=\left[f\left(\frac{t}{\sqrt{n}}\right)\right]^n,$$

而 $f(t)=\mathbf{E}\mathrm{e}^{it\xi_1}$.

式 (3) 中的正数 T 可以任意选. 设

$$T=\frac{\sqrt{n}}{5\beta_3}.$$

下面将要证明, 对于这样选择的 T, 有

$$|f_n(t)-\varphi(t)|\leqslant\frac{7\beta_3}{6\sqrt{n}}|t|^3\mathrm{e}^{-\frac{t^2}{4}},\quad |t|\leqslant T.\tag{4}$$

考虑到这一不等式, 由 (3) 式立即得到所要证明的不等式 (2), 而其中 C 是绝对常数. (更加精确的计算表明, C 的值小于 0.7655; 参见 [88 第 5 章 §4.3].)

我们现在证明不等式 (4).

由第二章 §12 的 (18) 式 ($n=3$, $\mathbf{E}\xi_1=0$, $\mathbf{E}\xi_1^2=1$, 而 $\mathbf{E}|\xi_1|^3<\infty$), 可见

$$f(t)=\mathbf{E}\mathrm{e}^{it\xi_1}=1-\frac{t^2}{2}+\frac{(it)^3}{6}[\mathbf{E}\xi_1^3(\cos\theta_1t\xi_1+i\sin\theta_2t\xi_1)],\tag{5}$$

其中 $|\theta_1|\leqslant 1$, $|\theta_2|\leqslant 1$. 因此

$$f\left(\frac{t}{\sqrt{n}}\right)=1-\frac{t^2}{2n}+\frac{(it)^3}{6n^{3/2}}\left[\mathbf{E}\xi_1^3\left(\cos\theta_1\frac{t}{\sqrt{n}}\xi_1+i\sin\theta_2\frac{t}{\sqrt{n}}\xi_1\right)\right].$$

如果 $|t|\leqslant T=\sqrt{n}/(5\beta_3)$, 则不等式 $\beta_3\geqslant\sigma^3=1$ (见第二章 §6(28) 式), 可见

$$1-\left|f\left(\frac{t}{\sqrt{n}}\right)\right|\leqslant\left|1-f\left(\frac{t}{\sqrt{n}}\right)\right|\leqslant\frac{t^2}{2n}+\frac{|t|^3\beta_3}{6n^{3/2}}\leqslant\frac{1}{25}.$$

从而, 当 $|t|\leqslant T$ 时, 有

$$\left[f\left(\frac{t}{\sqrt{n}}\right)\right]^n=\mathrm{e}^{n\ln f\left(\frac{t}{\sqrt{n}}\right)},\tag{6}$$

其中 $\ln z$ 应理解为复数 z 的对数的主值 ($\ln z=\ln|z|+i\arg z$, $-\pi<\arg z\leqslant\pi$).

由于 $\beta_3<\infty$, 则由带拉格朗日余项的泰勒公式 (亦对照第二章 §12 (35) 式), 因为 $s_{\xi_1}^{(1)}=\mathbf{E}\xi_1=0$, $s_{\xi_1}^{(2)}=\sigma^2=1$, 可见

$$\begin{aligned}\ln f\left(\frac{t}{\sqrt{n}}\right)&=\frac{it}{\sqrt{n}}s_{\xi_1}^{(1)}+\frac{(it)^2}{2n}s_{\xi_1}^{(2)}+\frac{(it)^3}{6n^{3/2}}(\ln f)'''\left(\frac{\theta t}{\sqrt{n}}\right)\\&=-\frac{t^2}{2n}+\frac{(it)^3}{6n^{3/2}}(\ln f)'''\left(\frac{\theta t}{\sqrt{n}}\right),\quad|\theta|\leqslant 1.\end{aligned}\tag{7}$$

其次

$$\begin{aligned}[\ln f(s)]'''&=\frac{f'''(s)f^2(s)-3f''(s)f'(s)f(s)+2[f'(s)]^3}{f^3(s)}\\&=\frac{\mathbf{E}[(i\xi_1)^3\mathrm{e}^{is\xi_1}]f^2(s)-3\mathbf{E}[(i\xi_1)^2\mathrm{e}^{is\xi_1}]\mathbf{E}[(i\xi_1)\mathrm{e}^{is\xi_1}]f(s)+2\mathbf{E}[(i\xi_1)\mathrm{e}^{is\xi_1}]^3}{f^3(s)}.\end{aligned}$$

因此, 注意到对于 $|t|\leqslant T$, $|f(s)|\leqslant 1$, 有 $|f(t/\sqrt{n})|\geqslant 24/25$, 得

$$\left|(\ln f)'''\left(\frac{\theta t}{\sqrt{n}}\right)\right|\leqslant\frac{\beta_3+3\beta_1\beta_2+2\beta_1^3}{\left(\frac{24}{25}\right)^3}\leqslant 7\beta_3\tag{8}$$

($\beta_k=\mathbf{E}|\xi_1|^k$, $k=1,2,3$; $\beta_1\leqslant\beta_2^{1/2}\leqslant\beta_3^{1/3}$; 见第二章 §6 (28) 式).

运用不等式 $|\mathrm{e}^z-1|\leqslant|z|\mathrm{e}^{|z|}$, 由 (6)~(8) 式, 对于 $|t|\leqslant T=\sqrt{n}/(5\beta_3)$, 得

$$\begin{aligned}\left|\left[f\left(\frac{t}{\sqrt{n}}\right)\right]^n-\mathrm{e}^{-\frac{t^2}{2}}\right|&=\left|\mathrm{e}^{n\ln f\left(\frac{t}{\sqrt{n}}\right)}-\mathrm{e}^{-\frac{t^2}{2}}\right|\\&\leqslant\frac{7\beta_3|t|^3}{6\sqrt{n}}\exp\left\{-\frac{t^2}{2}+\frac{7\beta_3|t|^3}{6\sqrt{n}}\right\}\leqslant\frac{7\beta_3|t|^3}{6\sqrt{n}}\mathrm{e}^{-\frac{t^2}{4}}.\end{aligned}$$ □

注 需要指出, 在缺乏关于可和随机变量的性质的补充假设的情况下, 估计式 (2) 的阶数和估计 $C\geqslant(2\pi)^{-1/2}$ 不能够改进. 事实上, 设 $\xi_1,\xi_2,\cdots$ 是独立同分布的伯努利随机变量, 且

$$\mathbf{P}\{\xi_k=+1\}=\mathbf{P}\{\xi_k=-1\}=\frac{1}{2}.$$

由对称性, 可见

$$2\mathbf{P}\left\{\sum_{k=1}^{2n}\xi_k<0\right\}+\mathbf{P}\left\{\sum_{k=1}^{2n}\xi_k=0\right\}=1,$$

因而, 根据斯特林公式 (第一章 §2 (6) 式), 有

$$\left|\mathbf{P}\left\{\sum_{k=1}^{2n}\xi_k<0\right\}-\frac{1}{2}\right|=\frac{1}{2}\mathbf{P}\left\{\sum_{k=1}^{2n}\xi_k=0\right\}=\frac{1}{2}C_{2n}^{n}2^{-2n}\sim\frac{1}{2\sqrt{\pi n}}=\frac{1}{\sqrt{(2\pi)(2n)}}.$$

特别, 由此可见, 常数 C 不小于 $(2\pi)^{-1/2}$, 且

$$\mathbf{P}\left\{\sum_{k=1}^{2n}\xi_k=0\right\}\sim\frac{1}{\sqrt{\pi n}},\quad n\to\infty. \tag{9}$$

2. 练习题

1. 证明 (8) 式.

2. 设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量, 且 $\mathbf{E}\xi_1=0,\mathrm{D}\xi_1^2=\sigma^2$, 而 $\mathbf{E}|\xi_1|^3<\infty$. 熟知, 对于一切 $-\infty<x<\infty$ 有如下不均匀不等式估计:

$$|F_n(x)-\Phi(x)|\leqslant\frac{C\mathbf{E}|\xi_1|^3}{\sigma^3\sqrt{n}}\times\frac{1}{(1+|x|)^3}.$$

证明此不等式, 至少对于伯努利随机变量证明之.

3. 设 $(\xi_k)_{k\geqslant 1}$ 是独立同分布随机变量序列, 分别以概率 $1/2$ 取 ± 1 为值. 设

$$\varphi_1(t)=\mathbf{E}e^{it\xi_1}=\frac{1}{2}(e^{it}+e^{-it}).$$

仿照拉普拉斯, 证明

$$\mathbf{P}\{S_{2n}=0\}=\frac{1}{\pi}\int_0^{\pi}\varphi_1^{2n}(t)\mathrm{d}t\sim\frac{1}{\sqrt{\pi n}},\quad n\to\infty,$$

其中 $S_k=\xi_1+\cdots+\xi_k$.

4. 设 $(\xi_k)_{k\geqslant 0}$ 是独立同分布随机变量序列, 各取 $2a+1$ 个整数值: $0,\pm1,\cdots,\pm a$. 设

$$\varphi_{2a+1}(t)=\mathbf{E}e^{it\xi_1}=\frac{1}{1+2a}\left(1+2\sum_{k=1}^{a}\cos tk\right).$$

仍仿照拉普拉斯, 证明

$$\mathbf{P}\{S_n=0\}=\frac{1}{\pi}\int_0^{\pi}\varphi_{2a+1}^{n}(t)\mathrm{d}t\sim\frac{\sqrt{3}}{\sqrt{2\pi(a+1)n}},\quad n\to\infty.$$

特别, 对于 $a=1$, 即对于 ξ_k 取 3 个可能值 $-1,0,1$ 的情形, 证明

$$\mathbf{P}\{S_n=0\}\sim\frac{\sqrt{3}}{2\sqrt{\pi n}},n\to\infty.$$

§12. 泊松定理的收敛速度

1. 泊松定理中收敛速度的估计 设 $\xi_1,\xi_2,\cdots,\xi_n$ 是独立伯努利随机变量, 取 1 和 0 两个可能值, 相应的概率为

$$\mathbf{P}(\xi_k=1)=p_k,\mathbf{P}(\xi_k=0)=q_k(=1-p_k),1\leqslant k\leqslant n.$$

记 $B=(B_0,B_1,\cdots,B_n)$ 为和 $S=\xi_1+\cdots+\xi_n$ 的二项分布的概率, 其中 $B_k=\mathbf{P}(S=k)$. 再设 $\Pi=(\Pi_0,\Pi_1,\cdots)$ 是参数为 λ 的泊松分布, 其中

$$\Pi_k=\frac{\lambda^k}{k!}e^{-\lambda},k\geqslant 0.$$

如第一章 §6 第 4 小节指出的, 若

$$p_1=\cdots=p_n,\lambda=np, \tag{1}$$

则对于 B 和 Π 间的变差距离有如下估计 (普罗霍罗夫):

$$\|B-\Pi\|=\sum_{k=0}^{\infty}|B_k-\Pi_k|\leqslant C_1(\lambda)p=C_1(\lambda)\times\frac{\lambda}{n}, \tag{2}$$

其中 $B_{n+1}=B_{n+2}=\cdots=0$, 而

$$C_1(\lambda)=2\min(2,\lambda). \tag{3}$$

对于 $\sum_{k=1}^{n}p_k=\lambda$ (但 p_k 未必相等) 的情形, 卡姆 (L. Le Cam) 证明, 得

$$\|B-\Pi\|=\sum_{k=0}^{\infty}|B_k-\Pi_k|\leqslant C_2(\lambda)\max_{1\leqslant k\leqslant n}p_k, \tag{4}$$

其中

$$C_2(\lambda)=2\min(9,\lambda). \tag{5}$$

由下面将证明的定理可得估计

$$\|B-\Pi\|\leqslant C_3(\lambda)\max_{1\leqslant k\leqslant n}p_k, \tag{6}$$

其中

$$C_3(\lambda)=2\lambda. \tag{7}$$

虽然当 $\lambda>9$ 时 $C_2(\lambda)<C_3(\lambda)$, 即估计 (6) 比估计 (4) 差, 但是我们还是证明 (6) 式, 因为估计 (6) 本质上是初等的, 然而要得到估计 (4) 中 “较好的” 常数, 证明的技术将复杂得多.

2. 估计式 (6) 的证明

定理 设 $\lambda=\sum_{k=1}^{n}p_k$, 则

$$\|B-\Pi\|\leqslant\sum_{k=1}^{n}\|B_k-\Pi_k\|\leqslant 2\sum_{k=1}^{n}p_k^2. \tag{8}$$

证明 利用 B 和 Π 中每一个分布是相应分布的卷积:

$$\begin{aligned}B&=B(p_1)*B(p_2)*\cdots*B(p_n),\\ \Pi&=\Pi(p_1)*\Pi(p_2)*\cdots*\Pi(p_n),\end{aligned} \tag{9}$$

理解为相应分布函数的卷积 (见第二章 §8 第 4 小节), 其中 $B(p_k)=(1-p_k,p_k)$ 是参数为 p_k 的在点 0 和 1 的伯努利分布, 而 $\Pi(p_k)$ 是集中在点 $0,1,\cdots$ 参数为 p_k 的泊松分布.

不难证明, 差 $B-\Pi$ 可以表示为:

$$B-\Pi=R_1+\cdots+R_n, \tag{10}$$

其中

$$R_k=[B(p_k)-\Pi_k(p_k)]*F_k, \tag{11}$$

而

$$\begin{aligned}&F_1=\Pi(p_2)*\cdots*\Pi(p_n),\\ &F_k=B(p_1)*\cdots*B(p_{k-1})*\Pi(p_{k+1})*\cdots*\Pi(p_n),\ 2\leqslant k\leqslant n-1,\\ &F_n=B(p_1)*\cdots*B(p_{n-1}).\end{aligned}$$

由于 §9 的练习题 6, 可见 $\|R_k\|\leqslant\|B(p_k)-\Pi(p_k)\|$. 因此, 由 (10) 式, 立即得

$$\|B-\Pi\|\leqslant\|B(p_k)-\Pi(p_k)\|. \tag{12}$$

注意到 §9 的 (12) 式, 可见计算变差 $\|B(p_k)-\Pi(p_k)\|$ 并不困难:

$$\begin{aligned}\|B(p_k)-\Pi(p_k)\|&=|(1-p_k)-e^{-p_k}|+|p_k-p_ke^{-p_k}|+\sum_{j\geqslant 2}\frac{p_k^je^{-p_k}}{j!}\\ &=|(1-p_k)-e^{-p_k}|+|p_k-p_ke^{-p_k}|+1-e^{-p_k}-p_ke^{-p_k}\\ &=2p_k(1-e^{-p_k})\leqslant 2p_k^2.\end{aligned}$$

于是, 由 (12) 式, 得所要求证明的不等式 (8). □

系 由于

$$\sum_{k=1}^{n}p_k^2\leqslant\lambda\max_{1\leqslant k\leqslant n}p_k,$$

可见估计式成立.

3. 练习题

1. 证明当 $\lambda_k=-\ln(1-p_k)$ 时

$$\|B(p_k)-\Pi(p_k)\|=2(1-e^{-\lambda_k}-\lambda_ke^{-\lambda_k})\leqslant\lambda_k^2,$$

因而 $\|B-\Pi\|\leqslant\sum_{k=1}^{n}\lambda_k^2$.

2. 证明表达式 (9) 和 (10).

§13. 数理统计的基本定理

1. 数理统计与概率论的关系 在第一章 §7 中对于伯努利概型, 曾介绍过根据对随机变量的观测, 估计 "成功" 的概率, 以及为其置信区间的问题. 对于在一定意义上研究概率论的反问题的数理统计来说, 这些问题是典型的和基本的. 例如, 假如概率论的基本兴趣是, 对于给定的概率模型, 计算各种概率指标 (事件的概率, 随机元的概率分布及其各种特征, 等等), 则在数理统计中, 我们的兴趣在于: 如何根据统计资料 (以一定的可靠性) 揭示相应的概率模型, 使之在经验资料统计性质的框架内, 最好地与生成这些统计资料的随机机制的概率性质一致.

下面引用的 (格里汶科 [В. И. Гливенко], 坎泰利, 柯尔莫戈洛夫和斯米尔诺夫 [Н. В. Смирнов]) 结果, 理应称为数理统计的基本定理, 因为有关成果不但奠定了由统计资料提取 (关于被观测随机变量的分布函数的) 概率信息的原则可能性, 而且可以估计经验资料与各种不同的概率模型拟合程度.

2. 分布函数与经验分布函数的拟合定理 设 $\xi_1,\xi_2,\cdots,\xi_n$ 是定义在某一概率空间 $(\Omega,\mathscr{F},\mathbf{P})$ 上的独立同分布随机变量序列, 而 $F=F(x),x\in\mathbb{R}$, 是其共同的分布函数 $(F(x)=\mathbf{P}\{\xi_k\leqslant x\})$. 对于每一个 $N\geqslant 1$, 定义经验分布函数:

$$F_N(x;\omega)=\frac{1}{N}\sum_{k=1}^{N}I(\xi_k\leqslant x),\quad x\in\mathbb{R}. \tag{1}$$

根据大数定律 (§3 定理 2), 对于每一个 $x\in\mathbb{R}$, 有

$$F_N(x;\omega)\xrightarrow{\mathbf{P}}F(x),\quad N\to\infty, \tag{2}$$

即 $\mathbf{P}$ 一依概率收敛.

由第四章 §3 的定理 1 和 2, 可见对于每一个 $x\in\mathbb{R}$, 当 $N\to\infty$ 时,依概率 1 收敛, 有

$$F_N(x;\omega)\to F(x),\quad(\mathbf{P}-\text{a.c.}). \tag{3}$$

这里有非常好的、更强的结果: (3) 式中的收敛关于 x 是一致的,

定理 (格里汶科和坎泰利) 在上述条件下, 随机变量

$$D_N(\omega)=\sup_{x\in\mathbb{R}}|F_N(x;\omega)-F(x)| \tag{4}$$

依概率 1 收敛于 0:

$$\mathbf{P}\{\lim_N D_N(\omega)=0\}=1. \tag{5}$$

证明 设 $\mathbb{Q}$ 是 $\mathbb{R}$ 中有理数的集合. 显然

$$\sup_{r\in\mathbb{Q}}|F_N(r;\omega)-F(r)|$$

是随机变量. 由于

$$D_N(\omega)=\sup_{x\in\mathbb{R}}|F_N(x;\omega)-F(x)|=\sup_{r\in\mathbb{Q}}|F_N(r;\omega)-F(r)|,$$

可见 $D_N(\omega)$ 也是随机变量, 因此它有概率分布.

设整数 $M\geqslant 2$, 而 $k=1,2,\cdots,M-1$. 定义数列

$$x_{M,k}=\min\{x\in\mathbb{R}:k/M\leqslant F(x)\},$$

同时设 $x_{M,0}=-\infty, x_{M,M}=\infty$.

假设 $x\in[x_{M,k},x_{M,k+1})$, 而且 $[x_{M,k},x_{M,k+1})\neq\varnothing$. 那么, 显然

$$\begin{aligned}
&F_N(x;\omega)-F(x)\leqslant F_N(x_{M,k+1}-0)-F(x_{M,k})\\
&=[F_N(x_{M,k+1}-0)-F(x_{M,k+1}-0)]+[F(x_{M,k+1}-0)-F(x_{M,k})]\\
&\leqslant F_N(x_{M,k+1}-0)-F(x_{M,k+1}-0)+\frac{1}{M}.
\end{aligned}$$

类似地, 再设 $x\in[x_{M,k},x_{M,k+1})$, 有

$$F_N(x;\omega)-F(x)\geqslant F_N(x_{M,k};\omega)-F(x_{M,k})-\frac{1}{M}.$$

因此, 对于每一个 $x\in\mathbf{R}$,

$$\begin{aligned}
&|F_N(x;\omega)-F(x)|\\
&\leqslant\max_{\substack{1\leqslant k\leqslant M-1\\ 1\leqslant l\leqslant M-1}}\{|F_N(x_{M,k};\omega)-F(x_{M,k})|,|F_N(x_{M,l}-0;\omega)-F(x_{M,l}-0)|\}+\frac{1}{M},
\end{aligned}$$

于是,

$$\varlimsup_{n\to\infty}\sup_x|F_N(x;\omega)-F(x)|\leqslant\frac{1}{M}\quad(\mathbf{P}-\text{a.c.}).$$

由于 M 的任意性, 由此可得结论 (5). □

格里汶科和坎泰利定理, 是数理统计的基本定理之一. 前面已经指出, 该定理奠定了由根据对独立同分布随机变量 $\xi_1,\xi_2,\cdots$ 的观测结果, 检验这些变量的分布函数是否恰好是函数 $F(x)$. 换句话说, 定理保障建立 "理论与试验" 一致的可能性.

3. 经验分布函数对分布函数的偏差 $D_N(\omega)$ 和 N 与 $F(x)$ 无关 尽管格里汶科和坎泰利定理有所指出的重要性, 然而它并未回答, 关于 "当 $N\to\infty$ 时, 偏差变量 $D_N(\omega)$ 收敛于 0 的速度" 问题. 因而, 不能推断, 独立观测结果与 "它 (观测结果) 具有假设的分布函数 $F=F(x), x\in\mathbb{R}$" 这一论断的似然程度.

由中心极限定理可见, 对于每一个固定的 $x\in\mathbb{R}$, 有

$$\sqrt{N}[F_N(x;\omega)-F(x)]\xrightarrow{law}\mathscr{N}(0,F(x)[1-F(x)]), \tag{6}$$

说明随机变量 $\sqrt{N}[F_N(x;\omega)-F(x)]$ 收敛于均值为 0, 方差为 $F(x)[1-F(x)]$ 的正态分布.

当然, 更重要的是得到 (一致) 统计量

$$D_N(\omega)=\sup_x|F_N(x;\omega)-F(x)|,$$

或相近的统计量

$$D_N^+(\omega)=\sup_x[F_N(x;\omega)-F(x)] \tag{7}$$

的极限分布.

在求这些统计量的极限分布时, 下面 (柯尔莫戈洛夫) 的研究结果是关键.

引理 1 设 $\mathscr{F}$ 是连续型分布函数 $F=F(x)$ 类.

对于每一个 $N\geqslant 1$ 和一切 $F\in\mathscr{F}$, 统计量 $D_N(\omega)$ 的概率分布是同一分布, 对于统计量 $D_N^+(\omega)$, 有类似的结果.

证明 设 $\eta_1,\eta_2,\cdots$ 是独立同分布随机变量序列, 且在 $[0,1]$ 上同 ($U=U(x)$) 服从均匀分布: $\mathbf{P}\{\eta_1\leqslant x\}=x, 0\leqslant x\leqslant 1$.

由于对于连续函数 $F=F(x)$, 统计量 $\sup_x|F_N(x;\omega)-F(x)|$ 的分布, 与随机变量 $\sup_x|U_N(x;\omega)-U(x)|$ 的分布相同, 其中

$$U_N(x;\omega)=\frac{1}{N}\sum_{k=1}^{N}I(\eta_k(\omega)\leqslant x)$$

是随机变量 $\eta_1,\cdots,\eta_N$ 的经验分布函数, 由此可以得到引理的结论.

以 $\mathcal{A}$ 表示分布函数 $F=F(x)$ 是常数的区间 $I=[a,b], -\infty<a<b<\infty$, 的集合, 即 $\mathbf{P}\{\xi_1\in I\}=0$.

$$D_N(\omega)=\sup_{x\in\mathbb{R}}|F_N(x;\omega)-F(x)|=\sup_{x\in\overline{\mathcal{A}}}|F_N(x;\omega)-F(x)|.$$

若引进随机变量 $\tilde{\eta}_k=F(\xi_k)$ 及其经验分布函数

$$\tilde{U}_N(x;\omega)=\frac{1}{N}\sum_{k=1}^{N}I(\tilde{\eta}_k(\omega)\leqslant x),$$

则对于任意 $x \in \overline{A}$,

$$U_N(F(x);\omega)=\frac{1}{N}\sum_{k=1}^{N} I(F(\xi_k(\omega))\leqslant F(x))=\frac{1}{N}\sum_{k=1}^{N} I(\xi_k(\omega)\leqslant x)=F_N(x;\omega),$$

因为对于这样的 x, $\{\omega:\xi_k(\omega)\leqslant x\}=\{\omega:F(\xi_k(\omega))\leqslant F(x)\}$.

这样

$$\begin{aligned} D_N(\omega) &= \sup_{x\in A}|F_N(x;\omega)-F(x)| = \sup_{x\in \overline{A}}|U_N(F(x);\omega)-F(x)| \\ &= \sup_{x\in \mathbf{R}}|U_N(F(x);\omega)-F(x)| = \sup_{y\in(0,1)}|U_N(y;\omega)-y| \\ &\overset{\mathbf{P}\text{-a.c.}}{=\!=} \sup_{y\in[0,1]}|U_N(y;\omega)-y|, \end{aligned}$$

其中最后一个等式 ($\mathbf{P}$ - a.c.) 成立, 因为

$$\mathbf{P}\{\tilde{\eta}_1=0\}=\mathbf{P}\{\tilde{\eta}_1=1\}=0. \tag{8}$$

现在证明, 随机变量 $\tilde{\eta}_k$ (在 $[0,1]$ 上) 均匀分布. 为此记

$$x(y)=\inf\{x\in\mathbf{R}:F(x)\geqslant y\},\quad y\in(0,1).$$

那么, 有 $F(x(y))=y$, $y\in(0,1)$, 且

$$\begin{aligned}\mathbf{P}\{\tilde{\eta}_1\leqslant y\}&=\mathbf{P}\{F(\xi_1)\leqslant y\}=\mathbf{P}\{F(\xi_1)\leqslant F(x(y))\}\\ &=\mathbf{P}\{\xi_1\leqslant x(y)\}=F(x(y))=y.\end{aligned}$$

由此连同 (8) 式就证明了, 随机变量 $\tilde{\eta}_1$ (因此 $\tilde{\eta}_2,\tilde{\eta}_3,\cdots$ 也都) 在 $[0,1]$ 上均匀分布. □

4. 统计量 $D_N(\omega)$ 和 $D_N^+(\omega)$ 的极限分布 上面的引理 1 表明, 为 (在对观测结果 $\xi_1,\xi_2,\cdots$ 的连续分布函数 $F=F(x)$ 类 $\mathcal{F}$ 中) 当 $N\to\infty$ 时求统计量 $D_N(\omega)$ 和 $D_N^+(\omega)$ 的极限分布, 只需假设 $\xi_1,\xi_2,\cdots$ 是独立同分布在 $[0,1]$ 上均匀分布的随机变量序列.

固定某个 $N\geqslant 1$, 并且 (对于每一个 $\omega\in\Omega$) 将变量 $\xi_1,\xi_2,\cdots,\xi_N$ 按递增顺序排列所得的新变量 (顺序统计量) 记作 $\xi_1^{(N)},\xi_2^{(N)},\cdots,\xi_N^{(N)}$, 其中

$$\xi_1^{(N)}=\min(\xi_1,\xi_2,\cdots,\xi_N),\cdots,\xi_N^{(N)}=\max(\xi_1,\xi_2,\cdots,\xi_N).$$

(两个这样有序变量相等的概率等于 0.)

设

$$U_N(y;\omega)=\frac{1}{N}\sum_{k=1}^{N} I(\xi_k(\omega)\leqslant y), \tag{9}$$

则

$$D_N(\omega)=\max_{y\in\mathbf{R}}|U_N(y;\omega)-y|. \tag{10}$$

不难看出, 上式右侧只有在函数 $U_N(y;\omega)$ 的跳跃点 $\xi_1^{(N)},\xi_2^{(N)},\cdots,\xi_N^{(N)}$ 上, 才可以达到最大值. 因此,

$$D_N(\omega)=\max_{k\leqslant N}\left|\frac{k}{N}-\xi_k^{(N)}\right|. \tag{11}$$

由此可见, 为求统计量 $D_N(\omega)$ 的分布, 需要知道顺序统计量 $\xi_1^{(N)},\xi_2^{(N)},\cdots,\xi_N^{(N)}$ 的联合分布.

为求此分布, 考虑独立指数分布的随机变量序列: $\zeta_1,\zeta_2,\cdots$: $\mathbf{P}\{\zeta_k>x\}=e^{-x}$, $x\geqslant 0$, 且设 $S_n=\zeta_1+\cdots+\zeta_n$, $n\geqslant 1$.

引理 2 向量 $(\xi_1^{(N)},\xi_2^{(N)},\cdots,\xi_N^{(N)})$ 的联合分布, 等同于向量

$$\left(\frac{S_1}{S_{N+1}},\frac{S_2}{S_{N+1}},\cdots,\frac{S_N}{S_{N+1}}\right)$$

的联合分布.

证明 一方面, 对于 $0\leqslant y_1\leqslant\cdots\leqslant y_N\leqslant 1$, 有

$$\mathbf{P}\{\xi_1^{(N)}\in \mathrm{d}y_1,\cdots,\xi_N^{(N)}\in \mathrm{d}y_N\}=\sum\mathbf{P}\{\xi_{k_1}\in \mathrm{d}y_1,\cdots,\xi_{k_N}\in \mathrm{d}y_N\}, \tag{12}$$

其中对整数 $(1,\cdots,N)$ 的一切排列 $(k_1,\cdots,k_N)$ 求和.(在 (12) 式中及下面使用有些随意的 "微分" 书写形式, 不过不难赋予其确切的含义.)

在 (12) 式右侧的随机变量独立服从 $[0,1]$ 上的均匀分布. 由于 $(1,\cdots,N)$ 的一切排列的总数为 $N!$, 可见 (12) 式右侧等于 $N!\mathrm{d}y_1\cdots\mathrm{d}y_N$. 从而, 如果 $(y_1,\cdots,y_N)\in\Delta_N$, 其中 $\Delta_N=\{y_1\in[0,1],\cdots,y_N\in[0,1]:0\leqslant y_1\leqslant\cdots\leqslant y_N\leqslant 1\}$, 则

$$\mathbf{P}\{\xi_1^{(N)}\in \mathrm{d}y_1,\cdots,\xi_N^{(N)}\in \mathrm{d}y_N\}=N!\mathrm{d}y_1\cdots\mathrm{d}y_N. \tag{13}$$

如果 $(y_1,\cdots,y_N)\notin\Delta_N$, 则

$$\mathbf{P}\{\xi_1^{(N)}\in \mathrm{d}y_1,\cdots,\xi_N^{(N)}\in \mathrm{d}y_N\}=0. \tag{14}$$

另一方面, 现在证明同样的结果

$$\mathbf{P}\left\{\frac{S_1}{S_{N+1}}\in \mathrm{d}y_1,\cdots,\frac{S_N}{S_{N+1}}\in \mathrm{d}y_N\right\}=\begin{cases}N!\mathrm{d}y_1\cdots\mathrm{d}y_N, & 若\ (y_1,\cdots,y_N)\in\Delta_N,\\ 0, & 若\ (y_1,\cdots,y_N)\notin\Delta_N.\end{cases} \tag{15}$$

为此, 考虑随机变量 $S_1,\cdots,S_{N+1}$ 的联合分布.

如果 $(x_1,\cdots,x_{N+1})\in\Delta_{N+1}$, 则

$$\begin{aligned}&\mathbf{P}\{S_1\in dx_1,S_2\in dx_2,\cdots,S_{N+1}\in dx_{N+1}\}\\&=\mathbf{P}\{\zeta_1\in dx_1,\zeta_2\in dx_2-x_1,\cdots,\zeta_{N+1}\in dx_{N+1}-x_N\}\\&=e^{-x_1}e^{-(x_2-x_1)}\cdots e^{-(x_{N+1}-x_N)}dx_1dx_2\cdots dx_{N+1}\\&=e^{-x_{N+1}}dx_1dx_2\cdots dx_{N+1}.\end{aligned}\tag{16}$$

如果 $(x_1,\cdots,x_{N+1})\notin\Delta_{N+1}$, 则

$$\mathbf{P}\{S_1\in dx_1,S_2\in dx_2,\cdots,S_{N+1}\in dx_{N+1}\}=0.\tag{17}$$

因为 $S_{N+1}=\zeta_1+\cdots+\zeta_{N+1}$, 其中 $\zeta_1,\cdots,\zeta_{N+1}$ 是独立指数分布随机变量, 则 (练习题 3)

$$\mathbf{P}\{S_{N+1}\in dx_{N+1}\}=\frac{x_{N+1}^Ne^{-x_{N+1}}}{N!}dx_{N+1}.\tag{18}$$

由 (16) 和 (18) 式, 可见对于 $(x_1,\cdots,x_{N+1})\in\Delta_{N+1}$, 有

$$\mathbf{P}\{S_1\in dx_1,\cdots,S_N\in dx_N\mid S_{N+1}\in dx_{N+1}\}=N!x_{N+1}^{-N}dx_1\cdots dx_N.\tag{19}$$

如果 $(x_1,\cdots,x_{N+1})\notin\Delta_{N+1}$, 则 (19) 式的左侧等于 0.

结果, 得

$$\begin{aligned}&\mathbf{P}\left\{\frac{S_1}{S_{N+1}}\in dy_1,\cdots,\frac{S_N}{S_{N+1}}\in dy_N\,\middle|\,S_{N+1}=x_{N+1}\right\}\\&=\begin{cases}N!dy_1\cdots dy_N, & 若\ (y_1,\cdots,y_N)\in\Delta_N,\\0, & 若\ (y_1,\cdots,y_N)\notin\Delta_N.\end{cases}\end{aligned}$$

而由于右侧对 x_{N+1} 的独立性, 可得表示式 (15) 的正确性, 将其与 (13) 和 (14) 式比较, 便可证明引理 2 的结论. ∎

由引理 2 可见

$$\sqrt{N}D_N(\omega)\overset{d}{=}\sqrt{N}\max_{k\leqslant N}\left|\frac{S_k}{S_{N+1}}-\frac{k}{N}\right|,\tag{20}$$

其中 "$\overset{d}{=}$" 表示左, 右两侧随机变量的分布等同.

由 (20) 式, 有

$$\sqrt{N}D_N(\omega)\overset{d}{=}\frac{N}{S_{N+1}}\max_{k\leqslant N}\left|\frac{S_k-k}{\sqrt{N}}-\frac{k}{N}\frac{S_{N+1}-N}{\sqrt{N}}\right|.\tag{21}$$

对于研究当 $N\to\infty$ 时统计量 $\sqrt{N}D_N(\omega)$ 的极限性质, 关系式 (20) 非常方便.

由于 $N/S_{N+1}\to1$ (P-a. c.), 则对于 $t\in[0,1]$, 设

$$X_t^{(N)}=\frac{S_{[Nt]}-[Nt]}{\sqrt{N}},$$

有

$$\lim_{N\to\infty}\mathbf{P}\{\sqrt{N}D_N(\omega)\leqslant y\}=\lim_{N\to\infty}\mathbf{P}\left\{\sup_{0\leqslant t\leqslant1}|X_t^{(N)}-tX_1^{(N)}|\leqslant y\right\}.$$

这样, 关于统计量 $\sqrt{N}D_N(\omega)$ 的极限分布问题, 归结为当 $N\to\infty$ 时研究统计量

$$\sup_{0\leqslant t\leqslant1}|X_t^{(N)}-tX_1^{(N)}|$$

的极限性质问题.

根据中心极限定理 (§4 定理 1), 对于每一固定的 $t\in[0,1]$, 随机变量 $X_t^{(N)}$ 的极限分布, 等同于随机变量 B_t 的极限分布 —— 参数为 $\mathbf{E}B_t=0$ 和 $\mathbf{E}B_t^2=t$ 的高斯分布.

实际上, 还可以得出更多结论. 具体地说, 考虑第二章 §13 引进的布朗运动 (维纳过程) $B=(B_t)_{0\leqslant t\leqslant1}$. 这样的过程在第二章 §13, 是作为 $B_0=0,\mathbf{E}B_t=0$ 且协方差矩阵为 $\mathbf{E}B_sB_t=\min(s,t)$ 的高斯过程引进的. 于是, 结果表明, 根据第七章 §9 定理 1 的注 3, 随机变量 $(X_{t_1}^N,\cdots,X_{t_k}^N)$ 的联合分布 $P_{t_1,\cdots,t_k}^{(N)}$ 弱收敛于随机变量 $(B_{t_1},\cdots,B_{t_k})$ 的联合分布 $P_{t_1,\cdots,t_k}$ (关于弱收敛见 §1, 2.) 自然想到, 统计量 $\sup_{0\leqslant t\leqslant1}|X_t^{(N)}-tX_1^{(N)}|$ 及其分布, 收敛于统计量 $\sup_{0\leqslant t\leqslant1}|B_t-tB_1|$ 及其分布.

一般说来, 有限维分布 $P_{t_1,\cdots,t_k}^{(N)}\to P_{t_1,\cdots,t_k},0\leqslant t_1<\cdots<t_k=1,k\geqslant1$, 的弱收敛 (即按 §1 练习题 3 的记号, $P^{(N)}\overset{f}{\Rightarrow}P$), 对于泛函

$$f(X^{(N)})=\sup_{0\leqslant t\leqslant1}|X_t^{(N)}-tX_1^{(N)}|$$

的分布收敛于泛函

$$f(X)=\sup_{0\leqslant t\leqslant1}|X_t-tX_1|$$

的分布还不充分, 其中 $X_t=B_t(0\leqslant t\leqslant1)$ 是布朗运动.

不过, 对于所考虑的情形确实是这样, 这由以下的讨论可见.

过程 $X^{(N)}$ 的轨道属于空间 $D=D[0,1]$, 而过程 X 的轨道属于空间 $C=C[0,1]\subseteq D$ (见第二章 §2 第 6,7 小节). 在空间 D 中可以引进 "普罗霍罗夫度量" ρ (在 [5, 第三章] 中 ρ 记作 d_0, 在 [87, 第四章] 中 $\rho=\delta$). 度量空间 $(D,\mathscr{D},\rho)$ 关于 ρ 是完全可分空间. (第二章 §2 第 7 小节定义的, 以 "斯科罗霍德度量" d 为度量的空间 $(D,\mathscr{D},\rho)$ 仅是可分的, 而对于以后在第三章 §2 中的应用普罗霍罗夫定理还需要 "完度").

关于度量 ρ, 泛函 $f(x)=\sup_{0\leqslant t\leqslant1}|x_t-tx_1|$ (显然与 $\sup_{0\leqslant t\leqslant1}|x_t-tx_1|$ 相等) 是连续的, 其中 $x=(x_t)_{0\leqslant t\leqslant1}\in D$. 从而为证明收敛性 $f(X^{(N)})\overset{d}{\to}f(X)$ (即按分布收敛), 只需证明 (空间 $(D,\mathscr{D},\rho)$ 的) 弱收敛 $P^{(N)}\overset{w}{\to}P$ (练习题 2).

由普罗霍罗夫定理直接引述的, 证明这一收敛性的一般方法, 基于如下蕴涵关系 (见练习题 3, 以及 §1 练习题 3 的记号):

$$\left(P^{(N)} \xrightarrow{f} P\right) \Leftrightarrow \left(\{P^{(N)}\} \text{ 的密度, 且 } \mathscr{K}_0(D) \text{ 是定义类}\right) \to \left(P^{(N)} \xrightarrow{w} P\right). \tag{22}$$

(这里, $\mathscr{K}_0(D)$ 是柱集类 §1 第 5 小节.)

在所考虑的情形下, 是有限维分布的收敛 $P^{(N)} \xrightarrow{f} P$, 而柱集类 $\mathscr{K}_0(D)$ 确实是收敛的定义类 (第二章 §2 第 7 小节). 当然, 在运用蕴涵关系 (22) 时, 最困难的是验证所考虑的测度族 $\{P^{(N)}\}$ 是稠密的. 这样的验证基于包含在测度族定义中 (§2 的 (1) 式) 空间 D 的紧集的性质描述, 而这已经超出了本书的范围. (相应的证明, 参见 [5, 定理 15.2], [87, 第六章, 3.21].)

在这些讨论的最后我们指出, 还存在证明收敛 $f(X^{(N)}) \xrightarrow{d} f(X)$ 的如下途径. 间断过程 $X^{(N)}$ 可以用如下连续过程 $\tilde{X}^{(N)}$ 逼近, 其中对于任意 $\varepsilon > 0$ 和一切 $\omega \in \Omega$, 以及充分大的 N, 有

$$\sup_{0 \leqslant t \leqslant 1} |X_t^{(N)}(\omega) - \tilde{X}_t^{(N)}(\omega)| \leqslant \varepsilon.$$

因此, 只需证明 $f(\tilde{X}^{(N)}) \xrightarrow{d} f(X)$, 而这略微简单, 因为这时可以不涉及空间 D, 而涉及比较简单的空间 C. 在空间 C 中, 至少对于所考虑的情形, "稠密性" 准则容易验证 ([5, 第二章], [87, 第六章]).

5. 柯尔莫戈洛夫分布 设 $B_t^\circ = B_t - tB_1 (0 \leqslant t \leqslant 1)$. 这一过程是高斯过程, $B_0^\circ = 0, \mathbf{E}B_t^\circ = 0$, 而 $\mathbf{E}B_s^\circ B_t^\circ = \min(s, t) - st$. 在第二章 §13 第 7 小节, 此过程曾经称做条件维纳过程或布朗桥.

于是, 由以上的讨论得出这样的结论:

$$\lim_{N \to \infty} \mathbf{P}\{\sqrt{N} D_N(\omega) \leqslant y\} = \mathbf{P}\left\{\sup_{0 \leqslant t \leqslant 1} |B_t^\circ| \leqslant y\right\}, \tag{23}$$

其中 $B^\circ = (B_t^\circ)_{0 \leqslant t \leqslant 1}$ 是布朗桥.

由布朗运动的理论 (例如, 见 [5, 第二章, §11]) 知, 分布

$$K(y) = \mathbf{P}\left\{\sup_{0 \leqslant t \leqslant 1} |B_t^\circ| \leqslant y\right\},$$

称做柯尔莫戈洛夫分布, 其中

$$K(y) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 y^2}, \quad y \geqslant 0. \tag{24}$$

这样, 有如下结果 (柯尔莫戈洛夫)

$$\lim_{N \to \infty} \mathbf{P}\{\sqrt{N} D_N(\omega) \leqslant y\} = K(y), \quad y \geqslant 0. \tag{25}$$

类似的讨论, 可得

$$\lim_{N \to \infty} \mathbf{P}\{\sqrt{N} D_N^+(\omega) \leqslant y\} = \mathbf{P}\left\{\sup_{0 \leqslant t \leqslant 1} B_t^\circ \leqslant y\right\}. \tag{26}$$

统计量 $\sup_{0 \leqslant t \leqslant 1} B_t^\circ$ 的分布, 比 $\sup_{0 \leqslant t \leqslant 1} |B_t^\circ|$ 的分布简单. 统计量 $\sup_{0 \leqslant t \leqslant 1} B_t^\circ$ 的分布为 (见 [5, 第二章, §11]):

$$\mathbf{P}\left\{\sup_{0 \leqslant t \leqslant 1} |B_t^\circ| \leqslant y\right\} = 1 - e^{-2y^2}, \quad y \geqslant 0. \tag{27}$$

有如下结果 (H. B. 斯米尔诺夫)

$$\lim_{N \to \infty} \mathbf{P}\{\sqrt{N} D_N^+(\omega) \leqslant y\} = 1 - e^{-2y^2}, \quad y \geqslant 0. \tag{28}$$

6. 试验与实际的一致性准则 现在讨论, 由 (25) 式的结果 (其中 $K(y)$ 决定于 (24) 式), 可以建立试验与实际一致性的准则. 为此, 我们首先给出分布函数 $K(y)$ 的一张不大的数值表:

y	$K(y)$	y	$K(y)$	y	$K(y)$
		1.10	0.822 282	2.10	0.999 705
0.28	0.000 001	1.20	0.887 750	2.20	0.999 874
0.30	0.000 009	1.30	0.931 908	2.30	0.999 949
0.40	0.002 808	1.40	0.960 318	2.40	0.999 980
0.50	0.036 055	1.50	0.977 782	2.50	0.999 992 5
0.60	0.135 718	1.60	0.988 048	2.60	0.999 997 4
0.70	0.288 765	1.70	0.993 828	2.70	0.999 999 0
0.80	0.455 857	1.80	0.996 932	2.80	0.999 999 7
0.90	0.607 270	1.90	0.998 536	2.90	0.999 999 90
1.00	0.730 000	2.00	0.999 329	3.00	0.999 999 97

如果 N 充分大, 则可以认为 $K(y)$ 给出 $\mathbf{P}\{\sqrt{N} D_N(\omega) \leqslant y\}$ 值的很好的近似值.

自然, 如果按 $\xi_1(\omega), \cdots, \xi_N(\omega)$ 的经验值计算的变量 $\sqrt{N} D_N(\omega)$ 的值充分大, 则应当否定关于 "这些变量的 (假设的) 概率分布是 (连续型) 分布函数 $F = F(x)$" 的假设.

由上面表中的数值, 可以得出关于 "这样得到的结论的可靠程度" 的印象. 例如, 若 $\sqrt{N} D_N(\omega) > 1.80$, 则 (由于 $K(1.80) = 0.996\,932$) 可以认为事件 $\{\sqrt{N} D_N(\omega) > 1.80\}$ 的概率大致等于 $0.003\,068 (= 1.000\,000 - 0.996\,932)$. 假如认为 "具有这样小的概率 ($= 0.003\,068$)" 的事件实际上出现的可能性很小, 则可以得出结论: 关于 "分布为 $\mathbf{P}\{\xi_1 \leqslant x\} = F(x)$" 的假设应当否定, 其中恰好是按函数 $F(x)$ 计算的 $\sqrt{N} D_N(\omega)$

的值. 而假如 $\sqrt{N}D_N(\omega)\leqslant 1.80$, 则 (根据大数定律) 可以认为在 1 000 000 次这样的情形下, 大致有 996 932 次 "经验与理论" 是一致的.

注 需要着重强调, 在运用基于 "柯尔莫戈洛夫分布和斯米尔诺夫分布" 的一致性准则时, (用于检验的) 分布函数 $F=F(x)$ 是完全确定的. 譬如, 只知道分布函数 $F=F(x)$ 属于分布函数 $G(x;\theta)$ 的某个参数族 $\mathbb{G}=\{G=G(x;\theta);\theta\in\Theta\}$, 而且 $G(x;\theta)$ 依赖于参数 $\theta\in\Theta$. 这时, 要求使用如下检验 "经验数据与符合真实分布函数 $F\in\mathbb{G}$" 途径: 首先根据 N 个观测值建立参数 θ 的估计, 然后计算

$$\hat{D}_N(\omega)=\sqrt{N}\sup_{x\in\mathbb{R}}|F_N(x;\omega)-G(x;\hat{\theta}_N(\omega))|,$$

并且像前面的例子一样采取相应的决定. 遗憾的是, 分布函数 $G_N(x;\hat{\theta}_N(\omega))$ 是随机的, 而统计量 $\hat{D}_N(\omega)$ 一般不再是柯尔莫戈洛夫分布.

因此, 在这种情形下, 关于如何检验假设 $F\in\mathbb{G}$, 参阅 [132].

7. 练习题

1. 证明 (18) 式.
2. 证明, 如果 (在空间 $(D,\mathscr{D},\rho)$ 中) $P^{(N)}\xrightarrow{w}P$, 则 $f(X^{(N)})\xrightarrow{d}f(X)$.
3. 证明蕴涵关系 (22).

图书文献资料

(第一章 ~ 第三章)

序言

在托德汉特 (I. Todhunter) 的专著 [68] 中, 陈述了拉普拉斯之前概率论的历史. 发表在《论文集》[45] 中的, 格涅坚科和舍伊宁 (O. B. Шейнин) 论文中, 介绍了从拉普拉斯到 19 世纪末的时期. 在斯蒂格列尔 (M. S. Stigler) 的书 [122] 中, 相当详细地介绍了 1900 年以前的概率论和数理统计史. 在迈斯特罗夫 (Л. Е. Майстров) 的书 [44] 中, 概率论的历史, 从概率论出现一直叙述到 20 世纪 30 年代. 在格涅坚科的教科书 [15] 中有概率论历史简介. 关于概率论许多术语的起源, 可参见亚历山大罗娃 (Н. В. Александрова) [2].

关于概率论的基本概念, 参见下列作者的书籍: 柯尔莫戈洛夫 [32], 格涅坚科 [15], 博罗夫科夫 [7], 格涅坚科和辛钦 [17], А. М. 亚格洛姆和 И. М. 亚格洛姆 (А. М. Яглом 和 И. М. Яглом) [83], 普罗霍罗夫和罗扎诺夫的手册 [56], 手册 [66] 以及译自英文的书: 费勒 [69], 奈曼 (Ю. Нейман) [51], 洛埃甫 (M. Loève) [42], 杜布 (J. L. Doob) [30]. 这里还要指出包含大量概率论题目的梅沙尔金 (Л. Д. Мешалкин) 和基辅 "高等学校" 出版的习题集 [46] 和 [67]①.

在编写这一部教学参考书时, 作者使用了各种不同的文献. 在英文教材和教学指南中, 我们特别指出如下图书: 布雷曼 (L. Breiman) [8], 比林斯利 (P. Billingsley) [106], 阿什 (R. B. Ash) [80], [81], 阿什和加德纳 (M. F. Gardner) [82], 达雷特 (R. Durrett) [108]~[110], 兰珀蒂 (J. Lamperti) [37]. 本书的作者认为这些文献是成功传授知识的

①乌克兰文版书——译者

典范.

关于概率论和数理统计方面的有益的参考资料, 读者可以引用《苏联大百科全书》,《苏联小百科全书》和《数学百科全书》① [121], 以及百科全书:《概率论和数理统计》[125].

于 1956 年创刊的杂志《概率论和数理统计》(出版社 "Наука"), 是俄罗斯出版的、概率论和数理统计方面的主要的刊物.

由全联盟科学与信息研究所 (莫斯科 ВИНИТИ)《文摘杂志》(《Реферативный журнал》), 刊登俄罗斯以及其他国家发表的 "概率论和数理统计方面" 论文的摘要.

有关概率统计大部分实际应用需要数值表, 可以使用博利舍夫 (Л. Н. Большев) 和斯米尔诺夫 (Н. В. Смирнов) 编著的《数理统计表》[6]. 在当前广泛使用电子计算机技术的条件下, 最好利用统计软件 MINITAB, SPSS, ··· 人们广泛采用 "Mathematica®" (见书 [126]).

第一章

§1. 关于概率模型的建立, 参见柯尔莫戈洛夫的论文 [31], 格涅坚科的书 [15]. 涉及 "按箱分配质点" 类型的问题大量的资料, 参见科尔钦 (В. Ф. Колчин), 塞伐斯季雅诺夫和奇斯佳科夫 (И. В. Чистяков) 的书 [34].

§2. 关于在统计物理中出现的概率模型 (特别, 一维伊金格 (Ezenger) 模型), 例如, 可以参见伊西哈尔 (A. Исихар) 的书 [25].

§3. 贝叶斯公式和定理, 是数理统计中所谓贝叶斯方法的基础. 例如, 可以参见德格鲁特 (M. H. deGroot) [18], 邦扎克斯 (S. Zacks) [22].

§4. 关于随机变量及其概率特征, 可以参见习题集 [46] 和 [67].

§5. J. 伯努利引进的大数定律的组合证明, 例如可以参见 [69,T. I]. 关于大数定律的试验解释, 见柯尔莫戈洛夫的论文 [31].

§6. 有关局部和积分定理, 以及泊松定理中的偏差, 见博罗夫科夫 [7] 和普罗霍罗夫 [54].

§7. 这里叙述的内容, 是通过伯努利概型的例子, 演示某些基本概念和数理统计方法. 关于详细资料, 例如参见克拉默 (H. Cramér) [35] 和范德瓦尔登 (B. L. van der Waerden) [10].

§8. 通过考虑关于分割的条件概率和条件数学期望, 可以更好地掌握下面将要引进的, 较为复杂的关于 $\sigma-$ 代数的条件概率和条件数学期望的概念.

§9. 实际上, 这里所引进的 "破产问题" 的形式拉普拉斯就已经考虑过. 关于这一问题亦可参见文集 [45] 中格涅坚科和舍甫宁的论文. 在费勒的书 [69,T.I] 中有该

① 中译本:《数学百科全书》, 第一卷 ~ 第五卷, 科学出版社, 1994 — 2000 年, 北京. —— 译者

问题的丰富的资料.

§10. 这里的内容基本上是按费勒的书 [69] 叙述的. 在论文 [19] 中有关系式 (10) 和 (11) 的证明.

§11. 在杜布的书 [20] 中, 详细地阐述了鞅论. 例如可以在费勒的书 [69,T.I] 中找到 "表决" 定理的其他证明.

§12. 马尔可夫链大量资料包含在下列书中: 费勒 [69,T.I], 邓肯 [21], 邓肯和尤什克维奇 (А. П. Юшкевич), 钟开莱 [75], [120], 雷夫尤兹 (Д. Revuz), 凯麦尼 (J. G. Kemeny) 和斯奈尔 (J. L. Snell) [27], 萨雷姆萨科夫 (Т. А. Сарымсаков) [61], 西拉日季诺夫 (С. Х. Сираждинов) [64], 塞伐斯季雅诺夫 (Б. А. Севастьянов) 的专著 [62]《分枝过程》.

第二章

§1. 关于柯尔莫戈洛夫公理化体系见 [32].

§2. 关于 $\sigma-$ 代数的资料, 亦见下列图书: 柯尔莫戈洛夫和福明 (С. В. Фомин) [33], 奈维尤 (J. Neveu) [49], 布赖曼 [8], 阿什 [81].

§3. 卡拉泰奥多里定理的证明, 参见下列书: 洛埃甫 (M. Loève) [42], 哈尔默斯 (P. R. Halmos) [70].

§4 ~ §5. 在哈尔默斯的书 [70] 中大量关于可测函数的资料.

§6. 亦见下列图书: 柯尔莫戈洛夫和福明 [33], 哈尔默斯 [70], 阿什 [81]. 在这些书中还有拉东 – 尼科迪姆定理的证明. 有时, 把

$$\mathbf{P}\{|\xi| \geqslant \varepsilon\} \leqslant \frac{\mathbf{E}\xi^2}{\varepsilon^2}$$

称做切比雪夫不等式, 而把

$$\mathbf{P}\{|\xi| \geqslant \varepsilon\} \leqslant \frac{\mathbf{E}|\xi|^r}{\varepsilon^r}, \quad r > 0,$$

称做马尔可夫不等式.

§7. 关于 $\sigma-$ 代数的条件概率和条件数学期望的定义, 是柯尔莫戈洛夫引进的 [32]. 在布赖曼 [8] 和阿什 [81] 的书中, 有所考虑问题的广泛资料.

§8. 亦见下列图书: 博罗夫科夫 [7], 阿什 [81], 克拉默 [35], 格涅坚科 [15].

§9. 柯尔莫戈洛夫关于 "具有给定有限维分布过程的存在性" 的定理见他的专著 [32]. 关于图尔恰 (I. Tulcea) 定理, 亦见奈维尤 [49], 阿什 [81]. 这里进行的证明遵循 [81].

§10 ~ 11. 亦见柯尔莫戈洛夫和福明 [33], 阿什 [81], 杜布 [20], 洛埃甫 [42].

§12. 特征函数理论在许多图书上都有, 例如: 格涅坚科 [15], 柯尔莫戈洛夫和格涅坚科 [16], 拉曼钱德兰 (B. Ramachandran) [57]. 所介绍的矩和半不变量的联系按照列昂诺夫 (В. П. Леонов) 和施利亚耶夫的论文 [40].

§13. 亦见下列图书: 易卜拉给莫夫 (И. А. Ибрагимов), 罗扎诺夫 [24], 布赖曼 [8], 列普彩尔 (Р. Ш. Липцер) 和施利亚耶夫 [41], 格里米特 (G. R. Grimmett) 和斯特扎克 (D. R. Stirzaker) [105], 兰珀蒂 [37].

第三章

§1. 关于概率测度和分布的弱收敛, 在下列书中有详细的陈述: 格涅坚科和柯尔莫戈洛夫 [16], 比林斯利 [5].

§2. 普罗霍罗夫定理在他的论文 [55] 中.

§3. 格涅坚科和柯尔莫戈洛夫的专著 [16] 的内容中, 有概率论的极限定理证明中的特征函数方法. 亦见比林斯利 [5]. 练习题第 2 题包括: J. 伯努利大数定律, 泊松大数定律, 假设满足条件: $\xi_1, \xi_2, \cdots$ 独立, 有 0 和 1 两个可能值, 但一般有不同的分布, 对 $\mathbf{P}\{\xi_i = 1\} = p_i, \mathbf{P}\{\xi_i = 0\} = 1 - p_i, i \geqslant 1$.

§4. 这里, 在林德伯格条件成立时, 进行独立随机变量之和的、中心极限定理的传统证明. 对照 [16],[92].

§5. 在没有极限可忽略的条件下, 中心极限定理成立的条件问题, 曾引起 P. 列维的注意. 在佐洛塔廖夫 (В. М. Золотарёв) 专著 [88] 中详细地介绍了, 极限定理理论在非经典提法下的现状. 罗塔里 (В. И. Ротарь) [96] 给出了定理 1 的提法和证明.

§6. 这一节利用了下列图书的资料: 格涅坚科和柯尔莫戈洛夫的专著 [16], 阿什 [82], 彼得罗夫 (В. В. Петров) [53],[92].

§7. 莱维—普罗霍罗夫度量, 是普罗霍罗夫在其著名的著作 [55] 中引进的. 定义在度量空间上的、测度之弱收敛的可度量性的结果也属于普罗霍罗夫. 关于度量 $\|P - \tilde{P}\|_{BL}^{*}$, 参见达德利 (R. M. Dudley) [85] 和波拉德 (D. Pollard) 的书 [93].

§8. 定理 1 属于斯科罗霍德 (А. В. Скороход). 关于 "一个概率空间方法" 的有益资料, 可以在下列文献在找到: 博罗夫科夫 1999 年教学参考书, 专著波拉德 [43].

§9 ~ 10. 我们列举包含 "关于所涉及问题" 大量资料的一系列图书: 扎克德 (J. Jacod) 和施利亚耶夫 [87], 卡姆 (L. Le Cam) [89], 格林伍得 (P. E. Greenwood) 和施利亚耶夫 [84], 莱斯 (F. Liese) 和威伊达 (I. Vajda) [90].

§11. 在彼得罗夫的专著 [92] 中, 有关于中心极限定理收敛速度估计的大量资料. 所引用的贝里和埃森定理的证明, 包含在格涅坚科和柯尔莫戈洛夫的专著 [16] 中.

§12. 证明借鉴普雷斯曼 (I. L. Pressman) 的文章 [94].

§13. 关于数理统计的基本定理的补充资料, 参见 [8],[35],[58],[106],[107].

参考文献

[1] Александров П. С. Введение в общую теорию множеств и функций. — М.: Гостехиздат, 1948.

[2] Александрова Н. В. Математические термины. — М.: Высшая школа, 1978.

[3] Бернштейн С. Н. О работах П. Л. Чебышева по теории вероятностей//Научное наследие П. Л. Чебышева. Вып. 1.Математика. 1945. — С. 59 — 60.

[4] Бернштейн С. Н. Теория вероятностей. — 4-е изд. — М.: Гостехиздат, 1946.

[5] Биллингсли П. Сходимость вероятностных мер. — М.: Наука, 1977.

[6] Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. — 3-е изд. — М.: Наука, 1983.

[7] Боровков А. А. Теория вероятностей. — 3-е изд. — М.:УРСС, 1999.

[8] Брейман (Breiman L.). Probability. — Reading, MA: Addison-Wesley, 1968.

[9] Вальд А. Последовательный анализ. — М.: Физматгиз, 1960.

[10] Ван дер Варден Б. Л. Математическая статистика. — М.: ИЛ, 1960.

[11] Вентцель А. Д. Курс теории случайных процессов. — М.: Наука, 1975.

[12] Гарсиа (Garsia A. M.). A simple proof of E. Hopf's maximal ergodic theorem // Journal of Mathematics and Mechanics. — 1965. — V 14, № 3. — P. 381 — 382.

[13] Гихман И. И., Скороход А. В. Введение в теорию случайных процессов. — М.: Наука. 1977.

[14] Гихман И. И., Скороход А. В. Теория случайных процессов: В 3 т. — М.: Наука, 1971 — 1975. 中译本:《随机过程论》, 第一卷 (邓永录等译), 第二卷 (冯概容译). 1986.

[15] Гнеденко Б. В. Курс теории вероятностей. —6-е изд. — М.: Наука, 1988. 中译本:《概率论教程》(丁寿田), (第三版). 高等教育出版社, 1961.

[16] Гнеденко Б. В., Колмогоров А. Н. Предельные распределения для сумм независимых случайных величин. — М.; Л.: Гостехиздат, 1949. 中译本:《相互独立随机变量之和的极限分布》(王寿仁译), 科学出版社, 1950.

[17] Гнеденко Б. В., Хинчин А. Я. Элементарное введение в теорию вероятностей. — 9-е изд. — М.: Наука, 1982.

[18] Де Гроот М. Оптимальные статистические решения. — М.: Мир, 1974.

[19] Доэрти (Doherty M.). An amusing proof in fluctuation theory// Combinatorial Mathematics, III: Proceedings of the Third Australian Conference, Univ. Queensland, St. Lucia 1974. — Berlin etc.: Springer-Verlag, 1975. — P. 101—104. — (Lecture Notes in Mathematics; V.452.)

[20] Дуб Дж. Л. Вероятностные процессы. — М.: ИЛ, 1956.

[21] Дынкин Е. Б. Марковские процессы. — М.: Физматгиз, 1963.

[22] Закс Ш. Теория статистических выводов. — М.: Мир, 1975.

[23] Ибрагимов И. А., Линник Ю. В. Независимые и стационарно связанные величины. — М.: Наука, 1965.

[24] Ибрагимов И. А., Розанов Ю. А. Гауссовские случайные процессы. — М.: Наука, 1970.

[25] Исихара А. Статистическая физика. — М.: Мир, 1973.

[26] Кабанов Ю. М., Липцер Р. Ш., Ширяев А. Н. К вопросу об абсолютной непрерывности и сингулярности вероятностных мер// Математический сборник. — 1977. — Т. 104, № 2. — С. 227 — 247.

[27] Кемени Дж., Снелл Дж. Конечные цепи Маркова. — М.: Наука, 1970.

[28] Колмогоров А. Н. Цепи Маркова со счетным числом возможных состояний//Бюллетень МГУ. — 1937. — Т.1, № 3. — С. 1 — 16.

[29] Колмогоров А. Н. Стационарные последовательности в гильбертовском пространстве//Бюллетень МГУ. — 1941. — Т. 2, № 6. — С. 1 — 40.

[30] Колмогоров А. Н. Роль русской науки в развитии теории вероятностей//Ученые записки МГУ. — 1947. — Вып. 91. — С. 53 — 64. 中译本:《概率论》(见《数学, 它的内容、方法和意义》(卷2)), 科学出版社, 1961.

[31] Колмогоров А. Н. Теория вероятностей//Математика, ее содержание, методы и значение.—М.: Изд-во АН СССР, 1956. — Т. II, — С. 252 — 284.

[32] Колмогоров А. Н. Основные понятия теории вероятностей. — М.; Л.: ОНТИ, 1936; 2-е изд. М.: Наука, 1974; 3-е изд. М.: Фазис, 1998.

[33] Колмогоров А. Н., Фомин С. В. Элементы теории функций и функционального анализа. — 6-е изд. —М.: Наука, 1989. 中译本:《概率论的基本概念》(丁寿田译), 商务印书馆, 1952.

[34] Колчин В. Ф., Севастьянов Б. А., Чистяков В. П. Случайные размещения. — М.: Наука, 1976.

[35] Крамер Г. Математические методы статистики. — 2-е изд. — М.: Мир, 1976. 中译本:《统计学数学方法》(魏宗舒译), 上海科技出版社, 1983.

[36] Кубилюс И. Вероятностные методы в теории чисел. — Вильнюс: Гос. изд-во полит. и науч. лит. ЛитССР, 1959.

[37] Ламперти Дж. Вероятность. — М.: Наука, 1973.

[38] Ламперти (Lamperti J.). Stochastic Processes. — Aarhus Univ., 1974. — (Lecture Notes Series; № 38).

[39] Ленглар (Lenglart E.). Relation de domination entre deux processus// Annales de l'Institut H. Poincaré Sect. B. (N. S.). — 1977. — V.13, № 2. — P. 171 — 179.

[40] Леонов В. П., Ширяев А. Н. К технике вычисления семиинвариантов// Теория вероятностей и ее применения. — 1959. — Т. IV, вып. 2. — С. 342 — 355.

[41] Липцер Р. Ш., Ширяев А. Н. Статистика случайных процессов. — М.: Наука, 1974. 中译本:《随机过程统计》.

[42] Лоэв М. Теория вероятностей. — М.: ИЛ, 1962. 中译本:《概率论及其应用》, 上册, (梁文骐译), 科学出版社, 1966.

[43] Марков А. А. Исчисление вероятностей. — 3-е изд. — СПб., 1913.

[44] Майстров Л. Е. Теория вероятностей (исторический очерк). — М.: Наука, 1967.

[45] Математика XIX века/ Под ред. А. Н. Колмогорова и А. П. Юшкевича. — М.: Наука, 1978.

[46] Мешалкин Л. Д. Сборник задач по теории вероятностей. — М.: Изд-во МГУ, 1963. 中译本:《概率论习题集》(盛骤等译), 高等教育出版社, 1984.

[47] Мейер (Meyer P.-A.). Martingales and Stochastic Integrals. I. — Berlin etc.: Springer-Verlag, 1972. — (Lecture Notes in Mathematics; V. 284).

[48] Мейер П.-А. Вероятность и потенциалы. — М.: Мир, 1973.

[49] Неве Ж. Математические основы теории вероятностей. — М.: Мир, 1969.

[50] Невё (Neveu J.). Discrete-Parameter Martingales. — Amsterdam etc.: North-Holland, 1975.

[51] Нейман Ю. Вводный курс теории вероятностей и математической статистики. — М.: Наука, 1968.

[52] Новиков А. А. Об оценках и асимптотическом поведении вероятностей непересечения подвижных границ суммами независимых случайных величин//Известия АН СССР. Серия математическая. — 1980. — Т. 40, вып. 4. — С. 868—885.

[53] Петров В. В. Суммы независимых случайных величин. — М.: Наука, 1972

[54] Прохоров Ю. В. Асимптотическое поведение биномиального распределения// Успехи математических наук. — 1953. — Т. VIII, вып 3(55). — С. 135—142.

[55] Прохоров Ю. В. Сходимость случайных процессов и предельные теоремы теории вероятностей// Теория вероятностей и ее применения. — 1956. — Т. I, вып. 2. — С. 177—238.

[56] Прохоров Ю. В., Розанов Ю. А. Теория вероятностей. — 2-е изд. — М.: Наука, 1973.

[57] Рамачандран Б. Теория характеристических функций. — М.: Наука, 1975.

[58] Реньи (Rényi A.) Probability Theory. — Amsterdam: North-Holland, 1970.

[59] Роббинс Г., Сигмунд Д., Чао И. Теория оптимальных правил остановки. — М.: Наука, 1977.

[60] Розанов Ю. А. Стационарные случайные процессы. — М.: Физматгиз, 1963.

[61] Сарымсаков Т. А. Основы теории процессов Маркова. — М.: Гостехиздат, 1954.

[62] Севастьянов Б. А. Ветвящиеся процессы. — М.: Наука, 1971.

[63] Синай Я. Г. Введение в эргодическую теорию. — Ереван: Изд-во Ереван. ун-та, 1973.

[64] Сираждинов С. Х. Предельные теоремы для однородных цепей Маркова. — Ташкент: Изд-во АН УзССР, 1955.

[65] Справочник по теории вероятностей и математической статистике/ Под ред. В. С. Королюка. — Киев: Наукова думка, 1978.

[66] Стаут (Stout W. F.). Almost Sure Convergence. — New York etc.: Academic Press, 1974.

[67] Теорія ймовірностей. — Київ: Вища школа, 1976.

[68] Тодхантер (Todhunter I.). A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace. — London: Macmillan, 1865.

[69] Феллер В. Введение в теорию вероятностей и ее приложения: В 2 х т. — М.: Мир, 1984. 中译本:《概率论导引及其应用》(胡迪鹤译), 人民邮电出版社, 2006.

[70] Халмош П. Теория меры. — М.: ИЛ, 1953. 中译本:《测度论》(王建华译), 科学出版社.

[71] Хеннан Э. Анализ временных рядов. — М.: Наука, 1964.

[72] Хеннан Э. Многомерные временные ряды. — М.: Мир, 1974.

[73] Чао, Тейчер (Chow Y. S., Teicher H.). Probability Theory. Independence, Interchangeability, Martingales. — 3rd ed. — New York: Springer-Verlag, 1997.

[74] Чебышев П. Л. Теория вероятностей: Лекции акад. П. Л. Чебышева, читанные в 1879, 1880 гг./ Издано А. Н. Крыловым по записи А. М. Ляпунова. — М.; Л., 1936.

[75] Чжун Кай-лай. Однородные цепи Маркова. — М.: Мир, 1964.

[76] Ширяев А. Н. Случайные процессы. — М.: Изд-во МГУ, 1972.

[77] Ширяев А. Н. Вероятность, статистика, случайные процессы: В 2-х т. — М.: Изд-во МГУ, 1973 — 1974.

[78] Ширяев А. Н. Статистический последовательный анализ. — 2-е изд. — М.: Наука, 1976.

[79] Энгельберт, Ширяев (Engelbert H.-J., Shiryaev A. N.). On the sets of convergence of generalized submartingales // Stochastics. — 1979. — V. 2, № 3. — P. 155 — 166.

[80] Эш (Ash R. B.). Basic Probability Theory. — New York etc.: Wiley, 1970.

[81] Эш (Ash R. B.). Real Analysis and Probability. — New York etc.: Academic Press, 1972.

[82] Эш, Гарднер (Ash R. B., Gardner M. F.). Topics in Stochastic Processes. — New York etc.: Academic Press, 1975.

[83] Яглом А. М., Яглом И. М. Вероятность и информация. — 3-е изд. — М.: Наука, 1973. 中译本:《概率与信息》，科学出版社.

[84] Гринвуд, Ширяев (Greenwood P. E., Shiryayev A. N.). Contiguity and the Statistical Invariance Principle. — London: Gordon & Breach, 1985.

[85] Дадли (Dudley R. M.)Distances of probability measures and random variables// Annals of Mathematical Statistics. — 1968. — V.39, № 5. — P. 1563 — 1572.

[86] Дакуна-Кастелль, Дюфло (Dacunha-Castelle D., Duflo M.). Probabilités et statistiques: 1,2. — Paris: Masson. — 1: Problèmes à temps fixe. — 1982; 2: Problèmes à temps mobile. — 1983. — Перев. на англ. яз.: Probability and Statistics: V. I, II. — Berlin etc.: Springer-Verlag, 1986.

[87] Жакод Ж., Ширяев А. Н. Предельные теоремы для случайных процессов: В 2-х т. — М.: Физматлит, 1994.

[88] Золотарев В. М. Современная Теория суммирования независимых случайных величин. — М.: Наука, 1986.

[89] Ле Кам (Le Cam L.). Asymptotic Methods in Statistical Decision Theory. — Berlin etc.: Springer-Verlag, 1986.

[90] Лизе, Вайда (Liese F., Vajda I.). Convex Statistical Distances. — Leipzig: Teubner, 1987.

[91] Липцер Р. Ш., Ширяев А. Н. Теория мартингалов. — М.: Наука, 1986

[92] Петров В. В. Предельные теоремы для сумм независимых случайных величин. — М.: Наука, 1987.

[93] Поллард (Pollard D.). Convergence of Stochastic Processes. — Berlin etc.: Springer-Verlag, 1984.

[94] Пресман Э. Л. О сближении по вариации распределения суммы независимых бернуллиевских величин с пуассоновским законом// Теория вероятностей и ее применения. — 1985. — Т. XXX, вып. 2. — С. 391 — 396.

[95] Розанов Ю. А. Теория вероятностей, случайные процессы и математическая статистика. — М.: Наука, 1985.

[96] Ротарь В. И. К обобщению теоремы Линдеберга — Феллера// Математические заметки. — 1975. — Т.18, вып. 1. — С. 129 — 135.

[97] Севастьянов Б. А. Курс теории вероятностей и математической статистики. — М.: Наука, 1982.

[98] Ширяев (Shiryayev A. N.) Probability. — 2nd ed. — Berlin etc.: Springer-Verlag, 1995.

[99] Ширяев (Shirjayev A. N.) Wahrscheinlichkeit. — Berlin: VEB Deutscher Verlag der Wissenschaften, 1988.

[100] Ширяев А. Н. Основы стохастической финансовой математики: В 2-х т. — М.: ФАЗИС, 1998.

[101] Фёллмер, Проттер, Ширяев (Föllmer H., Protter Ph., Shiryaev A. N.). Quadratic covariation and an extension of Itô's formula//Bernoulli. — 1995. — V 1, № 1/2. — P. 149 — 170.

[102] Дынкин Е. Б., Юшкевич А. А. Теоремы и задачи о процессах Маркова. — М.: Наука, 1967.

[103] Гнеденко, Колмогоров (Gnedenko B. V., Kolmogorov A. N). Limit Distributions for Sums of Independent Random Variables. — Reading, MA, etc.: Addison-Wesley, 1954.

[104] Боровков А. А. Эргодичность и устойчивость случайных процессов. — М.: УРСС, 1999.

[105] Гриммет, Стирзакер (Grimmet G. R., Stirzaker D. R.) Probability and Random Processes. — Oxford: Clarendon Press, 1993.

[106] Биллингсли (Billingsley P.). Probability and Measure. — 3rd ed. — New York: Wiley, 1995.

[107] Боровков А. А. Математическая статистика. — М.: Наука, 1984.

[108] Дарретт (Durrett R.). Probability: Theory and Examples. — Pacific Grove, CA: Wadsworth & Brooks/Cole, 1991.

[109] Дарретт (Durrett R.). Stochastic Calculus. — Boca Raton, FL: CRC Press, 1996.

[110] Дарретт (Durrett R.). Brownian Motion and Martingales in Analysis. — Belmont, CA: Wadsworth International Group, 1984.

[111] Калленберг (Kallenberg O.). Foundations of Modern Probability. — 2nd ed. — New York: Springer-Verlag, 2002.

[112] Карлин, Тейлор (Karlin S., Taylor H. M.). A First Course in Stochastic Processes. — 2nd ed. — New York etc.: Academic Press, 1975.

[113] Кашин Б. С., Саакян А. А. Ортогональные ряды. — 2-е изд. — М.: АФЦ, 1999.

[114] Жакод, Проттер (Jacod J., Protter Ph.). Probability Essentials. — Berlin etc.: Springer-Verlag, 2000.

[115] Нётс (Neuts M. F.) Probability. — Boston, MA: Allyn & Bacon, 1973.

[116] Плато (Plato J.). Creating Modern Probability. — Cambridge: Cambridge Univ. Press, 1998.

[117] Ревюз Д. Цепи Маркова. — М.: РФФИ, 1997.

[118] Уильямс (Williams D.). Probability with Martingales. — Cambridge: Cambridge Univ. Press, 1991.

[119] Холл, Хейде (Hall P., Heyde C. C.). Martingale Limit Theory and Its Applications. — New York etc.: Academic Press, 1980.

[120] Чжун Кай Лай (Chung Kai Lai). Elementary Probability Theory with Stochastic Processes. — 3rd ed. — Berlin etc.: Springer-Verlag, 1979. 中译本:《初等概率论和随机过程》(吕乃刚等译), 人民教育出版社.

[121] Математическая энциклопедия: В 5 т./Гл. ред. И. М. Виноградов. — М.: Советская энциклопедия, 1977 — 1985. 中译本:《数学百科全书》, 第一卷～第五卷, 科学出版社, 1984 — 2000.

[122] Стиглер (Stigler S. M.). The History of Statistics: The Measurement of Uncertainty Before 1900. — Cambridge: Belknap Press of Harvard Univ. Press, 1986.

[123] Гербер Х. Математика страхования жизни. — М.: Мир, 1995.

[124] Эренфесты П. и Т. (Ehrenfest P., Ehrenfest T.). Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem//Physikalische Zeitschrift. — 1907. — V.8. — P. 311 — 314.

[125] Теория вероятностей и математическая статистика: энциклопедия/Гл. ред. Ю. В. Прохоров. — М.: Большая Российская энциклопедия, 1999.

[126] Вольфрам (Wolfram S.). The Mathematica® Book. — 4th ed. — Champaign: Cambridge: Wolfram Media, Cambridge Univ. Press, 1999.

[127] Дуб (Doob J. L.). What is a martingale?//The American Mathematical Monthly. — 1971. — V.78. — P. 451 — 463.

[128] Синай Я. Г. Курс теории вероятностей. — М.: Изд-во МГУ, 1985. — 2-е изд., 1986.

[129] Синай (Sinai Ya. G.). Topics in Ergodic Theory. — Princeton, NJ: Princeton Univ. Press, 1993. — (Princeton Mathematical Series; V. 44.)

[130] Вальтерс (Walters P.) An Introduction to Ergodic Theory. — New York etc.: Springer-Verlag, 1982.

[131] Булинский А. В., Ширяев А. Н. Теория случайных процессов. — М.: Физматлит, 2003.

[132] Хмаладзе Э. В. Мартингальный подход в теории непараметрических критериев согласия// Теория вероятностей и ее применения. — 1981. — Т. XXVI, вып. 2. — С. 246 — 265.

[133] Гамильтон (Hamilton J. D.). Time Series Analysis. — Princeton, NJ: Princeton Univ. Press, 1994.

[134] Бернулли Я. О законе больших чисел. — Ч. 4: Искусство предположений. — М.: Наука, 1986.

[135] Лукач Е. Характеристические функции. — М.: Наука, 1979.

[136] Хренников (Khrennikov A.). Interpretations of Probability. — Utrecht: VSP, 1999.

名词索引

(汉语拼音为序)

C

D

H

J

K

L

M

N

P

Q

人名表

(汉语拼音为序)

A

阿什	R. B. Ash	Р. Эш
埃尔米特	Ch. Hermite	Ш. Эрмит
埃森	C. G. Esseen	К. Г. Эссеен
爱因斯坦	A. Einstein	А. Эйнштейн

B

巴拿赫	S. Banach	С. Банах
邦弗尔罗尼	Bonferroni	Бонферрони
鲍斯	S. N. Bose	Ш. Бозе
贝尔	A. G. Bell	А. Г. Белл
贝尔	R. L. Baire	Р. Л. Бэр
贝里	A. C. Berry	А. С. Берри
贝塞尔	F. W. Bessel	Ф. В. Бессель
贝叶斯	T. Bayes	Т. Байес
比林斯利	P. Billingsley	П. Биллингсли
彼得罗夫	W. W. Petrow	В. В. Петров
毕达哥拉斯	Pythagoras	Пифагор
波拉德	D. Pollard	Д. Пполлард
波利亚	G. Polya	Д. Пойа

伯恩斯坦	S. N. Bernstein	Н Бернштейн
伯克霍夫	C. D. Birkhoff	Дж. Д. Биркгоф
伯努利	J. Benoulli	Я. Бернулли
泊松	S. D. Poisson	С. Д. Пуассон
博赫纳	S. Bochner	С. Бохнер
博雷尔 (波莱尔)	E. Borel	Э. Борель
博利舍夫	L. N. Bol'shev	Л. Н. Большев
博罗夫科夫	A. A. Borowkov	А. А. Боровков
布尔	G. Boole	Дж. Буль
布耳兹曼	L. Boltzmann	Л. Больцман
布莱克韦尔	D. H. Blackwell	Л. Блэкуэлл
布赖曼	L. Breiman	Л. Брейман
布朗	E. T. Brown	Э. Т. Броун
布雷	J. R. Bray	Я. Р. Брей
布洛赫	A. Bloch	А. Блох
布尼亚科夫斯基	A. J. Buniakowsky	А. Я. Буняковский
布西	R. S. Busey	Р. С. Бьюсу

C~D

查普曼	D. G. Chapman	Д. Г. Чепман
达布	J. G. Darboux	Ж. Г. Дарбу
达德利	R. M. Dudley	Р. Дадли
达雷特	R. Durrett	Р. Даррет
德格鲁特	M. H. deGroot	М. Де Гроот
邓肯	E. B. Dynkin	И. Б. Дынкин
狄拉克	P. A. M. Dirac	П. А. М. Дирак
狄利克雷	P. G. L. Dirichlet	П. Г. А. Дирихле
笛卡尔	R. Descartes	Р. Декарт
棣莫弗	A. De Moivre	А. Дё Муавр
杜布	J. L. Doob	Дж. Л. Дуб

F

法图	F. Fatou	П. фату
范德瓦尔登	B. L. van der Waerden	Б. Л. Ван дер Варден
范基	Fan Ky	Ки Фан
费勒	W. Feller	В. Феллер
费马	P. Fermat	П. Ферма
费希尔	R. A. Fisher	Р. А. Фишер

弗雷歇	M. Fréchet	Ф. Фреше
费米	E. Fermi	Э. Ферми
福明	S. W. Fomin	С. В. Фомин
富比尼	G. Fubini	Г. Фубини
傅里叶	J. B. J. Fourier	Ж. Б. Ж. Фурье

G

冈贝尔	E. J. Gumbel	Э. Гумбель
高斯	C. F. Gauss	К. Ф. Гаусс
哥塞特	W. S. Gosset	В. С. Госсет
克拉默	G. P. Gram	Г. П. Грам
格里米特	G. R. Grimmett	Дж. Гриммет
格里汶科	W. I. Glivenko	В. И. Гливенко
格林伍德	P. E. Greenwood	П. Е. Гринвуд
格涅坚科	B. V. Gnedenko	Б. В. Гнеденко

H

哈恩	H. Hahn	Г. Хан 460
哈尔	A. Haar	А. Хаар
哈尔默斯	P. R. Halmos	П. Халмош
海林格	E. Helinger	Э. Хеллингер
海涅	H. E. Heine	Г. Э. Гейне
赫尔德	O. L. Hölder	О. Л. Гёльдер
赫利	E. Helly	Э. Хелли
惠更斯	Ch. Huyghens	Х. Гюйгенс
霍普夫	H. Hopf	Х. Хопф
霍奇	W. W. D. Hodge	У. В. Д. Ходж

J

加德纳	M. F. Gardner	М. Гарднер
加尔西亚	A. M. Garsia	А. М. Гарсиа

K

卡尔达诺	G. Cardano	Дж. Кардано
卡尔莱曼	T. Carleman	Т. Карлеман
卡库塔尼	S. Kakutani	С. Какутани
卡拉泰奥多里	C. Carathéodory	К. Каратеодори
卡姆	L. Le Cam	Л. Ле Кам
凯麦尼	J. G. Kemeny	Дж. Кемени

坎泰利	F. P. Cantelli	Ф. П. Кантелли
康托尔	G. Cantor	Г. Кантор
柯尔莫戈洛夫	A. N. Kolmogorov	А. Н. Колмогоров
柯西	A. L. Cauchy	О. Л. Коши
科尔钦	W. F. Kerchen	В. Ф. Колчин
克拉默	H. Cramer	Г. Крамер
克罗内克	L. Kronecker	Л. Кронекер
库尔贝克	S. Kullback	С. Кульбак

L

拉奥	C. R. Rao	С. Р. Рао
拉德马赫	H. Rademacher	Г. А. Радемахер
拉东	J. Radon	Дж. Радон
拉格朗日	J. L. Lagrange	Ж. Л. Лагранж
拉曼钱德兰	B. Ramachandran	Б. Рамачандран
拉普拉斯	P. S. Laplace	П. С. Лаплас
莱斯	F. Liese	Ф. Лизе
兰珀蒂	J. Lamperti	Дж. Ламперти
勒贝格	H. L. Lebesgue	А. Л. Лебег
雷夫尤兹	D. Revuz	Д. Ревюз
黎曼	G. F. B. Riemann	Г. Ф. Б. Риман
李雅普诺夫	A. M. Lyapunov	А. М. Ляпунов
里斯	F. Riesz	Ф. Рисс
利普彩尔	R. S. Lipchail	Р. Ш. Липцер
利普希茨	R. O. S. Lipschitz	Р. Липшиц
列昂诺夫	W. P. Leonov	В. П. Леонов
列维	P. P. Lévy	П. П. Леви
林德伯格	J. W. Linderberg	Дж. У. Линдеберг
洛埃甫	M. Loève	М. Лоэв
洛必达	L' Hospital	Г. Лопиталь
罗塔里	W. E. Rotarl	В. И. Ротарь
罗扎诺夫	Yu. A. Rozanov	Ю. А. Розанов

M

马尔可夫	A. A. Markov	А. А. Марков
马钦凯维奇	J. Matcinkiewicz	И. Марцинкевич
马哈拉诺比斯	P. S. Mahalanobis	П. С. Махаланобис
迈斯特罗夫	D. I. Mastrov	Д. Е. Майстров

麦克米兰	B. McMillan	B. Макмиллан
麦克斯韦	J. C. Maxwell	Д. К. Максвелл
曼	H. B. Mann	X. Б. Манн
梅沙尔金	L. D. Meshalkin	Л. Д. Мешалкин
米泽斯	R. Mises	Р. Мизес
闵可夫斯基	H. Minkowski	Г. Минковский
默瑟	J. Mercer	Ж. Мерсер

N

奈曼	Yu. Neyman	Ю. Нейман
奈维尤	J. Neveu	Р. Невё
尼科迪姆	O. M. Nikodym	О. М. Никодим

O

欧几里得	Euclid	Евклид
欧拉	L. Euler	Л. Эйлер

P

帕利	W. Pauli	В. Паули
帕乔利	L. Pacioli	Л. Пачоли
帕塞瓦尔	M. A. Parseval	М. А. Парсеваль
帕斯卡	B. Pascal	Б. Паскаль
庞加莱	J. H. Poincaré	Ж. Ан. Пуанкаре
蒲丰 (布丰)	G. L. L. Buffon	Ж. Л. Л. Бюффон
普拉特	Pratt	Пратт
普雷斯曼	I. L. Pressman	Э. Л. Пресман
普罗霍罗夫	Yu. B. Prokhorov	Ю. В. Прохоров

Q

奇斯佳科夫	W. P. Qisjiakv	В. П. Чистяков
切比雪夫	P. L. Chebyshev	П. Л. Чебышёв
切萨罗	E. Cesàro	Э. Чезаро 335

S

萨雷姆萨科夫	T. A. Sarımsakov	Т. А. Сарымсаков
塞维治	I. R. Sevage	И. Р. Севидж
沙利耶	C. L. Charlier	К. Л. Шарлье
绍德尔	A. Schauder	А. Шаудер
舍伊宁	O. W. Sheining	О. В. Шейнин

施利亚耶夫	A. N. Shiryaev	А. Н. Ширяев
施密特	E. Schmidt	Э. Шмидт
施瓦茨	L. Schwarz	Л. Шварц
斯梯格列尔	S. M. Stigler	С. М. Сти
斯蒂尔切斯	T. J. Stieltjes	Т. И. Стилтьес
斯捷克洛夫	W. A. Steclov	В. А. Стеклов
斯科罗霍德	A. V. Skorokhod	А. В. Скороход
斯莱皮恩	P. Slepian	П. Слепян 394
斯鲁茨基	E. Slutsky	Е. Е. Слуцкий
斯米尔诺夫	N. V. Smirnov	Н. В. Смирнов
斯奈尔	J. L. Snell	Дж. Снелл
斯皮策	F. Spitzer	Ф. Спицер
斯特林	J. Stirling	Дж. Стирлинг
斯特扎克	D. R. Stirzaker	Д. Стирзакер
斯蒂格列尔	M. S. Stigler	С. Стиглер
斯通	M. H. Stone	М. Г. Стоун

T

塔尔塔利亚	N. Tataylia	Н. Тарталья
图尔恰	I. Tulcea	И. Тулча
托德汉特	I. Todhunter	Э. Тодхантер
托德亨特	I. Todhunter	И. Тодхатер

W

威伊达	I. Vajda	И. Вайда
韦布尔	W. Weibull	В. Вейбулл
维尔斯特拉斯	K. T. W. Weierstrass	К. Т. В.
维纳	N. Weiner	Н. Винер
沃尔德	A. Wald	А. Вальд
沃尔夫	R. Wolf	Р. Вольф
乌沙科夫	W. G. Ushakoff	В. Г. Ушаков
伍拉姆	Woollam	Улам

X

西拉吉季诺夫	S. H. Sealarijinov	С. Х. Сираждинов
希尔伯特	D. Hilbert	Д. Гильберт
希奈	Y. G. Sinai	Я. Г. Синай

谢瓦斯契亚诺夫	B. A. Sevastyanov	Б. А. Севастьянов
辛钦	A. J. Khintchine	А. Я. Хинчин
休伊特	E. Hewitt	Э. Хьюитт

Y

雅可比	C. G. J. Jacobi	К. Г. Я. Якоби
亚当斯	J. G. Adams	Д. К. Адамс
亚格洛姆	A. M. Jaglom	А. М. Яглом
亚格洛姆	I. M. Jaglom	И. М. Яглом
亚历山大罗娃	N. W. Alexanderowa	Н. В. Александрова
延森	I. L. Jensen	И. Л. Иенсен
伊金格	Ezinger	Изинг
伊藤	K. Itô	К. Ито
伊西哈尔		А. Исихар
易卜拉金莫夫	E. A. Ibragamov	И. А. Ибрагимов
尤什克维奇	A. P. Yushkaivich	А. П. Юшкевич

Z

扎克德	J. Jacod	Ж. Жакод
扎克斯	S. Zacks	Ш. Закс
钟开莱	Kai Lai Chung	Чжун Кай-лай
祖布科夫	A. M. Zubkov	А. М. Зубков
佐洛塔廖夫	W. M. Zolotaileff	В. М. Золотарёв

常用数学符号

$\mathbb{R}=(-\infty,\infty)$ —— 实数的集合, 实直线, 一维欧几里得空间

$\mathbb{R}_+=[0,\infty)$

$\overline{\mathbb{R}}=[-\infty,\infty]$ —— 扩充实直线: $\overline{\mathbb{R}}=\mathbb{R}\bigcup\{-\infty\}\bigcup\{\infty\}$

$\overline{\mathbb{R}}_+=[0,\infty]$

$\mathbb{Q}$ —— 有理数的集合

$\mathbb{Q}_+=\mathbb{Q}\bigcap\mathbb{R}_+$

$\mathbb{R}^d$ —— d 维欧几里得空间

$\mathbb{N}$ —— 自然数: $\{0,1,2,\cdots\}$ 或 $\{1,2,\cdots\}$

$\mathbb{Z}$ —— 整数的集合: $\{0,\pm1,\pm2,\cdots\}$

$\mathbb{C}$ —— 复数的集合

$$(a,b)=\{x\in\overline{\mathbb{R}}:a<x<b\},\ [a,b]=\{x\in\overline{\mathbb{R}}:a\leqslant x\leqslant b\}$$

$$(a,b]=\{x\in\overline{\mathbb{R}}:a<x\leqslant b\},\ [a,b)=\{x\in\overline{\mathbb{R}}:a\leqslant x<b\}$$

$\inf X$ —— 集合 $X\subset\overline{\mathbb{R}}$ 的下界

$\sup X$ —— 集合 $X\subset\overline{\mathbb{R}}$ 的上界

$\inf\limits_{n\geqslant m}x_n$ —— 集合 $X=\{x_m,x_{m+1},\cdots\}$ 的下界

$\sup\limits_{n\geqslant m}x_n$ —— 集合 $X=\{x_m,x_{m+1},\cdots\}$ 的上界

如果 $x_n\in\mathbb{R}$, $n\geqslant1$, 则

$$\liminf_{n\to\infty}x_n=\underline{\lim_n}\,x_n\equiv\sup_{m\geqslant1}\inf_{n\geqslant m}x_n,\quad \limsup_{n\to\infty}x_n=\overline{\lim_n}\,x_n\equiv\inf_{m\geqslant1}\sup_{n\geqslant m}x_n$$

$$\lim x_n=x\Leftrightarrow\underline{\lim}\,x_n=\overline{\lim}\,x_n=x\Leftrightarrow\underline{\lim}\,x_n\geqslant x\geqslant\overline{\lim}\,x_n.$$

对 于 实 数

$x^+ = \max(x,0)$; $x^- = -\min(x,0)$

$$x^{\oplus} = \begin{cases} x^{-1}, & 若\ x \neq 0, \\ 0, & 若\ x = 0 \end{cases}$$

$x \vee y = \max(x,y)$; $x \wedge y = \min(x,y)$

$[x]$ 或 $\lfloor x \rfloor$ —— 不大于 x 的最大整数

$\lceil x \rceil$ —— 大于或等于 x 的最小整数

$\operatorname{sign} x$ —— 实数 x 的符号:

$$\operatorname{sign} x = \begin{cases} 1, & 若\ x > 0, \\ 0, & 若\ x = 0, \\ -1, & 若\ x < 0 \end{cases}$$

(有时, 当 $x \geqslant 0$ 时, 设 sign $= 1$; 当 $x < 0$ 时, 设 sign $= -1$)

$x_n \to x$, 其中 $n \in \{1,2,\cdots\}$, 表示 $\lim\limits_n x_n = x$

$x_n \uparrow$ 表示 $x_1 \leqslant x_2 \leqslant \cdots$; $x_n \uparrow x$ 表示 $x_n \uparrow$ 且 $\lim\limits_n x_n = x$

$x_n \downarrow$ 表示 $x_1 \geqslant x_2 \geqslant \cdots$; $x_n \downarrow x$ 表示 $x_n \downarrow$ 且 $\lim\limits_n x_n = x$

对于复数 $z = a + ib$, 其中 $a, b \in \mathbb{R}$, 而 $i = \sqrt{-1}$ 是虚单位

$\bar{z} = a - ib$ —— z 的共轭复数

$|z|$ —— z 的模 $(= \sqrt{a^2 + b^2})$

$\operatorname{Re} z$ 和 $\operatorname{Im} z$ —— z 的实部和虚部: $\operatorname{Re} z = a$, $\operatorname{Im} z = b$

对于 d- 维欧几里得空间 $\mathbb{R}^d$

$|x|$ —— $x = (x_1, \cdots, x_d)$ 的欧几里得范数, 即 $\sqrt{x_1^2 + \cdots + x_d^2}$

$x \cdot y$ 或 (x,y) —— $x = (x_1, \cdots, x_d)$ 和 $y = (y_1, \cdots, y_d)$ 的数量积, 即 $x_1y_1 + \cdots + x_dy_d$

集 合 论

$A_n \uparrow$ 表示 $A_1 \subseteq A_2 \subseteq \cdots$; $A_n \uparrow A$ 表示 $A_n \uparrow$ 且 $\bigcup A_n = A$

$A_n \downarrow$ 表示 $A_1 \supseteq A_2 \supseteq \cdots$; $A_n \downarrow A$ 表示 $A_n \downarrow$ 且 $\bigcap A_n = A$

$\limsup A_n$, 或 $\overline{\lim} A_n$, 或 $\bigcap\limits_{m \geqslant 1} \left(\bigcup\limits_{n \geqslant m} A_n \right)$ —— 属于无限多个集合 $A_n (n \geqslant 1)$ 的点的集合

$\liminf A_n$, 或 $\underline{\lim} A_n$, 或 $\bigcup\limits_{m \geqslant 1} \left(\bigcap\limits_{n \geqslant m} A_n \right)$ —— 属于所有 (仅可能有限个 A_n 除外) 集合 $A_n (n \geqslant 1)$ 的点的集合

I_A 或 $I(A)$ —— 集合 A 的示性函数

$\{\cdots\}$ —— 集合

数 学 符 号

$\ll$ —— 绝对连续

$\sim$ —— 等价

$\perp$ —— 垂直

$f = o(g) - \lim \left(\dfrac{f}{g} \right) = 0$

$f = O(g) - \limsup \left| \dfrac{f}{g} \right| < \infty$

$f \sim g - \lim \left(\dfrac{f}{g} \right) = 1$

$f \asymp g$ —— 比值 $\dfrac{f}{g}$ 自下与 0 分离, 且自上与 ∞ 分离

$f \circ g$ —— 复合

$f * g$ —— 卷积

[General Information]
书名=概率　　第1卷　　修订和补充第三版
作者=（俄罗斯）А．Н．施利亚耶夫著
页数=457
SS号=11852509
出版日期=2007.7

第四章　独立随机变量之和与独立随机变量序列

两个或多个试验的独立性的概念, 一定意义上在概率中占中心位置 …… 历史上试验和随机变量的独立性, 曾经是赋予概率论以印迹一种数学概念.

A. H. 柯尔莫戈洛夫,《概率论的基本概念》[32]

§1. 0–1 律

1. 柯尔莫戈洛夫 0–1 律 熟知, 下面两个级数:

$$\sum_{n=1}^{\infty}\frac{1}{n} \quad 和 \quad \sum_{n=1}^{\infty}(-1)^n\frac{1}{n},$$

其中前一个级数发散, 后一个级数收敛. 现在, 提出如下问题: 对于独立同分布伯努利随机变量序列$\xi_1,\xi_2,\cdots$ 关于级数

$$\sum_{n=1}^{\infty}\frac{\xi_n}{n}, \quad 其中 \quad \mathbf{P}\{\xi_1=+1\}=\mathbf{P}\{\xi_1=-1\}=\frac{1}{2},$$

的收敛性有何说法? 换句话说, 假如级数的一般项为 $\pm 1/n$, 其中符号 "+" 和 "−" 与上述序列 $\xi_1,\xi_2,\cdots$ 相对应, 并按随机顺序 "散布", 那么是否可以说 "以 ξ_n/n 为一般项的级数收敛"?

记

$$A_1=\left\{\omega:\sum_{n=1}^{\infty}\frac{\xi_n(\omega)}{n} \quad 收敛\right\}$$

是使其中的级数收敛 (于某些数值) 的基本事件 ω 的集合, 并且在事先并不知道该集合的概率 $\mathbf{P}(A_1)$ 取何值的情况下, 讨论概率 $\mathbf{P}(A_1)$.

不过, 非常精彩的是, 事先就可以断定: 这一概率只可能取 0 或 1 两个值之一. 这一结果是称做柯尔莫戈洛夫 (A. H. Колмогоров) 0–1 律的推论. 0–1 律的提法和证明是这一节的基本内容.

2. 柯尔莫戈洛夫 0–1 律的证明 设 $(\Omega,\mathscr{F},\mathbf{P})$ 是概率空间, $\xi_1,\xi_2,\cdots$ 是一随机变量序列. 以 $\mathscr{F}_n^{\infty}=\sigma(\xi_n,\xi_{n+1},\cdots)$ 表示由随机变量 $\xi_n,\xi_{n+1},\cdots$ 生成的 $\sigma-$ 代数, 并且设

$$\mathscr{X}=\bigcap_{n=1}^{\infty}\mathscr{F}_n^{\infty}.$$

由于 $\sigma-$ 代数的交仍然是 $\sigma-$ 代数, 可见 $\mathscr{X}$ 是 $\sigma-$ 代数. 因为对于任意有限数 n, 任何事件 $A\in\mathscr{X}$ 都不依赖于 $\xi_1,\cdots,\xi_n$, 而仅由"序列 $\xi_1,\xi_2,\cdots$ 的无穷远的性质" 所决定, 故我们把 $\sigma-$ 代数 $\mathscr{X}$ 称做 "尾部的" 或 "剩余的".

由于对于任何 $k\geqslant 1$, 有

$$A_1\equiv\left\{\sum_{n=1}^{\infty}\frac{\xi_n}{n} \quad 收敛\right\}=\left\{\sum_{n=k}^{\infty}\frac{\xi_n}{n} \quad 收敛\right\}\in\mathscr{F}_k^{\infty},$$

则 $A_1=\bigcap_k\mathscr{F}_k^{\infty}\equiv\mathscr{X}$. 同样, 假如 $\xi_1,\xi_2,\cdots$ 是任一随机变量序列, 则

$$A_2=\left\{\sum_{n=1}^{\infty}\xi_n \quad 收敛\right\}\in\mathscr{X}.$$

下列事件也是 "尾部的":

$$A_3=\{对于无限个\ n,\xi_n\in I_n\} \quad \left(=\overline{\lim_n}\{\xi_n\in I_n\}\right)$$

其中 $I_n\in\mathscr{B}(\mathbb{R}),n\geqslant 1$;

$$A_4=\left\{\overline{\lim_n}\,\xi_n<\infty\right\};$$

$$A_5=\left\{\overline{\lim_n}\frac{\xi_1+\cdots+\xi_n}{n}<\infty\right\};$$

$$A_6=\left\{\overline{\lim_n}\frac{\xi_1+\cdots+\xi_n}{n}<c\right\};$$

$$A_7=\left\{\frac{S_n}{n} \quad 收敛\right\};$$

$$A_8=\left\{\overline{\lim_n}\frac{S_n}{\sqrt{2n\ln n}}=1\right\},$$

其中 $S_n=\xi_1+\cdots+\xi_n$. 另一方面

$$B_1=\{对于一切\ \ n\geqslant 1,\xi_n=0\},$$

$$B_2=\left\{\lim_n(\xi_1+\cdots+\xi_n)\ \ 存在且小于\ c\right\}$$

都是不属于的 $\mathscr{X}$ 事件的例.

现在假设所考虑的随机变量是独立的. 在此条件下, 由博雷尔 – 坎泰利引理, 可见

$$\mathbf{P}(A_3)=0\Leftrightarrow\sum\mathbf{P}\{\xi_n\in I_n\}<\infty,$$

$$\mathbf{P}(A_3)=1\Leftrightarrow\sum\mathbf{P}\{\xi_n\in I_n\}=\infty.$$

这样, 以级数 $\sum\mathbf{P}\{\xi_n\in I_n\}$ 是否收敛为转移, 事件 A_3 的概率只有 0 或 1 两个可能值. 这一命题称做博雷尔 "0–1" 律, 它是如下命题的特殊情形.

定理 1 (柯尔莫戈洛夫 "0–1" 律) 设 $\xi_1, \xi_2, \cdots$ 是独立随机变量序列, 而 $A \in \mathscr{X}$. 那么, 概率 $\mathbf{P}(A)$ 只有 0 和 1 两个可能值.

证明 证明的思路如下: 为证明每一"尾部" 事件 A 与它自己独立, 只需证明 $\mathbf{P}(A \bigcap A) = \mathbf{P}(A) \times \mathbf{P}(A), \mathbf{P}(A) = \mathbf{P}^2(A)$, 故 $\mathbf{P}(A) = 0$ 或 1.

若 $A \in \mathscr{X}$, 则 $A \in \mathscr{F}_1^\infty = \sigma\{\xi_1, \xi_2, \cdots\} = \sigma(\bigcup \mathscr{F}_1^n)$, 其中 $\mathscr{F}_1^n = \sigma(\xi_1, \cdots, \xi_n)$, 并且存在集合 $A_n \in \mathscr{F}_1^n, n \geqslant 1$ (第二章 §3, 练习题 8), 使 $\mathbf{P}(A \triangle A_n) \to 0, n \to \infty$. 由此可见,

$$\mathbf{P}(A_n) \to \mathbf{P}(A), \quad \mathbf{P}(A_n \bigcap A) \to \mathbf{P}(A). \tag{1}$$

但是, 假如 $A \in \mathscr{X}$, 则对于每一个 $n \geqslant 1$, 事件 A_n 和 A独立:

$$\mathbf{P}(A \bigcap A_n) = \mathbf{P}(A)\mathbf{P}(A_n),$$

由于 (1), 由此可见 $\mathbf{P}(A) = \mathbf{P}^2(A)$, 故 $\mathbf{P}(A) = 0$ 或 1. □

系 设 η 是关于 "尾部" $\sigma-$ 代数 $\mathscr{X}$ 可测的随机变量, 即 $\{\eta \in B\} \in \mathscr{X}, B \in \mathscr{B}(\mathbb{R})$, 则 η 是退化随机变量, 即存在常数 c, 使 $\mathbf{P}\{\eta = c\} = 1$.

3. 柯尔莫戈洛夫 0–1 律的应用 下面引进的定理 2 是柯尔莫戈洛夫 "0–1 律" 的非平凡应用.

设 $\xi_1, \xi_2, \cdots$ 是独立伯努利随机变量序列: $\mathbf{P}\{\xi_n = 1\} = p, \mathbf{P}\{\xi_n = -1\} = q, p + q = 1, n \geqslant 1$, 而 $S_n = \xi_1 + \cdots + \xi_n$. 直观上明显, 对于对称 $(p = 1/2)$ 的情形, 随机游动 $S_n (n \geqslant 1)$ 的 "典型" 轨道无限多次地经过 0, 而对于 $p \neq 1/2$ 的情形 "趋向" 无穷. 下面是这一结果的确切表述.

定理 2 a) 如果 $p = 1/2$, 则 $\mathbf{P}\{$对于无限多个 $n, S_n = 0\} = 1$.

b) 如果 $p \neq 1/2$, 则 $\mathbf{P}\{$对于无限多个 $n, S_n = 0\} = 0$.

证明 首先注意到, $B = \{$对于无限多个 $n, S_n = 0\}$ 不是"尾部" 事件, 即 $B \notin \mathscr{X} = \bigcap \mathscr{F}_n^\infty, \mathscr{F}_n^\infty = \sigma\{\xi_n, \xi_{n+1}, \cdots\}$. 因此, 原则上不能说明事件 B 的概率只有 0 或 1 两个可能值.

利用博雷尔 – 坎泰利 (E. Borel – F. P. Cantelli) 引理 (的第一部分), 容易证明命题 b). 事实上, 如果 $B_{2n} = \{S_{2n} = 0\}$, 则可见斯特林 (J. Stirling) 公式 (第一章 §2 第 4 小节), 有

$$\mathbf{P}(B_{2n}) = \mathrm{C}_{2n}^n p^n q^n \sim \frac{(4pq)^n}{\sqrt{n\pi}},$$

从而 $\sum \mathbf{P}(B_{2n}) < \infty$. 因此, $\mathbf{P}\{$对于无限多个 $n, S_n = 0\} = 0$.

由于$A \subseteq B$, 为证明命题 a), 只需证明事件

$$A = \left\{\overline{\lim} \frac{S_n}{\sqrt{n}} = \infty, \quad \underline{\lim} \frac{S_n}{\sqrt{n}} = -\infty\right\}$$

的概率等于 1.

设 $A_c = A'_c \bigcap A''_c$, 其中

$$A'_c = \left\{\overline{\lim_n} \frac{S_n}{\sqrt{n}} \geqslant c\right\}, \quad A''_c = \left\{\underline{\lim_n} \frac{S_n}{\sqrt{n}} \leqslant -c\right\}.$$

那么, 当 $c \to \infty$ 时 $A_c \downarrow A$; 这时 A 以及所有 A'_c, A''_c 都是尾部事件. 现在验证, 对于每一个 $c, \mathbf{P}(A'_c) = \mathbf{P}(A''_c) = 1$. 由于 $A'_c \in \mathscr{X}, A''_c \in \mathscr{X}$, 故只需证明 $\mathbf{P}(A'_c) > 0, \mathbf{P}(A''_c) > 0$. 然而, 根据练习题 5 和棣莫弗 – 拉普拉斯定理 (A. DéMoivre- P. S. Laplace, 第一章 §6), 有

$$\mathbf{P}\left\{\underline{\lim_n} \frac{S_n}{\sqrt{n}} \leqslant -c\right\} = \mathbf{P}\left\{\overline{\lim_n} \frac{S_n}{\sqrt{n}} \geqslant c\right\} \geqslant \overline{\lim_n} \mathbf{P}\left\{\frac{S_n}{\sqrt{n}} \geqslant c\right\} > 0.$$

于是, 对于一切 $c > 0, \mathbf{P}(A_c) = 1$, 因而 $\mathbf{P}(A) = \lim_{c \to \infty} \mathbf{P}(A_c) = 1$. □

4. 休伊特和塞维治 0–1 律 我们再次强调, 事件 $B = \{$对于无限多个 $n, S_n = 0\}$ 不是 "尾部" 事件. 不过, 由定理 2 可见, 对于伯努利概型, 这一事件的概率同 "尾部" 事件的概率一样, 只有 0 或 1 两个可能值. 原来, 这种情况并不偶然, 它是所谓休伊特 (E. Hewitt) 和塞维治 (I. R. Sevage) 0–1 律的推论. 对于独立同分布随机变量, 休伊特和塞维治 0–1 律, 将定理 1 的结果推广到所谓 "可交换" 事件类 (包括 "尾部" 事件类).

现在给出必要的定义. 集合 $(1, 2, \cdots)$ 到自身的单值影射 $(\pi_1, \pi_2, \cdots)$ 称做有限可交换的, 如果对于一切 (只可能有有限个除外) $n, \pi_n = n$.

如果 $\xi = (\xi_1, \xi_2, \cdots)$ 是随机变量序列, 则以 $\pi(\xi)$ 表示 $(\xi_{\pi_1}, \xi_{\pi_2}, \cdots)$ 是随机变量序列.

如果事件 $A = \{\xi \in B\}, B \in \mathscr{B}(\mathbb{R}^\infty)$, 则以 $\pi(A)$ 表示事件 $\{\pi(\xi) \in B\}, B \in \mathscr{B}(\mathbb{R}^\infty)$.

称事件 $A = \{\xi \in B\}, B \in \mathscr{B}(\mathbb{R}^\infty)$ 为可交换的, 如果对于任意有限排列 π, 事件 $\pi(A)$ 与 A 重合.

事件 $A = \{$对于无限多个 $n, S_n = 0\}$, 其中 $S_n = \xi_1 + \cdots + \xi_n$, 是可交换事件的例. 此外可以证明 (练习题 4), 属于 "尾部" $\sigma-$ 代数 $\mathscr{X}(S) = \bigcap \mathscr{F}_n^\infty(S)$ 的每一个事件是可交换的, 其中 $\mathscr{F}_n^\infty(S) = \sigma\{\omega : S_n, S_{n+1}, \cdots\}$ 是随机变量 $S_1 = \xi_1, S_2 = \xi_1 + \xi_2, \cdots$ 生成的 $\sigma-$ 代数.

定理 3 (休伊特和塞维治 0–1 律) 设 $\xi = (\xi_1, \xi_2, \cdots)$ 是独立同分布随机变量序列, 而 $A = \{\xi \in B\}$ 是可交换事件, 则 $\mathbf{P}(A) = 0$ 或 1.

证明 设 $A = \{\xi \in B\}$ 是可交换事件. 选择一集合 $B_n \in \mathscr{B}(\mathbb{R}^n)$, 使 (见第二章 §3 练习题 8)

$$\mathbf{P}(A \triangle A_n) \to 0, \quad n \to \infty, \tag{2}$$

其中 $A_n = \{\omega : (\xi_1, \cdots, \xi_n) \in B_n\}$.

由于随机变量 $\xi_1, \xi_2, \cdots$ 独立且同分布, 可见概率分布 $P_\xi(B) \equiv \mathbf{P}\{\xi \in B\}$ 和 $P_{\pi_n(\xi)}(B) = \mathbf{P}\{\pi_n(\xi) \in B\}$, 其中 $\pi_n(\xi) = (\xi_{n+1}, \cdots, \xi_{2n}, \xi_1, \cdots, \xi_n, \xi_{2n+1}, \xi_{2n+2}, \cdots)$, 对于任何 $n \geqslant 1$ 相同. 因此,

$$\mathbf{P}(A \triangle A_n) = P_\xi(B \triangle B_n) = P_{\pi_n(\xi)}(B \triangle B_n). \tag{3}$$

因为 $A = \{\xi \in B\}$ 是可交换事件, 故

$$A \equiv \{\xi \in B\} \equiv \pi_n(A) \equiv \{\pi_n(\xi) \in B\}.$$

从而

$$\begin{aligned} P_{\pi_n(\xi)}(B \triangle B_n) &= \mathbf{P}(\{\pi_n(\xi) \in B\} \triangle \{\pi_n(\xi) \in B_n\}) \\ &= \mathbf{P}(\{\xi \in B\} \triangle \{\pi_n(\xi) \in B_n\}) = \mathbf{P}(A \triangle \pi_n(A_n)). \end{aligned} \tag{4}$$

这样, 由式 (3) 和 (4), 有

$$\mathbf{P}(A \triangle A_n) = \mathbf{P}(A \triangle \pi_n(A_n)). \tag{5}$$

因为 (2) 式, 由此可见

$$\mathbf{P}(A \triangle \{A_n \textstyle\bigcap \pi_n(A_n)\}) \to 0, \quad n \to \infty. \tag{6}$$

因而由式 (2), (5) 和 (6), 可得当 $n \to \infty$ 时, 有

$$\begin{gathered} \mathbf{P}(A_n) \to \mathbf{P}(A), \quad \mathbf{P}(\pi_n(A_n)) \to \mathbf{P}(A), \\ \mathbf{P}\{A_n \textstyle\bigcap \pi_n(A_n)\} \to \mathbf{P}(A). \end{gathered} \tag{7}$$

其次, 由于随机变量 $\xi_1, \xi_2, \cdots$ 的独立性, 有

$$\begin{aligned} &\mathbf{P}\{A_n \textstyle\bigcap \pi_n(A_n)\} = \mathbf{P}\{(\xi_1, \cdots, \xi_n) \in B_n, (\xi_{n+1}, \cdots, \xi_{2n}) \in B_n\} \\ &= \mathbf{P}\{(\xi_1, \cdots, \xi_n) \in B_n\}\mathbf{P}\{(\xi_{n+1}, \cdots, \xi_{2n}) \in B_n\} = \mathbf{P}(A_n)\mathbf{P}(\pi_n(A_n)). \end{aligned}$$

根据性质 (7), 由此可见

$$\mathbf{P}(A) = \mathbf{P}^2(A),$$

于是 $\mathbf{P}(A) = 0$ 或 1. □

5. 练习题

1. 证明定理 1 的系.

2. 设 $(\xi_n)_{n\geqslant 1}$ 是独立随机变量序列, 证明 $\overline{\lim}\xi_n$ 和 $\underline{\lim}\xi_n$ 是退化随机变量.

3. 设 $(\xi_n)_{n\geqslant 1}$ 是独立随机变量序列, $S_n = \xi_1 + \cdots + \xi_n$, 而 b_n 是常数且 $0 < b_n \uparrow \infty$. 证明

$$\overline{\lim_n} \frac{S_n}{b_n} \quad 和 \quad \underline{\lim_n} \frac{S_n}{b_n}$$

是退化随机变量.

4. 设 $S_n = \xi_1 + \cdots + \xi_n, n \geqslant 1$, 而, $\mathscr{X}(S) = \bigcap \mathscr{F}_n^\infty(S), \mathscr{F}_n^\infty(S) = \sigma\{\omega : S_n, S_{n+1}, \cdots\}$, 证明 $\mathscr{X}(S)$ 中的每一个事件都是可交换的.

5. 设 $(\xi_n)_{n\geqslant 1}$ 是独立随机变量序列, 证明对于任意常数 c, 有 $\{\overline{\lim}\xi_n \geqslant c\} \supseteq \{\underline{\lim}\xi_n \geqslant c\}$.

6. 举一概率严格大于 0 且严格小于 1 的《尾部》事件的例.

7. 设 $\xi_1, \xi_2, \cdots$ 是独立随机变量, $\mathbf{E}\xi_n = 0, \mathbf{E}\xi_n^2 = 1, n \geqslant 1$, 且服从中心极限定理 ($\mathbf{P}\{S_n/\sqrt{n} \leqslant x\} \to \Phi(x), x \in \mathbb{R}$, 其中 $S_n = \xi_1 + \cdots + \xi_n$, 而 $\Phi(x)$ 是标准正态分布函数), 证明

$$\overline{\lim_{n\to\infty}}\, n^{1/2} S_n = +\infty \quad (\mathbf{P}-\text{a.c.}).$$

特别, 对于独立同分布随机变量序列 (若 $\mathbf{E}\xi_1 = 0, \mathbf{E}\xi_1^2 = 1$), 这一性质仍然成立.

8. 设 $\xi_1, \xi_2, \cdots$ 独立同分布随机变量序列, 且 $\mathbf{E}|\xi_1| > 0$, 证明

$$\overline{\lim_{n\to\infty}} \left|\sum_{k=1}^n \xi_k\right| = +\infty \quad (\mathbf{P}-\text{a.c.}).$$

§2. 级数的收敛性

1. 独立随机变量的级数的敛散性准则 假设 $\xi_1, \xi_2, \cdots$ 是独立随机变量序列, $S_n = \xi_1 + \cdots + \xi_n$, 而 A 是使级数 $\sum \xi_n(\omega)$ 收敛于有限极限的基本事件 ω 的集合. 由柯尔莫戈洛夫 0–1 律知, 概率 $\mathbf{P}(A) = 0$ 或 1, 即级数 $\sum \xi_n$ 以概率 1 收敛或发散. 这一节的目的是, 给出确定独立随机变量的级数收敛还是发散的准则.

定理 1 (柯尔莫戈洛夫 – 辛钦) a) 设 $\mathbf{E}\xi_n = 0, n \geqslant 1$. 那么, 如果

$$\sum \mathbf{E}\xi_n^2 < \infty, \tag{1}$$

则级数 $\sum \xi_n$ 以概率 1 收敛.

b) 并且, 如果随机变量 $\xi_n, n \geqslant 1$, 一致有界 (对于某个 $c < \infty, \mathbf{P}\{|\xi_n| \leqslant c\} = 1$), 则逆命题也成立: 若级数 $\sum \xi_n$ 以概率 1 收敛, 则条件 (1) 成立.

该定理的证明本质上用到下面的不等式.

柯尔莫戈洛夫不等式 a) 设 $\xi_1, \xi_2, \cdots, \xi_n$ 是独立随机变量, $\mathbf{E}\xi_i = 0, \mathbf{E}\xi_i^2 < \infty, 0 \leqslant i \leqslant n$. 那么, 对于任意 $\varepsilon > 0$, 有

$$\mathbf{P}\left\{\max_{1\leqslant k\leqslant n} |S_k| \geqslant \varepsilon\right\} \leqslant \frac{\mathbf{E}S_n^2}{\varepsilon^2}. \tag{2}$$

b) 并且, 如果 $\mathbf{P}\{|\xi_i| \leqslant c\} = 1, 0 \leqslant i \leqslant n$, 则

$$\mathbf{P}\left\{\max_{1\leqslant k\leqslant n} |S_k| \geqslant \varepsilon\right\} \geqslant 1 - \frac{(c+\varepsilon)^2}{\mathbf{E}S_n^2}. \tag{3}$$

证明 a) 记

$$A=\left\{\max_{1\leqslant k\leqslant n}|S_k|\geqslant\varepsilon\right\},$$

$$A_k=\{|S_i|<\varepsilon, i=1,\cdots,k-1,|S_k|\geqslant\varepsilon\},\quad 1\leqslant k\leqslant n.$$

那么, $A=\sum A_k$, 而

$$\mathbf{E}S_n^2\geqslant\mathbf{E}S_n^2I_A=\sum\mathbf{E}S_k^2I_{A_k}.$$

因为

$$\begin{aligned}\mathbf{E}S_n^2I_{A_k}&=\mathbf{E}[S_k+(\xi_{k+1}+\cdots+\xi_n)]^2I_{A_k}\\&=\mathbf{E}S_k^2I_{A_k}+2\mathbf{E}S_k(\xi_{k+1}+\cdots+\xi_n)I_{A_k}+\mathbf{E}(\xi_{k+1}+\cdots+\xi_n)^2I_{A_k}\geqslant\mathbf{E}S_k^2I_{A_k},\end{aligned}$$

其中由于独立性假设和条件 $\mathbf{E}\xi_i=0, 0\leqslant i\leqslant n$, 可见

$$\mathbf{E}S_k(\xi_{k+1}+\cdots+\xi_n)I_{A_k}=\mathbf{E}S_kI_{A_k}\times\mathbf{E}(\xi_{k+1}+\cdots+\xi_n)=0,$$

所以

$$\mathbf{E}S_n^2\geqslant\sum\mathbf{E}S_k^2I_{A_k}\geqslant\varepsilon^2\sum\mathbf{P}(A_k)=\varepsilon^2\mathbf{P}(A),$$

于是, 不等式 (2) 得证.

为证明不等式 (3), 注意到

$$\mathbf{E}S_n^2I_A=\mathbf{E}S_n^2-\mathbf{E}S_n^2I_{\overline{A}}\geqslant\mathbf{E}S_n^2-\varepsilon^2\mathbf{P}(\overline{A})=\mathbf{E}S_n^2-\varepsilon^2+\varepsilon^2\mathbf{P}(A).\tag{4}$$

另一方面, 在集合 A_k 上, 有

$$|S_{k-1}|\leqslant\varepsilon,\quad |S_k|\leqslant|S_{k-1}|+|\xi_k|\leqslant\varepsilon+c,$$

因而

$$\begin{aligned}\mathbf{E}S_n^2I_{A_n}&=\sum_k\mathbf{E}S_k^2I_{A_k}+\sum_k\mathbf{E}[I_{A_k}(S_n-S_k)^2]\\&\leqslant(\varepsilon+c)^2\sum_k\mathbf{P}(A_k)+\sum_{k=1}^n\mathbf{P}(A_k)\sum_{j=k+1}^n\mathbf{E}\xi_j^2\\&\leqslant\mathbf{P}(A)\left[(\varepsilon+c)^2+\sum_{j=1}^n\mathbf{E}\xi_j^2\right]=\mathbf{P}(A)[(\varepsilon+c)^2+\mathbf{E}S_n^2].\end{aligned}\tag{5}$$

由式 (4) 和 (5), 得

$$\mathbf{P}(A)\geqslant\frac{\mathbf{E}S_n^2-\varepsilon^2}{(\varepsilon+c)^2+\mathbf{E}S_n^2-\varepsilon^2}=1-\frac{(\varepsilon+c)^2}{(\varepsilon+c)^2+\mathbf{E}S_n^2-\varepsilon^2}\geqslant1-\frac{(\varepsilon+c)^2}{\mathbf{E}S_n^2}.$$

于是, 不等式 (3) 得证. □

定理 1 的证明 a) 根据第二章 §10 定理 4, 序列 $(S_n)_{n\geqslant1}$ 以概率 1 收敛, 当且仅当该序列以概率 1 是基本序列. 根据第二章 §10 定理 1, 序列 $(S_n)_{n\geqslant1}$为基本序列 ($\mathbf{P}$ – a.c.) 的充分必要条件是, 对于任意 $\varepsilon>0$, 有

$$\mathbf{P}\left\{\sup_{k\geqslant1}|S_{n+k}-S_n|\geqslant\varepsilon\right\}\to0,\quad n\to\infty.\tag{6}$$

由 (2) 式, 有

$$\begin{aligned}\mathbf{P}\left\{\sup_{k\geqslant1}|S_{n+k}-S_n|\geqslant\varepsilon\right\}&=\lim_{N\to\infty}\mathbf{P}\left\{\max_{1\leqslant k\leqslant N}|S_{n+k}-S_n|\geqslant\varepsilon\right\}\\&\leqslant\lim_{N\to\infty}\frac{\sum_{k=n}^{n+N}\mathbf{E}\xi_k^2}{\varepsilon^2}=\frac{\sum_{k=n}^{\infty}\mathbf{E}\xi_k^2}{\varepsilon^2}.\end{aligned}$$

因此, 如果 $\sum_{k=1}^{\infty}\mathbf{E}\xi_k^2<\infty$, 则条件 (6) 成立, 于是级数 $\sum\xi_k$ 以概率 1 收敛.

b) 设级数 $\sum\xi_k$ 以概率 1 收敛, 则由 (6) 式知, 对于充分大的 n, 有

$$\mathbf{P}\left\{\sup_{k\geqslant1}|S_{n+k}-S_n|\geqslant\varepsilon\right\}<\frac{1}{2}.\tag{7}$$

由 (3) 式, 有

$$\mathbf{P}\left\{\sup_{k\geqslant1}|S_{n+k}-S_n|\geqslant\varepsilon\right\}\geqslant1-\frac{(c+\varepsilon)^2}{\sum_{k=n}^{\infty}\mathbf{E}\xi_k^2}.$$

从而, 如果满足 $\sum_{k=1}^{\infty}\mathbf{E}\xi_k^2=\infty$, 则

$$\mathbf{P}\left\{\sup_{k\geqslant1}|S_{n+k}-S_n|\geqslant\varepsilon\right\}=1,$$

而这与 (7) 式矛盾. □

例 如果 $\xi_1,\xi_2,\cdots$ 是独立同分布伯努利随机变量序列, $\mathbf{P}\{\xi_n=+1\}=\mathbf{P}\{\xi_n=-1\}=1/2$, 则级数 $\sum\xi_na_n(|a_n|\leqslant c)$ 以概率 1 收敛, 当且仅当 $\sum a_n^2<\infty$.

2. "两级数" 定理

定理 2 (柯尔莫戈洛夫 – 辛钦 "两级数" 定理) 独立随机变量的 $\xi_1,\xi_2,\cdots$ 的级数 $\sum\xi_n$ 以概率 1 收敛的充分条件是, 两级数 $\sum\mathbf{E}\xi_n$ 和 $\sum\mathbf{D}\xi_n$ 同时收敛. 并且, 如果对于某个 $c>0, \mathbf{P}\{|\xi_n|\leqslant c\}=1(n\geqslant1)$, 则此条件也是必要的.

证明 如果 $\sum\mathbf{D}\xi_n<\infty$, 则根据定理 1, 级数 $\sum(\xi_n-\mathbf{E}\xi_n)$ 以概率 1 收敛. 因为根据假设级数 $\sum\mathbf{E}\xi_n$ 收敛, 所以级数 $\sum\xi_n$ 也以概率 1 收敛.

为证明必要性, 利用 "对称化" 方法. 与 $\xi_1,\xi_2,\cdots$ 同时, 考虑依赖于它的独立随机变量序列 $\tilde{\xi}_1,\tilde{\xi}_2,\cdots$, 其中 $\tilde{\xi}_n$ 与 $\xi_n(n\geqslant1)$ 同分布. (假如原基本事件空间是充分

"丰富" 的, 则由第二章 §9 定理 1 的系 1, 可见这样序列的存在性. 同时, 可以证明此假设并不失普遍性).

那么, 如果级数 $\sum\xi_n$ 以概率 1 收敛, 则级数 $\sum\tilde{\xi}_n$ 以及级数 $\sum(\xi_n-\tilde{\xi}_n)$ 也以概率 1 收敛. 而 $\mathbf{E}(\xi_n-\tilde{\xi}_n)=0$ 且 $\mathbf{P}\{|\xi_n-\tilde{\xi}_n|\leqslant 2c\}=1$. 所以由定理 1 的命题 b) 知, $\sum\mathbf{D}(\xi_n-\tilde{\xi}_n)<\infty$. 此外,

$$\sum\mathbf{D}\xi_n=\frac{1}{2}\sum\mathbf{D}(\xi_n-\tilde{\xi}_n)<\infty.$$

因此, 根据定理 1 的命题 a), 级数 $\sum(\xi_n-\mathbf{E}\xi_n)$ 以概率 1 收敛, 说明级数 $\sum\mathbf{E}\xi_n$ 也收敛.

于是, 由于级数 $\sum\xi_n$ 以概率 1 收敛, 可见 (在假设 $\mathbf{P}\{|\xi_n|\leqslant c\}=1(n\geqslant 1)$ 下) 两级数 $\sum\mathbf{E}\xi_n$ 和 $\sum\mathbf{D}\xi_n$ 也同时收敛. □

3. "三级数" 定理

下面的定理, 在没有关于随机变量有界的假设条件下, 给出了级数 $\sum\xi_n$ 以概率 1 收敛的充分和必要条件.

对于某个常数 c, 设

$$\xi^c=\begin{cases}\xi, & 若\ |\xi|\leqslant c,\\ 0, & 若\ |\xi|>c.\end{cases}$$

定理 3 (柯尔莫戈洛夫 "三级数" 定理) *设 $\xi_1,\xi_2,\cdots$ 是独立随机变量序列. 级数 $\sum\xi_n$ 以概率 1 收敛的必要条件是, 对于任意 $c>0$, 三级数*

$$\sum\mathbf{E}\xi_n^c,\quad \sum\mathbf{D}\xi_n^c,\quad \sum\mathbf{P}\{|\xi_n|\geqslant c\}$$

收敛, 并且三级数对于某个 $c>0$ 收敛, 是级数 $\sum\xi_n^c$ 以概率 1 收敛的充分条件.

证明 (1) *充分性*. 根据 "两级数" 定理, 级数 $\sum\xi_n^c$ 以概率 1 收敛. 而如果对于某个 $c>0,\sum\mathbf{P}\{|\xi_n|\geqslant c\}<\infty$, 则根据博雷尔 – 坎泰利引理, 以概率 1 级数 $\sum I\{|\xi_n|\geqslant c\}<\infty$, 说明对于一切 n (只可能有有限个可能除外), 有 $\xi_n=\xi_n^c$. 因此, 级数 $\sum\xi_n$ 也以概率 1 收敛.

(2) *必要性*. 如果级数 $\sum\xi_n$ 以概率 1 收敛, 则以概率 1 有 $\xi_n\to 0$, 故对于任意 $c>0$, 以概率 1 最多有有限个事件 $\{|\xi_n|\geqslant c\}$ 出现. 因此, 以概率 1 有 $\sum I\{|\xi_n|\geqslant c\}<\infty$, 且根据第二博雷尔 – 坎泰利引理 $\sum\mathbf{P}\{|\xi_n|>c\}<\infty$. 此外, 由级数 $\sum\xi_n$ 收敛, 可见级数 $\sum\xi_n^c$ 收敛. 于是, 根据 "两级数" 定理, 两个级数 $\sum\mathbf{E}\xi_n^c$ 和 $\sum\mathbf{D}\xi_n^c$ 都收敛. □

系 *设 $\xi_1,\xi_2,\cdots$ 是独立随机变量, 且 $\mathbf{E}\xi_n=0$. 那么, 如果*

$$\sum\mathbf{E}\frac{\xi_n^2}{1+|\xi_n|}<\infty,$$

则级数 $\sum\xi_n$ 以概率 1 收敛.

为证明此结果, 我们注意到,

$$\sum\mathbf{E}\frac{\xi_n^2}{1+|\xi_n|}<\infty\Leftrightarrow\sum\mathbf{E}[\xi_n^2 I(|\xi_n|\leqslant 1)+|\xi_n|I(|\xi_n|>1)]<\infty.$$

因此, 如果 $\xi_n^1=\xi_n I(|\xi_n|\leqslant 1)$, 则

$$\sum\mathbf{E}(\xi_n^1)^2<\infty.$$

由于 $\mathbf{E}\xi_n=0$, 可见

$$\begin{aligned}\sum|\mathbf{E}\xi_n^1|&=\sum|\mathbf{E}\xi_n I(|\xi_n|\leqslant 1)|=\sum|\mathbf{E}\xi_n I(|\xi_n|>1)|\\&\leqslant\sum\mathbf{E}|\xi_n|I(|\xi_n|>1)<\infty.\end{aligned}$$

说明级数 $\sum\mathbf{E}\xi_n^1$ 和 $\sum\mathbf{D}\xi_n^1$ 都收敛. 此外, 根据切比雪夫不等式, 有

$$\mathbf{P}\{|\xi_n|>1\}=\mathbf{P}\{|\xi_n|I(|\xi_n|>1)|>1\}\leqslant\mathbf{E}|\xi_n|I(|\xi_n|>1).$$

因此, $\sum\mathbf{P}\{|\xi_n|>1\}<\infty$. 于是, 由 "三级数" 定理, 可见级数 $\sum\xi_n$ 以概率 1 收敛.

4. 练习题

1. 设 $\xi_1,\xi_2,\cdots$ 是独立随机变量序列, $S_n=\xi_1+\cdots+\xi_n$. 利用 "三级数" 定理, 证明:

a) 如果级数 $\sum\xi_n^2$ 以概率 1 收敛, 则级数 $\sum\xi_n$ 以概率 1 收敛, 当且仅当级数 $\sum\mathbf{E}\xi_n I(|\xi_n|\leqslant 1)$ 收敛.

b) 如果级数 $\sum\xi_n$ 以概率 1 收敛, 则级数 $\sum\xi_n^2$ 也以概率 1 收敛, 当且仅当

$$\sum[\mathbf{E}|\xi_n|I(|\xi_n|\leqslant 1)]^2<\infty.$$

2. 设 $\xi_1,\xi_2,\cdots$ 是独立随机变量序列, 证明级数 $\sum\xi_n^2$ 以概率 1 收敛, 当且仅当

$$\sum\mathbf{E}\frac{\xi_n^2}{1+\xi_n^2}<\infty.$$

3. 设 $\xi_1,\xi_2,\cdots$ 是独立随机变量序列, 证明如下三个条件等价:

a) 级数 $\sum\xi_n$ *以概率 1 收敛*;

b) 级数 $\sum\xi_n$ *依概率收敛*;

c) 级数 $\sum\xi_n$ *依分布收敛*.

4. 举例说明, 在定理 1 和定理 2 中, 一般不能去掉一致有界性条件: 对于某个 $c>0,\mathbf{P}\{|\xi_n|\leqslant c\}=1(n\geqslant 1)$.

5. 设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量, 且 $\mathbf{E}\xi_1=0,\mathbf{E}\xi_1^2<\infty,S_n=\xi_1+\cdots+\xi_n$. 证明如下柯尔莫戈洛夫不等式 (2) 的单侧类似 (马尔沙尔 [A. B. Маршалл]):

$$\mathbf{P}\left\{\max_{1\leqslant k\leqslant n}S_k\geqslant\varepsilon\right\}\leqslant\frac{\mathbf{E}S_n^2}{\varepsilon^2+\mathbf{E}S_n^2}.$$

6. 设 $\xi_1, \xi_2, \cdots$ 是 (任意) 随机变量序列. 证明, 如果 $\sum_{n \geqslant 1} \mathbf{E}|\xi_n| < \infty$, 则级数 $\sum_{n \geqslant 1} \xi_n$ 以概率 1 绝对收敛.

7. 设 $\xi_1, \xi_2, \cdots$ 是独立对称分布随机变量. 证明

$$\mathbf{E}\left[\left(\sum_n \xi_n\right)^2 \wedge 1\right] \leqslant \sum_n \mathbf{E}(\xi_n^2 \wedge 1).$$

8. 设 $\xi_1, \xi_2, \cdots$ 是具有有限二阶矩的、独立随机变量. 证明, 当且仅当级数 $\sum \mathbf{E}\xi_n$ 和 $\sum \mathbf{D}\xi_n$ 收敛时, 级数 $\sum \xi_n$ 在 L^2 中收敛.

9. 设 $\xi_1, \xi_2, \cdots$ 是独立随机变量, 级数 $\sum \xi_n$ 几乎必然 (处处) 收敛. 证明, 该级数的值几乎必然与项的求和顺序无关, 当且仅当 $\sum |\mathbf{E}(\xi_n; |\xi_n| \leqslant 1)| < \infty$.

10. 设 $\xi_1, \xi_2, \cdots$ 是独立随机变量, 且 $\mathbf{E}\xi_n = 0\ (n \geqslant 1)$, 而

$$\sum_{n=1}^{\infty} \mathbf{E}[\xi_n^2 I(|\xi_n| \leqslant 1) + |\xi_n| I(|\xi_n| > 1)] < \infty.$$

证明级数 $\sum_{n=1}^{\infty} \xi_n$ $(\mathbf{P}-\text{a.c.})$ 收敛.

11. 设 $A_1, A_2, \cdots$ 是独立事件, 满足: $\mathbf{P}(A_n) > 0 (n \geqslant 1)$, $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$. 证明当 $n \to \infty$ 时, 有

$$\sum_{j=1}^{n} I(A_j) \Big/ \sum_{j=1}^{n} \mathbf{P}(A_j) \to 1 \quad (\mathbf{P}-\text{a.c.}).$$

12. 设 $\xi_1, \xi_2, \cdots$ 是独立随机变量, 其数学期望为 $\mathbf{E}\xi_n$ 和方差 σ_n^2 满足: $\lim_n \mathbf{E}\xi_n = c$, 和 $\sum_{n=1}^{\infty} \sigma_n^{-2} = \infty$. 证明当 $n \to \infty$ 时, 有

$$\sum_{j=1}^{n} \frac{\xi_j}{\sigma_j^2} \Big/ \sum_{j=1}^{n} \frac{1}{\sigma_j^2} \to c \quad (\mathbf{P}-\text{a.c.}).$$

§3. 强大数定律

1. 坎泰利强大数定律 设 $\xi_1, \xi_2, \cdots$ 是具有有限二阶矩的独立随机变量序列, $S_n = \xi_1 + \cdots + \xi_n$. 根据第三章 §3 练习题 2, 如果方差 $\mathbf{D}\xi_i$ 一致有界, 则大数定律成立:

$$\frac{S_n - \mathbf{E}S_n}{n} \xrightarrow{\mathbf{P}} 0, \quad n \to \infty. \tag{1}$$

强大数定律 如果将 (1) 式中的 "依概率收敛" 换成 "以概率 1 收敛", 则相应的论断称做强大数定律.

下面的定理, 是强大数定律最早的结果之一.

定理 1 (坎泰利) 设 $\xi_1, \xi_2, \cdots$ 是具有有限 4 阶矩的独立随机变量, 且对于某个常数 C, 有

$$\mathbf{E}|\xi_n - \mathbf{E}\xi_n|^4 \leqslant C, \quad n \geqslant 1.$$

那么, 当 $n \to \infty$ 时, 有

$$\frac{S_n - \mathbf{E}S_n}{n} \to 0, \quad (\mathbf{P}-\text{a.c.}). \tag{2}$$

证明 不失普遍性, 可以认为 $\mathbf{E}\xi_n = 0, n \geqslant 1$. 根据第二章 §10 定理 1 的系, 为使 $S_n/n \to 0\ (\mathbf{P}-\text{a.c.})$, 只需对于任何 $\varepsilon > 0$, 有

$$\sum \mathbf{P}\left\{\left|\frac{S_n}{n}\right| \geqslant \varepsilon\right\} < \infty.$$

同样地, 由切比雪夫不等式, 为满足上述不等式只需满足条件

$$\sum \mathbf{E}\left|\frac{S_n}{n}\right|^4 < \infty.$$

现在证明, 在定理的条件下该不等式确实成立.

易见,

$$S_n^4 = (\xi_1 + \cdots + \xi_n)^4 = \sum_{i=1}^{n} \xi_i^4 + \sum_{\substack{i,j \\ i<j}} \frac{4!}{2!2!}\xi_i^2\xi_j^2 + \sum_{\substack{i \neq j \\ i \neq k \\ j<k}} \frac{4!}{2!1!1!}\xi_i^2\xi_j\xi_k + \sum_{i<j<k<l} 4!\xi_i\xi_j\xi_k\xi_l + \sum_{i \neq j} \frac{4!}{3!1!}\xi_i^3\xi_j.$$

由于 $\mathbf{E}\xi_k = 0, k \leqslant n$, 故由此可见

$$\mathbf{E}S_n^4 = \sum_{i=1}^{n} \mathbf{E}\xi_i^4 + 6\sum_{i,j=1}^{n} \mathbf{E}\xi_i^2\mathbf{E}\xi_j^2 \leqslant nC + 6\sum_{\substack{i,j=1 \\ i<j}}^{n} \sqrt{\mathbf{E}\xi_i^4\mathbf{E}\xi_j^4}$$
$$\leqslant nC + \frac{6n(n-1)}{2}C = (3n^2 - 2n)C < 3n^2C.$$

从而

$$\sum \mathbf{E}\left(\frac{S_n}{n}\right)^4 \leqslant 3C\sum \frac{1}{n^2} < \infty. \qquad \square$$

2. 强大数定律的柯尔莫戈洛夫准则 若引进更加精细的方法, 则可以本质上减弱定理 1 关于强大数定律成立的条件.

定理 2 (柯尔莫戈洛夫)① 设 $\xi_1, \xi_2, \cdots$ 是具有有限二阶矩的独立随机变量序列, 而 b_n 是满足如下条件的正数: $b_n \uparrow \infty$ 且

$$\sum \frac{\mathbf{D}\xi_n}{b_n^2} < \infty. \tag{3}$$

①该定理亦称做强大数定律的柯尔莫戈洛夫准则. —— 译者

那么,

$$\frac{S_n - \mathbf{E}S_n}{b_n} \to 0, \quad (\mathbf{P}-\text{a.c.}). \tag{4}$$

特别, 如果

$$\sum \frac{\mathbf{D}\xi_n}{n^2} < \infty, \tag{5}$$

则

$$\frac{S_n - \mathbf{E}S_n}{n} \to 0, \quad (\mathbf{P}-\text{a.c.}). \tag{6}$$

定理 2 和下面定理 3 的证明, 要用到两个辅助命题.

引理 1 (特普利茨 [O. Toeplitz]) 设 $(a_n)_{n\geqslant 1}$ 是非负数列, $b_n = \sum_{i=1}^{n} a_i, b_1 = a_1 > 0$, 且 $b_n \uparrow \infty, n \to \infty$; 设 $(x_n)_{n\geqslant 1}$ 是收敛于某一实数 x 的数列. 那么,

$$\frac{1}{b_n}\sum_{j=1}^{n} a_j x_j \to x, \quad n \to \infty. \tag{7}$$

特别, 如果 $a_n = 1$, 则

$$\frac{x_1 + \cdots + x_n}{n} \to x, \quad n \to \infty. \tag{8}$$

证明 设 $\varepsilon > 0$ 和 $n_0 = n_0(\varepsilon)$ 满足: 对于 $n \geqslant n_0, |x_n - x| \leqslant \varepsilon/2$. 选择 $n_1 > n_0$, 使

$$\frac{1}{b_{n_1}}\sum_{j=1}^{n_0} a_j|x_j - x| < \frac{\varepsilon}{2}.$$

因此, 对于 $n > n_1$, 有

$$\begin{aligned}\left|\frac{1}{b_n}\sum_{j=1}^{n} a_j x_j - x\right| &\leqslant \frac{1}{b_n}\sum_{j=1}^{n} a_j|x_j - x| \\ &= \frac{1}{b_n}\sum_{j=1}^{n_0} a_j|x_j - x| + \frac{1}{b_n}\sum_{j=n_0+1}^{n} a_j|x_j - x| \\ &\leqslant \frac{1}{b_{n_1}}\sum_{j=1}^{n_0} a_j|x_j - x| + \frac{1}{b_n}\sum_{j=n_0+1}^{n} a_j|x_j - x| \\ &\leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2}\frac{b_n - b_{n_0}}{b_n} \leqslant \varepsilon.\end{aligned}$$

□

引理 2 (克罗内克 [L. Kronecker]) 设 $(b_n)_{n\geqslant 1}$ 是正实数的递增数列, 并且 $b_n \uparrow \infty, n \to \infty$; 而 $(x_n)_{n\geqslant 1}$ 是使级数 $\sum x_n$ 收敛的某一实数列. 那么,

$$\frac{1}{b_n}\sum_{j=1}^{n} b_j x_j \to 0, \quad n \to \infty.$$

特别, 如果 $b_n = n, x_n = y_n/n$, 而级数 $\sum y_n/n$ 收敛, 则

$$\frac{y_1 + \cdots + y_n}{n} \to 0, \quad n \to \infty. \tag{9}$$

证明 设 $b_0 = 0, S_0 = 0, S_n = \sum_{j=1}^{n} x_j$. 那么 ("分部求和"), 有

$$\sum_{j=1}^{n} b_j x_j = \sum_{j=1}^{n} b_j(S_j - S_{j-1}) = b_n S_n - b_0 S_0 - \sum_{j=1}^{n} S_{j-1}(b_j - b_{j-1}),$$

设 $a_j = b_j - b_{j-1}$, 得

$$\frac{1}{b_n}\sum_{j=1}^{n} b_j x_j = S_n - \frac{1}{b_n}\sum_{j=1}^{n} S_{j-1}a_j \to 0,$$

因为若 $S_n \to x$, 则根据特普利茨引理, 有

$$\frac{1}{b_n}\sum_{j=1}^{n} S_{j-1}a_j \to x.$$

□

定理 2 的证明 因为

$$\frac{S_n - \mathbf{E}S_n}{b_n} = \frac{1}{b_n}\sum_{k=1}^{n} b_k\left(\frac{\xi_k - \mathbf{E}\xi_k}{b_k}\right),$$

则根据克罗内克引理, 为使 (4) 式成立, 只需级数 $\sum(\xi_k - \mathbf{E}\xi_k)/b_k$ 以概率 1 收敛. 实际上, 由于条件 (3) 和 §2 的定理 1 知, 该级数确实收敛.

例 1 设 $\xi_1, \xi_2, \cdots$ 是独立伯努利随机变量序列, 且 $\mathbf{P}\{\xi_n = 1\} = \mathbf{P}\{\xi_n = -1\} = 1/2$. 那么, 由

$$\sum \frac{1}{n\ln^2 n} < \infty, \quad \text{可见} \quad \frac{S_n}{\sqrt{n}\ln n} \to 0 \quad (\mathbf{P}-\text{a.c.}). \tag{10}$$

3. 独立同分布随机变量的强大数定律 对于随机变量 $\xi_1, \xi_2, \cdots$ 不但独立, 而且也同分布的情形, 为使强大数定律成立, 没有必要像定理 2 那样要求二阶矩存在, 只需要求一阶绝对矩存在.

定理 3 (柯尔莫戈洛夫) 设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量, 且 $\mathbf{E}|\xi_1| < \infty$. 那么,

$$\frac{S_n}{n} \to m \quad (\mathbf{P}-\text{a.c.}), \tag{11}$$

其中 $m = \mathbf{E}\xi_1, S_n = \xi_1 + \cdots + \xi_n$.

定理的证明用到下面的引理.

引理 3 设 ξ 是非负随机变量, 那么,

$$\sum_{n=1}^{\infty}\mathbf{P}\{\xi \geqslant n\} \leqslant \mathbf{E}\xi \leqslant 1 + \sum_{n=1}^{\infty}\mathbf{P}\{\xi \geqslant n\}. \tag{12}$$

引理 3 的证明. 下面的一系列不等式即可证明引理 3:

$$\begin{aligned}\sum_{n=1}^{\infty}\mathbf{P}\{\xi\geqslant n\}&=\sum_{n=1}^{\infty}\sum_{k\geqslant n}\mathbf{P}\{k\leqslant\xi<k+1\}=\sum_{k=1}^{\infty}k\mathbf{P}\{k\leqslant\xi<k+1\}\\&=\sum_{k=0}^{\infty}\mathbf{E}[kI\{k\leqslant\xi<k+1\}]\leqslant\sum_{k=0}^{\infty}\mathbf{E}[\xi I\{k\leqslant\xi<k+1\}]=\mathbf{E}\xi\\&\leqslant\sum_{k=0}^{\infty}\mathbf{E}[(k+1)I\{k\leqslant\xi<k+1\}]=\sum_{k=0}^{\infty}(k+1)\mathbf{P}\{k\leqslant\xi<k+1\}\\&=\sum_{n=1}^{\infty}\mathbf{P}\{\xi\geqslant n\}+\sum_{k=0}^{\infty}\mathbf{P}\{k\leqslant\xi<k+1\}=\sum_{n=1}^{\infty}\mathbf{P}\{\xi\geqslant n\}+1.\end{aligned}\quad\square$$

定理 3 的证明 由引理 3 和博雷尔 - 坎泰利引理 (第二章 §10), 可见

$$\begin{aligned}\mathbf{E}|\xi_1|<\infty&\Leftrightarrow\sum\mathbf{P}\{|\xi_1|\geqslant n\}<\infty\Leftrightarrow\sum\mathbf{P}\{|\xi_n|\geqslant n\}<\infty\\&\Leftrightarrow\sum\mathbf{P}\{\text{对于无限个 } n,|\xi_n|\geqslant n\}=0.\end{aligned}$$

因此, 以概率 1 对于一切 n (仅有限个 n 有可能除外), 有 $|\xi_n|<n$.

记

$$\widetilde{\xi}_n=\begin{cases}\xi_n, & |\xi_n|<n,\\ 0, & |\xi_n|\geqslant n,\end{cases}$$

且对于 $n\geqslant 1$, 认为 $\mathbf{E}\xi_n=0$. 那么,

$$\frac{\xi_1+\cdots+\xi_n}{n}\to 0,\quad n\to\infty(\mathbf{P}-\text{a.c.})\Leftrightarrow\frac{\widetilde{\xi}_1+\cdots+\widetilde{\xi}_n}{n}\to 0,\ n\to\infty(\mathbf{P}-\text{a.c.}).$$

注意, 一般 $\mathbf{E}\widetilde{\xi}_n\neq 0$, 但是

$$\mathbf{E}\widetilde{\xi}_n=\mathbf{E}\xi_nI\{|\xi_n|<n\}=\mathbf{E}\xi_1I\{|\xi_1|<n\}\to\mathbf{E}\xi_1=0.$$

因此, 由特普利茨引理, 有

$$\frac{1}{n}\sum_{k=1}^{n}\mathbf{E}\widetilde{\xi}_k\to 0,\quad n\to\infty,$$

因而

$$\begin{aligned}&\frac{\xi_1+\cdots+\xi_n}{n}\to 0,\quad n\to\infty(\mathbf{P}-\text{a.c.})\\\Leftrightarrow&\frac{(\widetilde{\xi}_1-\mathbf{E}\widetilde{\xi}_1)+\cdots+(\widetilde{\xi}_n-\mathbf{E}\widetilde{\xi}_n)}{n}\to 0,\quad n\to\infty(\mathbf{P}-\text{a.c.}).\end{aligned}\tag{13}$$

记 $\overline{\xi}_n=\widetilde{\xi}_n-\mathbf{E}\widetilde{\xi}_n$. 根据克罗内科引理, 为证明 (13) 式成立, 只需证明级数 $\sum\overline{\xi}_n/n$ 以概率 1 收敛. 而且由 §2 定理 1, 为此只需证明假设的条件 $\mathbf{E}|\xi_1|<\infty$ 可以保障级数 $\sum\mathbf{D}\overline{\xi}_n/n^2$ 收敛.

下面一系列不等式恰好证明这确实成立:

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{\mathbf{D}\overline{\xi}_n}{n^2}&\leqslant\sum_{n=1}^{\infty}\frac{\mathbf{E}\overline{\xi}_n^2}{n^2}=\sum_{n=1}^{\infty}\frac{1}{n^2}\mathbf{E}[\xi_nI(|\xi_n|<n)]^2\\&=\sum_{n=1}^{\infty}\frac{1}{n^2}\mathbf{E}[\xi_1^2I(|\xi_1|<n)]=\sum_{n=1}^{\infty}\frac{1}{n^2}\sum_{k=1}^{n}\mathbf{E}[\xi_1^2I(k-1\leqslant|\xi_1|<k)]\\&=\sum_{k=1}^{\infty}\mathbf{E}[\xi_1^2I(k-1\leqslant|\xi_1|<k)]\sum_{n=k}^{\infty}\frac{1}{n^2}\leqslant 2\sum_{k=1}^{n}\frac{1}{k}\mathbf{E}[\xi_1^2I(k-1\leqslant|\xi_1|<k)]\\&\leqslant 2\sum_{k=1}^{\infty}\mathbf{E}[|\xi_1|I(k-1\leqslant|\xi_1|<k)]=2\mathbf{E}|\xi_1|<\infty.\end{aligned}\quad\square$$

注 1 在如下的意义上, 定理的逆命题也成立: 设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量序列, 且以概率 1 满足

$$\frac{\xi_1+\cdots+\xi_n}{n}\to C,$$

其中 C 是某一 (有限) 常数. 那么, $\mathbf{E}|\xi_1|<\infty$ 且 $C=\mathbf{E}\xi_1$.

事实上, 如果 $S_n/n\to C(\mathbf{P}-\text{a.c.})$, 则

$$\frac{\xi_n}{n}=\frac{S_n}{n}-\left(\frac{n-1}{n}\right)\frac{S_{n-1}}{n-1}\to 0(\mathbf{P}-\text{a.c.}),$$

从而, $\mathbf{P}\{\text{对于无限个 } n,|\xi_n|>n\}=0$. 根据博雷尔 - 坎泰利引理 (第二章 §10), 且由引理 3 知 $\mathbf{E}|\xi_1|<\infty$, 有

$$\sum_{n=1}^{\infty}\mathbf{P}\{|\xi_1|>n\}<\infty.$$

那么, 由已经证明的定理, 可见 $C=\mathbf{E}\xi_1$.

于是, 对于独立同分布随机变量, 条件 $\mathbf{E}|\xi_1|<\infty$ 是 (以概率 1) 收敛于有限极限 S_n/n 的充分和必要条件.

注 2 如果数学期望 $m=\mathbf{E}\xi_1$ 存在, 但是未必有限, 则定理的断定 (9) 仍然成立.

事实上, 例如, 设 $\mathbf{E}\xi_1^-<\infty,\mathbf{E}\xi_1^+=\infty$. 对 $C>0$, 设

$$S_n^C=\sum_{i=1}^{n}\xi_iI(|\xi_i|\leqslant C).$$

那么, 以概率 1, 有

$$\varliminf_n\frac{S_n}{n}\geqslant\varliminf_n\frac{S_n^C}{n}=\mathbf{E}\xi_1I(\xi_1\leqslant C).$$

由于当 $C\to\infty$ 时, 有

$$\mathbf{E}\xi_1I(\xi_1\leqslant C)\to\mathbf{E}\xi_1=\infty,$$

可见 $S_n/n\to+\infty(\mathbf{P}-\text{a.c.})$.

注 3 定理 3 断定, $S_n/n \to m(\mathbf{P}-\text{a.c.})$. 需要指出, 这里除以概率 1 收敛之外, 平均收敛也成立:

$$\frac{S_n}{n} \xrightarrow{L^1} m, \quad \text{即} \quad \mathbf{E}\left|\frac{S_n}{n}-m\right| \to 0, \quad n \to \infty.$$

关于这一点, 由第五章 §3 的遍历性定理 3 的可以得到. 不过, 对于现在的独立同分布随机变量 $\xi_1, \xi_2, \cdots (S_n = \xi_1 + \cdots + \xi_n)$ 的情形, 可以直接证明 (练习题 7), 而无须借助于遍历性定理.

4. 应用强大数定律的例 现在考虑强大数定律的某些应用.

例 2 (用于数论) 设 $\Omega = [0,1), \mathscr{B}$ 是 Ω 的子集的 $\sigma-$ 代数, 而 $\mathbf{P}$ 是区间 $[0,1)$ 上的勒贝格测度. 考虑数 $\omega \in \Omega$ 的二进制分解 $\omega = 0.\ \omega_1\omega_2\cdots$ (含无限多个 0), 定义随机变量 $\xi_1(\omega), \xi_2(\omega), \cdots$, 其中 $\xi_n(\omega) = \omega_n$. 对于任意 $n \geqslant 1$ 和任意只取 0 或 1 为值的 $x_1, \cdots, x_n$, 有

$$\begin{aligned}&\{\omega : \xi_1(\omega) = x_1, \cdots, \xi_n(\omega) = x_n\}\\ &= \left\{\omega : \frac{x_1}{2} + \frac{x_2}{2^2} + \cdots + \frac{x_n}{2^n} \leqslant \omega < \frac{x_1}{2} + \frac{x_2}{2^2} + \cdots + \frac{x_n}{2^n} + \frac{1}{2^n}\right\},\end{aligned}$$

则该集合的 $\mathbf{P}-$ 测度等于 $1/2^n$. 说明 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量序列, 且

$$\mathbf{P}\{\xi_1 = 0\} = \mathbf{P}\{\xi_1 = 1\} = \frac{1}{2}.$$

由此和强大数定律得如下博雷尔的结果: *区间 $[0,1)$ 的几乎一切数按如下意义都是正规的: 其二进制分解中 "0 的比率" 和 "1 的比率" 各以概率 1 趋向 1/2*, 即.

$$\frac{1}{n}\sum_{k=1}^{n} I(\xi_k = 1) \to \frac{1}{2} \quad (\mathbf{P}-\text{a.c.}).$$

例 3 (应用于 "蒙特卡罗法") 设 $f(x)$ 是定义在区间 [0,1] 上、取值于 [0,1] 的连续函数. 下面的讨论基于积分 $\int_0^1 f(x)\mathrm{d}x$ 的数值计算的统计方法 ("蒙特卡罗法").

设 $\xi_1, \eta_1, \xi_2, \eta_2, \cdots$ 是独立同在 [0,1] 上均匀分布的随机变量序列; 而

$$\rho_i = \begin{cases} 1, & \text{若 } f(\xi_i) > \eta_i, \\ 0, & \text{若 } f(\xi_i) \leqslant \eta_i. \end{cases}$$

显然,

$$\mathbf{E}\rho_1 = \mathbf{P}\{f(\xi_1) > \eta_1\} = \int_0^1 f(x)\mathrm{d}x.$$

由强大数定律 (定理 3), 有

$$\frac{1}{n}\sum_{i=1}^{n} \rho_i \to \int_0^1 f(x)\mathrm{d}x \quad (\mathbf{P}-\text{a.c.}).$$

这样, 积分 $\int_0^1 f(x)\mathrm{d}x$ 的数值计算, 可以利用随机数偶 $(\xi_i, \eta_i), i \geqslant 1$, 模拟, 再计算 ρ_i 和 $n^{-1}\sum\limits_{i=1}^{n} \rho_i$ 的值来实现.

例 4 (更新过程的强大数定律) 设 $N = (N_t)_{t \geqslant 0}$ 是第二章 §9 第 4 小节引进的更新过程:

$$N_t = \sum_{n=1}^{\infty} I(T_n \leqslant t), \quad T_n = \sigma_1 + \cdots + \sigma_n,$$

其中 $\sigma_1, \sigma_2, \cdots, \sigma_n$ 是独立同分布正值随机变量序列. 此外, 假设 $\mu = \mathbf{E}\sigma_1 < \infty$.

在这样的假设条件下, 过程 N 服从强大数定律:

$$\frac{N_t}{t} \to \frac{1}{\mu}, \quad (\mathbf{P}-\text{a.c.}), \quad t \to \infty. \tag{14}$$

为证明这一结果, 首先由于 $T_{N_t} \leqslant t < T_{N_t+1} (t \geqslant 0)$, 可见在假设 $N_t > 0$ 的条件下, 不等式

$$\frac{T_{N_t}}{N_t} \leqslant \frac{t}{N_t} < \frac{T_{N_t+1}}{N_t+1}\left(1 + \frac{1}{N_t}\right) \tag{15}$$

成立.

显然, 当 $t \to \infty$ 时 $N_t = N_t(\omega) \to \infty$ $(\mathbf{P}-\text{a.c.})$. 同时, 根据定理 3, 当 $t \to \infty$ 时, 有

$$\frac{T_n(\omega)}{n} = \frac{\sigma_1(\omega) + \cdots + \sigma_n(\omega)}{n} \to \mu, \quad (\mathbf{P}-\text{a.c.}).$$

因此, 同样, 有

$$\frac{T_{N_t(\omega)}(\omega)}{N_t(\omega)} \to \mu, \quad (\mathbf{P}-\text{a.c.}), \quad n \to \infty,$$

从而, 由 (15) 式, 可见存在极限:

$$\lim_{t\to\infty} \frac{t}{N_t} = \mu \quad (\mathbf{P}-\text{a.c.}).$$

于是, 强大数定律 (14) 得证.

5. 练习题

1. 证明 $\mathbf{E}\xi^2 < \infty$, 当且仅当

$$\sum_{n=1}^{\infty} n\mathbf{P}\{|\xi| > n\} < \infty.$$

2. 假设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量, 证明:

(a) 若对于某个 $0 < \alpha < 1$, 有 $\mathbf{E}|\xi_1|^\alpha < \infty$, 则 $S_n/n^{1/\alpha} \to 0(\mathbf{P}-\text{a.c.})$.

(b) 若对于某个 $1 \leqslant \beta < 2$, 有 $\mathbf{E}|\xi_1|^\beta < \infty$, 则

$$\frac{S_n - n\mathbf{E}\xi_1}{n^{1/\beta}} \to 0 \quad (\mathbf{P}-\text{a.c.}).$$

3. 设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量序列, 且 $\mathbf{E}|\xi_1| = \infty$. 证明, 对于任意常数序列 $\{a_n\}$, 有

$$\overline{\lim_n}\left|\frac{S_n}{n} - a_n\right| = \infty \quad (\mathbf{P}-\text{a.c.}).$$

4. 问区间 $[0,1)$ 上的一切有理数 (在第 4 小节例 2) 是否都是正规的?

5. 举一独立随机变量序列 $\xi_1, \xi_2, \cdots$ 的例, 说明极限 $\lim_n S_n/n$ 在依概率的意义下存在, 但是在以概率 1 的意义下不存在,

6. (埃特麦迪 [N. Etemady].) 证明, 如果将随机变量 $\xi_1, \xi_2, \cdots$ "独立" 换成 "两两独立", 定理 3 的结论仍然成立.

7. 证明. 在定理 3 的条件下, 平均收敛成立:

$$\mathbf{E}\left|\frac{S_n}{n} - m\right| \to 0, \quad n \to \infty.$$

8. 设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量序列, 且 $\mathbf{E}|\xi_1|^2 < \infty$. 证明,

$$n\mathbf{P}\{|\xi_1| \geqslant \varepsilon\sqrt{n}\} \to 0 \quad 和 \quad \frac{1}{\sqrt{n}}\max_{k\leqslant n}|\xi_k| \xrightarrow{\mathbf{P}} 0.$$

9. 考虑区间 $[0,1)$ 上数 $\omega = 0.\omega_1\omega_2\cdots$ 的十进制分解.

(a) 将第 4 小节对于二进制分解的强大数定律, 移植到上面二进制分解的情形.

(b) 证明, 有理数 (按博雷尔) 不是正规的, 即对于有理数的十进制分解 $(\xi_k(\omega) = \omega_k, k \geqslant 1)$, 对于任意 $i = 0, 1, \cdots, 9$,

$$\frac{1}{n}\sum_{k=1}^{n} I(\xi_k(\omega) = i) \nrightarrow \frac{1}{10} \quad (\mathbf{P}-\text{a.c.}).$$

(c) 证明, 数 $\omega = 0.12345678910111213\cdots$, 其中在小数位上依次写的是一切非负整数, 是正规的 (见例 2).

10. (a) 设 $\xi_1, \xi_2, \cdots$ 是独立随机变量序列, 且 $\mathbf{P}\{\xi_n = \pm n^\alpha\} = 1/2$. 证明, 这一随机变量列服从大数定律, 当且仅当 $\alpha < 1/2$.

(b) 设 $f = f(x)$ 是 $(0, \infty)$ 上的有界连续函数. 证明, 对于任何 $a > 0$ 与一切 $x > 0$, 有

$$\lim_{n\to\infty}\sum_{k=1}^{\infty} f\left(x + \frac{k}{n}\right)\mathrm{e}^{-an}\frac{(an)^k}{k!} = f(x+a).$$

11. 证明, 可以赋予柯尔莫戈洛夫强大数定律 (定理 3) 如下形式: 设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量, 则

$$\mathbf{E}|\xi_1| < \infty \Leftrightarrow n^{-1}S_n \to \mathbf{E}\xi_1 \quad (\mathbf{P}-\text{a.c.}),$$
$$\mathbf{E}|\xi_1| = \infty \Leftrightarrow \overline{\lim}\, n^{-1}S_n = +\infty \quad (\mathbf{P}-\text{a.c.}).$$

证明, 如果将 "独立" 换成 "两两独立", 则第一个结论仍然成立.

12. 设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量序列. 证明,

$$\mathbf{E}\sup_n\left|\frac{\xi_n}{n}\right| < \infty \Leftrightarrow \mathbf{E}|\xi_1|\ln^+|\xi_1| < \infty.$$

13. 设 $S_n = \xi_1 + \cdots + \xi_n, n \geqslant 1$, 其中假设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量序列, 且 $\mathbf{E}\xi_1 = 0, \mathbf{E}|\xi_1| > 0$. 证明

$$\overline{\lim_n}\frac{S_n}{\sqrt{n}} = \infty, \quad \underline{\lim}\frac{S_n}{\sqrt{n}} = -\infty \quad (\mathbf{P}-\text{a.c.}).$$

14. 设 $S_n = \xi_1 + \cdots + \xi_n, n \geqslant 1$, 其中设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量序列. 证明. 对于任意 $\alpha \in (0, 1/2]$, 下面的性质之一成立:

(a) $n^{-\alpha}S_n \to \infty(\mathbf{P}-\text{a.c.})$;

(b) $n^{-\alpha}S_n \to -\infty(\mathbf{P}-\text{a.c.})$;

(c) $\overline{\lim}\, n^{-\alpha}S_n = \infty, \underline{\lim}\, n^{-\alpha}S_n = -\infty, (\mathbf{P}-\text{a.c.})$.

15. 设 $S_n = \xi_1 + \cdots + \xi_n, n \geqslant 1, S_0 = 0, \xi_1, \xi_2, \cdots$ 是独立同分布随机变量序列. 证明:

(a) 对于任意 $\varepsilon > 0$,

$$\sum_{n=1}^{\infty}\mathbf{P}\{|S_n| \geqslant n\varepsilon\} < \infty \Leftrightarrow \mathbf{E}\xi_1 = 0, \quad \mathbf{E}\xi_1^2 < \infty;$$

(b) 如果 $\mathbf{E}\xi_1 < 0$, 则对于 $p > 1$,

$$\mathbf{E}\left(\sup_{n\geqslant 0} S_n\right)^{p-1} < \infty \Leftrightarrow \mathbf{E}(\xi_1^+)^p < \infty;$$

(c) 如果 $\mathbf{E}\xi_1 = 0, 1 < p \leqslant 2$, 则对于某个常数 C_p,

$$\sum_{n=1}^{\infty}\mathbf{P}\left\{\max_{k\leqslant n} S_k \geqslant n\right\} \leqslant C_p\mathbf{E}|\xi_1|^p, \quad \sum_{n=1}^{\infty}\mathbf{P}\left\{\max_{k\leqslant n}|S_k| \geqslant n\right\} \leqslant 2C_p\mathbf{E}|\xi_1|^p;$$

(d) 如果 $\mathbf{E}\xi_1 = 0, \mathbf{E}\xi_1^2 < \infty$ 且 $M(\varepsilon) = \sup_{n\geqslant 0}(S_n - n\varepsilon), \varepsilon > 0$, 则

$$\lim_{\varepsilon\to 0}\varepsilon M(\varepsilon) = \frac{\sigma^2}{2}.$$

§4. 重对数定律

1. 辅助函数: 上函数和下函数 设 $\xi_1, \xi_2, \cdots$ 是独立同分布伯努利随机变量序列, 其中 $\mathbf{P}\{\xi_n = 1\} = \mathbf{P}\{\xi_n = -1\} = 1/2, S_n = \xi_1 + \cdots + \xi_n$. 由 §1 定理 2 的证明, 可见

$$\overline{\lim}\frac{S_n}{\sqrt{n}} = +\infty, \quad \underline{\lim}\frac{S_n}{\sqrt{n}} = -\infty. \tag{1}$$

另一方面, 根据 §3 的 (10) 式, 有

$$\frac{S_n}{\sqrt{n}\ln n} \to 0 \quad (\mathbf{P}-\text{a.c.}). \tag{2}$$

现在比较这两个结果.

由 (1) 式可见, 轨道 $(S_n)_{n\geqslant 1}$ 以概率 1 无限多次与"曲线" $\pm\varepsilon\sqrt{n}$ 相交, 其中 ε 是任意正数. 但是, 与此同时由 (2) 式可见, 只有有限多次越出有界曲线 $\pm\varepsilon\sqrt{n}\ln n$ 所围成的区域内部. 这两个关于对称随机游动 $(S_n)_{n\geqslant 1}$ 振荡"幅度"特点的结果, 提供了相当有用的信息. 下面引进的重对数定律, 本质上更精确地说明了关于随机游动 $(S_n)_{n\geqslant 1}$ 振荡"幅度"的概念.

首先引进上函数和下函数的概念.

定义 函数 $\varphi^* = \varphi^*(n), n \geqslant 1$, (对于 $(S_n)_{n\geqslant 1}$) 称做上函数, 如果自某个 $n = n_0(\omega)$ 开始对于一切 n, 以概率 1 有 $S_n \leqslant \varphi^*(n)$.

函数 $\varphi_* = \varphi_*(n), n \geqslant 1$ (对于 $(S_n)_{n\geqslant 1}$) 称做下函数, 如果对无限多个 n, 以概率 1 有 $S_n > \varphi_*(n)$.

根据这一定义, 由 (1) 式和 (2) 式可见, 每一个函数 $\varphi^* = \varepsilon\sqrt{n}\ln n(\varepsilon > 0)$, (对于 $(S_n)_{n\geqslant 1}$) 是上函数, 而函数 $\varphi_* = \varepsilon\sqrt{n}(\varepsilon > 0)$ 是下函数.

设 $\varphi = \varphi(n)$ 是某个函数, 而 $\varphi^*_\varepsilon = (1+\varepsilon)\varphi, \varphi_{*\varepsilon} = (1-\varepsilon)\varphi$, 其中 $\varepsilon > 0$. 那么, 易见

$$\left\{\overline{\lim_n}\frac{S_n}{\varphi(n)} \leqslant 1\right\} = \left\{\lim_n\left[\sup_{m\geqslant n}\frac{S_m}{\varphi(m)}\right] \leqslant 1\right\}$$
$$\Leftrightarrow \left\{\sup_{m\geqslant n_1(\varepsilon)}\frac{S_m}{\varphi(m)} \leqslant 1+\varepsilon,\ \text{对于一切}\ \varepsilon > 0\ \text{和某个}\ n_1(\varepsilon)\right\}$$
$$\Leftrightarrow \{S_m \leqslant (1+\varepsilon)\varphi(m)\ \text{对一切}\ \varepsilon > 0\ \text{和自某个}\ n_1(\varepsilon)\ \text{始的一切}\ m\}. \tag{3}$$

同样, 有

$$\left\{\overline{\lim_n}\frac{S_n}{\varphi(n)} \geqslant 1\right\} = \left\{\lim_n\left[\sup_{m\geqslant n}\frac{S_m}{\varphi(m)}\right] \geqslant 1\right\}$$
$$\Leftrightarrow \left\{\sup_{m\geqslant n_2(\varepsilon)}\frac{S_m}{\varphi(m)} \geqslant 1-\varepsilon,\ \text{对于一切}\ \varepsilon > 0\ \text{和某个}\ n_2(\varepsilon)\right\}$$
$$\Leftrightarrow \{S_m \geqslant (1-\varepsilon)\varphi(m)\ \text{对一切}\ \varepsilon > 0\ \text{和自某个}\ n_3(\varepsilon) \geqslant n_2(\varepsilon)\ \text{始的无限个}\ m\}. \tag{4}$$

由 (3) 式和 (4) 式可见, 为验证每一个函数 $\varphi^*_\varepsilon = (1+\varepsilon)\varphi, \varepsilon > 0$, 是上函数, 需要证明:

$$\mathbf{P}\left\{\overline{\lim_n}\frac{S_n}{\varphi(n)} \leqslant 1\right\} = 1. \tag{5}$$

而为证明每一个函数 $\varphi_{*\varepsilon} = (1-\varepsilon)\varphi, \varepsilon > 0$ 是下函数, 需要验证:

$$\mathbf{P}\left\{\overline{\lim_n}\frac{S_n}{\varphi(n)} \geqslant 1\right\} = 1. \tag{6}$$

2. 重对数定律

定理 1 (重对数定律) 设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量序列, 且 $\mathbf{E}\xi_i = 0$, $\mathbf{E}\xi_i^2 = \sigma^2 > 0$. 那么,

$$\mathbf{P}\left\{\overline{\lim_n}\frac{S_n}{\psi(n)} = 1\right\} = 1, \tag{7}$$

其中

$$\psi(n) = \sqrt{2\sigma^2 n\ln\ln n}. \tag{8}$$

对于一致有界随机变量, 1924 年由辛钦证明了重对数定律. 1929 年柯尔莫戈洛夫将这一结果, 推广到广泛的独立随机变量类. 在定理 1 所表述的条件下, 1941 年由哈特曼 (P. Hartman) 和温特纳 (A. Wintner) 证明了重对数定律.

由于该定理的证明相当复杂, 故我们仅限于考虑其特殊情形, 即随机变量 ξ_n 服从正态分布的情形: $\xi_n \sim N(0,1), n \geqslant 1$.

首先, 证明两个引理.

引理 1 设 $\xi_1, \cdots, \xi_n$ 是独立随机变量, 且有对称分布: 对于每一个事件 $B \in \mathscr{B}(\mathbb{R}), k \leqslant n$, 有 $\mathbf{P}\{\xi_k \in B\} = \mathbf{P}\{-\xi_k \in \mathbf{B}\}$. 那么, 对于任意实数 a, 有

$$\mathbf{P}\left\{\max_{1\leqslant k\leqslant n} S_k > a\right\} \leqslant 2\mathbf{P}\{S_n > a\}. \tag{9}$$

证明 引进记号: $A_k = \{S_i \leqslant a, i \leqslant k-1; S_k > a\}, A = \left\{\max_{1\leqslant k\leqslant n} S_k > a\right\}, B = \{S_n > a\}$. 由于在 $A_k \bigcap B \supseteq A_k \bigcap \{S_n \geqslant S_k\}$ 上, 则

$$\begin{aligned}\mathbf{P}(B\bigcap A_k) &\geqslant \mathbf{P}(A_k\bigcap\{S_n \geqslant S_k\}) = \mathbf{P}(A_k)\mathbf{P}\{S_n \geqslant S_k\}\\ &= \mathbf{P}(A_k)\mathbf{P}\{\xi_{k+1}+\cdots+\xi_n \geqslant 0\}.\end{aligned}$$

由于随机变量 $\xi_1, \cdots, \xi_n$ 的概率分布对称, 有

$$\mathbf{P}\{\xi_{k+1}+\cdots+\xi_n > 0\} = \mathbf{P}\{\xi_{k+1}+\cdots+\xi_n < 0\}.$$

因此 $\mathbf{P}\{\xi_{k+1}+\cdots+\xi_n \geqslant 0\} \geqslant 1/2$, 从而

$$\mathbf{P}(B) \geqslant \sum_{k=1}^n \mathbf{P}(A_k\bigcap B) \geqslant \frac{1}{2}\sum_{k=1}^n \mathbf{P}(A_k) = \frac{1}{2}\mathbf{P}(A),$$

于是 (9) 式得证. (对照第八章 §2 第 3 小节的证明.) □

引理 2 设 $S_n \sim N(0, \sigma^2(n)), \sigma^2(n) \uparrow \infty$, 函数 $a(n), n \geqslant 1$, 满足 $a(n)/\sigma(n) \to \infty, n \to \infty$. 那么

$$\mathbf{P}\{S_n > a(n)\} \sim \frac{\sigma(n)}{\sqrt{2\pi}a(n)}\mathrm{e}^{-\frac{a^2(n)}{2\sigma^2(n)}}. \tag{10}$$

证明 由于当 $x \to \infty$ 时

$$\frac{1}{\sqrt{2\pi}}\int_x^\infty \mathrm{e}^{-\frac{y^2}{2}}\mathrm{d}y \sim \frac{1}{\sqrt{2\pi}x}\mathrm{e}^{-\frac{x^2}{2}},$$

而 $S_n/\sigma^2(n) \sim N(0,1)$, 立即可得 (10) 式.

证明定理 1 现在证明 $\xi \sim N(0,1)$ 的情形. 假设 $\xi_i \sim N(0,1)$.

(a) 首先证明 (5) 式. 设 $\varepsilon > 0, \lambda = 1+\varepsilon, n_k = \lambda^k$, 其中 $k \geqslant k_0$, 且选择 k_0, 使 $\ln\ln k_0$ 有定义. 记

$$A_k = \{S_n > \lambda\psi(n),\ 对于某个\ n \in (n_k, n_{k+1}]\}, \tag{11}$$

亦设

$$A = \{A_k,\ 对无限多个\ k\} = \{S_n > \lambda\psi(n),\ 对于无限多个\ n\}.$$

由于 (3) 式, 为证明 (5) 式, 只需证明 $\mathbf{P}(A) = 0$.

为此证明 $\sum \mathbf{P}(A_k) < \infty$. 根据博雷尔 – 坎泰利引理 (第二章 §10), 有 $\mathbf{P}(A) = 0$.

由 (11),(9) 和 (10) 式, 有

$$\begin{aligned}&\mathbf{P}(A_k) \leqslant \mathbf{P}\{S_n > \lambda\psi(n_k),\ 对于某个\ n \in (n_k, n_{k+1}]\}\\
&\leqslant \mathbf{P}\{S_n > \lambda\psi(n_k),\ 对于某个\ n \leqslant n_{k+1}\} \leqslant 2\mathbf{P}\{S_{n_{k+1}} > \lambda\psi(n_k)\}\\
&\sim \frac{2}{\sqrt{2\pi}\frac{\lambda\psi(n_k)}{\sqrt{n_k}}}\mathrm{e}^{-\frac{1}{2}\left(\frac{\lambda\psi(n_k)}{\sqrt{n_k}}\right)^2} \leqslant C_1\mathrm{e}^{-\lambda\ln\ln\lambda^k} \leqslant C_2\mathrm{e}^{-\lambda\ln k} = C_2k^{-\lambda},\end{aligned}$$

其中 C_1 和 C_2 是常数. 由于 $\sum\limits_{k=1}^{\infty} k^{-\lambda} < \infty$, 故 $\sum \mathbf{P}(A_k) < \infty$.

这样, (5) 式得证.

(b) 其次, 证明 (6) 式. 根据 (4) 式, 需要证明对于 $\lambda = 1-\varepsilon, \varepsilon > 0$, 以概率 1 对无限多个 n, 有 $S_n \geqslant \lambda\psi(n)$. 现在将已证明的 (5) 式用于序列 $(-S_n)_{n\geqslant 1}$, 则有可能除有限个 n 的值之外的一切 n, ($\mathbf{P}$ – a.c.), 有 $-S_n \leqslant 2\psi(n)$. 从而, 如果 $n_k = N^k, N > 1$, 则对于充分大的 k, 有

$$S_{n_{k-1}} \geqslant -2\psi(n_{k-1}),$$

或

$$S_{n_k} \geqslant Y_k - 2\psi(n_{k-1}), \tag{12}$$

其中 $Y_k = S_{n_k} - S_{n_{k-1}}$.

因此, 如果证明了, 对无限多个 k,

$$Y_k > \lambda\psi(n_k) + 2\psi(n_{k-1}), \tag{13}$$

则这连同 (12) 式可以证明, 对于无限多个 k, 以概率 1 有 $S_{n_k} > \lambda\psi(n_k)$. 取某个 $\lambda' \in (\lambda, 1)$. 那么, 存在 $N > 1$, 使对一切 k,

$$\begin{aligned}&\lambda'[2(N^k - N^{k-1})\ln\ln N^k]^{1/2}\\
&> \lambda(2N^k\ln\ln N^k)^{1/2} + 2(2N^{k-1}\ln\ln N^{k-1})^{1/2}\\
&\equiv \lambda\psi(N^k) + 2\psi(N^{k-1}).\end{aligned}$$

现在只需证明, 对于无限多个 k,

$$Y_k > \lambda'[2(N^k - N^{k-1})\ln\ln N^k]^{1/2}. \tag{14}$$

显然, $Y_k \sim N(0, N^k - N^{k-1})$. 因此, 由于引理 2, 可见

$$\begin{aligned}&\mathbf{P}\{Y_k > \lambda'[2(N^k - N^{k-1})\ln\ln N^k]^{1/2}\}\\
&\sim \frac{1}{\sqrt{2\pi}\lambda'(2\ln\ln N^k)^{1/2}}\mathrm{e}^{(-\lambda')^2\ln\ln N^k}\\
&\geqslant \frac{C_1}{(\ln k)^{1/2}}k^{-(\lambda')^2} \geqslant \frac{C_2}{k\ln k}.\end{aligned}$$

因为 $\sum(k\ln k)^{-1} = \infty$, 则根据博雷尔 – 坎泰利引理的第 2 部分, 以概率 1 对于无限多个 k, (14) 式成立. 于是关系式 (6) 得证.

(c) 由 (5) 式和 (6) 式, 立即得所要证明的 (7) 式. □

注 1 将 (7) 式用于随机变量 $(-S_n)_{n\geqslant 1}$, 得 ($\mathbf{P}$ – a.c.)

$$\underline{\lim}\frac{S_n}{\psi(n)} = -1. \tag{15}$$

由 (7) 和 (15) 式可见, 重对数定律亦可表示为如下形式:

$$\mathbf{P}\left\{\overline{\lim}\frac{S_n}{\psi(n)} = 1\right\} = 1. \tag{16}$$

注 2 重对数定律说明, 对于任意 $\varepsilon > 0$, 每一个函数 $\psi_\varepsilon^* = (1+\varepsilon)\psi$ 都是上函数, 而函数 $\psi_{*\varepsilon} = (1-\varepsilon)\psi$ 是下函数.

重对数定律的结论 7 与下面的关系式等价: 对于任意 $\varepsilon > 0$,

$$\begin{aligned}&\mathbf{P}\{|S_n| \geqslant (1-\varepsilon)\psi(n),\ \ 对于无限多个\ n\} = 1,\\
&\mathbf{P}\{|S_n| \geqslant (1+\varepsilon)\psi(n),\ \ 对于无限多个\ n\} = 0.\end{aligned}$$

3. 练习题

1. 设 $\xi_1, \xi_2, \cdots$ 是独立随机变量序列, 且 $\xi_n \sim N(0,1)$. 证明

$$\mathbf{P}\left\{\overline{\lim_n}\frac{\xi_n}{\sqrt{2\ln n}} = 1\right\} = 1.$$

2. 设 $\xi_1, \xi_2, \cdots$ 是独立随机变量, 服从参数为 $\lambda > 0$ 的泊松分布. 证明 (与 λ 无关)

$$\mathbf{P}\left\{\overline{\lim_n} \frac{\xi_n \ln\ln n}{\ln n} = 1\right\} = 1.$$

3. 设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量序列, 其共同的特征函数为

$$\mathbf{E}e^{-it\xi_1} = e^{-|t|^\alpha}, \quad 0 < \alpha < 2.$$

证明

$$\mathbf{P}\left\{\overline{\lim_n} \left|\frac{S_n}{n^{1/\alpha}}\right|^{\frac{1}{\ln\ln n}} = e^{1/\alpha}\right\} = 1.$$

4. 证明不等式 (9) 的如下推广的正确性. 设 $\xi_1, \cdots, \xi_n$ 是独立随机变量, $S_0 = 0, S_k = \xi_1 + \cdots + \xi_k$, 则对于任意实数 a, 列维 (P. P. Lévy) 不等式成立:

$$\mathbf{P}\left\{\max_{1\leqslant k\leqslant n} [S_k + \mu(S_n - S_k)] > a\right\} \leqslant 2\mathbf{P}\{S_n > a\},$$

其中 $\mu(\xi)$ 是随机变量 ξ 的中位数, 即满足下列不等式的常数:

$$\mathbf{P}\{\xi \geqslant \mu(\xi)\} \geqslant \frac{1}{2}, \quad \mathbf{P}\{\xi \leqslant \mu(\xi)\} \geqslant \frac{1}{2}.$$

5. 设 $\xi_1, \cdots, \xi_n$ 是独立随机变量, $S_0 = 0, S_k = \xi_1 + \cdots + \xi_k$, 证明

(a) (第 4 题的补充)

$$\mathbf{P}\left\{\max_{1\leqslant k\leqslant n} |S_k + \mu(S_n - S_k)| > a\right\} \leqslant 2\mathbf{P}\{|S_n| \geqslant a\},$$

其中 $\mu(\xi)$ 是随机变量 ξ 的中位数.

(b) 如果 $\xi_1, \cdots, \xi_n$ 是同分布对称随机变量, 则

$$1 - e^{-n\mathbf{P}\{|\xi_1| > x\}} \leqslant \mathbf{P}\left\{\max_{1\leqslant k\leqslant n} |\xi_k| > x\right\} \leqslant 2\mathbf{P}\{|S_n| > x\}.$$

6. 设 $\xi_1, \cdots, \xi_n$ 是独立随机变量, 且 $\mathbf{E}\xi_i = 0 (1 \leqslant i \leqslant n), S_k = \xi_1 + \cdots + \xi_k$, 证明对于 $a > 0$,

$$\mathbf{P}\left\{\max_{1\leqslant k\leqslant n} S_k > a\right\} \leqslant 2\mathbf{P}\{S_n \geqslant a - \mathbf{E}|S_n|\}.$$

7. 设 $\xi_1, \cdots, \xi_n$ 是独立同分布随机变量, 且 $\mathbf{E}\xi_i = 0, \sigma^2 = \mathbf{E}\xi_i^2 < \infty, S_n = \xi_1 + \cdots + \xi_n, |\xi_i| \leqslant C(\mathbf{P} - \text{a.c.}), i \leqslant n$. 证明, 对于任意 $0 \leqslant x \leqslant 2C^{-1}$, 有

$$\mathbf{E}e^{xS_n} \leqslant \exp\{2^{-1}nx^2\sigma^2(1 + xC)\}.$$

在以上的条件下证明, 如果 (a_n) 是实数序列, 满足 $a_n/\sqrt{n} \to \infty$ 而 $a_n = o(n)$, 则对于任意 $\varepsilon > 0$ 和充分大的 n, 有

$$\mathbf{P}\{S_n > a_n\} > \exp\left\{-\frac{a_n^2}{2n\sigma^2}(1 + \varepsilon)\right\}.$$

8. 设 $\xi_1, \cdots, \xi_n$ 是独立同分布随机变量, 且 $\mathbf{E}\xi_i = 0, |\xi_i| \leqslant C(\mathbf{P} - \text{a.c.}), i \leqslant n; D_n = \sum_{i=1}^n \mathbf{D}\xi_i$. 证明, 对于 $S_n = \xi_1 + \cdots + \xi_n$, 普罗霍罗夫 (Ю. В. Прохоров) 不等式:

$$\mathbf{P}\{S_n \geqslant a_n\} \leqslant \exp\left\{-\frac{a}{2c} \arcsin \frac{ac}{2D_n}\right\}, \quad a \in \mathbb{R},$$

成立.

§5. 强大数定律的收敛速度和大偏差概率

1. 事件的频率向概率的收敛速度问题 现在考虑第一章 §6 中曾经讨论过的伯努利概型. 对于这种概型, 棣莫弗 – 拉普拉斯定理给出了标准 (正规) 偏差 $|S_n - np| \geqslant \varepsilon\sqrt{n}$ 的渐近式, 其中标准偏差即 S_n 对中心值 np 的偏差量级为 $\sqrt{n}$. 同是在第一章 §6 中, 还对所谓大偏差 $|S_n - np| \geqslant \varepsilon n$ 的概率作出了估计, 即 S_n 对 np 的 n 级偏差:

$$\mathbf{P}\left\{\left|\frac{S_n}{n} - p\right| \geqslant \varepsilon\right\} \leqslant 2e^{-2n\varepsilon^2} \tag{1}$$

(见第一章 §6 的 (42) 式). 由此当然可以得到不等式

$$\mathbf{P}\left\{\sup_{m\geqslant n}\left|\frac{S_m}{m} - p\right| \geqslant \varepsilon\right\} \leqslant \sum_{m\geqslant n} \mathbf{P}\left\{\left|\frac{S_m}{m} - p\right| \geqslant \varepsilon\right\} \leqslant \frac{2}{1 - e^{-2\varepsilon^2}} e^{-2n\varepsilon^2}, \tag{2}$$

不等式 (2) 给出了关于 "随机变量 S_n/n 以概率 1 收敛于 p 的速度" 的一定印象.

现在对于略微一般的情形, 即对于 $S_n = \xi_1 + \cdots + \xi_n$ 是独立同分布随机变量之和的情形, 讨论形如 (1), (2) 的公式正确性的问题.

2. 强大数定律中的收敛速度和大偏差概率 称随机变量 ξ 满足克拉默 (H. Cramer) 条件, 如果存在 0 的邻域, 使得对于该邻域的任何 λ, 有

$$\mathbf{E}e^{\lambda\xi} < \infty \tag{3}$$

(可以证明, 这一条件等价于 "概率 $\mathbf{P}\{|\xi| > x\}$ 以指数的速度下降".)

设

$$\varphi(\lambda) = \mathbf{E}e^{\lambda\xi}, \quad \psi(\lambda) = \ln\varphi(\lambda). \tag{4}$$

在集合

$$\Lambda = \{\lambda \in \mathbb{R} : \psi(\lambda) < \infty\} \tag{5}$$

的内部, 函数 $\psi(\lambda)$ 是凹 (向下凸) 函数, 且无限可微. 这时

$$\psi(0) = 0, \quad \psi'(0) = m (= \mathbf{E}\xi), \quad \psi''(\lambda) \geqslant 0.$$

建立函数

$$H(a) = \sup_\lambda [a\lambda - \psi(\lambda)], \quad a \in \mathbb{R}, \tag{6}$$

称做 (随机变量 ξ 的分布函数 $F=F(x)$ 的) 克拉默变换. 函数 $H(a)$ 也是凹函数, 并且在点 $a=m$ 达到其最小值 0.

如果 $a>m$, 则

$$H(a)=\sup_{\lambda>0}[a\lambda-\psi(\lambda)].$$

因此

$$\mathbf{P}\{\xi\geqslant a\}\leqslant\inf_{\lambda>0}\mathbf{E}e^{\lambda(\xi-a)}=\inf_{\lambda>0}e^{-[a\lambda-\psi(\lambda)]}=e^{-H(a)}.\tag{7}$$

完全同样, 对于 $a<m$, 有 $H(a)=\sup_{\lambda<0}[a\lambda-\psi(\lambda)]$ 和

$$\mathbf{P}\{\xi\leqslant a\}\leqslant e^{-H(a)}.\tag{8}$$

从而 (对照第一章 §6 的 (42) 式)

$$\mathbf{P}\{|\xi-m|\geqslant\varepsilon\}\leqslant 2e^{-\min\{H(m-\varepsilon),H(m+\varepsilon)\}}.\tag{9}$$

设 $\xi,\xi_1,\cdots,\xi_n$ 是独立同分布、满足克拉默条件 (3) 的随机变量, $S_n=\xi_1+\cdots+\xi_n$, $\psi_n(\lambda)=\ln\mathbf{E}\exp\left\{\lambda\frac{S_n}{n}\right\}$, $\psi(\lambda)=\ln\mathbf{E}e^{\lambda\xi}$,

$$H_n(a)=\sup_{\lambda}[a\lambda-\psi_n(\lambda)],\tag{10}$$

则

$$H_n(a)=nH(a)\left(=n\sup_{\lambda}[a\lambda-\psi(\lambda)]\right),$$

而不等式 (7), (8) 和 (9) 有如下形式:

$$\mathbf{P}\left\{\frac{S_n}{n}\geqslant a\right\}\leqslant e^{-nH(a)},\quad a>m,\tag{11}$$

$$\mathbf{P}\left\{\frac{S_n}{n}\leqslant a\right\}\leqslant e^{-nH(a)},\quad a<m,\tag{12}$$

$$\mathbf{P}\left\{\left|\frac{S_n}{n}-m\right|\geqslant\varepsilon\right\}\leqslant 2e^{-\min\{H(m-\varepsilon),H(m+\varepsilon)\}\times n}.\tag{13}$$

注 1 形如

$$\mathbf{P}\left\{\left|\frac{S_n}{n}-m\right|\geqslant\varepsilon\right\}\leqslant ae^{-bn}\tag{14}$$

的结果, 其中 $a>0$ 和 $b>0$, 称做关于常数 a 和 b 的 "正则" 指数收敛. 在大偏差理论中相应的结果常用略有不同的、比较 "粗糙" 形式表示为:

$$\overline{\lim_n}\frac{1}{n}\ln\mathbf{P}\left\{\left|\frac{S_n}{n}-m\right|\geqslant\varepsilon\right\}<0.\tag{15}$$

这种表示形式自然出于 (14) 式, 在于说明 "指数" 收敛速度, 但是不强调常数 a 和 b 的值.

现在考虑概率

$$\mathbf{P}\left\{\sup_{k\geqslant n}\frac{S_k}{k}>a\right\},\quad\mathbf{P}\left\{\inf_{k\geqslant n}\frac{S_k}{k}<a\right\},\quad\mathbf{P}\left\{\sup_{k\geqslant n}\left|\frac{S_k}{k}-m\right|>\varepsilon\right\}$$

自上侧估计的问题. 这些概率可以给出关于强大数定律中收敛速度的一定印象.

假设独立同分布非退化随机变量 $\xi,\xi_1,\xi_2,\cdots$ 满足克拉默条件 (3).

固定 $n\geqslant 1$, 设

$$\kappa=\inf\left\{k\geqslant n:\frac{S_k}{k}>a\right\},$$

当 $S_k/k\leqslant a, k\geqslant n$ 时, 设 $\kappa=\infty$.

其次, 设 a 和 $\lambda>0$, 满足

$$\lambda a-\ln\varphi(\lambda)\geqslant 0.\tag{16}$$

那么,

$$\begin{aligned}\mathbf{P}\left\{\sup_{k\geqslant n}\frac{S_k}{k}>a\right\}&=\mathbf{P}\left(\bigcup_{k\geqslant n}\left\{\frac{S_k}{k}>a\right\}\right)=\mathbf{P}\left\{\frac{S_\kappa}{\kappa}>a,\kappa<\infty\right\}\\&=\mathbf{P}\left\{e^{\lambda S_\kappa}>e^{\lambda a\kappa},\kappa<\infty\right\}=\mathbf{P}\left\{e^{\lambda S_k-\kappa\ln\varphi(\lambda)}>e^{\kappa[\lambda a-\ln\varphi(\lambda)]},\kappa<\infty\right\}\\&\leqslant\mathbf{P}\left\{e^{\lambda S_\kappa-\kappa\ln\varphi(\lambda)}>e^{n[\lambda a-\ln\varphi(\lambda)]},\kappa<\infty\right\}\\&\leqslant\mathbf{P}\left\{\sup_{k\geqslant n}e^{\lambda S_k-k\ln\varphi(\lambda)}\geqslant e^{n[\lambda a-\ln\varphi(\lambda)]}\right\}.\end{aligned}\tag{17}$$

注意, 在证明的最后一步, 随机变量序列

$$e^{\lambda S_k-k\ln\varphi(\lambda)},\quad k\geqslant 1,$$

关于 $\sigma-$ 代数流 $\mathscr{F}_k=\sigma(\xi_1,\cdots,\xi_k), k\geqslant 1$, 是鞅. (详见第七章, 特别见 §1 例 2.)

那么, 由第七章 §3 不等式 (8), 可以导出

$$\mathbf{P}\left\{\sup_{k\geqslant n}e^{\lambda S_k-k\ln\varphi(\lambda)}\geqslant e^{n[\lambda a-\ln\varphi(\lambda)]}\right\}\leqslant e^{-n[\lambda a-\ln\varphi(\lambda)]},$$

从而, (在 (16) 式的条件下) 得不等式:

$$\mathbf{P}\left\{\sup_{k\geqslant n}\frac{S_k}{k}>a\right\}\leqslant e^{-n[\lambda a-\ln\varphi(\lambda)]}.\tag{18}$$

设 $a>m$. 由于对函数 $f(\lambda)=\lambda a-\ln\varphi(\lambda)$, 有 $f(0)=0, f'(0)>0$, 则存在 $\lambda>0$ 使 (16) 式成立. 从而, 由 (18) 式可见: 如果 $a>m$, 则

$$\mathbf{P}\left\{\sup_{k\geqslant n}\frac{S_k}{k}>a\right\}\leqslant e^{-n\sup_{\lambda>0}[\lambda a-\ln\varphi(\lambda)]}=e^{-nH(a)}.\tag{19}$$

同样, 如果 $a < m$, 则

$$\mathbf{P}\left\{\inf_{k\geqslant n}\frac{S_k}{k} < a\right\} \leqslant \mathrm{e}^{-n\sup\limits_{\lambda<0}[\lambda a-\ln\varphi(\lambda)]} = \mathrm{e}^{-nH(a)}. \tag{20}$$

由 (19) 和 (20) 式, 可见

$$\mathbf{P}\left\{\sup_{k\geqslant n}\left|\frac{S_k}{k} - m\right| > \varepsilon\right\} \leqslant 2\mathrm{e}^{-\min\{H(m-\varepsilon),H(m+\varepsilon)\}\times n}. \tag{21}$$

注 2 由 (11) 和 (19) 式右侧的相同, 使我们想到这并非偶然. 事实上, 其解释在于, 对于任何 $n \leqslant N$, 序列 $(S_k/k)_{n\leqslant k\leqslant N}$ 是*可逆鞅* (见第七章 §1 练习题 5, 和第一章 §11 例 4).

3. 练习题

1. 证明不等式 (8) 和 (20).

2. 验证在集合 Λ (见 (5) 式) 的内部, 函数 $\psi(\lambda)$ 是凹 (向下凸) 函数 (若随机变量 ξ 是非退化的, 而且是*严格*凹函数), 并且无限可微.

3. 在随机变量 ξ 是非退化的假设条件下, 函数 $H(a)$ 在全直线上可微, 并且是凹函数.

4. 证明如下克拉姆变换的反演公式:

$$\psi(\lambda) = \sup_a[a\lambda - H(a)]$$

(对于一切 λ, 只有集合 $\Lambda = \{\lambda : \psi(\lambda) < \infty\}$ 的有限个点可能除外).

5. 设 $S_n = \xi_1 + \cdots + \xi_n$, 其中 $\xi_1, \cdots, \xi_n, n \geqslant 1$, 是独立同分布的简单随机变量, 且 $\mathbf{E}\xi_1 < 0, \mathbf{P}\{\xi_1 > 0\} > 0$. 假设 $\varphi(\lambda) = \mathbf{E}\mathrm{e}^{\lambda\xi_1}$, 而 $\inf_\lambda \varphi(\lambda) = \rho (0 < \rho < 1)$.

证明下面的 (切尔诺夫 [Чернов]) 定理:

$$\lim_n \frac{1}{n}\ln \mathbf{P}\{S_n \geqslant 0\} = \ln\rho. \tag{22}$$

6. 利用 (22) 式的结果, 证明, 对于伯努利概型 ($\mathbf{P}\{\xi_1 = 1\} = p, \mathbf{P}\{\xi_1 = 0\} = q$), 当 $p < x < 1$ 时,

$$\lim_n \frac{1}{n}\ln \mathbf{P}\{S_n \geqslant nx\} = -H(x), \tag{23}$$

其中 (参照第一章 §6 的记号)

$$H(x) = x\ln\frac{x}{p} + (1-x)\ln\frac{1-x}{1-p}.$$

7. 设 $S_n = \xi_1 + \cdots + \xi_n, n \geqslant 1$, 其中 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量, 且 $\mathbf{E}\xi_1 = 0, \mathbf{D}\xi_1 = 1$. 设数列 $(x_n)_{n\geqslant 1}$ 满足: $x_n \to \infty$, 而当 $n \to \infty$ 时 $x_n/\sqrt{n} \to 0$.

证明

$$\mathbf{P}\{S_n \geqslant x_n\sqrt{n}\} = \mathrm{e}^{-\frac{x_n^2}{2}}(1 + y_n),$$

其中当 $n \to \infty$ 时 $y_n \to 0$.

8. 由 (23) 式证明, 对于伯努利概型: $\mathbf{P}\{\xi_1 = 1\} = p, \mathbf{P}\{\xi_1 = 0\} = q$,

(a) 当 $p < x < 1$, 且 $x_n = n(x - p)$ 时,

$$\mathbf{P}\{S_n \geqslant np + x_n\} = \exp\left\{-nH\left(p + \frac{x_n}{n}\right)(1 + o(1))\right\}; \tag{24}$$

(b) 当 $x_n = a_n\sqrt{npq}$ 且 $a_n \to \infty, a_n/\sqrt{n} \to 0$ 时,

$$\mathbf{P}\{S_n \geqslant np + x_n\} = \exp\left\{-\frac{x_n^2}{2npq}(1 + o(1))\right\}. \tag{25}$$

将 (24) 和 (25) 两式与第一章 §6 中相应的结果进行比较.

第五章 强 (狭义) 平稳随机序列和遍历理论

在概率论的框架之外, 可以将强平稳随机序列理论, 表述为可测测度空间上的单参数保测变换群理论, 它与动态系统的一般理论和遍历性理论临近.

《数学百科全书》, (中译本) 第 4 卷, 第 985 页 [121]

§1. 强 (狭义) 平稳随机序列. 保测变换

1. 强平稳随机变量序列 设 $(\Omega, \mathscr{F}, \mathbf{P})$ 是一概率空间, $\xi = (\xi_1, \xi_2, \cdots)$ 是一随机变量序列, 或简称*随机序列*. 以 $\theta_k\xi$ 表示序列 $(\xi_{k+1}, \xi_{k+2}, \cdots)$.

定义 1 随机变量序列 ξ 称做*强平稳的*, 如果对于任意 $k \geqslant 1$, $\theta_k\xi$ 和 ξ 有相同的概率分布:

$$\mathbf{P}\{(\xi_1, \xi_2, \cdots) \in B\} = \mathbf{P}\{(\xi_{k+1}, \xi_{k+2}, \cdots) \in B\}, \quad B \in \mathscr{B}(\mathbb{R}^\infty).$$

由独立同分布随机变量 $\xi = (\xi_1, \xi_2, \cdots)$ 形成的序列, 是平稳序列 ξ 最简单的例子. 从这样的序列出发, 可以构造广泛的平稳序列类 $\eta = (\eta_1, \eta_2, \cdots)$, 其中 $\eta_k = g(\xi_k, \xi_{k+1}, \cdots, \xi_{k+n})$, 而 $g(x_1, \cdots, x_n)$ 是任意博雷尔函数.

如果 $\xi = (\xi_1, \xi_2, \cdots)$ 是独立同分布随机变量序列, $\mathbf{E}|\xi_1| < \infty, \mathbf{E}\xi_1 = m$, 则根据强大数定律, 以概率 1

$$\frac{\xi_1 + \cdots + \xi_n}{n} \to m, \quad n \to \infty.$$

伯克霍夫 (G. D. Birkhoff) 在 1931 年作为*力学定理*, 得到了这一结果出色的推广. 所谓力学定理, 涉及动态系统的 "相对到达时间" 的性质, 这里的动态系统是由微分方程描述的, 而假设微分方程有积分变式 ("守恒系统").

辛钦于 1932 年紧接着证明了, 实际上伯克霍夫定理可以推广到更一般情形: "多维空间在自身中的运动, 并且保持集合的测度不变".

我们将既在 "动态系统" 理论的框架内, 又在 "强平稳随机序列" 的框架内, 同时叙述伯克霍夫和辛钦的结果.

这时基本着重点放在该理论有关 "遍历性" 的结果上.

2. 保测变换 设 $(\Omega, \mathscr{F}, \mathbf{P})$ 是一 (完备) 概率空间.

定义 2 如果对于任何 $A \in \mathscr{F}$, 有

$$T^{-1}A = \{\omega : T\omega \in A\} \in \mathscr{F},$$

则空间 Ω 到自身的影射 T 称做*可测的*.

定义 3 如果对于任何 $A \in \mathscr{F}$, 有

$$\mathbf{P}(T^{-1}A) = \mathbf{P}(A),$$

则可测影射 T 称做保测变换 (morphism).

设 T 是保测变换, T^n 是其 n 次幂, 而 $\xi_1 = \xi_1(\omega)$ 是一随机变量. 设 $\xi_n = \xi_1(T^{n-1}\omega), n \geqslant 2$, 考虑序列 $\xi = (\xi_1, \xi_2, \cdots)$. 我们证明此序列是平稳的.

事实上, 设 $A = \{\omega : \xi \in B\}, A_1 = \{\omega : \theta_1 \xi \in B\}$, 其中 $B \in \mathscr{B}(\mathbb{R}^\infty)$, 则 $\omega \in A_1$ 当且仅当 $T\omega \in A$, 即 $A_1 = T^{-1}A$. 由于 $\mathbf{P}(T^{-1}A) = \mathbf{P}(A)$, 可见 $\mathbf{P}(A_1) = \mathbf{P}(A)$. 同样, 对于任意事件 $A_k = \{\omega : \theta_k \xi \in B\}, k \geqslant 2$, 有 $\mathbf{P}(A_k) = \mathbf{P}(A)$.

这样, 所引进的保测变换, 为建立强平稳序列提供了可能性.

在一定意义上有相反的结果: 对每一平稳序列 ξ, 对所考虑概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$, 存在新概率空间 $(\widetilde{\Omega}, \widetilde{\mathscr{F}}, \widetilde{\mathbf{P}})$, 以及随机变量 $\widetilde{\xi}_1(\widetilde{\omega})$ 和保测变换 $\widetilde{T}$, 使随机序列 $\widetilde{\xi} = \{\widetilde{\xi}_1(\widetilde{\omega}), \widetilde{\xi}_1(\widetilde{T}\widetilde{\omega}), \cdots\}$ 与随机序列 $\xi = \{\xi_1(\omega), \xi_2(\omega), \cdots\}$ 有相同的分布.

事实上, 作为 $\widetilde{\Omega}$ 取 "坐标" 空间 $\mathbb{R}^\infty$, 并设 $\widetilde{\mathscr{F}}(\mathbb{R}^\infty), \widetilde{\mathbf{P}} = P_\xi$, 其中 $P_\xi(B) = \mathbf{P}\{\omega : \xi \in B\}, B \in \mathscr{B}(\mathbb{R}^\infty)$. 空间 $\widetilde{\Omega}$ 的变换 $\widetilde{T}$, 决定于关系式 $\widetilde{T}(x_1, x_2, \cdots) = (x_2, x_3, \cdots)$. 对于 $\widetilde{\omega} = (x_1, x_2, \cdots)$, 记

$$\widetilde{\xi}_1(\widetilde{\omega}) = x_1, \quad \widetilde{\xi}_n(\widetilde{\omega}) = \widetilde{\xi}_1(\widetilde{T}^{n-1}\widetilde{\omega}), \quad n \geqslant 2.$$

现在设 $A = \{\widetilde{\omega} : (x_1, \cdots, x_k) \in B\}, B \in \mathscr{B}(\mathbb{R}^k)$, 且 $\widetilde{T}^{-1}A = \{\widetilde{\omega} : (x_2, \cdots, x_{k+1}) \in B\}$. 那么, 由平稳性, 可见

$$\widetilde{\mathbf{P}}(A) = \mathbf{P}\{\omega : (\xi_1, \cdots, \xi_k) \in B\} = \mathbf{P}\{\omega : (\xi_2, \cdots, \xi_{k+1}) \in B\} = \mathbf{P}(\widetilde{T}^{-1}A),$$

即 $\widetilde{T}$ 是保测变换. 由于对于任意 k,

$$\widetilde{\mathbf{P}}\{\widetilde{\omega} : (\widetilde{\xi}_1, \cdots, \widetilde{\xi}_k) \in B\} = \mathbf{P}\{\omega : (\xi_1, \cdots, \xi_k) \in B\},$$

则由此可见 ξ 和 $\widetilde{\xi}$ 有相同的分布.

下面是保测变换的例子.

例 1 设 $\Omega = \{\omega_1, \cdots, \omega_n\}, n \geqslant 2$, 是由有限个点构成的集合, $\mathscr{F}$ 是 Ω 中一切子集的 $\sigma-$ 代数, $T\omega_i = \omega_{i+1}, 1 \leqslant i \leqslant n-1$, 而 $T\omega_n = \omega_1$. 如果 $\mathbf{P}(\omega_i) = 1/n$, 则 T 是保测变换.

例 2 设 $\Omega = [0,1), \mathscr{F} = \mathscr{B}([0,1)), \mathbf{P}$ 是勒贝格测度, $\lambda \in [0,1)$, 则 $Tx = (x + \lambda)$ 是保测变换.

现在考虑导致研究保测变换的物理前提条件.

假设某一系统按照一定运动规律 (在离散时间) 演变, 并设想 Ω 是该系统的状态 ω 的相空间. 那么, 如果 ω 是系统在时刻 $n = 1$ 的状态, 则 $T^n\omega$ 是经过 n 步系统进入

的状态, 其中 T 是 (该运动规律诱导的) 推移算子. 其次, 假如 A 是某一 "状态 ω 的集合", 则 $T^{-1}A = \{\omega : T\omega \in A\}$ 根据其定义是经一步到达集合 A 的一切 "初始" 状态 ω 的集合. 因此, 假如把 Ω 视为 "不可压缩的液体", 则条件 $\mathbf{P}(T^{-1}A) = \mathbf{P}(A)$ 可以视为完全自然的 "体积" 保持不变的条件. (对于经典的封闭哈密顿 (W. R. Hamilton) 系统, 著名的刘维尔 (J. Liouville) 定理断定, 相应的变换 T 是保持勒贝格测度不变的变换.)

3. 关于保测变换的庞加莱定理 下面关于 "常返性" 的庞加莱 (J. H. Poincaré) 定理 (1912), 是有关保测变换最早的成果之一.

定理 *设 $(\Omega, \mathscr{F}, \mathbf{P})$ 是一概率空间, T 是保测变换, $A \in \mathscr{F}$. 那么, 对于无限多个 $n \geqslant 1$ 和几乎一切点 $\omega \in A$, 有 $T^n\omega \in A$.*

证明 记 $C = \{\omega \in A : T^n\omega \notin A,$ 对于一切 $n \geqslant 1\}$. 由于对于任意 $n \geqslant 1, C \bigcap T^{-n}C = \varnothing$, 则 $T^{-m}C \bigcap T^{-(m+n)}C = T^{-m}(C \bigcap T^{-n}C) = \varnothing$. 这样, 序列 $\{T^{-n}C\}$ 由有相同 $\mathbf{P}-$ 测度的不相交集合构成. 因此,

$$\sum_{n=1}^{\infty} \mathbf{P}(C) = \sum_{n=1}^{\infty} \mathbf{P}(T^{-n}C) \leqslant \mathbf{P}(\Omega) = 1,$$

从而 $\mathbf{P}(C) = 0$. 从而, 对于几乎一切点 $\omega \in A$ 和至少对于一个 $n \geqslant 1, T^n\omega \in A$. 由此可见, 对于无限个 $n \geqslant 1$, 有 $T^n\omega = A$.

将上面得到的结果用于变换 $T^k\omega, k \geqslant 1$. 那么, 对于每一个点 $\omega \in A \backslash N$, 其中 N 是 0 概率集合 (并且 N 是对应于不同 k 的相应集合的并), 存在这样的 n_k, 使 $(T^k)^{n_k}\omega \in A$. 由此显然可以得到, 对于无限多个 n, 有 $T^n\omega \in A$. □

系 *设 $\xi(\omega) \geqslant 0$, 则在集合 $\{\omega : \xi(\omega) > 0\}$ 上*

$$\sum_{k=0}^{\infty} \xi(T^k\omega) = \infty \quad (\mathbf{P}-\text{a.c.}).$$

实际上, 设 $A_n = \{\omega : \xi(\omega) \geqslant 1/n\}$, 那么, 根据上面的定理在集合 A_n 上, 有

$$\sum_{k=0}^{n} \xi(T^k\omega) = \infty \quad (\mathbf{P}-\text{a.c.}),$$

如果令当 $n \to \infty$ 时, 得所要证明的结果.

注 假如将概率测度 $\mathbf{P}$ 换成任意有限测度 $\mu(\mu(\Omega) < \infty)$, 则定理仍然成立.

4. 练习题

1. 设 T 是保测变换, $\xi = \xi(\omega)$ 是一随机变量, 并且有数学期望 $\mathbf{E}\xi(\omega)$. 证明 $\mathbf{E}\xi(\omega) = \mathbf{E}\xi(T\omega)$.

2. 证明, 例 1 和例 2 中变换 T 是保测变换.

3. 设 $\Omega = [0,1), \mathscr{F} = \mathscr{B}([0,1)), \mathbf{P}$ 是具有连续分布函数测度, 证明变换

$$Tx = \lambda x (0 < \lambda < 1) \quad \text{和} \quad Tx = x^2$$

都不是保测变换.

4. 设 Ω 是一切实数序列 $\omega=(\cdots,\omega_{-1},\omega_0,\omega_1,\cdots)$ 的集合, $\mathscr{F}$ 是由可测柱集诱导的 $\sigma-$ 代数, 其中的可测柱集为 $\{\omega:(\omega_k,\cdots,\omega_{k+n-1})\in B_n\}(n=1,2,\cdots;k=0,\pm1,\pm2,\cdots)$, 而 $B_n\in\mathscr{B}(\mathbb{R}^n)$. 以 $\mathbf{P}$ 表示 $(\Omega,\mathscr{F})$ 上的概率测度, 而双侧变换 T 决定于公式:

$$T(\cdots,\omega_{-1},\omega_0,\omega_1,\cdots)=(\cdots,\omega_0,\omega_1,\omega_2,\cdots).$$

证明, T 是保测变换, 当且仅当对于一切 $n=1,2,\cdots,k=0,\pm1,\pm2,\cdots$ 和 $B_n\in\mathscr{B}(\mathbb{R}^n)$

$$\mathbf{P}\{\omega:(\omega_0,\cdots,\omega_{n-1})\in B_n\}=\mathbf{P}\{\omega:(\omega_k,\cdots,\omega_{k+n-1})\in B_n\}.$$

5. 设 $\xi_0,\xi_1,\cdots$ 是一平稳随机元序列, $\xi_n(n\geqslant0)$ 的值属于博雷尔空间 S (见第二章 §7 定义 9). 证明, 可以构造 (有可能在原概率空间的扩展空间上) 在 S 中取值的随机元素 $\xi_{-1},\xi_{-2},\cdots$, 使双侧序列 $\cdots,\xi_{-1},\xi_0,\xi_1,\cdots$ 是平稳的.

6. 设 T 是 $(\Omega,\mathscr{F},\mathbf{P})$ 上的可测变换, $\mathscr{E}$ 是由 $\mathscr{F}(\pi(\mathscr{E})=\mathscr{F})$ 诱导的 Ω 子集的 $\pi-$ 系. 证明, 如果对于 $A\in\mathscr{E}$ 等式 $\mathbf{P}(T^{-1}A)=\mathbf{P}(A)$ 成立, 则此等式对于 $A\in\mathscr{F}(=\pi(\mathscr{E}))$ 也成立.

7. 设 T 是在 $(\Omega,\mathscr{F},\mathbf{P})$ 上的可测变换, $\mathscr{G}$ 是 $\mathscr{F}$ 的子 $\sigma-$ 代数. 证明对每个 $A\in\mathscr{F}$, 有

$$\mathbf{P}(A|\mathscr{G})(T\omega)=\mathbf{P}(T^{-1}A|T^{-1}\mathscr{G})(\omega)\quad(\mathbf{P}-\text{a.c.}).\tag{1}$$

特别, 设 $\Omega=\mathbb{R}^\infty$ 是数列 $\omega=(\omega_0,\omega_1,\cdots)$ 的空间, 而 $\xi_k(\omega)=\omega_k$ 的; T是推移变换: $T(\omega_0,\omega_1,\cdots)=(\omega_1,\omega_2,\cdots)$, 换句话说, 若 $\xi_k(\omega)=\omega_k$, 则 $\xi_k(T\omega)=\omega_{k+1}$. 那么, 等式 (1) 有如下形式

$$\mathbf{P}(A|\xi_n)(T\omega)=\mathbf{P}(T^{-1}A|\xi_{n+1})(\omega)\quad(\mathbf{P}-\text{a.c.}).$$

8. 设 T 是在 $(\Omega,\mathscr{F})$ 上的可测变换, $\mathscr{P}$ 是所有概率测度 $\mathbf{P}$ 的集合, 而关于概率测度 $\mathbf{P}$, 变换 T 是保持测度 $\mathbf{P}$ 不变的变换. 证明:

(a) 集合 $\mathscr{P}$ 是凸的;

(b) 变换 T 关于概率测度 $\mathbf{P}$ 是遍历变换, 当且仅当 $\mathbf{P}$ 是集合 $\mathscr{P}$ 的边界点 (即不能表示为如下形式的点: $\mathbf{P}=\lambda_1\mathbf{P}_1+\lambda_2\mathbf{P}_2,\lambda_1>0,\lambda_2>0,\lambda_1+\lambda_2=1,\mathbf{P}_1\neq\mathbf{P}_2,\mathbf{P}_1,\mathbf{P}_2\in\mathscr{P}$).

§2. 遍历性与混合性

1. 保测变换的遍历性 在整个这一节中, 我们都将以 T 表示作用于概率空间 $(\Omega,\mathscr{F},\mathbf{P})$ 上的保测变换.

定义 1 集合 $A\in\mathscr{F}$ 称做*不变的*, 如果 $T^{-1}A=A$. 集合 $A\in\mathscr{F}$ 称做*几乎不变的*, 如果 A 和 $T^{-1}A=A$ 只相差一 0 测集, 即 $\mathbf{P}(A\triangle T^{-1}A)=0$.

不难验证, 不变 (几乎不变) 集合类 $\mathscr{I}$ (相应地 $\mathscr{I}^*$) 构成 $\sigma-$ 代数.

定义 2 保测变换 T 称做*遍历的* (或*度量可递变换*), 如果每一个不变集合 A 的测度只能为 0 或 1.

定义 3 随机变量 $\eta=\eta(\omega)$ 称做*不变的* (或*几乎不变的*), 如果对于一切 $\omega\in\Omega$ (对于几乎一切 $\omega\in\Omega$)$\eta(\omega)=\eta(T\omega)$.

下面的引理建立不变集合与几乎不变集合之间的联系.

引理 1 *如果 A 是几乎不变集合, 则存在一几乎不变集合 B, 使 $\mathbf{P}(A\triangle B)=0$.*

证明 设 $B=\overline{\lim}T^{-n}A$, 则 $T^{-1}B=\overline{\lim}T^{-(n+1)}A=B$, 即 $B\in\mathscr{I}$. 不难验证

$$A\triangle B\subseteq\bigcup_{k=0}^{\infty}(T^{-k}A\triangle T^{-(k+1)}A).$$

由于 $\mathbf{P}(T^{-k}A\triangle T^{-(k+1)}A)=\mathbf{P}(A\triangle T^{-1}A)=0$, 可见 $\mathbf{P}(A\triangle B)=0$. □

引理 2 *变换 T 是遍历的, 当且仅当每一个几乎不变集合 A 的测度只能为 0 或 1.*

证明 设 $A\in\mathscr{I}^*$, 则根据引理 1, 存在一几乎不变集合 B, 使 $\mathbf{P}(A\triangle B)=0$. 但是, 由于变换 T 是遍历的, 故 $\mathbf{P}(B)=0$ 或 1. 由于 $\mathscr{I}\subseteq\mathscr{I}^*$, 故逆命题显然成立. □

定理 1 *设 T 是保测变换. 下列各个条件等价:*

(1) *T 是遍历的;*

(2) *每一个几乎不变随机变量以概率 $\mathbf{P}$ 为 1 是常数;*

(3) *每一个不变随机变量以概率 $\mathbf{P}$ 为 1 是常数.*

证明 (1)⇒(2). 设 T 是遍历的, 而 ξ 是几乎不变的, 即以概率 1 有 $\xi(\omega)=\xi(T\omega)$. 那么, 对于任何 $c\in\mathbb{R}$, 集合 $A_c=\{\omega:\xi(\omega)\leqslant c\}\in\mathscr{I}^*$, 而根据引理 2, $\mathbf{P}(A_c)=0$ 或 1. 设 $C=\sup\{c:\mathbf{P}(A_c)=0\}$. 当 $c\uparrow\infty$ 时 $A_c\uparrow\Omega$, 而当 $c\downarrow-\infty$ 时 $A_c\downarrow\varnothing$, 故 $C<\infty$.

那么,

$$\mathbf{P}\{\omega:\xi(\omega)<C\}=\mathbf{P}\left(\bigcup_{n=1}^{\infty}\left\{\xi(\omega)\leqslant C-\frac{1}{n}\right\}\right)=0;$$

类似可得, $\mathbf{P}\{\omega:\xi(\omega)>C\}=0$. 于是, $\mathbf{P}\{\omega:\xi(\omega)=C\}=1$.

(2)⇒(3). 显然.

(3)⇒(1). 设 $A\in\mathscr{I}$, 则 I_A 是不变随机变量, 因此以概率 1 有 $I_A=0$ 或 $I_A=1$, 故 $\mathbf{P}(A)=0$ 或 1. □

注 如果在定理中的随机变量有界, 则定理的结论仍然成立.

为演示该定理的应用, 我们看下面的例子.

例 设 $\Omega=[0,1),\mathscr{F}=\mathscr{B}([0,1)),\mathbf{P}$ 是勒贝格测度, 而 $T\omega=(\omega+1)\text{mod }1$. 证明 T 是遍历变换, 当且仅当 λ 是无理数.

设 $\xi=\xi(\omega)$ 是不变随机变量, $\mathbf{E}\xi^2(\omega)<\infty$. 熟知, 具有的 $\mathbf{E}\xi^2(\omega)<\infty$ 函数 $\xi(\omega)$ 的傅里叶级数 $\sum\limits_{n=-\infty}^{\infty} c_n e^{2\pi in\omega}$ 均方收敛, 且 $\sum|c_n|^2<\infty$. 由于 T 是保测变换 (§1 例 2), 则由于所假设的随机变量 $\xi=\xi(\omega)$ 的不变性 (§1 练习题 1). 可见

$$\begin{aligned}c_n&=\mathbf{E}\xi(\omega)e^{-2\pi in\omega}=\mathbf{E}\xi(T\omega)e^{-2\pi inT\omega}=e^{-2\pi in\lambda}\mathbf{E}\xi(T\omega)e^{-2\pi in\omega}\\&=e^{-2\pi in\lambda}\mathbf{E}\xi(\omega)e^{-2\pi in\omega}=c_ne^{-2\pi in\lambda}.\end{aligned}$$

因此 $c_n(1-e^{-2\pi in\lambda})=0$. 根据条件 λ 是无理数, 即对于一切 $n\neq 1, e^{-2\pi in\lambda}\neq 1$. 于是, $c_n=0, n\neq 1, \xi(\omega)=c_0(\mathbf{P}-\text{a.c.})$, 而根据定理 1, T 是遍历变换.

另一方面, 假如 λ 是有理数, 即 $\lambda=k/m$, 其中 k 和 m 是整数. 考虑集合

$$A=\bigcup_{k=0}^{2m-2}\left\{\omega:\frac{k}{2m}\leqslant\omega<\frac{k+1}{2m}\right\}.$$

显然, 集合 A 不变的, 然而 $\mathbf{P}(A)=1/2$. 从而, T 不是遍历的.

2. 变换的混合性及其与遍历性的关系

定义 4 称保测变换 T 为混合 (具有混合性的), 如果对于任意 $A,B\in\mathscr{F}$, 有

$$\lim_{n\to\infty}\mathbf{P}(A\cap T^{-n}B)=\mathbf{P}(A)\mathbf{P}(B).\qquad(1)$$

下面的定理指出了变换的混合性与遍历性的关系.

定理 2 *具有混合性的任何变换 T 都是遍历的.*

证明 设 $A\in\mathscr{F}, B\in\mathscr{S}$, 则 $B=T^{-n}B, n\geqslant 1$, 即对于一切 $n\geqslant 1$, 有

$$\mathbf{P}(A\cap T^{-n}B)=\mathbf{P}(A\cap B).$$

因此由于 (1) 式, 当 $A=B$ 时 $\mathbf{P}(B)=\mathbf{P}^2(B)$, 于是, $\mathbf{P}(B)=0$ 或 1. □

3. 练习题

1. 证明随机变量 $\xi=\xi(\omega)$ 是不变的, 当且仅当它 $\mathscr{S}-$ 可测.

2. 证明集合 A 是几乎不变的, 当且仅当 $\mathbf{P}(T^{-1}A\setminus A)=0$.

3. 证明变换 T 具有混合性, 当且仅当对于任何两个随机变量 ξ 和 η : $\mathbf{E}\xi^2<\infty, \mathbf{E}\eta^2<\infty$, 当 $n\to\infty$ 时, 有

$$\mathbf{E}\xi(T^{-1}\omega)\eta(\omega)\to\mathbf{E}\xi(\omega)\mathbf{E}\eta(\omega).$$

4. 举例说明, 保测遍历变换未必具有混合性质.

5. 设 T 是空间 $(\Omega,\mathscr{F},\mathbf{P})$ 上的可测变换, $\mathscr{A}$ 是 Ω 的子集的 $\sigma-$ 代数, 而 $\sigma(\mathscr{A})=\mathscr{F}$. 假如定义 4 仅要求性质

$$\lim_{n\to\infty}\mathbf{P}(A\cap T^{-n}B)=\mathbf{P}(A)\mathbf{P}(B)$$

对于 $A,B\in\mathscr{A}$ 成立. 证明此性质对于一切 $A,B\in\mathscr{F}=\sigma(\mathscr{A})$ 成立. (从而变换 T 是混合).

证明如果 $\mathscr{A}$ 是 $\pi-$ 系且 $\pi(\mathscr{A})=\mathscr{F}$, 则上述结论仍然成立.

6. 设 A 是几乎不变集合. 证明, 当且仅当对于一切 $n=1,2,\cdots,T^n\omega\in A$, 以概率 1, 有 $\omega\in A$. (对照 §1 的定理.)

7. 举一空间 $(\Omega,\mathscr{F},\mathbf{P})$ 上的保测变换 T 的例, 使得:

(a) 由 $A\in\mathscr{F}$ 绝对得不出 $TA\in\mathscr{F}$;

(b) 由 $A\in\mathscr{F}$ 和 $TA\in\mathscr{F}$ 绝对得不出 $\mathbf{P}(A)=\mathbf{P}(TA)$.

§3. 遍历性定理

1. 毕达哥拉斯 – 辛钦定理

定理 1 (毕达哥拉斯 [Pythagoras] – 辛钦) *设 T 保测变换, 而 $\xi=\xi(\omega)$ 是随机变量, $\mathbf{E}|\xi|<\infty$. 那么,*

$$\lim_n\frac{1}{n}\sum_{k=0}^{n-1}\xi(T^k\omega)=\mathbf{E}(\xi|\mathscr{S})\quad(\mathbf{P}-\text{a.c.}).\qquad(1)$$

而且, 若 T 是遍历的, 则

$$\lim_n\frac{1}{n}\sum_{k=0}^{n-1}\xi(T^k\omega)=\mathbf{E}\xi\quad(\mathbf{P}-\text{a.c.}).\qquad(2)$$

该定理的证明本质上基于下面的引理, 其简单的证明是加尔西亚 (A. M. Garsia, 1965 年) 得到的.

引理 (最大遍历性定理) *设 T 保测变换, 而 $\xi=\xi(\omega)$ 是随机变量, $\mathbf{E}|\xi|<\infty$, 且*

$$\begin{aligned}S_k(\omega)&=\xi(\omega)+\xi(T\omega)+\cdots+\xi(T^{k-1}\omega),\\M_k(\omega)&=\max\{0,S_1(\omega),\cdots,S_k(\omega)\}.\end{aligned}$$

那么, 对于任意 $n\geqslant 1$, 有

$$\mathbf{E}[\xi(\omega)I_{\{M_n>0\}}(\omega)]\geqslant 0.$$

证明 如果 $n>k$, 则 $M_n(T\omega)\geqslant S_k(T\omega)$, 从而

$$\xi(\omega)+M_n(T\omega)\geqslant\xi(\omega)+S_k(T\omega)=S_{k+1}(\omega).$$

因为显然 $\xi(\omega)\geqslant S_1(\omega)-M_n(T\omega)$, 所以

$$\xi(\omega)\geqslant\max\{S_1(\omega),\cdots,S_n(\omega)\}-M_n(T\omega).$$

注意到, 由于 $\{M_n(\omega)>0\}=\{\max[S_1(\omega),\cdots,S_n(\omega)]>0\}$, 可见

$$\begin{aligned}\mathbf{E}[\xi(\omega)I_{\{M_n>0\}}(\omega)] &\geqslant \mathbf{E}[\max\{S_1(\omega),\cdots,S_n(\omega)\}-M_n(T\omega)I_{\{M_n>0\}}(\omega)]\\ &\geqslant \mathbf{E}[M_n(\omega)-M_n(T\omega)I_{\{M_n(\omega)>0\}}(\omega)]\\ &\geqslant \mathbf{E}[M_n(\omega)-M_n(T\omega)]=0,\end{aligned}$$

其中用到, 如果 T 是保测变换, 则 $\mathbf{E}M_n(\omega)=\mathbf{E}M_n(T\omega)$ (§1 的练习题 1). □

证明定理 1 假设 $\mathbf{E}(\xi|\mathscr{I})=0$ (否则用 $\xi-\mathbf{E}(\xi|\mathscr{I})$ 替换 ξ). 设

$$\overline{\eta}=\overline{\lim_n}\frac{S_n}{n},\quad \underline{\eta}=\underline{\lim_n}\frac{S_n}{n}.$$

为证明 (1) 式, 只需证明以概率 1, 由

$$0\leqslant\underline{\eta}\leqslant\overline{\eta}\leqslant 0.$$

考虑随机变量 $\overline{\eta}=\overline{\eta}(\omega)$. 由于 $\overline{\eta}=\eta(T\omega)$, 可见 $\overline{\eta}$ 是不变的. 从而, 对于每一个 $\varepsilon>0$, 集合 $A_\varepsilon=\{\overline{\eta}(\omega)>\varepsilon\}$ 是不变的. 现在引进新的随机变量:

$$\xi^*(\omega)=[\xi(\omega)-\varepsilon]I_{A_\varepsilon}(\omega),$$

并设

$$\begin{aligned}S_k^*(\omega)&=\xi^*(\omega)+\xi^*(T\omega)+\cdots+\xi^*(T^{k-1}\omega),\\ M_k^*(\omega)&=\max\{0,S_1^*(\omega),\cdots,S_k^*(\omega)\}.\end{aligned}$$

那么, 根据引理, 对于任意 $n\geqslant1$, 有

$$\mathbf{E}[\xi^*I_{\{M_n^*>0\}}]\geqslant 0.$$

而当 $n\to\infty$ 时

$$\begin{aligned}\{M_n^*>0\}&=\left\{\sup_{1\leqslant k\leqslant n}S_k^*>0\right\}\uparrow\left\{\sup_{k\geqslant1}S_k^*>0\right\}\\ &=\left\{\sup_{k\geqslant1}\frac{S_k^*}{k}>0\right\}=\left\{\sup_{k\geqslant1}\frac{S_k}{k}>\varepsilon\right\}\bigcap A_\varepsilon=A_\varepsilon,\end{aligned}$$

其中由于 $\sup_{k\geqslant1}S_k/k>\overline{\eta}$, 而且 $A_\varepsilon=\{\omega:\overline{\eta}>\varepsilon\}$, 可见最后一等式成立.

此外, $\mathbf{E}|\xi^*|\leqslant\mathbf{E}|\xi|+\varepsilon$. 因此, 根据控制收敛定理, 有

$$0\leqslant\mathbf{E}[\xi^*I_{\{M_n^*>0\}}]\to\mathbf{E}[\xi^*I_{A_\varepsilon}].$$

于是,

$$\begin{aligned}0\leqslant\mathbf{E}[\xi^*I_{A_\varepsilon}]&=\mathbf{E}[(\xi-\varepsilon)I_{A_\varepsilon}]=\mathbf{E}[\xi I_{A_\varepsilon}]-\varepsilon\mathbf{P}(A_\varepsilon)\\ &=\mathbf{E}[\mathbf{E}(\xi|\mathscr{I})I_{A_\varepsilon}]-\varepsilon\mathbf{P}(A_\varepsilon)=-\varepsilon\mathbf{P}(A_\varepsilon),\end{aligned}$$

因此 $\mathbf{P}(A_\varepsilon)=0$, 从而, 有 $\mathbf{P}\{\overline{\eta}\leqslant0\}=1$.

类似地, 用 $-\xi(\omega)$ 代替 $\xi(\omega)$, 有

$$\overline{\lim_n}\left(-\frac{S_n}{n}\right)=-\underline{\lim_n}\frac{S_n}{n}=-\underline{\eta},$$

而 $\mathbf{P}\{-\underline{\eta}\leqslant0\}=1$, 即 $\mathbf{P}\{\underline{\eta}\geqslant0\}=1$. 因而, 以概率 1 有 $0\leqslant\underline{\eta}\leqslant\overline{\eta}\leqslant0$. 于是, 证明了定理的第一个结论.

为证明了定理的第二个结论, 只需注意到, 由于 $\mathbf{E}(\xi|\mathscr{I})$ 是不变随机变量, 可见若 T 是遍历的, 则以概率 1 有 $\mathbf{E}(\xi|\mathscr{I})=\mathbf{E}\xi$. □

系 *保测变换 T 是遍历的, 当且仅当对于任意 $A,B\in\mathscr{F}$, 有*

$$\lim_n\frac1n\sum_{k=0}^{n-1}\mathbf{P}(A\bigcap T^{-k}B)=\mathbf{P}(A)\mathbf{P}(B).\tag{3}$$

为证明 T 的遍历性, 在 (3) 式中设 $A=B\in\mathscr{I}$, 则 $A\bigcap T^{-k}B=B$, 故 $\mathbf{P}(B)=\mathbf{P}^2(B)$, 因此 $\mathbf{P}(B)=0$ 或 1. 相反, 设 T 是遍历的, 则对于随机变量 $\xi=I_B(\omega)$, 由 (2) 式可见以概率 1 有

$$\lim_n\frac1n\sum_{k=0}^{n-1}I_{T^{-k}B}(\omega)=\mathbf{P}(B).$$

于该式两侧在集合上同时求积分, 并利用控制收敛定理, 得所要证明的关系式 (3).

2. 在平均收敛意义下的毕达哥拉斯 – 辛钦定理 现在证明, 在定理 1 的条件中的 (1) 式和 (2) 式, 不仅以概率 1 收敛, 而且平均收敛也成立. (下面定理 3 的证明将要用到这一结果.)

定理 2 *设 T 保测变换, 而 $\xi=\xi(\omega)$ 是随机变量, $\mathbf{E}|\xi|<\infty$. 那么,*

$$\mathbf{E}\left|\frac1n\sum_{k=0}^{n-1}\xi(T^k\omega)-\mathbf{E}(\xi|\mathscr{I})\right|\to0,\quad n\to\infty.\tag{4}$$

而且, 若 T 是遍历的, 则

$$\mathbf{E}\left|\frac1n\sum_{k=0}^{n-1}\xi(T^k\omega)-\mathbf{E}(\xi)\right|\to0,\quad n\to\infty.\tag{5}$$

证明 对于任意 $\varepsilon>0$, 存在一有界随机变量 $\eta(|\eta(\omega)|\leqslant M)$, 使 $\mathbf{E}|\xi-\eta|\leqslant\varepsilon$. 那么,

$$\begin{aligned}\mathbf{E}\left|\frac1n\sum_{k=0}^{n-1}\xi(T^k\omega)-\mathbf{E}(\xi|\mathscr{I})\right|&\leqslant\mathbf{E}\left|\frac1n\sum_{k=0}^{n-1}[\xi(T^k\omega)-\eta(T^k\omega)]\right|\\ &+\mathbf{E}\left|\frac1n\sum_{k=0}^{n-1}[\eta(T^k\omega)-\mathbf{E}(\eta|\mathscr{I})]\right|+\mathbf{E}|\mathbf{E}(\xi|\mathscr{I})-\mathbf{E}(\eta|\mathscr{I})|.\end{aligned}\tag{6}$$

由于 $|\eta| \leqslant M$, 则由控制收敛定理和 (1) 式可见, 当 $n \to \infty$ 时, (6) 式中不等号的右侧第二项趋向 0. 至于第一项和第三项中, 每一项都小于或等于 ε. 因此, 对于充分大的 n, (6) 式小于 2ε, 从而 (4) 式得证. 最后, 如果若 T 是遍历的, 则由 (4) 式, 以及定理 1 之证明末尾的说明: "若 T 是遍历的, 则以概率 1 有 $\mathbf{E}(\xi|\mathscr{S}) = \mathbf{E}\xi$", 立即得 (5) 式. □

3. 遍历性定理 现在考虑遍历性对于强平稳随机序列正确性的问题. 设 $\xi = (\xi_1, \xi_2, \cdots)$ 是定义在概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 上的强平稳随机序列. 一般, 在 $(\Omega, \mathscr{F}, \mathbf{P})$ 上有可能并不存在保测变换, 故不能直接运用定理 1. 不过, 在 §1 中已经指出, 可以考虑建立这样的 (坐标) 概率空间 $(\widetilde{\Omega}, \widetilde{\mathscr{F}}, \widetilde{\mathbf{P}})$、随机序列 $\widetilde{\xi} = (\widetilde{\xi}_1, \widetilde{\xi}_2, \cdots)$ 和保测变换 $\widetilde{T}$, 其中 $\widetilde{\xi}_n(\widetilde{\omega}) = \widetilde{\xi}_1(\widetilde{T}^{-1}\widetilde{\omega})$, 使得 ξ 和 $\widetilde{\xi}$ 按分布相等. 由于这一性质, 像几乎必然收敛与平均收敛一样, 只与概率分布有关, 则对于某个随机变量 $\widetilde{\eta}$, 由以概率 1 收敛与平均收敛意义下

$$\frac{1}{n}\sum_{k=1}^{n}\widetilde{\xi}_1(\widetilde{T}^{k-1}\widetilde{\omega}) \to \widetilde{\eta},$$

可得 $n^{-1}\sum\limits_{k=1}^{n}\xi_k(\omega)$ 也以概率 1 收敛与平均收敛意义下收敛于某随机变量 η: $\eta \overset{d}{=} \widetilde{\eta}$. 由定理 1 可见, 如果 $\widetilde{\mathbf{E}}|\widetilde{\xi}_1| < \infty$, 则 $\widetilde{\eta} = \widetilde{\mathbf{E}}(\widetilde{\xi}_1|\widetilde{\mathscr{S}})$, 其中 $\widetilde{\mathscr{S}}$ 是不变集合的全体 ($\widetilde{\mathbf{E}}$ 是对测度 $\widetilde{\mathbf{P}}$求平均). 现在描绘随机变量 η的结构.

定义 1 称集合 $A \in \mathscr{F}$ 相对于序列 ξ *为不变的*, 如果存在集合 $B \in \mathscr{B}(\mathbb{R}^\infty)$, 使对于任何 $n \geqslant 1$, 有

$$A = \{\omega : (\xi_n, \xi_{n+1}, \cdots) \in B\}.$$

这样不变集合的全体构成 $\sigma-$ 代数, 记作 $\mathscr{S}_\xi$.

定义 2 平稳序列 ξ 称为*遍历的*, 如果任意不变集合的测度只有 0 或 1 两个可能值.

现在证明, 当 $n \to \infty$ 时, 随机变量 η 在以概率 1 收敛与平均收敛意义下, 是随机变量 $n^{-1}\sum\limits_{k=1}^{n}\xi_k(\omega)$ 的极限, 且可以取作 $\mathbf{E}(\xi_1|\mathscr{S})$. 为此首先注意到, 自然可以设

$$\eta(\omega) = \overline{\lim_n}\frac{1}{n}\sum_{k=1}^{n}\xi_n(\omega). \tag{7}$$

由上极限 $\overline{\lim}$ 的定义可见, 对于这样定义的随机变量 $\eta(\omega)$, 集合 $\{\omega : \eta(\omega) < y\}(y \in \mathbb{R})$ 是不变的, 从而 η 为 $\mathscr{S}_\xi-$ 可测. 其次, 设 $A \in \mathscr{S}_\xi$. 那么, 由于

$$\mathbf{E}\left|\frac{1}{n}\sum_{k=1}^{n-1}\xi_k - \eta\right| \to 0, \quad n \to \infty,$$

可见对于由 (7) 式定义的 η, 有

$$\frac{1}{n}\sum_{k=1}^{n}\int_A \xi_k \mathrm{d}\mathbf{P} \to \int_A \eta \mathrm{d}\mathbf{P}. \tag{8}$$

设 $B \in \mathscr{B}(\mathbb{R}^\infty)$ 满足条件: 对于任意 $k \geqslant 1$, 有 $A = \{\omega : (\xi_k, \xi_{k+1}, \cdots) \in B\}$. 那么, 由于 ξ 的平稳性, 可见

$$\int_A \xi_k \mathrm{d}\mathbf{P} = \int_{\{\omega:(\xi_k,\xi_{k+1},\cdots)\in B\}} \xi_k \mathrm{d}\mathbf{P} = \int_{\{\omega:(\xi_1,\xi_2,\cdots)\in B\}} \xi_1 \mathrm{d}\mathbf{P} = \int_A \xi_1 \mathrm{d}\mathbf{P}.$$

因此, 由 (8) 式可见, 对于任意 $A \in \mathscr{S}_\xi$, 有如下等式

$$\int_A \xi_1 \mathrm{d}\mathbf{P} = \int_A \eta \mathrm{d}\mathbf{P}.$$

该式表明 (见第二章 §7 的 (1) 式), ($\mathscr{S}_\xi-$ 可测) 随机变量 $\eta = \mathbf{E}(\xi_1|\mathscr{S}_\xi)$, 并且如果序列 ξ 是遍历的, 则 $\mathbf{E}(\xi_1|\mathscr{S}_\xi) = \mathbf{E}\xi_1$.

于是, 证明了下面的定理.

定理 3 (遍历性定理) *设 $\xi = (\xi_1, \xi_2, \cdots)$ 是强平稳序列, $\mathbf{E}|\xi_1| < \infty$. 那么, (在以概率 1 收敛与平均收敛意义下)*

$$\lim_n \frac{1}{n}\sum_{k=1}^{n}\xi_k(\omega) = \mathbf{E}(\xi_1|\mathscr{S}_\xi).$$

并且, 如果同时 ξ 又是遍历序列, 则 (在以概率 1 收敛与平均收敛意义下)

$$\lim_n \frac{1}{n}\sum_{k=1}^{n}\xi_k(\omega) = \mathbf{E}\xi_1.$$

4. 练习题

1. 设 $\xi = (\xi_1, \xi_2, \cdots)$ 是高斯平稳序列, $\mathbf{E}\xi_n = 0$, 而相关函数为 $R(n) = \mathbf{E}\xi_{k+n}\xi_k$. 证明, $R(n) \to 0$ 的充分条件是, 对应于随机变量序列 ξ 的保测变换为混合 (因而是*遍历的*).

2. 对于任何独立同分布随机变量序列 $\xi = (\xi_1, \xi_2, \cdots)$, 相应的保测变换都是混合.

3. 证明平稳序列 ξ 为遍历的, 当且仅当对于任意 $B \in \mathscr{B}(\mathbb{R}^k), k = 1, 2, \cdots$, 以概率 1,

$$\frac{1}{n}\sum_{i=1}^{n} I_B(\xi_i, \cdots, \xi_{i+k-1}) \to \mathbf{P}\{(\xi_1, \cdots, \xi_k) \in B\}.$$

4. 设 $\mathbf{P}$ 和 $\overline{\mathbf{P}}$ 是 $(\Omega, \mathscr{F})$ 上的两个概率测度, 且关于 $\mathbf{P}$ 和 $\overline{\mathbf{P}}$ 的保测变换 T 是遍历的. 证明, 这时要么 $\mathbf{P} = \overline{\mathbf{P}}$, 要么 $\mathbf{P} \perp \overline{\mathbf{P}}$.

5. 设 T 是 $(\Omega, \mathscr{F}, \mathbf{P})$ 上的保测变换, 而 $\mathscr{A}$ 是 Ω 的子集的代数, 且 $\sigma(\mathscr{A}) = \mathscr{F}$; 设

$$I_A^{(n)} = \frac{1}{n}\sum_{k=0}^{n-1} I_A(T^k\omega).$$

证明变换 T 是遍历的, 当且仅当满足下列条件之一:

(a) 对于任意 $A \in \mathscr{A}, I_A^{(n)} \xrightarrow{\mathbf{P}} \mathbf{P}(A)$;

(b) 对于一切 $A, B \in \mathscr{A}$,

$$\lim \frac{1}{n}\sum_{k=0}^{n-1}\mathbf{P}(A \bigcap T^{-k}B) = \mathbf{P}(A)\mathbf{P}(B);$$

(c) 对于任意 $A \in \mathscr{F}, I_A^{(n)} \xrightarrow{\mathbf{P}} \mathbf{P}(A)$.

6. 设 T 是 $(\Omega, \mathscr{F}, \mathbf{P})$ 上的保测变换. 证明变换 T (关于测度 $\mathbf{P}$)是遍历的, 当且仅当在 $(\Omega, \mathscr{F})$ 上不存在概率测度 $\overline{\mathbf{P}} \neq \mathbf{P}, \overline{\mathbf{P}} \ll \mathbf{P}$, 而且 T 关于测度 $\overline{\mathbf{P}}$ 是保测变换.

7. (伯努利推移变换) 假设 S 是一有限集合 (例如, $S = \{1, 2, \cdots, N\}$), 而 $\Omega = S^{\infty}$ 是序列 $\omega = (\omega_0, \omega_1, \cdots)$ 的全体, 其中 $\omega_i \in S$. 设 $\xi_k(\omega) = \omega_k$, 并且定义推移变换: $T(\omega_0, \omega_1, \cdots) = (\omega_1, \omega_2, \cdots)$, 即对于 $\xi_0, \xi_1, \cdots$, 如果 $\xi_k(\omega) = \omega_k$, 则 $\xi_k(T\omega) = \omega_{k+1}$. 假设在集合 $S = \{1, 2, \cdots, N\}$ 的元素 i 上给定正数 p_i, 且 $p_1 + p_2 + \cdots + p_N = 1$ (即数组 $(p_1, p_2, \cdots, p_N)$ 是概率分布). 利用这一变换可以在 $(S^{\infty}, \mathscr{B}(S^{\infty}))$ 上定义测度 $\mathbf{P}$ (见第二章 §3)

$$\mathbf{P}\{\omega : (\omega_1, \cdots, \omega_k) = (u_1, \cdots, u_k)\} = p_{u_1} \cdots p_{u_k}.$$

换句话说, 按 $\xi_0(\omega), \xi_1(\omega), \cdots$ 独立的原则引进概率. 关于这样建立的测度 $\mathbf{P}$, 推移变换 T 习惯上称做伯努利推移或伯努利变换.

证明伯努利变换具有混合性质.

8. 设 T 是 $(\Omega, \mathscr{F}, \mathbf{P})$ 上的保测变换. 引进记号 $T^{-n}\mathscr{F} = \{T^{-n}A : A \in \mathscr{F}\}$, 并且称 $\sigma-$ 代数

$$\mathscr{F}_{-\infty} = \bigcap_{n=1}^{\infty} T^{-n}\mathscr{F}$$

是平凡的 ($\mathbf{P}-$ 平凡的), 如果中的每一个集合的测度为 0 或 1 (这样的变换称做柯尔莫戈洛夫变换). 证明柯尔莫戈洛夫变换具有遍历性, 并且具有混合性.

9. 设 $1 \leqslant p < \infty, T$ 是概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 上的保测变换, 而随机变量 $\xi(\omega) \in L^p(\Omega, \mathscr{F}, \mathbf{P})$.

证明空间 $L^p(\Omega, \mathscr{F}, \mathbf{P})$ 上的如下 (冯・诺伊曼 [J. von Neuman]) 遍历性定理:

$$\mathbf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1}\xi(T^k\omega) - \eta(\omega)\right|^p \to 0, \quad n \to \infty.$$

10. 根据博雷尔定理 (第四章 §3 例 2): 区间 $[0, 1)$ 上数 ω 的二进制分解中 "1 的比率" 和 "0 的比率" (关于勒贝格测度) 几乎必然收敛于 1/2. 如果将这一结果视为由公式

$$T(\omega) = 2\omega (\bmod\ 1)$$

定义的区间 $[0, 1)$ 到自身的变换 $T : [0, 1) \to [0, 1)$, 试利用遍历性定理 1, 证明上述博雷尔定理.

11. 同上题一样, 设 $\omega \in [0, 1)$. 考虑由公式

$$T(\omega) = \begin{cases} 0, & 若\ \omega = 0, \\ \left\{\dfrac{1}{\omega}\right\}, & 若\ \omega \neq 0 \end{cases}$$

定义的变换 $T : [0, 1) \to [0, 1)$, 其中 $\{x\}$ 表示数 x 的分数部分.

设 $P = P(\cdot)$ 是区间 $[0, 1)$ 上的高斯测度, 由如下公式定义:

$$P(A) = \frac{1}{\ln 2}\int_A \frac{\mathrm{d}x}{1 + x}, \quad A \in \mathscr{B}([0, 1)).$$

证明变换 T 保持高斯测度不变.

12. 举例说明, 对于具有无限测度的可测空间, 庞加莱 "常返定理" (§1 第 3 小节) 一般不成立.

第六章　弱 (广义) 平稳随机序列. L^2 理论

谱表示在弱平稳随机过程理论的中心位置，是研究弱平稳随机过程的基础. 任何平稳过程可看成，具有随机振幅和相位的、各种频率的互不相关的调和振动的叠加.

《数学百科全书》(中译本) 第 4 卷, 第 914 页 [121]

§1. 协方差函数的谱表示

1. 平稳过程的概念与协方差函数 根据上一章给出的定义, 随机序列 $\xi=(\xi_1, \xi_2,\cdots)$ 称做*强平稳序列*①, 如果对于任意集合 $B\in\mathscr{B}(\mathbb{R}^\infty)$ 与任何 $n\geqslant 1$,

$$\mathbf{P}\{(\xi_1,\xi_2,\cdots)\in B\}=\mathbf{P}\{(\xi_{n+1},\xi_{n+2},\cdots)\in B\}. \tag{1}$$

特别, 由此可见, 若 $\mathbf{E}\xi_1^2<\infty$, 则 $\mathbf{E}\xi_n$ 不依赖于 n:

$$\mathbf{E}\xi_n=\mathbf{E}\xi_1, \tag{2}$$

而且协方差 $\operatorname{cov}(\xi_{n+m},\xi_n)=\mathbf{E}(\xi_{n+m}-\mathbf{E}\xi_{n+m})(\xi_n-\mathbf{E}\xi_n)$ 只依赖于 m:

$$\operatorname{cov}(\xi_{n+m},\xi_n)=\operatorname{cov}(\xi_{1+m},\xi_1). \tag{3}$$

这一章将研究 (具有有限二阶矩的) 所谓*弱 (广义) 平稳随机序列*, 这里将条件 (1) 换为条件 (2) 和 (3).

假设所考虑的随机变量 ξ_n, 对于 $n\in\mathbb{Z}=\{0,\pm1,\cdots\}$ 有定义, 而且取复数为值. 这一假设非但不使理论复杂化, 反而使之更加雅致. 这时, 实值随机变量的有关结果, 自然容易由复值随机变量的相应结果, 作为其特殊情形得到.

设 $H^2=H^2(\Omega,\mathscr{F},\mathbf{P})$ 是 (复值) 随机变量 ξ 的值空间: $\xi=\alpha+i\beta,\alpha,\beta\in\mathbb{R}$, 且$\mathbf{E}|\xi|^2<\infty$, 其中 $|\xi|^2=\alpha^2+\beta^2$. 如果 $\xi,\eta\in H^2$, 则设

$$(\xi,\eta)=\mathbf{E}\xi\overline{\eta}, \tag{4}$$

其中 $\overline{\eta}=\alpha-i\beta$ 是 $\eta=\alpha+i\beta$ 的共轭随机变量. 记

$$\|\xi\|=(\xi,\xi)^{1/2}. \tag{5}$$

①强平稳序列 (过程), 亦称严格意义下的平稳序列 (过程), 或严平稳序列 (过程). 弱平稳序列 (过程), 亦称广泛意义下的平稳序列 (过程), 或宽平稳序列 (过程). —— 译者

确切一点说, H^2 是等价随机变量类的空间 (见第二章 §10 和 §11). 像实值随机变量一样, 引进了内积 (ξ,η) 和范数 $\|\xi\|$ 的空间 H^2 是*完备的*. 按泛函分析的术语, 空间 H^2 称做所考虑概率空间 $(\Omega,\mathscr{F},\mathbf{P})$ 上的随机变量的*酉 (或复) 希尔伯特* (D. Hilbert) *空间*.

如果 $\xi,\eta\in H^2$, 则称

$$\operatorname{cov}(\xi,\eta)=\mathbf{E}(\xi-\mathbf{E}\xi)\overline{(\eta-\mathbf{E}\eta)} \tag{6}$$

为 ξ 和 η 的*协方差*.

由 (4) 和 (6) 式可见, 如果 $\mathbf{E}\xi=\mathbf{E}\eta=0$, 则

$$\operatorname{cov}(\xi,\eta)=(\xi,\eta). \tag{7}$$

定义 复值随机变量序列 $\xi=(\xi_n)_{n\in\mathbb{Z}},\mathbf{E}|\xi_n|^2<\infty,n\in\mathbb{Z}$, 称做*弱广义平稳的*, 如果对于所有 $n\in\mathbb{Z}$, 有

$$\begin{aligned}\mathbf{E}\xi_n&=\mathbf{E}\xi_0,\\ \operatorname{cov}(\xi_{n+k},\xi_n)&=\operatorname{cov}(\xi_n,\xi_0),\quad k\in\mathbb{Z}.\end{aligned} \tag{8}$$

为叙述简便, 以后总是假设 $\mathbf{E}\xi_0=0$. 这一假设并不影响一般性, 然而 (由于 (7) 式) 却可以使协方差与数积有相同的形式, 因此使希尔伯特空间理论的方法和结果用起来更加简单.

记

$$R(n)=\operatorname{cov}(\xi_n,\xi_0),\quad n\in\mathbb{Z}, \tag{9}$$

和 (在假设 $R(0)=\mathbf{E}|\xi_0|^2\neq 0$ 的条件下)

$$\rho(n)=\frac{R(n)}{R(0)},\quad n\in\mathbb{Z}. \tag{10}$$

我们把函数 $R(n)$ 称做 (弱平稳序列的) *协方差函数*, 而 $\rho(n)$ 称做*相关函数*.

直接由定义 (9) 可见, 协方差函数 $R(n)$ 是*非负定函数*, 即对于任意复数 $a_1,\cdots, a_m$, 与任何 $t_1,\cdots,t_m\in\mathbb{Z},m\geqslant 1$, 有

$$\sum_{i,j=1}^m a_i\overline{a}_jR(t_i-t_j)\geqslant 0. \tag{11}$$

同样, 由此 (或由 (9) 式) 不难得到 (练习题 1) 如下协方差函数的性质:

$$\begin{gathered}R(0)\geqslant 0,\quad R(-n)=\overline{R(n)},\quad |R(n)|\leqslant R(0),\\ |R(n)-R(m)|^2\leqslant 2R(0)[R(0)-\operatorname{Re}R(n-m)].\end{gathered} \tag{12}$$

2. 平稳序列的例 下面是一些平稳序列 $\xi=(\xi_n)_{n\in\mathbb{Z}}$ 的例. (为简便计, 下面常省略弱平稳过程的"弱"字, 以及$n\in\mathbb{Z}$).

例 1 设 $\xi_n=\xi_0 g(n)$, 其中 $\mathbf{E}\xi_0=0, \mathbf{E}|\xi_0|^2=1$, 而 $g=g(n)$ 是某一函数. 随机变量序列 $\xi=(\xi_n)$ 是平稳的, 当且仅当函数 $g(k+n)\overline{g(k)}$ 仅依赖于 n. 由此不难看到, 存在这样的 λ, 使

$$g(n)=g(0)\mathrm{e}^{-i\lambda n}.$$

因此, 随机变量序列

$$\xi_n=\xi_0 g(0)\mathrm{e}^{i\lambda n}$$

是平稳的, 且其协方差函数为

$$R(n)=|g(0)|^2\mathrm{e}^{i\lambda n}.$$

特别, "在时间上随机的常数" $\xi_n\equiv\xi_0$ 形成平稳序列.

注 我们关于例 1 指出, 由于 $\mathrm{e}^{i\lambda n}=\mathrm{e}^{in(\lambda+2\pi k)}, k=\pm1,\pm2,\cdots$, 则 (圆) 频率 λ 的值仅精确到任意被加项的 2π 倍. 按照习惯, 以后总假定 $\lambda\in[-\pi,\pi)$.

例 2 *殆周期性序列* 设

$$\xi_n=\sum_{k=1}^{N} z_k\mathrm{e}^{i\lambda_k n},\quad n\in\mathbb{Z}, \tag{13}$$

其中 $z_1,\cdots,z_N$ 是正交随机变量: $\mathbf{E}z_i\bar{z}_j=0, i\neq j$, 而且均值 $\mathbf{E}z_k=0; \mathbf{E}|z_k|^2=\sigma_k^2>0, -\pi\leqslant\lambda_k<\pi, k=1,\cdots,N; \lambda_i\neq\lambda_j(i\neq j)$.

因此, 随机变量序列 $\xi=(\xi_n)$ 是平稳的, 且其协方差函数为

$$R(n)=\sum_{k=1}^{N}\sigma_k^2\mathrm{e}^{i\lambda_k n}. \tag{14}$$

为推广 (13) 式, 现在设

$$\xi_n=\sum_{k=-\infty}^{\infty} z_k\mathrm{e}^{i\lambda_k n}, \tag{15}$$

其中 $z_k, k\in\mathbb{Z}$, 具有与 (13) 式中同样的性质. 如果假设 $\sum\limits_{k=-\infty}^{\infty}\sigma_k^2<\infty$, 则 (15) 式右侧的级数均方收敛, 并且

$$R(n)=\sum_{k=-\infty}^{\infty}\sigma_k^2\mathrm{e}^{i\lambda_k n}. \tag{16}$$

引进函数

$$F(\lambda)=\sum_{\{k:\lambda_k\leqslant\lambda\}}\sigma_k^2. \tag{17}$$

那么, 协方差函数 (16) 可以写成勒贝格 – 斯蒂尔切斯积分的形式

$$R(n)=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}\mathrm{d}F(\lambda)\left(=\int_{[-\pi,\pi)}\mathrm{e}^{i\lambda n}\mathrm{d}F(\lambda)\right). \tag{18}$$

平稳序列 (15) 是由"谐波" $\mathrm{e}^{i\lambda_k n}$ 之和构成, 其"频率"为 λ_k, 而其随机"振幅" z_k 的"强度"为 $\sigma_k^2=\mathbf{E}|z_k|^2$. 这样, 函数值 $F(\lambda)$ 关于序列 ξ 的"谱"的结构, 即关于构成表现 (15) 式之各种不同频率的强度值, 提供全面的信息. 根据 (18) 式, 函数 $F(\lambda)$ 的值亦完全决定协方差函数 $R(n)$ 的构造.

精确到常数因子, (非退化) 函数 $F(\lambda)$ 显然是分布函数, 并且在该例中函数 $F(\lambda)$ 是阶梯函数. 特别值得注意的是, 任意弱平稳序列的协方差函数可以表示为 (18) 式 (见第 3 小节的定理), 其中 $F(\lambda)$ (精确到规范因子) 是分布函数, 其承载子, 集中在区间 $[-\pi,\pi)$ 上, 即

$$F(\lambda)=\begin{cases}F(\pi-), & 若\ \lambda\geqslant\pi,\\ 0, & 若\ \lambda<-\pi.\end{cases}$$

由 (15) 和 (16) 两式形成的、协方差函数 $R(n)$ 的积分表示结果, 使我们产生一种想法: 任意平稳序列都可以用"积分"表示. 实际上也正是这样, 在 §3 利用按正交测度的随机积分将要证明这一点.

例 3 *白噪声* 设 $\varepsilon=(\varepsilon_n)$ 是正交规范随机变量序列, $\mathbf{E}\varepsilon_n=0, \mathbf{E}\varepsilon_i\varepsilon_j=\delta_{ij}$, 其中 δ_{ij} 是克罗内克符号. 显然 $\varepsilon=(\varepsilon_n)$ 是平稳序列, 其协方差函数为

$$R(n)=\begin{cases}1, & 若\ n=0,\\ 0, & 若\ n\neq0.\end{cases}$$

注意, 函数 $R(n)$ 可以表示为

$$R(n)=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}\mathrm{d}F(\lambda), \tag{19}$$

其中

$$F(\lambda)=\int_{-\pi}^{\lambda}f(v)\mathrm{d}v,\quad f(\lambda)=\frac{1}{2\pi},\quad -\pi\leqslant\lambda<\pi. \tag{20}$$

比较"谱"函数 (17) 和 (20) 式可见, 如果例 2 中的"谱"是*离散的*, 则在例 3 中它是*绝对连续的并有常数的谱密度* $f(\lambda)\equiv1/2\pi$. 实际上可以说, 序列 $\varepsilon=(\varepsilon_n)$ "由具有同一强度的振动构成". 正是因为这个原因, 与由同一强度的不同色调形成的物理"白噪声"类似, 才把 $\varepsilon=(\varepsilon_n)$ 称做"白噪声".

例 4 *移动平均序列* 由例 3 引进的白噪声 $\varepsilon=(\varepsilon_n)$ 出发, 组成新的序列

$$\xi_n=\sum_{k=-\infty}^{\infty}a_k\varepsilon_{n-k}, \tag{21}$$

其中 a_k 是复数, 满足 $\sum\limits_{k=-\infty}^{\infty}|a_k|^2<\infty$.

由 (21) 式, 可见

$$\operatorname{cov}(\xi_{n+m},\xi_m)=\operatorname{cov}(\xi_n,\xi_0)=\sum_{k=-\infty}^{\infty}a_{n+k}\bar{a}_k,$$

因此 $\xi=(\xi_k)$ 是平稳序列, 习惯上称做由序列 $\varepsilon=(\varepsilon_k)$ 形成的 (双侧) 移动平均序列.

特别, 当所有带负下标的 a_k 都等于 0 时, 即

$$\xi_n=\sum_{k=0}^{\infty}a_k\varepsilon_{n-k},$$

$\xi=(\xi_n)$ 称做单侧移动平均序列. 与此同时, 若对于一切 $k>p, a_k=0$, 即如果

$$\xi_n=a_0\varepsilon_n+a_1\varepsilon_{n-1}+\cdots+a_p\varepsilon_{n-p}, \tag{22}$$

则 $\xi=(\xi_n)$ 称做 p 阶移动平均序列.

可以证明 (练习题 3), 对于序列 (22), 其协方差函数 $R(n)$ 为:

$$R(n)=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}f(\lambda)\mathrm{d}\lambda,$$

其中谱密度 $f(\lambda)$ 为

$$f(\lambda)=\frac{1}{2\pi}|P(\mathrm{e}^{-i\lambda})|^2, \tag{23}$$

且

$$P(z)=a_0+a_1z+\cdots+a_pz^p.$$

例 5 自回归概型 设 $\varepsilon=(\varepsilon_n)$ 是白噪声. 称随机变量序列 $\xi=(\xi_n)$ 属于 q 阶自回归概型, 如果对于 $n\in\mathbb{Z}$, 有

$$\xi_n+b_1\xi_{n-1}+\cdots+b_q\xi_{n-q}=\varepsilon_n. \tag{24}$$

问当系数 $b_1,\cdots,b_q$ 满足何条件时, 方程 (24) 有平稳解? 为回答所提问题, 首先考虑 $q=1$ 的情形:

$$\xi_n=\alpha\xi_{n-1}+\varepsilon_n, \tag{25}$$

其中 $\alpha=-b_1$. 如果 $|\alpha|<1$, 则不难验证, 平稳序列 $\widetilde{\xi}=(\widetilde{\xi}_n)$:

$$\widetilde{\xi}_n=\sum_{j=0}^{\infty}\alpha^j\varepsilon_{n-j} \tag{26}$$

是方程 (25) 的解. (式 (26) 右侧的级数在均方意义下收敛.) 现在证明, 在具有有限二阶矩的平稳随机变量序列类中, $\widetilde{\xi}_n$ 是唯一解. 事实上, 由 (25) 式依次叠代, 得

$$\xi_n=\alpha\xi_{n-1}+\varepsilon_n=\alpha[\alpha\xi_{n-2}+\varepsilon_{n-1}]+\varepsilon_n=\cdots=\alpha^k\xi_{n-k}+\sum_{j=0}^{k-1}\alpha^j\varepsilon_{n-j}.$$

当 $k\to\infty$ 时, 由此得

$$\mathbf{E}\left[\xi_n-\sum_{j=0}^{k-1}\alpha^j\varepsilon_{n-j}\right]^2=\mathbf{E}[\alpha^k\xi_{n-k}]^2=\alpha^{2k}\mathbf{E}\xi_{n-k}^2=\alpha^{2k}\mathbf{E}\xi_0^2\to 0.$$

从而, 当 $|\alpha|<1$ 时方程 (25) 的平稳解存在, 且表示为 (26) 式的单侧移动平均的形式.

对于任意 $q>1$ 的情形, 有类似的结果: 假如多项式

$$Q(z)=1+b_1z+\cdots+b_qz^q \tag{27}$$

的一切 0 点都位于单位圆之外, 则自回归方程有解, 而且有可表示为单侧移动平均的形式 (24) 的唯一平稳解 (练习题 2). 这时, 协方差函数 $R(n)$ 可以表示为 (练习题 3):

$$R(n)=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}\mathrm{d}F(\lambda),\quad F(\lambda)=\int_{-\pi}^{\lambda}f(v)\mathrm{d}v, \tag{28}$$

其中

$$f(\lambda)=\frac{1}{2\pi}\times\frac{1}{|Q(\mathrm{e}^{-i\lambda})|^2}. \tag{29}$$

特别, 对于 $q=1$ 的情形, 由 (25) 式容易得到: $\mathbf{E}\xi_0=0$,

$$\mathbf{E}|\xi_0|^2=\frac{1}{1-|\alpha|^2},\quad R(n)=\frac{\alpha^n}{1-|\alpha|^2},\quad n\geqslant 0$$

(对于 $n<0, R(n)=\overline{R(-n)}$). 这时

$$f(\lambda)=\frac{1}{2\pi}\times\frac{1}{|1-\alpha\mathrm{e}^{-i\lambda}|^2}.$$

例 6 这个例子表明, 在水文学中建立概率模型时产生自回归概型. 考虑某水域 (例如, 里海), 并设法建立描绘, 由于径流和水表面的蒸发产生的振动, 引起的水平面对平均水平之偏差的概率模型.

假如以年做计量单位, 以 H_n 表示第 n 年水域的 "水平", 则可得如下平衡方程:

$$H_{n+1}=H_n-KS(H_n)+\Sigma_{n+1}, \tag{30}$$

其中以 Σ_{n+1} 表示第 $(n+1)$ 年的径流量, 以 $S(H)$ 表示在水平为 H 时水域的表面积, 而 K 是蒸发系数.

以 $\xi_n=H_n-\overline{H}$ 表示对平均水平 $\overline{H}$ 的偏差 ($\overline{H}$ 是根据多年的观测结果估计得来的), 假设 $S(H)=S(\overline{H})+c(H-\overline{H})$. 由平衡方程 (30), 可见随机变量满足方程:

$$\xi_{n+1}=\alpha\xi_n+\varepsilon_{n+1}, \tag{31}$$

其中 $\alpha = 1 - cK, \varepsilon_n = \Sigma_n - KS(\overline{H})$. 自然应认为随机变量 ε_n 的均值为 0, 并且作为初始逼近, 认为 $\{\varepsilon_n\}$是不相关和同分布的. 那么, 如同例 5 中曾证明的那样, (当 $|\alpha| < 1$ 时) 方程 (31) 有唯一平稳解, 应该将其视为所观测池中 (历年) 形成的水平面振动的状态.

作为可以由 (理论) 模型 (31) 得到的实际推论, 我们指出根据今年和往年的观测结果, 预测来年水面的水平的可能性. 结果恰好表明 (见下面 §6 的例 2), 由 $\cdots, \xi_{n-1}, \xi_n$ 的值对量 ξ_{n+1} 的最优均方线性估计, 就是 $\alpha\xi_n$.

例 7 *自回归和移动平均的混合模型* 如果在方程 (24) 的右侧将 ε_n 换成 $a_0\varepsilon_n + a_1\varepsilon_{n-1} + \cdots + a_p\varepsilon_{n-p}$, 则得所谓 (p,q) 阶自回归和移动平均的混合模型:

$$\xi_n + b_1\xi_{n-1} + \cdots + b_q\xi_{n-q} = a_0\varepsilon_n + a_1\varepsilon_{n-1} + \cdots + a_p\varepsilon_{n-p}. \tag{32}$$

在与例 5 相同的假设条件下, 关于多项式 $Q(z)$ 的 0 点, 下面将要证明 (§3 中定理 3 的系 2), 方程 (32) 有平稳解 $\xi = (\xi_n)$, 其协方差函数为:

$$R(n) = \int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n}\mathrm{d}F(\lambda), \quad F(\lambda) = \int_{-\pi}^{\lambda} f(v)\mathrm{d}v,$$

其中

$$f(\lambda) = \frac{1}{2\pi} \times \left|\frac{P(\mathrm{e}^{-i\lambda})}{Q(\mathrm{e}^{-i\lambda})}\right|^2.$$

3. 赫尔格洛茨定理

定理 (赫尔格洛茨 [G. Herglotz]) *设 $R(n)$ 是均值为 0 的 (弱) 平稳随机序列的协方差函数, 则在 $([-\pi,\pi), \mathscr{B}([-\pi,\pi))$ 上存在一有限测度 $F = F(B), B \in \mathscr{B}([-\pi,\pi))$, 使对于任何 $n \in \mathbb{Z}$, 有*

$$R(n) = \int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n}F(\mathrm{d}\lambda), \tag{33}$$

其中的积分 $\int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n}F(\mathrm{d}\lambda)$ 是集合 $[-\pi,\pi)$ 上的勒贝格 – 斯蒂尔切斯积分.

证明 对于 $N \geqslant 1, \lambda \in [-\pi,\pi]$, 设

$$f_N(\lambda) = \frac{1}{2\pi N}\sum_{k=1}^{N}\sum_{l=1}^{N} R(k-l)\mathrm{e}^{-ik\lambda}\mathrm{e}^{-il\lambda}. \tag{34}$$

由于 $R(n)$ 的非负定性, 可见 $f_N(\lambda)$ 是非负函数. 由于对于 $k - l = m$, 数偶 (k,l) 的个数等于 $N - |m|$, 则

$$f_N(\lambda) = \frac{1}{2\pi}\sum_{|m|<N}\left(1 - \frac{|m|}{N}\right)R(m)\mathrm{e}^{-im\lambda}. \tag{35}$$

设

$$F_N(B) = \int_B f_N(\lambda)\mathrm{d}\lambda, \quad B \in \mathscr{B}([-\pi,\pi)),$$

则

$$\int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n}F_N(\mathrm{d}\lambda) = \int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n}f_N(\lambda)\mathrm{d}\lambda = \begin{cases} \left(1 - \dfrac{|n|}{N}\right)R(n), & \text{若 } |n| < N, \\ 0, & \text{若 } |n| \geqslant N. \end{cases} \tag{36}$$

测度 $F_N, N \geqslant 1$, 集中在区间 $[-\pi,\pi]$上, 且对于任何 $N \geqslant 1, F_N[-\pi,\pi] = R(0) < \infty$. 从而, 测度 $F_N, N \geqslant 1$, 的族完备. 可见根据**普罗霍罗夫** (Ю. В. Прохоров) 定理 (第三章 §2 定理 1), 存在数列 $\{N_k\} \subseteq \{N\}$ 和测度 F, 使 $F_{N_k} \xrightarrow{w} F$. (密度、相对列紧性、弱收敛等概念, 以及普罗霍罗夫定理, 显然可以由概率测度移植到任何有限测度).

那么, 由 (36) 式可见

$$\int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n}F(\mathrm{d}\lambda) = \lim_{N_k\to\infty}\int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n}F_{N_k}(\mathrm{d}\lambda) = R(n).$$

所建立的测度 F 集中在区间 $[-\pi,\pi]$ 上. 在不改变积分 $\int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n}F(\mathrm{d}\lambda)$ 的情况下, 把集中在点 π 上的 "质量" $F(\{\pi\})$ 移到点 $-\pi$上, 可以**重新定义**测度 F. 这样得到的新测度 (仍记为 F) 就已经集中在区间 $[-\pi,\pi)$上. (关于恰好是在区间 $[-\pi,\pi)$ 上, 而不是在区间 $[-\pi,\pi]$ 上. 选择 λ 值的合理性, 见第 2 小节的例 1.) □

注 1 表达式 (33) 中的测度 $F = F(B)$, 称做**协方差函数**为 $R(n)$ 的平稳序列的**谱测度**, 而 $F(\lambda) = F([-\pi,\lambda])$ 称做**谱函数**.

在上面的例 2 中, 谱测度是离散的 (集中在点 $\lambda_k, k = 0, \pm1, \cdots$). 在例 3 ~ 6 中谱测度是绝对连续的.

注 2 谱测度 F 由协方差函数**唯一**决定. 事实上, 设 F_1 和 F_2 是两个谱测度, 且

$$\int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n}F_1(\mathrm{d}\lambda) = \int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n}F_2(\mathrm{d}\lambda), \quad n \in \mathbb{Z}.$$

由于任何有界连续函数 $g(\lambda)$, 可以在区间 $[-\pi,\pi)$ 上由三角多项式逼近, 则

$$\int_{-\pi}^{\pi} g(\lambda)F_1(\mathrm{d}\lambda) = \int_{-\pi}^{\pi} g(\lambda)F_2(\mathrm{d}\lambda),$$

由此可见 (对照第二章 §12 定理 2 的证明), 对于任何 $B \in \mathscr{B}([-\pi,\pi)), F_1(B) = F_2(B)$.

注 3 如果 $\xi = (\xi_n)$ 是由**实**随机变量 ξ_n 构成的平稳序列, 则 $R(n) = R(-n)$, 因而

$$R(n) = \frac{R(n) + R(-n)}{2} = \int_{-\pi}^{\pi} \cos\lambda n F(\mathrm{d}\lambda).$$

4. 练习题

1. 由 (11) 式导出 (12) 式.

2. 证明, 如果 (27) 式的多项式 $Q(z)$ 的一切 0 点都位于单位圆之外, 则自回归方程 (24) 有且唯一表示为单侧移动平均的平稳解.

3. 证明序列 (22) 和 (24) 的谱函数有密度, 且分别表示为公式 (23) 和 (29).

4. 证明, 如果 $\sum_{n=-\infty}^{\infty}|R(n)|^2<\infty$, 则谱函数 $F(\lambda)$ 有密度 $f(\lambda)$, 且由如下公式:

$$f(\lambda)=\frac{1}{2\pi}\sum_{n=-\infty}^{\infty}\mathrm{e}^{-in\lambda}R(n)$$

表示, 其中级数在复空间 $L^2=L^2([-\pi,\pi),\mathscr{B}([-\pi,\pi),\lambda))$中收敛, 而 λ 是勒贝格测度.

§2. 正交随机测度和随机积分

1. 扩充积分概念的必要性 在 §1 中已经讨论了协方差函数的积分表示, 并举出了具有两两正交随机变量 $z_k, k\in\mathbb{Z}$ 的平稳序列

$$\xi_n=\sum_{k=-\infty}^{\infty}z_k\mathrm{e}^{i\lambda_k n},\tag{1}$$

的例. 因此, 使我们产生一种想法: 是否可以将*任意*平稳序列表示为, 和式 (1) 之相应积分推广的形式.

假如设

$$Z(\lambda)=\sum_{\{k:z_k\leqslant\lambda\}}z_k,\tag{2}$$

则 (1) 式可以写为

$$\xi_n=\sum_{k=-\infty}^{\infty}\mathrm{e}^{i\lambda_k n}\Delta Z(\lambda_k),\tag{3}$$

其中 $\Delta Z(\lambda_k)\equiv Z(\lambda_k)-Z(\lambda_k-)=z_k$.

式 (3) 的右侧是 "勒贝格 — 斯蒂尔切斯积分"

$$\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}\mathrm{d}Z(\lambda)$$

的积分和. 然而, 对于现在所考虑的情形, 函数 $Z(\lambda)$ 也是 (依赖于 ω 的) 随机变量. 在这种情形下, 为说清楚任意平稳序列的积分表示, 不得不也考虑这样的函数 $Z(\lambda)$: 对于每一个 ω, $Z(\lambda)$ 有*无界变差*. 因此对于每一个 ω, 简单地把上述积分理解为黎曼 — 斯蒂尔切斯积分是不可行的.

2. 随机测度 与勒贝格积分, 勒贝格 — 斯蒂尔切斯积分以及黎曼 — 斯蒂尔切斯积分的一般概念类似 (第二章 §6) 从定义*随机测度*开始讨论我们所感兴趣的情形.

设 $(\Omega,\mathscr{F},\mathbf{P})$ 是概率空间, E 是某一集合, $\mathscr{E}_0$ 是集合 E 子集的代数, 而 $\mathscr{E}=\sigma(\mathscr{E}_0)$ 是 $\sigma-$ 代数.

定义 1 对于 $\omega\in\Omega$ 和 $\Delta\in\mathscr{E}_0$, 定义的复数值函数 $Z(\Delta)=Z(\omega,\Delta)$, 称做*有限 - 可加随机测度*, 如果

1) 对于任意 $\Delta\in\mathscr{E}_0, \mathbf{E}|Z(\Delta)|^2<\infty$;

2) 对于 $\mathscr{E}_0$ 中任意两个不相交集合 Δ_1 和 Δ_2.

$$Z(\Delta_1+\Delta_2)=Z(\Delta_1)+Z(\Delta_2)\quad(\mathbf{P}-\text{a.c.}).\tag{4}$$

定义 2 有限 — 可加随机测度 $Z(\Delta)$, 称做*初等随机测度*, 如果对于 $\mathscr{E}_0$ 中任意两两不相交集合 $\Delta_1,\Delta_2,\cdots$, 其中 $\Delta=\Delta_1+\Delta_2+\cdots\in\mathscr{E}_0$, 有

$$\mathbf{E}\left|Z(\Delta)-\sum_{k=1}^{n}Z(\Delta_k)\right|^2\to0,\quad n\to\infty.\tag{5}$$

注 1 在 $\mathscr{E}_0$ 的集合上的初等随机测度的定义中, 假设随机测度的值属于希尔伯特空间 $H^2=H^2(\Omega,\mathscr{F},\mathbf{P})$, 而该测度在 (5) 式的均方意义上是可数可加的. 尚存在随机测度的其他定义, 在这些定义中不要求存在二阶矩, 而且也不要求测度在 (5) 式的均方意义上可数 - 可加性, 例如要求在依概率收敛或以概率 1 收敛的意义上可数 — 可加.

注 2 与非随机测度的情形类似, 可以证明: 对于有限 — 可加随机测度, "在 (5) 式的均方意义上的可数可加性", 等价于 "在均方意义上在 '0' 的连续性":

$$\mathbf{E}|Z(\Delta_n)|^2\to0,\quad\Delta_n\downarrow\varnothing,\quad\Delta_n\in\mathscr{E}_0.\tag{6}$$

在初等随机测度类中, 按下面定义的正交测度特别重要.

定义 3 初等随机测度 $Z(\Delta),\Delta\in\mathscr{E}_0$, 称做*正交测度* (*或具有正交值的测度*), 如果对于 $\mathscr{E}_0$ 中任意两个不相交集合 Δ_1 和Δ_2, 有

$$\mathbf{E}Z(\Delta_1)\overline{Z(\Delta_2)}=0,\tag{7}$$

或者等价地, 如果对于 $\mathscr{E}_0$ 中任意两个集合 Δ_1 和 Δ_2, 有

$$\mathbf{E}Z(\Delta_1)\overline{Z(\Delta_2)}=\mathbf{E}|Z(\Delta_1\cap\Delta_2)|^2.\tag{8}$$

记

$$m(\Delta)=\mathbf{E}|Z(\Delta)|^2,\quad\Delta\in\mathscr{E}_0.\tag{9}$$

易见, 对于初等正交随机测度, 集函数 $m=m(\Delta),\Delta\in\mathscr{E}_0$, 是有限测度, 从而根据卡拉泰奥多里 (C. Carathéodory) 定理 (第二章 §3), $m(\Delta)$ 可以延拓到 $(E,\mathscr{E})$ 上. 这样,

得到的测度仍然记作 $m = m(\Delta)$, 并称之为 (初等正交随机测度 $Z = Z(\Delta), \Delta \in \mathscr{E}_0$, 的) *构造函数*.

现在自然产生如下问题: 既然定义在 $(E, \mathscr{E}_0)$ 上的集函数 $m = m(\Delta)$ 可以延拓到 $(E, \mathscr{E})$ 上, 其中 $\mathscr{E} = \sigma(\mathscr{E}_0)$, 那么是否可以将初等正交*随机测度* $Z = Z(\Delta), \Delta \in \mathscr{E}_0$, 延拓到集合 $\Delta \in \mathscr{E}$ 上, 并且使 $m(\Delta) = \mathbf{E}|Z(\Delta)|^2, \Delta \in \mathscr{E}$.

对该问题的回答是肯定的, 这可以由下面的构造证明. 与此同时由这一构造还可以建立, 平稳序列的积分表现所需要的随机积分.

3. 随机积分 设 $Z = Z(\Delta), \Delta \in \mathscr{E}_0$ 是初等正交随机测度, 而 $m = m(\Delta), \Delta \in \mathscr{E}_0$, 是其构造函数. 对于每一个只有有限个不同 (复数) 值的函数

$$f(\lambda) = \sum f_k I_{\Delta_k}(\lambda), \quad \Delta_k \in \mathscr{E}_0, \tag{10}$$

定义随机变量

$$\mathscr{S}(f) = \sum f_k Z(\Delta_k).$$

设 $L^2 = L^2(E, \mathscr{E}, m)$ 是希尔伯特复数值函数的空间, 其数积为

$$\langle f, g \rangle = \int_E f(\lambda)\overline{g(\lambda)} m(\mathrm{d}\lambda),$$

且范数为 $\|f\| = \langle f, f \rangle^{1/2}$, 而 $H^2 = H^2(\Omega, \mathscr{F}, \mathbf{P})$ 是希尔伯特复随机变量的空间, 相应的数积为

$$(\xi, \eta) = \mathbf{E}\xi\overline{\eta},$$

范数为 $\|\xi\| = \langle \xi, \xi \rangle^{1/2}$.

那么, 显然对于任意两个形如 (10) 式的函数 f 和 g, 有

$$(\mathscr{S}(f), \mathscr{S}(g)) = \langle f, g \rangle,$$

且

$$\|\mathscr{S}(f)\|^2 = \|f\|^2 = \int_E |f(\lambda)|^2 m(\mathrm{d}\lambda).$$

现在设 $f \in L^2$, 而 $\{f_n\}$ 是形如 (10) 式的函数, 满足 $\|f - f_n\| \to 0, n \to \infty$ (见练习题 2). 那么当 $n, m \to \infty$ 时, 有

$$\|\mathscr{S}(f_n) - \mathscr{S}(f_m)\| = \|f_n - f_m\| \to 0.$$

从而, 序列 $\{\mathscr{S}(f_n)\}$ 在均方意义上是基本的, 而由第二章 §10 的定理 7, 可见存在随机变量 (记作 $\mathscr{S}(f)$), 满足 $\mathscr{S}(f) \in H^2$ 且当 $n \to \infty$ 时 $\|\mathscr{S}(f_n) - \mathscr{S}(f)\| \to 0$.

这样建立的随机变量 $\mathscr{S}(f)$ (精确到随机等价性) 唯一, 而且与逼近序列 $\{f_n\}$ 的选择无关. 自然把 $\mathscr{S}(f)$ 称做函数 $f \in L^2$ 为关于初等正交随机测度 Z 的*随机积分*, 并且为了直观 (与 $\mathscr{S}(f)$ 同时) 使用 "积分" 记号:

$$\int_E f(\lambda) Z(\mathrm{d}\lambda).$$

现在指出随机积分 $\mathscr{S}(f)$ 的下列性质, 这些性质可以直接由其构造得出. 设函数 $g, f, f_n \in L^2$, 那么

$$(\mathscr{S}(f), \mathscr{S}(g)) = \langle f, g \rangle, \tag{11}$$

$$\|\mathscr{S}(f)\| = \|f\|, \tag{12}$$

$$\mathscr{S}(af + bg) = a\mathscr{S}(f) + b\mathscr{S}(g) \quad (\mathbf{P}-\text{a.c.}), \tag{13}$$

其中 a 和 b 是常数; 若当 $n \to \infty$ 时, $\|f_n - f\| \to 0$, 则

$$\|\mathscr{S}(f_n) - \mathscr{S}(f)\| \to 0. \tag{14}$$

4. 随机测度的延拓 将上面定义的随机积分, 用于初等正交测度 $Z(\Delta), \Delta \in \mathscr{E}_0$, 到 $\mathscr{E} = \sigma(\mathscr{E}_0)$ 中集合的*延拓*.

由于假设测度 m 是有限的, 则对于任意 $\Delta \in \mathscr{E}_0$, 函数 $I_\Delta = I_\Delta(\lambda) \in L^2$. 记 $\widetilde{Z}(\Delta) = \mathscr{S}(I_\Delta)$. 显然, 对于 $\Delta \in \mathscr{E}_0$, 有 $\widetilde{Z}(\Delta) = Z(\Delta)$. 由 (13) 式可见, 如果 $\Delta_1 \bigcap \Delta_2 = \varnothing, \Delta_1, \Delta_2 \in \mathscr{E}$, 则

$$\widetilde{Z}(\Delta_1 + \Delta_2) = \widetilde{Z}(\Delta_1) + \widetilde{Z}(\Delta_2) \quad (\mathbf{P}-\text{a.c.}),$$

而由 (12) 式, 得

$$\mathbf{E}|\widetilde{Z}(\Delta)|^2 = m(\Delta), \quad \Delta \in \mathscr{E}.$$

现在证明, 集合 $\widetilde{Z}(\Delta), \Delta \in \mathscr{E}$, 的随机函数, 是在均方意义上可数可加的. 事实上, 设 $\Delta_k \in \mathscr{E}, \Delta = \sum_{k=1}^{\infty} \Delta_k$, 则

$$\widetilde{Z}(\Delta) - \sum_{k=1}^{n} \widetilde{Z}(\Delta_k) = \mathscr{S}(g_n),$$

其中

$$g_n(\lambda) = I_\Delta(\lambda) - \sum_{k=1}^{n} I_{\Delta_k}(\lambda) = I_\Sigma(\lambda), \quad \Sigma = \sum_{k=n+1}^{\infty} \Delta_k.$$

但是

$$\mathbf{E}|\mathscr{S}(g_n)|^2 = \|g_n\|^2 = m\left(\sum_{k=n+1}^{\infty} \Delta_k\right) \downarrow 0, \quad n \to \infty,$$

即

$$\mathbf{E}\left|\widetilde{Z}(\Delta) - \sum_{k=1}^{n} \widetilde{Z}(\Delta_k)\right|^2 \to 0, \quad n \to \infty.$$

由 (11) 式可见, 对于 $\Delta_1 \bigcap \Delta_2 = \varnothing, \Delta_1, \Delta_2 \in \mathscr{E}$, 有

$$\mathbf{E}\widetilde{Z}(\Delta_1)\overline{\widetilde{Z}(\Delta_2)} = 0.$$

这样所建立定义在集合 $\Delta\in\mathscr{E}$ 上的随机函数 $\widetilde{Z}(\Delta)$, 是在均方意义上可数可加的, 并且在集合 $\Delta\in\mathscr{E}_0$ 上等于 $Z(\Delta)$. 我们称 $\widetilde{Z}(\Delta),\Delta\in\mathscr{E}$ (作为初等正交随机测度$Z(\Delta)$ 的延拓) 是以 $m(\Delta),\Delta\in\mathscr{E}$, 为构造函数的*正交随机测度*, 而上面定义的积分 $\int_E f(\lambda)Z(\mathrm{d}\lambda)$ 称做关于该测度的*随机积分*.

5. 正交增量随机过程 现在考虑对我们的目的最重要的情形, 即 $(E,\mathscr{E})=(\mathbb{R},\mathscr{B}(\mathbb{R}))$ 的情形. 由第二章 §3 定理 1 知, 空间 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ 上的任何有限测度 $m=m(\Delta)$, 都与某一 (广义) 分布函数 $G=G(x)$ 相互一一对应, 并且 $m(a,b]=G(b)-G(a)$.

结果表明, 对于正交测度有同样的情形. 下面引进正交增量随机过程的概念.

定义 4 概率空间 $(\Omega,\mathscr{F},\mathbf{P})$ 上 (复数值) 随机变量 $\{Z_\lambda\},\lambda\in\mathbb{R}$, 的全体, 称做*正交增量随机过程*, 如果

1) $\mathbf{E}|Z_\lambda|^2<\infty,\lambda\in\mathbb{R}$;

2) 对于每一个 $\lambda\in\mathbb{R}$,

$$\mathbf{E}|Z_\lambda-Z_{\lambda_n}|^2\to 0,\quad \lambda_n\downarrow\lambda,\quad \lambda_n\in\mathbb{R};$$

3) 对于任意 $\lambda_1<\lambda_2<\lambda_3<\lambda_4$,

$$\mathbf{E}(Z_{\lambda_4}-Z_{\lambda_3})\overline{(Z_{\lambda_2}-Z_{\lambda_1})}=0.$$

条件 3) 是增量的*正交性条件*. 条件 1) 表示 $Z_\lambda\in H^2$. 最后, 条件 2) 具有技术的特点, 表示要求在每一点 $\lambda\in\mathbb{R}$ (在均方意义上) 上是*右连续的*.

设 $Z=Z(\lambda)$ 是以 $m=m(\Delta)$ 为构造函数的正交随机测度, 是 (广义) 分布函数为 $G=G(\lambda)$ 的有限测度. 记

$$Z_\lambda=Z(-\infty,\lambda].$$

那么, $\mathbf{E}|Z_\lambda|^2=m(-\infty,\lambda]=G(\lambda)<\infty,\mathbf{E}|Z_\lambda-Z_{\lambda_n}|^2=m(\lambda_n,\lambda]\downarrow 0,\lambda_n\downarrow\lambda$, 并且条件 (3) 显然也成立. 因此, 所建立的过程 $Z(\lambda)$ 是*正交增量过程*.

另一方面, 如果 $G=G(\lambda)$ 是广义分布函数, 且 $G(-\infty)=0,G(+\infty)<\infty$; 而 $\{Z_\lambda\}$ 是正交增量过程, 且 $\mathbf{E}|Z_\lambda|^2=G(\lambda)$, 则对于 $\Delta\in(a,b]$, 设

$$Z(\Delta)=Z_b-Z_a.$$

设 $\mathscr{E}_0$ 是形如 $\Delta=\sum_{k=1}^n(a_k,b_k]$ 的集合产生的代数; 而 $Z(\Delta)=\sum_{k=1}^n Z(a_k,b_k]$. 显然

$$\mathbf{E}|Z(\Delta)|^2=m(\Delta),$$

其中 $m(\Delta)=\sum_{k=1}^n[G(b_k)-G(a_k)]$, 而对于不相交区间 $\Delta_1=(a_1,b_1]$ 和 $\Delta_2=(a_2,b_2]$, 有

$$\mathbf{E}Z(\Delta_1)\overline{Z(\Delta_2)}=0.$$

因为函数 $G(\lambda),\lambda\in\mathbb{R}$, 右连续, 故由此可见 $Z=Z(\Delta),\Delta\in\mathscr{E}_0$, 是具有正交值的初等随机测度. 集函数 $m=m(\Delta),\Delta\in\mathscr{E}_0$, 唯一地延拓到 $\mathscr{E}=\mathscr{B}(\mathbb{R})$ 上的测度, 而由前面的构造可见, 那么 $Z=Z(\Delta),\Delta\in\mathscr{E}_0$, 也可以延拓到 $\Delta\in\mathscr{E}=\mathscr{B}(\mathbb{R})$. 这时, 有

$$\mathbf{E}|Z(\Delta)|^2=m(\Delta),\quad \Delta\in\mathscr{B}(\mathbb{R}).$$

设 $\{Z_\lambda\},\lambda\in\mathbb{R}$, 是正交增量过程, $\mathbf{E}|Z_\lambda|^2=G(\lambda),G(-\infty)=0,G(+\infty)<\infty$; 而 $Z=Z(\Delta),\Delta\in\mathscr{B}(\mathbb{R})$, 是构造函数为 $m=m(\Delta)$ 的正交随机测度, 那么, 由以上的论述可见, 在 $\{Z_\lambda\}$ 与 $m=m(\Delta)$ 之间存在一一对应关系, 使

$$Z_\lambda=Z(-\infty,\lambda],\quad G(\lambda)=m(-\infty,\lambda]$$

和

$$Z(a,b]=Z_b-Z_a,\quad m(a,b]=G(b)-G(a).$$

根据在勒贝格 — 斯蒂尔切斯和黎曼 — 斯蒂尔切斯积分理论 (第二章 §6 第 9 和第 11 小节) 中通用的记号, 对于某一正交增量随机过程 $\{Z_\lambda\}$, 把随机积分 $\int_R f(\lambda)\mathrm{d}Z_\lambda$ 理解为与该过程 $\{Z_\lambda\}$ 相对应的、正交随机测度的随机积分 $\int_R f(\lambda)Z(\mathrm{d}\lambda)$.

6. 练习题

1. 证明条件 (5) 和 (6) 等价.

2. 设函数 $f\in L^2$. 利用第二章的结果 (第二章, §4 的定理 1, §6 定理 3 的系以及 §3 的练习题 8), 证明存在形如 (10) 式的函数序列 $\{f_n\}$, 使 $\|f_n-f\|\to 0,n\to\infty$.

3. 设 $Z(\Delta)$ 是以 $m(\Delta)$ 为构造函数的正交随机测度, 证明下列性质:

$$\mathbf{E}|Z(\Delta_1)-Z(\Delta_2)|^2=m(\Delta_1\triangle\Delta_2),$$

$$Z(\Delta_1\setminus\Delta_2)=Z(\Delta_1)-Z(\Delta_1\bigcap\Delta_2)\quad(\mathbf{P}-\text{a.c.}),$$

$$Z(\Delta_1\triangle\Delta_2)=Z(\Delta_1)+Z(\Delta_2)-2Z(\Delta_1\bigcap\Delta_2)\quad(\mathbf{P}-\text{a.c.}).$$

§3. 弱 (广义) 平稳序列的谱表示

1. 平稳随机序列谱表示 设 $\xi=(\xi_n)$ 是平稳随机序列, $\mathbf{E}\xi_n=0,n\in\mathbb{Z}$, 则根据 §1 的定理存在 $([-\pi,\pi),\mathscr{B}([-\pi,\pi)))$ 上的有限测度, 使协方差函数 $R(n)=\mathrm{cov}(\xi_{k+n},\xi_k)$ 有如下表现:

$$R(n)=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}F(\mathrm{d}\lambda).\tag{1}$$

下面的结果给出了序列 $\xi=(\xi_n),n\in\mathbb{Z}$, 的相应谱表示.

定理 1 *存在正交随机测度 $Z=Z(\Delta),\Delta\in\mathscr{B}([-\pi,\pi))$, 使对于每一个 $n\in\mathbb{Z}$, 以概率 1, 有*

$$\xi_n=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}Z(\mathrm{d}\lambda)\left(=\int_{[-\pi,\pi)}\mathrm{e}^{i\lambda n}Z(\mathrm{d}\lambda)\right).\tag{2}$$

这时 $\mathbf{E}Z(\Delta)=0, \mathbf{E}|Z(\Delta)|^2=F(\Delta)$.

证明 利用希尔伯特空间的有关结果, 进行证明最为简单.

设 $L^2(F)=L^2(E,\mathscr{E},F)$ 是复值函数的希尔伯特空间, $E=[-\pi,\pi), \mathscr{E}=\mathscr{B}([-\pi,\pi))$, 其中定义了内积

$$\langle f,g\rangle=\int_{-\pi}^{\pi} f(\lambda)\overline{g(\lambda)}F(\mathrm{d}\lambda), \tag{3}$$

而 $L_0^2(F)(L_0^2(F)\subseteq L^2(F))$ 是由 $e_n=e(\lambda), n\in\mathbb{Z}$, 生成的线性流形, 其中 $e_n(\lambda)=\mathrm{e}^{i\lambda n}$.

易见, 由于 $E=[-\pi,\pi)$, 而测度 F 有限, 可见流形 $L_0^2(F)$ 的闭包与 $L^2(F)$ 相等 (练习题 1):

$$\overline{L_0^2(F)}=L^2(F).$$

其次, 设 $L_0^2(\xi)$ 是由随机变量 $\xi_n, n\in\mathbb{Z}$, 生成的线性流形, 而 $L^2(\xi)(=\overline{L_0^2(\xi)})$ 是 (关于测度 $\mathbf{P}$ 的) 闭包.

我们现在建立随机元 $L_0^2(F)$ 与 $L_0^2(\xi)$ 之间一一对应 "$\leftrightarrow$" 关系,设

$$e_n\leftrightarrow\xi_n,\quad n\in\mathbb{Z}, \tag{4}$$

并且对于任意随机元 (确切地说, 对于任意等价随机元类), 按线性关系补充广义为

$$\sum\alpha_n e_n\leftrightarrow\sum\alpha_n\xi_n \tag{5}$$

(这里假设只有有限个复数 α_n 不为 0).

注意, 对应关系 (5) 式在如下意义上是适定的: 对于测度 F 几乎所有 $\sum\alpha_n e_n=0$, 当且仅当, 对于测度 $\mathbf{P}$ 几乎必然 $\sum\alpha_n\xi_n=0$.

这样定义的对应关系 "$\leftrightarrow$", 称做*等距的*, 即保持内积不变的. 事实上, 由 (3) 式, 有

$$\begin{aligned}\langle e_n,e_m\rangle&=\int_{-\pi}^{\pi}e_n(\lambda)\overline{e_m(\lambda)}F(\mathrm{d}\lambda)=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda(n-m)}F(\mathrm{d}\lambda)\\&=R(n-m)=\mathbf{E}\xi_n\bar{\xi}_m=(\xi_n,\xi_m),\end{aligned}$$

而且类似地有

$$\left\langle\sum\alpha_n e_n,\sum\beta_n e_n\right\rangle=\left(\sum\alpha_n\xi_n,\sum\beta_n\xi_n\right). \tag{6}$$

现在设 $\eta\in L^2(\xi)$. 由于 $L^2(\xi)=\overline{L_0^2}(\xi)$, 则存在随机变量序列 (η_n), 使 $\eta_n\in L_0^2(\xi)$ 和 $\|\eta_n-\eta\|\to0, n\to\infty$. 从而 (η_n) 是基本随机变量序列, 说明函数 (f_n) 也是基本随机变量序列, 其中 $f_n\in L_0^2(F)$ 且 $f_n\leftrightarrow\eta_n$. 由于空间 $L^2(F)$ 是完备的, 因而存在函数 $f\in L^2(F)$, 使 $\|f_n-f\|\to0, n\to\infty$.

显然相反的结果也成立: 如果 $f\in L^2(F)$, 且 $\|f_n-f\|\to0, f_n\in L_0^2(F)$, 则存在这样的元素 $\eta\in L^2(\xi)$, 使 $\eta_n\in L_0^2(\xi)$ 和 $\|\eta_n-\eta\|\to0, n\to\infty$ 且 $\eta_n\leftrightarrow f_n$.

至此, (等距) 对应关系 "$\leftrightarrow$", 暂时仅在 $L_0^2(\xi)$ 和 $L_0^2(F)$ 元素之间得以证明. 根据连续性, 将此对应关系补充定义: 设 $\eta\leftrightarrow f$, 其中 f 和 η 是上面所考虑的元素. 不难验证, 所建立的对应关系是 (等价随机变量类与函数类之间的) 线性的和保持内积不变的一一对应关系.

考虑函数 $f(\lambda)=I_\Delta(\lambda)$, 其中 $\Delta\in\mathscr{B}([-\pi,\pi)), \lambda\in[-\pi,\pi)$, 并设 $Z(\Delta)\in L^2(\xi)$ 是满足 $I_\Delta(\lambda)\leftrightarrow Z(\Delta)$ 的元素. 显然 $\|I_\Delta(\lambda)\|^2=F(\Delta)$, 因而 $\mathbf{E}|Z(\Delta)|^2=F(\Delta)$. 由于 $\mathbf{E}\xi_n=0, n\in\mathbb{Z}$, 故对于每一个元素 $L_0^2(\xi)$ (从而, 对于 $L^2(\xi)$), 其数学期望等于 0. 特别 $\mathbf{E}Z(\Delta)=0$. 此外, 如果 $\Delta_1\bigcap\Delta_2=\varnothing$, 则 $\mathbf{E}Z(\Delta_1)\overline{Z(\Delta_2)}=0$ 且

$$\mathbf{E}\left|Z(\Delta)-\sum_{k=1}^{n}Z(\Delta_k)\right|^2\to0,\quad n\to\infty,$$

其中 $\Delta=\Delta_1+\Delta_2+\cdots$.

这样, 元素的全体 $Z=Z(\Delta), \Delta\in\mathscr{B}([-\pi,\pi))$ 本身是正交随机测度, 因此 (由于 §2) 由该测度可以定义随机积分

$$\mathscr{S}(f)=\int_{-\pi}^{\pi}f(\lambda)Z(\mathrm{d}\lambda),\quad f\in L^2(F).$$

设 $f\in L^2(F)$ 和 $\eta\leftrightarrow f$. 以 $\Phi(f)$ 表示元素 η (确切地说, 在每一等价随机变量类与函数类中, 选一个元素作为代表). 现在证明, 以概率 1, 有

$$\mathscr{S}(f)=\Phi(f). \tag{7}$$

事实上, 如果

$$f(\lambda)=\sum\alpha_k I_{\Delta_k}(\lambda) \tag{8}$$

是函数 $I_{\Delta_k}(\lambda), \Delta_k=(a_k,b_k]$, 的有限线性组合, 则根据随机积分的定义, 显然 $\mathscr{S}(f)=\sum\alpha_k Z(\Delta_k)$ 等于$\Phi(f)$. 因此对于函数 (8), (7) 式成立. 但是, 如果 $f\in L^2(F)$ 且 $\|f_n-f\|\to0$, 其中 f_n 是形如 (8) 式的函数, 则 $\|\Phi(f_n)-\Phi(f)\|\to0$, 而根据 §2 的 (14) 式 $\|\mathscr{S}(f_n)-\mathscr{S}(f)\|\to0$. 从而, 以概率 1 有 $\Phi(f)=\mathscr{S}(f)$.

取函数 $f(\lambda)=\mathrm{e}^{i\lambda n}$, 则由 (4) 式可见 $\Phi(\mathrm{e}^{i\lambda n})=\xi_n$; 另一方面, $\mathscr{S}(f)=\int_{-\pi}^{\pi}f(\lambda)Z(\mathrm{d}\lambda)$. 于是, 由 (7) 式可见, 以概率 1, 有

$$\xi_n=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}Z(\mathrm{d}\lambda),\quad n\in\mathbb{Z}.$$

□

系 1 *设 $\xi=(\xi_n)$ 是平稳随机序列, 由实值随机变量 $\xi_n, n\in\mathbb{Z}$ 构成. 那么, 谱表达式 (2) 中的随机测度 $Z=Z(\Delta)$ 具有如下性质: 对于任何 $\Delta\in\mathscr{B}([-\pi,\pi))$,*

$$Z(\Delta)=\overline{Z(-\Delta)}, \tag{9}$$

其中集合 $-\Delta=\{\lambda:-\lambda\in\Delta\}$.

事实上, 设 $f(\lambda)=\sum\alpha_k e^{i\lambda k}$ 和 $\eta=\sum\alpha_k\xi_k$ (两个和式都是有限的). 那么, $f\leftrightarrow\eta$, 因此

$$\overline{\eta}=\sum\overline{\alpha}_k\xi_k\leftrightarrow\sum\overline{\alpha}_k e^{i\lambda k}=\overline{f(-\lambda)}. \tag{10}$$

由于 $I_\Delta(\lambda)\leftrightarrow Z(\Delta)$, 则由 (10) 式, 可见 $I_\Delta(-\lambda)\leftrightarrow\overline{Z(\Delta)}$ (或等价地 $I_{-\Delta}(\lambda)\leftrightarrow\overline{Z(\Delta)}$). 另一方面, 因为 $I_{-\Delta}(\lambda)\leftrightarrow Z(-\Delta)$, 所以 $\overline{Z(\Delta)}=Z(-\Delta)(\mathbf{P}-\text{a.c.})$.

系 2 设 $\xi=(\xi_n)$ 是平稳随机序列, 其中 ξ_n 是实值随机变量, 而 $Z(\Delta)=Z_1(\Delta)+iZ_2(\Delta)$. 那么, 对于任意 $\Delta_1,\Delta_2\in\mathscr{B}([-\pi,\pi))$, 有

$$\mathbf{E}Z_1(\Delta_1)Z_2(\Delta_2)=0, \tag{11}$$

并且, 如果 $\Delta_1\bigcap\Delta_2=\varnothing$ 和 $(-\Delta_1)\bigcap\Delta_2=\varnothing$, 则

$$\mathbf{E}Z_1(\Delta_1)Z_1(\Delta_2)=0,\quad \mathbf{E}Z_2(\Delta_1)Z_2(\Delta_2)=0. \tag{12}$$

事实上, 由于 $Z(\Delta)=\overline{Z(-\Delta)}$, 则

$$Z_1(-\Delta)=Z_1(\Delta),\quad Z_2(-\Delta)=-Z_2(\Delta). \tag{13}$$

其次, 因为 $\mathbf{E}Z(\Delta_1)\overline{Z(\Delta_2)}=\mathbf{E}|Z(\Delta_1\bigcap\Delta_2)|^2$,

$$\operatorname{Im}\mathbf{E}Z(\Delta_1)\overline{Z(\Delta_2)}=0,$$

即

$$\mathbf{E}Z_1(\Delta_1)Z_2(\Delta_2)-\mathbf{E}Z_2(\Delta_1)Z_1(\Delta_2)=0. \tag{14}$$

将 Δ_1 换成区间 $-\Delta_1$, 则得

$$\mathbf{E}Z_1(-\Delta_1)Z_2(\Delta_2)-\mathbf{E}Z_2(-\Delta_1)Z_1(\Delta_2)=0,$$

故由于 (13) 式, 由此得

$$\mathbf{E}Z_1(\Delta_1)Z_2(\Delta_2)+\mathbf{E}Z_2(\Delta_1)Z_1(\Delta_2)=0. \tag{15}$$

由 (14) 和 (15) 式得等式 (11).

如果 $\Delta_1\bigcap\Delta_2=\varnothing$ 和 $(-\Delta_1)\bigcap\Delta_2=\varnothing$, 则 $\mathbf{E}Z(\Delta_1)\overline{Z(\Delta_2)}=0$. 因此 $\operatorname{Re}\mathbf{E}Z(\Delta_1)\overline{Z(\Delta_2)}=0$ 和 $\operatorname{Re}\mathbf{E}Z(-\Delta_1)\overline{Z(\Delta_2)}=0$. 于是, 连同 (13) 式就证明了等式 (12).

系 3 设 $\xi=(\xi_n)$ 是高斯序列. 那么, 对于任意 $\Delta_1,\cdots,\Delta_k$, 向量 $(Z_1(\Delta_1),\cdots,Z_1(\Delta_k);Z_2(\Delta_1),\cdots,Z_2(\Delta_k))$ 服从高斯 (正态) 分布.

事实上, 由 (复数值) 高斯随机变量 η, 即由服从高斯分布的向量 $(\operatorname{Re}\eta,\operatorname{Im}\eta)$ 生成的线性流形 $L_0^2(\xi)$, 服从高斯分布. 那么, 根据第二章 §13 第 5 小节, 可见 $L_0^2(\xi)$ 的闭包 $\overline{L_0^2(\xi)}$ 也由高斯随机变量构成. 因此由系 2 知, 对于高斯序列 $\xi=(\xi_n)$,

实部 Z_1 和虚部 Z_2 在如下意义上独立: 任意随机变量组 $(Z_1(\Delta_1),\cdots,Z_1(\Delta_k))$ 和 $(Z_2(\Delta_1),\cdots,Z_2(\Delta_k))$ 相互独立. 由 (12) 知, 对于满足 $\Delta_i\bigcap\Delta_j=(-\Delta_i)\bigcap\Delta_j=\varnothing, i,j=1,\cdots,k, i\neq j$, 的集合 $\Delta_1,\cdots,\Delta_k$, 随机变量 $Z_i(\Delta_1),\cdots,Z_i(\Delta_k)(i=1,2)$ 全体独立.

系 4 设 $\xi=(\xi_n)$ 是平稳实值随机变量序列, 则以概率 1

$$\xi_n=\int_{-\pi}^{\pi}\cos\lambda n Z_1(\mathrm{d}\lambda)-\int_{-\pi}^{\pi}\sin\lambda n Z_2(\mathrm{d}\lambda). \tag{16}$$

注 如果 $\{Z_\lambda\},\lambda\in[-\pi,\pi)$ 是对应于正交随机测度 $Z=Z(\Delta)$ 的正交增量随机过程, 则谱表示 (2) (根据 §2) 可以表示为如下形式:

$$\xi_n=\int_{-\pi}^{\pi}e^{i\lambda n}\mathrm{d}Z_\lambda,\quad n\in\mathbb{Z}. \tag{17}$$

2. 由平稳序列经线性变换得到的随机变量的构造 设 $\xi=(\xi_n)$ 是具有谱分解 (2) 的平稳序列, 而 $\eta\in L^2(\xi)$. 下面的定理描绘这样随机变量 η 的构造.

定理 2 如果 $\eta\in L^2(\xi)$, 则存在函数 $\varphi\in L^2(F)$, 使以概率 1, 有

$$\eta=\int_{-\pi}^{\pi}\varphi(\lambda)Z(\mathrm{d}\lambda). \tag{18}$$

证明 如果

$$\eta_n=\sum_{|k|\leqslant n}\alpha_k\xi_k, \tag{19}$$

则由于 (2) 式, 有

$$\eta_n=\int_{-\pi}^{\pi}\left(\sum_{|k|\leqslant n}\alpha_k e^{i\lambda n}\right)Z(\mathrm{d}\lambda), \tag{20}$$

即对于函数

$$\varphi(\lambda)=\sum_{|k|\leqslant n}\alpha_k e^{i\lambda n}, \tag{21}$$

(18) 式成立. 在一般情形下, 如果 $\eta\in L^2(\xi)$, 则存在形如 (19) 式的随机变量 η_n, 使 $\|\eta-\eta_n\|\to0, n\to\infty$. 那么, $\|\varphi_n-\varphi_m\|=\|\eta_n-\eta_m\|\to0, n,m\to\infty$, 即序列 (φ_n) 在 $L^2(F)$ 中是基本序列. 因此, 存在函数 $\varphi\in L^2(F)$, 使 $\|\varphi-\varphi_n\|\to0, n\to\infty$.

根据 §2 的 (14) 式 $\|\mathscr{S}(\varphi_n)-\mathscr{S}(\varphi)\|\to0$, 而因为 $\eta_n=\mathscr{S}(\varphi_n)$, 所以以概率 1, 有 $\eta=\mathscr{S}(\varphi)$. □

注 设 $H_0(\xi)$ 和 $H_0(F)$ 是相应为变量 $\xi^0=(\xi_n)_{n\leqslant0}$ 和函数 $e^0=(e_n)_{n\leqslant0}$ 封闭线性流形. 那么, 如果 $\eta\in H_0(\xi)$, 则存在这样一个函数 $\varphi\in H_0(F)$, 使以概率 1, 有

$$\eta=\int_{-\pi}^{\pi}\varphi(\lambda)Z(\mathrm{d}\lambda).$$

3. 线性滤波器 公式 (18) 描绘经线性变换, 由 $\xi_n, n \in \mathbb{Z}$ 得到的随机变量, 即可以表示为形如 (19) 的随机变量及其均方极限的构造.

利用所谓 (线性) 滤波器表示的线性变换类, 是特殊然而重要的线性变换类. 假设在某系统 (滤波器) 入口于时刻 m 出现信号 x_m, 而这时系统对该信号的反映是: 在其出口于时刻 n 收到信号 $h(n-m)x_m$, 其中 $h(s), s \in \mathbb{Z}$, 是某一复数值函数, 称为 (滤波器的) *脉冲转移函数*.

这样, 在系统输出的和信号 y_n 可以表示为:

$$y_n = \sum_{m=-\infty}^{\infty} h(n-m)x_m. \tag{22}$$

对于*物理可实现的系统*, 输出信号的值只由 "过去" 的值决定, 即由 $x_m (m \leqslant n)$ 值决定. 如果对于一切 $s < 0$ 有 $h(s) = 0$, 或如果

$$y_n = \sum_{m=-\infty}^{\infty} h(n-m)x_m = \sum_{m=0}^{\infty} h(m)x_{n-m}, \tag{23}$$

则自然称脉冲转移函数为 $h = h(s)$ 的滤波器是*物理可实现的*.

称为滤波器的*频率特征*的、脉冲转移函数 h 的傅里叶变换

$$\varphi(\lambda) = \sum_{m=-\infty}^{\infty} \mathrm{e}^{-i\lambda m} h(m), \tag{24}$$

是以 h 为脉冲转移函数的滤波器的重要*谱特征*.

现在考虑关于 (22) 和 (24) 式中级数的收敛条件. 这些条件到目前为止尚未得到说明。假设在滤波器的入口发送具有协方差函数为 $R(n)$ 和谱分解 (2) 的、平稳随机序列 $h(s), s \in \mathbb{Z}$. 那么, 如果

$$\sum_{k,l=-\infty}^{\infty} h(k)R(l-k)\overline{h(l)} < \infty, \tag{25}$$

则级数 $\sum\limits_{m=-\infty}^{\infty} h(n-m)\xi_m$ 均方收敛, 因而平稳序列 $\eta = (\eta_n)$ 有定义, 其中

$$\eta_n = \sum_{m=-\infty}^{\infty} h(n-m)\xi_m = \sum_{m=-\infty}^{\infty} h(m)\xi_{n-m}. \tag{26}$$

按谱分析的术语, 条件 (25) 显然等价于 $\varphi(\lambda) \in L^2(F)$, 即

$$\int_{-\pi}^{\pi} |\varphi(\lambda)|^2 F(\mathrm{d}\lambda) < \infty. \tag{27}$$

在 (25) 式或 (27) 式的条件下, 由 (26) 式和 (2) 式, 可得序列 η 的谱表现

$$\eta_n = \int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n}\varphi(\lambda)Z(\mathrm{d}\lambda) < \infty, \quad n \in \mathbb{Z}. \tag{28}$$

从而, 序列 η 的协方差函数 $R_\eta(n)$ 由如下公式确定:

$$R_\eta(n) = \int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n}|\varphi(\lambda)|^2 F(\mathrm{d}\lambda) < \infty. \tag{29}$$

特别, 如果在频率特征为 $\varphi = \varphi(\lambda)$ 的滤波器入口传送白噪声 $\varepsilon = (\varepsilon_n)$, 则在其出口将收到平稳移动均值序列:

$$\eta_n = \sum_{m=-\infty}^{\infty} h(m)\varepsilon_{n-m}, \tag{30}$$

其谱密度为

$$f_\eta(\lambda) = \frac{1}{2\pi}|\varphi(\lambda)|^2.$$

下面的定理在一定意义上说明, 任何有谱密度的平稳序列可以由移动均值得到.

定理 3 *设 $\eta = (\eta_n)$ 是谱密度为 $f_\eta(\lambda)$ 的平稳序列, 则存在这样的白噪声序列 $\varepsilon = (\varepsilon_n)$ 和这样的滤波器 (必要时扩充原概率空间), 使表达式 (30) 成立.*

证明 由给定的 (非负) 函数 $f_\eta(\lambda)$, 存在这样的函数 $\varphi(\lambda)$, 使 $f_\eta(\lambda) = (2\pi)^{-1}|\varphi(\lambda)|^2$. 由于

$$\int_{-\pi}^{\pi} f_\eta(\lambda)\mathrm{d}\lambda < \infty, \quad \text{故} \quad \varphi(\lambda) \in L^2(\mathrm{d}\mu),$$

这里 $\mathrm{d}\mu$ 是 $[-\pi, \pi)$ 上的勒贝格测度. 因此, 函数 $\varphi(\lambda)$ 可以表示为傅里叶级数 (24), 其中

$$h(m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{im\lambda}\varphi(\lambda)\mathrm{d}\lambda,$$

并且收敛性应理解为

$$\int_{-\pi}^{\pi} \left|\varphi(\lambda) - \sum_{|m|\leqslant n} \mathrm{e}^{-im\lambda}h(m)\right|^2 \mathrm{d}\lambda \to 0, \quad n \to \infty.$$

设

$$\eta_n = \int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n} Z(\mathrm{d}\lambda) < \infty, \quad n \in \mathbb{Z}.$$

除了测度 $Z = Z(\Delta)$ 之外, 我们再引进不依赖于 $Z = Z(\Delta)$ 的、新的正交随机测度 $\widetilde{Z} = \widetilde{Z}(\Delta)$, 并满足 $\mathbf{E}|\widetilde{Z}(a,b]|^2 = (b-a)/(2\pi)$. (一般, 建立这样测度的可能性, 应以原概率空间充分 "丰富" 为前提.) 设

$$\overline{Z}(\Delta) = \int_\Delta \varphi^{\oplus}(\lambda)Z(\mathrm{d}\lambda) + \int_\Delta [1 - \varphi^{\oplus}(\lambda)\varphi(\lambda)]\widetilde{Z}(\mathrm{d}\lambda),$$

其中

$$a^{\oplus} = \begin{cases} a^{-1}, & \text{若 } a \neq 0, \\ 0, & \text{若 } a = 0. \end{cases}$$

随机测度 $\overline{Z}=\overline{Z}(\Delta)$ 是具有正交值的测度, 这时对于任何 $\Delta=(a,b]$,

$$\mathbf{E}|\overline{Z}(\Delta)|^2=\frac{1}{2\pi}\int_\Delta|\varphi^\oplus(\lambda)|^2|\varphi(\lambda)|^2\mathrm{d}\lambda+\frac{1}{2\pi}\int_\Delta|1-\varphi^\oplus(\lambda)\varphi(\lambda)|^2\mathrm{d}\lambda=\frac{|\Delta|}{2\pi},$$

其中 $|\Delta|=b-a$. 因此, 平稳序列 $\varepsilon=(\varepsilon_n),n\in\mathbb{Z}$, 是白噪声, 其中

$$\varepsilon_n=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}\overline{Z}(\mathrm{d}\lambda).$$

现在, 注意到,

$$\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}\varphi(\lambda)\overline{Z}(\mathrm{d}\lambda)=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}Z(\mathrm{d}\lambda)=\eta_n; \tag{31}$$

而另一方面, 根据 $\varphi(\lambda)$ 的定义和 §2 的性质 (14), 以概率 1, 有

$$\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}\varphi(\lambda)\overline{Z}(\mathrm{d}\lambda)=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}\left(\sum_{m=-\infty}^{\infty}\mathrm{e}^{-i\lambda m}h(m)\right)\overline{Z}(\mathrm{d}\lambda)$$
$$=\sum_{m=-\infty}^{\infty}h(m)\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda(n-m)}\overline{Z}(\mathrm{d}\lambda)=\sum_{m=-\infty}^{\infty}h(m)\varepsilon_{n-m},$$

于是, 注意到 (31) 式就证明了表达式 (30). □

注 如果 (按勒贝格测度几乎处处) $f_\eta(\lambda)>0$, 则引进辅助测度 $\widetilde{Z}=\widetilde{Z}(\Delta)$ 就没有必要 (因为这时勒贝格测度几乎处处 $1-\varphi^\oplus(\lambda)\varphi(\lambda)=0$), 而且 "原概率空间充分 '丰富' " 的前提条件也可以去掉.

系 1 假设谱密度 (按勒贝格测度几乎处处的) $f_\eta(\lambda)>0$, 并且

$$f_\eta(\lambda)=\frac{1}{2\pi}|\varphi(\lambda)|^2,$$

其中

$$\varphi(\lambda)=\sum_{k=0}^{\infty}\mathrm{e}^{-i\lambda k}h(k),\quad \sum_{k=0}^{\infty}|h(k)|^2<\infty,$$

则序列 η 可以表示为单侧移动平均的形式

$$\eta_n=\sum_{m=0}^{\infty}h(m)\varepsilon_{n-m}.$$

特别, 设 $P(z)=a_0+a_1z+\cdots+a_pz^p$, 则谱密度为

$$f_\eta(\lambda)=\frac{1}{2\pi}|P(\mathrm{e}^{-i\lambda})|^2$$

的序列 $\eta=(\eta_n)$ 可以表示为

$$\eta_n=a_0\varepsilon_n+a_1\varepsilon_{n-1}+\cdots+a_p\varepsilon_{n-p}.$$

系 2 假设 $\xi=(\xi_n)$ 是平稳序列, 具有有理谱密度

$$f_\xi(\lambda)=\frac{1}{2\pi}\left|\frac{P(\mathrm{e}^{-i\lambda})}{Q(\mathrm{e}^{-i\lambda})}\right|^2, \tag{32}$$

其中 $P(z)=a_0+a_1z+\cdots+a_pz^p, Q(z)=1+b_1z+\cdots+b_qz^q$.

如果多项式 $Q(z)$ 在集合 $\{z:|z|=1\}$ 上无 0 点, 则存在这样的白噪声 $\varepsilon=(\varepsilon_n)$, 使以概率 1, 有

$$\xi_n+b_1\xi_{n-1}+\cdots+b_q\xi_{n-q}=a_0\varepsilon_n+a_1\varepsilon_{n-1}+\cdots+a_p\varepsilon_{n-p}. \tag{33}$$

相反, 设 $\xi=(\xi_n)$ 是任意平稳序列, 并且满足这样带某白噪声 $\varepsilon=(\varepsilon_n)$ 的方程 (33), 以及在集合 $\{z:|z|=1\}$ 上无 0 点的多项式 $Q(z)$, 则 $\xi=(\xi_n)$ 有谱密度 (32) 式.

事实上, 设 $\eta_n=\xi_n+b_1\xi_{n-1}+\cdots+b_q\xi_{n-q}$. 那么,

$$f_\eta(\lambda)=\frac{1}{2\pi}|P(\mathrm{e}^{-i\lambda})|^2,$$

而由系 1 得所要求的表达式.

另一方面, 如果表达式 (33) 成立, 而 $F_\xi(\lambda)$ 和 $F_\eta(\lambda)$ 是序列 ξ 和 η 的谱函数, 则

$$F_\eta(\lambda)=\int_{-\pi}^{\lambda}|Q(\mathrm{e}^{-iv})|^2\mathrm{d}F_\xi(v)=\frac{1}{2\pi}\int_{-\pi}^{\lambda}|P(\mathrm{e}^{-iv})|^2\mathrm{d}v.$$

由于 $|Q(\mathrm{e}^{-iv})|^2>0$, 故由此得由 (32) 式确定的密度.

4. 遍历性定理 下面 (均方意义下的) 遍历性定理, 可以看做弱平稳随机序列的大数定律的类似.

定理 4 设 $\xi=(\xi_n),n\in\mathbb{Z}$, 是任意平稳序列, $\mathbf{E}\xi_n=0$, 且具有协方差函数 (1) 式和谱分解 (2) 式, 则

$$\frac{1}{n}\sum_{k=0}^{n-1}\xi_k\xrightarrow{L^2}Z(\{0\}) \tag{34}$$

和

$$\frac{1}{n}\sum_{k=0}^{n-1}R(k)\to F(\{0\}). \tag{35}$$

证明 由 (2) 式, 可见

$$\frac{1}{n}\sum_{k=0}^{n-1}\xi_k=\int_{-\pi}^{\pi}\frac{1}{n}\sum_{k=0}^{n-1}\mathrm{e}^{ik\lambda}Z(\mathrm{d}\lambda)=\int_{-\pi}^{\pi}\varphi_n(\lambda)Z(\mathrm{d}\lambda),$$

其中

$$\varphi_n(\lambda)=\frac{1}{n}\sum_{k=0}^{n-1}\mathrm{e}^{ik\lambda}=\begin{cases}1, & 若\ \lambda=0,\\ \dfrac{1}{n}\times\dfrac{\mathrm{e}^{in\lambda}-1}{\mathrm{e}^{i\lambda}-1}, & 若\ \lambda\neq0.\end{cases} \tag{36}$$

显然, $|\varphi_n(\lambda)| \leqslant 1$.

其次, 由于 $\varphi_n(\lambda) \xrightarrow{L^2(F)} I_{\{0\}}(\lambda)$, 因此由 §2 的性质 (14), 有

$$\int_{-\pi}^{\pi} \varphi_n(\lambda) Z(\mathrm{d}\lambda) \xrightarrow{L^2} \int_{-\pi}^{\pi} I_{\{0\}}(\lambda) Z(\mathrm{d}\lambda) = Z(\{0\}),$$

于是 (34) 式得证.

类似地可以证明结论 (35) 式. □

系 由于, 如果谱函数在 0 点连续, 即 $F(\{0\}) = 0$, 则 $Z(\{0\}) = 0(\mathbf{P} - \text{a.c.})$, 可见 (34) 式和 (35) 式, 有

$$\frac{1}{n}\sum_{k=0}^{n-1} R(k) \to 0 \Rightarrow \frac{1}{n}\sum_{k=0}^{n-1} \xi_k \xrightarrow{L^2} 0.$$

由于

$$\left|\frac{1}{n}\sum_{k=0}^{n-1} R(k)\right|^2 = \left|\mathbf{E}\left(\frac{1}{n}\sum_{k=0}^{n-1}\xi_k\right)\xi_0\right|^2 \leqslant \mathbf{E}|\xi_0|^2 \mathbf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1}\xi_k\right|^2,$$

可见相反得蕴涵关系也成立:

$$\frac{1}{n}\sum_{k=0}^{n-1} \xi_k \xrightarrow{L^2} 0 \Rightarrow \frac{1}{n}\sum_{k=0}^{n-1} R(k) \to 0.$$

这样, 算术平均值 $n^{-1}\sum\limits_{k=0}^{n-1}\xi_k$ (在均方意义下) 收敛于 0 的充分和必要条件是 $n^{-1}\sum\limits_{k=0}^{n-1} R(k) \to 0$. 由此可见, 如果原序列随机变量的数学期望为 $m(\mathbf{E}\xi_0 = m)$, 则

$$\frac{1}{n}\sum_{k=0}^{n-1} R(k) \to 0 \Leftrightarrow \frac{1}{n}\sum_{k=0}^{n-1} \xi_k \xrightarrow{L^2} 0, \tag{37}$$

其中 $R(n) = \mathbf{E}(\xi_n - \mathbf{E}\xi_n)\overline{(\xi_0 - \mathbf{E}\xi_0)}$.

我们还要指出, 如果以大于 0 的概率 $Z(\{0\}) \neq 0$, 则说明序列 ξ_n 包含 "随机常数 α":

$$\xi_n = \alpha + \eta_n,$$

其中 $\alpha = Z(\{0\})$, 而在谱表示

$$\eta_n = \int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n} Z_\eta(\mathrm{d}\lambda)$$

中测度 $Z_\eta = Z_\eta(\Delta)$ 已经满足 $Z_\eta(\{0\}) = 0(\mathbf{P} - \text{a.c.})$. 结论 (34) 式说明, 算术平均恰好均方收敛于这一随机常数 α.

5. 练习题

1. 证明 $\overline{L_0^2(F)} = L^2(F)$ (记号见定理 1 的证明).

2. 设 $\xi = (\xi_n)$是平稳序列, 具有性质: 对某个 N 和一切 n, 有 $\xi_{n+N} = \xi_n$. 证明这一序列的谱表示归结为 §1 的表示 (13) 式.

3. 设 $\xi = (\xi_n)$ 是平稳序列, 其中 $\mathbf{E}\xi_n = 0$, 且满足: 对于某 $C > 0, \alpha > 0$, 有

$$\frac{1}{N^2}\sum_{k=0}^{N-1}\sum_{l=0}^{N-1} R(k-l) = \frac{1}{N}\sum_{|k|\leqslant N-1} R(k)\left[1 - \frac{|k|}{N}\right] \leqslant \frac{C}{N^\alpha}.$$

利用博雷尔 – 康泰利引理证明

$$\frac{1}{N}\sum_{k=0}^{N}\xi_k \to 0 \quad (\mathbf{P} - \text{a.c.}).$$

4. 假设 $\xi = (\xi_m)$ 是随机变量序列, 具有有理谱密度

$$f_\xi(\lambda) = \frac{1}{2\pi}\frac{|P_{n-1}(\mathrm{e}^{-i\lambda})|}{|Q_n(\mathrm{e}^{-i\lambda})|}, \tag{38}$$

其中 $P_{n-1}(z) = a_0 + a_1 z + \cdots + a_{n-1}z^{n-1}, Q_n(z) = 1 + b_1 z + \cdots + b_n z^n$, 而且多项式 $Q_n(z)$ 的根不在单位圆上.

证明存在这样的白噪声 $\varepsilon = (\varepsilon_m), m \in \mathbb{Z}$, 使序列 (ξ_m) 是 n 维序列 $(\xi_m^1, \xi_m^2, \cdots, \xi_m^n), \xi_m^1 = \xi_m$ 的分量, 而 n 维序列 $(\xi_m^1, \xi_m^2, \cdots, \xi_m^n)$ 满足方程组:

$$\begin{aligned} \xi_{m+1}^i &= \xi_m^{i+1} + \beta_i \varepsilon_{m+1}, \quad i = 1, \cdots, n-1, \\ \xi_{m+1}^n &= -\sum_{j=0}^{n-1} b_{n-j}\xi_m^{j+1} + \beta_n \varepsilon_{m+1}, \end{aligned} \tag{39}$$

其中 $\beta_1 = a_0, \beta_i = a_{i-1} - \sum\limits_{k=1}^{i-1}\beta_k b_{i-k}$.

§4. 协方差函数和谱密度的统计估计

1. 协方差函数的估计及其性质 平稳随机序列概率分布的各种特征的统计估计问题, 出现在各种不同科学领域 (地球物理学, 医学, 经济学等等). 这一节讲述的内容, 建立估计的概念和方法以及这里可能出现的困难.

这样, 设 $\xi = (\xi_n), n \in \mathbb{Z}$, 是广义平稳随机序列 (为简便计, 假设是实数值序列), 且有数学期望 $\mathbf{E}\xi_n = m$, 而其协方差函数有如下表现:

$$R(n) = \int_{-\pi}^{\pi} \mathrm{e}^{i\lambda n} F(\mathrm{d}\lambda).$$

设在对随机变量 $\xi_0,\xi_1,\cdots,\xi_{N-1}$ 值观测过程中得到 (实现) $x_0,x_1\cdots,x_{N-1}$. 问如何根据观测结果建立 (未知) *均值的* "好" 估计?

设

$$m_N(x)=\frac{1}{N}\sum_{k=0}^{N-1}x_k. \tag{1}$$

那么, 由数学期望的初等性质, 可见该估计在如下意义上是数学期望 m 的 "好" 估计: "全部实现 $x_0,\cdots,x_{N-1}$ 的平均值" 是*无偏的*, 即

$$\mathbf{E}m_N(\xi)=\mathbf{E}\left(\frac{1}{N}\sum_{k=0}^{N-1}\xi_k\right)=m. \tag{2}$$

此外, 由 §3 定理 4 可见, 在

$$\frac{1}{N}\sum_{k=0}^{N}R(k)\to 0,\quad N\to\infty,$$

的条件下, 估计在均方意义上也是*相合的*, 即

$$\mathbf{E}|m_N(\xi)-m|^2\to 0,\quad N\to 0. \tag{3}$$

现在, 在 $m=0$ 条件下, 讨论*协方差函数* $R(n)$、*谱函数* $F(\lambda)=F([-\pi,\lambda))$ 和*谱密度* $f(\lambda)$ 的估计问题.

由于 $R(n)=\mathbf{E}\xi_{n+k}\xi_k$, 则作为对其根据 N 次观测结果 $x_0,x_1,\cdots,x_{N-1}$ 的估计, 自然假设对于 $0\leqslant n\leqslant N$,

$$\widehat{R}_N(n,x)=\frac{1}{N-n}\sum_{k=0}^{N-n-1}x_{n+k}x_k.$$

显然, 估计在如下意义上是无偏的:

$$\mathbf{E}\widehat{R}_N(n,\xi)=R(n),\quad 0\leqslant n<N.$$

现在讨论估计 $\widehat{R}_N(n,\xi)$ 的*相合性*问题. 将 §3 的 (37) 式中的 ξ_k 换成 $\xi_{n+k}\xi_k$, 并假设对于每一个整数 n, 序列 $\zeta=(\zeta_k)_{k\in\mathbb{Z}},\zeta_k=\xi_{n+k}\xi_k$ 是弱平稳的 (特别, 由此可见存在 4 阶矩 $\mathbf{E}\xi_0^4<\infty$). 因此条件

$$\frac{1}{N}\sum_{k=0}^{N-1}\mathbf{E}[\xi_{n+k}\xi_k-R(n)][\xi_n\xi_0-R(n)]\to 0,\quad N\to\infty, \tag{4}$$

对于

$$\mathbf{E}|\widehat{R}_N(n,\xi)-R(n)|^2\to 0,\quad N\to\infty \tag{5}$$

是充分和必要的.

假设原序列 $\xi=(\xi_n)$ 是 (均值为 0 和协方差为 $R(n)$ 的) 高斯序列, 则由于第二章 §12 的 (51) 式, 有

$$\begin{aligned}&\mathbf{E}[\xi_{n+k}\xi_k-R(n)][\xi_n\xi_0-R(n)]=\mathbf{E}\xi_{n+k}\xi_k\xi_n\xi_0-R^2(n)\\&=\mathbf{E}\xi_{n+k}\xi_k\mathbf{E}\xi_n\xi_0+\mathbf{E}\xi_{n+k}\xi_n\mathbf{E}\xi_k\xi_0+\mathbf{E}\xi_{n+k}\xi_0\mathbf{E}\xi_k\xi_n-R^2(n)\\&=R^2(k)+R(n+k)R(n-k).\end{aligned}$$

因此, 对于高斯情形, 条件 (4) 等价于如下条件:

$$\frac{1}{N}\sum_{k=0}^{N-1}[R^2(k)+R(n+k)R(n-k)]\to 0,\quad N\to\infty. \tag{6}$$

由于 $|R(n+k)R(n-k)|\leqslant|R(n+k)|^2+|R(n-k)|^2$, 故由条件

$$\frac{1}{N}\sum_{k=0}^{N-1}R^2(k)\to 0,\quad N\to\infty, \tag{7}$$

可得到条件 (6). 同时, 若条件 (6) 对于 $n=0$ 成立, 则条件 (7) 也成立.

于是, 我们证明了下面的定理.

定理 *设 $\xi=(\xi_n)$ 是高斯平稳序列, 其均值为 0, 而协方差函数为 $R(n)$. 则条件 (7) 是 "对于任何 $n\geqslant 0$, 估计 $\widehat{R}_N(n,\xi)$ 在均方意义下是协方差函数 $R(n)$ 的相合估计" 的充分必要条件, 即使 "条件 (5) 成立" 的充分必要条件.*

注 如果利用协方差函数的谱表达式, 则得

$$\frac{1}{N}\sum_{k=0}^{N-1}R^2(k)=\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\frac{1}{N}\sum_{k=0}^{N-1}\mathrm{e}^{i(\lambda-v)k}F(\mathrm{d}\lambda)F(\mathrm{d}v)=\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}f_N(\lambda,v)F(\mathrm{d}\lambda)F(\mathrm{d}v),$$

其中 (对照 §3 的 (36) 式)

$$f_N(\lambda,v)=\begin{cases}1, & \lambda=v,\\ \dfrac{1-\mathrm{e}^{i(\lambda-v)N}}{N[1-\mathrm{e}^{i(\lambda-v)}]}, & \lambda\neq v.\end{cases}$$

但当 $N\to\infty$

$$f_N(\lambda,v)\to f(\lambda,v)=\begin{cases}1, & \lambda=v,\\ 0, & \lambda\neq v.\end{cases}$$

从而

$$\frac{1}{N}\sum_{k=0}^{N-1}R^2(k)\to\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}f(\lambda,v)F(\mathrm{d}\lambda)F(\mathrm{d}v)=\int_{-\pi}^{\pi}F(\{\lambda\})F(\mathrm{d}\lambda)=\sum_{\lambda}F^2(\{\lambda\}),$$

其中由于测度 F 有限, 故对 λ 的和式中最多有可数项.

于是, 条件 (7) 与条件

$$\sum_{\lambda}F^2(\{\lambda\})=0, \tag{8}$$

等价. 而这说明谱函数 $F(\lambda)=F([-\pi,\lambda])$ 是*连续的*.

2. 谱函数和谱密度的估计的求法 现在讨论建立谱函数 $F(\lambda)$ 和谱密度 $f(\lambda)$ 的估计问题 (假设 $F(\lambda)$ 和 $f(\lambda)$ 存在).

自然, 由赫尔格洛茨定理 (第 55 页), 可以得出建立谱密度估计必经的途径. 回忆在 §1 引进的函数:

$$f_N(\lambda) = \frac{1}{2\pi} \sum_{|n|<N} \left(1 - \frac{|n|}{N}\right) R(n) \mathrm{e}^{-in\lambda}, \tag{9}$$

它具有如下性质, 由 $f_N(\lambda)$ 建立的函数

$$F_N(\lambda) = \int_{-\pi}^{\lambda} f_N(v)\mathrm{d}v$$

基本收敛于谱函数 $F(\lambda)$. 因此, 假如谱函数 $F(\lambda)$ 有密度 $f(\lambda)$, 则对于每个 $\lambda \in [-\pi,\pi)$, 有

$$\int_{-\pi}^{\lambda} f_N(v)\mathrm{d}v \to \int_{-\pi}^{\lambda} f(v)\mathrm{d}v. \tag{10}$$

由这些事实, 并注意到 $\widehat{R}_N(n,x)$ 是 (根据观测值 $x_0, x_1, \cdots, x_{N-1}$) 对 $R(n)$ 的估计值, 用函数

$$\widehat{f}_N(\lambda;x) = \frac{1}{2\pi} \sum_{|n|<N} \left(1 - \frac{|n|}{N}\right) \widehat{R}_N(n;x) \mathrm{e}^{-i\lambda n}, \tag{11}$$

做 $f(\lambda)$ 的估计值, 其中 $\widehat{R}_N(n;x) = \widehat{R}_N(|n|;x)$.

通常称函数 $\widehat{f}_N(\lambda;x)$ 为**周期图**, 并且不难验证函数 $\widehat{f}_N(\lambda;x)$ 也可以表示为如下较为方便的形式:

$$\widehat{f}_N(\lambda;x) = \frac{1}{2\pi N} \left| \sum_{n=0}^{N-1} x_n \mathrm{e}^{-i\lambda n} \right|^2. \tag{12}$$

由于 $\mathbf{E}\widehat{R}_N(n;\xi) = R(n), |n| < N$, 则

$$\mathbf{E}\widehat{f}_N(\lambda;\xi) = f_N(\lambda).$$

如果谱函数 $F(\lambda)$ 有密度 $f(\lambda)$, 则注意到 $f_N(\lambda)$ 可写成 §1 中 (34) 式的形式, 有

$$\begin{aligned} f_N(\lambda) &= \frac{1}{2\pi N} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} \int_{-\pi}^{\pi} \mathrm{e}^{iv(k-l)} \mathrm{e}^{i\lambda(l-k)} f(v)\mathrm{d}v \\ &= \int_{-\pi}^{\pi} \frac{1}{2\pi N} \left| \sum_{k=0}^{N-1} \mathrm{e}^{i(v-\lambda)k} \right|^2 f(v)\mathrm{d}v. \end{aligned}$$

函数

$$\Phi_N(\lambda) = \frac{1}{2\pi N} \left| \sum_{k=0}^{N-1} \mathrm{e}^{i\lambda k} \right|^2 = \frac{1}{2\pi N} \left| \frac{\sin \dfrac{\lambda}{2} N}{\sin \dfrac{\lambda}{2}} \right|^2$$

称做**费耶尔** (L. Féjer) **核**. 由该函数的性质知, 对于 (按勒贝格测度) 几乎所有的 λ, 有

$$\int_{-\pi}^{\pi} \Phi_N(\lambda - v) f(v)\mathrm{d}v \to f(\lambda). \tag{13}$$

因此对于几乎所有的 $\lambda \in [-\pi,\pi)$, 有

$$\mathbf{E}\widehat{f}_N(\lambda;\xi) \to f(\lambda), \tag{14}$$

换句话说, 谱函密度 $f(\lambda)$ 根据观测值 $x_0, x_1, \cdots, x_{N-1}$ 的估计 $\widehat{f}_N(\lambda;x)$, 是**渐近无偏的**.

在此意义上, 可以认为估计 $\widehat{f}_N(\lambda;x)$ 是充分 "好的". 然而, 周期图 $\widehat{f}_N(\lambda;x)$ 在个别观测值 $x_0, x_1, \cdots, x_{N-1}$ 上的值, 对真值 $f(\lambda)$ 的偏离往往是很大的. 例如, 设 $\xi = (\xi_n)$ 是独立平稳高斯随机变量序列, $\xi_n \sim N(0,1)$. 那么 $f(\lambda) \equiv 1/(2\pi)$, 而

$$\widehat{f}_N(\lambda;\xi) = \frac{1}{2\pi} \left| \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \xi_k \mathrm{e}^{-i\lambda k} \right|^2.$$

因此 $2\pi\widehat{f}_N(0;\xi)$ 与高斯随机变量 $\eta \sim N(0,1)$ 平方的同分布. 由此可见, 对于任意 N

$$\mathbf{E}|\widehat{f}_N(0;\xi) - f(0)|^2 = \frac{1}{4\pi^2}\mathbf{E}|\eta^2 - 1|^2 > 0.$$

此外, 经不复杂的计算可以证明, 如果 $f(\lambda)$ 是移动平均平稳序列 $\xi = (\xi_n)$:

$$\xi_n = \sum_{k=0}^{\infty} a_k \varepsilon_{n-k}, \tag{15}$$

的谱密度, 而且 $\sum_{k=0}^{\infty} |a_k| < \infty, \sum_{k=0}^{\infty} |a_k|^2 < \infty$, 其中 $\varepsilon = (\varepsilon_n)$ 是白噪声, 且 $\mathbf{E}\varepsilon_0^4 < \infty$, 则

$$\lim_{N\to\infty} \mathbf{E}|\widehat{f}_N(\lambda;\xi) - f(\lambda)|^2 = \begin{cases} 2f^2(0), & \lambda = 0, \pm\pi, \\ f^2(\lambda), & \lambda \neq 0, \pm\pi, \end{cases} \tag{16}$$

由此可见, 周期图不能做谱密度的满意的估计. 为矫正这种情况, 作为 $f(\lambda)$ 的估计常利用形如

$$\widehat{f}_N^W(\lambda;x) = \int_{-\pi}^{\pi} W_N(\lambda - v)\widehat{f}_N(v;x)\mathrm{d}v \tag{17}$$

的估计, 而 $\widehat{f}_N^W(\lambda;x)$ 是由周期图 $\widehat{f}_N(\lambda;x)$ 以及称做**谱窗**的 "光滑" 函数 $W_N(\lambda)$ 建立的估计. 对于函数 $W_N(\lambda)$ 自然的要求是:

a) $W_N(\lambda)$ 在点 $\lambda = 0$ 的邻域内有 "尖锐的" 极大值;

b) $\int_{-\pi}^{\pi} W_N(\lambda)\mathrm{d}\lambda = 1$;

c) $\mathbf{E}|\widehat{f}_N^W(\lambda;\xi) - f(\lambda)|^2 \to 0, N \to \infty, \lambda \in [-\pi,\pi)$.

由 (14) 式和条件 b) 知, 估计 $\widehat{f}_N^W(\lambda;\xi)$ 是渐近无偏的. 条件 c) 是估计为均方渐近相合性条件, 由上面的讨论知周期图不满足条件 c). 最后, 条件 a) 保障给定频率 λ 的周期图的 "尖锐度".

下面是形如 (17) 式的估计的一些例子.

巴特利特 (M. S. Bartlett) 估计, 基于谱窗

$$W_N(\lambda)=a_NB(a_N\lambda)$$

的选择, 其中 $a_N\uparrow\infty, a_N/N\to 0, N\to\infty$, 而

$$B(\lambda)=\frac{1}{2\pi}\left|\frac{\sin\frac{\lambda}{2}}{\frac{\lambda}{2}}\right|^2.$$

帕赞 (E. Parzen) 估计, 利用谱窗的函数

$$W_N(\lambda)=a_NP(a_N\lambda),$$

其中 a_N 像巴特利特的情形一样, 而

$$P(\lambda)=\frac{3}{8\pi}\left|\frac{\sin\frac{\lambda}{4}}{\frac{\lambda}{4}}\right|^4.$$

茹尔边科 (Журбенко) 估计利用形如

$$W_N(\lambda)=a_NZ(a_N\lambda)$$

的谱窗建立的, 其中

$$Z(\lambda)=\begin{cases}-\dfrac{\alpha+1}{2\alpha}|\lambda|^\alpha+\dfrac{\alpha+1}{2\alpha}, & |\lambda|\leqslant 1,\\ 0, & |\lambda|>1,\end{cases}$$

其中 $0<\alpha\leqslant 2$, 而 a_N 是专门选择的量.

我们不准备详细讨论谱密度的估计问题, 只是指出, 关于谱窗的建立, 以及与其相应的估计 $\widehat{f}_N^W(\lambda;x)$ 的性质的比较, 存在大量统计学文献. (例如, 见 [133],[71],[72]).

3. 谱函数的估计 现在考虑谱函数 $F(\lambda)=F([-\pi,\lambda])$ 的估计问题. 为此, 设

$$F_N(\lambda)=\int_{-\pi}^{\lambda}f_N(v)\mathrm{d}v,\quad \widehat{F}_N(\lambda;x)=\int_{-\pi}^{\lambda}\widehat{f}_N(v;x)\mathrm{d}v,$$

其中 $\widehat{f}_N(v;x)$ 是根据 $(x_0,x_1,\cdots,x_{N-1})$ 建立的周期图.

由赫尔格洛茨定理 (第 55 页) 的证明, 可见对于任意 $n\in\mathbb{Z}$, 当 $N\to\infty$ 时

$$\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}\mathrm{d}F_N(\lambda)\to\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}F(\mathrm{d}\lambda).$$

由此 (参照第三章 §3 定理 1 的系) 可得 $F_N\Rightarrow F$, 即在函数 $F(\lambda)$ 的每一个连续点上, $F_N(\lambda)$ 收敛于连续函数 $F(\lambda)$.

注意到, 对于一切 $|n|<N$, 有

$$\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}\mathrm{d}\widehat{F}_N(\lambda;\xi)=\widehat{R}_N(n;\xi)\left(1-\frac{|n|}{N}\right).$$

因此, 如果假设当 $N\to\infty$ 时 $\widehat{R}_N(n;\xi)$ 以概率 1 收敛于 $R(n)$, 则

$$\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}\mathrm{d}\widehat{F}_N(\lambda;\xi)\to\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}\mathrm{d}\widehat{F}(\lambda)\quad(\mathbf{P}-\text{a.c.}),$$

从而 $\widehat{F}_N(\lambda;\xi)\to F(\lambda)\quad(\mathbf{P}-\text{a.c.})$.

由此容易得到 (当证明必要性时, 需要考虑其子序列), $\widehat{R}_N(n;\xi)\xrightarrow{\mathbf{P}}R(n)$, 而也有 $\widehat{F}_N(\lambda;\xi)\xrightarrow{\mathbf{P}}F$.

4. 练习题

1. 设 (15) 式中的随机变量 $\varepsilon_n\sim N(0,1)$. 证明, 对于任何 n 与当 $N\to\infty$ 时, 有

$$(N-|n|)\mathbf{D}\widehat{R}_N(n;\xi)\to 2\pi\int_{-\pi}^{\pi}(1+\mathrm{e}^{2in\lambda})f^2(\lambda)\mathrm{d}\lambda.$$

2. 证明 (16) 式及其如下推广的正确性:

$$\lim_{N\to\infty}\operatorname{cov}(\widehat{f}_N(\lambda;\xi),\widehat{f}_N(v;\xi))=\begin{cases}2f^2(0), & \lambda=v=0,\pm\pi,\\ f^2(\lambda), & \lambda=v\neq 0,\pm\pi,\\ 0, & \lambda\neq v.\end{cases}$$

§5. 沃尔德分解

1. 平稳序列的正则分量和奇异分量 下面将要讨论的沃尔德 (A. Wold) 分解, 与 §3 的表达式 (2) 不同: (2) 式在*频率范围内*给出了平稳序列的分解, 而沃尔德分解是*在时间上的*分解. 沃尔德分解的实质在于, 平稳序列 $\xi=(\xi_n), n\in\mathbb{Z}$, 可以表示为两个平稳子列之和: 其中一个可以完全预测 (即它的值完全可以由 "过去" 的值复原), 而另一个却不可以预测.

我们首先引进若干记号. 设 $H_n(\xi)=\overline{L^2}(\xi^n)$ 和 $H(\xi)=\overline{L^2}(\xi)$ 相应为由随机变量 $\xi^n=(\cdots,\xi_{n-1},\xi_n)$ 和 $\xi=(\cdots,\xi_{n-1},\xi_n,\cdots)$ 生成的线性流形的闭包. 设

$$S(\xi)=\bigcap_n H_n(\xi).$$

对于任意元素 $\eta\in H(\xi)$, 以

$$\widehat{\pi}_n(\eta)=\widehat{\mathbf{E}}(\eta|H_n(\xi))$$

表示元素 η 在子空间 $H_n(\xi)$ 上的射影 (见第二章 §11). 记

$$\widehat{\pi}_{-\infty}(\eta)=\widehat{\mathbf{E}}[\eta|S(\xi)].$$

每一个元素 $\eta\in H(\xi)$, 可以表示为

$$\eta=\widehat{\pi}_{-\infty}(\eta)+[\eta-\widehat{\pi}_{-\infty}(\eta)],$$

其中 $\eta-\widehat{\pi}_{-\infty}(\eta)\perp\widehat{\pi}_{-\infty}(\eta)$. 因此, 空间 $H(\xi)$ 可以表示为正交和:

$$H(\xi)=S(\xi)\oplus R(\xi),$$

其中 $S(\xi)$ 由形如 $\widehat{\pi}_{-\infty}(\eta)[\eta\in H(\xi)]$ 的元素构成, 而 $R(\xi)$ 由形如 $\eta-\widehat{\pi}_{-\infty}(\eta)$ 的元素构成.

下面, 总是假设 $\mathbf{E}\xi_n=0,\mathbf{D}\xi_n>0$. 这样, 空间 $H(\xi)$ 显然是非平凡的 (包含非 0 元素).

定义 1 平稳序列 $\xi=(\xi_n)$ 称做*正则的*, 如果

$$H(\xi)=R(\xi),$$

而称做*奇异的*, 如果

$$H(\xi)=S(\xi).$$

注 1 奇异序列亦称为*确定序列*, 正则序列亦称为*纯粹或完全非确定序列*. 如果 $S(\xi)$ 是空间 $H(\xi)$ 的特征空间, 则序列 ξ 称做*非确定序列*.

定理 1 *任意弱平稳随机序列 ξ 可以分解为*

$$\xi_n=\xi_n^r+\xi_n^s,\tag{1}$$

其中 $\xi^r=(\xi_n^r)$ 是正则序列, 而 $\xi^s=(\xi_n^s)$ 是奇异序列. 而且 ξ^r 和 ξ^s 正交 (对于一切 n 和 m, $\xi_n^r\perp\xi_m^s$).

证明 根据定义, 设

$$\xi_n^s=\widehat{\mathbf{E}}[\xi_n|S(\xi)],\quad \xi_n^r=\xi_n-\xi_n^s.$$

由于对于任意 $n,\xi_n^r\perp S(\xi)$, 则 $S(\xi^r)\perp S(\xi)$. 另一方面, 因为 $S(\xi^r)\subseteq S(\xi)$, 说明 $S(\xi^r)$ 退化 (只含几乎必然为 0 的随机变量). 从而, 过程 ξ^r是正则的.

此外, 由于 $H_n(\xi)\subseteq H_n(\xi^s)\oplus H_n(\xi^r)$ 且 $H_n(\xi^s)\subseteq H_n(\xi),H_n(\xi^r)\subseteq H_n(\xi)$, 可见 $H_n(\xi)=H_n(\xi^s)\oplus H_n(\xi^r)$, 因而对于任意 n, 有

$$S(\xi)\subseteq H_n(\xi^s)\oplus H_n(\xi^r).\tag{2}$$

因为 $\xi_n^r\perp S(\xi)$, 所以由 (2) 式, 可见

$$S(\xi)\subseteq H_n(\xi^s),$$

因而 $S(\xi)\subseteq S(\xi^s)\subseteq H(\xi^s)$. 由于 $\xi_n^s\in S(\xi)$, 可见 $H(\xi^s)\subseteq S(\xi)$, 从而

$$S(\xi)=S(\xi^s)=H(\xi^s),$$

这说明序列 ξ^s 是奇异的.

由于 $\xi_n^s\in S(\xi)$, 而且 $\xi_n^r\perp S(\xi)$, 可见显然序列 ξ^r 和 ξ^s 正交. □

注 2 分解 (1) 式中, 将序列 ξ_n 分解为正则分量和奇异分量的唯一性, 见练习题 4.

2. 沃尔德分解

定义 2 设 $\xi=(\xi_n)$ 是非退化平稳序列. 称随机序列 $\varepsilon=(\varepsilon_n)$ (对于 ξ) 为*更新序列*, 如果:

a) $\varepsilon=(\varepsilon_n)$ 由 $\mathbf{E}\varepsilon_n=0,\mathbf{E}|\varepsilon_n|^2=1$ 的两两正交随机变量构成;

b) 对于任意 $n\in\mathbb{Z},H_n(\xi)=H_n(\varepsilon)$.

注 1 术语 "更新" 的含义与如下联想有关: ε_{n+1} 仿佛携带着 $H_n(\xi)$ 不包含的新的 "信息" (或者说, "更新" $H_n(\xi)$ 中形成 $H_{n+1}(\xi)$ 所必要的信息).

下面的重要定理是前面 (第 52 页例 4) 引进的, 单侧移动平均序列与正则序列之间建立联系.

定理 2 *对于非退化序列 ξ 为正则序列的充分和必要条件是, 存在一更新序列 $\varepsilon=(\varepsilon_n)$ 和复数的序列 $(a_n),n\geqslant 0$, 其中 $\sum\limits_{n=0}^{\infty}|a_n|^2<\infty$, 使*

$$\xi_n=\sum_{k=0}^{\infty}a_k\varepsilon_{n-k}\quad(\mathbf{P}-\text{a.c.}).\tag{3}$$

证明 (1) *必要性*. 将 $H_n(\xi)$ 表示为:

$$H_n(\xi)=H_{n-1}(\xi)\oplus B_n.$$

由于 $H_n(\xi)$ 是由 $H_{n-1}(\xi)$ 中形如 $\beta\xi_n$ (β 是复数) 的元素组成, 可见空间 B_n 的维数等于 1 或 0. 对于任何 n, $H_n(\xi)$ 都不会与 $H_{n-1}(\xi)$ 等同. 事实上, 假如对于某个 n, B_n 是平凡的, 则由于平稳性, 对于一切 k, B_k 也是平凡的, 而这表示 $H(\xi)=S(\xi)$, 然而这与序列 ξ 的正则性条件相矛盾. 于是, 空间 B_n 的维数等于 1.

假设 η_n 是 B_n 中的非 0 元素, 而

$$\varepsilon_n=\frac{\eta_n}{\|\eta_n\|},$$

其中 $\|\eta_n\|^2=\mathbf{E}|\eta_n|^2>0$.

对于固定的 n 和 k, 考虑分解:

$$H_n(\xi) = H_{n-k}(\xi) \oplus B_{n-k+1} \oplus \cdots \oplus B_n.$$

那么, $\varepsilon_{n-k}, \cdots, \varepsilon_n$ 在 $B_{n-k+1} \oplus \cdots \oplus B_n$ 中构成规范正交基, 且

$$\xi_n = \sum_{j=0}^{k-1} a_j \varepsilon_{n-j} + \widehat{\pi}_{n-k}(\xi_n), \tag{4}$$

其中 $a_j = \mathbf{E}\xi_n \overline{\varepsilon}_{n-j}$.

由贝塞尔不等式 (第二章 §11 的 (6) 式), 可见

$$\sum_{j=0}^{\infty} |a_j|^2 \leqslant \|\xi_n\|^2 < \infty.$$

由此可见, 级数 $\sum\limits_{j=0}^{\infty} a_j \varepsilon_{n-j}$ 在均方意义下收敛, 而由于 (4) 式, 为证明 (3) 式只需证明: 当 $k \to \infty$ 时 $\widehat{\pi}_{n-k}(\xi_n) \xrightarrow{L^2} 0$.

只需证明 $n = 0$ 的情形. 记 $\widehat{\pi}_i = \widehat{\pi}_i(\xi_0)$. 由于

$$\widehat{\pi}_{-k} = \widehat{\pi}_0 + \sum_{i=0}^{k} (\widehat{\pi}_{-i} - \widehat{\pi}_{-i+1}),$$

而和式中的各项正交, 故对于任意 $k \geqslant 0$, 有

$$\sum_{i=0}^{k} \|\widehat{\pi}_{-i} - \widehat{\pi}_{-i+1}\|^2 = \left\| \sum_{i=0}^{k} (\widehat{\pi}_{-i} - \widehat{\pi}_{-i+1}) \right\|^2 = \|\widehat{\pi}_{-k} - \widehat{\pi}_0\|^2 \leqslant 4\|\xi_0\|^2 < \infty.$$

于是, 存在均方极限 $\lim\limits_{k\to\infty} \widehat{\pi}_{-k}$. 对于每一个 $k, \widehat{\pi}_{-k} \in H_{-k}(\xi)$, 故所考虑的极限应属于空间 $\widehat{\pi}_{-k}$. 对于每个 $k, \widehat{\pi}_{-k} \in H_{-k}(\xi)$, 所考虑的极限应属于子空间 $\bigcap\limits_{k\geqslant 0} H_{-k}(\xi) = S(\xi)$. 因为根据假设 $S(\xi)$ 是退化的, 所以当 $k \to \infty$ 时 $\widehat{\pi}_{-k} \xrightarrow{L^2} 0$.

(2) *充分性*. 假设非退化序列 ξ 可以表示为 (3) 式的形式, 其中 $\varepsilon = (\varepsilon_n)$ 是正交系 (未必满足条件: $H_n(\xi) = H_n(\varepsilon), n \in \mathbb{Z}$). 那么, $H_n(\xi) \subseteq H_n(\varepsilon)$, 因而对于任意 n, 有 $S(\xi) = \bigcap\limits_k H_k(\xi) \subseteq H_n(\varepsilon)$. 由于 $\varepsilon_{n+1} \perp H_n(\varepsilon)$, 可见 $\varepsilon_{n+1} \perp S(\xi)$, 且 $\varepsilon = (\varepsilon_n)$ 是 $H(\xi)$ 的基. 由此可见, 子空间 $S(\xi)$ 是退化的, 所以序列 ξ 是正则的. □

注 2 由上面的证明, 可见非退化序列 ξ *是正则的*, 当且仅当它按照 §1 例 4 (第 52 页) 中的定义, 可以表示为*单侧移动平均*的形式:

$$\xi_n = \sum_{k=0}^{\infty} \tilde{a}_k \tilde{\varepsilon}_{n-k}, \tag{5}$$

其中 $\tilde{\varepsilon} = (\tilde{\varepsilon}_n)$ 是某一正交系. 在这种意义上, 由定理 2 的论断可以得到更多的结果. 具体地说, 对于正则序列 ξ, 存在这样的数列 $a = (a_n)$ 和正交系 $\varepsilon = (\varepsilon_n)$, 使与 (5) 式同时 (3) 式也成立, 且对于 (3) 式 $H_n(\xi) = H_n(\varepsilon), n \in \mathbb{Z}$.

由定理 1 和定理 2 直接得下面的定理.

定理 3 (沃尔德分解) *如果 $\xi = (\xi_n)$ 是非退化平稳序列, 则*

$$\xi_n = \xi_n^s + \sum_{k=0}^{\infty} a_k \varepsilon_{n-k}, \tag{6}$$

其中 $\sum\limits_{k=0}^{\infty} |a_k|^2 < \infty$, 而 $\varepsilon = (\varepsilon_n)$ (对于 ξ^r) 是一更新序列.

3. 外推和预测 在讨论下面的 (线性) 外推问题时, 上面引进的正则和奇异序列的含义, 就显得特别清晰. 利用沃尔德分解 (6) 非常有助于 (线性) 外推问题的一般解.

设 $H_0(\xi) = \overline{L^2}(\xi^0)$ 是由变量 $\xi^0 = (\cdots, \xi_{-1}, \xi_0)$ 生成的封闭线性流形. 考虑根据 "过去的" 观测结果 $\xi^0 = (\cdots, \xi_{-1}, \xi_0)$, 建立变量 ξ_n 的 (均方) 最优线性估计 $\widehat{\xi}_n$ 的问题.

由第二章 §11, 可见

$$\widehat{\xi}_n = \widehat{\mathbf{E}}[\xi_n | H_0(\xi)]. \tag{7}$$

(按第 1 小节的记号, $\widehat{\xi}_n = \widehat{\pi}_0(\xi_n)$). 由于 ξ^r 和 ξ^s 正交且 $H_0(\xi) = H_0(\xi^r) \oplus H_0(\xi^s)$, 则注意到 (6) 式, 有

$$\begin{aligned}
\widehat{\xi}_n &= \widehat{\mathbf{E}}[\xi_n^s + \xi_n^r | H_0(\xi)] = \widehat{\mathbf{E}}[\xi_n^s | H_0(\xi)] + \widehat{\mathbf{E}}[\xi_n^r | H_0(\xi)] \\
&= \widehat{\mathbf{E}}[\xi_n^s | H_0(\xi^r) \oplus H_0(\xi^s)] + \widehat{\mathbf{E}}[\xi_n^r | H_0(\xi^r) \oplus H_0(\xi^s)] \\
&= \widehat{\mathbf{E}}[\xi_n^s | H_0(\xi^s)] + \widehat{\mathbf{E}}[\xi_n^r | H_0(\xi^r)] = \xi_n^s + \widehat{\mathbf{E}}\left[\sum_{k=0}^{\infty} a_k \varepsilon_{n-k} \middle| H_0(\xi^r) \right].
\end{aligned}$$

在 (6) 式中, 对于 $\varepsilon = (\varepsilon_n), \xi^r = (\xi_n^r)$ 是更新序列, 从而 $H_0(\xi^r) = H_0(\varepsilon)$. 因此

$$\widehat{\xi}_n = \xi_n^s + \widehat{\mathbf{E}}\left[\sum_{k=0}^{\infty} a_k \varepsilon_{n-k} \middle| H_0(\varepsilon) \right] = \xi_n^s + \sum_{k=n}^{\infty} a_k \varepsilon_{n-k}, \tag{8}$$

而根据 $\xi^0 = (\cdots, \xi_{-1}, \xi_0)$ 预测 ξ_n 的均方误差等于

$$\sigma_n^2 = \mathbf{E}|\xi_n - \widehat{\xi}_n|^2 = \sum_{k=0}^{n-1} |a_k|^2. \tag{9}$$

由此可得如下两个重要结论:

a) 如果序列 ξ *是奇异的*, 则对于任意 $n \geqslant 1$, (外推) 误差 σ_n^2 等于 0. 换句话说, 可以根据 "过去" $\xi^0 = (\cdots, \xi_{-1}, \xi_0)$ 无误差地预测 ξ_n.

b) 如果序列 ξ *是正则的*, 则 $\sigma_n^2 \leqslant \sigma_{n+1}^2$, 而且

$$\lim_{n\to\infty} \sigma_n^2 = \sum_{k=0}^{\infty} |a_k|^2. \tag{10}$$

由于

$$\sum_{k=0}^{\infty}|a_k|^2=\mathbf{E}|\xi_n|^2,$$

则由 (10) 和 (9) 两式可见

$$\widehat{\xi}_n \xrightarrow{L^2} 0, \quad n\to\infty,$$

即随 n 的增大, 由 $\xi^0=(\cdots,\xi_{-1},\xi_0)$ 预测随机变量 ξ_n 是平凡的 (简单地就等于 $\mathbf{E}\xi_n=0$).

4. 正则序列的充分和必要条件 假设 ξ 是非退化正则平稳序列. 根据定理 2, 任何这样的序列可以表示为*单侧移动平均*的形式:

$$\xi_n=\sum_{k=0}^{\infty}a_k\varepsilon_{n-k}, \tag{11}$$

其中 $\sum_{k=0}^{\infty}|a_k|^2<\infty$, 规范正交序列 $\varepsilon=(\varepsilon_n)$ 具有如下重要性质:

$$H_n(\xi)=H_n(\varepsilon), \quad n\in\mathbb{Z}. \tag{12}$$

表达式 (11) 表示 (见 §3 第 3 小节), 当在*物理可实现的滤波器*入口传递序列 $\varepsilon=(\varepsilon_n)$ 时, 以 $a=(a_k),k\geqslant 0$, 为脉冲转移函数的滤波器、在出口的接收信号为 ξ_n.

像任何双侧移动平均一样, 正则序列具有谱密度 $f(\lambda)$. 但是, 由正则序列可以表示为*单侧移动平均*的情形, 可以得到关于谱密度的补充信息.

首先, 显然

$$f(\lambda)=\frac{1}{2\pi}|\varphi(\lambda)|^2,$$

其中

$$\varphi(\lambda)=\sum_{k=0}^{\infty}\mathrm{e}^{-i\lambda k}a_k, \quad \sum_{k=0}^{\infty}|a_k|^2<\infty. \tag{13}$$

设

$$\Phi(z)=\sum_{k=0}^{\infty}a_kz^k. \tag{14}$$

该函数在开区域 $\{z:|z|<1\}$ 是解析函数, 并且由于条件 $\sum_{n=0}^{\infty}|a_n|^2<\infty$, 属于所谓哈代 (G. H. Hardy) *函数类* H^2, 即在满足

$$\sup_{0\leqslant r<1}\frac{1}{2\pi}\int_{-\pi}^{\pi}|g(re^{i\theta})|^2\mathrm{d}\theta<\infty. \tag{15}$$

的、区域 $\{z:|z|<1\}$ 上的解析函数类 $g=g(z)$. 事实上,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}|\Phi(re^{i\theta})|^2\mathrm{d}\theta=\sum_{k=0}^{\infty}|a_k|^2r^{2k}$$

且

$$\sup_{0\leqslant r<1}\sum_{k=0}^{\infty}|a_k|^2r^{2k}\leqslant\sum_{k=0}^{\infty}|a_k|^2<\infty.$$

复变函数论中证明, 恒不等于 0 函数 $\Phi\in H^2, -\pi\leqslant\lambda<\pi$ 的、边界值 $\Phi(\mathrm{e}^{i\lambda})$ 满足如下性质:

$$\int_{-\pi}^{\pi}\ln|\Phi(\mathrm{e}^{-i\lambda})|\mathrm{d}\lambda>-\infty. \tag{16}$$

对于现在所考虑的情形,

$$f(\lambda)=\frac{1}{2\pi}|\Phi(\mathrm{e}^{-i\lambda})|^2,$$

其中 $\Phi\in H^2$. 因此

$$\ln f(\lambda)=-\ln 2\pi+2\ln|\Phi(\mathrm{e}^{-i\lambda})|,$$

从而, 正则过程的谱密度 $f(\lambda)$ 满足条件

$$\int_{-\pi}^{\pi}\ln f(\lambda)\mathrm{d}\lambda>-\infty. \tag{17}$$

另一方面, 假设谱密度 $f(\lambda)$ 满足条件 (17). 仍然由复变函数论的性质, 可见在哈代函数类 H^2 中, 存在函数 $\Phi(z)=\sum_{n=0}^{\infty}a_nz^k$, 使 (按勒贝格测度几乎处处)

$$f(\lambda)=\frac{1}{2\pi}|\Phi(\mathrm{e}^{-i\lambda})|^2.$$

因此, 设 $\varphi(\lambda)=\Phi(\mathrm{e}^{-i\lambda})$, 得

$$f(\lambda)=\frac{1}{2\pi}|\varphi(\lambda)|^2,$$

其中 $\varphi(\lambda)$ 是由 (13) 式中的函数. 那么, 由 §3 定理 3 系 1 可见, 序列 ξ 可以表示为单侧移动平均 (11) 的形式, 其中 $\varepsilon=(\varepsilon_n)$ 是某一规范正交序列. 由此以及第 2 小节注 2, 可见序列 ξ 是正则的.

于是, 我们证明了下面的定理.

定理 4 (柯尔莫戈洛夫) *设 ξ 是非退化正则平稳序列, 则存在满足*

$$\int_{-\pi}^{\pi}\ln f(\lambda)\mathrm{d}\lambda>-\infty. \tag{18}$$

谱密度 $f(\lambda)$. 特别, (按勒贝格测度几乎处处) $f(\lambda)>0$.

相反, 如果 ξ 是某一有谱密度且满足条件 (18) 的平稳序列, 则该序列是正则的.

5. 练习题

1. 证明具有离散谱 (谱函数 $F(\lambda)$ 是阶梯函数) 的平稳序列是奇异的.

2. 假设 $\sigma_n^2=\mathbf{E}|\xi_n-\widehat{\xi}_n|^2, \widehat{\xi}_n=\widehat{\mathbf{E}}[\xi_n|H_0(\xi)]$. 证明, 如果对于某个 $n\geqslant 1, \sigma_n^2=0$, 则序列 ξ 是奇异的; 而假如 $n\to\infty, \sigma_n^2\to R(0)$, 则序列 ξ 是正则的.

3. 证明平稳序列 $\xi=(\xi_n),\xi_n=\mathrm{e}^{in\varphi}$, 是正则的, 其中 φ 是在 $[0,2\pi]$ 上的均匀分布随机变量. 求估计 $\widehat{\xi}$ 和 σ_n^2 的值. 证明非线性估计

$$\widetilde{\xi}_n=\left(\frac{\xi_0}{\xi_{-1}}\right)^n$$

是根据"过去" $\xi^0=(\cdots,\xi_{-1},\xi_0)$ 对 ξ_n 的无误差预测, 即

$$\mathbf{E}|\xi_n-\widetilde{\xi}_n|^2=0,\quad n\geqslant 1.$$

4. 证明, 将 ξ_n 分为正则分量和奇异分量的分解 (1) 唯一.

§6. 外推、内插和过滤

1. 外推 对于奇异序列的情形, 由上一节结果, 可以根据"过去" $\xi^0=(\cdots,\xi_{-1},\xi_0)$, 进行无误差预测 (外推) 随机变量 $\xi_n,n\geqslant 1$. 因此, 自然在讨论任意随机序列的外推问题时, 首先研究正则序列的情形.

根据 §5 的定理 2, 任何正则序列 $\xi=(\xi_n)$ 都可以表示为单侧移动平均

$$\xi_n=\sum_{k=0}^{\infty}a_k\varepsilon_{n-k},\tag{1}$$

的情形, 其中 $\sum\limits_{k=0}^{\infty}|a_k|^2<\infty,\varepsilon=(\varepsilon_n)$ 是某一更新序列. 因为根据 §5 的 (8) 式, 有

$$\widehat{\xi}_n=\sum_{k=n}^{\infty}a_k\varepsilon_{n-k}\tag{2}$$

和

$$\sigma_n^2=\mathbf{E}|\xi_n-\widehat{\xi}_n|^2=\sum_{k=0}^{n-1}|a_k|^2.\tag{3}$$

所以由 §5 可见, 表达式 (1) 解决求最优 (线性) 估计 $\widehat{\xi}_n=\widehat{\mathbf{E}}(\xi_n|H_0(\xi))$ 的问题. 不过, 由于如下的原因, 该解只能是原则性的解.

通常所考虑的序列并没有 (1) 式表达, 而是由其协方差函数 $R(n)$ 或谱密度 $f(\lambda)$ 表达 (对于正则序列 $f(\lambda)$ 存在). 因此, 只有在系数 a_k 通过 $R(n)$ 或 $f(\lambda)$ 值表示, 而变量 ε_k 通过的 $\cdots,\xi_{k-1},\xi_k$ 值表示的情况下, 才可以认为解 (2) 是满意的.

我们不涉及这一问题的一般形式, 而局限于讨论一种 (对应用感兴趣的) 特殊情形, 即谱密度表示为

$$f(\lambda)=\frac{1}{2\pi}|\Phi(\mathrm{e}^{-i\lambda})|^2,\tag{4}$$

的情形, 其中函数 $\Phi(z)=\sum\limits_{k=0}^{\infty}b_kz^k$ 的收敛半径为 $r>1$, 而且在区域 $\{z:|z|\leqslant 1\}$ 内无 0 点.

设

$$\xi_n=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}Z(\mathrm{d}\lambda)\tag{5}$$

是序列 $\xi=(\xi_n),n\in\mathbb{Z}$, 的谱表示.

定理 1 如果序列 ξ 的谱密度表示为 (4), 则根据 $\xi^0=(\cdots,\xi_{-1},\xi_0)$ 对随机变量 ξ_n 的最优 (线性) 估计 $\widehat{\xi}_n$ 由公式

$$\widehat{\xi}_n=\int_{-\pi}^{\pi}\widehat{\varphi}_n(\lambda)Z(\mathrm{d}\lambda)\tag{6}$$

表示, 其中

$$\widehat{\varphi}_n(\lambda)=\mathrm{e}^{i\lambda n}\frac{\Phi_n(\mathrm{e}^{-i\lambda})}{\Phi(\mathrm{e}^{-i\lambda})}\tag{7}$$

而

$$\Phi_n(z)=\sum_{k=n}^{\infty}b_kz^k.$$

证明 根据 §3 定理 2 的注, 任意随机变量 $\widetilde{\xi}_n\in H_0(\xi)$ 可以表示为:

$$\widetilde{\xi}_n=\int_{-\pi}^{\pi}\widetilde{\varphi}_n(\lambda)Z(\mathrm{d}\lambda),\quad \widetilde{\varphi}_n\in H_0(F),\tag{8}$$

其中 $H_0(F)$ 是由函数

$$e_n=\mathrm{e}^{i\lambda n}\ (n\geqslant 0)\left[F(\lambda)=\int_{-\pi}^{\lambda}f(v)\mathrm{d}v\right].$$

生成的闭线性流形.

由于

$$\mathbf{E}|\xi_n-\widetilde{\xi}_n|^2=\left|\mathbf{E}\int_{-\pi}^{\pi}[\mathrm{e}^{i\lambda n}-\widetilde{\varphi}_n(\lambda)]Z(\mathrm{d}\lambda)\right|^2=\int_{-\pi}^{\pi}\left|\mathrm{e}^{i\lambda n}-\widetilde{\varphi}_n(\lambda)\right|^2f(\lambda)\mathrm{d}\lambda,$$

则估计 (6) 式之最优性的证明归结为证明

$$\inf_{\widetilde{\varphi}_n\in H_0(F)}\int_{-\pi}^{\pi}\left|\mathrm{e}^{i\lambda n}-\widetilde{\varphi}_n(\lambda)\right|^2f(\lambda)\mathrm{d}\lambda=\int_{-\pi}^{\pi}\left|\mathrm{e}^{i\lambda n}-\widehat{\varphi}_n(\lambda)\right|^2f(\lambda)\mathrm{d}\lambda.\tag{9}$$

由希尔伯特空间的理论 (第二章 §11) 可见, (在 (9) 式意义下的) 最优函数 $\widehat{\varphi}_n(\lambda)$ 决定于如下两个条件:

$$\begin{aligned}&1)\ \widehat{\varphi}_n(\lambda)\in H_0(F);\\&2)\ [\mathrm{e}^{i\lambda n}-\widehat{\varphi}_n(\lambda)]\perp H_0(F).\end{aligned}\tag{10}$$

由于

$$\mathrm{e}^{i\lambda n}\Phi_n(\mathrm{e}^{-i\lambda})=\mathrm{e}^{i\lambda n}[b_n\mathrm{e}^{-i\lambda n}+b_{n+1}\mathrm{e}^{-i\lambda(n+1)}+\cdots]\in H_0(F),$$

且类似地 $\Phi_n^{-1}(\mathrm{e}^{-i\lambda}) \in H_0(F)$, 则由 (7) 式定义的函数 $\widehat{\varphi}_n(\lambda)$ 属于函数类 $H_0(F)$. 因此, 为证明函数 $\widehat{\varphi}_n(\lambda)$ 的 "最优性", 只需对于任意 $m \geqslant 0$, 证明

$$[\mathrm{e}^{i\lambda n} - \widehat{\varphi}_n(\lambda)] \perp \mathrm{e}^{-i\lambda m},$$

即

$$I_{n,m} \equiv \int_{-\pi}^{\pi} [\mathrm{e}^{i\lambda n} - \widehat{\varphi}_n(\lambda)]\mathrm{e}^{i\lambda m} f(\lambda)\mathrm{d}\lambda, \quad m \geqslant 0.$$

下面的一系列等式就证明这一论断.

$$\begin{aligned}
I_{n,m} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{i\lambda(n+m)}\left[1 - \frac{\Phi_n(\mathrm{e}^{-i\lambda})}{\Phi(\mathrm{e}^{-i\lambda})}\right]|\Phi(\mathrm{e}^{-i\lambda})|^2\mathrm{d}\lambda \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{i\lambda(n+m)}[\Phi(\mathrm{e}^{-i\lambda}) - \Phi_n(\mathrm{e}^{-i\lambda})]\overline{\Phi(\mathrm{e}^{-i\lambda})}\mathrm{d}\lambda \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{i\lambda(n+m)}\left(\sum_{k=0}^{n-1} b_k \mathrm{e}^{-i\lambda k}\right)\left(\sum_{l=0}^{\infty} \overline{b}_l \mathrm{e}^{i\lambda l}\right)\mathrm{d}\lambda \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{i\lambda m}\left(\sum_{k=0}^{n-1} b_k \mathrm{e}^{i\lambda(n-k)}\right)\left(\sum_{l=0}^{\infty} \overline{b}_l \mathrm{e}^{i\lambda l}\right)\mathrm{d}\lambda = 0,
\end{aligned}$$

其中最后一等式成立, 因为对于任意 $m \geqslant 0$ 和 $r > 1$,有

$$\int_{-\pi}^{\pi} \mathrm{e}^{i\lambda m}\mathrm{e}^{i\lambda r}\mathrm{d}\lambda = 0.$$

□

注 1 如果将函数 $\widehat{\varphi}_n(\lambda)$ 展成傅里叶级数

$$\widehat{\varphi}_n(\lambda) = C_0 + C_{-1}\mathrm{e}^{-i\lambda} + C_{-2}\mathrm{e}^{-2i\lambda} + \cdots$$

则根据 "过去" $\xi^0 = (\cdots, \xi_{-1}, \xi_0)$, 对随机变量 $\xi_n, n \geqslant 1$, 的预测 (外推) $\widehat{\xi}_n$ 决定于如下公式:

$$\widehat{\xi}_n = C_0\xi_0 + C_{-1}\xi_{-1} + C_{-2}\xi_{-2} + \cdots$$

注 2 *有理函数*

$$f(\lambda) = \frac{1}{2\pi}\left|\frac{P(\mathrm{e}^{-i\lambda})}{Q(\mathrm{e}^{-i\lambda})}\right|^2,$$

是由 (4) 式表示的谱密度的典型例子, 其中多项式 $P(z) = a_0 + a_1 z + \cdots + a_p z^p$ 和 $Q(z) = 1 + b_1 z + \cdots + b_q z^q$ 在区域 $\{z : |z| \leqslant 1\}$ 内无 0 点.

事实上, 在这种情形下只需设 $\Phi(z) = P(z)/Q(z)$. 那么

$$\Phi(z) = \sum_{k=0}^{\infty} C_k z^k,$$

而且该级数的收敛半径大于 1.

举两个演示定理 1 的例子.

例 1 设谱密度为

$$f(\lambda) = \frac{1}{2\pi}(5 + 4\cos\lambda).$$

相应的协方差函数 $R(n)$ 具有 "三角形" 形状:

$$R(0) = 5, \quad R(\pm 1) = 2, \quad R(n) = 0(|n| \geqslant 2). \tag{11}$$

由于所考虑的谱密度可以表示为

$$f(\lambda) = \frac{1}{2\pi}|2 + \mathrm{e}^{-i\lambda}|^2,$$

故可以运用定理 1. 易见

$$\widehat{\varphi}_1(\lambda) = \mathrm{e}^{i\lambda}\frac{\mathrm{e}^{-i\lambda}}{2 + \mathrm{e}^{-i\lambda}}, \quad \widehat{\varphi}_n(\lambda) = 0(n \geqslant 2). \tag{12}$$

所以对于一切 $n \geqslant 2, \xi_n = 0$, 即根据 $\xi^0 = (\cdots, \xi_{-1}, \xi_0)$ 对 ξ_n 的 (线性) 预测是*平凡的*. 因为, 如果注意到根据 (11), $\xi_n (n \geqslant 2)$ 与 $\xi_0, \xi_{-1}, \cdots$, 中任意变量的相关性为 0, 则这一点也不奇怪.

对于 $n = 1$, 由 (6) 和 (12) 两式, 可见

$$\begin{aligned}
\widehat{\xi}_1 &= \int_{-\pi}^{\pi} \mathrm{e}^{i\lambda}\frac{\mathrm{e}^{-i\lambda}}{2 + \mathrm{e}^{-i\lambda}} Z(\mathrm{d}\lambda) = \frac{1}{2}\int_{-\pi}^{\pi} \frac{1}{1 + \frac{\mathrm{e}^{-i\lambda}}{2}} Z(\mathrm{d}\lambda) \\
&= \sum_{k=0}^{\infty} \frac{(-1)^k}{2^{k+1}}\int_{-\pi}^{\pi} \mathrm{e}^{-ik\lambda} Z(\mathrm{d}\lambda) = \sum_{k=0}^{\infty} \frac{(-1)^k \xi_k}{2^{k+1}} \\
&= \frac{1}{2}\xi_0 - \frac{1}{4}\xi_{-1} + \cdots
\end{aligned}$$

例 2 设协方差函数为

$$R(n) = a^n, \quad |a| < 1.$$

那么 (见第 53 页例 5)

$$f(\lambda) = \frac{1}{2\pi}\frac{1 - |a|^2}{|1 - a\mathrm{e}^{-i\lambda}|^2},$$

即

$$f(\lambda) = \frac{1}{2\pi}|\Phi(\mathrm{e}^{-i\lambda})|^2,$$

其中

$$\Phi(z) = \frac{(1 - |a|^2)^{1/2}}{1 - az} = (1 - |a|^2)^{1/2}\sum_{k=0}^{\infty}(az)^k,$$

由于 $\widehat{\varphi}_n(\lambda) = a^n$, 得

$$\widehat{\xi}_n = \int_{-\pi}^{\pi} a^n Z(\mathrm{d}\lambda) = a^n \xi_0.$$

换句话说, 为由观测值 $\xi^0 = (\cdots, \xi_{-1}, \xi_0)$ 预测 ξ_n, 只需知道**最后一个**观测值 ξ_0.

注 3 由正则序列 $\xi=(\xi_n)$ 的沃尔德分解

$$\xi_n=\sum_{k=0}^{\infty}a_k\varepsilon_{n-k}, \tag{13}$$

可见谱密度 $f(\lambda)$ 可以表示为

$$f(\lambda)=\frac{1}{2\pi}|\Phi(\mathrm{e}^{-i\lambda})|^2, \tag{14}$$

其中

$$\Phi(z)=\sum_{k=0}^{\infty}a_kz^k. \tag{15}$$

显然, 相反得命题也成立: 如果 $f(\lambda)$ 可以由形如 (15) 式的函数 $\Phi(z)$ 表示为 (14) 式, 则序列 $\xi=(\xi_n)$ 的沃尔德分解具有 (13) 式的形式. 从而, 谱密度 $f(\lambda)$ 形如 (14) 式的表现问题, 与沃尔德分解中求系数 a_k 的问题等价.

定理 1 中关于函数 $\Phi(z)$ 所作的假设: 对于 $r>1$, 在区域 $\{z:|z|\leqslant 1\}$ 内函数 $\Phi(z)$ 无 0 点, 实际上并不需要. 换句话说, 如果正则序列的谱密度由形如 (14) 的式子表示, 则由 $\xi^0=(\cdots,\xi_{-1},\xi_0)$ 建立的、变量 ξ_n 的均方最优估计由 $\widehat{\xi}_n$ 决定于 (6) 和 (7) 两式.

注 4 定理 1 (连同上面的注 3) 给出了*正则序列的预测问题的解*. 现在证明, 事实上对于*任意平稳序列*, 有同样的答案. 确切地说, 设 $\xi_n=\xi_n^s+\xi_n^r$,

$$\xi_n=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}Z(\mathrm{d}\lambda),\quad F(\Delta)=\mathbf{E}|Z(\Delta)|^2,\quad f^r(\lambda)=\frac{1}{2\pi}|\Phi(\mathrm{e}^{-i\lambda})|^2,$$

其中 $f^r(\lambda)$ 是正则序列 $\xi^r=(\xi_n^r)$ 的谱密度. 那么, 估计 $\widehat{\xi}_n$ 决定于 (6) 和 (7) 两式.

实际上 (见 §5 第 3 小节), 设

$$\widehat{\xi}_n=\int_{-\pi}^{\pi}\widehat{\varphi}_n(\lambda)Z(\mathrm{d}\lambda),\quad \widehat{\xi}_n^r=\int_{-\pi}^{\pi}\widehat{\varphi}_n^r(\lambda)Z^r(\mathrm{d}\lambda),$$

其中 $Z^r(\Delta)$ 是正则序列 ξ^r 表现中的正交随机测度. 那么,

$$\begin{aligned}\mathbf{E}|\xi_n-\widehat{\xi}_n|^2&=\int_{-\pi}^{\pi}|\mathrm{e}^{i\lambda n}-\widehat{\varphi}_n(\lambda)|^2F(\mathrm{d}\lambda)\geqslant\int_{-\pi}^{\pi}|\mathrm{e}^{i\lambda n}-\widehat{\varphi}_n(\lambda)|^2f^r(\lambda)\mathrm{d}\lambda\\&\geqslant\int_{-\pi}^{\pi}|\mathrm{e}^{i\lambda n}-\widehat{\varphi}_n^r(\lambda)|^2f^r(\lambda)\mathrm{d}\lambda=\mathbf{E}|\xi_n^r-\widehat{\xi}_n^r|^2.\end{aligned} \tag{16}$$

其次, 由于 $\mathbf{E}|\xi_n-\widehat{\xi}_n|^2=\mathbf{E}|\xi_n^r-\widehat{\xi}_n^r|^2$, 可见 $\xi_n-\widehat{\xi}_n=\xi_n^r-\widehat{\xi}_n^r$, 故由 (16) 式知可以将函数 $\widehat{\varphi}_n(\lambda)$ 选为 $\widehat{\varphi}_n^r(\lambda)$.

2. 内插 假设 $\xi=(\xi_n)$ 是谱密度为 $f(\lambda)$ 的正则序列. 根据对 ξ_0 的"过去"值 $\{\xi_n,n=\pm1,\pm2,\cdots\}$ 的观测结果, 建立 (均方) 最优线性估计是最简单的*内插问题*.

以 $H^0(\xi)$ 表示随机变量 $\xi_n(n\neq 0)$ 生成的封闭线性流形. 那么, 根据 §3 的定理 2, 任意随机变量 $\eta\in H^0(\xi)$ 可以表示为:

$$\eta=\int_{-\pi}^{\pi}\varphi(\lambda)Z(\mathrm{d}\lambda),$$

其中 $\varphi\in H^0(F)$ 是函数 $\mathrm{e}^{i\lambda n}(n\neq 0)$ 生成的封闭线性流形, 而估计

$$\check{\xi}=\int_{-\pi}^{\pi}\check{\varphi}(\lambda)Z(\mathrm{d}\lambda) \tag{17}$$

是最优的, 当且仅当

$$\begin{aligned}\inf_{\eta\in H^0(\xi)}\mathbf{E}|\xi_0-\eta|^2&=\inf_{\varphi\in H^0(F)}\int_{-\pi}^{\pi}|1-\varphi(\lambda)|^2F(\mathrm{d}\lambda)\\&=\int_{-\pi}^{\pi}|1-\check{\varphi}_n(\lambda)|^2F(\mathrm{d}\lambda)=\mathbf{E}|\xi_0-\check{\xi}_0|^2.\end{aligned}$$

由希尔伯特空间 $H^0(F)$ 中"垂线"性质, 可见 (对照 (10) 式) 函数 $\check{\varphi}(\lambda)$ 完全决定于如下两个条件:

$$\begin{aligned}&1)\ \check{\varphi}(\lambda)\in H^0(F);\\&2)\ 1-\check{\varphi}(\lambda)\perp H^0(F).\end{aligned} \tag{18}$$

定理 2 (柯尔莫戈洛夫) *设 $\xi=(\xi_n)$ 是谱密度为 $f(\lambda)$ 的正则序列, 且*

$$\int_{-\pi}^{\pi}\frac{\mathrm{d}\lambda}{f(\lambda)}=0. \tag{19}$$

那么,

$$\check{\varphi}(\lambda)=1-\frac{\alpha}{f(\lambda)}, \tag{20}$$

其中

$$\alpha=\frac{2\pi}{\displaystyle\int_{-\pi}^{\pi}\frac{\mathrm{d}\lambda}{f(\lambda)}}, \tag{21}$$

而内插的误差 $\delta^2=\mathbf{E}|\xi_0-\check{\xi}_0|^2$ 为 $\delta^2=2\pi\alpha$.

证明 证明将在关于谱密度相当严格的条件下进行, 假设

$$0<c\leqslant f(\lambda)\leqslant C<\infty. \tag{22}$$

由 (18) 式中的条件 2) 可见, 对于任意 $n\neq 0$, 有

$$\int_{-\pi}^{\pi}|1-\check{\varphi}(\lambda)|\mathrm{e}^{i\lambda n}f(\lambda)\mathrm{d}\lambda=0. \tag{23}$$

在 (22) 式的条件下, 函数 $|1-\check{\varphi}(\lambda)|f(\lambda)$ 属于带勒贝格测度 $d\mu$ 的、希尔伯特空间 $L^2([-\pi,\pi],\mathscr{B}([-\pi,\pi]),d\mu)$. 在此空间中函数系

$$\left\{\frac{e^{i\lambda n}}{\sqrt{2\pi}},\quad n=0,\pm1,\cdots\right\}$$

规范正交基 (第二章 §12 的练习题 10). 所以由 (23) 式可见, 函数 $|1-\check{\varphi}(\lambda)|f(\lambda)$ 是常数, 记作 α.

这样, 由 (18) 式的第二个条件, 得

$$\check{\varphi}(\lambda)=1-\frac{\alpha}{f(\lambda)}. \tag{24}$$

现在根据 (18) 式的第一个条件, 确定常数 α.

由于 (22) 式可见, $\check{\varphi}\in L^2$ 和 $\check{\varphi}\in H^0(F)$ 与如下的条件等价: $\check{\varphi}$ 属于函数 $e^{i\lambda n}(n\neq 0)$ 生成的、(在 L^2 中范数的意义上) 封闭线性流形. 由于在函数的分解式中零系数应等于 0, 故

$$0=\int_{-\pi}^{\pi}\check{\varphi}(\lambda)d\lambda=2\pi-\alpha\int_{-\pi}^{\pi}\frac{d\lambda}{f(\lambda)},$$

可见常数 α 决定于 (21) 式.

最后,

$$\delta^2=\mathbf{E}|\xi_0-\check{\xi}_0|^2=\int_{-\pi}^{\pi}|1-\check{\varphi}(\lambda)|^2f(\lambda)d\lambda=|\alpha|^2\int_{-\pi}^{\pi}\frac{f(\lambda)}{f^2(\lambda)}d\lambda=\frac{4\pi^2}{\displaystyle\int_{-\pi}^{\pi}\frac{d\lambda}{f(\lambda)}}.$$

于是, 在 (22) 式的补充条件下定理得证. □

系 如果

$$\check{\varphi}(\lambda)=\sum_{0<|k|\leqslant N}c_ke^{i\lambda k},$$

则

$$\check{\xi}_0=\sum_{0<|k|\leqslant N}c_k\int_{-\pi}^{\pi}e^{i\lambda k}Z(d\lambda)=\sum_{0<|k|\leqslant N}c_k\xi_k.$$

例 3 设 $f(\lambda)$ 是上面例 2 中的谱密度, 则通过简单的计算, 可得

$$\check{\xi}_0=\int_{-\pi}^{\pi}\frac{a}{1+|a|^2}(e^{i\lambda}+e^{-i\lambda})Z(d\lambda)=\frac{a}{1+|a|^2}(\xi_1+\xi_{-1}),$$

而内插的误差等于

$$\delta^2=\frac{1-|\alpha|^2}{1+|\alpha|^2}.$$

3. 过滤 设 $(\theta,\xi)=((\theta_n),(\xi_n)),n\in\mathbb{Z}$ 是部分观测序列, 其中 $\theta=(\theta_n)$ 是不被观测的分量, 而 $\xi=(\xi_n)$ 是被观测的分量. 假设 θ 和 ξ 是均值为 0 的 (弱) 平稳序列, 而

$$\theta_n=\int_{-\pi}^{\pi}e^{i\lambda n}Z_\theta(d\lambda),\quad \xi_n=\int_{-\pi}^{\pi}e^{i\lambda n}Z_\xi(d\lambda)$$

相应为 θ 和 ξ 的谱表示. 记

$$F_\theta(\Delta)=\mathbf{E}|Z_\theta(\Delta)|^2,\quad F_\xi(\Delta)=\mathbf{E}|Z_\xi(\Delta)|^2,$$

而

$$F_{\theta\xi}(\Delta)=\mathbf{E}Z_\theta(\Delta)\overline{Z_\xi(\Delta)}.$$

此外, 设 θ 和 ξ 是平稳相联系的, 即其协方差函数 $\operatorname{cov}(\theta_n,\xi_m)=\mathbf{E}\theta_n\overline{\xi}_m$ 只依赖于差 $n-m$, 记 $R_{\theta\xi}(n)=\mathbf{E}\theta_n\overline{\xi}_0$, 则

$$R_{\theta\xi}(n)=\int_{-\pi}^{\pi}e^{i\lambda n}F_{\theta\xi}(d\lambda).$$

所讨论的过滤问题, 在于根据对 ξ 各种不同的观测结果, 建立变量 θ_n 的均方最优线性估计 $\widehat{\theta}_n$.

如果假设是按 $\xi_m,m\in\mathbb{Z}$ 的所有值, 建立变量 θ_n 的估计 $\widehat{\theta}_n$, 则问题的解就特别简单. 事实上, 由于 $\widehat{\theta}_n=\widehat{\mathbf{E}}(\theta_n|H(\xi))$, 则存在这样的函数 $\widehat{\varphi}_n(\lambda)$, 使

$$\widehat{\theta}_n=\int_{-\pi}^{\pi}\widehat{\varphi}_n(\lambda)Z_\xi(d\lambda). \tag{25}$$

如同第 1,2 小节, "最优" 函数 $\widehat{\varphi}_n(\lambda)$ 应该满足的条件如下:

1) $\widehat{\varphi}_n(\lambda)\in H(F_\xi)$;

2) $(\theta_n-\widehat{\theta}_n)\perp H(\xi)$.

由最后一个条件, 可见对于任意 $m\in\mathbb{Z}$, 有

$$\int_{-\pi}^{\pi}e^{i\lambda(n-m)}F_{\theta\xi}(d\lambda)-\int_{-\pi}^{\pi}e^{-i\lambda m}\widehat{\varphi}_n(\lambda)F_\xi(d\lambda)=0. \tag{26}$$

因此, 如果假设函数 $F_{\theta\xi}(\lambda)$ 和 $F_\xi(\lambda)$ 有密度 $f_{\theta\xi}(\lambda)$ 和 $f_\xi(\lambda)$, 则由 (26) 式得

$$\int_{-\pi}^{\pi}e^{i\lambda(n-m)}[f_{\theta\xi}(\lambda)-e^{-i\lambda m}\widehat{\varphi}_n(\lambda)f_\xi(\lambda)]d\lambda=0.$$

如果 (按勒贝格测度几乎处处) $f_\xi(\lambda)>0$, 则由此立即得

$$\widehat{\varphi}_n(\lambda)=e^{i\lambda n}\widehat{\varphi}(\lambda), \tag{27}$$

其中

$$\widehat{\varphi}(\lambda)=f_{\theta\xi}(\lambda)\times f_\xi^{\oplus}(\lambda),$$

而 $f_\xi^{\oplus}(\lambda)$ 是 $f_\xi(\lambda)$ 的"广义逆", 即

$$f_\xi^{\oplus}(\lambda)=\begin{cases}[f_\xi(\lambda)]^{-1}, & f_\xi(\lambda)>0,\\ 0, & f_\xi(\lambda)=0.\end{cases}$$

这时, 过滤误差

$$\mathbf{E}|\theta_n-\widehat{\theta}_n|^2=\int_{-\pi}^{\pi}[f_\theta(\lambda)-f_{\theta\xi}^2(\lambda)f_\xi^{\oplus}(\lambda)]\mathrm{d}\lambda. \tag{28}$$

不难验证 $\widehat{\varphi}_n\in H(F_\xi)$, 从而估计 (25) 式是最优的, 其中函数 $\widehat{\varphi}_n(\lambda)$ 决定于 (27) 式.

例 4 从受到噪声干扰信号中提取信号. 设 $\xi_n=\theta_n+\eta_n$, 其中信号 $\theta=(\theta_n)$ 和噪声 $\eta=(\eta_n)$ 是不相关序列, 其谱密度相应为 $f_\theta(\lambda)$ 是 $f_\eta(\lambda)$. 那么

$$\widehat{\theta}_n=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}\widehat{\varphi}(\lambda)Z_\xi(\mathrm{d}\lambda),$$

其中

$$\widehat{\varphi}(\lambda)=f_\theta(\lambda)[f_\theta(\lambda)+f_\eta(\lambda)]^{\oplus},$$

而过滤误差

$$\mathbf{E}|\theta_n-\widehat{\theta}_n|^2=\int_{-\pi}^{\pi}[f_\theta(\lambda)f_\eta(\lambda)][f_\theta(\lambda)+f_\eta(\lambda)]^{\oplus}\mathrm{d}\lambda.$$

现在可以将上面得到的解 (25) 式, 用来由观测结果 $\xi_k(k\leqslant n)$ 建立变量 θ_{n+m} 的最优估计 $\widetilde{\theta}_{n+m}$, 其中 m 是 $\mathbb{Z}$ 中某给定的数. 假设 $\xi=(\xi_n)$ 是正则序列, 其谱密度为

$$f(\lambda)=\frac{1}{2\pi}|\Phi(\mathrm{e}^{-i\lambda})|^2,\quad 其中\quad \Phi(\mathrm{e}^{-i\lambda})=\sum_{k=0}^{\infty}a_k z^k.$$

根据沃尔德分解

$$\xi_n=\sum_{k=0}^{\infty}a_k\varepsilon_{n-k},$$

其中 $\varepsilon=(\varepsilon_n)$ 是白噪声, 其谱分解为

$$\varepsilon_n=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}Z_\varepsilon(\mathrm{d}\lambda).$$

因为

$$\widetilde{\theta}_{n+m}=\widehat{\mathbf{E}}[\theta_{n+m}|H_n(\xi)]=\widehat{\mathbf{E}}\left\{\widehat{\mathbf{E}}[\theta_{n+m}|H(\xi)]|H_n(\xi)\right\}=\widehat{\mathbf{E}}[\widehat{\theta}_{n+m}|H_n(\xi)],$$

而

$$\widehat{\theta}_{n+m}=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda(n+m)}\widehat{\varphi}(\lambda)\Phi(\mathrm{e}^{-i\lambda})Z_\varepsilon(\mathrm{d}\lambda)=\sum_{k=-\infty}^{\infty}\widehat{a}_{n+m-k}\varepsilon_k,$$

则

$$\widetilde{\theta}_{n+m}=\widehat{\mathbf{E}}\left[\sum_{k=-\infty}^{\infty}\widehat{a}_{n+m-k}\varepsilon_k|H_n(\xi)\right],$$

其中

$$\widehat{a}_k=\frac{1}{2\pi}\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda k}\widehat{\varphi}(\lambda)\Phi(\mathrm{e}^{-i\lambda})\mathrm{d}\lambda. \tag{29}$$

由于 $H_n(\xi)=H_n(\varepsilon)$, 可见

$$\begin{aligned}\widetilde{\theta}_{n+m}&=\sum_{k\leqslant n}\widehat{a}_{n+m-k}\varepsilon_k=\int_{-\pi}^{\pi}\left[\sum_{k\leqslant n}\widehat{a}_{n+m-k}\mathrm{e}^{i\lambda k}\right]Z_\varepsilon(\mathrm{d}\lambda)\\&=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}\left[\sum_{l=0}^{\infty}\widehat{a}_{l+m}\mathrm{e}^{-i\lambda l}\right]\Phi^{\oplus}(\mathrm{e}^{-i\lambda})Z_\xi(\mathrm{d}\lambda),\end{aligned}$$

其中 $\Phi^{\oplus}$ 是 Φ 的广义逆.

于是, 我们证明了如下定理.

定理 3 *如果被观测的 $\xi=(\xi_n)$ 是正则序列, 则由观测结果 $\xi_k(k\leqslant n)$ 对变量 θ_{n+m} (均方意义下) 的最优线性估计 $\widetilde{\theta}_{n+m}$, 由下面的公式表示:*

$$\widetilde{\theta}_{n+m}=\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda n}H_m(\mathrm{e}^{-i\lambda})Z_\xi(\mathrm{d}\lambda), \tag{30}$$

其中

$$H_m(\mathrm{e}^{-i\lambda})=\sum_{l=0}^{\infty}\widehat{a}_{l+m}\mathrm{e}^{-i\lambda l}\Phi^{\oplus}(\mathrm{e}^{-i\lambda}), \tag{31}$$

而系数 a_k 决定于 (29) 式.

4. 练习题

1. 在定理 1 中, 去掉假设: 级数 $\Phi(z)$ 的收敛半径为 $r>1$, 而且 $\Phi(z)$ 在区域 $\{z:|z|\leqslant 1\}$ 内无 0 点, 并且证明定理 1 仍然成立.

2. 证明, 对于正则过程, (4) 式中的函数 $\Phi(z)$ 可以表示为

$$\Phi(z)=\sqrt{2\pi}\exp\left\{\frac{1}{2}c_0+\sum_{k=1}^{\infty}c_kz^k\right\},\quad |z|<1,$$

其中

$$c_k=\frac{1}{2\pi}\int_{-\pi}^{\pi}\mathrm{e}^{i\lambda k}\ln f(\lambda)\mathrm{d}\lambda.$$

由此导出一步的预测误差 $\sigma_1^2=\mathbf{E}|\widehat{\xi}_1-\xi_1|^2$, 由塞格 — 柯尔莫戈洛夫 (G. Szegö – A. H. Колмогоров) 公式

$$\sigma_1^2=2\pi\exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln f(\lambda)\mathrm{d}\lambda\right\}$$

给出.

3. 不要求条件 (22) 成立, 证明定理 2.

4. 设不相关信号 θ 和噪声 η 的谱密度分别为:

$$f_\theta(\lambda)=\frac{1}{2\pi}\times\frac{1}{|1+b_1\mathrm{e}^{-i\lambda}|^2},\quad f_\eta(\lambda)=\frac{1}{2\pi}\times\frac{1}{|1+b_2\mathrm{e}^{-i\lambda}|^2}.$$

基于定理 3, 根据 $\xi_k(k\leqslant n)$ 的值求量 θ_{n+m} 的估计 $\widetilde{\theta}_{n+m}$, 其中 $\xi_k=\theta_k+\eta_k$. 对于谱密度

$$f_\theta(\lambda)=\frac{1}{2\pi}\times|2+\mathrm{e}^{-i\lambda}|^2,\quad f_\eta(\lambda)=\frac{1}{2\pi},$$

讨论同一问题.

§7. 卡尔曼 – 布西滤波器及其推广

1. 卡尔曼 – 布西概型, 卡尔曼 – 布西滤波器 从计算的角度, 上面给出的由对 ξ 的观测, 求解非观测分量 θ 的过滤问题并不适宜, 因为既然用谱的语言表述, 它对于自己的实现就要求进行*类似的*处理. 在卡尔曼 – 布西 (R. E. Kalman–R. S. Bucy) 提出的概型中, 最优滤波器的综合是用*递推的*方法实现的, 从而可能借助*数字*计算设备来实现. 决定卡尔曼 – 布西滤波器的广泛应用还有其他原因. 在*不假设* (θ,ξ) *是平稳序列*的情况下, 卡尔曼 – 布西滤波器仍然能 "运用" 就是其有广泛应用的原因之一.

我们在下面不仅讨论传统的卡尔曼 – 布西概型, 而且也研究其推广: 决定 (θ,ξ) 的递推方程中系数可能依赖于全部以往观测数据的情形.

这样, 假设 $(\theta,\xi)=((\theta_n),(\xi_n))$ 是*部分可观测序列*, 并且

$$\theta_n=(\theta_1(n),\cdots,\theta_k(n)),\quad \xi_n=(\xi_1(n),\cdots,\xi_l(n))$$

满足递推方程

$$\begin{aligned}\theta_{n+1}&=a_0(n,\xi)+a_1(n,\xi)\theta_n+b_1(n,\xi)\varepsilon_1(n+1)+b_2(n,\xi)\varepsilon_2(n+1),\\ \xi_{n+1}&=A_0(n,\xi)+A_1(n,\xi)\theta_n+B_1(n,\xi)\varepsilon_1(n+1)+B_2(n,\xi)\varepsilon_2(n+1),\end{aligned}\tag{1}$$

其中 $\varepsilon_1(n)=(\varepsilon_{11}(n),\cdots,\varepsilon_{1k}(n)),\varepsilon_2(n)=(\varepsilon_{21}(n),\cdots,\varepsilon_{2l}(n))$ 是具有独立分量的相互独立的高斯向量, 并且每一个分量服从参数 0 和 1 的高斯分布; $a_0(n,\xi)=(a_{01}(n,\xi),\cdots,a_{0k}(n,\xi))$ 和 $A_0(n,\xi)=(A_{01}(n,\xi),\cdots,A_{0l}(n,\xi))$ 是向量函数, 而且这些向量函数与 $\xi=(\xi_0,\xi_1,\cdots)$ 的独立性是 "不超前的", 也就是说, 对于任意固定的 $n,a_{01}(n,\xi),\cdots,A_{0l}(n,\xi)$, 仅依赖于$\xi_0,\cdots,\xi_n$. 矩阵函数

$$\begin{aligned}b_1(n,\xi)&=\|b_{ij}^{(1)}(n,\xi)\|,\quad &b_2(n,\xi)&=\|b_{ij}^{(2)}(n,\xi)\|,\\ B_1(n,\xi)&=\|B_{ij}^{(1)}(n,\xi)\|,\quad &B_2(n,\xi)&=\|B_{ij}^{(2)}(n,\xi)\|,\\ a_1(n,\xi)&=\|a_{ij}^{(1)}(n,\xi)\|,\quad &A_1(n,\xi)&=\|A_{ij}^{(1)}(n,\xi)\|\end{aligned}$$

的阶数相应地为 $k\times k,k\times l,l\times k,l\times l,k\times k,l\times k$, 并且亦 "不超前" 地依赖于 ξ. 此外, 假设向量 (θ_0,ξ_0) 不依赖于序列 $\varepsilon_1=(\varepsilon_1(n))$ 和 $\varepsilon_2=(\varepsilon_2(n))$.

为叙述简便, 以下将省略 "系数不依赖于 ξ" 的提示.

为使方程组 (1) 有解, 假设

$$\mathbf{E}(\|\theta_0\|^2+\|\xi_0\|^2)<\infty,\quad \left(\|x\|^2=\sum_{i=1}^k x_i^2, x=(x_1,\cdots,x_k)\right),$$

$|a_{ij}^{(1)}(n,\xi)|\leqslant C, A_{ij}^{(1)}(n,\xi)\leqslant C$, 而且如果 $g(n,\varepsilon)$ 是 $a_{0i},A_{0j},b_{ij}^{(1)},b_{ij}^{(2)},B_{ij}^{(1)},B_{ij}^{(2)}$ 任意函数, 则 $\mathbf{E}|g(n,\xi)|^2<\infty,n=1,2,\cdots$. 在这些条件下, 对于序列 (θ,ξ) 满足 $\mathbf{E}(\|\theta_n\|^2+\|\xi_n\|^2)<\infty,n\geqslant 1$.

其次, 设 $\mathscr{F}_n^\xi=\sigma\{\omega:\xi_0,\cdots,\xi_n\}$ 是随机变量 $\xi_0,\cdots,\xi_n$ 生成的最小 $\sigma-$ 代数, 并且

$$m_n=\mathbf{E}(\theta_n|\mathscr{F}_n^\xi),\quad \gamma_n=\mathbf{E}[(\theta_n-m_n)(\theta_n-m_n)^*|\mathscr{F}_n^\xi].$$

根据第二章 §8 定理 1, $m_n=(m_1(n),\cdots,m_k(n))$ 是向量 $\theta_n=(\theta_1(n),\cdots,\theta_k(n))$ 的均方最优估计量, 而 $\mathbf{E}\gamma_n=\mathbf{E}[(\theta_n-m_n)(\theta_n-m_n)^*]$ 是估计误差矩阵. 对于由方程 (1) 决定的任意序列 (θ,ξ), 求这些变量是相当困难的课题. 不过, 在如下关于 (θ_0,ξ_0) 一个补充条件下, 就可以导出 m_n 和 γ_n 递推方程组, 其中包含所谓*卡尔曼 — 布西滤波方程*. 关于 (θ_0,ξ_0) 的这个补充假设是: 条件分布 $\mathbf{P}\{\theta_0\leqslant a|\xi_0\}$ 是参数为高斯分布

$$\mathbf{P}\{\theta_0\leqslant a|\xi_0\}=\frac{1}{\sqrt{2\pi\gamma_0}}\int_{-\infty}^a \mathrm{e}^{-\frac{(x-m_0)^2}{2\gamma_0^2}}\mathrm{d}x,\tag{2}$$

其中 $m_0=m_0(\xi_0),\gamma_0=\gamma_0(\xi_0)$ 是分布参数.

首先证明一个辅助命题.

引理 1 *在上面引进的关于系数方程组* (1) *和条件* (2) *补充假设下, 序列* (θ,ξ) *服从条件高斯分布, 即条件分布函数*

$$\mathbf{P}\{\theta_0\leqslant a_0,\cdots,\theta_n\leqslant a_n|\mathscr{F}_n^\xi\}$$

以概率 1 是 n 维高斯向量的分布函数, 其分布的均值和协方差矩阵都依赖于 $\xi_0,\cdots,\xi_n$.

证明 我们仅限于证明分布 $\mathbf{P}\{\theta_n\leqslant a_n|\mathscr{F}_n^\xi\}$ 的高斯性, 因此只需导出 m_n 和 γ_n 的方程.

首先注意到, 由 (1) 式可见条件分布

$$\mathbf{P}\{\theta_{n+1}\leqslant a,\xi_{n+1}\leqslant x|\mathscr{F}_n^\xi,\theta_n=b\}$$

是高斯的, 其均值向量为

$$\mathbb{A}_0+\mathbb{A}_1 b=\begin{pmatrix}a_0+a_1 b\\ A_0+A_1 b\end{pmatrix},$$

而方差矩阵为

$$\mathbb{B} = \begin{pmatrix} b\circ b & b\circ B \\ (b\circ B)^* & (B\circ B) \end{pmatrix},$$

其中 $b\circ b = b_1b_1^* + b_2b_2^*, b\circ B = b_1B_1^* + b_2B_2^*, B\circ B = B_1B_1^* + B_2B_2^*$.

记 $\zeta_n = (\theta_n, \xi_n)$ 和 $t = (t_1, \cdots, t_{k+l})$, 则

$$\begin{aligned} &\mathbf{E}[\exp\{it^*\zeta_{n+1}\}|\mathscr{F}_n^\xi, \theta_n] \\ &= \exp\left\{it^*[\mathbb{A}_0(n,\xi) + \mathbb{A}_1(n,\xi)\theta_n] - \frac{1}{2}t^*\mathbb{B}(n,\xi)t\right\}. \end{aligned} \tag{3}$$

现在, 假设对于某各 $n \geqslant 0$, 引理的结论成立. 那么,

$$\begin{aligned} &\mathbf{E}[\exp\{it^*\mathbb{A}_1(n,\xi)\theta_n\}|\mathscr{F}_n^\xi] \\ &= \exp\left\{it^*\mathbb{A}_1(n,\xi)m_n - \frac{1}{2}t^*[\mathbb{A}_1(n,\xi)\gamma_n\mathbb{A}_1^*(n,\xi)]t\right\}. \end{aligned} \tag{4}$$

证明当将 n 换成 $n+1$ 时 (4) 式仍然成立.

由 (3) 式和 (4) 式, 有

$$\begin{aligned} &\mathbf{E}[\exp\{it^*\zeta_{n+1}\}|\mathscr{F}_n^\xi] \\ &= \exp\left\{it^*[\mathbb{A}_0(n,\xi) + \mathbb{A}_1(n,\xi)m_n] - \frac{1}{2}t^*\mathbb{B}(n,\xi)t - \frac{1}{2}t^*[\mathbb{A}_1(n,\xi)\gamma_n\mathbb{A}_1^*(n,\xi)]t\right\}. \end{aligned}$$

因此, 条件分布

$$\mathbf{P}\{\theta_{n+1} \leqslant a, \xi_{n+1} \leqslant x|\mathscr{F}_n^\xi\} \tag{5}$$

是高斯的.

像证明正态相关性定理 (第二章 §13 定理 2) 时一样, 验证存在矩阵 C, 使向量

$$\eta = [\theta_{n+1} - \mathbf{E}(\theta_{n+1}|\mathscr{F}_n^\xi)] - C[\xi_{n+1} - \mathbf{E}(\xi_{n+1}|\mathscr{F}_n^\xi)]$$

以概率 1 具有性质:

$$\mathbf{E}\left\{\eta[\xi_{n+1} - \mathbf{E}(\xi_{n+1}|\mathscr{F}_n^\xi)]^*|\mathscr{F}_n^\xi\right\} = 0.$$

由此可见, 在 $\mathscr{F}_n^\xi$ 的条件下, 条件高斯向量 η 和 ξ_{n+1} 相互独立, 即对于任意 $A \in \mathscr{B}(\mathbb{R}^k)$和 $B \in \mathscr{B}(\mathbb{R}^l)$, 以概率 1, 有

$$\mathbf{P}\{\eta \in A, \xi_{n+1} \in B|\mathscr{F}_n^\xi\} = \mathbf{P}\{\eta \in A|\mathscr{F}_n^\xi\} \times \mathbf{P}\{\xi_{n+1} \in B|\mathscr{F}_n^\xi\}.$$

因此, 对于任意 $s = (s_1, \cdots, s_k)$, 有

$$\begin{aligned} &\mathbf{E}[\exp(is^*\theta_{n+1})|\mathscr{F}_n^\xi, \xi_{n+1}] \\ &= \mathbf{E}\left\{\exp\left(is^*\left[\mathbf{E}(\theta_{n+1}|\mathscr{F}_n^\xi) + \eta + C[\xi_{n+1} - \mathbf{E}(\xi_{n+1}|\mathscr{F}_n^\xi)]\right]\right)\Big|\mathscr{F}_n^\xi, \xi_{n+1}\right\} \\ &= \exp\left\{is^*\left[\mathbf{E}(\theta_{n+1}|\mathscr{F}_n^\xi) + C[\xi_{n+1} - \mathbf{E}(\xi_{n+1}|\mathscr{F}_n^\xi)]\right]\right\}\mathbf{E}[\exp(is^*\eta)|\mathscr{F}_n^\xi, \xi_{n+1}] \\ &= \exp\left\{is^*\left[\mathbf{E}(\theta_{n+1}|\mathscr{F}_n^\xi) + C[\xi_{n+1} - \mathbf{E}(\xi_{n+1}|\mathscr{F}_n^\xi)]\right]\right\}\mathbf{E}[\exp(is^*\eta)|\mathscr{F}_n^\xi]. \end{aligned} \tag{6}$$

根据 (5) 式, 条件分布函数 $\mathbf{P}\{\eta \leqslant y|\mathscr{F}_n^\xi\}$是高斯的. 这连同 (6) 式, 就证明了条件分布 $\mathbf{P}\{\theta_{n+1} \leqslant a|\mathscr{F}_{n+1}^\xi\}$ 也是高斯的. □

定理 1 设 (θ, ξ) 是满足 (1) 式和 (2) 式的部分观测序列. 那么, (m_n, γ_n) 服从下列递推方程:

$$\begin{aligned} m_{n+1} &= [a_0 + a_1m_n] \\ &\quad + [b\circ B + a_1\gamma_nA_1^*][B\circ B + A_1\gamma_nA_1^*]^\oplus[\xi_{n+1} - A_0 - A_1m_n], \end{aligned} \tag{7}$$

$$\begin{aligned} \gamma_{n+1} &= [a_1\gamma_na_1^* + b\circ b] \\ &\quad - [b\circ B + a_1\gamma_nA_1^*][B\circ B + A_1\gamma_nA_1^*]^\oplus[b\circ B + a_1\gamma_nA_1^*]^*. \end{aligned} \tag{8}$$

证明 由 (1) 式, 有

$$\mathbf{E}(\theta_{n+1}|\mathscr{F}_n^\xi) = a_0 + a_1m_n, \quad \mathbf{E}(\xi_{n+1}|\mathscr{F}_n^\xi) = A_0 + A_1m_n \tag{9}$$

和

$$\begin{aligned} \theta_{n+1} - \mathbf{E}(\theta_{n+1}|\mathscr{F}_n^\xi) &= a_1(\theta_n - m_n) + b_1\varepsilon_1(n+1) + b_2\varepsilon_2(n+1), \\ \xi_{n+1} - \mathbf{E}(\xi_{n+1}|\mathscr{F}_n^\xi) &= A_1(\theta_n - m_n) + B_1\varepsilon_1(n+1) + B_2\varepsilon_2(n+1). \end{aligned} \tag{10}$$

引进记号:

$$\begin{aligned} d_{11} &= \mathrm{cov}(\theta_{n+1}, \theta_{n+1}|\mathscr{F}_n^\xi) \\ &= \mathbf{E}\{[\theta_{n+1} - \mathbf{E}(\theta_{n+1}|\mathscr{F}_n^\xi)][\theta_{n+1} - \mathbf{E}(\theta_{n+1}|\mathscr{F}_n^\xi)]^*|\mathscr{F}_n^\xi\}, \\ d_{12} &= \mathrm{cov}(\theta_{n+1}, \xi_{n+1}|\mathscr{F}_n^\xi) \\ &= \mathbf{E}\{[\theta_{n+1} - \mathbf{E}(\theta_{n+1}|\mathscr{F}_n^\xi)][\xi_{n+1} - \mathbf{E}(\xi_{n+1}|\mathscr{F}_n^\xi)]^*|\mathscr{F}_n^\xi\}, \\ d_{22} &= \mathrm{cov}(\xi_{n+1}, \xi_{n+1}|\mathscr{F}_n^\xi) \\ &= \mathbf{E}\{[\xi_{n+1} - \mathbf{E}(\xi_{n+1}|\mathscr{F}_n^\xi)][\xi_{n+1} - \mathbf{E}(\xi_{n+1}|\mathscr{F}_n^\xi)]^*|\mathscr{F}_n^\xi\}. \end{aligned}$$

那么, 由 (10) 式, 得

$$d_{11} = a_1\gamma_na_1^* + b\circ b, \quad d_{12} = a_1\gamma_nA_1^* + b\circ B, \quad d_{22} = A_1\gamma_nA_1^* + B\circ B. \tag{11}$$

由正态相关定理 (见第二章 §13 定理 2 和练习题 4), 有

$$m_{n+1} = \mathbf{E}(\theta_{n+1}|\mathscr{F}_n^\xi, \xi_{n+1}) = \mathbf{E}(\theta_{n+1}|\mathscr{F}_n^\xi) + d_{12}d_{22}^\oplus[\xi_{n+1} - \mathbf{E}(\xi_{n+1}|\mathscr{F}_n^\xi)]$$

和

$$\gamma_{n+1} = \mathrm{cov}(\theta_{n+1}, \theta_{n+1}|\mathscr{F}_n^\xi, \xi_{n+1}) = d_{11} - d_{12}d_{22}^\oplus d_{12}^*.$$

将 (9) 式中的 $\mathbf{E}(\theta_{n+1}|\mathscr{F}_n^\xi)$ 和 $\mathbf{E}(\xi_{n+1}|\mathscr{F}_n^\xi)$, 代入上面 m_{n+1} 的表达式, 将 (11) 式中的 d_{11}, d_{12}, d_{22}, 代入上面 γ_{n+1} 的表达式, 即可得到欲证明的递推方程 (7) 和 (8). □

系 1 如果在方程组 (1) 中, 所有系数 $a_0(n,\xi),\cdots,B_2(n,\xi)$ 都不依赖于 ξ, 则相应的概型称做卡尔曼 – 布西概型, 而关于 m_n 和 γ_n 的方程 (7) 和 (8) 称做卡尔曼 – 布西滤波器. 需要强调, 在这种情况下, 条件误差矩阵 γ_n 的等于无条件矩阵, 即

$$\gamma_n \equiv \mathbf{E}\gamma_n = \mathbf{E}[(\theta_n - m_n)(\theta_n - m_n)^*].$$

系 2 假设对于部分观测序列 (θ_n,ξ_n), θ_n 满足 (1) 式的第一个方程, 而 ξ_n 满足方程:

$$\begin{aligned}\xi_n = \widetilde{A}_0(n-1,\xi) &+\widetilde{A}_1(n-1,\xi)\theta_n\\ &+\widetilde{B}_1(n-1,\xi)\varepsilon_1(n)+\widetilde{B}_2(n-1,\xi)\varepsilon_2(n).\end{aligned}\tag{12}$$

那么, 显然

$$\begin{aligned}\xi_{n+1} = \widetilde{A}_0(n,\xi) &+ \widetilde{A}_1(n,\xi)[a_0(n,\xi)+a_1(n,\xi)\theta_n + b_1(n,\xi)\varepsilon_1(n+1)\\ &+ b_2(n,\xi)\varepsilon_2(n+1)] + \widetilde{B}_1(n,\xi)\varepsilon_1(n+1) + \widetilde{B}_2(n,\xi)\varepsilon_2(n+1).\end{aligned}$$

若记

$$\begin{aligned}A_0 &= \widetilde{A}_0 + \widetilde{A}_1 a_0, \quad A_1 = \widetilde{A}_1 a_1,\\ B_1 &= \widetilde{A}_1 b_1 + \widetilde{B}_1, \quad B_2 = \widetilde{A}_2 b_2 + \widetilde{B}_2,\end{aligned}$$

则所考虑的情形也属于概型 (1), 而且 m_n 和 γ_n 满足方程 (7) 和 (8).

2. 最优线性滤波器的结构 现在考虑线性概型 (对照 (1)式)

$$\begin{aligned}\theta_{n+1} &= a_0 + a_1\theta_n + a_2\xi_n + b_1\varepsilon_1(n+1) + b_2\varepsilon_2(n+1),\\ \xi_{n+1} &= A_0 + A_1\theta_n + A_2\xi_n + B_1\varepsilon_1(n+1) + B_2\varepsilon_2(n+1),\end{aligned}\tag{13}$$

其中所有系数 $a_0,\cdots,B_2$ 可能依赖于 n (但不依赖于 ξ), 而且 $\varepsilon_{ij}(n)$ 是 $\mathbf{E}\varepsilon_{ij}(n)=0$ 和 $\mathbf{E}\varepsilon_{ij}^2(n)=1$ 的独立高斯随机变量.

假设方程组 (13) 在初始条件 (θ_0,ξ_0) 下求解, 而且条件分布 $\mathbf{P}\{\theta_0 \leqslant a|\xi_0\}$ 是参数为 $m_0=\mathbf{E}(\theta_0|\xi_0)$ 和 $\gamma_0=\operatorname{cov}(\theta_0,\theta_0|\xi_0)=\mathbf{E}\gamma_0$ 的高斯分布. 那么, 由于正态相关定理以及 (7) 和 (8) 式, $m_n=\mathbf{E}(\theta_n|\mathscr{F}_n^{\xi})$ 的最优估计量是 $\xi_0,\xi_1,\cdots,\xi_n$ 的线性函数.

由以上讨论的结果, 在不要求高斯性的情形下, 可以证明关于 "线性滤波器结构" 的如下重要论断.

定理 2 设 $(\theta,\xi)=(\theta_n,\xi_n)_{n\geqslant 0}$ 是满足方程组 (13) 的部分被观测序列, 其中 $\varepsilon_{ij}(n)$ 是不相关随机变量, 且 $\mathbf{E}\varepsilon_{ij}(n)=0$ 和 $\mathbf{E}\varepsilon_{ij}^2(n)=1$, 而初值向量 (θ_0,ξ_0) 的分量具有有限二阶矩. 那么, $m_n=\mathbf{E}(\theta_n|\mathscr{F}_n^{\xi})$ 的最优线性估计量 $\widehat{m}_n=\widehat{\mathbf{E}}(\theta_n|\xi_0,\xi_1,\cdots,\xi_n)$ 满足方程 (7), 其中

$$a_0(n,\xi)=a_0(n)+a_2(n)\xi_n, \quad A_0(n,\xi)=A_0(n)+A_2(n)\xi_n,$$

而误差矩阵 $\widehat{\gamma}=\widehat{\mathbf{E}}[(\theta_n-\widehat{m}_n)(\theta_n-\widehat{m}_n)^*]$ 决定于方程 (8), 其初始条件为

$$\begin{aligned}\widehat{m}_0 &= \operatorname{cov}(\theta_0,\xi_0)\operatorname{cov}^{\oplus}(\xi_0,\xi_0)\xi_0,\\ \widehat{\gamma}_0 &= \operatorname{cov}(\theta_0,\theta_0) - \operatorname{cov}(\theta_0,\xi_0)\operatorname{cov}^{\oplus}(\xi_0,\xi_0)\operatorname{cov}^*(\theta_0,\xi_0).\end{aligned}\tag{14}$$

定理的证明需要下面的引理. 在建立最优线性估计时, 引理说明高斯性作用.

引理 2 设 (α,β) 是二维随机向量, 且 $\mathbf{E}(\alpha^2+\beta^2)<\infty$, 而 $(\widetilde{\alpha},\widetilde{\beta})$ 是与 (α,β) 具有相同一阶和二阶矩的二维高斯向量, 即

$$\mathbf{E}\widetilde{\alpha}^i=\mathbf{E}\alpha^i, \quad \mathbf{E}\widetilde{\beta}^i=\mathbf{E}\beta^i, \quad i=1,2, \quad \mathbf{E}\widetilde{\alpha}\widetilde{\beta}=\mathbf{E}\alpha\beta.$$

设 $\lambda(b)$ 是 b 的线性函数, 且

$$\lambda(b)=\mathbf{E}(\widetilde{\alpha}|\widetilde{\beta}=b).$$

那么, $\lambda(\beta)$ 是由 β 对 α 的均方最优线性估计, 即

$$\widehat{\mathbf{E}}(\alpha|\beta)=\lambda(\beta).$$

这时 $\mathbf{E}\lambda(\beta)=\mathbf{E}\alpha$.

证明 首先注意到, 由正态相关定理, 可见存在线性函数 $\lambda(b)$: $\lambda(b)=\mathbf{E}(\widetilde{\alpha}|\widetilde{\beta}=b)$. 其次, 假设 $\overline{\lambda}(b)$ 是另外一个线性估计, 那么

$$\mathbf{E}[\widetilde{\alpha}-\overline{\lambda}(\beta)]^2 \geqslant \mathbf{E}[\widetilde{\alpha}-\lambda(\widetilde{\beta})]^2,$$

而由于估计量 $\overline{\lambda}(b)$ 和 $\lambda(b)$ 线性的, 以及由引理的条件, 有

$$\mathbf{E}[\alpha-\overline{\lambda}(\widetilde{\beta})]^2=\mathbf{E}[\widetilde{\alpha}-\overline{\lambda}(\widetilde{\beta})]^2 \geqslant \mathbf{E}[\widetilde{\alpha}-\lambda(\widetilde{\beta})]^2=\mathbf{E}[\alpha-\lambda(\beta)]^2,$$

于是, 证明了 $\lambda(\beta)$ 在线性估计类中的最优性. 最后得

$$\mathbf{E}\lambda(\beta)=\mathbf{E}\lambda(\widetilde{\beta})=\mathbf{E}[\mathbf{E}(\widetilde{\alpha}|\widetilde{\beta})]=\mathbf{E}\widetilde{\alpha}=\mathbf{E}\alpha. \qquad \square$$

证明定理 2 与方程组 (13) 同时, 考虑方程组:

$$\begin{aligned}\widetilde{\theta}_{n+1} &= a_0 + a_1\widetilde{\theta}_n + a_2\widetilde{\xi}_n + b_1\widetilde{\varepsilon}_{11}(n+1) + b_2\widetilde{\varepsilon}_{12}(n+1),\\ \widetilde{\xi}_{n+1} &= A_0 + A_1\widetilde{\theta}_n + A_2\widetilde{\xi}_n + B_1\widetilde{\varepsilon}_{21}(n+1) + B_2\widetilde{\varepsilon}_{22}(n+1),\end{aligned}\tag{15}$$

其中 $\widetilde{\varepsilon}_{ij}(n)$ 是独立高斯随机变量, $\mathbf{E}\widetilde{\varepsilon}_{ij}(n)=0$ 和 $\mathbf{E}\widetilde{\varepsilon}_{ij}^2(n)=1$. 假设 $(\widetilde{\theta}_0,\widetilde{\xi}_0)$ 是高斯随机向量, 有与 (θ_0,ξ_0)相同的一阶矩和协方差, 而且不依赖于 $\widetilde{\varepsilon}_{ij}(n)$. 那么, 由于方程组 (15) 是线性的, 向量 $(\widetilde{\theta}_0,\cdots,\widetilde{\theta}_n,\widetilde{\xi}_0,\cdots,\widetilde{\xi}_n)$ 是高斯的, 即由引理 2 (确切一点说, 由定理 2 的明显的类似), 以及正态相关定理, 可以得到定理 2 的结论. $\square$

3. 例 下面是应用定理 1 和定理 2 的例子.

例 1 设 $\theta=(\theta_n)$ 和 $\eta=(\eta_n)$ 是两个 (弱) 平稳不相关随机序列, $\mathbf{E}\theta_n=\mathbf{E}\eta_n=0$, 而其谱密度相应为

$$f_\theta(\lambda)=\frac{1}{2\pi}\frac{1}{|1+b_1\mathrm{e}^{-i\lambda}|^2},\quad f_\eta(\lambda)=\frac{1}{2\pi}\frac{1}{|1+b_2\mathrm{e}^{-i\lambda}|^2},$$

其中 $|b_1|<1, |b_2|<1$.

下面将把 θ 视为有用信号, 而 η 视为噪声, 并且假设对序列 $\xi=(\xi_n)$ 进行观测:

$$\xi_n=\theta_n+\eta_n.$$

根据 §3 定理 3 的系 2, 存在 (互不相关的) 白噪声 $\varepsilon_1=(\varepsilon_1(n))$ 和 $\varepsilon_2=(\varepsilon_2(n))$, 使

$$\theta_{n+1}+b_1\theta_n=\varepsilon_1(n+1),\quad \eta_{n+1}+b_2\eta_n=\varepsilon_2(n+1).$$

那么,

$$\begin{aligned}\xi_{n+1}&=\theta_{n+1}+\eta_{n+1}=-b_1\theta_n-b_2\eta_n+\varepsilon_1(n+1)+\varepsilon_2(n+1)\\&=-b_2(\theta_n+\eta_n)-\theta_n(b_1-b_2)+\varepsilon_1(n+1)+\varepsilon_2(n+1)\\&=-b_2\xi_n-(b_1-b_2)\theta_n+\varepsilon_1(n+1)+\varepsilon_2(n+1).\end{aligned}$$

这样, 对于 θ 和 ξ 有递推公式

$$\begin{aligned}\theta_{n+1}&=-b_1\theta_n+\varepsilon_1(n+1),\\\xi_{n+1}&=-(b_1-b_2)\theta_n-b_2\xi_n+\varepsilon_1(n+1)+\varepsilon_2(n+1).\end{aligned}\tag{16}$$

而根据定理 2, $\widehat{m}_n=\widehat{\mathbf{E}}(\theta_n|\xi_0,\xi_1,\cdots,\xi_n)$ 和 $\widehat{\gamma}=\widehat{\mathbf{E}}(\theta_n-\widehat{m}_n)^2$ 满足如下最优线性滤波器的递推方程组:

$$\begin{aligned}m_{n+1}&=-b_1m_n+\frac{b_1(b_1-b_2)\gamma_n}{2+(b_1-b_2)^2\gamma_n}[\xi_{n+1}+(b_1-b_2)m_n+b_2\xi_n],\\\gamma_{n+1}&=b_1^2\gamma_n+1-\frac{[1+b_1(b_1-b_2)\gamma_n]^2}{2+(b_1-b_2)^2\gamma_n}.\end{aligned}\tag{17}$$

现在求为解该方程组所需要的初始条件 m_0 和 γ_0. 记 $d_{11}=\mathbf{E}\theta_n^2, d_{12}=\mathbf{E}\theta_n\xi_n$, $d_{22}=\mathbf{E}\xi_n^2$, 那么由 (16) 式, 可见

$$\begin{aligned}d_{11}&=b_1^2d_{11}+1,\\d_{12}&=b_1(b_1-b_2)d_{11}+b_1b_2d_{12}+1,\\d_{22}&=(b_1-b_2)^2d_{11}+b_2^2d_{22}+2b_2(b_1-b_2)d_{12}+2,\end{aligned}$$

因此,

$$d_{11}=\frac{1}{1-b_1^2},\quad d_{12}=\frac{1}{1-b_1^2},\quad d_{22}=\frac{2-b_1^2-b_2^2}{(1-b_1^2)(1-b_2^2)},$$

从而, 由 (14) 式得如下初始数据的值:

$$\begin{aligned}m_0&=\frac{d_{12}}{d_{22}}\xi_0=\frac{1-b_2^2}{2-b_1^2-b_2^2}\xi_0,\\\gamma_0&=d_{11}-\frac{d_{12}^2}{d_{22}}=\frac{1}{1-b_1^2}-\frac{1-b_2^2}{(1-b_1^2)(2-b_1^2-b_2^2)}=\frac{1}{2-b_1^2-b_2^2}.\end{aligned}\tag{18}$$

这样, 由 $\xi_0,\xi_1,\cdots,\xi_n$ 对信号 θ_n 的最优 (均方) 线性估计 m_n 和均方误差 γ_n, 决定于递推方程组 (17), 并且对于初始条件 (18) 求解. 需要指出, γ_n 的方程不含随机分量, 从而, 求 m_n 的数值所必须的 γ_n 的值可以事先 (在解该滤波之前) 计算出来.

例 2 这个例子从下面的角度可以借鉴: 它说明定理 2 的结果, 在序列 (θ,ξ) 服从不同于方程组 (13) 的 (非线性) 方程组时, 如何用于寻找最优线性滤波器的问题.

设 $\varepsilon_1=(\varepsilon_1(n))$ 和 $\varepsilon_2=(\varepsilon_2(n))$ 是两个独立高斯随机序列, 由独立随机变量构成, 且 $\mathbf{E}\varepsilon_i(n)=0, \mathbf{E}\varepsilon_i^2(n)=1, n\geqslant 1$. 考虑随机序列偶 $(\theta,\xi)=(\theta_n,\xi_n), n\geqslant 0$, 其中

$$\begin{aligned}\theta_{n+1}&=a\theta_n+(1+\theta_n)\varepsilon_1(n+1),\\\xi_{n+1}&=A\theta_n+\varepsilon_2(n+1).\end{aligned}\tag{19}$$

假设 θ_0 不依赖于 $(\varepsilon_1,\varepsilon_2)$, 且 $\theta_0\sim N(m_0,\gamma_0)$.

方程组 (19) 是非线性的, 故不能直接运用定理 2. 然而, 如果设

$$\widetilde{\varepsilon}_1(n+1)=\frac{1+\theta_n}{\sqrt{\mathbf{E}(1+\theta_n)^2}}\varepsilon_1(n+1),$$

则易见 $\mathbf{E}\widetilde{\varepsilon}_1(n)=0, \mathbf{E}\widetilde{\varepsilon}_1(n)\widetilde{\varepsilon}_1(m)=0(m\neq n), \mathbf{E}\widetilde{\varepsilon}_1^2(n)=1$. 所以与 (19) 式同时, 原序列 (θ,ξ) 也满足线性方程组

$$\begin{aligned}\theta_{n+1}&=a_1\theta_n+b_1(n)\widetilde{\varepsilon}_1(n+1),\\\xi_{n+1}&=A\theta_n+\varepsilon_2(n+1),\end{aligned}\tag{20}$$

其中 $b_1(n)=\sqrt{\mathbf{E}(1+\theta_n)^2}$, 而 $\{\widetilde{\varepsilon}_1(n)\}$ 是某一两两不相关随机变量序列.

方程组 (20) 是形如 (13) 的线性方程组, 即根据定理 2, 最优线性估计量 $\widehat{m}_n=\widehat{\mathbf{E}}(\theta_n|\xi_0,\xi_1,\cdots,\xi_n)$ 及其误差 $\widehat{\gamma}$ 可以由方程组 (7), (8) 决定, 在这种情形下方程组 (7), (8) 有如下形式:

$$\begin{aligned}\widehat{m}_{n+1}&=a_1\widehat{m}_n+\frac{a_1A_1\widehat{\gamma}_n}{1+A_1^2\widehat{\gamma}_n}[\xi_{n+1}-A_1\widehat{m}_n],\\\widehat{\gamma}_{n+1}&=[a_1^2\widehat{\gamma}_n+b_1^2(n)]-\frac{(a_1A_1\widehat{\gamma}_n)^2}{1+A_1^2\widehat{\gamma}_n},\end{aligned}$$

其中 $b_1(n)=\sqrt{\mathbf{E}(1+\theta_n)^2}$ 应由方程组 (19) 的第一个方程来求.

例 3 参数估计 设 $\theta = (\theta_1, \cdots, \theta_k)$ 是高斯向量, 且 $\mathbf{E}\theta = m$ 和 $\operatorname{cov}(\theta, \theta) = \gamma$. 假设 $\xi = (\xi_n), n \geqslant 0$, 是 l 维随机序列, 且

$$\xi_{n+1} = A_0(n, \xi) + A_1(n, \xi)\theta + B_1(n, \xi)\varepsilon_1(n+1), \quad \xi_0 = 0, \tag{21}$$

其中 ε_1 的含义与方程组 (1) 中的含义相同. 现在对于已知的 m 和 γ, 欲根据对 ξ 的观测结果, 求 θ 的最优估计.

那么, 对于 $m_n = \mathbf{E}(\theta|\mathscr{F}_n^\xi)$ 和 γ_n, 由 (7), (8) 式, 有

$$\begin{aligned} m_{n+1} = m_n + \gamma_n A_1^*(n,\xi)[(B_1B_1^*)(n,\xi) + A_1(n,\xi)\gamma_n A_1^*(n,\xi)]^{\oplus} \\ \times[\xi_{n+1} - A_0(n,\xi) - A_1(n,\xi)m_n], \end{aligned} \tag{22}$$
$$\gamma_{n+1} = \gamma_n - \gamma_n A_1^*(n,\xi)[(B_1B_1^*)(n,\xi) + A_1(n,\xi)\gamma_n A_1^*(n,\xi)]^{\oplus} A_1(n,\xi)\gamma_n.$$

如果 $B_1B_1^*$ 是非退化矩阵, 则方程组 (22) 的解为

$$\begin{aligned} m_{n+1} &= \left[E + \gamma\sum_{i=0}^n A_1^*(i,\xi)(B_1B_1^*)^{-1}(i,\xi)A_1^*(i,\xi)\right]^{-1} \times \\ &\quad \left[m + \gamma\sum_{i=0}^n A_1^*(i,\xi)(B_1B_1^*)^{-1}(i,\xi)(\xi_{i+1} - A_0(m,\xi))\right], \\ \gamma_{n+1} &= \left[E + \gamma\sum_{i=0}^n A_1^*(i,\xi)(B_1B_1^*)^{-1}(i,\xi)A_1(i,\xi)\right]^{-1}\gamma, \end{aligned} \tag{23}$$

其中 E 是单位矩阵.

4. 练习题

1. 证明, 对于概型 (1), 向量 m_n 和 $\theta_n - m_n$ 不相关:

$$\mathbf{E}m_n^*(\theta_n - m_n) = 0.$$

2. 设概型 (1) 中的 γ_0 和所有系数 (或许除系数 $a_0(n,\xi), A_0(n,\xi)$ 外) 都与 "随机性" 无关 (即不依赖于 ξ). 证明条件协方差 γ_n 也与 "随机性" 无关: $\gamma_n = \mathbf{E}\gamma_n$.

3. 证明方程组 (22) 的解由公式 (23) 表示.

4. 设 $(\theta, \xi) = (\theta_n, \xi_n)$ 是高斯序列, 满足概型 (1) 的如下特别形式:

$$\theta_{n+1} = a\theta_n + b\varepsilon_1(n+1), \quad \xi_{n+1} = A\theta_n + B\varepsilon_2(n+1).$$

证明, 如果 $A \neq 0, b \neq 0, B \neq 0$, 则极限过滤误差 $\gamma = \lim\limits_{n\to\infty}\gamma_n$ 存在, 且为方程

$$\gamma^2 + \left[\frac{B^2(1-a^2)}{A^2} - b^2\right]\gamma - \frac{b^2B^2}{A^2} = 0$$

的正根.

5. (内插; [41,13.3]). 设 (θ, ξ) 是部分可观测序列, 且服从递推关系式 (1) 和 (2). 设向量 θ_m 的条件分布

$$\pi_a(m,m) = \mathbf{P}\{\theta_m \leqslant a|F_m^\xi\}$$

是正态的.

(a) 证明, 对于 $n \geqslant m$, 条件分布

$$\pi_a(m,n) = \mathbf{P}\{\theta_m \leqslant a|\mathscr{F}_n^\xi\}$$

也是正态的, $\pi_a(m,n) \sim N(\mu(m,n), \gamma(m,n))$.

(b) 求随机变量 θ_m (关于 $\mathscr{F}_n^\xi$) 的内插估计 $\mu(m,n)$, 和矩阵 $\gamma(m,n)$.

6. (外推; [41,13.4]). 设关系式 (1) 和 (2) 中的变量为:

$$a_0(n,\xi) = a_0(n) + a_2(n)\xi_n, \quad a_1(n,\xi) = a_1(n),$$
$$A_0(n,\xi) = A_0(n) + A_2(n)\xi_n, \quad A_1(n,\xi) = A_1(n).$$

(a) 证明, 在这种情形下 $(n \geqslant m)$, 分布 $\pi_{a,b}(m,n) = \mathbf{P}\{\theta_m \leqslant a, \xi_n \leqslant b|\mathscr{F}_n^\xi\}$ 是正态的.

(b) 求外推估计量

$$\mathbf{E}(\theta_m|\mathscr{F}_m^\xi) \quad 和 \quad \mathbf{E}(\xi_n|\mathscr{F}_m^\xi).$$

7. (最优控制; [41,14.3]). 考虑 "被控制" 部分观测系统 $(\theta_n, \xi_n)_{0\leqslant n\leqslant N}$, 其中

$$\theta_{n+1} = u_n + \theta_n + b\varepsilon_1(n+1),$$
$$\xi_{n+1} = \theta_n + \varepsilon_2(n+1).$$

这里, "控制" u_n 为 $\mathscr{F}_n^\xi$– 可测, 并且对于一切 $0 \leqslant n \leqslant N-1, \mathbf{E}u_n^2 < \infty$. 随机变量 $\varepsilon_1(n)$ 和 $\varepsilon_2(n), n = 1, \cdots, N$ 的含义与方程组 (1) 及等式 (2) 相同; $\xi_0 = 0, \theta_0 \sim N(m, \gamma)$.

我们称 "控制" $u^* = (u_0^*, \cdots, u_{N-1}^*)$ 是最优的, 如果

$$V(u^*) = \sup_u V(u), \quad 其中 \quad V(u) = \mathbf{E}\left[\sum_{n=0}^{N-1}(\theta_n^2 + u_n^2) + \theta_N^2\right].$$

证明:

$$u_n^* = -[1 + P_{n+1}]^+ P_{n+1} m_n^*, \quad n = 0, \cdots, N-1,$$

其中

$$a^+ = \begin{cases} a^{-1}, & a \neq 0, \\ 0, & a = 0; \end{cases}$$

且 $(P_n)_{0\leqslant n\leqslant N}$ 决定于如下递推关系式:

$$P_n = 1 + P_{n+1} - P_{n+1}^2[1 + P_{n+1}]^+, \quad P_N = 1,$$

而 (m_n^*) 决定于关系式

$$m_{n+1}^* = u_n^* + \gamma_n^*(1+\gamma_n^*)^+(\xi_{n+1} - m_n^*), \quad 0 \leqslant n \leqslant N-1,$$

其中 $m_0^* = m$, 而 $\gamma_0^* = \gamma$,

$$\gamma_{n+1}^* = \gamma_n^* + 1 - (\gamma_n^*)^2(1+\gamma_n^*)^+, \quad 0 \leqslant n \leqslant N-1.$$

第七章　构成鞅的随机变量序列

[①]按 "全国科学技术名词审定委员会" 审定公布定名: submartingale 为 "下鞅". —— 译者

鞅论很好地描绘数学概率论形成的历史 —— 鞅论的基本概念, 是受博弈实践的启发引进的, 但是后来这些概念成为现代数学非常精细的工具之一……

J. L. 杜布 (J. L. Doob)《什么是鞅?》[127]

§1. 鞅和相关概念的定义

1. 这一节的研究对象 —— 鞅　在概率论中, 用各种不同的方法研究随机变量之间的相依性. 在 (弱) 随机序列的理论中, 基本特征是*协方差函数*, 而且该理论的一切结论完全决定于协方差函数的性质. 在马尔可夫链的理论中 (第一章 §12 和第八章), 基本特征是*转移函数*, 它决定由马尔可夫相依性联系的随机变量的演变.

在这一章 (亦见第一章 §11) 将引进相当广泛的随机变量序列类 (鞅及其推广), 而关于其相依性的研究, 利用基于*条件数学期望*性质的研究方法.

2. 鞅 (下鞅) 的概念　假设给定一带*过滤* (流) 的概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$, 所谓过滤 (流) 即 σ - 代数族 $(\mathscr{F}_n)$, 其中 $\mathscr{F}_n, n \geqslant 0$, 而且 $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \cdots \subseteq \mathscr{F}$ ("过滤概率空间").

设 $X_0, X_1, \cdots$ 是概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 上的随机变量. 如果对于每一个 $n \geqslant 0$, 变量 X_n 是 $\mathscr{F}_n$ - *可测的*, 则组合 $X = (X_n, \mathscr{F}_n)_{n \geqslant 0}$ 或简记 $X = (X_n, \mathscr{F}_n)$, 称之为*随机序列*.

假如随机序列 $X = (X_n, \mathscr{F}_n)$ 还具有性质: 每一个 $n \geqslant 1$, 变量 X_n 是 $\mathscr{F}_{n-1}$ - *可测的*, 则记作 $X = (X_n, \mathscr{F}_{n-1})$, 并且设 $\mathscr{F}_{-1} = \mathscr{F}_0$. 这时称 X 为*可预测随机序列*. 序列 $(X_n)_{n \geqslant 0}$ 称做递增的, 如果 $X_0 = 0$ 且 $X_n \leqslant X_{n+1} (\mathbf{P}$ - a.c.).

定义 1　随机序列 $X = (X_n, \mathscr{F}_n)$ 称做*鞅* (*下鞅*), 如果对于一切 $n \geqslant 0$, 有,

$$\mathbf{E}|X_n| < \infty, \tag{1}$$

$$\mathbf{E}(X_{n+1}|\mathscr{F}_n) \underset{(\geqslant)}{=} X_n \quad (\mathbf{P} - \text{a.c.}). \tag{2}$$

随机序列 $X = (X_n, \mathscr{F}_n)$ 称做*上鞅*, 如果序列 $-X = (-X_n, \mathscr{F}_n)$ 是*下鞅*.

在特殊情形下, 假如 $\mathscr{F}_n = \mathscr{F}_n^X$, 其中 $\mathscr{F}_n^X = \sigma(X_0, \cdots, X_n)$, 并且随机序列 $X = (X_n, \mathscr{F}_n)$ 是鞅 (下鞅), 则称序列 $(X_n)_{n \geqslant 0}$ 本身是鞅 (下鞅).

由条件数学期望的性质, 容易证明, 条件 (2) 等价于: 对于任何 $n \geqslant 0$ 和 $A \in \mathscr{F}_n$, 有

$$\int_A X_{n+1} \mathrm{d}\mathbf{P} \underset{(\geqslant)}{=} \int_A X_n \mathrm{d}\mathbf{P}. \tag{3}$$

例 1 设 $(\xi_n)_{n\geqslant 0}$ 是独立随机变量序列, $\mathbf{E}|\xi_n|<\infty, \mathbf{E}\xi_n=0, X_n=\xi_0+\cdots+\xi_n, \mathscr{F}_n=\sigma(\xi_0,\cdots,\xi_n)$, 则随机序列 $X=(X_n,\mathscr{F}_n)$ 是鞅.

例 2 设 $(\xi_n)_{n\geqslant 0}$ 是独立随机变量序列, $\mathbf{E}\xi_n=1$, 则随机序列 $X=(X_n,\mathscr{F}_n)$ 也是鞅, 其中

$$X_n=\prod_{k=0}^{n}\xi_k, \quad \mathscr{F}_n=\sigma(\xi_0,\cdots,\xi_n).$$

例 3 设 ξ 是随机变量序列, $\mathbf{E}|\xi|<\infty$, 而 $\mathscr{F}_0\subseteq\mathscr{F}_1\subseteq\cdots\subseteq\mathscr{F}$. 则随机序列 $X=(X_n,\mathscr{F}_n)$ 是鞅, 其中 $X_n=(\xi|\mathscr{F}_n)$ (称做列维 (P. P. Lévy) 鞅).

例 4 如果 $(\xi_n)_{n\geqslant 0}$ 是非负可积随机变量序列, 则序列 (X_n) 是鞅, 其中 $X_n=\xi_0+\cdots+\xi_n$.

例 5 设 $X=(X_n,\mathscr{F}_n)$ 是鞅, 而 $g(x)$ 是凹 (向下凸) 函数, 且 $\mathbf{E}|g(X_n)|<\infty, n\geqslant 0$, 则 (由第二章 §6 的延森 (I. L. Iensen) 不等式可见) 随机序列 $(g(X_n),\mathscr{F}_n)$ 是下鞅.

假如 $X=(X_n,\mathscr{F}_n)$ 是下鞅, 而 $g(x)$ 是凹函数, 且对于一切, $n\geqslant 0, \mathbf{E}|g(X_n)|<\infty$, 则随机序列 $(g(X_n),\mathscr{F}_n)$ 也是下鞅.

在定义 1 中所作的假设 $\mathbf{E}|X_n|<\infty$, 保障条件数学期望 $\mathbf{E}(X_{n+1}|\mathscr{F}_n), n\geqslant 0$, 的存在性. 不过, 在不要求 $\mathbf{E}|X_{n+1}|<\infty$ 的情况下, 这些条件数学期望也可能存在. 我们知道根据第二章 §7, $\mathbf{E}(X_{n+1}^+|\mathscr{F}_n)$ 和 $\mathbf{E}(X_{n+1}^-|\mathscr{F}_n)$ 总有定义, 如果 (若 $\mathbf{P}(A\triangle B)=0$, 则以 $A=B$ ($\mathbf{P}$ – a.c.) 表示)

$$\{\omega:\mathbf{E}(X_{n+1}^+|\mathscr{F}_n)<\infty\}\bigcup\{\omega:\mathbf{E}(X_{n+1}^-|\mathscr{F}_n)<\infty\}=\Omega \quad (\mathbf{P}-\text{a.c.}),$$

那么称 $\mathbf{E}(X_{n+1}|\mathscr{F}_n)$ 有定义, 且根据定义设

$$\mathbf{E}(X_{n+1}|\mathscr{F}_n)=\mathbf{E}(X_{n+1}^+|\mathscr{F}_n)-\mathbf{E}(X_{n+1}^-|\mathscr{F}_n).$$

由此可见, 下面的定义是自然的.

定义 2 随机序列 $X=(X_n,\mathscr{F}_n)$ 称做广义鞅 (下鞅), 如果对于一切 $n\geqslant 0$, 有 $\mathbf{E}|X_0|<\infty$, 而且 $\mathbf{E}(X_{n+1}|\mathscr{F}_n)$ 有定义, 并满足条件 (2).

注意, 由该定义可见, 对于广义下鞅 $\mathbf{E}(X_{n+1}^-|\mathscr{F}_n)<\infty$, 而对于广义鞅 $\mathbf{E}(|X_{n+1}||\mathscr{F}_n)<\infty$ ($\mathbf{P}$-a.c.).

3. 马尔可夫时间 下面的定义引进的马尔可夫时间的概念, 在以后介绍的全部理论中有十分重要的作用.

定义 3 在集合 $\{0,1,\cdots,+\infty\}$ 中取值的随机变量 $\tau=\tau(\omega)$, 称做 (关于 σ – 代数族 $(\mathscr{F}_n)$ 的) 马尔可夫时间或 "不依赖于将来的随机变量", 如果对于每一个 $n\geqslant 0$,

$$\{\tau=n\}\in\mathscr{F}_n. \tag{4}$$

对于 $\mathbf{P}\{\tau<\infty\}=1$ 的情形, 马尔可夫时间 τ, 称做停止时间.

设 $X=(X_n,\mathscr{F}_n)$ 是一随机序列, 而 τ (关于 σ – 代数族 $(\mathscr{F}_n)$) 是马尔可夫时间. 记

$$X_\tau(\omega)=\sum_{n=0}^{\infty}X_n(\omega)I_{\{\tau=n\}}(\omega)$$

(于是 $X_\infty=0$, 即在集合 $\{\omega:\tau=\infty\}$ 上 $X_\tau=0$).

那么, 对于每一个 $B\in\mathscr{B}(\mathbb{R})$,

$$\{\omega:X_\tau\in B\}=\{\omega:X_\infty\in B,\tau=\infty\}+\sum_{n=0}^{\infty}\{\omega:X_n\in B,\tau=n\}\in\mathscr{F},$$

从而 $X_\tau=X_{\tau(\omega)}(\omega)$ 是随机变量.

例 6 设 $X=(X_n,\mathscr{F}_n)$ 是一随机序列, 而 $B\in\mathscr{B}(\mathbb{R})$. 那么, (首次到达集合 B 中的) 时间

$$\tau_B=\inf\{n\geqslant 0:X_n\in B\}$$

(当 $\{\cdot\}=\varnothing$ 时 $\tau_B=+\infty$) 是马尔可夫时间, 因为对于任何 $n\geqslant 0$, 有

$$\{\tau_B=n\}=\{X_0\notin B,\cdots,X_{n-1}\notin B,X_n\in B\}\in\mathscr{F}_n.$$

例 7 设 $X=(X_n,\mathscr{F}_n)$ 是鞅 (下鞅), 而 τ (关于 σ – 代数族 $(\mathscr{F}_n)$) 是马尔可夫时间. 那么, "停止" 序列 $X^\tau=(X_{n\wedge\tau},\mathscr{F}_n)$ 也构成鞅 (下鞅).

事实上, 由关系式

$$X_{n\wedge\tau}=\sum_{m=0}^{n-1}X_mI_{\{\tau=m\}}=X_nI_{\{\tau\geqslant n\}}$$

可见, 随机变量 $X_{n\wedge\tau}$ 为 $\mathscr{F}_n$ – 可测的和可积的, 且

$$X_{(n+1)\wedge\tau}-X_{n\wedge\tau}=I_{\{\tau>n\}}(X_{n+1}-X_n),$$

由此可见

$$\mathbf{E}[X_{(n+1)\wedge\tau}-X_{n\wedge\tau}|\mathscr{F}_n]=I_{\{\tau>n\}}\mathbf{E}[X_{n+1}-X_n|\mathscr{F}_n]\underset{(\geqslant)}{=}0.$$

与每一个 σ – 代数族 $(\mathscr{F}_n)$, 及关于 $(\mathscr{F}_n)$ 的马尔可夫时间 τ, 可以与一集系

$$\mathscr{F}_\tau=\{A\in\mathscr{F}:A\bigcap\{\tau=n\}\in\mathscr{F}_n, \text{ 对于一切 } n\geqslant 0\}$$

相联系. 显然 $\Omega\in\mathscr{F}_\tau$ 和 $\mathscr{F}_\tau$ 关于求可数次并运算封闭. 此外, 如果 $A\in\mathscr{F}_\tau$, 则

$$\overline{A}\bigcap\{\tau=n\}=\{\tau=n\}\backslash(A\bigcap\{\tau=n\})\in\mathscr{F}_n,$$

故 $\overline{A}\in\mathscr{F}_\tau$. 由此可见 $\mathscr{F}_\tau$ 是 σ – 代数.

如果把 $\mathscr{F}_n$ 视为在时刻 n (包括 n) 之前观测到的事件的全体, 那么 $\mathscr{F}_\tau$ 可以视为在 "随机" 时间 τ 观测到的事件的全体.

不难证明 (练习题 3), 随机变量 τ 和 X_τ 是 $\mathscr{F}_\tau$ – 可测的.

4. 局部鞅和鞅变换

定义 4 称随机序列 $X=(X_n,\mathscr{F}_n)$ 为*局部鞅 (局部下鞅)*, 如果存在 (局部化的) 有限马尔可夫时间序列 $(\tau_k)_{k\geqslant 1}:\tau_k\leqslant\tau_{k+1}(\mathbf{P}-\text{a.c.})$, $\tau_k\uparrow\infty(\mathbf{P}-\text{a.c.})$, $k\to\infty$, 且每一个 ("停止") 序列 $X^{\tau_k}=(X_{\tau_k\wedge n}I_{\{\tau_k>0\}},\mathscr{F}_n)$ 是鞅 (下鞅).

下面在定理 1 中证明, 实际上局部鞅类与广义鞅类等同. 此外, 每一局部鞅可以借助于所谓鞅变换, 由某一鞅和某一可预测序列得到.

定义 5 设 $Y=(Y_n,\mathscr{F}_n)_{n\geqslant 0}$ 是随机序列, 而 $V=(V_n,\mathscr{F}_n)_{n\geqslant 0}$ 是可预测序列 ($\mathscr{F}_{-1}=\mathscr{F}_0$). 随机序列 $V\cdot Y=((V\cdot Y)_n,\mathscr{F}_n)$, 其中

$$(V\cdot Y)_n=V_0Y_0+\sum_{i=1}^{n}V_i\Delta Y_i,\tag{5}$$

而 $\Delta Y_i=Y_i-Y_{i-1}$, 称做 *Y 由 V 的变换*. 此外, 如果 Y 是鞅 (或局部鞅), 则称 $V\cdot Y$ 为*鞅变换*.

定理 1 *设 $X=(X_n,\mathscr{F}_n)_{n\geqslant 0}$ 是随机序列, 其中 $X_0=0(\mathbf{P}-\text{a.c.})$. 下列各条件等价:*

a) *X 是局部鞅;*

b) *X 是广义鞅;*

c) *X 是鞅变换, 即存在可预测序列 $V=(V_n,\mathscr{F}_n)_{n\geqslant 0}(V_0=0)$, 和鞅 $Y=(Y_n,\mathscr{F}_n)_{n\geqslant 0}(Y_0=0)$, 使 $X=V\cdot Y$.*

证明 a) $\Rightarrow$ b). 设 X 是局部鞅, 而 $(\tau_k)_{k\geqslant 1}$ 是局部化马尔可夫时间序列. 那么, 对于任何 $m\geqslant 0$,

$$\mathbf{E}[|X_{m\wedge\tau_k}|I_{\{\tau_k>0\}}]<\infty,\tag{6}$$

从而

$$\mathbf{E}[|X_{(n+1)\wedge\tau_k}|I_{\{\tau_k>n\}}]=\mathbf{E}[|X_{n+1}|I_{\{\tau_k>n\}}]<\infty.\tag{7}$$

随机变量 $I_{\{\tau_k>n\}}$ 是 $\mathscr{F}_n-$ 可测的. 因此由 (7) 式可见, 有

$$\mathbf{E}[|X_{n+1}|I_{\{\tau_k>n\}}|\mathscr{F}_n]=I_{\{\tau_k>n\}}\mathbf{E}[|X_{n+1}||\mathscr{F}_n]<\infty\ (\mathbf{P}-\text{a.c.}).$$

这里, $I_{\{\tau_k>n\}}\to 0\ (\mathbf{P}-\text{a.c.})$, $k\to\infty$, 因而

$$\mathbf{E}[|X_{n+1}||\mathscr{F}_n]<\infty\quad(\mathbf{P}-\text{a.c.}).\tag{8}$$

由于这一条件, 知 $\mathbf{E}[X_{n+1}|\mathscr{F}_n]$ 有定义, 故只剩下证明 $\mathbf{E}[X_{n+1}|\mathscr{F}_n]=X_n\ (\mathbf{P}-\text{a.c.})$.

为此, 需要对于 $A\in\mathscr{F}_n$, 证明

$$\int_A X_{n+1}\mathrm{d}\mathbf{P}=\int_A X_n\mathrm{d}\mathbf{P}.$$

由第二章 §7 练习题 7 知, $\mathbf{E}[|X_{n+1}||\mathscr{F}_n]<\infty(\mathbf{P}-\text{a.c.})$, 当且仅当测度

$$\int_A X_{n+1}\mathrm{d}\mathbf{P},\quad A\in\mathscr{F}_n$$

是 $\sigma-$ 有限的. 现在证明测度 $\int_A|X_n|\mathrm{d}\mathbf{P}, A\in\mathscr{F}_n$, 也是 $\sigma-$ 有限的.

由于 X^{τ_k} 是鞅, 故 $|X^{\tau_k}|=(|X_{\tau_k\wedge n}|I_{\{\tau_k>0\}},\mathscr{F}_n)$ 是下鞅, 从而 ($\{\tau_k>n\}\in\mathscr{F}_n$),

$$\begin{aligned}\int_{A\cap\{\tau_k>n\}}|X_n|\mathrm{d}\mathbf{P}&=\int_{A\cap\{\tau_k>n\}}|X_{n\wedge\tau_k}|I_{\{\tau_k>0\}}\mathrm{d}\mathbf{P}\\&\leqslant\int_{A\cap\{\tau_k>n\}}|X_{(n+1)\wedge\tau_k}|I_{\{\tau_k>0\}}\mathrm{d}\mathbf{P}=\int_{A\cap\{\tau_k>n\}}|X_{n+1}|\mathrm{d}\mathbf{P}.\end{aligned}$$

令 $k\to\infty$, 得

$$\int_A|X_n|\mathrm{d}\mathbf{P}\leqslant\int_A|X_{n+1}|\mathrm{d}\mathbf{P},$$

因此就证明了测度 $\int_A X_{n+1}\mathrm{d}\mathbf{P}, A\in\mathscr{F}$, 是 $\sigma-$ 有限的.

设 $A\in\mathscr{F}$, 使 $\int_A|X_{n+1}|\mathrm{d}\mathbf{P}<\infty$. 那么, 根据勒贝格控制收敛定理, 可以在下列式子中求极限:

$$\int_{A\cap\{\tau_k>n\}}X_n\mathrm{d}\mathbf{P}=\int_{A\cap\{\tau_k>n\}}X_{n+1}\mathrm{d}\mathbf{P},$$

而由于 X 是局部鞅, 可见求极限是合理的. 从而, 有

$$\int_A X_n\mathrm{d}\mathbf{P}=\int_A X_{n+1}\mathrm{d}\mathbf{P},$$

而且对于任意 $A\in\mathscr{F}_n$, $\int_A|X_{n+1}|\mathrm{d}\mathbf{P}<\infty$. 由此可见, 对于任意 $A\in\mathscr{F}_n$, 上面的等式成立, 从而 $\mathbf{E}[X_{n+1}|\mathscr{F}_n]=X_n(\mathbf{P}-\text{a.c.})$.

b) $\Rightarrow$ c). 设 $\Delta X_n=X_n-X_{n-1}, X_0=0; V_0=0, V_n=\mathbf{E}[|\Delta X_n||\mathscr{F}_{n-1}](n\geqslant 1)$; 记

$$W_n=V_n^{\oplus}\left(=\begin{cases}V_n^{-1}, & V_n\neq 0,\\ 0, & V_n\neq 0,\end{cases}\right),$$

$Y_0=0$ 和 $Y_n=\sum\limits_{i=1}^{n}W_i\Delta X_i(n\geqslant 1)$. 显然

$$\mathbf{E}[|\Delta Y_n||\mathscr{F}_{n-1}]\leqslant 1,\quad\mathbf{E}[\Delta Y_n|\mathscr{F}_{n-1}]=0,$$

从而, $Y=(Y_n,\mathscr{F}_n)$ 是鞅. 其次 $X_0=V_0\cdot Y_0=0, \Delta(V\cdot Y)_n=\Delta X_n$. 于是 $X=V\cdot Y$.

c) $\Rightarrow$ a). 设 $X=V\cdot Y$, 其中 V 是可预测序列, Y 是鞅, 而 $V_0=Y_0=0$. 记

$$\tau_k=\inf\{n\geqslant 0:|V_{n+1}|>k\},$$

且当集合 $\{\cdot\}=\varnothing$ 时 $\tau_\infty=\infty$. 由于 V_{n+1} 是 $\mathscr{F}_n-$ 可测的, 可见对于每一个 $k\geqslant 1$, 变量 τ_k 是马尔可夫时间.

考虑序列 $X^{\tau_k}=((V\cdot Y)_{n\wedge\tau_k}I_{\{\tau_k>0\}},\mathscr{F}_n)$.

在集合 $\{\tau_k>0\}$ 上 $|V_{n\wedge\tau_k}|\leqslant k$. 故对于任意 $n\geqslant 1$, 有 $\mathbf{E}|(V\cdot Y)_{n\wedge\tau_k}I_{\{\tau_k>0\}}|<\infty$. 其次, 对于任意 $n\geqslant 1$, 有

$$\begin{aligned}&\mathbf{E}\{[(V\cdot Y)_{(n+1)\wedge\tau_k}-(V\cdot Y)_{n\wedge\tau_k}]I_{\{\tau_k>0\}}|\mathscr{F}_n\}\\ =&I_{\{\tau_k>0\}}V_{(n+1)\wedge\tau_k}\mathbf{E}\{Y_{(n+1)\wedge\tau_k}-Y_{n\wedge\tau_k}|\mathscr{F}_n\}=0,\end{aligned}$$

因为 (见例 7) $\mathbf{E}\{Y_{(n+1)\wedge\tau_k}-Y_{n\wedge\tau_k}|\mathscr{F}_n\}=0$.

于是, 对于每一个 $k\geqslant 1$, "停止" 序列 X^{τ_k} 是鞅, 其中 $\tau_n\uparrow\infty$ ($\mathbf{P}-$a.c.), 从而 X 是局部鞅. □

5. 博弈与数学概念 "鞅" 的产生

例 8 $(\eta_n)_{n\geqslant 1}$ 是独立同分布的伯努利随机变量, 且 $\mathbf{P}\{\eta_n=1\}=p,\mathbf{P}\{\eta_n=-1\}=q, p+q=1$. 假如考虑某项博弈, 可以把事件 $\{\eta_n=1\}$ 视为某选手在第 n 局中 "成功" (赢), 而把事件 $\{\eta_n=-1\}$ 视为在第 n 局中 "失败" (输). 假设 V_n 是第 n 局的赌注. 那么, 赌徒经 n 局赢得的总金额 (*总收益*) 等于

$$X_n=\sum_{i=1}^n V_i\eta_i=X_{n-1}+V_n\eta_n,\quad X_0=0.$$

完全自然, 第 n 局的赌注 V_n 可依赖于上一局的结果, 即依赖于 $V_1,\cdots,V_{n-1}$ 和 $\eta_1,\cdots,\eta_{n-1}$. 换句话说, 如果设 $\mathscr{F}_0=(\varnothing,\Omega)$ 和 $\mathscr{F}_n=\sigma\{\eta_1,\cdots,\eta_n\}$, 则 V_n 是 $\mathscr{F}_{n-1}-$ 可测随机变量, 即决定赌徒 "策略" 的序列 $V=(V_n,\mathscr{F}_{n-1})$ 是*可以预测的*. 若设 $Y_n=\eta_1+\cdots+\eta_n$, 则有

$$X_n=\sum_{i=1}^n V_i\Delta Y_i,$$

则序列 $X=(X_n,\mathscr{F}_n)$, 其中 $X_0=0$, 是 Y 由 V 的变换.

按博弈的观点, 如果每一步的期望收益 $\mathbf{E}(X_{n+1}-X_n|\mathscr{F}_n)=0$ ($\geqslant 0$或 $\leqslant 0$), 则该博弈是*公平的* (*有利的*或*不利的*). 因此显然, 博弈是

公平的, 如果 $p=q=1/2$,

有利的, 如果 $p>q$,

不利的, 如果 $p<q$.

由于序列 $X=(X_n,\mathscr{F}_n)$ 构成

鞅, 如果 $p=q=1/2$,

下鞅, 如果 $p>q$,

上鞅, 如果 $p<q$,

则可以认为, 关于该博弈是*公平的* (*有利的*或*不利的*), 对应于关于序列 X 是鞅 (下鞅或上鞅).

现在考虑特别的 "策略" 类 $V_n=(V_n,\mathscr{F}_{n-1})_{n\geqslant 1}$, 其中 $V_1=1$ 和

$$V_n=\begin{cases}2^{n-1}, & \text{若 } \eta_1=-1,\cdots,\eta_{n-1}=-1,\\ 0, & \text{若不然},\end{cases}\quad n>1, \tag{9}$$

该式的含义是: 赌徒从赌注 $V_1=1$ 开始, 每当他赢一局时就将赌注增加一倍, 而每当他输一局时就终止博弈.

假如 $\eta_1=-1,\cdots,\eta_n=-1$, 则经 n 局赌徒的总损失等于

$$\sum_{i=1}^n 2^{i-1}=2^n-1.$$

因此, 假如再假设 $\eta_{n+1}=1$, 则

$$X_{n+1}=X_n+V_{n+1}=-(2^n-1)+2^n=1.$$

记 $\tau=\inf\{n\geqslant 1:X_n=1\}$. 如果 $p=q=1/2$, 即所考虑的博弈是公平的, 则

$$\mathbf{P}\{\tau=n\}=\frac{1}{2^n},\quad \mathbf{P}\{\tau<\infty\}=1,\quad \mathbf{P}\{X_\tau=1\}=1,\quad \mathbf{E}X_\tau=1.$$

这样, 甚至在公平博弈的情形下, 如果赌徒坚持运用 "策略" (9), 他 (以概率 1) 经过有限时间完全可以顺利的结束博弈, 并使自己的赌注再增加一个单位 ($\mathbf{E}X_\tau=1>X_0=0$).

在*博弈的实际中*, 在输局时加倍赌注而在首次赌赢时停止博弈, 所描绘博弈的系统称做*鞅*. 也正式由此产生了数学概念 "鞅".

注 在 $p=q=1/2$ 的情形下, 序列 $X=(X_n,\mathscr{F}_n)_{n\geqslant 0}$ 是鞅, 其中 $X_0=0$, 因而对于任何 $n\geqslant 1$,

$$\mathbf{E}X_n=\mathbf{E}X_0=0.$$

因此, 假如将考虑时刻 n 换成考虑随机时刻 τ, 则可以指望该关系式仍然成立. 以后 (§2 定理 1) 会清楚地看到, 在 "典型" 的场合 $\mathbf{E}X_\tau=\mathbf{E}X_0$ 成立. 而这一等式不成立 (如上面考虑的博弈) 的情形, 出现在所谓 "物理上不可实现" 的场合, 这时或者 τ, 或者 $|X_n|$ 取的值过分地大. (应该指出, 上面考虑的博弈的情形是物理上无法实现的, 因为它假定博弈时间无限, 以及赌徒的初始赌资无限.)

6. 鞅 – 差

定义 6 随机序列 $\xi=(\xi_n,\mathscr{F}_n)_{n\geqslant 0}$ 称做*鞅 – 差*, 如果对于一切 $n\geqslant 0, \mathbf{E}|\xi_n|<\infty$ 且

$$\mathbf{E}(\xi_{n+1}|\mathscr{F}_n)=0\ (\mathbf{P}-\text{a.c.}). \tag{10}$$

由定义 1 和 6 可以清楚地看到鞅与鞅 - 差之间的联系. 具体地说, 如果 $X = (X_n, \mathscr{F}_n)$ 是鞅, 则 $\xi = (\xi_n, \mathscr{F}_n)$ 是鞅 - 差, 其中 $\xi_0 = X_0, \xi_n = \Delta X_n, n \geqslant 1$. 同样地, 如果 $\xi = (\xi_n, \mathscr{F}_n)$ 是鞅 - 差, 则 $X = (X_n, \mathscr{F}_n)$ 是鞅, 其中 $X_n = \xi_0 + \cdots + \xi_n$.

按照这一术语, 如果 $\mathbf{E}\xi_n = 0, n \geqslant 0$, 而相应的 $\sigma-$ 代数为 $\mathscr{F}_n = \sigma(\xi_0, \cdots, \xi_n)$, 则任何独立可积随机变量序列 $\xi = (\xi_n)_{n \geqslant 0}$ 是鞅 - 差.

7. 下鞅 (上鞅) 的构造 下面的定理使下鞅 (上鞅) 的结构更加清晰.

定理 2 (杜布 (J. L. Doob)) *设 $X = (X_n, \mathscr{F}_n)$ 是下鞅, 则存在鞅 $m = (m_n, \mathscr{F}_n)$ 和这样的可预测递增序列 $A = (A_n, \mathscr{F}_{n-1})$, 使对于每一个 $n \geqslant 0$, 有杜布分解:*

$$X_n = m_n + A_n \quad (\mathbf{P}-\text{a.c.}). \tag{11}$$

而且类似形式的分解唯一.

证明 设 $m_0 = X_0, A_0 = 0$, 而

$$m_n = m_0 + \sum_{j=0}^{n-1} [X_{j+1} - \mathbf{E}(X_{j+1}|\mathscr{F}_j)], \tag{12}$$

$$A_n = \sum_{j=0}^{n-1} [\mathbf{E}(X_{j+1}|\mathscr{F}_j) - X_j]. \tag{13}$$

这样定义的 m 和 A, 显然具有所要求的性质. 其次, 亦设 $X_n = m'_n + A'_n$, 其中 $m' = (m'_n, \mathscr{F}_n)$ 是鞅, 而 $A' = (A'_n, \mathscr{F}_{n-1})$ 是可预测递增序列. 那么,

$$A'_{n+1} - A'_n = (A_{n+1} - A_n) + (m_{n+1} - m_n) - (m'_{n+1} - m'_n),$$

然后在等式两侧同求条件数学期望, 得 $A'_{n+1} - A'_n = A_{n+1} - A_n$ $(\mathbf{P}-\text{a.c.})$. 由于 $A_0 = A'_0 = 0$, 说明对于一切 $n \geqslant 0, A_n = A'_n$ $(\mathbf{P}-\text{a.c.})$ 和 $m_n = m'_n$ $(\mathbf{P}-\text{a.c.})$. □

由分解 (11) 可见, 序列 $A = (A_n, \mathscr{F}_{n-1})$ 补偿 $X = (X_n, \mathscr{F}_n)$ 成为鞅. 这说明如下定义是合理的.

定义 7 杜布分解 (11) 中的可预测递增序列 $A = (A_n, \mathscr{F}_{n-1})$, 称做 (下鞅 X 的) *补偿*.

如果 $\mathbf{E}M^2 < \infty, n \geqslant 0$, 则称鞅 $M = (M_n, \mathscr{F}_n)_{n \geqslant 0}$ 为平方可积的. 在研究平方可积鞅 $M = (M_n, \mathscr{F}_n)_{n \geqslant 0}$ 时, 杜布分解起核心作用, 因为根据上面的说明随机序列 $M^2 = (M_n^2, \mathscr{F}_n)$ 是下鞅. 根据定理 2, 存在鞅 $m = (m_n, \mathscr{F}_n)$ 和可预测递增序列 $\langle M \rangle = (\langle M \rangle_n, \mathscr{F}_{n-1})$, 使

$$M_n^2 = m_n + \langle M \rangle_n. \tag{14}$$

序列 $\langle M \rangle$ 称做鞅 M 的*二次特征*, 并在许多方面决定其构造和性质.

由 (13) 式可见,

$$\langle M \rangle_n = \sum_{j=1}^{n} \mathbf{E}[(M_j)^2 | F_{j-1}], \tag{15}$$

而且对于一切 $l \leqslant k$, 有

$$\mathbf{E}[(M_k - M_l)^2 | \mathscr{F}_l] = \mathbf{E}[M_k^2 - M_l^2 | \mathscr{F}_l] = \mathbf{E}[\langle M \rangle_k - \langle M \rangle_l | \mathscr{F}_l], \tag{16}$$

特别, 若 $M_0 = 0$ $(\mathbf{P}-\text{a.c.})$, 则

$$\mathbf{E}M_k^2 = \mathbf{E}\langle M \rangle_k. \tag{17}$$

有益地指出, 如果 $M_0 = 0, M_n = \xi_1 + \cdots + \xi_n$, 其中 (ξ_n) 是独立随机变量序列, 且 $\mathbf{E}\xi_i = 0, \mathbf{E}\xi_i^2 < \infty$, 则二次特征

$$\langle M \rangle_n = \mathbf{E}M_n^2 = \mathbf{D}\xi_1 + \cdots + \mathbf{D}\xi_n \tag{18}$$

是非随机的, 并且等于 M_n 的方差.

如果 $X = (X_n, \mathscr{F}_n)$ 和 $Y = (Y_n, \mathscr{F}_n)$ 是平方可积鞅, 则设

$$\langle X, Y \rangle_n = \frac{1}{4} [\langle X + Y \rangle_n - \langle X - Y \rangle_n]. \tag{19}$$

不难验证, $(X_n Y_n - \langle X, Y \rangle_n, \mathscr{F}_n)$ 是鞅, 因而对于任意 $l \leqslant k$,

$$\mathbf{E}[(X_k - X_l)(Y_k - Y_l) | \mathscr{F}_l] = \mathbf{E}[\langle X, Y \rangle_k - \langle X, Y \rangle_l | \mathscr{F}_l]. \tag{20}$$

当 $X_n = \xi_1 + \cdots + \xi_n$ 和 $Y_n = \eta_1 + \cdots + \eta_n$ 时, 其中 (ξ_n) 和 (η_n) 是独立随机变量序列, 而且 $\mathbf{E}\xi_i = \mathbf{E}\eta_i = 0, \mathbf{E}\xi_i^2 < \infty, \mathbf{E}\eta_i^2 < \infty$, 变量 $\langle X, Y \rangle_n$ 等于

$$\langle X, Y \rangle_n = \sum_{i=1}^{n} \operatorname{cov}(\xi_i, \eta_i).$$

序列 $\langle X, Y \rangle = (\langle X, Y \rangle_n, \mathscr{F}_{n-1})$ 常称为 (平方可积) 鞅 X 和 Y 的*相互特征*. 不难验证, 有 (对照 (15) 式)

$$\langle X, Y \rangle_n = \sum_{i=1}^{n} \mathbf{E}[\Delta X_i \Delta Y_i | \mathscr{F}_{i-1}].$$

下面两个量在鞅论中也有重要作用: 一个是*二次协方差*

$$[X, Y]_n = \sum_{i=1}^{n} \Delta X_i \Delta Y_i,$$

另一个是*二次变差*

$$[X]_n = \sum_{i=1}^{n} (\Delta X_i)^2.$$

这两个量对于任何随机变量序列 $X = (X_n)_{n \geqslant 1}$ 和 $Y = (Y_n)_{n \geqslant 1}$ 有定义.

8. 局部鞅是鞅的条件 鉴于定理 1, 自然产生一个问题, 局部鞅 (因而, 广义鞅或鞅变换) 实际上本身也是鞅.

定理 3 1) *设随机序列 $X=(X_n,\mathscr{F}_n)_{n\geqslant 0}$ 是局部鞅 (其中 $X_0=0$, 或更一般地 $\mathbf{E}|X_0|<\infty$).*

如果 $\mathbf{E}X_n^-<\infty, n\geqslant 0$, 或 $\mathbf{E}X_n^+<\infty, n\geqslant 0$, 则序列 $X=(X_n,\mathscr{F}_n)_{n\geqslant 0}$ 是鞅.

2) *设 $X=(X_n,\mathscr{F}_n)_{0\leqslant n\leqslant N}(N>0)$ 是局部鞅, 或 $\mathbf{E}X_N^-<\infty$, 或 $\mathbf{E}X_N^+<\infty$. 那么 $X=(X_n,\mathscr{F}_n)_{0\leqslant n\leqslant N}$ 是鞅.*

证明 1) 证明, 若条件 "$\mathbf{E}X_n^-<\infty, n\geqslant 0$" 和 "$\mathbf{E}X_n^+<\infty, n\geqslant 0$" 中任何一个条件成立, 则条件 "$\mathbf{E}|X_n|<\infty, n\geqslant 0$" 也成立.

事实上, 假如对于一切 $n\geqslant 0, \mathbf{E}X_n^-<\infty$. 那么, 由法图 (F. Fatou) 引理, 可见

$$\begin{aligned}\mathbf{E}X_n^+&=\mathbf{E}\varliminf_k X_{n\wedge\tau_k}^+\leqslant\varliminf_k\mathbf{E}X_{n\wedge\tau_k}^+=\varliminf_k[\mathbf{E}X_{n\wedge\tau_k}+\mathbf{E}X_{n\wedge\tau_k}^-]\\&=\mathbf{E}X_0+\varliminf_k\mathbf{E}X_{n\wedge\tau_k}^-\leqslant|\mathbf{E}X_0|+\sum_{k=0}^n\mathbf{E}X_k^-<\infty.\end{aligned}$$

从而, $\mathbf{E}|X_n|<\infty, n\geqslant 0$.

为证明 ($\mathbf{E}(X_{n+1}|\mathscr{F}_n)=X_n, n\geqslant 0$) 的鞅性, 注意到对于任意马尔可夫时间 τ_k, 有

$$|X_{(n+1)\wedge\tau_k}|=\sum_{i=0}^{n+1}|X_i|,$$

其中

$$\mathbf{E}\sum_{i=0}^{n+1}|X_i|<\infty.$$

因此, 由勒贝格控制收敛定理, 经在关系式 $\mathbf{E}(X_{(n+1)\wedge\tau_k}|\mathscr{F}_n)=X_{n\wedge\tau_k}$ 中极限过渡 ($k\to\infty, \tau_k\uparrow\infty(\mathbf{P}-\text{a.c.})$), 得 $\mathbf{E}(X_{n+1}|\mathscr{F}_n)=X_n(\mathbf{P}-\text{a.c.})$.

2) 假设 $\mathbf{E}X_N^-<\infty$, 证明对于一切 $n<N, \mathbf{E}X_n^-<\infty$.

事实上, 由于局部鞅也是广义鞅, 可见 $X_n=\mathbf{E}(X_{n+1}|\mathscr{F}_n)$, 其中 $\mathbf{E}(|X_{n+1}||\mathscr{F}_n)<\infty$ ($\mathbf{P}-\text{a.c.}$). 那么, 根据对于条件数学期望的延森 (I. L. Iensen) 不等式 (见第二章 §7 练习题 5), 有 $X_n^-\leqslant\mathbf{E}(X_{n+1}^-|\mathscr{F}_n)$. 因此, $\mathbf{E}X_n^-\leqslant\mathbf{E}X_{n+1}^-\leqslant\mathbf{E}X_N^-<\infty$.

于是, 由命题 1), 可得所要证明的局部鞅 $X=(X_n,\mathscr{F}_n)_{0\leqslant n\leqslant N}$ 的鞅性.

9. 练习题

1. 证明条件 (2) 和 (3) 的等价性.

2. 设 σ 和 τ 是马尔可夫时间, 证明 $\tau+\sigma, \tau\vee\sigma, \tau\wedge\sigma$ 也是马尔可夫时间, 并且如果 $\sigma\leqslant\tau$, 则 $\mathscr{F}_\sigma\subseteq\mathscr{F}_\tau$.

3. 证明 τ 和 X_τ 是 $\mathscr{F}_\tau$ - 可测的.

4. 设 $Y=(Y_n,\mathscr{F}_n)$ 是鞅 (下鞅), $V=(V_n,\mathscr{F}_{n-1})$ 是可预测序列, $(V\cdot Y)_n, n\geqslant 0$ 是可积随机变量. 证明 $V\cdot Y$ 是鞅 (下鞅).

5. 设 $\mathscr{G}_1\supseteq\mathscr{G}_2\supseteq\cdots$ 是非增 $\sigma-$ 代数族, 而 ξ 是可积随机变量. 证明序列 $(X_n)_{n\geqslant 1}$, 其中 $X_n=\mathbf{E}(\xi|\mathscr{G}_n)$, 是逆鞅, 即对于任何 $n\geqslant 1$, 有

$$\mathbf{E}(X_n|X_{n+1},X_{n+2},\cdots)=X_{n+1}\quad(\mathbf{P}-\text{a.c.}).$$

6. 设 $\xi_1,\xi_2,\cdots$ 是独立随机变量, 且

$$\mathbf{P}\{\xi_i=0\}=\mathbf{P}\{\xi_i=2\}=\frac{1}{2}\quad\text{和}\quad X_n=\prod_{i=1}^n\xi_i.$$

证明不存在可积随机变量 ξ 和非降 $\sigma-$ 代数族 ($\mathscr{F}_n$), 使 $X_n=\mathbf{E}(\xi|\mathscr{F}_n)$. (这个例子说明, 并不是每一个鞅 $(X_n)_{n\geqslant 1}$ 都可以表示为 $(\mathbf{E}(\xi|\mathscr{F}_n))_{n\geqslant 1}$; 对照第一章 §11 例 3).

7. (a) 设 $\xi_1,\xi_2,\cdots$ 是独立随机变量, 且对于 $n\geqslant 1, \mathbf{E}|\xi_n|<\infty, \mathbf{E}\xi_n=0$, 证明对于每一个 $k\geqslant 1$, 序列

$$X_n^{(k)}=\sum_{1\leqslant i_1<\cdots<i_k\leqslant n}\xi_{i_1}\cdots\xi_{i_k},\quad n\geqslant k$$

是鞅.

(b) 设 $\xi_1,\xi_2,\cdots$ 是可积随机变量, 满足

$$\mathbf{E}(\xi_{n+1}|\xi_1,\cdots,\xi_n)=\frac{\xi_1+\cdots+\xi_n}{n}(=X_n).$$

证明序列 $X_1,X_2,\cdots$ 是鞅.

8. 举一鞅的例 $X=(X_n,\mathscr{F}_n)_{n\geqslant 1}$, 而随机变量族 $(X_n, n\geqslant 1)$ 不一致可积.

9. 设 $X=(X_n)_{n\geqslant 0}$ 是马尔可夫链 (第八章 §1), 且其状态的集合 $E=\{i,j,\cdots\}$ 是可数的, 而 p_{ij} 是它的转移概率. 此外, 假设 $\psi=\psi(x), x\in E$, 是有界函数, 并且满足条件

$$\sum_{j\in E}p_{ij}\psi(j)\leqslant\lambda\psi(i),\quad\lambda>0,\quad i\in E.$$

证明序列 $(\lambda^{-n}\psi(X_n))_{n\geqslant 0}$ 是上鞅.

§2. 在时间变量为随机时间时鞅性的不变性

1. 杜布定理 如果 $X=(X_n,\mathscr{F}_n)_{n\geqslant 0}$ 是鞅, 则对于一切 $n\geqslant 1$,

$$\mathbf{E}X_n=\mathbf{E}X_0.\tag{1}$$

假如将时间 n 换成随机时间 (例如, 马尔可夫时间) τ, 那么上述性质是否仍然成立? 在上一节例 8 中举的例子说明, 一般并非如此: 存在鞅 X 和 (以概率 1 有限的) 马尔可夫时间 τ, 使得

$$\mathbf{E}X_\tau\neq\mathbf{E}X_0.\tag{2}$$

下面的重要定理描绘一些 "典型" 情形, 其中包括 $\mathbf{E}X_\tau=\mathbf{E}X_0$ 的情形,

定理 1 (杜布) 设 $X=(X_n,\mathscr{F}_n)$ 是鞅 (下鞅), τ_1 和 τ_2 是停止时间, 满足

$$\mathbf{E}|X_{\tau_i}|<\infty,\quad i=1,2. \tag{3}$$

$$\varliminf_{n\to\infty}\int_{\{\tau_2>n\}}|X_n|\mathrm{d}\mathbf{P}=0. \tag{4}$$

那么,

$$\mathbf{E}(X_{\tau_2}|\mathscr{F}_{\tau_1})\underset{(\geqslant)}{=}X_{\tau_1}\quad(\{\tau_2\geqslant\tau_1\};\mathbf{P}-\text{a.c.}). \tag{5}$$

并且, 如果 $\mathbf{P}\{\tau_1\leqslant\tau_2\}=1$, 则

$$\mathbf{E}\tau_2\underset{(\geqslant)}{=}\mathbf{E}\tau_1. \tag{6}$$

证明 只需证明, 对于任意 $A\in\mathscr{F}_{\tau_1}$, 有

$$\int_{A\cap\{\tau_2\geqslant\tau_1\}}X_{\tau_2}\mathrm{d}\mathbf{P}\underset{(\geqslant)}{=}\int_{A\cap\{\tau_2\geqslant\tau_1\}}X_{\tau_1}\mathrm{d}\mathbf{P}. \tag{7}$$

同样, 为此只需证明, 对于任意 $n\geqslant0$, 有

$$\int_{A\cap\{\tau_2\geqslant\tau_1\}\cap\{\tau_1=n\}}X_{\tau_2}\mathrm{d}\mathbf{P}\underset{(\geqslant)}{=}\int_{A\cap\{\tau_2\geqslant\tau_1\}\cap\{\tau_1=n\}}X_{\tau_1}\mathrm{d}\mathbf{P},$$

或者

$$\int_{B\cap\{\tau_2\geqslant n\}}X_{\tau_2}\mathrm{d}\mathbf{P}\underset{(\geqslant)}{=}\int_{B\cap\{\tau_2\geqslant n\}}X_n\mathrm{d}\mathbf{P}. \tag{8}$$

其中 $B=A\bigcap\{\tau_1=n\}\in\mathscr{F}_n$.

因此, 有

$$\begin{aligned}\int_{B\cap\{\tau_2\geqslant n\}}X_n\mathrm{d}\mathbf{P}&=\int_{B\cap\{\tau_2=n\}}X_n\mathrm{d}\mathbf{P}+\int_{B\cap\{\tau_2>n\}}X_n\mathrm{d}\mathbf{P}\\&\underset{(\leqslant)}{=}\int_{B\cap\{\tau_2=n\}}X_n\mathrm{d}\mathbf{P}+\int_{B\cap\{\tau_2>n\}}\mathbf{E}(X_{n+1}|\mathscr{F}_n)\mathrm{d}\mathbf{P}\\&=\int_{B\cap\{\tau_2=n\}}X_{\tau_2}\mathrm{d}\mathbf{P}+\int_{B\cap\{\tau_2\geqslant n+1\}}X_{n+1}\mathrm{d}\mathbf{P}\\&\underset{(\leqslant)}{=}\int_{B\cap\{n\leqslant\tau_2\leqslant n+1\}}X_{\tau_2}\mathrm{d}\mathbf{P}+\int_{B\cap\{\tau_2\geqslant n+2\}}X_{n+2}\mathrm{d}\mathbf{P}\\&\underset{(\leqslant)}{=}\int_{B\cap\{n\leqslant\tau_2\leqslant m\}}X_{\tau_2}\mathrm{d}\mathbf{P}+\int_{B\cap\{\tau_2>m\}}X_m\mathrm{d}\mathbf{P}.\\&\underset{(\leqslant)}{=}\cdots\end{aligned}$$

从而, 有

$$\int_{B\cap\{n\leqslant\tau_2\leqslant m\}}X_{\tau_2}\mathrm{d}\mathbf{P}\underset{(\geqslant)}{=}\int_{B\cap\{n\leqslant\tau_2\}}X_n\mathrm{d}\mathbf{P}-\int_{B\cap\{m<\tau_2\}}X_m\mathrm{d}\mathbf{P},$$

并且由于 (4) 式和等式 $X_m=2X_m^+-|X_m|$, 得

$$\begin{aligned}\int_{B\cap\{\tau_2\geqslant n\}}X_{\tau_2}\mathrm{d}\mathbf{P}&\underset{(\geqslant)}{=}\lim_{m\to\infty}\left[\int_{B\cap\{n\leqslant\tau_2\}}X_n\mathrm{d}\mathbf{P}-\int_{B\cap\{m<\tau_2\}}X_m\mathrm{d}\mathbf{P}\right]\\&=\int_{B\cap\{n\leqslant\tau_2\}}X_{\tau_2}\mathrm{d}\mathbf{P}-\varliminf_{m\to\infty}\int_{B\cap\{m<\tau_2\}}X_m\mathrm{d}\mathbf{P}\\&=\int_{B\cap\{n\leqslant\tau_2\}}X_{\tau_2}\mathrm{d}\mathbf{P}.\end{aligned}$$

于是 (8) 式得证, 故 (5) 式也得证. 最后, 由 (5) 式得 (6) 式. □

系 1 如果存在常数 N, 使 $\mathbf{P}\{\tau_1\leqslant N\}=1,\mathbf{P}\{\tau_2\leqslant N\}=1$, 则满足条件 (3),(4). 因此, 如果同时 $\mathbf{P}\{\tau_1\leqslant\tau_2\}=1$ 且 X 是鞅, 则

$$\mathbf{E}X_0=\mathbf{E}\tau_1=\mathbf{E}\tau_2=\mathbf{E}X_N.$$

系 2 如果随机变量族 $\{X_n\}$ 一致可积 (特别, 如果以概率 1, 有 $|X_n|\leqslant C<\infty,n\geqslant0$), 则满足条件 (3), (4).

事实上, 由于 $\mathbf{P}\{\tau_i>n\}\to0,n\to\infty$, 故由第二章 §6 引理 2 可以得出条件 (4). 其次, 由于随机变量族 $\{X_n\}$ 一致可积, 可见 (见第二章 §6 (16) 式)

$$\sup_N\mathbf{E}|X_N|<\infty. \tag{9}$$

如果 τ 是某一停时, 而 X 是下鞅, 则将系 1 用于有界停时 $\tau_N=\tau\wedge N$, 可见

$$\mathbf{E}X_0\leqslant\mathbf{E}X_{\tau_N}.$$

因此,

$$\mathbf{E}|X_{\tau_N}|=2\mathbf{E}\tau_{\tau_N}^+-\mathbf{E}X_{\tau_N}\leqslant2\mathbf{E}\tau_{\tau_N}^+-\mathbf{E}X_0. \tag{10}$$

序列 $X^+=(X_n^+,\mathscr{F}_n)$ 是下鞅 (§1 例 5), 从而

$$\begin{aligned}\mathbf{E}X_{\tau_N}^+&=\sum_{j=0}^N\int_{\{\tau_N=j\}}X_j^+\mathrm{d}\mathbf{P}+\int_{\{\tau>N\}}X_N^+\mathrm{d}\mathbf{P}\\&\leqslant\sum_{j=0}^N\int_{\{\tau_N=j\}}X_N^+\mathrm{d}\mathbf{P}+\int_{\{\tau>N\}}X_N^+\mathrm{d}\mathbf{P}=\mathbf{E}X_N^+\leqslant\mathbf{E}|X_N|\leqslant\sup_N\mathbf{E}|X_N|.\end{aligned}$$

由此连同 (10) 式, 得

$$\mathbf{E}|X_{\tau_N}|\leqslant3\sup_N\mathbf{E}|X_N|.$$

从而根据法图引理, 得

$$\mathbf{E}|X_\tau|\leqslant3\sup_N\mathbf{E}|X_N|.$$

于是, 若选择 $\tau=\tau_i(i=1,2)$, 并考虑到 (9) 式, 得 $\mathbf{E}|X_{\tau_i}|<\infty,i=1,2$.

注 在上一节考虑的例 8 中,

$$\int_{\{\tau>n\}}|X_n|\mathrm{d}\mathbf{P}=(2^n-1)\mathbf{P}\{\tau>n\}=(2^n-1)2^{-n}\to 1,\quad n\to\infty,$$

从而, (对于 $\tau_2=\tau$) 条件 (4) 不成立.

2. 杜布定理的特例 对于应用, 由定理 1 引出的条件常是重要的.

定理 2 设 $X=(X_n)$ 是鞅 (下鞅), 而 τ 关于 $(\mathscr{F}_n^X)$ 是停时, 其中 $\mathscr{F}_n^X=\sigma\{X_0,\cdots,X_n\}$. 假设

$$\mathbf{E}\tau<\infty,$$

并且对于任意 $n\geqslant 0$ 和某个常数 $C>0$, 有

$$\mathbf{E}\{|X_{n+1}-X_n||\mathscr{F}_n^X\}\leqslant C\quad(\{\tau\geqslant n\};\mathbf{P}-\text{a.c.}).$$

那么,

$$\mathbf{E}|X_\tau|<\infty,$$

并且

$$\mathbf{E}X_\tau\underset{(\geqslant)}{=}\mathbf{E}X_0. \tag{11}$$

证明 现在验证, 对于 $\tau_2=\tau$, 定理 1 的条件 (3) 和 (4) 成立.

设 $Y_0=|X_0|, Y_j=|X_j-X_{j-1}|, j\geqslant 1$, 则

$$|X_\tau|\leqslant\sum_{j=0}^{\tau}Y_j,$$

$$\begin{aligned}\mathbf{E}|X_\tau|\leqslant\mathbf{E}\left(\sum_{j=0}^{\tau}Y_j\right)&=\int_\Omega\sum_{j=0}^{\tau}Y_j\mathrm{d}\mathbf{P}=\sum_{n=0}^{\infty}\int_{\{\tau=n\}}\sum_{j=0}^{n}Y_j\mathrm{d}\mathbf{P}\\&=\sum_{n=0}^{\infty}\sum_{j=0}^{n}\int_{\{\tau=n\}}Y_j\mathrm{d}\mathbf{P}=\sum_{j=0}^{\infty}\sum_{n=j}^{\infty}\int_{\{\tau=n\}}Y_j\mathrm{d}\mathbf{P}=\sum_{j=0}^{\infty}\int_{\{\tau\geqslant j\}}Y_j\mathrm{d}\mathbf{P},\end{aligned}$$

其中集合 $\{\tau\geqslant j\}=\Omega\backslash\{\tau<j\}\in\mathscr{F}_{j-1}^X$. 因此对于 $j\geqslant 1$, 有

$$\int_{\{\tau\geqslant j\}}Y_j\mathrm{d}\mathbf{P}=\int_{\{\tau\geqslant j\}}\mathbf{E}(Y_j|X_0,\cdots,X_{j-1})\mathrm{d}\mathbf{P}\leqslant C\mathbf{P}\{\tau\geqslant j\}.$$

于是,

$$\mathbf{E}|X_\tau|\leqslant\mathbf{E}\left(\sum_{j=0}^{\tau}Y_j\right)\leqslant C\sum_{j=1}^{\infty}\mathbf{P}\{\tau\geqslant j\}+\mathbf{E}|X_0|=C\mathbf{E}\tau+\mathbf{E}|X_0|<\infty. \tag{12}$$

其次, 如果 $\tau>n$, 则

$$\sum_{j=0}^{n}Y_j\leqslant\sum_{j=0}^{\tau}Y_j,$$

所以

$$\int_{\{\tau>n\}}|X_n|\mathrm{d}\mathbf{P}\leqslant\int_{\{\tau>n\}}\sum_{j=0}^{\tau}Y_j\mathrm{d}\mathbf{P}.$$

由此并考虑到 (根据 (12) 式), $\mathbf{E}\left(\sum\limits_{j=0}^{\tau}Y_j\right)<\infty$, 且当 $n\to\infty$ 时 $\{\tau>n\}\downarrow\varnothing$, 根据控制收敛定理, 有

$$\varliminf_{n\to\infty}\int_{\{\tau>n\}}|X_n|\mathrm{d}\mathbf{P}\leqslant\varliminf_{n\to\infty}\int_{\{\tau>n\}}\sum_{j=0}^{\tau}Y_j\mathrm{d}\mathbf{P}.$$

于是, 定理 1 的条件成立, 故由此得 (11) 式. □

3. 基本瓦尔德恒等式 现在介绍刚证明的定理的某些应用,

定理 3 (瓦尔德恒等式) 设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量, 且 $\mathbf{E}|\xi_i|<\infty$, 而 τ 关于 $(\mathscr{F}_n^\xi)$ 是停时, 其中 $\mathscr{F}_n^\xi=\sigma\{\xi_1,\cdots,\xi_n\},\tau\geqslant 1,\mathbf{E}\tau<\infty$. 那么,

$$\mathbf{E}(\xi_1+\cdots+\xi_\tau)=\mathbf{E}\xi_1\times\mathbf{E}\tau. \tag{13}$$

如果同时又 $\mathbf{E}\xi_1^2<\infty$, 则

$$\mathbf{E}[(\xi_1+\cdots+\xi_\tau)-\tau\mathbf{E}\xi_1]^2=\mathbf{D}\xi_1\times\mathbf{E}\tau. \tag{14}$$

证明 显然, $X=(X_n,\mathscr{F}_n^\xi)_{n\geqslant 1}$ 是鞅, 其中 $X_n=(\xi_1+\cdots+\xi_\tau)-n\mathbf{E}\xi_1$, 且

$$\begin{aligned}\mathbf{E}[|X_{n+1}-X_n||X_1,\cdots,X_n]&=\mathbf{E}[|\xi_{n+1}-\mathbf{E}\xi_1||\xi_1,\cdots,\xi_n]\\&=\mathbf{E}|\xi_{n+1}-\mathbf{E}\xi_1|\leqslant 2\mathbf{E}|\xi_1|<\infty.\end{aligned}$$

因此, 根据定理 2, 得 $\mathbf{E}X_\tau=\mathbf{E}X_0=0$, 从而 (13) 式得证.

现在给出 “第二个瓦尔德恒等式” (14) 的三个证明,

证明 I 设 $\eta_i=\xi_i-\mathbf{E}\xi_i, S_n=\eta_1+\cdots+\eta_n$. 需要证明

$$\mathbf{E}S_\tau^2=\mathbf{E}\eta_1^2\times\mathbf{E}\tau.$$

设 $\tau(n)=\tau\wedge n(=\min\{\tau,n\})$. 由于

$$S_n^2=\sum_{i=1}^{n}\eta_i^2+2\sum_{1\leqslant i<j\leqslant n}\eta_i\eta_j,$$

可见序列 $\left(S_n^2-\sum\limits_{i=1}^{n}\eta_i^2,\mathscr{F}_n^\xi\right)_{n\geqslant 1}$ 是均值为 0 的鞅.

由定理 1 的系 1, 可见

$$\mathbf{E}S_{\tau(n)}^2=\mathbf{E}\sum_{i=1}^{\tau(n)}\eta_i^2.$$

根据“第一个瓦尔德恒等式” (13), 有

$$\mathbf{E}\sum_{i=1}^{\tau(n)}\eta_i^2=\mathbf{E}\eta_1^2\times\mathbf{E}\tau(n),$$

因而有 $\mathbf{E}S_\tau^2=\mathbf{E}\eta_1^2\times\mathbf{E}\tau$.

由于根据假设当 $m,n\to\infty$ 时, 类似地可得

$$\mathbf{E}(S_{\tau(n)}-S_{\tau(m)})^2=\mathbf{E}\eta_1^2\times\mathbf{E}[\tau(n)-\tau(m)]\to 0.$$

于是, 序列 $\{S_{\tau(n)}\}_{n\geqslant 1}$ 是 L^2 中的基本序列 (柯西序列, 见第二章 §10 第 5 小节), 从而根据第二章 §10 定理 7, 存在随机变量 S, 使 $\mathbf{E}(S_{\tau(n)}-S)^2\to 0, n\to\infty$. 由此可见 (第二章 §11 练习题 1), $\mathbf{E}S_{\tau(n)}^2\to\mathbf{E}S^2, n\to\infty$. 上面已证明 $\mathbf{E}S_{\tau(n)}^2=\mathbf{E}\eta_1^2\times\mathbf{E}\tau(n)$, 故当 $n\to\infty$ 时得 $\mathbf{E}S^2=\mathbf{E}\eta_1^2\times\mathbf{E}\tau$.

现在只剩下识别变量 S. 为此只需注意到存在一序列 $\{n'\}\subseteq\{n\}$, 使以概率 1 同时有收敛: $S_{\tau(n')}\to S$ 和 $\tau(n')\to\tau$. 那么, 以概率 1 也有收敛 $S_{\tau(n')}\to S_\tau$. 从而, 以概率 1 有 $S=S_\tau$, 故 $\mathbf{E}S_\tau^2=\mathbf{E}\eta_1^2\times\mathbf{E}\tau$, 而这正是要证明的.

证明 II 由已证明的等式 $\mathbf{E}S_{\tau(n)}^2=\mathbf{E}\eta_1^2\times\mathbf{E}\tau(n)$ 和法图引理 (见第二章 §6 定理 2 中的 a)), 可见

$$\mathbf{E}S_\tau^2=\mathbf{E}\underline{\lim}S_{\tau(n)}^2\leqslant\underline{\lim}\mathbf{E}S_{\tau(n)}^2\leqslant\mathbf{E}\eta_1^2\times\mathbf{E}\tau.$$

如果能证明对所有 $n\geqslant 1$, 不等式 $\mathbf{E}S_{\tau(n)}^2\leqslant\mathbf{E}S_\tau^2$ 成立, 则欲证的等式 $\mathbf{E}S_\tau^2=\mathbf{E}\eta_1^2\times\mathbf{E}\tau$ 就随之得证.

为此注意到, 由“第一个瓦尔德恒等式” (13), 有

$$\mathbf{E}|S_\tau|=\mathbf{E}|\eta_1+\cdots+\eta_\tau|\leqslant\mathbf{E}(|\eta_1|+\cdots+|\eta_\tau|)=\mathbf{E}|\eta_1|\times\mathbf{E}\tau<\infty.$$

因而, 当 $n\to\infty$ 时, 有

$$\begin{aligned}\mathbf{E}|S_n|I(\tau>n)&=\mathbf{E}|\eta_1+\cdots+\eta_n|I(\tau>n)\\&\leqslant\mathbf{E}(|\eta_1|+\cdots+|\eta_n|)I(\tau>n)\leqslant\mathbf{E}(|\eta_1|+\cdots+|\eta_\tau|)I(\tau>n)\to 0.\end{aligned}$$

运用定理 1 (设 $\tau_1=n,\tau_2=\tau$ 和下鞅 $(|S_n|,\mathscr{F}_n^\xi)_{n\geqslant 1}$) 可得, 在集合 $\{\tau\geqslant n\}$ 上, 以概率 1 有

$$\mathbf{E}(|S_\tau||\mathscr{F}_n^\xi)\geqslant|S_n|.$$

由 (关于条件数学期望得) 延森不等式 (第二章 §7 练习题 5) 可见, 在集合 $\{\tau\geqslant n\}$ 上, 以概率 1 有

$$\mathbf{E}(S_\tau^2|\mathscr{F}_n^\xi)\geqslant S_n^2=S_{\tau(n)}^2.$$

但在集合 $\{\tau<n\}$ 上 $\mathbf{E}(S_\tau^2|\mathscr{F}_n^\xi)=S_\tau^2=S_{\tau(n)}^2$. 所以以概率 1 有

$$\mathbf{E}(S_\tau^2|\mathscr{F}_n^\xi)\geqslant S_{\tau(n)}^2.$$

因此, $\mathbf{E}S_\tau^2\geqslant S_{\tau(n)}^2$. 而这正是要证明的.

证明 III 由“第一个证明”可见 $\left(S_n^2-\sum_{i=1}^n\eta_i^2,\mathscr{F}_n^\xi\right)_{n\geqslant 1}$ 是鞅, 而对于 $\tau(n)=\tau\wedge n$, 有

$$\mathbf{E}S_{\tau(n)}^2=\mathbf{E}\eta_1^2\times\mathbf{E}\tau(n).$$

因为 $\mathbf{E}\tau(n)\to\mathbf{E}\tau$, 故只需证明 $\mathbf{E}S_{\tau(n)}^2\to\mathbf{E}S_\tau^2$. 而为此只需证明

$$\mathbf{E}\sup_n S_{\tau(n)}^2<\infty,$$

因为, 这时由勒贝格控制收敛定理 (第二章 §6 定理 3) 可得要求证明的收敛性.

为了证明不等式 $\mathbf{E}\sup_n S_{\tau(n)}^2<\infty$, 我们利用将在下面 §3 中引进的“最大不等式” (14). 若将该不等式用于鞅 $(S_{\tau(k)},\mathscr{F}_k^\xi)_{k\geqslant 1}$, 则得

$$\mathbf{E}\left[\sup_{1\leqslant k\leqslant n}S_{\tau(k)}^2\right]\leqslant 4\mathbf{E}S_{\tau(n)}^2\leqslant 4\sup_n\mathbf{E}S_{\tau(n)}^2,$$

由此利用单调收敛定理 (第二章 §6 定理 1), 得

$$\mathbf{E}\sup_{k\geqslant 1}S_{\tau(k)}^2\leqslant 4\sup_n\mathbf{E}S_{\tau(n)}^2.$$

因为

$$\mathbf{E}S_{\tau(n)}^2=\mathbf{E}\eta_1^2\times\mathbf{E}\tau(n)\leqslant\mathbf{E}\eta_1^2\times\mathbf{E}\tau<\infty,$$

所以

$$\mathbf{E}\sup_n S_{\tau(n)}^2\leqslant 4\mathbf{E}\eta_1^2\times\mathbf{E}\tau<\infty,$$

而这正是需要证明的. □

系 设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量序列, 且 $\mathbf{P}\{\xi_i=1\}=\mathbf{P}\{\xi_i=-1\}=1/2, S_n=\xi_1+\cdots+\xi_n$, 而 $\tau=\inf\{n\geqslant 1:S_n=1\}$. 那么, $\mathbf{P}\{\tau<\infty\}=1$ (例如, 参见第一章 §9 的 (20) 式), 因此 $\mathbf{P}\{S_\tau=1\}=1, \mathbf{E}S_\tau=1$. 从而, 由 (13) 式可见 $\mathbf{E}\tau=\infty$.

定理 4 (瓦尔德基本恒等式) 设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量序列, 且 $S_n=\xi_1+\cdots+\xi_n, n\geqslant 1$. 假设 $\varphi(t)=\mathbf{E}e^{t\xi_1}, t\in\mathbb{R}$, 而且对于某个 $t_0\neq 0, \varphi(t_0)$ 存在, 并且 $\varphi(t_0)\geqslant 1$.

如果 $\tau(\tau\geqslant 1)$ (关于 $\mathscr{F}_n^\xi=\sigma(\xi_1,\cdots,\xi_n), n\geqslant 1$) 是停时, 使 $|S_n|\leqslant C(\{\tau\geqslant n\};\mathbf{P}-\text{a.c.})$, 而且 $\mathbf{E}\tau<\infty$, 则

$$\mathbf{E}\left[\frac{e^{t_0S_\tau}}{[\varphi(t_0)]^\tau}\right]=1. \tag{15}$$

证明 记

$$Y_n = e^{t_0 S_n}[\varphi(t_0)]^{-n}.$$

那么, $Y=(Y_n,\mathscr{F}^{\xi}_n)_{n\geqslant 1}$ 是鞅, 其中 $\mathbf{E}Y_n=1$, 而在集合 $\{\tau=n\}$ 上, 有

$$\begin{aligned}\mathbf{E}[|Y_{n+1}-Y_n||Y_1,\cdots,Y_n] &= Y_n\mathbf{E}\left\{\left|\frac{e^{t_0\xi_{n+1}}}{\varphi(t_0)}-1\right|\Big|\xi_1,\cdots,\xi_n\right\}\\ &= Y_n\mathbf{E}\left|e^{t_0\xi_1}[\varphi(t_0)]^{-1}-1\right|\leqslant B<\infty,\end{aligned}$$

其中 B 是某一常数. 于是, 由于 $\mathbf{E}Y_n=1$, 故由定理 2, 得 (15) 式. □

例 1 这个例子在于演示, 将上述结果用于博弈问题, 包括求破产概率和平均博弈时间 (见第一章 §9).

设 $\xi_1,\xi_2,\cdots$ 是独立伯努利随机变量序列: $\mathbf{P}\{\xi_i=1\}=p,\mathbf{P}\{\xi_i=-1\}=q,p+q=1$, 而 $S_n=\xi_1+\cdots+\xi_n$, 且

$$\tau=\inf\{n\geqslant 1:S_n=B\text{ 或 }A\},\tag{16}$$

其中 $(-A)$ 和 B 是正整数.

由第一章 §9 的 (20) 式可见, $\mathbf{P}\{\tau<\infty\}=1$ 和 $\mathbf{E}\tau<\infty$. 那么, 如果 $\alpha=\mathbf{P}\{S_\tau=A\},\beta=\mathbf{P}\{S_\tau=B\}$, 则 $\alpha+\beta=1$, 并且当 $p=q=1/2$ 时, 由 (13) 式, 可见

$$0=\mathbf{E}S_\tau=\alpha A+\beta B,$$

因此

$$\alpha=\frac{B}{B+|A|},\quad \beta=\frac{|A|}{B+|A|}.$$

利用 (14) 式, 得

$$\mathbf{E}\tau=\mathbf{E}S_\tau^2=\alpha A^2+\beta B^2=|AB|.$$

假如 $p\neq q$, 那么, 若考虑 $((q/p)^{S_n})_{n\geqslant 1}$ 鞅, 则可得

$$\mathbf{E}\left(\frac{q}{p}\right)^{S_\tau}=\mathbf{E}\left(\frac{q}{p}\right)^{S_1}=1,$$

因此

$$\alpha\left(\frac{q}{p}\right)^{A}+\beta\left(\frac{q}{p}\right)^{B}=1.$$

由此连同等式 $\alpha+\beta=1$, 得

$$\alpha=\frac{\left(\frac{q}{p}\right)^B-1}{\left(\frac{q}{p}\right)^B-\left(\frac{q}{p}\right)^A},\quad \beta=\frac{1-\left(\frac{q}{p}\right)^A}{\left(\frac{q}{p}\right)^B-\left(\frac{q}{p}\right)^A}.\tag{17}$$

注意到 $\mathbf{E}S_\tau=(p-q)\mathbf{E}\tau$, 最后得

$$\mathbf{E}\tau=\frac{\mathbf{E}S_\tau}{p-q}=\frac{\alpha A+\beta B}{p-q},$$

其中 α 和 β 由 (17) 式决定.

例 2 假设在例 1 中 $p=q=1/2$. 证明, 对于任意 $\lambda,0<\lambda<\pi/(B+|A|)$ 和由 (16) 式决定的停时 τ,

$$\mathbf{E}(\cos\lambda)^{-\tau}=\frac{\cos\left(\lambda\frac{B+A}{2}\right)}{\cos\left(\lambda\frac{B+|A|}{2}\right)}.\tag{18}$$

为此, 考虑鞅 $X=(X_n,\mathscr{F}^{\xi}_n)_{n\geqslant 0}$, 其中

$$X_n=(\cos\lambda)^{-n}\cos\left[\lambda\left(S_n-\frac{B+A}{2}\right)\right],\tag{19}$$

而 $S_0=0$. 显然

$$\mathbf{E}X_n=\mathbf{E}X_0=\mathbf{E}\cos\left(\lambda\frac{B+A}{2}\right),\tag{20}$$

现在证明, 变量族 $\{X_{n\wedge\tau}\}$ 一致可积. 为此注意到, 由于定理 1 的系 1, 对于任意 $\lambda,0<\lambda<\pi/(B+|A|)$, 有

$$\begin{aligned}\mathbf{E}X_0=\mathbf{E}X_{n\wedge\tau}&=\mathbf{E}(\cos\lambda)^{-(n\wedge\tau)}\cos\left[\lambda\left(S_{n\wedge\tau}-\frac{B+A}{2}\right)\right]\\&\geqslant\mathbf{E}(\cos\lambda)^{-(n\wedge\tau)}\cos\left(\lambda\frac{B-A}{2}\right).\end{aligned}$$

因此, 由 (20) 式, 可见

$$\mathbf{E}(\cos\lambda)^{-(n\wedge\tau)}\leqslant\frac{\cos\left(\lambda\frac{B+A}{2}\right)}{\cos\left(\lambda\frac{B+|A|}{2}\right)},$$

故根据法图引理, 有

$$\mathbf{E}(\cos\lambda)^{-\tau}\leqslant\frac{\cos\left(\lambda\frac{B+A}{2}\right)}{\cos\left(\lambda\frac{B+|A|}{2}\right)}.\tag{21}$$

从而, 根据 (19) 式

$$|X_{n\wedge\tau}|=(\cos\lambda)^{-\tau},$$

故连同 (21) 式便证明了变量族 $\{X_{n\wedge\tau}\}$ 一致可积. 那么, 由定理 1 的系 2, 可见

$$\cos\left(\lambda\frac{B+A}{2}\right)=\mathbf{E}X_0=\mathbf{E}X_\tau=\mathbf{E}(\cos\lambda)^{-\tau}\cos\left(\lambda\frac{B-A}{2}\right),$$

于是, 等式 (18) 得证.

4. 更新理论的基本定理 作为瓦尔德恒等式 (13) 的一种应用, 我们证明更新理论的所谓基本定理: 如果 $N=(N_t)_{t\geqslant 0}$ 是更新过程, 其中

$$N_t=\sum_{n=1}^{\infty} I(T_n\leqslant t),\quad T_n=\sigma_1+\cdots+\sigma_n,$$

而 $\sigma_1,\sigma_2,\cdots$ 是独立同分布正值随机变量序列 (见第二章 §9 第 4 小节), $\mu=\mathbf{E}\sigma_1<\infty$, 则更新函数 $m(t)=\mathbf{E}N_t$ 具有如下性质:

$$\frac{m(t)}{t}\to\frac{1}{\mu},\quad t\to\infty. \tag{22}$$

回忆, 过程 $N=(N_t)_{t\geqslant 0}$ 本身服从强大数定律:

$$\frac{N_t}{t}\to\frac{1}{\mu}(\mathrm{P}-\text{a.c.}),\quad t\to\infty.$$

(例如, 参见第四章 §3 的例 4.)

为证明 (22) 式, 只需证明

$$\varliminf_{t\to\infty}\frac{m(t)}{t}\geqslant\frac{1}{\mu}\quad 和\quad \varlimsup_{t\to\infty}\frac{m(t)}{t}\leqslant\frac{1}{\mu}. \tag{23}$$

为此, 注意到

$$T_{N_t}\leqslant t<T_{N_t+1}.\quad t>0. \tag{24}$$

因为对于任意 $n\geqslant 1$,

$$\{N_t+1\leqslant n\}=\{N_t\leqslant n-1\}=\{N_t<n\}=\{T_n>t\}=\left\{\sum_{k=1}^{n}\sigma_k>t\right\}\in\mathscr{F}_n,$$

其中 $\mathscr{F}_n$ 是由变量 $\sigma_1,\sigma_2,\cdots$ 诱导的 $\sigma-$ 代数, $\mu=\mathbf{E}\sigma_1<\infty$, 则 (对于每一个固定的 $t>0$) 时刻 N_t+1 (但不是 N_t) 是马尔可夫时间. 那么, 由瓦尔德恒等式 (13), 可以得到

$$\mathbf{E}T_{N_t+1}=\mu[m(t)+1], \tag{25}$$

因此, 由 (24) 式的第一个不等式, 得

$$t<\mu[m(t)+1],\quad 即\quad \frac{m(t)}{t}>\frac{1}{\mu}-\frac{1}{t}, \tag{26}$$

故当 $t\to\infty$ 时得 (23) 式的第一个不等式.

其次, 由不等式 (24) 的右侧可见, $t\geqslant\mathbf{E}T_{N_t}$. 由于 $T_{N_t+1}=T_{N_t}+\sigma_{N_t+1}$, 则

$$t\geqslant\mathbf{E}T_{N_t}=\mathbf{E}(T_{N_t+1}-\sigma_{N_t+1})=\mu[m(t)+1]-\mathbf{E}\sigma_{N_t+1}. \tag{27}$$

如果假设变量 σ_i 都有上界 $(\sigma_i\leqslant c)$, 则由 (27) 得 $t\geqslant\mu[m(t)+1]-c$, 因而

$$\frac{m(t)}{t}\leqslant\frac{1}{\mu}-\frac{1}{t}\cdot\frac{c-\mu}{\mu}. \tag{28}$$

那么, 由此得 (23) 式得第二个不等式.

为去掉限制 $\sigma_i\leqslant c, i\geqslant 1$, 对于某一 $c>0$, 引进变量

$$\sigma_i^c=\sigma_i I(\sigma_i<c)+cI(\sigma_i\geqslant c),$$

并且将其与更新过程 $N^c=(N_t^c)_{t\geqslant 0}$ 相联系, 其中

$$N_t^c=\sum_{n=1}^{\infty} I(T_n^c\leqslant t),\quad T_n^c=\sigma_1^c+\cdots+\sigma_n^c.$$

由于 $\sigma_i^c\leqslant\sigma_i(i\geqslant 1)$, 则 $N_t^c\geqslant N_t$, 因而 $m^c(t)=\mathbf{E}N_t^c\geqslant\mathbf{E}N_t=m(t)$. 那么, 由 (28) 式可见

$$\frac{m(t)}{t}\leqslant\frac{m^c(t)}{t}\leqslant\frac{1}{\mu^c}+\frac{1}{t}\cdot\frac{c-\mu^c}{\mu^c}.$$

其中 $\mu^c=\mathbf{E}\sigma_1^c$.

从而

$$\varlimsup_{t\to\infty}\frac{m(t)}{t}\leqslant\frac{1}{\mu^c}.$$

现在设 $c\to\infty$, 并考虑到当 $c\to\infty$ 时 $\mu^c\to\mu$, 由此得 (23) 式得第二个不等式.

于是, 性质 (22) 得证.

注 关于更新理论更一般的结果, 例如, 可以参见 [7, 第 9 章],[69, 卷 1, 第 XIII 章].

5. 练习题

1. 对于下鞅的情形, 证明, 如果将条件 (4) 换成条件

$$\varliminf_{n\to\infty}\int_{\{\tau_2>n\}}X_n^+\mathrm{d}\mathbf{P}=0,$$

则定理 1 仍然成立.

2. 设 $X=(X_n,\mathscr{F}_n)_{n\geqslant 0}$ 是平方可积鞅, τ 是停时, $\mathbf{E}X_0=0$,

$$\varliminf_{n\to\infty}\int_{\{\tau_2>n\}}X_n^2\mathrm{d}\mathbf{P}=0.$$

证明

$$\mathbf{E}X_\tau^2=\mathbf{E}\langle X\rangle_\tau\quad\left(=\mathbf{E}\sum_{j=0}^{\tau}(\Delta X_j)^2\right),$$

其中 $\Delta X_0=X_0,\Delta X_j=X_j-X_{j-1}, j\geqslant 1$.

3. 证明, 对于每一个鞅 $X=(X_n,\mathscr{F}_n)_{n\geqslant 0}$ 或非负下鞅, 以及停时 τ, 有

$$\mathbf{E}|X_\tau|\leqslant \varliminf_{n\to\infty}\mathbf{E}|X_n|.$$

4. 设 $X=(X_n,\mathscr{F}_n)_{n\geqslant 0}$ 是这样的上鞅, 使 $X_n\geqslant \mathbf{E}(\xi|\mathscr{F}_n)(\mathbf{P}-\text{a.c.}), n\geqslant 0$, 其中 $\mathbf{E}|\xi|<\infty$. 证明, 如果 τ_1 和 τ_2 是停时, 且 $\mathbf{P}\{\tau_1\leqslant\tau_2\}=1$, 则

$$X_{\tau_1}\geqslant \mathbf{E}(X_{\tau_2}|\mathscr{F}_{\tau_1})\ \ (\mathbf{P}-\text{a.c.}).$$

5. 设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量序列, 且 $\mathbf{P}\{\xi_i=1\}=\mathbf{P}\{\xi_i=-1\}=1/2$, 而 a 和 $b(b>a)$ 是正数,

$$X_n=a\sum_{k=1}^n I(\xi_k=+1)-b\sum_{k=1}^n I(\xi_k=-1),$$

而

$$\tau=\inf\{n\geqslant 1: X_n\leqslant -r\},\quad r>0.$$

证明: 当 $\lambda\leqslant\alpha_0$ 时, $\mathbf{E}\mathrm{e}^{\lambda\tau}<\infty$, 而当 $\lambda>\alpha_0$ 时, $\mathbf{E}\mathrm{e}^{\lambda\tau}=\infty$, 其中

$$\alpha_0=\frac{b}{a+b}\ln\frac{2b}{a+b}+\frac{a}{a+b}\ln\frac{2a}{a+b}.$$

6. 设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量序列, 且 $\mathbf{E}\xi_i=0, \mathbf{D}\xi_i=\sigma_i^2, S_n=\xi_1+\cdots+\xi_n, \mathscr{F}_n^\xi=\sigma(\xi_1,\cdots,\xi_n)$. 对于推广瓦尔德恒等式 (13) 和 (14), 证明如下论断的正确性:

1) 如果 $\mathbf{E}\sum\limits_{j=1}^{\tau}\mathbf{E}|\xi_j|<\infty$, 则 $\mathbf{E}S_\tau=0$;

2) 如果 $\mathbf{E}\sum\limits_{j=1}^{\tau}\mathbf{E}\xi_j^2<\infty$, 则

$$\mathbf{E}S_\tau^2=\mathbf{E}\sum_{j=1}^{\tau}\xi_j^2=\mathbf{E}\sum_{j=1}^{\tau}\sigma_j^2. \tag{29}$$

7. 设 $X=(X_n,\mathscr{F}_n)_{n\geqslant 1}$ 是平方可积鞅, 而 τ 是停时. 证明

$$\mathbf{E}X_\tau^2\leqslant\mathbf{E}\sum_{n=1}^{\tau}(\Delta X_n)^2.$$

证明, 如果

$$\varliminf_{n\to\infty}\mathbf{E}[X_n^2 I(\tau>n)]<\infty\quad \text{或}\quad \varliminf_{n\to\infty}\mathbf{E}[|X_n|I(\tau>n)]=0,$$

则

$$\mathbf{E}(\Delta X_\tau)^2=\mathbf{E}\sum_{n=1}^{\tau}X_n^2.$$

8. 设 $X=(X_n,\mathscr{F}_n)_{n\geqslant 1}$ 是下鞅, 而对于停时 $\tau_1\leqslant\tau_2\leqslant\cdots$, $\mathbf{E}X_{\tau_m}$ 有定义. 证明

$$\varliminf_{n\to\infty}\mathbf{E}[X_n^+(\tau_m>n)]=0,\quad m\geqslant 1.$$

证明序列 $(X_{\tau_m},\mathscr{F}_{\tau_m})_{m\geqslant 1}$ 是下鞅 (通常 $\mathscr{F}_{\tau_m}=\{A\in\mathscr{F}: A\bigcap\{\tau_m=j\}\in\mathscr{F}_j, j\geqslant 1\}$).

§3. 一些基本不等式

1. 概率的最大不等式和 L^p 中的最大不等式 如果 $X=(X_n,\mathscr{F}_n)_{n\geqslant 0}$ 是随机序列,

$$X_n^*=\max_{0\leqslant j\leqslant n}|X_j|,\quad \|X_n\|_p=(\mathbf{E}|X_n|^p)^{1/p},\quad p>0.$$

下面是属于杜布的三个定理: 定理 1 ~定理 3 . 这些定理, 对于下鞅、上鞅和鞅, 给出了基本的 "概率的最大不等式" 和 " L^p 中的最大不等式".

定理 1 Ⅰ. 设 $X=(X_n,\mathscr{F}_n)_{n\geqslant 0}$ 是下鞅, 则对于任意 $\lambda>0$, 有

$$\lambda\mathbf{P}\left\{\max_{k\leqslant n}X_k\geqslant\lambda\right\}\leqslant\mathbf{E}\left[X_n^+I\left(\max_{k\leqslant n}X_k\geqslant\lambda\right)\right]\leqslant\mathbf{E}X_n^+, \tag{1}$$

$$\lambda\mathbf{P}\left\{\min_{k\leqslant n}X_k\leqslant-\lambda\right\}\leqslant\mathbf{E}\left[X_nI\left(\min_{k\leqslant n}X_k>-\lambda\right)\right]-\mathbf{E}X_0\leqslant\mathbf{E}X_n^+-\mathbf{E}X_0, \tag{2}$$

$$\lambda\mathbf{P}\left\{\max_{k\leqslant n}|X_k|\geqslant\lambda\right\}\leqslant 3\max_{k\leqslant n}\mathbf{E}|X_k|. \tag{3}$$

Ⅱ. 设 $Y=(Y_n,\mathscr{F}_n)_{n\geqslant 0}$ 是上鞅, 则对于任意 $\lambda>0$, 有

$$\lambda\mathbf{P}\left\{\max_{k\leqslant n}Y_k\geqslant\lambda\right\}\leqslant\mathbf{E}Y_0-\mathbf{E}\left[Y_nI\left(\max_{k\leqslant n}Y_k<\lambda\right)\right]\leqslant\mathbf{E}Y_0+\mathbf{E}Y_n^-, \tag{4}$$

$$\lambda\mathbf{P}\left\{\min_{k\leqslant n}Y_k\leqslant-\lambda\right\}\leqslant-\mathbf{E}\left[Y_nI\left(\min_{k\leqslant n}Y_k\leqslant-\lambda\right)\right]\leqslant\mathbf{E}Y_n^- \tag{5}$$

$$\lambda\mathbf{P}\left\{\max_{k\leqslant n}|Y_k|\geqslant\lambda\right\}\leqslant 3\max_{k\leqslant n}\mathbf{E}|Y_k|. \tag{6}$$

Ⅲ. 设 $X=(X_n,\mathscr{F}_n)_{n\geqslant 0}$ 是非负上鞅, 则对于任意 $\lambda>0$, 有

$$\lambda\mathbf{P}\left\{\max_{k\leqslant n}Y_k\geqslant\lambda\right\}\leqslant\mathbf{E}Y_0, \tag{7}$$

$$\lambda\mathbf{P}\left\{\sup_{k\geqslant n}Y_k\geqslant\lambda\right\}\leqslant\mathbf{E}Y_n. \tag{8}$$

定理 2 设 $X=(X_n,\mathscr{F}_n)_{n\geqslant 0}$ 是非负下鞅, 则对于任意 $p\geqslant 1$, 下列不等式成立:

如果 $p>1$, 则

$$\|X_n\|_p \leqslant \|X_n^*\|_p \leqslant \frac{p}{p-1}\|X_n\|_p; \tag{9}$$

如果 $p=1$, 则

$$\|X_n\|_1 \leqslant \|X_n^*\|_1 \leqslant \frac{\mathrm{e}}{\mathrm{e}-1}\{1+\|X_n \ln^+ X_n\|_1\}. \tag{10}$$

定理 3 设 $X=(X_n,\mathscr{F}_n)_{n\geqslant 0}$ 是鞅, 而 $\lambda>0, p\geqslant 1$, 则

$$\mathbf{P}\left\{\max_{k\leqslant n}|X_k|\geqslant\lambda\right\}\leqslant\frac{\mathbf{E}|X_n|^p}{\lambda^p}. \tag{11}$$

而如果 $p>1$, 则

$$\|X_n\|_p \leqslant \|X_n^*\|_p \leqslant \frac{p}{p-1}\|X_n\|_p. \tag{12}$$

特别, 当 $p=2$ 时, 有

$$\mathbf{P}\left\{\max_{k\leqslant n}|X_k|\geqslant\lambda\right\}\leqslant\frac{\mathbf{E}|X_n|^2}{\lambda^2}, \tag{13}$$

$$\mathbf{E}\left[\max_{k\leqslant n}X_k^2\right]\leqslant 4\mathbf{E}|X_n|^2. \tag{14}$$

证明定理 1 由于带相反符号的下鞅是上鞅, 故由不等式 (4)~(6) 得不等式 (1)~(3). 因此, 我们考虑上鞅 $Y=(Y_n,\mathscr{F}_n)_{n\geqslant 0}$ 的情形.

设 $\tau=\inf\{k\leqslant n: Y_k\geqslant\lambda\}$, 如果 $\max\limits_{k\leqslant n}Y_k<\lambda$, 则认为 $\tau=n$. 那么, 由 §2 性质 (6) 可见,

$$\begin{aligned}\mathbf{E}Y_0\geqslant\mathbf{E}Y_\tau&=\mathbf{E}\left[Y_\tau;\max_{k\leqslant n}Y_k\geqslant\lambda\right]+\mathbf{E}\left[Y_\tau;\max_{k\leqslant n}Y_k<\lambda\right]\\&\geqslant\lambda\mathbf{P}\left[\max_{k\leqslant n}Y_k\geqslant\lambda\right]+\mathbf{E}\left[Y_n;\max_{k\leqslant n}Y_k<\lambda\right],\end{aligned}$$

从而, 证明了 (4) 式.

现在, 设 $\sigma=\inf\{k\leqslant n: Y_k\leqslant-\lambda\}$, 如果 $\max\limits_{k\leqslant n}Y_k>-\lambda$, 则认为 $\sigma=n$. 那么, 仍然由 §2 性质 (6) 可见,

$$\begin{aligned}\mathbf{E}Y_n\leqslant\mathbf{E}Y_\tau&=\mathbf{E}\left[Y_\tau;\min_{k\leqslant n}Y_k\leqslant-\lambda\right]+\mathbf{E}\left[Y_\tau;\min_{k\leqslant n}Y_k>-\lambda\right]\\&\leqslant-\lambda\mathbf{P}\left[\min_{k\leqslant n}Y_k\leqslant-\lambda\right]+\mathbf{E}\left[Y_n;\min_{k\leqslant n}Y_k>-\lambda\right].\end{aligned}$$

由此, 得

$$\lambda\mathbf{P}\left[\min_{k\leqslant n}Y_k\leqslant-\lambda\right]\leqslant-\mathbf{E}\left[Y_n;\min_{k\leqslant n}Y_k\leqslant-\lambda\right]\leqslant\mathbf{E}Y_n^-,$$

从而, 证明了 (5) 式.

为证明不等式 (6), 注意到 $Y^-=(-Y)^+$ 是下鞅, 从而根据 (4) 式和 (1) 式, 有

$$\begin{aligned}\lambda\mathbf{P}\left[\max_{k\leqslant n}|Y_k|\geqslant\lambda\right]&\leqslant\lambda\mathbf{P}\left[\max_{k\leqslant n}Y_k^+\geqslant\lambda\right]+\lambda\mathbf{P}\left[\max_{k\leqslant n}Y_k^-\geqslant\lambda\right]\\&=\lambda\mathbf{P}\left\{\max_{k\leqslant n}Y_k\geqslant\lambda\right\}+\lambda\mathbf{P}\left\{\max_{k\leqslant n}Y_k^-\geqslant\lambda\right\}\\&\leqslant\mathbf{E}Y_0+2\mathbf{E}Y_n^-\leqslant 3\max_{k\leqslant n}\mathbf{E}|Y_k|.\end{aligned}$$

由 (4) 式得不等式 (7).

为证明不等式 (8), 如果当 $k\geqslant n$ 时 $Y_k<\lambda$, 则设 $\gamma=\inf\{k\geqslant n: Y_k\geqslant\lambda\}$, 认为 $\gamma=\infty$. 又设 $n<N<\infty$. 那么, 由 §2 (6) 式, 有

$$\mathbf{E}Y_n\geqslant\mathbf{E}Y_{\gamma\wedge N}\geqslant\mathbf{E}[Y_{\gamma\wedge N}I(\gamma\leqslant N)]\geqslant\lambda\mathbf{P}\{\gamma\leqslant N\},$$

于是, 当 $N\to\infty$ 时, 得

$$\mathbf{E}Y_n\geqslant\lambda\mathbf{P}\{\gamma<\infty\}=\lambda\mathbf{P}\left\{\sup_{k\geqslant n}Y_k\geqslant n\right\}.$$ □

证明定理 2 (9) 式和 (10) 式的前两个不等式显然.

为证明 (9) 式的后两个不等式, 先假设

$$\|X_n^*\|_p<\infty, \tag{15}$$

并且利用如下事实: 对于任意非负随机变量 ξ 和 $r>0$, 有

$$\mathbf{E}\xi^r=r\int_0^\infty t^{r-1}\mathbf{P}\{\xi\geqslant t\}\mathrm{d}t. \tag{16}$$

那么, 对于 $p>1$, 由不等式 (1) 和傅比尼 (G. Fubini) 定理, 可见

$$\begin{aligned}\mathbf{E}(X_n^*)^p&=p\int_0^\infty t^{p-1}\mathbf{P}\{X_n^*\geqslant t\}\mathrm{d}t\leqslant p\int_0^\infty t^{p-2}\left(\int_{\{X_n^*\geqslant t\}}X_n\mathrm{d}\mathbf{P}\right)\mathrm{d}t\\&=p\int_0^\infty t^{p-2}\left[\int_\Omega X_nI\{X_n^*\geqslant t\}\mathrm{d}\mathbf{P}\right]\mathrm{d}t=p\int_\Omega X_n\left[\int_0^{X_n^*}t^{p-2}\mathrm{d}t\right]\mathrm{d}\mathbf{P}\\&=\frac{p}{p-1}\mathbf{E}[X_n(X_n^*)^{p-1}].\end{aligned} \tag{17}$$

从而, 根据赫尔德 (O. L. Hölder) 不等式 (第二章 §6 第 7 小节), 有

$$\mathbf{E}(X_n^*)^p\leqslant q\|X_n\|_p\cdot\|(X_n^*)^{p-1}\|_q=q\|X_n\|_p[\mathbf{E}(X_n^*)^p]^{1/q}, \tag{18}$$

其中 $q=p/(p-1)$.

假如不等式 (15) 成立, 则由 (18) 式立即得到 (9) 式的第二个不等式.

假如不等式 (15) 不成立, 则应该按如下方式操作. 在 (17) 式中不是考虑 X_n^* 而是考虑变量 $(X_n^*\wedge L)$, 其中 L 是某一常数. 那么, 有

$$\mathbf{E}(X_n^*\wedge L)^p\leqslant q\mathbf{E}[X_n(X_n^*\wedge L)^{p-1}]\leqslant q\|X_n\|_p[\mathbf{E}(X_n^*\wedge L)^p]^{1/q},$$

因此, 由不等式 $\mathbf{E}(X_n^*\wedge L)^p\leqslant L^p<\infty$ 可见

$$\mathbf{E}(X_n^*\wedge L)^p\leqslant q^p\mathbf{E}X_n^p=q^p\|X_n\|_p^p,$$

从而

$$\mathbf{E}(X_n^*)^p=\lim_{L\to\infty}\mathbf{E}(X_n^*\wedge L)^p=q^p\|X_n\|_p^p.$$

现在证明 (10) 式的第二个不等式.

仍然利用 (1) 式, 有

$$\mathbf{E}X_n^*-1\leqslant\mathbf{E}(X_n^*-1)^+=\int_0^\infty\mathbf{P}\{X_n^*-1\geqslant t\}\mathrm{d}t$$
$$\leqslant\int_0^\infty\frac{1}{1+t}\left[\int_{\{X_n^*\geqslant 1+t\}}X_n\mathrm{d}\mathbf{P}\right]\mathrm{d}t=\mathbf{E}X_n\int_0^{X_n^*-1}\frac{\mathrm{d}t}{1+t}=\mathbf{E}X_n\cdot\ln X_n^*.$$

由于对于任何 $a\geqslant 0$ 和 $b>0$, 有

$$a\ln b\leqslant a\ln^+ a+b\mathrm{e}^{-1},\tag{19}$$

因此

$$\mathbf{E}X_n^*-1\leqslant\mathbf{E}X_n\ln X_n^*\leqslant\mathbf{E}X_n\ln^+X_n+\mathrm{e}^{-1}\mathbf{E}X_n^*.$$

如果 $\mathbf{E}X_n^*<\infty$, 则由此立即得 (10) 式的第二个不等式.

如果 $\mathbf{E}X_n^*=\infty$, 则像上面一样将 X_n^* 换成 $X_n^*\wedge L$.

证明定理 3 由于 (若 $\mathbf{E}|X|^p<\infty, n\geqslant 0$) $|X|^p(p\geqslant 1)$是非负下鞅, 故由 (1) 式和 (9) 式可见定理 3 成立.

定理 3 的系 设 $X_n=\xi_0+\cdots+\xi_n, n\geqslant 0$, 其中 $(\xi_k)_{k\geqslant 0}$ 是独立随机变量序列, 且 $\mathbf{E}\xi_k=0, \mathbf{E}\xi_k^2<\infty$. 那么, 不等式 (13) 就是柯尔莫戈洛夫不等式 (第四章 §2).

2. 最大概率及 L^p 中最大范数的估计式 设 $X=(X_n,\mathscr{F}_n)_{n\geqslant 0}$ 是非负下鞅, 而

$$X_n=M_n+A_n$$

是杜布分解 (§1 的 (11) 式). 那么, 由于 $\mathbf{E}M_n=0$, 故由 (1) 式可见

$$\mathbf{P}\{X_n^*\geqslant\varepsilon\}\leqslant\frac{\mathbf{E}A_n}{\varepsilon}.$$

下面的定理 4 证明此不等式不仅对于下鞅成立, 而且对于控制性的更广泛的一类序列成立. 下面的定义说明了这里所谓控制性 (优势性) 的含义.

定义 设 $X=(X_n,\mathscr{F}_n)$ 是某一非负随机序列, 而 $A=(A_n,\mathscr{F}_{n-1})$ 是一递增可预测序列. 称序列 A 控制序列 X (或 A 是 X 的优势序列), 如果对于任何停时 τ, 有

$$\mathbf{E}X_\tau\leqslant\mathbf{E}A_\tau.\tag{20}$$

定理 4 设 $A=(A_n,\mathscr{F}_{n-1})$ 是一递增可预测序列, 而 $X=(X_n,\mathscr{F}_n)$ 是被序列 A 控制的非负随机序列, 那么, 对于 $\lambda>0, a>0$ 及任何停时 τ, 有.

$$\mathbf{P}\{X_\tau^*\geqslant\lambda\}\leqslant\frac{\mathbf{E}A_\tau}{\lambda},\tag{21}$$

$$\mathbf{P}\{X_\tau^*\geqslant\lambda\}\leqslant\frac{1}{\lambda}\mathbf{E}(A_\tau\wedge a)+\mathbf{P}\{A_\tau\geqslant a\},\tag{22}$$

$$\|X_\tau^*\|_p\leqslant\left(\frac{2-p}{1-p}\right)^{1/p}\|A_\tau\|_p,\quad 0<p<1.\tag{23}$$

证明 设

$$\sigma_n=\min\{j\leqslant\tau\wedge n: X_j\geqslant\lambda\},$$

而若 $\{\cdot\}=\varnothing$, 则认为 $\sigma_n=\tau\wedge n$. 那么,

$$\mathbf{E}A_\tau\geqslant\mathbf{E}A_{\sigma_n}\geqslant\mathbf{E}X_{\sigma_n}\geqslant\int_{\{X_{\tau\wedge n}^*>\lambda\}}X_{\sigma_n}\mathrm{d}\mathbf{P}\geqslant\lambda\mathbf{P}\{X_{\tau\wedge n}^*>\lambda\},$$

因而

$$\mathbf{P}\{X_{\tau\wedge n}^*>\lambda\}\leqslant\frac{1}{\lambda}\mathbf{E}A_\tau,$$

于是, 由法图引理可得不等式 (21).

为证明 (22) 式, 引进时间

$$\gamma=\inf\{j: A_{j+1}\geqslant a\},$$

而若 $\{\cdot\}=\varnothing$, 则认为 $\gamma=\infty$. 那么,

$$\mathbf{P}\{X_\tau^*\geqslant\lambda\}=\mathbf{P}\{X_\tau^*\geqslant\lambda, A_\tau<a\}+\mathbf{P}\{X_\tau^*\geqslant\lambda, A_\tau\geqslant a\}$$
$$\leqslant\mathbf{P}\{I_{\{A_\tau<a\}}X_\tau^*\geqslant\lambda\}+\mathbf{P}\{A_\tau\geqslant a\}\leqslant\mathbf{P}\{X_{\tau\wedge\gamma}^*\geqslant\lambda\}+\mathbf{P}\{A_\tau\geqslant a\}$$
$$\leqslant\frac{1}{\lambda}\mathbf{E}A_{\tau\wedge\gamma}+\mathbf{P}\{A_\tau\geqslant a\}\leqslant\frac{1}{\lambda}\mathbf{E}(A_\tau\wedge a)+\mathbf{P}\{A_\tau\geqslant a\},$$

其中用到不等式 (21) 和 $I_{\{A_\tau<a\}}X_\tau^*\leqslant X_{\tau\wedge\gamma}^*$. 最后 (注意到 (22) 式), 由下面的一系列关系式, 得不等式 (23):

$$\|X_\tau^*\|_p^p=\mathbf{E}(X_\tau^*)^p=\int_0^\infty\mathbf{P}\{(X_\tau^*)^p\geqslant t\}\mathrm{d}t=\int_0^\infty\mathbf{P}\{X_\tau^*\geqslant t^{1/p}\}\mathrm{d}t$$
$$\leqslant\int_0^\infty t^{-1/p}\mathbf{E}[A_\tau\wedge t^{1/p}]\mathrm{d}t+\int_0^\infty\mathbf{P}\{A_\tau^p\geqslant t\}\mathrm{d}t$$
$$=\mathbf{E}\int_0^{A_\tau^p}\mathrm{d}t+\mathbf{E}\int_{A_\tau^p}^\infty A_\tau t^{-1/p}\mathrm{d}t+\mathbf{E}X_\tau^p=\frac{2-p}{1-p}\mathbf{E}X_\tau^p.\qquad\square$$

注 假设定理 4 的条件成立, 但序列 $A=(A_n,\mathscr{F}_{n\geqslant 0})$ 未必是可预测的, 不过假定对于某个常数 $c>0$, 有

$$\mathbf{P}\left\{\sup_{k\geqslant 1}|\Delta A_k|\leqslant c\right\}=1,$$

其中 $\Delta A_k=A_k-A_{k-1}$. 那么, 如下不等式成立 (对照 (22) 式):

$$\mathbf{P}\{X_\tau^*\geqslant\lambda\}\leqslant\frac{1}{\lambda}\mathbf{E}[A_\tau\wedge(a+c)]+\mathbf{P}\{A_\tau\geqslant a\}. \tag{24}$$

这一事实证明与不等式 (22) 的证明类似, 只需要把时间 $\gamma=\inf\{j:A_{j+1}\geqslant a\}$ 换成时间 $\gamma=\inf\{j:A_j\geqslant a\}$, 并且注意到 $A_\gamma\leqslant a+c$.

系 设 $X^k=(X_n^k,\mathscr{F}_n^k)$ 和 $A^k=(A_n^k,\mathscr{F}_n^k), n\geqslant 0, k\geqslant 1$, 满足定理 4 的条件, 或者满足上面的注. 假设 $(\tau^k)_{k\geqslant 1}$ 关于 $\mathscr{F}^k=(\mathscr{F}_n^k)$ 是停时序列, 而且 $A_{\tau^k}^k\xrightarrow{\mathbf{P}}0$, 那么 $(X^k)_{\tau_k}^*\xrightarrow{\mathbf{P}}0$.

3. 鞅的不等式 在这一小节将 (不加证明, 但是有其应用) 列举一系列非常好的鞅的不等式. 这些不等式, 有些是下面的介绍的辛钦不等式的推广, 有些是独立随机变量之和的马尔钦凯维奇 (J. Matcinkiewicz) 和齐格蒙特 (A. Zygmund) 不等式的推广.

辛钦不等式 设 $\xi_1,\xi_2,\cdots$ 是独立同分布伯努利随机变量, 且 $\mathbf{P}\{\xi_i=1\}=\mathbf{P}\{\xi_i=-1\}=1/2$, 而 $(c_n)_{n\geqslant 1}$ 是一数列.

那么, 对于任意 $0<p<1$ 存在 (不依赖于 (c_n) 的) 通用常数 A_p 和 B_p, 使得对于任意 $n\geqslant 1$, 有

$$A_p\left(\sum_{j=1}^n c_j^2\right)^{1/2}\leqslant\left\|\sum_{j=1}^n c_j\xi_j\right\|_p\leqslant B_p\left(\sum_{j=1}^n c_j^2\right)^{1/2}. \tag{25}$$

马尔钦凯维奇和齐格蒙特不等式 设 $\xi_1,\xi_2,\cdots$ 是独立可积随机变量序列, 且 $\mathbf{E}\xi_i=0$, 则对于任意 $p\geqslant 1$, 存在这样的 (不依赖于 (ξ_n) 的) 通用常数 A_p 和 B_p, 使得对于任意 $n\geqslant 1$, 有

$$A_p\left\|\left(\sum_{j=1}^n \xi_j^2\right)^{1/2}\right\|_p\leqslant\left\|\sum_{j=1}^n \xi_j\right\|_p\leqslant B_p\left\|\left(\sum_{j=1}^n \xi_j^2\right)^{1/2}\right\|_p. \tag{26}$$

在不等式 (25) 和 (26) 中, 序列 $X=(X_n)$ 是鞅, 其中相应地

$$X_n=\sum_{j=1}^n c_j\xi_j \quad 和 \quad X_n=\sum_{j=1}^n \xi_j.$$

自然地提出问题, 是否可以将这些不等式推广到任意鞅? 在此方向上最早的结果属于伯克霍尔德 (D. L. Burkholder).

伯克霍尔德不等式 如果 $X=(X_n,\mathscr{F}_n)$ 是鞅, 则对于任意 $p>1$, 存在这样的 (不依赖于 X 的) 通用常数 A_p 和 B_p, 使得对于任意 $n\geqslant 1$, 有

$$A_p\left\|\sqrt{[X]_n}\right\|_p\leqslant\|X_n\|_p\leqslant B_p\left\|\sqrt{[X]_n}\right\|_p, \tag{27}$$

其中 $[X]_n$ 表示 X_n 的二次变差 (第 118 页),

$$[X]_n=\sum_{j=i}^n(\Delta X_j)^2,\quad X_0=0. \tag{28}$$

特别, 常数 A_p 和 B_p 可以取为

$$A_p=[18p^{3/2}/(p-1)]^{-1} \quad 和 \quad B_p=18p^{3/2}/(p-1)^{1/2}.$$

考虑到 (12) 式, 由 (27) 式可见

$$A_p\left\|\sqrt{[X]_n}\right\|_p\leqslant\|X_n^*\|_p\leqslant B_p^*\left\|\sqrt{[X]_n}\right\|_p, \tag{29}$$

其中

$$A_p=[18p^{3/2}/(p-1)]^{-1} \quad 和 \quad B_p^*=18p^{5/2}/(p-1)^{3/2}.$$

当 $p>1$ 时, 伯克霍尔德不等式 (27) 成立; 而且马尔钦凯维奇和齐格蒙特不等式 (26), 也对于 $p=1$ 成立. 问对于 $p=1$, 不等式 (27) 是否成立? 下面的例子表明, 当 $p=1$ 时, 不等式 (27) 不能直接推广.

例 设 $\xi_1,\xi_2,\cdots$ 是独立伯努利随机变量, 而 $\mathbf{P}\{\xi_i=1\}=\mathbf{P}\{\xi_i=-1\}=1/2$, 且

$$X_n=\sum_{j=1}^{n\wedge\tau}\xi_j,\quad 其中\quad \tau=\inf\left\{n\geqslant 1:\sum_{j=1}^n\xi_j=1\right\}.$$

序列 $X=(X_n,\mathscr{F}_n)$ 是鞅, 其中

$$\|X_n\|_1=\mathbf{E}|X_n|=2\mathbf{E}X_n^+\to 2,\quad n\to\infty.$$

但是

$$\|\sqrt{[X]_n}\|_1=\mathbf{E}\sqrt{[X]_n}=\mathbf{E}\left(\sum_{j=1}^{n\wedge\tau}1\right)^{1/2}=\mathbf{E}\sqrt{\tau\wedge n}\to\infty.$$

从而, (27) 式的第一个不等式不成立.

结果表明, 对于 $p=1$ 的情形, 虽然不等式 (27) 不能推广, 然而 (对于 $p>1$, 与不等式 (27) 等价的) 不等式 (29) 可以推广.

戴维斯 (H. T. Davis) 不等式 如果 $X=(X_n,\mathscr{F}_n)$ 是鞅, 则存在这样的通用常数 A 和 $B(0<A<B<\infty)$, 使

$$A\left\|\sqrt{[X]_n}\right\|_1\leqslant\|X_n^*\|_1\leqslant B\left\|\sqrt{[X]_n}\right\|_1, \tag{30}$$

即

$$A\mathbf{E}\sqrt{\sum_{j=1}^{n}(\Delta X_j)^2} \leqslant \mathbf{E}\left[\max_{1\leqslant j\leqslant n}|X_j|\right] \leqslant B\mathbf{E}\sqrt{\sum_{j=1}^{n}(\Delta X_j)^2}.$$

系 1 设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量, $S_n = \xi_1 + \cdots + \xi_n$. 如果 $\mathbf{E}|\xi_1| < \infty, \mathbf{E}\xi_1 = 0$, 则根据 §2 中瓦尔德不等式 (13), 对于任意 (关于 $(\mathscr{F}_n^\xi)$ 的) 停时 $\tau(\tau < \infty)$, 如下等式成立:

$$\mathbf{E}S_\tau = 0. \tag{31}$$

如果补充假设 $\mathbf{E}|\xi_1|^r < \infty(1 < r \leqslant 2)$, 则等式 $\mathbf{E}S_\tau = 0$ 成立的充分条件是 $\mathbf{E}\tau^{1/r} < \infty$.

为证明系 1 的结论, 引进记号: $\tau_n = \tau \wedge n, Y = \sup\limits_n |S_{\tau_n}|$, 并且假设 $t > 0, m = [t^r]$ 是 t^r 的整数部分. 由于 §2 定理 1 的系 1, 可知 $\mathbf{E}S_{\tau_n} = 0$. 因此, 为关系式 $\mathbf{E}S_\tau = 0$ 成立, (根据控制收敛定理) 只需验证 $\mathbf{E}\sup\limits_n|S_{\tau_n}| < \infty$.

由不等式 (1) 和 (27), 有

$$\begin{aligned}\mathbf{P}\{Y \geqslant t\} &= \mathbf{P}\{\tau \geqslant t^r, Y \geqslant t\} + \mathbf{P}\{\tau < t^r, Y \geqslant t\} \\ &\leqslant \mathbf{P}\{\tau \geqslant t^r\} + \mathbf{P}\left\{\max_{1\leqslant j\leqslant m}|S_{\tau_j}| \geqslant t\right\} \leqslant \mathbf{P}\{\tau \geqslant t^r\} + t^{-r}\mathbf{E}|S_{\tau_m}|^r \\ &\leqslant \mathbf{P}\{\tau \geqslant t^r\} + t^{-r}B_r^r\mathbf{E}\left(\sum_{j=1}^{\tau_m}\xi_j^2\right)^{r/2} \leqslant \mathbf{P}\{\tau \geqslant t^r\} + t^{-r}B_r^r\mathbf{E}\sum_{j=1}^{\tau_m}|\xi_j|^r.\end{aligned}$$

注意到, (记 $\mathscr{F}_0^\xi = \{\varnothing, \Omega\}$)

$$\begin{aligned}\mathbf{E}\sum_{j=1}^{\tau_m}|\xi_j|^r &= \mathbf{E}\sum_{j=1}^{\infty}I(j \leqslant \tau_m)|\xi_j|^r = \sum_{j=1}^{\infty}\mathbf{E}\mathbf{E}\left[I(j \leqslant \tau_m)|\xi_j|^r\middle|\mathscr{F}_{j-1}^\xi\right] \\ &= \mathbf{E}\sum_{j=1}^{\infty}I(j \leqslant \tau_m)\mathbf{E}\left[|\xi_j|^r\middle|\mathscr{F}_{j-1}^\xi\right] = \mathbf{E}\sum_{j=1}^{\tau_m}\mathbf{E}|\xi_j|^r = \mu_r\mathbf{E}\tau_m,\end{aligned}$$

其中 $\mu_r = \mathbf{E}|\xi_1|^r$. 因此,

$$\begin{aligned}\mathbf{P}\{Y \geqslant t\} &\leqslant \mathbf{P}\{\tau \geqslant t^r\} + t^{-r}B_r^r\mu_r\mathbf{E}\tau_m \\ &= \mathbf{P}\{\tau \geqslant t^r\} + B_r^r\mu_r t^{-r}\left[m\mathbf{P}\{\tau \geqslant t^r\} + \int_{\{\tau<t^r\}}\tau\mathrm{d}\mathbf{P}\right] \\ &\leqslant (1 + B_r^r\mu_r)\mathbf{P}\{\tau \geqslant t^r\} + B_r^r\mu_r t^{-r}\int_{\{\tau<t^r\}}\tau\mathrm{d}\mathbf{P},\end{aligned}$$

于是,

$$\begin{aligned}\mathbf{E}Y &= \int_0^\infty \mathbf{P}\{Y \geqslant t\}\mathrm{d}t \leqslant (1 + B_r^r\mu_r)\mathbf{E}\tau^{1/r} + B_r^r\mu_r\int_0^\infty t^{-r}\left[\int_{\{\tau<t^r\}}\tau\mathrm{d}\mathbf{P}\right]\mathrm{d}t \\ &= (1 + B_r^r\mu_r)\mathbf{E}\tau^{1/r} + B_r^r\mu_r\int_\Omega\tau\left[\int_{\tau^{1/r}}^\infty t^{-r}\mathrm{d}t\right]\mathrm{d}\mathbf{P} \\ &= \left(1 + B_r^r\mu_r + \frac{B_r^r\mu_r}{r-1}\right)\mathbf{E}\tau^{1/r} < \infty.\end{aligned}$$

系 2 设 $M = (M_n)$ 是鞅, 满足条件: 对于某一 $r \geqslant 1, \mathbf{E}|M_n|^{2r} < \infty$, 且

$$\sum_{n=1}^{\infty}\frac{\mathbf{E}|\Delta M_n|^{2r}}{n^{1+r}} < \infty, \quad (M_0 = 0). \tag{32}$$

那么, 强大数定律成立 (对照第四章 §3 定理 2):

$$\frac{M_n}{n} \to 0 \quad (\mathbf{P}-\text{a.c.}), \quad n \to \infty. \tag{33}$$

当 $r = 1$ 时, (32) 式的证明与第四章 §3 定理 2 的证明方法一样. 具体地说, 设

$$m_n = \sum_{k=1}^{n}\frac{\Delta M_k}{k}.$$

那么,

$$\frac{M_n}{n} = \frac{\sum\limits_{k=1}^{n}\Delta M_k}{n} = \frac{1}{n}\sum_{k=1}^{n}k\Delta m_k,$$

而根据克罗内克引理 (第四章 §3), 级数收敛 $(\mathbf{P}-\text{a.c.})$

$$\frac{1}{n}\sum_{k=1}^{n}k\Delta m_k \to 0, \quad n \to \infty,$$

的充分条件是, 存在有限极限 $\lim\limits_n m_n(\mathbf{P}-\text{a.c.})$. 而极限 $\lim\limits_n m_n(\mathbf{P}-\text{a.c.})$ 存在, 当且仅当对于任意 $\varepsilon > 0$ (第二章 §10 定理 1 和定理 4), 有

$$\mathbf{P}\left\{\sup_{k\geqslant 1}|m_{n+k} - m_n| \geqslant \varepsilon\right\} \to 0, \quad n \to \infty. \tag{34}$$

由于不等式 (1), 有

$$\mathbf{P}\left\{\sup_{k\geqslant 1}|m_{n+k} - m_n| \geqslant \varepsilon\right\} \leqslant \frac{1}{\varepsilon^2}\sum_{k=n}^{\infty}\frac{\mathbf{E}(\Delta M_k)^2}{k^2}.$$

从而, 当 $r = 1$ 时, (32) 式和 (34) 式得证.

现在假设 $r>1$. 由第二章 §10 定理 1, 可见关系式 (33) 等价于: 对于任意 $\varepsilon>0$,

$$\varepsilon^{2r}\mathbf{P}\left\{\sup_{j\geqslant n}\frac{|M_j|}{j}\geqslant\varepsilon\right\}\to 0,\quad n\to\infty. \tag{35}$$

由下面 (第 7 小节) 练习题 1 的不等式 (52), 可见

$$\varepsilon^{2r}\mathbf{P}\left\{\sup_{j\geqslant n}\frac{|M_j|}{j}\geqslant\varepsilon\right\}=\varepsilon^{2r}\lim_{m\to\infty}\mathbf{P}\left\{\max_{n\leqslant j\leqslant m}\frac{|M_j|^{2r}}{j^{2r}}\geqslant\varepsilon^{2r}\right\}$$
$$\leqslant\frac{1}{n^{2r}}\mathbf{E}|M_n|^{2r}+\sum_{j\geqslant n+1}\frac{1}{j^{2r}}\mathbf{E}(|M_j|^{2r}-|M_{j-1}|^{2r}).$$

由克罗内克引理, 可得

$$\lim_{n\to\infty}\frac{1}{n^{2r}}\mathbf{E}|M_n|^{2r}=0.$$

从而, 为证明 (35) 式只需验证

$$\sum_{j\geqslant 2}\frac{1}{j^{2r}}\mathbf{E}(|M_j|^{2r}-|M_{j-1}|^{2r})<\infty. \tag{36}$$

易见, 有

$$I_N\equiv\sum_{j=2}^{N}\frac{1}{j^{2r}}\mathbf{E}[|M_j|^{2r}-\mathbf{E}|M_{j-1}|^{2r}]$$
$$\leqslant\sum_{j=2}^{N}\left[\frac{1}{(j-1)^{2r}}-\frac{1}{j^{2r}}\right]\mathbf{E}|M_{j-1}|^{2r}+\frac{\mathbf{E}|M_N|^{2r}}{N^{2r}}.$$

由伯克霍尔德不等式 (27) 和赫尔德 (O. L. Hölder) 不等式, 可见

$$\mathbf{E}|M_j|^{2r}\leqslant B_{2r}^{2r}\mathbf{E}\left[\sum_{i=1}^{j}(\Delta M_i)^2\right]^r\leqslant B_{2r}^{2r}\mathbf{E}\sum_{i=1}^{j}(\Delta M_i)^{2r}.$$

因而

$$I_N\leqslant\sum_{j=2}^{N-1}B_{2r}^{2r}\left[\frac{1}{j^{2r}}-\frac{1}{(j+1)^{2r}}\right]j^{r-1}\sum_{i=1}^{j}\mathbf{E}|\Delta M_i|^{2r}\frac{\mathbf{E}|M_N|^{2r}}{N^{2r}}$$
$$\leqslant C_1\sum_{j=2}^{N-1}\frac{1}{j^{r+2}}\sum_{i=1}^{j}\mathbf{E}|\Delta M_i|^{2r}\frac{\mathbf{E}|M_N|^{2r}}{N^{2r}}\leqslant C_2\sum_{j=2}^{N}\frac{\mathbf{E}|\Delta M_j|^{2r}}{j^{r+1}}+C_3.$$

其中 $C_i(i=1,2,3)$ 是常数. 于是, 由于 (32) 式, 估计式 (36) 得证.

4. 下鞅的极限平均"振动"次数的上界 随机变量序列 $(X_n)_{n\geqslant1}$ 以概率 1 有 (有限或无限) 极限 $\lim X_n$, 当且仅当"在任意两个 (有理) 数 a 和 $b(a<b)$ 之间振动的"次数以概率 1 有限. 对于下鞅, 下面的定理 5 给出了"振动的"平均次数这上侧估计, 定理 5 将用来证明其收敛性的基本结果.

固定两个 (有理) 数 a 和 $b(a<b)$, 并且对于序列 $X=(X_n)_{n\geqslant1}$ 定义停时:

$$\tau_0=0,$$
$$\tau_1=\min\{n>0:X_n\leqslant a\},$$
$$\tau_2=\min\{n>\tau_1:X_n\geqslant b\},$$
$$\cdots\cdots\cdots$$
$$\tau_{2m-1}=\min\{n>\tau_{2m-2}:X_n\leqslant a\},$$
$$\tau_{2m}=\min\{n>\tau_{2m-1}:X_n\geqslant b\},$$

若相应的集合 $\{\cdot\}$ 是空集, 则设 $\tau_k=\infty$.

其次, 对于每一个 $n\geqslant1$, 定义随机变量

$$\beta_n(a,b)=\begin{cases}0, & \text{若 }\tau_2>n,\\ \max\{m:\tau_{2m}\leqslant n\}, & \text{若 }\tau_2\leqslant n.\end{cases}$$

随机变量 $\beta_n(a,b)$ 的含义是: 序列 $X_1,\cdots,X_n$ 与区间 $[a,b]$ (自下而上) 相交的次数.

定理 5 (杜布) *设 $X=(X_n,\mathscr{F}_n)_{n\geqslant1}$ 是下鞅. 那么, 对于任意 $n\geqslant1$, 有*

$$\mathbf{E}\beta_n(a,b)\leqslant\frac{\mathbf{E}|X_n-a|^+}{b-a}. \tag{37}$$

证明 下鞅 $X=(X_n,\mathscr{F}_n)$ 与区间 $[a,b]$ 相交的次数, 等于非负下鞅 $X^+=((X_n-a)^+,\mathscr{F}_n)$ 与区间 $[0,b-a]$ 相交的次数. 因此, 如果把下鞅 X 视为非负的且 $a=0$, 则需要证明

$$\mathbf{E}\beta_n(0,b)\leqslant\frac{\mathbf{E}X_n}{b}. \tag{38}$$

设 $X_0=0, F_0=\{\varnothing,\Omega\}$, 而对于 $i=1,2,\cdots$ 和某个数 m, 记

$$\varphi_i=\begin{cases}1, & \text{若 }\tau_m<i\leqslant\tau_{m+1}\text{ 且 }m\text{ 为奇数},\\ 0, & \text{若 }\tau_m<i\leqslant\tau_{m+1}\text{ 且 }m\text{ 为偶数}.\end{cases}$$

不难看到,

$$b\beta_n(0,b)\leqslant\sum_{i=1}^{n}\varphi_i[X_i-X_{i-1}]$$

和

$$\{\varphi_i=1\}=\bigcup_{m\text{ 为奇数}}[\{\tau_m<i\}\backslash\{\tau_{m+1}<i\}]\in\mathscr{F}_{i-1}.$$

从而

$$
\begin{aligned}
b\mathbf{E}\beta_n(0,b) &\leqslant \mathbf{E}\sum_{i=1}^{n}\varphi_i[X_i - X_{i-1}] = \sum_{i=1}^{n}\int_{\{\varphi_i=1\}}(X_i - X_{i-1})\mathrm{d}\mathbf{P} \\
&= \sum_{i=1}^{n}\int_{\{\varphi_i=1\}}\mathbf{E}(X_i - X_{i-1}|\mathscr{F}_{i-1})\mathrm{d}\mathbf{P} = \sum_{i=1}^{n}\int_{\{\varphi_i=1\}}[\mathbf{E}(X_i|\mathscr{F}_{i-1}) - X_{i-1}]\mathrm{d}\mathbf{P} \\
&\leqslant \sum_{i=1}^{n}\int_{\Omega}[\mathbf{E}(X_i|\mathscr{F}_{i-1}) - X_{i-1}]\mathrm{d}\mathbf{P} = \mathbf{E}X_n,
\end{aligned}
$$

于是, 不等式 (38) 得证. □

5. 二次可积鞅大偏差概率的估计 在这一小节对于平方可积鞅, 我们将讨论大偏差概率的不等式的某些简单不等式.

设 $M = (M_n, \mathscr{F}_n)_{n\geqslant 0}$ 是平方可积鞅, 而 $\langle M\rangle = (\langle M\rangle_n, \mathscr{F}_{n-1}), M_0 = 0$ 是其二次特征. 如果将不等式 (22) 用于 $X_n = M_n^2, A_n = \langle M\rangle_n$, 则对于 $a>0, b>0$, 得

$$
\begin{aligned}
&\mathbf{P}\left\{\max_{k\leqslant n}|M_k| \geqslant an\right\} = \mathbf{P}\left\{\max_{k\leqslant n}M_k^2 \geqslant (an)^2\right\} \\
&\leqslant \frac{1}{(an)^2}\mathbf{E}[\langle M\rangle_n \wedge (bn)] + \mathbf{P}\{\langle M\rangle_n \geqslant an\}. \qquad (39)
\end{aligned}
$$

事实上, 利用第四章 §5, 在估计独立同分布随机变量之和的、大偏差概率的估计时, 所阐述的思想, 至少当 $|\Delta M_n| \leqslant C$ 时, 对于一切 n 和 $\omega\in\Omega$, 可以本质上改进该不等式.

注意, 在第四章 §5 中推导相应的不等式时, 这样的环节是利用了如下事实: 序列

$$
(\mathrm{e}^{\lambda S_n}/[\varphi(\lambda)]^n, \mathscr{F}_n)_{n\geqslant 1}, \quad \mathscr{F}_n = \sigma(\xi_1,\cdots,\xi_n), \qquad (40)
$$

形成非负鞅, 然后再对其运用本节的不等式 (8). 如果现在将 S_n 取作 M_n, 则与 (40) 式类似的将是非负鞅

$$
(\mathrm{e}^{\lambda M_n}/\mathscr{E}_n(\lambda), \mathscr{F}_n)_{n\geqslant 1},
$$

其中

$$
\mathscr{E}_n(\lambda) = \prod_{j=1}^{n}\mathbf{E}\left(\mathrm{e}^{\lambda\Delta M_j}|\mathscr{F}_{j-1}\right) \qquad (41)
$$

是所谓随机分量 (亦见第二章 §6 第 13 小节).

这一表达式相当复杂. 然而, 为使所形成的序列是鞅, 完全没有必要利用不等式 (8). 只需使它形成非负上鞅. 我们在这里也正是这样做的: 首先构造序列 $(Z_n(\lambda), \mathscr{F}_n)$ (见下面的 (43) 式), 然后运用第四章 §5 使用的方法.

引理 1 设 $M = (M_n, \mathscr{F}_n)_{n\geqslant 0}$ 是二次可积鞅, $M_0 = 0, \Delta M_0 = 0$, 而对于一切 n 和 ω, $|\Delta M_n(\omega)| \leqslant c$. 假设对于 $\lambda>0$, 有

$$
\psi_c(\lambda) = \begin{cases} \dfrac{\mathrm{e}^{\lambda c} - 1 - \lambda c}{c^2}, & c>0, \\ \dfrac{\lambda^2}{2}, & c\leqslant 0, \end{cases} \qquad (42)
$$

而

$$
Z_n(\lambda) = \mathrm{e}^{\lambda M_n - \psi_c(\lambda)\langle M\rangle_n}. \qquad (43)
$$

那么, 对于每一个 $c\geqslant 0$, 序列 $Z(\lambda) = (Z_n(\lambda), \mathscr{F}_n)_{n\geqslant 0}$ 是非负上鞅.

证明 对于 $|x|\leqslant c$,

$$
\mathrm{e}^{\lambda x} - 1 - \lambda x = (\lambda x)^2\sum_{m\geqslant 2}\frac{(\lambda x)^{m-2}}{m!} \leqslant (\lambda x)^2\sum_{m\geqslant 2}\frac{(\lambda c)^{m-2}}{m!} \leqslant x^2\psi_c(\lambda).
$$

由于该不等式以及表达式 $(Z_n = Z_n(\lambda))$

$$
\Delta Z_n = Z_{n-1}[(\mathrm{e}^{\lambda\Delta M_n} - 1)\mathrm{e}^{-\Delta\langle M\rangle_n\psi_c(\lambda)} + (\mathrm{e}^{-\Delta\langle M\rangle_n\psi_c(\lambda)} - 1)],
$$

得

$$
\begin{aligned}
&\mathbf{E}(\Delta Z_n|\mathscr{F}_{n-1}) \\
&= Z_{n-1}[\mathbf{E}(\mathrm{e}^{\lambda\Delta M_n} - 1|\mathscr{F}_{n-1})\mathrm{e}^{-\Delta\langle M\rangle_n\psi_c(\lambda)} + (\mathrm{e}^{-\Delta\langle M\rangle_n\psi_c(\lambda)} - 1)] \\
&= Z_{n-1}[\mathbf{E}(\mathrm{e}^{\lambda\Delta M_n} - 1 - \lambda\Delta M_n|\mathscr{F}_{n-1})\mathrm{e}^{-\Delta\langle M\rangle_n\psi_c(\lambda)} + (\mathrm{e}^{-\Delta\langle M\rangle_n\psi_c(\lambda)} - 1)] \\
&\leqslant Z_{n-1}[\psi_c(\lambda)\mathbf{E}((\Delta M_n)^2|\mathscr{F}_{n-1})\mathrm{e}^{-\Delta\langle M\rangle_n\psi_c(\lambda)} + (\mathrm{e}^{-\Delta\langle M\rangle_n\psi_c(\lambda)} - 1)] \\
&= Z_{n-1}[\psi_c(\lambda)\Delta\langle M\rangle_n\mathrm{e}^{-\Delta\langle M\rangle_n\psi_c(\lambda)} + (\mathrm{e}^{-\Delta\langle M\rangle_n\psi_c(\lambda)} - 1)] \leqslant 0, \qquad (44)
\end{aligned}
$$

其中亦用到如下不等式: 对于 $x\geqslant 0$, 有

$$
x\mathrm{e}^{-x} + (\mathrm{e}^{-x} - 1) \leqslant 0.
$$

由 (44) 式可见

$$
\mathbf{E}(Z_n|\mathscr{F}_{n-1}) \leqslant Z_{n-1},
$$

即 $Z(\lambda) = (Z_n(\lambda), \mathscr{F}_n)$ 是上鞅. □

假设满足引理 1 的条件. 那么, 总存在 $\lambda>0$, 使 (对于给定的 $a>0, b>0$), 有

$a\lambda - b\psi_c(\lambda) > 0$. 由此, 可见

$$\begin{aligned}\mathbf{P}\left\{\max_{k\leqslant n} M_k \geqslant an\right\} &= \mathbf{P}\left\{\max_{k\leqslant n} \mathrm{e}^{\lambda M_k} \geqslant \mathrm{e}^{\lambda an}\right\} \\ &\leqslant \mathbf{P}\left\{\max_{k\leqslant n} \mathrm{e}^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} \geqslant \mathrm{e}^{\lambda an - \psi_c(\lambda)\langle M\rangle_n}\right\} \\ &= \mathbf{P}\left\{\max_{k\leqslant n} \mathrm{e}^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} \geqslant \mathrm{e}^{\lambda an - \psi_c(\lambda)\langle M\rangle_n}, \langle M\rangle_n \leqslant bn\right\} \\ &\quad + \mathbf{P}\left\{\max_{k\leqslant n} \mathrm{e}^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} \geqslant \mathrm{e}^{\lambda an - \psi_c(\lambda)\langle M\rangle_n}, \langle M\rangle_n > nb\right\} \\ &\leqslant \mathbf{P}\left\{\max_{k\leqslant n} \mathrm{e}^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} \geqslant \mathrm{e}^{\lambda an - \psi_c(\lambda)bn}\right\} + \mathbf{P}\{\langle M\rangle_n > bn\} \\ &\leqslant \mathrm{e}^{-n[\lambda a - b\psi_c(\lambda)]} + \mathbf{P}\{\langle M\rangle_n > bn\}, \end{aligned} \tag{45}$$

其中由 (7) 式得最后一个不等式.

记 (对照第四章 §5 的函数 $H(a)$)

$$H_c(a,b) = \sup_{\lambda>0}[\lambda a - b\psi_c(\lambda)].$$

那么, 由 (45) 式可见

$$\mathbf{P}\left\{\max_{k\leqslant n} M_k \geqslant an\right\} \leqslant \mathbf{P}\{\langle M\rangle_n > bn\} + \mathrm{e}^{-nH_c(a,b)}. \tag{46}$$

将鞅 M 换成 $-M$ 可见, 不等式 (46) 的右侧也从上侧估计概率 $\mathbf{P}\left\{\min_{k\leqslant n} M_k \leqslant -an\right\}$. 因此,

$$\mathbf{P}\left\{\max_{k\leqslant n} |M_k| \geqslant an\right\} \leqslant 2\mathbf{P}\{\langle M\rangle_n > bn\} + 2\mathrm{e}^{-nH_c(a,b)}. \tag{47}$$

于是, 证明了如下定理.

定理 6 *设 $M = (M_n, \mathscr{F}_n)$ 是具有一致有界跃度的鞅, 即对于某个常数 $c > 0$ 以及一切 n 和 ω, $|\Delta M_n(\omega)| \leqslant c$. 那么, 对于任意 $a > 0, b > 0$, 不等式 (46) 和 (47) 成立.*

注 函数

$$H_c(a,b) = \frac{1}{c}\left(a + \frac{b}{c}\right)\ln\left(1 + \frac{ac}{b}\right) - \frac{a}{c}. \tag{48}$$

6. 二次可积鞅与其二次特征最大比值概率的估计 在定理 6 的条件下, 我们现在考虑形如

$$\mathbf{P}\left\{\sup_{k\geqslant n} \frac{M_k}{\langle M\rangle_k} > a\right\}$$

的概率的估计问题. 特别, 这种形式的概率, 在关于鞅的强大数定律中 (下面见 §5 的定理 4), 表征收敛的速度.

由与第四章 §5 同样的方法, 可见对于任意 $a > 0$, 存在 $\lambda > 0$, 使 $a\lambda - \psi_c(\lambda) > 0$. 那么, 对于任意 $b > 0$, 有

$$\begin{aligned}\mathbf{P}\left\{\sup_{k\geqslant n} \frac{M_k}{\langle M\rangle_k} > a\right\} &\leqslant \mathbf{P}\left\{\sup_{k\geqslant n} \mathrm{e}^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} > \mathrm{e}^{[a\lambda - \psi_c(\lambda)]\langle M\rangle_n}\right\} \\ &\leqslant \mathbf{P}\left\{\sup_{k\geqslant n} \mathrm{e}^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} > \mathrm{e}^{[a\lambda - \psi_c(\lambda)]bn}\right\} + \mathbf{P}\{\langle M\rangle_n < bn\} \\ &\leqslant \mathrm{e}^{-bn[a\lambda - \psi_c(\lambda)]} + \mathbf{P}\{\langle M\rangle_n < bn\}. \end{aligned} \tag{49}$$

由此, 得

$$\mathbf{P}\left\{\sup_{k\geqslant n} \frac{M_k}{\langle M\rangle_k} > a\right\} \leqslant \mathbf{P}\{\langle M\rangle_n < bn\} + \mathrm{e}^{-nH_c(ab,b)}, \tag{50}$$

$$\mathbf{P}\left\{\sup_{k\geqslant n} \left|\frac{M_k}{\langle M\rangle_k}\right| > a\right\} \leqslant 2\mathbf{P}\{\langle M\rangle_n < bn\} + 2\mathrm{e}^{-nH_c(ab,b)}. \tag{51}$$

于是, 证明了下面的定理.

定理 7 *假设满足定理 6 的条件, 则对于任意 $a > 0, b > 0$, 不等式 (50) 和 (51) 成立.*

注 将 (51) 式的估计, 与第四章 §5 中的 (21) 式比较, 其中对于伯努利概型: $p = 1/2, M_n = S_n - n/2, b = 1/4, c = 1/2$, 则可以看到当 $\varepsilon > 0$ 二者产生同样的结果:

$$\mathbf{P}\left\{\sup_{k\geqslant n} \left|\frac{M_k}{\langle M\rangle_k}\right| > \varepsilon\right\} = \mathbf{P}\left\{\sup_{k\geqslant n} \left|\frac{S_k - k/2}{k}\right| > \frac{\varepsilon}{4}\right\} \leqslant 2\mathrm{e}^{-4\varepsilon^2 n}.$$

7. 练习题

1. 设 $X = (X_n, \mathscr{F}_n)$ 是非负下鞅, $V = (V_n, \mathscr{F}_{n-1})$ 是可预测序列, 其中以概率 1, 有 $0 \leqslant V_{n+1} \leqslant V_n \leqslant C$, 而 C 是常数, 证明不等式 (1) 有如下推广:

$$\varepsilon\mathbf{P}\left\{\max_{1\leqslant j\leqslant n} V_j X_j \geqslant \varepsilon\right\} + \int_{\left\{\max\limits_{1\leqslant j\leqslant n} V_j X_j < \varepsilon\right\}} V_n X_n \mathrm{d}\mathbf{P} \leqslant \sum_{j=1}^{n} \mathbf{E}V_j \Delta X_j. \tag{52}$$

2. 证明克里克伯格 (Krickbeg) 分解: 任何鞅 $X = (X_n, \mathscr{F}_n)$, 只要 $\sup \mathbf{E}|X_n| < \infty$, 都可以表示为两个非负鞅之差.

3. 设 $\xi_1, \xi_2, \cdots$ 是独立随机变量序列, $S_n = \xi_1 + \cdots + \xi_n$, 而 $S_{m,n} = \sum_{j=m+1}^{n} \xi_j$. 证明如下奥塔维安尼 (Ottawiani) 不等式:

$$\mathbf{P}\left\{\max_{1\leqslant j\leqslant n} |S_j| > 2\varepsilon\right\} \leqslant \frac{\mathbf{P}\{|S_n| > \varepsilon\}}{\max\limits_{1\leqslant j\leqslant n} \mathbf{P}\{|S_{j,n}| \leqslant \varepsilon\}},$$

并 (在 $\mathbf{E}\xi_i = 0, i \geqslant 1$, 的条件下) 导出不等式:

$$\int_0^\infty \mathbf{P}\left\{\max_{1\leqslant j\leqslant n} |S_j| > 2t\right\} \mathrm{d}t \leqslant 2\mathbf{E}|S_n| + 2\int_{2\mathbf{E}|S_n|}^\infty \mathbf{P}\{|S_n| > t\} \mathrm{d}t. \tag{53}$$

4. 设 $\xi_1, \xi_2, \cdots$ 是独立随机变量序列, 且 $\mathbf{E}\xi_i = 0$. 利用 (53) 式证明, 在这种情形下有不等式 (10) 的加强:

$$\mathbf{E}S_n^* \leqslant 8\mathbf{E}|S_n|.$$

5. 证明 (16) 式.

6. 证明 (19) 式.

7. 设 $\sigma-$ 代数 $\mathscr{F}_0, \cdots, \mathscr{F}_n$ 满足 $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \cdots \subseteq \mathscr{F}_n$, 而事件 $A_k \in \mathscr{F}_k (k = 1, \cdots, n)$. 利用 (22) 式证明如下德沃列茨基 (Дворецкий) 不等式: 对于任意 $\varepsilon > 0$,

$$\mathbf{P}\left(\bigcup_{k=1}^{n} A_k\right) \leqslant \varepsilon + \mathbf{P}\left\{\sum_{k=1}^{n} \mathbf{P}(A_k | \mathscr{F}_{k-1}) > \varepsilon\right\}.$$

8. 设 $X = (X_n)_{n \geqslant 1}$ 是二次可积鞅, 而 $(b_n)_{n \geqslant 1}$ 是正实数不减序列. 证明如下哈伊克 – 雷内伊 (J. Hajek– A. Rènyi) 不等式: 对于任意 $\lambda > 0$,

$$\mathbf{P}\left\{\max_{1 \leqslant j \leqslant n} \left|\frac{X_k}{b_k}\right| \geqslant \lambda\right\} \leqslant \frac{1}{\lambda^2} \sum_{k=1}^{n} \frac{\mathbf{E}(\Delta X_k)^2}{b_k^2}, \quad \Delta X_k = X_k - X_{k-1}, \quad X_0 = 0.$$

9. 设 $X = (X_n)_{n \geqslant 1}$ 是下鞅, 而 $g(x)$ 是非负递增凹 (下凸) 函数. 那么, 对于任意正 t, 和实数 x, 有

$$\mathbf{P}\left\{\max_{1 \leqslant j \leqslant n} X_k \geqslant x\right\} \leqslant \frac{\mathbf{E}g(tX_n)}{g(tx)}.$$

特别,

$$\mathbf{P}\left\{\max_{1 \leqslant j \leqslant n} X_k \geqslant x\right\} \leqslant \mathrm{e}^{-tx} \mathbf{E}\mathrm{e}^{tX_n}.$$

10. 设 $\xi_1, \xi_2, \cdots$ 是独立随机变量序列, 且 $\mathbf{E}\xi_m = 0, \mathbf{E}\xi_n^2 = 1, n \geqslant 1$. 记

$$\tau = \inf\left\{n \geqslant 1 : \sum_{i=1}^{n} \xi_i > 0\right\}.$$

证明 $\mathbf{E}\tau^{1/2} < \infty$.

11. 设 $\xi = (\xi_n)_{n \geqslant 1}$ 是鞅 – 差, 而 $1 < p \leqslant 2$. 证明

$$\mathbf{E}\sup_{n \geqslant 1}\left|\sum_{j=1}^{n} \xi_j\right|^p \leqslant C_p \sum_{j=1}^{n} \mathbf{E}|\xi_j|^p,$$

其中 C_p 是常数.

12. 设 $X = (X_n)_{n \geqslant 1}$ 是鞅, $\mathbf{E}\xi_m = 0, \mathbf{E}\xi_n^2 < \infty, n \geqslant 1$. 证明 (第四章 §2 练习题 5), 对于任意 $n \geqslant 1$ 和 $\varepsilon > 0$, 有

$$\mathbf{P}\left\{\max_{1 \leqslant k \leqslant n} X_k \geqslant \varepsilon\right\} \leqslant \frac{\mathbf{E}X_n^2}{\varepsilon^2 + \mathbf{E}X_n^2}.$$

§4. 下鞅和鞅收敛的基本定理

1. 有界单调鞅序列极限的存在性 下面的结果在整个下鞅收敛性问题中, 可以视为数学分析中著名事实 "有界单调数列有 (有限) 极限" 的概率类似,

定理 1 (杜布) *设 $X = (X_n, \mathscr{F}_n)_{n \geqslant 1}$ 是下鞅, 且*

$$\sup_n \mathbf{E}|X_n| < \infty, \tag{1}$$

则以概率 1 存在极限 $\lim X_n = X_\infty$, 并且 $\mathbf{E}|X_\infty| < \infty$.

证明 假设

$$\mathbf{P}\{\overline{\lim} X_n > \underline{\lim} X_n\} > 0. \tag{2}$$

那么, 由于

$$\{\overline{\lim} X_n > \underline{\lim} X_n\} = \bigcup_{a<b} \{\overline{\lim} X_n > b > a > \underline{\lim} X_n\}$$

(a, b 是有理数), 故存在 a 和 b, 使

$$\mathbf{P}\{\overline{\lim} X_n > b > a > \underline{\lim} X_n\} > 0. \tag{3}$$

设 $\beta_n(a,b)$ 是序列 $X_1, \cdots, X_n$ 自下而上与区间 (a,b) 相交的次数, 而 $\beta_\infty(a,b) = \lim_n \beta_n(a,b)$. 根据 §3 的 (37) 式, 有

$$\mathbf{E}\beta_n(a,b) = \frac{\mathbf{E}|X_n - a|^+}{b-a} \leqslant \frac{\mathbf{E}X_n^+ + |a|}{b-a},$$

因此,

$$\mathbf{E}\beta_\infty(a,b) = \lim_n \mathbf{E}\beta_n(a,b) \leqslant \frac{\sup_n \mathbf{E}X_n^+ + |a|}{b-a} < \infty.$$

由 (1) 式以及 (由于 $\mathbf{E}X_n^+ \leqslant \mathbf{E}|X_n| = 2\mathbf{E}X_n^+ - \mathbf{E}X_n \leqslant 2\mathbf{E}X_n^+ - \mathbf{E}X_1$) 对于下鞅, 有

$$\sup_n \mathbf{E}|X_n| < \infty \Leftrightarrow \sup_n \mathbf{E}X_n^+ < \infty.$$

然而, 条件 $\mathbf{E}\beta_\infty(a,b) < \infty$ 与假设 (3) 矛盾. 于是, 以概率 1 存在极限 $\lim X_n = X_\infty$, 而由法图引理可见

$$\mathbf{E}|X_\infty| \leqslant \sup_n \mathbf{E}|X_n| < \infty. \qquad \square$$

系 1 *如果 X 是非负下鞅, 则以概率 1 存在有限极限 $\lim X_n = X_\infty$.*

系 2 *如果 $X = (X_n, \mathscr{F}_n)_{n \geqslant 1}$ 是非正下鞅, 则序列*

$$\overline{X} = (X_n, \mathscr{F}_n), \quad 1 \leqslant n \leqslant \infty, \quad X_\infty = \lim X_n \quad 和 \quad \mathscr{F}_\infty = \sigma\left(\bigcup_n \mathscr{F}_n\right)$$

构成 (非正) 下鞅.

事实上, 根据法图引理,

$$\mathbf{E}X_\infty = \mathbf{E}\lim X_n \geqslant \overline{\lim}\mathbf{E}X_n > \mathbf{E}X_1 > -\infty,$$

且以概率 1, 有

$$\mathbf{E}(X_\infty|\mathscr{F}_m) = \mathbf{E}(\lim X_n|\mathscr{F}_m) \geqslant \overline{\lim}\mathbf{E}(X_n|\mathscr{F}_m) \geqslant X_m.$$

系 3 如果 $X=(X_n,\mathscr{F}_n)$ 是非负上鞅, 则以概率 1 存在有限极限 $\lim X_n$.

事实上, 这时有

$$\sup_n \mathbf{E}|X_n| = \sup_n \mathbf{E}X_n = \mathbf{E}X_1 < \infty,$$

从而, 可以运用定理 1.

2. 鞅几乎必然收敛也是在 L^1 上平均收敛的条件 设 $\xi_1,\xi_2,\cdots$ 是独立随机变量序列, 且 $\mathbf{P}\{\xi_i=0\}=\mathbf{P}\{\xi_i=2\}=1/2$. 那么, $X=(X_n,\mathscr{F}_n^\xi)$ 是鞅, 其中

$$X_n = \prod_{i=1}^n \xi_i, \quad \mathscr{F}_n^\xi = \sigma(\xi_1,\cdots,\xi_n),$$

且 $\mathbf{E}X_n=1$ 和 $X_n\to X_\infty\equiv 0$ $(\mathbf{P}-\text{a.c.})$. 同样, 显然 $\mathbf{E}|X_n-\mathbf{E}X_\infty|=1$, 从而 $X_n \nrightarrow X_\infty$. 这样, 条件 (1) 一般不能保障在 L^1 的意义上 X_n 收敛于 X_∞.

下面将要介绍的定理 2 表明, 如果将条件 (1) 加强为 "随机变量族 $\{X_n\}$ 具有一致可积性" (那么, 根据第二章 §6 的第 5 小节知, 性质 (16) 成立), 则 $\{X_n\}$ 在 L_1 意义上几乎处处收敛, 同时也在 L^1 的意义上平均收敛.

定理 2 设 $X=(X_n,\mathscr{F}_n)$ 是下鞅, 且随机变量族 $\{X_n\}$ 一致可积. 那么, 存在随机变量 X_∞, 且 $\mathbf{E}|X_\infty|<\infty$, 使当 $n\to\infty$ 时, 有

$$X_n \to X_\infty \ (\mathbf{P}-\text{a.c.}), \tag{4}$$

$$X_n \xrightarrow{L^1} X_\infty. \tag{5}$$

这时, 序列 $\overline{X}=(X_n,\mathscr{F}_n), 1\leqslant n\leqslant\infty, \mathscr{F}_\infty=\sigma\left(\bigcup_n \mathscr{F}_n\right)$ 也构成下鞅.

证明 由定理 1 得 (4) 式, 由 (4) 式和第二章 §6 的定理 4 的 (5) 式.

其次, 如果 $A\in\mathscr{F}_n$ 和 $m\geqslant n$, 则

$$\mathbf{E}I_A|X_m-X_\infty|\to 0, \quad m\to\infty,$$

从而,

$$\lim_{n\to\infty}\int_A X_m \mathrm{d}\mathbf{P} = \int_A X_\infty \mathrm{d}\mathbf{P}.$$

由于序列

$$\left(\int_A X_m \mathrm{d}\mathbf{P}\right)_{m\geqslant n}$$

是非降的, 可见

$$\int_A X_n \mathrm{d}\mathbf{P} \leqslant \int_A X_m \mathrm{d}\mathbf{P} \leqslant \int_A X_\infty \mathrm{d}\mathbf{P},$$

于是, 对于一切 $n\geqslant 1, X_n \leqslant (X_\infty|\mathscr{F}_n)(\mathbf{P}-\text{a.c.})$. □

系 设 $X=(X_n,\mathscr{F}_n)$ 是下鞅, 且对于某个 $p>1$,

$$\sup_n \mathbf{E}|X_n| < \infty, \tag{6}$$

则存在可积随机变量 X_∞, 使 (4) 式和 (5) 式成立.

对于证明只需注意到, 根据第二章 §6 的引理 3, 条件 (6) 可以保障随机变量族 $\{X_n\}$ 的一致可积性.

3. 鞅一致可积的充分和必要条件 现在引进关于条件数学期望连续性的定理, 是关于鞅的收敛性的最早的成果之一.

定理 3 (P. 列维 (P. P. Lévy)) 设 $(\Omega,\mathscr{F},\mathbf{P})$ 是概率空间, $(\mathscr{F}_n)_{n\geqslant 1}$ 是非降 $\sigma-$ 代数族: $\mathscr{F}_1\subseteq\mathscr{F}_2\subseteq\cdots\subseteq\mathscr{F}$. 假设 ξ 是随机变量, 且 $\mathbf{E}|\xi|<\infty$, 而 $\mathscr{F}_\infty=\sigma\left(\bigcup_n \mathscr{F}_n\right)$. 那么, 以概率 1 并且在 L^1 收敛意义上, 有

$$\mathbf{E}(\xi|\mathscr{F}_n)\to\mathbf{E}(\xi|\mathscr{F}_\infty), \quad n\to\infty. \tag{7}$$

证明 设 $X_n=(\xi,\mathscr{F}_n), n\geqslant 1$. 那么, 对于 $a>0, b>0$, 有

$$\begin{aligned}\int_{\{|X_n|\geqslant a\}}|X_n|\mathrm{d}\mathbf{P} &\leqslant \int_{\{|X_n|\geqslant a\}}\mathbf{E}(|\xi|\big|\mathscr{F}_n)\mathrm{d}\mathbf{P} = \int_{\{|X_n|\geqslant a\}}|\xi|\mathrm{d}\mathbf{P}\\ &= \int_{\{|X_n|\geqslant a, |\xi|\leqslant b\}}|\xi|\mathrm{d}\mathbf{P} + \int_{\{|X_n|\geqslant a, |\xi|>b\}}|\xi|\mathrm{d}\mathbf{P}\\ &\leqslant b\mathbf{P}\{|X_n|\geqslant a\} + \int_{\{|\xi|>b\}}|\xi|\mathrm{d}\mathbf{P}\\ &\leqslant \frac{b}{a}\mathbf{E}|\xi| + \int_{\{|\xi|>b\}}|\xi|\mathrm{d}\mathbf{P}.\end{aligned}$$

先令 $a\to\infty$, 再令 $b\to\infty$, 得

$$\lim_{a\to\infty}\sup_n \int_{\{|X_n|\geqslant a\}}|X_n|\mathrm{d}\mathbf{P} = 0.$$

这说明随机变量族 $\{X_n\}$ 一致可积. 那么, 根据定理 2, 存在随机变量 X_∞, 使以概率 1 收敛并且在 L^1 中的收敛, 有 $X_n=\mathbf{E}(\xi|\mathscr{F}_n)\to X_\infty$. 因此, 只需设

$$X_\infty = \mathbf{E}(\xi|\mathscr{F}_\infty) \ (\mathbf{P}-\text{a.c.}).$$

设 $m \geqslant n, A \in \mathscr{F}_n$, 则

$$\int_A X_m \mathrm{d}\mathbf{P} = \int_A X_n \mathrm{d}\mathbf{P} = \int_A \mathbf{E}(\xi|\mathscr{F}_n)\mathrm{d}\mathbf{P} = \int_A \xi \mathrm{d}\mathbf{P}.$$

由于随机变量族 $\{X_n\}$ 的一致可积, 且由第二章 §6 定理 5 知: 当 $m \to \infty$ 时, 有 $\mathbf{E}I_{\mathbf{A}}|X_m - X_\infty| \to 0$, 从而

$$\int_A X_\infty \mathrm{d}\mathbf{P} = \int_A \xi \mathrm{d}\mathbf{P}. \tag{8}$$

该等式对于任意 $A \in \mathscr{F}_n$ 成立, 因而对于任意 $A \in \bigcup_{n=1}^{\infty} \mathscr{F}_n$ 成立. 由于 $\mathbf{E}|X_\infty| < \infty, \mathbf{E}|\xi| < \infty$, 则 (8) 式的左右两侧都是亦可能取负值的 $\sigma-$ 可加测度, 不过这样的测度是有限的, 并且在代数 $\bigcup_{n=1}^{\infty} \mathscr{F}$ 上重合. 根据卡拉泰奥多里 (C. Carathéodory) 定理 (第二章 §3), 由于测度自代数到 $\sigma-$ 代数上的延拓的唯一性, 可见对于集合 $A \in F_\infty = \sigma(\bigcup F_n)$ 下面的等式 (9) 成立:

$$\int_A X_\infty \mathrm{d}\mathbf{P} = \int_A \xi \mathrm{d}\mathbf{P} = \int_A \mathbf{E}(\xi|\mathscr{F}_\infty)\mathrm{d}\mathbf{P}, \quad A \in \mathscr{F}_\infty. \tag{9}$$

随机变量 X_∞ 和 $\mathbf{E}(\xi|\mathscr{F}_\infty)$ 是 $\mathscr{F}_\infty-$ 可测的, 因此由 (9) 式根据第二章 §6 第 3 小节的性质 I, 可见 $X_\infty = \mathbf{E}(\xi|\mathscr{F}_\infty)(\mathbf{P}-\text{a.c.})$. □

系 *随机序列 $X = (X_n, \mathscr{F}_n)$ 是一致可积鞅, 当且仅当存在随机变量 $\xi, \mathbf{E}|\xi| < \infty$, 使对于一切 $n \geqslant 1, X_n = (\xi, \mathscr{F}_n)$, (即 $X = (X_n, \mathscr{F}_n)$ 是列维鞅). 并且当 $n \to \infty$ 时 $X_n \to (\xi, \mathscr{F}_n)$ (以概率 1 收敛且在 L^1 中的收敛).*

事实上, 如果 $X = (X_n, \mathscr{F}_n)$ 是一致可积鞅, 则根据定理 2 存在可积随机变量 X_∞, 使 $X_n \to X_\infty$ (以概率 1 收敛且在 L^1 中的收敛), 并且 $X_n = (X_\infty|\mathscr{F}_n)$. 这样, 可以将 ($\mathscr{F}_\infty-$ 可测) 随机变量 ξ 取为 X_∞.

由定理 3 可得逆命题.

4. 应用列维定理的例 下面举几个列维定理应用的例.

例 1 "0–1" 律 设 $\xi_1, \xi_2, \cdots$ 是独立随机变量序列, 而 $\mathscr{F}_n^\xi = \sigma(\xi_1, \cdots, \xi_n), \mathscr{X}$ 是 "尾部" 事件的 $\sigma-$ 代数, 而 $A \in \mathscr{X}$. 由定理 3 可见

$$\mathbf{E}(I_A|\mathscr{F}_n^\xi) \to \mathbf{E}(I_A|\mathscr{F}_\infty^\xi) = I_A \quad (\mathbf{P}-\text{a.c.}).$$

由于 I_A 与 $(\xi_1, \cdots, \xi_n)$ 独立, 可见 $\mathbf{E}(I_A|\mathscr{F}_n^\xi) = \mathbf{E}I_A$, 因而 $I_A = \mathbf{E}I_A(\mathbf{P}-\text{a.c.})$, 故 $\mathbf{P}(A) = 0$ 或 $\mathbf{P}(A) = 1$.

下面的两个例子, 说明在数学分析中上面引进的关于收敛性的定理.

例 2 如果 $f = f(x)$ 是区间 $[0,1)$ 上的函数, 并且满足利普希茨 (R. O. S. Lipschitz) 条件, 则 $f = f(x)$ 绝对连续, 并且由数学分析熟知, 存在一勒贝格可积函数 $g = g(x)$, 使

$$f(x) - f(0) = \int_0^x g(y)\mathrm{d}y. \tag{10}$$

(在这种意义上 $g(x)$ 是 $f(x)$ 的 "导数").

现在说明由定理 1 可能得出这一结果. 设 $\Omega = [0,1), \mathscr{F} = \mathscr{B}([0,1))$, 而 $\mathbf{P}$ 是勒贝格测度. 记

$$\xi_n(x) = \sum_{k=1}^{2^n} \frac{k-1}{2^n} I\left\{\frac{k-1}{2^n} \leqslant x < \frac{k}{2^n}\right\},$$

$\mathscr{F}_n = \sigma(\xi_1, \cdots, \xi_n) = \sigma\{\xi_n\}$ 而设

$$X_n = \frac{f(\xi_n + 2^{-n}) - f(\xi_n)}{2^{-n}}.$$

因为, 对于给定的 ξ_n 的值, 随机变量 ξ_{n+1} 只有 ξ_n 和 $\xi_n + 2^{-(n+1)}$ 两个可能值, 且相应的条件概率都是 1/2, 所以

$$\begin{aligned}\mathbf{E}[X_{n+1}|\mathscr{F}_n] &= \mathbf{E}(X_{n+1}|\xi_n) = 2^{n+1}\mathbf{E}[f(\xi_{n+1} + 2^{-(n+1)}) - f(\xi_{n+1})|\xi_n] \\ &= 2^{n+1}\left\{\frac{1}{2}[f(\xi_n + 2^{-(n+1)}) - f(\xi_n)] + \frac{1}{2}[f(\xi_n + 2^{-n}) - f(\xi_n + 2^{-(n+1)})]\right\} \\ &= 2^n[f(\xi_n + 2^{-n}) - f(\xi_n)] = X_n.\end{aligned}$$

由此可见, $X = (X_n, \mathscr{F}_n)$ 是鞅, 并且由于 $|X_n| \leqslant L$, 其中 L 是利普希茨条件的常数: $|f(x) - f(y)| \leqslant L|x - y|$. 注意, $\mathscr{F} = \mathscr{B}([0,1)) = \sigma(\bigcup \mathscr{F}_n)$. 因此, 根据定理 3 的系, 存在 $\mathscr{F}-$ 可测函数 $g = g(x)$, 使 $X_n \to g$ $(\mathbf{P}-\text{a.c.})$, 并且

$$X_n = \mathbf{E}(g|\mathscr{F}_n). \tag{11}$$

取集合 $B = [0, k/2^n]$, 则由 (11) 式, 有

$$f\left(\frac{k}{2^n}\right) - f(0) = \int_0^{k/2^n} x_n \mathrm{d}x = \int_0^{k/2^n} g(x)\mathrm{d}x.$$

从而, 因为 n 和 k 的任意性, 由此得所要求证明的等式 (10).

例 3 设 $\Omega = [0,1), \mathscr{F} = \mathscr{B}([0,1))$, 而 $\mathbf{P}$ 是勒贝格测度. 考虑哈尔 (A. Haar) 函数系 (由第二章 §11 例 3 的定义). 设 $\mathscr{F}_n = \sigma(H_1, \cdots, H_n)$, 并注意到 $\mathscr{F} = \mathscr{B}([0,1)) = \sigma(\bigcup \mathscr{F}_n)$. 由条件数学期望的性质以及哈尔函数的构造, 不难导出, 对于博雷尔函数 $f \in L$, 有

$$\mathbf{E}[f(x)|\mathscr{F}_n] = \sum_{k=1}^n \alpha_k H_k(x) \quad (\mathbf{P}-\text{a.c.}), \tag{12}$$

其中

$$\alpha_k = (f, H_k) = \int_0^1 f(x)H_k(x)\mathrm{d}x.$$

换句话说, 在函数 $f(x)$ 按哈尔函数系的展开时, 条件数学期望 $\mathbf{E}[f(x)|\mathscr{F}_n]$ 是傅里叶级数的部分和. 那么, 将定理 3 用于鞅 $(\mathbf{E}(f|\mathscr{F}_n),\mathscr{F}_n)$ 可得, 当 $n\to\infty$ 时

$$\sum_{k=1}^{n}(f,H_k)H_k(x)\to f(x)\ (\mathbf{P}-\text{a.c.}),$$

和

$$\int_0^1\left|\sum_{k=1}^{n}(f,H_k)H_k(x)-f(x)\right|\mathrm{d}x\to 0.$$

例 4 设 $(\xi_n)_{n\geqslant 1}$ 是随机变量序列. 根据第二章 §10 的定理 2, 如果级数 $\sum\xi_n$ 以概率 1 收敛, 则它也依概率收敛和按分布收敛. 结果表明, 假如随机变量 $\xi_1,\xi_2,\cdots$ 独立, 则逆命题也成立: 若由独立随机变量 $\xi_1,\xi_2,\cdots$ 构成的级数 $\sum\xi_n$ 按分布收敛, 则它也依概率收敛和以概率 1 收敛.

这一性质可以用如下方式证明. 设 $S_n=\xi_1+\cdots+\xi_n, n\geqslant 1$, 且 $S_n\xrightarrow{d}S$, 则对于每一个实数 $t, \mathbf{E}\mathrm{e}^{itS_n}\to\mathbf{E}\mathrm{e}^{itS}$, 显然, 存在 $\delta>0$, 使对于所有 $|t|<\delta$, 有 $|\mathbf{E}\mathrm{e}^{itS}|>0$. 取某一 t_0, 使之满足 $|t_0|<\delta$, 则存在 $n_0=n_0(t_0)$, 使得对于一切 $n\geqslant n_0$, 有 $|\mathbf{E}\mathrm{e}^{it_0S_n}|\geqslant c>0$, 其中 c 是某一常数.

对于 $n\geqslant n_0$, 建立一序列 $X=(X_n,\mathscr{F}_n)$, 其中

$$X_n=\frac{\mathrm{e}^{it_0S_n}}{\mathbf{E}\mathrm{e}^{it_0S_n}},\quad \mathscr{F}_n=\sigma(\xi_1,\cdots,\xi_n).$$

由于假设随机变量 $\xi_1,\xi_2,\cdots$ 独立, 则序列 $X=(X_n,\mathscr{F}_n)$ 是鞅, 且

$$\sup_{n\geqslant n_0}\mathbf{E}|X_n|\leqslant c^{-1}<\infty.$$

那么, 由定理 1 可见, 以概率 1 极限 $\lim\limits_n X_n$ 存在并且有限. 因此极限 $\lim\limits_n \mathrm{e}^{it_0S_n}$ 也以概率 1 存在. 因而可以断定: 存在 $\delta>0$, 使得对于集合 $T=\{t:|t|<\delta\}$ 中每一个 t, 极限 $\lim\limits_n \mathrm{e}^{itS_n}$ 以概率 1 存在.

设 $T\times\Omega=\{(t,\omega):t\in T,\omega\in\Omega\}, \overline{\mathscr{B}}(T)$, 是 T 上勒贝格集合的 $\sigma-$ 代数, 而 λ 是在 $(T,\overline{\mathscr{B}}(T))$ 上的勒贝格测度. 其次, 记

$$C=\{(t,\omega)\in T\times\Omega:\lim_n \mathrm{e}^{itS_n(\omega)}\ \text{存在}\}.$$

显然, $C\in\overline{\mathscr{B}}(T)\otimes\mathscr{F}$.

上面曾经证明, 对于每一个 $t\in T, \mathbf{P}(C_t)=1$, 其中 $C_t=\{\omega\in\Omega:(t,\omega)\in C\}$ 是在点 t 处集合 C 的截线. 根据傅比尼定理 (第二章 §6 的定理 8):

$$\int_{T\times\Omega}I_C(t,\omega)\mathrm{d}(\lambda\times\mathbf{P})=\int_T\left(\int_\Omega I_C(t,\omega)\mathrm{d}\mathbf{P}\right)\mathrm{d}\lambda=\int_T\mathbf{P}(C_t)\mathrm{d}\lambda=\lambda(T)=2\delta>0.$$

另一方面, 仍然根据傅比尼定理:

$$\lambda(T)=\int_{T\times\Omega}I_C(t,\omega)\mathrm{d}(\lambda\times\mathbf{P})=\int_\Omega\left(\int_T I_C(t,\omega)\mathrm{d}\lambda\right)\mathrm{d}\mathbf{P}=\int_\Omega\lambda(C_\omega)\mathrm{d}\mathbf{P},$$

其中 $C_\omega=\{t:(t,\omega)\in C\}$.

由此可见, 存在集合 $\widetilde{\Omega}$, 且 $\mathbf{P}(\widetilde{\Omega})=1$, 使得对于所有 $\omega\in\widetilde{\Omega}$, 有 $\lambda(C_\omega)=\lambda(T)=2\delta>0$.

从而, 可以断定: 对于每一个 $\omega\in\widetilde{\Omega}$ 和对于所有 $t\in C_\omega$, 极限 $\lim\limits_n \mathrm{e}^{itS_n}$ 存在; 并且集合 C_ω 的勒贝格测度大于 0. 由此以及练习题 8, 对于每一个 $\omega\in\widetilde{\Omega}$, 极限 $\lim\limits_n S_n$ 存在且有限. 由于 $\mathbf{P}(\widetilde{\Omega})=1$, 可见极限 $\lim\limits_n S_n$ 以概率 1 存在并且有限.

5. 练习题

1. 设 $\{\mathscr{G}_n\}$ 是不增 $\sigma-$ 代数族, $\mathscr{G}_1\supseteq\mathscr{G}_2\supseteq\cdots,\mathscr{G}_\infty=\bigcap\mathscr{G}_n$, 而 η 是某一可积随机变量. 证明定理 3 之如下的类似成立: 当 $n\to\infty$ 时, $\mathbf{P}-$ 几乎处处, 以及在 L^1 收敛的意义上, 有

$$\mathbf{E}(\eta|\mathscr{G}_n)\to\mathbf{E}(\eta|\mathscr{G}_\infty).$$

2. 设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量序列, $\mathbf{E}|\xi_1|<\infty, \mathbf{E}\xi_1=m, S_n=\xi_1+\cdots+\xi_n$. 首先证明 (见第二章 §7 练习题 2)

$$\mathbf{E}(\xi_1|S_n,S_{n+1},\cdots)=\mathbf{E}(\xi_1|S_n)=\frac{S_n}{n}\quad(\mathbf{P}-\text{a.c.}),$$

然后利用练习题 1 的结果证明强大数定律: 当 $n\to\infty$ 时, $\mathbf{P}-$ 几乎处处以及在 L^1 的意义上, 有

$$\frac{S_n}{n}\to m.$$

3. 证明如下反映 "勒贝格控制收敛定理与 P. 列维定理" 联系的结果. 假设 $(\xi_n)_{n\geqslant 1}$ 是随机变量序列, 满足: $\xi_n\to\xi(\mathbf{P}-\text{a.c.}), |\xi_n|\leqslant\eta, \mathbf{E}\eta<\infty$, 而 $(\mathscr{F}_m)_{m\geqslant 1}$ 是非降 $\sigma-$ 代数族, $\mathscr{F}_\infty=\sigma(\bigcup\mathscr{F}_m)$. 那么, $(\mathbf{P}-\text{a.c.})$

$$\lim_{\substack{m\to\infty\\ n\to\infty}}\mathbf{E}(\xi_n|\mathscr{F}_m)=\mathbf{E}(\xi|\mathscr{F}_\infty).$$

4. 证明 (12) 式.

5. 设 $\Omega=[0,1),\mathscr{F}=\mathscr{B}([0,1))$, $\mathbf{P}$ 是勒贝格测度, 而 $f=f(x)\in L^1$. 记

$$f_n(x)=2^n\int_{k2^{-n}}^{(k+1)2^{-n}}f(y)\mathrm{d}y,\quad k2^{-n}\leqslant x<(k+1)2^{-n}.$$

证明 $f_n(x)\to f(x)(\mathbf{P}-\text{a.c.})$.

6. 设 $\Omega=[0,1),\mathscr{F}=\mathscr{B}([0,1))$,$\mathbf{P}$ 是勒贝格测度, 而 $f=f(x)\in L^1$. 假设该函数在 $[0,2)$ 上是周期的, 并设

$$f_n(x)=\sum_{i=1}^{2^n}2^{-n}f(x+i2^{-n}).$$

证明 $f_n(x)\to f(x)$ ($\mathbf{P}$ – a.c.).

7. 证明, 如果 $X=(X_n,\mathscr{F}_n)$ 是广义下鞅 (§1 定义 2), 满足

$$\inf_n\sup_{n\geqslant m}\mathbf{E}(X_n^+|\mathscr{F}_m)<\infty\ (\mathbf{P}-\text{a.c.}),$$

则对于 $X=(X_n,\mathscr{F}_n)$ 定理 1 仍然成立,

8. 设 $(a_n)_{n\geqslant1}$ 是一数列, 对于一切实数 $t:|t|<\delta(\delta>0)$, 存在极限 $\lim\limits_n e^{ita_n}$. 证明极限 $\lim\limits_n a_n$ 存在并且有限.

9. 设 $F=F(x),x\in\mathbb{R}$, 是分布函数, 而 $\alpha\in(0,1)$; 假设存在 $\theta\in\mathbb{R}$, 使 $F(\theta)=\alpha$. 构造序列 $X_1,X_2,\cdots$ (罗宾斯 – 门罗 [H. Robbins-Monroe] 方法), 使

$$X_{n+1}=X_n-n^{-1}(Y_n-\alpha),$$

其中 $Y_1,Y_2,\cdots$ 是满足

$$\mathbf{P}(Y_n=y|X_1,\cdots,X_n;Y_1,\cdots,Y_{n-1})=\begin{cases}F(X_n), & 若\ y=1,\\ 1-F(X_n), & 若\ y=0,\end{cases}$$

的随机变量.

证明 "随机逼近" 理论的如下结果: $\mathbf{E}|X_n-\theta|^2\to0,n\to\infty$.

10. 设 $X=(X_n,\mathscr{F}_n)_{n\geqslant1}$ 是下鞅, 且对于每一个停时 τ, 有 $\mathbf{E}(X_\tau I(\tau<\infty))\neq\infty$. 证明以概率 1 极限 $\lim\limits_n X_n$ 存在.

11. 设 $X=(X_n,\mathscr{F}_n)_{n\geqslant1}$ 是鞅, 而 $\mathscr{F}_\infty=\sigma\left(\bigcup\limits_{n=1}^{\infty}\mathscr{F}_n\right)$. 证明, 如果序列 $(X_n)_{n\geqslant1}$ 一致可积, 则极限 $X_\infty=\lim\limits_n X_n(\mathbf{P}-\text{a.c.})$ 存在, 且 "封闭" 序列 $\overline{X}=(X_n,\mathscr{F}_n)_{1\leqslant n\leqslant\infty}$ 是鞅.

12. 设 $X=(X_n,\mathscr{F}_n)_{n\geqslant1}$ 是下鞅, 而 $\mathscr{F}_\infty=\sigma(\bigcup\limits_{n=1}^{\infty}\mathscr{F}_n)$. 证明, 如果序列 $(X_n^+)_{n\geqslant1}$ 一致可积, 则极限 $X_\infty=\lim\limits_n X_n(\mathbf{P}-\text{a.c.})$ 存在, 且 "封闭" 序列 $\overline{X}=(X_n,\mathscr{F}_n)_{1\leqslant n<\infty}$ 是下鞅.

§5. 下鞅和鞅的收敛集

1. 随机序列类 $\mathbf{C}^+$ 设 $X=(X_n,\mathscr{F}_n)$ 是随机序列以 $\{X_n\to\}$ 或 $\{-\infty<\lim X_n<\infty\}$ 表示 "使 $\lim X_n$ *存在和有限*的基本结局的集合". 对于任意二事件 A 和 B, 如果 $\mathbf{P}\{I_A\leqslant I_B\}=1$, 则亦称 $A\subseteq B(\mathbf{P}-\text{a.c.})$.

如果 X 是下鞅且 $\sup\mathbf{E}|X_n|<\infty$ (或者等价地 $\sup\mathbf{E}X_n^+<\infty$), 则根据 §4 定理 1, 有

$$\{X_n\to\}=\Omega\quad(\mathbf{P}-\text{a.c.}),\quad 即\ \{X_n\nrightarrow\}=0.$$

现在讨论, 在条件 $\sup\mathbf{E}|X_n|<\infty$ 不成立的情况下, 关于下鞅的收敛集合 $\{X_n\to\}$ 的结构问题.

设 $a>0$,

$$\tau_a=\begin{cases}\inf\{n\geqslant1:X_n>a\}, & 若\ \{\cdot\}\neq\varnothing,\\ \infty, & 若\ \{\cdot\}=\varnothing.\end{cases}$$

定义 如果对于任意 $a>0$, 有

$$\mathbf{E}(\Delta X_{\tau_a})^+I(\tau_a<\infty)<\infty,\tag{1}$$

则称随机序列 $X=(X_n,\mathscr{F}_n)$ 属于类 $\mathbf{C}^+(X\in\mathbf{C}^+)$, 其中 $\Delta X_n=X_n-X_{n-1},X_0=0$.

显然, 如果

$$\mathbf{E}\sup_n|\Delta X_n|<\infty\tag{2}$$

则 $X\in\mathbf{C}^+$; 特别, 如果 ($\mathbf{P}$ – a.c.) 对于一切 $n\geqslant1$,

$$|\Delta X_n|\leqslant C<\infty.\tag{3}$$

定理 1 如果下鞅 $X\in\mathbf{C}^+$, 则 ($\mathbf{P}$ – a.c.)

$$\{\sup X_n<\infty\}=\{X_n\to\}.\tag{4}$$

证明 包含关系 $\{X_n\to\}\subseteq\{\sup X_n<\infty\}$ 显然. 为证明相反的包含关系, 考虑 "停止" 下鞅 $X^{\tau_a}=(X_{\tau_a\wedge n},\mathscr{F}_n)$. 那么, 由于 (1) 式, 有

$$\sup_n\mathbf{E}X_{\tau_a\wedge n}^+\leqslant a+\mathbf{E}[X_{\tau_a}^+I\{\tau_a<\infty\}]\leqslant 2a+\mathbf{E}[(\Delta X_{\tau_a})^+I\{\tau_a<\infty\}]<\infty.\tag{5}$$

从而, 由 §4 定理 1 ($\mathbf{P}$ – a.c.), 有

$$\{\tau_a=\infty\}\subseteq\{X_n\to\}.$$

由于

$$\bigcup_{a>0}\{\tau_a=\infty\}=\{\sup X_n<\infty\},$$

可见 $\{\sup X_n<\infty\}\subseteq\{X_n\to\}(\mathbf{P}-\text{a.c.})$. □

系 设 X 是鞅, 且 $\mathbf{E}\sup|\Delta X_n|<\infty$. 那么,

$$\{X_n\to\}\bigcup\{\underline{\lim}X_n=-\infty,\overline{\lim}X_n=+\infty\}=\Omega\ (\mathrm{P-a.c.}). \tag{6}$$

实际上, 将定理 1 用于 X 和 $-X$, 可得

$$\{\overline{\lim}X_n<\infty\}=\{\sup X_n<\infty\}=\{X_n\to\}\ (\mathrm{P-a.c.}),$$
$$\{\underline{\lim}X_n>-\infty\}=\{\inf X_n>-\infty\}=\{X_n\to\}\ (\mathrm{P-a.c.}).$$

因此,

$$\{\overline{\lim}X_n<\infty\}\bigcup\{\underline{\lim}X_n>-\infty\}=\{X_n\to\}\ (\mathrm{P-a.c.}),$$

于是, (6) 式得证,

命题 (6) 表示, 对于满足条件 $\mathbf{E}\sup|\Delta X_n|<\infty$ 的鞅 X 的几乎一切轨道, 或者存在有限极限, 或者这些轨道按如下意义上, 表现 "不好": $\overline{\lim}X_n=+\infty,\underline{\lim}X_n=-\infty$.

2. C^+类非负鞅的性质 如果 $\xi_1,\xi_2,\cdots$ 是独立随机变量序列, 且 $\mathbf{E}\xi_i=0,|\xi_i|\leqslant c<\infty$, 则根据第四章 §2 定理 1, 级数 $\sum\xi_i$ 收敛 $(\mathbf{P}-\mathrm{a.c.})$, 当且仅当 $\sum\xi_i^2<\infty$. 序列

$$X=(X_n,\mathscr{F}_n),\quad X_n=\xi_1+\cdots+\xi_n,\quad \mathscr{F}_n=\sigma\{\xi_1,\cdots,\xi_n\},$$

是平方可积鞅, 且 $\langle X\rangle_n=\sum\limits_{i=1}^n\xi_i^2$, 而可以将命题表述为如下形式:

$$\{\langle X\rangle_\infty<\infty\}=\{X_n\to\}=\Omega,\quad (\mathbf{P}-\mathrm{a.c.}),$$

其中 $\langle X\rangle_\infty=\lim\limits_n\langle X\rangle_n$.

下面的命题将这一结果推广到鞅和下鞅的更一般的情形.

定理 2 设 $X=(X_n,\mathscr{F}_n)$ 是下鞅, 而

$$X_n=m_n+A_n$$

是其杜布分解.

a) 如果 X 是非负下鞅, 则

$$\{A_\infty<\infty\}\subseteq\{X_n\to\}\subseteq\{\sup X_n<\infty\}\quad(\mathbf{P}-\mathrm{a.c.}). \tag{7}$$

b) 如果 $X\in\mathrm{C}^+$, 则

$$\{X_n\to\}=\{\sup X_n<\infty\}\subseteq\{A_\infty<\infty\}\quad(\mathbf{P}-\mathrm{a.c.}). \tag{8}$$

c) 如果 X 是非负下鞅, 且 $X\in\mathrm{C}^+$, 则

$$\{X_n\to\}=\{\sup X_n<\infty\}=\{A_\infty<\infty\}\quad(\mathbf{P}-\mathrm{a.c.}). \tag{9}$$

证明 a) (7) 式的第二个包含关系显然. 为证明 (7) 式的第一个包含关系, 引进时间

$$\sigma_a=\begin{cases}\inf\{n\geqslant1:A_{n+1}>a\}, & 若\ a>0,\\ +\infty, & 若\ \{\cdot\}=\varnothing.\end{cases}$$

那么, $A_{\sigma_a}\leqslant a$, 且由于 §2 定理 1 的系 1, 可见

$$\mathbf{E}X_{n\wedge\sigma_a}=\mathbf{E}A_{n\wedge\sigma_a}\leqslant a.$$

设 $Y_n^a=X_{n\wedge\sigma_a}$, 则 $Y^a=(Y_n^a,\mathscr{F}_n)$ 是下鞅, 其中 $\sup\mathbf{E}Y_n^a\leqslant a<\infty$, 而由于其非负性, 由 §4 定理 1, 可见.

$$\{A_\infty<\infty\}=\{\sigma_a=\infty\}\subseteq\{X_n\to\}\ (\mathbf{P}-\mathrm{a.c.}).$$

因此

$$\{A_\infty<\infty\}=\bigcup_{a>0}\{A_\infty\leqslant a\}\subseteq\{X_n\to\}\ (\mathbf{P}-\mathrm{a.c.}).$$

b) 由定理 1 可得 (8) 式的第一个等式. 为证明 (8) 式中的包含关系, 注意到, 根据关系式 (5), 有

$$\mathbf{E}A_{\tau_a\wedge n}=\mathbf{E}X_{\tau_a\wedge n}\leqslant\mathbf{E}X^+_{\tau_a\wedge n}\leqslant 2a+\mathbf{E}[(\Delta X_{\tau_a})^+I\{\tau_a<\infty\}],$$

从而

$$\mathbf{E}A_{\tau_a}=\mathbf{E}\lim_n A_{\tau_a\wedge n}<\infty.$$

于是, 由

$$\bigcup_{a>0}\{\tau_a=\infty\}=\{\sup X_n<\infty\},$$

得 $\{\tau_a=\infty\}\subseteq\{A_\infty<\infty\}$ 和所要证明的 (8) 式.

c) 该命题是命题 a) 和 b) 的直接推论. □

注 可以将 "X 的非负性" 条件换成 $\sup\limits_n\mathbf{E}X_n^-<\infty$.

系 1 设 $X_n=\xi_1+\cdots+\xi_n$, 其中 $\xi_i\geqslant0,\mathbf{E}\xi_i<\infty$, 且 ξ_i 为 $\mathscr{F}_i-$ 可测的, 而 $\mathscr{F}_0=\{\varnothing,\Omega\}$. 那么, $(\mathbf{P}\text{-a.c.})$

$$\left\{\sum_{n=1}^\infty\mathbf{E}(\xi_n|\mathscr{F}_{n-1})<\infty\right\}\subseteq\{X_n\to\}, \tag{10}$$

并且. 如果 $\mathbf{E}\sup\limits_n\xi_n<\infty$, 则$(\mathbf{P}\text{-a.c.})$

$$\left\{\sum_{n=1}^\infty\mathbf{E}(\xi_n|\mathscr{F}_{n-1})<\infty\right\}=\{X_n\to\}. \tag{11}$$

系 2 (博雷尔 – 坎泰利 – 列维引理) 如果事件 $B_n \in \mathscr{F}_n$, 则在 (11) 式中设 $\xi_n = I_{B_n}$, 得 ($\mathbf{P}$-a.c.)

$$\left\{\sum_{n=1}^{\infty} \mathbf{P}(B_n|\mathscr{F}_{n-1}) < \infty\right\} = \left\{\sum_{n=1}^{\infty} I_{B_n} < \infty\right\}. \tag{12}$$

3. 平方可积鞅的性质

定理 3 设 $M = (M_n, \mathscr{F}_n)_{n\geqslant 1}$ 是平方可积鞅, 则 ($\mathbf{P}$-a.c.)

$$\{\langle M\rangle_\infty < \infty\} \subseteq \{M_n \to\}. \tag{13}$$

假如还满足 $\mathbf{E}\sup|\Delta M_n|^2 < \infty$, 那么 ($\mathbf{P}$-a.c.)

$$\{\langle M\rangle_\infty < \infty\} = \{M_n \to\}, \tag{14}$$

其中

$$\langle M\rangle_\infty = \sum_{n=1}^{\infty} \mathbf{E}[(\Delta M_n)^2|\mathscr{F}_{n-1}], \tag{15}$$

而 $M_0 = 0$, $\mathscr{F}_0 = \{\varnothing, \Omega\}$.

证明 考虑两个下鞅: $M^2 = (M_n^2, \mathscr{F}_n)$ 和 $(M+1)^2 = ((M+1)_n^2, \mathscr{F}_n)$. 那么, 在它们的杜布分解

$$M^2 = m_n' + A_n' \quad \text{和} \quad (M+1)^2 = m_n'' + A_n''$$

中随机变量 A_n' 和 A_n'' 相等, 因为

$$A_n' = \sum_{k=1}^{n} \mathbf{E}(\Delta M_n^2|\mathscr{F}_{k-1})$$

而

$$A_n'' = \sum_{k=1}^{n} \mathbf{E}[\Delta(M_k+1)^2|\mathscr{F}_{k-1}] = \sum_{k=1}^{n} \mathbf{E}(\Delta M_n^2|\mathscr{F}_{k-1}) = A_n'.$$

从而, 由 (7) 式, 以概率 1 有

$$\{\langle M\rangle_\infty < \infty\} = \{A_\infty' < \infty\} \subseteq \{M_n^2 \to\} \bigcap \{(M_n+1)^2 \to\} = \{M_n \to\}.$$

由于 (9) 式, 为证明 (14) 式只需验证, 条件 $\mathbf{E}\sup|\Delta M_n|^2 < \infty$ 可以保障下鞅 M^2 属于 $\mathbf{C}^+$ 类: $M^2 \in \mathbf{C}^+$.

设 $\tau_a = \inf\{n \geqslant 1 : M_n^2 > a\}, a > 0$, 则在集合 $\{\tau_a < \infty\}$ 上, 有

$$\begin{aligned}|M_{\tau_a}^2| = |M_{\tau_a}^2 - M_{\tau_a-1}^2| &\leqslant |M_{\tau_a} - M_{\tau_a-1}|^2 + 2|M_{\tau_a-1}|\cdot|M_{\tau_a} - M_{\tau_a-1}| \\ &\leqslant (\Delta M_{\tau_a})^2 + 2a^{1/2}|\Delta M_{\tau_a}|,\end{aligned}$$

由此, 得

$$\begin{aligned}\mathbf{E}|M_{\tau_a}^2|I\{\tau_a < \infty\} &\leqslant \mathbf{E}(\Delta M_{\tau_a})^2 I\{\tau_a < \infty\} + 2a^{1/2}\sqrt{\mathbf{E}(\Delta M_{\tau_a})^2 I\{\tau_a < \infty\}} \\ &\leqslant \mathbf{E}\sup|\Delta M_n|^2 + 2a^{1/2}\sqrt{\mathbf{E}\sup|\Delta M_n|^2} < \infty.\end{aligned}$$

□

作为应用这一定理的例子, 我们引进下面的结果, 可以视为平方可积鞅的强大数定律的独特形式 (对照第四章 §3 定理 2, 以及第四章第 3 小节系 2).

定理 4 设 $M = (M_n, \mathscr{F}_n)$ 是平方可积鞅, 而 $A = (A_n, \mathscr{F}_{n-1})$ 是可预测的递增序列, 且 $A_1 \geqslant 1, A_\infty = \infty(\mathbf{P} - \text{a.c.})$.

如果 ($\mathbf{P}$ – a.c.)

$$\sum_{i=1}^{\infty} \frac{\mathbf{E}[(\Delta M_i)^2|\mathscr{F}_{i-1}]}{A_i^2} < \infty, \tag{16}$$

则以概率 1, 有

$$\frac{M_n}{A_n} \to 0, \quad n \to \infty. \tag{17}$$

特别, 如果 $\langle M\rangle = (\langle M\rangle_n, \mathscr{F}_{n-1})$ 是平方可积鞅 $M = (M_n, \mathscr{F}_n)$ 的二次特征, 且 $\langle M\rangle_\infty = \infty(\mathbf{P} - \text{a.c.})$, 则以概率 1, 有

$$\frac{M_n}{\langle M\rangle_n} \to 0, \quad n \to \infty. \tag{18}$$

证明 考虑平方可积鞅 $m = (m_n, \mathscr{F}_n)$, 且

$$m_n = \sum_{i=1}^{n} \frac{\Delta M_i}{A_i}.$$

那么,

$$\langle m\rangle_n = \sum_{k=1}^{n} \frac{\mathbf{E}[(\Delta M_i)^2|\mathscr{F}_{i-1}]}{A_i^2}. \tag{19}$$

由于

$$\frac{M_n}{A_n} = \frac{\sum_{k=1}^{n} A_k \Delta m_k}{A_n},$$

可见, 根据克罗内克引理 (第四章 §3), 如果以概率 1 存在有限极限 $\lim m_n$, 则

$$\frac{M_n}{A_n} \to 0, \quad n \to \infty \ (\mathbf{P} - \text{a.c.}).$$

而由于 (13) 式, 有

$$\{\langle m\rangle_\infty < \infty\} \leqslant \{m_n \to\}, \tag{20}$$

因此由 (19) 式可见, (16) 式是 (17) 式成立的充分条件.

最后, 假如 $A_n = \langle M\rangle_n$, 则条件 (16) 自然成立 (练习题 6). □

例 考虑独立随机变量序列 $\xi_1, \xi_2, \cdots$, 且 $\mathbf{E}\xi_i = 0, \mathbf{D}\xi_i = D_i > 0$, 而序列 $X = (X_n)_{n\geqslant 0}$ 决定于如下递推方程:

$$X_{n+1} = \theta X_n + \xi_{n+1}, \tag{21}$$

其中 X_0 不依赖于 $\xi_1, \xi_2 \cdots$, 而 $\theta(-\infty < \theta < \infty)$ 是未知参数.

我们把 X_n 视为在时刻 n 的观测结果, 并且要求估计未知参数 θ. 以量

$$\widehat{\theta}_n = \frac{\sum_{k=0}^{n-1} \frac{X_k X_{k+1}}{D_{k+1}}}{\sum_{k=0}^{n-1} \frac{X_k^2}{D_{k+1}}}, \tag{22}$$

为根据观测结果 $X_0, X_1, \cdots, X_n$, 对 θ 的估计, 且当分母为 0 时, 设 $\widehat{\theta}_n = 0$. ($\widehat{\theta}_n$ 是由**最小二乘法**得来的估计.)

由 (21) 和 (22) 式, 显然有

$$\widehat{\theta}_n = \theta + \frac{M_n}{A_n},$$

其中

$$M_n = \sum_{k=0}^{n-1} \frac{X_k \xi_{k+1}}{D_{k+1}}, \quad A_n = \langle M \rangle_n = \sum_{k=0}^{n-1} \frac{X_k^2}{D_{k+1}}.$$

因此, 假如未知参数的真值为 θ, 则

$$\mathbf{P}\{\widehat{\theta}_n \to \theta\} = 1 \tag{23}$$

当且仅当 ($\mathbf{P}$ – a.c.)

$$\frac{M_n}{A_n} \to 0, \quad n \to \infty. \tag{24}$$

现在证明, 条件

$$\sup_n \frac{D_{n+1}}{D_n} < \infty, \quad \sum_{n=1}^{\infty} \left(\frac{\xi_n^2}{D_n} \wedge 1\right) = \infty, \tag{25}$$

对于 (24) 式是充分的, 因此对于 (23) 式也是充分的. 有

$$\begin{aligned}\sum_{n=1}^{\infty} \left(\frac{\xi_n^2}{D_n} \wedge 1\right) &\leqslant \sum_{n=1}^{\infty} \frac{\xi_n^2}{D_n} = \sum_{n=1}^{\infty} \frac{(X_n - \theta X_{n-1})^2}{D_n} \\ &\leqslant 2\left[\sum_{n=1}^{\infty} \frac{X_n^2}{D_n} + \theta^2 \sum_{n=1}^{\infty} \frac{X_{n-1}^2}{D_n}\right] \leqslant 2\left[\sup \frac{D_{n+1}}{D_n} + \theta^2\right] \langle M \rangle_\infty.\end{aligned}$$

于是,

$$\left\{\sum_{n=1}^{\infty} \left(\frac{\xi_n^2}{D_n} \wedge 1\right) = \infty\right\} \subseteq \{\langle M \rangle_\infty = \infty\}.$$

根据 “三级数” 定理 (第四章 §2 定理 3), 若级数

$$\sum_{n=1}^{\infty} \mathbf{E}\left(\frac{\xi_n^2}{D_n} \wedge 1\right)$$

发散, 则级数

$$\sum_{n=1}^{\infty} \left(\frac{\xi_n^2}{D_n} \wedge 1\right),$$

以概率 1 也发散. 从而, $\mathbf{P}\{\langle M \rangle_\infty = \infty\} = 1$, 故由定理 4 的论断 (7) 直接得 (24) 式.

具有性质 (23) 的估计量 $\widehat{\theta}_n, n \geqslant 1$, 称做**强相合的** (对照第一章 §7 中 “相合性” 的概念).

在下一节 §6 的第 5 小节, 对于高斯序列 $\xi_1, \xi_2 \cdots$ 的情形, 将继续这个例子.

定理 5 *设 $X = (X_n, \mathscr{F}_n)$ 是下鞅, 而*

$$X_n = m_n + A_n$$

是其杜布分解. 如果 $|\Delta X_n| \leqslant C$, 则 ($\mathbf{P}$ – a.c.)

$$\{\langle m \rangle_\infty + A_\infty < \infty\} = \{X_n \to\}, \tag{26}$$

或同样地,

$$\left\{\sum_{n=1}^{\infty} \mathbf{E}[\Delta X_n + (\Delta X_n)^2 | \mathscr{F}_{n-1}]\right\} = \{X_n \to\}. \tag{27}$$

证明 由于

$$A_n = \sum_{n=1}^{\infty} \mathbf{E}[(\Delta X_n)^2 | \mathscr{F}_{n-1}], \tag{28}$$

$$m_n = \sum_{k=1}^{n} \mathbf{E}[\Delta X_k - (\Delta X_n | \mathscr{F}_{k-1})], \tag{29}$$

则由于假设 $|\Delta X_n| \leqslant C$, 鞅 $m = (m_n, \mathscr{F}_n)$ 平方可积, 且 $|\Delta m_n| \leqslant 2C$. 那么, 由 (13) 式, 有

$$\{\langle m \rangle_\infty + A_\infty < \infty\} \subseteq \{X_n \to\}, \tag{30}$$

而且, 根据 (8) 式, 有

$$\{X_n \to\} \subseteq \{A_\infty < \infty\}.$$

因此, 由 (14) 式和 (30) 式, 可见

$$\begin{aligned}\{X_n \to\} &= \{X_n \to\} \bigcap \{A_\infty < \infty\} = \{X_n \to\} \bigcap \{A_\infty < \infty\} \bigcap \{m_n \to\} \\ &= \{X_n \to\} \bigcap \{A_\infty < \infty\} \bigcap \{\langle m \rangle_\infty < \infty\} \\ &= \{X_n \to\} \bigcap \{A_\infty + \langle m \rangle_\infty < \infty\} = \{A_\infty + \langle m \rangle_\infty < \infty\}.\end{aligned}$$

最后, 由于 (29) 式, 有

$$\langle m\rangle_n=\sum_{k=1}^{n}\left\{\mathbf{E}[(\Delta X_k)^2|\mathscr{F}_{k-1}]-[\mathbf{E}(\Delta X_k|\mathscr{F}_{k-1})]^2\right\};$$

可见 (26) 式和 (27) 式等价. 而由非负项级数

$$\sum_{k=1}^{\infty}\mathbf{E}(\Delta X_k|\mathscr{F}_{k-1})$$

的收敛性, 可见级数

$$\sum_{k=1}^{\infty}[\mathbf{E}(\Delta X_k|\mathscr{F}_{k-1})]^2$$

的收敛性. □

4. 级数 $\sum\xi_n$ 的收敛集合 柯尔莫戈洛夫 "三级数" 定理 (第四章 §2 定理 3) 给出了, 独立随机变量的级数 $\sum\xi_n$ 以概率 1 收敛的充分且必要条件. 下面的定理 6, 在不假设随机变量 $\xi_1,\xi_2,\cdots$ 独立的情形下, 给出了级数 $\sum\xi_n$ 之收敛集合的描述, 而其证明基于定理 2 和定理 3.

定理 6 *假设 $\xi=(\xi_n,\mathscr{F}_n)_{n\geqslant1}$ 是随机变量序列, $\mathscr{F}_0=(\varnothing,\Omega)$ 而 c 是正常数. 那么, 级数 $\sum\xi_n$ 在集合 A 上收敛, 其中 A 是使三个级数*

$$\sum\mathbf{P}(|\xi_n|\geqslant c|\mathscr{F}_{n-1}),\quad\sum\mathbf{E}(\xi_n^c|\mathscr{F}_{n-1}),\quad\sum\mathbf{D}(\xi_n^c|\mathscr{F}_{n-1}),$$

同时收敛的集合, 其中 $\xi_n^c=\xi_nI(|\xi_n|\leqslant c)$.

证明 设级数 $X_n=\sum\limits_{k=1}^{n}\xi_k$. 由于级数 $\sum\mathbf{P}(|\xi_n|\geqslant c|\mathscr{F}_{n-1})$ (在集合 A 上) 收敛, 则根据定理 2 的系 2, 以及级数 $\sum\mathbf{E}(\xi_n^c|\mathscr{F}_{n-1})$ 收敛, 有

$$\begin{aligned}A\bigcap\{X_n\to\}&=A\bigcap\left\{\sum_{k=1}^{n}\xi_kI(|\xi_k|\leqslant c)\to\right\}\\&=A\bigcap\left\{\sum_{k=1}^{n}[\xi_kI(|\xi_k|\leqslant c)-\mathbf{E}(\xi_kI(|\xi_k|\leqslant c)|\mathscr{F}_{n-1})]\to\right\}.\end{aligned}\tag{31}$$

设 $\eta_k=\xi_k^c-\mathbf{E}(\xi_k^c|\mathscr{F}_{n-1})$ 和 $Y_n=\sum\limits_{k=1}^{n}\eta_k$. 那么, $Y=(Y_n,\mathscr{F}_n)$ 是平方可积鞅, 其中 $|\eta_k|\leqslant2c$. 根据定理 3,

$$A\subseteq\left\{\sum\mathbf{D}(\xi_n^c|\mathscr{F}_{n-1})<\infty\right\}=\{\langle Y\rangle_\infty<\infty\}=\{Y_n\to\}.$$

因此, 由 (31) 可见

$$A\bigcap\{X_n\to\}=A.$$

于是, $A\subseteq\{X_n\to\}$. □

5. 练习题

1. 证明, 如果下鞅 $X=(X_n,\mathscr{F}_n)$ 满足条件 $\mathbf{E}\sup\limits_n|X_n|<\infty$, 则它属于 $\mathbf{C}^+$ 类.
2. 证明, 对于广义下鞅, 定理 1 和定理 2 仍然成立.
3. 证明, 对于广义下鞅, 以概率 1 有包含关系:

$$\left\{\inf_m\sup_{n\geqslant m}\mathbf{E}(X_n^+|\mathscr{F}_m)<\infty\right\}\subseteq\{X_n\to\}.$$

4. 证明, 对于广义鞅, 定理 1 的系仍然成立.
5. 证明, $\mathbf{C}^+$ 类的任何广义下鞅是局部下鞅.
6. 设 $a_n>0,n\geqslant1;b_n=\sum\limits_{k=1}^{n}a_k$, 证明

$$\sum_{k=1}^{\infty}\frac{a_n}{b_n^2}<\infty.$$

7. 设 $\xi_0,\xi_1,\xi_2,\cdots$ 是一致有界随机变量序列: $|\xi_n|\leqslant c,n\geqslant1$, 证明级数

$$\sum_{n\geqslant0}\xi_n\quad\text{和}\quad\sum_{n\geqslant1}\mathbf{E}(\xi_n|\xi_1,\cdots,\xi_{n-1}),$$

以概率 1 或者同时收敛, 或者同时发散.

§6. 概率测度在带滤子可测空间上的绝对连续性和奇异性

1. 测度的局部绝对连续性和奇异性 设 $(\Omega,\mathscr{F}_n)$ 是某一可测空间, 而 $(\mathscr{F}_n)_{n\geqslant1}$ 是该空间上的 $\sigma-$ 代数族, 且 $\mathscr{F}_1\subseteq\subseteq\mathscr{F}_2\subseteq\cdots\subseteq\mathscr{F}$, 其中

$$\mathscr{F}=\sigma\left(\bigcup_{n=1}^{\infty}\mathscr{F}_n\right).\tag{1}$$

假设在 $(\Omega,\mathscr{F}_n)$ 上有两个概率测度 $\mathbf{P}$ 和 $\widetilde{\mathbf{P}}$. 记

$$\mathbf{P}_n=\mathbf{P}|\mathscr{F}_n\quad\widetilde{\mathbf{P}}_n=\widetilde{\mathbf{P}}|\mathscr{F}_n$$

为两个概率测度在上 $\mathscr{F}_n$ 的收缩, 即 $\mathbf{P}_n$ 和 $\widetilde{\mathbf{P}}_n$ 是 $(\Omega,\mathscr{F}_n)$ 上的概率测度, 并且对于 $B\in\mathscr{F}_n$, 有

$$\mathbf{P}_n(B)=\mathbf{P}(B),\quad\widetilde{\mathbf{P}}_n(B)=\widetilde{\mathbf{P}}(B).$$

熟知, 测度 $\widetilde{\mathbf{P}}$ 称做关于测度 $\mathbf{P}$ *绝对连续的* (记作 $\widetilde{\mathbf{P}}\ll\mathbf{P}$), 如果 $\mathbf{P}_n(A)=0,A\in\mathscr{F}_n$, 则必有 $\widetilde{\mathbf{P}}_n(A)=0$.

如果 $\widetilde{\mathbf{P}}\ll\mathbf{P}$ 且 $\mathbf{P}\ll\widetilde{\mathbf{P}}$, 则称测度 $\mathbf{P}$ 和 $\widetilde{\mathbf{P}}$ 为*等价的* (记作 $\widetilde{\mathbf{P}}\sim\mathbf{P}$).

测度 $\widetilde{\mathbf{P}}$ 和 $\mathbf{P}$ 称做*奇异的*或*正交的* (记作 $\widetilde{\mathbf{P}}\perp\mathbf{P}$), 如果存在集合 $A\in\mathscr{F}$, 使 $\widetilde{\mathbf{P}}(A)=1$和 $\mathbf{P}(\overline{A})=1$.

定义 1 称测度 $\widetilde{\mathbf{P}}$ 关于测度 $\mathbf{P}$ 为局部绝对连续的 (记作 $\widetilde{\mathbf{P}} \overset{\mathrm{loc}}{\ll} \mathbf{P}$), 如果对于任意 $n \geqslant 1$, 有

$$\widetilde{\mathbf{P}}_n \ll \mathbf{P}_n. \tag{2}$$

这一节考虑的基本问题在于说明, 由局部绝对连续性 $\widetilde{\mathbf{P}} \overset{\mathrm{loc}}{\ll} \mathbf{P}$, 说明 $\widetilde{\mathbf{P}} \ll \mathbf{P}, \widetilde{\mathbf{P}} \sim \mathbf{P}, \widetilde{\mathbf{P}} \perp \mathbf{P}$ 等性质成立的条件. 我们以后将清楚地看到, 鞅论是可以全面地回答这些问题的数学工具.

回忆, 在第三章 §9 中. 对于任意概率测度, 讨论过概率测度的绝对连续性和奇异性问题. 曾经说明, 利用海林格 (E. Helinger) 积分可以表述相应的准则 (定理 2 和定理 3). 下面引进的关于概率测度的绝对连续性和奇异性的结果, 亦可由这些准则得到 (在专著 [84], [87] 中陈述了有关方法). 为更完整地演示运用 §5 中关于下鞅的收敛集合, 我们在此倾向于略有不同的叙述方法. (注意, 这一节全部叙述, 假设局部绝对连续性成立的. 这样做仅仅是为了简便. 对于一般情形, 请读者参阅 [84], [87]).

这样, 假设 $\widetilde{\mathbf{P}} \overset{\mathrm{loc}}{\ll} \mathbf{P}$. 设

$$z_n = \frac{d\widetilde{\mathbf{P}}_n}{d\mathbf{P}_n}$$

是 $\widetilde{\mathbf{P}}_n$ 关于 $\mathbf{P}_n$ 的拉东 – 尼科迪姆导数. 显然, z_n 为 $\mathscr{F}_n$– 可测, 并且如果 $A \in \mathscr{F}_n$, 则

$$\int_A z_{n+1} d\mathbf{P} = \int_A \frac{d\widetilde{\mathbf{P}}_{n+1}}{d\mathbf{P}_{n+1}} d\mathbf{P} = \widetilde{\mathbf{P}}_{n+1}(A) = \widetilde{\mathbf{P}}_n(A) = \int_A \frac{d\widetilde{\mathbf{P}}_n}{d\mathbf{P}_n} d\mathbf{P} = \int_A z_n d\mathbf{P}.$$

由此可见, 随机序列 $z = (z_n, \mathscr{F}_n)_{n\geqslant 1}$ 关于测度 $\mathbf{P}$ 是鞅.

在 "绝对连续性和奇异性" 的全部问题中, 关键环节是下面的定理.

定理 1 *设* $\widetilde{\mathbf{P}} \overset{\mathrm{loc}}{\ll} \mathbf{P}$.

a) $(\mathbf{P} + \widetilde{\mathbf{P}})/2$ *以概率 1 存在极限* $\lim\limits_n z_n$, *记为* z_∞, *使*

$$\mathbf{P}\{z_\infty = \infty\} = 0.$$

b) *有如下勒贝格分解:*

$$\widetilde{\mathbf{P}}(A) = \int_A z_\infty d\mathbf{P} + \widetilde{\mathbf{P}}(A \cap \{z_\infty = \infty\}), \quad A \in \mathscr{F}, \tag{3}$$

而且测度 $\widetilde{\mathbf{P}}(A \cap \{z_\infty = \infty\})$ *和* $\mathbf{P}(A), A \in \mathscr{F}$, *是奇异的.*

证明 我们首先指出, 根据经典的勒贝格分解 (第三章 §9 的 (29) 式), 任意概率测度 $\widetilde{\mathbf{P}}$ 关于概率测度 $\mathbf{P}$ 有如下表现:

$$\widetilde{\mathbf{P}}(A) = \int_A \frac{\widetilde{\mathfrak{z}}}{\mathfrak{z}} d\mathbf{P} + \widetilde{\mathbf{P}}(A \cap \{\mathfrak{z} = 0\}), \quad A \in \mathscr{F}, \tag{4}$$

其中

$$\mathfrak{z} = \frac{d\mathbf{P}}{d\mathbf{Q}}, \quad \widetilde{\mathfrak{z}} = \frac{d\widetilde{\mathbf{P}}}{d\mathbf{Q}},$$

而其中测度 $\mathbf{Q}$, 例如可以取作 $\mathbf{Q} = (\mathbf{P} + \widetilde{\mathbf{P}})/2$. 这样, (3) 式可以视为一般分解 (4) 的具体化, 与所考虑情形的如下特征相联系: $\widetilde{\mathbf{P}} \overset{\mathrm{loc}}{\ll} \mathbf{P}$ 即 $\widetilde{\mathbf{P}}_n \ll \mathbf{P}_n, n \geqslant 1$. 设

$$\mathfrak{z}_n = \frac{d\mathbf{P}_n}{d\mathbf{Q}_n}, \quad \widetilde{\mathfrak{z}}_n = \frac{d\widetilde{\mathbf{P}}_n}{d\mathbf{Q}_n}, \quad \mathbf{Q}_n = \frac{1}{2}(\mathbf{P}_n + \widetilde{\mathbf{P}}_n).$$

序列 $(\mathfrak{z}_n, \mathscr{F}_n)$ 和 $(\widetilde{\mathfrak{z}}_n, \mathscr{F}_n)$ 关于测度 $\mathbf{Q}$ 是鞅, 并且满足 $0 \leqslant \mathfrak{z}_n \leqslant 2, 0 \leqslant \widetilde{\mathfrak{z}}_n \leqslant 2$. 因此 (根据 §4 定理 2), 既关于测度 $\mathbf{Q}$ 几乎处处, 又在 $L^1(\Omega, \mathscr{F}, \mathbf{Q})$ 的收敛意义上存在极限:

$$\mathfrak{z}_\infty = \lim_n \mathfrak{z}_n, \quad \widetilde{\mathfrak{z}}_\infty = \lim_n \widetilde{\mathfrak{z}}_n. \tag{5}$$

特别, 由在 $L^1(\Omega, \mathscr{F}, \mathbf{Q})$ 的意义上收敛, 可见对于任意 $A \in \mathscr{F}_m$, 有

$$\int_A \widetilde{\mathfrak{z}}_\infty d\mathbf{Q} = \lim_{n\uparrow\infty} \int_A \widetilde{\mathfrak{z}}_n d\mathbf{Q} = \int_A \widetilde{\mathfrak{z}}_m d\mathbf{Q} = \widetilde{\mathbf{P}}_m(A) = \widetilde{\mathbf{P}}(A).$$

那么, 由卡拉泰奥多里 (C. Carathéodory) 定理 (第二章 §3) 可见, 对于任意 $A \in \mathscr{F} = \sigma\left(\bigcup\limits_n \mathscr{F}_n\right)$, 有

$$\int_A \widetilde{\mathfrak{z}}_\infty d\mathbf{Q} = \widetilde{\mathbf{P}}(A),$$

即 $d\widetilde{\mathbf{P}}/d\mathbf{Q} = \widetilde{\mathfrak{z}}_\infty$; 类似地, 有

$$\int_A \mathfrak{z}_\infty d\mathbf{Q} = \mathbf{P}(A),$$

即 $d\mathbf{P}/d\mathbf{Q} = \mathfrak{z}_\infty$.

从而, 这就证明了自然期待的结果: *假如测度* $\mathbf{P}$ *和* $\mathbf{Q}$ *定义在* $\mathscr{F} = \sigma\left(\bigcup\limits_n \mathscr{F}_n\right)$ *上, 而* $\mathbf{P}_n$ *和* $\mathbf{Q}_n$ *是这些测度到* $\mathscr{F}_n$ *上的收缩, 则关于测度* $\mathbf{Q}$ *几乎处处和在* $L^1(\Omega, \mathscr{F}, \mathbf{Q})$ *收敛的意义上, 有*

$$\lim_n \frac{d\mathbf{P}_n}{d\mathbf{Q}_n} = \frac{d\mathbf{P}}{d\mathbf{Q}}.$$

类似地, 有

$$\lim_n \frac{d\widetilde{\mathbf{P}}_n}{d\mathbf{Q}_n} = \frac{d\widetilde{\mathbf{P}}}{d\mathbf{Q}}.$$

在我们所讨论的特别情形下, 即当 $\widetilde{\mathbf{P}}_n \ll \mathbf{P}_n, n \geqslant 1$, 时, 不难证明, ($\mathbf{Q}$–a.c.) 有

$$z_n = \frac{\widetilde{\mathfrak{z}}}{\mathfrak{z}}, \tag{6}$$

这时, 由于

$$\mathbf{Q}\{\mathfrak{z}_n = 0, \widetilde{\mathfrak{z}}_n = 0\} \leqslant \frac{1}{2}[\mathbf{P}\{\mathfrak{z}_n = 0\} + \widetilde{\mathbf{P}}\{\widetilde{\mathfrak{z}}_n = 0\}] = 0,$$

因此, 可见在 (6) 式中 (**Q**−a.c.) 不会出现形如 0/0 不定式的情形.

形如 2/0 的记号, 像通常一样认为等于 $+\infty$. 需要指出, 因为 $(z_n,\mathscr{F}_n)$ 是非负鞅, 所以由 §2 的 (5) 式可见, 如果 $z_\tau=0$, 则对于一切 $n\geqslant\tau$ (**Q**−a.c.), $z_n=0$. 当然, 同样对于 $(\widetilde{z}_n,\mathscr{F}_n)$ 也是一样. 由此可见, 点 0 和 $+\infty$ 也是序列 $(z_n)_{n\geqslant1}$ 的 "吸收状态".

由于 (5) 式和 (6) 式可见, **Q**−a.c. 存在极限

$$z_\infty\equiv\lim_n z_n=\frac{\lim\limits_n\widetilde{\mathfrak{z}}_n}{\lim\limits_n\mathfrak{z}_n}=\frac{\widetilde{\mathfrak{z}}_\infty}{\mathfrak{z}_\infty}. \tag{7}$$

因为

$$\mathbf{P}\{\mathfrak{z}_\infty=0\}=\int_{\{\mathfrak{z}_\infty=0\}}\mathfrak{z}_\infty\mathrm{d}\mathbf{Q}=0,$$

所以 $\mathbf{P}\{z_\infty=\infty\}=0$, 因而定理的命题 a) 得证.

为证明 (3) 式, 利用 (4) 式的一般分解. 对于现在所考虑的情形, 由已经证明的结果:

$$\mathfrak{z}=\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}=\mathfrak{z}_\infty,\quad \widetilde{\mathfrak{z}}=\frac{\mathrm{d}\widetilde{\mathbf{P}}}{\mathrm{d}\mathbf{Q}}=\widetilde{\mathfrak{z}}_\infty\quad(\mathbf{Q}-\text{a.c.}),$$

知由 (4) 式, 得

$$\widetilde{\mathbf{P}}(A)=\int_A\frac{\widetilde{\mathfrak{z}}_\infty}{\mathfrak{z}_\infty}\mathrm{d}\mathbf{P}+\widetilde{\mathbf{P}}(A\bigcap\{\mathfrak{z}_\infty=0\});$$

由 (7) 式, 以及由于 $\widetilde{\mathbf{P}}\{\widetilde{\mathfrak{z}}_\infty=0\}=0$, 得所要证明的 (3) 式. 最后注意到, 由于 $\mathbf{P}\{z_\infty<\infty\}=1$, 可见对于 $A\in\mathscr{F}$, 则测度

$$\mathbf{P}(A)\equiv\mathbf{P}(A\bigcap\{z_\infty<\infty\})\quad\text{和}\quad\widetilde{\mathbf{P}}(A\bigcap\{z_\infty=\infty\})$$

是奇异的. □

由勒贝格分解 (3), 可得如下关于局部绝对连续概率测度的, 绝对连续性和奇异性重要的准则.

定理 2 设 $\widetilde{\mathbf{P}}\overset{\text{loc}}{\ll}\mathbf{P}$, 即 $\widetilde{\mathbf{P}}_n\ll\mathbf{P}_n, n\geqslant1$. 那么,

$$\widetilde{\mathbf{P}}\ll\mathbf{P}\Leftrightarrow\mathbf{E}z_\infty=1\Leftrightarrow\widetilde{\mathbf{P}}\{z_\infty<\infty\}=1, \tag{8}$$

$$\widetilde{\mathbf{P}}\perp\mathbf{P}\Leftrightarrow\mathbf{E}z_\infty=0\Leftrightarrow\widetilde{\mathbf{P}}\{z_\infty=\infty\}=1, \tag{9}$$

其中 **E** 表示对测度 **P** 的平均 (数学期望).

证明 在 (3) 式中设 $A=\Omega$, 得

$$\mathbf{E}z_\infty=1\Leftrightarrow\widetilde{\mathbf{P}}\{z_\infty=\infty\}=0, \tag{10}$$

$$\mathbf{E}z_\infty=0\Leftrightarrow\widetilde{\mathbf{P}}\{z_\infty=\infty\}=1. \tag{11}$$

如果 $\widetilde{\mathbf{P}}\{z_\infty=\infty\}=0$, 则仍然由 (3) 式可见 $\widetilde{\mathbf{P}}\ll\mathbf{P}$.

相反, 假设 $\widetilde{\mathbf{P}}\ll\mathbf{P}$. 那么, 由于 $\mathbf{P}\{z_\infty=\infty\}=0$, 可见 $\widetilde{\mathbf{P}}\{z_\infty=\infty\}=0$.

其次, 假如 $\widetilde{\mathbf{P}}\perp\mathbf{P}$, 则存在集合 $B\in\mathscr{F}$, 使 $\widetilde{\mathbf{P}}(B)=1$ 和 $\mathbf{P}(B)=0$. 那么, 由 (3) 式可见 $\widetilde{\mathbf{P}}(B\bigcap\{z_\infty=\infty\})=1$, 因此 $\widetilde{\mathbf{P}}\{z_\infty=\infty\}=1$. 相反, 假如 $\widetilde{\mathbf{P}}\{z_\infty=\infty\}=1$, 则因为 $\widetilde{\mathbf{P}}\{z_\infty=\infty\}=0$, 所以性质 $\widetilde{\mathbf{P}}\perp\mathbf{P}$ 显然. □

2. 应用绝对连续性和奇异性准则的例 由定理 2 知, 绝对连续性和奇异性准则, 或者可以用测度 **P** 的术语表示 (并验证等式 $\mathbf{E}z_\infty=1$ 或 $\mathbf{E}z_\infty=0$), 或者可以通过测度 $\widetilde{\mathbf{P}}$ 的术语表示 (并验证等式 $\widetilde{\mathbf{P}}\{z_\infty<\infty\}=1$ 或 $\widetilde{\mathbf{P}}\{z_\infty=\infty\}=1$).

由第二章 §6 定理 5 知, 条件 $\mathbf{E}z_\infty=1$ 等价于条件 "随机变量族 $\{z_n\}_{n\geqslant1}$ (关于测度 **P**) 一致可积". 这一情况, 使得可以给出绝对连续性 $\widetilde{\mathbf{P}}\ll\mathbf{P}$ 简单的充分条件. 例如, 若

$$\sup_n\mathbf{E}[z_n\ln^+z_n]<\infty \tag{12}$$

或如果

$$\sup_n\mathbf{E}[z_n^{1+\varepsilon}]<\infty,\quad\varepsilon>0, \tag{13}$$

则由第二章 §6 引理 3 知, 随机变量族 $\{z_n\}_{n\geqslant1}$ 一致可积, 因此 $\widetilde{\mathbf{P}}\ll\mathbf{P}$.

在许多情形下, 更倾向于运用通过测度 $\widetilde{\mathbf{P}}$ 的术语表示的准则, 因为这时可以归结为 "尾部" 事件 $\{z_\infty<\infty\}$ 之 $\widetilde{\mathbf{P}}$− 概率的研究, 而这里可以利用 "0–1" 律类型的论点. 作为例子, 我们现在讨论如何由定理 2 导出角谷择一性.

设 $(\Omega,\mathscr{F},\mathbf{P})$ 是某一概率空间, $(\mathbb{R}^\infty,\mathscr{B}_\infty)$ 是数列 $x=(x_1,x_2,\cdots)$ 的可测空间, 其中 $\mathscr{B}_\infty=\mathscr{B}(\mathbb{R}^\infty)$, 而 $\mathscr{B}_n=\sigma(x_1,\cdots,x_n)$. 假设 $\xi=(\xi_1,\xi_2,\cdots)$ 和 $\widetilde{\xi}=(\widetilde{\xi}_1,\widetilde{\xi}_2,\cdots)$ 是两个独立随机变量序列.

记 P 和 $\widetilde{P}$ 相应为随机变量 ξ 和 $\widetilde{\xi}$ 在 $(\mathbb{R}^\infty,\mathscr{B}_\infty)$ 上的概率分布, 即

$$P(B)=\mathbf{P}\{\xi\in B\},\quad\widetilde{P}(B)=\mathbf{P}\{\widetilde{\xi}\in B\},\quad B\in\mathscr{B}_\infty.$$

亦假设

$$P_n=P|\mathscr{B}_n,\widetilde{P}_n=\widetilde{P}|\mathscr{B}_n$$

是 P 和 $\widetilde{P}$ 测度在 $\mathscr{B}_n$ 上的收缩, 且对于 $A\in\mathscr{B}(\mathbb{R}^1)$,

$$P_{\xi_n}(A)=\mathbf{P}\{\xi_n\in A\},$$

$$P_{\widetilde{\xi}_n}(A)=\mathbf{P}\{\widetilde{\xi}_n\in B\}.$$

定理 3 (角谷 (S. Kakutani) 择一性) 假设 $\xi=(\xi_1,\xi_2,\cdots)$ 和 $\widetilde{\xi}=(\widetilde{\xi}_1,\widetilde{\xi}_2,\cdots)$ 是两个独立随机变量序列, 且满足条件

$$P_{\widetilde{\xi}_n}\ll P_{\xi_n},\quad n\geqslant1. \tag{14}$$

那么, 要么 $\widetilde{P}\ll P$, 要么 $\widetilde{P}\perp P$.

证明 显然, 条件 (14) 等价于条件: $\widetilde{P}_n \ll P_n, n \geqslant 1$, 即 $\widetilde{P} \overset{\text{loc}}{\ll} P$.

$$z_n = \frac{\mathrm{d}\widetilde{P}_n}{\mathrm{d}P_n} = q_1(x_1)\cdots q_n(x_n),$$

其中

$$q_i(x_i) = \frac{\mathrm{d}P_{\widetilde{\xi}_i}}{\mathrm{d}P_{\xi_i}}(x_i). \tag{15}$$

从而

$$\{x : z_\infty < \infty\} = \{x : \ln z_\infty < \infty\} = \left\{x : \sum_{i=1}^{\infty} \ln q_i(x_i) < \infty\right\}.$$

事件

$$\left\{x : \sum_{i=1}^{\infty} \ln q_i(x_i) < \infty\right\}$$

是 "尾部的". 因此, 由柯尔莫戈洛夫 "0–1" 律 (第四章 §1 定理 1) 知, 概率 $\widetilde{P}\{x : z_\infty < \infty\}$ 只有 "0 或 1" 两个可能值. 于是, 根据定理 2, 要么 $\widetilde{P} \ll P$, 要么 $\widetilde{P} \perp P$.□

3. 测度的绝对连续性和奇异性与 "可预测性" 下面的定理用 "可预测的" 语言, 给出绝对连续性和奇异性的准则.

定理 4 设 $\widetilde{\mathbf{P}} \overset{\text{loc}}{\ll} \mathbf{P}$,

$$\alpha_n = z_n z_{n-1}^{\oplus}, \quad n \geqslant 1,$$

其中 $z_0 = 1$. 那么 ($\mathscr{F}_0 = \{\varnothing, \Omega\}$),

$$\widetilde{\mathbf{P}} \ll \mathbf{P} \Leftrightarrow \widetilde{\mathbf{P}}\left\{\sum_{n=1}^{\infty}[1 - \mathbf{E}(\sqrt{\alpha_n}|\mathscr{F}_{n-1})] < \infty\right\} = 1, \tag{16}$$

$$\widetilde{\mathbf{P}} \perp \mathbf{P} \Leftrightarrow \widetilde{\mathbf{P}}\left\{\sum_{n=1}^{\infty}[1 - \mathbf{E}(\sqrt{\alpha_n}|\mathscr{F}_{n-1})] = \infty\right\} = 1. \tag{17}$$

证明 因为

$$\widetilde{\mathbf{P}}_n\{z_n = 0\} = \int_{\{z_n=0\}} z_n \mathrm{d}\mathbf{P} = 0,$$

所以 ($\widetilde{\mathbf{P}}$ – a.c.)

$$z_n = \prod_{k=1}^{n} \alpha_k = \exp\left\{\sum_{k=1}^{n} \ln \alpha_k\right\}. \tag{18}$$

在 (3) 式中设 $A = \{z_\infty = 0\}$, 得 $\widetilde{\mathbf{P}}\{z_\infty = 0\} = 0$. 因此, 由 (18) 式, ($\mathbf{P}$ – a.c.) 有

$$\begin{aligned}\{z_\infty < \infty\} &= \{0 < z_\infty < \infty\} = \{0 < \lim z_n < \infty\} \\ &= \left\{-\infty < \lim \sum_{k=1}^{n} \ln \alpha_k < \infty\right\}.\end{aligned} \tag{19}$$

引进函数

$$u(x) = \begin{cases} x, & |x| \leqslant 1, \\ \operatorname{sign} x, & |x| > 1. \end{cases}$$

那么,

$$\left\{-\infty < \lim \sum_{k=1}^{n} \ln \alpha_k < \infty\right\} = \left\{-\infty < \lim \sum_{k=1}^{n} u(\ln \alpha_k) < \infty\right\}. \tag{20}$$

以 $\widetilde{\mathbf{E}}$ 表示对测度 $\widetilde{\mathbf{P}}$ 的数学期望, 而 η 是 $\mathscr{F}_n$– 可测可积随机变量. 由条件数学期望的性质知 (练习题 4),

$$z_{n-1}\widetilde{\mathbf{E}}(\eta|\mathscr{F}_{n-1}) = \mathbf{E}(\eta z_n|\mathscr{F}_{n-1}) (\mathbf{P} - \text{ 和 } \widetilde{\mathbf{P}} - \text{a.c.}), \tag{21}$$

$$\widetilde{\mathbf{E}}(\eta|\mathscr{F}_{n-1}) = z_{n-1}^{\oplus}\mathbf{E}(\eta z_n|\mathscr{F}_{n-1}) \ (\widetilde{\mathbf{P}} - \text{a.c.}). \tag{22}$$

注意到, $\alpha_n = z_{n-1}^{\oplus} z_n$, 由 (22) 式得 "条件数学期望的换算公式" 的如下重要变式 (第二章 §7 的 (44) 式):

$$\widetilde{\mathbf{E}}(\eta|\mathscr{F}_{n-1}) = \mathbf{E}(\alpha_n \eta|\mathscr{F}_{n-1}) \ \ (\widetilde{\mathbf{P}} - \text{a.c.}). \tag{23}$$

特别, 由 (23) 式可得

$$\mathbf{E}(\alpha_n|\mathscr{F}_{n-1}) = 1 \ \ (\widetilde{\mathbf{P}} - \text{a.c.}). \tag{24}$$

由 (23) 式, 有

$$\widetilde{\mathbf{E}}[u(\ln \alpha_n)|\mathscr{F}_{n-1}] = \mathbf{E}[\alpha_n u(\ln \alpha_n)|\mathscr{F}_{n-1}] \ \ (\widetilde{\mathbf{P}} - \text{a.c.}).$$

因为对于一切 $x \geqslant 0$, 有 $xu(\ln x) \geqslant x - 1$, 故由 (24) 式, 有

$$\widetilde{\mathbf{E}}[u(\ln \alpha_n)|\mathscr{F}_{n-1}] \geqslant 0 \ (\widetilde{\mathbf{P}} - \text{a.c.}).$$

由此可见, 随机序列 $X = (X_n, \mathscr{F}_n)$, 其中

$$X_n = \sum_{k=1}^{n} u(\ln \alpha_k),$$

关于测度 $\widetilde{\mathbf{P}}$ 是下鞅, 且 $|\Delta X_n| = |u(\ln \alpha_k)| \leqslant 1$.

那么, 根据 §5 的定理 5 ($\widetilde{\mathbf{P}}$ – a.c.) 有

$$\begin{aligned}&\left\{-\infty < \lim \sum_{k=1}^{n} u(\ln \alpha_k) < \infty\right\} \\ &= \left\{\sum_{k=1}^{\infty} \widetilde{\mathbf{E}}[u(\ln \alpha_k) + u^2(\ln \alpha_k)|\mathscr{F}_{k-1}] < \infty\right\}.\end{aligned} \tag{25}$$

这样, 由 (19), (20), (22) 和 (25) 式, ($\widetilde{\mathbf{P}}$ − a.c.) 有

$$\{z_\infty<\infty\}=\left\{\sum_{k=1}^{\infty}\widetilde{\mathbf{E}}[u(\ln\alpha_k)+u^2(\ln\alpha_k)|\mathscr{F}_{k-1}]<\infty\right\}$$
$$=\left\{\sum_{k=1}^{\infty}\mathbf{E}[\alpha_k u(\ln\alpha_k)+\alpha_k u^2(\ln\alpha_k)|\mathscr{F}_{k-1}]<\infty\right\},$$

从而, 由定理 2, 有

$$\widetilde{\mathbf{P}}\ll\mathbf{P}\Leftrightarrow\widetilde{\mathbf{P}}\left\{\sum_{k=1}^{\infty}\mathbf{E}[\alpha_k u(\ln\alpha_k)+\alpha_k u^2(\ln\alpha_k)|\mathscr{F}_{k-1}]<\infty\right\}=1, \tag{26}$$

$$\widetilde{\mathbf{P}}\perp\mathbf{P}\Leftrightarrow\widetilde{\mathbf{P}}\left\{\sum_{k=1}^{\infty}\mathbf{E}[\alpha_k u(\ln\alpha_k)+\alpha_k u^2(\ln\alpha_k)|\mathscr{F}_{k-1}]=\infty\right\}=1. \tag{27}$$

现在注意到, 由 (24) 式, 有

$$\widetilde{\mathbf{E}}[(1-\sqrt{\alpha_k})^2|\mathscr{F}_{k-1}]=2\widetilde{\mathbf{E}}[1-\sqrt{\alpha_k}|\mathscr{F}_{k-1}]\ (\widetilde{\mathbf{P}}-\text{a.c.}),$$

且对于一切 $x\geqslant 0$, 存在常数 A 和 $B(0<A<B<\infty)$, 使

$$A(1-\sqrt{x})^2\leqslant xu(\ln x)+xu^2(\ln x)+1-x\leqslant B(1-\sqrt{x})^2. \tag{28}$$

于是, 由 (26), (27) 和 (24), (28) 式, 得结论 (16) 和 (17)式. □

系 1 假如对于 $\widetilde{\mathbf{P}}\overset{\text{loc}}{\ll}\mathbf{P}$, 及任意 $n\geqslant 1$, σ− 代数 $\sigma(\alpha_n)$ 和 $\mathscr{F}_{n-1}$ (关于测度 $\mathbf{P}$ (或 $\widetilde{\mathbf{P}}$)) 独立, 则 $\widetilde{\mathbf{P}}\ll\mathbf{P}$ 或 $\widetilde{\mathbf{P}}\perp\mathbf{P}$ 两种情形必有一个并且只有一个成立. 这时

$$\widetilde{\mathbf{P}}\ll\mathbf{P}\Leftrightarrow\sum_{m=1}^{\infty}[1-\widetilde{\mathbf{E}}\sqrt{\alpha_m}]<\infty,$$
$$\widetilde{\mathbf{P}}\perp\mathbf{P}\Leftrightarrow\sum_{m=1}^{\infty}[1-\widetilde{\mathbf{E}}\sqrt{\alpha_m}]=\infty.$$

特别, 对于角谷的情形 (见定理 3), $\alpha_n=q_n$ 而

$$\widetilde{\mathbf{P}}\ll\mathbf{P}\Leftrightarrow\sum_{m=1}^{\infty}[1-\mathbf{E}\sqrt{q_m(x_m)}]<\infty.$$
$$\mathbf{P}\perp\mathbf{P}\Leftrightarrow\sum_{m=1}^{\infty}[1-\mathbf{E}\sqrt{q_m(x_m)}]=\infty.$$

系 2 设 $\widetilde{\mathbf{P}}\overset{\text{loc}}{\ll}\mathbf{P}$, 则

$$\widetilde{\mathbf{P}}\left\{\sum_{m=1}^{\infty}\mathbf{E}(\alpha_m\ln\alpha_m|\mathscr{F}_{m-1})<\infty\right\}=1\Rightarrow\widetilde{\mathbf{P}}\ll\mathbf{P}.$$

为证明, 只需注意到, 对于任意 $x\geqslant 0$, 有

$$x\ln x+\frac{3}{2}(1-x)\geqslant 1-x^{1/2}, \tag{29}$$

并利用 (16) 和 (17) 式.

系 3 因为 ($\widetilde{\mathbf{P}}$ − a.c.) 非负项级数

$$\sum_{n=1}^{\infty}[1-\mathbf{E}(\sqrt{\alpha_n}|\mathscr{F}_{n-1})]\quad 与\quad\sum_{n=1}^{\infty}|\ln\mathbf{E}(\sqrt{\alpha_n}|\mathscr{F}_{n-1})|$$

同时收敛或发散, 则定理 4 的论断 (16) 和 (17) 式可以写成如下的形式:

$$\widetilde{\mathbf{P}}\ll\mathbf{P}\Leftrightarrow\widetilde{\mathbf{P}}\left\{\sum_{n=1}^{\infty}|\ln\mathbf{E}(\sqrt{\alpha_n}|\mathscr{F}_{n-1})|<\infty\right\}=1, \tag{30}$$

$$\widetilde{\mathbf{P}}\perp\mathbf{P}\Leftrightarrow\widetilde{\mathbf{P}}\left\{\sum_{n=1}^{\infty}|\ln\mathbf{E}(\sqrt{\alpha_n}|\mathscr{F}_{n-1})|=\infty\right\}=1. \tag{31}$$

系 4 设存在常数 $0\leqslant A<1$ 和 $B\geqslant 0$, 使

$$\mathbf{P}\{1-A\leqslant\alpha_n\leqslant 1+B\}=1,\quad n\geqslant 1.$$

那么, 如果 $\widetilde{\mathbf{P}}\overset{\text{loc}}{\ll}\mathbf{P}$, 则

$$\widetilde{\mathbf{P}}\ll\mathbf{P}\Leftrightarrow\widetilde{\mathbf{P}}\left\{\sum_{n=1}^{\infty}\mathbf{E}[(1-\alpha_n)^2|\mathscr{F}_{n-1}]<\infty\right\}=1,$$
$$\widetilde{\mathbf{P}}\perp\mathbf{P}\Leftrightarrow\widetilde{\mathbf{P}}\left\{\sum_{n=1}^{\infty}\mathbf{E}[(1-\alpha_n)^2|\mathscr{F}_{n-1}]=\infty\right\}=1.$$

为证明, 只需注意到, 对于 $x\in[1-A,1+B]$, $0\leqslant A<1$ 和 $B\geqslant 0$, 存在常数 c 和 $C(0<c<C<\infty)$, 使

$$c(1-x)^2\leqslant(1-\sqrt{x})^2\leqslant C(1-x)^2. \tag{32}$$

4. 哈伊克 – 费里德曼择一性 假设 $\xi=(\xi_1,\xi_2,\cdots)$ 和 $\widetilde{\xi}=(\widetilde{\xi}_1,\widetilde{\xi}_2,\cdots)$ 是两个高斯随机变量序列, 且 (在第 2 小节的记号下) $\widetilde{P}_n\sim P_n, n\geqslant 1$. 我们说明, 对于这样的序列, 如何由上面得到的 "可预测" 准则, 得出 "哈伊克 – 费里德曼 (J. Hajek–H. M. Friedman) 择一性": $\widetilde{P}\sim P$ 或 $\widetilde{P}\perp P$ 二者必居其一.

根据正态相关定理 (第二章 §13 定理 2), 条件数学期望 $\mathbf{E}(x_n|\mathscr{B}_{n-1})$ 和 $\widetilde{\mathbf{E}}(x_n|\mathscr{B}_{n-1})$ 是 $x_1,\cdots,x_{n-1}$ 的线性函数, 其中 $\mathbf{E}$ 和 $\widetilde{\mathbf{E}}$ 相应为对测度 P 和 $\widetilde{P}$ 求数学期望. 以 $a_{n-1}(x)$ 和 $\widetilde{a}_{n-1}(x)$ 分别表示这些 (线性) 函数 ($a_0(x)=a_0$ 和 $\widetilde{a}_0(x)=\widetilde{a}_0$ 是常数). 记

$$b_{n-1}=(\mathbf{E}[x_n-a_{n-1}(x)]^2)^{1/2},$$
$$\widetilde{b}_{n-1}=(\widetilde{\mathbf{E}}[x_n-\widetilde{a}_{n-1}(x)]^2)^{1/2}.$$

根据上面提到的正态相关定理, 存在均值为 0、方差为 1 的, 高斯随机变量序列 $\varepsilon=(\varepsilon_1,\varepsilon_2,\cdots)$ 和 $\widetilde{\varepsilon}=(\widetilde{\varepsilon}_1,\widetilde{\varepsilon}_2,\cdots)$, 使得 ($\mathbf{P}$ − a.c.)

$$\begin{aligned}\xi_n&=a_{n-1}(\xi)+b_{n-1}\varepsilon_n,\\ \widetilde{\xi}_n&=\widetilde{a}_{n-1}(\widetilde{\xi})+\widetilde{b}_{n-1}\widetilde{\varepsilon}_n.\end{aligned}\tag{33}$$

注意, 在 $b_{n-1}=0(\widetilde{b}_{n-1}=0)$ 情形下, 为构造随机变量 $\varepsilon_n(\widetilde{\varepsilon}_n)$, 一般不得不扩充概率空间. 不过, 假如 $b_{n-1}=0$, 则随机向量 $(x_1,\cdots,x_n)$ 的分布 (P − a.c.) 集中在线性流形 $x_n=a_{n-1}(x)$ 上. 因为根据假设 $\widetilde{P}_n\sim P_n$, 所以 $\widetilde{b}_{n-1}=0, a_{n-1}(x)=\widetilde{a}_{n-1}(x)$ 和 $\alpha_n(x)=1$ (P− 及 $\widetilde{P}$ − a.c.). 因此, 不失普遍性可以认为, 对于一切 $n\geqslant 1$, 有 $b_n^2>0,\widetilde{b}_n^2>0$, 因为在相反的情形下, 和式

$$\sum_{n=1}^{\infty}[1-\mathbf{E}(\sqrt{\alpha_n}|\mathscr{B}_{n-1})]$$

中相应项的 "贡献" 等于 0 (见 (16) 式和 (17) 式).

利用高斯性假设, 由 (33) 式可见, 对于 $n\geqslant 1$, 有

$$\alpha_n=d_{n-1}^{-1}\exp\left\{-\frac{[x_n-a_{n-1}(x)]^2}{2b_{n-1}^2}+\frac{[x_n-\widetilde{a}_{n-1}(x)]^2}{2\widetilde{b}_{n-1}^2}\right\},\tag{34}$$

其中 $d_n=|b_n/\widetilde{b}_n|$, 而

$$\begin{aligned}a_0=\mathbf{E}\xi_1,\quad \widetilde{a}_0=\mathbf{E}\widetilde{\xi}_1,\\ b_0^2=\mathbf{D}\xi_1,\quad \widetilde{b}_0^2=\mathbf{D}\widetilde{\xi}_1.\end{aligned}$$

由 (34), 有

$$\ln\mathbf{E}(\alpha_n^{1/2}|\mathscr{B}_{n-1})=\frac{1}{2}\ln\frac{2d_{n-1}}{1+d_{n-1}^2}-\frac{d_{n-1}^2}{1+d_{n-1}^2}\left(\frac{a_{n-1}(x)-\widetilde{a}_{n-1}(x)}{b_{n-1}}\right)^2.$$

由于

$$\ln\frac{2d_{n-1}}{1+d_{n-1}^2}\leqslant 0$$

则 (30) 式有如下形式

$$\widetilde{P}\ll P\Leftrightarrow\widetilde{P}\left\{\sum_{n=1}^{\infty}\left[\frac{1}{2}\ln\frac{1+d_{n-1}^2}{2d_{n-1}}+\frac{d_{n-1}^2}{1+d_{n-1}^2}\left(\frac{a_{n-1}(x)-\widetilde{a}_{n-1}(x)}{b_{n-1}}\right)^2\right]<\infty\right\}=1.\tag{35}$$

由于级数

$$\sum_{n=1}^{\infty}\ln\frac{1+d_{n-1}^2}{2d_{n-1}}\quad 和\quad\sum_{n=1}^{\infty}(d_{n-1}^2-1)^2$$

或者同时收敛, 或者同时发散, 则由 (35) 式可见

$$\widetilde{P}\ll P\Leftrightarrow\widetilde{P}\left\{\sum_{n=1}^{\infty}\left[\left(\frac{\Delta_n(x)}{b_n}\right)^2+\left(\frac{\widetilde{b}_n^2}{b_n^2}-1\right)^2\right]<\infty\right\}=1,\tag{36}$$

其中 $\Delta_n(x)=a_n(x)-\widetilde{a}_n(x)$.

因为 $a_n(x)$ 和 $\widetilde{a}_n(x)$ 是线性函数, 所以随机变量序列 $\{\Delta_n(x)/b_n\}_{n\geqslant 0}$ (既关于测度 $\widetilde{P}$, 也关于测度 P) 构成高斯系统. 由下面将要证明的引理, 可见

$$\widetilde{P}\left\{\sum\left(\frac{\Delta_n(x)}{b_n}\right)^2<\infty\right\}=1\Leftrightarrow\sum\widetilde{\mathbf{E}}\left(\frac{\Delta_n(x)}{b_n}\right)^2<\infty.\tag{37}$$

从而, 由 (36) 式, 可见

$$\widetilde{P}\ll P\Leftrightarrow\sum_{n=0}^{\infty}\left[\widetilde{\mathbf{E}}\left(\frac{\Delta_n(x)}{b_n}\right)^2+\left(\frac{\widetilde{b}_n^2}{b_n^2}-1\right)^2\right]<\infty.$$

类似地, 可得

$$\begin{aligned}\widetilde{P}\perp P&\Leftrightarrow\widetilde{P}\left\{\sum_{n=1}^{\infty}\left[\left(\frac{\Delta_n(x)}{b_n}\right)^2+\left(\frac{\widetilde{b}_n^2}{b_n^2}-1\right)^2\right]<\infty\right\}=0\\ &\Leftrightarrow\sum_{n=0}^{\infty}\left[\widetilde{\mathbf{E}}\left(\frac{\Delta_n(x)}{b_n}\right)^2+\left(\frac{\widetilde{b}_n^2}{b_n^2}-1\right)^2\right]=\infty.\end{aligned}$$

由此可见, 如果测度 P 与 $\widetilde{P}$ 非奇异, 则 $\widetilde{P}\ll P$. 因为根据假设 $\widetilde{P}_n\sim P, n\geqslant 1$, 所以由对称性知 $P\ll\widetilde{P}$. 于是有下面的定理.

定理 5 (哈伊克 − 费里德曼择一性) 设 $\xi=(\xi_1,\xi_2,\cdots)$ 和 $\widetilde{\xi}=(\widetilde{\xi}_1,\widetilde{\xi}_2,\cdots)$ 是两个高斯随机变量序列, 且其有限维分布等价: $\widetilde{P}_n\sim P_n, n\geqslant 1$. 那么, $\widetilde{P}\sim P$ 或 $\widetilde{P}\perp P$ 二者必居其一. 这时

$$\begin{aligned}\widetilde{P}\sim P&\Leftrightarrow\sum_{n=0}^{\infty}\left[\widetilde{\mathbf{E}}\left(\frac{\Delta_n(x)}{b_n}\right)^2+\left(\frac{\widetilde{b}_n^2}{b_n^2}-1\right)^2\right]<\infty,\\ \widetilde{P}\perp P&\Leftrightarrow\sum_{n=0}^{\infty}\left[\widetilde{\mathbf{E}}\left(\frac{\Delta_n(x)}{b_n}\right)^2+\left(\frac{\widetilde{b}_n^2}{b_n^2}-1\right)^2\right]=\infty.\end{aligned}\tag{38}$$

引理 设 $\beta=(\beta_n)_{n\geqslant 1}$ 是定义在 $(\Omega,\mathscr{F},\mathbf{P})$ 上的高斯序列. 那么,

$$\mathbf{P}\left\{\sum_{n=1}^{\infty}\beta_n^2<\infty\right\}>0\Leftrightarrow\mathbf{P}\left\{\sum_{n=1}^{\infty}\beta_n^2<\infty\right\}=1\Leftrightarrow\sum_{n=1}^{\infty}\mathbf{E}\beta_n^2<\infty.\tag{39}$$

证明 蕴涵关系 (⇐) 显然. 现在假设 $\mathbf{E}\beta_n^2=0, n\geqslant 1$, 并证明蕴涵关系 (⇒). 为此, 只需证明

$$\mathbf{E}\sum_{n=1}^{\infty}\beta_n^2\leqslant\left[\mathbf{E}\exp\left(-\sum_{n=1}^{\infty}\beta_n^2\right)\right]^{-2}, \tag{40}$$

因为, 由 $\mathbf{P}\{\sum\beta_n^2<\infty\}>0$ 条件可见, (40) 式的右侧小于 ∞. 即 $\sum\limits_{n=1}^{\infty}\mathbf{E}\beta_n^2<\infty$, 而由已证明的蕴涵关系, 可见

$$\mathbf{P}\left\{\sum_{n=1}^{\infty}\beta_n^2<\infty\right\}=1.$$

固定某一 $n\geqslant 1$, 则由第二章 §11 和 §13 知, 存在独立高斯随机变量 $\beta_{k,n}(k=1,\cdots,r\leqslant n)$, 且 $\mathbf{E}\beta_{k,n}=0$, 使得

$$\sum_{k=1}^{n}\beta_k^2=\sum_{n=1}^{r}\beta_{k,n}^2.$$

如果记 $\mathbf{E}\beta_{k,n}^2=\lambda_{k,n}$, 则易见

$$\mathbf{E}\sum_{k=1}^{r}\beta_{k,n}^2=\sum_{k=1}^{r}\lambda_{k,n} \tag{41}$$

而

$$\mathbf{E}\exp\left(-\sum_{k=1}^{r}\beta_{k,n}^2\right)=\prod_{k=1}^{r}(1+2\lambda_{k,n})^{-1/2}. \tag{42}$$

比较 (41) 和 (42) 式右侧, 得

$$\mathbf{E}\sum_{k=1}^{n}\beta_k^2=\mathbf{E}\sum_{k=1}^{r}\beta_{k,n}^2\leqslant\left[\mathbf{E}\exp\left(-\sum_{k=1}^{r}\beta_{k,n}^2\right)\right]^{-2}=\left[\mathbf{E}\exp\left(-\sum_{k=1}^{n}\beta_k^2\right)\right]^{-2},$$

由此, 当 $n\to\infty$ 时经极限过渡, 得所要证明得不等式 (40).

现在假设 $\mathbf{E}\beta_n\neq 0$.

考虑与 $\beta=(\beta_n)_{n\geqslant 1}$ 独立并且具有同一分布的序列 $\widetilde{\beta}=(\widetilde{\beta}_n)_{n\geqslant 1}$ (必要时扩展原概率空间). 那么, 如果

$$\mathbf{P}\left\{\sum_{n=1}^{\infty}\beta_n<\infty\right\}>0,\quad 则\quad \mathbf{P}\left\{\sum_{n=1}^{\infty}(\beta_n-\widetilde{\beta}_n)^2<\infty\right\}>0,$$

且由已经证明的, 有

$$2\sum_{n=1}^{\infty}\mathbf{E}(\beta_n-\mathbf{E}\beta_n)^2=\sum_{n=1}^{\infty}\mathbf{E}(\beta_n-\widetilde{\beta}_n)^2<\infty.$$

因为

$$(\mathbf{E}\beta_n)^2\leqslant 2\beta_n^2+2(\beta_n-\mathbf{E}\beta_n)^2,\quad 则\quad \sum_{n=1}^{\infty}(\mathbf{E}\beta_n)^2<\infty.$$

于是,

$$\sum_{n=1}^{\infty}\mathbf{E}\beta_n^2=\sum_{n=1}^{\infty}(\mathbf{E}\beta_n)^2=\sum_{n=1}^{\infty}\mathbf{E}(\beta_n-\mathbf{E}\beta_n)^2<\infty. \qquad \square$$

5. 例 我们继续讨论 §5 第 3 小节的例. 假设 $\xi_0,\xi_1,\cdots$ 是独立高斯随机变量序列, $\mathbf{E}\xi_i=0, \mathbf{D}\xi_i=V_i>0$.

仍设

$$X_{n+1}=\theta X_n+\xi_{n+1},\quad n\geqslant 0,$$

其中 $X_0=\xi_0$, 而 $\theta(-\infty<\theta<\infty)$ 是待估计未知参数. 设 $\widehat{\theta}_n$ 是用最小二乘法得到的估计.

定理 6 *序列 $\widehat{\theta}_n(n\geqslant 1)$ 是强相合估计的, 充分和必要条件是:*

$$\sum_{n=0}^{\infty}\frac{V_n}{1+V_n}=\infty. \tag{43}$$

证明 *充分性.* 当未知参数的值为 θ 时, 以 $\mathbf{P}_\theta$ 表示 $(\mathbb{R}^\infty,\mathscr{B}^\infty)$ 上对应于序列 $(X_0,X_2,\cdots)$ 的概率分布; 以 $\mathbf{E}_\theta$ 表示对测度 $\mathbf{P}_\theta$ 的数学期望.

我们已经知道,

$$\widehat{\theta}_n=\theta+\frac{M_n}{\langle M\rangle_n},$$

其中

$$M_n=\sum_{k=0}^{n-1}\frac{X_k\xi_{k+1}}{V_{k+1}},\quad \langle M\rangle_n=\sum_{k=0}^{n-1}\frac{X_k^2}{V_{k+1}}.$$

根据上一小节的引理

$$\mathbf{P}_\theta\{\langle M\rangle_\infty=\infty\}=1\Leftrightarrow\mathbf{E}_\theta\langle M\rangle_\infty=\infty.$$

从而 $\langle M\rangle_\infty=\infty$ ($\mathbf{P}_\theta$ – a.c.), 当且仅当

$$\sum_{k=1}^{\infty}\frac{\mathbf{E}_\theta X_k^2}{V_{k+1}}=\infty. \tag{44}$$

而且

$$\mathbf{E}_\theta X_k^2=\sum_{i=0}^{k}\theta^{2i}V_{k-i},$$

$$\begin{aligned}\sum_{k=0}^{\infty}\frac{\mathbf{E}_\theta X_k^2}{V_{k+1}}&=\sum_{k=0}^{\infty}\frac{1}{V_{k+1}}\left(\sum_{i=0}^{k}\theta^{2i}V_{k-i}\right)=\sum_{i=k}^{\infty}\theta^{2k}\sum_{i=k}^{\infty}\frac{V_{i-k}}{V_{i+1}}\\&=\sum_{i=0}^{\infty}\frac{V_i}{V_{i+1}}+\sum_{k=1}^{\infty}\theta^{2k}\left(\sum_{i=k}^{\infty}\frac{V_{i-k}}{V_{i+1}}\right).\end{aligned} \tag{45}$$

因此由 (43) 式得 (44) 式, 而根据定理 4, 对于每一个 θ, 估计 $\widehat{\theta}_n (n \geqslant 1)$ 序列为强相合的.

必要性 假设对于一切 $\theta, \mathbf{P}_\theta(\widehat{\theta}_n \to \theta) = 1$. 现在证明, 假如 $\theta_1 \neq \theta_2$, 则 $\mathbf{P}_{\theta_1}$ 和 $\mathbf{P}_{\theta_2}$ 是奇异的 $(\mathbf{P}_{\theta_1} \perp \mathbf{P}_{\theta_2})$. 事实上, 由于 $(X_0, X_1, \cdots)$ 是高斯序列, 则根据定理 5 测度 $\mathbf{P}_{\theta_1}$ 和 $\mathbf{P}_{\theta_2}$ 要么奇异要么等价. 然而测度 $\mathbf{P}_{\theta_1}$ 和 $\mathbf{P}_{\theta_2}$ 不可能等价, 因为, 假如 $\mathbf{P}_{\theta_1} \sim \mathbf{P}_{\theta_2}$, 但是由于 $\mathbf{P}_{\theta_1}(\widehat{\theta}_n \to \theta_1) \to 1$, 则 $\mathbf{P}_{\theta_2}(\widehat{\theta}_n \to \theta_1) = 1$. 而由于 $\mathbf{P}_{\theta_2}(\widehat{\theta}_n \to \theta_2) = 1$, 故应该 $\theta_1 = \theta_2$, 而这与假设 $\theta_1 \neq \theta_2$ 矛盾.

于是, 对于 $\theta_1 \neq \theta_2, \mathbf{P}_{\theta_1} \perp \mathbf{P}_{\theta_2}$.

根据 (38) 式, 对于 $\theta_1 \neq \theta_2$, 有

$$\mathbf{P}_{\theta_1} \perp \mathbf{P}_{\theta_2} \Leftrightarrow (\theta_1 - \theta_2)^2 \sum_{k=0}^{\infty} \mathbf{E}_{\theta_1} \left(\frac{X_k^2}{V_{k+1}} \right) = \infty.$$

设 $\theta_1 = 0$ 和 $\theta_2 \neq 0$, 则由 (45) 式得

$$\mathbf{P}_0 \perp \mathbf{P}_{\theta_2} \Leftrightarrow \sum_{i=0}^{\infty} \frac{V_i}{V_{i+1}} = \infty.$$

于是, 这就证明了必要性. □

6. 练习题

1. 证明等式 (6).

2. 设 $\widetilde{\mathbf{P}}_n \sim \mathbf{P}_n, n \geqslant 1$. 证明

$$\widetilde{\mathbf{P}} \sim \mathbf{P} \Leftrightarrow \widetilde{\mathbf{P}}\{z_\infty < \infty\} = \mathbf{P}\{z_\infty > 0\} = 1,$$
$$\widetilde{\mathbf{P}} \perp \mathbf{P} \Leftrightarrow \widetilde{\mathbf{P}}\{z_\infty = \infty\} = 1 \text{ 或 } \mathbf{P}\{z_\infty = 0\} = 1.$$

3. 设 $\widetilde{\mathbf{P}}_n \ll \mathbf{P}_n, n \geqslant 1$, 而 τ (关于 $(\mathscr{F}_n)$) 是停时, $\widetilde{\mathbf{P}}_\tau = \widetilde{\mathbf{P}}|\mathscr{F}_\tau$ 和 $\mathbf{P}_\tau = \mathbf{P}|\mathscr{F}_\tau$ 是测度 $\widetilde{\mathbf{P}}$ 和 $\mathbf{P}$ 到 $\sigma-$ 代数 $\mathscr{F}_\tau$ 的收缩. 证明 $\widetilde{\mathbf{P}}_\tau \ll \mathbf{P}_\tau$, 当且仅当 $\{\tau = \infty\} = \{z_\infty < \infty\}(\mathbf{P}-\text{a.c.})$. 特别, 如果 $\mathbf{P}\{\tau < \infty\} = 1$, 则 $\widetilde{\mathbf{P}}_\tau \ll \mathbf{P}_\tau$.

4. 证明 "换算公式" (21) 和 (22).

5. 证明不等式 (28), (29), (32).

6. 证明公式 (34).

7. 假设在第 2 小节中, 序列 $\xi = (\xi_1, \xi_2, \cdots)$ 和 $\widetilde{\xi} = (\widetilde{\xi}_1, \widetilde{\xi}_2, \cdots)$ 由独立同分布随机变量构成. 证明, 1) 如果 $\mathbf{P}_{\widetilde{\xi}_1} \ll \mathbf{P}_{\xi_1}$, 则 $\widetilde{\mathbf{P}} \ll \mathbf{P}$, 当且仅当测度 $\mathbf{P}_{\widetilde{\xi}_1} = \mathbf{P}_{\xi_1}$; 2) 如果 $\mathbf{P}_{\widetilde{\xi}_1} \ll \mathbf{P}_{\xi_1}$ 且 $\mathbf{P}_{\widetilde{\xi}_1} \neq \mathbf{P}_{\xi_1}$, 则 $\widetilde{\mathbf{P}} \perp \mathbf{P}$.

§7. 随机游动越出曲线边界的概率的渐近式

1. 概率 $\mathbf{P}\{\tau > n\}$ 的渐近式 设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量序列, $S_n =$

$\xi_1 + \cdots + \xi_n$; $g = g(n)$, $n \geqslant 1, g(1) < 0$, 是某一"边界"; 而

$$\tau = \inf\{n \geqslant 1 : S_n < g(n)\}$$

是随机游动 $(S_n)_{n \geqslant 1}$ 首次处于边界 $g = g(n)$ 以下的时间. (像通常一样, 如果 $\{\cdot\} = \varnothing$, 则认为 $\tau = \infty$). 求时间 τ 的分布是相当困难的. 在这一节, 对于广泛的一类边界 $g = g(n)$, 并且 ξ_i 在服从正态分布的条件下, 求当 $n \to \infty$ 时概率 $\mathbf{P}\{\tau > n\}$ 的渐近式. 所用方法基于 "测度绝对连续替换" 的思想, 并且利用前面介绍的鞅和马尔可夫时间的一系列性质.

定理 1 *设 $\xi_1, \xi_2, \cdots$ 是独立同分布的随机变量, $\xi_i \sim N(0,1)$. 假设边界 $g = g(n)$ 满足条件: $g(1) < 0$, 而对于 $n \geqslant 2$,*

$$0 \leqslant \Delta g(n+1) \leqslant \Delta g(n), \tag{1}$$

其中 $\Delta g(n) = g(n) - g(n-1)$, 并且

$$\ln n = o\left(\sum_{k=2}^{n} [\Delta g(k)]^2\right), \quad n \to \infty. \tag{2}$$

那么,

$$\mathbf{P}\{\tau > n\} = \exp\left\{-\frac{1}{2}\sum_{k=2}^{n} [\Delta g(k)]^2 (1 + o(1))\right\}, \quad n \to \infty. \tag{3}$$

在进行证明之前, 我们首先指出, 条件 (1) 和 (2) 成立, 例如, 若

$$g(n) = an^\nu + b, \quad \frac{1}{2} < \nu \leqslant 1, \quad a + b < 0, \quad a > 0,$$

或者 (对于充分大的 n)

$$g(n) = an^\nu L(n), \quad \frac{1}{2} \leqslant \nu \leqslant 1,$$

其中 $L(n)$ 是某一缓慢变化的函数 (例如, $L(n) = C(\ln n)^\beta$, 对于 $1/2 < \nu < 1$, β 是任意的, 而对于 $\nu = 1/2, C > 0, \beta > 0$).

2. 定理 1 的证明 在证明定理 1 时用到下面两个辅助命题.

假设 $\xi_1, \xi_2, \cdots$ 是独立同分布的随机变量序列, $\xi_i \sim N(0,1)$. 引进记号: $\mathscr{F}_0 = \{\varnothing, \Omega\}, \mathscr{F}_n = \sigma\{\xi_1, \cdots, \xi_n\}$, 而设 $\alpha = (\alpha_n, \mathscr{F}_{n-1})$ 是可预测序列, 且 $\mathbf{P}\{|\alpha_n| \leqslant C\} = 1, n \geqslant 1$, 其中 C 是某一常数. 建立序列 $z = (z_n, \mathscr{F}_n)$, 其中

$$z_n = \exp\left\{\sum_{k=1}^{n} \alpha_k \xi_k - \frac{1}{2}\sum_{k=1}^{n} \alpha_k^2\right\}, \quad n \geqslant 1. \tag{4}$$

不难验证, 序列 (关于测度 $\mathbf{P}$) $z = (z_n, \mathscr{F}_n)$ 是鞅, 且 $\mathbf{E}z_n = 1, n \geqslant 1$.

固定某个 $n \geqslant 1$, 并且在可测空间上 $(\Omega, \mathscr{F}_n)$引进概率测度 $\widetilde{\mathbf{P}}_n$:

$$\widetilde{\mathbf{P}}_n(A) = \mathbf{E}I(A)z_n, \quad A \in \mathscr{F}_n. \tag{5}$$

引理 1 (“士尔萨诺夫 (И. Гирсанов) 定理” 的离散变式) *随机变量 $\widetilde{\xi}_k = \xi_k - \alpha_k, 1 \leqslant k \leqslant n$, 关于测度 $\widetilde{\mathbf{P}}_n$, 独立且服从正态分布 $\xi_k \sim N(0,1)$.*

证明 设 $\widetilde{\mathbf{E}}_n$ 表示对测度 $\widetilde{\mathbf{P}}_n$ 的数学期望, 则对于 $\lambda_k \in \mathbb{R}, 1 \leqslant k \leqslant n$, 有

$$\begin{aligned}\widetilde{\mathbf{E}}_n \exp\left\{i\sum_{k=1}^n \lambda_k\widetilde{\xi}_k\right\} &= \mathbf{E}\exp\left\{i\sum_{k=1}^n \lambda_k\widetilde{\xi}_k\right\} z_n \\ &= \mathbf{E}\left[\exp\left\{i\sum_{k=1}^n \lambda_k\widetilde{\xi}_k\right\} z_{n-1}\mathbf{E}\left\{\exp\left(i\lambda_n(\xi_n-\alpha_n)+\alpha_n\xi_n-\frac{\alpha_n^2}{2}\right)\Big|\mathscr{F}_{n-1}\right\}\right] \\ &= \mathbf{E}\left[\exp\left\{i\sum_{k=1}^n \lambda_k\widetilde{\xi}_k\right\} z_{n-1}\right]\exp\left\{-\frac{\lambda_n^2}{2}\right\} = \cdots = \exp\left\{-\frac{1}{2}\sum_{k=1}^n \lambda_k^2\right\}.\end{aligned}$$

于是, 由第二章 §12 定理 4 得所需要证明的结论. □

引理 2 *设 $X = (X_n, \mathscr{F}_n)_{n\geqslant 1}$ 是均值为 0 的平方可积鞅, 而*

$$\sigma = \inf\{n \geqslant 1 : X_n \leqslant -b\},$$

其中常数 $b > 0$. 假设

$$\mathbf{P}\{X_1 \leqslant -b\} > 0.$$

那么, 存在常数 $C > 0$, 使对于一切 $n \geqslant 1$,

$$\mathbf{P}\{\sigma > n\} \geqslant \frac{C}{\mathbf{E}X_n^2}. \tag{6}$$

证明 根据 §2 定理 1 的系 1, $\mathbf{E}X_{\sigma\wedge n} = 0$, 由此可见

$$-\mathbf{E}I(\sigma \leqslant n)X_\sigma = \mathbf{E}I(\sigma > n)X_n. \tag{7}$$

在集合 $\{\sigma \leqslant n\}$ 上, 有

$$-X_\sigma \geqslant b > 0.$$

因此, 对于 $n \geqslant 1$

$$-\mathbf{E}I(\sigma \leqslant n)X_\sigma \geqslant b\mathbf{P}\{\sigma \leqslant n\} \geqslant b\mathbf{P}\{\sigma = 1\} = b\mathbf{P}\{X_1 < -b\} > 0. \tag{8}$$

另一方面, 由于可西 – 布尼亚科夫斯基 (A. L. Cauchy – A.Я.Буняковский) 不等式, 可见

$$\mathbf{E}I(\sigma > n)X_n \leqslant [\mathbf{P}\{\sigma > n\}\cdot\mathbf{E}X_n^2]^{1/2}. \tag{9}$$

于是, 由 (9) 式及 (7) 和 (8) 式, 所要证明的不等式 (6), 其中 $C = [b\mathbf{P}\{X_1 < -b\}]^2$.

证明定理 1 只需证明:

$$\varliminf_n \ln\mathbf{P}\{\tau > n\}\Big/\sum_{k=2}^n[\Delta g(k)]^2 \geqslant -\frac{1}{2} \tag{10}$$

和

$$\varlimsup_n \ln\mathbf{P}\{\tau > n\}\Big/\sum_{k=2}^n[\Delta g(k)]^2 \leqslant -\frac{1}{2}. \tag{11}$$

为此, 考虑 (非随机) 序列 $(\alpha_n)_{n\geqslant 1}$, 其中

$$\alpha_1 = 0, \quad \alpha_n = \Delta g(n), \quad n \geqslant 2,$$

而概率测度 $(\widetilde{\mathbf{P}}_n)_{n\geqslant 1}$ 决定于 (5) 式. 那么, 根据赫尔德 (O. L. Hölder) 不等式, 有

$$\widetilde{\mathbf{P}}_n\{\tau > n\} = \mathbf{E}I(\tau > n)z_n \leqslant (\mathbf{P}\{\tau > n\})^{1/q}(\mathbf{E}z_n^p)^{1/p}, \tag{12}$$

其中 $p > 1$ 和 $q = p/(p-1)$.

式 (12) 中最末尾一个因子的明显表达式为:

$$(\mathbf{E}z_n^p)^{1/p} = \exp\left\{\frac{p-1}{2}\sum_{k=2}^n[\Delta g(k)]^2\right\}. \tag{13}$$

现在估计 (12) 式左侧的概率 $\widetilde{\mathbf{P}}_n\{\tau > n\}$. 有

$$\widetilde{\mathbf{P}}_n\{\tau > n\} = \widetilde{\mathbf{P}}_n\{S_k \geqslant g(k), 1 \leqslant k \leqslant n\} = \widetilde{\mathbf{P}}_n\{\widetilde{S}_k \geqslant g(1), 1 \leqslant k \leqslant n\},$$

其中 $\widetilde{S}_k = \sum_{i=1}^n \widetilde{\xi}_i$, $\widetilde{\xi}_i = \xi_i - \alpha_i$. 根据引理 1, 随机变量 $\widetilde{\xi}_1, \cdots, \widetilde{\xi}_n$, 关于测度 $\widetilde{\mathbf{P}}_n$ 独立且服从正态分布 $\widetilde{\xi}_k \sim N(0,1)$. 将引理 2 用于 $b = -g(1)$, $\mathbf{P} = \widetilde{\mathbf{P}}_n$, $X_n = \widetilde{S}_k$, 得

$$\widetilde{\mathbf{P}}_n\{\tau > n\} \geqslant \frac{C}{n}, \tag{14}$$

其中 C 是某一常数.

那么, 由 (12)~(14) 式, 可见对于任何 $p > 1$, 有

$$\mathbf{P}\{\tau > n\} \geqslant C_p\exp\left\{-\frac{p}{2}\sum_{k=2}^n[\Delta g(k)]^2 - \frac{p}{p-1}\ln n\right\}, \tag{15}$$

其中 C_p 是某一常数. 鉴于定理的条件和 $p > 1$ 的任意性, 由 (15) 式得的下侧估计 (10).

为得到 (11) 式的上侧估计, 首先注意到由于 $z_n > 0$ ($\mathbf{P}$– 且 $\widetilde{\mathbf{P}}$ – a.c.), 则由 (5) 式可见

$$\mathbf{P}\{\tau > n\} = \widetilde{\mathbf{E}}_n I(\tau > n)z_n^{-1}, \tag{16}$$

其中 $\widetilde{\mathbf{E}}_n$ 表示对测度 $\widetilde{\mathbf{P}}_n$ 的数学期望.

对于现在考虑的情形, $\alpha_1 = 0$, $\alpha_n = \Delta g(n)$, $n \geqslant 2$, 因此对于 $n \geqslant 2$,

$$z_n^{-1} = \exp\left\{-\sum_{k=2}^n \Delta g(k)\cdot\xi_k + \frac{1}{2}\sum_{k=2}^n[\Delta g(k)]^2\right\}.$$

由分部求和公式 (见第四章 §3 引理 2 的证明), 得

$$\sum_{k=2}^{n}\Delta g(k)\cdot\xi_k=\Delta g(n)\cdot S_n-\frac{1}{2}\sum_{k=2}^{n}S_{k-1}\Delta[\Delta g(k)];$$

由此并注意到, 根据定理的条件 $\Delta g(k)\geqslant 0,\Delta[\Delta g(k)]\leqslant 0$, 可见在集合 $\{\tau>n\}=\{S_k\geqslant g(k),1\leqslant k\leqslant n\}$ 上, 有

$$\sum_{k=2}^{n}\Delta g(k)\cdot\xi_k\geqslant\Delta g(n)\cdot g(n)-\sum_{k=3}^{n}g(k-1)\Delta[\Delta g(k)]-\xi_1\Delta g(2)$$
$$=\sum_{k=2}^{n}[\Delta g(k)]^2+g(1)\Delta g(2)-\xi_1\Delta g(2).$$

因而, 由 (16) 式, 有

$$\mathbf{P}\{\tau>n\}\leqslant\exp\left\{-\frac{1}{2}\sum_{k=2}^{n}[\Delta g(k)]^2-g(1)\Delta g(2)\right\}\widetilde{\mathbf{E}}_nI(\tau>n)\mathrm{e}^{-\xi_1\Delta g(2)}$$
$$=\exp\{-g(1)\Delta g(2)\}\exp\left\{-\frac{1}{2}\sum_{k=2}^{n}[\Delta g(k)]^2\right\}\widetilde{\mathbf{E}}_nI(\tau>n)\mathrm{e}^{-\xi_1\Delta g(2)},$$

其中

$$\widetilde{\mathbf{E}}_nI(\tau>n)\mathrm{e}^{-\xi_1\Delta g(2)}\leqslant\mathbf{E}z_n\mathrm{e}^{-\xi_1\Delta g(2)}=\mathbf{E}\mathrm{e}^{-\xi_1\Delta g(2)}<\infty.$$

因此,

$$\mathbf{P}\{\tau>n\}\leqslant C\exp\left\{-\frac{1}{2}\sum_{k=2}^{n}[\Delta g(k)]^2\right\},$$

其中 C 是某一正常数, 于是证明了 (11) 式的上侧估计.

3. 双侧界限的情形 "测度绝对连续替换" 的思想, 亦可对于双侧边界的情形, 研究类似的问题. 我们 (不加证明地) 给出这一方向上的结果之一.

定理 2 设 $\xi_1,\xi_2,\cdots$ 是独立同分布的随机变量序列, $\xi_i\sim N(0,1)$. 假设 $f=f(n)$ 是正值函数, 满足条件:

$$f(n)\to\infty,\quad n\to\infty,$$

且

$$\sum_{k=2}^{n}[\Delta f(k)]^2=o\left(\sum_{k=1}^{n}f^{-2}(k)\right),\quad n\to\infty$$

那么, 如果

$$\sigma=\inf\{n\geqslant 1:|S_n|\geqslant f(n)\},$$

则

$$\mathbf{P}\{\sigma>n\}=\exp\left\{-\frac{\pi^2}{8}\sum_{k=1}^{n}f^{-2}(k)[1+o(1)]\right\},\quad n\to\infty.\tag{17}$$

4. 练习题

1. 证明由 (4) 式决定的序列是鞅. 问在不要求条件 $\mathbf{P}\{|\alpha_n|\leqslant c=1,n\geqslant 1\}$ 成立的情况下, 这是否成立?

2. 证明公式 (13).

3. 证明公式 (17).

§8. 相依随机变量之和的中心极限定理

1. 函数的中心极限定理 在第三章 §4 中, 我们对于随机变量 $\xi_{1n},\cdots,\xi_{nn}$ 之和 $S_n=\xi_{1n}+\cdots+\xi_{nn}(n\geqslant 1)$, 证明了中心极限定理, 当时假设: $\xi_{1n},\cdots,\xi_{nn}$ 相互独立, 其二阶矩有限, 以及各被加项的极限可忽略性. 在这一节, 我们将不再假设满足独立性条件, 甚至也不再要求一阶绝对矩有限. 不过, 仍然假设被加项的极限可忽略性.

这样, 假设在概率空间 $(\Omega,\mathscr{F},\mathbf{P})$ 上给定随机序列

$$\xi^n=(\xi_{nk},\mathscr{F}^n_k),\quad 0\leqslant k\leqslant n,\quad n\geqslant 1,$$

其中 $\xi_{n0}=0,\mathscr{F}^n_0=\{\varnothing,\Omega\},\mathscr{F}^n_k\subseteq\mathscr{F}^n_{k+1}\subseteq\mathscr{F}(k+1\leqslant n)$. 设

$$X^n_t=\sum_{k=0}^{[nt]}\xi_{nk},\quad 0\leqslant t\leqslant 1.$$

定理 1 设对于固定的 $0<t\leqslant 1$, 满足下列条件: 对于任意 $\varepsilon\in(0,1]$, 当 $n\to\infty$ 时, 有

$$\text{(A)}\qquad\sum_{k=1}^{[nt]}\mathbf{P}\{|\xi_{nk}|>\varepsilon|\mathscr{F}^n_{k-1}\}\xrightarrow{\mathbf{P}}0,$$

$$\text{(B)}\qquad\sum_{k=1}^{[nt]}\mathbf{E}[\xi_{nk}I(|\xi_{nk}|\leqslant\varepsilon|\mathscr{F}^n_{k-1})]\xrightarrow{\mathbf{P}}0,$$

$$\text{(C)}\qquad\sum_{k=1}^{[nt]}\mathbf{D}[\xi_{nk}I(|\xi_{nk}|\leqslant\varepsilon)|\mathscr{F}^n_{k-1}]\xrightarrow{\mathbf{P}}\sigma^2_t.$$

那么,

$$X^n_t\xrightarrow{d}N(0,\sigma^2_t).$$

注 1 条件 (A) 和 (B) 保障随机变量 X^n_t 可以表示为

$$X^n_t=Y^n_t+Z^n_t,\quad 而\ Z^n_t\xrightarrow{\mathbf{P}}0,\quad Y^n_t=\sum_{k=0}^{[nt]}\eta_{nk},$$

其中序列 $\eta^n=(\eta_{nk},\mathscr{F}^n_k)$ 是鞅 – 差 $\mathbf{E}(\eta_{nk}|\mathscr{F}^n_{k-1})=0$ 其中对 $0\leqslant k\leqslant n, n\geqslant 1$ 一致 $|\eta_{nk}|\leqslant c$. 因此 (在所考虑的条件下) 定理的证明实际上归结为, 形成鞅 – 差序列的极限定理的证明.

对于随机变量 $\xi_{1n},\cdots,\xi_{nn}$ 独立的情形, 条件 (A), (B) 和 (C) 当 $t=1$ 时就是条件 $(\sigma^2=\sigma_1^2)$:

$$\text{(a)}\qquad \sum_{k=1}^{n}\mathbf{P}\{|\xi_{nk}|>\varepsilon\}\to 0,$$

$$\text{(b)}\qquad \sum_{k=1}^{n}\mathbf{E}[\xi_{nk}I(|\xi_{nk}|\leqslant\varepsilon)]\to 0,$$

$$\text{(c)}\qquad \sum_{k=1}^{n}\mathbf{D}[\xi_{nk}I(|\xi_{nk}|\leqslant\varepsilon)]\to \sigma^2.$$

这由格涅坚科 (Б. В. Гнеденко) 和柯尔莫戈洛夫的专著 [16] 是熟知的. 因而由定理 1 得如下推论.

系 如果 $\xi_{1n},\cdots,\xi_{nn}(n\geqslant 1)$ 是独立随机变量, 则

$$\text{(a), (b), (c)}\Rightarrow X_1^n=\sum_{k=1}^{n}\xi_{nk}\xrightarrow{d}N(0,\sigma^2).$$

注 2 在条件 (C) 中不排除 $\sigma_t^2=0$ 的情形. 这样, 定理 1 中也包括收敛于退化分布的情形 $(X_t^n\xrightarrow{d}0)$.

注 3 用定理 1 的方法可以表述和证明更加一般的命题.

假设 $0<t_1<t_2<\cdots<t_j\leqslant 1, \sigma_{t_1}\leqslant\sigma_{t_2}\leqslant\cdots\leqslant\sigma_{t_j}$, 而 $\varepsilon_1,\cdots,\varepsilon_j$ 是均值为 0 的独立高斯随机变量, 且 $\mathbf{E}\varepsilon_k^2=\sigma_{t_k}^2-\sigma_{t_{k-1}}^2$. 建立高斯随机向量

$$(W_{t_1},\cdots,W_{t_j}),\quad W_{t_k}=\varepsilon_1+\cdots+\varepsilon_k.$$

假如条件 (A),(B) 和 (C) 当 $t=t_1,\cdots,t_j$ 时成立, 则随机变量 $X^n_{t_1},\cdots,X^n_{t_j}$ 的联合分布 $P^n_{t_1,\cdots,t_j}$ 弱收敛于随机变量 $(W_{t_1},\cdots,W_{t_j})$ 的高斯分布 $P_{t_1,\cdots,t_j}$: $P^n_{t_1,\cdots,t_j}\xrightarrow{w}P_{t_1,\cdots,t_j}$.

注 4 设 $(\sigma_t^2)_{0\leqslant t\leqslant 1}$ 是连续不减函数, $\sigma_t^2=0$. 以 $W=(W_t)_{0\leqslant t\leqslant 1}$ 表示布朗运动过程 (维纳过程), 且 $\mathbf{E}W_t=0,\mathbf{E}W_t^2=\sigma_t^2$. 在第二章 §13 中对于 $\sigma_t^2=t$ 定义了该过程. 在没有这一条件的情况下, 该过程可以类似地定义为, 具有独立增量的高斯过程 $W=(W_t)_{0\leqslant t\leqslant 1}$, 且 $W_0=0$, 而其协方差函数为 $r(s,t)=\min(\sigma_s^2,\sigma_t^2)$. 在随机过程的一般理论中, 证明具有连续轨道的这样的过程总是存在的. (当 $\sigma_t^2=t$ 时这样的过程称为标准布朗运动).

如果以 P^n 和 P 表示过程 X^n 和 W 在函数空间 $(D,\mathscr{B}(D))$ (见第二章 §2 第 7 小节) 中的概率分布, 则可以断定: 对于一切 $0<t\leqslant 1$ 成立的条件 (A), (B) 和 (C), 不仅可以保证上述有限维分布的弱收敛 $(P^n_{t_1,\cdots,t_j}\xrightarrow{w}P_{t_1,\cdots,t_j}, t_1<t_2<\cdots<t_j\leqslant t, j=1,2,\cdots)$, 而且也保证函数的收敛性, 即过程 X^n 的分布 P^n 弱收敛于过程 W 的分布 P. (有关细节, 参见 [5], [91], [87]). 这一结果, 一般称做函数的中心极限定理, 或者 (对于 $n\geqslant 1$, 当随机变量 $\xi_{1n},\cdots,\xi_{nn}$ 独立时) 称做 (汤斯凯 [M. Donsker] – 普罗霍罗夫) 不变原理.

2. 一致渐近可忽略性

定理 2 (1) 条件 (A) 等价于一致极限可忽略条件 (一致渐近可忽略性)

$$(\text{A}^*)\qquad \max_{1\leqslant k\leqslant[nt]}|\xi_{nk}|\xrightarrow{\mathbf{P}}0.$$

(2) 在条件 (A) 或 (A*) 下, 条件 (C) 等价于条件

$$(\text{C}^*)\qquad \sum_{k=1}^{[nt]}[\xi_{nk}-\mathbf{E}\xi_{nk}I(|\xi_{nk}|\leqslant 1)|\mathscr{F}^n_{k-1}]^2\xrightarrow{\mathbf{P}}\sigma_t^2.$$

((A*) 和 (C*) 中与 (A) 和 (C) 中 t 的值相同.)

定理 3 假设对于每一个 $n\geqslant 1$, 序列

$$\xi^n=(\xi_{nk},\mathscr{F}^n_k),\quad 0\leqslant k\leqslant n,$$

是平方可积鞅 — 差, 即 $\mathbf{E}\xi_{nk}^2<\infty, \mathbf{E}(\xi_{nk}|\mathscr{F}^n_{k-1})=0$.

假如满足林德伯格 (J. W. Lindeberg) 条件: 对于任意 $\varepsilon>0$, 有

$$(\text{L})\qquad \sum_{k=1}^{[nt]}\mathbf{E}[\xi_{nk}^2I(|\xi_{nk}|\geqslant\varepsilon)|\mathscr{F}^n_{k-1}]\xrightarrow{\mathbf{P}}0.$$

那么, 条件 (C) 等价于条件

$$\langle X^n\rangle_t\xrightarrow{\mathbf{P}}\sigma_t^2, \tag{1}$$

其中

$$\langle X^n\rangle_t=\sum_{k=1}^{[nt]}\mathbf{E}(\xi_{nk}^2|\mathscr{F}^n_{k-1}) \tag{2}$$

是二次特征, 而条件 (C*) 等价于条件

$$[X^n]_t\xrightarrow{\mathbf{P}}\sigma_t^2, \tag{3}$$

其中

$$[X^n]_t=\sum_{k=1}^{[nt]}\xi_{nk}^2 \tag{4}$$

是二次特征.

由定理 1~3, 得下面的定理.

定理 4 假设对于平方可积鞅 - 差 $\xi^n=(\xi_{nk},\mathscr{F}_k^n), n\geqslant 1$ 和 (对给定的 $0<t\leqslant 1$), 满足条件林德伯格 (L), 那么,

$$\sum_{k=1}^{[nt]}\mathbf{E}(\xi_{nk}^2|\mathscr{F}_{k-1}^n)\xrightarrow{\mathbf{P}}\sigma_t^2\Rightarrow X_t^n\xrightarrow{d}N(0,\sigma_t^2), \tag{5}$$

$$\sum_{k=1}^{[nt]}\xi_{nk}^2\xrightarrow{\mathbf{P}}\sigma_t^2\Rightarrow X_t^n\xrightarrow{d}N(0,\sigma_t^2). \tag{6}$$

3. 定理 1 的证明 1) 将 X_t^n 表示为如下形式:

$$\begin{aligned}X_t^n&=\sum_{k=1}^{[nt]}\xi_{nk}I(|\xi_{nk}|\leqslant 1)+\sum_{k=1}^{[nt]}\xi_{nk}I(|\xi_{nk}|>1)\\&=\sum_{k=1}^{[nt]}\mathbf{E}[\xi_{nk}I(|\xi_{nk}|\leqslant 1)|\mathscr{F}_{k-1}^n]+\sum_{k=1}^{[nt]}\xi_{nk}I(|\xi_{nk}|>1)\\&\quad+\sum_{k=1}^{[nt]}\{\xi_{nk}I(|\xi_{nk}|\leqslant 1)-\mathbf{E}[\xi_{nk}I(|\xi_{nk}|\leqslant 1)|\mathscr{F}_{k-1}^n]\}.\end{aligned}\tag{7}$$

记

$$\begin{aligned}B_t^n&=\sum_{k=1}^{[nt]}\mathbf{E}[\xi_{nk}I(|\xi_{nk}|\leqslant 1)|\mathscr{F}_{k-1}^n],\\\mu_k^n(\Gamma)&=I(\xi_{nk}\in\Gamma),\\\nu_k^n(\Gamma)&=\mathbf{P}(\xi_{nk}\in\Gamma|\mathscr{F}_{k-1}^n),\end{aligned}\tag{8}$$

其中 Γ 是属于集系 $\mathscr{A}_0$ 生成的最小 $\sigma-$ 代数 $\mathscr{B}_0=\sigma(\mathscr{A}_0)$ 的集合, 而 $\mathscr{A}_0$ 是 $\mathbb{R}_0=\mathbb{R}\backslash\{0\}$ 中有限个形如 $(a,b]$ (不包含点 $\{0\}$) 的、不相交区间之和形成的集系; $\mathbf{P}(\xi_{nk}\in\Gamma|\mathscr{F}_{k-1}^n)$ 是 ξ_{nk} 关于 $\sigma-$ 代数 $\mathscr{F}_{k-1}^n$ 的正则条件分布.

那么, (7) 式可以写成

$$X_t^n=B_t^n+\sum_{k=1}^{[nt]}\int_{\{|x|>1\}}x\mathrm{d}\mu_k^n+\sum_{k=1}^{[nt]}\int_{\{|x|\leqslant 1\}}x\mathrm{d}(\mu_k^n-\nu_k^n). \tag{9}$$

表达式 (9) 称做序列 $(X_t^n,\mathscr{F}_{[nt]}^n)$ 的典范分解式. (所有积分应理解为对于每一个基本事件由定义, 勒贝格 — 斯蒂尔切斯积分.)

根据条件 (B), $B_t^n\xrightarrow{\mathbf{P}}0$. 现在证明, 由于条件 (A), 有

$$\sum_{k=1}^{[nt]}\int_{\{|x|>1\}}|x|\mathrm{d}\mu_k^n\xrightarrow{\mathbf{P}}0. \tag{10}$$

有

$$\sum_{k=1}^{[nt]}\int_{\{|x|>1\}}|x|\mathrm{d}\mu_k^n=\sum_{k=1}^{[nt]}|\xi_{nk}|I(|\xi_{nk}|>1). \tag{11}$$

对于任意 $\delta\in(0,1)$,

$$\left\{\sum_{k=1}^{[nt]}|\xi_{nk}|I(|\xi_{nk}|>1)>\delta\right\}\subseteq\left\{\sum_{k=1}^{[nt]}I(|\xi_{nk}|>1)>\delta\right\}. \tag{12}$$

显然,

$$\sum_{k=1}^{[nt]}I(|\xi_{nk}|>1)=\sum_{k=1}^{[nt]}\int_{\{|x|>1\}}\mathrm{d}\mu_k^n\ (\equiv U_{[nt]}^n).$$

由条件 (A), 有

$$V_{[nt]}^n\equiv\sum_{k=1}^{[nt]}\int_{\{|x|>1\}}\mathrm{d}\nu_k^n\xrightarrow{\mathbf{P}}0, \tag{13}$$

并且 $V_{[nt]}^n$ 是 $\mathscr{F}_{k-1}^n-$ 可测的.

那么, 由 §3 定理 4 的系, 有

$$V_{[nt]}^n\xrightarrow{\mathbf{P}}0\Rightarrow U_{[nt]}^n\xrightarrow{\mathbf{P}}0. \tag{14}$$

注意, 由 §3 定理 4 的系, 以及不等式 $\Delta U_{[nt]}^n\leqslant 1$, 有相反的蕴含关系

$$U_{[nt]}^n\xrightarrow{\mathbf{P}}0\Rightarrow V_{[nt]}^n\xrightarrow{\mathbf{P}}0, \tag{15}$$

这在证明定理 2 时将要用到.

由 (11)~(14) 式的所要证明的 (10) 式.

这样,

$$X_t^n=Y_t^n+Z_t^n, \tag{16}$$

其中

$$Y_t^n=\sum_{k=1}^{[nt]}\int_{\{|x|\leqslant 1\}}x\mathrm{d}(\mu_k^n-\nu_k^n), \tag{17}$$

而

$$Z_t^n=B_t^n+\sum_{k=1}^{[nt]}\int_{\{|x|>1\}}x\mathrm{d}\mu_k^n\xrightarrow{\mathbf{P}}0. \tag{18}$$

2) 由于练习题 1, 由此可见为证明收敛性 $X_t^n\xrightarrow{d}N(0,\sigma_t^2)$, 只需证明

$$Y_t^n\xrightarrow{d}N(0,\sigma_t^2). \tag{19}$$

把 Y_t^n 表示为 $(\varepsilon \in (0,1])$

$$Y_t^n = \gamma_{[nt]}^n(\varepsilon) + \Delta_{[nt]}^n(\varepsilon),$$

其中

$$\gamma_{[nt]}^n(\varepsilon) = \sum_{k=1}^{[nt]} \int_{\{\varepsilon<|x|\leqslant 1\}} x\mathrm{d}(\mu_k^n - \nu_k^n), \tag{20}$$

$$\Delta_{[nt]}^n(\varepsilon) = \sum_{k=1}^{[nt]} \int_{\{|x|\leqslant \varepsilon\}} x\mathrm{d}(\mu_k^n - \nu_k^n), \tag{21}$$

像 (10) 式的证明一样, 由于条件 (A) 容易证明 $\gamma_{[nt]}^n(\varepsilon) \xrightarrow{\mathbf{P}} 0, n \to \infty$. 序列 $\Delta^n(\varepsilon) = (\Delta_k^n(\varepsilon), \mathscr{F}_k^n), 1 \leqslant k \leqslant n$ 是平方可积鞅, 且具有二次特征

$$\begin{aligned}\langle \Delta^n(\varepsilon)\rangle_k &= \sum_{i=1}^k \left[\int_{\{|x|\leqslant\varepsilon\}} x^2 \mathrm{d}\nu_i^n - \left(\int_{\{|x|\leqslant\varepsilon\}} x\mathrm{d}\nu_i^n\right)^2\right] \\ &= \sum_{i=1}^k \mathbf{D}[\xi_{ni} I(|\xi_{ni}| \leqslant \varepsilon|\mathscr{F}_{i-1}^n)].\end{aligned}$$

由于条件 (C), 有

$$\langle \Delta^n(\varepsilon)\rangle_{[nt]} \xrightarrow{\mathbf{P}} \sigma_t^2.$$

从而, 对于任意 $\varepsilon \in (0,1]$

$$\max\{\gamma_{[nt]}^n(\varepsilon), |\langle \Delta^n(\varepsilon)\rangle_{[nt]} - \sigma_t^2|\} \xrightarrow{\mathbf{P}} 0.$$

根据练习题 2, 存在数列 $\varepsilon_n \downarrow 0$, 使

$$\gamma_{[nt]}^n(\varepsilon_n) \xrightarrow{\mathbf{P}} 0, \quad \langle \Delta^n(\varepsilon_n)\rangle_{[nt]} \xrightarrow{\mathbf{P}} \sigma_t^2.$$

因此, 仍然由于练习题 1, 只需证明,

$$M_{[nt]}^n \xrightarrow{d} N(0, \sigma_t^2). \tag{22}$$

其中

$$M_k^n = \Delta_k^n(\varepsilon_n) = \sum_{i=1}^k \int_{\{|x|\leqslant\varepsilon_n\}} x\mathrm{d}(\mu_i^n - \nu_i^n). \tag{23}$$

假设对于 $\Gamma \in \mathscr{B}_0$,

$$\widetilde{\mu}_k^n(\Gamma) = I(\Delta M_k^n \in \Gamma), \quad \widetilde{\nu}_k^n(\Gamma) = \mathbf{P}(\Delta M_k^n \in \Gamma|\mathscr{F}_{k-1}^n)$$

是正则条件概率, $\Delta M_k^n = M_k^n - M_{k-1}^n, k \geqslant 1, M_0^n = 0$. 那么, 平方可积鞅 $M^n = (M_k^n, \mathscr{F}_k^n)$, $1 \leqslant k \leqslant n$, 显然可以写成下面的形式:

$$M_k^n = \sum_{i=1}^k \Delta M_i^n = \sum_{i=1}^k \int_{\{|x|\leqslant 2\varepsilon_n\}} x\mathrm{d}\widetilde{\mu}_i^n.$$

(注意, 由于 (23), 可见 $|\Delta M_i^n| \leqslant 2\varepsilon_n$.)

3) 根据第三章 §3 定理 1, 为证明 (22) 式, 需要对于任意实数 λ, 证明

$$\mathbf{E}\mathrm{e}^{i\lambda M_{[nt]}^n} \to \mathrm{e}^{-\frac{\lambda^2\sigma_t^2}{2}}. \tag{24}$$

记

$$G_k^n = \sum_{j=1}^k \int_{\{|x|\leqslant 2\varepsilon_n\}} (\mathrm{e}^{i\lambda x} - 1)\mathrm{d}\widetilde{\nu}_j^n$$

和

$$\mathscr{E}_t^n(G^n) = \prod_{j=1}^k (1 + \Delta G_j^n).$$

注意到

$$1 + \Delta G_k^n = 1 + \int_{\{|x|\leqslant 2\varepsilon_n\}} (\mathrm{e}^{i\lambda x} - 1)\mathrm{d}\widetilde{\nu}_k^n = \mathbf{E}(\mathrm{e}^{i\lambda \Delta M_k^n}|\mathscr{F}_{k-1}^n),$$

从而

$$\mathscr{E}_k^n(G^n) = \prod_{j=1}^k \mathbf{E}(\mathrm{e}^{i\lambda\Delta M_j^n}|\mathscr{F}_{j-1}^n).$$

4) 根据第 4 小节已证明的引理 4, 为验证 (24) 式只需证明, 对于任意实数 λ, 有

$$|\mathscr{E}_{[nt]}^n(G^n)| = \left|\prod_{j=1}^{[nt]} \mathbf{E}(\mathrm{e}^{i\lambda\Delta M_j^n}|\mathscr{F}_{j-1}^n)\right| \geqslant c(\lambda) > 0 \tag{25}$$

和

$$|\mathscr{E}_{[nt]}^n(G^n)| \xrightarrow{\mathbf{P}} \mathrm{e}^{-\frac{\lambda^2\sigma_t^2}{2}}. \tag{26}$$

为此, 将 $\mathscr{E}_k^n(G^n)$ 表示为

$$\mathscr{E}_k^n(G^n) = \mathrm{e}^{G_k^n} \prod_{j=1}^k (1 + \Delta G_j^n)\mathrm{e}^{-\Delta G_j^n}.$$

(对照第二章 §6 中由 (76) 式定义的函数 $\mathscr{E}_t(A)$).

由于

$$\int_{\{|x|\leqslant 2\varepsilon_n\}} x\mathrm{d}\widetilde{\nu}_j^n = \mathbf{E}(\Delta M_j^n|\mathscr{F}_{j-1}^n) = 0,$$

可见

$$G_k^n=\sum_{j=1}^{k}\int_{\{|x|\leqslant 2\varepsilon_n\}}(e^{i\lambda x}-1-i\lambda x)d\widetilde{\nu}_j^n. \tag{27}$$

从而,

$$|\Delta G_k^n|\leqslant\int_{\{|x|\leqslant 2\varepsilon_n\}}|e^{i\lambda x}-1-i\lambda x|d\widetilde{\nu}_k^n\leqslant\frac{\lambda^2}{2}\int_{\{|x|\leqslant 2\varepsilon_n\}}x^2d\widetilde{\nu}_k^n\leqslant\frac{\lambda^2}{2}(2\varepsilon_n)^2\to 0 \tag{28}$$

和

$$\sum_{j=1}^{k}|\Delta G_j^n|\leqslant\frac{\lambda^2}{2}\sum_{j=1}^{k}\int_{\{|x|\leqslant 2\varepsilon_n\}}x^2d\widetilde{\nu}_j^n=\frac{\lambda^2}{2}\langle M^n\rangle_k. \tag{29}$$

根据条件 (C),

$$\langle M^n\rangle_{[nt]}\xrightarrow{\mathbf{P}}\sigma_t^2. \tag{30}$$

首先, 设 $\langle M^n\rangle_{[nt]}\leqslant a$ ($\mathbf{P}$-a.c.), 则由 (28), (29) 式以及练习题 3, 显然

$$\prod_{k=1}^{[nt]}(1+\Delta G_k^n)e^{-\Delta G_k^n}\xrightarrow{\mathbf{P}}1,\quad n\to\infty,$$

因此, 为证明 (26) 式, 只需证明

$$G_{[nt]}^n\xrightarrow{\mathbf{P}}-\frac{\lambda^2\sigma_t^2}{2} \tag{31}$$

或 (由 (27), (29) 和 (30) 各式)

$$\sum_{k=1}^{[nt]}\int_{\{|x|\leqslant 2\varepsilon_n\}}\left(e^{i\lambda x}-1-i\lambda x+\frac{\lambda^2x^2}{2}\right)d\widetilde{\nu}_k^n\xrightarrow{\mathbf{P}}0. \tag{32}$$

由于

$$\left|e^{i\lambda x}-1-i\lambda x+\frac{\lambda^2x^2}{2}\right|\leqslant\frac{|\lambda x|^3}{6},$$

可见

$$\begin{aligned}\sum_{k=1}^{[nt]}\int_{\{|x|\leqslant 2\varepsilon_n\}}\left|e^{i\lambda x}-1-i\lambda x+\frac{\lambda^2x^2}{2}\right|d\widetilde{\nu}_k^n&\leqslant\frac{|\lambda|^3}{6}\times 2\varepsilon_n\times\sum_{k=1}^{[nt]}\int_{\{|x|\leqslant 2\varepsilon_n\}}x^2d\widetilde{\nu}_k^n\\&=\frac{|\lambda|^3\varepsilon_n}{3}\langle M^n\rangle_{[nt]}\leqslant\frac{|\lambda|^3\varepsilon_n}{3}a\to 0,\quad n\to\infty.\end{aligned}$$

这样, 假如 $\langle M^n\rangle_{[nk]}\leqslant a$ ($\mathbf{P}$-a.c.), 则 (31) 式成立, 从而 (26) 式得证.

5) 现在证明性质 (25). 因为

$$|e^{i\lambda x}-1-i\lambda x|\leqslant\frac{(\lambda x)^2}{2},$$

所以, 由 (28) 式可见, 对于充分大的 n, 有

$$\begin{aligned}|\mathscr{E}_k^n(G^n)|&=\left|\prod_{j=1}^{k}(1+\Delta G_j^n)\right|\geqslant\prod_{j=1}^{k}\left(1-\frac{\lambda^2}{2}\Delta\langle M^n\rangle_j\right)\\&=\exp\left\{\sum_{j=1}^{k}\ln\left(1-\frac{\lambda^2}{2}\Delta\langle M^n\rangle_j\right)\right\}.\end{aligned}$$

而 (对于充分大的 n)

$$\ln\left(1-\frac{\lambda^2}{2}\Delta\langle M^n\rangle_j\right)\geqslant-\frac{\frac{\lambda^2}{2}\Delta\langle M^n\rangle_j}{1-\frac{\lambda^2}{2}\Delta\langle M^n\rangle_j},$$

其中 $\Delta\langle M^n\rangle_j\leqslant(2\varepsilon_n)^2\downarrow 0, n\to\infty$, 因此存在 $n_0=n_0(\lambda)$, 使对于一切 $n\geqslant n_0(\lambda)$, 有

$$|\mathscr{E}_k^n(G^n)|\geqslant e^{-\lambda^2\langle M^n\rangle_k},$$

从而

$$|\mathscr{E}_{[nt]}^n(G^n)|\geqslant e^{-\lambda^2\langle M^n\rangle_{[nt]}}\geqslant e^{-\lambda^2a}.$$

于是, 在 $\langle M^n\rangle_{[nk]}\leqslant a$ ($\mathbf{P}$-a.c.) 的条件下, 定理得证.

6) 为去掉条件 $\langle M^n\rangle_{[nk]}\leqslant a$ ($\mathbf{P}$-a.c.), 我们作如下处理, 设

$$\tau^n=\begin{cases}\inf\{k\leqslant[nt]:\langle M^n\rangle_k\geqslant\sigma_t^2+1\}, & 若\ \langle M^n\rangle_{[nt]}\geqslant\sigma_t^2+1,\\ \infty, & 若\ \langle M^n\rangle_{[nt]}<\sigma_t^2+1.\end{cases}$$

那么, 对于 $\overline{M}^n=M_{k\wedge\tau^n}^n$, 有

$$\langle\overline{M}^n\rangle_{[nt]}=M_{[nt]\wedge\tau^n}\leqslant 1+\sigma_t^2+2\varepsilon_n^2\leqslant 1+\sigma_t^2+2\varepsilon_1^2(=a),$$

且根据已证明的 (见 (24) 式)

$$\mathbf{E}e^{i\lambda\overline{M}_{[nt]}^n}\to e^{-\frac{\lambda^2\sigma_t^2}{2}}.$$

但是

$$\lim_n|\mathbf{E}(e^{i\lambda M_{[nt]}^n}-e^{i\lambda\overline{M}_{[nt]}^n})|\leqslant 2\lim_n\mathbf{P}\{\tau^n<\infty\}=0.$$

因此,

$$\lim_n\mathbf{E}e^{i\lambda M_{[nt]}^n}=\lim_n\mathbf{E}(e^{i\lambda M_{[nt]}^n}-e^{i\lambda\overline{M}_{[nt]}^n})+\lim_n\mathbf{E}e^{i\lambda\overline{M}_{[nt]}^n}=e^{-\frac{\lambda^2\sigma_t^2}{2}}.$$

□

注 为证明定理 1 的注 2 中提出的论断, 需要 (根据克拉默 - 沃尔德方法 [5]) 证明, 对于任意实数 $\lambda_1, \cdots, \lambda_j$, 证明

$$\mathbf{E}\exp\left\{i\left[\lambda_1 M^n_{[nt_1]} + \sum_{k=2}^{j}\lambda_k(M^n_{[nt_k]} - M^n_{[nt_{k-1}]})\right]\right\}$$
$$\to \exp\left\{-\frac{\lambda_1^2\sigma_{t_1}^2}{2} - \sum_{k=2}^{j}\frac{\lambda_k^2(\sigma_{t_k}^2 - \sigma_{t_{k-1}}^2)}{2}\right\}.$$

该式的证明与 (24) 式一样, 只是用平方可积鞅 $(\widehat{M}^n_k, \mathscr{F}^n_k)$ 代替 $(M^n_k, \mathscr{F}^n_k)$, 其中

$$\widehat{M}^n_k = \sum_i^k \nu_i \Delta M^n_i,$$

而

$$\nu_i = \begin{cases}\lambda_1, & 若\ i \leqslant [nt_1],\\ \lambda_k, & 若\ [nt_{k-1}] < i \leqslant [nt_k], 2 \leqslant k \leqslant j.\end{cases}$$

4. 辅助命题 在这一小节将证明一个简单的引理, 它可以将验证 (25) 式和 (26) 式, 归结为验证 (24) 式.

设 $\eta^n = (\eta_{nk}, \mathscr{F}^n_k), 1 \leqslant k \leqslant n, n \geqslant 1$ 是随机序列, $Y^n = \sum_{k=1}^{n}\eta_{nk}$,

$$\mathscr{E}^n(\lambda) = \prod_{k=1}^{n}\mathbf{E}(\mathrm{e}^{i\lambda\eta_{nk}}|\mathscr{F}^n_{k-1}), \quad \lambda \in \mathbb{R},$$

Y 是随机变量, 其中

$$\mathscr{E}^n(\lambda) = \mathbf{E}\mathrm{e}^{i\lambda Y}, \quad \lambda \in \mathbb{R}.$$

引理 如果 (对于给定的 λ) $|\mathscr{E}^n(\lambda)| \geqslant c(\lambda) > 0, n \geqslant 1$, 则收敛

$$\mathbf{E}\mathrm{e}^{i\lambda Y^n} \to \mathbf{E}\mathrm{e}^{i\lambda Y} \tag{33}$$

的充分条件为收敛

$$\mathscr{E}^n(\lambda) \xrightarrow{\mathbf{P}} \mathscr{E}(\lambda). \tag{34}$$

证明 设

$$m^n(\lambda) = \frac{\mathrm{e}^{i\lambda Y^n}}{\mathscr{E}^n(\lambda)},$$

则 $|m^n(\lambda)| \leqslant c^{-1}(\lambda) < \infty$, 且容易验证

$$\mathbf{E}m^n(\lambda) = 1.$$

因此, 由 (34) 式以及勒贝格控制收敛定理, 有

$$|\mathbf{E}\mathrm{e}^{i\lambda Y^n} - \mathbf{E}\mathrm{e}^{i\lambda Y}| = |\mathbf{E}(\mathrm{e}^{i\lambda Y^n} - \mathscr{E}(\lambda))| \leqslant |\mathbf{E}\{m^n(\lambda)[\mathscr{E}^n(\lambda) - \mathscr{E}(\lambda)]\}|$$
$$\leqslant c^{-1}(\lambda)\mathbf{E}|\mathscr{E}^n(\lambda) - \mathscr{E}(\lambda)| \to 0, \quad n \to \infty.$$

□

注 由 (33) 式和假设 $|\mathscr{E}^n(\lambda)| \geqslant c(\lambda) > 0$, 可见 $\mathscr{E}(\lambda) \neq 0$. 实际上, 在不要求满足条件 $|\mathscr{E}^n(\lambda)| \geqslant c(\lambda) > 0$ 的情况下, 引理的论点仍然成立; 具体地可以表述为: 如果 $\mathscr{E}^n(\lambda) \xrightarrow{\mathbf{P}} \mathscr{E}(\lambda)$ 且 $\mathscr{E}(\lambda) \neq 0$, 则收敛性 (33) 式成立 (练习题 5).

5. 定理 2 的证明 1) 设 $\varepsilon > 0, \delta \in (0, \varepsilon)$, 并且为简便计设 $t = 1$. 由于

$$\max_{1 \leqslant k \leqslant n}|\xi_{nk}| \leqslant \varepsilon + \sum_{k=1}^{n}|\xi_{nk}|I(|\xi_{nk}| > \varepsilon).$$

和

$$\left\{\sum_{k=1}^{n}|\xi_{nk}|I(|\xi_{nk}| > \varepsilon) > \delta\right\} \subseteq \left\{\sum_{k=1}^{n}I(|\xi_{nk}| > \varepsilon) > \delta\right\},$$

则

$$\mathbf{P}\left\{\max_{1 \leqslant k \leqslant n}|\xi_{nk}| > \varepsilon + \delta\right\} \leqslant \mathbf{P}\left\{\sum_{k=1}^{n}I(|\xi_{nk}| > \varepsilon) > \delta\right\} = \mathbf{P}\left\{\sum_{k=1}^{n}\int_{\{|x|>\varepsilon\}}\mathrm{d}\mu^n_k > \delta\right\}.$$

如果满足条件 (A), 即当 $n \to \infty$ 时

$$\mathbf{P}\left\{\sum_{k=1}^{n}\int_{\{|x|>\varepsilon\}}\mathrm{d}\nu^n_k > \delta\right\} \to 0$$

则 (对照 (10) 式)

$$\mathbf{P}\left\{\sum_{k=1}^{n}\int_{\{|x|>\varepsilon\}}\mathrm{d}\mu^n_k > \delta\right\} \to 0.$$

从而 (A) ⇒ (A*).

相反, 假设条件 (A*) 成立. 对于任意 $\varepsilon > 0$, 设

$$\sigma_n = \begin{cases}\min\left\{k \leqslant n : |\xi_{nk}| > \dfrac{\varepsilon}{2}\right\}, & 若\ \max\limits_{1 \leqslant k \leqslant n}|\xi_{nk}| \geqslant \dfrac{\varepsilon}{2},\\ \infty, & 若\ \max\limits_{1 \leqslant k \leqslant n}|\xi_{nk}| < \dfrac{\varepsilon}{2}.\end{cases}$$

由 (A*) 式 $\lim_n \mathbf{P}\{\sigma_n < \infty\} = 0$.

我们现在指出, 对于任意 $\delta \in (0, 1)$, 集合

$$\left\{\sum_{k=1}^{n\wedge\sigma_n}I\left(|\xi_{nk}| \geqslant \frac{\varepsilon}{2}\right) > \delta\right\} \quad 和 \quad \left\{\max_{1 \leqslant k \leqslant n\wedge\sigma_n}|\xi_{nk}| \geqslant \frac{\varepsilon}{2}\right\}$$

相等, 而根据条件 (A*), 有

$$\sum_{k=1}^{n\wedge\sigma_n}I\left(|\xi_{nk}| \geqslant \frac{\varepsilon}{2}\right) = \sum_{k=1}^{n\wedge\sigma_n}\int_{\{|x|\geqslant\frac{\varepsilon}{2}\}}\mathrm{d}\mu^n_k \xrightarrow{\mathbf{P}} 0.$$

因此, 由于 (15) 式, 有

$$\sum_{k=1}^{n\wedge\sigma_n}\int_{\{|x|\geqslant\varepsilon\}}\mathrm{d}\nu_k^n\leqslant\sum_{k=1}^{n\wedge\sigma_n}\int_{\{|x|\geqslant\frac{\varepsilon}{2}\}}\mathrm{d}\nu_k^n\xrightarrow{\mathbf{P}}0;$$

由该式连同性质 $\lim_n\mathbf{P}\{\sigma_n<\infty\}=0$ 最后证明蕴涵关系 $(\mathrm{A}^*)\Rightarrow(\mathrm{A})$.

2) 仍然假设 $t=1$. 固定某个 $\varepsilon\in(0,1]$, 对于任意 $\delta\in(0,\varepsilon]$, 考虑平方可积鞅 (见 (21) 式),

$$\Delta^n(\delta)=(\Delta_k^n(\delta),\mathscr{F}_k^n),\quad 1\leqslant k\leqslant n.$$

由于条件 (C), 对于给定的 $\varepsilon\in(0,1]$, 有

$$\langle\Delta^n(\varepsilon)\rangle_n\xrightarrow{\mathbf{P}}\sigma_1^2.$$

从而, 由于条件 (A), 容易导出, 对于任意 $\delta\in(0,\varepsilon]$, 有

$$\langle\Delta^n(\delta)\rangle_n\xrightarrow{\mathbf{P}}\sigma_1^2.\tag{35}$$

现在证明, 由条件 (C^*) 和 (A), 或等价地, 由条件 (C^*) 和 (A^*) 可见, 对于任意 $\delta\in(0,\varepsilon]$, 有

$$[\Delta^n(\delta)]_n\xrightarrow{\mathbf{P}}\sigma_1^2,\tag{36}$$

其中

$$[\Delta^n(\delta)]_n=\sum_{k=1}^{n}\left[\xi_{nk}I(|\xi_{nk}|\leqslant\delta)-\int_{\{|x|\leqslant\delta\}}x\mathrm{d}\nu_k^n\right]^2.$$

事实上, 由于 (A) 式, 容易验证,

$$[\Delta^n(\delta)]_n-[\Delta^n(1)]_n\xrightarrow{\mathbf{P}}0.\tag{37}$$

$$\begin{aligned}&\left|\sum_{k=1}^{n}\left[\xi_{nk}-\int_{\{|x|\leqslant1\}}x\mathrm{d}\nu_k^n\right]^2-\sum_{k=1}^{n}\left[\xi_{nk}I(|\xi_{nk}|\leqslant1)-\int_{\{|x|\leqslant1\}}x\mathrm{d}\nu_k^n\right]^2\right|\\&\leqslant\sum_{k=1}^{n}I(|\xi_{nk}|>1)\left[\xi_{nk}^2+2|\xi_{nk}|\times\left|\int_{\{|x|\leqslant1\}}x\mathrm{d}(\mu_k^n-\nu_k^n)\right|\right]\\&\leqslant5\sum_{k=1}^{n}I(|\xi_{nk}|>1)\xi_{nk}^2\leqslant5\max_{1\leqslant k\leqslant n}\xi_{nk}^2\times\sum_{k=1}^{n}\int_{\{|x|>1\}}\mathrm{d}\mu_k^n\xrightarrow{\mathbf{P}}0.\end{aligned}\tag{38}$$

因此, 由 (37) 式和 (38) 式得 (36) 式.

这样, 为证明条件 (C) 与条件 (C^*) 等价, 只需证明: 在 (对于给定的 $\varepsilon\in(0,1]$) 条件 (C) 满足时, 同时, 若对于任意 $a>0$, 条件 (C^*) 成立, 则

$$\lim_{\delta\to0}\overline{\lim_n}\mathbf{P}\{|[\Delta^n(\delta)]_n-\langle\Delta^n(\delta)\rangle_n|>a\}=0.\tag{39}$$

记 $m_k^n(\delta)=[\Delta^n(\delta)]_k-\langle\Delta^n(\delta)\rangle_k,1\leqslant k\leqslant n$. 序列 $m^n(\delta)=(m_k^n(\delta),\mathscr{F}_k^n)$ 是平方可积鞅, 这时序列 $(m^n(\delta))^2$ 被序列 $[\Delta^n(\delta)]_n$ 和 $\langle\Delta^n(\delta)\rangle_n$ 控制 (由 §3 定义).

显然,

$$\begin{aligned}[m^n(\delta)]_n&=\sum_{k=1}^{n}[\Delta m_k^n(\delta)]^2\leqslant\max_{1\leqslant k\leqslant n}|\Delta m_k^n(\delta)|\times\{[\Delta^n(\delta)]_n+\langle\Delta^n(\delta)\rangle_n\}\\&\leqslant3\delta^2\{[\Delta^n(\delta)]_n+\langle\Delta^n(\delta)\rangle_n\}.\end{aligned}\tag{40}$$

由于 $[\Delta^n(\delta)]$ 和 $\langle\Delta^n(\delta)\rangle$ 相互控制, 故由 (40) 式可见, $(m^n(\delta))^2$ 受序列 $6\delta^2[\Delta^n(\delta)]$ 和 $6\delta^2\langle\Delta^n(\delta)\rangle$ 控制.

因此, 如果满足条件 (C), 则对于充分小的 δ (例如, $\delta^2<\min(\varepsilon,b(\sigma_1^2+1)/6)$), 有

$$\overline{\lim_n}\mathbf{P}\{6\delta^2\langle\Delta^n(\delta)\rangle_n>b\}=0,$$

从而, 由 §3 定理 4 的系知 (39) 式成立. 而假如满足条件 (C^*), 则对于同一 δ, 有

$$\overline{\lim_n}\mathbf{P}\{6\delta^2[\Delta^n(\delta)]_n>b\}=0.\tag{41}$$

仍然由 §3 定理 4 的系知, 由于 $|\Delta[\Delta^n(\delta)]_k|\leqslant(2\delta)^2$, 故由 (41) 式得 (39) 式. □

6. 定理 3 的证明 考虑林德伯格条件 (L), 可以通过直接计算验证条件 (C) 和 (1), 以及条件 (C^*) 和 (3) 的等价性 (练习题 6).

7. 定理 4 的证明 由林德伯格条件 (L), 可见条件 (A) 成立. 关于条件 (B), 只需注意到, 在 ξ^n 构成鞅 – 差的情形下, 典型分解式 (9) 中的随机变量 B_t^n 可以表为如下形式:

$$B_t^n=-\sum_{k=1}^{[nt]}\int_{\{|x|>1\}}x\mathrm{d}\nu_k^n.$$

所以, 由林德伯格条件 (L), 可见 $B_t^n\xrightarrow{\mathbf{P}}0$.

8. 平方可积鞅 – 差的情形 这一节的基本定理 —— 定理 1, 是在被加项之和一致渐近小的条件下证明的. 自然提出的问题是, 在没有一致渐近小条件的情况下, 中心极限定理成立的条件如何? 对于独立随机变量的情形, 第三章 §5 的定理 1 就是这样的例子 (其条件是: 二阶矩有限).

我们现在 (不加证明地) 给出这一定理的类似, 但是局限于序列 $\xi^n=(\xi_{nk},\mathscr{F}_k^n)$, $1\leqslant k\leqslant n$, 是平方可积鞅 – 差的情形 ($\mathbf{E}(\xi_{nk}^2<\infty,\mathbf{E}(\xi_{nk}|\mathscr{F}_{k-1}^n)=0$).

记 $F_{nk}(x)=\mathbf{P}\{\xi_{nk}\leqslant x|\mathscr{F}_{k-1}^n\}$ 是 ξ_{nk} 关于 $\mathscr{F}_{k-1}^n$ 的正则分布函数. 设 $\Delta_{nk}=\mathbf{E}(\xi_{nk}^2|\mathscr{F}_{k-1}^n)$.

定理 5 如果对于平方可积鞅 – 差 $\xi^n=(\xi_{nk},\mathscr{F}_k^n)$, $0\leqslant k\leqslant n$, $n\geqslant1$, 满足条件:

$$\sum_{k=1}^{[nt]}\Delta_{nk}\xrightarrow{\mathbf{P}}\sigma_t^2,\quad 0\leqslant\sigma_t^2<\infty,\quad 0\leqslant t\leqslant1,$$

而对于任意 $\varepsilon > 0$,

$$\sum_{k=1}^{[nt]} \int_{\{|x|>\varepsilon\}} |x| \left| F_{nk}(x) - \Phi\left(\frac{x}{\sqrt{\Delta_{nk}}}\right) \right| \mathrm{d}x \xrightarrow{\mathbf{P}} 0,$$

且

$$X_t^n \xrightarrow{d} N(0, \sigma_t^2).$$

9. 练习题

1. 设 $\xi_n = \eta_n + \zeta_n, n \geqslant 1$, 其中 $\eta_n \xrightarrow{d} \eta$, 而 $\zeta_n \xrightarrow{d} 0$. 证明 $\xi_n \xrightarrow{d} \eta$.

2. 设 $(\xi(\varepsilon)), n \geqslant 1, \varepsilon > 0$, 是随机变量族, 且对于每一个 $\varepsilon > 0$, 当 $n \to \infty$ 时, 有 $\xi_n \xrightarrow{\mathbf{P}} 0$. 例如, 利用第二章 §10 练习题 11 的论断证明, 存在序列 $\varepsilon_n \downarrow 0$, 使 $\xi_n(\varepsilon_n) \xrightarrow{\mathbf{P}} 0$.

3. 设 $(\alpha_k^n), 1 \leqslant k \leqslant n, n \geqslant 1$, 是复数值随机变量, ($\mathbf{P}$-a.c.)

$$\sum_{k=1}^{n} |\alpha_k^n| \leqslant C, \quad |\alpha_k^n| \leqslant a_n \downarrow 0.$$

证明 ($\mathbf{P}$-a.c.)

$$\lim_n \prod_{k=1}^{n} (1 + \alpha_k^n) \mathrm{e}^{-\alpha_k^n} = 1.$$

4. 证明定理 1 的注 2 中提出的命题.

5. 证明引理的注中提出的命题.

6. 证明定理 3.

7. 证明定理 5.

§9. 伊藤公式的离散版本

1. 引言 在布朗运动、以及与之同类的过程 (鞅、局部鞅、下鞅 ……) 的随机分析中, *伊藤清* (Itô Kiyosi) *变量替换公式*有特别的作用. 在这一节考虑该公式的*离散型变式*, 并且说明如何经极限过渡, 可以得到*布朗运动的伊藤清公式*.

2. 二次协方差的积分表示 设 $X = (X_n)_{0 \leqslant n \leqslant N}$ 和 $Y = (Y_n)_{0 \leqslant n \leqslant N}$ 是概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 上的两个随机变量序列, 且 $X_0 = Y_0 = 0$ 而

$$[X, Y] = ([X, Y]_n)_{0 \leqslant n \leqslant N},$$

其中

$$[X, Y]_n = \sum_{k=1}^{n} \Delta X_k \Delta Y_k \tag{1}$$

是序列 X 和 Y 的*二次协方差* (见 §1).

假设给定一绝对连续函数 $F = F(x)$:

$$F(x) = F(0) + \int_0^x f(y) \mathrm{d}y, \tag{2}$$

其中 $f = f(y)$ 是 $\mathbb{R}$ 上的博雷尔函数, 对于任意 $c > 0$, 满足条件

$$\int_{|y| \leqslant c} |f(y)| \mathrm{d}y < \infty.$$

下面提到的变量替换公式, 用序列 $X = (X_n)_{0 \leqslant n \leqslant N}$ "自然" 泛函的语言, 给出序列

$$F(X) = (F(X_n))_{0 \leqslant n \leqslant N}. \tag{3}$$

表示的概念.

考虑序列 X 和 $f(X) = (f(X_n))_{0 \leqslant n \leqslant N}$ 的二次协方差 $[X, f(X)]$, 其中 $f = (x), x \in \mathbb{R}$, 是 (2) 式中的函数. 根据 (1) 式

$$[X, f(X)]_n = \sum_{k=1}^{n} \Delta f(X_k) \Delta X_k = \sum_{k=1}^{n} [f(X_k) - f(X_{k-1})](X_k - X_{k-1}). \tag{4}$$

如果引进两个 "离散积分" (对照 §1 定义 5)

$$I_n(X, f(X)) = \sum_{k=1}^{n} f(X_{k-1}) \Delta X_k, \quad 1 \leqslant n \leqslant N, \tag{5}$$

和

$$\widetilde{I}_n(X, f(X)) = \sum_{k=1}^{n} f(X_k) \Delta X_k, \quad 1 \leqslant n \leqslant N, \tag{6}$$

则二次协方差可以表示为

$$[X, f(X)]_n = \widetilde{I}_n(X, f(X)) - I_n(X, f(X)). \tag{7}$$

(当 $n = 0$ 时, 设 $I_0 = \widetilde{I}_0 = 0$).

在我们所考虑的情形下, 对于*固定*的实数 N, 引进新的 ("逆向") 序列 $\widetilde{X} = (\widetilde{X}_n)_{0 \leqslant n \leqslant N}$, 得

$$\widetilde{X}_n = X_{N-n}. \tag{8}$$

显然,

$$\widetilde{I}_N(X, f(X)) = -I_N(\widetilde{X}, f(\widetilde{X})),$$

且类似地

$$\widetilde{I}_n(X, f(X)) = -\{I_N(\widetilde{X}, f(\widetilde{X})) - I_{N-n}(\widetilde{X}, f(\widetilde{X}))\},$$

因此, 由 (7) 式可得

$$[X,f(X)]_N=-\{I_N(\widetilde{X},f(\widetilde{X}))+I_N(X,f(X))\},$$

且对于 $1<n<N$, 有

$$\begin{aligned}[X,f(X)]_n&=-\{I_N(\widetilde{X},f(\widetilde{X}))-I_{N-n}(\widetilde{X},f(\widetilde{X}))\}-I_n(X,f(X))\\&=-\left\{\sum_{k=N-n+1}^{N}f(\widetilde{X}_{k-1})\Delta\widetilde{X}_k+\sum_{k=1}^{n}f(X_{k-1})\Delta X_k\right\}.\end{aligned}\tag{9}$$

注 有益地注意到, 由 (7) 式和 (9) 式给出的二次协方差 $[X,f(X)]_n$ 表达式的不同结构. (7) 式中的 "积分"

$$I_n(X,f(X))=\sum_{k=1}^{n}f(X_{k-1})\Delta X_k,$$

是这样形成的: 在区间 $[k-1,k]$ ("左" 端点的) 值 $f(X_{k-1})$, 在此整个区间上都乘以 $\Delta X_k=X_k-X_{k-1}$. 然而, "积分" $\widetilde{I}_n(X,f(X))$ 的组成却不同: 是增量 $\Delta X_k=X_k-X_{k-1}$ 乘以区间 $[k-1,k]$ 的 "右" 端点上的值, 即值 X_k.

这样, 可以说,(7) 式既包含 "正向积分 $I_n(X,f(X))$", 又包含 "反向积分 $\widetilde{I}_n(X,f(X))$". 然而, 在 (9) 式中 (对于序列 X 和 $\widetilde{X}$) 所有 "积分" 都是 "正向的".

3. 伊藤公式的离散版本 由于对于每一个函数 $g=g(x)$, 有

$$g(X_{k-1})+\frac{1}{2}[g(X_k)-g(X_{k-1})]-\frac{1}{2}[g(X_k)+g(X_{k-1})]=0,$$

则显然

$$\begin{aligned}F(X_n)=F(X_0)&+\sum_{k=1}^{n}g(X_{k-1})\Delta X_k+\frac{1}{2}[X,g(X)]_n\\&+\sum_{k=1}^{n}\left\{[F(X_k)-F(X_{k-1})]-\frac{g(X_{k-1})+g(X_k)}{2}\Delta X_k\right\}.\end{aligned}\tag{10}$$

特别, 如果 $g(x)=f(x)$, 其中 $f=f(x)$ 是 (2) 式中的函数, 则

$$F(X_n)=F(X_0)+I_n(X,f(X))+\frac{1}{2}[X,f(X)]_n+R_n(X,f(X)),\tag{11}$$

其中

$$R_n(X,f(X))=\sum_{k=1}^{n}\int_{X_{k-1}}^{X_k}\left[f(x)-\frac{f(X_{k-1})+f(X_k)}{2}\right]\mathrm{d}x.\tag{12}$$

由数学分析中熟知的性质, 如果函数 $f''(x)$连续, 则有 "梯形公式":

$$\begin{aligned}\int_a^b\left[f(x)-\frac{f(a)+f(b)}{2}\right]\mathrm{d}x&=\int_a^b(x-a)(x-b)\frac{f''(\xi(x))}{2!}\mathrm{d}x\\&=\frac{(b-a)^3}{2}\int_0^1x(x-1)f''(\xi[a+(b-a)x])\mathrm{d}x\\&=\frac{(b-a)^3}{2}f''(\xi[a+(b-a)\overline{x}])\int_0^1x(x-1)\mathrm{d}x=-\frac{(b-a)^3}{12}f''(\eta),\end{aligned}$$

其中 $\xi(x),\overline{x}$ 和 η 是区间 $[a,b]$ "中间" 的点.

因此, 由 (12) 式, 有

$$R_n(X,f(X))=-\frac{1}{12}\sum_{k=1}^{n}f''(\eta_k)(\Delta X_k)^3,$$

其中 $X_{k-1}\leqslant\eta_k\leqslant X_k$. 由上面的表达式, 易见

$$|R_n(X,f(X))|\leqslant\frac{1}{12}\sup f''(\eta)\times\sum_{k=1}^{n}|\Delta X_k|^3,\tag{13}$$

其中 sup 对于满足 $\min(X_0,X_1,\cdots,X_n)\leqslant\eta\leqslant\max(X_0,X_1,\cdots,X_n)$ 的一切 η 值来求:

我们把 (11) 式称做*伊藤公式的离散版本*. 需要强调, 该公式的右侧由如下三部分构成: "离散积分" $I_n(X,f(X))$, 二次协方差 $[X,f(X)]_n$, 和 "余" 项 $R_n(X,f(X))$. "余" 项的名称在于说明, 当由相应的极限过程向连续情形过渡时, $R_n(X,f(X))$ 的极限为 0 (详见第 5 小节).

4. 例

例 1 如果 $f(x)=a+bx$, 则 $R_n(X,f(X))=0$, 这时 (11) 式有如下形式:

$$F(X_n)=F(X_0)+I_n(X,f(X))+\frac{1}{2}[X,f(X)]_n.\tag{14}$$

(对照下面的 (19) 式).

例 2 设

$$f(x)=\operatorname{sign}x=\begin{cases}1,&x>0,\\0,&x=0,\\-1,&x<0,\end{cases}$$

而 $F(x)=|x|$.

假设 $X_k=S_k$, 其中 $S_k=\xi_1+\cdots+\xi_k, k\geqslant1$, 而 $\xi_1,\xi_2,\cdots$ 是独立伯努利随机变量序列, 且每一个 $\xi_k(k\geqslant1)$ 都以概率 1/2 分别取 ±1 为值.

如果设 $S_0=0$, 则直接由 (11) 式可以求得

$$|S_n|=\sum_{k=1}^{n}(\operatorname{sign}S_{k-1})\Delta S_k+N_n,\tag{15}$$

其中 $N_n = \#\{0 \leqslant k < n, S_k = 0\}$ 是序列 $S_0, S_1, \cdots$ 中 "0" 的个数.

包含在 (14) 式中的 "离散积分"

$$\left(\sum_{k=1}^{n} (\mathrm{sign} S_{k-1}) \Delta S_k\right)_{n \geqslant 1}$$

是鞅. 因此, 由 (15) 式可见

$$\mathbf{E}|S_n| = \mathbf{E} N_n. \tag{16}$$

因为 (练习题 2)

$$\mathbf{E}|S_n| \sim \sqrt{\frac{2n}{\pi}}, \quad n \to \infty, \tag{17}$$

所以由 (16) 式可见

$$\mathbf{E} N_n \sim \sqrt{\frac{2n}{\pi}}, \quad n \to \infty. \tag{18}$$

换句话说, 在随机游动 $S_0, S_1, \cdots, S_n$ 中 "平局" 的平均次数按量级 $\sqrt{n}$ 增长, 而不是像最初感觉的那样好像 "按量级 n 增长". 需要指出, 性质 (18) 十分直接地与*反正弦定律* (见第一章 §10) 相联系, 并且实际上可以由反正弦定律得到.

5. 布朗运动的伊藤变量替换公式 设 $B = (B_t)_{0 \leqslant t \leqslant 1}$ 是标准 ($B_0 = 0, \mathbf{E} B_t = 0, \mathbf{E} B_t^2 = t$) 布朗运动 (见第二章 §13), 而 $X_k = B_{k/n}, k = 0, 1, \cdots, n$.

运用 (11) 式, 可得如下结果

$$\begin{aligned} F(B_1) = F(B_0) &+ \sum_{k=1}^{n} f(B_{(k-1)/2}) \Delta B_{k/n} \\ &+ \frac{1}{2} [f(B_{\cdot/n}), B_{\cdot/n}]_n + R_n(B_{\cdot/n}, f(B_{\cdot/n})). \end{aligned} \tag{19}$$

由布朗运动的理论 (例如, 可参见 [11], [17], [77]), 知

$$\sum_{k=1}^{n} |B_{k/n} - B_{(k-1)/n}|^3 \xrightarrow{\mathbf{P}} 0, \quad n \to 0. \tag{20}$$

从而, 如果函数 $f = f(x)$ 有*二阶导数*, 且对于某个常数 $C > 0, f''(x) \leqslant C, x \in \mathbb{R}$, 则由估计 (13) 式, 可见 $R_n(B_{\cdot/n}, f(B_{\cdot/n})) \xrightarrow{\mathbf{P}} 0$.

仍然由布朗运动的理论知, 对于任意博雷尔函数 $f = f(x) \in L^2_{\mathrm{loc}}$ (即对于任意常数 $C > 0$, 满足

$$\int_{|x| \leqslant C} f^2(x) dx < \infty$$

的函数), "离散积分"

$$\sum_{k=1}^{n} f(B_{(k-1)/n}) \Delta B_{k/n}$$

(在依概率收敛的意义上显然) 存在极限, 记作 $\int_0^1 f(B_s) \mathrm{d}B_s$, 并且称做*布朗运动的伊藤清随机积分*.

这样, 由 (19) 式可见, 该式中的 "余" 项 $R_n(B_{\cdot/n}, f(B_{\cdot/n})) \xrightarrow{\mathbf{P}} 0$, "离散积分"

$$\sum_{k=1}^{n} f(B_{(k-1)/n}) \Delta B_{k/n}$$

(依概率) 收敛于 "随机积分" $\int_0^1 f(B_s) \mathrm{d}B_s$, 从而二次协方差

$$[B_{\cdot/n}, f(B_{\cdot/n})] \quad (= [f(B_{\cdot/n}), B_{\cdot/n}])$$

依概率收敛的极限, 自然应当记为

$$[B, f(B)]_1.$$

这样, 如果函数 $f = f(x)$ 由*二阶导数*, 且对于某个常数 $C > 0, |f''(x)| \leqslant C, x \in \mathbb{R}$, 且 $f \in L^2_{\mathrm{loc}}$, 则有如下公式:

$$F(B_1) = F(0) + \int_0^1 f(B_s) \mathrm{d}B_s + \frac{1}{2} [B, f(B)]_1. \tag{21}$$

这时,

$$[B, f(B)]_1 = \int_0^1 f'(B_s) \mathrm{d}s. \tag{22}$$

于是,

$$F(B_1) = F(0) + \int_0^1 f(B_s) \mathrm{d}B_s + \frac{1}{2} \int_0^1 f'(B_s) \mathrm{d}s, \tag{23}$$

或者更标准的形式为:

$$F(B_1) = F(0) + \int_0^1 F'(B_s) \mathrm{d}B_s + \frac{1}{2} \int_0^1 F''(B_s) \mathrm{d}s, \tag{24}$$

正是这一公式 (对于 $F \in C^2$) 称做*布朗运动的伊藤变量替换公式*.

6. 练习题

1. 证明公式 (15).
2. 证明渐近式 (17).
3. 证明公式 (22).
4. 试证明对于任意函数 $F \in C^2$ (24) 式成立.

§10. 保险中破产概率的计算. 鞅方法

1. 破产概率 这一节将要介绍的内容可以很好地演示, *鞅方法*如何 (例如为保险公司) 提供破产概率的简单估计.

设 $X=(X_t)_{t\geqslant 0}$ 是描绘所考察保险公司资本演变的随机过程. 以 $X_0=u>0$ 的值表示公司的初始资本. 假设保费收入由公司以固定的速度 $c>0$ (即经时间 Δt, 收入为 $c\Delta t$) 连续地积累. 假设偿付保费的诉求出现在随机的时间 $T_1,T_2,\cdots(0<T_1<T_2<\cdots)$, 而赔偿的相应金额表现为非负随机变量 $\xi_1,\xi_2,\cdots$

由以上的说明可见, 在时间 $t>0$ 保险公司的资本为

$$X_t=u+ct-S_t, \tag{1}$$

其中

$$S_t=\sum_{i\geqslant 1}\xi_i I(T_i\leqslant t). \tag{2}$$

设

$$T=\inf\{t\geqslant 0:X_t\leqslant 0\}$$

是保险公司的资本首次变为 0 或负值的时间.

如果对于一切 $t\geqslant 0, X_t>0$, 则设 T 等于 $+\infty$. 由完全明显的理由, 自然把时间 T 称为公司的 "破产时间". 我们在下面的主要兴趣是, 求 (或者估计) *破产的概率* $\mathbf{P}\{T<\infty\}$, 或对于任意 $t>0$, *在时刻 t 前破产的概率* $\mathbf{P}\{T\leqslant t\}$.

2. 克拉默 – 伦德伯格模型 求这些破产概率是相当不容易的问题. 不过, 对于*克拉默 – 伦德伯格* (H. Cramér – G. A. Lundeberg) *模型*, 该问题有 (部分) 解. 而所谓克拉默 – 伦德伯格模型的特点是, 它满足如下条件.

A. 假设时间 $\sigma_i=T_i-T_{i-1}, i\geqslant 1 (T_0=0)$, 是独立随机变量, 并且服从指数分布: $\mathbf{P}\{\sigma_i>t\}=\lambda \mathrm{e}^{-\lambda t}, t\geqslant 0, i\geqslant 1$. (见第二章 §3 表 3).

B. 随机变量 $\xi_1,\xi_2,\cdots$ 是独立同分布, 其分布函数 $F(x)=\mathbf{P}\{\xi_1\leqslant x\}$ 满足条件: $F(0)=0$,

$$\mu=\int_0^\infty x\mathrm{d}F(x)<\infty.$$

C. 序列 $(T_1,T_2,\cdots)$ 和 $(\xi_1,\xi_2,\cdots)$ 是独立序列 (按第二章 §5 定义 6 的含义).

设

$$N_t=\sum_{i\geqslant 1}I(T_i\leqslant t),\quad t>0, \tag{3}$$

是描绘 t 时前 (包括时刻 t), 偿付保费诉求数量的过程, 且 $N_0=0$.

因为对于 $k\geqslant 1$,

$$\{T_k>t\}=\{\sigma_1+\cdots+\sigma_k>t\}=\{N_t<k\},$$

所以注意到条件 **A**, 且根据第二章 §8 练习题 6, 有

$$\mathbf{P}\{N_t<k\}=\mathbf{P}\{\sigma_1+\cdots+\sigma_k>t\}=\sum_{i=0}^{k-1}\mathrm{e}^{-\lambda t}\frac{(\lambda t)^i}{i!}.$$

从而

$$\mathbf{P}\{N_t=k\}=\mathrm{e}^{-\lambda t}\frac{(\lambda t)^k}{k!}\quad k=0,1,\cdots \tag{4}$$

这样, 随机变量 N_t 有泊松分布 (参见第二章 §3 表2), 参数为 λt, 恰好等于数学期望 $\mathbf{E}N_t$.

按 (3) 式构造的过程 $N=(N_t)_{t\geqslant 0}$, 特别更新过程的情况 (第二章 §9 第 4 小节), 称做*泊松过程*. 该过程的轨道是间断的 (确切地说, 是阶梯函数, 右连续且跃度皆等于 1). 与轨道是连续函数的布朗运动 (第二章 §13) 一样, 泊松过程在随机过程论中重要作用. 具体地说, 利用这两个过程, 可以建立具有相当复杂概率结构的随机过程.(独立增量过程就是这种情形的典型例子: 例如. 见 [13], [11], [76]).

3. 克拉默 – 伦德伯格模型下的破产概率 由条件 **C** 可见,

$$\begin{aligned}\mathbf{E}(X_t-X_0)&=ct-\mathbf{E}S_t=ct-\mathbf{E}\sum_i\xi_iI(T_i\leqslant t)=ct-\sum_i\mathbf{E}\xi_iI(T_i\leqslant t)\\&=ct-\sum_i\mathbf{E}\xi_i\mathbf{E}I(T_i\leqslant t)=ct-\mu\sum_i\mathbf{P}\{T_i\leqslant t\}\\&=ct-\mu\sum_i\mathbf{P}\{N_i\geqslant i\}=ct-\mu\mathbf{E}N_t=t(c-\lambda\mu).\end{aligned}$$

由此清楚地看到, 公司经营赢利 (即 $\mathbf{E}(X_t-X_0)>0$) 的条件, 表现为

$$c>\lambda\mu. \tag{5}$$

在下面的分析中, 起重要作用的下面的函数

$$h(z)=\int_0^\infty(\mathrm{e}^{zx}-1)\mathrm{d}F(x),\quad z\geqslant 0, \tag{6}$$

其中 $h(z)=\widehat{F}(-z)-1$, 而

$$\widehat{F}(s)=\int_0^\infty\mathrm{e}^{-sx}\mathrm{d}F(x)$$

是拉普拉斯 – 斯蒂尔切斯 (P. S. Laplace – T. J. Stieltjes) 变换 (s 是复数).

设

$$g(z)=\lambda h(z)-cz,\quad \xi_0=0,$$

则对于任何使 $h(r)<\infty$ 的 $r>0$, 有

$$\begin{aligned}\mathbf{E}\mathrm{e}^{-r(X_t-X_0)}&=\mathbf{E}\mathrm{e}^{-r(X_t-\mu)}=\mathrm{e}^{-rct}\mathbf{E}\mathrm{e}^{r\sum\limits_{i=0}^{N_t}\xi_i}=\mathrm{e}^{-rct}\sum_{n=0}^\infty\mathbf{E}\mathrm{e}^{r\sum\limits_{i=0}^{N_t}\xi_i}\mathbf{P}\{N_t=n\}\\&=\mathrm{e}^{-rct}\sum_{n=0}^\infty[1+h(r)]^n\frac{\mathrm{e}^{-\lambda t}(\lambda t)^n}{n!}=\mathrm{e}^{-rct}\mathrm{e}^{\lambda th(r)}=\mathrm{e}^{t[\lambda h(r)-cr]}=\mathrm{e}^{tg(r)}.\end{aligned}$$

对于 $s<t$, 类似地可以得到

$$\mathbf{E}e^{-r(X_t-X_s)}=e^{(t-s)g(r)}. \tag{7}$$

设 $\mathscr{F}_0^X=\sigma(X_s,s\leqslant t)$. 由于过程 $X=(X_t)_{t\geqslant 0}$ 是独立增量过程 (练习题 2), 则 ($\mathbf{P}$ - a.c.)

$$\mathbf{E}(e^{-r(X_t-X_s)}|\mathscr{F}_s^X)=\mathbf{E}e^{-r(X_t-X_s)}=e^{(t-s)g(r)},$$

从而

$$\mathbf{E}(e^{-rX_t-tg(r)}|\mathscr{F}_s^X)=e^{-rX_s-sg(r)}, \tag{8}$$

设

$$Z_t=e^{-rX_s-sg(r)},\quad t\geqslant 0. \tag{9}$$

由 (8) 式可见, 显然

$$\mathbf{E}(Z_t|\mathscr{F}_s^X)=Z_s,\quad s\leqslant t. \tag{10}$$

与 §1 定义 1 类似, 自然称过程 $Z=(Z_t)_{t\geqslant 0}$ (关于 $\sigma-$ 代数 "流" $(\mathscr{F}_t^X)_{t\geqslant 0}$) 是*鞅*. 注意, 对于所考虑的情形, $\mathbf{E}|Z_t|<\infty, t\geqslant 0$ (对照 §1 的性质 1). 与 §1 定义 3 类似, 值域为 $[0,+\infty]$ 的随机变量 $\tau=\tau(\omega)$ 是*马尔可夫时间*, 或 (关于 $\sigma-$ 代数 "流" $(\mathscr{F}_t^X)_{t\geqslant 0}$) *不依赖于将来的随机变量*, 假如对于每一个 $t\geqslant 0$, 有

$$\{\tau(\omega)\leqslant t\}\in\mathscr{F}_t^X.$$

对于现在考虑的连续时间的情形, §2 的定理 1 (在明显地改变记号的情形下) 仍然成立. 特别, 对于马尔可夫时间 τ, 有

$$\mathbf{E}Z_{t\wedge\tau}=\mathbf{E}Z_0. \tag{11}$$

设 $T=\tau$. 那么, 由 (9) 式和 (11) 式, 可见对于任意 $t>0$, 有

$$\begin{aligned}e^{-ru}&=\mathbf{E}e^{-rX_{t\wedge T}-(t\wedge T)g(r)}\geqslant\mathbf{E}[e^{-rX_{t\wedge T}-(t\wedge T)g(r)}|T\leqslant t]\mathbf{P}\{T\leqslant t\}\\&=\mathbf{E}[e^{-rX_T-Tg(r)}|T\leqslant t]\mathbf{P}\{T\leqslant t\}\geqslant\mathbf{E}[e^{-Tg(r)}|T\leqslant t]\mathbf{P}\{T\leqslant t\}\\&\geqslant\min_{0\leqslant s\leqslant t}e^{-sg(r)}\mathbf{P}\{T\leqslant t\}.\end{aligned}$$

从而

$$\mathbf{P}\{T\leqslant t\}\leqslant\frac{e^{-ru}}{\min\limits_{0\leqslant s\leqslant t}e^{-sg(r)}}=e^{-ru}\max_{0\leqslant s\leqslant t}e^{sg(r)}. \tag{12}$$

我们现在更进一步考虑函数

$$g(r)=\lambda h(r)-cr.$$

显然, $g(0)=0$, (由于 (5) 式) $g'(0)=\lambda\mu-c<0$, 且 $g''(r)=\lambda h''(r)\geqslant 0$. 因此存在唯一正值 $r=R$, 使 $g(R)=0$.

注意, 对于 $r>0$,

$$\begin{aligned}\int_0^\infty e^{rx}[1-F(x)]dx&=\int_0^\infty\int_x^\infty e^{rx}dF(y)dx=\int_0^\infty\left(\int_0^y e^{rx}dx\right)dF(y)\\&=\frac{1}{r}\int_0^\infty(e^{ry}-1)dF(y)=\frac{1}{r}h(r).\end{aligned}$$

由此且由等式 $\lambda h(R)-cR=0$ 可见, R 的值是方程

$$\frac{\lambda}{c}\int_0^\infty e^{rx}[1-F(x)]dx=1 \tag{13}$$

(且在这种情况下是唯一) 的根.

现在, 如果在 (12) 式中设 $r=R$, 则对于每一个 $t>0$, 有

$$\mathbf{P}\{T\leqslant t\}\leqslant e^{-Ru}, \tag{14}$$

从而

$$\mathbf{P}\{T<\infty\}\leqslant e^{-Ru}, \tag{15}$$

于是, 证明了下面的定理.

定理 *假设对于克拉默 - 伦德伯格模型, 满足如下条件 A,B,C 和 $\lambda\mu<c$.*

那么, 破产概率 $\mathbf{P}\{T\leqslant t\}$ 和 $\mathbf{P}\{T<\infty\}$ 满足不等式 (14) 和 (15), 其中 R 是方程 (13) 的 (唯一) 正根.

4. 连续时间的情形 在上面进行的证明中用到 (11) 式, 如曾经指出的那样, 对于连续情形, (11) 式的正确性, 可以由 §2 定理 1 相应的类似得到 (§2 定理 1, 是关于在把时间换为随机马尔可夫时间时鞅性的不变性).(这一结果的证明, 例如, 参见专著 [41] 的 §3.2). 不过, 假如将 (条件 **A** 中的) 变量 $\sigma_i, i=1,2,\cdots$, "服从指数分布", 换成 "服从 (离散型) 几何分布: $\mathbf{P}\{\sigma_i=k\}=q^{k-1}p, k\geqslant 1$", 那么只需引用 §2 中证明的定理 1.

以上引进的要求借鉴连续时间随机过程理论的结果, 是有益的至少在于, 在应用中出现的模型, 并不是在离散时间而是在*连续时间*起作用的模型.

5. 练习题

1. 证明过程 $N=(N_t)_{t\geqslant 0}$ (在条件 **A** 下) 是独立增量过程.

2. 证明过程 $X=(X_t)_{t\geqslant 0}$ 也是独立增量过程.

3. 考虑克拉默 - 伦德伯格模型, 并且假设变量 $\sigma_i, i=1,2,\cdots$, 独立, 并且 "服从几何分布: $\mathbf{P}\{\sigma_i=k\}=q^{k-1}p, k\geqslant 1$". 试表述本节的定理.

§11. 随机金融数学的基本定理. 无仲裁的鞅特征

1. 引言 上一节介绍了鞅论用于证明保险理论的基本定理之一 —— 伦德伯格 - 克拉默定理. 这一节再考虑鞅论的一种应用 —— 鞅论用于在不确定条件下运行的 "无仲裁金融市场" 问题. 下面引进的定理 1 和定理 2, 在随机金融数学中习惯上称为制裁理论的 "基本定理". 这两个定理的重要性在于, 其用鞅的术语给出了保障所考察的金融市场无仲裁性的条件 (其含义将在下面说明), 以及保障达到确定的金融目标的条件. (关于金融数学, 详见 [100]).

2. 金融数学的某些概念 这里, 将要给出几个必要的定义.

对于讨论的所有问题, 都假设给定某个固定的概率空间 $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbf{P})$, 它在价格、金融指数以及金融市场的其他指标的演变中, 是描绘随机不确定性的基础. 这里, 我们把 $\mathscr{F}_n$ 中的全体事件视为在时间 n (包括 n) 得到的 "信息". 例如, $\mathscr{F}_n$ 可能包含关于一定金融证卷的价格、金融指数等数值的 "信息".

与 "基本定理" 有关的, 基本对象是 $(B, S)-$ 市场的概念, 其定义如下.

设 $B = (B_n)_{n\geqslant 0}$ 和 $S = (S_n)_{n\geqslant 0}$ 是正值随机变量序列, 其中假设对于每个 $n \geqslant 0$, 随机变量 B_n 是 $\mathscr{F}_{n-1}-$ 可测的 ($\mathscr{F}_{-1} = \mathscr{F}_0$), 而变量 S_n 是 $\mathscr{F}_n-$ 可测的. 为简便计, 在下面的讨论中, 假设 $\sigma-$ 代数 $\mathscr{F}_0$ 是平凡的, 即 $\mathscr{F}_0 = (\varnothing, \Omega)$ (见第二章 §2). 这样, B_0 和 S_0 都是常数. 于是, 按 §1 的术语, 两个序列 $B = (B_n)_{n\geqslant 0}$ 和 $S = (S_n)_{n\geqslant 0}$ 都是随机子序列, 并且序列 $B = (B_n)_{n\geqslant 0}$ 同时又是可预测的 (因为 B_n 是 $\mathscr{F}_{n-1}-$ 可测的).

按其金融意义, 序列 $B = (B_n)_{n\geqslant 0}$ 是描绘银行账户 ("bank account", "meney account")"单位" 演变的序列. 这时, 随机变量 B_n 是 $\mathscr{F}_{n-1}-$ 可测性表示, 在时间 n (例如, 今天) 银行账户的值, 在时间 $n-1$ (例如, 昨天) 已经是完全已知的.

如果对于 $n \geqslant 1$, 记

$$r_n = \frac{\Delta B_n}{B_{n-1}}, \tag{1}$$

其中 $\Delta B_n = B_n - B_{n-1}$, 则显然 B_n 可以表示为

$$B_n = (1 + r_n)B_{n+1}, \quad n \geqslant 1, \tag{2}$$

其中随机变量 r_n 是 $\mathscr{F}_{n-1}-$ 可测的, 并且 $r_n > -1$ (因为根据条件 $B_n > 0$). 在金融学文献中, 量 r_n 称做 (银行) 利率.

序列 $S = (S_n)_{n\geqslant 0}$ 与序列 $B = (B_n)_{n\geqslant 0}$ 的差别在于, S_n 是 $\mathscr{F}_n-$ 可测的, 而 B_n 是 $\mathscr{F}_{n-1}-$ 可测的. 实际情况正是这样, 例如, 股票 ("stock", "stocks") 价格: 对于股票在时间 n 的真实价格, 只有在其报价时 (即 "今天", 像银行账户那样而不是 "昨天").

类似于 (银行) 利率, 对于股票 $S = (S_n)_{n\geqslant 0}$, 可以引进所谓 "市场" 利率:

$$\rho_n = \frac{\Delta S_n}{S_{n-1}}, \quad n \geqslant 1. \tag{3}$$

由此明显可见, 对于一切 $\rho_n > -1$, 有

$$S_n = (1 + \rho_n)S_{n-1}, \tag{4}$$

因为 (根据假设) 所有 $S_n > 0$.

由 (2) 式和 (4) 式, 可见

$$B_n = B_0 \prod_{k=1}^{n} (1 + r_k), \tag{5}$$

$$S_n = S_0 \prod_{k=1}^{n} (1 + \rho_k). \tag{6}$$

在金融学文献中, 通常称这些公式按单利的类型形成. 在许多问题中, 按复利

$$B_n = B_0 \exp\left\{\sum_{k=1}^{n} \widehat{r}_k\right\}, \quad S_n = S_0 \exp\left\{\sum_{k=1}^{n} \widehat{\rho}_k\right\}, \tag{7}$$

的类型形成的概念也很重要, 其中

$$\widehat{r}_n = \ln(1 + r_k) = \ln\left(1 + \frac{\Delta B_k}{B_{k-1}}\right), \tag{8}$$

$$\widehat{\rho}_n = \ln(1 + \rho_k) = \ln\left(1 + \frac{\Delta S_k}{S_{k-1}}\right). \tag{9}$$

习惯上把这些变量称为 "对数利润", "归还", "偿还".

上面用描述方法引进的两过程 $B = (B_n)_{n\geqslant 0}$ 和 $S = (S_n)_{n\geqslant 0}$, 按定义构成 $(B, S)-$ 金融市场, 由银行账户 B 和股票 S 两种资产组成.

注 显然, 因为现实金融市场通常包含大量不同性质的资产 (例如, 见 [100]), 这样的 $(B, S)-$ 市场仅仅是现实金融市场最简单的模型. 然而, 就是在这种简单的情形下, 在研究许多纯金融 - 经济来源的问题时, 已经可以仔细研究并且作为例子说明, 鞅论方法的有效性. (例如, 关于在 $(B, S)-$ 市场上无仲裁可能性的问题, 就属于这种情形. 而关于这一问题, 将在下面引进的 "第一基本定理" 中给出答案).

3. 有价证卷总存和自筹资金总存 现在给出有价证券的总存量及其资本的定义, 以及引进自筹资金总存量的重要概念.

设 $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbf{P})$ 是所考察的过滤概率空间, 其中 $\mathscr{F}_0 = (\varnothing, \Omega)$; 而 $\pi = (\beta, \gamma)$ 是可预测序列偶: $\beta = (\beta_n)_{n\geqslant 0}$和 $\gamma = (\gamma_n)_{n\geqslant 0}$.

除变量 β_n 和 $\gamma_n, n \geqslant 0$ 的可预测性, 即 β_n 和 γ_n 为 $\mathscr{F}_{n-1}-$ 可测的条件之外, 其中 ($\mathscr{F}_{-1} = \mathscr{F}_0$), 对于 β_n 和 $\gamma_n (n \geqslant 0)$ 的可能值不再加任何限制. 特别, 这些变量可以取分数和负数为值.

随机变量 β_n 和 γ_n 的含义相应为, 在时间 n 银行账户的 "个数" 和股票的 "张数".

我们称在所考察的 (B,S)– 市场上, $\pi=(\beta,\gamma)$ 是有价证券的总存量.

与每一总存量 $\pi=(\beta,\gamma)$, 有与其相应的资本 $X^\pi=(X_n^\pi)_{n\geqslant 0}$ 相联系, 其中

$$X_n^\pi=\beta_n B_n+\gamma_n S_n, \tag{10}$$

把 $\beta_n B_n$ 解释为银行账户上的货币资金, 而 $\gamma_n S_n$ 是时间 n 股票的价值. 序列 β 和 γ 的可预测性也是清楚的:"明天" 有价证卷的总存量, 应该在 "今天" 编制.

特别, 下一个重要的概念 —— "自筹资金" 总量的概念, 反映分析这样 (B,S)–市场: 在这样的市场上 "无论资金流出, 还是自外部流入" 都不存在. 从形式的观点看, 相应的定义是由下面的方式给出的.

由 "离散微分公式" $\Delta(a_nb_n)=a_n\Delta b_n+b_{n-1}\Delta a_n$, 可见资金的增量 ΔX_n^π $(=X_n^\pi-X_{n-1}^\pi)$ 可以表示为

$$\Delta X_n^\pi=[\beta_n\Delta B_n+\gamma_n\Delta S_n]+[B_{n-1}\Delta\beta_n+S_{n-1}\Delta\gamma_n]. \tag{11}$$

资金的实际的变化只与银行账户和股票价格值的 "市场" 变化有关, 即与量 $\beta_n\Delta B_n+\gamma_n\Delta S_n$ 的变化有关. (11) 式右侧的第二项, 即 $B_{n-1}\Delta\beta_n+S_{n-1}\Delta\gamma_n$ 是 $\mathscr{F}_{n-1}$– 可测变量, 并且在时间 n 变量 X_{n-1}^π 任何增加或减少都不可能. 于是它应该等于 0.

原则上, 资本的可能发生变化, 不仅因为利率 (r_n 和 $\rho_n, n\geqslant 1$) 的 "市场" 变化, 例如还可能因为资本从外部流入, 由于业务费用的支出, 等等.

以后, 这些可能性都不考虑, 假设全部考察的总存 $\pi=(\beta,\gamma)$ 满足条件: 对于一切 $n\geqslant 1$,

$$\Delta X_n^\pi=\beta_n\Delta B_n+\gamma_n\Delta S_n. \tag{12}$$

在随机金融数学中, 通常把这样的总存称为*自筹的* (self-financing).

4. 随机金融数学的第一基本定理 由 (12) 式可见, 对于自筹总存量 $\pi=(\beta,\gamma)$,

$$X_n^\pi=X_0^\pi+\sum_{k=1}^n(\beta_k\Delta B_k+\gamma_k\Delta S_k), \tag{13}$$

且由于

$$\Delta\left(\frac{X_n^\pi}{B_n}\right)=\gamma_n\Delta\left(\frac{S_n}{B_n}\right), \tag{14}$$

则

$$\frac{X_n^\pi}{B_n}=\frac{X_0^\pi}{B_0}+\sum_{k=1}^n\gamma_k\Delta\left(\frac{S_k}{B_k}\right). \tag{15}$$

固定某个 $N\geqslant 1$, 并考察 (B,S)– 市场在时间 $n=0,1,\cdots,N$ 的变化.

定义 1 称在时间 N 的自筹资金总存量 (或自筹资金策略) $\pi=(\beta,\gamma)$, *实施仲裁, 或实施仲裁的可能性*, 如果 $X_0^\pi=0, X_N^\pi\geqslant 0$ ($\mathbf{P}$ – a.c.), 且以大于 0 的 $\mathbf{P}$– 概率, 有 $X_N^\pi>0$, 即 $\mathbf{P}\{X_N^\pi>0\}>0$.

定义 2 称 (在时间 N) 在 (B,S)– 市场上*无仲裁或无仲裁可能性*, 如果对于 $X_0^\pi=0$ 和 $\mathbf{P}\{X_N^\pi\geqslant 0\}=1$ 的任何总存量 $\pi=(\beta,\gamma)$, 实际上 $\mathbf{P}\{X_N^\pi=0\}=1$, 即仅以等于 0 的 $\mathbf{P}$ – 概率可能使 $X_N^\pi>0$.

直观上, 在无仲裁市场上不可能出现如下情况: 对于某一总存量存在得到无风险收入的可能性.

显然, 关于某一 (B,S)– 市场是否无仲裁, 因而在一定意义上是 "正确的"、"合理的" 解的问题, 与序列 $B=(B_n)_{n\geqslant 0}$ 和 $S=(S_n)_{n\geqslant 0}$ 的概率 – 统计性质有关, 从而与包含在过滤概率空间 $(\Omega,\mathscr{F},(\mathscr{F}_n)_{n\leqslant N},\mathbf{P})$ 上结构中的假设条件有关.

值得注意的是, 鞅论可以相当有效地描绘保障无仲裁可能性的条件. 可以得到的甚至更多. 具体地说, 有下面的定理.

定理 1 ("第一基本定理") *假设过滤概率空间 $(\Omega,\mathscr{F},(\mathscr{F}_n)_{n\leqslant N},\mathbf{P})$ 描绘随机不确定性, 其中 $\mathscr{F}_0=(\varnothing,\Omega),\mathscr{F}_N=\mathscr{F}$.*

为使定义在 $(\Omega,\mathscr{F},(\mathscr{F}_n)_{n\leqslant N},\mathbf{P})$ 上的 (B,S)– 市场是无仲裁的, 必要且充分条件是, 在 $(\Omega,\mathscr{F})$ 上存在与测度 $\mathbf{P}$ 等价的这样一个测度 $\widetilde{\mathbf{P}}(\widetilde{\mathbf{P}}\sim\mathbf{P})$, 使贴现序列

$$\frac{S}{B}=\left(\frac{S_n}{B_n}\right)_{n\leqslant N}$$

关于测度 $\widetilde{\mathbf{P}}$ 构成鞅:

$$\widetilde{\mathbf{E}}\left|\frac{S_n}{B_n}\right|<\infty,\quad n\leqslant N,$$

且

$$\widetilde{\mathbf{E}}\left(\frac{S_n}{B_n}\middle|\mathscr{F}_{n-1}\right)=\frac{S_{n-1}}{B_{n-1}},\quad n\leqslant N,$$

其中 $\widetilde{\mathbf{E}}$ 表示关于测度 $\widetilde{\mathbf{P}}$ 的数学期望.

注 1 对于向量过程 $S=(S^1,\cdots,S^d), d<\infty$, 定理的结论仍然成立 (见 [100] 中第 V 章 §2b).

注 2 根据完全明显的原因, 定理中的测度 $\widetilde{\mathbf{P}}$ 通常称做*鞅测度*.

我们以

$$\mathbf{M}(\mathbf{P})=\left\{\widetilde{\mathbf{P}}\sim\mathbf{P}:\frac{S}{B}\text{ 是 }\widetilde{\mathbf{P}}\text{– 鞅}\right\}$$

表示测度 $\widetilde{\mathbf{P}}$ 的全体: 测度 $\widetilde{\mathbf{P}}$ 与测度 $\mathbf{P}$ 等价且序列

$$\frac{S}{B}=\left(\frac{S_n}{B_n}\right)_{n\leqslant N}$$

关于测度 $\widetilde{\mathbf{P}}$ 是鞅.

以记号 **NA** 表示*无仲裁* (No Arbitrage).

那么, 定理 1 的结论可以表示为:

$$\mathbf{NA}\Leftrightarrow\mathbf{M}(\mathbf{P})\neq\varnothing. \tag{16}$$

证明 *充分性* 设 $\widetilde{\mathbf{P}}$ 是 $\mathbf{M}(\mathbf{P})$ 中的鞅测度, $\pi=(\beta,\gamma)$ 是总存量, 且 $X_0^\pi=\beta_0B_0+\gamma_0S_0=0$. 由 (15) 式可见, 对于 $1\leqslant n\leqslant N$, 有

$$\frac{X_n^\pi}{B_n}=\sum_{k=1}^n\gamma_k\Delta\left(\frac{S_k}{B_k}\right). \tag{17}$$

序列

$$\frac{S}{B}=\left(\frac{S_n}{B_n}\right)_{n\leqslant N}$$

关于测度 $\widetilde{\mathbf{P}}$ 是鞅. 因此, 序列 $G=(G_n^\pi)_{0\leqslant n\leqslant N}(G_0^\pi=0)$, 和

$$G_n^\pi=\sum_{k=1}^n\gamma_k\Delta\left(\frac{S_k}{B_k}\right),\quad 1\leqslant n\leqslant N,$$

是鞅变换. 因此序列 $(X_n^\pi/B_n)_{0\leqslant n\leqslant N}$ 也是鞅变换.

在仲裁测试或无仲裁的情况下, 不仅需要考虑 $X_0^\pi=0$ 的总存量 π 的情形, 而且需要考虑 $X_0^\pi\geqslant 0$ ($\mathbf{P}$–a.c.) 的总存量 π 的情形. 由于 $\widetilde{\mathbf{P}}\sim\mathbf{P}$, 且 $B_N>0$ ($\mathbf{P}$ – 和 $\widetilde{\mathbf{P}}$–a.c.), 可得

$$\widetilde{\mathbf{P}}\left\{\frac{X_N^\pi}{B_N}\geqslant 0\right\}=1.$$

那么, 将 §1 的定理 3 用于鞅变换 $(X_n^\pi/B_n)_{0\leqslant n\leqslant N}$ 可见, 该序列实际上关于测度 $\widetilde{\mathbf{P}}$ 是鞅. 从而

$$\widetilde{\mathbf{E}}\frac{X_N^\pi}{B_N}=\widetilde{\mathbf{E}}\frac{X_0^\pi}{B_0}=0,$$

而由于

$$\widetilde{\mathbf{P}}\left\{\frac{X_N^\pi}{B_N}\geqslant 0\right\}=1,\quad 故\ \widetilde{\mathbf{P}}\left\{\frac{X_N^\pi}{B_N}=0\right\}=1.$$

由此可见, $X_N^\pi=0$ ($\widetilde{\mathbf{P}}$ – 和 $\mathbf{P}$–a.c.). 从而对于任意自筹总存量 π, 其中 $X_0^\pi=0, X_N^\pi\geqslant 0$ ($\mathbf{P}$–a.c.), 实际上 $X_N^\pi=0$ ($\mathbf{P}$–a. c.). 于是, 根据定义 2 无仲裁可能性.

必要性 我们之准备仅对于一阶段模型 (B,S)– 市场进行证明, 即只证明 $N=1$ 的情形. 即便通过这个简单的例子, 就已经清楚地看出证明的好思路: 利用无仲裁, 明显地建立无论是什么样的鞅测度. 我们将基于 (下面将引进的) *埃舍* (Esher) *变换*, 建立这样的测度.(关于的一般情形 ($N\geqslant 1$) 的证明, [100] 中第 V 章 §2b).

不失普遍性可认为 $B_0=B_1=1$. 这里, 由无仲裁可能性的假设可见 (练习题 1)

$$\mathbf{P}\{\Delta S_1>0\}>0\quad 和\quad \mathbf{P}\{\Delta S_1<0\}>0. \tag{18}$$

(我们排除了平凡的情形 $\mathbf{P}\{\Delta S_1=0\}=1$).

由此需要证明, 存在满足如下条件的等价鞅测度 $\widetilde{\mathbf{P}}$ 即 $\widetilde{\mathbf{P}}\sim\mathbf{P}$, 且 $\widetilde{\mathbf{E}}|\Delta S_1|<\infty, \widetilde{\mathbf{E}}|\Delta S_1|=0$.

这由下面的引理即可得到. 不过该引理尚有一般概率的应用.

引理 1 *设 $(\Omega,\mathscr{F})=(\mathbb{R},\mathscr{B}(\mathbb{R}))$, 而 $X=X(\omega)$ 是由坐标给定的随机变量 $(X(\omega)=\omega)$; $\mathbf{P}$ 是 $(\Omega,\mathscr{F})$ 上的概率测度:*

$$\mathbf{P}\{X>0\}>0,\quad 和\quad \mathbf{P}\{X<0\}>0. \tag{19}$$

那么, 在 $(\Omega,\mathscr{F})$ 上存在概率测度 $\widetilde{\mathbf{P}}\sim\mathbf{P}$, 使对于任意实数 a, 有

$$\widetilde{\mathbf{E}}\mathrm{e}^{aX}<\infty. \tag{20}$$

特别, $\widetilde{\mathbf{E}}|X|<\infty$, 并且

$$\widetilde{\mathbf{E}}X=0. \tag{21}$$

证明 引进测度 $\mathbf{Q}=\mathbf{Q}(\mathrm{d}x)$, 而 $\mathbf{Q}(\mathrm{d}x)=c\mathrm{e}^{-x^2}\mathbf{P}(\mathrm{d}x)$, 其中 $c=(\mathbf{E}\mathrm{e}^{-X^2})^{-1}$ 规范化常数.

对于任意实数 a, 设

$$\varphi(a)=\mathbf{E}_{\mathbf{Q}}\mathrm{e}^{aX}. \tag{22}$$

其中 $\mathbf{E}_{\mathbf{Q}}$ 表示对于测度 $\mathbf{Q}$ 的数学期望.

设

$$Z_a(x)=\frac{\mathrm{e}^{aX}}{\varphi(a)}. \tag{23}$$

由于 $Z_a(x)>0$, 且 $\mathbf{E}_{\mathbf{Q}}Z_a(X)=1$, 则对于任意实数 a, 测度 $\widetilde{\mathbf{P}}_a$:

$$\widetilde{\mathbf{P}}_a(\mathrm{d}x)=Z_a(x)\mathbf{Q}_a(\mathrm{d}x) \tag{24}$$

是概率测度. 显然 $\widetilde{\mathbf{P}}_a\sim\mathbf{Q}\sim\mathbf{P}$.

注 3 变换 $x\rightsquigarrow\mathrm{e}^{aX}/\varphi(a)$ 常称做*埃舍变换*. 在下面将要看到, 对于 a 的某个特殊值 a_*, 测度 $\widetilde{\mathbf{P}}=\widetilde{\mathbf{P}}_{a_*}$ 具有 (鞅) 性质 (21). 正是这个测度通常称为*埃舍测度* (或*埃舍鞅测度*).

由于 $\varphi''(a)>0$, 可见对于一切实数 a 定义的函数 $\varphi=\varphi(a)$ 是严格凹 (即向下凸) 的.

设 $\varphi_*=\inf\{\varphi(a):a\in\mathbb{R}\}$. 有两种可能的情形: 1) 存在 a_*, 使 $\varphi(a_*)=\varphi_*$; 2) 这样的有限 a_* 不存在.

对于第一种情形, $\varphi'(a_*)=0$. 因此

$$\mathbf{E}_{\widetilde{\mathbf{P}}_{a_*}}X=\mathbf{E}_{\mathbf{Q}}\frac{X\mathrm{e}^{a_*X}}{\varphi(a_*)}=\frac{\varphi'(a_*)}{\varphi(a_*)}=0,$$

且可以取测度 $\mathbf{P}_{a^*}$ 做所要求的测度 $\widetilde{\mathbf{P}}$.

到现在为止, 我们还没有用到条件 "无仲裁性" (19). 不难看到 (练习题 2), "无仲裁性" 条件排除可能性 2). 这样只剩下可能性 1), 而这种情形已经讨论过.

于是, 当 $N=1$ 时 (存在鞅测度的) "必要性" 得证. 关于的一般情形 $N\geqslant 1$, 我们已经给读者指出, 可以借鉴 [100] 中第 V 章 §2d 的内容. □

5. 例 下面举一些无仲裁 (B,S)– 市场的例.

例 1 假设 (B,S)– 市场由 (5) 式和 (6) 式描绘, 其中 $1\leqslant k\leqslant N$, 且对于一切 $1\leqslant k\leqslant N, r_k=r$ (常数); 而 $\rho=(\rho_1,\rho_2,\cdots,\rho_N)$, 是独立同分布、只有 a 和 $b(a<b)$ 两个可能值的 (伯努利) 随机变量序列: $\mathbf{P}\{\rho_1=a\}=q,\mathbf{P}\{\rho_1=b\}=p,p+q=1,0<p<1$. 设, 这时

$$-1<a<r<b. \tag{25}$$

这样描绘的 (B,S)– 市场称做 CRR – 模型, 其中是模型的作者姓氏的字头: 考克斯 (J. C. Cox), 罗斯 (R. A. Ross), 鲁宾斯坦 (M. Rubinstein), (详见 [100]).

由于在该模型中

$$\frac{S_n}{B_n}=\left(\frac{1+\rho_n}{1+r}\right)\frac{S_{n-1}}{B_{n-1}},$$

则显然鞅测度 $\widetilde{\mathbf{P}}$ 应当满足

$$\widetilde{\mathbf{E}}\frac{1+\rho_n}{1+r}=1,$$

即应当满足 $\widetilde{\mathbf{E}}\rho_n=r$.

如果记 $\widetilde{p}=\widetilde{\mathbf{P}}\{\rho_n=b\},\widetilde{q}=\widetilde{\mathbf{P}}\{\rho_n=a\}$, 则对于任意 $n\geqslant 1$, 有

$$\widetilde{p}+\widetilde{q}=1,\quad b\widetilde{p}+a\widetilde{q}=r.$$

由此, 得

$$\widetilde{p}=\frac{r-a}{b-a},\quad \widetilde{q}=\frac{b-r}{b-a}. \tag{26}$$

在上面的情形下, 全部 "随机性" 决定于伯努利序列 $\rho=(\rho_1,\rho_2,\cdots,\rho_N)$. 假设 $\Omega=\{a,b\}^N$, 即基本事件空间由序列 $(x_1,\cdots,x_N)$ 构成, 其中 $x_i=a$ 或 b. (关于空间 Ω 构造的这一特别的, (确切一点说) "坐标的" 构造的假设, 并不失研究的普遍性; 关于这一点, 亦见第 6 小节定理 2 中充分性证明的末尾).

作为练习 (练习题 3), 需要证明, 由

$$\widetilde{\mathbf{P}}(x_1,\cdots,x_N)=\widetilde{p}^{\nu_b(x_1,\cdots,x_N)}\widetilde{q}^{N-\nu_b(x_1,\cdots,x_N)}, \tag{27}$$

定义的测度 $\widetilde{\mathbf{P}}(x_1,\cdots,x_N)$ 是鞅测度, 并且是唯一的, 其中

$$\nu_b(x_1,\cdots,x_N)=\sum_{i=1}^N I_b(x_i)$$

是等于 b 的 x_i 的个数.

由 (27) 式明显可见 $\widetilde{\mathbf{P}}\{\rho_n=a\}=\widetilde{q},\mathbf{P}\{\rho_n=b\}=\widetilde{p}$. 这样, 由定理 1 可见, CRR – 模型是无仲裁 (B,S)– 市场的例.

例 2 假设 (B,S)– 市场有如下构造: 对于一切 $n=0,1,\cdots,N$, 有 $B_n=1$, 而

$$S_n=S_0\exp\left\{\sum_{k=1}^n\widehat{\rho}_k\right\},\quad 1\leqslant n\leqslant N. \tag{28}$$

设 $\widehat{\rho}=\mu_k+\sigma_k\varepsilon_k$ 为 $\mathscr{F}_{n-1}$– 可测, 其中 $\mu_k>0,\sigma_k>0$, 而 $(\varepsilon_1,\cdots\varepsilon_N)$ 是独立平稳高斯随机变量序列, $\varepsilon_k\sim N(0,1)$.

我们现在利用埃舍条件变换, 在 $(\Omega,\mathscr{F}_N)$ 上建立所要求的鞅测度 $\widetilde{\mathbf{P}}$. 具体地说, 设 $\widetilde{\mathbf{P}}(\mathrm{d}\omega)=Z_N(\omega)\mathbf{P}(\mathrm{d}\omega)$, 其中

$$Z_N(\omega)=\prod_{1\leqslant k\leqslant N}z_k(\omega),\quad 而\quad z_k(\omega)=\frac{\mathrm{e}^{a_k\widehat{\rho}_k}}{\mathbf{E}(\mathrm{e}^{a_k\widehat{\rho}_k}|\mathscr{F}_{k-1})}; \tag{29}$$

现在, 应当这样选择 $\mathscr{F}_{n-1}$– 可测随机变量 $a_k=a_k(\omega)$, 使得序列 $(S_n)_{0\leqslant n\leqslant N}$ 关于测度 $\widetilde{\mathbf{P}}$ 是鞅, 其中 $\mathscr{F}_0=(\varnothing,\Omega)$.

由于表现 (28), 关于测度 $\widetilde{\mathbf{P}}$ 的鞅性等价于: 对于一切 $1\leqslant n\leqslant N$, (关于原测度 $\mathbf{P}$) 有

$$\mathbf{E}[\mathrm{e}^{(a_n+1)\widehat{\rho}_n}|\mathscr{F}_{n-1}]=\mathbf{E}[\mathrm{e}^{a_n\widehat{\rho}_n}|\mathscr{F}_{n-1}]. \tag{30}$$

由于 $\widehat{\rho}_n=\mu_n+\sigma_n\varepsilon_n$, 故由 (30) 式可见, 应这样选择变量 a_n, 使

$$\mu_n+\frac{\sigma_n^2}{2}=-a_n\sigma_n^2,$$

即

$$a_n=-\frac{\mu_n}{\sigma_n^2}-\frac{1}{2}.$$

对于这样选择变量 $a_n(1\leqslant n\leqslant N)$, 密度 $Z_N(\omega)$ 由下面的公式表示:

$$Z_N(\omega)=\exp\left\{-\sum_{n=1}^N\left[\left(\frac{\mu_n}{\sigma_n}+\frac{\sigma_n}{2}\right)\varepsilon_n+\frac{1}{2}\left(\frac{\mu_n}{\sigma_n}+\frac{\sigma_n}{2}\right)^2\right]\right\}. \tag{31}$$

如果一开始对于 $1\leqslant n\leqslant N$, 设 $\mu_n=-\sigma_n^2/2$, 则 $\widetilde{\mathbf{P}}=\mathbf{P}$. 换句话说, 原来的测度 $\mathbf{P}$ 本身就是鞅.

于是, 对于所分析的 (B,S) – 市场 $(B=(B_n)_{0\leqslant n\leqslant N})$ 有: $B_n\equiv 1$, 而 $S=(S_n)_{0\leqslant n\leqslant N}$ 同例 1 一样是无仲裁的, 其中 S_n 由 (28) 式表示. 作为习题 (练习题 4), 要求分析如下问题: 上面建立的鞅测度 $\widetilde{\mathbf{P}}$ 是否唯一.

6. 随机金融数学的第二基本定理 下面将要引进的 (B,S) – 市场完全性的概念, 对于随机金融数学十分重要, 因为 (无论所考察的市场是无仲裁的, 还是有仲裁的) 它与如下很自然的问题有关: 假如对于给定的 $\mathscr{F}_N$– 可测 "自筹资金委托" f_N, 存在自筹资金总存量 π, 使其资本 X_N^π 准确地再现 (或者至少不小于) f_N.

定义 3 称 (B,S)– 市场 (关于时间 N) 为完全的, 或 N– 完全的, 如果任意有界 $\mathscr{F}_N$– 可测 "自筹资金委托" f_N 是可以再现的, 即存在自筹资金总存量 π, 使 $X_N^\pi = f_N$ ($\mathbf{P}$-a.c.).

定理 2 (第二基本定理) 同定理 1 一样, 假设 $(\Omega,\mathscr{F},(\mathscr{F}_n)_{0\leqslant n\leqslant N},\mathbf{P})$ 是过滤概率空间, $\mathscr{F}_0=(\varnothing,\Omega),\mathscr{F}_N=\mathscr{F}$; 在 $(\Omega,\mathscr{F},(\mathscr{F}_n)_{n\leqslant N},\mathbf{P})$ 上给定的 (B,S)– 市场是无仲裁的 $(\mathbf{M}(\mathrm{P})\neq\varnothing)$.

为使此市场是完全的, 必要和充分条件是存在唯一鞅测度 $(|\mathbf{M}(\mathrm{P})|=1)$.

证明 必要性 假设所考虑的市场是完全的, 即对于任意 $\mathscr{F}_N$– 可测有界 "自筹资金委托" f_N, 存在这样的自筹资金总存量 $\pi=(\beta,\gamma)$, 使 $X_N^\pi=f_N$ ($\mathbf{P}$-a.c.). 不失普遍性可以认为 $B_n=1, 0\leqslant n\leqslant N$. 因而, 由 (13) 式可见

$$f_N=X_N^\pi=X_0^\pi+\sum_{k=1}^N\gamma_k\Delta S_k. \tag{32}$$

由于所作的无仲裁性假设, 鞅测度的集合 $\mathbf{M}(\mathrm{P})\neq\varnothing$. 现在证明, 由于完全性假设, 可见鞅测度的唯一性 $(|\mathbf{M}(\mathrm{P})|=1)$.

设 $\mathbf{P}^1$ 和 $\mathbf{P}^2$ 是两个鞅测度, 则关于其中任何一个测度, 序列

$$\left(\sum_{k=1}^n\gamma_k\Delta S_k\right)_{1\leqslant n\leqslant N}$$

是鞅变换.

对于某个集合 $A\in\mathscr{F}_N$, 设 $f_N(\omega)=I_A(\omega)$. 由于 ($\mathbf{P}$-a.c.) 对于某个 π, 有

$$I_A(\omega)=X_N^\pi=X_0^\pi+\sum_{k=1}^N\gamma_k\Delta S_k,$$

则由 §1 定理 3 可见, 序列

$$\left(\sum_{k=1}^n\gamma_k\Delta S_k\right)_{1\leqslant n\leqslant N}$$

是关于每一个测度 $\mathbf{P}^1$ 和 $\mathbf{P}^2$ 鞅. 从而

$$\mathbf{E}_{\mathbf{P}^i}I_A(\omega)=x,\quad i=1,2, \tag{33}$$

其中 $\mathbf{E}_{\mathbf{P}^i}$ 表示对测度 $\mathbf{P}^i$ 的数学期望, 而因为 $\mathscr{F}_0=(\varnothing,\Omega), x=X_0^\pi$ 是常数.

由 (33) 式可见, 对于任意集合 $A\in\mathscr{F}_N$, 有 $\mathbf{P}^1(A)=\mathbf{P}^2(A)$. 从而, 鞅测度的唯一性得证.

充分性 充分性的证明比较复杂, 我们将分为几阶段进行.

1) 考虑无仲裁 (B,S)– 市场 $(\mathbf{M}(\mathrm{P})\neq\varnothing)$, 同时假设鞅测度具有唯一性 $(|\mathbf{M}(\mathrm{P})|=1)$.

注意, 关于鞅测度唯一的假设, 以及关于鞅测度完全性的假设, 都是很强的限制. 此外, 结果表明, 由这些假设自然地可见, 轨道 $S=(S_n)_{0\leqslant n\leqslant N}$ 具有 "条件两点" 的构造. 关于这一点将在下面证明. (具体的例子: 设 $\Delta S_n=\rho_nS_{n-1}$ 为 CRR – 模型, 其中 ρ_n 只有两个可能值, 因此条件概率 $\mathbf{P}(\Delta S_n\in\cdot|\mathscr{F}_{n-1})$ 仅集中在 aS_{n-1} 和 bS_{n-1} 两个点上).

鞅测度的唯一性 $(|\mathbf{M}(\mathrm{P})|=1)$ 也是加在滤波 $(\mathscr{F}_n)_{n\leqslant N}$ 构造上的限制. 原来, σ– 代数 $\mathscr{F}_n$ 自然应该是价格 $S_0,S_1,\cdots,S_n$ 产生的 σ– 代数 (只是现在假设 $B_k\equiv 1, k\leqslant n$). 关于这一点见 [100] 第 610 页的图形, 以及 [100] 第五章的 §4e.

2) 作为证明蕴涵关系 "$|\mathbf{M}(\mathrm{P})|=1\Rightarrow$ 完全性" 路径的中间结果之一, 我们证明如下的重要命题, 它给出了无仲裁市场上完全性的等价特征.

引理 2 为了使无仲裁 (B,S)– 市场是完全的, 必要而且充分的条件是: 在所有鞅测度的集合 $\mathbf{M}(\mathbf{P})$ 上, 存在具有如下性质的测度 $\widetilde{\mathbf{P}}$: 任何有界鞅 $m=(m_n,\mathscr{F}_n,\widetilde{\mathbf{P}})_{0\leqslant n\leqslant N}$, 有 "$(S/B)$– 表现"

$$m_n=m_0+\sum_{k=1}^n\gamma_k^*\Delta\left(\frac{S_k}{B_k}\right), \tag{34}$$

其中 $\gamma_k^*(1\leqslant k\leqslant n)$ 是某一可预测随机变量.

证明引理 2 a) 必要性 设所考察的 (B,S)– 市场是无仲裁的和完全的. (不失普遍性, 可以认为 $B_n=1, 0\leqslant n\leqslant N$).

取 $\mathbf{M}(\mathbf{P})$ 中的任意测度 $\widetilde{\mathbf{P}}$, 并且设 $m=(m_n,\mathscr{F}_n,\widetilde{\mathbf{P}})_{0\leqslant n\leqslant N}$ 是某一有限鞅 $(|m_n|\leqslant c, 0\leqslant n\leqslant N)$. 记 $f_N=m_N$. 那么, 根据完全性的定义 (见定义 3), 存在这样的总存量 $\pi^*=(\beta^*,\gamma^*)$, 使 $X_N^{\pi^*}=f_N$, 且对于一切 $0\leqslant n\leqslant N$, 有

$$X_N^{\pi^*}=x+\sum_{k=1}^n\gamma_k^*\Delta S_k, \tag{35}$$

其中 $x=X_0^{\pi^*}$.

由于 $X_N^{\pi^*}=f_N\leqslant c$, 可见序列 $X^{\pi^*}=(X_n^{\pi^*},\mathscr{F}_n,\widetilde{\mathbf{P}})_{0\leqslant n\leqslant N}$ 是鞅 (见 §1 定理 3). 这样, 就有具有同一最终值 $f_N(X_N^{\pi^*}=m_N=f_N)$ 的两个鞅 m 和 X^{π^*}. 但是, 根据鞅性质的定义 $m_n=\mathbf{E}(m_N|\mathscr{F}_n)$ 和 $X_n^{\pi^*}=\mathbf{E}(X_N^{\pi^*}|\mathscr{F}_n), 0\leqslant n\leqslant N$. 因此, 列维 (P. P. Lévy) 鞅 m 和 X^{π^*} 相等. 从而, 由 (35) 式知, 对于鞅 $m=(m_n,\mathscr{F}_n,\widetilde{\mathbf{P}})_{0\leqslant n\leqslant N}$, 有 "$S$–表现":

$$m_n=x+\sum_{k=1}^n\gamma_k^*\Delta S_k,\quad 1\leqslant n\leqslant N, \tag{36}$$

其中 $x=m_0$.

b) 充分性 现在证明相反的结果 ("S–表现" $\Rightarrow$ 完全性).

根据假设的条件, 存在测度 $\widetilde{\mathbf{P}}\in\mathbf{M}(\mathbf{P})$, 使任何有限 $\widetilde{\mathbf{P}}$– 鞅有 "S– 表现".

作为这样的鞅 $X=(X_n,\mathscr{F}_n,\widetilde{\mathbf{P}})_{0\leqslant n\leqslant N}$, 取 $X_n=\widetilde{\mathbf{E}}(f_N|\mathscr{F}_n)$ 的鞅, 其中 $\widetilde{\mathbf{E}}$ 表示对于测度 $\widetilde{\mathbf{P}}$ 的数学期望, 而 f_N 是定义 3 中提到的 "支付委托", 且对于 f_N 需要求自筹资金总量 $\pi: X_N^{\pi}=f_N(\widetilde{\mathbf{P}}$ - 和 $\mathbf{P}$-a.c.).

对于有界鞅 $X=(X_n,\mathscr{F}_n,\mathbf{P})_{0\leqslant n\leqslant N}$, 考虑其 "$S-$ 表现":

$$X_n=X_0+\sum_{k=1}^{n}\gamma_k\Delta S_k, \tag{37}$$

其中 γ_k 是某 $\mathscr{F}_{k-1}-$ 可测随机变量.

现在证明, 由此可见, 存在自筹资金总存量 $\widetilde{\pi}=(\widetilde{\beta},\widetilde{\gamma})$, 使得对于一切 $0\leqslant n\leqslant N$, 特别对于 $f_N=X_N=X_N^{\widetilde{\pi}}$, 有定义 3 所要求的表现:

$$f_N=X_0^{\widetilde{\pi}}+\sum_{k=1}^{n}\widetilde{\gamma}_k\Delta S_k. \tag{38}$$

有了 (37) 式, 现在设 $\widetilde{\gamma}_k=\gamma_k$, 并且定义

$$\widetilde{\beta}_n=X_n-\gamma_nS_n. \tag{39}$$

由 (37) 式可见, 随机变量 $\widetilde{\beta}_n$ 是 $\mathscr{F}_{n-1}-$ 可测的. 这时, 有

$$\begin{aligned}S_{n-1}\Delta\widetilde{\gamma}_n+\Delta\widetilde{\beta}_n&=S_{n-1}\Delta\gamma_n+\Delta X_n-\Delta(\gamma_nS_n)\\&=S_{n-1}\Delta\gamma_n+\gamma_n\Delta S_n-\Delta(\gamma_nS_n)=0.\end{aligned}$$

这样, 根据第三小节, 所建立的总存量 $\widetilde{\pi}=(\widetilde{\beta},\widetilde{\gamma})$ 是自筹的, 并且 $X_N^{\widetilde{\pi}}=f_N$, 即*无仲裁 $(B,S)-$ 市场是完全的得证.*

于是, 引理 3 得证.□

3) 由引理 3 可见, 为完成定理 2 的证明, 需要证明如下一系列蕴涵关系:

$$\boxed{|\mathbf{M}(\mathbf{P})|=1}\overset{\{3\}}{\Longrightarrow}\boxed{S\text{ 表现}}\overset{\{2\}}{\Longleftrightarrow}\boxed{\text{完全性}}\overset{\{1\}}{\Longrightarrow}\boxed{|\mathbf{M}(\mathbf{P})|=1}.$$

在证明引理 2 的 "必要性" 时, "蕴涵关系 {1} " 已经得证; "蕴涵关系 {2}" 就是引理 2.

为使蕴涵关系 {3} 的证明更加清晰, 考虑由 CRR - 模型描绘的 $(B,S)-$ 市场的特殊情形.

上面 (例 1) 曾指出, 对于这样的模型, 鞅测度 $\widetilde{\mathbf{P}}$ 是唯一的 $(|\mathbf{M}(\mathbf{P})|=1)$. 因此, 需要弄清楚, 这里为什么 (关于鞅测度 $\widetilde{\mathbf{P}}$) "*$S-$ 表现*" 成立. 原来在上面已经指出, 关键是 (4) 式中的变量 ρ_n 只有 a 和 b 两个可能值, 而作为这一事实的推论, 条件分布 $\mathbf{P}\{\Delta S\in\cdot|\mathscr{F}_{n-1}\}$ 全部仅集中在两个点上 ("条件两点").

这样, 考虑例 1 中引进的 CRR - 模型, 并且补充假设对于 $1\leqslant n\leqslant N,\mathscr{F}_n=\sigma(\rho_1,\cdots,\rho_n)$, 而 $\mathscr{F}_0=(\varnothing,\Omega)$. 以 $\widetilde{\mathbf{P}}$ 表示由 (27) 式定义在 $(\Omega,\mathscr{F}_N)$ 上的鞅测度.

设 $X=(X_n,\mathscr{F}_n,\widetilde{\mathbf{P}})_{0\leqslant n\leqslant N}$ 是有界鞅, 则存在函数 $g_n=g_n(x_1,\cdots,x_n)$, 使 $X_n(\omega)=g_n(\rho_1(\omega),\cdots,\rho_n(\omega))$, 因此

$$\Delta X_n=g_n(\rho_1,\cdots,\rho_n)-g_{n-1}(\rho_1,\cdots,\rho_{n-1}).$$

由于 $\widetilde{\mathbf{E}}(\Delta X_n|\mathscr{F}_{n-1})=0$, 可见

$$\widetilde{p}g_n(\rho_1,\cdots,\rho_{n-1},b)-\widetilde{q}g_n(\rho_1,\cdots,\rho_{n-1},a)=g_{n-1}(\rho_1,\cdots,\rho_{n-1}),$$

即

$$\begin{aligned}&\frac{g_n(\rho_1,\cdots,\rho_{n-1},b)-g_{n-1}(\rho_1,\cdots,\rho_{n-1})}{\widetilde{q}}\\&=\frac{g_{n-1}(\rho_1,\cdots,\rho_{n-1})-g_n(\rho_1,\cdots,\rho_{n-1},a)}{\widetilde{p}}.\end{aligned} \tag{40}$$

由于 $\widetilde{p}=(r-a)/(b-a),\widetilde{q}=(b-r)/(b-a)$, 故由 (40) 式可见

$$\begin{aligned}&\frac{g_n(\rho_1,\cdots,\rho_{n-1},b)-g_{n-1}(\rho_1,\cdots,\rho_{n-1})}{b-r}\\&=\frac{g_n(\rho_1,\cdots,\rho_{n-1},a)-g_{n-1}(\rho_1,\cdots,\rho_{n-1})}{a-r}.\end{aligned} \tag{41}$$

设 $\mu_n(\{a\};\omega)=I(\rho_n(\omega)=a),\mu_n(\{b\};\omega)=I(\rho_n(\omega)=b)$; 并且设

$$W_n(\omega,x)=g_n(\rho_1(\omega),\cdots,\rho_{n-1}(\omega),x)-g_{n-1}(\rho_1(\omega),\cdots,\rho_{n-1}(\omega),x),$$
$$W_n^*(\omega,x)=\frac{W_n(\omega,x)}{x-r}.$$

由于这些记号, 可见

$$\Delta X_n(\omega)=W_n(\omega,\rho_n(\omega))=\int W_n(\omega,x)\mu_n(\mathrm{d}x;\omega)=\int(x-r)W_n^*(\omega,x)\mu_n(\mathrm{d}x;\omega).$$

由于 (41) 式, 可见 *$W_n^*(\omega,x)$ 与 x 无关*. 如果将 (41) 式的左侧 (或等价地将右侧) 的式子, 记作 $\gamma_n^*(\omega)$, 则得

$$\Delta X_n(\omega)=\gamma_n^*(\omega)(\rho_n(\omega)-r). \tag{42}$$

从而

$$X_n(\omega)=X_0(\omega)+\sum_{k=1}^{n}\gamma_k^*(\omega)[\rho_k(\omega)-r]. \tag{43}$$

易见, 有

$$\Delta\left(\frac{S_n}{B_n}\right)=\frac{S_{n-1}}{B_{n-1}}\cdot\frac{\rho_n-r}{1+r}.$$

由此, 得

$$\rho_n-r=(1+r)\frac{B_{n-1}}{S_{n-1}}\Delta\left(\frac{S_n}{B_n}\right),$$

故由 (43) 式, 可见

$$X_n(\omega) = X_0(\omega) + \sum_{k=1}^{n} \gamma_k(\omega)\Delta\left(\frac{S_k(\omega)}{B_k}\right), \tag{44}$$

其中

$$\gamma_k(\omega) = \gamma_k^*(\omega)(1+r)\frac{B_{k-1}}{S_{k-1}}.$$

序列

$$\frac{S}{B} = \left(\frac{S_n}{B_n}\right)_{0\leqslant n\leqslant N}$$

关于测度 $\widetilde{\mathbf{P}}$ 是鞅. 于是, 上面引进的关系式 (44) 恰好是, X 关于 (基本) $\widetilde{\mathbf{P}}-$ 鞅 S/B 的 "$S/B-$ 表现".

4) 在 CRR – 模型 (其中 $|\mathbf{M}(\mathbf{P})| = 1$) 的蕴涵关系 {3} 的证明中, 关键是变量 ρ_n 只有 a 和 b 两个可能值. 然而, 结果表明, 关于鞅测度 $\widetilde{\mathbf{P}}$ 唯一性的假设是相当强的条件, 以至于由此可以得到变量 $\rho_n = \Delta S_n/S_{n-1}$ 具有 "两点" 的结构: 存在可以预测的量 $a_n = a_n(\omega)$ 和 $b_n = b_n(\omega)$, 使

$$\widetilde{\mathbf{P}}(\rho_n = a_n|\mathscr{F}_{n-1}) + \widetilde{\mathbf{P}}(\rho_n = b_n|\mathscr{F}_{n-1}) = 1. \tag{45}$$

如果相信这条性质, 那么上面所作的 CRR – 模型中 "$S/B-$ 表现" 的证明, 在一般情形下仍然 "适用". 这样只剩下证明 (45) 式. 在假定读者独立证明 (45) 式的正确性 (练习题 5) 的情况下, 我们仍然引进某些启发性想法, 说明具备鞅测度的唯一性导致 "条件两点" 的结构.

假设 $\mathbf{Q} = \mathbf{Q}(dx)$ 是 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 上的某一概率分布, 而 $\xi = \xi(x)$ 是一坐标给定的随机变量 $(\xi(x) = x)$, 并且满足条件: $\mathbf{E_Q}|\xi| < \infty, \mathbf{E_Q}\xi = 0$ ("鞅"); 而测度 $\mathbf{Q}$ 具有性质: 假如 $\widetilde{\mathbf{Q}}$ 是满足条件: $\mathbf{E}_{\widetilde{\mathbf{Q}}}|\xi| < \infty, \mathbf{E}_{\widetilde{\mathbf{Q}}}\xi = 0$ 的另一测度, 那么必有 $\widetilde{\mathbf{Q}} = \mathbf{Q}$ ("鞅测度的唯一性").

可以证明, 测度 $\mathbf{Q}$ 的承载子*最多集中在两点* ($a \leqslant 0$ 和 $b \geqslant 0$) 上, 也可能两点都 "粘合" 在 "零" 点 $(a = b = 0)$.

5) 上面提到的启发性想法, 使得可以将上面最后表述的、十分近乎情理的命题, 可以叙述如下.

假设测度 $\mathbf{Q}$ 集中在 x_-, x_0, x_+ 等*三个点*上, 而且三个点是有序的: $x_- \leqslant x_0 \leqslant x_+$, 其质量 (权重) 相应为 q_-, q_0, q_+. 条件 $\mathbf{E_Q}\xi = 0$ 表示

$$q_-x_- + q_0x_0 + q_+x_+ = 0.$$

假如 $x_0 = 0$, 则 $q_-x_- + q_+x_+ = 0$. 设

$$\widetilde{q}_- = \frac{q_-}{2}, \quad \widetilde{q}_0 = \frac{1}{2} + \frac{q_0}{2}, \quad \widetilde{q}_+ = \frac{q_+}{2}, \tag{46}$$

即把在点 x_- 和 x_+ 上的部分质量 q_- 和 q_+ "汲取到" 点 x_0 上.

由 (46) 式可见, 相应的测度 $\widetilde{\mathbf{Q}} \sim \mathbf{Q}$, 且 $\mathbf{E}_{\widetilde{\mathbf{Q}}}\xi = 0$, 而且 $\widetilde{\mathbf{Q}} \neq \mathbf{Q}$.

但是, 这与具有性质 $\mathbf{E_Q}\xi = 0$ 的测度 $\mathbf{Q}$ 的*唯一性*矛盾.

从而, 测度 $\mathbf{Q}$ 不可能集中在 (x_-, x_0, x_+) 三个点上, 且 $x_0 = 0$. 类似地可以基于 "汲取质量" 的思路, 讨论 $x_0 \neq 0$ 的情形. (详见 [100] 的第 V 章 §4e). □

7. 练习题

1. 假设 $\mathbf{P}\{\Delta S_1 = 0\} < 1$, 对于 $N = 1$ 的情形, 证明无仲裁条件等价于不等式 (18) 成立等价.

2. 证明, 第 4 小节引理 1 的证明中, 条件 (19) 排除可能性 2).

3. 证明, 第 5 小节例 1 中的测度 $\widetilde{\mathbf{P}}$ 是*鞅测度*, 并且在这种情形下在 $\mathbf{M}(\mathbf{P})$ 类中是唯一的.

4. 讨论第 5 小节例 2 建立的鞅测度的唯一性.

5. 证明, 对于变量 $S_n/B_n (1 \leqslant n \leqslant N)$ 的分布, 由 $(B, S)-$ 模型中的条件 $|\mathbf{M}(\mathbf{P})| = 1$, 得 "条件两点".

§12. 无仲裁模型中与 "套头交易" 有关的核算

1. 购货保留权 (选择权) 进行套头交易 (hedge, 即买即卖), 是有价证卷总存量的动态管理的基本方法之一. 下面通过所谓*选择权合同* (简称*选择权*) 核算的例子, 阐述这种方法某些基本原理和结果.

作为任意有价证券, (金融工程的工具) *选择权* 具有非常高的*风险* . 然而与此同时这些工具 (及其与其他有价证券的组合, 例如, *期货*) 被成功地利用, 不仅为通过 "市场" 价格的变化而获得收入的目的, 而且在 "市场" 价格出现戏剧性变化的情况下是 (选择权的) *保护工具*.

选择权 (option, 选择) 指金融机构发行的有价证券 (合同), 在协议的时间内或者在事先约定条件下的时间, 赋予买方购买或出售一定的价值 (例如, 股票, 债券, 货币) 的*权限*.

设计选择权核算的基本问题之一, 在于按何种*价格*出售选择权? 显然, 卖方希望获得 "尽量多", 而买方希望付出 "尽量少". 问什么价格是 "正确的", "合理的", 是买方和卖方应该都可以接受的价格?

自然, 这一 "合理的" 价格应该是 "明智的". 具体地说, 买方应该明白, 用较低的价格购买选择权, 不能得到卖方的兑现自己义务的*承诺*, 因为它所得到的 "奖励", 有可能简直不足以保障构成 "供货委托" 总存量.

与此同时, 这些 "奖励" 不应给卖方提供 "free lunch (免费午餐)" 一类的仲裁机会, 即得到*无风险收入*的可能性.

在给出应如何理解选择权的“正确”价格的定义之前，我们首先考虑选择权一些被普遍采用的分类.

2. 购货保留权 (选择权) 的类型 设 $(\Omega,\mathscr{F},(\mathscr{F}_n)_{0\leqslant n\leqslant N},\mathbf{P})$ 是过滤概率空间，而 $\mathscr{F}_0=\{\varnothing,\Omega\}$, 且 $\mathscr{F}_N=\mathscr{F}$. 考虑定义在给定概率空间上，在时间 $n=0,1,\cdots,N$ 运转的 $(B,S)-$ 市场，其中 $B=(B_n)_{0\leqslant n\leqslant N}, S=(S_n)_{0\leqslant n\leqslant N}$.

下面考虑的选择权，建立在*股票*上的情形，假设其价格由序列 $S=(S_n)_{0\leqslant n\leqslant N}$ 描绘的.

选择权按的执行时间分为两种类型：*欧洲型*和*美国型*.

如果选择权可以表现为仅在合同固定的时间 N 进行，则称 N 为执行时间，这样的选择权称为*欧洲型的*.

如果选择权可以表现为，可以在任意马尔可夫时间或者停止时间 $\tau=\tau(\omega)$ 进行 (见 §1 定义 3)，且 $\tau=\tau(\omega)$ 在合同的条件中，预先说明的集合 $\{0,1,\cdots,N\}$ 中的值，则所考虑的选择权是*美国型的*.

按照通用的术语，区分如下两类选择权：

(1) *买方选择权* (call option①, 由此产生了的术语 опцион–колл (购买权))

(2) *卖方选择权* (put option②, 由此产生了的术语 опцион–пут (出售权)).

为确定计，我们考虑*欧洲型标准选择权*的例子.

这样的选择权由两份合同体现：执行时间 N 和购买价格 K (对于买方)，或出售价格 K (对于卖方).

假如，万一在时间 N “市场”价格 $S_N>K$，则根据合同的条件，选择权的买方有权按价格 K 购买股票. 买方同样有权按“市场”价格 S_N 出售，这样可获利 S_N-K. 假如结果为 $S_N<K$，则协议商定的按价格 K 的购买权应毫无疑义地采用，因为可以用较低于“市场”价的价格 S_N 购入股票.

于是，综合这两种情形可见，买方在时间 N 的收入决定于量：

$$f_N=(S_N-K)^+, \tag{1}$$

其中对于任意实数 $a, a^+=\max(a,0)$. 其“纯”收入等于 f_N 减去他支付给股票卖方的“奖励”.

类似地，出售权卖方的收入决定于量：

$$f_N=(K-S_N)^+, \tag{2}$$

3. 完全市场和不完全市场 在定义无仲裁 $(B,S)-$ 市场的“合理”价值，应当区分两种情形：*完全市场和不完全市场*.

①买证券的特权，买期货的权利. 俄语译为“опцион–колл”. —— 译者

②买期货的权利. 俄语译为“опцион–пут”. —— 译者

定义 1 *设 $(B,S)-$ 市场是无仲裁的和完全的. 完全套头交易的价格，即量*

$$\mathbb{C}(f_N;\mathbf{P})=\inf\{x:\exists\pi,\ 使\ X_0^\pi=x,\quad X_N^\pi=f_N(\mathbf{P}-\text{a.c.})\}, \tag{3}$$

称做欧洲型选择权的“合理”价格，其中 f_N 是 $\mathscr{F}_N-$ 可测有界 (非负) 支付函数.

关于这一定义我们指出，如果依 $\mathbf{P}-$ 概率 1 有 $X_N^\pi\geqslant f_N$, 则称总存量 π 为支付委托的费用. 由 §11 可见，对于完全无仲裁市场的情形，存在有界“支付费用”的完全套头交易 π, 即满足 $X_N^\pi=f_N(\mathbf{P}-\text{a.c.})$ 的 π. 这恰好说明，为什么在定义 3 中，要考虑具有性质 $X_N^\pi=f_N$ $(\mathbf{P}-\text{a.c.})$ 的 (非空) 总存量类.

对于不完全无仲裁市场的情形，自然地有如下定义.

定义 2 *设 $(B,S)-$ 市场是无仲裁的. 超级套头交易的价格，即价格*

$$\mathbb{C}(f_N;\mathbf{P})=\inf\{x:\exists\pi,\quad 使\ X_0^\pi=x,\quad X_N^\pi\geqslant f_N(\mathbf{P}-\text{a.c.})\}, \tag{4}$$

称做具有 $\mathscr{F}_N-$ 可测有界 (非负) 支付函数 f_N 的，欧洲型选择权的“合理”价格.

注意，该定义是适定的：对于任何有界函数 f_N, 一定存在具有初始资本 x 的总存量 π, 使 $X_N^\pi\geqslant f_N(\mathbf{P}-\text{a.c.})$.

4. 价格公式 现在引进 $\mathbb{C}(f_N;\mathbf{P})$ 的公式，并且对于完全市场的情形给予证明，而对于不完全市场的情形，介绍专门的文献 ([100] 第 VI 章 §1).

定理 1 *1) 对于完全的无仲裁 $(B,S)-$ 市场的情形，支付函数为 f_N 的欧洲型选择权的“合理”价格，决定于公式*

$$\mathbb{C}(f_N;\mathbf{P})=B_0\mathbf{E}_{\widetilde{\mathbf{P}}}\frac{f_N}{B_N}, \tag{5}$$

其中 $\mathbf{E}_{\widetilde{\mathbf{P}}}$ 是对 (唯一) 鞅测度 $\widetilde{\mathbf{P}}$ 的数学期望.

2) 对于一般不完全的无仲裁 $(B,S)-$ 市场的情形，支付函数为 f_N 的欧洲型选择权的“合理”价格，决定于公式

$$\mathbb{C}(f_N;\mathbf{P})=\sup_{\widetilde{\mathbf{P}}\in\mathbf{M}(\mathbf{P})}B_0\mathbf{E}_{\widetilde{\mathbf{P}}}\frac{f_N}{B_N}, \tag{6}$$

其中 sup 在一切鞅测度 $\mathbf{M}(\mathbf{P})$ 的集合上来求.

证明 1) 设 π 是某完全套头交易，其中 $X_0^\pi=x, X_N^\pi=f_N(\mathbf{P}-\text{a.c.})$. 那么，有 (见 §11 的 (15) 式)

$$\frac{f_N}{B_N}=\frac{X_N^\pi}{B_N}=\frac{x}{B_0}+\sum_{k=1}^{N}\gamma_k\Delta\left(\frac{S_k}{B_k}\right), \tag{7}$$

因而，由于 §1 的定理 3, 得

$$\mathbf{E}_{\widetilde{\mathbf{P}}}\frac{f_N}{B_N}=\frac{x}{B_0}, \tag{8}$$

因为，由于鞅变换

$$\left(\frac{x}{B_0}+\sum_{k=1}^{n}\gamma_k\Delta\left(\frac{S_k}{B_k}\right)\right)_{1\leqslant n\leqslant N},$$

可见在“最末”时间 N, 有

$$\frac{x}{B_0}+\sum_{k=1}^{N}\gamma_k\Delta\left(\frac{S_k}{B_k}\right)=\frac{f_N}{B_N}\geqslant 0. \tag{9}$$

注意,(8) 式的左侧不依赖于, 所考虑的初始值为 $X_0^\pi=x$ 的套头交易 π. 假如现在 π' 是初始值为 $X_0^{\pi'}$ 的另外一起套头交易, 则根据 (8) 式其值仍等于 $B_0\mathbf{E}_{\widetilde{\mathbf{P}}}(f_N/B_N)$. 由此显然, 对于一切完全套头交易, 初始值 x 是同一个. 于是, 就证明了 (5) 式.

2) 我们在这里只证明不等式

$$\sup_{\widetilde{\mathbf{P}}\in\mathbf{M}(\mathbf{P})}B_0\mathbf{E}_{\widetilde{\mathbf{P}}}\frac{f_N}{B_N}\leqslant\mathbb{C}(f_N;\mathbf{P}). \tag{10}$$

(相反不等式的证明, 要用到所谓“选择权”的分解, 而这已经超出本书的范围; 见 [100] 第 VI 章的 §1c) 和 §2d)).

假设对套头交易 π, 有 $X_0^\pi=x, X_N^\pi\geqslant f_N(\mathbf{P}-\text{a.c.})$.

那么, 由 (7) 式可见

$$\frac{x}{B_0}+\sum_{k=1}^{N}\gamma_k\Delta\left(\frac{S_k}{B_k}\right)\geqslant\frac{f_N}{B_N}\geqslant 0.$$

因此, 对于任何 测度 $\widetilde{\mathbf{P}}\in\mathbf{M}(\mathbf{P})$, 有

$$B_0\mathbf{E}_{\widetilde{\mathbf{P}}}\frac{f_N}{B_N}\leqslant x$$

(对照 (8) 式和 (9) 式). 于是, 如果该式的左侧对于所有 $\widetilde{\mathbf{P}}\in\mathbf{M}(\mathbf{P})$ 求 sup, 则得所要证明的 (10) 式. □

5. 美国型购货保留权 (选择权) 考虑涉及美国型选择权的一些定义和结果. 对于这样的选择权需要作如下假设: 给定涉及时间 N 的不只一个支付函数 f_N, 而是整整一组函数 $f_0,f_1,\cdots,f_N$; 这些支付函数的含义是, 如果选择权表明买方出现在时间 n, 则 (卖方给买方选择权的) 相应的支付决定于 ($\mathscr{F}_n$– 可测) 函数 $f_n=f_n(\omega)$.

以 $\tau=\tau(\omega)$ 表示值域为 $\{0,1,\cdots,N\}$ 的马尔可夫时间. 假如选择权的买方决定在时间 $\tau=\tau(\omega)$ 提出要求执行选择权, 则支付函数等于 $f_{\tau(\omega)}(\omega)$, 因而有价证券 π 选择权的卖方永远应当预见到, 使得对于任何 τ, 都能满足套头交易的如下条件: $X_\tau^\pi\geqslant f_\tau(\mathbf{P}-\text{a.c.})$.

这说明了如下定义的合理性.

定义 3 设 (B,S)– 市场是无仲裁的, 而 $f=(f_n)_{0\leqslant n\leqslant N}$ 是 $\mathscr{F}_n$– 可测非负支付函数 f_n 系. 超级套头交易的最高价格,即价格

$$\overline{\mathbb{C}}(f;\mathbf{P})=\inf\{x:\exists\pi,X_0^\pi=x,X_n^\pi\geqslant f_n(\mathbf{P}-\text{a.c.}),0\leqslant n\leqslant N\}, \tag{11}$$

称做以 $f=(f_n)_{0\leqslant n\leqslant N}$ 为支付函数系的、美国型选择权的“合理”价格.

对于美国型选择权的情形, 我们 (不加证明地) 引进定理 1 的如下类似.

定理 2 1) 对于完全的无仲裁 (B,S)– 市场的情形, 以 $f=(f_n)_{0\leqslant n\leqslant N}$ 为支付函数系的、美国型选择权的“合理”价格, 决定于公式

$$\overline{\mathbb{C}}(f;\mathbf{P})=\sup_{\tau\in\mathfrak{M}_0^N}B_0\mathbf{E}_{\widetilde{\mathbf{P}}}\frac{f_\tau}{B_\tau}, \tag{12}$$

其中 $\mathfrak{M}_0^N=\{\tau:\tau\leqslant N\}$ 是 (关于 $(\mathscr{F}_n)_{0\leqslant n\leqslant N}$) 的停时类, 而 $\widetilde{\mathbf{P}}$ 是唯一鞅测度. $\mathbf{E}_{\widetilde{\mathbf{P}}}$ 是对 (唯一) 鞅测度 $\widetilde{\mathbf{P}}$ 的数学期望.

2) 对于一般不完全的无仲裁 (B,S)– 市场的情形, 支付函数为 $f=(f_n)_{0\leqslant n\leqslant N}$ 的美国型选择权的“合理”价格, 决定于公式

$$\overline{\mathbb{C}}(f;\mathbf{P})=\sup_{\tau\in\mathfrak{M}_0^N,\widetilde{\mathbf{P}}\in\mathbf{M}(\mathbf{P})}B_0\mathbf{E}_{\widetilde{\mathbf{P}}}\frac{f_\tau}{B_\tau}, \tag{13}$$

其中 $\mathbf{M}(\mathbf{P})$ 是一切鞅测度 $\widetilde{\mathbf{P}}$ 的集合.

证明见 [100] 第 VI 章的 §2c.

6. 欧洲型购货保留权 (选择权) 上面的两个定理回答选择权如何决定其“合理”价格的问题.

在得到“奖励” $\mathbb{C}(f_N;\mathbf{P})$ 和 $\overline{\mathbb{C}}(f;\mathbf{P})$ 之后, 选择权的卖方如何建立套头存量 π^* 的问题也同样重要.

为简便计, 我们的叙述仅限于考虑欧洲型购货选择权完全 (B,S)– 市场的情形.

定理 3 设 (B,S)– 市场是无仲裁的和完全的.

存在自筹总存量 $\pi^*(\beta^*,\gamma^*)$, 其初始资本为 $X_0^{\pi^*}=\mathbb{C}(f_N;\mathbf{P})$, 并且实行完全支付义务 f_N 的套头交易:

$$X_0^{\pi^*}=f_N(\mathbf{P}-\text{a.c.}).$$

资本 $X_n^{\pi^*}=\beta_n^*B_n+\gamma_n^*S_n(0\leqslant n\leqslant N)$ 的变动决定于如下公式

$$X_n^{\pi^*}=B_n\mathbf{E}_{\widetilde{\mathbf{P}}}\left(\frac{f_N}{B_N}\middle|\mathscr{F}_n\right). \tag{14}$$

自筹总存量 $\pi^*=(\beta^*,\gamma^*)$ 中的分量 $\gamma^*=(\gamma_n^*)_{0\leqslant n\leqslant N}$, 根据 $X^{\pi^*}=(X_n^{\pi^*})_{0\leqslant n\leqslant N}$ 的值由公式

$$\Delta\left(\frac{X_n^{\pi^*}}{B_n}\right)=\gamma_n^*\Delta\left(\frac{S_n}{B_n}\right), \tag{15}$$

来求, 而分量 $\beta^*=(\beta_n^*)_{0\leqslant n\leqslant N}$ 的值由公式

$$X_n^{\pi^*}=\beta_n^*B_n+\gamma_n^*S_n, \tag{16}$$

来求.

定理的证明可以直接仿照 §11 引理 2 中蕴涵关系

$$\text{“完全性”} \Rightarrow \text{“}\frac{S}{B}-\text{表现”}$$

的证明得到. 为此, 只需将上面蕴涵关系的证明用于鞅

$$m=(m_n)_{0\leqslant n\leqslant N},\quad m_n=\mathbf{E}_{\widetilde{\mathbf{P}}}\left(\frac{f_N}{B_N}\bigg|\mathscr{F}_n\right).$$

7. 选择权实际核算的例 作为选择权实际核算的例子, 考虑由 CCR – 模型

$$\begin{aligned}B_n&=B_{n-1}(1+r),\\ S_n&=S_{n-1}(1+\rho_n),\end{aligned}\tag{17}$$

描绘的 (B,S)– 市场, 其中 $\rho_1,\cdots,\rho_N$ 是只有 a 和 $b(-1<a<r<b)$ 两个可能值的、独立同分布随机变量.

这样的市场是*无仲裁的和完全的* (见 §11 练习题 3), 且与其相联系的鞅测度 $\widetilde{\mathbf{P}}$ 满足: $\widetilde{\mathbf{P}}\{\rho_n=b\}=\widetilde{p},\widetilde{\mathbf{P}}\{\rho_n=a\}=\widetilde{q}$, 其中

$$\widetilde{p}=\frac{r-a}{b-a},\quad \widetilde{q}=\frac{b-r}{b-a},\tag{18}$$

(见 §11 第 5 小节例 1).

根据定理 1 的 (5) 式, 对于所考虑的 (B,S)– 市场, “合理” 价格为

$$\mathbb{C}(f_N;\mathbf{P})=\mathbf{E}_{\widetilde{\mathbf{P}}}\frac{f_N}{(1+r)^N}.\tag{19}$$

故根据定理 3 求完全套头交易总存量 $\pi^*(\beta^*,\gamma^*)$, 需要首先计算

$$X_0^{\pi^*}=\mathbf{E}_{\widetilde{\mathbf{P}}}\left(\frac{f_N}{(1+r)^N}\bigg|\mathscr{F}_n\right),\tag{20}$$

其中 $\mathscr{F}_n=\sigma(\rho_1,\cdots,\rho_n),1\leqslant n\leqslant N$, 且 $\mathscr{F}_0=\{\varnothing,\Omega\}$; 然后由公式 (15) 和 (16) 求 β^* 和 γ^*.

由于 $X_0^{\pi^*}=\mathbb{C}(f_N;\mathbf{P})$, 可见这一切归结为, 对于 $n=0,1,\cdots,N$, 求 (20) 式右侧的条件数学期望.

假设 $\mathscr{F}_n$– 可测函数 f_N 有 “马尔可夫” 结构, 即 $f_N=f(S_N)$, 其中 $f=f(x)$ 是 $x\geqslant 0$ 的某一非负函数.

记

$$F_n(x;p)=\sum_{k=0}^{n}f(x(1+b)^k(1+a)^{n-k})\mathrm{C}_n^k p^k(1-p)^{n-k}.\tag{21}$$

由于

$$\prod_{n<k\leqslant N}(1+\rho_k)=(1+b)^{\Delta_N-\Delta_n}(1+a)^{(N-n)-(\Delta_N-\Delta_n)},$$

其中 $\Delta_k=\delta_1+\cdots+\delta_n$, 而 $\delta_k=(p_k-a)/(b-a)$, 可见

$$\mathbf{E}_{\widetilde{\mathbf{P}}}f\left(x\prod_{n<k\leqslant N}(1+\rho_k)\right)=F_{N-n}(x,\widetilde{p}),\quad \text{其中 } \widetilde{p}=\frac{r-a}{b-a}.\tag{22}$$

而考虑到

$$S_N=S_n\prod_{n<k\leqslant N}(1+\rho_k),$$

故由式 (21) 式和 (22) 式, 得

$$X_0^{\pi^*}=\mathbf{E}_{\widetilde{\mathbf{P}}}\left(\frac{f_N}{(1+r)^N}\bigg|\mathscr{F}_n\right)=(1+r)^{-N}F_{N-n}(S_n,\widetilde{p}).\tag{23}$$

特别,

$$\mathbb{C}(f_N;\mathbf{P})=X_0^{\pi^*}=(1+r)^{-N}F_N(S_0,\widetilde{p}).\tag{24}$$

于是, 由 (15) 式, 并注意到 (23) 式, 可见

$$\gamma_n^*=\Delta\left(\frac{X_n^{\pi^*}}{B_n}\right)\bigg/\Delta\left(\frac{S_n}{B_n}\right)$$

决定于

$$\gamma_n^*=(1+r)^{-(N-n)}\frac{F_{N-n}(S_{n-1}(1+b);\widetilde{p})-F_{N-n}(S_{n-1}(1+a);\widetilde{p})}{S_{n-1}(b-a)}.\tag{25}$$

为求 β_n^*, 注意到 $B_{n-1}\Delta\beta_n^*+S_{n-1}\Delta\gamma_n^*=0$. 所以

$$X_{n-1}^{\pi^*}=\beta_n^*B_{n-1}+\gamma_n^*S_{n-1},\tag{26}$$

因此

$$\beta_n^*=\frac{X_{n-1}^{\pi^*}-\gamma_n^*S_{n-1}}{B_{n-1}}.\tag{27}$$

由 (23) 式和 (25) 式, 可见

$$\begin{aligned}\beta_n^*=\frac{1}{B_N}\bigg\{&F_{N-n+1}(S_{n-1};\widetilde{p})-\frac{1+r}{1+b}[F_{N-n}(S_{n-1}(1+b);\widetilde{p})\\&-F_{N-n}(S_{n-1}(1+a);\widetilde{p})]\bigg\}.\end{aligned}\tag{28}$$

最后, 我们讨论一下, 例如, 在买方的标准选择权 (购买权) 的情形下, 即当函数 $f_N=(S_N-K)^+$ 时, “合理” 价格 $\mathbb{C}(f_N,\mathbf{P})$ 的公式具有何种形式.

设 $K_0=K_0(a,b,N;s_0/K)$ 是满足如下条件的最小整数:

$$S_0(1+a)^N\left(\frac{1+b}{1+a}\right)^{K_0}>K,\tag{29}$$

即设

$$K_0 = 1 + \left[\ln\left(\frac{K}{S_0(1+a)^N}\right) \Big/ \ln\left(\frac{1+b}{1+a}\right)\right], \tag{30}$$

其中 $[x]$ 表示 x 的整数部分.

如果记

$$p^* = \frac{1+b}{1+r}\widetilde{P}, \quad 其中 \quad \widetilde{P} = \frac{r-a}{b-a},$$

而

$$\mathbb{B}(K_0, N, p) = \sum_{k=K_0}^{N} C_N^k p^k (1-p)^{N-k}, \tag{31}$$

则对于标准购买权, 由 (24) 式不难推导出“合理”价格 (记作 $\mathbb{C}_N$) 的如下 (考科斯 – 罗斯 – 鲁宾斯坦) 公式:

$$\mathbb{C}_N = S_0\mathbb{B}(K_0, N; p^*) - K(1+r)^{-N}\mathbb{B}(K_0, N; \widetilde{p}). \tag{32}$$

当 $K_0 > N$ 时 $\mathbb{C}_N = 0$.

注 由于

$$(K - S_N)^+ = (S_N - K)^+ - S_N + K,$$

可见卖方的标准选择权 (出售权)“合理”价格 $\mathbb{P}_N (= \mathbb{C}(f_N, \mathbf{P})$, 其中 $f_N = (K - S_N)^+)$ 决定于如下公式:

$$\mathbb{P}_N = \widetilde{\mathbf{E}}(1+r)^{-N}(K - S_N)^+ = \mathbb{C}_N - \widetilde{\mathbf{E}}(1+r)^{-N}S_N + K(1+r)^{-N}.$$

因为 $\widetilde{\mathbf{E}}(1+r)^{-N}S_N = S_0$, 所以显然有“平价购买 – 出售权”恒等式:

$$\mathbb{P}_N = \mathbb{C}_N - S_0 + K(1+r)^{-N}. \tag{33}$$

8. 练习题

1. 对于 §11 第 5 小节例 2 中的 (B, S)– 市场模型, 求标准选择权的出售权的“合理”价格: $\mathbb{C}(f_N, \mathbf{P})$, 其中 $f_N = (S_N - K)^+$.
2. 试证明 (10) 式中相反不等式的正确性.
3. 证明 (12) 式,并试证明 (13) 式.
4. 证明 (23) 式详细类似的结论.
5. 证明 (25) 式和 (28) 式.
6. 对 (32) 式进行详细的推导.

§13. 最优停时问题. 鞅方法

1. “合理” 价格 在这一节中, 我们假设给定一固定的概率空间 $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbf{P})$ 和 $\mathscr{F}_0 = (\varnothing, \Omega)$. 设 $\tau = \tau(\omega)$ 是马尔可夫时间 (或停止时间), 其值域是集合 $\{0, 1, \cdots, N\}$. 设 $\mathfrak{M}_0^N$ 是由属于集合

$$\{\omega : \tau(\omega) = n\} \in \mathscr{F}_n \tag{1}$$

的值域为 $\{0, 1, \cdots, N\}$ 的随机变量 $\tau = \tau(\omega)$ 类. 在描绘美国型选择权“合理”价格时, 我们已经遇到过有关停止时间的例子. 具体地说, §12 的 (12) 式表明, 对于简化了的条件: $B_n = 1, 0 \leqslant n \leqslant N$ 和 $\widetilde{\mathbf{P}} = \mathbf{P}$, 为求这一价格, 需要寻找量 (也称做“价格”)

$$V_0^N = \sup_{\tau\in\mathfrak{M}_0^N} \mathbf{E}f_\tau, \tag{2}$$

其中 $f = (f_0, f_1 \cdots, f_N)$ 是 $\mathscr{F}_n$– 可测非负函数 f_n 的序列.

设 $\tau = \tau(\omega) \in \mathfrak{M}_0^N$ 是马尔可夫时间. 与问题 (1) 同时, 对寻求量 (“价格”)

$$V_0^\infty = \sup_{\tau\in\mathfrak{M}_0^\infty} \mathbf{E}f_\tau \tag{3}$$

感兴趣, 其中 $\mathfrak{M}_0^N = \{\tau : \tau < \infty\}$, 而 $f = (f_0, f_1, \cdots)$ 是 $\mathscr{F}_n$– 可测随机变量 f_n, $n \geqslant 0$ 的随机序列, 且 $\mathbf{E}|f_\tau| < \infty$.

像 (2)式的情形一样, 对于情形 (3), 除寻求“价格” V_0^N 和 V_0^∞ 之外, (若上确界存在) 还要寻求达到上确界的时间.

在许多问题中, 允许考虑可以取 $+\infty$ 为值的马尔可夫时间也是适宜的. 在这种情形下, 讨论 $\mathbf{E}f_\tau$ 时应当约定如何理解 f_∞. 一种自然的方法是把 f_∞ 视为的 $\overline{\lim}_n f_n$ 的值. 另一种方法是, 允许对于 τ 值和无限大, 把“价格”定义为

$$\overline{V}_0^\infty = \sup_{\tau\in\overline{\mathfrak{M}}_0^\infty} \mathbf{E}f_\tau I(\tau < \infty), \tag{4}$$

其中 $\overline{\mathfrak{M}}_0^\infty$ 是马尔可夫时间类, 即一切马尔可夫时间的集合 $\overline{\mathfrak{M}}_0^\infty = \{\tau : \tau \leqslant \infty\}$. 显然, 若 $f_\infty = 0$, 则

$$\overline{V}_0^\infty = \sup_{\tau\in\overline{\mathfrak{M}}_0^\infty} \mathbf{E}f_\tau$$

(对照 §1 的第 3 小节).

我们在下面将只考虑问题 (2). (关于 $N = \infty$ 的情形, 见第 VIII 章 §9). 假如不考虑序列 $f = (f_0, f_1, \cdots, f_N)$ 的具体购造, 则问题 (2) 和 (3) 的解法是下面描绘的“鞅”方法. (不失普遍性, 我们以后都假设对于一切 $n \leqslant N$, 有 $\mathbf{E}|f_n| < \infty$, 并且每次不再特别说明).

2. 价格最高的时间 设 $N < \infty$. 现在考虑的情形可以看成"向后归纳法", 这里用下面的方式实现.

与价格 V_0^N 同时引进"价格"

$$V_n^N = \sup_{\tau \in \mathfrak{M}_n^N} \mathbf{E} f_\tau, \tag{5}$$

其中 $\mathfrak{M}_n^N = \{\tau : n \leqslant \tau \leqslant N\}$ 是对于一切 $\omega \in \Omega$, 满足 $n \leqslant \tau(\omega) \leqslant N$ 的停时类.

对于 $n = N-1, \cdots, 0$, 同样直观地引进随机序列 $v^N = (v_n^N)_{0 \leqslant n \leqslant N}$, 其中

$$v_N^N = f_N, \quad v_n^N = \max\{f_n, \mathbf{E}(v_{n+1}^N | \mathscr{F}_n)\}. \tag{6}$$

对于 $0 \leqslant n \leqslant N$, 设

$$\tau_n^N = \min\{n \leqslant k \leqslant N : f_k = v_k^N\}. \tag{7}$$

下面的命题可以完整地描绘, 利用所引进的量, 求解 (2) 式 (5) 式的最优停止的问题.

定理 1 *假设对于序列 $f = (f_0, f_1, \cdots, f_N)$, 随机变量 $f_n, n \geqslant 0$ 为 $\mathscr{F}_n$– 可测.*

1) 对于每一个 $n (0 \leqslant n \leqslant N)$, 时间

$$\tau_n^N = \min\{n \leqslant k \leqslant N : v_k^N = f_k\}. \tag{8}$$

在类 $\mathfrak{M}_n^N$ 中是最优的:

$$\mathbf{E} f_{\tau_n^N} = \sup_{\tau \in \mathfrak{M}_n^N} \mathbf{E} f_\tau \quad (= V_n^N). \tag{9}$$

2) 时间 $\tau_n^N (0 \leqslant n \leqslant N)$ 在如下的"条件"意义上也是最优的: $(\mathbf{P}-\text{a.c.})$

$$\mathbf{E}(f_{\tau_n^N} | \mathscr{F}_n) = \operatorname*{ess\,sup}_{\tau \in \mathfrak{M}_n^N} \mathbf{E}(f_\tau | \mathscr{F}_n). \tag{10}$$

"随机价格" $\operatorname*{ess\,sup}_{\tau \in \mathfrak{M}_n^N} \mathbf{E}(f_\tau | \mathscr{F}_n)$ 等于 v_n^N:

$$\operatorname*{ess\,sup}_{\tau \in \mathfrak{M}_n^N} \mathbf{E}(f_\tau | \mathscr{F}_n) = v_n^N \quad (\mathbf{P}-\text{a.c.}) \tag{11}$$

且

$$V_n^N = \mathbf{E} v_n^N. \tag{12}$$

如果 $n = 0$, 则

$$V_0^N = v_0^N. \tag{13}$$

如果 $n = N$, 则

$$V_N^N = \mathbf{E} f_N. \tag{14}$$

3. 随机变量族的本质上确界 在开始证明之前我们回忆, 关于 (10) 式中使用的 $\mathscr{F}_n$– 可测随机变量族 $\{\xi_\alpha(\omega), \alpha \in \mathfrak{U}\}$ 的、*本质上确界* $\operatorname*{ess\,sup}_{\alpha \in \mathfrak{U}} \xi_\alpha(\omega)$ 的概念.

下面的情况决定了引进这一概念的必要性: 在不可数集合 $\mathfrak{U}$ 的情形下, 函数 $\sup_{\alpha \in \mathfrak{U}} \xi_\alpha(\omega) (\omega \in \Omega)$ 一般有可能是 $\mathscr{F}$– 不可测的.

事实上, 对于任意 $c \in \mathbb{R}$,

$$\left\{\omega : \sup_{\alpha \in \mathfrak{U}} \xi_\alpha(\omega) \leqslant c\right\} = \bigcap_{\alpha \in \mathfrak{U}} \{\omega : \xi_\alpha(\omega) \leqslant c\}.$$

这里, 集合 $A_\alpha = \{\omega : \xi_\alpha(\omega) \leqslant c\} \in \mathscr{F}$ (即 A_α 是事件). 然而, 由于集合 $\mathfrak{U}$ 的不可数性, 故不能保障 $\bigcap_{\alpha \in \mathfrak{U}} A_\alpha = \{\omega : \xi_\alpha(\omega) \leqslant c\} \in \mathscr{F}$.

定义 设 $\{\xi_\alpha(\omega) : \alpha \in \mathfrak{U}\}$ 是随机变量族 (即值域为 $(-\infty, +\infty)$ 的 $\mathscr{F}_n$– 可测函数族). 称广义随机变量 $\xi(\omega)$ (即值域为 $(-\infty, +\infty)$ 的 $\mathscr{F}_n$– 可测函数) 是随机变量族 $\{\xi_\alpha(\omega) : \alpha \in \mathfrak{U}\}$ 的本质上确界, 记作 $\xi(\omega) = \operatorname*{ess\,sup}_{\alpha \in \mathfrak{U}} \xi_\alpha(\omega)$, 如果

a) 对于一切 $\alpha \in \mathfrak{U}, \xi(\omega) \geqslant \xi_\alpha(\omega) (\mathbf{P}-\text{a.c.})$,

b) 由于对于, (广义) 随机变量 $\eta(\omega)$, 对一切 $\alpha \in \mathfrak{U}$ 满足 $\eta(\omega) \geqslant \xi_\alpha(\omega) (\mathbf{P}-\text{a.c.})$, 可见 $\xi(\omega) \leqslant \eta(\omega) (\mathbf{P}-\text{a.c.})$.

换句话说, 在一切控制随机变量 $\xi_\alpha(\omega), \alpha \in \mathfrak{U}$ 的 (广义) 随机变量中, $\xi(\omega)$ 是最小的 (广义) 随机变量.

当然, 首先应该证明这一定义富有内容, 而这由下面的引理可以看到.

引理 *对于任意随机变量族 $\{\xi_\alpha(\omega) : \alpha \in \mathfrak{U}\}$, 存在具有定义中性质 a) 和 b) 的随机变量 $\xi(\omega)$ (记作 $\xi(\omega) = \operatorname*{ess\,sup}_{\alpha \in \mathfrak{U}} \xi_\alpha(\omega)$, 一般为广义随机变量).*

存在具有同样性质的子集 $\mathfrak{U}_0 \subseteq \mathfrak{U}$, 并且作为这样的随机变量, 可以取随机变量

$$\xi(\omega) = \sup_{\alpha \in \mathfrak{U}_0} \xi_\alpha(\omega).$$

证明 首先假设所有变量 $\xi_\alpha(\omega), \alpha \in \mathfrak{U}$,一致有界 $(|\xi_\alpha(\omega)| \leqslant c, \omega \in \Omega, \alpha \in \mathfrak{U})$.

设 A 是下标 $\alpha \in \mathfrak{U}$ 的有限集合. 记

$$S(A) = \left(\max_{\alpha \in A} \xi_\alpha(\omega)\right).$$

其次, 设 $S = \sup S(A)$, 其中上确界对一切有限子集 $A \subseteq \mathfrak{U}$ 来求.

对于 $n \geqslant 1$, 以 A_n 表示满足

$$\mathbf{E}\left(\max_{\alpha \in A} \xi_\alpha(\omega)\right) \geqslant S - \frac{1}{n}$$

的有限集合 A_n.

记 $\mathfrak{U}_0 = \bigcap_{n\geqslant 1} A_n$. 由于这一集合的可数, 函数

$$\xi(\omega) = \sup_{\alpha\in\mathfrak{U}_0} \xi_\alpha(\omega)$$

$\mathscr{F}_n-$ 可测, 因此是随机变量. (注意, 由于 $|\xi(\omega)| \leqslant c$, 可见 $\xi(\omega)$ 是普通的, 而不是广义随机变量).

由随机变量 $\xi(\omega)$ 如上的构造可见 (练习题 1), 它满足上面定义中的条件 a) 和 b).

因而, 在随机变量族 $\{\xi_\alpha(\omega) : \alpha \in \mathfrak{U}\}$ 一致有界的情形下, 就证明了上确界的存在性.

在一般情形下, 需要首先由随机变量 $\xi_\alpha(\omega)$, 转换为有界随机变量 $\widetilde{\xi}_\alpha(\omega) = \arctan\xi_\alpha(\omega)$, 使之满足 $|\widetilde{\xi}_\alpha(\omega)| \leqslant \pi/2, \alpha\in\mathfrak{U}, \omega\in\Omega$, 然后建立 $\widetilde{\xi}(\omega) = \operatorname{ess\,sup}_{\alpha\in\mathfrak{U}}\widetilde{\xi}_\alpha(\omega)$.

随机变量 $\xi_\alpha(\omega) = \tan\widetilde{\xi}_\alpha(\omega)$ 满足本质上确界定义 (练习题 1) 的条件 a) 和 b).

4. 定理 1 的证明

证明 固定上标 N, 而为简便计现在将其略去.

如果 $n = N$, 则 $v_N = f_N$ 且 $\tau_N = N$, 故性质 (9) $\sim$ (12), (14) 显然. 现在用归纳法证明.

设对于 $n = N, N-1, \cdots, k$, 定理的论断已经得证. 现在证明对于 $n = k-1$, 定理的论断也成立.

设 $\tau \in \mathfrak{M}_{k-1}(= \mathfrak{M}_{k-1}^N)$ 和 $A \in \mathscr{F}_{k-1}$. 定义时间 $\overline{\tau} \in \mathfrak{M}_k$, 设 $\overline{\tau} = \max(\tau, k)$. 由于 $\overline{\tau} \in \mathfrak{M}_k$ 且事件 $\{\tau \geqslant k\} \in \mathscr{F}_{k-1}$, 可见

$$\begin{aligned}
\mathbf{E}[I_A f_\tau] &= \mathbf{E}[I_{A\cap\{\tau=k-1\}} f_\tau] + \mathbf{E}[I_{A\cap\{\tau\geqslant k\}} f_\tau] \\
&= \mathbf{E}[I_{A\cap\{\tau=k-1\}} f_\tau] + \mathbf{E}[I_{A\cap\{\tau\geqslant k\}} \mathbf{E}(f_\tau|\mathscr{F}_{k-1})] \\
&= \mathbf{E}[I_{A\cap\{\tau=k-1\}} f_\tau] + \mathbf{E}[I_{A\cap\{\tau\geqslant k\}} \mathbf{E}(\mathbf{E}(f_{\overline{\tau}}|\mathscr{F}_k)|\mathscr{F}_{k-1})] \\
&\leqslant \mathbf{E}[I_{A\cap\{\tau=k-1\}} f_{k-1}] + \mathbf{E}[I_{A\cap\{\tau\geqslant k\}} \mathbf{E}(v_k|\mathscr{F}_{k-1})] \leqslant \mathbf{E}[I_A v_{k-1}].
\end{aligned} \tag{15}$$

由于集合 A 的 $\mathscr{F}_{k-1}-$ 可测性, 由此可见, 对于任意设 $\tau \in \mathfrak{M}_{k-1}(\mathbf{P}-\text{a.c.})$,

$$\mathbf{E}(f_\tau|\mathscr{F}_{k-1}) \leqslant v_{k-1}. \tag{16}$$

现在证明, 对于时间 τ_{k-1}, 依 $\mathbf{P}-$ 概率 1, 有

$$\mathbf{E}(f_{\tau_{k-1}}|\mathscr{F}_{k-1}) = v_{k-1}. \tag{17}$$

(假如能证明该式, 则由 (16) 式可见, 对于 $n = k-1$, 关系式 (10) 和 (11) 也成立).

为此, 只需证明, 对于 $\tau = \tau_{k-1}$, 实际上 (15) 式到处为等式.

像 (15) 式那样开始, 并注意到根据 (5) 式的定义在集合 $\{\tau_{k-1} \geqslant k\}$ 上 $\tau = \tau_k$, 则 (根据归纳法的假设) 由 $\mathbf{E}(f_{\tau_k}|\mathscr{F}_k) = v_k(\mathbf{P}-\text{a.c.})$, 可得

$$\begin{aligned}
\mathbf{E}[I_A f_{\tau_{k-1}}] &= \mathbf{E}[I_{A\cap\{\tau_{k-1}=k-1\}} f_{k-1}] + \mathbf{E}[I_{A\cap\{\tau_{k-1}\geqslant k\}} \mathbf{E}(f_{\tau_{k-1}}|\mathscr{F}_{k-1})] \\
&= \mathbf{E}[I_{A\cap\{\tau_{k-1}=k-1\}} f_{k-1}] + \mathbf{E}[I_{A\cap\{\tau_{k-1}\geqslant k\}} \mathbf{E}(f_{\tau_k}|\mathscr{F}_{k-1})] \\
&= \mathbf{E}[I_{A\cap\{\tau_{k-1}=k-1\}} f_{k-1}] + \mathbf{E}[I_{A\cap\{\tau_{k-1}\geqslant k\}} \mathbf{E}(v_k|\mathscr{F}_{k-1})] = \mathbf{E}[I_A v_{k-1}],
\end{aligned}$$

其中在证明最后一个等式时, 根据 (6) 式, 有: $v_{k-1} = \max(f_{k-1}, \mathbf{E}(v_k|\mathscr{F}_{k-1}))$; 由此可见, 在集合 $\{\tau_{k-1} = k-1\}$ 上 $v_{k-1} = f_{k-1}$, 而在集合 $\{\tau_{k-1} > k-1\} = \{\tau_{k-1} \geqslant k\}$ 上 $v_{k-1} > f_{k-1}$ (因此, 在这一集合上 $v_{k-1} = \mathbf{E}(v_k|\mathscr{F}_{k-1})$).

这样, 性质 (17) 得证. 像前面已经指出的那样, 该性质连同 (16) 式就证明了欲证的关系式 (10) 和 (11).

由这些关系式可见, 对于 $\tau \in \mathfrak{M}_n(= \mathfrak{M}_n^N), (\mathbf{P}-\text{a.c.})$ 有

$$v_n = \mathbf{E}(f_{\tau_n}|\mathscr{F}_n) \geqslant \mathbf{E}(f_\tau|\mathscr{F}_n). \tag{18}$$

从而, 注意到 $v_n^N = v_n$, 由此可得

$$\mathbf{E}v_n^N = \mathbf{E}f_{\tau_n} \geqslant \sup_{\tau\in\mathfrak{M}_n^N} \mathbf{E}f_\tau = V_n^N, \tag{19}$$

而这就证明了欲证的关系式 (9) 和 (12).

性质 (13) 是如下两个性质的特殊情形: 1) 性质 (13) 是性质 (12) 当 $n = 0$ 时的情形, 2) v_0^N 是常数 (因为由 (11) 式, 以及 $\sigma-$ 代数 $\mathscr{F}_0 = (\varnothing, \Omega)$ 是平凡的). 最后, 等式 (14) 当 $n = N$ 时是定义 (5) 的推论. □

5. 最优停时的 "鞅" 视角 为显示所研究的最优停止问题的 "鞅" 视角, 现在讨论序列 $v^N = (v_0^N, v_1^N, \cdots, v_N^N)$, 的递推关系式 (6), 其中 $v_N^N = f_N$ 是 "边界" 条件.

由 (6) 式可见, 对于每个 $n = 0, 1, \cdots, N-1, (\mathbf{P}-\text{a.c.})$ 有

$$v_n^N \geqslant f_n, \tag{20}$$

$$v_n^N \geqslant \mathbf{E}(v_{n+1}^N|\mathscr{F}_n). \tag{21}$$

第一个不等式在这里表示, 序列 v^N 控制序列 $f = (f_0, f_1, \cdots, f_N)$. 第二个不等式表示, 序列 v^N 是 "术语" 值为 $v_N^N = f_N$ 的上鞅. 这样, 可以说, 序列 $v^N = (v_0^N, v_1^N, \cdots, v_N^N)$, 其中由 (6) 式或 (11) 式决定于的量 v_n^N, 是上鞅的控制序列 $f = (f_0, f_1, \cdots, f_N)$.

换一种说法, 这表示序列 v^N 属于序列 $\gamma^N = (\gamma_0^N, \gamma_1^N, \cdots, \gamma_N^N)$ 类, 其中 $\gamma_n^N \geqslant f_N$, 且对于一切 $n = 0, 1, \cdots, N-1, (\mathbf{P}-\text{a.c.})$ 满足 "变分不等式"

$$\gamma_n^N \geqslant \max(f_n, \mathbf{E}(\gamma_{n+1}^N|\mathscr{F}_n)). \tag{22}$$

但是序列 v^N 具有如下补充性质: (22) 式不仅是"不严格"不等式"$\geqslant$", 而且就是等式"$=$" (见 (6) 式). 根据这条性质, 就可以用如下方式从序列 $\gamma^N=(\gamma_0^N,\gamma_1^N,\cdots,\gamma_N^N)$ 类 (其中 $\gamma_N^N\geqslant f_N$) 中, 划分出序列 v^N.

定理 2 *序列 v^N 是最小上鞅的控制序列 $f=(f_0,f_1,\cdots,f_N)$.*

证明 事实上, 由于 $v_N^N=f_N$, 而 $\gamma_N^N\geqslant f_N$, 可见 $\gamma_N^N\geqslant v_N^N$. 由此以及 (22) 式和 (6) 式, ($\mathbf{P}$-a.c.) 有

$$\gamma_{N-1}^N\geqslant\max(f_{N-1},\mathbf{E}(\gamma_N^N|\mathscr{F}_{N-1}))\geqslant\max(f_{N-1},\mathbf{E}(v_N^N|\mathscr{F}_{N-1}))=v_{N-1}^N.$$

类似地, 对于其余 $n<N-1$, 可得 $\gamma_n^N\geqslant v_n^N$.

注 该定理得结果可以表述为如下形式: *递推方程组*

$$v_n^N=\max(f_n,\mathbf{E}(v_{n+1}^N|\mathscr{F}_n)),\quad n<N,$$

的解 $v^N=(v_0^N,v_1^N,\cdots,v_N^N)$, 其中 $v_N^N=f_N$, 是递推不等式组

$$\gamma_n^N\geqslant\max(f_n,\mathbf{E}(\gamma_{n+1}^N|\mathscr{F}_n)),\quad n<N,\tag{23}$$

一切可能解 $\gamma^N=(\gamma_0^N,\gamma_1^N,\cdots,\gamma_N^N)$ 中最小者, 其中 $\gamma_N^N\geqslant f_N$.

6. 停时集与继续观测集 定理 1 和定理 2, 不仅描绘寻找价格 $V_0^N=\sup\mathbf{E}f_\tau$ 的方法, 其中上确界在马尔可夫类时间 $\mathfrak{M}_0^N$ 上求, 但同时表明如何求最优时间 τ_0^N, 即使 $\mathbf{E}f_{\tau_0^N}=V_0^N$ 的时间.

根据 (8) 式

$$\tau_0^N=\min\{0\leqslant k\leqslant N:v_k^N=f_k\}.\tag{24}$$

在求解关于最优停时问题时, 如下这一停时 τ_0^N 的等价描绘很重要. 设

$$D_n^N=\{\omega:v_n^N(\omega)=f_n(\omega)\}\tag{25}$$

和

$$C_n^N=\Omega\backslash D_n^N=\{\omega:v_n^N(\omega)=\mathbf{E}(v_{n+1}^N|\mathscr{F}_n)(\omega)\}.$$

显然, $D_N^N=\Omega,C_N^N=\varnothing$ 和

$$D_0^N\subseteq D_1^N\subseteq,\cdots,\subseteq D_N^N=\Omega,$$
$$C_0^N\supseteq C_1^N\supseteq,\cdots,\supseteq C_N^N=\varnothing.$$

由 (24) 和 (25) 式可见, 时间 τ_0^N可由下列形式定义

$$\tau_0^N=\min\{0\leqslant k\leqslant N:\omega\in D_k^N\}.\tag{26}$$

自然, 称区域 D_k^N 为"停时集合", 称区域 C_k^N 为"继续观测集合". 这些术语的正确性有如下根据.

假设时间 $n=0$, 并将集合 Ω 划分为两个集合 D_0^N 和 C_0^N, 使 $D_0^N\bigcup C_0^N=\Omega,D_0^N\bigcap C_0^N=\varnothing$. 如果 $\omega\in D_0^N$, 则 $\tau_0^N(\omega)=0$. 换句话说, "停止" 发生在时刻 $n=0$. 而若 $\omega\in C_0^N$, 则说明对于这样的 ω, 时间 $\tau_0^N(\omega)\geqslant1$. 对于所考虑的 $\omega\in D_0^N\bigcap C_0^N$ 的情形, 时间 $\tau_0^N(\omega)=1$. 以下各个阶段的情形类似. 在时刻 N 观测自然应该停止.

7. 例 现在举几个例子.

例 1 设序列 $f=(f_0,f_1,\cdots,f_N)$ 是鞅, 其中 $f_0=1$. 那么, 根据 §2 定理 1 的系 1, 对于一切马尔可夫时间 $\tau\in\mathfrak{M}_0^N,\mathbf{E}f_\tau=1$. 因此, 在所考虑的情形下

$$V_0^N=\sup_{\tau\in\mathfrak{M}_0^N}\mathbf{E}f_\tau=1.$$

对于一切 $1\leqslant n\leqslant N$, 函数 $v_n^N=f_n$ 和 $v_0^N=1$. 易见, $\tau_0^N=\min\{0\leqslant k\leqslant N:f_k=v_k^N\}=0$, 而对于任意 $1\leqslant n\leqslant N,\tau_n^N=n$.

于是, 鞅序列的最优停时问题的解法, 实质上是平凡的: 最优停时就是时间 $\tau_0^N(\omega)=0,\omega\in\Omega$ (任何其他时间, 如 $\tau_n^N(\omega)=n,\omega\in\Omega,1\leqslant n\leqslant N$ 也一样).

例 2 设序列 $f=(f_0,f_1,\cdots,f_N)$ 是下鞅, 则对于任何 $\tau\in\mathfrak{M}_0^N$, 有 $\mathbf{E}f_\tau\leqslant\mathbf{E}f_N$ (§2 定理 1). 因此, 这里的最优停时 $\tau^*\equiv N$. 由于 $v_k^N=\mathbf{E}(f_N|\mathscr{F}_k)\geqslant f_k(\mathbf{P}-\text{a.c.})$, 故完全可能停时 $\tau_0^N(\omega)$ 对于某个 $\omega\in\Omega$, 小于 N. 然而, 在任何情形下停时 τ_0^N和停时 $\tau^*\equiv N$ 都是最优的. 虽然停时 $\tau^*\equiv N$ 有简单的构造, 但是停时 τ_0^N 具有一定的优越性: 它在全部可能的最优停时中是*最小*的, 即如果 $\tilde{\tau}$ 也是类 $\mathfrak{M}_0^N$ 中的最优停时, 则 $\mathbf{P}\{\tau_0^N\leqslant\tilde{\tau}\}=1$.

例 3 设序列 $f=(f_0,f_1,\cdots,f_N)$ 是上鞅. 那么, 对于一切 $0\leqslant n\leqslant N$, 有 $v_n^N=f_n$. 从而, 像鞅的情形一样, 最优停时为 $\tau_0^N=0$.

所引进的例子相当简单, 解所考虑的停时的最优性问题, 本质上并未用到在定理 1 和定理 2 中阐述的理论. 只用到如下熟知的结果: 在对马尔可夫时间作时间变换时 (见 §2), 鞅性, 下鞅性和上鞅性的不变性. 然而, 在一般情形下, 寻求价格 V_0^N, 以及最优停时 τ_0^N 的问题, 可能是相当困难的问题.

非常重要意义的是, 函数 f_n 具有如下形式的情形:

$$f_n(\omega)=f(X_n(\omega)),$$

其中 $X=(X_n)_{n\geqslant0}$ 是某一*马尔可夫链*. 由第八章 §9 可见, 这时求解最优停时问题, 实质上归结为求解*变分不等式*, 求解*瓦尔德—贝尔曼* (A. Wald – R. E. Bellman) *动态规划方程*问题.

在第八章的 §9 中还将要给出, 非平凡例子, 在这些例子中, 将给出关于马尔可夫序列最优停时一系列问题的全解.

8. 练习题

1. 设 $\xi(\omega)=\sup\limits_{\alpha\in\mathfrak{U}_0}\xi_\alpha(\omega)$ 是, 在证明第 3 小节的引理时引进的随机变量. 证明满足本质上确界定义中的条件 a) 和 b). (提示: 在 $\alpha\notin\mathfrak{U}_0$ 的情形下考虑 $\mathbf{E}\max(\xi(\omega),\xi_\alpha(\omega))$).

2. 证明, 1) 随机变量 $\xi(\omega)=\tan\widetilde{\xi}(\omega)$ (见第 3 小节引理的证明), 2) 随机变量 $\xi(\omega)$ 也满足条件 a) 和 b).

3. 设 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量序列, 其中 $\mathbf{E}|\xi_1(\omega)|<\infty$. 考虑最优停时问题 (在类 $\mathfrak{M}_1^\infty=\{\tau:1\leqslant\tau<\infty\}$ 中):

$$V^*=\sup_{\tau\in\mathfrak{M}_1^\infty}\mathbf{E}\left(\max_{i\leqslant\tau}\xi_i-c\tau\right).$$

设 $\tau^*=\inf\{n\geqslant1:\xi_n\geqslant A^*\}$, 其中 A^* 是方程 $\mathbf{E}(\xi_1-A^*)=c$ 的唯一解.

证明, 如果 $\mathbf{E}\{\tau^*<\infty\}=1$, 则在 $\mathbf{E}(\max_{i\leqslant\tau}\xi_i(\omega)-c\tau)$ 存在的所有有限停时 τ 类中, 停时 τ^* 是最优的. 同样证明 $V^*=A^*$.

4. 假设在该题和下题中, 有

$$\mathfrak{M}_n^\infty=\{\tau:n\leqslant\tau<\infty\},$$
$$V_n^\infty=\sup_{\tau\in\mathfrak{M}_n^\infty}\mathbf{E}f_\tau,$$
$$v_n^\infty=\operatorname*{ess\,sup}_{\tau\in\mathfrak{M}_n^\infty}\mathbf{E}(f_\tau|\mathscr{F}_n),$$
$$\tau_n^\infty=\inf\{k\geqslant n:v_n^\infty=f_n\}.$$

假设

$$\mathbf{E}\sup f_n^-<\infty,$$

证明, 对于随机变量的极限

$$\widetilde{v}_n=\lim_{N\to\infty}v_n^N,$$

下列命题成立:

(a) 如果 $\tau\in\mathfrak{M}_n^\infty$ 停时, 则

$$\widetilde{v}_n\geqslant\mathbf{E}(f_\tau|\mathscr{F}_n);$$

(b) 如果 $\tau_n^\infty\in\mathfrak{M}_n^\infty$ 停时, 则

$$\widetilde{v}_n=\mathbf{E}(f_{\tau_n^\infty}|\mathscr{F}_n),$$
$$\widetilde{v}_n=v_0^\infty\left(=\operatorname*{ess\,sup}_{\tau\in\mathfrak{M}_n^\infty}\mathbf{E}(f_\tau|\mathscr{F}_n)\right).$$

5. 设 $\tau_n^\infty\in\mathfrak{M}_n^\infty$. 由上题的论断 (a) 和 (b) 证明, 停时 τ_n^∞ 按如下意义是最优的:

$$\operatorname*{ess\,sup}_{\tau\in\mathfrak{M}_n^\infty}\mathbf{E}(f_\tau|\mathscr{F}_n)=\mathbf{E}(f_{\tau_n^\infty}|\mathscr{F}_n)\quad(\mathbf{P}-\text{a.c.}),$$

而

$$\sup_{\tau\in\mathfrak{M}_n^\infty}\mathbf{E}f_\tau=\mathbf{E}f_{\tau_n^\infty},$$

即 $V_n^\infty=\mathbf{E}f_{\tau_n^\infty}$.

第八章　形成马尔可夫链的随机变量序列

现代马尔可夫过程论的源泉, 一方面是, 马尔可夫 (A. A. Марков) 有关 “联系为链” 的试验序列 (1906 年 — 1917 年) 的工作; 另一方面是, 著名布朗运动的物理现象之数学描述的尝试 (巴切利耶 [L. Bachelier], 1900 年; 爱因斯坦 [A. Einstein], 1905 年).

邓肯 (E. Б. Дынкин)《**马尔可夫过程**》**[21]**

§1. 定义和基本性质

1. 引言 在第一章 §12 中, 对于有限概率空间, 阐述了基于对随机变量马尔可夫相依性的理解和原理 (第一章 §12 中性质 (7)), 马尔可夫相依性用来描绘, 具有无后效性系统的演变. 在这一节, 对更加一般的概率空间进行相应的研究.

马尔可夫理论的基本问题之一, 就是研究无后效系统 (随时间的发展) 的渐近性质. 非常值得注意的是, 在相当广泛的条件下, 这样系统的演变似乎 “忘掉了” 它初始的状态, 其状态 “趋于稳定”, 系统进入 “平稳状态”. 接下来对用 “具有可数种状态的马尔可夫链”, 详细地描绘这样系统的渐近状态问题. 为此, 我们需要对 “马尔可夫链” 的状态, 按转移概率的代数性质和渐近性质进行分类.

2. 广义马尔可夫链 假设 $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbf{P})$ 是固定的概率空间, 即具有另外选定附加构造的概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$, 而所谓附加构造就是 $\sigma-$ 代数 $\mathscr{F}_n (n \geqslant 0)$ 的、过滤结构 (流) $(\mathscr{F}_n)_{n\geqslant 0}$, 其中 $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \cdots \subseteq \mathscr{F}$. 直观上, $\mathscr{F}_n$ 是在时间 n 之前 (包括 n) 获得的 “信息”.

设 $(E, \mathscr{E})$ 是某一可测空间, 是所研究的系统取值的状态空间. 由于 “技术” 的原因 (例如, 为了使随机元素 $X_0(\omega)$ 和 $x \in E$ 的集合 $\{\omega : X_0(\omega) = x\}$ 属于 $\mathscr{F}$), 我们假设 $\sigma-$ 代数 $\mathscr{E}$ 包含 E 中的全部单点集.(关于这一假设, 亦见下面的第 6 小节).

在这样的假设下, 可测空间 $(E, \mathscr{E})$ 通常称为 (所研究系统的) 相空间或状态空间.

定义 1 (广义马尔可夫链) 设 $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbf{P})$ 是固定的概率空间, 而 $(E, \mathscr{E})$ 是相空间.

定义在空间 $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbf{P})$ 上、取值于 E 的 $\mathscr{F}/\mathscr{E}-$ 可测随机元 $X_n = X_n(\omega), n \geqslant 0$, 的序列 $X = (X_n)_{n\geqslant 0}$, 称为由马尔可夫相依性联系的广义随机变量序列 (马尔可夫链), 如果对于任意 $n \geqslant 0$ 和 $B \in \mathscr{E}$ 满足如下广义马尔可夫性[1]:

$$\mathbf{P}(X_{n+1} \in B | \mathscr{F}_n)(\omega) = \mathbf{P}(X_{n+1} \in B | X_n(\omega)) \quad (\mathbf{P}-\text{a.c.}). \tag{1}$$

[1] 广义马尔可夫性(wide – sense Markov proberty) 亦称弱马尔可夫性(weak Markov proberty). —— 译者

如果 $\mathscr{F}_n^X=\sigma(X_0,X_1,\cdots,X_n)$ 是由随机变量 $X_0,X_1,\cdots,X_n$ 生成的 $\sigma-$ 代数，则由于 $\mathscr{F}_n^X\subseteq\mathscr{F}_n$，且 X_n 为 $\mathscr{F}_n^X-$ 可测，故由 (1) 式得强马尔可夫性 (或简称马尔可夫性):

$$\mathbf{P}(X_{n+1}\in B|\mathscr{F}_n^X)(\omega)=\mathbf{P}(X_{n+1}\in B|X_n(\omega))\quad(\mathbf{P}-\text{a.c.}).\tag{2}$$

为直观计 (对照第一章 §12)，强马尔可夫性常表示为:

$$\mathbf{P}(X_{n+1}\in B|X_0(\omega),\cdots,X_n(\omega))=\mathbf{P}(X_{n+1}\in B|X_n(\omega))\ (\mathbf{P}-\text{a.c.}).\tag{3}$$

受由 (1) 式引出的强马尔可夫性 (2) 的启发，可见在事先不考虑 $\sigma-$ 代数流 $(\mathscr{F}_n)_{n\geqslant0}$ 的情况下，引进马尔可夫相依性的概念较为适宜.

定义 2 (马尔可夫链) 设 $(\Omega,\mathscr{F},\mathbf{P})$ 是概率空间，而 $(E,\mathscr{E})$ 是相空间. 取值于 E 且 $\mathscr{F}/\mathscr{E}-$ 可测随机元 $X_n=X_n(\omega)$ 序列，$X=(X_n)_{n\geqslant0}$，称为由马尔可夫相依性联系的广义随机变量序列 (马尔可夫链)，如果对于任意 $n\geqslant0$ 和 $B\in\mathscr{E}$，强马尔可夫性 (2) 成立.

注 在一开始引进的过滤概率空间上，在此空间上根据各种不同的"信息流" $(\mathscr{F}_n)_{n\geqslant0}$ 研究系统的性质. 在此空间上定义的广义马尔可夫链，在许多问题中是有效的，例如，可能出现这样的情况: 对于"二维"过程 $(X,Y)=(X_n,Y_n)_{n\geqslant0}$，其第一个分量 $X=(X_n)_{n\geqslant0}$ 按定义 (2) 不是马尔可夫过程，然而按定义 (1) 是马尔可夫过程，其中 $\mathscr{F}_n=\mathscr{F}_n^{X,Y},n\geqslant0$.

在介绍马尔可夫链的初等理论时，通常并不引进 $\sigma-$ 代数流 $(\mathscr{F}_n)_{n\geqslant0}$，而且以定义 2 作为基础. 这一章的内容就是讲马尔可夫链的初等理论.

3. 马尔可夫性 在由序列 $X=(X_n)_{n\geqslant0}$ 描绘的系统状态的演变中，马尔可夫性的特征是"无后效性". 在有限空间 Ω 的情形下，在第一章 §12 中，表示为性质

$$\mathbf{P}(\mathrm{J}|\mathrm{GX})=\mathbf{P}(\mathrm{J}|\mathrm{X}),\tag{4}$$

其中 J —— 将来，G —— 过去，X —— 现在. 这在第一章 §12 中曾经指出，马尔可夫系统还具有如下性质:

$$\mathbf{P}(\mathrm{GJ}|\mathrm{X})=\mathbf{P}(\mathrm{G}|\mathrm{X})\mathbf{P}(\mathrm{J}|\mathrm{X}),\tag{5}$$

表示在"现在"固定的情形下，"过去"和"将来"独立.

在一般情形下，下面的定理，(在定义 2 意义上) 给出了马尔可夫性的各种不同表述，并指出性质 (6) 和 (7) 是性质 (4) 和 (5) 的类似，使用了如下记号:

$$\mathscr{F}_{[0,n]}^X=\sigma(X_0,X_1,\cdots,X_n),$$
$$\mathscr{F}_{[n,\infty)}^X=\sigma(X_n,X_{n+1},\cdots),$$
$$\mathscr{F}_{(n,\infty)}^X=\sigma(X_{n+1},X_{n+2},\cdots).$$

定理 1 马尔可夫性 (2)，等价于如下两条性质中的任意一条:

1) 对于 $n\geqslant0$ 和任意"将来的"事件 $\mathrm{J}\in\mathscr{F}_{[0,\infty]}^X$，有

$$\mathbf{P}(\mathrm{J}|\mathscr{F}_{[0,n]}^X)(\omega)=\mathbf{P}(\mathrm{J}|X_n(\omega))\quad(\mathbf{P}-\text{a.c.}),\tag{6}$$

2) 对于 $n\geqslant1$ 以及任意"将来的"事件 $\mathrm{J}\in\mathscr{F}_{(n,\infty)}^X$ 和"过去的"事件 $\mathrm{G}\in\mathscr{F}_{[0,n-1]}^X$，有

$$\mathbf{P}(\mathrm{GJ}|X_n(\omega))=\mathbf{P}(\mathrm{G}|X_n(\omega))\mathbf{P}(\mathrm{J}|X_n(\omega))\quad(\mathbf{P}-\text{a.c.}).\tag{7}$$

证明 (a) 我们首先证明性质 (6) 和 (7) 的等价性.

(6)⇒(7). $(\mathbf{P}-\text{a.c.})$ 有

$$\begin{aligned}&\mathbf{P}(\mathrm{G}|X_n(\omega))\mathbf{P}(\mathrm{J}|X_n(\omega))=\mathbf{E}(I_\mathrm{G}|X_n(\omega))\mathbf{E}(I_\mathrm{J}|X_n(\omega))\\&=\mathbf{E}\{I_\mathrm{G}\mathbf{E}(I_\mathrm{J}|X_n(\omega)|X_n(\omega))\}=\mathbf{E}\{I_\mathrm{G}\mathbf{E}(I_\mathrm{J}|\mathscr{F}_{[0,n]}^X)(\omega)|X_n(\omega)\}\\&=\mathbf{E}\{\mathbf{E}(I_\mathrm{G}I_\mathrm{J}|\mathscr{F}_{[0,n]}^X)(\omega)|X_n(\omega)\}=\mathbf{E}\{I_\mathrm{G}I_\mathrm{J}|X_n(\omega)\}=\mathbf{P}(\mathrm{GJ}|X_n(\omega)).\end{aligned}$$

(7)⇒(6). 需要证明，对于任意集合 $C\in\mathscr{F}_{[0,n]}^X$，有

$$\mathbf{E}[I_C\mathbf{P}(\mathrm{J}|X_n)]=\mathbf{E}[I_C\mathbf{P}(\mathrm{J}|\mathscr{F}_{[0,n]}^X)].\tag{6'}$$

为此，首先考虑这样集合的特殊情形，即集合 GX，其中 $\mathrm{G}\in\mathscr{F}_{[0,n-1]}^X$ 和 $X\in\sigma(X_n)$，并且证明这时由 (7) 式可得 (6′) 式.

事实上，

$$\begin{aligned}&\mathbf{E}[I_{\mathrm{GX}}\mathbf{P}(\mathrm{J}|X_n)]=\mathbf{E}[I_\mathrm{G}I_\mathrm{X}\mathbf{E}(\mathrm{J}|X_n)]=\mathbf{E}\{I_\mathrm{X}\mathbf{E}[I_\mathrm{G}\mathbf{E}(I_\mathrm{J}|X_n)|X_n]\}\\&=\mathbf{E}\{I_\mathrm{X}\mathbf{E}(I_\mathrm{G}|X_n)\mathbf{E}(I_\mathrm{J}|X_n)\}=\mathbf{E}\{I_\mathrm{X}\mathbf{P}(\mathrm{G}|X_n)\mathbf{P}(\mathrm{J}|X_n)\}\\&\overset{(7)}{=}\mathbf{E}\{I_\mathrm{X}\mathbf{P}(\mathrm{GJ}|X_n)\}=\mathbf{P}(\mathrm{GXJ})=\mathbf{E}[I_{\mathrm{GX}}\mathbf{P}(\mathrm{J}|\mathscr{F}_{[0,n]}^X)],\end{aligned}\tag{8}$$

即对于形如 GX 的集合 C，(6′) 式成立，其中 $\mathrm{G}\in\mathscr{F}_{[0,n-1]}^X$ 和 $\mathrm{X}\in\sigma(X_n)$. 根据"单调类"的性质 (见第二章 §2)，由此可见，对于任意集合 $C\in\mathscr{F}_{[0,n]}^X$，性质 (6′) 成立. 因为函数 $\mathbf{P}(\mathrm{J}|X_n)$ 为 $\mathscr{F}_{[0,n]}^X-$ 可测，所以由 (6′) 式可见. $\mathbf{P}(\mathrm{J}|X_n)$ 是条件概率 $\mathbf{P}(\mathrm{J}|\mathscr{F}_{[0,n]}^X)$ 的变式，于是，性质 (6) 式成立.

(b) 现在证明性质 (2) 和 (6) 等价，因此由已证明的，可见性质 (2) 和 (7) 等价. (6)⇒(2) 显然. 现在证明 (2)⇒(6). 仍然利用"单调类"的性质证明之.

式 (6) 中集合 J 是 $\sigma-$ 代数 $\mathscr{F}_{(n,\infty)}^X=\mathscr{F}_{[n+1,\infty)}^X$ 中的子集，而 $\mathscr{F}_{[n+1,\infty)}^X$ 是代数 $\bigcup\limits_{n=1}^{\infty}\mathscr{F}_{[n+1,n+k]}^X$ 诱导的 $\sigma-$ 代数，其中 $\mathscr{F}_{[n+1,n+k]}^X=\sigma(X_{n+1},\cdots,X_{n+k})$. 因此，自然首先对于 $\sigma-$ 代数 $\mathscr{F}_{[n+1,n+k]}^X$ 中的集合 J 证明性质 (6).

我们用归纳法来证明. 假如 $k=1$，则 $\mathscr{F}_{[n+1,n+1]}^X=\sigma(X_{n+1})$，且 (6) 式恰好是 (2) 式，故命题成立.

现在假设 (6) 式对于某个 $k \geqslant 1$ 成立, 我们证明 (6) 式对于 $k+1$ 成立.

为此取形如 $\mathrm{J}=\mathrm{J}^1+\mathrm{J}^2$ 的集合 $\mathrm{J}\in\mathscr{F}^X_{[n+1,n+k+1]}$, 其中 $\mathrm{J}^1\in\mathscr{F}^X_{[n+1,n+k]}$ 和 $\mathrm{J}^2\in\sigma(X_{n+k+1})$. 那么, 根据归纳法的假设, $(\mathbf{P}-\text{a.c.})$ 有

$$\begin{aligned}\mathbf{P}(\mathrm{J}|\mathscr{F}^X_{[0,n]}) &= \mathbf{E}(I_{\mathrm{J}}|\mathscr{F}^X_{[0,n]}) = \mathbf{E}(I_{\mathrm{J}^1\cap\mathrm{J}^2}|\mathscr{F}^X_{[0,n]}) = \mathbf{E}[I_{\mathrm{J}^1}\mathbf{E}(I_{\mathrm{J}^2}|\mathscr{F}^X_{[0,n+k]}|\mathscr{F}^X_{[0,n]})]\\ &= \mathbf{E}[I_{\mathrm{J}^1}\mathbf{E}(I_{\mathrm{J}^2}|X_{n+k})|\mathscr{F}^X_{[0,n]}] = \mathbf{E}[I_{\mathrm{J}^1}\mathbf{E}(I_{\mathrm{J}^2}|X_{n+k})|X_n]\\ &= \mathbf{E}[I_{\mathrm{J}^1}\mathbf{E}(I_{\mathrm{J}^2}|\mathscr{F}_{[n,n+k]})|X_n] = \mathbf{E}[\mathbf{E}(I_{\mathrm{J}^1}I_{\mathrm{J}^2}|\mathscr{F}_{[n,n+k]})|X_n]\\ &= \mathbf{E}[I_{\mathrm{J}^1}I_{\mathrm{J}^2}|X_n] = \mathbf{P}(\mathrm{J}^1\bigcap\mathrm{J}^2|X_n) = \mathbf{P}(\mathrm{J}|X_n).\end{aligned}\tag{9}$$

这样, 对形如 $\mathrm{J}=\mathrm{J}^1+\mathrm{J}^2$ 的集合 $\mathrm{J}\in\mathscr{F}^X_{[n+1,n+k+1]}$, 其中 $\mathrm{J}^1\in\mathscr{F}^X_{[n+1,n+k]}$ 和 $\mathrm{J}^2\in\sigma(X_{n+k+1})$, 已经证明了性质 (9). 由此可见 (练习题 1a), 性质 (9) 对于任意集合 $\mathrm{J}\in\mathscr{F}^X_{[n+1,n+k+1]}$ 成立. 因而 (练习题 1b), 性质 (9) 对于任意集合 $\mathrm{J}\in\bigcup_{k=1}^{\infty}\mathscr{F}^X_{[n+1,n+k]}$ 代数也成立. 于是, 由此同样地可见 (练习题 1c), 性质 (9) 对于 $\sigma-$ 代数

$$\sigma\left(\bigcup_{k=1}^{\infty}\mathscr{F}^X_{[n+1,n+k]}\right)=\mathscr{F}^X_{(n,\infty)},$$

也成立. □

注 这一证明的论据是 *适当集合原理* 的应用 (首先对于 "简单" 构造的集合进行证明), 然后利用单调类的结果 (第二章 §2). 以后, 这一证明方法还将不止一次地使用 (例如, 见定理 2 和定理 3 的证明, 特别由这些定理的证明, 可以得到以上定理 1 的证明中, 引用练习题 1a, 1b, 1c 的地方).

4. 随机游动 马尔可夫链的经典例子是随机游动 $X=(X_n)_{n\geqslant0}$, 其中

$$X_n=X_0+S_n,\quad n\geqslant1,\tag{10}$$

而 $S_n=\xi_1+\cdots+\xi_n$, 且定义在概率空间 $(\Omega,\mathscr{F},\mathbf{P})$ 上的随机变量 $X_0,\xi_1,\xi_2,\cdots$ 相互独立.

定理 2 *假设 $\mathscr{F}_0=\sigma(X_0),\mathscr{F}_n=\sigma(X_0,\xi_1,\cdots,\xi_n),n\geqslant1$; 而 $(\Omega,\mathscr{F},(\mathscr{F}_n)_{n\geqslant0},\mathbf{P})$ 是过滤概率空间, 且所研究的序列 $X=(X_n)_{n\geqslant0}$ 是该空间上的 (广义及狭义) 马尔可夫链: 对于任意 $n\geqslant0$ 和 $B\in\mathscr{B}(\mathbb{R})$,*

$$\mathbf{P}(X_{n+1}\in B|\mathscr{F}_n)(\omega)=\mathbf{P}(X_{n+1}\in B|X_n(\omega))\quad(\mathbf{P}-\text{a.c.}),\tag{11}$$

并且

$$\mathbf{P}(X_{n+1}\in B|X_n(\omega))=P_{n+1}(B-X_n(\omega))\quad(\mathbf{P}-\text{a.c.}),\tag{12}$$

其中

$$P_{n+1}(A)=\mathbf{P}\{\xi_{n+1}\in A\},\tag{13}$$

而

$$B-X_n(\omega)=\{y:y+X_n(\omega)\in B\},\quad B\in\mathscr{B}(\mathbb{R}).$$

证明 我们同时证明 (11) 式和 (12) 式. 对于离散型概率空间的情形, 在第一章 §12 中曾经作过类似的证明. 初看来, 似乎这里的证明也同样简单. 然而实际上, 我们通过所作证明仍然可以体会到 "什么是证明".

考虑集合 $A\in\{X_0\in B_0,\xi_1\in B_1,\cdots,\xi_n\in B_n\}$, 其中 $B_i\in\mathscr{B}(\mathbb{R}),i=0,1,\cdots,n$. 根据条件概率 $\mathbf{P}(X_{n+1}\in B|\mathscr{F}_n)(\omega)$ 的定义 (见第二章 §7)

$$\begin{aligned}&\int_A\mathbf{P}(X_{n+1}\in B|\mathscr{F}_n)(\omega)\mathbf{P}(\mathrm{d}\omega)=\int_A I_{\{X_{n+1}\in B\}}(\omega)\mathbf{P}(\mathrm{d}\omega)\\ &=\mathbf{P}\{X_0\in B_0,\xi_1\in B_1,\cdots,\xi_n\in B_n,X_{n+1}\in B\}\\ &=\int_{B_0\times B_1\times\cdots\times B_n}P_{n+1}(B-[x_0+x_1+\cdots+x_n])P_0(\mathrm{d}x_0)\cdots P_n(\mathrm{d}x_n)\\ &=\int_A P_{n+1}(B-X_n(\omega))\mathbf{P}(\mathrm{d}\omega).\end{aligned}\tag{14}$$

这样, 对于 $\mathscr{F}_n$ 中形如 $A=\{X_0\in B_0,\xi_1\in B_1,\cdots,\xi_n\in B_n\}$ 的集合, 有

$$\int_A\mathbf{P}(X_{n+1}\in B|\mathscr{F}_n)(\omega)\mathbf{P}(\mathrm{d}\omega)=\int_A P_{n+1}(B-X_n(\omega))\mathbf{P}(\mathrm{d}\omega).\tag{15}$$

显然, 上面的集合A 的集系 $\mathscr{A}_n$ 是 $\pi-$ 系 ($\Omega\in\mathscr{A}_n$ 且若 $A_1\in\mathscr{A}_n$ 和 $A_2\in\mathscr{A}_n$, 则 $A_1\bigcap A_2\in\mathscr{A}_n$; 见第二章 §2 定义 2). 其次, 以 $\mathscr{L}$ 表示一切使 (15) 式成立的集合 $A\in\mathscr{F}_n$ 的全体.

现在证明 $\mathscr{L}$ 是 $\lambda-$ 系 (见第二章 §2 定义 2). 显然 $\Omega\in\mathscr{L}$, 即满足上述定义的性质 (λ_a). 由于勒贝格积分的可加性, 由上述定义也可得性质 (λ_b). 由于勒贝格积分的单调收敛定理 (见第二章 §6), 由 $\lambda-$ 系的定义可以得到第三条性质 (λ_c).

从而, $\mathscr{L}$ 是 $\lambda-$ 系. 运用第二章 §2 定理 2 的命题 c), 可得 $\sigma(\mathscr{A}_n)\subseteq\mathscr{L}$. 注意到 $\sigma(\mathscr{A}_n)=\mathscr{F}_n$, 从而性质 (15) 对于 $A\in\mathscr{F}_n$ 集合也成立.

这样, 注意到 $P_{n+1}(B-X_n(\omega))$ 作为 ω 的函数 $\mathscr{F}_n-$ 可测 (练习题 2), 故 (根据条件概率的定义) 由 (15), 可见 $P_{n+1}(B-X_n(\omega))$ 是条件概率 $\mathbf{P}(X_{n+1}\in B|\mathscr{F}_n)(\omega)$ 的变式. 最后, 由条件数学期望的 "望远性" 性质 (见第二章 §7 性质 $\mathbf{H}^*$) 可见, $(\mathbf{P}-\text{a.c.})$ 有

$$\begin{aligned}\mathbf{P}(X_{n+1}\in B|X_n)(\omega)&=\mathbf{E}[I_{\{X_{n+1}\in B\}}|X_n](\omega)=\mathbf{E}\{\mathbf{E}[(I_{\{X_{n+1}\in B\}}|\mathscr{F}_n)|X_n]\}(\omega)\\ &=\mathbf{E}[P_{n+1}(B-X_n)|X_n(\omega)]=P_{n+1}(B-X_n(\omega)).\end{aligned}\tag{16}$$

于是, 两条性质 (11) 和 (12) 得证. □

注 性质 (11) 和 (12) 的正确性, 或许也可以直接由第二章 §2 引理 3 推出 (练习题 3). 我们对这些 "几乎显然" 的性质作了详细的证明, 在一定意义上是为了演示类似命题的证明技术, 证明基于*适当集合原理*以及*单调类*的有关结果.

5. 广义马尔可夫性 考虑马尔可夫性 (1). 如果 $(E,\mathscr{E})$ 是博雷尔空间, 则由第二章 §7 定理 5 可见, 对于每一个 $n\geqslant 0$, 存在正则条件分布 $P_{n+1}(x;B)$, 使 ($\mathbf{P}$ – a.c.) 有

$$\mathbf{P}(X_{n+1}\in B|X_n(\omega))=P_{n+1}(X_n(\omega);B), \tag{17}$$

其中 $P_{n+1}(x;B),B\in\mathscr{E},x\in E$, 具有如下性质 (见第二章 §7 定义 7):

(a) 对于每一个 x, 集函数 $P_{n+1}(X_n;\cdot)$ 是空间 $(E,\mathscr{E})$ 上的测度;

(b) 对于每一个 $B\in\mathscr{E}$, 函数 $P_{n+1}(\cdot;B)$ 为 $\mathscr{E}$– 可测.

函数 $P_n=P_n(x;B),n\geqslant 1$, 称做转移函数 (亦称马尔可夫核).

特别重要的是下面的情形: 这些转移函数全相等 $P_1=P_2=\cdots$, 确切地说, 对这些条件概率 $\mathbf{P}(X_{n+1}\in B|X_n(\omega))$, $n\geqslant 0$, 存在同一个正则条件分布的变式 $P(x;B)$, 使 ($\mathbf{P}$ – a.c.) 对于一切 $n\geqslant 0,B\in\mathscr{E}$, 有

$$\mathbf{P}(X_{n+1}\in B|X_n(\omega))=P(X_n(\omega);B), \tag{18}$$

如果变式 $P=P(x;B)$ 存在 (那么可以认为对于一切 $n\geqslant 0,P_n=P$), 则马尔可夫链称做 (对时间) 齐次的, 其转移函数为 $P(x;B),x\in E,B\in\mathscr{E}$.

马尔可夫链的齐次性直观意义很明显: 相应系统的运动在如下意义上均匀地进行, 控制系统变化的概率机制在所有时间 $n\geqslant 0$ 都保持不变(在动态系统理论中, 这种性质等同于保守性).

除转移概率 $P_1,P_2,\cdots$ 之外, 而对于齐次马尔可夫链, 除转移概率 P 之外, 马尔可夫链的重要特征是初始分布 $\pi=\pi(B),B\in\mathscr{E}$, 即由等式 $\pi(B)=\mathbf{P}\{X_0\in B\},B\in\mathscr{E}$, 所决定的概率分布.

数组 $(\pi,P_1,P_2,\cdots)$ 完全决定序列 $X=(X_n)_{n\geqslant 0}$ 的概率性质, 因为该序列的一切有限维分布决定于如下公式:

$$\mathbf{P}\{X_0\in B\}=\pi(B),\quad B\in\mathscr{E},$$

而对于任意 $n\geqslant 1$ 和 $B\in\mathscr{B}(E^{n+1})(=\mathscr{E}^{n+1}=\mathscr{E}\otimes\cdots\otimes\mathscr{E}(n+1)$ 次), 有

$$\begin{aligned}&\mathbf{P}\{(X_0,X_1,\cdots,X_n)\in B\}\\&=\int_{E\times\cdots\times E}I_B(x_0,x_1,\cdots,x_n)\pi(\mathrm{d}x_0)P_1(x_0,\mathrm{d}x_1)\cdots P_n(x_{n-1},\mathrm{d}x_n).\end{aligned} \tag{19}$$

事实上, 首先考虑形如 $B=B_0\times\cdots\times B_n$ 的集合 B. 那么, 当 $n=1$ 时, 由全概

率公式 (见第二章 §7(5) 式)

$$\begin{aligned}&\mathbf{P}\{X_0\in B_0,X_1\in B_1\}\\&=\int_\Omega I_{\{X_0\in B_0\}(\omega)}(\omega)\mathbf{P}(X_1\in B_1|X_0(\omega))\mathbf{P}(\mathrm{d}\omega)\\&=\int_\Omega I_{\{X_0\in B_0\}(\omega)}(\omega)P_1(B_1;X_0(\omega))\mathbf{P}(\mathrm{d}\omega)\\&=\int_E I_{B_0}(x_0)P_1(B_1;x_0)\pi(\mathrm{d}x_0)\\&=\int_{E\times E}I_{B_0\times B_1}(x_0,x_1)P_1(\mathrm{d}x_1;x_0)\pi(\mathrm{d}x_0).\end{aligned}$$

下面用归纳法证明:

$$\begin{aligned}&\mathbf{P}\{X_0\in B_0,X_1\in B_1,\cdots,X_n\in B_n\}\\&=\int_\Omega I_{\{X_0\in B_0,\cdots,X_{n-1}\in B_{n-1}\}}(\omega)\mathbf{P}(X_n\in B_n|X_0(\omega),\cdots,X_{n-1}(\omega))\mathbf{P}(\mathrm{d}\omega)\\&=\int_\Omega I_{\{X_0\in B_0,\cdots,X_{n-1}\in B_{n-1}\}}(\omega)\mathbf{P}(X_n\in B_n|X_{n-1}(\omega))\mathbf{P}(\mathrm{d}\omega)\\&=\int_\Omega I_{\{X_0\in B_0,\cdots,X_{n-1}\in B_{n-1}\}}(\omega)P_n(B_n;X_{n-1}(\omega))\mathbf{P}(\mathrm{d}\omega)\\&=\int_{E\times\cdots\times E}I_{B_0\times B_1\times\cdots\times B_{n-1}}(x_0,x_1,\cdots,x_{n-1})\\&\quad\times P_n(B_n;x_{n-1})\mathbf{P}\{X_0\in\mathrm{d}x_1,\cdots,X_{n-1}\in\mathrm{d}x_n\}\\&=\int_{E\times\cdots\times E}I_{B_0\times B_1\times\cdots\times B_{n-1}\times B_n}(x_0,x_1,\cdots,x_{n-1},x_n)\\&\quad\times P_n(\mathrm{d}x_n;x_{n-1})P_{n-1}(\mathrm{d}x_{n-1};x_{n-2})\cdots P_1(\mathrm{d}x_1;x_0)\pi(\mathrm{d}x_0).\end{aligned}$$

这样, 在集合 B 为 $B=B_0\times\cdots\times B_n$ 情形下, 所得结果恰好为 (19) 式. 对于集合 $B\in\mathscr{B}(E^{n+1})$ 的一般情形, 可以用与定理 2 证明中类似地方的方法实现.

基于单调类的结果 (见第二章 §2), 由性质 (19) 可以导出 (练习题 4), 对于任意有界 $\mathscr{B}(E^{n+1})$– 可测函数 $h=h(x_0,x_1,\cdots,x_n)$, 有

$$\begin{aligned}&\mathbf{E}h(X_0,X_1,\cdots,X_n)\\&=\int_{E^{n+1}}h(x_0,x_1,\cdots,x_n)\pi(\mathrm{d}x_0)P_1(\mathrm{d}x_1;x_0)\cdots P_n(\mathrm{d}x_n;x_{n-1}).\end{aligned} \tag{20}$$

6. 数组 $(\pi,P_1,P_2,\cdots)$ 决定的马尔可夫链 对于 (广义或狭义) 马尔可夫链, 利用 (19) 式, 根据其初始分布 $\pi=\pi(B)$, 其中 $\pi(B)=\mathbf{P}\{X_0\in B\},B\in\mathscr{E}$, 以及转移概率 $P_n(x,B),n\geqslant 1,x\in E,B\in\mathscr{E}$, 就可以完整地建立任意随机变量组 $X_1,X_2,\cdots,X_n(n\geqslant 1)$ 的分布律 $\mathrm{Law}(X_0,X_1,\cdots,X_n)$.

我们现在以完全给定的数组 $(\pi,P_1,P_2,\cdots)$ 为基础, 调换关于马尔可夫链的定义全部观点. 从数组 $(\pi,P_1,P_2,\cdots)$ 出发, 把其中 π 的概率意义理解为系统初始状态的

概率分布, 而把满足第 5 小节中定义的性质 (a) 和 (b) 的函数, $P_{n+1}=P_{n+1}(x,B)$, $n\geqslant 0$, 当作转移概率, 即系统在时间 n 处于状态 x, 在时间 $n+1$ 处于集合 $B\in\mathscr{E}$ 的概率. 自然, 假如给定的是数组 $(\pi,P_1,P_2,\cdots)$, 就产生一个问题: 它一般是否对应着某一马尔可夫链, 而且以给定的 π 为其初始分布, 以给定的函数 $P_1,P_2,\cdots$ 为其转移概率?

对这一问题的答案是 "正确". 实际上, 至少对于的 $E=\mathbb{R}^d$ 情形 (见第二章 §9 的定理 1 及其系 3), 答案包含在柯尔莫戈洛夫定理中; 对于任意可测空间 $(E,\mathscr{E})$ 的情形. 包含在 I. 图尔恰 (I. Tulcea) 定理中 (见第二章 §9 的定理 2).

按照这些定理证明, 我们首先定义可测空间 $(\Omega,\mathscr{F})$: 设 $(\Omega,\mathscr{F})=(E^\infty,\mathscr{B}(E^\infty))$, 其中 $E^\infty=E\times E\times\cdots,\mathscr{B}(E^\infty)=\mathscr{E}\otimes\mathscr{E}\otimes\cdots$; 换句话说, 把 "点" $\omega=(x_0,x_1,\cdots)$ 视为基本事件, 其中 $x_i\in E$.

设 $\mathscr{F}_n=\sigma(x_0,x_1,\cdots,x_n)$, 定义流 $(\mathscr{F}_n)_{n\geqslant 0}$. 我们以 "典则" 方式定义 $X_n(\omega)=x_n$ 的值: 若 $\omega=(x_0,x_1,\cdots)$, 则设 $X_n(\omega)=x_n$.

I. 图尔恰定理: 对于任意可测空间 $(E,\mathscr{E})$ (其中也包括所考虑的相空间), 在 $(\Omega,\mathscr{F})$ 上存在概率测度 $\mathbf{P}_\pi$, 使

$$\mathbf{P}_\pi\{X_0\in B\}=\pi(B),\quad B\in\mathscr{E}, \tag{21}$$

而对于一切 $n\geqslant 1$, 其有限维分布为

$$\begin{aligned}&\mathbf{P}_\pi\{(X_0,X_1,\cdots,X_n)\in B\}\\&=\int_E\pi(\mathrm{d}x_0)\int_E P_1(x_0;\mathrm{d}x_1)\cdots\int_E I_B(x_0,\cdots,x_n)P_n(x_{n-1};\mathrm{d}x_n).\end{aligned} \tag{22}$$

定理 3 关于 (在 I. 图尔恰定理中) 引进的测度 $\mathbf{P}_\pi$, 典则给定的随机变量序列 $X=(X_n)_{n\geqslant 0}$ (在定义 2 的意义上) 是马尔可夫序列.

证明 需要证明对于 $n\geqslant 0,B\in\mathscr{E}$, $(\mathbf{P}_\pi-\text{a.c.})$ 有

$$\mathbf{P}_\pi(X_{n+1}\in B|\mathscr{F}_n)(\omega)=\mathbf{P}_\pi(X_{n+1}\in B|X_n(\omega)), \tag{23}$$

而且, 这时对于 $n\geqslant 0$, $(\mathbf{P}-\text{a.c.})$ 有

$$\mathbf{P}_\pi(X_{n+1}\in B|X_n(\omega))=P_n(X_n(\omega);B). \tag{24}$$

仍然利用适当集合原理和单调类的结果 (第二章 §2) 进行证明.

像前面的作法一样, 作为适当集合, 考虑 "简单" 构造的集合 $A\in\mathscr{F}_n$, 其中 A 具有如下形式:

$$A=\{\omega:X_0(\omega)\in B_0,\cdots,X_n(\omega)\in B_n\},$$

而 $B_i\in\mathscr{E},i=0,1,\cdots,n$, 并设 $B\in\mathscr{E}$.

那么, 由于测度 $\mathbf{P}_\pi$ 的构造 (见 (22) 式), 有

$$\begin{aligned}&\int_A I_{\{X_{n+1}\in B\}}(\omega)\mathbf{P}_\pi(\mathrm{d}\omega)=\mathbf{P}_\pi\{X_0\in B_0,\cdots,X_n\in B_n,X_{n+1}\in B\}\\&=\int_{B_0}\pi(\mathrm{d}x_0)\int_{B_1}P_1(x_0;\mathrm{d}x_1)\cdots\int_{B_n}P_n(x_{n-1};\mathrm{d}x_n)\int_B P_{n+1}(x_n;\mathrm{d}x_{n+1})\\&=\int_B P_{n+1}(X_n(\omega);B)\mathbf{P}_\pi(\mathrm{d}\omega).\end{aligned} \tag{25}$$

现在通过像定理 2 的证明类似的讨论 (对于集合 $A\in\mathscr{F}_n$, 见性质 (15) 的证明), 可见这里性质 (25) 对于集合 $A\in\mathscr{F}_n$ 成立, 即对于形如 $A=\{\omega:(X_0(\omega),\cdots,X_n(\omega))\in C\}$ 的集合 $A\in\mathscr{F}_n$ 成立, 其中 $C\in\mathscr{B}(E^{n+1})$.

由于条件概率的定义 (第二章 §7), 可见

$$\int_A I_{\{X_{n+1}\in B\}}(\omega)\mathbf{P}_\pi(\mathrm{d}\omega)=\int_A\mathbf{P}_\pi(X_{n+1}\in B|\mathscr{F}_n)(\omega)\mathbf{P}_\pi(\mathrm{d}\omega), \tag{26}$$

而且函数 $P_{n+1}(x_n(\omega);B)$ 为 $\mathscr{F}_n-$ 可测, 故由 (25) 式和条件数学期望的 "望远" 性质 (见第二章 §7 第 4 小节性质 $\mathbf{H}^*$) 可见, 关系式 (23) 和 (24) 成立. □

7. 马尔可夫链 (族) 上一节证明了, 与每一个给定的数组 $(\pi,P_1,P_2,\cdots)$, 有一个马尔可夫链与之相联系 (为了直观, 记作 $X^\pi=(X_n,\mathbf{P}_\pi)_{n\geqslant 0}$), 并且以 π 为其初始分布, 而以 $P_1,P_2,\cdots$ 为其转移概率 (即满足性质 (21), (23) 和 (24) 的链).

在初始时间 $n=0$, 根据分布 π 随机地 "抽签决定" 初始状态的值. 例如, 结果 X_0 的值等于 x, 则系统在下一时间按照分布 $P_1(\cdot;x)$, 系统自该状态转移到某个状态 x_1, 等等.

这样, 初始分布 π 的作用仅仅表现在时间 $n=0$, 而后系统的发展就决定于转移概率 $P_1,P_2,\cdots$ 因此, 假如对于两个初始分布 π_1 和 π_2, 相应 "抽签" 的结果为同一状态 x, 那么系统的发展 (在概率意义上) 将是一样的, 都仅决定于转移概率 $P_1,P_2,\cdots$ 亦可将这一情况表示如下.

仍以 $(E,\mathscr{E})$ 表示相空间. 假设以 $\mathbf{P}_x$ 表示对应于如下情况的分布 $\mathbf{P}_\pi$: 分布 π 集中在点 x, 而 $\pi(\mathrm{d}y)=\delta_x(\mathrm{d}y)$, 即 $\pi(\{x\})=1$, 其中 $\{x\}$ 是属于 $\sigma-$ 代数 $\mathscr{E}$ 的单点集.

那么, 由性质 (22) 可见 (练习题 4), 对于每一个 $A\in\mathscr{B}(E^\infty)$ 和 $x\in E$, (对于每一个 π) 概率 $\mathbf{P}_x(A)$ 是条件概率 $\mathbf{P}_\pi(A|X_0=x)$ 的变式, 即 $\mathbf{P}_\pi-\text{a.c.}$ 有

$$\mathbf{P}_\pi(A|X_0=x)=P_x(A). \tag{27}$$

对于每一个 $x\in E$, 概率 $\mathbf{P}_x(\cdot)$ 完全决定于转移概率 $(P_1,P_2,\cdots)$ 组.

于是, 假如基本着眼点是系统对转移概率 $(P_1,P_2,\cdots)$ 的依赖关系, 则只需利用概率 $\mathbf{P}_x(\cdot),x\in E$, 而且如果需要得到概率 $\mathbf{P}_\pi(\cdot)$, 则只需进行简单的积分运算:

$$\mathbf{P}_\pi(A)=\int_E\mathbf{P}_x(A)\pi(\mathrm{d}x),\quad A\in\mathscr{B}(E^\infty). \tag{28}$$

这些思路导致如下情形: "马尔可夫过程的一般理论" 中 (见 [21]), 对于这里研究的离散时间的情形, 认为基本研究对象并不是哪个具体的马尔可夫链, 而是马尔可夫链族 $X^x = (X_n, \mathbf{P}_x)_{n\geqslant 0}, x \in E$. (不过, 通常不提 "马尔可夫链族", 仍然简称 "马尔可夫链", 并且通常使用记号 " $X = (X_n, \mathscr{F}_n, \mathbf{P}_x)$", 而不是 " $X^x = (X_n, \mathbf{P}_x)_{n\geqslant 0}, x \in E$".)

需要强调, 这些论述都假设, 是用 "典则" 方法建立的链: 以 $(E^\infty, \mathscr{E}^\infty)$ 做 $(\Omega, \mathscr{F})$, 其中 $\mathscr{E}^\infty = \mathscr{E} \otimes \mathscr{E} \otimes \cdots$ 而若 $\omega = (x_0, x_1, \cdots)$, 则一切随机变量 $X_n(\omega)$ 定义为 $X_n(\omega) = x_n$. 这样, 在 $X^x = (X_n, \mathbf{P}_x)$ 中只有 $\mathbf{P}_x$ 依赖于 x, 而不假设 X_n 本身的值对 x 的任何特别的依赖条件, 在这种情况下, 按测度 $\mathbf{P}_x$ 轨道 $(x_n)_{n\geqslant 0}$ 自动从点 x "开始", 即 $\mathbf{P}_x\{X_0 = x\} = 1$.

8. 柯尔莫戈洛夫 – 查普曼方程 (任意状态空间) 对于有限马尔可夫链的情形 (第一章 §12), 主要注意力放在, 利用转移概率 $p_{ij}^{(n)} = \mathbf{P}(X_n = j|X_0 = i)$ 的研究, 分析这种链的行为和性质方面, 证明了 $p_{ij}^{(n)}$ 满足柯尔莫戈洛夫 – 查普曼 (D. G. Chapman) 方程 (见第一章 §12 的 (13) 式), 并且由此同样的可以得到柯尔莫戈洛夫前向方程和后向方程 (第一章 §12 的 (15) 和 (16) 式).

我们现在讨论, 在具有任意相空间 $(E, \mathscr{E})$ 的情形下, 柯尔莫戈洛夫 – 查普曼方程正确性的问题, 但是只局限于齐次链的情形, 即 $P_1 = P_2 = \cdots = P$ 的情形.

在这种情形下, 由 (22) 式, 有

$$\begin{aligned}&\mathbf{P}_\pi\{(X_0, X_1, \cdots, X_n) \in B\}\\ &= \int_E \pi(dx_0) \int_E P(x_0; dx_1) \cdots \int_E I_B(x_0, x_1, \cdots, x_n) P(x_{n-1}; dx_n).\end{aligned} \tag{29}$$

特别, 如果 $n = 2$, 则

$$\mathbf{P}_\pi\{X_0 \in B_0, X_2 \in B_2\} = \int_{B_0} \int_E P(x_1; B_2) P(x_0; dx_1) \pi(dx_0). \tag{30}$$

由此, 根据拉东 – 尼科迪姆定理 (第二章 §6), 以及条件概率的定义, (π – a.c.) 有

$$\mathbf{P}_\pi\{X_2 \in B_2 | X_0 = x\} = \int_E P(x; dx_1) P(x_1; B_2). \tag{31}$$

现在注意到, 由于 (27) 式, 可见

$$\mathbf{P}_\pi\{X_2 \in B_2 | X_0 = x\} = \mathbf{P}_x\{X_2 \in B_2\} \quad (\pi - \text{a.c.}),$$

其中概率 $\mathbf{P}_x\{X_2 \in B_2\}$ 有简单的含义: 系统由在时间 $n = 0$ 状态 x, 于时间 $n = 2$ 转移到状态集合 B_2 的概率, 即经两步转移的概率.

记 $P^{(n)}(x; B_n) = \mathbf{P}_x\{X_n \in B_n\}$ 是经 n 步转移的概率. 那么, 由于所考虑的链的齐性, 有 $P^{(1)}(x; B_1) = P(x; B_1)$, 从而由 (31) 式, 可见 ($\pi$ – a.c.) 有

$$P^{(2)}(x; B) = \int_E P^{(1)}(x; dx_1) P^{(1)}(x_1; B), \tag{32}$$

其中 $B \in \mathscr{E}$.

类似地可以证明 (练习题 5), 对于任意 $n \geqslant 0, m \geqslant 0$, ($\pi$– a.c.) 有

$$P^{(n+m)}(x; B) = \int_E P^{(n)}(x; dy) P^{(m)}(y; B). \tag{33}$$

该式就是著名的

柯尔莫戈洛夫 – 查普曼方程,

其直观含义十分清楚: 为计算系统 "自点 $x \in E$ 经过 $m + n$ 步转移到集合 $B \in \mathscr{E}$ 中的概率 $P^{(n+m)}(x; B)$", 需要将 "自点 x 经过 n 步转移到点 $y \in E$ 的 '无穷小' 区域 dy 概率" $P^{(n)}(x; dy)$, 乘以 "自点 y 经过 m 步转移到集合 B 的概率", 然后对可能的 "中间点" y 积分.

柯尔莫戈洛夫 – 查普曼方程 (33), 把不同转移次数的概率相联系. 应该注意, 该式仅仅是精确到 " π– 几乎处处" 成立. 特别, 由此可见, (33) 式并不是对于一切 $x \in E$ 都成立的关系式. 不应当觉得这是怪异的, 因为我们在前面不止一次地处理选择条件概率各种不同变式 (异说) 的问题. 一般不能指望, 所研究的性质 "对于这些变式 (关于 x) 恒成立", 而事实上 "仅仅 π– 几乎必然 (处处) 成立".

然而, 可以指出这样一些变式, 使柯尔莫戈洛夫 – 查普曼方程 (33) 已经对于一切 $x \in E$ 成立.

这由如下命题 (见练习题 6) 可以得到.

设 "转移概率" $P^{(n)}(x; B)$ 定义为:

$$P^{(1)}(x; B) = P(x; B),$$

而对于 $n > 1$,

$$P^{(n)}(x; B) = \int_E P(x; dy) P^{(n-1)}(y; B).$$

那么,

(i) 对每一个给定的 x, $P^{(n)}(x; B), n \geqslant 1$, 是 $\mathscr{E}$ 上的正则条件概率;

(ii) $P^{(n)}(x; B)$ 等于 $\mathbf{P}_x\{X_n \in B\}$, 从而 $P^{(n)}(x; B)$ (π– a.c.) 是条件概率 $\mathbf{P}_\pi(X_n \in B | X_0 = x)$ 的变式;

(iii) 对于这样定义的函数 $P^{(n)}(x; B), n \geqslant 1$, 柯尔莫戈洛夫 – 查普曼方程 (33) 对于一切 $x \in E$ 恒成立.

9. 练习题

1. 证明在定理 1 的证明中提到的问题 1a, 1b, 1c 的命题,
2. 证明在定理 2 中的函数 $P_{n+1}(B - X_n(\omega))$ 对 ω 是 $\mathscr{F}_n$– 可测的.
3. 证明第二章 §2 引理 3 的命题中性质 (11) 和 (12).
4. 证明性质 (20) 和 (27).

5. 证明关系式 (33).

6. 证明第 8 小节中的命题 (i), (ii), (iii).

7. 问由马尔可夫性 (3) 是否可以导出如下性质:

$$\mathbf{P}(X_{n+1} \in B | X_0 \in B_0, X_1 \in B_1, \cdots, X_n \in B_n) = \mathbf{P}(X_{n+1} \in B | X_n \in B_n),$$

其中 $B_0, B_1, \cdots, B_n$ 和 B 是属于 $\mathscr{E}$ 的集合, 而 $\mathbf{P}\{X_0 \in B_0, X_1 \in B_1, \cdots, X_n \in B_n\} > 0$.

§2. 推广马尔可夫性和强马尔可夫性

1. 推广马尔可夫性 这一节主要研究齐次马尔可夫链族 $X^x = (X_n, P_x)_{n\geqslant 0}, x \in E$, 而链是 "典则地" 定义在坐标空间 $(\Omega, \mathscr{F}) = (E^\infty, \mathscr{E}^\infty)$ 上并且决定于转移函数 $P(x; B), x \in E, B \in \mathscr{E}$.

在 $(\Omega, \mathscr{F})$ 上定义**推移算子** $\theta_n : \Omega \to \Omega$ (对照第五章 §1), 对状态 $\omega = (x_0, x_1, \cdots)$, 设

$$\theta_n(\omega) = (x_n, x_{n+1}, \cdots).$$

如果 $H = H(\omega)$ 是 $\mathscr{F}-$ 可测函数, 则以 $H \circ \theta_n$ 表示由等式

$$(H \circ \theta_n)(\omega) = H(\theta_n(\omega)) \tag{1}$$

定义的函数 $(H \circ \theta_n)(\omega)$. 从而, 如果 $\omega = (x_0, x_1, \cdots)$ 和 $H = H(x_0, x_1, \cdots)$, 则

$$(H \circ \theta_n)(x_0, x_1, \cdots) = H(x_n, x_{n+1}, \cdots).$$

下面的定理实质上是 §1 中的 (6) 应用于现在考虑的齐次马尔可夫链族命题的重新表述.

定理 1 *设 $X^x = (X_n, \mathbf{P}_x)_{n\geqslant 0}, x \in E$ 是由转移函数 $P(x; B), x \in E, B \in \mathscr{E}$ 生成的齐次马尔可夫链族. 假设对于 $B \in \mathscr{B}(E^{n+1})$ 和 $n \geqslant 0$, 由 §1 中的公式 (22) 由 P 决定测度 P_x 的值 $\mathbf{P}_x\{(X_0, X_1, \cdots, X_n) \in B\}$, 其中 $\pi(\mathrm{d}y) = \delta_{(x)}(\mathrm{d}y)$, 而 $P_1 = P_2 = \cdots = P$.*

那么, 对于任意初始分布 π, 任何 $n \geqslant 0$ 以及任意有界 (或非负) $\mathscr{F}-$ 可测函数 $H = H(\omega)$, 有如下的推广马尔可夫性[①]:

$$\mathbf{E}_\pi(H \circ \theta_n | \mathscr{F}_n^X) = \mathbf{E}_{X_n(\omega)} H \quad (\mathbf{P}_\pi - \text{a.c.}). \tag{2}$$

注 虽然所使用记号自身 "无须解释", 然而我们还是指出 $\mathbf{E}_\pi$ 是按测度

$$\mathbf{P}_\pi(\cdot) = \int_E \mathbf{P}_x(\cdot) \pi(\mathrm{d}x)$$

①这里将要介绍的广义马尔可夫性 (обобщённое марковское свойство) 与一般广义马尔可夫性有所不同, 因此我们将其译为 "推广马尔可夫性", 以示区别. —— 译者

求平均, 而 $\mathbf{E}_{X_n(\omega)} H$ 应该按如下的方式理解: 求数学期望 $\mathbf{E}_x H$ (即按测度 $\mathbf{P}_x$ 求对 H 平均), 而然后向此式 (记作 $\psi(x)$) 中 x 的位置 "代入" 随机变量 $X_n(\omega)$, 即 $\mathbf{E}_{X_n(\omega)} H = \psi(X_n(\omega))$. (注意, $\mathbf{E}_x H$ 是 x 的 $\mathscr{E}-$ 可测函数 (练习题 1), 故 $\mathbf{E}_{X_n(\omega)} H$ 是随机变量, 即 $\mathscr{F}/\mathscr{E}-$ 可测函数). 定理的证明仍然利用*适当集合与函数原理*, 然后运用*单调类*的结果.

为证明性质 (2), 我们需要验证对属于 $\mathscr{F}_n^X = \sigma(x_0, x_1, \cdots, x_n)$ 的任意集合 A, 有

$$\int_A (H \circ \theta_n)(\omega) \mathbf{P}_\pi(\mathrm{d}\omega) = \int_A (\mathbf{E}_{X_n(\omega)} H) \mathbf{P}_\pi(\mathrm{d}\omega), \tag{3}$$

或者对于更加紧凑的形式, 有

$$\mathbf{E}_\pi(H \circ \theta_n; A) = \mathbf{E}_\pi(\mathbf{E}_{X_n} H; A), \tag{4}$$

其中 $\mathbf{E}_\pi(\xi; A)$ 表示 $\mathbf{E}_\pi(\xi I_A)$ 适当集合与函数原理 (见第二章 §6 第 2 小节).

根据*适当集合原理*, 我们考虑形如 $A = \{\omega : X_0 \in B_0, \cdots, X_n \in B_n\} (B_i \in \mathscr{E}_i)$ 的 "简单" 构成集合 A; 考虑函数 $H = H(x_0, x_1, \cdots, x_m) (m \geqslant 0)$ (更确切地说, 设 H 是 $\mathscr{F}_n^X-$ 可测函数). 那么, 性质 (4) 有如下形式:

$$\mathbf{E}_\pi[H(X_n, X_{n+1}, \cdots, X_{n+m}); A] = \mathbf{E}_\pi[\mathbf{E}_{X_n} H(X_0, X_1, \cdots, X_m); A]. \tag{5}$$

利用 §1 中的表示 (22), 可得

$$\begin{aligned}
&\mathbf{E}_\pi[H(X_n, X_{n+1}, \cdots, X_{n+m}); A] = \mathbf{E}_\pi[I_A(X_0, \cdots, X_n) H(X_n, \cdots, X_{n+m})] \\
&= \int_{E^{n+m+1}} I_A(x_0, \cdots, x_n) H(x_n, \cdots, x_{n+m}) \pi(\mathrm{d}x_0) P(x_0; \mathrm{d}x_1) \cdots P(x_{n+m-1}; \mathrm{d}x_{n+m}) \\
&= \int_{E^{n+1}} I_A(x_0, \cdots, x_n) \pi(\mathrm{d}x_0) P(x_0; \mathrm{d}x_1) \cdots P(x_{n-1}; \mathrm{d}x_n) \\
&\qquad \times \left[\int_{E^m} H(x_n, \cdots, x_{n+m}) P(x_n; \mathrm{d}x_{n+1}) \cdots P(x_{n+m-1}; \mathrm{d}x_{n+m})\right] \\
&= \int_{E^{n+1}} I_A(x_0, \cdots, x_n) \pi(\mathrm{d}x_0) P(x_0; \mathrm{d}x_1) \cdots P(x_{n-1}; \mathrm{d}x_n) \\
&\qquad \times \left[\int_{E^m} H(x_0, \cdots, x_m) \mathbf{P}_x(\mathrm{d}x_1, \cdots, \mathrm{d}x_m)\right] \\
&= \mathbf{E}_\pi[\mathbf{E}_{X_n} H(X_0, X_1, \cdots, X_m); A],
\end{aligned}$$

其中 $\mathbf{P}_x(\mathrm{d}x_1, \cdots, \mathrm{d}x_m) = P(x_0; \mathrm{d}x_1) P(x_1; \mathrm{d}x_2) \cdots P(x_{m-1}; \mathrm{d}x_m)$.

这样, 对形如 $A = \{\omega : x_0 \in B_0, \cdots, x_n \in B_n\}$ 的集合 A, 和函数 $H = H(x_0, x_1, \cdots, x_m)$, 性质 (5) 得证. 集合 $A \in \mathscr{F}_n^X$ 的一般情形 (对于固定的 m) 的证明, 同 §1 中定理 2 的证明是一样的.

只剩下证明, 已证明的性质对于一切 $\mathscr{F}(= \mathscr{E}^\infty)-$ 可测有界函数 $H = H(x_0, x_1, \cdots)$ 也仍然成立.

为证明这一事实, 只需证明, 如果 $A \in \mathscr{F}_n^X$, 则对于函数

$$\mathbf{E}_\pi[H(X_n, X_{n+1}, \cdots); A] = \mathbf{E}_\pi[\mathbf{E}_{X_n} H(X_0, X_1, \cdots); A], \tag{6}$$

性质 (5) 成立.

为了利用适当集合与函数原理 (第二章 §2), 以 $\mathscr{H}$ 表示满足性质 (5) 的一切有界 $\mathscr{F}-$ 可测函数 $H = H(x_0, x_1, \cdots)$ 的全体.

设 J 是形如 $I_m = \{\omega : x_0 \in B_0, \cdots, x_m \in B_m\}$ 的集合的全体, 其中 $B_i \in E, i = 0, 1, \cdots, m (m) \geqslant 0$. 显然, 这一集系 J 是 $\mathscr{F}(\mathscr{E}^\infty)$ 中集合的 $\pi-$ 系.

现在考虑第二章 §2 中定理 3 的条件.

条件 (h_1) 成立, 因为当 $A \in J$ 时, 根据上面已证明的结果有 $I_A \in \mathscr{H}$ (需要在 (5) 式中设 $H(x_0, \cdots, x_m) = I_A(x_0, \cdots, x_m)$). 由勒贝格积分的可加性, 可得条件 (h_2), 而由勒贝格积分中的单调收敛定理, 可以得到现在 (h_3).

根据上述定理 3(第二章 §2), $\mathscr{H}$ 包含一切关于 $\sigma-$ 代数 $\sigma(J)$ 的可测函数, 而根据定义 $\sigma(J)$ 就是 $\sigma-$ 代数 $\mathscr{E}^\infty = \mathscr{B}(E^\infty)$ (见第二章 §2 的第 4 和第 8 小节).

2. 强马尔可夫性的另一种推广 现在考虑马尔可夫性的第二种推广, 即将 "时间 n" 换成 "随机时间 τ" 的强马尔可夫性. (这一节开始引进的所有前提条件保持不变, 例如, $(\Omega, \mathscr{F}) = (E^\infty, \mathscr{E}^\infty), \cdots$)

以 $\tau = \tau(\omega)$ 表示有限随机变量 $\tau(\omega)$, 且对于每一个 $n \geqslant 0$, 满足

$$\{\omega : \tau(\omega) = n\} \in \mathscr{F}_n^X.$$

按照第七章 §1 引进的术语 (见定义 3), 这样的随机变量称做 (有限) *马尔可夫时间*或*停止时间*.

将停时 τ 与 $\sigma-$ 代数流 $(\mathscr{F}_n^X)_{n \geqslant 0}$:

$$\mathscr{F}_\tau^X = \{A \in \mathscr{F}^X : A \bigcap \{\tau = n\} \in \mathscr{F}_n^X \quad 对于一切\ n \geqslant 0\},$$

相联系, 其中 $\mathscr{F}^X = \sigma(\bigcup \mathscr{F}_n^X)$ 是在 "随机区间" $[0, \tau]$ 上观测到的事件的 $\sigma-$ 代数.

定理 2 *设定理 1 中提出的全部条件成立, 而 $\tau = \tau(\omega)$ 是有限马尔可夫时间. 那么, 有如下强马尔可夫性:*

$$\mathbf{E}_\pi(H \circ \theta_\tau | \mathscr{F}_\tau^X) = \mathbf{E}_{X_\tau} H \quad (\mathbf{P}_\pi - \text{a.c.}). \tag{7}$$

在进行证明之前, 首先关于应当如何理解 $\mathbf{E}_{X_\tau} H$ 和 $H \circ \theta_\tau$, 作一些说明.

记 $\psi(x) = \mathbf{E}_x H$. (在第 1 小节已经指出, $\psi(x)$ 是 $\mathscr{E}-$ 可测函数.) 把 $\mathbf{E}_{X_\tau} H$ 理解为 $\psi(X_\tau) = \psi(X_{\tau(\omega)}(\omega))$ 的值. 关于 $(H \circ \theta_\tau)(\omega)$, 应理解为 $(H \circ \theta_{\tau(\omega)})(\omega) = H(\theta_{\tau(\omega)}(\omega))$.

证明 设集合 $A \in \mathscr{F}_\tau$. 像定理 1 一样, 为证明 (7) 式, 需要证明

$$\mathbf{E}_\pi(H \circ \theta_\tau; A) = \mathbf{E}_\pi(\mathbf{E}_{X_\tau} H; A). \tag{8}$$

等式的左侧为

$$\begin{aligned} \mathbf{E}_\pi(H \circ \theta_\tau; A) &= \sum_{n=0}^{\infty} \mathbf{E}_\pi(H \circ \theta_\tau; A \bigcap \{\tau = n\}) \\ &= \sum_{n=0}^{\infty} \mathbf{E}_\pi(H \circ \theta_n; A \bigcap \{\tau = n\}). \end{aligned} \tag{9}$$

由 (8) 式的右侧, 得

$$\mathbf{E}_\pi(\mathbf{E}_{X_\tau} H; A) = \sum_{n=0}^{\infty} \mathbf{E}_\pi(\mathbf{E}_{X_n} H; A \bigcap \{\tau = n\}). \tag{10}$$

事件 $A \bigcap \{\tau = n\} \in \mathscr{F}_n^X$. 由于 (4) 式, (9) 式和 (10) 式的右侧相等, 可见强马尔可夫性 (7) 得证. □

系 *设函数 $H(x_0, x_1, \cdots) = I_A(x_0, x_1, \cdots)$, 而 $A = \{\omega : (x_0, x_1, \cdots) \in B\}, B \in \mathscr{E}^\infty = \mathscr{B}(E^\infty)$, 则由 (7) 式得强马尔可夫性的如下常用形式:*

$$\begin{aligned} &\mathbf{P}_\pi\{\omega : (X_\tau, X_{\tau+1}, \cdots) \in B | X_0, X_1, \cdots, X_\tau\} \\ &= \mathbf{P}_{X_\tau}\{(X_0, X_1, \cdots) \in B\} \quad (\mathbf{P}_\pi - \text{a.c.}). \end{aligned} \tag{11}$$

注 1 如果分析一下强马尔可夫性 (7) 式的证明, 就会注意到, 实际上有如下性质.

设对于任意 $n \geqslant 0$, 定义在 $\Omega = E^\infty$ 上的实函数 $H_n = H_n(\omega)$ 为 $\mathscr{F}-$ 可测的 $(\mathscr{F} = \mathscr{E}^\infty)$, 并且一致有界 (即 $|H_n(\omega)| \leqslant c, n \geqslant 0, \omega \in \Omega$). 那么, 对于每一个有限马尔可夫时间 $\tau = \tau(\omega) (\tau(\omega) < \infty, \omega \in \Omega)$, 有如下形式的强马尔可夫性:

$$\mathbf{E}_\pi(H_\tau \circ \theta_\tau | \mathscr{F}_\tau^X) = \psi(\tau, X_\tau) \quad (\mathbf{P}_\pi - \text{a.c.}). \tag{12}$$

其中 $\psi(n, x) = \mathbf{E}_x H_n$, 而 $H_\tau \circ \theta_\tau = (H_\tau \circ \theta_\tau)(\omega) = H_{\tau(\omega)}(\theta_{\tau(\omega)}(\omega))$.

注 2 上面假设 $\tau = \tau(\omega)$ 是有限马尔可夫时间. 假如不是这样, 即 $\tau(\omega) \leqslant \infty, \omega \in \Omega$, 那么 (12) 式应当改为如下形式 (练习题 3):

$$\mathbf{E}_\pi(H_\tau \circ \theta_\tau | \mathscr{F}_\tau^X) = \psi(\tau, X_\tau) \quad (\{\tau < \infty\}; \mathbf{P}_\pi - \text{a.c.}). \tag{13}$$

换句话说, 在此情形下, (12) 式在集合 $\{\tau < \infty\}$ 上 $\mathbf{P}_\pi - \text{a.c.}$ 成立,

3. 强马尔可夫性的例 在研究重对数定律时, 我们曾经用到一个不等式 (第四章 §4 中的引理 1; 亦见下面 (14) 式), 对于布朗运动 $B = (B_t)_{t \leqslant T}$, 等式

$$\mathbf{P}\left\{\max_{0 \leqslant t \leqslant T} B_t > a\right\} = 2\mathbf{P}\{|B_T| > a\}$$

是上述不等式的类似 (见 [131] 的第 III 章).

设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量序列, 并且各随机变量共同的分布 (关于 0) 对称. 记 $X_0 = x \in \mathbb{R}, X_m = X_0 + (\xi_1 + \cdots + \xi_m), m \geqslant 1$. 像前面一样, 以 $\mathbf{P}_x$ 表示序列 $X = (X_m)_{m \geqslant 0}, (X_0 = x)$ 的概率分布. (假设空间 $\Omega = \{\omega\}$, 是由坐标给定的, 其中 $\omega = (x_0, x_1, \cdots)$, 而 $X_m(\omega) = x_m$.)

由 (很容易更改的) 第四章 §4 引理 1 的 (9) 式, 可见对于任意 $a > 0$, 有

$$\mathbf{P}_0\left\{\max_{0 \leqslant m \leqslant n} X_m > a\right\} \leqslant 2\mathbf{P}_0\{X_n > a\}. \tag{14}$$

引进马尔可夫时间 $\tau = \tau(\omega)$, 设

$$\tau(\omega) = \inf\{0 \leqslant m \leqslant n : X_m(\omega) > a\}. \tag{15}$$

(一般, 设 $\inf \varnothing = \infty$.) 我们现在说明, 假如允许像处理确定性变量一样, 对待这样的 (随机) 时间, 应如何利用所引进的马尔可夫时间, 或许可以给予不等式 (14) "很容易的证明". (对照第四章 §4 引理 1 的证明) 易见

$$\begin{aligned}\mathbf{P}_0\{X_n > a\} &= \mathbf{P}_0\{(X_n - X_{\tau\wedge n}) + X_{\tau\wedge n} > a\} \\ &\geqslant \mathbf{P}_0\{X_n - X_{\tau\wedge n} \geqslant 0, X_{\tau\wedge n} > a\} = \mathbf{P}_0\{X_n - X_{\tau\wedge n} \geqslant 0\}\mathbf{P}_0\{X_{\tau\wedge n} > a\} \\ &\geqslant \frac{1}{2}\mathbf{P}_0\{X_{\tau\wedge n} > a\} = \frac{1}{2}\mathbf{P}_0\{\tau \leqslant n\} = \frac{1}{2}\mathbf{P}_0\left\{\max_{0 \leqslant m \leqslant n} X_m > a\right\},\end{aligned} \tag{16}$$

其中我们利用了 "'仿佛' 几乎显然的" 性质: 随机变量 $X_n - X_{\tau\wedge n}$ 和 $X_{\tau\wedge n}$ 相互独立. 当然, 对于确定性时间 τ 这是对的; 然而, 对于随机时间 τ, 一般说来这是不对的 (练习题 4). (于是, 这里的 "很容易的证明" 不能认为是适当的.)

我们现在基于强马尔可夫性 (13) 式的应用, 引进不等式 (14) 的真正 "正确的证明".

由于 $\{X_n > a\} \subseteq \{\tau \leqslant n\}$, 可见

$$\mathbf{P}_0\{X_n > a\} = \mathbf{E}_0(I_{\{X_n > a\}}; \tau \leqslant n). \tag{17}$$

引进函数 $H_m = H_m(x_0, x_1, \cdots)$, 设

$$H_m(x_0, x_1, \cdots) = \begin{cases} 1, & \text{若 } m \leqslant n, \text{ 且 } x_{n-m} > a, \\ 0, & \text{其他.} \end{cases}$$

由相应的定义可见, 在集合 $\{\tau \leqslant n\}$ 上, 有

$$(H_\tau \circ \theta_\tau)(x_0, x_1, \cdots) = \begin{cases} 1, & \text{若 } x_n > a, \\ 0, & \text{其他,} \end{cases} \tag{18}$$

因此, 注意到 (17) 式, 由于 $\{X_n > a\} \subseteq \{\tau \leqslant n\}$, 以及 $\{\tau \leqslant n\} \in \mathscr{F}_\tau$, 可得

$$\mathbf{P}_0\{X_n > a\} = \mathbf{E}_0(H_\tau \circ \theta_\tau; \tau \leqslant n) = \mathbf{E}_0[\mathbf{E}_0(H_\tau \circ \theta_\tau | \mathscr{F}_\tau^X); \tau \leqslant n]. \tag{19}$$

由强马尔可夫性 (13), 在集合 $\{\tau \leqslant \infty\}$ 上, 有

$$\mathbf{E}_0(H_\tau \circ \theta_\tau | \mathscr{F}_\tau) = \psi(\tau, X_\tau) \quad (\mathbf{P}_0 - \text{a.c.}). \tag{20}$$

根据定义 $\psi(m, x) = \mathbf{E}_x H_m$, 而对于 $x \geqslant a$,

$$\mathbf{E}_x H_m = \mathbf{P}_x\{X_{n-m} > a\} \geqslant \mathbf{P}_x\{X_{n-m} > x\} \geqslant \frac{1}{2}$$

(由于随机变量 $\xi_1, \xi_2, \cdots$ 的分布的对称性, 得最后的不等式).

于是, 在集合 $\{\tau \leqslant n\}$ 上

$$\mathbf{E}_0(H_\tau \circ \theta_\tau | \mathscr{F}_\tau) \geqslant \frac{1}{2} \quad (\mathbf{P}_0 - \text{a.c.}). \tag{21}$$

由此以及由 (19), (20) 式得所要求证明的不等式 (14).

4. 柯尔莫戈洛夫 – 查普曼方程 如果考虑柯尔莫戈洛夫 – 查普曼方程 (13) 和方程 (38) (第一章 §12), 则可以注意到它们非常相似. 自然应分析它们的表述与它们的结论中的共同点和差异. (我们仅限于讨论具有离散状态集合 $\boldsymbol{E}$ 的齐次马尔可夫链.)

对于 $n \geqslant 1, 1 \leqslant k \leqslant n, i, j \in E$, (注意到 (1) 式和 (2) 式), 有

$$\begin{aligned}\mathbf{P}_i\{X_n = j\} &= \sum_{\alpha \in E} \mathbf{P}_i\{X_n = j, X_k = \alpha\} + \sum_{\alpha \in E} \mathbf{E}_i I(X_n = j) I(X_k = \alpha) \\ &= \sum_{\alpha \in E} \mathbf{E}_i\{\mathbf{E}_i[I(X_n = j) I(X_k = \alpha) | \mathscr{F}_k]\} = \sum_{\alpha \in E} \mathbf{E}_i\{I(X_k = \alpha)\mathbf{E}_i[I(X_{n-k} = j) | \mathscr{F}_k]\} \\ &\overset{(1)}{=} \sum_{\alpha \in E} \mathbf{E}_i\{I(X_k = \alpha)\mathbf{E}_i[I(X_{n-k} = j) \circ \theta_k | \mathscr{F}_k]\} \\ &\overset{(2)}{=} \sum_{\alpha \in E} \mathbf{E}_i\{I(X_k = \alpha)\mathbf{E}_{X_k} I(X_{n-k} = j)\} \\ &= \sum_{\alpha \in E} \mathbf{E}_i\{I(X_k = \alpha)\mathbf{E}_\alpha I(X_{n-k} = j)\} = \sum_{\alpha \in E} \mathbf{E}_i I(X_k = \alpha)\mathbf{E}_\alpha I(X_{n-k} = j) \\ &= \sum_{\alpha \in E} \mathbf{P}_i(X_k = \alpha)\mathbf{P}_\alpha(X_{n-k} = j),\end{aligned} \tag{22}$$

这就是柯尔莫戈洛夫 – 查普曼方程 (13), 而在第一章 §12 中表示为

$$p_{ij}^{(n)} = \sum_{\alpha \in E} p_{i\alpha}^{(k)} p_{\alpha j}^{(n-k)}.$$

假如在 (22) 式中, 将时间 k 换成 (取 $1, 2, \cdots, n$ 为值的) 马尔可夫时间 τ, 并且将采用马尔可夫性 (2) 换成使用强马尔可夫性 (7), 则得 (练习题 5) 柯尔莫戈洛夫 – 查普曼方程如下自然的形式:

$$\mathbf{P}_i\{X_n = j\} = \sum_{\alpha \in E} \mathbf{P}_i\{X_\tau = \alpha\}\mathbf{P}_\alpha\{X_{n-\tau} = j\}. \tag{23}$$

在 (22) 式和在 (23) 式中都是对相变量 $\alpha \in E$ 求和. 在第一章 §12 的 (38) 式中是对时间变量求和.

注意到上述情况, 现在假设 τ 是取 $1, 2, \cdots, n$ 为值的马尔可夫时间. 像上述 (38) 式一样, 我们开始推导下面的式子, 有

$$\begin{aligned}
\mathbf{P}_i\{X_n = j\} &= \sum_{k=1}^n \mathbf{P}_i\{X_n = j, \tau = k\} + \mathbf{P}_i\{X_n = j, \tau \geqslant n+1\} \\
&= \sum_{k=1}^n \mathbf{E}_i I(X_n = j) I(\tau = k) + \mathbf{P}_i\{X_n = j, \tau \geqslant n+1\} \\
&= \sum_{k=1}^n \mathbf{E}_i\{\mathbf{E}_i[I(X_n = j) I(\tau = k) | \mathscr{F}_k]\} + \mathbf{P}_i\{X_n = j, \tau \geqslant n+1\} \\
&= \sum_{k=1}^n \mathbf{E}_i\{I(\tau = k) \mathbf{E}_i I[(X_n = j) | \mathscr{F}_k]\} + \mathbf{P}_i\{X_n = j, \tau \geqslant n+1\} \\
&= \sum_{k=1}^n \mathbf{E}_i\{I(\tau = k) \mathbf{E}_i[I(X_{n-k} = j) \circ \theta_k | \mathscr{F}_k]\} + \mathbf{P}_i\{X_n = j, \tau \geqslant n+1\} \\
&= \sum_{k=1}^n \mathbf{E}_i\{I(\tau = k) \mathbf{E}_i I(X_{n-k} = j)\} + \mathbf{P}_i\{X_n = j, \tau \geqslant n+1\}.
\end{aligned} \tag{24}$$

在第一章 §12 的第 7 小节, 如果集合 $\{\cdot\} = \varnothing$, 则时间 $\tau = \tau_j$, 其中

$$\tau_j = \min\{1 \leqslant k \leqslant n : X_k = j\}$$

满足条件 $\tau_j = n+1$. 从而, 在这种情形下, (24) 式自然可以简化为:

$$\begin{aligned}
\mathbf{P}_i\{X_n = j\} &= \sum_{k=1}^n \mathbf{E}_i[I(\tau_j = k) \mathbf{E}_{X_{\tau_j}} I(X_{n-k} = j)] \\
&= \sum_{k=1}^n \mathbf{E}_i[I(\tau_j = k) \mathbf{E}_j I(X_{n-k} = j)] \\
&= \sum_{k=1}^n \mathbf{E}_i I(\tau_j = k) \mathbf{E}_j I(X_{n-k} = j) \\
&= \sum_{k=1}^n \mathbf{P}_i\{\tau_j = k\} \mathbf{P}_j\{X_{n-k} = j\},
\end{aligned}$$

成为第一章 §12 的 (38) 式

$$p_{ij}^{(n)} = \sum_{k=1}^n f_{ij}^{(k)} p_{jj}^{(n-k)}. \tag{25}$$

在 (24) 式中也可以得到另一有用的公式, 其中 (与柯尔莫戈洛夫 – 查普曼方程不同) 是对时间变量求和. 例如, 假设马尔可夫时间

$$\tau(\alpha) = \min\{1 \leqslant k \leqslant n : X_k = \alpha(k)\},$$

以及 (确定性) 函数 $\alpha = \alpha(k), 1 \leqslant k \leqslant n$, 而马尔可夫链满足条件: 对于给定的 i 和 $n, \mathbf{P}_i\{\tau(\alpha) \leqslant n\} = 1$. 那么, 由 (24) 式可见

$$\begin{aligned}
\mathbf{P}_i\{X_n = j\} &= \sum_{k=1}^n \mathbf{E}_i[I(\tau(\alpha) = k) \mathbf{E}_{X_{\tau(\alpha)}} I(X_{n-k} = j)] \\
&= \sum_{k=1}^n \mathbf{E}_i I(\tau(\alpha) = k) \mathbf{E}_{\alpha(k)} I(X_{n-k} = j),
\end{aligned}$$

即 (对照 (23) 式)

$$\mathbf{P}_i\{X_n = j\} = \sum_{k=1}^n \mathbf{P}_i\{\tau(\alpha) = k\} \mathbf{P}_{\alpha(k)}\{X_{n-k} = j\}.$$

5. 练习题

1. 证明第 1 小节的注中的函数 $\psi(x) = \mathbf{E}_x H$ 为 $\mathscr{E}-$ 可测.
2. 证明性质 (12).
3. 证明性质 (13).
4. 问第 3 小节的例子中, 随机变量 $X_n - X_{\tau \wedge n}$ 和 $X_{\tau \wedge n}$ 是否独立?
5. 证明性质 (23).

§3. 马尔可夫链的极限、遍历和平稳概率分布问题

1. 广义马尔可夫性 我们在 §1 中已经指出, 用马尔可夫链描绘的*无后效*随机系统的*渐近性质*问题, 是马尔可夫过程理论的中心课题之一. 这在一定程度上与下面的情况有关, 在相当广泛的条件下, 这样的系统好像趋向于 “*稳定*”, 进入平稳 “*状态*”.

可以从不同的角度研究齐次马尔可夫链 $X = (X_n)_{n \geqslant 0}$ 的极限性质. 例如, 可以像狭义平稳序列的遍历性定理那样 (第五章 §3 定理 3), 对不同函数 $f = f(x)$, 研究当 $n \to \infty$ 时形如 $n^{-1} \sum\limits_{m=0}^{n-1} f(X_m)$ 泛函 $\mathbf{P}_\pi-$ 几乎必然收敛问题. 像第一章那样, 大数定律成立的条件也很重要.

进一步的叙述, 主要的注意力, 并不放在诸如*几乎必然收敛*或*依概率收敛*上, 而放在当 $n \to \infty$ 时 n *步转移概率* $P^{(n)}(x; A)$ 的*渐近性质*的问题上 (见 §1 的 10 式), 以及*平稳 (不变) 测度* $q = q(A)$ 存在性的问题上, 其中测度

$$q(A) = \int P(x; A) q(\mathrm{d}x), \tag{1}$$

其中 $P(x; A)$ 是 (一步) 转移函数.

需要强调, 在 (1) 式定义中, 一般将完全*没有*假定测度 $q = q(A)$ 是概率测度 (没有要求 $q(E) = 1$).

假如 $q=q(A)$ 是概率测度, 则习惯上称之为*平稳分布*或*不变分布*. 这一术语的含义是完全清楚的: 如果把分布 q 取为初始分布 π, 即认为 $\mathbf{P}_q\{X_0\in A\}=q(A)$, 则由 (1) 式知, 对于任意 $n\geqslant 1,\mathbf{P}_q\{X_n\in A\}=q(A)$, 即该分布随时间是*不变的*.

不难举出不存在平稳分布 $q=q(A)$, 但是*存在平稳测度*的例子.

例 设 $X=(X_n)_{n\geqslant 0}$ 是由伯努利概型生成的马尔可夫链, 即 $X_{n+1}=X_n+\xi_{n+1}$, 其中 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量序列: $\mathbf{P}\{\xi_n=+1\}=p,\mathbf{P}\{\xi_n=-1\}=q$. 设 $X_0=x$, 其中 $x\in\{0,\pm1,\cdots\}$. 显然这时转移函数为

$$P(x;\{x+1\})=p,\quad P(x;\{x-1\})=q.$$

不难验证, 对于任意 $x\in\{0,\pm1,\cdots\}$, 满足 $q(\{x\})=1$ 测度 $q(A)$, 是 (1) 式的解之一. 如果 $p\neq q$, 则满足 $q(\{x\})=(p/q)^x$ 测度 $q(A)$ 是第二不变测度. 显然, 这些测度都非*概率测度*, 而且这里不存在概率不变测度.

这一简单的例子说明, 为*存在*平稳 (不变) 分布, 需要所研究的马尔可夫链满足一定条件. 当 $n\to\infty$ 时, 关于转移概率 $P^{(n)}(x;A)$ 的*极限值*的问题之所以重要, 首先是因为*不依赖于初始状态 x 的极限的存在问题*. 这时应该注意到完全可能出现任何极限分布都不存在的情形. 例如, 可能出现这样的情形: 对于任何 $A\in\mathscr{E}$ 和任意初始状态 $x\in E,\lim P^{(n)}(x;A)=0$. 例如, 在上面的例子中只需设 $p=1$, 即考虑确定性的向右运动 (亦见 §8 中的例 4 和例 5; 对照 §5 的练习题 6).

对于*任意*相空间 $(E,\mathscr{E})$ 的情形, 寻找平稳 (不变) 分布, 以及寻找 (具有某种性质的) 转移概率极限值的分布, 是相当困难的课题 (例如, 见 [104]). 然而, 对于*可数*状态空间 (“可数马尔可夫链”) 的情形, 这里得到了重要的和相当明显表达的结果. 有关结果将在 §6 和 §7 中介绍. 不过, 我们需要首先对可数状态马尔可夫链, 按转移概率的代数性质和渐近性质, 进行详细的分类.

需要指出, 所考虑的关于平稳分布和极限 $\lim P^{(n)}(x;A)$ 的存在性问题, 二者密切相关. 事实上, 如果 $\lim P^{(n)}(x;A)(=v(A))$ *存在, 与 x 无关*, 而且 (对于 $A\in\mathscr{E}$) 是*测度*, 则由*柯尔莫戈洛夫 – 查普曼方程*

$$P^{(n+1)}(x;A)=\int P^{(n)}(x;\mathrm{d}y)P(y;A)$$

当 $n\to\infty$ 时, 经 (形式地) 极限过程, 得

$$v(A)=\int P(y;A)v(\mathrm{d}y).$$

于是, $v=v(A)$ 是*平稳 (不变) 测度*.

2. 要研究的主要问题 以下到处假设, 所考虑的马尔可夫链 $X=(X_n)_{n\geqslant 0}$, 在*可数*相空间 $E=\{1,2,\cdots\}$ 取值. 为简便计, 我们将用 $p_{ij}(i,j\in E)$ 表示转移函数 $P(i,\{j\})$. (为直观计, 求游动 “质点”) 从状态 i 到状态 j 的转移概率记为 $p_{ij}^{(n)}$.

我们关心的主要问题涉及如下条件的说明:

A. 对于一切 $j\in E$, *存在不依赖于初始状态 $i\in E$ 的极限*

$$\pi_j=\lim_n p_{ij}^{(n)};$$

B. 这些 $\Pi=(\pi_1,\pi_2,\cdots)$ 的极限值是*概率分布*, 即 $\pi_j\geqslant 0,\sum\limits_{j\in E}\pi_j=1$;

C. 链 $X=(X_n)_{n\geqslant 0}$ 是*遍历的*, 换句话说, 极限值 $\Pi=(\pi_1,\pi_2,\cdots)$ 满足条件一切 $\pi_j>0,\sum\limits_{j\in E}\pi_j=1$;

D. *平稳 (不变) 概率分布* $\mathbb{Q}=(q_1,q_2,\cdots)$ 存在并且唯一, 即 $\mathbb{Q}=(q_1,q_2,\cdots)$ 满足条件 $q_j\geqslant 0,\sum\limits_{j\in E}q_j=1$, 并且对于一切 $j\in E$, 有

$$q_j=\sum_{i\in E}q_ip_{ij}.$$

注 这里用到的术语 “遍历性” 在第五章已经出现过 (*遍历性*作为度量的可递性, 辛钦 – 博赫纳*遍历性定理*). 从字面上讲, 这些词的含义涉及不同的对象, 然而它们也有共同点, 就是它们都反映, 在时间参数趋向无穷的情况下, 各种概率特征的极限性质.

3. 练习题

1. 列举马尔可夫链的例子, 使得 $\pi_j=\lim\limits_n p_{ij}^{(n)}$ 存在, 并且: (a) 不依赖于初始状态 j; (b) 依赖于初始状态 j.
2. 列举遍历链和非遍历链的例子.
3. 举出不是遍历分布的平稳分布例子.

§4. 马尔可夫链的状态按转移概率矩阵的代数性质分类

1. 转移概率矩阵 假设所考虑的马尔可夫链有*可数个*状态集 $E=\{1,2,\cdots\}$, 而其转移概率为 $p_{ij},i,j\in E$. 以 $\mathbb{P}=(p_{ij})$ 表示这些转移概率构成的矩阵 (表), 或者用展开的形式将 $\mathbb{P}$ 写成

$$\mathbb{P}=\begin{pmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ \vdots & \vdots & \vdots & \\ p_{i1} & p_{i2} & p_{i3} & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

下面要引进的马尔可夫链状态的分类, 完全决定于转移概率矩阵 $\mathbb{P}$ 及其 $n\geqslant 1$ 次幂 $\mathbb{P}^{(n)}$ 的代数性质.

转移概率矩阵 $\mathbb{P}$ 完全决定于状态到状态的一步转移. 而由于马尔可夫性, 矩阵 $\mathbb{P}^{(n)}=(p_{ij}^{(n)})$ 决定 n 步的转移. 例如, 矩阵

$$\mathbb{P}=\begin{pmatrix}1/2 & 1/2\\ 0 & 0\end{pmatrix}$$

和与其相对应的图形 (第一章 §12) 表明, 它所决定的 "质点" 的运动是这样的: "质点" 在状态 0 和 1 上游动: 经一步以概率可能 $0\to 1$ 移动 (以概率 1/2), 但是移动 $1\to 0$ 不可能. 由 n 步转移概率矩阵

$$\mathbb{P}^{(n)}=\begin{pmatrix}2^{-n} & 1-2^{-n}\\ 0 & 1\end{pmatrix}$$

可见对于任意 $n\geqslant 1, p_{10}^{(n)}=0$. 因此, 移动 $1\to 0$ 显然不可能, 可见无论经过任何步移动, $1\to 0$ 当然也不可能.

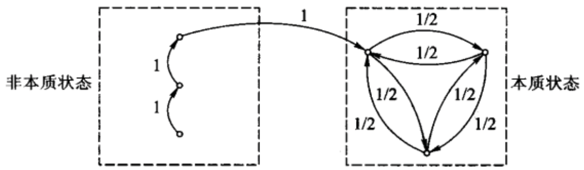


图 36

在该例中从状态 0 进入状态 1, 但是不可能再越出状态 1.

由图 36, 可见根据该图很容易复原转移概率矩阵 $\mathbb{P}$. 由此图的形态可以看出, 这里有三个状态 (图的左半部分), 从某一状态越出后, 就再也不能返回.

从按该图形游动 "质点" 的 "将来" 的性质看, 这三个状态*不重要* (故称之为*非本质状态*), 因为, 从这样的状态*可以越出*, 但是*不可能再返回*.

这样的 "非本质" 状态不值得研究, 立即可以去掉, 并把全部注意力放在剩下的 "本质" 状态的分类上. (可以用转移概率性质的语言: $p_{ij}^{(n)}, i,j\in E, n\geqslant 1$, 给 "本质" 状态和 "非本质" 状态的描述性定义以确切的形式, 练习题 1.)

2. 可达状态和可通状态 为将本质状态或这样的状态组分类, 需要用到如下概念.

定义 1 称状态 j 为由状态 i *可达的*, 记作 $i\to j$, 如果存在 $n\geqslant 0$, 使 $p_{ij}^{(n)}>0$ (若 $i=j$, 则 $p_{ij}^{(0)}=1$; 若 $i\neq j$, 则 $p_{ij}^{(0)}=0$).

称状态 i 和 j 为*互通的*, 记作 $i\leftrightarrow j$, 如果 $i\to j$ 且 $j\to i$, 即状态 i 和 j *互为可达的*.

引理 1 *互通性 $i\leftrightarrow j$ (相互可达性) 是 (转移矩概率阵为 $\mathbb{P}$ 的马尔可夫链的) 状态的等价关系.*

证明 根据等价关系 $i\leftrightarrow j$ 的定义, 需要验证其*自反性* ($i\leftrightarrow i$), *对称性* (即若 $i\leftrightarrow j$, 则 $j\leftrightarrow i$) 和*可递性* (即若 $i\leftrightarrow j, j\leftrightarrow k$, 则 $i\leftrightarrow k$).

前两条性质可以由状态的*互通性*得到. 关于可递性由可尔莫戈洛夫 -- 查普曼方程, 当 $p_{ij}^{(n)}>0, p_{jk}^{(m)}>0$ 时, 由

$$p_{jk}^{(n+m)}=\sum_{l\in E}p_{il}^{(n)}p_{lk}^{(m)}\geqslant p_{ij}^{(n)}p_{jk}^{(m)}>0,$$

可见 $i\to k$. 类似地 $k\to i$. 于是 $i\leftrightarrow k$. □

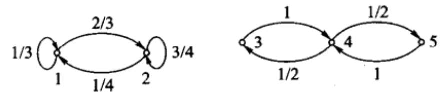
我们把所有互通状态 $i,j,k,\cdots(i\leftrightarrow j, j\leftrightarrow k, k\leftrightarrow i,\cdots)$ 都归为一类. 那么, 所有这样的状态类或者重合, 或者不相交. 从而, *互通性*关系把 (本质) 状态的整个集合 E, 分割为有限或可数个不相交集合 $E_1,E_2,\cdots(E=E_1+E_2+\cdots)$.

我们把这些集合 $E_1,E_2,\cdots$ 称为 (本质连通) 状态的*不可约类*. 所有状态形成一个不可约类的马尔可夫链称做*不可约的*.

为演示上面引进的概念, 我们考虑马尔可夫链: 其状态空间为 $E=\{1,2,3,4,5\}$, 而转移概率矩阵为

$$\mathbb{P}^{(n)}=\left(\begin{array}{cc|ccc}1/3 & 2/3 & 0 & 0 & 0\\ 1/4 & 3/4 & 0 & 0 & 0\\ \hline 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 1/2 & 0 & 1/2\\ 0 & 0 & 0 & 1 & 0\end{array}\right)=\left(\begin{array}{c|c}\mathbb{P}_1 & 0\\ \hline 0 & \mathbb{P}_2\end{array}\right).$$

这一链具有 5 个状态的图有如下形状:



显然, 该马尔可夫链有*两个不可约的状态类*: $E_1=\{1,2\}$ 和 $E_2=\{3,4,5\}$. 研究该链的性质归结为, 研究两个链中每一个链的性质: 两个链的状态空间相应为 E_1 和 E_2, 而转移概率矩阵分别等于 $\mathbb{P}_1$ 和 $\mathbb{P}_2$.

现在考虑任何一个不可约状态类 E. 作为例子, 假设将状态类描述在图 37 中. 注意, 这里只有经过*偶数步*返回每一个状态才是可能的, 经*奇数步*只能转移到相邻状态, 而转移概率矩阵具有分块的结构:

$$\mathbb{P}=\left(\begin{array}{cc|cc}0 & 0 & 1/2 & 1/2\\ 0 & 0 & 1/2 & 1/2\\ \hline 1/2 & 1/2 & 0 & 0\\ 1/2 & 1/2 & 0 & 0\end{array}\right)$$

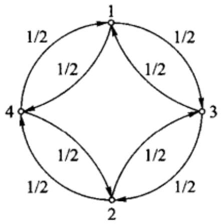


图 37 周期 $d=2$ 的马尔可夫链

由此可见状态类 $E=\{1,2,3,4\}$ 划分为两个子类 $C_0=\{1,2\}$ 和 $C_1=\{3,4\}$, 且具有如下周期性: 质点一步从子类 C_0 越出, 一定转移到子类 C_1, 而由 C_1 转移到 C_0.

3. 按周期对状态分类 上面的例子表明, 看来在一般情形下, 也可以进行不可约状态类按周期子类的分类.

为此需要引进某些概念, 以及数论中的一个结果.

定义 2 设 $\varphi=(\varphi_1,\varphi_2,\cdots)$ 是某一非负数 $\varphi_n\geqslant 0(n\geqslant 1)$ 的序列; M_φ 是使 $\varphi_n>0(n\geqslant 1)$ 的下标 n 的集合, 并且当 $\varphi_n=0(n\geqslant 1)$ 时, 设 $M_\varphi=\varnothing$, 而设 $\mathrm{GCD}(M_\varphi)=0$, 其中 $\mathrm{GCD}(M_\varphi)$ 是集合 M_φ 的最大公约数①. 那么, 称数

$$d(\varphi)=\mathrm{GCD}\{n\geqslant 1:\varphi_n>0\}$$

为序列 φ 的周期.

换句话说, 称序列 φ 的周期为 $d(\varphi)$, 如果由 $\varphi_n>0$ 可见 n 被 $d(\varphi)$ 整除 (即对于某个 $k\geqslant 1$, n 应具有 $d(\varphi)k$ 的形式), 则 $d(\varphi)$ 在一切具有该性质 (即对于某个整数 $l\geqslant 1, n=dl$) 的数中是最大的.

例如, 假如序列 $\varphi=(\varphi_1,\varphi_2,\cdots)$ 满足:

$$\varphi_n=\begin{cases}\varphi_{4k}>0, & 若\ n=4k,k=1,2,\cdots,\\ 0, & 若\ n\neq 4k,\end{cases}$$

则序列 φ 的周期为 $d(\varphi)=4$; 虽然 $\varphi_{2l}>0(l=2,4,8)$, 但是 $d(\varphi)\neq 2$.

定义 3 称序列 $\varphi=(\varphi_1,\varphi_2,\cdots)$ 为非周期的, 如果其周期 $d(\varphi)=1$.

下面在根据周期性对状态分类时, 将要用到数论中的如下初等结果.

引理 2 设 M 是关于加法运算封闭的非负整数的集合 ($M\subseteq E$), 并且 $\mathrm{GCD}(M)=1$. 那么, 对于某一 n_0, 所有数 $n\geqslant n_0$ 都属于集合 M.

把序列 $\varphi=(\varphi_1,\varphi_2,\cdots)$, 取为序列 $(p_{jj}^{(1)},p_{jj}^{(2)},\cdots)$ 或 $(p_{jj}^{(d)},p_{jj}^{(2d)},\cdots)$, $d\geqslant 1$, 其中 j 是马尔可夫链的某个状态, 而马尔可夫链转移概率矩阵为 $\mathbb{P}=(p_{ij})$, 且 $p_{jj}^{(n)}$ 是矩阵 $\mathbb{P}^{(n)}(n\geqslant 1)$ 的元素, $\mathbb{P}^{(1)}=\mathbb{P}$. (这时, 如果 $d(j)$ 是序列 $(p_{jj}^{(1)},p_{jj}^{(2)},\cdots)$ 的周期, 则称状态 j 有周期 $d(j)$.) 那么, 就有如下结果.

①GCD是 "Greatest Common Divisor" (最大公约数) 的缩写. —— 译者

定理 1 设状态 j 的周期 $d=d(j)$.

如果 $d=1$, 则存在 $n_0=n_0(j,d)$, 使对所有 $n\geqslant n_0$ 转移概率 $p_{jj}^{(n)}>0$.

如果 $d>1$, 则存在 $n_0=n_0(j,d)$, 使对所有 $n\geqslant n_0$ 转移概率 $p_{jj}^{(nd)}>0$.

如果 $d\geqslant 1$ 且对于某个 $i\in E$ 和 $m>1, p_{ii}^{(m)}>0$, 则存在 $n_0=n_0(j,d,m)$, 使对所有 $n\geqslant n_0$ 转移概率 $p_{ij}^{(m+nd)}>0$.

现在给出一定理, 它说明不可约类的状态的周期, 具有 "单一类型" 的特性.

定理 2 设 $E_*=\{i,j,\cdots\}$ 是集合 E 中 (互通) 状态的某一不可约类.

这样状态类的一切状态按如下意义是 "单一类型的": 它们有同一周期 (记作 $d(E_*)$, 并称为 E_* 类的周期.)

证明 设 $i,j\in E_*$, 则存在 k 和 l, 使 $p_{ij}^{(k)}>0$ 和 $p_{ji}^{(l)}>0$. 而根据柯尔莫戈洛夫–查普曼方程

$$p_{ii}^{(k+l)}=\sum_{\alpha\in E}p_{i\alpha}^{(k)}p_{\alpha i}^{(l)}\geqslant p_{ij}^{(k)}p_{ji}^{(l)}>0,$$

因此 $k+l$ 应被状态 $i\in E_*$ 的周期 $d(i)$ 整除. 设 $d(j)$ 是状态 $j\in E_*$ 的周期, 由于 $d(j)=\mathrm{GCD}\{n:p_{jj}^{(n)}>0\}$, 可见 n 应该被 $d(j)$ 整除, 而因为

$$p_{ii}^{(n+k+l)}\geqslant p_{ij}^{(k)}p_{jj}^{(n)}p_{ji}^{(l)}>0,$$

故 $n+k+l$ 被 $d(i)$ 整除; 由于 $k+l$ 被整除, 而且 $d(j)=\mathrm{GCD}\{n:p_{ij}^{(n)}>0\}$, 因而 $d(i)\leqslant d(j)$.

由对称性有 $d(j)\leqslant d(i)$, 于是 $d(i)=d(j)$. □

4. 不可约马尔可夫链 假如状态集合组成 (互通状态的) 不可约类 $E_*\subseteq E$, 且 $d(E_*)=1$, 则这样的类常称做非周期状态类.

现在考虑 $d(E_*)>1$ 的情形.

在这样的类内, 由状态到状态的转移, 可以以相当出奇别致的方式实现 (就像上面讨论过的、周期为 $d(E_*)=2$ 的马尔可夫链的例子一样; 见图 37). 然而, 结果表明在由一组状态向另一组状态的转移中, 具有完全的 "周期性 (循环性)".

定理 3 设 $E_*(E_*\subseteq E)$ 是周期为 $d=d(E_*)>1$ 的不可约状态类.

那么, 存在称为循环子类的 d 个状态组 $C_0,C_1,\cdots,C_{d-1}(E_*=C_0+C_1+\cdots+C_{d-1})$, 其特征是: 在时间 $n=p+kd(p=0,1,\cdots,d-1;k=0,1,\cdots)$, "质点" 将处于子类 C_p 中, 并且在下一时间进入 C_{p+1}, 然后进入 $C_{p+2},\cdots$, 进入 C_{d-1}, 由 C_{d-1} 进入 C_0 等.

证明 固定某个状态 $i_0\in E_*$, 并引进如下子类:

$$C_0=\{j\in E_*:\ 若\ p_{i_0j}^{(n)}>0,\ \ 则\ n=kd,\ \ k=0,1,\cdots\},$$

$$C_1=\{j\in E_*:\ 若\ p_{i_0j}^{(n)}>0,\ \ 则\ n=kd+1,k=0,1,\cdots\},$$

$$\cdots\cdots\cdots$$

$$C_{d-1}=\{j\in E_*:\ 若\ p_{i_0j}^{(n)}>0,\ \ 则\ n=kd+(d-1),\ \ k=0,1,\cdots\}.$$

显然, $E_* = C_0 + C_1 + \cdots + C_{d-1}$. 现在证明, "质点" 从子类到子类的运动是按定理 3 所描述的方式进行的; 见图 38.

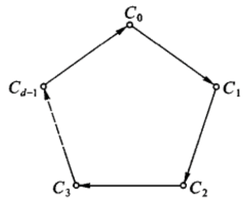
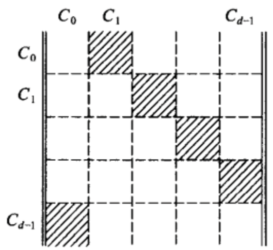


图 38 在循环子类中的运动

事实上, 我们考虑某个状态 $i \in C_p$, 并假设对于状态 $j \in E_*, p_{ij} > 0$, 则证明必有 $j \in C_{(p+1)(\mathrm{mod})}$.

假设 n 使 $p_{i_0 j}^{(n)} > 0$. 那么, n 可以表示为 $n = p + kd, p = 0, 1, \cdots, d-1$ 和 $k = 0, 1, \cdots$, 故 $n \equiv p\ (\mathrm{mod})$, 所以 $n+1 \equiv p+1\ (\mathrm{mod})$. 因此 $p_{i_0 j}^{(n+1)} > 0$ (根据周期的定义 $d = d(E_*)$), 故 $j \in C_{(p+1)(\mathrm{mod})}$, 而这正是要求证明的. □

需要指出, 由以上的讨论可见, 转移概率矩阵 $\mathbb{P}$ 具有分块结构:



现在假设游动 "质点" 的运动由矩阵 $\mathbb{P}$ 控制, "质点" 自子类 C_0 的某种状态开始运动. 那么, (由于子类 $C_0, C_1, \cdots, C_{d-1}$ 的定义) 此 "质点" 在 $n = p + kd$ 中的每一时间处于集合 C_p 中.

从而, 与每一个这样的状态集合 C_p, 都可以与一新的马尔可夫链相联系, 其转移概率矩阵为 $(p_{ij}^{(d)}), i, j \in C_p$. 这一新马尔可夫链是*不可约的和非周期的*.

于是, 由上述 (非本质状态和本质状态, 以及不可约类和循环子类) 的分类 (见综合图 39), 可以得出这样的结论:

设 $p_{ij}^{(n)}, n \geqslant 1, i, j \in E$, 是决定 "马尔可夫质点" 游动的转移概率. 在研究上述转移概率的极限性质时, 可以仅局限于 "相空间 E 本身是*唯一不可约非周期状态类*" 情形的研究.

在此假设下, 具有这样相空间和转移概率矩阵 $\mathbb{P}$ 的马尔可夫链 $X = (X_n)_{n \geqslant 0}$ 称做*不可约的和非周期的*.

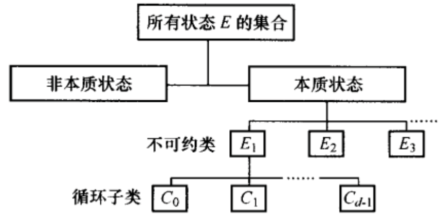


图 39 马尔可夫链的状态按转移概率 $p_{ij}^{(n)}$ 的渐近性分类

5. 练习题

1. 在第 1 小节的末尾, 讨论了非本质状态和本质状态的描述性定义. 试用转移概率 $p_{ij}^{(n)}, i, j \in E, n \geqslant 1$ 的术语, 给出非本质状态和本质状态定义的确切表述.

2. 设 $\mathbb{P}$ 是不可约的有限马尔可夫链的转移概率矩阵. 且 $\mathbb{P}^2 = \mathbb{P}$. 讨论矩阵 $\mathbb{P}$ 的构造.

3. 设 $\mathbb{P}$ 是有限状态马尔可夫链 $X = (X_n)_{n \geqslant 0}$ 的转移概率矩阵; $\sigma_1, \sigma_2, \cdots$ 是独立同分布非负整数随机变量序列, 并且与 X 独立; $\tau_0 = 0, \tau_n = \sigma_1 + \cdots + \sigma_n, n \geqslant 1$.

(a) 证明: 序列 $\widetilde{X} = (\widetilde{X}_n)_{n \geqslant 0}$ 是马尔可夫链, 其中 $\widetilde{X}_n = X_{\tau_n}$;

(b) 求马尔可夫链 $\widetilde{X} = (\widetilde{X}_n)_{n \geqslant 0}$ 的转移概率矩阵 $\widetilde{\mathbb{P}}$;

(c) 证明, 如果对于马尔可夫链 X, 状态 i 和 j 互通, 则对于马尔可夫链 $\widetilde{X}$, 状态 i 和 j 也互通.

4. 考虑马尔可夫链: 其状态空间为 $E = \{0, 1\}$, 而转移概率矩阵为

$$\mathbb{P} = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix}, \quad 0 < \alpha < 1, \quad 0 < \beta < 1.$$

试描绘矩阵 $\mathbb{P}^{(n)}, n \geqslant 2$ 的构造.

§5. 马尔可夫链的状态按转移概率矩阵的渐近性质分类

1. 常返和非常返状态的概念与准则 设 $X = (X_n)_{n \geqslant 0}$ 是有可数个状态集合 $E = \{1, 2, \cdots\}$ 的齐次马尔可夫链, 其转移概率为 $p_{ij} = \mathbf{P}_i\{X_1 = j\}, i, j \in E$.

设

$$f_{ii}^{(n)} = \mathbf{P}_i\{X_n = i, X_k \neq i, 1 \leqslant k \leqslant n-1\}, \tag{1}$$

而对于 $i \neq j$,

$$f_{ij}^{(n)} = \mathbf{P}_i\{X_n = j, X_k \neq j, 1 \leqslant k \leqslant n-1\}. \tag{2}$$

显然, $f_{ii}^{(n)}$ 是"质点"恰好在第 n 步首返状态 i 的概率, 而 $f_{ij}^{(n)}$ 是"质点"在 $X_0=i$ 的条件下, 恰好在第 n 步首达状态 j 的概率.

假如设

$$\sigma_i(\omega)=\inf\{n\geqslant 1: X_n(\omega)=i\}, \tag{3}$$

其中当上式右侧大括号 $\{\cdot\}=\varnothing$ 时 $\sigma_i(\infty)=\infty$, 则概率 $f_{ii}^{(n)}$ 和 $f_{ij}^{(n)}$ 亦可表示为下面的形式:

$$f_{ii}^{(n)}=\mathbf{P}_i\{\sigma_i=n\},\quad f_{ij}^{(n)}=\mathbf{P}_i\{\sigma_j=n\}. \tag{4}$$

对于 $i,j\in E$, 引进变量

$$f_{ij}=\sum_{n=1}^{\infty}f_{ij}^{(n)}. \tag{5}$$

由 (4) 式可见

$$f_{ij}=\mathbf{P}_i\{\sigma_j<\infty\}. \tag{6}$$

换句话说, f_{ij} 是"质点"自状态 i 出发开始游动, 迟早到达状态 j 的概率.

在以后特别重要的是, "质点"自状态 i 出发开始游动, 迟早返回状态 i 的概率 f_{ii}. 下面给出这些概率的定义.

定义 1 如果 $f_{ii}=1$, 则称状态i 为常返的 (循环的, 返回的 [recurrent], 回归的 [persistent]).

定义 2 如果 $f_{ii}<1$, 则称状态 i 为非常返的 (非可迁的, 非返回的, 非回归的).

下面的定理给出了常返性和非常返性的准则.

定理 1 a) 状态 $i\in E$ 的常返性等价于如下两条性质中的任何一条:

$$\mathbf{P}_i\{X_n=i \text{ 对于无限多个 } n\}=1 \quad \text{或} \quad \sum_i p_{ii}^{(n)}=\infty.$$

b) 状态 $i\in E$ 的非常返性等价于如下两条性质中的任何一条:

$$\mathbf{P}_i\{X_n=i \text{ 对于无限多个 } n\}=0 \quad \text{或} \quad \sum_i p_{ii}^{(n)}<\infty.$$

于是, 由定理 1, 可见

$$f_{ii}=1\Leftrightarrow \mathbf{P}_i\{X_n=i \text{ 对于无限多个 } n\}=1\Leftrightarrow \sum_n p_{ii}^{(n)}=\infty, \tag{7}$$

$$f_{ii}<1\Leftrightarrow \mathbf{P}_i\{X_n=i \text{ 对于无限多个 } n\}=0\Leftrightarrow \sum_n p_{ii}^{(n)}<\infty. \tag{8}$$

注 注意, 根据第二章 §1 表 1, 事件 $\{X_n=i$ 对于无限多个 $n\}$, 是对于无限多个 n 使 $X_n(\omega)=i$ 的 ω 集合的. 如果这时 $A_n=\{\omega: X_n(\omega)=i\}$, 则 (见第二章 §1 表 1).

$$\{X_n=i \text{ 对于无限多个 } n\}=\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty}A_k.$$

证明 立即可以指出如下蕴涵关系:

$$\sum_n p_{ii}^{(n)}<\infty\Rightarrow \mathbf{P}_i\{X_n=i \text{ 对于无限多个 } n\}=0, \tag{9}$$

因为, 由于 $p_{ii}^{(n)}=\mathbf{P}_i\{X_n=i\}$, 可见上面的蕴含关系是博雷尔 – 坎泰利引理 1 的推论 (第二章 §10 命题 a)).

现在证明,

$$f_{ii}=1\Leftrightarrow \sum_n p_{ii}^{(n)}=\infty. \tag{10}$$

由齐次性和马尔可夫性, 可见对于任何数组 $(i_1,\cdots,i_k)$ 和 $(j_1,\cdots,j_n)$, 有

$$\begin{aligned}&\mathbf{P}_i\{(X_1,\cdots,X_k)=(i_1,\cdots,i_k),(X_{k+1},\cdots,X_{k+n})=(j_1,\cdots,j_n)\}\\ &=\mathbf{P}_i\{(X_1,\cdots,X_k)=(i_1,\cdots,i_k)\}\mathbf{P}_{i_k}\{(X_1,\cdots,X_n)=(j_1,\cdots,j_n)\}.\end{aligned}$$

由此直接得 (对照第一章 §12 公式 (38) 和第一章 §2 公式 (25) 的推导):

$$\begin{aligned}p_{ij}^{(n)}=\mathbf{P}_i\{X_n=j\}&=\sum_{k=0}^{n-1}\mathbf{P}_i\{X_1\neq j,\cdots,X_{n-k-1}\neq j,X_{n-k}=j,X_n=j\}\\ &=\sum_{k=0}^{n-1}\mathbf{P}_i\{X_1\neq j,\cdots,X_{n-k-1}\neq j,X_{n-k}=j\}\mathbf{P}_j\{X_k=j\}\\ &=\sum_{k=0}^{n-1}f_{ij}^{(n-k)}p_{jj}^{(k)}=\sum_{k=1}^{n}f_{ij}^{(k)}p_{jj}^{(n-k)}.\end{aligned}$$

这样, 得如下 $n(n\geqslant 1)$ 步转移概率的公式:

$$p_{ij}^{(n)}=\sum_{k=1}^{n}f_{ij}^{(k)}p_{jj}^{(n-k)}. \tag{11}$$

设 $j=i$, 有

$$\begin{aligned}\sum_{n=1}^{\infty}p_{ii}^{(n)}&=\sum_{n=1}^{\infty}\sum_{k=1}^{n}f_{ii}^{(k)}p_{ii}^{(n-k)}=\sum_{n=1}^{\infty}f_{ii}^{(k)}\sum_{n=k}^{\infty}p_{ii}^{(n-k)}\\ &=f_{ii}\sum_{n=0}^{\infty}p_{ii}^{(n)}=f_{ii}\left(1+\sum_{n=1}^{\infty}p_{ii}^{(n)}\right),\end{aligned} \tag{12}$$

其中 $p_{ii}^{(0)}=1$. 由此可见

$$\sum_{n=1}^{\infty}p_{ii}^{(n)}<\infty\Rightarrow f_{ii}=\frac{\sum_{n=1}^{\infty}p_{ii}^{(n)}}{1+\sum_{n=1}^{\infty}p_{ii}^{(n)}}. \tag{13}$$

现在设 $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$, 则

$$\sum_{n=1}^{N} p_{ii}^{(n)} = \sum_{n=1}^{N}\sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)} = \sum_{k=1}^{N} f_{ii}^{(k)} \sum_{n=k}^{N} p_{ii}^{(n-k)} \leqslant \sum_{k=1}^{N} f_{ii}^{(k)} \sum_{l=0}^{N} p_{ii}^{(l)},$$

因而

$$f_{ii} = \sum_{k=1}^{\infty} f_{ii}^{(k)} \geqslant \sum_{k=1}^{N} f_{ii}^{(k)} \geqslant \frac{\sum\limits_{n=1}^{N} p_{ii}^{(n)}}{\sum\limits_{l=0}^{N} p_{ii}^{(l)}} \to 1, \quad N \to \infty.$$

从而

$$\sum_{n=1}^{N} p_{ii}^{(n)} = \infty \Rightarrow f_{ii} = 1. \tag{14}$$

由蕴含关系 (13) 和 (14), 立即得到互蕴涵关系:

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty \Leftrightarrow f_{ii} < 1, \tag{15}$$

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty \Leftrightarrow f_{ii} = 1. \tag{16}$$

为完成定理的证明, 只需验证下面的互蕴涵关系:

$$f_{ii} < 1 \Leftrightarrow \mathbf{P}_i\{X_n = i \text{ 对于无限多个 } n\} = 0, \tag{17}$$

$$f_{ii} = 1 \Leftrightarrow \mathbf{P}_i\{X_n = i \text{ 对于无限多个 } n\} = 1. \tag{18}$$

直观上, 这些性质是很清楚的. 例如, 如果 $f_{ii} = 1$, 则说明 $\mathbf{P}_i\{\sigma_i < \infty\} = 1$, 即 "质点" 迟早要返回它开始游动的状态 i. 但这是根据强马尔可夫性, 从这一 (随机) 时刻起, "质点的寿命" 仿佛重新开始. 假如继续观察, 则我们将看到, 事件 $\{X_n = i\}$ 将对于无限多个下标 n 出现, 即 $\mathbf{P}_i\{X_n = i, \text{ 对于无限多个 } n\} = 1$.

我们现在对性质 (17) 和 (18) 进行严格证明.

对于给定的状态 $i \in E$, 考虑回返状态 i 的次数不少于 m 的概率 α_m. 我们证明此概率等于 $\alpha_m = (f_{ii})^m$.

事实上, 如果 $m = 1$, 则 f_{ii} 由定义可知 $\alpha_1 = f_{ii}$. 现在假设 $\alpha_{m-1} = (f_{ii})^{m-1}$, 证明 $\alpha_m = (f_{ii})^m$.

由强马尔可夫性 (见 §2 的 (8) 式), 并注意到事件 $\{\sigma_i = k\} \in \mathscr{F}_{\sigma_i}$, 可见

$$\begin{aligned}
\alpha_m &= \mathbf{P}_i\{\text{返回状态 } i \text{ 的次数不少于 } m\} \\
&= \sum_{k=1}^{\infty} \mathbf{P}_i\{\sigma_i = k \text{ 且在 } k \text{ 时后返回状态 } i \text{ 的次数不少于 } m-1\} \\
&= \sum_{k=1}^{\infty} \mathbf{P}_i\{\sigma_i = k\}\mathbf{P}_i\{X_{\sigma_i+1}, X_{\sigma_i+2}, \cdots \text{ 中至少有 } m-1 \text{ 个值等于 } i|\sigma_i = k\} \\
&= \sum_{k=1}^{\infty} \mathbf{P}_i\{\sigma_i = k\}\mathbf{P}_i\{X_1, X_2, \cdots \text{ 中至少有 } m-1 \text{ 个值等于 } i\} \\
&= \sum_{k=1}^{\infty} f_{ii}^{(k)}(f_{ii})^{m-1} = f_{ii}(f_{ii})^{m-1} = (f_{ii})^m.
\end{aligned}$$

由此得

$$\mathbf{P}_i\{X_n = i \text{ 对于无限多个 } n\} = \lim_{m\to\infty}(f_{ii})^m = \begin{cases} 1, & \text{若 } f_{ii} = 1, \\ 0, & \text{若 } f_{ii} = 0. \end{cases} \tag{19}$$

该公式表明, 如果 $A = \{A_n \text{ 对于无限多个 } n\}\ (= \overline{\lim} A_n)$, 其中 $A_n = \{X_n = i\}$, 则对于概率 $\mathbf{P}_i(A)$, "0 – 1 律" 成立, 即 $\mathbf{P}_i(A)$ 只有 0 和 1 两个可能值. 注意, 这一性质不能直接由博雷尔 – 坎泰利引理 (第二章 §10) 直接得到, 因为事件 $A_n (n \geqslant 1)$ 一般不独立.

由 (19) 式, 以及性质: $\mathbf{P}_i(A)$ 只有 0 和 1 两个可能值, 可得所要证明的 (17) 式和 (18) 式中的蕴涵关系. □

2. 非常返状态 由定理 1 可以得到非常返状态如下简单而重要的性质.

定理 2 如果状态 j 非常返, 则对于任意状态 $i \in E$,

$$\sum_{n=1}^{\infty} p_{ij}^{(n)} < \infty, \tag{20}$$

因此, 任意状态 $i \in E$,

$$p_{ij}^{(n)} \to 0, \quad n \to \infty. \tag{21}$$

证明 由 (11) 式 (其中 $p_{jj}^{(0)} = 1$), 有

$$\begin{aligned}
\sum_{n=1}^{\infty} p_{ij}^{(n)} &= \sum_{n=1}^{\infty}\sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)} = \sum_{k=1}^{\infty} f_{ij}^{(k)} \sum_{n=0}^{\infty} p_{jj}^{(n)} \\
&= f_{ij} \sum_{n=0}^{\infty} p_{jj}^{(n)} \leqslant \sum_{n=0}^{\infty} p_{jj}^{(n)} < \infty,
\end{aligned}$$

其中我们注意到

$$f_{ij} = \sum_{k=1}^{\infty} f_{ij}^{(k)} \leqslant 1$$

(f_{ij} 是质点从状态 i 出发, 迟早到达状态 j 的概率).

由 (20) 式显然得 (21) 式. □

3. 常返状态 现在考虑常返状态. 以首返某状态 $i \in E$ 的平均首返时间

$$\mu_i = \sum_{n=1}^{\infty} f_{ii}^{(n)} \leqslant 1 \quad (= \mathbf{E}_i\sigma_i) \tag{22}$$

有限还是无限为转移, 每一个常返状态 $i \in E$ 分为两种类型: 正常返或零常返状态. 我们知道, 根据 (1) 式 $f_{ii}^{(n)}$ 表示恰好经过 n 步首返的概率.

定义 3 如果

$$\mu_i^{-1} = \left(\sum_{n=1}^{\infty} n f_{ii}^{(n)}\right)^{-1} > 0, \tag{23}$$

则称状态 $i \in E$ 为正常返的, 而如果

$$\mu_i^{-1} = \left(\sum_{n=1}^{\infty} n f_{ii}^{(n)}\right)^{-1} = 0, \tag{24}$$

则称状态 $i \in E$ 为零常返的.

这样, 根据定义首返零常返状态, 平均经过无限时间出现. 平均首返正常返状态的时间是有限的.

4. 马尔可夫链的状态按转移概率矩阵的渐近性质分类 下面的框图直观地表示, 马尔可夫链的状态, 根据常返和非常返, 正常返和零常返分类.

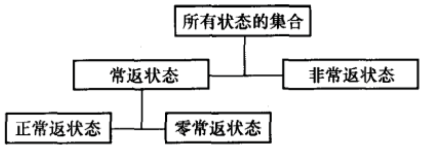


图 40 马尔可夫链的状态按概率 $p_{ij}^{(n)}$ 的渐近性质分类

5. 常返和非周期状态下转移概率矩阵的渐近性质

定理 3 设马尔可夫链的状态 $i \in E$ 是常返的和非周期的 ($d(j) = 1$). 那么, 对于任意 $i \in E$, 有

$$p_{ij}^{(n)} \to \frac{f_{ij}}{\mu_j}, \quad n \to \infty. \tag{25}$$

此外, 如果状态 i 和 j 是互通的 ($i \leftrightarrow j$), 即状态 i 和 j 属于同一不可约类, 则

$$p_{ij}^{(n)} \to \frac{1}{\mu_j}, \quad n \to \infty. \tag{26}$$

在下面将要进行的证明中, 本质地依赖于引理 1 的论点, 而该论点是 "离散更新理论" 的关键结果. 关于定理 3 的另外一个, 基于耦合 (coupling, 第三章 §8) 的思想, 例如参见 [104], [105].

引理 1 ("离散更新理论" 的基本引理) 设 $\varphi = (\varphi_1, \varphi_2, \cdots)$ 是非负数的非周期 ($d(j) = 1$) 序列, 根据该序列建立一对应于如下递推规则的序列 $u = (u_1, u_2, \cdots)$: $u_0 = 1$, 对于任意 $n \geqslant 1$,

$$u_n = \varphi_1 u_{n-1} + \varphi_2 u_{n-2} + \cdots + \varphi_n u_0. \tag{27}$$

那么, 当 $n \to \infty$ 时

$$u_n \to \mu^{-1},$$

其中 $\mu = \sum_{n=1}^{\infty} n\varphi_n$.

关于该引理的证明, 例如参见 [69] (第 1 卷, 第 VIII 章, 10).

定理 3 的证明 首先设 $i = j$. 证明

$$p_{ij}^{(n)} \to \frac{1}{\mu_j}, \quad n \to \infty. \tag{28}$$

为此将 (11) 式 (对于 $i = j$ 的情形) 写成下面的形式:

$$p_{jj}^{(n)} = f_{jj}^{(1)} p_{jj}^{(n-1)} + f_{jj}^{(2)} p_{jj}^{(n-2)} + \cdots + f_{jj}^{(n)} p_{jj}^{(0)}, \tag{29}$$

其中设 $p_{jj}^{(0)} = 1$, 并且显然 $f_{jj}^{(1)} = p_{jj}^{(1)}$. 如果设

$$u_k = p_{jj}^{(k)}, \quad \varphi_k = f_{jj}^{(k)}, \tag{30}$$

则 (29) 式有如下形式

$$u_n = \varphi_1 u_{n-1} + \varphi_2 u_{n-2} + \cdots + \varphi_n u_0,$$

因此恰好是引理 1 的递推公式.

假如我们能证明 (定理 3 中假设的条件) 序列 $(p_{jj}^{(1)}, p_{jj}^{(2)}, \cdots)$ 的周期等于 1, 则序列 $(f_{jj}^{(1)}, f_{jj}^{(2)}, \cdots)$ 的周期 $d_f(j)$ 等于 1, 那么所要证明的 (28) 式的结果可以直接由引理 1 得到.

而这又是下面 (引理 2 的) 一般命题的推论.

引理 2 对于任意 $j \in E$,

$$\mathrm{GCD}\{n \geqslant 1 : p_{jj}^{(n)} > 0\} = \mathrm{GCD}\{n \geqslant 1 : f_{jj}^{(n)} > 0\}, \tag{31}$$

即周期 $d_f(j) = d(j)$.

证明 设

$$M = \{n : p_{jj}^{(n)} > 0\} \quad 和 \quad M_f = \{n : f_{jj}^{(n)} > 0\}.$$

由于 $M_f \subseteq M$, 可见

$$\mathrm{GCD}(M) \leqslant \mathrm{GCD}(M_f),$$

即 $d(j) \leqslant d_f(j)$.

相反的不等式由如下概率的意义 $p_{jj}^{(n)}$ 和 $f_{jj}^{(n)}, n \geqslant 1$ 可以得到.

如果 "质点" 从状态 j 出发经过 n 步仍然处于该状态 ($p_{jj}^{(n)} > 0$), 则说明其游动是这样进行的: 首先经 k_1 步 ($f_{jj}^{(k_1)} > 0$) 由状态 j 首次返回 j, 然后经 k_2 步 ($f_{jj}^{(k_2)} > 0$), $\cdots$, 经 k_l 步 ($f_{jj}^{(k_l)} > 0$) 由状态 j 首次返回 j.

因而 $n = k_1 + k_2 + \cdots + k_l$. 数 $d_f(j)$ 整除 $k_1, k_2, \cdots, k_l$, 故也整除 n. 由于 $d(j)$ 是被 n 整除且使 $p_{jj}^{(n)} > 0$ 的数中最大者, 可见 $d(j) \geqslant d_f(j)$.

于是 $d(j) = d_f(j)$. 顺便指出, 在利用公式 $d(j) = \mathrm{GCD}\{n \geqslant 1 : p_{jj}^{(n)} > 0\}$ 定义状态 j 的周期 $d(j)$ 时, 也可以利用公式 $d(j) = \mathrm{GCD}\{n \geqslant 1 : f_{jj}^{(n)} > 0\}$. 引理 2 得证. □

完成定理 3 的证明 1) (25) 式的证明 ($i \neq j$ 的情形).

将 (11) 式表示为如下形式:

$$p_{ij}^{(n)} = \sum_{k=1}^{\infty} f_{ij}^{(k)} p_{jj}^{(n-k)}, \tag{32}$$

其中 $p_{jj}^{(l)} = 0, l < 0$.

这里, 由于当 $n \to \infty$ 时

$$p_{ij}^{(n)} \to \frac{1}{\mu_j}, \quad 且 \quad \sum_{k=1}^{\infty} f_{ij}^{(k)} \leqslant 1,$$

可见根据控制收敛定理 (第二章 §6 定理 3)

$$\lim_n \sum_{k=1}^{\infty} f_{ij}^{(k)} p_{jj}^{(n-k)} = \sum_{k=1}^{\infty} f_{ij}^{(k)} \lim_n p_{jj}^{(n-k)} = \frac{1}{\mu_j} \sum_{k=1}^{\infty} f_{ij}^{(k)} = \frac{f_{ij}}{\mu_j}. \tag{33}$$

由 (32) 和 (33) 式得

$$\lim_n p_{ij}^{(n)} = \frac{f_{ij}}{\mu_j}, \tag{34}$$

即命题 (25) 得证.

2) (26) *式的证明* ($i \leftrightarrow j$ *的情形*). 假如在条件 $i \leftrightarrow j$ 下 (即状态 i 和 j 属于同一互通状态的不可约类) 概率 $f_{ij} = 1$, 则由 (34) 式即可得性质 (26) 式.

假设状态 j 是常返的, 则由定理 1 的命题 a), 有 $\mathbf{P}_j\{X_n = j$ 对于无限多个 $n\} = 1$. 因此, 对于任意 m,

$$\begin{aligned} p_{ji}^{(m)} &= \mathbf{P}_j(\{X_m = i\} \bigcap \{X_n = j \text{ 对于无限多个 } n\}) \\ &\leqslant \sum_{n>m} \mathbf{P}_j\{X_m = i, X_{m+1} \neq j, \cdots, X_{n-1} \neq j, X_n = j\} \\ &= \sum_{n>m} p_{ji}^{(m)} f_{ij}^{(n-m)} = p_{ji}^{(m)} f_{ij}, \end{aligned} \tag{35}$$

其中中间的不等式是广义马尔可夫性的推论 (见 §2 的 (2) 式). 由于 E 是互通状态类, 可见存在 m, 使 $p_{ji}^{(m)} > 0$. 因此由 (35) 式推得 $f_{ij} = 1$. □

6. 状态周期任意的情形 对于感兴趣的状态 j 的周期 d 任意 ($d = d(j) > 0$) 的情形, 可以自然地表述上面的定理 3 的类似. 看下面的定理.

定理 4 *设马尔可夫链的状态 $j \in E$ 是常返, 且其周期的 $d = d(j) \geqslant 1$, 而 $i \in E$ 也是马尔可夫链的状态 j (有可能与状态 j 相同).*

a) 设 i 和 j 属于同一不可分解状态类 $C \subseteq E$, 且 $C_0, C_1, \cdots, C_{d-1}$ 是 (循环的) 子类, 按如下顺序编号: $j \in C_0, i \in C_a$ 其中 $a \in \{0, 1, \cdots, d-1\}$, 而在此子类中的运动是按如下顺序: $C_0 \to C_1 \to \cdots \to C_a \to \cdots \to C_{d-1} \to C_0$. 那么, 当 $n \to \infty$ 时

$$p_{ij}^{(nd+a)} \to \frac{d}{\mu_j}, \tag{36}$$

b) 在一般情形下, 当 i 和 j 可能属于不同的不可约状态类时, 当 $n \to \infty$ 时, 对于任意 $a = 0, 1, \cdots, d-1$,

$$p_{ij}^{(nd+a)} \to \frac{d}{\mu_j} \sum_{k=0}^{n} f_{jj}^{(kd+a)}. \tag{37}$$

证明 首先设 $a = 0$, 即 i 和 j 可能属于同一不可约状态类 C, 并且同一循环子类 C_0.

考虑转移概率 $p_{ij}^{(d)}, i, j \in C$, 并根据 $p_{ij}^{(d)}$ (按照 §1 的构造) 建立新的马尔可夫链.

对于新链, 状态 j 是常返的和非周期的. 新链的状态 i 和 j 仍然是互通的 ($i \leftrightarrow j$). 这就证明了定理 3 的性质 (26):

$$p_{ij}^{(nd)} \to \frac{1}{\sum_{k=1}^{\infty} k f_{jj}^{(kd)}} = \frac{d}{\sum_{k=1}^{\infty} (kd) f_{jj}^{(kd)}} = \frac{d}{\mu_j},$$

其中最后一等式成立, 因为, 对于一切不能被 d 整除的 l, 有 $f_{jj}^{(l)} = 0$, 并且根据定义.

$$\mu_j = \sum_{l=1}^{\infty} l f_{jj}^{(l)}.$$

现在假设, 对于 $a=0,1,\cdots,r(\leqslant d-2)$, (36) 式已得证.

根据控制收敛定理 (第二章 §6 定理 3),

$$p_{ij}^{(nd+r+1)}=\sum_{k=1}^{\infty}p_{ik}p_{kj}^{(nd+r)}\to\sum_{k=1}^{\infty}p_{ik}\frac{d}{\mu_j}=\frac{d}{\mu_j}.$$

于是, 对于 $a=r+1(\leqslant d-1)$, 欲证明的 (36) 式得证. 从而, 根据归纳法, 对于一切 $a=0,1,\cdots,d-1$, (36) 式得证.

b) 对于任何 $i,j\in E$, 下面的公式成立 (见 (11) 式):

$$p_{ij}^{(nd+a)}=\sum_{k=1}^{nd+a}f_{ij}^{(k)}p_{jj}^{(nd+a-k)},\quad a=0,1,\cdots,d-1.$$

根据假设状态 j 的周期等于 d. 因此, 除当 $k-a$ 具有形式 rd 时之外, 有 $p_{jj}^{(nd+a-k)}=0$. 于是, 有

$$p_{ij}^{(nd+a)}=\sum_{r=0}^{n}f_{ij}^{(rd+a)}p_{jj}^{((n-r)d)}.$$

由此以及上面证明的 (36) 式, 仍然运用控制收敛定理, 最后得到需要证明的公式 (37). □

7. 非周期马尔可夫链的完全分类 像在 §4 的末尾指出的那样, 在根据转移概率的渐近性质, 研究马尔可夫链的状态的分类时, 只需局限于考虑非周期不可约链. 在定理 1~3 中陈述的结果, 实际上包含了这样链的完全分类的全部必要内容.

首先引进一个辅助命题, 它是关于不可约链的所有状态属于同一种 ("常返" 或 "非常返") 类型. (对照 §4 定理 2 中的 "单一类型性".)

引理 3 *设 E 是 (互通状态的) 的不可约类. 那么, 其全部状态或者都是常返的, 或者都是非常返的.*

证明 假设链至少有一个非常返状态, 例如状态 i, 则根据定理 1, $\sum_n p_{ii}^{(n)}<\infty$.

现在假设是 j 任意另一个状态. 由于 E 互通状态 $(i\leftrightarrow j)$ 的不可约类, 存在这样的状态 k 和 l, 使 $p_{ij}^{(k)}>0$ 和 $p_{ij}^{(l)}>0$. 那么, 由明显的不等式

$$p_{ii}^{(n+k+l)}\geqslant p_{ij}^{(k)}p_{jj}^{(n)}p_{ji}^{(l)},$$

可见

$$\sum_n p_{ii}^{(n+k+l)}\geqslant p_{ij}^{(k)}p_{ji}^{(l)}\sum_n p_{jj}^{(n)}.$$

根据假设 $\sum_n p_{ii}^{(n)}<\infty$, 而 k,l 满足 $p_{ij}^{(k)}p_{ji}^{(l)}>0$, 从而 $\sum_n p_{jj}^{(n)}<\infty$.

鉴于定理 1 的 b), 由此可见, 状态 j 也是非常返状态. 换句话说, 假如不可约类只要有一个非常返状态, 则其余状态也都是非常返状态.

现在设 i 是常返状态, 我们证明其余状态也都是常返状态.

假设 (除 i 之外) 至少有一个非常返状态. 那么根据上面已证明的结果, 其余状态也都是非常返状态. 因此与 "i 是常返状态" 的假设矛盾.

因此哪怕有一个常返状态, 自动导致 (不可约链的) 全部状态都是常返状态. 从而引理得证. □

该引理的结论, 完全证明了对于不可约链,"常返链" 和 "非常返链" 这两个 (被普遍接受的) 术语是合适的. (注意, 这里说的 "链", 而不仅是个别状态.)

定理 5 *假设马尔可夫链由非周期状态的一个不可约类 E 构成. 对于这样的链仅属于如下三种可能的类型之一:*

(i) 非常返链. 在这种情形下, 对于一切 $i,j\in E$,

$$\lim_n p_{ij}^{(n)}=0,$$

并且在

$$\sum_n p_{ij}^{(n)}<\infty$$

的意义上充分 "快" 地收敛于 0.

(ii) 常返与零常返链. 在这种情形下, 对于一切 $i,j\in E$,

$$\lim_n p_{ij}^{(n)}=0,$$

并且收敛在

$$\sum_n p_{ij}^{(n)}=\infty$$

的意义上充分地 "慢", 此外从 j 到 j 的平均首返时间 μ_j 等于 ∞.

(iii) 常返与正常返链. 在这种情形下, 对于一切 $i,j\in E$,

$$\lim_n p_{ij}^{(n)}=\frac{1}{\mu_j}>0,$$

其中 μ_j 是从 j 到 j 的平均首返时间, 并且有限.

证明 命题 (i) 在定理 1 的 b) 和定理 2 已经证明. 命题 (ii) 和 (iii) 可以直接由定理 1 的 a) 和定理 3 得到. □

8. 有限马尔可夫链 现在考虑有限马尔可夫链的情形, 即状态集合 E 由有限个元素组成的情形.

结果, 在这种情形下, 定理 5 中的三种可能性 (i), (ii), (iii) 只有 (iii) 仍然成立.

定理 6 *设有限马尔可夫链是不可约的和非周期的. 那么, 这样的链是常返的和正的. 这时*

$$\lim_n p_{ij}^{(n)}=\frac{1}{\mu_j}>0.$$

证明 用反证法. 假设此链是非常返的. 那么, 如果链的状态数 r 是有限的 $(E=1,2,\cdots,r)$, 则

$$\lim_n \sum_{j=1}^{r} p_{ij}^{(n)} = \sum_{j=1}^{r} \lim_n p_{ij}^{(n)}. \tag{38}$$

该式的右侧显然等于 1. 但是我们假设它是非常返的, 那么, 这时 (根据定理 5 的 (i)) 右侧应该等于 0. 由此得到的矛盾将 (i) 排除.

现在假设链是常返的.

由于定理 5 的命题只剩下两种可能性 (ii) 和 (iii). 需要排除 (ii). 事实上, 由于对于所有 $i,j\in E, \lim_n p_{ij}^{(n)}=0$, 则像非常返状态的情形一样, 出现矛盾.

于是, 只有第三种可能性 (iii). □

9. 练习题

1. 考虑不可约链, 其状态集为 $0,1,2,\cdots$ 那么, 它是非常返的, 当且仅当方程组

$$u_j=\sum_i u_i p_{ij}, \quad j=0,1,\cdots$$

有有界解, 而且 $u_i\neq c$ (常数), $i=0,1,\cdots$

2. 只要序列 $u=(u_0,u_1,\cdots), u_i\to\infty, i\to\infty$, 使得对于一切 $j\neq 0, u_i\geqslant \sum_i u_i p_{ij}$, 则状态集为 $0,1,2,\cdots$ 的不可约链是常返的.

3. 状态集为 $0,1,2,\cdots$ 的不可约链是常返的和正的, 当且仅当方程组

$$u_j=\sum_i u_i p_{ij}, \quad j=0,1,\cdots$$

有不恒等于 0 的解, 而且 $\sum_i |u_i|<\infty$.

4. 假设马尔可夫链的状态为 $0,1,\cdots$, 而转移概率为

$$p_{00}=r_0, \quad p_{01}=p_0>0,$$

$$p_{ij}=\begin{cases} p_i>0, & j=i+1,\\ r_i\geqslant 0, & j=i,\\ q_i>0, & j=i-1,\\ 0, & \text{其他}. \end{cases}$$

设

$$\rho_0=1, \quad \rho_m=\frac{q_1\cdots q_m}{p_1\cdots p_m}.$$

证明如下的等价关系:

$$\text{常返链} \Leftrightarrow \sum \rho_m=\infty,$$
$$\text{非常返链} \Leftrightarrow \sum \rho_m<\infty,$$
$$\text{正链} \Leftrightarrow \sum \rho_m=\infty, \quad \sum \frac{1}{p_m\rho_m}<\infty,$$
$$\text{零链} \Leftrightarrow \sum \rho_m=\infty, \quad \sum \frac{1}{p_m\rho_m}=\infty.$$

5. 证明:

$$f_{ik}\geqslant f_{ij}f_{jk},$$
$$\sup_n p_{ij}^{(n)} \leqslant f_{ij} \leqslant \sum_{n=1}^{\infty} p_{ij}^{(n)}.$$

6. 证明, 对于任意具有可数状态集的马尔可夫链, 对于转移概率 $p_{ij}^{(n)}$, 在切萨罗 (E. Cesàro) 意义上永远存在极限:

$$\lim_n \frac{1}{n}\sum_{k=1}^{n} p_{ij}^{(k)}=\frac{f_{ij}}{\mu_j}.$$

7. 考虑马尔可夫链 $\xi_0,\xi_1,\cdots$, 且 $\xi_{k+1}=(\xi_k)^+ +\eta_{k+1}, k\geqslant 0$, 其中 $\eta_1,\eta_2,\cdots$ 是独立同分布随机变量序列, 其概率分布为: $\mathbf{P}\{\eta_k=j\}=p_j, j=0,1,\cdots$

(a) 试写出转移概率矩阵;

(b) 证明, 如果 $p_0>0, p_0+p_1<1$, 则链常返当且仅当

$$\sum_k kp_k\leqslant 1.$$

§6. 可数马尔可夫链的极限分布、遍历分布和平稳分布

1. 极限值 $\mathbb{\Pi}$ 与平稳分布 $\mathbb{Q}$ 的联系 我们从一个一般结果开始, 它首先与非常清晰的极限值 $\mathbb{\Pi}=(\pi_1,\pi_2,\cdots)$ 相联系, 其中 $\pi_j=\lim_n p_{ij}^{(n)} (j=1,2,\cdots)$, 以及与平稳分布 $\mathbb{Q}=(q_1,q_2,\cdots)$ 相联系.

定理 1 *考虑具有可数状态集 $E=\{1,2,\cdots\}$ 的马尔可夫链, 其转移概率 $p_{ij}^{(n)} (i,j\in E)$ 有不依赖于初始状态 $i\in E$ 的极限:*

$$\pi_j=\lim_n p_{ij}^{(n)}, \quad j\in E.$$

那么,

(a) $\sum_{j=1}^{\infty}\pi_j\leqslant 1, \sum_{j=1}^{\infty}\pi_i p_{ij}=\pi_j, j\in E$;

(b) 下面的式子二者必居其一:

$$\sum_{j=1}^{\infty}\pi_j=0(\text{ 即所有 }\pi_j=0,j\in E)\quad\text{或}\quad\sum_{j=1}^{\infty}\pi_j=1;$$

(c) 如果 $\sum\limits_{k=1}^{\infty}\pi_j=0$, 则马尔可夫链无平稳分布; 如果 $\sum\limits_{j=1}^{\infty}\pi_j=1$, 则极限值的向量 $\mathbb{\Pi}=(\pi_1,\pi_2,\cdots)$, 就是马尔可夫链的平稳分布, 而且该马尔可夫链再无其他平稳分布.

证明 有

$$\sum_{j=1}^{\infty}\pi_j=\sum_{j=1}^{\infty}\lim_n p_{ij}^{(n)}\leqslant\varliminf_n\sum_{j=1}^{\infty}p_{ij}^{(n)}=1,\tag{1}$$

而对于任意 $j\in E,k\in E$, 有

$$\sum_{i=1}^{\infty}\pi_i p_{ij}=\sum_{i=1}^{\infty}\lim_n p_{ki}^{(n)}p_{ij}\leqslant\varliminf_n\sum_{i=1}^{\infty}p_{ki}^{(n)}p_{ij}=\varliminf_n p_{kj}^{(n+1)}=\pi_j.\tag{2}$$

注 这里出现的不等式和下极限, 当然是 "法图引理" 的推论. 不过应该注意, 在这里法图引理, 并不是像第二章 §6 那样, 用于由概率测度定义的勒贝格积分, 而是用于按 $\sigma-$ 有限 (非负) 测度进行积分的情形.

这样, 极限值的向量 $\mathbb{\Pi}=(\pi_1,\pi_2,\cdots)$, 具有如下性质:

$$\sum_{j=1}^{\infty}\pi_j\leqslant 1,\quad\sum_{i=1}^{\infty}\pi_i p_{ij}\leqslant\pi_j,\quad j\in E.\tag{3}$$

现在证明, 最后一个不等式实际上是等式.

设对于任意 $j_0\in E$, 有

$$\sum_{i=1}^{\infty}\pi_i p_{ij_0}\leqslant\pi_{j_0}.\tag{4}$$

那么,

$$\sum_{j=1}^{\infty}\pi_j>\sum_{j=1}^{\infty}\left(\sum_{i=1}^{\infty}\pi_i p_{ij}\right)=\sum_{i=1}^{\infty}\pi_i\sum_{j=1}^{\infty}p_{ij}=\sum_{i=1}^{\infty}\pi_i.$$

所得矛盾说明, (3) 式中最后一个不等式实际上是等式. 由于对于 $j\in E$, 有

$$\sum_{j=1}^{\infty}\pi_j\leqslant 1,\quad\sum_{i=1}^{\infty}\pi_i p_{ij}=\pi_j,$$

可见性质 (a) 成立.

为证明性质 (b), 注意到 (3) 式 $\sum\limits_{i=1}^{\infty}\pi_i p_{ij}=\pi_j$, 经积分, 对于任意 $n\geqslant 1$ 和任意 $j\in E$, 得

$$\sum_{i=1}^{\infty}\pi_i p_{ij}^{(n)}=\pi_j.$$

根据勒贝格控制收敛定理 (第二章 §6 定理 3), 由此可见

$$\pi_j=\lim_n\sum_{i=1}^{\infty}\pi_i p_{ij}^{(n)}=\sum_{j=1}^{\infty}\pi_i\lim_n p_{ij}^{(n)}=\left(\sum_{i=1}^{\infty}\pi_i\right)\pi_j,$$

即

$$\pi_j\left(1-\sum_{i=1}^{\infty}\pi_i\right)=0,\quad j\in E,$$

因此

$$\left(\sum_{j=1}^{\infty}\pi_j\right)\left(1-\sum_{i=1}^{\infty}\pi_i\right)=0.$$

这样, $a(1-a)=0$, 而 $a=\sum\limits_{i=1}^{\infty}\pi_i$. 所以 $a=1$ 或 $a=0$, 这就证明了命题 (b).

最后证明 (c). 为此, 假设 $\mathbb{Q}=(q_1,q_2,\cdots)$ 是某一平稳分布, 则

$$\sum_{i=1}^{\infty}q_i p_{ij}^{(n)}=q_j,\quad\text{且}\quad\left(\sum_{i=1}^{\infty}q_i\right)\pi_j=q_j,\quad j\in E,$$

其中后一个等式成立, 是根据控制收敛定理.

所以, 如果 $\mathbb{Q}$ 是平稳分布, 则 $\sum\limits_{i=1}^{\infty}q_i=1$, 从而对于此平稳分布一定 (应该) 满足条件: 对于一切 $j\in E$, 有 $q_j=\pi_j$. 这样一来, 假如 $\sum\limits_{j=1}^{\infty}\pi_j=0$, 则性质 $\sum\limits_{i=1}^{\infty}q_i=1$ 就不可能成立, 说明在这种情形下没有平稳分布.

根据 (b) 还有一种可能性 $\sum\limits_{j=1}^{\infty}\pi_j=1$. 在这种情形下, 根据 (a), $\mathbb{\Pi}=(\pi_1,\pi_2,\cdots)$ 本身是平稳分布, 而由以上的叙述知, 如果 $\mathbb{Q}$ 是另外某个平稳分布, 则它应与 $\mathbb{\Pi}$ 重合. 于是, 在 $\sum\limits_{j=1}^{\infty}\pi_j=1$ 的情形下, 平稳分布的唯一性也得到证明. □

2. 平稳分布和遍历分布的基本定理 定理 1 给了平稳分布存在 (同时也唯一) 的充分条件. 该条件在于对于一切 $j\in E$, 存在不依赖于 $i\in E$ 的极限值 $\pi_j=\lim\limits_n p_{ij}^{(n)}$, 并且至少对于一个状态 $j\in E$, 使得 $\pi_j>0$.

同时, 在 §5 中相当详细地研究了极限 $\lim\limits_n p_{ij}^{(n)}$ 存在的更为一般的问题, 与如下一些链的 "内在" 性质相联系, 诸如不可约性, 周期性, 常返性和非常返性, 正常返性和零常返性. 所以, 自然正是用这些由转移概率矩阵 $p_{ij}(i,j\in E)$ 决定的、"内在" 性质的术语, 来表述平稳分布存在的条件. 显然同样地, 如果用这些术语指出一切极限值 $\pi_j>0(j\in E)$ 的条件, 则根据定义 (见 §3 的性质 **C**) 向量 $\mathbb{\Pi}=(\pi_1,\pi_2,\cdots)$ 将形成遍历极限分布.

下面的两个定理将给出这些问题的答案.

定理 2 (平稳分布的基本定理) *考虑具有可数状态集 E 的马尔可夫链. 存在唯一平稳分布的必要和充分条件是:*

(a) 恰好存在一个不可约子类,

(b) 一切状态都是正常返的.

定理 3 (遍历分布的基本定理) *考虑具有可数状态集的马尔可夫链. 存在遍历分布的必要和充分条件: 链是*

(a) 不可约的,

(b) 正常返的,

(c) 非周期的.

3. 定理 2 的证明 *必要性.* 假设所考虑的马尔可夫链有唯一平稳分布, 记作 $\widetilde{\mathbb{Q}}_s$. 那么, 我们证明, 在状态集 E 上存在并且唯一正常返子类.

以 $N(0 \leqslant N \leqslant \infty)$ 表示这样子类的潜在可能数.

设 $N=0$, 而 j 是 E 中的某个状态. 由于不存在正常返类, 则状态 j 可能或者是非常返的, 或者是零常返的.

在第一种情形中, 由 §5 的定理 2 可见, 对于 $i \in E$, 极限 $\pi_j = \lim\limits_n p_{ij}^{(n)}$ 存在, 并且等于 0.

而在第二种情形中, 由 §5 的性质 (37) 以及由于 $\mu_j = \infty$ (因为状态 j 是零常返的) 可见, 这些极限 $\pi_j = \lim\limits_n p_{ij}^{(n)}$ 也存在, 并且等于 0.

这样, 当 $N=0$ 时对于一切 $i, j \in E$, 极限 $\pi_j = \lim\limits_n p_{ij}^{(n)}$ 存在, 并且等于 0. 因此在这种情形下, 根据定理 1 的命题 (c), 在这种情形下没有平稳分布, 从而 $N=0$ 的情形被 "存在平稳分布 $\widetilde{\mathbb{Q}}_s$" 的假设排除.

现在假设 $N=1$. 以 C 表示唯一正常返类.

如果该类的周期 $d(C)=1$, 则根据 §5 定理 3 的性质 (26), 可见对于一切 $i, j \in C$,

$$p_{ij}^{(n)} \to \mu_j^{-1}, \quad n \to \infty.$$

而若 $j \notin C$, 则这一状态非常返. 那么, 根据 §5 定理 2 的性质 (21), 可见对于一切 $i \in E$, 有

$$p_{ij}^{(n)} \to 0, \quad n \to \infty.$$

设

$$q_j = \begin{cases} \mu_j^{-1}(>0), & \text{若 } j \in C, \\ 0, & \text{若 } j \notin C. \end{cases} \tag{5}$$

那么, 由于集合 $C \neq \varnothing$, 则根据定理 1 数组 $\mathbb{Q} = (q_1, q_2, \cdots)$ 是唯一平稳分布, 从而 $\mathbb{Q} = \widetilde{\mathbb{Q}}_s$.

现在假设周期 $d(C) > 1$.

设 $C_0, C_1, \cdots, C_{d-1}$ 是 (正常返) 类的 C 的循环子类.

每一个子类 $C_k, k=0,1,\cdots,d-1$, 关于转移概率矩阵 $p_{ij}^{(d)}, i, j \in C$, 是常返的和非周期的. 那么, 如果 $i, j \in C_k$, 则根据 §5 公式 (36), 有

$$p_{ij}^{(nd)} \to \frac{d}{\mu_j} > 0.$$

所以在每一个集合 C_k 上, (由于定理 1 的性质 (b)) 数组 $\{d/\mu_j, j \in C_k\}$ (关于矩阵 $p_{ij}^{(d)}, i, j \in C$) 是唯一平稳分布.

特别, 由此可见

$$\sum_{j \in C_k} \frac{d}{\mu_j} = 1, \quad \text{即} \quad \sum_{j \in C_k} \frac{1}{\mu_j} = \frac{1}{d}.$$

设

$$q_j = \begin{cases} \mu_j^{-1}, \ j \in C = C_0 + \cdots + C_{d-1}, \\ 0, \quad j \notin C. \end{cases} \tag{6}$$

因而, 就证明了, 对于所考虑的马尔可夫链, 数组 $\mathbb{Q} = (q_1, q_2, \cdots)$ 是唯一平稳分布.

事实上, 如果 $i \in C$, 则

$$p_{ii}^{(nd)} = \sum_{j \in C} P_{ij}^{(nd-1)} p_{ji}.$$

那么, 如同 (1) 式, 有

$$\frac{d}{\mu_i} = \lim_n p_{ii}^{(nd)} \geqslant \sum_{j \in C} \underline{\lim_n} p_{ij}^{(nd-1)} p_{ji} = \sum_{j \in C} \frac{d}{\mu_j} p_{ji},$$

从而

$$\frac{1}{\mu_i} \geqslant \sum_{j \in C} \frac{1}{\mu_j} p_{ji}. \tag{7}$$

但是

$$\sum_{i \in C} \frac{1}{\mu_i} = \sum_{k=0}^{d-1} \left(\sum_{i \in C_k} \frac{1}{\mu_i} \right) = \sum_{k=0}^{d-1} \frac{1}{d} = 1. \tag{8}$$

像定理 1 的证明 (见 (3) 式和 (4) 式) 一样, 由 (7) 式和 (8) 式, 可见 (7) 式中实际上是等式:

$$\frac{1}{\mu_i} = \sum_{j \in C} \frac{1}{\mu_j} p_{ji}. \tag{9}$$

由于 $q_i = \mu_i^{-1} > 0$, 则 (9) 式表示数组 $\mathbb{Q} = (q_1, q_2, \cdots)$ 是平稳分布, 而由于定理 1 是唯一平稳分布. 从而 $\mathbb{Q} = \mathbb{Q}_s$.

最后假设 $2 \leqslant N < \infty$ 或 $N = \infty$. 如果 $2 \leqslant N < \infty$, 则以 $C^1, \cdots, C^N$ 表示正常返子类; 如果 $N = \infty$, 则以 $C^1, C^2, \cdots$ 表示正常返子类.

设数组 $\mathbb{Q}^k=(q_1^k,q_2^k,\cdots)$ 是类 C^k 中平稳分布, 由公式 (对照 (5) 式和 (6) 式)

$$q_j^k=\begin{cases}\mu_j^{-1}>0, & j\in C^k,\\ 0, & j\notin C^k\end{cases}$$

建立的平稳分布. 那么, 对于任意非负数 $a_1,a_2,\cdots: a_1+a_2+\cdots=1$(若 $N<\infty$, 则 $a_{N+1}=\cdots=0$), 数组 $\{a_1\mathbb{Q}^1,\cdots,a_N\mathbb{Q}^N,\cdots\}$ 显然是平稳分布. 因而, 若假设 $2\leqslant N<\infty$, 则导致 "一切平稳分布的集合" 是连续统, 但这与所作假设 "平稳分布唯一" 矛盾.

因而, 所作证明表示, 只可能是 $N=1$. 换句话说, *如果存在 (唯一) 平稳分布, 则链恰好有一个由正常返状态形成的不可约类.*

充分性 假如马尔可夫链有不可约正常返状态子类, 即有情形 $N=1$, 则由前面的叙述可见, (由定理 1 的命题 (c)) 平稳分布存在而且唯一.

于是, 定理 2 完全地得证. □

4. 定理 3 的证明 实质上, 这一定理的证明全部必要的, 都包含在定理 2 及其证明的论述中.

充分性 如果利用定理 2 的证明中记号, 则由定理的条件, 有 $N=1, C=E$ 和 $d(E)=1$ (非周期性). 那么, 由定理 2 的证明中关于 "$N=1$ 的情形" 的讨论, 可见 $\mathbb{Q}=(q_1,q_2,\cdots)$, 其中 $q_j=\mu_j^{-1}, j\in E$, 是平稳分布, 同时又是遍历分布, 因为所有 $\mu_j^{-1}<\infty, j\in E$.

因此, 遍历分布 $\Pi=(\pi_1,\pi_2,\cdots)(=\mathbb{Q})$ 的存在性得证.

必要性 如果存在遍历分布 $\Pi=(\pi_1,\pi_2,\cdots)$, 则根据定理 1, 存在并且唯一等同于 Π 的平稳分布 $\mathbb{Q}$.

由定理 2 的论断 (及其证明), 可见 $N=0$ 和 $2\leqslant N<\infty$ 不可能出现, 即只有 $N=1$, 并且仅存在唯一由正常返类组成的不可约类 C. 只剩下证明 $C=E$ 和 $d(E)=1$.

用反证法, 假设 $C\neq E$ 和 $d(C)=1$, 则仍然由于在定理 2 的证明中关于 "$N=1$ 的情形" 的讨论, 可见存在状态 $j\notin C$, 使对一切 $i\in E$, 由 $p_{ij}^{(n)}\to 0$. 然而, 这与对于一切 $i\in E, \pi_j=\lim\limits_n p_{ij}^{(n)}>0$ 矛盾.

这样, 在 $d(C)=1$ 的情形下, 有 $C=E$ 和 $d(E)=1$ (非周期性).

最后, 如果 $C\neq E$ 和 $d(C)>1$, 则仍然由于在定理 2 的证明中关于 "$N=1$ 的情形" 的讨论, 可见存在平稳分布 $\mathbb{Q}=(q_1,q_2,\cdots)$, 其中某些 $q_j=0$, 这与 "$\Pi=\mathbb{Q}$, 而 $\Pi=(\pi_1,\pi_2,\cdots)$ 是遍历分布且 (根据定义) 所有 $\pi_j>0, j\in E$" 矛盾. □

5. 定理 3 的推广 根据平稳 (不变) 分布 $\mathbb{Q}=(q_1,q_2,\cdots)$ 本身的定义, 是满足条件

$$q_j\geqslant 0,\quad j\in E=\{1,2,\cdots\},\quad \sum_{j=1}^{\infty}q_j=1 \tag{10}$$

的一组数, 并且服从方程:

$$q_j=\sum_{i=1}^{\infty}q_ip_{ij},\quad j\in E. \tag{11}$$

换一种提法, 可以认为, 平稳分布 $\mathbb{Q}=(q_1,q_2,\cdots)$ 是方程组

$$x_j=\sum_{i=1}^{\infty}x_ip_{ij},\quad j\in E \tag{12}$$

的一个解, 其中方程组服从*非负性条件和规范性条件*:

$$x_j\geqslant 0,\quad j\in E \quad 和 \quad \sum_{j=1}^{\infty}x_j=1.$$

如果满足定理 3 的条件, 则平稳解存在, 并且同时也是遍历的. 因此, 由定理 1 的命题 (c) 可见, 方程组 (12) 在序列类

$$(x_1,x_2,\cdots),\quad x_j\geqslant 0,\quad j\in E,\quad \sum_{j=1}^{\infty}x_j=1$$

中有解并且唯一.

实际上, 由此可以得到更多结论. 具体地说, 这里定理 3 的条件也成立, 从而存在遍历分布 $\Pi=(\pi_1,\pi_2,\cdots)$.

在此条件下, 我们考虑在如下 (*更广泛的*) 序列类

$$(x_1,x_2,\cdots),\quad x_j\in\mathbb{R},\quad j\in E,\quad \sum_{j=1}^{\infty}|x_j|<\infty,\quad \sum_{j=1}^{\infty}x_j=1$$

中, 方程组 (12) 解的存在性问题. 我们证明在该类中, 解唯一并且遍历分布 Π 就是该唯一解.

事实上, 如果 $(x_1,x_2,\cdots)$ 是解, 则由于 $\sum\limits_{j=1}^{\infty}|x_j|<\infty$, 可以得到如下一系列等式:

$$\begin{aligned}x_j=\sum_{i=1}^{\infty}x_ip_{ij}&=\sum_{i=1}^{\infty}\left(\sum_{k=1}^{\infty}x_kp_{kj}\right)p_{ij}=\sum_{k=1}^{\infty}x_k\left(\sum_{i=1}^{\infty}p_{ki}p_{ij}\right)\\&=\sum_{k=1}^{\infty}x_kp_{kj}^{(2)}=\cdots=\sum_{k=1}^{\infty}x_kp_{kj}^{(n)},\end{aligned}$$

其中 $n\geqslant 1$. 当 $n\to\infty$ 时求极限, (根据控制收敛定理) 对于任意 $k\in E$, 由此得

$$x_j=\left(\sum_{k=1}^{\infty}x_k\right)\pi_j,\quad \pi_j=\lim_n p_{kj}^{(n)}.$$

由于根据假设 $\sum\limits_{k=1}^{\infty}|x_k|=1$, 故 $x_j=\pi_j, j\in E$, 而这正是需要证明的.

6. 练习题

1. 对于转移概率矩阵为

$$\mathbb{P}=\begin{pmatrix}1/2 & 0 & 1/2 & 0\\ 0 & 0 & 0 & 1\\ 1/4 & 1/2 & 1/4 & 0\\ 0 & 1/2 & 1/2 & 0\end{pmatrix}$$

的马尔可夫链, 讨论其平稳分布, 极限分布, 遍历分布问题.

2. 设 $\mathbb{P}=(p_{ij})$ 是有限双随机矩阵 (即对于 $i=1,\cdots,m, \sum\limits_{j=1}^{m}p_{ij}=1$; 而对于 $i=1,\cdots,m, \sum\limits_{j=1}^{m}p_{ij}=1$). 证明, 对于相应的马尔可夫链, 向量 $\mathbb{Q}=(1/m,\cdots,1/m)$ 是平稳分布.

3. 设 $E=\{0,1\}$ 是马尔可夫链的状态空间, 而

$$\mathbb{P}=\begin{pmatrix}\alpha & 1-\alpha\\ 1-\beta & \beta\end{pmatrix},\quad 0<\alpha<1,\quad 0<\beta<1$$

是其转移概率矩阵. 讨论该马尔可夫链的, 极限分布, 遍历分布, 平稳分布问题.

§7. 有限马尔可夫链的极限分布、遍历分布和平稳分布

1. 有限链转移概率的渐近性质 根据 §5 定理 6, 任意不可约和非周期的、具有有限状态集的马尔可夫链, 都是正常返的. 这一事实使得有可能, 给 §6 的定理 3 以如下表述.(对照 §3 的问题, **A**, **B**, **C** 和 **D**).

定理 1 *假设 $X=(X_n)_{n\geqslant 0}$ 是具有有限状态集 $E=\{1,2,\cdots,r\}$ 的马尔可夫链, 并且是不可约的和非周期的.*

那么, 如下命题成立.

(a) *对于一切 $j\in E$, 存在不依赖于初始状态 $i\in E$ 的极限值 $\pi_j=\lim\limits_n p_{ij}^{(n)}$.*

(b) *极限值 $\Pi=(\pi_1,\cdots,\pi_r)$ 是概率分布, 即 $\pi_j\geqslant 0, \sum\limits_{i=1}^{r}\pi_i=1, j\in E$.*

(c) *此外, 对于一切 $j\in E$, 极限值 $\pi_j=\mu_j^{-1}>0$, 其中 μ_j 是首返状态 j 前度过的平均时间:*

$$\mu_j=\sum_{n=1}^{\infty}nf_{jj}^{(n)}=\mathbf{E}_j\tau(j),$$

其中 $\tau(j)=\inf\{n\geqslant 1: X_n=j\}$. 从而, 数组 $\Pi=(\pi_1,\cdots,\pi_r)$ 是遍历分布.

(d) *平稳分布 $\mathbb{Q}=(q_1,\cdots,q_r)$ 存在、唯一并且等于 $\Pi=(\pi_1,\cdots,\pi_r)$.*

2. 不可约性和非周期性对有限链的意义 作为定理 1 的补充, 我们引进如下定理, 其作用体现在 "不可约性" 和 "非周期性" 上.

定理 2 *假设 $X=(X_n)_{n\geqslant 0}$ 是具有有限状态集 $E=\{1,2,\cdots,r\}$ 的马尔可夫链, 并且是不可约的和非周期的.*

那么, 如下命题成立.

(a) *链是不可约的和非周期的 ($d=1$);*

(b) *链是不可约的, 非周期的 ($d=1$), 正常返的;*

(c) *链是遍历的;*

(d) *存在 n_0, 使对于一切 $n\geqslant n_0$, 有*

$$\min_{i,j\in E}p_{ij}^{(n)}>0.$$

证明 蕴含关系 (d)⇒(c), 在第一章 §12 定理 1 中已经证明. 相反的蕴含关系 (c)⇒(d) 显然. 蕴含关系(a)⇒(b), 由 §5 定理 6 可见; 而蕴含关系 (b)⇒(a) 显然. 最后, 命题 (b) 与 (c) 等价, 包含在 §6 定理 3 中. □

§8. 作为马尔可夫链的简单随机游动

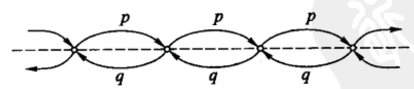
1. 简单随机游动 · 波利亚 (G. Polia) 定理 $d-$ 维简单随机游动指, 描绘 "质点" 在格子的结点 $\mathbb{Z}^d=\{0,\pm1,\pm2,\cdots\}^d$ 上运动的、齐次马尔可夫链, 这是该 "质点" 可能以某一概率逗留在每一状态, 也可能以一定概率转移到相邻的主题之一.

例 1 设 $d=1$, 而链的状态集为 $E=\mathbb{Z}=\{0,\pm1,\pm2,\cdots\}$, 其转移概率矩阵有如下形式:

$$p_{ij}=\begin{cases}p, & j=i+1,\\ q, & j=i-1,\\ 0, & \text{其他},\end{cases}$$

其中 $p+q=1$.

该矩阵对应于图



它直观上描绘该链可能的转移.

如果 $p=0$, 则指点肯定向左移; 如果 $p=1$, 则指点肯定向右移.

这些 "确定性" 情形没有什么意义, 况且这时的一切状态都是非本质的. 因此, 我们将假设 $0<p<1$.

在这样的条件下, 链由一个本质互通状态类组成. 换句话说, 在假设 $0<p<1$ 下, 链是不可约的 (见 §4).

对于任意 $j \in E$, 由二项分布的公式 (第一章 §2), 可见

$$p_{jj}^{(2n)} = \mathrm{C}_{2n}^{n}(pq)^n = \frac{(2n)!}{(n!)^2}(pq)^n. \tag{1}$$

根据司特林公式 (第一章 §2 的 (6) 式; 亦见练习题 1):

$$n! \sim \sqrt{2\pi n}n^n \mathrm{e}^{-n}.$$

从而, 由 (1) 式可见

$$p_{jj}^{(2n)} \sim \frac{(4pq)^n}{\sqrt{\pi n}}, \tag{2}$$

因此

$$\sum_{n=1}^{\infty} p_{jj}^{(2n)} = \infty, \quad 若\ p = q. \tag{3}$$

$$\sum_{n=1}^{\infty} p_{jj}^{(2n)} < \infty, \quad 若\ p \neq q. \tag{4}$$

由这些公式和 §5 的定理 1, 得到如下结果:

在集合 $E = \mathbb{Z} = \{0, \pm 1, \pm 2, \cdots\}$ 上的简单一维随机游动, 若 $p = q = 1/2$, 则是常返的和对称的; 而若 $p \neq q$, 则是非常返的.

在第一章 §10 曾经证明对于 $p = q = 1/2$ 的情形, 当 n 充分大时

$$f_{jj}^{(2n)} \sim \frac{1}{2\sqrt{\pi}n^{3/2}}. \tag{5}$$

因此

$$\mu_j = \sum_{n=1}^{\infty}(2n)f_{jj}^{(2n)} = \infty, \quad j \in E. \tag{6}$$

从而, 这时一切状态是*常返的和零的*. 因此根据 §5 的定理 5, 在一切 $0 < p < 1$ 的情形下, 对于任意 i 和 j, 当 $n \to \infty$ 时 $p_{ij}^{(n)} \to 0$. 由此可见 (§6 的定理 1), 极限分布, 以及平稳分布和遍历分布都*不存在*.

例 2 设 $d = 2$. 考虑对称的情形 (对应于例 1 中 $p = q = 1/2$ 的情形): 质点向左、向右的、向上或向下移动概率都等于1/4.

为确定计, 固定零状态 $\mathbf{0} = (0, 0)$. 假设开始处于零状态, 并讨论 "质点" 常返或者非常返零状态的问题.

为此, 我们讨论游动质点如下的那些 "轨道": 质点向右移动 i 步、向左 i 步、向上 j 步和向下 j 步. 假如 $2i + 2j = 2n$, 则说明 "质点" 从零点出发, 经过 $2n$ 步必然又返回该状态. 显然, 经奇数步 "质点" 不可能返回零状态.

由此可见, 从状态 $\mathbf{0}$ 仍然返回状态 $\mathbf{0}$ 的转移概率决定于下面的公式:

$$p_{\mathbf{00}}^{(2n+1)} = 0, \quad n = 0, 1, 2, \cdots$$

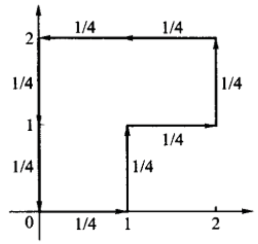


图 41 平面上的游动

且 (由全概率公式)

$$p_{\mathbf{00}}^{(2n)} = \sum_{(i,j):i+j=n} \frac{(2n)!}{(i!)^2(j!)^2}\left(\frac{1}{4}\right)^{2n}, \quad n = 1, 2, \cdots \tag{7}$$

(亦见第一章 §2 第 2 小节).

在 (7) 式中的和号下分式的分子与分母同乘以 $(n!)^2$, 得

$$p_{\mathbf{00}}^{(2n)} = \left(\frac{1}{4}\right)^{2n} \mathrm{C}_{2n}^{n} \sum_{i=0}^{n} \mathrm{C}_n^i \mathrm{C}_n^{n-i} = \left(\frac{1}{4}\right)^{2n} (\mathrm{C}_{2n}^{n})^2, \tag{8}$$

其中用到公式:

$$\sum_{i=0}^{n} \mathrm{C}_n^i \mathrm{C}_n^{n-i} = \mathrm{C}_{2n}^{n}$$

(第一章 §2 练习题 4).

根据司特林公式, 由 (8) 式可得 $p_{\mathbf{00}}^{(2n)} \sim \dfrac{1}{\pi n}$, 因而

$$\sum_{n=0}^{\infty} p_{\mathbf{00}}^{(2n)} = \infty. \tag{9}$$

根据对称性, 类似的论断, 当然不仅对零状态成立, 而且对于任何状态 (i, j) 也成立.

像 $d = 1$ 的情形一样, 由 §5 定理 1 的公式 (9), 得到如下结果:

在集合 $E = \mathbb{Z}^2 = \{0, \pm 1, \pm 2, \cdots\}^2$ 上的简单双对称随机游动是常返的.

例 3 结果表明, 当 $d \geqslant 3$ 时, 在状态 $E = \mathbb{Z}^d = \{0, \pm 1, \pm 2, \cdots\}^d$ 上对称随机游动, 非常明显的不同于上面讨论的 $d = 1$ 和 $d = 2$ 的情形.

具体地说,

在集合 $E = \mathbb{Z}^d = \{0, \pm 1, \pm 2, \cdots\}^d$ 上的简单 $d-$ 维对称随机游动, 对于任意 $d \geqslant 3$ 是非常返的.

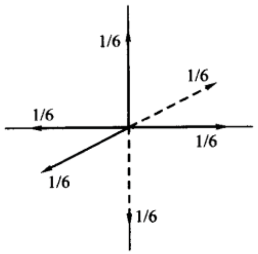
这一事实的证明基于如下考虑: 对于 $d \geqslant 3$, 当 $n \to \infty$ 时, 概率 $p_{jj}^{(2n)}$ 有如下渐近式:

$$p_{ij}^{(2n)} \sim \frac{c(d)}{n^{d/2}}, \tag{10}$$

其中 $c(d)$ 是某一依赖于维数 d 的正常数.

我们现在证明 $d=3$ 的情形, 而 $d>3$ 的情形的证明留做练习题.

由于随机游动对称性的假设, 可见"质点"以概率 $1/6$ 沿 (空间 3 条) 坐标轴的六个方向之一移动一步:



假设"质点"从点 $\mathbf{0}=(0,0,0)$ 出发. 那么, 像 $d=2$ 的情形, 由多项分布公式 (第一章 §2), 有

$$\begin{aligned} p_{\mathbf{00}}^{(2n)} &= \sum_{(i,j):0\leqslant i+j\leqslant n} \frac{(2n)!}{(i!)^2(j!)^2[(n-i-j)!]^2}\left(\frac{1}{6}\right)^{2n} \\ &= 2^{-2n}\mathrm{C}_{2n}^{n} \sum_{(i,j):0\leqslant i+j\leqslant n}\left[\frac{n!}{i!j!(n-i-j)!}\right]^2\left(\frac{1}{3}\right)^{2n} \\ &\leqslant \mathrm{C}_n 2^{-2n}\mathrm{C}_{2n}^{n}3^{-n} \sum_{(i,j):0\leqslant i+j\leqslant n}\frac{n!}{i!j!(n-i-j)!}\left(\frac{1}{3}\right)^{2n} \\ &= \mathrm{C}_n 2^{-2n}\mathrm{C}_{2n}^{n}3^{-n}, \end{aligned} \tag{11}$$

其中

$$\mathrm{C}_n = \max_{(i,j):0\leqslant i+j\leqslant n}\left[\frac{n!}{i!j!(n-i-j)!}\right], \tag{12}$$

并且利用了明显的公式:

$$\sum_{(i,j):0\leqslant i+j\leqslant n}\frac{n!}{i!j!(n-i-j)!}\left(\frac{1}{3}\right)^{2n} = 1.$$

下面将要证明

$$\mathrm{C}_n \sim \frac{n!}{[(n/3)!]^3}. \tag{13}$$

运用司特林公式, 由 (13) 式, 可得

$$\mathrm{C}_n 2^{-2n}\mathrm{C}_{2n}^{n}3^{-n} \sim \frac{3\sqrt{3}}{2\pi^{3/2}n^{3/2}}. \tag{14}$$

因而, 由 (11) 式, 可得

$$\sum_{n=1}^{\infty} p_{\mathbf{00}}^{(2n)} < \infty \tag{15}$$

于是, 根据 §5 的定理 1, 状态 $\mathbf{0}=(0,0,0)$ 是*非常返的*. 由于对称性, 对于 $E=\mathbb{Z}^3$ 中的任何其他状态, 有类似的结果.

只剩下证明 (13) 式.

设

$$m_n(i,j) = \frac{n!}{i!j!(n-i-j)!},$$

而 $i_0 = i_0(n), j_0 = j_0(n)$ 的值满足

$$\max_{(i,j):0\leqslant i+j\leqslant n} m_n(i,j) = m_n(i_0,j_0).$$

取 4 个点 $(i_0-1,j_0),(i_0+1,j_0),(i_0,j_0-1),(i_0,j_0+1)$. 由于相应的值 $m_n(i_0-1,j_0), m_n(i_0+1,j_0), m_n(i_0,j_0-1)$ 和 (i_0,j_0+1) 不大于 $m_n(i_0,j_0)$, 可得 4 个不等式:

$$\begin{aligned} n-i_0-1 &\leqslant 2j_0 \leqslant n-i_0+1, \\ n-j_0-1 &\leqslant 2i_0 \leqslant n-j_0+1. \end{aligned}$$

由这些不等式, 可得

$$i_0(n) \sim \frac{n}{3}, \quad j_0(n) \sim \frac{n}{3},$$

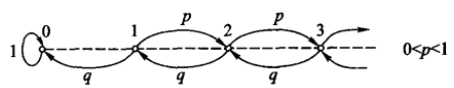
由此得所要证明得 (13) 式.

归纳所分析 $d=1,2,3$ 的情形, 得如下波利亚 (G. Polia) 的结果.

定理 *当 $d=1$ 或 $d=2$ 时, 在状态集 $E=\mathbb{Z}^d=\{0,\pm1,\pm2,\cdots\}^d$ 上的简单对称随机游动是常返的, 而当 $d=3$ (及 $d\geqslant 3$) 时是非常返的.*

2. 简单随机游动的例 ($E\subset\mathbb{Z}^d, d=1$) 上面的例子讲的是, 在"整个"空间 $\mathbb{Z}^d$ 中的简单随机游动. 在这一小节, 将要讨论的简单随机游动, 其相空间 E 严格包含在 $\mathbb{Z}^d$ 之内. 这里, 我们讲局限于 $d=1$ 的情形.

例 4 考虑简单随机游动, 其相空间 $E=\{0,1,2,\cdots\}$, 其中状态 0 是*吸收的*, 而其转移概率如下图所示:



这里, 状态 0 是正常返状态, 并且构成唯一不可约子类. (其余状态都是非常返的.) 根据 §6 的定理 2, 存在并且唯一平稳分布 $\mathbb{Q} = (q_0, q_1, \cdots), q_0 = 1, q_i = 0 (i = 1, 2, \cdots)$.

所考虑的随机游动直观上提供了一个例子: 对于某 i 和 j, 极限 $\lim\limits_n p_{ij}^{(n)}$ 存在, 但是依赖于初始状态, 同时说明在此随机游动的例子中不存在遍历分布..

显然, $p_{00}^{(n)} = 1$, 而对于 $j = 1, 2, \cdots, p_{0j}^{(n)} = 0$. 经简单运算, 可见对于一切 $i, j = 1, 2, \cdots, p_{ij}^{(n)} \to 0$.

现在证明对于一切 $i = 1, 2, \cdots$, 极限 $\alpha(i) = \lim\limits_n p_{i0}^{(n)}$ 存在, 并且对其有如下公式:

$$\alpha(i) = \begin{cases} (q/p)^i, & p > q, \\ 1, & p \leqslant q. \end{cases} \tag{16}$$

由该式可见, 对于 $p > q$ 的情形 ("有向右运动的倾向"), 由状态 $i (i = 1, 2, \cdots)$ 向状态 0 转移的极限概率 $\alpha(i) = \lim\limits_n p_{i0}^{(n)}$, 事实上依赖于 i, 并且随着 i 的增长以几何的速度减小.

为证明 (16) 式, 首先注意到, 由于状态 0 是吸收状态, 故 $p_{i0}^{(n)} = \sum\limits_{k \leqslant n} f_{i0}^{(k)}$, 从而极限 $\alpha(i) = \lim\limits_n p_{i0}^{(n)}$ 存在, 并且等于 f_{i0}, 即概率 $\alpha(i)$ 是 "质点" 由状态 i 出发, 迟早到达状态 "零" 的概率. 对于这些概率, 在第一章 §12 (亦见第七章 §2), 当 $\alpha(i) = 1$ 时导出了递推公式:

$$\alpha(i) = p\alpha(i+1) + q\alpha(i-1). \tag{17}$$

该方程的通解为

$$\alpha(i) = a + b(q/p)^i, \tag{18}$$

而由条件 $\alpha(0) = 1$, 可见常数 a 和 b 满足条件 $a + b = 1$.

如果假设 $q > p$, 则因为 $\alpha(i)$ 有界, 立即得 $b = 0$, 所以 $\alpha(i) = 1$. 这一结论十分清楚, 因为当 $q > p$ 时, "质点" 有沿到达零状态运动的倾向.

假如 $p > q$, 则情况相反: 有向右移动的倾向, 且自然想到

$$\alpha(i) \to 0, \quad i \to \infty, \tag{19}$$

即 $\alpha = 0$, 而

$$\alpha(i) = (q/p)^i. \tag{20}$$

为证明该等式, 我们并不先证明 (19) 式, 而是通过其他途径.

除在点 0 的吸收屏幕外, 我们在引进在点 N 的吸收屏幕. 以 $\alpha_N(i)$ 表示, 从点 i 出发的 "质点", 到达状态 0 早于到达状态 N 的概率. 概率 $\alpha_N(i)$ 满足边界条件为

$$\alpha_N(0) = 1, \quad \alpha_N(N) = 0$$

的方程 (17), 而且在第一章 §9 已经证明

$$\alpha_N(i) = \frac{\left(\dfrac{q}{p}\right)^i - \left(\dfrac{q}{p}\right)^N}{1 - \left(\dfrac{q}{p}\right)^N}, \quad 0 \leqslant i \leqslant N. \tag{21}$$

由此可见 $\lim\limits_N \alpha_N(i) = (q/p)^i$, 从而为证明 (20) 式的结果, 只需证明

$$\alpha(i) = \lim_N \alpha_N(i). \tag{22}$$

这在直观上很清楚. 严格的证明有如下一些途径.

假设 "质点" 从固定 i 状态出发. 那么,

$$\alpha(i) = \mathbf{P}_i(A), \tag{23}$$

其中 A 是事件 "存在 N, 使由 i 点出发的 '质点' 到达状态 0 早于到达状态 N". 如果

$$A = \{\text{"质点" 到达状态 0 早于状态} N\},$$

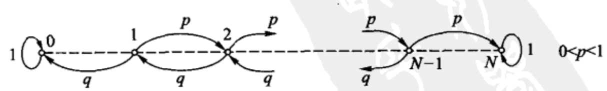
则 $A = \bigcup\limits_{N=i+1}^{\infty} A_N$. 显然, $A_N \subseteq A_{N+1}$, 且

$$\mathbf{P}_i\left(\bigcup_{N=i+1}^{\infty} A_N\right) = \lim_{N \to \infty} \mathbf{P}_i(A_N). \tag{24}$$

因为 $\alpha_N(i) = \mathbf{P}_i(A_N)$, 所以由 (23) 式和 (24) 式, 立即得 (22) 式.

这样, 如果 $p > q$, 则极限值 $\lim\limits_n p_{i0}^{(n)}$ 依赖于 i. 如果 $p \leqslant q$, 则对于任意 i, 极限值 $\lim\limits_n p_{i0}^{(n)} = 1$, 且 $\lim\limits_n p_{ij}^{(n)} = 0, j \geqslant 1$. 于是, 在这种情形下存在不依赖于 i 的极限分布 $\Pi = (\pi_0, \pi_1, \cdots)$, 其中 $\pi_j = \lim\limits_n p_{ij}^{(n)}$. 这时 $\Pi = (1, 0, 0, \cdots)$.

例 5 考虑简单随机游动, 其相空间 $E = \{0, 1, \cdots, N\}$, 其中 "边界" 状态 0 和 N 是吸收的:



这里, 存在两个不可约的正常返类: $\{0\}$ 和 $\{N\}$, 而所有其他状态类 $1, 2, \cdots, N-1$ 都是非常返的. 由 §6 定理 2 的证明可见, 存在平稳分布 $\mathbb{Q} = (q_0, q_1, \cdots, q_N)$ 的连续统, 且其平稳分布都具有如下形式: $q_1 = \cdots = q_{N-1} = 0, q_0 = a, q_N = b, a \geqslant 0, b \geqslant 0, a + b = 1$.

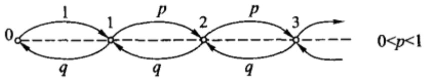
根据第一章 §9 第 2 小节, 有

$$\lim_n p_{i0}^{(n)} = \begin{cases} \dfrac{\left(\dfrac{q}{p}\right)^i - \left(\dfrac{q}{p}\right)^N}{1-\left(\dfrac{q}{p}\right)^N}, & p \neq q, \\ 1-\dfrac{i}{N}, & p=q=\dfrac{1}{2}, \end{cases} \tag{25}$$

$$\lim_n p_{iN}^{(n)} = 1-\lim_n p_{i0}^{(n)} \text{ 和 } \lim_n p_{ij}^{(n)} = 0, 1 \leqslant j \leqslant N-1.$$

需要强调, 这里及前面例子中, 极限值 $\lim_n p_{ij}^{(n)}$ 依赖于初始状态.

例 6 考虑简单随机游动, 其相空间 $E=\{0,1,\cdots\}$, 且在状态 0是*反射壁*:



所考虑的链本质上依赖于 p 和 q.

如果 $p>q$, 则游动的 "质点" 有向右移动的倾向. 而在 "零" 状态存在*反射壁*的情形, 与例 4 中游动相比仅仅在于, 对于例 4 的情形在 "零" 状态可以越过 "壁障". 这是全部状态都是非常返的; 对于一切 $i,j\in E, p_{ij}^{(n)}\to 0, n\to\infty$; 平稳分布和遍历分布都不存在.

如果 $p<q$, 则游动的质点有向左移动的倾向. 这时, 链是常返的. 当 $p=q$ 时链也是常返的.

现在给出平稳分布 $\mathbb{Q}=(q_0,q_1,\cdots)$ 应满足的方程组 (对照 §6 的 (12) 式):

$$q_0 = q_1 q,$$
$$q_1 = q_0 + q_2 q,$$
$$q_2 = q_1 p + q_3 q,$$
$$\cdots\cdots\cdots\cdots$$

由此, 有

$$q_1 = q(q_1+q_2),$$
$$q_2 = q(q_2+q_3),$$
$$\cdots\cdots\cdots\cdots$$

于是

$$q_j = \left(\frac{p}{q}\right) q_{j-1}, \quad j=2,3,\cdots$$

假如 $p=q$, 则 $q_1=q_2=\cdots$, 从而方程组没有满足条件

$$\sum_{j=1}^{\infty} q_j = 1, \quad q_0 = q_1 q$$

的非负解.

这意味着, 当 $p=q=1/2$ 时*不存在平稳解*. 这时链的全部状态都是*常返的*.

最后, 设 $p<q$. 由条件 $\sum\limits_{j=0}^{\infty} q_j = 1$, 可见

$$q_1\left[q+1+\frac{p}{q}+\left(\frac{p}{q}\right)^2+\cdots\right]=1.$$

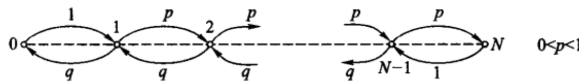
因此

$$q_1 = \frac{q-p}{2q}, \quad q_0 = q_1 q = \frac{q-p}{2},$$

而

$$q_j = \frac{q-p}{2q}\left(\frac{p}{q}\right)^{j-1}, \quad j \geqslant 2.$$

例 7 考虑简单随机游动, 相空间为 $E=\{0,1,\cdots,N\}$, 而状态 0 和 N 是*反射壁*:



这里, 链的状态是一个不可约类. 状态是周期为 $d=2$ 的正常返的. 根据 §6 的定理 2, 该链*有唯一平稳分布* $\mathbb{Q}=(q_0,q_1,\cdots,q_N)$. 在条件

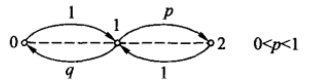
$$\sum_{i=0}^{\infty} q_i = 1, \quad q_j \geqslant 0, \quad j\in E,$$

下, 解方程组 $q_j = \sum\limits_{i=0}^{N} q_i p_{ij}$, 得

$$q_j = \frac{\left(\dfrac{p}{q}\right)^{j-1}}{1+\sum\limits_{i=1}^{N-1}\left(\dfrac{p}{q}\right)^{i-1}}, \quad 1 \leqslant j \leqslant N-1, \tag{26}$$

而 $q_0 = q_1 q, q_N = q_{N-1} q$.

由 §6 的定理 3, 以及所考虑链的周期为 $d=2$, 可见遍历分布不存在. 可以直接得出结论, 这里无遍历分布. 例如, 对于 $N=2$



那么, 易见 $p_{11}^{(2n)}=1$, 而 $p_{11}^{(2n+1)}=0$. 因而 $\lim\limits_{n} p_{11}^{(2n)}$ 不存在. 与此同时, 由 (26) 式可见, 平稳分布:

$$\mathbb{Q}=(q_0,q_1,q_2),$$

并且具有如下形式:

$$q_0=\frac{1}{2}q,\quad q_1=\frac{1}{2},\quad q_2=\frac{1}{2}p.$$

3. 利用简单随机游动描绘现实物理过程的示例 书中叙述的资料表明, *简单随机游动*是一个经典模型. 基于这一模型, "精练了" 概率的思想体系, 使得概率的技术更加完美, 发现了许多概率 - 统计的规律性. 例如, 考虑只有两个可能值的伯努利随机变量 $\xi_1,\xi_2,\cdots$ 之和 $X_n=\xi_1+\cdots+\xi_n$, 从而, $X=(X_n)_{n\geqslant 1}$ 是*简单随机游动* (也是马尔可夫链). 基于 $X=(X_n)_{n\geqslant 1}$, 发现了许多规律性, 诸如, 大数定律 (第一章 §5), 棣莫弗 - 拉普拉斯定理 (第一章 §6), 反正弦定律 (第一章 §10), 等等.

在这一小节, 我们考虑两个*离散扩散模型*, 它们是 "利用简单随机游动, 如何反映现实的物理过程" 的很好示例.

A. 埃伦弗斯特 (P. Ehrenfest, T. Ehrenfest) 模型

像例 7 一样, 我们将考虑简单随机游动, 假设其相空间为 $E=\{0,1,\cdots,N\}$, 且在状态 0 和 N 各有一个反射壁.

在状态 0 和 N 的转移概率为 $p_{01}=1$ 和 $p_{N,N-1}=1$. 在其余状态 $i=1,\cdots,N-1$, "质点" 只可能以概率

$$p_{ij}=\begin{cases}1-\dfrac{i}{N}, & j=i+1,\\ \dfrac{i}{N}, & j=i-1\end{cases}\tag{27}$$

向左或向右移动一步.

在 1907 年 P. 埃伦弗斯特和 T. 埃伦弗斯特 [124], 在研究统计力学的一种模型时, 得到了具有上述转移概率的马尔可夫链. 这一*统计力学中描绘气体分子的模型*是: 仪器有两个 (封闭的) 箱子 A 和 B, 由薄膜连接在一起, 气体分子由一个箱子 (A 或 B), 通过薄膜的 (微) 小孔移动到另一个箱子 (B 或 A).

假设在所观察的两个箱子中, 分子的总数等于 N, 且每一步从一个箱子向另一个箱子的移动是以如下的方式积极地进行的: 以概率 $1/N$ *随机地*选一个分子, 使它转移到另一个箱子; 并且每一步欲转移分子的选取, 与过去的 "历史" 无关.

设 X_n 是这样分子的个数: (例如) 分子一 "开始" 于时刻 n 在某一个箱子 A 中, 为描绘分子的移动的机制有马尔可夫性 (练习题 2):

$$\begin{aligned}&\mathbf{P}\{X_{n+1}=j|X_0=i_0,X_1=i_1,\cdots,X_{n-1}=i_{n-1},X_n=i\}\\&\qquad=\mathbf{P}\{X_{n+1}=j|X_n=i\},\end{aligned}\tag{28}$$

此外,

$$\mathbf{P}\{X_{n+1}=j|X_n=i\}=p_{ij},\tag{29}$$

其中 p_{ij} 由 (27) 式决定.

对于该模型存在由如下二项公式

$$q_j=\mathrm{C}_N^j\left(\frac{1}{2}\right)^N,\quad j=0,1,\cdots,N\tag{30}$$

决定的 (练习题3) *平稳分布* $\mathbb{Q}=(q_0,q_1,\cdots,q_N)$. 这里所有考虑的链是常返的 (练习题 4).

需要指出, 例如当 N 为偶数时. 概率 $q_j(j=0,1,\cdots,N)$ 在 "中间" 值 $N/2$ 上达到最大值, 这恰好对应于最可能的 "平衡" 状态, 即这时两个箱子中的分子数相同.

可以想到, 这样建立的随时间的 "平衡", 带有 (由上面引进的分布 $\mathbb{Q}$ 描绘的) *概率 - 统计*特点.

我们还要指出, 在直观上分子的数量随时间 "稳定的" 可能性十分明显: 状态 i 离 "中心" 值越远, (由于 (27) 式) 向该值*方向运动的概率就越大*.

B. D. 伯努利 - 拉普拉斯模型

考虑在一定意义上与埃伦弗斯特模型类似的模型. 该模型是 D. 伯努利 (1769 年) 提出的, 后来拉普拉斯 (1812 年) 在描绘*不可压缩液体* (而不是质点) 过程时提出来的.

更精确地说, 有两个样品箱 A 和 B, 含有 $2N$ 个质点, 其中N 个质点是 "白色" 的, 而另外 N 个质点是 "黑色" 的.

称 "系统" 处于状态 i, 其中 $i\in E=\{0,1,\cdots,N\}$, 如果在箱 A 中恰好有 i 个 "白色" 质点和 $N-i$ 个 "黑色" 质点. 假设 "不相容性" 是指, 对于所观测的状态 i, 在箱 B 中有 $N-i$ 个 "白色" 质点, 有 i 个 "黑色" 质点. 每箱中质点的总数保持为常数, 并且等于 N.

在每一步 n, 从每箱中*随机地* (即以概率 $1/N$) 选一个质点, 并且这些点改变位置. 假设这一 (随机) 从箱中选择质点的过程是独立进行的, 并且以后的选择按同样的模式进行, 而且与以往的阶段无关.

以 X_n 表示 A 箱中 "白" 色质点的个数. 那么, 质点变换的机制导致马尔可夫性 (28) 的成立, 这是 (29) 式中转移概率 p_{ij} 决定于如下表达式 (练习题 5):

$$p_{ij}=\begin{cases}\left(\dfrac{i}{N}\right)^2, & j=i-1,\\ \left(1-\dfrac{i}{N}\right)^2, & j=i+1,\\ 2\dfrac{i}{N}\left(1-\dfrac{i}{N}\right), & j=i,\end{cases} \tag{31}$$

其中 $p_{ij}=0$, 如果 $|i-j|>1, i=0,1,\cdots,N$.

像埃伦弗斯特模型一样, 这里的所有状态也都是常返的. 平稳分布 $\mathbb{Q}=(q_0, q_1,\cdots,q_N)$ 存在而且唯一, 并且由如下公式表示 (练习题 5):

$$q_j=\frac{(\mathrm{C}_N^j)^2}{(\mathrm{C}_{2N}^N)^2},\quad j=0,1,\cdots,N. \tag{32}$$

4. 现实物理过程的示例 (更加复杂的情形) 在这一章的开始曾经指出, 这一章的基本内容是, "(随着 n 的增长) 无后效系统的渐近性质", 上一节内容表明, 研究的是, 具有可数状态集 $E=\{0,1,\cdots,N,\cdots\}$ 的马尔可夫链的, 转移概率 p_{ij} 当 n 充分大时的性质, 包括简单随机游动, 转移只可能发生在相邻状态之间的情形.

很重要的是, 对于具有更加复杂状态空间的马尔可夫链研究类似的问题. 关于这类内容, 例如可以见 [75], [117].

5. 关于术语 "离散扩散模型" 的说明 上面讨论的两个模型 (埃伦弗斯特模型和伯努利 - 拉普拉斯模型) 曾称之为*离散扩散模型*.

现在关于这一名称做一些说明, 为此考虑 $\mathbb{R}$ 中随机游动的极限性质. 设 $S_n=\xi_1+\cdots+\xi_n, n\geqslant 1, S_0=0$, 其中 $\xi_1,\xi_2,\cdots$ 是独立同分布随机变量序列, 且 $\mathbf{E}\xi_i=0$, $\mathbf{D}\xi_i=1$. 设 $X_0^n=0$, 而

$$X_t^n=\frac{S_{[nt]}}{\sqrt{n}}\quad\left(=\frac{1}{\sqrt{n}}\sum_{k=1}^{[nt]}\xi_k\right),\quad 0<t\leqslant 1.$$

显然, 序列 $(0, X_{1/n}^N, X_{2/n}^N,\cdots,X_1^N)$ 可以视为在时刻 $\Delta, 2\Delta,\cdots,1$, 其中 $\Delta=1/n$, 跃度的量级为 $\sqrt{\Delta}(\Delta X_{k\Delta}^n\equiv X_{k\Delta}^n-X_{(k-1)\Delta}^n=\xi_k\sqrt{\Delta})$ 的简单随机游动.

像在第七章 §8 注 4 已经指出的那样, 这一随机游动 $X^n=(X_t^n)_{0\leqslant t\leqslant 1}$ 的全部有限维分布, 弱收敛于维纳过程 (布朗运动) $W=(W_t)_{0\leqslant t\leqslant 1}$; 此外, 在第七章 §8 注 4 中还指出, 函数收敛性也成立, 即过程 X^n 之分布向过程 W 之分布的弱收敛 (收敛的意义, 仍然是经验过程向布朗桥的收敛性; 见第三章 §13 的第 4 小节). 维纳过程是*扩散过程*典型的 (并且是基本的) 例子 (见 [69, V. II], [21], [131]). 这一事实恰好说明, X^n 类型的过程, 以及出现在埃伦弗斯特模型和伯努利 - 拉普拉斯模型中的, 过程为什么称为*离散扩散模型*.

6. 练习题

1. 利用如下概率的思想 ([106], 题 27.18), 证明斯特林公式: $n!\sim\sqrt{2\pi}n^{n+\frac12}\mathrm{e}^{-n}$. 设 $S_n=X_1+\cdots+X_n(n\geqslant 1)$, 其中 $X_1,X_2,\cdots$ 是独立且服从参数 $\lambda=1$ 的泊松分布的随机变量. 依次证明:

(a) $$\mathbf{E}\left(\frac{S_n-n}{\sqrt{n}}\right)^-=\mathrm{e}^{-n}\sum_{k=0}^{n}\left(\frac{n-k}{\sqrt{n}}\right)\frac{n^k}{k!}=\frac{n^{n+\frac12}}{n!}\mathrm{e}^{-n};$$

(b) $$\mathrm{Law}\left[\left(\frac{S_n-n}{\sqrt{n}}\right)^-\right]\to\mathrm{Law}\left[N^-\right],$$

其中 N 是正态分布的随机变量;

(c) $$\mathbf{E}\left(\frac{S_n-n}{\sqrt{n}}\right)^-\to\mathbf{E}N^-=\frac{1}{\sqrt{2\pi}};$$

(d) $$n!\sim\sqrt{2\pi}n^{n+\frac12}\mathrm{e}^{-n}.$$

2. 证明马尔可夫性 (28) 式.
3. 证明 (30) 式.
4. 证明埃伦弗斯特模型中马尔可夫链的所有状态都是常返的.
5. 验证 (31) 式和 (32) 式的正确性.

§9. 马尔可夫链的最优停时问题

1. 这一节的基本内容 下面讨论的内容, 与第七章 §13 紧密衔接, 这部分阐述任意随机序列的、最优停时问题求解的 "鞅" 方法. 这一节的基本内容, 涉及由马尔可夫链之状态的函数产生的随机序列, 可以赋予第七章 §13 的结果简单而直观的形式和解释.

2. 一步转移算子 假设 $X=(X_n,\mathscr{F}_n,\mathbf{P}_x)$ 是具有相空间 $(E,\mathscr{E})$ 的、离散时间齐次马尔可夫链.

亦假设 $(\Omega,\mathscr{F})$ 是坐标空间 (见 §1 的第 6 小节), 而 $X_n=X_n(\omega), n\geqslant 0$, 是定义在 $(\Omega,\mathscr{F})$ 上的随机变量, 并且 $X_n=X_n(\omega)$ 本身是以坐标的形式给定的: 如果 $\omega=(x_0,x_1,\cdots)\in\Omega$, 则 $X_n(\omega)=x_n$; 这里 $\mathscr{F}$ 理解为 $\sigma-$ 代数: $\mathscr{F}=\sigma(\bigcup\mathscr{F}_n)$, 其中 $\mathscr{F}_n=\sigma(x_0,\cdots,x_n), n\geqslant 0$.

注 在"最优停止规则的一般理论中"完全不要求 Ω 是坐标空间. 然而, 就是在"一般理论中"仍然需要假设 Ω 是充分"丰富"的 (详见 [78]).

对于我们来讲, 假设"坐标性"可以简化讨论, 特别对于广义马尔可夫性情形 (§2 定理 1), 正是在此假设下进行的.

像上一节一样, 以 $P(x;B)$ 表示所考虑的链的转移函数: $P(x;B) = \mathbf{P}_x\{X_1 \in B\}, x \in E, B \in \mathscr{E}$.

设 T 是作用于 $\mathscr{E}-$ 可测函数 $f = f(x)$ 的一步转移算子,

$$(Tf)(x) = \mathbf{E}_x f(x_1) \left(= \int_E f(y)P(x;\mathrm{d}y) \right), \tag{1}$$

假设其中 $\mathbf{E}_x|f(x_1)| < \infty, x \in E$. (为简便计, 记号 $(Tf)(x)$ 往往写成 $Tf(x)$. 在其他相近的情形下, 也采用类似的约定.)

3. 最优停时 为表述马尔可夫链 X 的最优停时问题, 假设给定某一 $\mathscr{E}-$ 可测实函数 $g = g(x)$, 满足条件: $\mathbf{E}_x|g(x_n)| < \infty, x \in E$, 对一切 $n \geqslant 0$ (或 $0 \leqslant n \leqslant N$, 如果事先存在某个"最后的"值 N, 而且在 N 之前应得到"最优解").

以 $\mathfrak{M}_0^n$ 表示 (关于过滤 $(\mathscr{F}_k)_{0\leqslant k\leqslant N}$ 的) 马尔可夫时间 $\tau = \tau(\omega)$ 类, 且在"停时"集合 $\{0, 1, \cdots, n\}$ 上取值.

下面的定理是, 第七章 §13 中定理 1 和定理 2 的"马尔可夫"提法.

定理 1 *设对于 $0 \leqslant n \leqslant N$ 和 $x \in E$, "(价) 值"为*

$$s_n(x) = \sup_{\tau\in\mathfrak{M}_0^n} \mathbf{E}_x g(X_\tau), \tag{2}$$

其中 $\mathbf{E}_x$ 表示对测度 $\mathbf{P}_x$ 求平均.

假设

$$\tau_0^n = \min\{0 \leqslant k \leqslant n : s_{n-k}(X_k) = g(X_k)\}, \tag{3}$$

而

$$Qg(x) = \max\{g(x), Tg(x)\}. \tag{4}$$

那么, 下面的命题成立,

1) 时间 τ_0^n 在类 $\mathfrak{M}_0^n$ 中是最优停时: 对于一切 $x \in E$,

$$\mathbf{E}_x g(X_{\tau_0^n}) = s_n(x). \tag{5}$$

2) 函数 $s_n(x)$ 可以按如下公式来求:

$$s_n(x) = Q^n g(x), \quad x \in E, \tag{6}$$

其中对于 $n = 0, Q^0 g(x) = g(x)$.

3) 函数 $s_n(x) (n \leqslant N)$ 满足递推关系式 $(s_0(x) = g(x))$:

$$s_n(x) = \max\{g(x), Ts_{n-1}(x)\}, \quad x \in E, \quad 1 \leqslant n \leqslant N. \tag{7}$$

证明 将第七章 §13 定理 1 和 2 的结果, 用于函数 $f_n = g(X_n), 0 \leqslant n \leqslant N$.

为此固定某个"初始"状态 $x \in E$, 并且考虑上面提到的 §13 中引进的函数 V_n^N 和 v_n^N. 这时, 为强调与初始状态有关, 我们特别使用记号 $V_n^N = V_n^N(x)$. 这样,

$$V_n^N(x) = \sup_{\tau\in\mathfrak{M}_n^N} \mathbf{E}_x g(X_\tau), \tag{8}$$

其中 $\mathfrak{M}_0^n$ 表示 (关于过滤 $(\mathscr{F}_k)_{k\leqslant N}$ 的) 所有马尔可夫时间 $\tau = \tau(\omega)$ 类, 且在"停时"集合 $\{n, n+1, \cdots, N\}$ 上取值.

由于第七章 §13 的 (6) 式, 函数 v_n^N 决定于如下递推公式:

$$v_N^N = g(X_N), \quad v_n^N = \max\left[g(X_n), \mathbf{E}_x(v_{n+1}^N|\mathscr{F}_n)\right]. \tag{9}$$

由于广义马尔可夫性 (§2 定理 1), ($\mathbf{P}_x$ – a.c.) 有

$$\mathbf{E}_x(v_N^N|\mathscr{F}_{N-1}) = \mathbf{E}_x(g(X_N)|\mathscr{F}_{N-1}) = \mathbf{E}_{X_{N-1}} g(X_1), \tag{10}$$

其中 $\mathbf{E}_{X_{N-1}} g(X_1)$ 应作如下理解 (见 §2): 取函数 $\psi(x) = \mathbf{E}_x g(X_1)$, 即 $\psi(x) = (Tg)(x)$, 而根据定义认为

$$\mathbf{E}_{X_{N-1}} g(X_1) \equiv \psi(X_{N-1}) = (Tg)(X_{N-1}).$$

这样, $v_N^N = g(X_N)$, 而

$$v_{N-1}^N = \max[g(X_{N-1}), (Tg)(X_{N-1})] = (Qg)(X_{N-1}). \tag{11}$$

类似地继续推导, 可见对于任意 $0 \leqslant n \leqslant N-1$, 有

$$v_n^N = (Q^{N-n}g)(X_n), \tag{12}$$

特别

$$v_0^N = (Q^N g)(X_0) = (Q^N g)(x), \quad (\mathbf{P}_x - \text{a.c.}).$$

根据第七章 §13 的 (13) 式, $v_0^N = V_0^N$. 由于 $V_0^N = V_0^N(x) = s_N(x)$, 可见 $s_N(x) = (Q^N g)(x)$. 于是, 对于 $n = N$ 就证明了 (6) 式. 类似地, 对于 $n < N$ 可以证明 (6) 式.

由 (6) 式和算子 Q 的定义可得 (7) 式.

现在证明, 当 $n = N$ 时, 由 (3) 式定义的停时 τ_0^n 在类 $\mathfrak{M}_0^N$ 中是最优的. 同样, 对于 $n < N, \tau_0^n$ 在类 $\mathfrak{M}_0^n$ 中也是最优的.

根据第七章 §13 的定理 1, 最优停时为

$$\tau_0^N = \min\{0 \leqslant k \leqslant N : v_k^N = g(X_k)\}.$$

由 (12) 式以及上面证明的事实: 对于任意 $n \geqslant 0, s_n(x) = (Q^n g)(x)$, 可见

$$v_k^N = (Q^{N-k}g)(X_k) = s_{N-k}(X_k), \tag{13}$$

从而

$$\tau_0^N = \min\{0 \leqslant k \leqslant N : s_{N-k}(X_k) = g(X_k)\}. \tag{14}$$

于是, 停时 τ_0^n 在类 $\mathfrak{M}_0^N$ 中的最优性得证. □

4. 停止区域和继续观测区域 记

$$\mathbb{D}_k^N = \{x \in E : s_{N-k}(x) = g(x)\}. \tag{15}$$

$$\mathbb{C}_n^k = E \setminus \mathbb{D}_k^N = \{x \in E : s_{N-k}(x) > g(x)\}. \tag{16}$$

那么, 由 (14) 式, 可见

$$\tau_0^N(\omega) = \min\{0 \leqslant k \leqslant N : X_k(\omega) \in \mathbb{D}_k^N\}, \tag{17}$$

而与第七章 §13 的第 6 小节引进的、 Ω 中的集合 D_k^N 和 C_k^N 类似, 集合

$$\mathbb{D}_0^N \subseteq \mathbb{D}_1^N \subseteq \cdots \subseteq \mathbb{D}_N^N = E, \tag{18}$$

$$\mathbb{C}_0^N \supseteq \mathbb{C}_1^N \supseteq \cdots \supseteq \mathbb{C}_N^N = \varnothing, \tag{19}$$

可以相应地称为 E 中的 "停止" 区域和 "继续" 观测区域.

我们指出, 所讨论的马尔可夫链的最优停时问题的特点. 与一般情形不同, 在马尔可夫链的情形下, 对 "停止观测还是继续观测" 问题的回答, 决定于马尔可夫链本身的状态 ($\tau_0^N(\omega) = \min\{0 \leqslant k \leqslant N : X_k(\omega) \in \mathbb{D}_k^N\}$), 换句话说, 决定于游动的 "质点" 所处的位置. 这时, 从原则的观点出发, 最优停时问题的完全解决 (即描绘 "(价) 值" $s_N(x)$ 和最优停时 τ_0^N), 在于从递推 "动态规划方程" (7) 式依次寻找函数 $s_0(x) = g(x), s_1(x), \cdots, s_N(x)$.

5. 类 $\mathfrak{M}_0^\infty$ 中的最优停时 设 $\mathfrak{M}_0^\infty$ 是一切有限马尔可夫停时类. 现在, 在假设 $\tau \in \mathfrak{M}_0^\infty$ 的条件下, 讨论最优停时问题.(对于所有 $\omega \in \Omega$, 当 $\tau \in \mathfrak{M}_0^N$ 时, 停时 $\tau \leqslant N$; 当 $\tau \in \mathfrak{M}_0^\infty$ 时, 停时 $\tau = \tau(\omega) < \infty$.)

这样, 假设 "(价) 值"

$$s(x) = \sup_{\tau \in \mathfrak{M}_0^\infty} \mathbf{E}_x g(X_\tau). \tag{20}$$

为使这里不出现数学期望 $\mathbf{E}_x g(X_\tau)$ 的存在性问题, 例如可以假设

$$\mathbf{E}_x \left[\sup_n g^-(X_n)\right] < \infty, \quad x \in E. \tag{21}$$

显然, 这样以来, 该式一定成立, 如果函数 $g = g(x)$ 有界 ($|g(x)| \leqslant C, x \in E$); 特别, 如果链的状态空间有限, 则条件 (21) 成立.

由 "(价) 值" $s_N(x)$ 和 $s(x)$ 的定义, 可见对于一切 $x \in E$, 有

$$s_N(x) \leqslant s_{N+1}(x) \leqslant \cdots \leqslant s(x). \tag{22}$$

自然, 应想到 $\lim\limits_{N\to\infty} s_N(x)$ 等于 $s(x)$. 而假如是这样, 则在 (7) 式中求极限, 将要得到的 "(价) 值" $s(x)$ 应满足方程:

$$s(x) = \max\{g(x), Ts(x)\}, \quad x \in E. \tag{23}$$

顺便指出, 由该方程可见, 对于 $s(x), x \in E$, 有 "变分不等式":

$$s(x) \geqslant g(x), \tag{24}$$

$$s(x) \geqslant Ts(x). \tag{25}$$

不等式 (24) 说明, "(价) 值" $s(x)$ 是 $g(x)$ 的控制函数. 第二个不等式 (25) 说明, 根据马尔可夫过程一般理论的定义, 函数 $s(x)$ 是*峰态的或上调和的*.

这样, 假如可以证明函数 $s(x)$ 满足 (23) 式, 则可以得出结论, "(价)值" $s(x)$ 是 $g(x)$ *峰态强函数*.

我们现在指出如下情形. 假如某一函数 $v(x)$ 是函数 $g(x)$ 峰态强函数, 那么显然有 "变分不等式":

$$v(x) \geqslant \max\{g(x), Tv(x)\}, \quad x \in E. \tag{26}$$

不过, 结果表明, 如果补充假设, 函数 $v(x)$ 是*最小峰态强函数*, 则在 (26) 式中将是等式, 即 $v(x)$ 满足方程:

$$v(x) = \max\{g(x), Tv(x)\}, \quad x \in E. \tag{27}$$

引理 1 *函数 $g(x)$ 的任何最小峰态强函数 $v(x)$, 都满足方程* (27).

证明 证明相当简单. 显然 $v(x)$ 满足不等式 (26). 记 $v_1(x) = \max\{g(x), Tv(x)\}$. 由于 $v_1(x) \geqslant g(x)$ 和 $v_1(x) \leqslant v(x)$, 可见

$$Tv_1(x) \leqslant Tv(x) \leqslant \max\{g(x), Tv(x)\} = v_1(x).$$

从而 $v_1(x)$ 是 $g(x)$ 的峰态强函数. 由于 $v(x)$ 是最小峰态强函数, 故 $v(x) \leqslant v_1(x)$, 即 $v(x) \leqslant \max\{g(x), Tv(x)\}$. 于是, 这连同不等式 (26) 就证明了等式 (27). □

所进行的基于假设 $s(x) = \lim\limits_{N\to\infty} s_N(x)$ 的预先讨论, 并且导出了 (23) 式, 以及引理 1 的命题提示表征 "(价) 值" 的途径, 看来 $s(x)$ 是函数 $g(x)$ 的*最小峰态强函数*. 而事实上确有下面的定理.

定理 2 *设函数 $g = g(x)$ 满足条件* (21):

$$\mathbf{E}_x \left[\sup_n g^-(X_n)\right] < \infty, \quad x \in E.$$

那么, 下列各命题成立:

(a) *价值 $s(x)$ 是函数 $g(x)$ 的最小峰态强函数.*

(b) 价值 $s(x)$ 等于 $\lim\limits_{N\to\infty} s_N(x) = \lim\limits_{N\to\infty} Q^N g(x)$, 并且满足 "瓦尔德 – 贝尔曼 (A. Wald – R. E. Bellman) 动态规划方程"

$$s(x) = \max\{g(x), Ts(x)\}, \quad x \in E.$$

(c) 如果

$$\mathbf{E}_x\left[\sup_n |g(X_n)|\right] < \infty, \quad x \in E,$$

则对于每一个 $\varepsilon > 0$, 停时

$$\tau_\varepsilon^* = \inf\{n \geqslant 0 : s(X_n) \leqslant g(X_n) + \varepsilon\}$$

在类 $\mathfrak{M}_0^\infty$ 中是 ε– 最优的, 即

$$s(x) - \varepsilon \leqslant \mathbf{E}_x g(X_{\tau_\varepsilon^*}), \quad x \in E.$$

如果 $\mathbf{P}_x\{\tau_0^* < \infty\} = 1, x \in E$, 则 τ_0^* 是最优的 (0 – 最优的), 即

$$s(x) = \mathbf{E}_x g(X_{\tau_0^*}), \quad x \in E. \tag{28}$$

(d) 如果集合 E 有限, 则停时 τ_0^* 属于 $\mathfrak{M}_0^\infty$, 并且是最优的.

注 完全可能, 对于某个状态 $x \in E$, 停时 $\tau_0^* = \inf\{n \geqslant 0 : s(X_n) = g(X_n)\}$ 以正概率等于 $+\infty, \mathbf{P}_x\{\tau_0^* = \infty\} > 0$. (甚至有可能对于可数个状态出现这样的情形; 练习题 1.) 因而, 需要约定, 在 "值 X_∞" 无定义的情况下, 应如何理解 "停时 τ 取 $+\infty$ 为值".

常规定 $g(X_\infty) \equiv \overline{\lim\limits_n} g(X_n)$ (见第七章 §13 第 1 小节; 亦见 [78]). 还有其他可能的解决办法: 用 $g(X_\tau)I(\tau < \infty)$ 代替 $g(X_\tau)$. 那么, 如果以 $\overline{\mathfrak{M}}_0^\infty$ 表示一切马尔可夫时间类, 其中有可能包含取 $+\infty$ 为值者, 则 "(价) 值"

$$\overline{s}(x) = \sup_{\tau \in \overline{\mathfrak{M}}_0^\infty} \mathbf{E}_x g(X_\tau) I(\tau < \infty) \tag{29}$$

有定义. 从而, 就可以在类 $\overline{\mathfrak{M}}_0^\infty$ 中讨论最优停时问题.

定理 2 的证明 我们仅对于有限集合 E 进行证明. 这种情形比较简单, 而且可以很好地在最优停时问题中显示峰态函数的产生. 关于一般情形的证明, 参见 [78], [102].

(a) 证明 $s(x)$ 是峰态函数, 即 $s(x) \geqslant Ts(x), x \in E$.

显然, 对于每一个状态 $y \in E$ 和 $\varepsilon > 0$, 存在 (一般依赖于 $\varepsilon > 0$ 的) ($\mathbf{P}$ – a. c.) 有限的停时 $\tau_y \in \mathfrak{M}_0^\infty$, 使

$$\mathbf{E}_y g(X_{\tau_y}) \geqslant s(y) - \varepsilon. \tag{30}$$

根据这一停时 $\tau_y, y \in E$, 建立一新停时 $\widehat{\tau}$: 形象地说, 使之决定下一步选择 "停时" 的 "策略".

假设开始时 "质点" 位于状态 E. "质点" 在此状态不会 "逗留", 然而却实现了一次观测. 设在时刻 $n = 1$, "质点" 处于状态 $y \in E$. 那么, 由 $\widehat{\tau}$ 表征的 "策略" 在于, 认为 "质点的生命" 仿佛重新开始, 而决定其停止的规则, 由停时 τ_y 控制.

而停时 $\widehat{\tau}$ 形式上由如下方式决定.

设 $y \in E$. 考虑事件 $\{\omega : \tau_y(\omega) = n\}, n \geqslant 0$. 由于 τ_y 是马尔可夫时间, 则该事件属于 $\mathscr{F}_n$. 假设 Ω 是由序列 $\omega = (x_0, x_1, \cdots), x_i \in E$, 生成的坐标空间, 而 $\mathscr{F}_n = \sigma(\omega : x_0, \cdots, x_n), x \in E$. 由此可见, 集合 $\{\omega : \tau_y(\omega) = n\}$ 可以写成 $\{\omega : (X_0(\omega), \cdots, X_n(\omega)) \in B_y(n)\}$, 其中 $B_y(n)$ 是 $\mathscr{E}^{n+1} = \mathscr{E} \otimes \cdots \otimes \mathscr{E}$ $(n+1$个 $\mathscr{E})$ 中的某个集合. (亦见第二章 §2 的定理 4.)

停时 $\widehat{\tau} = \widehat{\tau}(\omega)$ 是这样定义的, 其值形如 $n + 1 (n \geqslant 0)$, 这时, 在集合

$$\widehat{A}_n = \sum_{y \in E} \{\omega : X_1(\omega) = y, (X_1(\omega), \cdots, X_{n+1}(\omega)) \in B_y(n)\}$$

上, $\widehat{\tau}(\omega) = n + 1$. (直观上, 关于停时 $\widehat{\tau}$ 可以作如下说明: 对于任何状态 x, 在时刻 $n = 0$肯定进行观测; 假如这时 $X_1 = y$, 则下一步使用停时 τ_y.)

由于 $\sum\limits_{n \geqslant 0} \widehat{A}_n = \Omega$, 则停时 $\widehat{\tau} = \widehat{\tau}(\omega)$ 确实对于一切 $\omega \in \Omega$ 定义, 并且是马尔可夫的 (练习题 2).

由此构造, 广义马尔可夫性 (§2 中 (2) 式) 和 (30) 式可见, 对任意 $x \in E$, 有

$$\begin{aligned}
\mathbf{E}_x g(X_{\widehat{\tau}}) &= \sum_{n \geqslant 0} \sum_{y \in E} \sum_{z \in E} \mathbf{P}_x\{X_1 = y, (X_1, \cdots, X_{n+1}) \in B_y(n), X_{n+1} = z\} g(z) \\
&= \sum_{n \geqslant 0} \sum_{y \in E} \sum_{z \in E} p_{xy} \mathbf{P}_y\{X_0 = y, (X_0, \cdots, X_n) \in B_y(n), X_n = z\} g(z) \\
&= \sum_{n \geqslant 0} \sum_{y \in E} \sum_{z \in E} p_{xy} \mathbf{P}_y\{(X_0, \cdots, X_n) \in B_y(n), X_n = z\} g(z) \\
&= \sum_{y \in E} p_{xy} \mathbf{E}_y g(X_{\tau_y}) \geqslant \sum_{y \in E} p_{xy} [s(y) - \varepsilon] = Ts(y) - \varepsilon.
\end{aligned}$$

这样,

$$s(x) = \mathbf{E}_x g(X_{\widehat{\tau}}) \geqslant Ts(y) - \varepsilon, \quad x \in E,$$

而由于 $\varepsilon > 0$ 的任意性, 有

$$s(x) \geqslant Ts(x), \quad x \in E,$$

因此函数 $s = s(x), x \in E$ 的峰态性得证.

由上面证明的峰态性 (上调和性), 立即可以得到下面的重要结果.

系 1 对于任意 $x \in E$, 过程 (序列)

$$s = (s(X_n))_{n \geqslant 0} \tag{31}$$

(关于 $\mathbf{P}_x$ - 概率) 是上鞅.

将由第七章 §2 的定理 1 用于该上鞅, 可见对于任意停时 $\tau \in \mathfrak{M}_0^\infty$, 有如下不等式:

$$s(x) \geqslant \mathbf{E}_x s(X_\tau), \quad x \in E, \tag{32}$$

而且, 如果对于中的两个马尔可夫停时 σ 和 τ, 有 $\sigma \leqslant \tau(\mathbf{P}_x - \text{a.c.}, x \in E)$, 则

$$\mathbf{E}_x s(X_\sigma) \geqslant \mathbf{E}_x s(X_\tau), \quad x \in E. \tag{33}$$

(注意, 由于空间 E 有限, 故对于所考虑的情形, 上述第七章 §2 定理 1 的所有条件都成立).

由 (32) 式, 得

系 2 假设对于 (20) 式的最优停时问题, 函数 $g = g(x), x \in E$, 是峰态的 (上鞅的). 那么, 停时 $\tau_0^* \equiv 0$ 是最优停时.

(b) 证明 $s(x) = \lim\limits_N s_N(x), x \in E$.

因为 $s_N(x) \leqslant s_{N+1}(x)$, 所以极限 $\lim\limits_N s_N(x)$存在, 记作 $\overline{s}(x)$. 由于 E 有限, 且对于 $s_N(x)(N \geqslant 0)$, 有递推关系式,

$$s_N(x) = \max\{g(x), Ts_{N-1}(x)\},$$

则当 $N \to \infty$时求极限, 得

$$\overline{s}(x) = \max\{g(x), T\overline{s}(x)\}.$$

由此可见, $\overline{s}(x)$ 是函数 $g(x)$ 的峰态强函数. 但是, 由于 $\overline{s}(x)$ 是最小峰态强函数, 可见 $s(x) \leqslant \overline{s}(x)$. 此外, 由于对任意 $N \geqslant 0$, 显然 $s_N(x) \leqslant s(x)$, 可见 $\overline{s}(x) \leqslant s(x)$.

这样 $\overline{s}(x) = s(x)$, 从而证明了欲证的定理 2 的命题 (b).

(c,d) 证明命题 (c) 和 (d). 最后, 证明停时

$$\tau_0^* = \inf\{n \geqslant 0 : s(X_n) = g(X_n)\}, \tag{34}$$

即

$$\tau_0^* = \inf\{n \geqslant 0 : X_n \in \mathbb{D}^*\}, \tag{35}$$

首达 (停时) 集合

$$\mathbb{D}^* = \{x \in E : s(x) = g(x)\} \tag{36}$$

的时间, (在集合 E 有限的情形下) 在类 $\mathfrak{M}_0^\infty$ 中是最优的.

为此, 首先注意到, 由于满足 $g(\widetilde{x}) = \max\limits_{x \in E} g(x)$ 的 $\widetilde{x}$ 肯定属于 $\mathbb{D}^*$, 可见 $\mathbb{D}^*$ 不空. 由于在状态 $\widetilde{x}$ 下 $s(\widetilde{x}) = g(\widetilde{x})$, 则显然在最优策略应该是: 在状态 $\widetilde{x}$ 下立刻 "停止". 也正是这一点决定了停时 τ_0^*.

为从类 $\mathfrak{M}_0^\infty$ 中最优的角度考虑停时 τ_0^*, 首先需要断定此时间 τ_0^* 属于类 $\mathfrak{M}_0^\infty$, 即

$$\mathbf{P}_x\{\tau_0^* < \infty\} = 1, x \in E. \tag{37}$$

由所作关于状态集合 E 有限性假设, 确实可以证明这一点. (对于状态集合 E 可数的情形, 一般已经不是这样; 练习题 1).

为了证明上述事实, 我们指出, 我们感兴趣的事件 $\{\tau_0^* = \infty\}$ 等于事件

$$A = \bigcap_{n \geqslant 0} \{X_n \notin \mathbb{D}^*\}.$$

这样, 需要证明对于一切 $x \in E$, 有 $\mathbf{P}_x(A) = 0$.

如果 $\mathbb{D}^* = E$, 则这显然.

设 $\mathbb{D}^* \neq E$. 由于集合 E 有限性, 存在 $\alpha > 0$, 使对于一切 $y \in E \backslash \mathbb{D}^*, g(y) \leqslant s(y) - \alpha$.

那么, 对于任意 $\tau \in \mathfrak{M}_0^\infty$, 有

$$\begin{aligned}
\mathbf{E}_x g(X_\tau) &= \sum_{n=0}^{\infty} \sum_{y \in E} \mathbf{P}_x\{\tau = n, X_n = y\} g(y) \\
&= \sum_{n=0}^{\infty} \sum_{y \in \mathbb{D}^*} \mathbf{P}_x\{\tau = n, X_n = y\} g(y) + \sum_{n=0}^{\infty} \sum_{y \in E \backslash \mathbb{D}^*} \mathbf{P}_x\{\tau = n, X_n = y\} g(y) \\
&\leqslant \sum_{n=0}^{\infty} \sum_{y \in \mathbb{D}^*} \mathbf{P}_x\{\tau = n, X_n = y\} s(y) + \sum_{n=0}^{\infty} \sum_{y \in E \backslash \mathbb{D}^*} \mathbf{P}_x\{\tau = n, X_n = y\} [s(y) - \alpha] \\
&\leqslant \mathbf{E}_x s(X_\tau) - \alpha \mathbf{P}_x(A) \leqslant s(x) - \alpha \mathbf{P}_x(A),
\end{aligned} \tag{38}$$

其中最后一个不等式, 由函数 $s(x)$ 的峰态性 (上鞅性), 以及对其成立的不等式 (32) 得到. 在 (38) 式左侧对 $\tau \in \mathfrak{M}_0^\infty$ 取上确界, 得不等式

$$s(x) \leqslant s(x) - \alpha \mathbf{P}_x(A), \quad x \in E.$$

但是 $|s(x)| < \infty, \alpha > 0$. 因此 $\mathbf{P}_x(A) = 0, x \in E$, 从而就证明了停时 τ_0^* 的有限性.

现在证明停时 τ_0^* 在类 $\mathfrak{M}_0^\infty$ 中的最优性.

根据 τ_0^* 的定义

$$s(X_{\tau_0^*}) = g(X_{\tau_0^*}), \tag{39}$$

鉴于这一性质, 我们考虑函数 $\gamma(x) = \mathbf{E}_x g(X_{\tau_0^*}) = \mathbf{E}_x s(X_{\tau_0^*})$. 下面证明这一函数 $\gamma(x)$

1) 是峰态的;

2) 控制函数 $g(x): g(x) \leqslant \gamma(x), x \in E$;

3) $\gamma(x) \leqslant s(x)$,

其中最后一个不等式显然.

由 1) 和 2) 可见, $\gamma(x)$控制函数 $s(x)$, 同时又是函数 $g(x)$ 的最小峰态强函数. 因此, 由 3) 可见 $\gamma(x) = s(x), x \in E$, 从而

$$s(x) = \mathbf{E}_x g(X_{\tau_0^*}), \quad x \in E$$

由此可证 τ_0^* 在类 $\mathfrak{M}_0^\infty$ 中的最优性.

现在来证性质 1) 记 $\overline{\tau} = \inf\{n \geqslant 1 : X_n \in \mathbb{D}^*\}$. 该时刻是马尔可夫停时, $\tau_0^* \leqslant \overline{\tau}, \overline{\tau} \in \mathfrak{M}_1^\infty$, 由于函数 $s(x)$ 是峰态的, 故由性质 (33) 得

$$\mathbf{E}_x s(X_{\overline{\tau}}) \leqslant \mathbf{E}_x s(X_{\tau_0^*}), x \in E, \tag{40}$$

此外, 由于广义马尔可夫性 (见 §2 定理 1 的 (2) 式), 可见

$$\begin{aligned}\mathbf{E}_x s(X_{\overline{\tau}}) &= \sum_{n=1}^{\infty} \sum_{y \in \mathbb{D}^*} \mathbf{P}_x\{X_1 \notin \mathbb{D}^*, \cdots, X_{n-1} \notin \mathbb{D}^*, X_n = y\} s(y) \\ &= \sum_{n=1}^{\infty} \sum_{y \in \mathbb{D}^*} \sum_{z \in E} p_{xz} \mathbf{P}_z\{X_0 \notin \mathbb{D}^*, \cdots, X_{n-2} \notin \mathbb{D}^*, X_{n-1} = y\} s(y) \\ &= \sum_{z \in E} p_{xz} \mathbf{E}_z s(X_{\tau_0^*}). \end{aligned} \tag{41}$$

因此, 由 (40) 式, 可见

$$\mathbf{E}_x s(X_{\tau_0^*}) \geqslant \sum_{z \in E} p_{xz} \mathbf{E}_z s(X_{\tau_0^*}),$$

即

$$\gamma(x) \geqslant \sum_{z \in E} p_{xz} \gamma(z), \quad x \in E,$$

这就证明了函数的 $\gamma(x)$ 峰态性.

只剩下证明函数 $\gamma(x)$ 是 $g(x)$ 的强函数.

如果 $x \in \mathbb{D}^*$, 则 $\tau_0^* = 0$, 且显然 $\gamma(x) = \mathbf{E}_x g(X_{\tau_0^*}) = g(x)$.

考虑集合 $E \backslash \mathbb{D}^*$, 并且设 $E_0^* = \{x \in E \backslash \mathbb{D}^* : \gamma(x) < g(x)\}$. 这一集合 E_0^* 是有限的. 设 x_0^* 是函数 $g(x) - \gamma(x)$ 在集合 E_0^*上的极大值点:

$$g(x_0^*) - \gamma(x_0^*) = \max_{x \in E_0^*} [g(x) - \gamma(x)].$$

引进一新函数

$$\widetilde{\gamma}(x) = \gamma(x) + [g(x_0^*) - \gamma(x_0^*)], \quad x \in E. \tag{42}$$

该函数 (作为峰态函数与常数之和) 显然是峰态的, 并且对于一切 $x \in E$,

$$\widetilde{\gamma}(x) - g(x) = [g(x_0^*) - \gamma(x_0^*)] - [g(x) - \gamma(x)] \geqslant 0.$$

这样, 函数 $\widetilde{\gamma}(x)$ 是 $g(x)$ 的峰态强函数. 由于函数 $s(x)$ 是 $g(x)$ 的最小峰态强函数, 故 $\widetilde{\gamma}(x) \geqslant s(x)$.

由此可见,

$$\widetilde{\gamma}(x_0^*) \geqslant s(x_0^*).$$

但是, 由于根据 (42) 式 $\widetilde{\gamma}(x_0^*) = g(x_0^*)$, 故 $g(x_0^*) \geqslant s(x_0^*)$. 因为对于一切 $x \in E, s(x) \geqslant g(x)$, 所以 $g(x_0^*) = s(x_0^*)$, 说明点 x_0^* 属于集合 $\mathbb{D}^*$. 然而, 根据假设 $x_0^* \in E \backslash \mathbb{D}^*$.

所得矛盾说明集合 $E \backslash \mathbb{D}^* = \varnothing$. 由此可见, 对于一切 $x \in E, \gamma(x) \geqslant g(x)$. □

6. 例 现在举几个例子.

例 1 考虑 §8 例 5 中描绘的、具有两个 "吸收" 状态 0 和 N 的简单随机游动. 这时, 我们假设 $p = q = 1/2$ (对称游动). 对于所考虑的随机游动, 如果函数 $\gamma(x), x \in E = \{0, 1, \cdots, N\}$, 是峰态的, 则对于一切 $x = 1, \cdots, N-1$, 有

$$\gamma(x) \geqslant \frac{1}{2}\gamma(x-1) + \frac{1}{2}\gamma(x+1). \tag{43}$$

设给定一函数 $g = g(x), x \in \{0, 1, \cdots, N\}$. 由于 0 和 N 是吸收状态可见, 应当在满足条件 (43) 和边界条件 $\gamma(0) = g(0), \gamma(N) = g(N)$ 的所有函数 $\gamma(x)$ 中求函数 $s(x)$.

条件 (43) 表示函数 $\gamma(x)$ (在集合 $\{1, 2, \cdots, N-1\}$ 上) 的凸性. 因此, 可以得出如下结论: 在该问题中, "(价) 值" $s(x)$, 其中

$$s(x) = \sup_{\tau \in \mathfrak{M}_0^\infty} \mathbf{E}_x g(X_\tau)$$

是最小凸函数, 满足边界条件 $\gamma(0) = g(0), \gamma(N) = g(N)$. 直观上, 为求函数 $s(x)$的值, 需要按如下方式进行. 在函数 $g(x)$ 的值上 "栓" 上 "绷紧的线". 在图 42 "绷紧的线" 通过 $(0, a), (1, b), (4, c), (6, d)$ 4 个点, 其中 0,1,4,6 构成的 $\mathbb{D}^*$ 剩余集, 且在这些点上 $s(x) = g(x)$. 在其余状态 $x = 2, 3, 5$ 上, 要求 $s(x)$ 的值决定于线性内插. "凸流形" $s(x)$ 的值, 对于属于 E 的一切状态 x, 在一般情形下类似地决定.

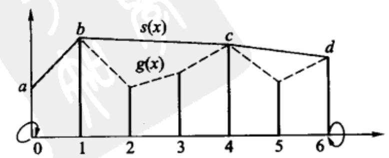


图 42 函数 $g(x)$ (虚线) 及其凸流形 $s(x), x = 0, 1, \cdots, 6$.

例 2 像 §8 中例 7 和例 8 一样, 考虑在状态集合 $E=\{0,1,\cdots,N\}$ 上且在 0 和 N 是反射屏幕的、对称 $(p=q=1/2)$ 简单随机游动. 所考虑的游动是正常返的. 由此可见, 在此最优停时问题中,

$$s(x)=\sup_{\tau\in\mathfrak{M}_0^\infty}\mathbf{E}_x g(X_\tau)$$

具有相当简单和自然的构造: 需要等到 "函数 $g(x)$ 达到最大值的任何一个状态" 的时刻, 并且在此时刻停止观测.

例 3 假设对于在集合 $E=\{0,1,\cdots,N\}$ 上的、简单对称随机游动, 0 是吸收状态而 N 是反射状态. 设 x_0是函数 $g(x)$ 达到最大值, 而且是游动离 N 最近的状态. 那么, 最优停时具有如下形式: 如果 $x_0\leqslant x\leqslant N$, 则当 (以 $\mathbf{P}_x$- 概率 1) 达到状态 x_0 时游动停止.

如果假设 $E=\{0,1,\cdots,x_0\}$, 而 x_0 是吸收状态, 则在状态 0 和 x_0 之间停止问题的解与例 1 一样.

7. 最优对象的选择问题 最后讨论有广泛知名度的 "最优对象的选择问题", 亦称 "选择审慎的未婚女问题", "选秘书问题", …… (参阅 [59], [78], [102], [106]). 为叙述的直观计, 我们选择 "选择审慎的未婚女问题" 的形式.

假设 "未婚女" 希望自 N 个候选人中选出最好的 "未婚夫". 假设 N 事先已知, 并且所有候选人事先已经按 "素质" 排序. 为确定计, 假设序号最大者, 即序号为 N 者是最优秀的, 按 "素质" 排第二位者的序号为 $N-1,\cdots$, 最差者序号为 1.

候选 "未婚夫" 按*随机*的顺序会见 "未婚女", 并且按如下方式实现.

设 $(a_1,a_2,\cdots,a_N)$ 是 $1,2,\cdots,N$ 的排列. 排列的总数等于 $N!$, 并且假设它们都是 "随机" 的, 即每一个排列出现的概率都等于 $1/N!$.

在该问题中, 样本 $(a_1,a_2,\cdots,a_N)$ 的*有序性*在于, (在会见全部候选人的潜在可能性的情况下) 第一个会见 "未婚女" 者是编号为 a_1 的候选人, 然后是 $a_2,\cdots$, 最后是编号为 a_N 的候选人.

加在 "未婚女" 可能 "策略" 由如下的想法形成.

"未婚女" 并 "不了解" 她会见的 "未婚夫" 候选人的完全的素质. 她关于候选人的全部了解, 仅仅是通过*两两比较*的结果, 知道哪个较好或较坏.

其次, 如果 "未婚女" 拒绝了某个 "未婚夫" 候选人, 则他就*不再与她会面* (然而被拒绝者实际上可能是最优秀的).

"未婚女" 的策略应该是, 根据顺序会见候选人的结果 (并记忆; 两两比较的结果并考虑 "素质"), 这样来选择停时 τ^*, 使得

$$\mathbf{P}\{a_{\tau^*}=N\}=\sup_\tau\mathbf{P}\{a_\tau=N\},\tag{44}$$

其中 τ 属于某一停时类 $\mathfrak{M}_1^\infty$, 而类 $\mathfrak{M}_1^\infty$ 决定于由 "未婚女" 在会见 "未婚夫" 候选人的过程中得到的 "信息".

为更确切地描绘所考虑的停时类 $\mathfrak{M}_1^\infty$, 我们根据序列 $\omega=(a_1,a_2,\cdots,a_N)$ 建立 "秩" 序列 $X=(X_1,X_2,\cdots)$, 该序列是由上面描绘的 "未婚女" 的行动方式自然产生的.

具体地说, 设 $X_1=1$, 并设 X_2 是 "未婚夫" 的大于以上所有编号的编号 (即到达的时刻). 这样, 如果 $X_2=3$, 则说明对于所讨论的序列 $\omega=(a_1,a_2,\cdots,a_N)$, 值 $a_1>a_2$, 而 $a_3>a_1(>a_2)$. 类似地继续定义 $X_3,X_4,\cdots$, 例如设 $X_3=5$. 因而 $a_3>a_4$, 而 $a_5>a_3(>a_4)$.

最多有 N 个强函数 (在 $(a_1,a_2,\cdots,a_N)=(1,2,\cdots,N)$ 的情形下). 假如对于序列 $\omega=(a_1,a_2,\cdots,a_N)$, 强函数的个数等于 m, 则设 $X_{m+1}=X_{m+2}=\cdots=N+1$.

所考虑的停时类 $\mathfrak{M}_1^\infty$, 由具有性质

$$\{\omega:\tau(\omega)=n\}\in\mathscr{F}_n^X$$

的停时 $\tau=\tau(\omega)$ 构成, 其中 $\mathscr{F}_n^X=\sigma(X_1,\cdots,X_n),1\leqslant n\leqslant N$.

现在比较详细地讨论 "秩" 序列 $X=(X_1,X_2,\cdots)$ 的构造.

不难证明 (练习题 3), 该序列是相空间为 $E=\{1,2,\cdots,N+1\}$ 齐次马尔可夫链. 该链的转移概率决定于公式:

$$p_{ij}=\frac{i}{j(j-1)},\quad 1\leqslant i<j\leqslant N,\tag{45}$$

$$p_{i,N+1}=\frac{i}{N},\quad 1\leqslant i\leqslant N,\tag{46}$$

$$p_{N+1,N+1}=1.\tag{47}$$

由此可见, $N+1$ 是*吸收状态*, 而且只有沿集合 E "向上" 转移是可能的, 即只可能转移 $i\to j$, 并且 $j>i$.

注 考虑到每一个序列 $\omega=(a_1,a_2,\cdots,a_N)$ 出现的概率都等于 $1/N!$, 故可由如下简单地讨论得到 (45) 式.

对于 $1\leqslant i<j\leqslant N$, 转移概率为

$$p_{ij}=\mathbf{P}\{X_{n+1}=j|X_n=i\}=\frac{\mathbf{P}\{X_n=i,X_{n+1}=j\}}{\mathbf{P}\{X_n=i\}}.\tag{48}$$

事件 $\{X_n=i,X_{n+1}=j\}$ 表示, a_j的值在 $a_1,\cdots,a_j$ 各值中最大的, 这时 $a_j>a_i$. 上述事件的概率等于

$$\mathbf{P}\{X_n=i,X_{n+1}=j\}=\frac{(j-2)!}{j!}=\frac{1}{j(j-1)}.$$

同样, 事件 $\{X_n=i\}$ 表示, a_i的值在 $a_1,\cdots,a_i$各值中是最大的, 而事件 $\{X_n=i\}$ 的概率等于

$$\mathbf{P}\{X_n=i\}=\frac{(i-1)!}{i!}=\frac{1}{i}.$$

由这些讨论和 (48) 式, 得 (45) 式.

为证明 (46) 式, 只需注意到, 如果 $X_n = i$, 则 $X_{n+1} = N+1$, 说明 a_i的值在 $a_1, \cdots, a_{i-1}$ 各值中, 以及在 $a_{i+1}, \cdots, a_{N+1}$ 中都是最大的.

最后, 公式 (47) 显然.

现在假设 "未婚女" 选择 (关于 σ- 代数系 $(\mathscr{F}_n^X)$ 的) 某一停时 τ, 并且使 $X_\tau = i$. 那么, 根据 (46) 式, 此时间 τ 结果是成功的 (即 $a_\tau = N$) 条件概率等于 $X_\tau/N(=i/N)$. 从而,

$$\mathbf{P}\{a_\tau = N\} = \mathbf{E}\frac{X_\tau}{N}.$$

因此, 求最优停时 τ^*, 即求满足

$$\mathbf{P}\{a_{\tau^*} = N\} = \sup_\tau \mathbf{P}\{a_\tau = N\}$$

的时间 τ^*, 归结为解最优停时问题

$$V^* = \sup_\tau \mathbf{E}\frac{X_\tau}{N}, \tag{49}$$

其中 τ 关于 σ - 代数系 $(\mathscr{F}_n^X)$ 是马尔可夫时间.

在 (49) 式中设 $X_1 = 1$. 根据对马尔可夫序列, 求解最优停时问题的一般方法, 记

$$v(i) = \sup_\tau \mathbf{E}_i g(X_\tau),$$

其中 $\mathbf{E}_i$ 是在 $X_n = i$ 的条件下的数学期望, 且

$$g(i) = \frac{i}{N}, \quad i \leqslant N, \quad g(N+1) = 0.$$

由定理 2 已知, 函数 $v(i), 1 \leqslant i \leqslant N+1$, 是函数 $g(i), 1 \leqslant i \leqslant N+1$, 的**峰态强函数**:

$$v(i) \geqslant Tv(i) = \sum_{j=i+1}^{N} \frac{i}{j(j-1)} v(j), \tag{50}$$

$$v(i) \geqslant g(i), \tag{51}$$

并且 (函数 $v(i)$) 是同类函数中最小的. 由同一定理 2, 可见满足方程:

$$v(i) = \max\{g(i), Tv(i)\}, \quad 1 \leqslant i \leqslant N+1. \tag{52}$$

这时, 不难看出, 所求函数 $v(i)$ 应当满足:

$$v(N+1) = 0, \quad v(N) = g(N) = 1.$$

以 $\mathbb{D}^*$ 表示停止观测的状态 $i \in E$ 的集合. 根据定理 1, 该集合可以用如下方式描绘:

$$\mathbb{D}^* = \{i \in E : v(i) = g(i)\}.$$

相应地, 继续观测的集合为

$$\mathbb{C}^* = \{i \in E : v(i) > g(i)\}.$$

这样, 如果 $i \in \mathbb{D}^*$, 则

$$g(i) = v(i) \geqslant Tv(i) = \sum_{j=i+1}^{N} \frac{i}{j(j-1)} v(j) \geqslant \sum_{j=i+1}^{N} \frac{i}{j(j-1)} g(j)$$
$$= \sum_{j=i+1}^{N} \frac{i}{j(j-1)} \cdot \frac{j}{N} = g(i) \sum_{j=i+1}^{N} \frac{1}{j-1}.$$

从而, 如果 $i \in \mathbb{D}^*$, 则应该成立不等式:

$$\sum_{j=i+1}^{N} \frac{1}{j-1} \leqslant 1.$$

此外, 如果该不等式成立, 且值 $i+1, \cdots, N$ 都属于 $\mathbb{D}^*$, 则

$$Tv(i) = \sum_{j=i+1}^{N} \frac{i}{j(j-1)} g(j) = g(i) \sum_{j=i+1}^{N} \frac{1}{j-1} \leqslant g(i),$$

因而, 状态 i 也属于集合 $\mathbb{D}^*$.

由 $v(N) = g(N)$, 可见 $N \in \mathbb{D}^*$. 因此, 上面的讨论说明, 集合 $\mathbb{D}^*$ 应该有如下形式:

$$\mathbb{D}^* = \{i^*, i^*+1, \cdots, N, N+1\},$$

其中 $i^* = i^*(N)$ 决定于不等式

$$\frac{1}{i^*} + \frac{1}{i^*+1} + \cdots + \frac{1}{N-1} \leqslant 1 < \frac{1}{i^*-1} + \frac{1}{i^*} + \cdots + \frac{1}{N-1}, \tag{53}$$

而对于充分大的 N, 由此得

$$i^*(N) \sim \frac{N}{\mathrm{e}}. \tag{54}$$

事实上, 对于任意 $n \geqslant 2$, 有

$$\ln(n+1) - \ln n < \frac{1}{n} < \ln n - \ln(n-1).$$

因此,

$$\ln\frac{N}{n} < \frac{1}{n} + \cdots + \frac{1}{N-1} < \ln\frac{N-1}{n-1},$$

由此连同 (53) 式得不等式

$$\ln\frac{N}{i^*(N)} < 1 < \ln\frac{N-1}{i^*(N)-2}.$$

于是, 由此不等式得渐近式 (54).

现在对于 $i \in E = \{1, 2, \cdots, N+1\}$, 求函数 $v = v(i)$.

如果 $i \in \mathbb{D}^* = \{i^*, i^*+1, \cdots, N, N+1\}$, 则 $v(i) = g(i) = i/N$.

设 $i = i^* - 1$, 则

$$v(i^*-1) = Tv(i^*-1) = \sum_{j=i^*}^{N} \frac{i^*-1}{j(j-1)} g(j) = \frac{i^*-1}{N}\left(\frac{1}{i^*-1} + \cdots + \frac{1}{N-1}\right).$$

设 $i = i^* - 2$, 则

$$\begin{aligned} v(i^*-2) = Tv(i^*-2) &= \frac{i^*-2}{(i^*-1)(i^*-2)} v(i^*-1) + \sum_{j=i^*}^{N} \frac{i^*-2}{j(j-1)} g(j) \\ &= \frac{1}{N}\left(\frac{1}{i^*-1} + \cdots + \frac{1}{N-1}\right) + \frac{i^*-2}{N} \sum_{j=i^*}^{N} \frac{1}{j-1} \\ &= \frac{i^*-1}{N}\left(\frac{1}{i^*-1} + \cdots + \frac{1}{N-1}\right). \end{aligned}$$

由归纳法可得, 对于一切 $1 \leqslant i < i^*$, 有

$$v(i) = v^*(N) = \frac{i^*-1}{N}\left(\frac{1}{i^*-1} + \cdots + \frac{1}{N-1}\right). \tag{55}$$

从而, 对于 $i \in \{1, 2, \cdots, N\}$, 有

$$v(i) = \begin{cases} v^*(N), & 1 \leqslant i < i^*(N), \\ g(i) = \dfrac{i}{N}, & i \leqslant N. \end{cases} \tag{56}$$

注意到 (55) 式, 由于

$$\lim_{N\to\infty}\left(\frac{1}{i^*(N)-1} + \cdots + \frac{1}{N-1}\right) = 1, \tag{57}$$

则当 $N \to \infty$ 时

$$\lim_{N\to\infty} v^*(N) = \lim_{N\to\infty} \frac{i^*(N)-1}{N} = \frac{1}{\mathrm{e}} \approx 0.0368. \tag{58}$$

初看来, 所得结果可能有些奇怪, 因为由此可见, 如果候选人数非常大, 则 "未婚女" 以很大的概率

$$V^* = \sup_{\tau} \mathbf{P}\{a_\tau = N\} = v^*(N) \approx 0.368,$$

存在从他们之中选择最优秀者的策略. 这时, 最优停时为

$$\tau^* = \inf\{n : X_n \in \mathbb{D}^*\},$$

其中 $\mathbb{D}^* = \{i^*, i^*+1, \cdots, N, N+1\}$.

这样, "未婚女" 最优策略是, 观察 $i^* - 1$个候选人, 其中

$$i^* = i^*(N) \sim \frac{N}{\mathrm{e}}, \quad n \to \infty,$$

然后, 在随后会见的候选人中, 选择第一个比以前会见的都优秀者.

在 $N = 10$ 的情形下, 较细致地分析表明 (例如, 见 [102] 的第三章 §1), $i^*(10) = 4$. 换句话说, 这时需要会见 3 个候选人, 并在随后会见的候选人中, 选择选择第一个比前 3 个都优秀者. 选中最优秀 "未婚夫" 相应的概率 (即 v^* (10) 的值) 等于 0.399.

8. 练习题

1. 举例说明, 对于具有可数状态集的马尔可夫链, (在类 $\mathfrak{M}_0^\infty$ 中) 有可能不存在最优停时.

2. 验证在定理 2 的证明中引进的时间 τ_y, 是马尔可夫时间.

3. 证明在第 7 小节, 讨论 "选择审慎的未婚女问题" 时引进的序列 $X = (X_1, X_2, \cdots)$, 是齐次马尔可夫链.

4. 设 $X = (X_n)_{n\geqslant 0}$ 是在 $\mathbb{R}$ 中取值的齐次马尔可夫链, 而其转移函数为 $P = P(x; B), x \in \mathbb{R}, B \in \mathscr{B}(\mathbb{R})$. 称 $\overline{\mathbb{R}}$- 函数 $f = f(x), x \in \mathbb{R}$, 是 P - 调和函数 (或关于 P 是调和函数), 如果

$$\mathbf{E}_x|f(X_1)| = \int_{\mathbb{R}} |f(y)| P(x; \mathrm{d}y) < \infty, \quad x \in \mathbb{R},$$

而

$$f(x) = \int_{\mathbb{R}} f(y) P(x; \mathrm{d}y), \quad x \in \mathbb{R}. \tag{59}$$

(如果将 (59) 式中的 "=" 换成 "$\geqslant$", 则函数 $f = f(x)$ 为上调和函数).

证明如果 f 是上调和函数, 则对于任意 $x \in \mathbb{R}$, 序列 $(f(X_n))_{n\geqslant 0} (X_0 = x)$ (关于测度 $\mathbf{P}_x$) 是上鞅.

5. 证明 (38) 式中的停时 $\overline{\tau}$ 属于类 $\mathfrak{M}_1^\infty$.

6. 仿照第 6 小节中的例 1, 对于 §8 的例中的所有简单随机游动, 考虑最优停时问题:

$$s_N(x) = \sup_{\tau \in \mathfrak{M}_0^N} \mathbf{E}_x g(X_\tau)$$

和

$$s(x) = \sup_{\tau \in \mathfrak{M}_0^\infty} \mathbf{E}_x g(X_\tau).$$

概率的数学理论形成的简史

在叙述概率论的历史问题时，可以程式化地分为如下的几个阶段 (对照 [26] *)，[43])：

史前

第一时期 (17 世纪 —18 世纪初)

第二时期 (18 世纪 —19 世纪初)

第三时期 (19 世纪后半叶)

第四时期 (20 世纪初和 20 世纪中叶)

史前. 关于随机性的直观印象，以及关于可能机会的各种不同的论断 (涉及宗教祭祀活动，解决纠纷，预测等)，进入世纪的纵深. 在前科学时代，还涉及一些人类尚未被认识的现象，以及尚无合理解释的现象，并且只有在几个世纪之前才开始思考和真正进行逻辑研究的现象.

考古资料表明，第一个 "随机工具" 的难得器件是，很久就被用于原始 (极简单) 的博弈的色子 (astragalus) **). 可以肯定地说，在如下一些年代，这样的色子在桌上的博弈中已经使用：埃及的第一个王朝 (大约公元前 3500 年)，之后在古希腊和古罗马. 众所周知 [20]，罗马皇帝奥古斯特 (August, 公元前 63 年 — 公元 14 年)，克劳迪厄斯 (Claudius, 公元前 10 年 — 公元 54 年)，曾经是古怪的色子博弈者.

与博弈相联系，那时已经出现了关于 "有利结局" 和 "不利结局" 的个数问题. 除博弈之外，在保险业和商业中出现了类似的问题. 已知保险业最早的形式，是在巴比伦的记事中发现的海运合同，大致在公元前 4000 年 — 公元前 3000 年. 后来，类

*) 这一部分的引文见第 894 — 898 页上文献索引.

**) astragalus 是一种刻有点 "•" 的 "双蹄目" 的骨头，其形状为：外形如正四棱柱，有 4 个平面，上下底面为球面. 掷到桌面上时，只有其中一个平面朝上.

似契约的实际经腓尼基① 人传给希腊人，罗马人，印度人. 其踪迹可以在罗马文化的早期法典和拜占庭帝国② 的法律中找到. 鉴于寿命保险的需要，罗马法学家尤尔皮安 (Ulpian) 于 (公元前 220 年) 编制了第一个死亡表. 在意大利的城市 - 共和国 (罗马，威尼斯，热那亚，比萨，佛罗伦萨)，鉴于保险的实际，出现了简单统计及实际核算的必要性. 众所周知，第一份确切注明日期的寿命保险的合同，于 1347 年在热那亚签订.

城市 - 共和国开创了文艺复兴时期 (Renaissance, 14 世纪末 — 17 世纪初) —— 在西欧的改造和更新时期. 看来，正是在意大利的文艺复兴时期，出现了或多或少严重的争议，主要是 L. 帕乔里 (Luca Pacioli, 1445 年 — 1517 年 (?))，C. 卡利卡尼尼 (Celio Calcagnini, 1479 年 — 1541 年) 和 N. F. 塔尔塔利亚 (Nicola Fontana Tartaglea, 1500 年 — 1557 年) 关于 "概率" 论断的争议，基本上是哲学性质的争议 (见 [43], [20]).

看来，G. 卡尔达诺 (Gerolamo Cardano, 1501 年 — 1576 年) 是最早开始从数学上分析博弈结局者之一，他是广为人知的 "卡尔达诺轴" 的发明者，并且求解了 3 次方程. 他的文献抄本 (大约 1525 年)，只有在 1563 年，以 "Liber de Ludo Aleæ" (《关于博弈的书》) 书名才出版，该书并不仅仅是某种赌徒的实际参考书. 在该书中最早阐述了组合的思想，利用组合方法便于描绘一切可能结局的集合 (在不同次数地，投掷不同特点的色子时，可能结局的集合). G. 卡尔达诺还发现，对于均匀对称的色子，"'有利组合数' 与 '一切可能组合总数' 的比值，与博弈的实际十分一致" [20].

1. 第一时期 (17 世纪 — 18 世纪初) 许多人，例如，拉普拉斯 [39] (亦见 [61]) 把《概率演算》(《概率论》) 1654年登记与帕斯卡 (Blaise Pascal,1623 年 — 1662 年) 及费马 (Pierre de Fermat, 1601 年 — 1665 年) 的通信相联系. 通信与勋章获得者 de 梅尔 (Chevalier de Méré, 他也是 Antoine Gombaud —— 作家和道学先生，1607 年 — 1684 年) 给帕斯卡提出的某些问题有关.

这些问题之一就是，在中断的博弈中，如何公平地分配赌注. 具体地说，假设两个选手 A 和 B 约定，在对弈中 (例如，共对弈 5 局) 获胜者将得到全部赌注. 假设当选手 A 获 4 胜而选手 B 获 3 胜时，则对弈强行停止. 问在此被停止的对弈中，选手应按什么比例分配赌注？仿佛该问题的 "自然" 答案之一就是，应当按 $2:1$ 的比例分配赌注. 实际上，对弈经两步肯定结束，这时选手 A 只需赢一局，而选手 B 需要赢两局. 由此可见，仍导致 $2:1$ 的比例.

然而，按局数赢得对弈的选手，"自然" 也可以认为是 $4:3$. 如帕斯卡和费马认为两个都不是：应当按 $3:1$ 的比例分配赌注.

另一个问题是，哪一种情形更有可能：将一枚均匀对称的色子掷 4 次，"6 个点至少出现一次"，还是将两枚均匀对称的色子同时掷 24 次，"数偶 (6,6) 至少出现一次".

①腓尼基 (Phocnicia) 地中海东岸的古国，曾在地中海开辟了许多殖民地，公元前6 世纪被波斯征服，公元前 332 年被马其顿王朝亚历山大占领. —— 译者

②拜占庭 (Byzantium) 帝国，亦称东罗马帝国 (公元 4 — 5 世纪) 首都君士坦丁堡. —— 译者

对于上述问题，帕斯卡和费马给出了正确的答案：第一种情形比第二种情形可能性更大. 两种情形的概率分别等于

$$p_1 = 1 - \left(\frac{5}{6}\right)^4 \approx 0.516 \quad 和 \quad p_2 = 1 - \left(\frac{35}{36}\right)^{24} \approx 0.491.$$

在求解这些问题时，帕斯卡和费马 (像卡尔达诺一样) 广泛运用了组合分析的方法. 组合分析法，在计算不同结局的个数时，成为 "概率演算" 的基本方法之一. 较早就著名的帕斯卡三角形①，在这里也找到了其 "应用" 的位置.

在 1657 年出版了 C. 惠更斯 (Christianus Huygens, 1629 年 — 1695 年) 的书《De Retiociniis in Lugo Aleæ》(《博弈中的计算》)，被认为是 "概率演算" 的第一部系统的著作. 书中以明显的形式，表述了许多基本概念，概率的演算原则，引进了概率的加法法则和乘法法则，还包括关于数学期望概念的讨论. 在很常时间里，该书是《初等概率论》的主要参考书.

所涉及的 "概率论" 形成时期的核心代表人物是，J. 伯努利 (Jacob〈Jakob, James, Jacques〉Benoulli, 1654 年 — 1705 年). J. 伯努利的功勋在于，他把 "有关所考虑事件的可能结局数" 与 "可能结局的总数" 的比值，作为 "事件概率" 的 "古典型" 概念引进了科学.

J. 伯努利的与其名字联系在一起的基本成果，当然是作为概率论一切应用基础的*大数定律*.

J. 伯努利于 1713 年，将这一定律用极限定理的正式形式，(在其侄子 N. 伯努利 [Nikolaus Benoulli] 的参与下) 出版了的专著《Ars Conjectandi》(《假设的艺术》)，人们把该专著出版的日期当作大数定律的生日 (见 [3] 第 9, 27, 75, 83 页). 像 A. A. 马尔可夫在纪念大数定律 200 周年时的发言中 (见 [53],[3]) 指出的，J. 伯努利在 (1703 年 10 月 3 日和 1704 年 4 月 20 日) 给 G.W. 莱布尼茨 (Gottfried Wilhelm Leibniz, 1646 年 — 1716 年) 信中写到，他 "在 20 年前已经知道" 这个定理. ("大数定律" 这一术语是泊松 1835 年建议启用的.)

伯努利家族的另一名代表，是 D. 伯努利 (Daniel Benoulli, 1667 年 — 1748 年)，因为他在概率论中关于所谓 "彼得堡奇论" 的争议而知名. 对于该问题的解，他使用了 "道德期望" 的概念.

概率论形成的第一个时期，正是数学科学建立的时代. 使用诸如连续性，无限大和无限小等概念，正是属于这个时期. I. 牛顿 (Isaac Newton, 1642 年 — 1727 年) 和 G. W. 莱布尼茨创建微分学与积分学，也是在这个时期. 像 A. H. 柯尔莫戈洛夫 [26] 指出的那样，这个时期的任务是，"理解研究因果关系的、数学方法不寻常的广度和深度 (而当时似乎是万能的). 作为根据系统现在的状态，唯一决定系统将来的变化的定律的微分方程的思想研究，与现在相比在数学科学中还占有相当特殊的地位. 在

①即杨辉三角形. 杨辉 (中国数学家, 13 世纪, 南宋)，帕斯卡 (17 世纪, 法国数学家). —— 译者

数学科学中，在微分方程的确定性概型不能应用的地方，需要概率论. 在当时具体的自然科学的资料，对于计算和事物性的应用，尚未用到概率论.

然而，对于微分方程组一类的确定性模型已经十分明显，引进实际现象的、深入的模式化是不可避免的. 同样明显的是，对于无法单独考虑，且相互之间没有关联的大量现象混乱的基础上，'平均' 来讲可能出现完全清晰的规律性"，这充分地揭示了 J. 伯努利的极限定理 —— 大数定律的意义.

必须指出，J. 伯努利意识到，考虑重复试验结果的无限序列的重要性. 关于事件在试验中出现频率的*极限性质*提法本身，在局限于初等算术和简单组合方法的、概率论的研究中曾经是新的 ("无限的") 思想. 恰好是导致大数定律的、问题这样的提法，不但显示出，"事件出现的*概率*" 与 "事件在有限次重复试验中出现的*频率*" 之间的差异，而且显示出，用试验次数充分大时频率的值，以一定的准确性确定事件概率的可能性.

2. 第二时期 (18 世纪 — 19 世纪初) 这个时期主要有这样一些代表人物：P. R. 蒙特莫尔特 (Pierre-Remond de Montmort, 1678 年 — 1719 年), Dé. 棣莫弗 (Abraham Dé Moivre, 1667 年 — 1754 年), T. 贝叶斯 (Thomas Bayes, 1702 年 — 1761 年); P. S. 拉普拉斯 (Pierre Simon de Laplace, 1749 年 — 1827 年), C. F. 高斯 (Carl Friedrich Gauss, 1777 年 — 1855 年); S. D. 泊松 (Siméon Denis Poisson, 1781 年 — 1840 年).

如果说第一时期实质上带有*哲学*的特点，那么在第二时期是发展和*分析方法*的精练. 在各种不同领域出现了进行计算的必要性，将概率统计方法用于观测误差理论，射击理论等.

蒙特莫尔特，以及棣莫弗受到 J. 伯努利在 "概率的演算" 工作的强烈影响. 蒙特莫尔特在他的书《Essai d' Analyse sur les jeux de Hasard》(《随机博弈试验分析》, 1708 年) 中，重点就是不同博弈中核算方法的展开.

棣莫弗在他的两部书中：《Doctrine of Chances》(《偶然性的理论》; 1718 年),《Miscellanea Analytica Supplementum》(《分析方法》或《分析混合体》, 1730 年), 相当详细地给出了如下一些定义：事件的*独立性*，*期望*，*条件概率*.

棣莫弗最著名的是二项分布的正态逼近. 假如说伯努利大数定律，是频率在多方面 "平均" 服从某种明显的规律性 (按收敛的形式，事件出现的频率在一定意义上收敛于其概率)，那么棣莫弗发现的正态逼近，揭示了关于*偏差*对平均水平性质的另一种广泛的规律性. 棣莫弗的结果及其后来的推广是如此地卓著，使得人们把 "积分极限定理" 称为概率论的*中心极限定理*.(使用这一术语，是 G. 波利亚 (George Pólya, 1887 年 — 1985 年) 1920 年建议的, [55].)

在所涉及的时期，最著名的无疑是拉普拉斯. 他在 1812 年出版的专著《Théorie Analytique des Probabilités》(《概率的分析理论》)，是 19 世纪概率论的主要参考书. 除天文学和数学分析方面的工作之外，他还撰写了若干关于概率演算的基础，哲

学问题和具体问题的论文. 拉普拉斯在误差理论方面的贡献卓著. 具体地说, 在误差理论中引进正态律自然的思想属于他和高斯, 正态律是大量独立的、最简单的误差叠加的总效应而产生的结果. 拉普拉斯不仅给棣莫弗积分定理以更加一般的提法 ("棣莫弗 – 拉普拉斯定理"), 并且提出了新的分析证明.

继 J. 伯努利之后, 在有限个可能结局的情形下, 拉普拉斯明确地依照导致概率概念的 "古典型" 定义的、"等可能性原则" 或 "无差异原则".

然而, 在这一时期就已经出现了, 不能置于古典概型中的, "非古典型" 概率分布. 例如, 出现了正态分布和泊松分布. 不过, 这两个分布长时间仅仅被视为某种逼近, 而不是 (像现代对术语的理解) 看成概率分布.

另一个 "非古典型" 分布的例子, 就是 "几何概率的" 问题 (例如, 见牛顿 1665 年, [52] 的 p.60). 这里还有著名的 "蒲丰针" 问题. 与贝叶斯公式相联系, 亦出现了不同的概率. *贝叶斯公式*发表在 1763 年的论文 "An Essay Towards Solving a Problem in the Doctrine of Chances" 中, 在试验出现了某一事件的情况下, 该公式给出了重新计算先验概率的规则 (贝叶斯认为两个概率是相同的)① . 由该公式在统计学中产生了一个完整研究方向, 如今称为 "贝叶斯方法".

综上所述可见, 在概率论的 "古典型" (有限) 框架内本质上制约了其发展和应用的可能性, 而对于正态分布, 泊松分布, 以及其他分布的解释只能局限于某种*极限*形式, 因而产生一种不完善的感觉. 在这个时期, 在概率论中缺乏抽象的数学概念, 而且它还没有被当作应用数学. 况且, 其方法还局限在 (诸如, 赌博, 误差理论, 射击理论, 保险, 人口学 ······) 的具体应用.

3. 第三时期 (19 世纪后半叶) 在这个时期, 彼得堡占据了概率论一般问题的基本位置: П. Л. 切比雪夫 (П. Л. Чебышёв, 亦译 "切贝绍夫", 1821 年 — 1894 年), A. A. 马尔可夫 (A. A. Марков, 1856 年 — 1922 年), 和 A. M. 李雅普诺夫 (A. M. Ляпунов , 1857 年 — 1918 年), 在概率论的整个系统的扩展和深入方面, 作了重要贡献. 具体地说, 由于他们的工作, 突破了 "古典型" 概率的机会的框架. 切比雪夫非常明确地评价了随机变量的概念, 数学期望的概念的作用, 并且有效地演示了这些观念的适用性, 当然这些在现在看来都是很平常的.

大数定律, 棣莫弗 – 拉普拉斯定理涉及仅有两个可能值的随机变量. П. Л. 切比雪夫本质上扩展了这些定理的适用范围 (使之适用于更一般的随机变量). 例如, 他的第一个成果是, 证明了大数定律对于任意独立随机变量之和成立, 只要这些随机变量的绝对值都不大于某一常数. (下一步工作是 A. A. 马尔可夫完成的, 证明用到 "切比雪夫 – 马尔可夫不等式".)

在大数定律之后, П. Л. 切比雪夫转向 "对于独立随机变量之和, 棣莫弗 – 拉普拉斯定理的正确性", 为此他提出了新的证明方法 —— *矩法*, 后来由 A. A. 马尔可夫实现.

① "априори" (a priori) 汉语常译为 "先验", 有的译为 "验前". —— 译者

在寻找棣莫弗 – 拉普拉斯定理成立的一般条件的过程中, A. M. 李雅普诺夫迈出了意想不到的一步, 他用由拉普拉斯开始的特征函数方法证明了, 只要存在 $\delta > 0$ 使被加独立随机变量有 $2 + \delta$ 阶矩, 而没必要存在一切阶矩, 即可证明该定理. 此条件称做 "李雅普诺夫条件".

作为原则上新的概念应该指出, A. A. 马尔可夫引进的、具有 "无后效" 性的、相依 (*非独立*) 随机变量, 现在称做 "马尔可夫链". 并且对于马尔可夫链, 首先严格证明了 "遍历性" 定理.

可以确切地断言, П. Л. 切比雪夫, A. A. 马尔可夫和 A. M. 李雅普诺夫 ("彼得堡学派") 的工作, 为后来概率论的全部发展奠定了坚实的基础.

19 世纪的后半叶, 在西欧由于发现概率论与纯数学、统计物理的深刻联系, 以及数理统计蓬勃发展, 对于概率论的兴趣开始急速增长.

在这个时期, 概率论本身的发展, 越来越明显地受到其 "古典型" 假设 (结局的有限性及其等可能性) 的强烈制约, 并且需要在纯数学中寻找相应的扩展. (这里应注意, 当时集合论才刚刚建立, 而测度论刚刚处于创立的 "门槛").

与此同时, 在纯数学中, 特别在数论中, 科学仿佛离概率论十分遥远. 人们开始利用概率的概念, 并且得到了纯 "概率" 本性的结果, 开始对于概率的直观产生兴趣.

例如在 1890 年, J. 庞加莱 (Jules Henri Poincaré, 1854 年 — 1912 年) 在其关于 "三物体" 的论文中得出如下结果: 描绘保持 "体积" 变换 T 的、动态系统运动的返回性的结果说明, 如果 A 是初始状态 ω 的集合, 则对于 "典型的" $\omega \in A$, 运动的轨道 $T^n\omega$ 将无限多次返回集合 A. (按现代语言, 回返不是对于所有初始状态, 而仅对于*几乎所有*初始状态.)

在这个时期的研究中, 人们常提到 "随机抽样", "典型情形", "特殊情形". 在 J. 庞加莱的教科书《Calcul des Probabilités》([56], 1896 年) 中, 提出一个问题 "在区间 [0,1] 上*随机*选取的一点, 恰好是有理数的概率如何?"

1888 年天文学家 J. A. H. 盖尔登 (Johan August Hugo Gyldén, 1841 年 — 1896 年) 发表了论文 [18], 论文的起源 (像 J. 庞加莱 [57], 1890 年的论文一样) 与星球的稳定性有关, 而现在应纳入概率数论. 论文的内容如下.

以 "随机地" 选取一个数 $\omega \in [0,1)$, 并且设 $\omega = (a_1, a_2, \cdots)$ 是 ω 的连续分数分解, 其中 $a_n = a_n(\omega)$ 是整数. (对于有理数 $\omega \in [0,1)$, 在此分解中仅有有限个 a_n 不为 0, 并且由 $(a_1, a_2, \cdots)$ 形成的数 $\omega^{(k)} = (a_1, a_2, \cdots, a_k, 0, 0, \cdots)$, 用于做 ω 的最优初始逼近.) 问在 "典型" 情形下, 当 n 的值充分大时量 $a_n(\omega)$ 的性质如何?

虽然并不严格, J. A. H. 盖尔登证明, 当 n 充分大时在分解 $\omega = (a_1, a_2, \cdots)$ 中, $a_n = k$ 的值在 "多少" 程度上与 k^2 呈反比. 略晚些, T. 布罗登 (T. Brodén, [12]) 和 A. 威曼 (A. Wiman, [62]) 证明, 利用几何概率, 如果把 "随机" 抽取 $\omega \in [0,1)$ 看作 ω 在 [0,1) 上 "均匀分布", 则当 $n \to \infty$ 时 $a_n(\omega) = k$ 的概率收敛于值

$$(\ln 2)^{-1} \times \ln\left[\left(1+\frac{1}{k}\right)\Big/\left(1+\frac{1}{k+1}\right)\right];$$

由此可见, 对于充分大 k 的, 该式与 k^2 呈反比, 而这实质上具有盖尔登的形式.

19 世纪的后半叶, 概率的概念和论断, 开始广泛地应用于经典物理和统计力学. 例如, 只需注意到, 分子运动速度的麦克斯韦分布 (James Clerk Maxwell, 1831 年 — 1879 年; 见 [44]); 布耳兹曼 (Ludwig Boltzmann; 1844 年 — 1906 年) 时间平均值和遍历假说 (见 [6], [7]).

总体的概念与他们的名字相联系, 后来总体的概念在吉普斯 (Josiah Willard Gibbs, 1839 年 — 1903 年) 的工作中得到进一步发展 (见 [17]).

对于概率论以后全部的发展, 以及关于对概率方法与概念作用的认识的深化, 如下一些学者都起了重要作用: 1827 年 R. 布朗 (Robert Brown, 1773 年 — 1858 年) 发现的被称为布朗运动的现象 (对该现象的描绘, 1828 年发表在抨击性文章《A Brief Account of Microscopical Observation ···》 [11] 中); 在研究铀的性质时, 1896 年 A. 贝克尔雷尔 ((Antoine-) Henri Becquerel, 1852 年 — 1908 年) 发现了放射性衰变现象; 1900 年 L. 巴彻里耶 (Louis bachelier, 1870 年 — 1946年, [2]) 利用布朗运动对股票的价格进行数学描绘 (详见 [74]).

A. 爱因斯坦 (Albert Einstein, 1879 年 — 1955 年, [75]) 和 M. 斯莫卢霍夫斯基 (Marian Smoluchowski, 1872 年 — 1917 年, [59]) 后来对布朗运动, 作了定性的解释和定量的描述. 放射性现象在量子力学的框架内得到了说明, 量子力学是 20 世纪的 20 年代创立的.

由以上所述知, 新概率概型和模型的出现, 以及概率的思想体系都超出了 "古典型概率" 的范围, 并且要求有新的概念, 以便对诸如 "来自区间 $[0,1)$ 的样本点" 的含义, 给予确切的数学意义, 更不用说对于 "随机" 布朗运动的解释了. 从这种观点看来, 非常适时的是, 产生了集合论, 和由 E. 博雷尔 (Emile Borel, 1871 年 — 1976 年; [8]) 于 1898 年引进的 "博雷尔测度" 的概念; 以及 H. 勒贝格 (Henri Lebesgue, 1875 年 — 1941 年) 积分理论, 包含在其 1904 年的书中 [40]; 由 E. 博雷尔作为长度概念的推广, 将测度引进欧几里得空间. 遵循 M. 弗雷歇 (Maurice Fréchet, 1878 年 — 1973 年; 1915 年 [71]) 现代讲述测度论, 是在抽象可测空间上 (关于测度论和积分的历史, 例如见 [72]).

实际上, 立即可以察觉, 博雷尔测度论和勒贝格积分理论概念的基础, 不但可以为许多研究奠定基础, 而且为诸如 "在区间 $[0,1)$ 是随机选点" 之类的许多直观提法赋予确切的意义. 而且, E. 博雷尔本人很快 (1905 年) 就将理论 - 集合方法应用于概率论, 证明了 —— 强大数定律 —— 关于实数的某些性质, "以概率 1" 成立或 "几乎必然" 成立.

这些定理给出一定的印象, "或多或少" 的实数, 具有 (在如下指出意义上) "特殊" 性质. 这些定理的实质如下.

假设实数 $\omega \in [0,1)$, 而 $\omega = 0.\alpha_1\alpha_2\cdots$ 是二进制分解, 其中 $\alpha_n = 0, 1$ (对照上面的 ω 分解为连续分数 $\omega = (\alpha_1, \alpha_2, \cdots)$). 那么, 若 $v_n(\omega)$ 是在 ω 的前 n 个值 $\alpha_1, \cdots, \alpha_n$ 中 "1" 的个数, 则满足条件: 当 $n \to \infty$ 时 $v_n(\omega) \to 1/2$ 的 ω (博雷尔称之为 "正规的") 集合的博雷尔测度等于 1, 而对于那些 ("特殊的") $\omega, v_n(\omega)$ 不收敛, 且相应集合的博雷尔测度等于 0.

这一结果 ("博雷尔强大数定律") 外观上像 J. 伯努利定理 ("大数定律"). 然而, 在二者之间, 既有形式上数学的不同, 又有哲学概念上的差异. 事实上, 在大数定律中仅仅断定: 对于任意 $\varepsilon > 0$, 当 $n \to \infty$ 时事件 $\{\omega : |v_n(\omega) - 1/2| \geqslant \varepsilon\}$ 的概率收敛于 0. 而在强大数定律中结果更多: 事件

$$\left\{\omega : \sup_{m \geqslant n} |v_m(\omega) - 1/2| \geqslant \varepsilon\right\}$$

的概率趋向 0. 此外, 在第一种情形下, 命题涉及有限序列 $(\alpha_1, \alpha_2, \cdots, \alpha_n), n \geqslant 1$ 概率的某些性质, 以及这些概率的极限. 而在第二种情形下, 命题涉及定义在无限序列 $(\alpha_1, \alpha_2, \cdots, \alpha_n, \cdots)$ 概率的性质, 以及这些概率的极限. (与概率论方法渗透到数论中, 以及构建现代概率论, 有关的数学和哲学问题的广泛资料详细阐述, 参见专著 "Creating Modern Probability", [54], 作者是 J. v. 普拉托 (Jan von Plato)).

4. 第四时期 (20 世纪初和 20 世纪中叶) 19 世纪末, 表现出概率论与纯数学的联系, 使得 D. 希尔伯特 (David Hilbert, 1862 年 — 1943 年) 提出概率论数学化的问题. 1900 年 8 月 8 日, 在巴黎召开的第二届数学学术会议上, D. 希尔伯特在他的提纲性报告中提出上述问题. 在他的著名课题中 (第一个是关于连续统 - 假说), 第六个是数学起决定性作用的物理学科的公理化问题. D. 希尔伯特把概率论和力学归为这样的学科, 他还指出在物理学, 其中包括气体动能理论, 严格的和尚未满意的发展平均值方法的必要性. D. 希尔伯特指出, 格丁根 (Göttingen) 大学①的 G. 布耳曼 (Georg BoHlmann, 1869 年 — 1928 年) 副教授, 由他提出了 "关于概率论的公理化问题的动议", 他 1900 年在巴黎召开的保险统计人员会议上作了关于公理化问题的发言 (见 [5], [19]). G. 布耳曼引进的概率定义为事件上的 (有限 - 可加) 函数, 然而没有 "事件系" 的充分清晰的定义, 其实他自己也承认这一点.

概率论形成历史的第四个时期 —— 是概率论的逻辑奠基和形成数学学科的时期.

在 D. 希尔伯特的报告之后, 很快出现了几个建立概率的数学理论的尝试, 其共同的特点是, 以集合论与测度论为基础.

例如, 1904 年 R. 拉默尔 (R. Lämmel, [41]; 亦见 [19]) 为描绘结局的集合, 他试图利用集合论, 不过, 概率的概念 (使用术语 "content", 且结合体积, 面积, 长度 ······)

①格丁根是德国城市, 人口十几万. 格丁根大学是一所综合性大学 1737 年创立, 也是该城唯一一所高校, 学生两万人左右. 格丁根科学院 1751 年成立, 有近百名院士和百余名通讯院士. —— 译者

本身仍然停留在以往时期的直观水平上.

另一个作者 U. 布罗吉 (Ugo Broggi, 1880 年 — 1965 年) 在其由 D. 希尔伯特指导的学位论文 (1907 年, [10]; 亦见 [19]) 中, 也运用博雷尔和勒贝格测度论 (基于它在 1904 年勒贝格的书 [40]) 中的概念, 但是 (有限 - 可加) 概率的概念本身 (在最简单的场合) 需要运用 "相对测度", "相对频率" 和 (在一般情形下) 某种人为的极限过程.

伯恩斯坦 在随后关于概率论的逻辑奠基工作作者中, 首先应当提到的是, C. H. 伯恩斯坦 (С. Н. Бернштейн, 1880 年 — 1968 年) 和 R. von 米泽斯 (Richard von Mises, 1883 年 — 1953 年).

C. H. 伯恩斯坦的公理化体系 ([4], 1917 年), 以事件按其大或小的似然程度的质量比较为基础, 而概率的数值本身表现为某个任意值.

费内提 后来, 在 B. de 费内提 (Bruno de Finetti, 1906 年 — 1985 年) 20 年代末 — 30 年代初的工作, 用基于与*主观*数量判断十分相似的方法 ("主观知识体系"), 得到广泛的发展 (例如, 见 [65]~[70]).

B. de 费内提的思想, 得到统计学中许多贝叶斯方向代表的很大支持, 例如, L. J. 塞维奇 (Leonard Jimmie Savage, 1917 年 — 1971 年; [60]). B. de 费内提的思想, 在有相当大的作用的对策论和判决理论中, 也得到同样的支持.

米泽斯 1919 年 R. 米泽斯提出了([49], [50]), 论证概率论的所谓*频率方法* (亦称*统计方法*或*经验方法*). 他的所谓*频率方法*以如下的思想为基础, 概率的概念只可能用于所谓 "集体", 即*个别无限有序数列*, 且这种有序数列具有其形成的某种 "随机" 性质.

R. 米泽斯的一般概形描绘如下.

考虑 "试验" 结局的某个样本空间, 并且假设可以进行无限多次试验, 得序列 $x = (x_1, x_2, \cdots)$, 其中 x_n 是第 n 次 "试验" 的结局. 其次, 设 A 是试验结局的集合的子集, 而

$$v_n(A; x) = \frac{1}{n}\sum_{i=1}^{n} I_A(x_i)$$

是 "事件" A 在前 n 次 "试验" 中出现的频率.

序列 $x = (x_1, x_2, \cdots)$ 称做*集体*, 如果它满足如下 (称为*米泽斯泽一性条件* (见 [49]~ [51]) 的) 两个假说.

I (对于*序列*频率的极限*存在*) 对于一切 "容许" 集合 A, 频率的极限存在:

$$\lim_n v_n(A; x)\ (= p(A; x));$$

II (对于*子序列*频率的极限*存在*) 对于一切由序列 $x = (x_1, x_2, \cdots)$, 利用某一事先约定的 ("容许") 其形成规则体系 (米泽斯称之为*位选择函数* [Place-selection functions]), 得到的子序列 $x' = (x_1', x_2', \cdots)$, 频率 $\lim_n v_n(A; x')$ 的极限与关于序列 $x = (x_1, x_2, \cdots)$ 本身的极限一样, 即等于 $\lim_n v_n(A; x)$.

按照米泽斯, 只有联系具体的 "集体" 才可以称 "集合 A 的概率", 而且 (根据假说 I) 这一概率 ($P(A;x)$) 定义为频率 $\lim_n v_n(A;x)$的极限. 需要强调, 假如该极限不存在 (即按照定义 x, 不是 "集体"), 那么相应的概率也就没有定义. 第二个假说, 米泽斯用来表示 (相应的直观, 恰好是一切概率研究的基础) 在形成 "集体" $x = (x_1, x_2, \cdots)$ 时 "随机性" 概念, 反映该序列 "非正则性" 思想, 以及对于任意 $n \geqslant 1$, 根据 "过去" $(x_1, x_2, \cdots, x_{n-1})$, 其 "将来" 值 $(x_n, x_{n+1}, \cdots)$ 的 "不可预测性". (赞同第二章 §1 介绍的, 柯尔莫戈洛夫公理化体系的, 概率论的代表人物, 这样的序列应当联想到, 对独立同分布随机变量观测结局的 "典型" 序列; 见第一章 §5 第 4 小节.)

米泽斯, 像他自己 ([51] 第 1 页) 说的那样, 在建立 "a mathematical theory of repetitive events" (重复事件的数学理论) 时提出的假说, (特别是在 30 年代) 引起了很大的争议和批评. 基本的不同意见如下: 实际中人们遇到的一般是*有限序列*, 而不是无限序列. 因而, 实际上无法确定极限 $\lim_n v_n(A;x)$ 是否存在; 在从序列 x 转移到序列 x' 时, 实际上也无法确定该极限《敏感性》. 米泽斯定义的子序列形成 "容许" 规则的概念, 也受到严重的批评, 使得在非此即彼的条件 II 中, 出现许多 ("模棱两可") 规则定义的模糊不清晰性.

如果考虑由 0 和 1 形成的序列 $x = (x_1, x_2, \cdots)$, 而且对于该序列, 极限 $\lim_n v_n(x; \{1\})$ 的值属于区间 (0,1), 则该序列既含无限个 0 又含无限个 1. 因此, 如果容许*任何*形成子序列的规则, 那么, 例如可以由 x 组成仅含 "1" 的子序列 x', 显然 $\lim_n v_n(x'; \{1\}) = 1$. 由此可见, 关于一切形成子序列的方法的非平凡 "集体" *不存在*.

在证明 "集体" 类 "非空" 的第一步, 是由 A. 瓦尔德 (Abraham Wald, 1902 年 — 1950 年) 1937 年在其论文 [13] 中实现的. 在他由序列 $x = (x_1, x_2, \cdots)$ 构造子序列 $x' = (x_1', x_2', \cdots)$ 时, 利用只取 0 和 1 两个可能值的可数个函数组 $f_i = f_i(x_1, \cdots, x_i), i \geqslant 1$: 若 $f_i(x_1, \cdots, x_i) = 1$, 则元素 x_{n+1} 属于子序列 x'; 若 $f_i(x_1, \cdots, x_i) = 0$, 则元素 $x_{n+1} \notin x'$. 1940 年 A. 乔尔奇 (Alonzo Church, 1903 年 — 1995 年) 提出形成子序列的另一种方法 (见 [73]), 方法基于如下想法: 这样的形成方法应该是实际上 "可有效计算的". 乔尔奇的这一思想导致, 他为序列的建立而提出的, 函数计算*算法*的概念. (例如, 设 x_i 有两个可能值: $\omega_1 = 0, \omega_2 = 1$. 将序列与一正数

$$\lambda = \sum_{k=1}^{n} i_k 2^{k-1}$$

相对应, 其中决定于 $i_k = \omega_{i_k}$. 如果 $\varphi = \varphi(\lambda)$ 是定义在集合 $\{0, 1, 2, \cdots\}$ 上的, 给定的二进制 $\{0, 1\}$ - 函数, 则当 $\varphi(\lambda_n) = 1$ 时 x_{n+1} 包含新序列 x' 中, 而当 $\varphi(\lambda_n) = 0$ 时 x_{n+1} 不包含 x' 中).

"集体" 作为具有 "随机性" 序列的说明和证据之一, 米泽斯引进了直观的论据: 对于这样的序列不能构造 "能获胜的对策体系".

在 J. 维尔 (Jean Ville, 1910 年 — 1988 年) 1939 年不大的专著 [14] 中, 对这样的论断进行了批评性分析, J. 维尔在 [14] 中给米泽斯的结果以严格的数学形式. 需

要指出, 也正是在此专著中, 首次 (作为数学概念) 使用术语 "鞅".

由以上引进的概率论公理化不同方法 (……, 伯恩斯坦, de 费内提, 米泽斯) 的描述可见, 在他们那里留有概念的复杂化与过分超负荷错误的印记, 其中多是出于建立尽可能接近实际应用的、概率概形的愿望. 正像 A. H. 柯尔莫戈洛夫在其专著《概率论的基本概念》[23] 中指出的, 这无法产生简单的公理化体系.

A. H. 柯尔莫戈洛夫, 提到他对于概率论逻辑基础的兴趣的, 发表的第一篇作品, 是 (非广为已知的) 论文 "测度与概率演算的一般理论" [27]. 论文的名称及其内容表明, A. H. 柯尔莫戈洛夫把集合论与测度论, 看作概率论逻辑基础的可能性所在. 由以上的叙述可见, 这种情况完全不是新的, 并且对于莫斯科数学学派是完全自然的. 因为对于莫斯科数学学派, 集合论及度量函数理论, 是数学研究的基本领域之一.

在这篇论文 (1929 年) 与《概率论的基本概念》([23], 1933 年) 之间, A. H. 柯尔莫戈洛夫发表了他著名的论文之一 "概率论的分析方法" [29]. 关于这篇论文, П.С. 亚历山大罗夫 (П. С. Александров) 和辛钦 [1] 写道: "在整个 20 世纪的整个概率论中, 很难指出对于未来科学的发展如此奠基性的研究成果 ……".

这篇论文的重大价值在于, 它不仅奠定了*马尔可夫随机过程*的理论基础, 而且它表明概率论整体上与数学分析 (特别是与常微分和偏微分方程论) 的联系, 以及与经典力学和经典物理学等的联系.

鉴于所考虑的概率的数学理论基础问题, 我们指出, 譬如说论文 "分析方法" [29], 可以看作逻辑上建立随机过程基础必要性的、"物理的" 动机综合体, 是 (除 "公理化" 外) "基本概念" 链中的一环.

A. H. 柯尔莫戈洛夫提出的 (非正式的) 概率论的公理化, 基础是*概率空间*

$$(\Omega, \mathscr{F}, \mathbf{P})$$

的概念, 其中 $(\Omega, \mathscr{F})$ 是 ("基本" 结局和 "事件" 的) 某一 (抽象) 可测空间, 而 $\mathbf{P}$ 是 $\mathscr{F}$ 中的非负可数 - 可加集函数, 且满足规范性条件 $\mathbf{P}(\Omega) = 1$ ("概率"); 见第二章 §1.

关于*随机变量* $\xi = \xi(\omega)$ 理解为 $\mathscr{F}$ - 可测函数 $\xi = \xi(\omega)$, 而按测度 $\mathbf{P}$ 定义的 $\xi(\omega)$ 的勒贝格积分是其*数学期望*.

关于 $\mathscr{G} \subseteq \mathscr{F}$ 的 σ- 子代数*条件数学期望* $\mathbf{E}(\xi|\mathscr{G})$ 是新概念 (见 A. H. 柯尔莫戈洛夫的《概率论的基本概念》(第二版) 的前言 [24]).

在《概率论的基本概念》中, 有一定理, A. H. 柯尔莫戈洛夫称之为*基本的*, 这本身就强调它所包含 (关于具有给定有限维分布的存在性的过程) 论断的重要性. 这里, 问题的实质如下.

在 A. H. 柯尔莫戈洛夫的论文 "概率论的分析方法" [29] 中, 马尔可夫过程用于 "随机确定系统" 的发展变化, 这用满足 "柯尔莫戈洛夫 - 查普曼方程" 的、函数 $P(s, x; t, A)$ 性质的语言描述. 函数 $P(s, x; t, A)$ 称做转移概率函数, 因为它表示 "在时间 s '系统' 处于状态 x, 而在时间 t '系统' 处于其状态相空间的集合 A 中".

同样, 同一时间, 在 [30], [31], [66], [67] 关于 "齐次独立增量随机过程" 的工作中, 所有的研究都用满足函数方程

$$P_{s+t}(x) = \int P_s(x - y) \mathrm{d}P_t(y)$$

的函数 $P_t(x)$ 性质的语言进行的, 在说明 $P_t(x)$ 是 "经时间 t 过程的增量不大于 x 的概率" 时, 上面的方程是自然产生的.

不过, 从形式逻辑的观点看, 可以称为 "具有给定转移概率 $P(s, x; t, A)$ 或给定分布 $P_t(x)$ 的 '过程' " 的对象之*存在性*问题, 仍然没有解决.

就是在求解该问题时, 涉及如下*基本定理*: 对于每个*一致*有限维概率分布组

$$F_{t_1, t_2, \cdots, t_n}(x_1, x_2, \cdots, x_n), \quad 0 \leqslant t_1 < t_2 < \cdots < t_n, x_i \in \mathbb{R},$$

可以建立一个概率空间 $(\Omega, \mathscr{F}, \mathbf{P})$ 和随机变量组 $X = (X_t)_{t \geqslant 0}, X_t = X_t(\omega)$, 使得

$$\mathbf{P}\{X_{t_1} \leqslant x_1, X_{t_2} \leqslant x_2, \cdots, X_{t_n} \leqslant x_n\} = F_{t_1, t_2, \cdots, t_n}(x_1, x_2, \cdots, x_n).$$

作为 Ω 取实函数 $\omega = (\omega_t)_{t \geqslant 0}$ 的空间 $\mathbb{R}^{[0,\infty)}$, 作为 $\mathscr{F}$ 取柱集生成的最小 σ - 代数, 而测度 $\mathbf{P}$ 是由柱集代数上的测度 (在柱集代数上, 该测度是自然地根据有限维概率分布建立的), 开拓到 $\mathscr{F}$ 上概率测度. 随机变量组 $X_t(\omega)$ 是用坐标方式建立的: 若 $\omega = (\omega_t)_{t \geqslant 0}$, 则 $X_t(\omega) = \omega_t$. (这一构造说明了, 为何 "随机过程" 的概念, 常等同于 (它) 在函数空间 $\mathbb{R}^{[0,\infty)}$ 中的*测度*).

在《概率论的基本概念》中, 以不大的篇幅讲概率论的*适用性*问题.

在描绘将概率论用于 "现实试验世界" 的*条件模式*时, A. H. 柯尔莫戈洛夫许多方面是按米泽斯做的, 而且表明在概率论的解释与应用性问题上它与米泽斯方法并不是格格不入的.

关于这一*条件模式*的实质可以描述如下.

假设有某一条件的*综合体*, 使得可以进行无限次重复试验.

设 $(x_1, x_2, \cdots, x_n)$ 表示 n 次试验的结果, 其中例如可以假设 $x_i (1 \leqslant i \leqslant n)$ 属于集合 X. 此外, 设 A 是我们感兴趣的 X 的某一子集.

如果 $x_i \in A$, 则称第 i 次试验出现了事件 A. (注意, *事先*不作诸如: 试验是 "随机和独立地进行的" 等, 关于试验的 "概率" 性质之类的任何假设; 对于导致事件 A 的 "情形" 也不作任何假设, …….)

其次, 假设可以赋予事件 A 某个数 (记作 $\mathbf{P}(A)$), 使得*实际上可以认为*, 在 n 次试验中事件 A 出现的频率 $v_n(A)$, 对于充分大的 n, 与 $\mathbf{P}(A)$ 的差异很小. 而如果 $\mathbf{P}(A)$ 较小, 那么*实际上可以认为*, 在一次试验中事件 A 不出现.

在《概率论的基本概念》中, A. H. 柯尔莫戈洛夫, 在未讨论概率论对于《现实世界》可应用性条件细节的情况下, 写道 "我们 …… 有意识地把经验世界中, 关于

概率概念的高深哲学著作放在一边", 但是在第一章引言中指出, 存在概率论的应用领域, 其中 "与 '偶然性' 和 '概率' 概念等词的本意毫无关系" (见 [24]).

三十年后, A. H. 柯尔莫戈洛夫又回到概率的可应用性问题 (见 [32]~[37]), 对于该问题的解, 他提出两种 ("第一种" 和 "第二种") 处理方法, 相应为 "公理化随机性" 的概念和 "算法的复杂性" 的概念. 他这时特别强调 [37], 与定义了*无限序列* $(x_1, x_2, \cdots)$ 的 P. 米泽斯和 A. 乔尔奇不同, 他对于 "随机性" 概念的处理方法, 带有*严格有限性*, 即涉及*有限长度序列* $(x_1, x_2, \cdots, x_N), N \geqslant 1$ (下面 [38] 的情形称为*链*), 这实际上是现实中的情形.

"公理化随机性" 是这样引进的.

设 $(x_1, x_2, \cdots, x_N)$ 是以长为 $N(n \leqslant N)$ 的二进制 $x_i = 0, 1$ 链. 称此链关于 (有限) *容许算法*的全体 Φ 是 (n, ε) - *随机的*, 如果存在这样一个数 $p (= \mathbf{P}(\{1\}))$, 使对于利用某个算法 $A \in \Phi$, 由 $(x_1, x_2, \cdots, x_N)$ 得到的任何链 $(x'_1, x'_2, \cdots, x'_m), n \leqslant m \leqslant N$, 出现 "1" 的频率与 $v_m(x'; \{1\})$ 离 p 的差异不大于 ε. (算法的全体 Φ 中导致长为 $m < n$ 的链, 不予考虑).

A. H. 柯尔莫戈洛夫在 [32] 中证明, 假如对于给定的 n 和 $0 < \varepsilon < 1$, 容许算法的个数不大于

$$\frac{1}{2} \exp\{2n\varepsilon^2(1 - \varepsilon)\},$$

则对于每一个 $0 < p < 1$和任意 $N \geqslant n$, 存在具有 (n, ε) - 随机性 ("公理化随机性") 的链 $(x_1, x_2, \cdots, x_N)$.

由于容许描绘和选择算法的不确定性, 在上面描绘的分出 "随机" 链 (像米泽斯情形一样) 的方法, 有一定任意性. 这时这一算法类明显不会太大, 否则 "公理化随机" 链的集合成为空集. 与此同时, 总是希望容许算法类构造简单 (例如可以用表格表示).

在概率论中形成一种完全确定、加强不同类型的、关于 "典型随机实现相当不正规的, 十分复杂的" 概率论点的观念.

因此, 如果倾向于, 使链及序列 "随机性" 的*算法*定义, 最大限度地接近关于随机实现构造的*概率*概念, 则 Φ 中的算法 (的全体) 容许挑出 "非典型的, 容易剔除的" 链, 把那些充分非正则的, 十分复杂的定为 "随机的".

这种理由导致 A. H. 柯尔莫戈洛夫的 "第二种" 处理方法, 导致 "随机性" 概念, 其中着重点不是所考虑的算法 "简单", 而是链本身的 "复杂性" 和 "径直地" 引进 "复杂性的" 某个数字特征, 使之表示形成该链时 "非正则性" 程度.

这一数字特征就是单个链 x 关于算法 A 的所谓 "*算法的*" (或 "*柯尔莫戈洛夫的*") *复杂性* $K_A(x)$; 形象地说, $K_A(x)$ 定义为满足如下条件的二进制链的长度: 在算法 (机器, 计算机 ……) 的 "入口" 输入 A, 容许同一链在 "出口" 再现.

正式的定义如下.

设 $\sum$ 是一切有限二进制链 $(x_1, x_2, \cdots, x_n)$ 的全体, 以 $|x| (= n)$ 表示链的长度,

而 Φ 是某一算法类. 实数

$$K_A(x) = \min\{|\rho| : A(\rho) = x\},$$

称做链 $x \in \sum$ 关于算法 $A \in \Phi$ 的*复杂性*. 数 $K_A(x)$ 是在算法 A 的 "入口" 的二进制链 ρ 的最小长度 ($|\rho|$), 并且在链的 "出口" $x (A(\rho) = x)$ 将得到恢复.

A. H. 柯尔莫戈洛夫在 [34] 中证明, (对于某一重要算法类 Φ) 有如下结果: 存在这样的*通用*算法 $U \in \Phi$, 对于任意 $A \in \Phi$ 存在常数 $C(A)$, 使对于任意链 $x \in \sum$, 有

$$K_U(x) \leqslant K_A(x) + C(A),$$

而对于通用算法 U' 和 U'',

$$|K_{U'}(x) - K_{U''}(x)| \leqslant C, \quad x \in \sum,$$

其中 C 不依赖于 $x \in \sum$. (A. H. 柯尔莫戈洛夫在 [34] 中指出, P. 所罗门诺夫 [P. Соломонов] 同时证明了类似的结果).

这一事实 (以及对于 "典型" 链 $x, K_U(x)$ 的 $|x|$ 值随增长而增长) 决定如下定义: 量 $K(x) = K_U(x)$ 称做*链* $x \in \sum$ *关于算法类* Φ *的复杂性*, 其中 U 是 Φ 中的某一通用算法.

量 $K(x)$ 通常称做 "对象" x *算法复杂性*或*柯尔莫戈洛夫复杂性*. A. H. 柯尔莫戈洛夫把该量看作包含在 "有限对象" x 中的*算法信息质量的度量*, 并且称之为 x 的*熵*. 并且认为这个概念是比*概率信息量*的概念更基本, 而为概率信息量的要求知道在 "对象" x 上的概率分布.

量 $K(x)$ 还可以视为 "文本" x 的压缩程度的指标. 如果在类 Φ 中包含元素的简单计数算法, 则很明显 (精确到常数) 链 x 的 "复杂性" $K(x)$ 不大于其长度 $|x|$. 另一方面, 由简单的讨论可见, "复杂性" 小于 K 的 (二进制) 链 x 的数量不大于 $2^K - 1$, 等于小于长度 K 的不同 "输入" 二进制序列的个数:

$$1 + 2 + \cdots + 2^{K-1} = 2^K - 1.$$

其次, 通过简单的论证 (例如, 见 [15]) 可以证明, *存在这样的链* x: 其 "复杂性" (精确到常数) 等于 (且不可能大于)长度 $|x|$, 并且容许强压缩 ("复杂性" 为 $n - a$ 的链的比率不大于 2^{-a}). 由全部所作的这些论述, 自然地导致如下的定义: 算法的 "复杂性" $K(x)$ 接近 $|x|$ 的链 x 称为 (关于算法类 Φ) "算法上随机的".

换句话说, 算法的处理方法①判定这样一些链 x 为 "随机的", 假如其 "复杂性" 是最大的 ($K(x) \sim |x|$).

A. H. 柯尔莫戈洛夫引进的 "复杂性" 的概念, 产生了整个算法随机性的方向, 称为 "柯尔莫戈洛夫复杂性", 在最广泛的数学及其应用领域得到众多的应用 (例如, 详见 [38], [45]~[48], [22]).

①即前面提到的, A. H. 柯尔莫戈洛夫 "第二种" 处理方法, 对应于 "算法的复杂性". —— 译者

在概率论中, 这些新概念是一系列工作的开端: 说明对于何种链和序列的 "算法随机性", 概率—统计规律性 (诸如, 强大数定律, 重对数定律 ……) 成立 (例如, 见 [16]), 从而有可能运用概率论的方法及其结果, 而在前面已经指出的那些领域 (见 [24]), "与 '偶然性' 和 '概率' 概念等词的本意毫无关系".

"概率的数学理论形成的简史" 的参考文献

[1] Александров П. С., Хинчин А. Я. Андрей Николаевич Колмогоров (к пятидесятилетию со дня рождения) // Успехи математических наук. — 1953. — Т. 8. №3. — С. 177 — 200.

[2] Башелье (Bachelier L.). Théorie de la spéculation // Annales de l'École Normale Supérieure. — 1900. — V. 17. — P. 21 — 86.

[3] Бернулли Я. О законе больших чисел. Ч.4: Искусство предположений. — М.: Наука, 1986. — С. 23—59.

[4] Бернштейн С. Н. Опыт аксиоматического обоснования теории вероятностей // Сообщения Харьковского математического общества. Сер. 2. — 1917. — Т. 15. — С. 209 — 274.

[5] Больман (Bohlmann G.). Lebensversicherungsmathematik // Encyklopaedieder mathematischen Wissenschaften. — Bd. I, Heft 2. — Artikel ID4b. — Leipzig: Teubner, 1903.

[6] Больцман (Boltzmann L.). Wissenschaftliche Abhandlungen. — V.1 — 3. — Leipzig: Barth, 1909.

[7] Больцман, Набл (Boltzmann L., Nabl J.). Kinetische Theorie der Materie // Encyklopaedie der mathematischen Wissenschaften. — Bd. V, Heft 4. — Leipzig: Teubner, 1907. — S. 493 — 557.

[8] Борель (Borel É.). Leçons sur la théorie des fonctions. — Paris: Gauthier-Villars, 1898; Éd. 2. — Paris: Gauthier-Villars, 1914.

[9] Борель (Borel É.). Quelques remarques sur les principes de la theorie desensembles // Mathematische Annalen. — 1905. — V. 60. — P. 194 — 195.

[10] Брогги(Broggi U.). Die Axiome der Wahrscheinlichkeitsrechnung. Dissertation. Göttingen, 1907. (См. также [19].)

[11] Браун (Brown R.). A brief account of microscopical observations made in the months of June, July, and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. — Philosophical Magazine N. S. — 1828. V. 4. — P. 161 — 173.

[12] Броден (Brodén T.). Wahrscheinlichkeitsbestimmungen bei der gewöhnlichen Kettenbruchentwicklung reeller Zahlen // Akad. Förh. Stockholm. — 1900. — V. 57. — P. 239 — 266.

[13] Вальд(Wald A.). Die Widerspruchsfreiheit des Kollektivbegriffes der Wahrscheinlichkeitsrechnung // Ergebnisse eines mathematischen Kolloquiums. —1937. — V. 8. — P. 38 — 72.

[14] Виль (Ville J. A.). Étude critique de la notion de collectif. — Paris: Gauthier-Villars, 1939.

[15] Витаньи П., Ли М. Колмогоровская сложность: двадцать лет спустя // Успехи математических наук. — 1988. — Т. 43, № 6. — С. 129 — 166.

[16] Вовк В. Г. Закон повторного логарифма для случайных по Колмогорову, или хаотических, последовательностей // Теория вероятностей и ее применения. — 1987. — Т. 32, № 3. — С. 456 — 468.

[17] Гиббс (Gibbs J. W.). Elementary Principles in Statistical Mechanics. Developed with especial reference to the rational foundation of thermodynamics. — New Haven: Yale Univ. Press, 1902; New York: Dover, 1960.

[18] Гюлден (Gyldén H.). Quelques remarques relativement à la représentation de nombres irrationnels au moyen des fractions continues // Comptes Rendus. Paris. — 1888. — V. 107. — P. 1584 — 1587.

[19] Die Entwicklung der Wahrscheinlichkeitstheorie von den Anfängen bis 1933 / Ed. I. Schneider. — Berlin: Akademie-Verlag, 1989.

[20] Дэвид (David F. N.). Games, Gods and Gambling. The Origin and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era. — London: Griffin, 1962.

[21] Звонкин А. К., Левин Л. А. Сложность конечных объектов и обоснование понятий информации и случайности с помощью теории алгоритмов // Успехи математических наук. — 1970. — Т. 25, № 6. — С. 85 — 127.

[22] Кирхгер, Ли, Витаньи (Kircherr W., Li M., Vitányi P.). The miraculous universal distribution // Mathematical Intelligencer. — 1997. — V. 19, № 4. — P. 7 — 15.

[23] Колмогоров (Kolmogoroff A.). Grundbegriffe der Wahrscheinlichkeitsrechnung. — Berlin: Springer, 1933; Berlin — New York: Springer, 1973.

[24] Колмогоров А. Н. Основные понятия теории вероятностей. — М. — Л.: ОНТИ, 1936; 2-е изд. — М.: Наука, 1974; 3-е изд. — М.: ФАЗИС, 1998. 汉译本：《概率论的基本概念》(丁寿田译), 商务印书馆, 1952.

[25] Колмогоров (Kolmogorov A. N.). Foundations of the Theory of Probability. — New York: Chelsea, 1950; 2nd ed. — New York: Chelsea, 1956.

[26] Колмогоров А. Н. Роль русской науки в развитии теории вероятностей // Роль русской науки в развитии мировой науки и культуры. — Т. I, кн. 1. — М.: Изд-во Моск. ун-та, 1947. — С. 53 — 64.

[27] Колмогоров А. Н. Общая теория меры и исчисление вероятностей // Коммунистическая академия. Секция естественных и точных наук. Сборник работ математического раздела. — Т. 1. — М., 1929. — С. 8 — 21. (См. также [28], с. 48 — 58.)

[28] Колмогоров А. Н. Теория вероятностей и математическая статистика. М.: Наука, 1986.

[29] Колмогоров (Kolmogoroff A.). Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung // Mathematische Annalen. — 1931. — V. 104. — P. 415 — 458. (См. также [28], с. 60 — 105.)

[30] Колмогоров (Kolmogoroff A.). Sulla forma generale di un processo stochastico omogeneo. (Un problema di Bruno de Finetti.) // Atti della Accademia Nazionale dei Lincei. — 1932. — V. 15. — P. 805 — 808. (См. также [28].)

[31] Колмогоров (Kolmogoroff A.). Ancora sulla forma generale di un processo omogeneo // Atti della Accademia Nazionale dei Lincei. — 1932. — V. 15. — P. 866 — 869. (См. также [28].)

[32] Колмогоров (Kolmogorov A. N.). On tables of random numbers // Sankhyā A. — 1963 — V.25, № 4. — P. 369 — 376.

[33] Колмогоров А. Н. Теория информации и теория алгоритмов. М.: Наука, 1987.

[34] Колмогоров А. Н. Три подхода к определению понятия 《количество информации》 // Проблемы передачи информации. — 1965. — Т. 1, №1. — С. 3 — 11. (См. также [33], с. 213 — 223.)

[35] Колмогоров (Kolmogorov A. N.). Logical basis for information theory and probability theory // IEEE Transactions on Information Theory. — 1968. — V. 14, № 5. — P. 662 — 664. (См. также [33], с. 232 — 237.)

[36] Колмогоров А. Н. Комбинаторные основания теории информации и исчисления вероятностей // Успехи математических наук. — 1983. — Т. 38, № 4. — С. 27 — 36.

[37] Колмогоров(Kolmogorov A. N.). On logical foundations of probability theory // Probability Theory and Mathematical Statistics (Tbilisi, 1982). Berlin etc.: Springer-Verlag, 1983. — P. 1 — 5. — (Lecture Notes in Mathematics; V. 1021) (См. также [28], с. 467 — 471.)

[38] Колмогоров А. Н., Успенский В. А. Алгоритмы и случайность // Теория вероятностей и ее применения. — 1987 — Т. 32, № 3. — С. 425 — 455.

[39] Лаплас (Laplace P. S., de). A Philosophical Essay on Probabilities. — New York: Dover, 1951; — Первое издание: La Place P. S. Essai philosophique surles probabilités. — Paris, 1814.

[40] Лебег (Lebesgue H.). Leçons sur l'intégration et la recherche des fonctions primitives. — Paris: Gauthier-Villars, 1904.

[41] Леммель (Lämmel R.). Untersuchungen über die ermittlung der Wahrscheinlichkeiten. Dissertation. Zürich, 1904. (См. также [19].)

[42] Ли, Витаньи (Li M., Vitányi P. M. B.). An Introduction to Kolmogorov Complexity and its Applications. — 2nd ed. — Berlin — New York: Springer-Verlag, 1997.

[43] Майстров Л. Е. Теория вероятностей. Исторический очерк. — М.: Наука, 1967.

[44] Максвелл (Maxwell J. C.). The scientific letters and papers of James Clerk Maxwell. — V. I: 1846 — 1862. — V. II: 1862 —1873. — V. III: 1874 — 1879 / Ed. P. M. Harman. — Cambridge: Cambridge Univ. Press, 1990, 1995, 2002.

[45] Мартин-Лёф П. О понятии случайной последовательности // Теорияв-ероятностей и ее применения. — 1966. — Т. 11. № 1. — С. 198 — 200.

[46] Мартин-Лёф (Martin-Löf P.). The definition of random sequences // Information and Control. — 1966. — V. 9, № 6. — P. 602 — 619.

[47] Мартин-Лёф (Martin-Löf P.). On the notion of randomness // Intuitionism and Proof Theory: Proceedings of the conference at Buffalo, NY, 1968 / Ed. A. Kino et al. Amsterdam: North-Holland, 1970. — P.73 — 78.

[48] Мартин-Лёф (Martin-Löf P.). Complexity oscillations in infinite binary sequences // Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete. — 1971. — V. 19. — P. 225 — 230.

[49] фон Мизес (Mises R., von). Fundamentalsätze der Wahrscheinlichkeitsrechnung // Mathematische Zeitschrift. — 1919. — V. 4. — P. 1 — 97.

[50] фон Мизес (Mises R., von). Grundlagen der Wahrscheinlichkeitsrechnung // Mathematische Zeitschrift. — 1919. — V. 5. — P. 52 — 99; 1920. — V. 7. — P. 323.

[51] фон Мизес (Mises R., von). Mathematical Theory of Probability and Statistics. — New York — London: Academic Press, 1964.

[52] Ньютон (Newton I.). The Mathematical Works of Isaac Newton / Ed. D. T. Whiteside. — V. 1. — New York: Johnson, 1967.

[53] О теории вероятностей и математической статистике (переписка А. А. Маркова и А. А. Чупрова). — М.: Наука, 1977.

[54] Плато (Plato J., von). Creating Modern Probability. Its Mathematics, Physics and Philosophy in Historical Perspective. — Cambridge: Cambridge Univ. Press, 1994.

[55] Пойя (Pólya G.). Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem // Mathematische Zeitschrift. — 1920. — V. 8. — P. 171 — 181.

[56] Пуанкаре (Poincaré H.). Calcul des probabilités. — Paris: G. Carré, 1896.

[57] Пуанкаре (Poincaré H.). Sur le problème des trois corps et les équations de la dynamique. I, II // Acta Mathematica. — 1890. — V. 13. — P. 1 — 270.

[58] Сельванатан и др. (Selvanathan A., Selvanathan S., Keller G., Warrack B., Bartel H.). Australian Business Statistics. — Melbourne: Nelson, An International Thomson Publ. Co., 1994.

[59] Смолуховский (Smoluchowski M. R., von). Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen // Annalen der Physik. — 1906. — V. 21. — P. 756 — 780.

[60] Сэвидж (Savage L. J.). The Foundations of Statistics. — New York: Wiley; London :Chapman & Hall, 1954.

[61] Тодхантер (Todhunter I.). A History of the Mathematical Theory of Probability from the Time of Pascal to That of Laplace. — New York: Chelsea, 1949; — Первое издание: Cambridge: Macmillan, 1865.

[62] Уиман (Wiman A.). Über eine Wahrscheinlichkeitsaufgabe bei Kettenbruchentwicklungen // Akad. Förh. Stockholm. — 1900. — V. 57. — P. 829 — 841.

[63] Успенский, Семёнов (Uspensky V. A., Semenov A. L.). What are the gains of the theory of algorithms: basic developments connected with the concept of algorithm and with its application in mathematics // Algorithms in Modern Mathematics and Computer Science (Urgench, 1979). — Berlin etc.: Springer-Verlag, 1981. — P. 100 — 234. — (Lecture Notes in Computer Science; V. 122.)

[64] Файн (Fine T. L.). Theories of Probability. An Examination of Foundations. — New York — London: Academic Press, 1973.

[65] де Финетти (Finetti B., de). Sulle probabilità numerabili e geometriche // Istituto Lombardo. Accademia di Scienze e Lettere. Rendiconti (2). — 1928. — V. 61. — P. 817 — 824.

[66] де Финетти (Finetti B., de). Sulle funzioni a incremento aleatorio // Accademia Nazionale dei Lincei. Rendiconti (6). — 1929. — V. 10. — P. 163 — 168.

[67] де Финетти (Finetti B., de). Integrazione delle funzioni a incremento aleatorio // Accademia Nazionale dei Lincei. Rendiconti (6). — 1929. — V. 10. — P. 548 — 553.

[68] де Финетти (Finetti B., de). Probabilismo: saggio critico sulla teoria delle probabilità e sul valore della scienza. — Napoli: Perrella, 1931; // Logos. — 1931. — V. 14. — P. 163 — 219. — English transl.: // Erkenntnis. The International Journal of Analytic Philosophy. — 1989. — V. 31. — P. 169 — 223.

[69] де Финетти (Finetti B., de). Probability, Induction and Statistics. The Art of Guessing. — New York etc.: Wiley, 1972.

[70] де Финетти (Finetti B., de). Teoria delle probabilità: sintesi introduttiva con appendice critica. — V. 1, 2. — Turin: Einaudi, 1970. — English transl.: Theory of Probability: A Critical Introductory Treatment. — V. 1, 2. — New York etc.: Wiley, 1974, 1975.

[71] Фреше (Fréchet M.). Sur l'intégrale d'une fonctionnelle étendue à un ensemble abstrait // Bulletin de la Société Mathématique de France. — 1915. — V. 43. — P. 248 — 265.

[72] Хокинс (Hawkins T.). Lebesgue's Theory of Integration. Its Origin and Development. — Madison, Wis. — London: Univ. Wisconsin Press, 1970.

[73] Чёрч (Church A.) On the concept of a random sequence // American Mathematical Society. Bulletin. — 1940. — V. 46, № 2. — P. 130 — 135.

[74] Ширяев А. Н. Основы стохастической финансовой математики: В 2-х т. — М.: ФАЗИС, 1998.

[75] Эйнштейн (Einstein A.). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen // Annalen der Physik. — 1905. — V. 17. — P. 549 — 560.

图书文献资料

(第四章 ~ 第八章)

第四章

§1. A. H. 柯尔莫戈洛夫 "0 – 1" 律在他的书 [32] 中. 关于 E. 休伊特 (E. Hewitt) 和塞维奇 "0 – 1" 律, 亦见 A. A. 博罗夫科夫 (A. A. Боровков) [7], L. 布赖曼 (Л. Брейман) [8], R. B. 阿什 (R. B. Ash) [81].

§2~§4. 这里是 A. H. 柯尔莫戈洛夫和 A. Я. 辛钦得到的基本结果 (见 [32] 以及其中的文献). 亦见 B. B. 彼得罗夫 (W. W. Petrov) [53] 和, W. F. 斯托特 (W. F. Stout) [66]. 关于数论中的概率方法, 见 Й. 库比柳斯 (Й. Кубилюс)[①][36].

我们现在对于伯努利概形, 回忆 "强大数定律和重对数定律" 的历史.

关于强大数定律最早的工作, 出现在 E. 博雷尔如下论文中: "关于集合 $[0,1)$ 中数的正态性" (É. Borel. Les probabilittés dénombrables et leurs applications arithmétiqes // Rendiconti del Circolo Matematico di Palermo. — 1909. — V.27 — P.247~271). 如果使用 §3 例 2 的记号, 则对于量

$$S_n = \sum_{k=1}^{n} \left(I(\xi_k = 1) - \frac{1}{2} \right),$$

E. 博雷尔所得结果是: (关于勒贝格测度) 对于几乎一切 $\omega \in [0,1)$, 存在 $N = N(\omega)$, 当 $n \geqslant N(\omega)$ 时, 有

$$\left| \frac{S_n(\omega)}{n} \right| \leqslant \frac{\ln(n/2)}{\sqrt{2n}}.$$

①立陶宛数学家 (1921 年 ~), 立陶宛科学院院士, 立陶宛维尔纽斯大学校长. —— 译者

这样, 特别, 几乎必然 (处处) $S_n = o(n)$.

下一步由 F. 豪斯多夫 (F. Hausdorff, Grundzüge der Mengenlehre, — Leipzig: Veit, 1914), 证明了, 对于任意 $\varepsilon > 0$, 几乎必然 $S_n = o(n^{1/2+\varepsilon})$.

在 1914 年, G. H. 哈代 (G. H. Hardy) 和 J. E. 李特尔伍德 (J. E. Littlewood, Some Problems of Diophantine approximation//Acta Mathematica, — 1914. —, V.37, — P.155~239) 证明, 几乎必然 $S_n = O((n \ln n)^{1/2})$.

在 1922 年, H. D. 斯坦因豪斯 (H. D. Steinhaus, Les probabilités dénombrables et leur rappot à la théorie de la mesure//Fundamenta Mathematicae, — 1923. — V4, — P.286~310) 将 G. H. 哈代和 J. E. 李特尔伍德的结果精确化, 证明几乎必然, 有

$$\limsup_n \frac{S_n}{\sqrt{2n \ln n}} \leqslant 1.$$

1923 年 A. Я. 辛钦 (A. J. Khinchin, Über dyadische Brüche//Mathematische Zeitschrift — 1923. — V18, P.109~116) 断定, 几乎必然 $S_n = O(\sqrt{n \ln \ln n})$.

最后, 1924 年 A. Я. 辛钦 (A. J. Khinchin, Über einen Satzder Wahrscheinlichkeitsrechnung//Fundamenta Mathematicae, — 1924. — V.6. — P.9~20 到最终的结果 ("重对数定律"): 几乎必然, 有

$$\limsup_n \frac{S_n}{\sqrt{(n/2) \ln \ln n}} = 1.$$

(需要指出, 对于所考虑的情形,

$$\sigma^2 = \mathbf{E}\left[I(\xi_k = 1) - \frac{1}{2} \right]^2 = \frac{1}{4},$$

因为出现了因子 $n/2$, 而不是因子 $2n$; 对照 §4 定理 1 的提法).

如同 §4 提到的, 下一步就是对于广泛的独立随机变量类, 证明重对数定律, 这是由 A. H. 柯尔莫戈洛夫于 1922 年实现的 (A. N. Kolmogoroff, Über das Gesets des iterierten Logarithmus//Mathematische Annalen, — 1929. — V.101, — P.126~135).

§5. 关于这些问题, 见 B. B. 彼得罗夫 [92], A. A. 博罗夫科夫 [7], D. 达昆纳-卡斯特里 (D. Ducunna-Castelle) 和 M. 杜弗劳 (M. Duflo) [86].

第五章

§1~§3. 在叙述 (强) 平稳随机序列时, 借鉴了书籍: L. 布赖曼 [8], Я. Г. 希奈 (Я. Г. Синаи) [63], J. 兰珀蒂 (J. Lamperti) [38]. A. M. 加尔西亚 (A. M. Garsia) [12] 给出了最大遍历性定理的简单证明.

第六章

§1. 关于 (广义) 平稳随机序列理论见下列图书: Ю. А. 罗扎诺夫 (Ю. А. Розанов) [60], И. И. 基赫曼 (И. И. Гихман) 和 А. В. 斯科罗霍德 (А. В. Скороход) [13], [14]. А. Н. 柯尔莫戈洛夫常在讲课时举例 6.

§2. 关于正交随机测度和随机积分, 亦见 J. L. 杜布 (J. L. Doob) [20], И. И. 基赫曼和 А. В. 斯科罗霍德 [14], Ю. А. 罗扎诺夫 [60], R. B. 阿什和 M. F. 加德纳 M. F. Cardner [82].

§3. 谱表现 (2) 是 H. 克拉默 (H. Cramer) 和 M. 洛埃甫(M. Loéve) 得到的 (例如, 见 [42]). 在 А. Н. 柯尔莫戈洛夫的著作 [29] 中, (在其他术语下) 含这样的表现. 亦见如下著作: J. L. 杜布 [20], Ю. А. 罗扎诺夫 [60], R. B. 阿什和 M. F. 加德纳 M. F. Cardner [82].

§4. 在 E. J. 汉南 (E. J. Hannan) 的书 [71] 和 [72] 中, 有协方差函数和谱密度的统计估计问题的详细叙述.

§5~§6. 亦见下列图书: Ю. А. 罗扎诺夫 [60], J. 兰珀蒂 [38], И. И. 基赫曼和 А. В. 斯科罗霍德 [13], [14].

§7. 这部分内容, 是按 Р. Ш. 利普彩尔 (Р. Ш. Липцер) 和 А. Н. 施利亚耶夫的书 [41] 叙述的.

第七章

§1. 鞅论的多数基本结果是 J. L. 杜布得到的 [20]. 定理 1 包含在 P. -A. 麦耶 (P. -A. Meyer) 的专著 [47] 中. 亦见, P. -A. 麦耶的书 [48], Р. Ш. 利普彩尔和 А. Н. 施利亚耶夫的书 [41], И. И. 基赫曼和 А. В. 斯科罗霍德 [14], J. 扎克德 (J. Jacod) 和 А. Н. 施利亚耶夫[87].

§2. 定理 1 常称为 "关于自由选择的变换" 定理, [20]. 关于恒等式 (13) 和 (14), 以及 A. 瓦尔德基本恒等式, 见书 [9].

§3. 第一个不等式 (25), 是 А. Я. 辛钦于 1923 年在其论文 (A. J. Khintchine, Über dyadische Brüche//Mathematische Zeischrift — 1923. — V18, P.109~116) 中, 在证明 "重对数定律" 时得到的. 为说明是什么使 А. Я. 辛钦必然会得到该不等式, 我们提醒注意, E. 博雷尔和 F. 豪斯多夫证明强大数定律的概形 (亦见上面对第四章 §2~§4 的评述).

设 $\xi_1, \xi_2, \cdots$ 是独立同分布随机变量列, 且 $\mathbf{P}\{\xi_1 = 1\} = \mathbf{P}\{\xi_2 = -1\} = 1/2$ (伯努利概形), $S_n = \xi_1 + \cdots + \xi_n$.

E. 博雷尔的证明 "几乎必然 $S_n = o(n)$" 的实质如下: 由于对于任意 $\delta > 0$, 有

$$\mathbf{P}\left\{\left|\frac{S_n}{n}\right| \geqslant \delta\right\} \leqslant \frac{\mathbf{E}S_n^4}{n^4\delta^4} \leqslant \frac{3n^2}{n^4\delta^4} = \frac{3}{n^2\delta^4},$$

则当 $n \to \infty$ 时

$$\mathbf{P}\left\{\sup_{k\geqslant n}\left|\frac{S_k}{k}\right| \geqslant \delta\right\} \leqslant \sum_{k\geqslant n}\mathbf{P}\left\{\left|\frac{S_k}{k}\right| \geqslant \delta\right\} \leqslant \frac{3}{\delta^4}\sum_{k\geqslant n}\frac{1}{k^2} \to 0,$$

因此, 根据博雷尔 - 坎泰利引理 (第二章 §10), 几乎必然 $S_n/n \to 0$.

F. 豪斯多夫的证明, 对于任意 $\varepsilon > 0$, "几乎必然 $S_n = o(n^{1/2+\varepsilon})$", 是类似地进行的: 因为 $\mathbf{E}S_n^{2r} = O(n^r)$, 对于任意整数$r > 1/(2\varepsilon)$, 则当 $n \to \infty$ 时

$$\mathbf{P}\left\{\sup_{k\geqslant n}\left|\frac{S_k}{k^{1/2+\varepsilon}}\right| \geqslant \delta\right\} \leqslant \sum_{k\geqslant n}\mathbf{P}\left\{\left|\frac{S_k}{k^{1/2+\varepsilon}}\right| \geqslant \delta\right\}$$
$$\leqslant \frac{1}{\delta^{2r}}\sum_{k\geqslant n}\mathbf{E}\left|\frac{S_k}{k^{1/2+\varepsilon}}\right|^{2r} \leqslant \frac{c}{\delta^{2r}}\sum_{k\geqslant n}\frac{k^r}{k^{r+2\varepsilon r}} \to 0,$$

其中 c 是常数. 由此 (仍然根据博雷尔 - 坎泰利引理) 可得, 几乎必然

$$\frac{S_n}{n^{1/2+\varepsilon}} \to 0.$$

由以上的讨论可见, 证明的关键是对于概率 $\mathbf{P}\{|S_n| \geqslant t(n)\}$ 得到 "好" 估计, 其中对伯雷尔, $t(n) = n$, 豪斯多夫, $t(n) = n^{1/2+\varepsilon}$, (对哈代和李特尔伍德, $t(n) = (n\ln n)^{1/2}$).

正是为得到概率 $\mathbf{P}\{|S_n| \geqslant t(n)\}$ 的 "好" 估计. А. Я. 辛钦用到他的 "辛钦不等式" (25) (确切地说, 这些不等式中的第一个).

关于辛钦 (右和左) 不等式对于任意 $p > 0$ 的推导, 以及在 (25) 式中关于常数 A_p 和 B_p 之最优性的证明, 见 Г. 佩舍基尔 (Г. Пешкир) 和 А. Н. 施利亚耶夫的综述性文章: Неравен-ства Хинчина и мартингальное расширение сферы их действия (辛钦不等式及其应用范围的扩展)//Успехи математических наук. —. 1995 年, Т.50, 5. стр. 3~62.

由 (25) 式的第一个不等式, 当 $p = 2m$ 时, А. Я. 辛钦得到, 对于任意 $t > 0$,

$$\mathbf{P}\{|X_n| > t\} \leqslant t^{-2m}\mathbf{E}|X_n|^{2m} \leqslant \frac{(2m)!}{2^m m!}t^{-2m}[X]_n^{2m}.$$

由斯特林公式

$$\frac{(2m)!}{2^m m!} \leqslant D\left(\frac{2}{\mathrm{e}}\right)^m m^m,$$

其中 $D = \sqrt{2}$. 因此, 若设 $m = [t^2/(2[X]_n^2)]$, 则有

$$\mathbf{P}\{|X_n| > t\} \leqslant D\left(\frac{2m[X]_n^2}{\mathrm{e}t^2}\right)^m \leqslant D\mathrm{e}^{-m} \leqslant D\exp\left\{1 - \frac{t^2}{2[X]_n^2}\right\}$$
$$= D\mathrm{e}\exp\left\{\frac{-t^2}{2[X]_n^2}\right\} = c\exp\left\{\frac{-t^2}{2[X]_n^2}\right\},$$

其中 $c = De = \sqrt{2}e$.

由此估计, 可得不等式

$$\mathbf{P}\{|S_n| > t\} \leqslant e^{-\frac{t^2}{2n^2}}.$$

A. Я. 辛钦曾经利用该式, 证明 $S_n = O(\sqrt{n \ln\ln n})$ (a.c.).

在 Y. S. 乔 (Y. S. Chow) 和泰切尔 (H. Teicher) 的书中 [73], 有大量在这一节引用的不等式. 定理 2 属于 E. Lenglavt [39].

§4. 见 J. L. 杜布的专著 [20].

§5. 这里叙述的内容遵循如下读者的文章: Ю. М. 卡巴诺夫 (Ю. М. Кабанов), Р. Ш. 利普彩尔, А. Н. 施利亚耶夫 [26], H. -J. 恩格尔伯特 (H. -J. Engelbert), А. Н. 施利亚耶夫 [79], 以及奈维尤 (J. Neveu) 的书 [50]. 定理 4 和例是 Р. Ш. 利普彩尔提供的.

§6. 这里对 "绝对连续性和奇异性" 问题提供的处理方法, 以及所介绍的结果, 包含在 Ю. М. 卡巴诺夫, Р. Ш. 利普彩尔, А. Н. 施利亚耶夫的工作 [26] 中.

§7. 定理 1 和定理 2 世界语属于 А. А. 诺维科夫 (А. А. Новиков) [52]. 引理 1 是著名的吉尔萨诺夫 (Гирсанов) 定理的 "离散" 类似 (见 [41]).

§8. 亦见 Р. Ш. 利普彩尔和 А. Н. 施利亚耶夫的书 [91], J. 扎克德和 А. Н. 施利亚耶夫 [87], 在该书中讲述相当一般情形的随机过程论 (鞅, 半鞅 ……).

§9. 这里叙述遵从 [98],[100]. 对伊藤清 (Itô Kiyosi) 公式的推广, 所叙述的方法的发展, 见 H. 福尔默 (H. Föllmer), Ph. 普罗泰尔 (Ph. Protter) 和 А. Н. 施利亚耶夫的文章 [101].

§10. 关于保险中的鞅方法, 见格贝尔 (H. Gerber) 的书 [123]. 所作证明接近 [98] 中相应的证明.

§11~§12. 涉及鞅方法, 在金融数学和金融工程学中的应用问题, 较为详细的叙述, 见 [100].

§13. 最优停止规则主要专著是: Е. В. 邓肯 (E. B. Dynkin) 和 А. А. 尤什克维奇 (А. А. Юшкевич) [102], H. R. 罗宾斯 (H. R. Robbins), D. 西格蒙德 (D. Sigmund) 和 Y. S. 乔 [59], А. Н. 施利亚耶夫 [78].

第八章

§1~§2. 关于马尔可夫链的定义和基本性质, 亦见下列著作: Е. В. 邓肯和 А. А. 尤什克维奇 [102], Е. В. 邓肯 [21], А. Д. 温策尔 (А. Д. Вентцель) [11], J. L. 杜布 [20], И. И. 基赫曼和 А. В. 斯科罗霍德 [14], L. 布赖曼 [8], 钟开莱 (Kai Lai Chung) [75],[120], D. 雷夫尤兹 (D. Revuz) [117].

§3~§7. 关于马尔可夫链的, 极限, 遍历和平稳概率分布问题, 见 А. Н. 柯尔莫戈洛夫文章 [28], 和如下书籍: W. 费勒 (W. Feller) [69], А. А. 博罗夫科夫 [7], [104], R. B. 埃什 (R. B. Ashi), 钟开莱 [120], D. 雷夫尤兹 [117], Е. В. 邓肯和 А. А. 尤什克维奇 [102].

§8. 简单随机游动, 是最简单的马尔可夫链的经典例子, 对此情形曾经发现了许多规律性 (例如, 常返性于非常返性, 遍历性等等). 有关内容在许多图书中都有, 例如, 上面引用过的书 [7],[80],[120],[117].

§9. 统计序列分析问题 (А. 瓦尔德 [9], M. de 格鲁特 (M. de Gloot) [18], S. 萨克斯 (S. Sacks) [22], А. Н. 施利亚耶夫 [78]), 决定了对于最优停时问题的兴趣. 关于马尔可夫链的最优停时规则本身, 有如下著作: Е. В. 邓肯和 А. А. 尤什克维奇 [102], А. Н. 施利亚耶夫 [78]), P. 比林斯利 (P. Billingsley) [106] 的某些章节. 最优停时问题的鞅方法, 见 H. R. 罗宾斯, D. 西格蒙德和 Y. S. 乔的专著 [59].

概率的数学理论形成的简史

该简史是本书的作者, 为 А. Н. 柯尔莫戈洛夫的专著《概率论的基本概念》[32] 第三版, 作为补充而写的.

参考文献

[1] Александров П. С. Введение в общую теорию множеств и функций. — М.: Гостехиздат, 1948.

[2] Александрова Н. В. Математические термины. — М.: Высшая школа, 1978.

[3] Бернштейн С. Н. О работах П. Л. Чебышева по теории вероятностей//Научное наследие П. Л. Чебышева. Вып.1: Математика. — 1945. — С. 59 — 60.

[4] Бернштейн С. Н. Теория вероятностей. — 4-е изд. — М.: Гостехиздат, 1946.

[5] Биллингсли П. Сходимость вероятностных мер. — М.: Наука, 1977.

[6] Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. — 3-е изд. — М.: Наука, 1983.

[7] Боровков А. А. Теория вероятностей. — 3-е изд. — М.: УРСС, 1999.

[8] Брейман (Breiman L.). Probability. — Reading, MA: Addison-Wesley, 1968.

[9] Вальд А. Последовательный анализ. — М.: Физматгиз, 1960.

[10] Ван дер Варден Б. Л. Математическая статистика. — М.: ИЛ, 1960.

Под номерами 85 — 101 идет литература, добавленная во втором издании к той, которая была приведена в первом издании книги. Под номерами 102 — 136 идет литература, добавленная в настоящем издании.

[11] Вентцель А. Д. Курс теории случайных процессов. — М.: Наука, 1975.

[12] Гарсиа (Garsia A. M.). A simple proof of E. Hopf's maximal ergodic theorem // Journal of Mathematics and Mechanics. — 1965. — V.14, № 3. — P. 381 — 382.

[13] Гихман И. И., Скороход А. В. Введение в теорию случайных процессов. — М.: Наука, 1977.

[14] Гихман И. И., Скороход А. В. Теория случайных процессов: В3т. — М.: Наука, 1971 — 1975. 汉译本:《随机过程论》, 第一卷 (邓永录等译), 第二卷 (周概容译), 1986.

[15] Гнеденко Б. В. Курс теории вероятностей. — 6-е изд. — М.: Наука, 1988. 汉译本:《概率论教程》(丁寿田), (第三版), 高等教育出版社, 1961.

[16] Гнеденко Б. В., Колмогоров А. Н. Предельные распределения для сумм независимых случайных величин. — М.: Л.: Гостехиздат, 1949. 汉译本:《相互独立随机变量之和的极限分布》(王寿仁译), 科学出版社, 1950.

[17] Гнеденко Б. В., Хинчин А. Я. Элементарное введение в теорию вероятностей. — 9-е изд. — М.: Наука, 1982.

[18] Де Гроот М. Оптимальные статистические решения. — М.: Мир, 1974.

[19] Дохерти (Doherty M.). An amusing proof in fluctuation theory// Combinatorial Mathematics, III: Proceedings of the Third Australian Conference, Univ. Queensland, St. Lucia 1974. — Berlin etc.: Springer-Verlag, 1975. — P.101 — 104. — (Lecture Notes in Mathematics; V.452.)

[20] Дуб Дж. Л. Вероятностные процессы. — М.: ИЛ, 1956.

[21] Дынкин Е. Б. Марковские процессы. — М.: Физматгиз, 1963.

[22] Закс Ш. Теория статистических выводов. — М.: Мир, 1975.

[23] Ибрагимов И. А., Линник Ю. В. Независимые и стационарно связанные величины. — М.: Наука, 1965.

[24] Ибрагимов И. А., Розанов Ю. А. Гауссовские случайные процессы. — М.: Наука, 1970.

[25] Исихара А. Статистическая физика. — М.: Мир, 1973.

[26] Кабанов Ю. М., Липцер Р. Ш., Ширяев А. Н. К вопросу об абсолютной непрерывности и сингулярности вероятностных мер// Математический сборник. — 1977, — Т. 104. № 2. — С. 227 — 247.

[27] Кемени Дж., Снелл Дж. Конечные цепи Маркова. — М.: Наука, 1970.

[28] Колмогоров А. Н. Цепи Маркова со счетным числом возможных состояний//Бюллетень МГУ. — 1937. — Т.1, № 3. — С. 1 — 16.

[29] Колмогоров А. Н. Стационарные последовательности в гильбертовском пространстве//Бюллетень МГУ. — 1941. — Т. 2, № 6. — С. 1 — 40.

[30] Колмогоров А. Н. Роль русской науки в развитии теории вероятностей//Ученые записки МГУ. — 1947. — Вып. 91. — С. 53 — 64. 汉译本:《概率论》(见 "数学, 它的内容、方法和意义" (卷2)), 科学出版社, 1961.

[31] Колмогоров А. Н. Теория вероятностей//Математика, ее содержание, методы и значение. — М.: Изд-во АН СССР, 1956. — Т. II. — С. 252 — 284.

[32] Колмогоров А. Н. Основные понятия теории вероятностей. — М.; Л.: ОНТИ, 1936; 2-е изд. М.: Наука, 1974; 3-е изд. М.: Фазис, 1998. 汉译本:《概率论的基本概念》(丁寿田译), 商务印书馆, 1952.

[33] Колмогоров А. Н., Фомин С. В. Элементы теории функций и функционального анализа. — 6-е изд. —М.: Наука, 1989.

[34] Колчин В. Ф., Севастьянов Б. А., Чистяков В. П. Случайные размещения. — М.: Наука, 1976.

[35] Крамер Г. Математические методы статистики. — 2-е изд. — М.: Мир, 1976. 汉译本:《统计学数学方法》(魏宗舒译), 上海科技出版社, 1983.

[36] Кубилюс Й. Вероятностные методы в теории чисел. — Вильнюс: Гос. изд-во полит. и науч. лит. ЛитССР, 1959.

[37] Ламперти Дж. Вероятность. — М.: Наука, 1973.

[38] Ламперти (Lamperti J.). Stochastic Processes. — Aarhus Univ., 1974. — (Lecture Notes Series; № 38).

[39] Ленгляр (Lenglart E.). Relation de domination entre deux processus// Annales de l'Institut H. Poincaré Sect. B. (N. S.). — 1977. — V.13, № 2. — P. 171 — 179.

[40] Леонов В. П., Ширяев А. Н. К технике вычисления семиинвариантов// Теория вероятностей и ее применения. — 1959. — Т. IV, вып. 2. — С. 342 — 355.

[41] Липцер Р. Ш., Ширяев А. Н. Статистика случайных процессов. — М.: Наука, 1974. 汉译本:《随机过程统计》.

[42] Лоэв М. Теория вероятностей. — М.: ИЛ, 1962. 汉译本:《概率论》, 上册, (梁文骐译), 科学出版社, 1966.

[43] Марков А. А. Исчисление вероятностей. — 3-е изд. — СПб., 1913.

[44] Майстров Л. Е. Теория вероятностей (исторический очерк). — М.: Наука, 1967.

[45] Математика XIX века/ Под ред. А. Н. Колмогорова и А. П. Юшкевича. — М.: Наука, 1978.

[46] Мешалкин Л. Д. Сборник задач по теории вероятностей. — М.: Изд-во МГУ, 1963. 汉译本:《概率论习题集》(盛骤等译), 高等教育出版社, 1984.

[47] Мейер (Meyer P.-A.). Martingales and Stochastic Integrals. I. — Berlin etc.: Springer-Verlag, 1972. — (Lecture Notes in Mathematics; V. 284).

[48] Мейер П. -А. Вероятность и потенциалы. — М.: Мир, 1973.

[49] Неве Ж. Математические основы теории вероятностей. — М.: Мир, 1969.

[50] Невё (Neveu J.). Discrete-Parameter Martingales. — Amsterdam etc.: North-Holland, 1975.

[51] Нейман Ю. Вводный курс теории вероятностей и математической статистики. — М.: Наука, 1968.

[52] Новиков А. А. Об оценках и асимптотическом поведении вероятностей непересечения подвижных границ суммами независимых случайных величин//Известия АН СССР. Серия математическая. — 1980. — Т. 40, вып. 4. — С. 868 — 885.

[53] Петров В. В. Суммы независимых случайных величин. — М.: Наука, 1972

[54] Прохоров Ю. В. Асимптотическое поведение биномиального распределения// Успехи математических наук. — 1953. — Т. VIII, вып. 3(55). — С. 135—142.

[55] Прохоров Ю. В. Сходимость случайных процессов и предельные теоремы теории вероятностей// Теория вероятностей и ее применения. — 1956. — Т. I, вып. 2. — С. 177—238.

[56] Прохоров Ю. В., Розанов Ю. А. Теория вероятностей. — 2-е изд. — М.: Наука, 1973.

[57] Рамачандран Б. Теория характеристических функций. — М.: Наука, 1975.

[58] Реньи (Rényi A.) Probability Theory. — Amsterdam: North-Holland, 1970.

[59] Роббинс Г., Сигмунд Д., Чао И. Теория оптимальных правил остановки. — М.: Наука, 1977.

[60] Розанов Ю. А. Стационарные случайные процессы. — М.: Физматгиз, 1963.

[61] Сарымсаков Т. А. Основы теории процессов Маркова. — М.: Гостехиздат, 1954.

[62] Севастьянов Б. А. Ветвящиеся процессы. — М.: Наука, 1971.

[63] Синай Я. Г. Введение в эргодическую теорию. — Ереван: Изд-во Ереван. ун-та, 1973.

[64] Сираждинов С. Х. Предельные теоремы для однородных цепей Маркова. — Ташкент: Изд-во АН УзССР, 1955.

[65] Справочник по теории вероятностей и математической статистике/ Под ред. В. С. Королюка. — Киев: Наукова думка, 1978.

[66] Стоут (Stout W. F.). Almost Sure Convergence. — New York etc.: Academic Press, 1974.

[67] Теорія имовірностей. — Київ: Вища школа, 1976.

[68] Тодхантер (Todhunter I.). A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace. — London: Macmillan, 1865.

[69] Феллер В. Введение в теорию вероятностей и ее приложения: В 2-х т. — М.: Мир, 1984. 汉译本:《概率论导引》(胡迪鹤译), 人民邮电出版社, 2006.

[70] Халмош П. Теория меры. — М.: ИЛ, 1953. 汉译本:《测度论》(王建华译), 科学出版社.

[71] Хеннан Э. Анализ временных рядов. — М.: Наука, 1964.

[72] Хеннан Э. Многомерные временные ряды. — М.: Мир, 1974.

[73] Чао, Тейчер (Chow Y. S., Teicher H.). Probability Theory. Independence, Interchangeability, Martingales. — 3rd ed. — New York: Springer-Verlag, 1997.

[74] Чебышев П. Л. Теория вероятностей: Лекции акад. П. Л. Чебышева, читанные в 1879, 1880 гг./ Издано А. Н. Крыловым по записи А. М. Ляпунова. — М.; Л., 1936.

[75] Чжун Кай-лай. Однородные цепи Маркова. — М.: Мир, 1964.

[76] Ширяев А. Н. Случайные процессы. — М.: Изд-во МГУ, 1972.

[77] Ширяев А. Н. Вероятность, статистика, случайные процессы: В 2-х т. — М.: Изд-во МГУ, 1973 — 1974.

[78] Ширяев А. Н. Статистический последовательный анализ. — 2-е изд. — М.: Наука, 1976.

[79] Энгельберт, Ширяев (Engelbert H.-J., Shiryaev A. N.). On the sets of convergence of generalized submartingales // Stochastics. — 1979. — V. 2, № 3. — P. 155 — 166.

[80] Эш (Ash R. B.). Basic Probability Theory. — New York etc.: Wiley, 1970.

[81] Эш (Ash R. B.). Real Analysis and Probability.—New York etc.: Academic Press, 1972.

[82] Эш, Гарднер (Ash R. B., Gardner M. F.). Topics in Stochastic Processes. — New York etc.: Academic Press, 1975.

[83] Яглом А. М., Яглом И. М. Вероятность и информация. — 3-е изд. — М.: Наука, 1973. 汉译本:《概率与信息》, 科学出版社.

[84] Гринвуд, Ширяев (Greenwood P. E., Shiryayev A. N.). Contiguity and the Statistical Invariance Principle. — London: Gordon & Breach, 1985.

[85] Дадли (Dudley R. M.)Distances of probability measures and random variables// Annals of Mathematical Statistics. — 1968. — V.39, № 5. — P. 1563 — 1572.

[86] Дакуна-Кастелль, Дюфло (Dacunha-Castelle D., Duflo M.). Probabilités et statistiques: 1, 2. — Paris: Masson. — 1: Problèmes à temps fixe. — 1982; — 2: Problèmes à temps mobile. — 1983. — Перев. на англ. яз.: Probability and Statistics: V. I, II. — Berlin etc.: Springer-Verlag, 1986.

[87] Жакод Ж., Ширяев А. Н. Предельные теоремы для случайных процессов: В 2-х т. — М.: Физматлит, 1994.

[88] Золотарев В. М. Современная Теория суммирования независимых случайных величин. — М.: Наука, 1986.

[89] Ле Кам (Le Cam L.). Asymptotic Methods in Statistical Decision Theory. — Berlin etc.: Springer-Verlag, 1986.

[90] Лизе, Вайда (Liese F., Vajda I.). Convex Statistical Distances. — Leipzig: Teubner, 1987.

[91] Липцер Р. Ш., Ширяев А. Н. Теория мартингалов. — М.: Наука, 1986.

[92] Петров В. В. Предельные теоремы для сумм независимых случайных величин. — М.: Наука, 1987.

[93] Поллард (Pollard D.). Convergence of Stochastic Processes. — Berlin etc.: Springer-Verlag, 1984.

[94] Пресман Э. Л. О сближении по вариации распределения суммы независимых бернуллиевских величин с пуассоновским законом// Теория вероятностей и ее применения. — 1985. — Т. XXX, вып. 2. — С. 391 — 396.

[95] Розанов Ю. А. Теория вероятностей, случайные процессы и математическая статистика. — М.: Наука, 1985.

[96] Ротарь В. И. Кобобщению теоремы Линдеберга — Феллера// Математические заметки.— 1975. — Т.18, вып. 1. — С. 129 — 135.

[97] Севастьянов Б. А. Курс теории вероятностей и математической статистики. — М.: Наука, 1982.

[98] Ширяев (Shiryayev A. N.) Probability. — 2nd ed. — Berlin etc.: Springer-Verlag, 1995.

[99] Ширяев (Shirjayev A. N.) Wahrscheinlichkeit. — Berlin: VEB Deutscher Verlag der Wissenschaften, 1988.

[100] Ширяев А. Н. Основы стохастической финансовой математики: В 2-х т. — М.: ФАЗИС, 1998.

[101] Фёллмер, Проттер, Ширяев (Föllmer H., Protter Ph., Shiryaev A. N.). Quadratic covariation and an extension of Itô's formula//Bernoulli. — 1995. — V.1, № 1/2. — P. 149 — 170.

[102] Дынкин Е. Б., Юшкевич А. А. Теоремы и задачи о процессах Маркова. — М.: Наука, 1967.

[103] Гнеденко, Колмогоров (Gnedenko B. V., Kolmogorov A. N). Limit Distributions for Sums of Independent Random Variables. — Reading, MA, etc.: Addison-Wesley, 1954.

[104] Боровков А. А. Эргодичность и устойчивость случайных процессов. — М.: УРСС, 1999.

[105] Гриммет, Стирзакер (Grimmet G. R., Stirzaker D. R.). Probability and Random Processes. — Oxford: Clarendon Press, 1993.

[106] Биллингсли (Billingsley P.). Probability and Measure. — 3rd ed. — New York: Wiley, 1995.

[107] Боровков А. А. Математическая статистика. — М.: Наука, 1984.

[108] Дарретт (Durrett R.). Probability: Theory and Examples. — Pacific Grove, CA: Wadsworth & Brooks/Cole, 1991.

[109] Дарретт (Durrett R.). Stochastic Calculus. — Boca Raton, FL: CRC Press, 1996.

[110] Дарретт (Durrett R.). Brownian Motion and Martingales in Analysis. — Belmont, CA: Wadsworth International Group, 1984.

[111] Калленберг (Kallenberg O.). Foundations of Modern Probability. — 2nd ed. — New York: Springer-Verlag, 2002.

[112] Карлин, Тейлор (Karlin S., Taylor H. M.). A First Course in Stochastic Processes. — 2nd ed. — New York etc.: Academic Press, 1975.

[113] Кашин Б. С., Саакян А. А. Ортогональные ряды. — 2-е изд. — М.: АФЦ, 1999.

[114] Жакод, Проттер (Jacod J., Protter Ph.). Probability Essentials. — Berlin etc.: Springer-Verlag, 2000.

[115] Нётс (Neuts V. F.) Probability. — Boston, MA: Allyn & Bacon, 1973.

[116] Плато (Plato J.). Creating Modern Probability. — Cambridge: Cambridge Univ. Press, 1998.

[117] Ревюз Д. Цепи Маркова. — М.: РФФИ, 1997.

[118] Уильямс (Williams D.). Probability with Martingales. — Cambridge: Cambridge Univ. Press, 1991.

[119] Холл, Хейде (Hall P., Heyde C. C.). Martingale Limit Theory and Its Applications. — New York etc.: Academic Press, 1980.

[120] Чжун Кай-Лай (Chung Kai Lai). Elementary Probability Theory with Stochastic Processes. — 3rd ed. — Berlin etc.: Springer-Verlag, 1979. 汉译本:《初等概率论和随机过程》(吕乃刚等译), 人民教育出版社.

[121] Математическая энциклопедия: В 5 т./Гл. ред. И. М. Виноградов. — М.: Советская энциклопедия, 1977 — 1985. 汉译本:《数学百科全书》, 第一卷 ~ 第五卷, 科学出版社, 1984 — 2000.

[122] Стиглер (Stigler S. M.).The History of Statistics: The Measurement of Uncertainty Before 1900. — Cambridge: Belknap Press of Harvard Univ. Press, 1986.

[123] Гербер Х. Математика страхования жизни. — М.: Мир, 1995.

[124] Эренфесты П. и Т. (Ehrenfest P., Ehrenfest T.). Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem//Physikalische Zeitschrift. — 1907. — V.8. — P. 311 — 314.

[125] Теория вероятностей и математическая статистика: энциклопедия/ Гл. ред. Ю. В. Прохоров. — М.: Большая Российская энциклопедия, 1999.

[126] Вольфрам (Wolfram S.). The Mathematica® Book. — 4th ed. — Champaign; Cambridge: Wolfram Media; Cambridge Univ. Press, 1999.

[127] Дуб (Doob J. L.). What is a martingale?//The American Mathematical Monthly. — 1971. — V.78. — P. 451 — 463.

[128] Синай Я. Г. Курс теории вероятностей. — М.: Изд-во МГУ, 1985. — 2-е изд., 1986.

[129] Синай (Sinaĭ Ya. G.). Topics in Ergodic Theory. — Princeton, NJ: Princeton Univ. Press, 1999. — (Princeton Mathematical Series; V. 44.)

[130] Вальтерс (Walters P.) An Introduction to Ergodic Theory. — New York etc.: Springer-Verlag, 1982.

[131] Булинский А. В., Ширяев А. Н. Теория случайных процессов. — М.: Физматлит, 2003.

[132] Хмаладзе Э. В. Мартингальный подход в теории непараметрических критериев согласия// Теория вероятностей и ее применения. — 1981. — Т. XXVI. вып. 2. — С. 246 — 265.

[133] Гамильтон (Hamilton J. B.). Time Series Analysis. — Princeton, NJ: Princeton Univ. Press, 1994.

[134] Бернулли Я. О законе больших чисел. — Ч. 4: Искусство предположений. — М.: Наука, 1986.

[135] Лукач Е. Характеристические функции. — М.: Наука, 1979.

[136] Хренников (Khrennikov A.). Interpretations of Probability. — Utrecht: VSP, 1999.

名词索引

(汉语拼音为序, 右上角带 “1” 者为第一卷页码)

T

W

X

Y

Z

人名表

(汉语拼音为序)

A

阿什	R. B. Ash	Р. Эш
埃尔米特	Ch. Hermite	Ш. Эрмит
埃伦弗斯特	P. Ehrenfest	П. Эренфест
埃伦弗斯特	T. Ehrenfest	Т. Эренфест
埃森	C. G. Esseen	К. Г. Эссеен
爱因斯坦	A. Einstein	А. Эйнштейн
埃什	R. B. Ashi	Р. Б. Эш
埃舍	Esher	Эшер
埃特麦迪	N. Etemady	Н. Этемади
奥塔维安尼	Ottawiani	Оттавиани

B

巴特利特	M. S. Bartlett	М. С. Бартлетт
巴拿赫	S. Banach	С. Банах
巴彻里耶	L. bachelier	Л. Башелье
邦弗尔罗尼	Bonferroni	Бонферрони
鲍斯	S. N. Bose	Ш. Бозе
贝尔	A. G. Bell	А. Г. Белл
贝尔	R. L. Baire	Р. Л. Бэр

贝尔曼	R. E. Bellman	Р. Э. Беллман
贝克尔雷尔	A. H. Becquerel	А. Х. Беккерелем
贝里	A. C. Berry	А. С. Берри
贝塞尔	F. W. Bessel	Ф. В. Бессель
贝叶斯	T. Bayes	Т. Байес
比林斯利	P. Billingsley	П. Биллингсли
彼得罗夫	W. W. Petrov	В. В. Петров
毕达哥拉斯	Pythagoras	Пифагор
波拉德	D. Pollard	Д. Пполлард
波利亚	G. Pólya	Д. Пойа
伯恩斯坦	S. N. Bernstein	С. Н. Бернштейн
伯克霍尔德	D. L. Burkholder	Д. Л. Буркхольдер
伯克霍夫	G. D. Birkhoff	Дж. Д. Биркгоф
伯努利	D. Benoulli	Д. Бернулли
伯努利	J. Benoulli	Я. Бернулли
泊松	S. D. Poisson	С. Д. Пуассон
博赫纳	S. Bochner	С. Бохнер
博雷尔 (波莱尔)	E. Borel	Э. Борель
博利舍夫	L. N. Bolishev	Л. Н. Большев
博罗夫科夫	A. A. Borowkov	А. А. Боровков
布罗登	T. Brodén	Т. Броден
布尔	G. Boole	Дж. Буль
布耳曼	G. BoHlmann	Г. Больман
布耳兹曼	L. Boltzmann	Л. Больцман
布莱克韦尔	D. H. Blackwell	Д. Блэкуэлл
布赖曼	L. Breiman	Л. Брейман
布朗	E. T. Brown	Э. Т. Броун
布雷	J. R. Bray	Я. Р. Брэй
布罗吉	U. Broggi	У. Брогги
布洛赫	A. Bloch	А. Блох
布尼亚科夫斯基	A. J. Buniakowsky	А. Я. Буняковский
布西	R. S. Busey	Р. С. Вьюсу

C~D

查普曼	D. G. Chapman	Д. Г. Чепман
达布	J. G. Darboux	Ж. Г. Дарбу
达德利	R. M. Dudley	Р. Дадли

达昆纳-卡斯特里	D. Ducunna-Castelle	Д. Дакуна-Кастелле
达雷特	R. Durrett	Р. Даррет
戴维斯	H. T. Davis	Х. Т. Дэвис
德沃列茨基	—	Дворецкий
德格鲁特	M. H. deGroot	М. Де Гроот
邓肯	E. B. Dynkin	И. Б. Дынкин
狄拉克	P. A. M. Dirac	П. А. М. Дирак
狄利克雷	P. G. L. Dirichlet	П. Г. А. Дирихле
笛卡儿	R. Descartes	Р. Декарт
棣莫弗	A. Dé Moivre	А. Дé Муавр
杜布	J. L. Doob	Дж. Л. Дуб
杜弗劳	M. Duflo	М. Дюфло

E

恩格尔伯特	H.-J. Engelbert	Г.-Ю. Энгельберт

F

法图	F. Fatou	П. фату
范德瓦尔登	B. L. van der Waerden	Б. Л. Ван дер Варден
范基	Fan Ky	Ки Фан
费勒	W. Feller	В. Феллер
费马	P. Fermat	П. Ферма
费内提	B. de Finetti	Б. де Финетти
费希尔 (费歇尔)	R. A. Fisher	Р. А. Фишер
费叶尔	L. Féjer	Л. Фейер
弗雷歇	M. Fréchet	Ф. Фреше
弗米	E. Fermi	Э. Ферми
冯・诺伊曼	J. von Neuman	Дж. Фон Нейман
福尔默	H. Föllmer	Г. Фёллмер
福明	S. W. Fomin	С. В. Фомин
傅比尼	G. Fubini	Г. фубини
费里德曼	H. M. Friedman	Г. М. Фельдман
傅里叶	J. B. J. Fourier	Ж. Б. Ж. Фурье

G

盖尔登	J. A. H. Gyldén	Дж. А. Х. Гюлден
冈贝尔	E. J. Gumbel	Э. Гумбель
高斯	G. F. Gauss	К. Ф. Гаусс
格贝尔	H. Gerber	Х. Гербер
格拉姆	G. P. Gram	Г. П. Грам
格里米特	G. R. Grimmit	Дж. Гриммет
格里汶科	W. I. Glivenko	В. И. Гливенко
格林伍得	P. E. Greenword	П. Е. Гринвуд
格鲁特	M. de Gloot	М. де Гроот
格涅坚科	B. V. Gnedenko	Б. В. Гнеденко
哥塞特	W. S. Gosset	В. С. Госсет

H

哈恩	H. Hahn	Г. Хан
哈尔	A. Haar	А. Хаар
哈伊克	J. Hajek	Дж. Гаек
哈代	G. H. Hardy	Г. Х. Харди
哈尔默斯	P. R. Halmos	П. Халмош
哈密顿	W. R. Hamilton	У. Р. Гамильтон
哈特曼	P. Hartman	П. Хартман
汉南	E. J. Hannan	Э. Дж. Хеннан
海林格	E. Helinger	Э. Хеллингер
海涅	H. E. Heine	Г. Э. Нейне
豪斯多夫	F. Hausdorff	Ф. Хаусдрф
赫尔德	O. L. Hölder	О. Л. Гёльдер
赫尔格洛茨	G. Herglotz	Г. Герглотц
赫利	E. Helly	Э. Хелли
惠更斯	Ch. Huyghens	Х. Гюйгенс
霍普夫	H. Hopf	Х. Хопф
霍奇	W. W. D. Hodge	У. В. Д. Ходе

J

吉尔萨诺夫		И. Гирсанов
基赫曼	E. E. Gihman	И. И. Гихман
加德钠	M. F. Cardner	М. Гарднер
加尔西亚	A. M. Garsia	А. М. Гарсиа
角谷	S. Kakutani	С. Какутани

K

卡巴诺夫	Y. M. Kabanov	Ю. М. Кабанов
卡尔达诺	G. Cardanno	Дж. Кардано
卡尔莱曼	T. Carleman	Т. Карлеман
卡尔曼	R. E. Kalman	Р. И. Калман
卡拉泰奥多里	C. Carathéodory	К. Каратеодори
卡利卡尼尼	C. Calcagnini	Ч. Калькανини
卡姆	L. Le Cam	Л. Ле Кам
凯麦尼	J. G. Kemeny	Дж. Кемени
坎泰利	F. P. Cantelli	Ф. П. Кантелли
康托尔	G. Cantor	Г. Контор
考克斯	J. C. Cox	Дж. К. Кокс
柯尔莫戈洛夫	A. N. Kolmogorov	А. Н. Колмогоров
柯西	A. L. Cauchy	О. Л. Коши
科尔钦	W. F. Kerchen	В. Ф. Колчин
克里克伯格	Krickbeg	Крикберг
克拉默	H. Cramer	Г. Крамер
克罗内克	L. Kronecker	Л. Кронекер
库尔贝克	S. Kullback	С. Кульбак

L

拉奥	C. R. Rao	С. Р. Рао
拉德马赫	H. Rademacher	Г. А. Радемахер
拉东	J. Radon	Дж. Радон
拉格朗日	J. L. Lagrange	Ж. Л. Лагранж
拉曼钱德兰	B. Lamanchandran	Б. Раманчадран
拉普拉斯	P. S. Laplace	П. С. Лаплас
莱布尼茨	G. W. Lebniz	Г. В. Лейбниц
莱斯	F. Liese	Ф. Лизе
莱维	E. E. Levy	И. И. Леви
兰珀蒂	J. Lamperti	Дж. Ламперти
勒贝格	H. L. Lebesgue	А. Л. Лебег
雷夫尤兹	D. Revuz	Д. Ревюз
雷内伊	A. Rènyi	А. Ренуи
黎曼	G. F. B. Riemann	Г. Ф. Б. Риман
李雅普诺夫	A. M. Lyapunov	А. М. Ляпунов

李特尔伍德	J. E. Littlewood	Дж. И. Литтлвуд
里斯	F. Riesz	Ф. Рисс
利普彩尔	R. S. Lipchail	Р. Ш. Липцер
利普希茨	R. O. S. Lipschitz	Р. Липшиц
刘维尔	J. Liouville	Дж. Лиувилль
列昂诺夫	W. P. Leonov	В. П. Леонов
列维	P. P. Lévy	П. П. Леви
林德伯格	J. W. Linderberg	Дж. У. Линдеберг
林格拉特	E. Linglart	Э. Лингляр
洛埃甫	M. Loéve	М. Лоэв
洛必达	L' Hospital	Г. Лопиталь
伦德伯格	G. A. Lundeberg	Г. А. Лундберг
鲁宾斯坦	M. Rubinstein	М. Рубинштейн
罗塔里	W. E. Rotari	В. И. Ротарь
罗扎诺夫	Yu. A. Rozanov	Ю. А. Розанов
罗宾斯	H. R. Robbins	Г. Р. Робенс
罗斯	R. A. Ross	Р. А. Росс

M

马尔可夫	A. A. Markov	А. А. Марков
马歇尔	A. W. Marshall	А. В. Маршалл
马钦凯维奇	J. Matcinkiewicz	Й. Марцинкевич
马哈拉诺比斯	P. S. Mahalanobis	П. С. Махаланобис
迈斯特罗夫	D. I. Mastrov	Д. Е. Майстров
麦耶	P. -A. Meyer	П. -А. Мейер
麦克米兰	B. McMillan	Б. Макмиллан
麦克斯韦	J. C. Maxwell	Д. К. Максвелл
曼	H. B. Mann	Х. Б. Манн
梅沙尔金	L. D. Mesharlkin	Л. Д. Мешалкин
门罗	Monroe	Монро
米泽斯	R. Mises	Р. Мизесу
闵可夫斯基	H. Minkowski	Г. Минковский
默瑟	J. Mercer	Ж. Мерсер

N

奈曼	Yu. Neyman	Ю. Нейман
奈维尤	J. Neveu	Р. Невё
尼科迪姆	O. M. Nikodym	О. М. Никодим
牛顿	I. Newton	И. Ньютон
诺维科夫		А. А. Новиков

O

欧几里得	Euclid	Евклид
欧拉	L. Euler	Л. Эйлер

P

帕赞	E. Parzen	И. Парзен
帕利	W. Pauli	В. Паули
帕乔里	L. Papcioli	Л. Пачоли
帕塞瓦尔	M. A. Parseval	М. А. Пасеваль
帕斯卡	B. Pascal	Б. Паскаль
庞加莱	J. H. Poincaré	Ж. Ан. Пуанкаре
蒲丰 (布丰)	G. L. L. Buffon	Ж. Л. Л. Бюффон
普拉托	Jan von Plato	Ян Фон Плато
普拉特	Pratt	Пратт
普雷斯曼	I. L. Pressman	Э. Л. Пресман
普罗泰尔	Ph. Protter	Ф. Проттер
普罗霍罗夫	Yu. V. Prokhorov	Ю. В. Прохров

Q~R

奇斯佳科夫	W. P. Qisjiakov	В. П. Чистяков
乔	Y. S. Chow	Ю. Ш. Чао
切比雪夫 (切贝绍夫)	P. L. Chebyshev	П. Л. Чебышёв
齐格蒙特	A. Zygmund	А. Зигмунд
乔尔奇	A. Church	А. Чёрч
切萨罗	E. Cesàro	Э. Чеэаро
切尔诺夫		Чернов
茹尔边科		Журбенко

S

萨雷姆萨科夫	T. A. Saramsakov	Т. А. Сарымсаков
塞维治	I. R. Sevage	И. Р. Сэвидж
塞维治	L. J. Sevage	Л. И. Сэвидж
塞格	G. Szegö	Г. Сеге
沙利耶	C. L. Charlier	К. Л. Шарлье
绍德尔	A. Schauder	А. Шаудер
舍伊宁	O. W. Sherining	О. В. Шейнин
施利亚耶夫	A. N. Shiryaev	А. Н. Ширяев
施密特	E. Schimidt	Э. Шмидт
施瓦茨	L. Schwarz	Л. Шварц
斯莫卢霍夫斯基	M. Smoluchowski	М. Смолуховский
斯梯格列尔	S. M. Stigler	С. М. Стиглер
斯蒂尔切斯	T. J. Stieltjes	Т. И. Стилтьес
斯捷克洛夫	V. A. Steclov	В. А. Стеклов
斯科罗霍德	A. W. Skorokhod	А. В. Скороход
斯莱皮恩	P. Slepian	П. Слепян
斯鲁斯基	E. Slutsky	Е. Е. Слуцкий
斯米尔诺夫	N. V. Smirnov	Н. В. Смирнов
斯奈尔	J. L. Snell	Дж. Снелл
斯皮策	F. Spitzer	Ф. Спицер
斯坦因豪斯	H. D. Steihaus	Г. Д. Штейнгаус
斯特林	J. Stirling	Дж. Стирлинг
斯特扎克	D. R. Stirzaker	Д. Стирэакер
斯梯格列尔	M. S. Stigler	С. Стиглер
斯通	M. H. Stone	М. Г. Стоун
斯托特	W. F. Stout	В. Ф. Стоут
所罗门诺夫	R. Solomonov	Р. Соломонов

T

塔尔塔利亚	N. Tartalya	Н. Тарталья
泰切尔	H. Teicher	Г. Тейчер
汤斯凯	M. Donsker	М. Донскер
特普利茨	O. Toeplitz	О. Тёплиц
图尔恰	I. T ulcea	И. Тулча
托德汉特	I. Todhanter	Э. Тодхантер

托德亨特	I. Todhunter	И. Тодхатер
托洛佐娃	T. B. Tolozowa	Т. Б. Толозова

W

威曼	A. Wiman	А. Уиман
威伊达	I. Vajda	И. Вайда
韦布尔	W. Weibull	В. Вейбулл
维尔	J. Ville	Ж. Вилль
维尔斯特拉斯	K. T. W. Weirstrass	К. Т. В. Вейерштрасс
维纳	N. Weiner	Н. Винер
温策尔	A. D. Wentzel	А. Д. Вентцель
温特纳	A. Wintner	А. Винтнер
瓦尔德	A. Wald	А. Вальд
沃尔德	H. Wold	Г. Вольд
沃尔夫	R. Wolf	Р. Вольф
乌沙科夫	W. G. Ushakoff	В. Г. Ушаков
伍拉姆	Woollam	Улам

X

西拉日季诺夫	S. H. Sealaijinov	С. Х. Сиражддинов
西格蒙德	D. Sigmund	Д. Сигмунд
希尔伯特	D. Hilbert	Д. Гильберт
希奈	Y. G. Sinai	Я. Г. Синай
谢瓦斯契亚诺夫	B. A. Sevastyanov	Б. А. Севастьянов
辛钦	A. J. Khintchine	А. Я. Хичин
休伊特	E. Hewitt	Э. Хьюитт

Y

雅可比	G. G. J. Jaccbi	К. Г. Я. Якоби
亚当斯	J. G. Adams	Д. К. Адамс
亚格洛姆	A. M. Jaglom	А. М. Яглом
亚格洛姆	E. M. Jaglom	И. М. Яглом
亚历山大罗夫	A. S. Alexaderdrov	П. С. Алексадров
亚历山大罗娃	N. W. Alexaderdrova	Н. В. Алексадрова
延森	I. L. Iensen	И. Л. Иенсен
伊金格	Ezenger	Изинг

伊藤清	Î to Kiyosi	К. Ито
伊西哈尔	—	А. Исихар
易卜拉给莫夫	E. A. Ibragemov	И. А. Ибрагимов
尤什克维奇	A. P. Yushikaivici	А. П. Юшкевич

Z

扎克德	J. Jacod	Ж. Жакод
扎克斯	S. Zacks	Ш. Закс
钟开莱	Kai Lai Chung	Чжун Кай -лай
祖布科夫	A. M. Zubkoff	А. М. Зубков
佐洛塔廖夫	W. M. Zolotaileoff	В. М. Золотарёв

记号索引

常用数学符号

$\mathbb{R}=(-\infty,\infty)$ —— 实数的集合, 实直线, 一维欧几里得空间
$\mathbb{R}_+=[0,\infty)$
$\overline{\mathbb{R}}=[-\infty,\infty]$ —— 扩充实直线: $\overline{\mathbb{R}}=\mathbb{R}\bigcup\{-\infty\}\bigcup\{\infty\}$
$\overline{\mathbb{R}}_+=[0,\infty]$
$\mathbb{Q}$ —— 有理数的集合
$\mathbb{Q}_+=\mathbb{Q}\bigcap\mathbb{R}_+$
$\mathbb{R}^d$ ——d 维欧几里得空间
$\mathbb{N}$ —— 自然数: $\{0,1,2,\cdots\}$ 或 $\{1,2,\cdots\}$
$\mathbb{Z}$ —— 整数的集合: $\{0,\pm1,\pm2,\cdots\}$
$\mathbb{C}$ —— 复数的集合

$$(a,b)=\{x\in\overline{\mathbb{R}}:a<x<b\},[a,b]=\{x\in\overline{\mathbb{R}}:a\leqslant x\leqslant b\}$$

$$(a,b]=\{x\in\overline{\mathbb{R}}:a<x\leqslant b\},[a,b)=\{x\in\overline{\mathbb{R}}:a\leqslant x<b\}$$

$\inf X$ —— 集合 $X\subseteq\overline{\mathbb{R}}$ 的下界
$\sup X$ —— 集合 $X\subseteq\overline{\mathbb{R}}$ 的上界
$\inf\limits_{n\geqslant m}x_n$ —— 集合 $X=\{x_m,x_{m+1},\cdots\}$ 的下界
$\sup\limits_{n\geqslant m}x_n$ —— 集合 $X=\{x_m,x_{m+1},\cdots\}$ 的上界

如果$x_n\in\overline{\mathbb{R}},n\geqslant1$, 则

$$\liminf_{n\to\infty}x_n=\underline{\lim}_n x_n\equiv\sup_{m\geqslant1}\inf_{n\geqslant m}x_n,\quad \limsup_{n\to\infty}x_n=\overline{\lim}_n x_n\equiv\inf_{m\geqslant1}\sup_{n\geqslant m}x_n,$$

$$\lim x_n=x\Leftrightarrow\underline{\lim}x_n=\overline{\lim}x_n=x\Leftrightarrow\underline{\lim}x_n\geqslant x\geqslant\overline{\lim}x_n.$$

对 于 实 数

$x^+=\max(x,0),x^-=-\min(x,0)$
$x^\oplus=\begin{cases}x^{-1}, & 若\ x\neq0,\\ 0, & 若\ x=0\end{cases}$
$x\vee y=\max(x,y),x\wedge y=\min(x,y)$
$[x]$ 或 $\lfloor x\rfloor$ —— 不大于 x 的最大整数
$\lceil x\rceil$ —— 大于或等于 x 的最小整数
$\text{sign}\ x$ —— 实数 x 的符号:

$$\text{sign}\ x=\begin{cases}1, & 若\ x>0,\\ 0, & 若\ x=0,\\ -1, & 若\ x<0\end{cases}$$

(有时, 当 $x\geqslant0$ 时, 设 sign $=1$; 当 $x<0$ 时, 设 sign $=-1$)
$x_n\to x$, 其中 $n\in\{1,2,\cdots\}$, 表示 $\lim\limits_n x_n=x$
$x_n\uparrow$ 表示 $x_1\leqslant x_2\leqslant\cdots$; $x_n\uparrow x$ 表示 $x_n\uparrow$ 且 $\lim\limits_n x_n=x$
$x_n\downarrow$ 表示 $x_1\geqslant x_2\geqslant\cdots$; $x_n\downarrow x$ 表示 $x_n\downarrow$ 且 $\lim\limits_n x_n=x$

对于复数 $z=a+ib$, 其中 $a,b\in\mathbb{R}$, 而 $i=\sqrt{-1}$ 是虚单位

$\bar{z}=a-ib$ —— z 的共轭复数
$|z|$ —— z 的模 $(=\sqrt{a^2+b^2})$
$\text{Re}\ z$ 和 $\text{Im}\ z$ —— z 的实部和虚部: $\text{Re}\ z=a,\text{Im}\ z=b$

对于 $d-$ 维欧几里得空间 $\mathbb{R}^d$

$|x|$ —— $x=(x_1,\cdots,x_d)$ 的欧几里得范数, 即 $\sqrt{x_1^2+\cdots+x_d^2}$
$x\cdot y$ 或 (x,y) —— $x=(x_1,\cdots,x_d)$ 和 $y=(y_1,\cdots,y_d)$ 的数量积, 即 $x_1y_1+\cdots+x_dy_d$

集 合 论

$A_n\uparrow$ 表示 $A_1\subseteq A_2\subseteq\cdots$; $A_n\uparrow A$ 表示 $A_n\uparrow$ 且 $\bigcup A_n=A$
$A_n\downarrow$ 表示 $A_1\supseteq A_2\supseteq\cdots$; $A_n\downarrow A$ 表示 $A_n\downarrow$ 且 $\bigcap A_n=A$
$\limsup A_n$, 或 $\overline{\lim}A_n$, 或 $\bigcap\limits_{m\geqslant1}\left(\bigcup\limits_{n\geqslant m}A_n\right)$ 表示 {有无限多个 A_n} —— 属于无限多个集合 $A_n(n\geqslant1)$ 的点的集合
$\liminf A_n$, 或 $\underline{\lim}A_n$, 或 $\bigcup\limits_{m\geqslant1}\left(\bigcap\limits_{n\geqslant m}A_n\right)$ 表示 —— 属于所有 (仅可能有限个 A_n 除外)

集合 $A_n(n \geqslant 1)$ 的点的集合
I_A 或 $I(A)$ —— 集合 A 的示性函数
$\{\cdots\}$ —— 集合

数 学 符 号

$\ll$ —— 绝对连续
$\sim$ —— 等价
$\perp$ —— 垂直
$f = o(g)$ —— $\lim\left(\frac{f}{g}\right) = 0$
$f = O(g)$ —— $\limsup\left|\frac{f}{g}\right| < \infty$
$f \sim g$ —— $\lim\left(\frac{f}{g}\right) = 1$
$f \asymp g$ —— 比值 $\frac{f}{g}$ 自下与 0 分离, 且自上与 ∞ 分离
$f \circ g$ —— f 与 g 的复合
$f * g$ —— f 与 g 的卷积