Graduate Texts in Mathematics

GTM

Albert N. Shiryaev

Probability-2

Third Edition

Graduate Texts in Mathematics

Graduate Texts in Mathematics bridge the gap between passive study and creative understanding, offering graduate-level introductions to advanced topics in mathematics. The volumes are carefully written as teaching aids and highlight characteristic features of the theory. Although these books are frequently used as textbooks in graduate courses, they are also suitable for individual study.

More information about this series at http://www.springer.com/series/136

Albert N. Shiryaev

Probability-2

Third Edition

Translated by R.P. Boas† and D.M. Chibisov

Springer

Albert N. Shiryaev
Department of Probability Theory
and Mathematical Statistics
Steklov Mathematical Institute and
Lomonosov Moscow State University
Moscow, Russia

Translated by R.P. Boas† and D.M. Chibisov

ISSN 0072-5285 ISSN 2197-5612 (electronic)
Graduate Texts in Mathematics
ISBN 978-0-387-72207-8 ISBN 978-0-387-72208-5 (eBook)
https://doi.org/10.1007/978-0-387-72208-5

Library of Congress Control Number: 2018953349

Mathematics Subject Classification: 60Axx, 60Exx, 60Fxx, 60Gxx, 60Jxx, 62Lxx

Originally published in one volume.
Translation from the Russian language edition: *Veroiãtnost′ – 2* (fourth edition) by Albert N. Shiryaev

This Springer imprint is published by the registered company Springer Science+Business Media, LLC part of Springer Nature.
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Preface to the Third English Edition

The present edition is a translation of the fourth Russian edition of 2007, with the previous three published in 1980, 1989, and 2004. The English translations of the first two appeared in 1984 and 1996. The third and fourth Russian editions, extended compared to the second edition, were published in two volumes titled *Probability-1* and *Probability-2*. Accordingly, the present edition consists of two volumes: this Vol. 2, titled *Probability-2*, contains Chaps. 4–8, and Chaps. 1–3 are contained in Vol. 1, titled *Probability-1*, which was published in 2016.

This English translation of *Probability-2* was prepared by the editor and translator Prof. D. M. Chibisov, Professor of the Steklov Mathematical Institute. A former student of N. V. Smirnov, he has a broad view of probability and mathematical statistics, which enabled him not only to translate the parts that had not been translated before, but also to edit both the previous translation and the Russian text, making in them quite a number of corrections and amendments.

The author is sincerely grateful to D. M. Chibisov for the translation and scientific editing of this book.

Moscow, Russia A. Shiryaev
2018

Preface to the Fourth Russian Edition

A university course on probability and statistics usually consists of three one-semester parts: probability theory, random processes, and mathematical statistics.

The book *Probability-1* covered the material normally included in probability theory.

This book, *Probability-2*, contains extensive material for a course on random processes in the part dealing with *discrete time* processes, i.e., random sequences. (The reader interested in random processes with *continuous* time may refer to [12], which is closely related to *Probability-1* and *Probability-2*.)

Chapter 4, which opens this book, is focused mostly on the properties of sums of independent random variables that hold with *probability one* (e.g., "zero–one" laws, the strong law of large numbers, the law of the iterated logarithm).

Chapters 5 and 6 treat the strict and wide sense *stationary* random sequences.

In Chaps. 7 and 8, we set out random sequences that form *martingales* and *Markov chains*. These classes of processes enable us to study the behavior of various stochastic systems in the "future", depending on their "past" and "present" thanks to which these processes play a very important role in modern probability theory and its applications.

The book concludes with a Historical Review of the Development of Mathematical Theory of Probability.

Moscow, Russia A. Shiryaev
2003

Contents

Table of Contents of *Probability-1*

Chapter 4
Sequences and Sums of Independent Random Variables

1. Zero–One Laws

> The concept of mutual *independence* of two or more experiments holds, in a certain sense, a central position in the theory of probability. ... Historically, the independence of experiments and random variables represents the very mathematical concept that has given the theory of probability its peculiar stamp.
>
> A. N. Kolmogorov, *Foundations of Probability Theory* [50]

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

$$A_1 = \left\{ \omega \colon \sum_{n=1}^{\infty} \frac{\xi_n}{n} \text{ converges} \right\}$$

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 this 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

A. N. Shiryaev, *Probability-2*, Graduate Texts in Mathematics 95,
https://doi.org/10.1007/978-0-387-72208-5_1

$$\mathscr{X} = \bigcap_{n=1}^{\infty} \mathscr{F}_n^{\infty}.$$

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,

$$A_2 = \left\{ \sum_{n=1}^{\infty} \xi_n \text{ converges} \right\} \in \mathscr{X}.$$

The following events are also tail events:

$$A_3 = \{\xi_n \in I_n \text{ for infinitely many } n\} \quad (= \limsup \{\xi_n \in I_n\}),$$

where $I_n \in \mathscr{B}(R)$, $n \geq 1$;

$$A_4 = \left\{ \limsup_n \xi_n < \infty \right\};$$

$$A_5 = \left\{ \limsup_n \frac{\xi_1 + \cdots + \xi_n}{n} < \infty \right\};$$

$$A_6 = \left\{ \limsup_n \frac{\xi_1 + \cdots + \xi_n}{n} < c \right\};$$

$$A_7 = \left\{ \frac{S_n}{n} \text{ converges} \right\}, \quad \text{where} \quad S_n = \xi_1 + \cdots + \xi_n;$$

$$A_8 = \left\{ \limsup_n \frac{S_n}{\sqrt{2n \log n}} = 1 \right\}.$$

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 a value of 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*, which is a particular case of the following theorem.

Theorem 1 (Kolmogorov's Zero–One Law). *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables, and let $A \in \mathscr{X}$. Then $\mathsf{P}(A)$ can only have a value of 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 find (Problem 8, Sect. 3, Chap. 2, Vol. 1) sets $A_n \in \mathscr{F}_1^n$, $n \geq 1$, 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,

$$\mathsf{P}(A \cap A_n) = \mathsf{P}(A)\ \mathsf{P}(A_n),$$

for every $n \geq 1$. Hence (1) implies 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 $\mathsf{P}(S_n = 0\ i.o.) = 1$.*
(b) *If $p \neq \frac{1}{2}$, then $\mathsf{P}(S_n = 0\ i.o.) = 0$.*

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 a value of 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 (see (6), Sect. 2, Chap. 1, Vol. 1),

$$\mathsf{P}(B_{2n}) = C_{2n}^n p^n q^n \sim \frac{(4pq)^n}{\sqrt{\pi n}}$$

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

$$A = \left\{ \limsup \frac{S_n}{\sqrt{n}} = \infty,\ \liminf \frac{S_n}{\sqrt{n}} = -\infty \right\}$$

has probability 1 (since $A \subseteq B$).

Let $A_c = A_c' \cap A_c''$, where

$$A_c' = \left\{ \limsup_n \frac{S_n}{\sqrt{n}} \geq c \right\}, \quad A_c'' = \left\{ \liminf_n \frac{S_n}{\sqrt{n}} \leq -c \right\}.$$

Then $A_c \downarrow A$, $c \to \infty$, and all the events A, 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,

$$\mathsf{P}\left(\liminf_n \frac{S_n}{\sqrt{n}} \leq -c \right) = \mathsf{P}\left(\limsup_n \frac{S_n}{\sqrt{n}} \geq c \right) \geq \limsup_n \mathsf{P}\left(\frac{S_n}{\sqrt{n}} \geq c \right) > 0,$$

where the last inequality follows from the de Moivre–Laplace theorem (Sect. 6, Chap. 1, Vol. 1).

Thus, $\mathsf{P}(A_c) = 1$ for all $c > 0$, and therefore $\mathsf{P}(A) = \lim_{c\to\infty} \mathsf{P}(A_c) = 1$.

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, it can be shown (Problem 4) that every event in the tail σ-algebra $\mathscr{X}(S) = \bigcap \mathscr{F}_n^\infty(S)$, where $\mathscr{F}_n^\infty(S) = \sigma\{S_n, S_{n+1}, \ldots\}$, generated by $S_1 = \xi_1, S_2 = \xi_1 + \xi_2, \ldots$ is symmetric.

Theorem 3 (Hewitt–Savage Zero–One Law). *Let $\xi_1, \xi_2, \dots$ be a sequence of independent identically distributed random variables and $A = \{\xi \in B\}$ a symmetric event. Then $\mathsf{P}(A) = 0$ or 1.*

PROOF. Let $A = \{\xi \in B\}$ be a symmetric event. Choose sets $B_n \in \mathscr{B}(R^n)$ (see Problem 8 in Sect. 3, Chap. 2, Vol. 1) such that, for $A_n = \{\omega\colon (\xi_1, \dots, \xi_n) \in B_n\}$,

$$\mathsf{P}(A \triangle A_n) \to 0, \qquad n \to \infty. \tag{2}$$

Since the random variables $\xi_1, \xi_2, \dots$ are independent identically distributed, the probability distributions $P_\xi(B) = \mathsf{P}(\xi \in B)$ and $P_{\pi_n(\xi)}(B) = \mathsf{P}(\pi_n(\xi) \in B)$ coincide, where $\pi_n(\xi) = (\xi_{n+1}, \dots, \xi_{2n}, \xi_1, \dots, \xi_n, \xi_{2n+1}, \xi_{2n+2}, \dots)$ for all $n \geq 1$. 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),

$$\mathsf{P}(A \triangle A_n) = \mathsf{P}(A \triangle \pi_n(A_n)). \tag{5}$$

By (2), this implies that

$$\mathsf{P}(A \triangle (A_n \cap \pi_n(A_n))) \to 0, \quad n \to \infty. \tag{6}$$

Therefore we conclude from (2), (5), and (6) that

$$\begin{gathered} \mathsf{P}(A_n) \to \mathsf{P}(A), \quad \mathsf{P}(\pi_n(A_n)) \to \mathsf{P}(A), \\ \mathsf{P}(A_n \cap \pi_n(A_n)) \to \mathsf{P}(A). \end{gathered} \tag{7}$$

Moreover, since $\xi_1, \xi_2, \dots$ are independent,

$$\begin{aligned} \mathsf{P}(A_n \cap \pi_n(A_n)) &= \mathsf{P}\{(\xi_1, \dots, \xi_n) \in B_n, (\xi_{n+1}, \dots, \xi_{2n}) \in B_n\} \\ &= \mathsf{P}\{(\xi_1, \dots, \xi_n) \in B_n\} \cdot \mathsf{P}\{(\xi_{n+1}, \dots, \xi_{2n}) \in B_n\} \\ &= \mathsf{P}(A_n)\, \mathsf{P}(\pi_n(A_n)), \end{aligned}$$

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 $(\xi_n)_{n\geq 1}$ is a sequence of independent random variables, then the random variables $\limsup \xi_n$ and $\liminf \xi_n$ 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 $\limsup (S_n/b_n)$ and $\liminf (S_n/b_n)$ are degenerate.
4. Let $S_n = \xi_1 + \cdots + \xi_n$, $n \geq 1$, and

$$\mathscr{X}(S) = \bigcap \mathscr{F}_n^{\infty}(S), \quad \mathscr{F}_n^{\infty}(S) = \sigma\{S_n, S_{n+1}, \ldots\}.$$

   Show that every event in $\mathscr{X}(S)$ is symmetric.
5. Let (ξ_n) be a sequence of random variables. Show that $\{\limsup \xi_n > c\} \supseteq \limsup \{\xi_n > c\}$ for each $c > 0$.
6. Give an example of a tail event whose probability is strictly greater than 0 and less than 1.
7. Let $\xi_1, \xi_2, \ldots$ be independent random variables with $\mathsf{E}\,\xi_1 = 0$, $\mathsf{E}\,\xi_1^2 = 1$ that satisfy the central limit theorem ($\mathsf{P}\{S_n/\sqrt{n} \leq x\} \to \Phi(x)$, $x \in R$, where $S_n = \xi_1 + \cdots + \xi_n$). Show that

$$\limsup_{n\to\infty} n^{-1/2} S_n = +\infty \quad (\mathsf{P}\text{-a.s.}).$$

   In particular, this property holds for a sequence of independent identically distributed random variables (with $\mathsf{E}\,\xi_1 = 0$, $\mathsf{E}\,\xi_1^2 = 1$).
8. Let $\xi_1, \xi_2, \ldots$ be independent identically distributed random variables with $\mathsf{E}\,|\xi_1| > 0$. Show that

$$\limsup_{n\to\infty} \left| \sum_{k=1}^{n} \xi_k \right| = +\infty \quad (\mathsf{P}\text{-a.s.}).$$

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 this 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 ξ_n are uniformly bounded (i.e., $\mathsf{P}(|\xi_n| \le c) = 1$, $c < \infty$, $n \ge 1$), then the converse is true*: *the convergence of $\sum \xi_n$ with probability* 1 *implies* (1).

The proof depends on

Kolmogorov's Inequalities. (a) *Let $\xi_1, \xi_2, \dots, \xi_n$ be independent random variables with $\mathsf{E}\,\xi_i = 0$, $\mathsf{E}\,\xi_i^2 < \infty$, $1 \le i \le n$. Then for every $\varepsilon > 0$*

$$\mathsf{P}\left\{\max_{1\le k\le n} |S_k| \ge \varepsilon\right\} \le \frac{\mathsf{E}\,S_n^2}{\varepsilon^2}. \tag{2}$$

(b) *If also $\mathsf{P}(|\xi_i| \le c) = 1$, $1 \le i \le n$, then*

$$\mathsf{P}\left\{\max_{1\le k\le n} |S_k| \ge \varepsilon\right\} \ge 1 - \frac{(c+\varepsilon)^2}{\mathsf{E}\,S_n^2}. \tag{3}$$

PROOF. (a) Put

$$A = \{\max_{1\le k\le n} |S_k| \ge \varepsilon\},$$

$$A_k = \{|S_i| < \varepsilon,\ i = 1, \dots,\ k-1,\ |S_k| \ge \varepsilon\}, \quad 1 \le k \le n.$$

Then $A = \sum A_k$ and

$$\mathsf{E}\,S_n^2 \ge \mathsf{E}\,S_n^2 I_A = \sum \mathsf{E}\,S_n^2 I_{A_k}.$$

But

$$\begin{aligned}\mathsf{E}\,S_n^2 I_{A_k} &= \mathsf{E}\,(S_k + (\xi_{k+1} + \cdots + \xi_n))^2 I_{A_k} \\ &= \mathsf{E}\,S_k^2 I_{A_k} + 2\mathsf{E}\,S_k(\xi_{k+1} + \cdots + \xi_n) I_{A_k} + \mathsf{E}\,(\xi_{k+1} + \cdots + \xi_n)^2 I_{A_k} \\ &\ge \mathsf{E}\,S_k^2 I_{A_k},\end{aligned}$$

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 $\mathsf{E}\,\xi_i = 0$, $1 \le i \le n$. Hence

$$\mathsf{E}\,S_n^2 \ge \sum \mathsf{E}\,S_k^2 I_{A_k} \ge \varepsilon^2 \sum \mathsf{P}(A_k) = \varepsilon^2\,\mathsf{P}(A),$$

which proves the first inequality.

(b) To prove (3), we observe that

$$\mathsf{E}\,S_n^2 I_A = \mathsf{E}\,S_n^2 - \mathsf{E}\,S_n^2 I_{\bar{A}} \ge \mathsf{E}\,S_n^2 - \varepsilon^2\,\mathsf{P}(\bar{A}) = \mathsf{E}\,S_n^2 - \varepsilon^2 + \varepsilon^2\,\mathsf{P}(A). \tag{4}$$

On the other hand, on the set A_k,

$$|S_{k-1}| \le \varepsilon, \quad |S_k| \le |S_{k-1}| + |\xi_k| \le \varepsilon + c$$

and therefore

$$\begin{aligned}\mathsf{E}\, S_n^2 I_A &= \sum_k \mathsf{E}\, S_k^2 I_{A_k} + \sum_k \mathsf{E}\,(I_{A_k}(S_n - S_k)^2) \\ &\le (\varepsilon + c)^2 \sum_k \mathsf{P}(A_k) + \sum_{k=1}^{n} \mathsf{P}(A_k) \sum_{j=k+1}^{n} \mathsf{E}\,\xi_j^2 \\ &\le \mathsf{P}(A)\left[(\varepsilon + c)^2 + \sum_{j=1}^{n} \mathsf{E}\,\xi_j^2\right] = \mathsf{P}(A)[(\varepsilon + c)^2 + \mathsf{E}\, S_n^2]. \qquad (5)\end{aligned}$$

From (4) and (5) we obtain

$$\mathsf{P}(A) \ge \frac{\mathsf{E}\, S_n^2 - \varepsilon^2}{(\varepsilon + c)^2 + \mathsf{E}\, S_n^2 - \varepsilon^2} = 1 - \frac{(\varepsilon + c)^2}{(\varepsilon + c)^2 + \mathsf{E}\, S_n^2 - \varepsilon^2} \ge 1 - \frac{(\varepsilon + c)^2}{\mathsf{E}\, S_n^2}.$$

This completes the proof of (3).

□

PROOF OF THEOREM 1. (a) By Theorem 4 in Sect. 10, Chap. 2, Vol. 1, the sequence $(S_n)_{n\ge1}$ converges with probability 1 if and only if it is fundamental with probability 1. By Theorem 1 of Sect. 10, Chap. 2, Vol. 1, the sequence $(S_n)_{n\ge1}$, is fundamental (P-a.s.) if and only if

$$\mathsf{P}\left\{\sup_{k\ge1} |S_{n+k} - S_n| \ge \varepsilon\right\} \to 0, \quad n \to \infty. \qquad (6)$$

By (2),

$$\begin{aligned}\mathsf{P}\left\{\sup_{k\ge1} |S_{n+k} - S_n| \ge \varepsilon\right\} &= \lim_{N\to\infty} \mathsf{P}\left\{\max_{1\le k\le N} |S_{n+k} - S_n| \ge \varepsilon\right\} \\ &\le \lim_{N\to\infty} \frac{\sum_{k=n}^{n+N} \mathsf{E}\,\xi_k^2}{\varepsilon^2} = \frac{\sum_{k=n}^{\infty} \mathsf{E}\,\xi_k^2}{\varepsilon^2}.\end{aligned}$$

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,

$$\mathsf{P}\left\{\sup_{k\ge1} |S_{n+k} - S_n| \ge \varepsilon\right\} < \tfrac{1}{2}. \qquad (7)$$

By (3),

$$\mathsf{P}\left\{\sup_{k\ge1} |S_{n+k} - S_n| \ge \varepsilon\right\} \ge 1 - \frac{(c + \varepsilon)^2}{\sum_{k=n}^{\infty} \mathsf{E}\,\xi_k^2}.$$

Therefore if we suppose that $\sum_{k=1}^{\infty} \mathsf{E}\,\xi_k^2 = \infty$, then we obtain

$$\mathsf{P}\left\{\sup_{k\ge1} |S_{n+k} - S_n| \ge \varepsilon\right\} = 1,$$

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 $|a_n| \le c$, converges with probability 1 if and only if $\sum a_n^2 < \infty$.

2. Theorem 2 (Kolmogorov–Khinchin's 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 $\sum \operatorname{Var} \xi_n$ converge. If $\mathsf{P}(|\xi_n| \le c) = 1$* for some $c > 0$, *this condition is also necessary*.

PROOF. If $\sum \operatorname{Var} \xi_n < \infty$, 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 $\sum \xi_n$ also 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, $n \ge 1$. (When the original sample space is sufficiently rich, the existence of such a sequence follows from Corollary 1 to Theorem 1 of Sect. 9, Chap. 2, Vol. 1. 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 $\mathsf{P}(|\xi_n - \tilde{\xi}_n| \le 2c) = 1$. Therefore $\sum \operatorname{Var}(\xi_n - \tilde{\xi}_n) < \infty$ by Theorem 1 (b). In addition,

$$\sum \operatorname{Var} \xi_n = \tfrac{1}{2} \sum \operatorname{Var}(\xi_n - \tilde{\xi}_n) < \infty.$$

Consequently, by Theorem 1 (a), $\sum(\xi_n - \mathsf{E}\,\xi_n)$ converges with probability 1, and therefore $\sum \mathsf{E}\,\xi_n$ converges.

Thus, if $\sum \xi_n$ converges (P-a.s.) (and $\mathsf{P}(|\xi_n| \le c) = 1$, $n \ge 1$), then it follows that both $\sum \mathsf{E}\,\xi_n$ and $\sum \operatorname{Var} \xi_n$ 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 $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables. A necessary condition for the convergence of $\sum \xi_n$ with probability* 1 *is that the series*

$$\sum \mathsf{E}\,\xi_n^c, \qquad \sum \operatorname{Var} \xi_n^c, \qquad \sum \mathsf{P}\,(|\xi_n| \ge c)$$

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 $\sum \mathsf{P}(|\xi_n| \geq c) < \infty$, then $\sum I(|\xi_n| \geq c) < \infty$ with probability 1 by the Borel–Cantelli lemma. Consequently, $\xi_n = \xi_n^c$ 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 $\{|\xi_n| \geq c\}$ can occur (P-a.s.). Therefore $\sum I(|\xi_n| \geq c) < \infty$ (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 $\sum \operatorname{Var} \xi_n^c$ converge.

This completes the proof of the theorem.

□

Corollary. Let $\xi_1, \xi_2, \ldots$ *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

$$\sum \mathsf{E}\,\frac{\xi_n^2}{1 + |\xi_n|} < \infty \Leftrightarrow \sum \mathsf{E}\left[\xi_n^2 I(|\xi_n| \leq 1) + |\xi_n| I(|\xi_n| > 1)\right] < \infty.$$

Therefore if $\xi_n^1 = \xi_n I(|\xi_n| \leq 1)$, then we have

$$\sum \mathsf{E}\,(\xi_n^1)^2 < \infty.$$

Since $\mathsf{E}\,\xi_n = 0$, we have

$$\begin{aligned}\sum |\mathsf{E}\,\xi_n^1| &= \sum |\mathsf{E}\,\xi_n I(|\xi_n| \leq 1)| = \sum |\mathsf{E}\,\xi_n I(|\xi_n| > 1)| \\ &\leq \sum \mathsf{E}\,|\xi_n| I(|\xi_n| > 1) < \infty.\end{aligned}$$

Therefore both $\sum \mathsf{E}\,\xi_n^1$ and $\sum \operatorname{Var} \xi_n^1$ converge. Moreover, by Chebyshev's inequality,

$$\mathsf{P}\{|\xi_n| > 1\} = \mathsf{P}\{|\xi_n| I(|\xi_n| > 1) > 1\} \leq \mathsf{E}\,(|\xi_n| I(|\xi_n| > 1).$$

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. Then the following three conditions are equivalent:
   (a) The series $\sum \xi_n$ converges *with probability* 1;
   (b) The series $\sum \xi_n$ converges *in probability*;
   (b) The series $\sum \xi_n$ converges *in distribution.*
4. Give an example showing that in Theorems 1 and 2 we cannot dispense with the uniform boundedness condition ($\mathsf{P}\{|\xi_n| \le c\} = 1$ for some $c > 0$).
5. Let $\xi_1, \ldots, \xi_n$ be independent identically distributed random variables such that $\mathsf{E}\, \xi_1 = 0$, $\mathsf{E}\, \xi_1^2 < \infty$, and let $S_n = \xi_1 + \cdots + \xi_n$. Prove the following one-sided analog (A. V. Marshall) of *Kolmogorov's inequality* (2):

$$\mathsf{P}\Big\{ \max_{1 \le k \le n} S_k \ge \varepsilon \Big\} \le \frac{\mathsf{E}\, S_n^2}{\varepsilon^2 + \mathsf{E}\, S_n^2}.$$

6. Let $\xi_1, \xi_2, \ldots$ be a sequence of (arbitrary) random variables. Show that if $\sum_{n \ge 1} \mathsf{E}\, |\xi_n| < \infty$, then $\sum_{n \ge 1} \xi_n$ absolutely converges with probability 1.
7. Let $\xi_1, \xi_2, \ldots$ be independent random variables with a symmetric distribution. Show that

$$\mathsf{E}\left[\Big(\sum_n \xi_n \Big)^2 \wedge 1 \right] \le \sum_n \mathsf{E}\, (\xi_n^2 \wedge 1).$$

8. Let $\xi_1, \xi_2, \ldots$ be independent random variables with finite second moments. Show that $\sum \xi_n$ converges in L^2 if and only if $\sum \mathsf{E}\, \xi_n$ and $\sum \operatorname{Var} \xi_n$ converge.
9. Let $\xi_1, \xi_2, \ldots$ be independent random variables and the series $\sum \xi_n$ converge a.s. Show that the value of this series is independent of the order of its terms if and only if $\sum |\, \mathsf{E}\, (\xi_n; |\xi_n| \le 1)| < \infty$.
10. Let $\xi_1, \xi_2, \ldots$ be independent random variables with $\mathsf{E}\, \xi_n = 0$, $n \ge 1$, and

$$\sum_{n=1}^{\infty} \mathsf{E}\, [\xi_n^2 I(|\xi_n| \le 1) + |\xi_n| I(|\xi_n| > 1)] < \infty.$$

    Then $\sum_{n=1}^{\infty} \xi_n$ converges P-a.s.

11. Let $A_1, A_2, \dots$ be independent events with $\mathsf{P}(A_n) > 0$, $n \geq 1$, and $\sum_{n=1}^{\infty} \mathsf{P}(A_n) = \infty$. Show that

$$\sum_{j=1}^{n} I(A_j) \Big/ \sum_{j=1}^{n} \mathsf{P}(A_j) \to 1 \quad (\mathsf{P}\text{-a.s.}) \quad \text{as } n \to \infty.$$

12. Let $\xi_1, \xi_2, \dots$ be independent random variables with expectations $\mathsf{E}\,\xi_n$ and variances σ_n^2 such that $\lim_n \mathsf{E}\,\xi_n = c$ and $\sum_{n=1}^{\infty} \sigma_n^{-2} = \infty$. Show that in this case

$$\sum_{j=1}^{n} \frac{\xi_j}{\sigma_j^2} \Big/ \sum_{j=1}^{n} \frac{1}{\sigma_j^2} \to c \quad (\mathsf{P}\text{-a.s.}) \quad \text{as } n \to \infty.$$

13. Let $\xi_1, \xi_2, \dots, \xi_n$ be independent random variables with $\mathsf{E}\,\xi_i = 0$, $i \leq n$, and let $S_k = \xi_1 + \xi_2 + \cdots + \xi_k$. Prove Etemadi's inequality

$$\mathsf{P}\Big(\max_{1\leq k\leq n} |S_k| \geq 3\varepsilon\Big) \leq 3 \max_{1\leq k\leq n} \mathsf{P}(|S_k| \geq \varepsilon)$$

and deduce from it Kolmogorov's inequality (with an extra factor 27):

$$\mathsf{P}\Big(\max_{1\leq k\leq n} |S_k| \geq 3\varepsilon\Big) \leq \frac{27}{\varepsilon^2}\,\mathsf{E}\,S_n^2.$$

3. Strong Law of Large Numbers

1. Let $\xi_1, \xi_2, \dots$ be a sequence of independent random variables with finite second moments: $S_n = \xi_1 + \cdots + \xi_n$. By Problem 2 in Sect. 3, Chapter 3, Vol. 1, if the variances $\operatorname{Var} \xi_i$ are uniformly bounded, we have the (weak) law of large numbers:

$$\frac{S_n - \mathsf{E}\,S_n}{n} \xrightarrow{\mathsf{P}} 0, \quad n \to \infty. \tag{1}$$

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 $\xi_1, \xi_2, \dots$ be independent random variables with finite fourth moments, and let*

$$\mathsf{E}\,|\xi_n - \mathsf{E}\,\xi_n|^4 \leq C, \quad n \geq 1,$$

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 $n \geq 1$. By the corollary to Theorem 1, Sect. 10 of Chap. 2, Vol. 1, we will have $S_n/n \to 0$ (P-a.s.), provided that

$$\sum \mathsf{P}\left\{\left|\frac{S_n}{n}\right| \geq \varepsilon\right\} < \infty$$

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 $\mathsf{E}\,\xi_k = 0$, $k \geq 1$, we then obtain

$$\mathsf{E}\,S_n^4 = \sum_{i=1}^n \mathsf{E}\,\xi_i^4 + 6\sum_{i,j=1}^n \mathsf{E}\,\xi_i^2\,\mathsf{E}\,\xi_j^2 \leq nC + 6\sum_{\substack{i,j=1 \\ i<j}}^n \sqrt{\mathsf{E}\,\xi_i^4 \cdot \mathsf{E}\,\xi_j^4}$$
$$\leq nC + \frac{6n(n-1)}{2}C = (3n^2 - 2n)C < 3n^2C.$$

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.

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*

$$\sum \frac{\operatorname{Var}\xi_n}{b_n^2} < \infty. \tag{3}$$

Then

$$\frac{S_n - \mathsf{E}\,S_n}{b_n} \to 0 \quad (\mathsf{P}\text{-}a.s.). \tag{4}$$

In particular, if

$$\sum \frac{\operatorname{Var} \xi_n}{n^2} < \infty \tag{5}$$

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 3 in what follows, we need two lemmas.

Lemma 1 (Toeplitz). *Let $\{a_n\}$ be a sequence of nonnegative numbers, $b_n = \sum_{i=1}^{n} a_i$, $b_1 = a_1 > 0$, and $b_n \uparrow \infty$, $n \to \infty$. Let $\{x_n\}_{n\geq 1}$ 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 $n_1 > n_0$ such 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*

$$\frac{1}{b_n}\sum_{j=1}^{n} b_j x_j \to 0, \quad n \to \infty.$$

In particular, if $b_n = n$, $x_n = y_n/n$ *and* $\sum(y_n/n)$ *converges, then*

$$\frac{y_1 + \cdots + y_n}{n} \to 0, \quad n \to \infty. \tag{9}$$

PROOF. Let $b_0 = 0$, $S_0 = 0$, $S_n = \sum_{j=1}^{n} x_j$. Then (by summation by parts)

$$\sum_{j=1}^{n} b_j x_j = \sum_{j=1}^{n} b_j(S_j - S_{j-1}) = b_n S_n - b_0 S_0 - \sum_{j=1}^{n} S_{j-1}(b_j - b_{i-1})$$

and therefore (setting $a_j = b_j - b_{j-1}$),

$$\frac{1}{b_n}\sum_{j=1}^{n} b_j x_j = S_n - \frac{1}{b_n}\sum_{j=1}^{n} S_{j-1} a_j \to 0,$$

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 2. 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) and Theorem 1 of Sect. 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 $\mathsf{P}(\xi_n = 1) = \mathsf{P}(\xi_n = -1) = \frac{1}{2}$. Then, since $\sum[1/(n\log^2 n)] < \infty$, we have

$$\frac{S_n}{\sqrt{n}\log n} \to 0 \quad (\mathsf{P}\text{-a.s.}). \tag{10}$$

3. In the case where 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*

$$\frac{S_n}{n} \to m \quad (\mathsf{P}\text{-}a.s.) \tag{11}$$

where $m = \mathsf{E}\,\xi_1$.

For the proof we need the following lemma.

Lemma 3. *Let ξ be a nonnegative random variable. Then*

$$\sum_{n=1}^{\infty} \mathsf{P}(\xi \geq n) \leq \mathsf{E}\,\xi \leq 1 + \sum_{n=1}^{\infty} \mathsf{P}(\xi \geq n). \tag{12}$$

The *proof* consists of the following chain of inequalities:

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathsf{P}(\xi \geq n) &= \sum_{n=1}^{\infty} \sum_{k \geq n} \mathsf{P}(k \leq \xi < k+1) \\
&= \sum_{k=1}^{\infty} k\,\mathsf{P}(k \leq \xi < k+1) = \sum_{k=0}^{\infty} \mathsf{E}[kI(k \leq \xi < k+1)] \\
&\leq \sum_{k=0}^{\infty} \mathsf{E}[\xi I(k \leq \xi < k+1)] \\
&= \mathsf{E}\,\xi \leq \sum_{k=0}^{\infty} \mathsf{E}[(k+1)I(k \leq \xi < k+1)] \\
&= \sum_{k=0}^{\infty} (k+1)\,\mathsf{P}(k \leq \xi < k+1) \\
&= \sum_{n=1}^{\infty} \mathsf{P}(\xi \geq n) + \sum_{k=0}^{\infty} \mathsf{P}(k \leq \xi < k+1) = \sum_{n=1}^{\infty} \mathsf{P}(\xi \geq n) + 1.
\end{aligned}$$

(Or one can use formula (69) with $n = 1$ of Sect. 6, Chap. 2, Vol. 1.) □

PROOF OF THEOREM 3. By Lemma 3 and the Borel–Cantelli lemma (Sect. 10, Chap. 2, Vol. 1),

$$\begin{aligned}
\mathsf{E}\,|\xi_1| < \infty &\Leftrightarrow \sum \mathsf{P}\{|\xi_1| \geq n\} < \infty \\
&\Leftrightarrow \sum \mathsf{P}\{|\xi_n| \geq n\} < \infty \Leftrightarrow \mathsf{P}\{|\xi_n| \geq n \text{ i.o.}\} = 0.
\end{aligned}$$

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 $\xi_n \neq \tilde{\xi}_n$ for finitely many n (P-a.s.), and therefore $(\xi_1 + \cdots + \xi_n)/n \to 0$ (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' lemma,

$$\frac{1}{n}\sum_{k=1}^{n} \mathsf{E}\,\tilde{\xi}_k \to 0, \quad n \to \infty,$$

and consequently, $(\xi_1 + \cdots + \xi_n)/n \to 0$ (P-a.s.) as $n \to \infty$ if and only if

$$\frac{(\tilde{\xi}_1 - \mathsf{E}\,\tilde{\xi}_1) + \cdots + (\tilde{\xi}_n - \mathsf{E}\,\tilde{\xi}_n)}{n} \to 0 \quad (\mathsf{P}\text{-a.s.}). \tag{13}$$

Write $\bar{\xi}_n = \tilde{\xi}_n - \mathsf{E}\,\tilde{\xi}_n$. By Kronecker's lemma, (13) will be established if $\sum(\tilde{\xi}_n/n)$ converges (P-a.s.). In turn, by Theorem 1 of Sect. 2, this will follow if we show that, when $\mathsf{E}\,|\xi_1| < \infty$, the series $\sum(\mathrm{Var}\,\bar{\xi}_n/n^2)$ converges.

We have

$$\begin{aligned}
\sum \frac{\mathrm{Var}\,\bar{\xi}_n}{n^2} &\leq \sum_{n=1}^{\infty} \frac{\mathsf{E}\,\tilde{\xi}_n^2}{n^2} = \sum_{n=1}^{\infty} \frac{1}{n^2}\,\mathsf{E}\,[\xi_n I(|\xi_n| < n)]^2 \\
&= \sum_{n=1}^{\infty} \frac{1}{n^2}\,\mathsf{E}\,[\xi_1^2 I(|\xi_1| < n)] = \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{k=1}^{n} \mathsf{E}\,[\xi_1^2 I(k-1 \leq |\xi_1| < k)] \\
&= \sum_{k=1}^{\infty} \mathsf{E}\,[\xi_1^2 I(k-1 \leq |\xi_1| < k)] \cdot \sum_{n=k}^{\infty} \frac{1}{n^2} \\
&\leq 2\sum_{k=1}^{\infty} \frac{1}{k}\,\mathsf{E}\,[\xi_1^2 I(k-1 \leq |\xi_1| < k)] \\
&\leq 2\sum_{k=1}^{\infty} \mathsf{E}\,[|\xi_1| I(k-1 \leq |\xi_1| < k)] = 2\,\mathsf{E}\,|\xi_1| < \infty.
\end{aligned}$$

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

$$\frac{\xi_n}{n} = \frac{S_n}{n} - \left(\frac{n-1}{n}\right)\frac{S_{n-1}}{n-1} \to 0 \quad (\mathsf{P}\text{-a.s.})$$

and therefore $\mathsf{P}(|\xi_n| > n \text{ i.o.}) = 0$. By the Borel–Cantelli lemma (Sect. 10, Chap. 2, Vol. 1),

$$\sum \mathsf{P}(|\xi_1| > n) < \infty,$$

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 (9) 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.)

$$\liminf_n \frac{S_n}{n} \ge \liminf_n \frac{S_n^C}{n} = \mathsf{E}\,\xi_1 I(\xi_1 \le C).$$

But as $C \to \infty$,

$$\mathsf{E}\,\xi_1 I(\xi_1 \le C) \to \mathsf{E}\,\xi_1 = \infty,$$

and therefore $S_n/n \to +\infty$ (P-a.s.).

Remark 3. Theorem 3 asserts the convergence $\frac{S_n}{n} \to m$ (P-a.s.). Note that, besides the convergence almost surely (a.s.), in this case, the *convergence in the mean* $\left(\frac{S_n}{n} \xrightarrow{L^1} m\right)$ also holds, i.e., $\mathsf{E}\left|\frac{S_n}{n} - m\right| \to 0$, $n \to \infty$. This follows from the ergodic Theorem 3 of Sect. 3, Chap. 5. But in the case under consideration of *independent* identically distributed random variables $\xi_1, \xi_2, \ldots$ and $S_n = \xi_1 + \xi_2 + \cdots + \xi_n$, this can be proved directly (Problem 7) without invoking the ergodic theorem.

4. Let us give some applications of the strong law of large numbers.

EXAMPLE 2 (Application to number theory). Let $\Omega = [0, 1)$, let $\mathscr{B}$ be the sigma-algebra of Borel subsets of Ω, and let P be a Lebesgue measure on $[0, 1)$. Consider the binary expansions $\omega = 0.\omega_1\omega_2\ldots$ of numbers $\omega \in \Omega$ (with infinitely many 0s), and define random variables $\xi_1(\omega)$, $\xi_2(\omega), \ldots$ by putting $\xi_n(\omega) = \omega_n$. Since, for all $n \ge 1$ and all $x_1, \ldots, x_n$ taking a value 0 or 1,

$$\{\omega\colon \xi_1(\omega) = x_1, \ldots, \xi_n(\omega) = x_n\}$$
$$= \left\{\omega\colon \frac{x_1}{2} + \frac{x_2}{2^2} + \cdots + \frac{x_n}{2^n} \le \omega < \frac{x_1}{2} + \cdots + \frac{x_n}{2^n} + \frac{1}{2^n}\right\},$$

the P-measure of this set is $1/2^n$. It follows that $\xi_1, \xi_n, \dots$ 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 zeroes and ones in its binary expansion tends to* $\frac{1}{2}$, *i.e.*,

$$\frac{1}{n}\sum_{k=1}^{n} I(\xi_k = 1) \to \frac{1}{2} \quad (\mathsf{P}\text{-a.s.}).$$

EXAMPLE 3 (The Monte Carlo method). Let $f(x)$ be a continuous function defined on $[0,1]$, with values in $[0,1]$. The following idea is the foundation of the statistical method of calculating $\int_0^1 f(x)\,dx$ (the Monte Carlo method). Let $\xi_1, \eta_1, \xi_2, \eta_2, \dots$ be a sequence of independent random variables uniformly distributed on $[0,1]$. Put

$$\rho_i = \begin{cases} 1 & \text{if } f(\xi_i) > \eta_i, \\ 0 & \text{if } f(\xi_i) \le \eta_i. \end{cases}$$

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),

$$\frac{1}{n}\sum_{i=1}^{n} \rho_i \to \int_0^1 f(x)\,dx \quad (\mathsf{P}\text{-a.s.}).$$

Consequently, we can approximate an integral $\int_0^1 f(x)\,dx$ by taking a simulation consisting of pairs of random variables (ξ_i, η_i), $i \ge 1$, and then calculating ρ_i and $(1/n)\sum_{i=1}^{n} \rho_i$.

EXAMPLE 4 (The strong law of large numbers for a renewal process). Let $N = (N_t)_{t\ge 0}$ be a renewal process introduced in Subsection 4 of Sect. 9, Chap. 2, Vol. 1: $N_t = \sum_{n=1}^{\infty} I(T_n \le t)$, $T_n = \sigma_1 + \cdots + \sigma_n$, where $\sigma_1, \sigma_2, \dots$ is a sequence of independent identically distributed positive random variables. We assume now that $\mu = \mathsf{E}\,\sigma_1 < \infty$.

Under this condition, the process N satisfies the strong law of large numbers:

$$\frac{N_t}{t} \to \frac{1}{\mu} \quad (\mathsf{P}\text{-a.s.}), \quad t \to \infty. \tag{14}$$

For the proof, we observe that the assumption $N_t > 0$ and the fact that $T_{N_t} \le t < T_{N_t+1}$, $t \ge 0$, imply the inequalities

$$\frac{T_{N_t}}{N_t} \le \frac{t}{N_t} < \frac{T_{N_t+1}}{N_t+1}\Big(1 + \frac{1}{N_t}\Big). \tag{15}$$

Clearly, $N_t = N_t(\omega) \to \infty$ (P-a.s.) as $t \to \infty$. At the same time, by Theorem 3,

$$\frac{T_n(\omega)}{n} = \frac{\sigma_1(\omega) + \cdots + \sigma_n(\omega)}{n} \to \mu \quad (\mathsf{P}\text{-a.s.}), \quad n \to \infty.$$

Therefore we also have

$$\frac{T_{N_t(\omega)}(\omega)}{N_t(\omega)} \to \mu \quad (\mathsf{P}\text{-a.s.}), \quad n \to \infty,$$

and hence we see from (15) that there exists (P-a.s.) the limit $\lim_{t\to\infty} t/N_t$, which is equal to μ, which proves the strong law of large numbers (14).

5. Problems

1. Show that $\mathsf{E}\,\xi^2 < \infty$ if and only if $\sum_{n=1}^{\infty} n\,\mathsf{P}(|\xi| > n) < \infty$.
2. Supposing that $\xi_1, \xi_2, \ldots$ are independent identically distributed, show that if $\mathsf{E}\,|\xi_1|^\alpha < \infty$ for some α, $0 < \alpha < 1$, then $S_n/n^{1/\alpha} \to 0$ (P-a.s.), and if $\mathsf{E}\,|\xi_1|^\beta < \infty$ for some β, $1 \le \beta < 2$, then $(S_n - n\,\mathsf{E}\,\xi_1)/n^{1/\beta} \to 0$ (P-a.s.).
3. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables, and let $\mathsf{E}\,|\xi_1| = \infty$. Show that

$$\limsup_n \left| \frac{S_n}{n} - a_n \right| = \infty \quad (\mathsf{P}\text{-a.s.})$$

   for every sequence of constants $\{a_n\}$.
4. Are all rational numbers in $[0, 1)$ normal (in the sense of Example 3)?
5. Give an example of a sequence of independent random variables $\xi_1, \xi_2, \ldots$ such that the limit $\lim_{n\to\infty}(S_n/n)$ does exist in probability but does not exist with probability 1.
6. (N. Etemadi) Show that Theorem 3 remains valid with the independence condition of $\xi_1, \xi_2, \ldots$ replaced by their pairwise independence.
7. Show that under the conditions of Theorem 3, convergence in the mean (i.e., $\mathsf{E}\,|(S_n/n) - m| \to 0$, $n \to \infty$) also holds.
8. Let $\xi_1, \xi_2, \ldots$ be independent identically distributed random variables with $\mathsf{E}\,\xi_1^2 < \infty$. Show that

$$n\,\mathsf{P}\{|\xi_1| \ge \varepsilon\sqrt{n}\} \to 0 \quad \text{and} \quad \frac{1}{\sqrt{n}} \max_{k\le n} |\xi_k| \xrightarrow{\mathsf{P}} 0.$$

9. Consider *decimal* expansions of the numbers $\omega = 0.\omega_1\omega_2 \ldots$ in $[0, 1)$.
   (a) Carry over to this case the strong law of large numbers obtained in Subsection 4 for binary expansions.
   (b) Show that rational numbers are not normal (in the Borel sense), i.e., in their decimal expansion ($\xi_k(\omega) = \omega_k$, $k \ge 1$),

$$\frac{1}{n}\sum_{k=1}^{n} I(\xi_k(\omega) = i) \nrightarrow \frac{1}{10} \quad (\mathsf{P}\text{-a.s.}) \text{ for any } i = 0, 1, \ldots, 9.$$

(c) Show that the Champernowne number $\omega = 0.12345678910111213\ldots$, containing all the integers in a row, is *normal* (Example 3).

10. (a) Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables such that $\mathsf{P}\{\xi_n = \pm n^a\} = 1/2$. Show that this sequence satisfies the strong law of large numbers if and only if $a < 1/2$.
(b) Let $f = f(x)$ be a bounded continuous function on $(0, \infty)$. Show that, for any $a > 0$ and $x > 0$,

$$\lim_{n\to\infty} \sum_{k=1}^{\infty} f\Big(x + \frac{k}{n}\Big) e^{-an} \frac{(an)^k}{k!} = f(x + a).$$

11. Prove that *Kolmogorov's law of large numbers* (Theorem 3) can be restated in the following form: Let $\xi_1, \xi_2, \ldots$ be independent identically distributed random variables; then

$$\mathsf{E}\,|\xi_1| < \infty \iff n^{-1} S_n \to \mathsf{E}\,\xi_1 \quad (\mathsf{P}\text{-a.s.}),$$
$$\mathsf{E}\,|\xi_1| = \infty \iff \limsup n^{-1} S_n = +\infty \quad (\mathsf{P}\text{-a.s.}).$$

Prove that the first statement remains true with independence replaced by *pairwise* independence.

12. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables. Show that

$$\mathsf{E} \sup_n \Big|\frac{\xi_n}{n}\Big| < \infty \iff \mathsf{E}\,|\xi_1| \log^+ |\xi_1| < \infty.$$

13. Let $S_n = \xi_1 + \cdots + \xi_n$, $n \geq 1$, where $\xi_1, \xi_2, \ldots$ is a sequence of independent identically distributed random variables with $\mathsf{E}\,\xi_1 = 0$, $\mathsf{E}\,|\xi_1| > 0$. Show that $\limsup n^{-1/2} S_n = \infty$, $\liminf n^{-1/2} S_n = -\infty$ (P-a.s.).

14. Let $S_n = \xi_1 + \cdots + \xi_n$, $n \geq 1$, where $\xi_1, \xi_2, \ldots$ is a sequence of independent identically distributed random variables. Show that for any $\alpha \in (0, 1/2]$ one of the following properties holds:
(a) $n^{-\alpha} S_n \to \infty$ (P-a.s.);
(b) $n^{-\alpha} S_n \to -\infty$ (P-a.s.);
(c) $\limsup n^{-\alpha} S_n = \infty$, $\liminf n^{-\alpha} S_n = -\infty$ (P-a.s.).

15. Let $S_0 = 0$ and $S_n = \xi_1 + \cdots + \xi_n$, $n \geq 1$, where $\xi_1, \xi_2, \ldots$ is a sequence of independent identically distributed random variables. Show that:
(a) For any $\varepsilon > 0$

$$\sum_{n=1}^{\infty} \mathsf{P}\{|S_n| \geq n\varepsilon\} < \infty \iff \mathsf{E}\,\xi_1 = 0,\ \mathsf{E}\,\xi_1^2 < \infty;$$

(b) If $\mathsf{E}\,\xi_1 < 0$, then for $p > 1$

$$\mathsf{E}\Big(\sup_{n\geq 0} S_n\Big)^{p-1} < \infty \iff \mathsf{E}(\xi_1^+)^p < \infty;$$

(c) If $\mathsf{E}\,\xi_1 = 0$ and $1 < p \le 2$, then for a constant C_p

$$\sum_{n=1}^{\infty} \mathsf{P}\Big\{\max_{k\le n} S_k \ge n\Big\} \le C_p\,\mathsf{E}\,|\xi_1|^p, \quad \sum_{n=1}^{\infty} \mathsf{P}\Big\{\max_{k\le n} |S_k| \ge n\Big\} \le 2C_p\,\mathsf{E}\,|\xi_1|^p;$$

(d) If $\mathsf{E}\,\xi_1 = 0$, $\mathsf{E}\,\xi_1^2 < \infty$, and $M(\varepsilon) = \sup_{n\ge 0}(S_n - n\varepsilon)$, $\varepsilon > 0$, then

$$\lim_{\varepsilon\to 0} \varepsilon M(\varepsilon) = \sigma^2/2.$$

4. Law of the Iterated Logarithm

1. Let $\xi_1, \xi_2, \dots$ be a sequence of independent Bernoulli random variables with $\mathsf{P}(\xi_n = 1) = \mathsf{P}(\xi_n = -1) = \frac{1}{2}$; let $S_n = \xi_1 + \cdots + \xi_n$. It follows from the proof of Theorem 2, Sect. 1, that

$$\limsup \frac{S_n}{\sqrt{n}} = +\infty, \quad \liminf \frac{S_n}{\sqrt{n}} = -\infty, \tag{1}$$

with probability 1. On the other hand, by (10) of Sect. 3,

$$\frac{S_n}{\sqrt{n}\log n} \to 0 \quad (\mathsf{P}\text{-a.s.}). \tag{2}$$

Let us compare these results.

It follows from (1) that with probability 1 the paths of $(S_n)_{n\ge 1}$ intersect the "curves" $\pm\varepsilon\sqrt{n}$ *infinitely* often for any given $\varepsilon > 0$; 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 $(S_n)_{n\ge 1}$. The law of the iterated logarithm, which we present in what follows, improves this picture of the amplitude of the oscillations of $(S_n)_{n\ge 1}$.

Definition. We call a function $\varphi^* = \varphi^*(n)$, $n \ge 1$, *upper* (for $(S_n)_{n\ge 1}$) if, with probability 1, $S_n \le \varphi^*(n)$ for all n from some $n = n_0(\omega)$ on.

We call a function $\varphi_* = \varphi_*(n)$, $n \ge 1$, *lower* (for $(S_n)_{n\ge 1}$) 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

$$\left\{\limsup \frac{S_n}{\varphi(n)} \le 1\right\} = \left\{\lim_n \left[\sup_{m\ge n} \frac{S_m}{\varphi(m)}\right] \le 1\right\}$$

$$\Leftrightarrow \left\{ \sup_{m \geq n_1(\varepsilon,\omega)} \frac{S_m}{\varphi(m)} \leq 1+\varepsilon \text{ for any } \varepsilon > 0 \text{ and some } n_1(\varepsilon,\omega) \right\}$$
$$\Leftrightarrow \{S_m \leq (1+\varepsilon)\varphi(m) \text{ for any } \varepsilon > 0, \text{ from some } n_1(\varepsilon,\omega) \text{ on}\}. \quad (3)$$

In the same way,

$$\left\{ \limsup \frac{S_n}{\varphi(n)} \geq 1 \right\} = \left\{ \lim_n \left[\sup_{m \geq n} \frac{S_m}{\varphi(m)} \right] \geq 1 \right\}$$
$$\Leftrightarrow \left\{ \sup_{m \geq n_2(\varepsilon,\omega)} \frac{S_m}{\varphi(m)} \geq 1-\varepsilon \text{ for any } \varepsilon > 0 \text{ and some } n_2(\varepsilon,\omega) \right\}$$
$$\Leftrightarrow \left\{ \begin{array}{l} S_m \geq (1-\varepsilon)\varphi(m) \text{ for any } \varepsilon > 0 \text{ and} \\ \text{for } m \text{ larger than some } n_3(\varepsilon,\omega) \geq n_2(\varepsilon,\omega). \end{array} \right\} \quad (4)$$

It follows from (3) and (4) that to verify that each function $\varphi_\varepsilon^* = (1+\varepsilon)\varphi$, $\varepsilon > 0$, is upper, we must show that

$$\mathsf{P}\left\{ \limsup \frac{S_n}{\varphi(n)} \leq 1 \right\} = 1, \quad (5)$$

and to show that $\varphi_{*\varepsilon} = (1-\varepsilon)\varphi$, $\varepsilon > 0$, is lower, we must show that

$$\mathsf{P}\left\{ \limsup \frac{S_n}{\varphi(n)} \geq 1 \right\} = 1. \quad (6)$$

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*

$$\mathsf{P}\left\{ \limsup \frac{S_n}{\psi(n)} = 1 \right\} = 1, \quad (7)$$

where

$$\psi(n) = \sqrt{2\sigma^2 n \log\log n}. \quad (8)$$

For *uniformly bounded* random variables, the law of the iterated logarithm was established in 1924 by Khinchin [46]. In 1929 Kolmogorov [48] 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 [40].

Since the proof of Theorem 1 is rather complicated, we shall confine ourselves to the special case where 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 $a > 0$*

$$\mathsf{P}\left(\max_{1\le k\le n} S_k > a\right) \le 2\,\mathsf{P}(S_n > a). \tag{9}$$

PROOF. Let $A_k = \{S_i \le a,\ i \le k-1;\ S_k > a\}$, $A = \{\max_{1\le k\le n} S_k > a\}$, and $B = \{S_n > a\}$. Since $A_k \cap B \supseteq A_k \cap \{S_n \ge S_k\}$, we have

$$\begin{aligned}\mathsf{P}(A_k \cap B) &\ge \mathsf{P}(A_k \cap \{S_n \ge S_k\}) = \mathsf{P}(A_k)\,\mathsf{P}(S_n \ge S_k)\\ &= \mathsf{P}(A_k)\,\mathsf{P}(\xi_{k+1} + \cdots + \xi_n \ge 0).\end{aligned}$$

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 $\mathsf{P}(\xi_{k+1} + \cdots + \xi_n \ge 0) \ge \frac{1}{2}$, and therefore

$$\mathsf{P}(B) \ge \sum_{k=1}^{n} \mathsf{P}(A_k \cap B) \ge \frac{1}{2}\sum_{k=1}^{n} \mathsf{P}(A_k) = \frac{1}{2}\,\mathsf{P}(A),$$

which establishes (9) (cf. proof in Subsection 3 of Sect. 2, Chap. 8).

□

Lemma 2. *Let* $S_n \sim \mathcal{N}(0, \sigma^2(n))$, $\sigma^2(n) \uparrow \infty$, *and let* $a(n)$, $n \ge 1$, *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

$$\frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-y^2/2}\,dy \sim \frac{1}{\sqrt{2\pi}x}e^{-x^2/2}, \quad x \to \infty,$$

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 $k \ge k_0$, and k_0 is chosen so that $\log\log k_0$ 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 (Sect. 10, Chap. 2, Vol. 1).

From (11), (9), and (10) we find that

$$\begin{aligned}\mathsf{P}(A_k) &\le \mathsf{P}\{S_n > \lambda\psi(n_k) \text{ for some } n \in (n_k,\, n_{k+1})\}\\ &\le \mathsf{P}\{S_n > \lambda\psi(n_k) \text{ for some } n \le n_{k+1}\}\\ &\le 2\,\mathsf{P}\{S_{n_{k+1}} > \lambda\psi(n_k)\} \sim \frac{2\sqrt{n_k}}{\sqrt{2\pi}\lambda\psi(n_k)} \exp\{-\tfrac{1}{2}\lambda^2[\psi(n_k)/\sqrt{n_k}]^2\}\\ &\le C_1 \exp(-\lambda \log\log \lambda^k) \le C_2 e^{-\lambda \log k} = C_2 k^{-\lambda},\end{aligned}$$

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 $S_n \ge \lambda\psi(n)$ for infinitely many n.

Let us apply (5), which we just proved, to the sequence $(-S_n)_{n\ge 1}$. Then we find that for all n, with finitely many exceptions, $-S_n \le 2\psi(n)$ (P-a.s.). Consequently, if $n_k = N^k$, $N > 1$, then for sufficiently large k, either

$$S_{n_{k-1}} \ge -2\psi(n_{k-1})$$

or

$$S_{n_k} \ge Y_k - 2\psi(n_{k-1}), \tag{12}$$

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

$$\begin{aligned}\lambda'[2(N^k - N^{k-1})\log\log N^k]^{1/2} &> \lambda(2N^k \log\log N^k)^{1/2}\\ &\quad + 2(2N^{k-1}\log\log N^{k-1})^{1/2} \equiv \lambda\psi(N^k) + 2\psi(N^{k-1}).\end{aligned}$$

It is now enough to show that

$$Y_k > \lambda'[2(N^k - N^{k-1})\log\log N^k]^{1/2} \tag{14}$$

for infinitely many k. Evidently $Y_k \sim \mathcal{N}(0,\, N^k - N^{k-1})$. Therefore, by Lemma 2,

$$\begin{aligned}\mathsf{P}\{Y_k > \lambda'[2(N^k - N^{k-1})\log\log N^k]^{1/2}\} &\sim \frac{1}{\sqrt{2\pi}\lambda'(2\log\log N^k)^{1/2}} e^{-(\lambda')^2 \log\log N^k}\\ &\ge \frac{C_1}{(\log k)^{1/2}} k^{-(\lambda')^2} \ge \frac{C_2}{k\log k}.\end{aligned}$$

Since $\sum(1/k\log k) = \infty$, 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 (P-a.s.)

$$\liminf \frac{S_n}{\varphi(n)} = -1. \tag{15}$$

It follows from (7) and (15) that the law of the iterated logarithm can be put in the form

$$\mathsf{P}\left\{\limsup \frac{|S_n|}{\varphi(n)} = 1\right\} = 1. \tag{16}$$

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 $\varepsilon > 0$,

$$\begin{aligned}\mathsf{P}\{|S_n| \geq (1-\varepsilon)\psi(n) \text{ i.o.}\} &= 1,\\ \mathsf{P}\{|S_n| \geq (1+\varepsilon)\psi(n) \text{ i.o.}\} &= 0.\end{aligned}$$

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
$$\mathsf{P}\left\{\limsup \frac{\xi_n}{\sqrt{2\log n}} = 1\right\} = 1.$$

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 (regardless of λ)
$$\mathsf{P}\left\{\limsup \frac{\xi_n \log\log n}{\log n} = 1\right\} = 1.$$

3. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables with
$$\mathsf{E}\, e^{it\xi_1} = e^{-|t|^\alpha}, \quad 0 < \alpha < 2.$$
Show that
$$\mathsf{P}\left\{\limsup \left|\frac{S_n}{n^{1/\alpha}}\right|^{1/(\log\log n)} = e^{1/\alpha}\right\} = 1.$$

4. Establish the following generalization of (9). Let $\xi_1, \ldots, \xi_n$ be independent random variables, and let $S_0 = 0$, $S_k = \xi_1 + \cdots + \xi_k$. Then *Lévy's inequality*
$$\mathsf{P}\left\{\max_{0\leq k\leq n}[S_k + \mu(S_n - S_k)] > a\right\} \leq 2\,\mathsf{P}(S_n > a)$$

holds for every real $a > 0$, where $\mu(\xi)$ is the median of ξ, i.e., a constant such that

$$\mathsf{P}(\xi \geq \mu(\xi)) \geq \tfrac{1}{2}, \quad \mathsf{P}(\xi \leq \mu(\xi)) \geq \tfrac{1}{2}.$$

5. Let $\xi_1, \ldots, \xi_n$ be independent random variables, and let $S_0 = 0$, $S_k = \xi_1 + \cdots + \xi_k$. Prove that:
   (a) (In addition to Problem 4)

$$\mathsf{P}\Big\{\max_{1\leq k\leq n} |S_k + \mu(S_n - S_k)| \geq a\Big\} \leq 2\,\mathsf{P}\{|S_n| \geq a\},$$

   where $\mu(\xi)$ is the median of ξ;
   (b) If $\xi_1, \ldots, \xi_n$ are identically distributed and symmetric, then

$$1 - e^{-n\,\mathsf{P}\{|\xi_1|>x\}} \leq \mathsf{P}\Big\{\max_{1\leq k\leq n} |\xi_k| > x\Big\} \leq 2\,\mathsf{P}\{|S_n| > x\}.$$

6. Let $\xi_1, \ldots, \xi_n$ be independent random variables with $\mathsf{E}\,\xi_i = 0$, $1 \leq i \leq n$, and let $S_k = \xi_1 + \cdots + \xi_k$. Show that

$$\mathsf{P}\Big\{\max_{1\leq k\leq n} S_k > a\Big\} \leq 2\,\mathsf{P}\{S_n \geq a - \mathsf{E}\,|S_n|\} \quad \text{for } a > 0.$$

7. Let $\xi_1, \ldots, \xi_n$ be independent random variables such that $\mathsf{E}\,\xi_i = 0$, $\sigma^2 = \mathsf{E}\,\xi_i^2 < \infty$, and $|\xi_i| \leq C$ (P-a.s.), $i \leq n$. Let $S_n = \xi_1 + \cdots + \xi_n$. Show that

$$\mathsf{E}\,e^{xS_n} \leq \exp\{2^{-1}nx^2\sigma^2(1 + xC)\} \quad \text{for any } 0 \leq x \leq 2C^{-1}.$$

   Under the same assumptions, show that if (a_n) is a sequence of real numbers such that $a_n/\sqrt{n} \to \infty$, but $a_n = o(n)$, then for any $\varepsilon > 0$ and sufficiently large n

$$\mathsf{P}\{S_n > a_n\} > \exp\Big\{-\frac{a_n^2}{2n\sigma^2}(1 + \varepsilon)\Big\}.$$

8. Let $\xi_1, \ldots, \xi_n$ be independent random variables such that $\mathsf{E}\,\xi_i = 0$, $|\xi_i| \leq C$ (P-a.s.), $i \leq n$. Let $D_n = \sum_{i=1}^n \operatorname{Var}\xi_i$. Show that $S_n = \xi_1 + \cdots + \xi_n$ satisfies the inequality (Yu. V. Prohorov)

$$\mathsf{P}\{S_n \geq a\} \leq \exp\Big\{-\frac{a}{2C}\arcsin\frac{aC}{2D_n}\Big\}, \quad a \in R.$$

5. Probabilities of Large Deviations

1. Consider the Bernoulli scheme treated in Sect. 6, Chap. 1, Vol. 1. For this scheme, the de Moivre–Laplace theorem provides an approximation for the probabilities of *standard (normal)* deviations $|S_n - np| \geq \varepsilon\sqrt{n}$, i.e., deviations of S_n from the *central* value np by a quantity of order $\sqrt{n}$. In the same Sect. 6, Chap. 1, Vol. 1 we gave a

bound for probabilities of so-called *large* deviations $|S_n - np| \geq \varepsilon n$, i.e., deviations of S_n from np of order n:

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

(see (42) in Sect. 6, Chap. 1, Vol. 1). From this, of course, there follow the inequalities

$$\mathsf{P}\left\{\sup_{m\geq n}\left|\frac{S_m}{m} - p\right| \geq \varepsilon\right\} \leq \sum_{m\geq n}\mathsf{P}\left\{\left|\frac{S_m}{m} - p\right| \geq \varepsilon\right\} \leq \frac{2}{1 - e^{-2\varepsilon^2}}e^{-2n\varepsilon^2}, \tag{2}$$

which provide an idea 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 a more general situation, when $S_n = \xi_1 + \cdots + \xi_n$ is a sum of independent identically distributed random variables.

2. We say that a random variable ξ satisfies *Cramér's condition* if there is a neighborhood of zero such that for any λ in this neighborhood

$$\mathsf{E}\, e^{\lambda|\xi|} < \infty \tag{3}$$

(it can be shown that this condition is equivalent to an exponential decrease of $\mathsf{P}(|\xi| > x)$, as $x \to \infty$).

Let

$$\varphi(\lambda) = \mathsf{E}\, e^{\lambda\xi} \quad \text{and} \quad \psi(\lambda) = \log\varphi(\lambda). \tag{4}$$

On the interior of the set

$$\Lambda = \{\lambda \in R\colon \psi(\lambda) < \infty\} \tag{5}$$

the function $\psi(\lambda)$ is convex (from below) and infinitely differentiable. We also notice that

$$\psi(0) = 0, \quad \psi'(0) = m\ (= \mathsf{E}\,\xi), \quad \psi''(\lambda) \geq 0.$$

We define the function

$$H(a) = \sup_{\lambda}[a\lambda - \psi(\lambda)], \quad a \in R, \tag{6}$$

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 $a = m$.

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

$$\mathsf{P}\{\xi \le a\} \le e^{-H(a)}. \tag{8}$$

Consequently (cf. (42) in Sect. 6, Chap. 1, Vol. 1)

$$\mathsf{P}\{|\xi - m| \ge \varepsilon\} \le e^{-\min\{H(m-\varepsilon),H(m+\varepsilon)\}}. \tag{9}$$

If $\xi, \xi_1, \dots, \xi_n$ are independent identically distributed random variables that satisfy Cramér's condition (3), $S_n = \xi_1 + \cdots + \xi_n$, $\psi_n(\lambda) = \log \mathsf{E} \exp(\lambda S_n/n)$, $\psi(\lambda) = \log \mathsf{E}\, e^{\lambda\xi}$, and

$$H_n(a) = \sup_{\lambda}[a\lambda - \psi_n(\lambda)], \tag{10}$$

then

$$H_n(a) = nH(a) \ \big(= n \sup_{\lambda}[a\lambda - \psi(\lambda)]\big)$$

and inequalities (7), (8), and (9) assume the following forms:

$$\mathsf{P}\left\{\frac{S_n}{n} \ge a\right\} \le e^{-nH(a)}, \quad a > m, \tag{11}$$

$$\mathsf{P}\left\{\frac{S_n}{n} \le a\right\} \le e^{-nH(a)}, \quad a < m, \tag{12}$$

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - m\right| \ge \varepsilon\right\} \le 2e^{-\min\{H(m-\varepsilon),H(m+\varepsilon)\}\cdot n}. \tag{13}$$

Remark 1. Results of the type

$$\mathsf{P}\left\{\left|\frac{S_n}{n} - m\right| \ge \varepsilon\right\} \le ae^{-bn}, \tag{14}$$

where $a > 0$ and $b > 0$, indicate exponential convergence "adjusted" by the constants a and b. In the theory of *large deviations*, such results are often presented in a somewhat different, "cruder," form,

$$\limsup_n \frac{1}{n} \log \mathsf{P}\left\{\left|\frac{S_n}{n} - m\right| \ge \varepsilon\right\} < 0, \tag{15}$$

that clearly arises from (14) and refers to the "exponential" rate 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

$$\mathsf{P}\left\{\sup_{k\ge n}\frac{S_k}{k} > a\right\}, \quad \mathsf{P}\left\{\inf_{k\ge n}\frac{S_k}{k} < a\right\}, \quad \mathsf{P}\left\{\sup_{k\ge n}\left|\frac{S_k}{k} - m\right| > \varepsilon\right\},$$

which can provide definite bounds *on the rate of convergence in the strong law of large numbers*.

Let us suppose that the independent identically distributed nondegenerate random variables ξ, ξ_1, $\xi_2, \ldots$ satisfy Cramér's condition (3).

We fix $n \geq 1$ and set

$$\kappa = \min\left\{k \geq n\colon \frac{S_k}{k} > a\right\},$$

taking $\kappa = \infty$ if $S_k/k < a$ for all $k \geq n$.

In addition, let a and $\lambda > 0$ satisfy

$$\lambda a - \log \varphi(\lambda) \geq 0. \tag{16}$$

Then

$$\begin{aligned}
\mathsf{P}\left\{\sup_{k\geq n} \frac{S_k}{k} > a\right\} &= \mathsf{P}\left\{\bigcup_{k\geq n}\left\{\frac{S_k}{k} > a\right\}\right\} \\
&= \mathsf{P}\left\{\frac{S_\kappa}{\kappa} > a,\ \kappa < \infty\right\} = \mathsf{P}\{e^{\lambda S_\kappa} > e^{\lambda a \kappa}, \kappa < \infty\} \\
&= \mathsf{P}\{e^{\lambda S_\kappa - \kappa \log \varphi(\lambda)} > e^{\kappa(\lambda a - \log \varphi(\lambda))},\ \kappa < \infty\} \\
&\leq \mathsf{P}\{e^{\lambda S_\kappa - \kappa \log \varphi(\lambda)} > e^{n(\lambda a - \log \varphi(\lambda))},\ \kappa < \infty\} \\
&\leq \mathsf{P}\left\{\sup_{k\geq n} e^{\lambda S_k - k \log \varphi(\lambda)} \geq e^{n(\lambda a - \log \varphi(\lambda)}\right\}.
\end{aligned} \tag{17}$$

To take the final step, we notice that the sequence of random variables

$$e^{\lambda S_k - k \log \varphi(\lambda)}, \quad k \geq 1,$$

with respect to the flow of σ-algebras $\mathscr{F}_k = \sigma\{\xi_1, \ldots, \xi_k\}$, $k \geq 1$, forms a *martingale*. (For more details, see Chap. 7 and, in particular, Example 2 in Sect. 1 therein.) Then it follows from inequality (8) in Sect. 3, Chap. 7, that

$$\mathsf{P}\left\{\sup_{k\geq n} e^{\lambda S_k - k \log \varphi(\lambda)} \geq e^{n(\lambda a - \log \varphi(\lambda))}\right\} \leq e^{-n(\lambda a - \log \varphi(\lambda))},$$

and consequently (assuming (16)) we obtain the inequality

$$\mathsf{P}\left\{\sup_{k\geq n} \frac{S_k}{k} > a\right\} \leq e^{-n(\lambda a - \log \varphi(\lambda))}. \tag{18}$$

Let $a > m$. Since the function $f(\lambda) = a\lambda - \log \varphi(\lambda)$ 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 $a > m$, then*

$$\mathsf{P}\left\{\sup_{k\geq n} \frac{S_k}{k} > a\right\} \leq e^{-n \sup_{\lambda>0}[\lambda a - \log \varphi(\lambda)]} = e^{-nH(a)}. \tag{19}$$

Similarly, *if* $a < m$, *then*

$$\mathsf{P}\left\{\sup_{k\geq n}\frac{S_k}{k} < a\right\} \leq e^{-n\sup_{\lambda<0}[\lambda a-\log\varphi(\lambda)]} = e^{-nH(a)}. \tag{20}$$

From (19) and (20) we obtain

$$\mathsf{P}\left\{\sup_{k\geq n}\left|\frac{S_k}{k} - m\right| > \varepsilon\right\} \leq 2e^{-\min[H(m-\varepsilon),H(m+\varepsilon)]\cdot n}. \tag{21}$$

Remark 2. The fact that the right-hand sides of inequalities (11) and (19) are the same leads us to suspect that this situation is not accidental. In fact, this expectation is concealed in the property that the sequences $(S_k/k)_{n\leq k\leq N}$ form, for every $n \leq N$, *reversed martingales* (see Problem 5 in Sect. 1, Chap. 7, and Example 4 in Sect. 11, Chap. 1, Vol. 1).

2. Problems

1. Carry out the proof of inequalities (8) and (20).
2. Verify that under condition (3), the function $\psi(\lambda)$ is convex (from below) on the interior of the set Λ (see (5)) (and *strictly* convex provided ξ is nondegenerate) and infinitely differentiable.
3. Assuming that ξ is nondegenerate, prove that the function $H(a)$ is differentiable on the whole real line and is convex (from below).
4. Prove the following inversion formula for Cramér's transform:

$$\psi(\lambda) = \sup_a[\lambda a - H(a)]$$

(for all λ, except, possibly, the endpoints of the set $\Lambda = \{\lambda\colon \psi(\lambda) < \infty\}$).

5. Let $S_n = \xi_1 + \cdots + \xi_n$, where $\xi_1, \ldots, \xi_n$, $n \geq 1$, are independent identically distributed simple random variables with $\mathsf{E}\,\xi_1 < 0$, $\mathsf{P}\{\xi_1 > 0\} > 0$. Let $\varphi(\lambda) = \mathsf{E}\,e^{\lambda\xi_1}$ and $\inf_\lambda \varphi(\lambda) = \rho$ $(0 < \rho < 1)$.
Show that the following result (*Chernoff's theorem*) holds:

$$\lim\frac{1}{n}\log\mathsf{P}\{S_n \geq 0\} = \log\rho. \tag{22}$$

6. Using (22), prove that in the Bernoulli scheme ($\mathsf{P}\{\xi_1 = 1\} = p$, $\mathsf{P}\{\xi_1 = 0\} = q$)

$$\lim\frac{1}{n}\log\mathsf{P}\{S_n \geq nx\} = -H(x), \tag{23}$$

for $p < x < 1$, where (cf. notation in Sect. 6, Chap. 1, Vol. 1)

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

7. Let $S_n = \xi_1 + \cdots + \xi_n$, $n \geq 1$, where $\xi_1, \xi_2, \ldots$ are independent identically distributed random variables with $\mathsf{E}\,\xi_1 = 0$, $\operatorname{Var}\xi_1 = 1$. Let $(x_n)_{n\geq 1}$ be a sequence such that $x_n \to \infty$ and $\frac{x_n}{\sqrt{n}} \to 0$ as $n \to \infty$. Show that

$$\mathsf{P}\{S_n \geq x_n\sqrt{n}\} = e^{-\frac{x_n^2}{2}(1+y_n)},$$

where $y_n \to 0$, $n \to \infty$.

8. Derive from (23) that in the Bernoulli case ($\mathsf{P}\{\xi_1 = 1\} = p$, $\mathsf{P}\{\xi_1 = 0\} = q$) we have:

   (a) For $p < x < 1$ and $x_n = n(x - p)$,

$$\mathsf{P}\{S_n \geq np + x_n\} = \exp\Big\{-nH\Big(p + \frac{x_n}{n}\Big)(1 + o(1))\Big\}; \tag{24}$$

   (b) For $x_n = a_n\sqrt{npq}$ with $a_n \to \infty$, $a_n/\sqrt{n} \to 0$,

$$\mathsf{P}\{S_n \geq np + x_n\} = \exp\Big\{-\frac{x_n^2}{2npq}(1 + o(1))\Big\}. \tag{25}$$

   Compare (24) with (25) and both of them with the corresponding results in Sect. 6 of Chap. 1, Vol. 1.

Chapter 5
Stationary (Strict Sense) Random Sequences and Ergodic Theory

> In the strict sense, the theory [of stationary stochastic processes] can be stated outside the framework of probability theory as the theory of one-parameter groups of transformations of a measure space that preserve the measure; this theory is very close to the general theory of dynamical systems and to ergodic theory.
>
> Encyclopaedia of Mathematics [42, Vol. 8, p. 479].

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 $k \geq 1$:

$$\mathsf{P}((\xi_1, \xi_2, \ldots) \in B) = \mathsf{P}((\xi_{k+1}, \xi_{k+2}, \ldots) \in B), \quad B \in \mathscr{B}(R^\infty).$$

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 $\eta_k = g(\xi_k, \xi_{k+1}, \ldots, \xi_{k+n-1})$.

If $\xi = (\xi_1, \xi_2, \ldots)$ is a sequence of *independent identically* distributed random variables with $\mathsf{E}\,|\xi_1| < \infty$ and $\mathsf{E}\,\xi_1 = m$, then the strong law of large numbers tells us that, with probability 1,

$$\frac{\xi_1 + \cdots + \xi_n}{n} \to m, \quad n \to \infty.$$

A. N. Shiryaev, *Probability-2*, Graduate Texts in Mathematics 95,
https://doi.org/10.1007/978-0-387-72208-5_2

In 1931, Birkhoff [6] obtained a remarkable generalization of this fact, which was stated as a theorem of *statistical mechanics* dealing with the behavior of the "relative residence time" of dynamical systems described by differential equations admitting an integral invariant ("conservative systems"). Soon after, in 1932, Khinchin [47] obtained an extension of Birkhoff's theorem to a more general case of "stationary motions of a multidimensional space within itself preserving the measure of a set."

The following presentation of Birkhoff's and Khinchin's results will combine the ideas of the theory of "dynamical systems" and the theory of "stationary in a strict sense random sequences."

In this presentation we will primarily concentrate on the "ergodic" results of these theories.

2. Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a (complete) probability space.

Definition 2. A transformation T of Ω into itself is *measurable* if, for every $A \in \mathscr{F}$,

$$T^{-1}A = \{\omega\colon T\omega \in A\} \in \mathscr{F}.$$

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. Set $\xi_n(\omega) = \xi_1(T^{n-1}\omega)$, $n \geq 2$, and consider the sequence $\xi = (\xi_1, \xi_2, \ldots)$. We claim that this sequence is stationary.

In fact, let $A = \{\omega\colon \xi \in B\}$ and $A_1 = \{\omega\colon \theta_1\xi \in B\}$, where $B \in \mathscr{B}(R^\infty)$. Then $\omega \in A_1$ if and only if $T\omega \in A$, i.e., $A_1 = T^{-1}A$. But $\mathsf{P}(T^{-1}A) = \mathsf{P}(A)$, hence $\mathsf{P}(A_1) = \mathsf{P}(A)$. Similarly, $\mathsf{P}(A_k) = \mathsf{P}(A)$ for any $A_k = \{\omega\colon \theta_k\xi \in B\}$, $k \geq 2$.

Thus we can use measure-preserving transformations to construct stationary (in strict sense) random sequences.

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 $\xi = \{\xi_1(\omega), \xi_2(\omega), \ldots\}$.

In fact, take $\tilde{\Omega}$ to be the coordinate space R^∞, and set $\tilde{\mathscr{F}} = \mathscr{B}(R^\infty)$, $\tilde{\mathsf{P}} = P_\xi$, 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 $\tilde{\omega} = (x_1, x_2, \ldots)$, set

$$\tilde{\xi}_1(\tilde{\omega}) = x_1, \quad \tilde{\xi}_n(\tilde{\omega}) = \tilde{\xi}_1(\tilde{T}^{n-1}\tilde{\omega}), \quad n \geq 2.$$

Now let $A = \{\tilde{\omega}\colon (x_1, \ldots, x_k) \in B\}$, $B \in \mathscr{B}(R^k)$, and

$$\tilde{T}^{-1}A = \{\tilde{\omega}\colon (x_2, \ldots, x_{k+1}) \in B\}.$$

Then the property of being stationary means that

$$\tilde{\mathsf{P}}(A) = \mathsf{P}\{\omega\colon (\xi_1, \ldots, \xi_k) \in B\} = \mathsf{P}\{\omega\colon (\xi_2, \ldots, \xi_{k+1}) \in B\} = \tilde{\mathsf{P}}(\tilde{T}^{-1}A),$$

i.e., $\tilde{T}$ is a measure-preserving transformation. Since $\tilde{\mathsf{P}}\{\tilde{\omega}\colon (\tilde{\xi}_1, \ldots, \tilde{\xi}_k) \in B\} = \mathsf{P}\{\omega\colon (\xi_1, \ldots, \xi_k) \in B\}$ for every k, it follows that ξ and $\tilde{\xi}$ have the same distribution.

What follows 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 $\mathsf{P}(\omega_i) = 1/n$, then the transformation T is measure-preserving.

EXAMPLE 2. If $\Omega = [0, 1)$, $\mathscr{F} = \mathscr{B}([0, 1))$, P is the Lebesgue measure, $\lambda \in [0, 1)$, then $Tx = (x + \lambda) \bmod 1$ is a measure-preserving transformation.

Let us consider the physical hypotheses that lead to the consideration of measure-preserving transformations.

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 $T^{-1}A = \{\omega\colon T\omega \in A\}$ 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 the Lebesgue measure.)

3. One of the earliest results on measure-preserving transformations was *Poincaré's recurrence theorem* [63].

Theorem 1. *Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space, let T be a measure-preserving transformation, and let $A \in \mathscr{F}$. Then, for almost every point $\omega \in A$, we have $T^n\omega \in A$ for infinitely many $n \geq 1$.*

PROOF. Let $C = \{\omega \in A\colon T^n\omega \notin A \text{ for all } n \geq 1\}$. 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. But $\sum_{n=0}^{\infty} \mathsf{P}(C) = \sum_{n=0}^{\infty} \mathsf{P}(T^{-n}C) \leq \mathsf{P}(\Omega) = 1$, and consequently $\mathsf{P}(C) = 0$. Therefore, for almost every point $\omega \in A$, for at least one $n \geq 1$, we have $T^n\omega \in A$. We will show that, consequently, $T^n\omega \in A$ for infinitely many n.

Let us apply the preceding result to T^k, $k \geq 1$. Then for every $\omega \in A \setminus N$, where N is a set of probability zero, which is the union of the corresponding sets related 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 $\{\omega\colon \xi(\omega) > 0\}$.

In fact, let $A_n = \{\omega\colon \xi(\omega) \geq 1/n\}$. Then, by 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 $\mathsf{E}\,\xi(\omega)$ exists. Show that $\mathsf{E}\,\xi(\omega) = \mathsf{E}\,\xi(T\omega)$.
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.
4. Let Ω be the set of all sequences $\omega = (\dots, \omega_{-1}, \omega_0, \omega_1, \dots)$ of real numbers, $\mathscr{F}$ the σ-algebra generated by measurable cylinders $\{\omega\colon (\omega_k, \dots, \omega_{k+n-1}) \in B_n\}$, where $n = 1, 2, \dots$, $k = 0, \pm 1, \pm 2, \dots$, and $B_n \in \mathscr{B}(R^n)$. Let P be a probability measure on $(\Omega, \mathscr{F})$, and let T be the two-sided transformation defined by

   $$T(\dots, \omega_{-1}, \omega_0, \omega_1, \dots) = (\dots, \omega_0, \omega_1, \omega_2, \dots).$$

   Show that T is measure-preserving if and only if

   $$\mathsf{P}\{\omega\colon (\omega_0, \dots, \omega_{n-1}) \in B_n\} = \mathsf{P}\{\omega\colon (\omega_k, \dots, \omega_{k+n-1}) \in B_n\}$$

   for all $n = 1, 2, \dots$, $k = 0, \pm 1, \pm 2, \dots$, and $B_n \in \mathscr{B}(R^n)$.
5. Let $\xi_0, \xi_1, \dots$ be a stationary sequence of random elements taking values in a Borel space S (see Definition 9 in Sect. 7, Chap. 2, Vol. 1). Show that one can construct (maybe on an extended probability space) random elements $\xi_{-1}, \xi_{-2}, \dots$ with values in S such that the two-sided sequence $\dots, \xi_{-1}, \xi_0, \xi_1, \dots$ is stationary.
6. Let T be a measurable transformation on $(\Omega, \mathscr{F}, \mathsf{P})$, and let $\mathscr{E}$ be a π-system of subsets of Ω that generates $\mathscr{F}$ (i.e., $\pi(\mathscr{E}) = \mathscr{F}$). Prove that if the equality $\mathsf{P}(T^{-1}A) = \mathsf{P}(A)$ holds for all $A \in \mathscr{E}$, then it holds also for all $A \in \mathscr{F}$ $(= \pi(\mathscr{E}))$.
7. Let T be a measure-preserving transformation on $(\Omega, \mathscr{F}, \mathsf{P})$, and let $\mathscr{G}$ be a sub-σ-algebra of $\mathscr{F}$. Show that for any $A \in \mathscr{F}$

   $$\mathsf{P}(A \,|\, \mathscr{G})(T\omega) = \mathsf{P}(T^{-1}A \,|\, T^{-1}\mathscr{G})(\omega) \quad (\mathsf{P}\text{-a.s.}). \tag{1}$$

   In particular, let $\Omega = R^\infty$ be the space of numerical sequences $\omega = (\omega_0, \omega_1, \dots)$ and $\xi_k(\omega) = \omega_k$. Let T be the shift transformation $T(\omega_0, \omega_1, \dots) = (\omega_1, \omega_2, \dots)$

(in other words, if $\xi_k(\omega) = \omega_k$, then $\xi_k(T\omega) = \omega_{k+1}$). Then (1) becomes

$$\mathsf{P}(A \mid \xi_n)(T\omega) = \mathsf{P}(T^{-1}A \mid \xi_{n+1})(\omega) \quad (\mathsf{P}\text{-a.s.}).$$

2. Ergodicity and Mixing

1. In this 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 $\eta = \eta(\omega)$ is *invariant* (or *almost invariant*) if $\eta(\omega) = \eta(T\omega)$ 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 an almost invariant set, then there is an invariant set B such that $\mathsf{P}(A \triangle B) = 0$.*

PROOF. Let $B = \limsup T^{-n}A$. Then $T^{-1}B = \limsup T^{-(n+1)}A = B$, 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, by 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 $\mathscr{I} \subseteq \mathscr{I}^*$.

□

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., $\xi(\omega) = \xi(T\omega)$ (P-a.s.). Then for every $c \in R$ we have $A_c = \{\omega\colon \xi(\omega) \le c\} \in \mathscr{I}^*$, and then $\mathsf{P}(A_c) = 0$ or 1 by Lemma 2. Let $C = \sup\{c\colon \mathsf{P}(A_c) = 0\}$. Since $A_c \uparrow \Omega$ as $c \uparrow \infty$ and $A_c \downarrow \varnothing$ as $c \downarrow -\infty$, we have $|C| < \infty$. Then

$$\mathsf{P}\{\omega\colon \xi(\omega) < C\} = \mathsf{P}\left\{\bigcup_{n=1}^{\infty}\left\{\xi(\omega) \le C - \frac{1}{n}\right\}\right\} = 0.$$

And, similarly, $\mathsf{P}\{\omega\colon \xi(\omega) > C\} = 0$. Consequently, $\mathsf{P}\{\omega\colon \xi(\omega) = C\} = 1$.

(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 1. The conclusion of the theorem remains valid in the case where "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 the 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 an invariant 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, Sect. 1), we have (Problem 1, Sect. 1) that, since the random variable ξ is invariant,

$$\begin{aligned} c_n = \mathsf{E}\,\xi(\omega)e^{-2\pi i n\xi(\omega)} &= \mathsf{E}\,\xi(T\omega)e^{-2\pi i nT\omega} = e^{-2\pi i n\lambda}\,\mathsf{E}\,\xi(T\omega)e^{-2\pi i n\omega} \\ &= e^{-2\pi i n\lambda}\,\mathsf{E}\,\xi(\omega)e^{-2\pi i n\omega} = c_n e^{-2\pi i n\lambda}. \end{aligned}$$

Thus, $c_n(1 - e^{-2\pi i n\lambda}) = 0$. 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

$$A = \bigcup_{k=0}^{2m-2}\left\{\omega\colon \frac{2k}{2m} \le \omega < \frac{2k+1}{2m}\right\}.$$

It is clear that this set is invariant; but $\mathsf{P}(A) = \frac{1}{2}$. Consequently, T is not ergodic.

2. Definition 4. A measure-preserving transformation is *mixing* (or has the mixing property) if, for all A and $B \in \mathscr{F}$,

$$\lim_{n\to\infty} \mathsf{P}(A \cap T^{-n}B) = \mathsf{P}(A)\,\mathsf{P}(B). \tag{1}$$

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 $B = T^{-n}B$, $n \geq 1$, and therefore

$$\mathsf{P}(A \cap T^{-n}B) = \mathsf{P}(A \cap B)$$

for all $n \geq 1$. 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 $\mathsf{P}(T^{-1}A \setminus A) = 0$.
3. Show that a transformation T is mixing if and only if, for all random variables ξ and η with $\mathsf{E}\,\xi^2 < \infty$ and $\mathsf{E}\,\eta^2 < \infty$,

$$\mathsf{E}\,\xi(T^n\omega)\eta(\omega) \to \mathsf{E}\,\xi(\omega)\,\mathsf{E}\,\eta(\omega), \quad n \to \infty.$$

4. Give an example of a measure-preserving *ergodic* transformation that is not *mixing*.
5. Let T be a measure-preserving transformation on $(\Omega, \mathscr{F}, \mathsf{P})$. Let $\mathscr{A}$ be an algebra of subsets of Ω and $\sigma(\mathscr{A}) = \mathscr{F}$. Suppose that Definition 1 requires *only* that the property
$$\lim_{n\to\infty} \mathsf{P}(A \cap T^{-n}B) = \mathsf{P}(A)\ \mathsf{P}(B)$$
be satisfied for sets A and B in $\mathscr{A}$. Show that this property will then hold for all A and B in $\mathscr{F} = \sigma(\mathscr{A})$ (and therefore the transformation T is mixing).
Show that this statement remains true if $\mathscr{A}$ is a π-system such that $\pi(\mathscr{A}) = \mathscr{F}$.
6. Let A be an almost invariant set. Show that $\omega \in A$ (P-a.s.) if and only if $T^n\omega \in A$ for all $n = 1, 2, \ldots$ (cf. Theorem 1 in Sect. 1.)
7. Give examples of measure-preserving transformations T on $(\Omega, \mathscr{F}, \mathsf{P})$ such that (a) $A \in \mathscr{F}$ does not imply that $TA \in \mathscr{F}$ and (b) $A \in \mathscr{F}$ and $TA \in \mathscr{F}$ do not imply that $\mathsf{P}(A) = \mathsf{P}(TA)$.
8. Let T be a measurable transformation on $(\Omega, \mathscr{F})$, and let $\mathscr{P}$ be the set of probability measures P with respect to which T is measure-preserving. Show that:
   (a) The set $\mathscr{P}$ is convex;
   (b) T is an ergodic transformation with respect to P if and only if P is an extreme point of $\mathscr{P}$ (i.e., P cannot be represented as $\mathsf{P} = \lambda_1\,\mathsf{P}_1 + \lambda_2\,\mathsf{P}_2$ with $\lambda_1 > 0$, $\lambda_2 > 0$, $\lambda_1 + \lambda_2 = 1$, $\mathsf{P}_1 \neq \mathsf{P}_2$, and $\mathsf{P}_1, \mathsf{P}_2 \in \mathscr{P}$).

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* ($\mathsf{P}$-*a.s.*)

$$\lim_n \frac{1}{n}\sum_{k=0}^{n-1} \xi(T^k\omega) = \mathsf{E}(\xi \mid \mathscr{I}), \tag{1}$$

where $\mathscr{I}$ is the invariant σ-algebra. 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 Garsia [28].

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 $n > k$, 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 $\xi(\omega) = S_1(\omega) \geq S_1(\omega) - M_n(T\omega)$, we have

$$\xi(\omega) \geq \max\{S_1(\omega), \ldots, S_n(\omega)\} - M_n(T\omega).$$

Therefore, since $\{M_n(\omega) > 0\} = \{\max(S_1(\omega), \ldots, S_n(\omega)) > 0\}$,

$$\begin{aligned}\mathsf{E}[\xi(\omega) I_{\{M_n>0\}}(\omega)] &\geq \mathsf{E}[(\max(S_1(\omega), \ldots, S_n(\omega)) - M_n(T\omega)) I_{\{M_n>0\}}(\omega)] \\ &\geq \mathsf{E}\{(M_n(\omega) - M_n(T\omega)) I_{\{M_n(\omega)>0\}}\} \geq \mathsf{E}\{M_n(\omega) - M_n(T\omega)\} = 0,\end{aligned}$$

where we have used the fact that if T is a measure-preserving transformation, then $\mathsf{E}\,M_n(\omega) = \mathsf{E}\,M_n(T\omega)$ (Problem 1 in Sect. 1).

This completes the proof of the lemma.

□

PROOF OF THEOREM 1. Let us suppose that $\mathsf{E}(\xi \mid \mathscr{I}) = 0$ (otherwise, replace ξ by $\xi - \mathsf{E}(\xi \mid \mathscr{I})$).

Let $\overline{\eta} = \limsup(S_n/n)$ and $\underline{\eta} = \liminf(S_n/n)$. It will be enough to establish that

$$0 \leq \underline{\eta} \leq \overline{\eta} \leq 0 \quad (\mathsf{P}\text{-a.s.}).$$

Consider the random variable $\overline{\eta} = \overline{\eta}(\omega)$. Since $\overline{\eta}(\omega) = \overline{\eta}(T\omega)$, the variable $\overline{\eta}$ is invariant, and consequently, for every $\varepsilon > 0$, the set $A_\varepsilon = \{\overline{\eta}(\omega) > \varepsilon\}$ is also invariant. Let us introduce the new random variable

$$\xi^*(\omega) = (\xi(\omega) - \varepsilon) I_{A_k}(\omega),$$

and set

$$S_k^*(\omega) = \xi^*(\omega) + \cdots + \xi^*(T^{k-1}\omega), \quad M_k^*(\omega) = \max(0,\ S_1^*,\ \ldots,\ S_k^*).$$

Then, by the lemma,

$$\mathsf{E}[\xi^* I_{\{M_n^*>0\}}] \geq 0$$

for every $n \geq 1$. But as $n \to \infty$,

$$\begin{aligned}\{M_n^* > 0\} &= \left\{\max_{1\leq k\leq n} S_k^* > 0\right\} \uparrow \left\{\sup_{k\geq 1} S_k^* > 0\right\} = \left\{\sup_{k\geq 1} \frac{S_k^*}{k} > 0\right\} \\ &= \left\{\sup_{k\geq 1} \frac{S_k}{k} > \varepsilon\right\} \cap A_\varepsilon = A_\varepsilon,\end{aligned}$$

where the last equation follows because $\sup_{k\geq 1}(S_k^*/k) \geq \overline{\eta}$ and $A_\varepsilon = \{\omega\colon \overline{\eta} > \varepsilon\}$.

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,

$$\begin{aligned}0 \leq \mathsf{E}[\xi^* I_{A_\varepsilon}] &= \mathsf{E}[(\xi-\varepsilon) I_{A_\varepsilon}] = \mathsf{E}[\xi I_{A_\varepsilon}] - \varepsilon\,\mathsf{P}(A_\varepsilon) \\ &= \mathsf{E}[\mathsf{E}(\xi|\mathscr{I}) I_{A_\varepsilon}] - \varepsilon\,\mathsf{P}(A_\varepsilon) = -\varepsilon\,\mathsf{P}(A_\varepsilon),\end{aligned}$$

so that $\mathsf{P}(A_\varepsilon) = 0$, and therefore $\mathsf{P}(\overline{\eta} \leq 0) = 1$.

Similarly, if we consider $-\xi(\omega)$ instead of $\xi(\omega)$, we find that

$$\limsup\left(-\frac{S_n}{n}\right) = -\liminf\frac{S_n}{n} = -\underline{\eta}$$

and $\mathsf{P}(-\underline{\eta} \leq 0) = 1$, i.e., $\mathsf{P}(\underline{\eta} \geq 0) = 1$. Therefore $0 \leq \underline{\eta} \leq \overline{\eta} \leq 0$ (P-a.s.), and the first part of the theorem is established.

To prove the second part, we observe that since $\mathsf{E}(\xi \mid \mathscr{I})$ is an invariant random variable, we have $\mathsf{E}(\xi \mid \mathscr{I}) = \mathsf{E}\,\xi$ (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 let $A = B \in \mathscr{I}$ 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 the mean. (This result will be used subsequently 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*

$$\mathsf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1}\xi(T^k\omega) - \mathsf{E}(\xi \mid \mathscr{I})\right| \to 0, \quad n \to \infty. \tag{4}$$

If also T is ergodic, then

$$\mathsf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1}\xi(T^k\omega) - \mathsf{E}\,\xi\right| \to 0, \quad n \to \infty. \tag{5}$$

PROOF. For every $\varepsilon > 0$ there is a bounded random variable η ($|\eta(\omega)| \le M$) such that $\mathsf{E}\,|\xi - \eta| \le \varepsilon$. Then

$$\begin{aligned}\mathsf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1}\xi(T^k\omega) - \mathsf{E}(\xi \mid \mathscr{I})\right| &\le \mathsf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1}(\xi(T^k\omega) - \eta(T^k\omega))\right| \\ &+ \mathsf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1}\eta(T^k\omega) - \mathsf{E}(\eta \mid \mathscr{I})\right| + \mathsf{E}\,|\mathsf{E}(\xi \mid \mathscr{I}) - \mathsf{E}(\eta \mid \mathscr{I})|. \end{aligned} \tag{6}$$

Since $|\eta| \le M$, by the dominated convergence theorem and using (1), we find that the second term on the right-hand side 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 3ε, so that (4) is proved. Finally, if T is ergodic, then (5) follows from (4) and the remark that $\mathsf{E}(\xi \mid \mathscr{I}) = \mathsf{E}\,\xi$ (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* (in 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 Sect. 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 $(1/n)\sum_{k=1}^{n}\tilde{\xi}_1(T^{k-1}\tilde{\omega})$ (P-a.s. and in the 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 the mean) to a random variable η such that $\eta \stackrel{d}{=} \tilde{\eta}$. It follows from Theorem 1 that if $\tilde{\mathsf{E}}|\tilde{\xi}_1| < \infty$, then $\tilde{\eta} = \tilde{\mathsf{E}}(\tilde{\xi}_1 \mid \tilde{\mathscr{I}})$, where $\tilde{\mathscr{I}}$ is a collection of invariant sets ($\tilde{\mathsf{E}}$ is the expectation 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 $n \geq 1$

$$A = \{\omega\colon (\xi_n,\ \xi_{n+1},\ \ldots) \in B\}.$$

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 if the random variable η is the limit (P-a.s. and in the mean) of $\frac{1}{n}\sum_{k=1}^{n}\xi_k(\omega)$, $n \to \infty$, then it can be taken equal to $\mathsf{E}(\xi_1 \mid \mathscr{I}_{\xi})$. To this end, notice that we can set

$$\eta(\omega) = \limsup_n \frac{1}{n}\sum_{k=1}^{n}\xi_k(\omega). \tag{7}$$

It follows from the definition of $\limsup$ that for the random variable $\eta(\omega)$ so defined, the sets $\{\omega\colon \eta(\omega) < y\}$, $y \in R$, are invariant, and therefore η is $\mathscr{I}_{\xi}$-*measurable*. Now, 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 for η defined by (7)

$$\frac{1}{n}\sum_{k=1}^{n}\int_A \xi_k\, d\mathsf{P} \to \int_A \eta\, d\mathsf{P}. \tag{8}$$

Let $B \in \mathscr{B}(R^{\infty})$ be such that $A = \{\omega\colon (\xi_k, \xi_{k+1}, \ldots) \in B\}$ for all $k \geq 1$. Then since ξ is stationary,

$$\int_A \xi_k\, d\mathsf{P} = \int_{\{\omega\colon (\xi_k,\xi_{k+1},\ldots)\in B\}} \xi_k\, d\mathsf{P} = \int_{\{\omega\colon (\xi_1,\xi_2,\ldots)\in B\}} \xi_1\, d\mathsf{P} = \int_A \xi_1\, d\mathsf{P}.$$

Hence it follows from (8) that for all $A \in \mathscr{I}_{\xi}$,

$$\int_A \xi_1\, d\mathsf{P} = \int_A \eta\, d\mathsf{P},$$

which implies (see (1) in Sect. 7, Chap. 2, Vol. 1) that (η being $\mathscr{I}_{\xi}$-measurable) $\eta = \mathsf{E}(\xi_1 \mid \mathscr{I}_{\xi})$. Here $\mathsf{E}(\xi_1 \mid \mathscr{I}_{\xi}) = \mathsf{E}\,\xi_1$ if ξ is ergodic.

Therefore we have proved the following theorem.

Theorem 3 (Ergodic Theorem). *Let $\xi = (\xi_1, \xi_2, \dots)$ 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 $\xi = (\xi_1,\ \xi_2,\ \dots)$ 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 the measure-preserving transformation related to ξ to be *mixing* (and, hence, *ergodic*).
2. Show that for every sequence $\xi = (\xi_1, \xi_2, \dots)$ of independent identically distributed random variables the corresponding measure-preserving transformation is mixing.
3. Show that a stationary sequence ξ is ergodic if and only if

$$\frac{1}{n} \sum_{i=1}^{n} I_B(\xi_i, \dots, \xi_{i+k-1}) \to \mathsf{P}((\xi_1, \dots, \xi_k) \in B) \quad (\mathsf{P}\text{-a.s.})$$

   for every $B \in \mathscr{B}(R^k)$, $k = 1, 2, \dots$.
4. Let P and $\bar{\mathsf{P}}$ be two measures on the space $(\Omega, \mathscr{F})$ such that the measure-preserving transformation T is ergodic with respect to each of them. Prove that, then, either $\mathsf{P} = \bar{\mathsf{P}}$ or $\mathsf{P} \perp \bar{\mathsf{P}}$.
5. Let T be a measure-preserving transformation on $(\Omega, \mathscr{F}, \mathsf{P})$ and $\mathscr{A}$ an algebra of subsets of Ω such that $\sigma(\mathscr{A}) = \mathscr{F}$. Let

$$I_A^{(n)} = \frac{1}{n} \sum_{k=0}^{n-1} I_A(T^k\omega).$$

   Prove that T is ergodic if and only if one of the following conditions holds:
   (a) $I_A^{(n)} \xrightarrow{\mathsf{P}} \mathsf{P}(A)$ for any $A \in \mathscr{A}$;
   (b) $\lim \frac{1}{n} \sum_{k=0}^{n-1} \mathsf{P}(A \cap T^{-k}B) = \mathsf{P}(A)\,\mathsf{P}(B)$ for all $A, B \in \mathscr{A}$;
   (c) $I_A^{(n)} \xrightarrow{\mathsf{P}} \mathsf{P}(A)$ for any $A \in \mathscr{F}$.
6. Let T be a measure-preserving transformation on $(\Omega, \mathscr{F}, \mathsf{P})$. Prove that T is ergodic (with respect to P) if and only if there is no measure $\bar{\mathsf{P}} \neq \mathsf{P}$ on $(\Omega, \mathscr{F})$ such that $\bar{\mathsf{P}} \ll \mathsf{P}$ and T is measure-preserving with respect to $\bar{\mathsf{P}}$.
7. (*Bernoullian shifts*.) Let S be a finite set (say, $S = \{1, 2, \dots, N\}$), and let $\Omega = S^\infty$ be the space of sequences $\omega = (\omega_0, \omega_1, \dots)$ with $\omega_i \in S$. Set $\xi_k(\omega) = \omega_k$,

and define the shift transformation $T(\omega_0, \omega_1, \dots) = (\omega_1, \omega_2, \dots)$, or, in terms of ξ_k, $\xi_k(T\omega) = \omega_{k+1}$ if $\xi_k(\omega) = \omega_k$. Suppose that for $i \in \{1, 2, \dots, N\}$ there are nonnegative numbers p_i such that $\sum_{i=1}^{N} p_i = 1$ (i.e., $(p_1, \dots, p_N)$ is a probability distribution). Define the probability measure P on $(S^\infty, \mathscr{B}(S^\infty))$ (see Sect. 3, Chap. 2, Vol. 1) such that

$$\mathsf{P}\{\omega\colon (\omega_1, \dots, \omega_k) = (u_1, \dots, u_k)\} = p_{u_1} \dots p_{u_k}.$$

In other words, this probability measure is introduced to provide the independence of $\xi_0(\omega), \xi_1(\omega), \dots$. The shift transformation T (relative to this measure P) is called the *Bernoullian shift* or the *Bernoulli transformation*.
Show that the Bernoulli transformation is mixing.

8. Let T be a measure-preserving transformation on $(\Omega, \mathscr{F}, \mathsf{P})$. Use the notation $T^{-n}\mathscr{F} = \{T^{-n}A\colon A \in \mathscr{F}\}$. We say that the σ-algebra

$$\mathscr{F}_{-\infty} = \bigcap_{n=1}^{\infty} T^{-n}\mathscr{F}$$

is *trivial* (P-*trivial*) if every set in $\mathscr{F}_{-\infty}$ has measure 0 or 1 (such transformations are referred to as *Kolmogorov transformations*). Prove that the Kolmogorov transformations are ergodic and, what is more, mixing.

9. Let $1 \le p < \infty$, and let T be a measure-preserving transformation on a probability space $(\Omega, \mathscr{F}, \mathsf{P})$. Consider a random variable $\xi(\omega) \in L^p(\Omega, \mathscr{F}, \mathsf{P})$.
Prove the following ergodic theorem in $L^p(\Omega, \mathscr{F}, \mathsf{P})$ (von Neumann). There exists a random variable $\eta(\omega)$ such that

$$\mathsf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1} \xi(T^k\omega) - \eta(\omega)\right|^p \to 0, \quad n \to \infty.$$

10. Borel's normality theorem (Example 3 in Sect. 3, Chap. 4) states that the fraction of ones and zeros in the binary expansion of a number ω in $[0, 1)$ converges to $\frac{1}{2}$ almost everywhere (with respect to the Lebesgue measure). Prove this result by considering the transformation $T\colon [0, 1) \to [0, 1)$ defined by

$$T(\omega) = 2\omega \pmod 1,$$

and using the ergodic Theorem 1.

11. As in Problem 10, let $\omega \in [0, 1)$. Consider the transformation $T\colon [0, 1) \to [0, 1)$ defined by

$$T(\omega) = \begin{cases} 0, & \text{if } \omega = 0, \\ \{1/\omega\}, & \text{if } \omega \ne 0, \end{cases}$$

where $\{x\}$ is the fractional part of x.

Show that T preserves the *Gaussian measure* $P = P(\cdot)$ on $[0, 1)$ defined by

$$P(A) = \frac{1}{\log 2} \int_A \frac{dx}{1+x}, \quad A \in \mathscr{B}([0, 1)).$$

12. Show by an example that Poincaré's recurrence theorem (Subsection 3 of Sect. 1) is, in general, false for measurable spaces with infinite measure.

Chapter 6
Stationary (Wide Sense) Random Sequences: L^2-Theory

> The [spectral] decomposition provides grounds for considering any stationary stochastic process in the wide sense as a superposition of a set of non-correlated harmonic oscillations with random amplitudes and phases.
>
> Encyclopaedia of Mathematics [42, Vol. 8, p. 480].

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 $\mathsf{E}\,\xi_n$ does not depend on n:

$$\mathsf{E}\,\xi_n = \mathsf{E}\,\xi_1, \tag{2}$$

and the covariance $\operatorname{Cov}(\xi_{n+m}\xi_n) = \mathsf{E}(\xi_{n+m} - \mathsf{E}\,\xi_{n+m})(\xi_n - \mathsf{E}\,\xi_n)$ depends only on m:

$$\operatorname{Cov}(\xi_{n+m}, \xi_n) = \operatorname{Cov}(\xi_{1+m}, \xi_1). \tag{3}$$

In this chapter we study sequences that are stationary in the wide sense (having 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

A. N. Shiryaev, *Probability-2*, Graduate Texts in Mathematics 95,
https://doi.org/10.1007/978-0-387-72208-5_3

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 $\xi = \alpha + i\beta$, $\alpha,\ \beta \in R$, with $\mathsf{E}\,|\xi|^2 < \infty$, where $|\xi|^2 = \alpha^2 + \beta^2$. If ξ and $\eta \in H^2$, then we set

$$(\xi, \eta) = \mathsf{E}\,\xi\overline{\eta}, \tag{4}$$

where $\overline{\eta} = \gamma - i\delta$ is the complex conjugate of $\eta = \gamma + i\delta$, 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; cf. Sects. 10 and 11 of Chap. 2, Vol. 1) 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

$$\mathrm{Cov}(\xi, \eta) = \mathsf{E}(\xi - \mathsf{E}\,\xi)(\overline{\eta - \mathsf{E}\,\eta}). \tag{6}$$

It follows from (4) and (6) that if $\mathsf{E}\,\xi = \mathsf{E}\,\eta = 0$, then

$$\mathrm{Cov}(\xi, \eta) = (\xi, \eta). \tag{7}$$

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 $n \in \mathbb{Z}$,

$$\begin{aligned} \mathsf{E}\,\xi_n &= \mathsf{E}\,\xi_0, \\ \mathrm{Cov}(\xi_{k+n},\ \xi_k) &= \mathrm{Cov}(\xi_n,\ \xi_0), \quad k \in \mathbb{Z}. \end{aligned} \tag{8}$$

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

$$R(n) = \mathrm{Cov}(\xi_n,\ \xi_0), \quad n \in \mathbb{Z}, \tag{9}$$

and (assuming $R(0) = \mathsf{E}\,|\xi_0|^2 \neq 0$)

$$\rho(n) = \frac{R(n)}{R(0)}, \quad n \in \mathbb{Z}. \tag{10}$$

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 *positive semidefinite*, i.e., for all complex numbers $a_1, \dots, a_m$ and $t_1, \dots, t_m \in \mathbb{Z}$, $m \geq 1$, we have

$$\sum_{i,j=1}^{m} a_i \bar{a}_j R(t_i - t_j) \geq 0, \tag{11}$$

since the left-hand side of (11) is equal to $\| \sum (\alpha_i \xi_{t_i}) \|^2$. It is then easy to deduce (either from (11) or directly from (9)) the following properties of the covariance function (see Problem 1):

$$\begin{aligned} &R(0) \geq 0, \quad R(-n) = \overline{R(n)}, \quad |R(n)| \leq R(0), \\ &|R(n) - R(m)|^2 \leq 2R(0)[R(0) - \operatorname{Re} R(n-m)]. \end{aligned} \tag{12}$$

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 often 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" $\xi_n \equiv \xi_0$ is a stationary sequence.

Remark. In connection with this example, notice that, since $e^{i\lambda n} = e^{in(\lambda + 2\pi k)}$, $k = \pm 1, \pm 2, \dots$, the (circular) frequency λ is defined up to a multiple of 2π. Following tradition, we will assume henceforth that $\lambda \in [-\pi, \pi]$.

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 $\mathsf{E}\,|z_k|^2 = \sigma_k^2 > 0$; $-\pi \leq \lambda_k < \pi$, $k = 1, \dots, N$; $\lambda_i \neq \lambda_j$, $i \neq j$. 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-hand side 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:

$$R(n) = \int_{-\pi}^{\pi} e^{i\lambda n}\, dF(\lambda) \quad \Big(= \int_{[-\pi,\pi)} e^{i\lambda n}\, dF(\lambda) \Big). \tag{18}$$

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 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 Sect. 3 using what we shall learn to call stochastic integrals with respect to orthogonal stochastic measures (Sect. 2).

EXAMPLE 3 (White noise). Let $\varepsilon = (\varepsilon_n)$ be an orthonormal sequence of random variables, $\mathsf{E}\,\varepsilon_n = 0$, $\mathsf{E}\,\varepsilon_i\overline{\varepsilon}_j = \delta_{ij}$, where δ_{ij} is the Kronecker delta. Such a sequence is evidently stationary, and

$$R(n) = \begin{cases} 1, & n = 0, \\ 0, & n \ne 0. \end{cases}$$

Observe that $R(n)$ can be represented in the form

$$R(n) = \int_{-\pi}^{\pi} e^{i\lambda n} dF(\lambda), \tag{19}$$

where

$$F(\lambda) = \int_{-\pi}^{\lambda} f(v)dv; \quad f(\lambda) = \frac{1}{2\pi}, \quad -\pi \le \lambda < \pi. \tag{20}$$

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" $f(\lambda) \equiv 1/2\pi$. 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 $\sum_{k=-\infty}^{\infty} |a_k|^2 < \infty$. From (21) we obtain

$$\operatorname{Cov}(\xi_{n+m}, \xi_m) = \operatorname{Cov}(\xi_n, \xi_0) = \sum_{k=-\infty}^{\infty} a_{n+k}\overline{a}_k,$$

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 where 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*.

It can be shown (Problem 3) 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 $|\alpha| < 1$, then 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-hand side 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

$$\mathsf{E}\left[\xi_n - \sum_{j=0}^{k-1} \alpha^j \varepsilon_{n-j}\right]^2 = \mathsf{E}[\alpha^k \xi_{n-k}]^2 = \alpha^{2k}\,\mathsf{E}\,\xi_{n-k}^2 = \alpha^{2k}\,\mathsf{E}\,\xi_0^2 \to 0, \quad k \to \infty.$$

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 3) in the form

$$R(n) = \int_{-\pi}^{\pi} e^{i\lambda n} dF(\lambda), \quad F(\lambda) = \int_{-\pi}^{\lambda} f(v)dv, \tag{28}$$

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 $\mathsf{E}\,\xi_0 = 0$,

$$\mathsf{E}\,\xi_0^2 = \frac{1}{1-|\alpha|^2}, \quad \text{and} \quad R(n) = \frac{\alpha^n}{1-|\alpha|^2}, \quad n \geq 0$$

(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 $\xi_n = H_n - \overline{H}$ be the deviation from the mean level (which is obtained from observations over many years), and suppose that $S(H) = S(\overline{H}) + c(H - \overline{H})$. Then it follows from the balance equation that ξ_n satisfies

$$\xi_{n+1} = \alpha\xi_n + \varepsilon_{n+1} \tag{31}$$

with $\alpha = 1 - cK$, $\varepsilon_n = \Sigma_n - KS(\overline{H})$. It is natural to assume that the random variables ε_n have zero means and, as a first order approximation, are uncorrelated and identically distributed. Then, as we showed in Example 5, Eq. (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 Sect. 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 of $Q(z)$ (see (27)) it will be shown later (Corollary 6, Sect. 3) that (32) has a 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

$$f(\lambda) = \frac{1}{2\pi} \cdot \left| \frac{P(e^{-i\lambda})}{Q(e^{-i\lambda})} \right|^2$$

with P and Q as in (23) and (27).

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 $n \in \mathbb{Z}$*

$$R(n) = \int_{-\pi}^{\pi} e^{i\lambda n} F(d\lambda), \tag{33}$$

where the integral is understood as the Lebesgue–Stieltjes integral over $[-\pi,\pi)$.

PROOF. For $N \geq 1$ and $\lambda \in [-\pi,\pi]$, set

$$f_N(\lambda) = \frac{1}{2\pi N} \sum_{k=1}^{N} \sum_{l=1}^{N} R(k-l)\, e^{-ik\lambda}\, e^{il\lambda}. \tag{34}$$

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

$$F_N(B) = \int_B f_N(\lambda)\, d\lambda, \quad B \in \mathscr{B}([-\pi,\pi)).$$

Then

$$\int_{-\pi}^{\pi} e^{i\lambda n} F_N(d\lambda) = \int_{-\pi}^{\pi} e^{i\lambda n} f_N(\lambda)\, d\lambda = \begin{cases} \left(1 - \frac{|n|}{N}\right) R(n), & |n| < N, \\ 0, & |n| \geq N. \end{cases} \tag{36}$$

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 Sect. 2, Chap. 3, Vol. 1) 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

$$\int_{-\pi}^{\pi} e^{i\lambda n} F(d\lambda) = \lim_{N_k \to \infty} \int_{-\cdot\pi}^{\pi} e^{i\lambda n} F_{N_k}(d\lambda) = R(n).$$

The measure F so constructed is supported on $[-\pi,\pi]$. Without changing the integral $\int_{-\pi}^{\pi} e^{i\lambda n} F(d\lambda)$, 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 $[-\pi, \pi)$. (Regarding the choice of $[-\pi, \pi)$ as the domain of λ see the Remark to Example 1.)

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 the preceding Example 2, the spectral measure was discrete (concentrated at $\lambda_k,\ k = 0,\ \pm 1,\ \ldots$). 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

$$\int_{-\pi}^{\pi} e^{i\lambda n} F_1(d\lambda) = \int_{-\pi}^{\pi} e^{i\lambda n} F_2(d\lambda), \quad n \in \mathbb{Z}.$$

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 (cf. proof of Theorem 2 in Sect. 12, Chap. 2, Vol. 1) 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 $R(n) = R(-n)$, and therefore

$$R(n) = \frac{R(n) + R(-n)}{2} = \int_{-\pi}^{\pi} \cos \lambda n\, F(d\lambda).$$

4. Problems

1. Derive (12) from (11).
2. Prove that the autoregression Eq. (24) has a unique stationary solution representable as a one-sided moving average if all the zeros of the polynomial $Q(z)$ defined by (27) lie *outside* the unit disk.
3. Show that the spectral functions of sequences (22) and (24) have densities specified by (23) and (29), respectively.
4. Show that if $\sum_{n=-\infty}^{+\infty} |R(n)|^2 < \infty$, then the spectral function $F(\lambda)$ has a density $f(\lambda)$ given by

$$f(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-i\lambda n} R(n),$$

where the series converges in the complex space $L^2 = L^2([-\pi, \pi), \mathscr{B}([-\pi, \pi)), \lambda)$ with λ the Lebesgue measure.

2. Orthogonal Stochastic Measures and Stochastic Integrals

1. As we observed in Sect. 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 set

$$Z(\lambda) = \sum_{\{k:\ \lambda_k \le \lambda\}} z_k, \tag{2}$$

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 the Riemann–Stieltjes type. However, in the present case, $Z(\lambda)$ is a random function (it also depends on ω). And it will be seen 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 (Sect. 6, Chap. 2, Vol. 1), we begin by defining *stochastic measure*.

Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space, and let E be a set, with an algebra $\mathscr{E}_0$ of its subsets and the σ-algebra $\mathscr{E}$ generated by $\mathscr{E}_0$, $\mathscr{E} = \sigma(\mathscr{E}_0)$.

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 $\Delta = \sum_{k=1}^{\infty} \Delta_k \in \mathscr{E}_0$,

$$\mathsf{E}\left| Z(\Delta) - \sum_{k=1}^{n} Z(\Delta_k) \right|^2 \to 0, \quad n \to \infty. \tag{5}$$

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 1.

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":

$$\mathsf{E}\,|Z(\Delta_n)|^2 \to 0, \quad \Delta_n \downarrow \varnothing,\ \Delta_n \in \mathscr{E}_0. \tag{6}$$

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

$$m(\Delta) = \mathsf{E}\,|Z(\Delta)|^2, \quad \Delta \in \mathscr{E}_0. \tag{9}$$

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 (Sect. 3, Chap. 2, Vol. 1), 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 $\mathscr{E} = \sigma(\mathscr{E}_0)$, can 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 that 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

$$f(\lambda) = \sum f_k I_{\Delta_k}(\lambda), \quad \Delta_k \in \mathscr{E}_0, \tag{10}$$

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

$$\langle f, g\rangle = \int_E f(\lambda)\overline{g(\lambda)}\, m(d\lambda)$$

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

$$(\xi, \eta) = \mathsf{E}\,\xi\overline{\eta}$$

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 the type (10) such that $\|f - f_n\| \to 0$, $n \to \infty$ (Problem 2). Consequently,

$$\|\mathscr{I}(f_n) - \mathscr{I}(f_m)\| = \|f_n - f_m\| \to 0, \quad n, m \to \infty.$$

Therefore the sequence $\{\mathscr{I}(f_n)\}$ is fundamental in the mean-square sense and, by Theorem 7 in Sect. 10, Chap. 2, Vol. 1, 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. 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 measure 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

$$\tilde{Z}(\Delta_1 + \Delta_2) = \tilde{Z}(\Delta_1) + \tilde{Z}(\Delta_2) \quad (\mathsf{P}\text{-a.s.})$$

and it follows from (12) that

$$\mathsf{E}\,|\tilde{Z}(\Delta)|^2 = m(\Delta), \quad \Delta \in \mathscr{E}.$$

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

$$g_n(\lambda) = I_\Delta(\lambda) - \sum_{k=1}^{n} I_{\Delta_k}(\lambda) = I_{\Sigma_n}(\lambda), \quad \Sigma_n = \sum_{k=n+1}^{\infty} \Delta_k.$$

But

$$\mathsf{E}\,|\mathscr{I}(g_n)|^2 = \|g_n\|^2 = m(\Sigma_n) \downarrow 0, \quad n \to \infty,$$

i.e.,

$$\mathsf{E}\,|\tilde{Z}(\Delta) - \sum_{k=1}^{n} \tilde{Z}(\Delta_k)|^2 \to 0, \quad n \to \infty.$$

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 earlier, 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 (Theorem 1, Sect. 3, Chap. 2, Vol. 1), there is a one-to-one correspondence between finite measures $m = m(\Delta)$ on $(R, \mathscr{B}(R))$ and (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 $\lambda \in R$

$$\mathsf{E}\,|Z_\lambda - Z_{\lambda_n}|^2 \to 0, \quad \lambda_n \downarrow \lambda,\ \lambda_n \in R;$$

(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 $m = m(\Delta)$, which is a finite measure with (generalized) distribution function $G(\lambda)$. Let us set

$$Z_\lambda = Z(-\infty, \lambda].$$

Then

$$\mathsf{E}\,|Z_\lambda|^2 = m(-\infty, \lambda] = G(\lambda) < \infty, \quad \mathsf{E}\,|Z_\lambda - Z_{\lambda_n}|^2 = m(\lambda, \lambda_n] \downarrow 0,\ \lambda_n \downarrow \lambda,$$

and (evidently) (3) is also satisfied. Thus, $\{Z_\lambda\}$ is a process *with orthogonal increments*.

On the other hand, let $G(\lambda)$ be a generalized distribution function, $G(-\infty) = 0$, $G(+\infty) < \infty$, and let $\{Z_\lambda\}$ be a process with orthogonal increments such that $\mathsf{E}\,|Z_\lambda|^2 = G(\lambda)$. Set

$$Z(\Delta) = Z_b - Z_a$$

when $\Delta = (a, b]$. Let $\mathscr{E}_0$ be the algebra generated by the sets $\Delta = \sum_{k=1}^n (a_k, b_k]$ with disjoint $(a_k, b_k]$ and

$$Z(\Delta) = \sum_{k=1}^n Z(a_k,\ b_k].$$

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 $\Delta_2 = (a_2, b_2]$.

Due to continuity on the right of $G(\lambda)$, $\lambda \in R$, this implies that $Z = Z(\Delta)$, $\Delta \in \mathscr{E}_0$, 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 sets $\Delta \in \mathscr{E}$, 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

$$Z_\lambda = Z(-\infty, \lambda], \quad G(\lambda) = m(-\infty, \lambda]$$

and

$$Z(a, b] = Z_b - Z_a, \quad m(a, b] = G(b) - G(a).$$

By analogy with the usual notation of the theory of Lebesgue–Stieltjes and Riemann–Stieltjes integration (Subsections 9 and 11 of Sect. 6, Chap. 2, Vol. 1), 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 orthogonal stochastic measure corresponding to $\{Z_\lambda\}$.

6. PROBLEMS

1. Prove the equivalence of (5) and (6).
2. Let $f \in L^2$. Using the results of Chap. 2, Vol. 1 (Theorem 1 in Sect. 4, the Corollary to Theorem 3 of Sect. 6, and Problem 8 of Sect. 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 $m(\Delta)$:

$$\begin{aligned} \mathsf{E}\,|Z(\Delta_1) - Z(\Delta_2)|^2 &= m(\Delta_1 \triangle \Delta_2), \\ Z(\Delta_1 \setminus \Delta_2) &= Z(\Delta_1) - Z(\Delta_1 \cap \Delta_2) \quad (\mathsf{P}\text{-a.s.}), \\ Z(\Delta_1 \triangle \Delta_2) &= Z(\Delta_1) + Z(\Delta_2) - 2Z(\Delta_1 \cap \Delta_2) \quad (\mathsf{P}\text{-a.s.}). \end{aligned}$$

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 Sect. 1, there is a finite measure $F = F(\Delta)$ on $([-\pi, \pi), \mathscr{B}([-\pi, \pi)))$ such that the covariance function $R(n) = \mathrm{Cov}(\xi_{k+n}, \xi_k)$ 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.*)

$$\xi_n = \int_{-\pi}^{\pi} e^{i\lambda n}\, Z(d\lambda) \quad \left(= \int_{[-\pi,\pi)} e^{i\lambda n}\, Z(d\lambda) \right). \tag{2}$$

Moreover, $\mathsf{E}\,Z(\Delta) = 0$, $\mathsf{E}\,|Z(\Delta)|^2 = F(\Delta)$.

PROOF. 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

$$\langle f, g\rangle = \int_{-\pi}^{\pi} f(\lambda)\overline{g}(\lambda)\, F(d\lambda), \tag{3}$$

and let $L_0^2(F)$ be the linear manifold ($L_0^2(F) \subseteq L^2(F)$) spanned by the functions $e_n = e_n(\lambda)$, $n \in \mathbb{Z}$, 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

$$e_n \leftrightarrow \xi_n, \quad n \in \mathbb{Z}, \tag{4}$$

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 $\Sigma\alpha_n e_n = 0$ 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),

$$\begin{aligned}\langle e_n, e_m\rangle = \int_{-\pi}^{\pi} e_n(\lambda)\overline{e_m(\lambda)}\, F(d\lambda) &= \int_{-\pi}^{\pi} e^{i\lambda(n-m)}\, F(d\lambda) \\ &= R(n-m) = \mathsf{E}\,\xi_n\overline{\xi}_m = (\xi_n, \xi_m)\end{aligned}$$

and similarly,

$$\Big\langle \sum \alpha_n e_n, \sum \beta_n e_n \Big\rangle = \Big(\sum \alpha_n \xi_n, \sum \beta_n \xi_n \Big). \tag{6}$$

Now let $\eta \in L^2(\xi)$. Since $L^2(\xi) = \overline{L_0^2}(\xi)$, 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 earlier. 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 $\Delta \in \mathscr{B}([-\pi, \pi))$, $\lambda \in [-\pi, \pi)$, and let $Z(\Delta)$ be the element of $L^2(\xi)$ such that $I_\Delta(\lambda) \leftrightarrow Z(\Delta)$. It is clear that $\|I_\Delta(\lambda)\|^2 = F(\Delta)$, and therefore $\mathsf{E}\,|Z(\Delta)|^2 = F(\Delta)$. Since $\mathsf{E}\,\xi_n = 0$, $n \in \mathbb{Z}$, we have for every element of $L_0^2(\xi)$ (and hence of $L^2(\xi)$) that it has zero expectation. In particular, $\mathsf{E}\,Z(\Delta) = 0$. Moreover, if $\Delta_1 \cap \Delta_2 = \varnothing$, we have $\mathsf{E}\,Z(\Delta_1)Z(\Delta_2) = 0$ and $\mathsf{E}\left|Z(\Delta) - \sum_{k=1}^n Z(\Delta_k)\right|^2 \to 0$, $n \to \infty$, 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 Sect. 2) we can define the stochastic integral

$$\mathscr{I}(f) = \int_{-\pi}^{\pi} f(\lambda)\, Z(d\lambda), \quad f \in L^2(F).$$

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 (14) of Sect. 2). 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

$$\xi_n = \int_{-\pi}^{\pi} e^{i\lambda n}\, Z(d\lambda), \quad n \in \mathbb{Z} \quad (\mathsf{P}\text{-a.s.})$$

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 $-\Delta = \{\lambda\colon\ -\lambda \in \Delta\}$.

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

$$\overline{\eta} = \sum \overline{\alpha}_k \xi_k \leftrightarrow \sum \overline{\alpha}_k e^{i\lambda k} = \overline{f(-\lambda)}. \tag{10}$$

Since $I_\Delta(\lambda) \leftrightarrow Z(\Delta)$, it follows from (10) that $I_\Delta(-\lambda) \leftrightarrow \overline{Z(\Delta)}$ (or, equivalently, $I_{-\Delta}(\lambda) \leftrightarrow \overline{Z(\Delta)}$). On the other hand, $I_{-\Delta}(\lambda) \leftrightarrow Z(-\Delta)$. Therefore $\overline{Z(\Delta)} = Z(-\Delta)$ (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 in $\mathscr{B}([-\pi, \pi))$, and if $\Delta_1 \cap \Delta_2 = \varnothing$ and $(-\Delta_1) \cap \Delta_2 = \varnothing$, then

$$\mathsf{E}\, Z_1(\Delta_1)Z_1(\Delta_2) = 0, \quad \mathsf{E}\, Z_2(\Delta_1)Z_2(\Delta_2) = 0. \tag{12}$$

In fact, since $Z(\Delta) = \overline{Z(-\Delta)}$, we have

$$Z_1(-\Delta) = Z_1(\Delta), \quad Z_2(-\Delta) = -Z_2(\Delta). \tag{13}$$

Moreover, since $\mathsf{E}\, Z(\Delta_1)\overline{Z(\Delta_2)} = \mathsf{E}\, |Z(\Delta_1 \cap \Delta_2)|^2$, we have $\operatorname{Im} \mathsf{E}\, Z(\Delta_1)\overline{Z(\Delta_2)} = 0$, 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).

When $\Delta_1 \cap \Delta_2 = \varnothing$ and $(-\Delta_1) \cap \Delta_2 = \varnothing$, we have $\mathsf{E}\, Z(\Delta_1)\overline{Z(\Delta_2)} = 0$, whence $\operatorname{Re} \mathsf{E}\, Z(\Delta_1)\overline{Z(\Delta_2)} = 0$ and $\operatorname{Re} \mathsf{E}\, Z(-\Delta_1)\overline{Z(\Delta_2)} = 0$, which, with (13), provides an evident proof of (12).

Corollary 3. *Let $\xi = (\xi_n)$ be a Gaussian sequence. Then, for any $\Delta_1, \ldots, \Delta_k$, the vector $(Z_1(\Delta_1), \ldots, Z_1(\Delta_k), Z_2(\Delta_1), \ldots, Z_2(\Delta_k))$ 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 of Sect. 13, Chap. 2, Vol. 1, 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 if $\Delta_i \cap \Delta_j = (-\Delta_i) \cap \Delta_j = \varnothing$, $i, j = 1, \ldots, k$, $i \neq j$, the random variables $Z_i(\Delta_1), \ldots, Z_i(\Delta_k)$ are mutually 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 Sect. 2, the spectral representation (2) can also be written in the following form:

$$\xi_n = \int_{-\pi}^{\pi} e^{i\lambda n} \, dZ_\lambda, \quad n \in \mathbb{Z}. \tag{17}$$

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 $\eta \in L^2(\xi)$, then 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 by

$$\varphi_n(\lambda) = \sum_{|k| \le n} \alpha_n e^{i\lambda k}. \tag{21}$$

In the general case, where $\eta \in L^2(\xi)$, 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 (14) of Sect. 2, 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 $\xi^0 = (\xi_n)_{n \le 0}$ and by the functions $e^0 = (e_n)_{n \le 0}$. 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 is 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 at the output 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

$$y_n = \sum_{m=-\infty}^{\infty} h(n-m)x_m = \sum_{m=0}^{\infty} h(m)x_{n-m}. \tag{23}$$

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

$$\sum_{k,l=-\infty}^{\infty} h(k)R(k-l)\overline{h(l)} < \infty, \tag{25}$$

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

$$\eta_n = \int_{-\pi}^{\pi} e^{i\lambda n}\varphi(\lambda)\, Z(d\lambda), \quad n \in \mathbb{Z}, \tag{28}$$

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, in a certain sense, 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 $\varphi(\lambda) \in L^2(d\mu)$, where $d\mu$ is the 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

$$\int_{-\pi}^{\pi} \left| \varphi(\lambda) - \sum_{|m|\le n} e^{-i\lambda m} h(m) \right|^2 d\lambda \to 0, \quad n \to \infty.$$

Let

$$\eta_n = \int_{-\pi}^{\pi} e^{i\lambda n} \, Z(d\lambda), \quad n \in \mathbb{Z}.$$

Besides the measure $Z = Z(\Delta)$, we introduce another, independent of Z, 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 set

$$\overline{Z}(\Delta) = \int_\Delta \varphi^{\oplus}(\lambda)\, Z(d\lambda) + \int_\Delta [1 - \varphi^{\oplus}(\lambda)\varphi(\lambda)]\, \tilde{Z}(d\lambda),$$

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

$$\int_{-\pi}^{\pi} e^{i\lambda n}\varphi(\lambda)\,\overline{Z}(d\lambda) = \int_{-\pi}^{\pi} e^{i\lambda n}\,Z(d\lambda) = \eta_n \tag{31}$$

and, on the other hand, by definition of $\varphi(\lambda)$ and property (14) in Sect. 2, we have (P-a.s.)

$$\begin{aligned}\int_{-\pi}^{\pi} e^{i\lambda n}\varphi(\lambda)\,\overline{Z}(d\lambda) &= \int_{-\pi}^{\pi} e^{i\lambda n}\Bigg(\sum_{m=-\infty}^{\infty} e^{-i\lambda m}h(m)\Bigg)\overline{Z}(d\lambda)\\ &= \sum_{m=-\infty}^{\infty} h(m)\int_{-\pi}^{\pi} e^{i\lambda(n-m)}\,\overline{Z}(d\lambda) = \sum_{m=-\infty}^{\infty} h(m)\varepsilon_{n-m},\end{aligned}$$

which, together with (31), establishes 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 5. *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

$$\varphi(\lambda) = \sum_{k=0}^{\infty} e^{-i\lambda k}h(k), \quad \sum_{k=0}^{\infty}|h(k)|^2 < \infty.$$

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 $P(z) = a_0 + a_1 z + \cdots + a_p z^p$. *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 6. *Let* $\xi = (\xi_n)$ *be a stationary 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 $P(z) = a_0 + a_1 z + \cdots + a_p z^p$, $Q(z) = 1 + b_1 z + \cdots + b_q z^q$.

If $Q(z)$ *has no zeros on* $\{z\colon |z| = 1\}$, *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* $\{z\colon |z| = 1\}$ *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 5.

On the other hand, if (33) holds and $F_\xi(\lambda)$ and $F_\eta(\lambda)$ are the spectral functions of ξ and η, then

$$F_\eta(\lambda) = \int_{-\pi}^{\lambda} |Q(e^{-iv})|^2\, dF_\xi(v) = \frac{1}{2\pi}\int_{-\pi}^{\lambda} |P(e^{-iv})|^2\, dv.$$

Since $|Q(e^{-iv})|^2 > 0$, 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 representation* (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

$$\varphi_n(\lambda) = \frac{1}{n}\sum_{k=0}^{n-1} e^{ik\lambda} = \begin{cases} 1, & \lambda = 0, \\ \frac{1}{n}\frac{e^{in\lambda}-1}{e^{i\lambda}-1}, & \lambda \neq 0. \end{cases} \tag{36}$$

It is clear that $|\varphi_n(\lambda)| \leq 1$.

Moreover, $\varphi_n(\lambda) \xrightarrow{L^2(F)} I_{\{0\}}(\lambda)$, and therefore, by (14) of Sect. 2,

$$\int_{-\pi}^{\pi} \varphi_n(\lambda)\, Z(d\lambda) \xrightarrow{L^2} \int_{-\pi}^{\pi} I_{\{0\}}(\lambda)\, Z(d\lambda) = Z(\{0\}),$$

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

$$\left|\frac{1}{n}\sum_{k=0}^{n-1} R(k)\right|^2 = \left|\mathsf{E}\left(\frac{1}{n}\sum_{k=0}^{n-1}\xi_k\right)\xi_0\right|^2 \leq \mathsf{E}\,|\xi_0|^2\,\mathsf{E}\left|\frac{1}{n}\sum_{k=0}^{n-1}\xi_k\right|^2,$$

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 sequence $\xi = (\xi_n)$ 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 $Z(\{0\}) \neq 0$ with a positive probability and $m = 0$, then ξ_n "contains a random constant α":

$$\xi_n = \alpha + \eta_n,$$

where $\alpha = Z(\{0\})$ and the measure $Z_\eta = Z_\eta(\Delta)$ in the spectral representation $\eta_n = \int_{-\pi}^{\pi} e^{i\lambda n} Z_\eta(d\lambda)$ 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 (13) of Sect. 1.
3. Let $\xi = (\xi_n)$ be a stationary sequence such that $\mathsf{E}\,\xi_n = 0$ and

$$\frac{1}{N^2}\sum_{k=0}^{N-1}\sum_{l=0}^{N-1} R(k-l) = \frac{1}{N}\sum_{|k|\le N-1} R(k)\left[1 - \frac{|k|}{N}\right] \le CN^{-\alpha}$$

for some $C > 0,\ \alpha > 0$. Use the Borel–Cantelli lemma to show that then

$$\frac{1}{N}\sum_{k=0}^{N} \xi_k \to 0 \quad (\mathsf{P}\text{-a.s.}).$$

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 no zeros of $Q_n(z)$ lie on the unit circle.

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 $(\xi_m^1, \xi_m^2, \ldots, \xi_m^n)$, $\xi_m^1 = \xi_m$, satisfying the system of equations

$$\begin{aligned} \xi_{m+1}^i &= \xi_m^{i+1} + \beta_i \varepsilon_{m+1}, \quad i = 1, \ldots, n-1, \\ \xi_{m+1}^n &= -\sum_{j=0}^{n-1} b_{n-j}\xi_m^{j+1} + \beta_n \varepsilon_{m+1}, \end{aligned} \tag{39}$$

where $\beta_1 = a_0,\ \beta_i = a_{i-1} - \sum_{k=1}^{i-1} \beta_k b_{i-k}$.

4. Statistical Estimation of Covariance Function and 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 (e.g., geophysics, medicine, economics). 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 $R(n) = \int_{-\pi}^{\pi} e^{i\lambda n} F(d\lambda)$.

Suppose we have the results $x_0, x_1, \ldots, x_{N-1}$ 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 set

$$m_N(x) = \frac{1}{N}\sum_{k=0}^{N-1} x_k. \tag{1}$$

Then it follows from the elementary properties of the expectation that this is a "good" estimator of m in the sense that "in the average over all possible realizations of data $x_0, \ldots, x_{N-1}$" it is *unbiased*, 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 Sect. 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.,

$$\mathsf{E}\,|m_N(\xi) - m|^2 \to 0, \quad N \to \infty. \tag{3}$$

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

$$\mathsf{E}\,\hat{R}_N(n;\xi) = R(n), \quad 0 \le n < N.$$

Let us now consider the question of its *consistency*. If we replace ξ_k in (37) of Sect. 3 by $\zeta_k = \xi_{n+k}\xi_k$ and suppose that for each integer n the sequence $\zeta = (\zeta_k)_{k\in\mathbb{Z}}$ is wide-sense stationary (which implies, in particular, that $\mathsf{E}\,\xi_0^4 < \infty$), we find that the condition

$$\frac{1}{N}\sum_{k=0}^{N-1}\mathsf{E}[\xi_{n+k}\xi_k - R(n)][\xi_n\xi_0 - R(n)] \to 0, \quad N \to \infty, \tag{4}$$

is necessary and sufficient for

$$\mathsf{E}\,|\hat{R}_N(n;\xi) - R(n)|^2 \to 0, \quad N \to \infty. \tag{5}$$

Let us suppose that the original sequence $\xi = (\xi_n)$ is Gaussian (with zero mean and covariance $R(n)$). Then, proceeding analogously to (51) of Sect. 12, Chap. 2, Vol. 1, we obtain

$$\begin{aligned}\mathsf{E}[\xi_{n+k}\xi_k - R(n)][\xi_n\xi_0 - R(n)] &= \mathsf{E}\,\xi_{n+k}\xi_k\xi_n\xi_0 - R^2(n)\\ &= \mathsf{E}\,\xi_{n+k}\xi_k \cdot \mathsf{E}\,\xi_n\xi_0 + \mathsf{E}\,\xi_{n+k}\xi_n \cdot \mathsf{E}\,\xi_k\xi_0\\ &\quad + \mathsf{E}\,\xi_{n+k}\xi_0 \cdot \mathsf{E}\,\xi_k\xi_n - R^2(n)\\ &= R^2(k) + R(n+k)R(n-k).\end{aligned}$$

Therefore, in the Gaussian case, condition (4) is equivalent to

$$\frac{1}{N}\sum_{k=0}^{N-1}[R^2(k) + R(n+k)R(n-k)] \to 0, \quad N \to \infty. \tag{6}$$

Since $|R(n+k)R(n-k)| \le |R(n+k)|^2 + |R(n-k)|^2$, the condition

$$\frac{1}{N}\sum_{k=0}^{N-1} R^2(k) \to 0, \quad N \to \infty, \tag{7}$$

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

$$\begin{aligned}\frac{1}{N}\sum_{k=0}^{N-1} R^2(k) &= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\frac{1}{N}\sum_{k=0}^{N-1} e^{i(\lambda-\nu)k}\,F(d\lambda)F(d\nu)\\ &= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f_N(\lambda,\nu)\,F(d\lambda)F(d\nu),\end{aligned}$$

where (cf. (36) of Sect. 3)

$$f_N(\lambda,\nu) = \begin{cases} 1, & \lambda = \nu,\\ \frac{1-e^{i(\lambda-\nu)N}}{N[1-e^{i(\lambda-\nu)}]}, & \lambda \ne \nu.\end{cases}$$

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's 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 Sect. 1 has the property that the function

$$F_N(\lambda) = \int_{-\pi}^{\lambda} f_N(\nu)\, d\nu$$

converges on the whole to the spectral function $F(\lambda)$. Therefore, if $F(\lambda)$ has a density $f(\lambda)$, then we have

$$\int_{-\pi}^{\lambda} f_N(\nu)\, d\nu \to \int_{-\pi}^{\lambda} f(\nu)\, d\nu \tag{10}$$

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

$$\hat{f}_N(\lambda;x) = \frac{1}{2\pi}\sum_{|n|<N}\left(1 - \frac{|n|}{N}\right)\hat{R}_N(n;x)e^{-i\lambda n}, \tag{11}$$

setting $\hat{R}_N(n;x) = \hat{R}_N(|n|;x)$ 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 (34) of Sect. 1, we find that

$$\begin{aligned} f_N(\lambda) &= \frac{1}{2\pi N}\sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\int_{-\pi}^{\pi} e^{i\nu(k-l)}e^{i\lambda(l-k)}f(\nu)\,d\nu \\ &= \int_{-\pi}^{\pi}\frac{1}{2\pi N}\left|\sum_{k=0}^{N-1} e^{i(\nu-\lambda)k}\right|^2 f(\nu)\,d\nu. \end{aligned}$$

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)

$$\int_{-\pi}^{\pi}\Phi_N(\lambda-\nu)f(\nu)\,d\nu \to f(\lambda). \tag{13}$$

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 $\hat{f}_N(\lambda; x)$ could be considered "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

$$\hat{f}_N(\lambda;\xi) = \frac{1}{2\pi}\left|\frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}\xi_k e^{-i\lambda k}\right|^2.$$

Therefore for $\lambda = 0$ we have that $2\pi\hat{f}_N(0,\xi)$ coincides 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 $\hat{f}_N(\lambda;x)$ by means of a smoothing function $W_N(\lambda)$, which we call a *spectral window*. Natural requirements on $W_N(\lambda)$ are as follows:

(a) $W_N(\lambda)$ has a sharp maximum at $\lambda = 0$;
(b) $\int_{-\pi}^{\pi} W_N(\lambda)\,d\lambda = 1$;
(c) $\mathsf{P}\,|\hat{f}_N^W(\lambda;\xi) - f(\lambda)|^2 \to 0, \quad N \to \infty,\ \lambda \in [-\pi,\ \pi)$.

By (14) and (b), the estimators $\hat{f}_N^W(\lambda;\xi)$ are asymptotically unbiased. Condition (c) is the condition of 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 uses the spectral window

$$W_N(\lambda) = a_N P(a_N\lambda),$$

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

$$Z(\lambda) = \begin{cases} -\frac{\alpha+1}{2\alpha}|\lambda|^\alpha + \frac{\alpha+1}{2\alpha}, & |\lambda| \le 1, \\ 0, & |\lambda| > 1, \end{cases}$$

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 $\hat{f}_N^W(\lambda; x)$. (See, e.g., [36, 37, 38].)

3. We now consider the problem of estimating the spectral function $F(\lambda) = F([-\pi, \lambda])$. We begin by defining

$$F_N(\lambda) = \int_{-\pi}^{\lambda} f_N(\nu)\, d\nu, \quad \hat{F}_N(\lambda; x) = \int_{-\pi}^{\lambda} \hat{f}_N(\nu; x)\, d\nu,$$

where $\hat{f}_N(\nu; x)$ is the periodogram constructed with $(x_0, x_1, \ldots, x_{N-1})$.

It follows from the proof of Herglotz's theorem (Sect. 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 (cf. corollary to Theorem 1 of Sect. 3, Chap. 3, Vol. 1) 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 1 (or in mean square) as $N \to \infty$, we have

$$\int_{-\pi}^{\pi} e^{i\lambda n} d\hat{F}_N(\lambda; \xi) \to \int_{-\pi}^{\pi} e^{i\lambda n}\, dF(\lambda) \quad (\mathsf{P}\text{-a.s.})$$

and therefore $\hat{F}_N(\lambda; \xi) \Rightarrow F(\lambda)$ (P-a.s.) (or in mean square).

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

$$(N - |n|) \operatorname{Var} \hat{R}_N(n, \xi) \to 2\pi \int_{-\pi}^{\pi} (1 + e^{2in\lambda}) f^2(\lambda)\, d\lambda$$

for every n, as $N \to \infty$.

2. Establish (16) and the following generalization:

$$\lim_{N\to\infty} \operatorname{Cov}(\hat{f}_N(\lambda;\xi),\hat{f}_N(\nu;\xi)) = \begin{cases} 2f^2(0), & \lambda=\nu=0,\ \pm\pi,\\ f^2(\lambda), & \lambda=\nu\neq 0,\ \pm\pi,\\ 0, & \lambda\neq\nu. \end{cases}$$

5. Wold's Expansion

1. In contrast to representation (2) of Sect. 3, 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 notation. 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 $\xi = (\ldots,\xi_{n-1},\xi_n,\ldots)$. Let

$$S(\xi) = \bigcap_n H_n(\xi).$$

For every $\eta \in H(\xi)$, denote by

$$\hat{\pi}_n(\eta) = \hat{\mathsf{E}}(\eta \,|\, H_n(\xi))$$

the projection of η on the subspace $H_n(\xi)$ (Sect. 11, Chap. 2, Vol. 1). We also write

$$\hat{\pi}_{-\infty}(\eta) = \hat{\mathsf{E}}(\eta \,|\, S(\xi)).$$

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 $\operatorname{Var}\xi_n > 0$. 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 1. 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

$$\xi_n^s = \hat{\mathsf{E}}(\xi_n \,|\, S(\xi)), \quad \xi_n^r = \xi_n - \xi_n^s.$$

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 $\xi_n^s \in S(\xi)$; 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 $\xi_n^r \perp S(\xi)$.

This completes the proof of the theorem.

□

Remark 2. Decomposition (1) into regular and singular parts is unique (Problem 4).

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 3. 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 important theorem establishes a connection between one-sided moving averages (Example 4 in Sect. 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 $\beta\xi_n$, where β is a complex number, the dimension of B_n is either zero or one. But the space $H_n(\xi)$ is different from $H_{n-1}(\xi)$ 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, hence $H(\xi) = S(\xi)$, 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. Set

$$\varepsilon_n = \frac{\eta_n}{\|\eta_n\|},$$

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 (6), Sect. 11, Chap. 2, Vol. 1,

$$\sum_{j=0}^{\infty} |a_j|^2 \leq \|\xi_n\|^2 < \infty.$$

It follows that $\sum_{j=0}^{\infty} a_i \varepsilon_{n-j}$ converges in mean square, and then, by (4), Eq. (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 $n = 0$. Let $\hat{\pi}_i = \hat{\pi}_i(\xi_0)$. 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 $k \geq 0$

$$\sum_{i=0}^{k} \|\hat{\pi}_{-i} - \hat{\pi}_{-i+1}\|^2 = \left\| \sum_{i=0}^{k} (\hat{\pi}_{-i} - \hat{\pi}_{-i+1}) \right\|^2$$
$$= \|\hat{\pi}_{-k} - \hat{\pi}_0\|^2 \le 4\|\xi_0\|^2 < \infty.$$

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 $\bigcap_{k\ge 0} H_{-k}(\xi) = S(\xi)$. 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 4. 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 (see the definition in Example 4 of Sect. 1). In this sense, the conclusion of Theorem 2 says more, 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 $H_0(\xi) = \overline{L^2}(\xi^0)$ 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 Sect. 11, Chap. 2, Vol. 1, 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),

$$\begin{aligned}\xi_n &= \hat{\mathsf{E}}(\xi_n^s + \xi_n^r \,|\, H_0(\xi)) = \hat{\mathsf{E}}(\xi_n^s \,|\, H_0(\xi)) + \hat{\mathsf{E}}(\xi_n^r \,|\, H_0(\xi)) \\ &= \hat{\mathsf{E}}(\xi_n^s \,|\, H_0(\xi^r) \oplus H_0(\xi^s)) + \hat{\mathsf{E}}(\xi_n^r \,|\, H_0(\xi^r) \oplus H_0(\xi^s)) \\ &= \hat{\mathsf{E}}(\xi_n^s \,|\, H_0(\xi^s)) + \hat{\mathsf{E}}(\xi_n^r \,|\, H_0(\xi^r)) \\ &= \xi_n^s + \hat{\mathsf{E}}\Big(\sum_{k=0}^{\infty} a_k \varepsilon_{n-k} \,|\, H_0(\xi^r)\Big).\end{aligned}$$

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

$$\hat{\xi}_n = \xi_n^s + \hat{\mathsf{E}}\Big(\sum_{k=0}^{\infty} a_k \varepsilon_{n-k} \,|\, H_0(\varepsilon)\Big) = \xi_n^s + \sum_{k=n}^{\infty} a_k \varepsilon_{n-k} \tag{8}$$

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

$$\hat{\xi}_n \xrightarrow{L^2} 0, \quad n \to \infty,$$

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

$$H_n(\xi) = H_n(\varepsilon), \quad n \in \mathbb{Z}. \tag{12}$$

The representation (11) means (Subsection 3, Sect. 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 the properties of the spectral density.

In the first place, it is clear that

$$f(\lambda) = \frac{1}{2\pi} |\varphi(\lambda)|^2,$$

where

$$\varphi(\lambda) = \sum_{k=0}^{\infty} e^{-i\lambda k} a_k, \quad \sum_{k=0}^{\infty} |a_k|^2 < \infty. \tag{13}$$

Set

$$\Phi(z) = \sum_{k=0}^{\infty} a_k z^k. \tag{14}$$

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

$$\sup_{0 \leq r < 1} \frac{1}{2\pi} \int_{-\pi}^{\pi} |g(re^{i\theta})|^2 d\theta < \infty. \tag{15}$$

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

$$\sup_{0 \leq r < 1} \sum |a_k|^2 r^{2k} \leq \sum |a_k|^2 < \infty.$$

It is shown in the theory of functions of a complex variable (e.g., [64]) that the boundary function $\Phi(e^{i\lambda})$, $-\pi \leq \lambda < \pi$, of $\Phi \in H^2$, not identically zero, has the property that

$$\int_{-\pi}^{\pi} \log |\Phi(e^{-i\lambda})| \, d\lambda > -\infty. \tag{16}$$

In our case,

$$f(\lambda) = \frac{1}{2\pi} |\Phi(e^{-i\lambda})|^2,$$

where $\Phi \in H^2$. Therefore

$$\log f(\lambda) = -\log 2\pi + 2\log|\Phi(e^{-i\lambda})|,$$

and consequently the spectral density $f(\lambda)$ of a regular process satisfies

$$\int_{-\pi}^{\pi} \log f(\lambda)\, d\lambda > -\infty. \tag{17}$$

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 set $\varphi(\lambda) = \Phi(e^{-i\lambda})$, we obtain

$$f(\lambda) = \frac{1}{2\pi}|\varphi(\lambda)|^2,$$

where $\varphi(\lambda)$ is given by (13). Then it follows from Corollary 5, Sect. 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 Remark 4 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*

$$\int_{-\pi}^{\pi} \log f(\lambda)\, d\lambda > -\infty. \tag{18}$$

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 its mean-square error σ_n^2, and show that the *nonlinear* estimator

$$\tilde{\xi}_n = \left(\frac{\xi_0}{\xi_{-1}}\right)^n$$

provides an *error-free* prediction of ξ_n by the "past" $\xi^0 = (\dots, \xi_{-1}, \xi_0)$, i.e.,

$$\mathsf{E}\,|\tilde{\xi}_n - \xi_n|^2 = 0, \quad n \geq 1.$$

4. Prove that decomposition (1) into regular and singular components is unique.

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," $\xi^0 = (\dots, \xi_{-1}, \xi_0)$. 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 Sect. 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 Sect. 5 that the representation (1) solves the problem of finding the optimal (linear) estimator $\hat{\xi}_n = \hat{\mathsf{E}}(\xi_n \,|\, H_0(\xi))$ since, by (8) of Sect. 5,

$$\hat{\xi}_n = \sum_{k=n}^{\infty} a_k \varepsilon_{n-k} \tag{2}$$

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 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 the ε_k in terms of $\dots, \xi_{k-1}, \xi_k$.

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 form (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 Remark 4 on Theorem 2 of Sect. 3, every variable $\tilde{\xi}_n \in H_0(\xi)$ admits a representation in the form

$$\tilde{\xi}_n = \int_{-\pi}^{\pi} \tilde{\varphi}_n(\lambda)\, Z(d\lambda), \quad \tilde{\varphi}_n \in H_0(F), \tag{8}$$

where $H_0(F)$ is the closed linear manifold spanned by the functions $e_n = e^{i\lambda n}$ for $n \le 0$ $\left(F(\lambda) = \int_{-\pi}^{\lambda} f(\nu)\, d\nu\right)$.

Since (Sect. 2)

$$\begin{aligned} \mathsf{E}\,|\xi_n - \tilde{\xi}_n|^2 &= \mathsf{E}\left|\int_{-\pi}^{\pi} (e^{i\lambda n} - \tilde{\varphi}_n(\lambda))\, Z(d\lambda)\right|^2 \\ &= \int_{-\pi}^{\pi} |e^{i\lambda n} - \tilde{\varphi}_n(\lambda)|^2 f(\lambda)\, d\lambda, \end{aligned}$$

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 (Sect. 11, Chap. 2, Vol. 1) that the optimal function $\hat{\varphi}_n(\lambda)$ (in the sense of (9)) is determined by the two conditions

$$\begin{aligned} &\text{(i) } \hat{\varphi}_n(\lambda) \in H_0(F), \\ &\text{(ii) } e^{i\lambda n} - \hat{\varphi}_n(\lambda) \perp H_0(F). \end{aligned} \tag{10}$$

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.,

$$I_{n,m} \equiv \int_{-\pi}^{\pi} [e^{i\lambda n} - \hat{\varphi}_n(\lambda)]e^{i\lambda m} f(\lambda)\, d\lambda = 0, \quad m \geq 0.$$

The following chain of equations shows that this is actually the case:

$$\begin{aligned} I_{n,m} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\lambda(n+m)}\left[1 - \frac{\Phi_n(e^{-i\lambda})}{\Phi(e^{-i\lambda})}\right]|\Phi(e^{-i\lambda})|^2\, d\lambda \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\lambda(n+m)}[\Phi(e^{-i\lambda}) - \Phi_n(e^{-i\lambda})]\overline{\Phi(e^{-i\lambda})}\, d\lambda \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\lambda(n+m)}\left(\sum_{k=0}^{n-1} b_k e^{-i\lambda k}\right)\left(\sum_{l=0}^{\infty}\overline{b}_l e^{i\lambda l}\right)d\lambda \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\lambda m}\left(\sum_{k=0}^{n-1} b_k e^{i\lambda(n-k)}\right)\left(\sum_{l=0}^{\infty}\overline{b}_l e^{i\lambda l}\right)d\lambda = 0, \end{aligned}$$

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

$$\hat{\varphi}_n(\lambda) = C_0 + C_{-1}e^{-i\lambda} + C_{-2}e^{-2i\lambda} + \cdots,$$

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 $\{z\colon |z| \leq 1\}$.

In fact, in this case it is enough to set $\Phi(z) = P(z)/Q(z)$. 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

$$R(0) = 5, \quad R(\pm 1) = 2, \quad R(n) = 0 \ \text{ for } |n| \geq 2. \tag{11}$$

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

$$\hat{\varphi}_1(\lambda) = e^{i\lambda}\frac{e^{-i\lambda}}{2 + e^{-i\lambda}}, \quad \hat{\varphi}_n(\lambda) = 0 \ \text{ for } n \geq 2. \tag{12}$$

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

$$\begin{aligned}\hat{\xi}_1 &= \int_{-\pi}^{\pi} e^{i\lambda}\frac{e^{-i\lambda}}{2 + e^{-i\lambda}}\, Z(d\lambda) \\ &= \frac{1}{2}\int_{-\pi}^{\pi} \frac{1}{(1 + \frac{1}{2}e^{-i\lambda})}\, Z(d\lambda) = \sum_{k=0}^{\infty} \frac{(-1)^k}{2^{k+1}} \int_{-\pi}^{\pi} e^{-ik\lambda}\, Z(d\lambda) \\ &= \sum_{k=0}^{\infty} \frac{(-1)^k \xi_k}{2^{k+1}} = \frac{1}{2}\xi_0 - \frac{1}{4}\xi_{-1} + \cdots .\end{aligned}$$

EXAMPLE 2. Let the covariance function be

$$R(n) = a^n, \quad |a| < 1.$$

Then (see Example 5 in Sect. 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 $\hat{\varphi}_n(\lambda) = a^n$, and therefore

$$\hat{\xi}_n = \int_{-\pi}^{\pi} a^n \, Z(d\lambda) = a^n \xi_0.$$

In other words, 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 a 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 $|z| \le 1$ 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 3) 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

$$\xi_n = \xi_n^s + \xi_n^r, \quad \xi_n = \int_{-\pi}^{\pi} e^{i\lambda n} \, Z(d\lambda), \quad F(\Delta) = \mathsf{E}\,|Z(\Delta)|^2,$$

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 of Sect. 5)

$$\hat{\xi}_n = \int_{-\pi}^{\pi} \hat{\varphi}_n(\lambda)\, Z(d\lambda), \quad \hat{\xi}_n^r = \int_{-\pi}^{\pi} \hat{\varphi}_n^r(\lambda)\, Z^r(d\lambda),$$

where $Z^r(\Delta)$ is the orthogonal stochastic measure in the representation of the regular sequence ξ^r. Then

$$\begin{aligned} \mathsf{E}\,|\xi_n - \hat{\xi}_n|^2 &= \int_{-\pi}^{\pi} |e^{i\lambda n} - \hat{\varphi}_n(\lambda)|^2\, F(d\lambda) \\ &\geq \int_{-\pi}^{\pi} |e^{i\lambda n} - \hat{\varphi}_n(\lambda)|^2 f^r(\lambda)\, d\lambda \geq \int_{-\pi}^{\pi} |e^{i\lambda n} - \hat{\varphi}_n^r(\lambda)|^2 f^r(\lambda)\, d\lambda \\ &= \mathsf{E}\,|\xi_n^r - \hat{\xi}_n^r|^2. \end{aligned} \tag{16}$$

But $\xi_n - \hat{\xi}_n = \hat{\xi}_n^r - \hat{\xi}^r$. 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 for ξ_0 from the results of the measurements $\{\xi_n,\ n = \pm 1, \pm 2, \ldots\}$ with omitted ξ_0.

Let $H^0(\xi)$ be the closed linear manifold spanned by $\xi_n,\ n \neq 0$. Then, according to Theorem 2 of Sect. 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

$$\begin{aligned} \inf_{\eta \in H^0(\xi)} \mathsf{E}\,|\xi_0 - \eta|^2 &= \inf_{\varphi \in H^0(F)} \int_{-\pi}^{\pi} |1 - \varphi(\lambda)|^2\, F(d\lambda) \\ &= \int_{-\pi}^{\pi} |1 - \check{\varphi}(\lambda)|^2\, F(d\lambda) = \mathsf{E}\,|\xi_0 - \check{\xi}_0|^2. \end{aligned}$$

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

$$\begin{aligned} &\text{(i) } \check{\varphi}(\lambda) \in H^0(F), \\ &\text{(ii) } 1 - \check{\varphi}(\lambda) \perp H^0(F). \end{aligned} \tag{18}$$

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

$$\alpha = \frac{2\pi}{\int_{-\pi}^{\pi} \frac{d\lambda}{f(\lambda)}}, \tag{21}$$

and the interpolation error $\delta^2 = \mathsf{E}\,|\xi_0 - \check{\xi}_0|^2$ is given by $\delta^2 = 2\pi\alpha$.

PROOF. We shall give the proof only under very stringent hypotheses on the spectral density, specifically that

$$0 < c \le f(\lambda) \le C < \infty. \tag{22}$$

It follows from (2) in (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 10, Sect. 12, Chap. 2, Vol. 1). 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

$$\check{\varphi}(\lambda) = 1 - \frac{\alpha}{f(\lambda)}. \tag{24}$$

Starting from the first condition (18), we now determine α.

By (22), we have $\check{\varphi} \in L^2$, and the condition $\check{\varphi} \in H^0(F)$ is equivalent to the condition that $\check{\varphi}$ 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 $\check{\varphi}(\lambda)$ must be zero. Therefore

$$0 = \int_{-\pi}^{\pi} \check{\varphi}(\lambda)\, d\lambda = 2\pi - \alpha \int_{-\pi}^{\pi} \frac{d\lambda}{f(\lambda)}$$

and hence α is determined by (21).

Finally,

$$\delta^2 = \mathsf{E}\,|\xi_0 - \check{\xi}_0|^2 = \int_{-\pi}^{\pi} |1 - \check{\varphi}(\lambda)|^2 f(\lambda)\, d\lambda$$

$$= |\alpha|^2 \int_{-\pi}^{\pi} \frac{f(\lambda)}{f^2(\lambda)}\, d\lambda = \frac{4\pi^2}{\int_{-\pi}^{\pi} \frac{d\lambda}{f(\lambda)}}.$$

This completes the proof (under condition (22)).

□

Corollary. *If*

$$\check{\varphi}(\lambda) = \sum_{0<|k|\le N} c_k e^{i\lambda k},$$

then

$$\check{\xi}_0 = \sum_{0<|k|\le N} c_k \int_{-\pi}^{\pi} e^{i\lambda k}\, Z(d\lambda) = \sum_{0<|k|\le N} c_k \xi_k.$$

EXAMPLE 3. Let $f(\lambda)$ be the spectral density in Example 2 above. Then an easy calculation shows that

$$\check{\xi}_0 = \int_{-\pi}^{\pi} \frac{a}{1+|a|^2}[e^{i\lambda} + e^{-i\lambda}]\, Z(d\lambda) = \frac{a}{1+|a|^2}[\xi_1 + \xi_{-1}],$$

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 observed components. Each of the sequences θ and ξ will be supposed stationary (wide sense) with zero means and spectral representations

$$\theta_n = \int_{-\pi}^{\pi} e^{i\lambda n}\, Z_\theta(d\lambda) \quad \text{and} \quad \xi_n = \int_{-\pi}^{\pi} e^{i\lambda n}\, Z_\xi(d\lambda).$$

We write

$$F_\theta(\Delta) = \mathsf{E}\,|Z_\theta(\Delta)|^2, \quad F_\xi(\Delta) = \mathsf{E}\,|Z_\xi(\Delta)|^2$$

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 function $\mathrm{Cov}(\theta_n, \xi_m) = \mathsf{E}\,\theta_n\overline{\xi}_m$ depends only on the difference $n-m$. Let $R_{\theta\xi}(n) = \mathsf{E}\,\theta_n\overline{\xi}_0$; then

$$R_{\theta\xi}(n) = \int_{-\pi}^{\pi} e^{i\lambda n}\, F_{\theta\xi}(d\lambda).$$

The filtering problem of interest 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 $\hat{\theta}_n = \hat{\mathsf{E}}(\theta_n \,|\, H(\xi))$, 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

(i) $\hat{\varphi}_n(\lambda) \in H(F_\xi)$,
(ii) $\theta_n - \hat{\theta}_n \perp H(\xi)$.

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

$$\hat{\varphi}(\lambda) = f_{\theta\xi}(\lambda) \cdot f_\xi^{\oplus}(\lambda)$$

and $f_\xi^{\oplus}(\lambda)$ is the "pseudoinverse" of $f_\xi(\lambda)$, i.e.,

$$f_\xi^{\oplus}(\lambda) = \begin{cases} [f_\xi(\lambda)]^{-1}, & f_\xi(\lambda) > 0, \\ 0, & f_\xi(\lambda) = 0. \end{cases}$$

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, $\hat{\varphi} \in H(F_\xi)$, 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 $f_n(\lambda)$. Then

$$\hat{\theta}_n = \int_{-\pi}^{\pi} e^{i\lambda n}\hat{\varphi}(\lambda)\, Z_\xi(d\lambda),$$

where

$$\hat{\varphi}(\lambda) = f_\theta(\lambda)[f_0(\lambda) + f_\eta(\lambda)]^{\oplus},$$

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 earlier can now be used to construct an optimal estimator $\tilde{\theta}_{n+m}$ of θ_{n+m} based on observations ξ_k, $k \le n$, where m is a given number in $\mathbb{Z}$. 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 $\varepsilon = (\varepsilon_n)$ is white noise with the spectral decomposition

$$\varepsilon_n = \int_{-\pi}^{\pi} e^{i\lambda n}\, Z_\varepsilon(d\lambda).$$

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

$$\hat{\theta}_{n+m} = \int_{-\pi}^{\pi} e^{i\lambda(n+m)}\hat{\varphi}(\lambda)\Phi(e^{-i\lambda})\, Z_\varepsilon(d\lambda) = \sum_{k=-\infty}^{\infty} \hat{a}_{n+m-k}\varepsilon_k,$$

where

$$\hat{a}_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\lambda k}\hat{\varphi}(\lambda)\Phi(e^{-i\lambda})\, d\lambda, \tag{29}$$

we have

$$\tilde{\theta}_{n+m} = \hat{\mathsf{E}}\left[\sum_{k=-\infty}^{\infty} \hat{a}_{n+m-k}\varepsilon_k \,|\, H_n(\xi)\right].$$

But $H_n(\xi) = H_n(\varepsilon)$, and therefore

$$\begin{aligned}\tilde{\theta}_{n+m} &= \sum_{k\le n} \hat{a}_{n+m-k}\varepsilon_k = \int_{-\pi}^{\pi}\left[\sum_{k\le n}\hat{a}_{n+m-k}e^{i\lambda k}\right] Z_\varepsilon(d\lambda)\\ &= \int_{-\pi}^{\pi} e^{i\lambda n}\left[\sum_{l=0}^{\infty}\hat{a}_{l+m}e^{-i\lambda l}\right]\Phi^{\oplus}(e^{-i\lambda})\, Z_\xi(d\lambda),\end{aligned}$$

where $\Phi^{\oplus}$ is the pseudoinverse 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. Show that the conclusion of Theorem 1 remains valid even without the hypotheses that $\Phi(z)$ has a radius of convergence $r > 1$ and that the zeros of $\Phi(z)$ all lie in $|z| > 1$.
2. Show that, for a regular process, the function $\Phi(z)$ involved in (4) can be represented in the form

$$\Phi(z) = \sqrt{2\pi} \exp\left\{\frac{1}{2}c_0 + \sum_{k=1}^{\infty} c_k z^k\right\}, \quad |z| < 1,$$

where

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\lambda} \log f(\lambda)\, d\lambda.$$

Deduce from this formula that the one-step prediction error $\sigma_1^2 = \mathsf{E}\,|\hat{\xi}_1 - \xi_1|^2$ is given by the *Szegő–Kolmogorov formula*

$$\sigma_1^2 = 2\pi \exp\left\{\frac{1}{2\pi} \int_{-\pi}^{\pi} \log f(\lambda)\, d\lambda\right\}.$$

3. Prove Theorem 2 without assuming (22).
4. 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 earlier 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 *analog* devices. 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 its generalizations in which the recurrent equations for (θ, ξ) have coefficients that may 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

$$\theta_n = (\theta_1(n), \dots, \theta_k(n)) \quad \text{and} \quad \xi_n = (\xi_1(n), \dots, \xi_l(n))$$

be governed by the recurrent equations

$$\begin{aligned} \theta_{n+1} = &\quad a_0(n,\xi) + a_1(n,\xi)\theta_n + b_1(n,\xi)\varepsilon_1(n+1) + b_2(n,\xi)\varepsilon_2(n+1), \\ \xi_{n+1} = &\quad A_0(n,\xi) + A_1(n,\xi)\theta_n + B_1(n,\xi)\varepsilon_1(n+1) + B_2(n,\xi)\varepsilon_2(n+1). \end{aligned} \tag{1}$$

Here

$$\varepsilon_1(n) = (\varepsilon_{11}(n), \dots, \varepsilon_{1k}(n)) \quad \text{and} \quad \varepsilon_2(n) = (\varepsilon_{21}(n), \dots, \varepsilon_{2l}(n))$$

are independent Gaussian vectors with independent components, each of which is normally distributed with parameters 0 and 1; $a_0(n,\xi) = (a_{01}(n,\xi), \dots, \alpha_{0k}(n,\xi))$ and $A_0(n,\xi) = (A_{01}(n,\xi), \dots, A_{0l}(n,\xi))$ are vector functions with nonanticipative dependence on $\xi = \{\xi_0, \dots, \xi_n)$, i.e., for a given n the functions $a_0(n,\xi), \dots, A_{0l}(n,\xi)$ depend only on $\xi_0, \dots, \xi_n$; the matrix functions

$$\begin{aligned} b_1(n,\xi) &= \|b_{ij}^{(1)}(n,\xi)\|, \quad b_2(n,\xi) = \|b_{ij}^{(2)}(n,\xi)\|, \\ B_1(n,\xi) &= \|B_{ij}^{(1)}(n,\xi)\|, \quad B_2(n,\xi) = \|B_{ij}^{(2)}(n,\xi)\|, \\ a_1(n,\xi) &= \|a_{ij}^{(1)}(n,\xi)\|, \quad A_1(n,\xi) = \|A_{ij}^{(1)}(n,\xi)\| \end{aligned}$$

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 nonanticipatively. 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 $\mathsf{E}(\|\theta_0\|^2 + \|\xi_0\|^2) < \infty$ $\left(\text{with } \|x\|^2 = \sum_{i=1}^k x_i^2 \text{ for } x = (x_1, \dots, x_k)\right)$, $|a_{ij}^{(1)}(n,\xi)| \le C$, $|A_{ij}^{(1)}(n,\xi)| \le C$, 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 $\mathsf{E}\,|g(n,\xi)|^2 < \infty$, $n = 0, 1, \dots$. With these assumptions, (θ, ξ) has $\mathsf{E}(\|\theta_n\|^2 + \|\xi_n\|^2) < \infty$, $n \ge 0$.

Now let $\mathscr{F}_n^\xi = \sigma\{\xi_0, \dots, \xi_n\}$ be the smallest σ-algebra generated by $\xi_0, \dots, \xi_n$ and

$$m_n = \mathsf{E}(\theta_n \mid \mathscr{F}_n^\xi), \quad \gamma_n = \mathsf{E}[(\theta_n - m_n)(\theta_n - m_n)^* \mid \mathscr{F}_n^\xi].$$

According to Theorem 1, Sect. 8, Chap. 2, Vol. 1, $m_n = (m_1(n), \dots, m_k(n))$ is an optimal estimator (in the mean-square sense) for the vector $\theta_n = (\theta_1(n), \dots, \theta_k(n))$, and $\mathsf{E}\,\gamma_n = \mathsf{E}[(\theta_n - m_n)(\theta_n - m_n)^*]$ is the matrix of errors of observation. Determining these matrices for arbitrary sequences (θ, ξ) governed by Eqs. (1) is a very difficult problem. However, under a further supplementary condition on (θ_0, ξ_0), namely, that the conditional distribution $\mathsf{P}(\theta_0 \le a \,|\, \xi_0)$ is Gaussian,

$$\mathsf{P}(\theta_0 \le a \,|\, \xi_0) = \frac{1}{\sqrt{2\pi\gamma_0}} \int_{-\infty}^{a} \exp\left\{-\frac{(x - m_0)^2}{2\gamma_0}\right\} dx, \tag{2}$$

with parameters $m_0 = m_0(\xi_0)$, $\gamma_0 = \gamma_0(\xi_0)$, we can derive a system of recurrent equations for m_n and γ_m that also include the *Kalman–Bucy filter* equations.

To begin with, let us establish an important auxiliary result.

Lemma 1. *Under the assumptions made earlier about the coefficients of* (1), *together with* (2), *the sequence* (θ, ξ) *is conditionally Gaussian, i.e., the conditional distribution function*

$$\mathsf{P}\{\theta_0 \le a_0, \dots, \theta_n \le a_n \,|\, \mathscr{F}_n^{\xi}\}$$

is (P-*a.s.*) *the distribution function of an n-dimensional Gaussian vector whose mean and covariance matrix depend on* $(\xi_0, \dots, \xi_n)$.

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 $t = (t_1, \dots, t_{k+l})$. 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

$$\mathsf{E}[\exp(it^*\mathbb{A}_1(n,\xi)\theta_n) \,|\, \mathscr{F}_n^{\xi}] = \exp\left\{it^*\mathbb{A}_1(n,\xi)m_n - \tfrac{1}{2}t^*(\mathbb{A}_1(n,\xi)\gamma_n\mathbb{A}_1^*(n,\xi))t\right\}. \tag{4}$$

Let us show that (4) is also valid when n is replaced by $n + 1$.

From (3) and (4) we have

$$\mathsf{E}[\exp(it^*\zeta_{n+1}) \,|\, \mathscr{F}_n^{\xi}] = \exp\big\{it^*(\mathbb{A}_0(n,\xi) + \mathbb{A}_1(n,\xi)m_n) - \tfrac{1}{2}t^*\mathbb{B}(n,\xi)t - \tfrac{1}{2}t^*(\mathbb{A}_1(n,\xi)\gamma_n\mathbb{A}_1^*(n,\xi))t\big\}.$$

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 in Sect. 13, Chap. 2, Vol. 1) 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.)

$$\mathsf{E}[\eta(\xi_{n+1} - \mathsf{E}(\xi_{n+1} \,|\, \mathscr{F}_n^{\xi}))^* \,|\, \mathscr{F}_n^{\xi}] = 0.$$

This implies that the conditionally Gaussian vectors η and ξ_{n+1}, considered under the condition $\mathscr{F}_n^{\xi}$, are independent, i.e., (P-a.s.)

$$\mathsf{P}(\eta \in A,\ \xi_{n+1} \in B \,|\, \mathscr{F}_n^{\xi}) = \mathsf{P}(\eta \in A \,|\, \mathscr{F}_n^{\xi}) \cdot \mathsf{P}(\xi_{n+1} \in B \,|\, \mathscr{F}_n^{\xi})$$

for all $A \in \mathscr{B}(R^k),\ B \in \mathscr{B}(R^l)$.

Therefore, if $s = (s_1, \ldots, s_n)$, then

$$\begin{aligned}
&\mathsf{E}[\exp(is^*\theta_{n+1}) \,|\, \mathscr{F}_n^{\xi},\ \xi_{n+1}] \\
&\quad = \mathsf{E}\{\exp(is^*[\mathsf{E}(\theta_{n+1} \,|\, \mathscr{F}_n^{\xi}) + \eta + C[\xi_{n+1} - \mathsf{E}(\xi_{n+1} \,|\, \mathscr{F}_n^{\xi})]]) \,|\, \mathscr{F}_n^{\xi},\ \xi_{n+1}\} \\
&\quad = \exp\{is^*[\mathsf{E}(\theta_{n+1} \,|\, \mathscr{F}_n^{\xi}) + C[\xi_{n+1} - \mathsf{E}(\xi_{n+1} \,|\, \mathscr{F}_n^{\xi})]]\} \\
&\qquad \times \mathsf{E}[\exp(is^*\eta) \,|\, \mathscr{F}_n^{\xi},\ \xi_{n+1}] \\
&\quad = \exp\{is^*[\mathsf{E}(\theta_{n+1} \,|\, \mathscr{F}_n^{\xi}) + C[\xi_{n+1} - \mathsf{E}(\xi_{n+1} \,|\, \mathscr{F}_n^{\xi})]]\} \\
&\qquad \times \mathsf{E}(\exp(is^*\eta) \,|\, \mathscr{F}_n^{\xi}).
\end{aligned} \tag{6}$$

By (5), the conditional distribution $\mathsf{P}(\eta \le y \,|\, \mathscr{F}_n^{\xi})$ is Gaussian. With (6), this shows that the conditional distribution $\mathsf{P}(\theta_{n+1} \le a \,|\, \mathscr{F}_{n+1}^{\xi})$ is also Gaussian.

This completes the proof of the lemma.

□

Theorem 1. *Let (θ, ξ) be a partially observed sequence that satisfies the system (1) and condition (2). Then (m_n, γ_n) obey the following recursion relations*:

$$\begin{aligned}
m_{n+1} &= [a_0 + a_1 m_n] + [b \circ B + a_1\gamma_n A_1^*][B \circ B + A_1\gamma_n A_1^*]^{\oplus} \\
&\quad \times [\xi_{n+1} - A_0 - A_1 m_n],
\end{aligned} \tag{7}$$

$$\begin{aligned}
\gamma_{n+1} &= [a_1\gamma_n a_1^* + b \circ b] - [b \circ B + a_1\gamma_n A_1^*][B \circ B + A_1\gamma_n A_1^*]^{\oplus} \\
&\quad \times [b \circ B + a_1\gamma_n A_1^*]^*.
\end{aligned} \tag{8}$$

PROOF. From (1),

$$\mathsf{E}(\theta_{n+1} \mid \mathscr{F}_n^{\xi}) = a_0 + a_1 m_n, \quad \mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^{\xi}) = A_0 + A_1 m_n \tag{9}$$

and

$$\begin{aligned} \theta_{n+1} - \mathsf{E}(\theta_{n+1} \mid \mathscr{F}_n^{\xi}) &= a_1[\theta_n - m_n] + b_1\varepsilon_1(n+1) + b_2\varepsilon_2(n+1), \\ \xi_{n+1} - \mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^{\xi}) &= A_1[\theta_n - m_n] + B_1\varepsilon_1(n+1) + B_2\varepsilon_2(n+1). \end{aligned} \tag{10}$$

Let us write

$$\begin{aligned} d_{11} &= \operatorname{Cov}(\theta_{n+1}, \theta_{n+1} \mid \mathscr{F}_n^{\xi}) \\ &= \mathsf{E}\{[\theta_{n+1} - \mathsf{E}(\theta_{n+1} \mid \mathscr{F}_n^{\xi})][\theta_{n+1} - \mathsf{E}(\theta_{n+1} \mid \mathscr{F}_n^{\xi})]^* \mid \mathscr{F}_n^{\xi}\}, \\ d_{12} &= \operatorname{Cov}(\theta_{n+1}, \xi_{n+1} \mid \mathscr{F}_n^{\xi}) \\ &= \mathsf{E}\{[\theta_{n+1} - \mathsf{E}[\theta_{n+1} \mid \mathscr{F}_n^{\xi})][\xi_{n+1} - \mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^{\xi})]^* \mid \mathscr{F}_n^{\xi}\}, \\ d_{22} &= \operatorname{Cov}(\xi_{n+1}, \xi_{n+1} \mid \mathscr{F}_n^{\xi}) \\ &= \mathsf{E}\{[\xi_{n+1} - \mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^{\xi})][\xi_{n+1} - \mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^{\xi})]^* \mid \mathscr{F}_n^{\xi}\}. \end{aligned}$$

Then, by (10),

$$d_{11} = a_1\gamma_n a_1^* + b \circ b, \quad d_{12} = a_1\gamma_n A_1^* + b \circ B, \quad d_{22} = A_1\gamma_n A_1^* + B \circ B. \tag{11}$$

By the theorem on normal correlation (see Theorem 2 and Problem 4 in Sect. 13, Chap. 2, Vol. 1),

$$m_{n+1} = \mathsf{E}(\theta_{n+1} \mid \mathscr{F}_n^{\xi}, \xi_{n+1}) = \mathsf{E}(\theta_{n+1} \mid \mathscr{F}_n^{\xi}) + d_{12}d_{22}^{\oplus}(\xi_{n+1} - \mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^{\xi}))$$

and

$$\gamma_{n+1} = \operatorname{Cov}(\theta_{n+1}, \theta_{n+1} \mid \mathscr{F}_n^{\xi}, \xi_{n+1}) = d_{11} - d_{12}d_{22}^{\oplus}d_{12}^*.$$

If we then use the expressions from (9) for $\mathsf{E}(\theta_{n+1} \mid \mathscr{F}_n^{\xi})$ and $\mathsf{E}(\xi_{n+1} \mid \mathscr{F}_n^{\xi})$ 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 7. *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 Eqs. (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 8. *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

$$\begin{aligned} A_0 &= \tilde{A}_0 + \tilde{A}_1 a_0, \quad & A_1 &= \tilde{A}_1 a_1, \\ B_1 &= \tilde{A}_1 b_1 + \tilde{B}_1, \quad & B_2 &= \tilde{A}_2 b_2 + \tilde{B}_2, \end{aligned}$$

we find that the case under consideration also obeys the model (1) *and that* m_n *and* γ_n *satisfy* (7) *and* (8).

2. We now consider a *linear* model (cf. (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 $a_0, \dots, B_2$ 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 with initial values (θ_0, ξ_0) such 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 $\gamma = \operatorname{Cov}(\theta_0, \theta_0 \,|\, \xi_0) = \mathsf{E}\,\gamma_0$. 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 $\xi_0, \xi_1, \dots, \xi_n$.

This remark makes it possible to prove the following important statement about the structure of the optimal linear filter without the assumption that the random variables involved are 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* $\hat{m}_n = \hat{\mathsf{E}}(\theta_n \,|\, \xi_0, \dots, \xi_n)$ *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*

$$\hat{\gamma}_n = \mathsf{E}[(\theta_n - \hat{m}_n)(\theta_n - \hat{m}_n)^*]$$

satisfies (8) *with initial values*

$$\begin{aligned}\hat{m}_0 &= \operatorname{Cov}(\theta_0, \xi_0)\operatorname{Cov}^{\oplus}(\xi_0, \xi_0)\cdot\xi_0, \\ \hat{\gamma}_0 &= \operatorname{Cov}(\theta_0, \theta_0) - \operatorname{Cov}(\theta_0, \xi_0)\operatorname{Cov}^{\oplus}(\xi_0, \xi_0)\operatorname{Cov}^{*}(\theta_0, \xi_0).\end{aligned} \tag{14}$$

For the proof of this theorem, 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* $\mathsf{E}(\alpha^2 + \beta^2) < \infty$, and $(\tilde{\alpha}, \tilde{\beta})$ *a two-dimensional Gaussian vector with the same first and second moments as* (α, β), *i.e.*,

$$\mathsf{E}\,\tilde{\alpha}^i = \mathsf{E}\,\alpha^i, \quad \mathsf{E}\,\tilde{\beta}^i = \mathsf{E}\,\beta^i, \quad i = 1, 2; \quad \mathsf{E}\,\tilde{\alpha}\tilde{\beta} = \mathsf{E}\,\alpha\beta.$$

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 $\overline{\lambda}(b)$ be any other linear estimator. Then

$$\mathsf{E}[\tilde{\alpha} - \overline{\lambda}(\tilde{\beta})]^2 \geq \mathsf{E}[\tilde{\alpha} - \lambda(\tilde{\beta})]^2$$

and since $\overline{\lambda}(b)$ and $\lambda(b)$ are linear and the hypotheses of the lemma are satisfied, we have

$$\mathsf{E}[\alpha - \overline{\lambda}(\beta)]^2 = \mathsf{E}[\tilde{\alpha} - \overline{\lambda}(\tilde{\beta})]^2 \geq \mathsf{E}[\tilde{\alpha} - \lambda(\tilde{\beta})]^2 = \mathsf{E}[\alpha - \lambda(\beta)]^2,$$

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

$$\begin{aligned} \tilde{\theta}_{n+1} &= a_0 + a_1\tilde{\theta}_n + a_2\tilde{\xi}_n + b_1\tilde{\varepsilon}_{11}(n+1) + b_2\tilde{\varepsilon}_{12}(n+1), \\ \tilde{\xi}_{n+1} &= A_0 + A_1\tilde{\theta}_n + A_2\hat{\xi}_n + B_1\tilde{\varepsilon}_{21}(n+1) + B_2\tilde{\varepsilon}_{22}(n+1), \end{aligned} \tag{15}$$

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 that has the same first moments 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 correlation.

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

$$f_\theta(\lambda) = \frac{1}{2\pi|1 + b_1 e^{-i\lambda}|^2} \quad \text{and} \quad f_\eta(\lambda) = \frac{1}{2\pi} \cdot \frac{1}{|1 + b_2 e^{-i\lambda}|^2},$$

where $|b_1| < 1,\ |b_2| < 1$.

We shall interpret θ as an informative 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 in Sect. 3, there are (mutually uncorrelated) white noises $\varepsilon_1 = (\varepsilon_1(n))$ and $\varepsilon_2 = (\varepsilon_2(n))$ such that

$$\theta_{n+1} + b_1\theta_n = \varepsilon_1(n+1), \quad \eta_{n+1} + b_2\eta_n = \varepsilon_2(n+1).$$

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

$$d_{11} = \frac{1}{1 - b_1^2}, \quad d_{12} = \frac{1}{1 - b_2^2}, \quad d_{22} = \frac{2 - b_1^2 - b_2^2}{(1 - b_1^2)(1 - b_2^2)},$$

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 $\xi_0, \dots, \xi_n$ 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 contains no random components, and consequently the numbers γ_n, which are needed for finding m_n, can be calculated in advance, before solving the filtering problem.

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 that 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 $\theta_0 \sim \mathcal{N}(m_0, \gamma_0)$.

System (19) is *nonlinear*, and Theorem 2 is not immediately applicable. However, if we set

$$\tilde{\varepsilon}_1(n+1) = \frac{1 + \theta_n}{\sqrt{\mathsf{E}(1 + \theta_n)^2}}\varepsilon_1(n+1),$$

we can observe that $\mathsf{E}\,\tilde{\varepsilon}_1(n) = 0$, $\mathsf{E}\,\tilde{\varepsilon}_1(n)\tilde{\varepsilon}_1(m) = 0$, $n \neq m$, $\mathsf{E}\,\varepsilon_1^2(n) = 1$. 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 $\hat{m}_n = \hat{\mathsf{E}}(\theta_n \,|\, \xi_0, \dots, \xi_n)$ 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:

$$\begin{aligned} \hat{m}_{n+1} &= a_1\hat{m}_n + \frac{a_1A_1\hat{\gamma}_n}{1 + A_1^2\hat{\gamma}_n}[\xi_{n+1} - A_1\hat{m}_n], \\ \hat{\gamma}_{n+1} &= (a_1^2\hat{\gamma}_n + b_1^2(n)) - \frac{(a_1A_1\hat{\gamma}_n)^2}{1 + A_1^2\hat{\gamma}_n}, \end{aligned}$$

where $b_1(n) = \sqrt{\mathsf{E}(1 + \theta_n)^2}$ must be found from the first equation in (19).

EXAMPLE 3 (Estimators for parameters). Let $\theta = (\theta_1, \dots, \theta_k)$ be a Gaussian vector with $\mathsf{E}\,\theta = m_0$ and $\mathrm{Cov}(\theta, \theta) = \gamma_0$. Suppose that (with known m_0 and γ_0) we look for the optimal estimator of θ in terms of observations on an l-dimensional sequence $\xi = (\xi_n)$, $n \geq 0$, with

$$\xi_{n+1} = A_0(n,\ \xi) + A_1(n,\ \xi)\theta + B_1(n,\ \xi)\varepsilon_1(n+1), \quad \xi_0 = 0, \tag{21}$$

where ε_1 is as in (1).

Then we have from (7) and (8) that $m_n = \mathsf{E}(\theta \mid \mathscr{F}_n^{\xi})$ and γ_n can be found from

$$\begin{aligned} m_{n+1} &= m_n + \gamma_n A_1^*(n,\xi)[(B_1B_1^*)(n,\xi) + A_1(n,\xi)\gamma_n A_1^*(n,\xi)]^{\oplus} \\ &\quad \times [\xi_{n+1} - A_0(n,\xi) - A_1(n,\xi)m_n], \\ \gamma_{n+1} &= \gamma_n - \gamma_n A_1^*(n,\xi)[(B_1B_1^*)(n,\xi) + A_1(n,\xi)\gamma_n A_1^*(n,\xi)]^{\oplus} A_1(n,\xi)\gamma_n. \end{aligned} \tag{22}$$

If the matrices $B_1B_1^*$ are nonsingular for all n and ξ, the solution of (22) is given by

$$\begin{aligned} m_{n+1} &= \left[E + \gamma \sum_{i=0}^{n} A_1^*(i,\xi)(B_1B_1^*)^{-1}(i,\xi)A_1^*(i,\xi)\right]^{-1} \\ &\quad \times \left[m + \gamma \sum_{i=0}^{n} A_1^*(i,\xi)(B_1B_1^*)^{-1}(i,\xi)(\xi_{i+1} - A_0(i,\xi))\right], \\ \gamma_{n+1} &= \left[E + \gamma \sum_{i=0}^{n} A_1^*(i,\xi)(B_1B_1^*)^{-1}(i,\xi)A_1(i,\xi)\right]^{-1} \gamma, \end{aligned} \tag{23}$$

where E is the identity 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)–(2), 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):

$$\theta_{n+1} = a\theta_n + b\varepsilon_1(n+1), \quad \xi_{n+1} = A\theta_n + B\varepsilon_2(n+1).$$

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

$$\gamma^2 + \left[\frac{B^2(1-a^2)}{A^2} - b^2\right]\gamma - \frac{b^2B^2}{A^2} = 0.$$

5. (*Interpolation*, [54, 13.3]) Let (θ, ξ) be a partially observed sequence governed by recurrence relations (1) and (2). Suppose that the conditional distribution

$$\pi_a(m,m) = \mathsf{P}(\theta_m \le a \,|\, \mathscr{F}_m^\xi)$$

of θ_m is Gaussian.

(a) Show that the conditional distribution

$$\pi_a(m,n) = \mathsf{P}(\theta_m \le a \,|\, \mathscr{F}_n^\xi), \quad n \ge m,$$

is also Gaussian, $\pi_a(m,n) \sim \mathscr{N}(\mu(m,n), \gamma(m,n))$.

(b) Find the interpolation estimator $\mu(m,n)$ (of θ_m given $\mathscr{F}_n^\xi$) and the matrix $\gamma(m,n)$.

6. (*Extrapolation*, [54, 13.4]) In (1) and (2), let

$$\begin{aligned} a_0(n,\xi) &= a_0(n) + a_2(n)\xi_n, \quad & a_1(n,\xi) &= a_1(n), \\ A_0(n,\xi) &= A_0(n) + A_2(n)\xi_n, \quad & A_1(n,\xi) &= A_1(n). \end{aligned}$$

(a) Show that in this case the distribution $\pi_{a,b}(m,n) = \mathsf{P}(\theta_n \le a, \xi_n \le b \,|\, \mathscr{F}_m^\xi)$ is Gaussian ($n \ge m$).

(b) Find the extrapolation estimators

$$\mathsf{E}(\theta_n \,|\, \mathscr{F}_m^\xi) \quad \text{and} \quad \mathsf{E}(\xi_n \,|\, \mathscr{F}_m^\xi).$$

7. (*Optimal control*, [54, 14.3]) Consider a "controlled" partially observed system $(\theta_n, \xi_n)_{0 \le n \le N}$, where

$$\begin{aligned} \theta_{n+1} &= u_n + \theta_n + b\varepsilon_1(n+1), \\ \xi_{n+1} &= \theta_n + \varepsilon_2(n+1). \end{aligned}$$

Here the "control" u_n is $\mathscr{F}_n^\xi$-measurable and satisfies $\mathsf{E}\, u_n^2 < \infty$ for all $0 \le n \le N-1$. The variables $\varepsilon_1(n)$ and $\varepsilon_2(n)$, $n = 1, \dots, N$, are the same as in (1), (2); $\xi_0 = 0$, $\theta_0 \sim \mathscr{N}(m, \gamma)$.

We say that the "control" $u^* = (u_0^*, \dots, u_{N-1}^*)$ is optimal if $V(u^*) = \sup_u V(u)$, where

$$V(u) = \mathsf{E}\left[\sum_{n=0}^{N-1} (\theta_n^2 + u_n^2) + \theta_N^2\right].$$

Show that

$$u_n^* = -[1 + P_{n+1}]^+ P_{n+1} m_n^*, \quad n = 0, \dots, N-1,$$

where

$$a^+ = \begin{cases} a^{-1}, & a \ne 0, \\ 0, & a = 0, \end{cases}$$

$(P_n)_{0\le n\le N}$ are found from the recurrence relations

$$P_n = 1 + P_{n+1} - P_{n+1}^2[1 + P_{n+1}]^+, \quad P_N = 1,$$

and (m_n^*) are determined by

$$m_{n+1}^* = u_n^* + \gamma_n^*(1 + \gamma_n^*)^+(\xi_{n+1} - m_n^*), \quad 0 \le n \le N-1,$$

with $m_0^* = m$ and (γ_n^*) by

$$\gamma_{n+1}^* = \gamma_n^* + 1 - (\gamma_n^*)^2(1 + \gamma_n^*)^+, \quad 0 \le n \le N-1,$$

with $\gamma_0^* = \gamma$.

Chapter 7
Martingales

1. Definitions of Martingales and Related Concepts

> Martingale theory illustrates the history of mathematical probability; the basic definitions are inspired by crude notions of gambling, but the theory has become a sophisticated tool of modern abstract mathematics, drawing from and contributing to other fields.
>
> J. L. Doob [19]

1. The study of the dependence between 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 (Sect. 12 of Chap. 1, Vol. 1 and Chap. 8) the basic dependence is supplied by the transition function, which completely determines the development of the random variables involved in Markov dependence.

In this chapter (see also Sect. 11 of Chap. 1, Vol. 1) 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 the properties of *conditional expectations.*

2. Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a given probability space with a *filtration* (*flow*), i.e., with a family $(\mathscr{F}_n)$ of σ-algebras $\mathscr{F}_n$, $n \geq 0$, such that $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \ldots \subseteq \mathscr{F}$ ("filtered probability space").

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*, the set $X = (X_n, \mathscr{F}_n)_{n\geq 0}$, or simply $X = (X_n, \mathscr{F}_n)$, is called 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.).

A. N. Shiryaev, *Probability-2*, Graduate Texts in Mathematics 95,
https://doi.org/10.1007/978-0-387-72208-5_4

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,

$$\begin{gathered} \mathsf{E}(X_{n+1} \mid \mathscr{F}_n) = X_n \quad (\mathsf{P}\text{-a.s.}) \text{ (martingale)} \\ \text{or} \\ \mathsf{E}(X_{n+1} \mid \mathscr{F}_n) \geq X_n \quad (\mathsf{P}\text{-a.s.}) \text{ (submartingale).} \end{gathered} \tag{2}$$

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 where $\mathscr{F}_n = \mathscr{F}_n^X$, where $\mathscr{F}_n^X = \sigma\{X_0, \ldots, X_n\}$, 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 $A \in \mathscr{F}_n$,

$$\begin{gathered} \int_A X_{n+1}\, d\,\mathsf{P} = \int_A X_n\, d\,\mathsf{P} \\ \text{or} \\ \int_A X_{n+1}\, d\,\mathsf{P} \geq \int_A X_n\, d\,\mathsf{P}\,. \end{gathered} \tag{3}$$

EXAMPLE 1. If $(\xi_n)_{n\geq 0}$ is a sequence of independent random variables such that $\mathsf{E}\,|\xi_n| < \infty$, $\mathsf{E}\,\xi_n = 0$, and $X_n = \xi_0 + \cdots + \xi_n$, $\mathscr{F}_n = \sigma\{\xi_0, \ldots, \xi_n\}$, 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 such that $\mathsf{E}\,|\xi_n| < \infty$ and $\mathsf{E}\,\xi_n = 1$, the stochastic sequence $(X_n, \mathscr{F}_n)$ with $X_n = \prod_{k=0}^n \xi_k$, $\mathscr{F}_n = \sigma\{\xi_0, \ldots, \xi_n\}$ 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 $X_n = \mathsf{E}(\xi \mid \mathscr{F}_n)$, is a martingale called *Levy's 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; see Sect. 6 of Chap. 2, Vol. 1).

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, according to Sect. 7 of Chap. 2, Vol. 1, $\mathsf{E}(X_{n+1}^+ \,|\, \mathscr{F}_n)$ 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

$$\{\omega\colon\ \mathsf{E}(X_{n+1}^+ \,|\, \mathscr{F}_n) < \infty\} \cup \{\omega\colon\ \mathsf{E}(X_n^- \,|\, \mathscr{F}_n) < \infty\} = \Omega \quad (\mathsf{P}\text{-a.s.})$$

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 the corresponding condition (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 that $\mathsf{E}(|X_{n+1}| \,|\, \mathscr{F}_n) < \infty$ (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

$$X_\tau(\omega) = \sum_{n=0}^{\infty} X_n(\omega) I_{\{\tau=n\}}(\omega)$$

(hence we set $X_\infty = 0$ and $X_\tau = 0$ on the set $\{\omega\colon \tau = \infty\}$).

Then, for every $B \in \mathscr{B}(R)$,

$$\{\omega\colon X_\tau \in B\} = \{\omega\colon X_\infty \in B, \tau = \infty\} + \sum_{n=0}^{\infty} \{X_n \in B,\ \tau = n\} \in \mathscr{F},$$

and consequently, $X_\tau = X_{\tau(\omega)}(\omega)$ 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,

$$\tau_B = \min\{n \geq 0 \colon X_n \in B\}$$

(with $\tau_B = +\infty$ if $\{.\} = \varnothing$) is a Markov time, since

$$\{\tau_B = n\} = \{X_0 \notin B,\ \ldots,\ X_{n-1} \notin B,\ X_n \in B\} \in \mathscr{F}_n$$

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" sequence $X^\tau = (X_{n\wedge\tau}, \mathscr{F}_n)$ 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 flow $(\mathscr{F}_n)$ and Markov time τ corresponding to it generate a collection of sets

$$\mathscr{F}_\tau = \{A \in \mathscr{F} \colon A \cap \{\tau = n\} \in \mathscr{F}_n \text{ for all } n \geq 0\}.$$

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 $\overline{A} \cap \{\tau = n\} = \{\tau = n\}\backslash(A \cap \{\tau = n\}) \in \mathscr{F}_n$, and therefore $\overline{A} \in \mathscr{F}_\tau$. Hence it follows that $\mathscr{F}_\tau$ is a σ-algebra.

If we think of $\mathscr{F}_n$ as the collection of events observed up to time n (inclusive), then $\mathscr{F}_\tau$ can be thought of as the collection of events observed until 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 finite 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 $X^{\tau_k} = (X_{\tau_k\wedge n} I_{\{\tau_k>0\}}, \mathscr{F}_n)$ 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 as a "martingale transform" from a martingale and a predictable sequence.

Definition 5. Let $Y = (Y_n, \mathscr{F}_n)_{n\geq 0}$ be a stochastic sequence and $V = (V_n, \mathscr{F}_{n-1})_{n\geq 0}$ 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 (or a local 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 localizing 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

$$\mathsf{E}[|X_{n+1}| I_{\{\tau_k>n\}} \,|\, \mathscr{F}_n] = I_{\{\tau_k>n\}}\, \mathsf{E}[|X_{n+1}| \,|\, \mathscr{F}_n] < \infty \quad (\mathsf{P}\text{-a.s.}).$$

Here $I_{\{\tau_k>n\}} \to 1$ (P-a.s.) as $k \to \infty$, and therefore

$$\mathsf{E}[|X_{n+1}| \,|\, \mathscr{F}_n] < \infty \quad (\mathsf{P}\text{-a.s.}). \tag{8}$$

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, Sect. 7, Chap. 2, Vol. 1, 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 $\{\tau_k > n\} \in \mathscr{F}_n$)

$$\int_{A\cap\{\tau_k>n\}} |X_n|\,d\mathsf{P} = \int_{A\cap\{\tau_k>n\}} |X_{n\wedge\tau_k}| I_{\{\tau_k>0\}}\,d\mathsf{P}$$
$$\le \int_{A\cap\{\tau_k>n\}} |X_{(n+1)\wedge\tau_k}| I_{\{\tau_k>0\}}\,d\mathsf{P} = \int_{A\cap\{\tau_k>n\}} |X_{n+1}|\,d\mathsf{P}\,.$$

Letting $k\to\infty$, we have

$$\int_A |X_n|\,d\mathsf{P} \le \int_A |X_{n+1}|\,d\mathsf{P},$$

from which there follows the required σ-finiteness of the measure $\int_A |X_n|\,d\mathsf{P}$, $A\in\mathscr{F}_n$.

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 $\mathsf{E}(X_{n+1}\mid\mathscr{F}_n)=X_n$ (P-a.s.).

(b)$\Rightarrow$ (c). Let $\Delta X_n = X_n - X_{n-1},\ X_0=0$, and $V_0=0,\ V_n=\mathsf{E}[|\Delta X_n|\,|\,\mathscr{F}_{n-1}]$, $n\ge 1$. We set

$$W_n = V_n^{\oplus}\ \left(=\begin{cases} V_n^{-1}, & V_n\ne 0\\ 0, & V_n=0\end{cases}\right),$$

$Y_0=0$, and $Y_n=\sum_{i=1}^n W_i\Delta X_i,\ n\ge 1$. It is clear that

$$\mathsf{E}[|\Delta Y_n|\,|\,\mathscr{F}_{n-1}]\le 1,\quad \mathsf{E}[\Delta Y_n\mid\mathscr{F}_{n-1}]=0,$$

and 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 $V_0=Y_0=0$. Set

$$\tau_k=\min\{n\ge 0\colon |V_{n+1}|>k\}$$

letting $\tau_k=\infty$ if the set $\{\cdot\}=\varnothing$. Since V_{n+1} is $\mathscr{F}_n$-measurable, the variables τ_k are Markov times for every $k\ge 1$.

Consider the 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 $|V_{n\wedge t_k}| \le k$ is in effect. Hence it follows that $\mathsf{E}\,|(V \cdot Y)_{n\wedge\tau_k} I_{\{\tau_k>0\}}| < \infty$ for every $n \ge 1$. In addition, for $n \ge 1$,

$$\begin{aligned}\mathsf{E}\{[(V \cdot Y)_{(n+1)\wedge\tau_k} - (V \cdot Y)_{n\wedge\tau_k}]\, I_{\{\tau_k>0\}} \mid \mathscr{F}_n\} \\ = I_{\{\tau_k>0\}} V_{(n+1)\wedge\tau_k} \cdot \mathsf{E}\{Y_{(n+1)\wedge\tau_k} - Y_{n\wedge\tau_k} \mid \mathscr{F}_n\} = 0\end{aligned}$$

since (Example 7) $\mathsf{E}\{Y_{(n+1)\wedge\tau_k} - Y_{n\wedge\tau_k} \mid \mathscr{F}_n\} = 0$.

Thus for every $k \ge 1$ the "stopped" sequence X^{τ_k} is a martingale, $\tau_k \uparrow \infty$ (P-a.s.), and consequently X is a local martingale.

This completes the proof of the theorem.

□

5. EXAMPLE 8. Let $(\eta_n)_{n\ge1}$ be a sequence of independent identically distributed Bernoulli random variables with $\mathsf{P}(\eta_n = 1) = p$, $\mathsf{P}(\eta_n = -1) = q$, $p + q = 1$. We interpret the event $\{\eta_n = 1\}$ as the success (gain) and $\{\eta_n = -1\}$ as the 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 $V_1, \dots, V_{n-1}$ and on $\eta_1, \dots, \eta_{n-1}$. In other words, if we put $F_0 = \{\varnothing, \Omega\}$ and $F_n = \sigma\{\eta_1, \dots, \eta_n\}$, 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

$$\mathsf{E}(X_{n+1} - X_n \mid \mathscr{F}_n) = 0 \ (\text{or} \ge 0 \text{ or} \le 0).$$

Moreover, it is clear that the game is

$$\begin{aligned}&\textit{fair} \text{ if } p = q = \tfrac{1}{2},\\ &\textit{favorable} \text{ if } p > q,\\ &\textit{unfavorable} \text{ if } p < q.\end{aligned}$$

Since $X = (X_n, \mathscr{F}_n)$ is a

$$\begin{aligned} &\textit{martingale} \text{ if } p = q = \tfrac{1}{2}, \\ &\textit{submartingale} \text{ if } p > q, \\ &\textit{supermartingale} \text{ if } p < q, \end{aligned}$$

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 $\eta_{n+1} = 1$, then we have

$$X_{n+1} = X_n + V_{n+1} = -(2^n - 1) + 2^n = 1.$$

Let $\tau = \min\{n \geq 1 \colon X_n = 1\}$. 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 1) 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

$$\mathsf{E}\, X_n = \mathsf{E}\, X_0 = 0 \qquad \text{for every} \qquad n \geq 1.$$

We may therefore expect that this equation will be preserved if the instant n is replaced by a *random* instant τ. It will appear later (Theorem 1 in Sect. 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. That is, 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 $\mathscr{F}_n = \sigma\{\xi_0,\ \xi_1, \ldots, \xi_n\}$).

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 are 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*

$$X_n = m_n + A_n \quad (\mathsf{P}\text{-}a.s.) \tag{11}$$

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 justifies 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 $M = (M_n, \mathscr{F}_n)$, 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 *quadratic characteristic* of M and, in many respects, determines its structure and properties.

It follows from (13) 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 characteristic

$$\langle M \rangle_n = \mathsf{E}\, M_n^2 = \operatorname{Var} \xi_1 + \cdots + \operatorname{Var} \xi_n \tag{18}$$

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 $\mathsf{E}\, \xi_i^2 < \infty$, $\mathsf{E}\, \eta_i^2 < \infty$, the variable $\langle X, Y \rangle_n$ is given by

$$\langle X, Y \rangle_n = \sum_{i=1}^{n} \operatorname{Cov}(\xi_i, \eta_i).$$

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

(cf. (15)) that

$$\langle X, Y\rangle_n = \sum_{i=1}^{n} \mathsf{E}[\Delta X_i \Delta Y_i \mid \mathscr{F}_{i-1}].$$

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. In connection with Theorem 1, it is natural to ask when a local martingale (and hence a generalized martingale or a martingale transform) is in fact a *martingale*.

Theorem 3. (1) *Suppose that a stochastic sequence* $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ *is a local martingale* (*with* $X_0 = 0$ *or, more generally, with* $\mathsf{E}\,|X_0| < \infty$).

If $\mathsf{E}\,X_n^- < \infty$, $n \geq 0$, *or* $\mathsf{E}\,X_n^+ < \infty$, $n \geq 0$, *then* $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ *is a martingale.*

(2) *Let* $X = (X_n, \mathscr{F}_n)_{0\leq n\leq N}$ *be a local martingale,* $N < \infty$, *and either* $\mathsf{E}\,X_N^- < \infty$ *or* $\mathsf{E}\,X_N^+ < \infty$. *Then* $X = (X_n, \mathscr{F}_n)_{0\leq n\leq N}$ *is a martingale.*

PROOF. (1) Let us show that either of the conditions $\mathsf{E}\,X_n^- < \infty$, $n \geq 0$, or $\mathsf{E}\,X_n^+ < \infty$, $n \geq 0$, implies that $\mathsf{E}\,|X_n| < \infty$, $n \geq 0$.

Indeed, let, for example, $\mathsf{E}\,X_n^- < \infty$ for all $n \geq 0$. Then, by the Fatou lemma,

$$\begin{aligned}\mathsf{E}\,X_n^+ &= \mathsf{E}\liminf_k X^+_{n\wedge\tau_k} \leq \liminf_k \mathsf{E}\,X^+_{n\wedge\tau_k} = \liminf_k [\mathsf{E}\,X_{n\wedge\tau_k} + \mathsf{E}\,X^-_{n\wedge\tau_k}] \\ &= \mathsf{E}\,X_0 + \liminf_k \mathsf{E}\,X^-_{n\wedge\tau_k} \leq |\,\mathsf{E}\,X_0| + \sum_{k=0}^{n} \mathsf{E}\,X_k^- < \infty.\end{aligned}$$

Therefore $\mathsf{E}\,|X_n| < \infty$, $n \geq 0$.

To prove the martingale property $\mathsf{E}(X_{n+1} \mid \mathscr{F}_n) = X_n$, $n \geq 0$, let us observe that *for any* Markov time τ_k we have

$$|X_{(n+1)\wedge\tau_k}| \leq \sum_{i=0}^{n+1} |X_i|,$$

where

$$\mathsf{E} \sum_{i=0}^{n+1} |X_i| < \infty.$$

Therefore, taking the limit as $k \to \infty$, $\tau_k \uparrow \infty$ (P-a.s.) in the equality $\mathsf{E}(X_{(n+1)\wedge\tau_k} \,|\, \mathscr{F}_n) = X_{n\wedge\tau_k}$, we obtain by Lebesgue's dominated convergence theorem that $\mathsf{E}(X_{n+1} \,|\, \mathscr{F}_n) = X_n$ (P-a.s.).

(2) Assume, for example, that $\mathsf{E}\, X_N^- < \infty$. We will then show that $\mathsf{E}\, X_n^- < \infty$ for all $n < N$.

Indeed, since a local martingale is a generalized martingale, we have $X_n = \mathsf{E}(X_{n+1} \,|\, \mathscr{F}_n)$, where $\mathsf{E}(|X_{n+1}| \,|\, \mathscr{F}_n) < \infty$ (P-a.s.). Then, by Jensen's inequality for conditional expectations (see Problem 5 in Sect. 7, Chap. 2, Vol. 1), $X_n^- \le \mathsf{E}(X_{n+1}^- \,|\, \mathscr{F}_n)$. Therefore $\mathsf{E}\, X_n^- \le \mathsf{E}\, X_{n+1}^- \le \mathsf{E}\, X_N^- < \infty$.

Thus the desired martingale property of the local martingale $X = (X_n, \mathscr{F}_n)_{0\le n\le N}$ follows from conclusion (1).

□

9. 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; in addition, if $\mathsf{P}(\sigma \le \tau) = 1$, then $\mathscr{F}_\sigma \subseteq \mathscr{F}_\tau$ (see Example 7 for the definition of $\mathscr{F}_\tau$).
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, $n \ge 0$. Show that $V \cdot Y$ is a martingale (or submartingale).
5. 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 $(X_n)_{n\ge1}$ 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 $n \ge 1$.
6. Let $\xi_1, \xi_2, \ldots$ be independent random variables,

$$\mathsf{P}(\xi_i = 0) = \mathsf{P}(\xi_i = 2) = \tfrac{1}{2} \quad \text{and} \quad X_n = \prod_{i=1}^{n} \xi_i.$$

   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 $(X_n)_{n\ge1}$ can be represented in the form $(\mathsf{E}(\xi \,|\, \mathscr{F}_n))_{n\ge1}$ (cf. Example 3 in Sect. 11, Chap. 1, Vol. 1).
7. (a) Let $\xi_1, \xi_2, \ldots$ be independent random variables with $\mathsf{E}\,|\xi_n| < \infty$, $\mathsf{E}\,\xi_n = 0$, $n \ge 1$. Show that for any $k \ge 1$ the sequence

$$X_n^{(k)} = \sum_{1\le i_1<\cdots<i_k\le n} \xi_{i_1} \ldots \xi_{i_k}, \quad n \ge k,$$

   is a martingale.

(b) Let $\xi_1, \xi_2, \dots$ be integrable random variables such that

$$\mathsf{E}(\xi_{n+1} \,|\, \xi_1, \dots, \xi_n) = \frac{\xi_1 + \dots + \xi_n}{n} \quad (= X_n).$$

Prove that the sequence $X_1, X_2, \dots$ is a martingale.

8. Give an example of a martingale $(X_n, \mathscr{F}_n)_{n\geq 1}$ such that the family $\{X_n,\ n \geq 1\}$ is not uniformly integrable.
9. Let $X = (X_n)_{n\geq 0}$ be a Markov chain (Sect. 1, Chap. 8) with a countable state space $E = \{i, j, \dots\}$ and transition probabilities p_{ij}. Let $\psi = \psi(x)$, $x \in E$, be a bounded function such that $\sum\limits_{j\in E} p_{ij}\psi(j) \leq \lambda\psi(i)$ for $\lambda > 0$ and $i \in E$. Show that the sequence $(\lambda^{-n}\psi(X_n))_{n\geq 0}$ is a supermartingale.

2. Preservation of Martingale Property Under a Random Time Change

1. If $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ is a martingale, then 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 1 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, $\mathsf{E}\, X_\tau = \mathsf{E}\, X_0$. (We let $X_\tau = 0$ on the set $\{\tau = \infty\}$.)

Theorem 1 (Doob). (a) *Let $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ be a submartingale, and τ and σ finite* (P-*a.s.*) *stopping times for which $\mathsf{E}\, X_\tau$ and $\mathsf{E}\, X_\sigma$ are defined* (*e.g., such that $\mathsf{E}\,|X_\tau| < \infty$ and $\mathsf{E}\,|X_\sigma| < \infty$*). *Assume that*

$$\liminf_{m\to\infty} \mathsf{E}[X_m^+ I(\tau > m)] = 0. \tag{3}$$

Then

$$\mathsf{E}(X_\tau \,|\, \mathscr{F}_\sigma) \geq X_{\tau\wedge\sigma} \quad (\mathsf{P}\text{-}a.s.) \tag{4}$$

or, equivalently,

$$\mathsf{E}(X_\tau \,|\, \mathscr{F}_\sigma) \geq X_\sigma \quad (\{\tau \geq \sigma\};\ \mathsf{P}\text{-}a.s.).$$

(b) *Let $M = (M_n, \mathscr{F}_n)_{n\geq 0}$ be a martingale, and τ and σ finite* (P-*a.s.*) *stopping times for which $\mathsf{E}\, M_\tau$ and $\mathsf{E}\, M_\sigma$ are defined* (*e.g., such that $\mathsf{E}\,|M_\tau| < \infty$ and $\mathsf{E}\,|M_\sigma| < \infty$*). *Assume that*

$$\liminf_{m\to\infty} \mathsf{E}[|M_m| I(\tau > m)] = 0. \tag{5}$$

Then

$$\mathsf{E}(M_\tau \mid \mathscr{F}_\sigma) = M_{\tau\wedge\sigma} \quad (\mathsf{P}\text{-}a.s.) \tag{6}$$

or, equivalently,

$$\mathsf{E}(M_\tau \mid \mathscr{F}_\sigma) = M_\sigma \quad (\{\tau \geq \sigma\};\ \mathsf{P}\text{-}a.s.).$$

PROOF. (a) We must show that, for every $A \in \mathscr{F}_\sigma$,

$$\mathsf{E}\, X_\tau I(A, \tau \geq \sigma) \geq \mathsf{E}\, X_\sigma I(A, \tau \geq \sigma), \tag{7}$$

where $I(A, \tau \geq \sigma)$ is the indicator function of the set $A \cap \{\tau \geq \sigma\}$.

To prove (7), it suffices to show that for any $n \geq 0$

$$\mathsf{E}\, X_\tau I(A, \tau \geq \sigma, \sigma = n) \geq \mathsf{E}\, X_\sigma I(A, \tau \geq \sigma, \sigma = n),$$

i.e., that

$$\mathsf{E}\, X_\tau I(B, \tau \geq n) \geq \mathsf{E}\, X_n I(A, \tau \geq n), \quad B = A \cap \{\sigma = n\}.$$

Using the property $B \cup \{\tau > n\} \in \mathscr{F}_n$ and the fact that the process $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ is a submartingale, we find by iterating in n that for any $m \geq n$

$$\begin{aligned}
\mathsf{E}\, X_n I(B, \tau \geq n) &= \mathsf{E}\, X_n I(B, \tau = n) + \mathsf{E}\, X_n I(B, \tau > n) \\
&\leq \mathsf{E}\, X_n I(B, \tau = n) + \mathsf{E}\big[\mathsf{E}(X_{n+1} \mid \mathscr{F}_n)\, I(B, \tau > n)\big] \\
&= \mathsf{E}\, X_n I(B, \tau = n) + \mathsf{E}\big[X_{n+1}\, I(B, \tau \geq n+1)\big] \\
&= \mathsf{E}\, X_\tau I(B, n \leq \tau \leq n+1) + \mathsf{E}\, X_{n+1} I(B, \tau > n+1) \\
&\leq \mathsf{E}\, X_\tau I(B, n \leq \tau \leq n+1) + \mathsf{E}\, X_{n+2} I(B, \tau \geq n+2) \\
&\leq \cdots \leq \mathsf{E}\, X_\tau I(B, n \leq \tau \leq m) + \mathsf{E}\, X_m I(B, \tau > m).
\end{aligned}$$

Consequently,

$$\mathsf{E}\, X_\tau I(B, n \leq \tau \leq m) \geq \mathsf{E}\, X_n I(B, \tau \geq n) - \mathsf{E}\, X_m I(B, \tau > m). \tag{8}$$

By assumption, $\mathsf{E}\, X_\tau$ is defined. Therefore the function $Q(C) = \mathsf{E}\, X_\tau\, I(C)$ of Borel sets $C \in \mathscr{B}(R)$ is countably additive (Subsection 8 in Sect. 6, Chap. 2, Vol. 1), and hence there exists the limit $\lim_{m\to\infty} \mathsf{E}\, X_\tau I(B, n \leq \tau \leq m)$. Therefore, since the Markov time τ is finite (P-a.s.), inequality (8) implies that

$$\begin{aligned}
\mathsf{E}\, X_\tau I(B, \tau \geq n) &\geq \limsup_{m\to\infty} \big[\mathsf{E}\, X_n I(B, \tau \geq n) - \mathsf{E}\, X_m I(B, \tau > m)\big] \\
&= \mathsf{E}\, X_n I(B, \tau \geq n) - \liminf_{m\to\infty} \mathsf{E}\, X_m I(B, \tau > m) \\
&\geq \mathsf{E}\, X_n I(B, \tau \geq n) - \lim_{m\to\infty} \mathsf{E}\, X_m^+ I(B, \tau > m) \\
&= \mathsf{E}\, X_n I(B, \tau \geq n).
\end{aligned}$$

Thus, we have

$$\mathsf{E}\, X_\tau I(B, \sigma = n, \tau \geq n) \geq \mathsf{E}\, X_n I(B, \sigma = n, \tau \geq n)$$

or

$$\mathsf{E}\, X_\tau I(A, \tau \geq \sigma, \sigma = n) \geq \mathsf{E}\, X_\sigma I(A, \tau \geq \sigma, \sigma = n).$$

Hence, using the assumption $\mathsf{P}\{\sigma < \infty\} = 1$ and the fact that the expectations $\mathsf{E}\, X_\tau$ and $\mathsf{E}\, X_\sigma$ are defined, we obtain the desired inequality (7).

(b) Let $M = (M_n, \mathscr{F}_n)_{n\geq 0}$ be a martingale satisfying (5). This condition implies that

$$\liminf_{m\to\infty} \mathsf{E}\big[M_m^+ I(\tau > m)\big] = \liminf_{m\to\infty} \mathsf{E}\big[M_m^- I(\tau > m)\big] = 0.$$

Setting $X = M$ and $X = -M$ in (a) we find that (P-a.e.)

$$\mathsf{E}[M_\tau \,|\, \mathscr{F}_\sigma] \geq M_{\tau\wedge\sigma} \quad \text{and} \quad \mathsf{E}[-M_\tau \,|\, \mathscr{F}_\sigma] \geq -M_{\tau\wedge\sigma}$$

with the latter inequality telling us that $\mathsf{E}[M_\tau \,|\, \mathscr{F}_\sigma] \leq M_{\tau\wedge\sigma}$. Hence $\mathsf{E}[M_\tau \,|\, \mathscr{F}_\sigma] = M_{\tau\wedge\sigma}$ (P-a.s.), which is precisely equality (6).

□

Corollary 1. *Let τ and σ be stopping times such that*

$$\mathsf{P}\{\sigma \leq \tau \leq N\} = 1$$

for some N. Then for a submartingale X we have

$$\mathsf{E}\, X_0 \leq \mathsf{E}\, X_\sigma \leq \mathsf{E}\, X_\tau \leq \mathsf{E}\, X_N,$$

and for a martingale M

$$\mathsf{E}\, M_0 = \mathsf{E}\, M_\sigma = \mathsf{E}\, M_\tau = \mathsf{E}\, M_N.$$

Corollary 2. *Let $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ be a submartingale. If the family of random variables $\{X_n, n \geq 0\}$ is uniformly integrable (in particular, if $|X_n| \leq c$ (P-a.s.), $n \geq 0$, for some c), then for any finite (P-a.s.) stopping times τ and σ inequality (4) holds, and if $\mathsf{P}\{\sigma \leq \tau\} = 1$, then*

$$\mathsf{E}\, X_0 \leq \mathsf{E}\, X_\sigma \leq \mathsf{E}\, X_\tau.$$

Moreover, if $X = M$ is a martingale, then equality (6) holds, and if $\mathsf{P}\{\sigma \leq \tau\} = 1$, then

$$\mathsf{E}\, M_0 = \mathsf{E}\, M_\sigma = \mathsf{E}\, X_\tau.$$

For the proof, let us observe that the properties (3) and (5) follow from Lemma 2 in Subsection 5, Sect. 6, Chap. 2, Vol. 1, and the fact that $\mathsf{P}\{\tau > m\} \to 0$ as $m \to \infty$.

We will now show that the expectations $\mathsf{E}\,|X_\tau|$ and $\mathsf{E}\,|X_\sigma|$ are finite. To prove this, it suffices to show that

$$\mathsf{E}\,|X_\tau| \leq 3 \sup_N \mathsf{E}\,|X_N| \tag{9}$$

(and similarly for σ) because, due to inequality (16) of Sect. 6, Chap. 2, Vol. 1, the assumption of uniform integrability of $\{X_n, n \geq 0\}$ implies that $\sup_N \mathsf{E}\,|X_N| < \infty$;

hence the required inequality $\mathsf{E}\,|X_\tau| < \infty$ (and, similarly, $\mathsf{E}\,|X_\sigma| < \infty$). will follow from (9).

Corollary 1 applied to the bounded stopping time $\tau_N = \tau \wedge N$ implies

$$\mathsf{E}\,X_0 \le \mathsf{E}\,X_{\tau_N}.$$

Therefore

$$\mathsf{E}\,|X_{\tau_N}| = 2\,\mathsf{E}\,X^+_{\tau_N} - \mathsf{E}\,X_{\tau_N} \le 2\,\mathsf{E}\,X^+_{\tau_N} - \mathsf{E}\,X_0. \tag{10}$$

The sequence $X^+ = (X_n^+, \mathscr{F}_n)_{n\ge 0}$ is a submartingale (see Example 5 in Sect. 1); hence

$$\begin{aligned}\mathsf{E}\,X^+_{\tau_N} &= \sum_{j=0}^{N} \mathsf{E}\big[X_j^+\, I(\tau_N = j)\big] + \mathsf{E}\big[X_N^+\, I(\tau > N)\big] \\ &\le \sum_{j=0}^{N} \mathsf{E}\big[X_N^+\, I(\tau_N = j)\big] + \mathsf{E}\big[X_N^+\, I(\tau > N)\big] = \mathsf{E}\,X_N^+ \le \mathsf{E}\,|X_N| \le \sup_m \mathsf{E}\,|X_m|,\end{aligned}$$

which, combined with the inequality in (10), yields

$$\mathsf{E}\,|X_{\tau_N}| \le 3 \sup_m \mathsf{E}\,|X_m|.$$

Hence we obtain by Fatou's lemma (Theorem 2 (a) in Sect. 6, Chap. 2, Vol. 1)

$$\mathsf{E}\,|X_\tau| = \mathsf{E} \lim_N |X_{\tau_N}| = \mathsf{E} \liminf_N |X_{\tau_N}| \le \liminf_N \mathsf{E}\,|X_{\tau_N}| \le 3 \sup_N \mathsf{E}\,|X_N|,$$

which proves (9).

Remark 1. The martingale $X = (X_n, \mathscr{F}_n)_{n\ge 0}$ (with $p = q = 1/2$) in Example 8 of the previous section was shown to satisfy

$$\mathsf{E}\,|X_m|\, I(\tau > m) = (2^m - 1)\ \mathsf{P}\{\tau > m\} = (2^m - 1)\cdot 2^{-m} \to 1, \quad m \to \infty.$$

Therefore condition (5) fails here. It is of interest to notice that the property (6) fails here as well since, as was shown in that example, there is a stopping time τ such that $\mathsf{E}\,X_\tau = 1 > X_0 = 0$. In this sense, condition (5) (together with the condition that $\mathsf{E}\,X_\sigma$ and $\mathsf{E}\,X_\tau$ are defined) is not only sufficient for (6), but also "almost necessary."

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 $\mathscr{F}_n^X = \sigma\{X_0, \dots, X_n\}$). 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

$$\mathsf{E}\,X_\tau \underset{(\geq)}{=} \mathsf{E}\,X_0. \tag{11}$$

PROOF. We first verify that the stopping time τ has the properties

$$\mathsf{E}\,|X_\tau| < \infty \quad \text{and} \quad \liminf_{n\to\infty} \int\limits_{\{\tau>n\}} |X_n|\,d\,\mathsf{P} = 0,$$

which by Theorem 1 imply (11).

Let

$$Y_0 = |X_0|, \qquad Y_j = |X_j - X_{j-1}|, \quad j \geq 1.$$

Then $|X_\tau| \leq \sum_{j=0}^{\tau} Y_j$ and

$$\begin{aligned}\mathsf{E}\,|X_\tau| &\leq \mathsf{E}\left(\sum_{j=0}^{\tau} Y_j\right) = \int_\Omega \sum_{j=0}^{\tau} Y_j\,d\,\mathsf{P} = \sum_{n=0}^{\infty} \int_{\{\tau=n\}} \sum_{j=0}^{n} Y_j\,d\,\mathsf{P} \\ &= \sum_{n=0}^{\infty}\sum_{j=0}^{n} \int_{\{\tau=n\}} Y_j\,d\,\mathsf{P} = \sum_{j=0}^{\infty}\sum_{n=j}^{\infty} \int_{\{\tau=n\}} Y_j\,d\,\mathsf{P} = \sum_{j=0}^{\infty} \int_{\{\tau\geq j\}} Y_j\,d\,\mathsf{P}.\end{aligned}$$

The set $\{\tau \geq j\} = \Omega\backslash\{\tau < j\} \in \mathscr{F}^X_{j-1}$, $j \geq 1$. Therefore

$$\int_{\{\tau\geq j\}} Y_j\,d\,\mathsf{P} = \int_{\{\tau\geq j\}} \mathsf{E}[Y_j \,|\, X_0, \ldots, X_{j-1}]\,d\,\mathsf{P} \leq\ C\,\mathsf{P}\{\tau \geq j\}$$

for $j \geq 1$; and hence

$$\mathsf{E}\,|X_\tau| \leq \mathsf{E}\left(\sum_{j=0}^{\tau} Y_j\right) \leq \mathsf{E}\,|X_0| + C\sum_{j=1}^{\infty} \mathsf{P}\{\tau \geq j\} = \mathsf{E}\,|X_0| + C\,\mathsf{E}\,\tau < \infty. \tag{12}$$

Moreover, if $\tau > n$, then

$$\sum_{j=0}^{n} Y_j \leq \sum_{j=0}^{\tau} Y_j,$$

and therefore

$$\int_{\{\tau>n\}} |X_n|\,d\,\mathsf{P} \leq \int_{\{\tau>n\}} \sum_{j=0}^{\tau} Y_i\,d\,\mathsf{P}.$$

Hence, since (by (12)) $\mathsf{E}\sum_{j=0}^{\tau} Y_j < \infty$ and $\{\tau > n\} \downarrow \varnothing$, $n \to \infty$, the dominated convergence theorem yields

$$\liminf_{n\to\infty}\int_{\{\tau>n\}}|X_n|\,d\mathsf{P}\le\liminf_{n\to\infty}\int_{\{\tau>n\}}\sum_{j=0}^{\tau}Y_j\,d\mathsf{P}=0.$$

Hence the hypotheses of Theorem 1 are satisfied, and (11) 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 $\xi_1,\xi_2,\dots$ 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 $\mathscr{F}_n^\xi=\sigma\{\xi_1,\dots,\xi_n\}$, $\tau\ge 1$), and $\mathsf{E}\,\tau<\infty$. Then*

$$\mathsf{E}(\xi_1+\cdots+\xi_\tau)=\mathsf{E}\,\xi_1\cdot\mathsf{E}\,\tau. \tag{13}$$

If also $\mathsf{E}\,\xi_i^2<\infty$, then

$$\mathsf{E}\{(\xi_1+\cdots+\xi_\tau)-\tau\,\mathsf{E}\,\xi_1\}^2=\operatorname{Var}\xi_1\cdot\mathsf{E}\,\tau. \tag{14}$$

PROOF. Let $X=(X_n,\mathscr{F}_n^\xi)_{n\ge1}$, where $X_n=(\xi_1+\cdots+\xi_n)-n\,\mathsf{E}\,\xi_1$. It is clear that X is a martingale with

$$\begin{aligned}\mathsf{E}[|X_{n+1}-X_n|\,|\,X_1,\dots,X_n]&=\mathsf{E}[|\xi_{n+1}-\mathsf{E}\,\xi_1|\,|\,\xi_1,\dots,\xi_n]\\&=\mathsf{E}\,|\xi_{n+1}-\mathsf{E}\,\xi_1|\le 2\,\mathsf{E}\,|\xi_1|<\infty.\end{aligned}$$

Therefore, by Theorem 2, $\mathsf{E}\,X_\tau=\mathsf{E}\,X_0=0$, and (13) is established.

We will give *three* proofs of Wald's second identity (14).

The first proof. Let $\eta_i=\xi_i-\mathsf{E}\,\xi_i$, $S_n=\eta_1+\cdots+\eta_n$. We must show that

$$\mathsf{E}\,S_\tau^2=\mathsf{E}\,\eta_1^2\cdot\mathsf{E}\,\tau.$$

Put $\tau(n)=\tau\wedge n\ (=\min(\tau,n))$.

Since

$$S_n^2=\sum_{i=1}^n\eta_i^2+2\sum_{1\le i<j\le n}\eta_i\eta_j,$$

the sequence $(S_n^2-\sum_{i=1}^n\eta_i^2,\mathscr{F}_n^\xi)_{n\ge1}$ is a martingale with zero expectation.

By Corollary 1 we have

$$\mathsf{E}\,S_{\tau(n)}^2=\mathsf{E}\sum_{i=1}^{\tau(n)}\eta_i^2$$

and by Wald's first identity (13)

$$\mathsf{E}\sum_{i=1}^{\tau(n)}\eta_i^2=\mathsf{E}\,\eta_1^2\cdot\mathsf{E}\,\tau(n),$$

so that $\mathsf{E}\,S_{\tau(n)}^2=\mathsf{E}\,\eta_1^2\cdot\mathsf{E}\,\tau(n)$.

In a similar way we obtain that

$$\mathsf{E}(S_{\tau(n)} - S_{\tau(m)})^2 = \mathsf{E}\,\eta_1^2 \cdot \mathsf{E}(\tau(n) - \tau(m)) \to 0$$

as $m, n \to \infty$, since $\mathsf{E}\,\tau < \infty$ by assumption. Hence the sequence $\{S_{\tau(n)}\}_{n\geq 1}$ is *fundamental* (or a Cauchy sequence) in L^2 (see Subsection 5 of Sect. 10, Chap. 2, Vol. 1), so, by Theorem 7 of Sect. 10, Chap. 2, Vol. 1, there is a random variable S such that $\mathsf{E}(S_{\tau(n)} - S)^2 \to 0$, $n \to \infty$. This implies (Problem 1 in Sect. 11, Chap. 2, Vol. 1) that $\mathsf{E}\,S^2_{\tau(n)} \to \mathsf{E}\,S^2$, $n \to \infty$. As was shown earlier, $\mathsf{E}\,S^2_{\tau(n)} = \mathsf{E}\,\eta_1^2 \cdot \mathsf{E}\,\tau(n)$; therefore, letting $n \to \infty$, we obtain that $\mathsf{E}\,S^2 = \mathsf{E}\,\eta_1^2 \cdot \mathsf{E}\,\tau$.

It remains to identify the random variable S. Let us observe that with probability 1 there is a subsequence $\{n'\} \subseteq \{n\}$ such that both $S_{\tau(n')} \to S$ and $\tau(n') \to \tau$. But then it is clear that $S_{\tau(n')} \to S_\tau$ with probability 1. Therefore S and S_τ are the same almost surely; hence $\mathsf{E}\,S_\tau^2 = \mathsf{E}\,\eta_1^2 \cdot \mathsf{E}\,\tau$, which was to be proved.

The second proof. By Fatou's lemma (Theorem 2 (a), Sect. 6, Chap. 2, Vol. 1), we obtain from the equality $\mathsf{E}\,S^2_{\tau(n)} = \mathsf{E}\,\eta_1^2 \cdot \mathsf{E}\,\tau(n)$ established above that

$$\mathsf{E}\,S_\tau^2 = \mathsf{E}\liminf S^2_{\tau(n)} \leq \liminf \mathsf{E}\,S^2_{\tau(n)} = \mathsf{E}\,\eta_1^2 \cdot \mathsf{E}\,\tau.$$

The required equality $\mathsf{E}\,S_\tau^2 = \mathsf{E}\,\eta_1^2 \cdot \mathsf{E}\,\tau$ will be proved if we show that

$$\mathsf{E}\,S^2_{\tau(n)} \leq \mathsf{E}\,S_\tau^2$$

for any $n \geq 1$.

Notice, using Wald's first identity (13), that

$$\mathsf{E}\,|S_\tau| = \mathsf{E}\,|\eta_1 + \cdots + \eta_\tau| \leq \mathsf{E}(|\eta_1| + \cdots + |\eta_\tau|) = \mathsf{E}\,|\eta_1| \cdot \mathsf{E}\,\tau < \infty,$$

so

$$\begin{aligned}\mathsf{E}\,|S_n| I(\tau > n) = \mathsf{E}\,|\eta_1 + \cdots + \eta_n|\, I(\tau > n) &\leq \mathsf{E}(|\eta_1| + \cdots + |\eta_n|)\, I(\tau > n) \\ &\leq \mathsf{E}(|\eta_1| + \cdots + |\eta_\tau|)\, I(\tau > n) \to 0 \quad \text{as} \ \ n \to \infty.\end{aligned}$$

Applying Theorem 1 to the submartingale $(|S_n|, \mathscr{F}_n^\xi)_{n\geq 1})$, we find that, on the set $\{\tau \geq n\}$,

$$\mathsf{E}(|S_\tau| \,|\, \mathscr{F}_n^\xi) \geq |S_n| \quad (\mathsf{P}\text{-a.s.}).$$

Hence, by Jensen's inequality for conditional expectations (Problem 5, Sect. 7, Chap. 2, Vol. 1), we obtain that on the set $\{\tau \geq n\}$

$$\mathsf{E}(S_\tau^2 \,|\, \mathscr{F}_n^\xi) \geq S_n^2 = S^2_{\tau(n)} \quad (\mathsf{P}\text{-a.s.}).$$

And on the complementary set $\{\tau < n\}$ we have $\mathsf{E}(S_\tau^2 \,|\, \mathscr{F}_n^\xi) = S_\tau^2 = S^2_{\tau(n)}$. Thus ($\mathsf{P}$-a.s.)

$$\mathsf{E}(S_\tau^2 \,|\, \mathscr{F}_n^\xi) \geq S^2_{\tau(n)}$$

and hence $\mathsf{E}\,S_\tau^2 \geq \mathsf{E}\,S^2_{\tau(n)}$, as required.

The third proof. We see from the first proof that $(S_n^2 - \sum_{i=1}^n \eta_i^2, \mathscr{F}_n^\xi)_{n\geq 1}$ is a martingale and

$$\mathsf{E}\, S_{\tau(n)}^2 = \mathsf{E}\, \eta_1^2 \cdot \mathsf{E}\, \tau(n)$$

for $\tau(n) = \tau \wedge n$. Since $\mathsf{E}\, \tau(n) \to \mathsf{E}\, \tau$, we only have to show that $\mathsf{E}\, S_{\tau(n)}^2 \to \mathsf{E}\, S_\tau^2$. For that, it suffices to establish that

$$\mathsf{E} \sup_n S_{\tau(n)}^2 < \infty,$$

because the required convergence will then follow by Lebesgue's dominated convergence theorem (Theorem 3, Sect. 6, Chap. 2, Vol. 1).

For the proof of this inequality we will use the "maximal inequality" (13) to be given in the next Sect. 3. This inequality applied to the martingale $(S_{\tau(k)}, \mathscr{F}_k^\xi)_{k\geq 1}$ yields

$$\mathsf{E}\left[\sup_{1\leq k\leq n} S_{\tau(k)}^2\right] \leq 4\, \mathsf{E}\, S_{\tau(n)}^2 \leq 4 \sup_n \mathsf{E}\, S_{\tau(n)}^2.$$

Hence, using the monotone convergence theorem (Theorem 1 of Sect. 6, Chap. 2, Vol. 1), we obtain that

$$\mathsf{E} \sup_{k\geq 1} S_{\tau(k)}^2 \leq 4 \sup_n \mathsf{E}\, S_{\tau(n)}^2.$$

But

$$\mathsf{E}\, S_{\tau(n)}^2 = \mathsf{E}\, \eta_1^2 \cdot \mathsf{E}\, \tau(n) \leq \mathsf{E}\, \eta_1^2 \cdot \mathsf{E}\, \tau < \infty.$$

Therefore

$$\mathsf{E} \sup_n S_{\tau(n)}^2 \leq 4\, \mathsf{E}\, \eta_1^2 \cdot \mathsf{E}\, \tau < \infty,$$

as was to be shown.

□

Corollary. *Let $\xi_1, \xi_2, \ldots$ be independent identically distributed random variables with*

$$\mathsf{P}(\xi_i = 1) = \mathsf{P}(\xi_i = -1) = \tfrac{1}{2}, \quad S_n = \xi_1 + \cdots + \xi_n,$$

and $\tau = \inf\{n \geq 1 \colon S_n = 1\}$. Then $\mathsf{P}\{\tau < \infty\} = 1$ (see, for example, (20) in Sect. 9, Chap. 1, Vol. 1) and therefore $\mathsf{P}(S_\tau = 1) = 1$, $\mathsf{E}\, S_\tau = 1$. Hence it follows from (13) 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, $S_n = \xi_1 + \cdots + \xi_n$, $n \geq 1$. Let $\varphi(t) = \mathsf{E}\, e^{t\xi_1}$, $t \in R$, and let $\varphi(t_0)$ exist for some $t_0 \neq 0$ and $\varphi(t_0) \geq 1$.*

If τ is a stopping time (with respect to $(\mathscr{F}_n^\xi)$, $\mathscr{F}_n^\xi = \sigma\{\xi_1, \ldots, \xi_n\}, \tau \geq 1$), such that $|S_n| \leq C$ $(\{\tau \geq n\}; \mathsf{P}$-a.s.) and $\mathsf{E}\, \tau < \infty$, then

$$\mathsf{E}\left[\frac{e^{t_0 S_\tau}}{(\varphi(t_0))^\tau}\right] = 1. \tag{15}$$

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 $\{\tau \geq n\}$,

$$\begin{aligned}\mathsf{E}\{|Y_{n+1} - Y_n| \,|\, Y_1, \ldots, Y_n\} &= Y_n \,\mathsf{E}\left\{\left|\frac{e^{t_0\xi_{n+1}}}{\varphi(t_0)} - 1\right| \,\Big|\, \xi_1, \ldots, \xi_n\right\} \\ &= Y_n \cdot \mathsf{E}\,|e^{t_0\xi_1}(\varphi(t_0))^{-1} - 1| \leq C < \infty \quad (\mathsf{P}\text{-a.s.}),\end{aligned}$$

where C is a constant. Therefore Theorem 2 is applicable, and (15) 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 *mean duration* in games (Sect. 9, Chap. 1, Vol. 1).

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

$$\tau = \min\{n \geq 1 \colon S_n = B \text{ or } A\}, \tag{16}$$

where $(-A)$ and B are positive integers.

It follows from (20) (Sect. 9, Chap. 1, Vol. 1) 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 from (13)

$$0 = \mathsf{E}\, S_\tau = \alpha A + \beta B,$$

whence

$$\alpha = \frac{B}{B + |A|}, \quad \beta = \frac{|A|}{B + |A|}.$$

Applying (14), we obtain

$$\mathsf{E}\,\tau = \mathsf{E}\, S_\tau^2 = \alpha A^2 + \beta B^2 = |AB|.$$

However, if $p \neq q$, then 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

$$\alpha = \frac{\left(\frac{q}{p}\right)^{B} - 1}{\left(\frac{q}{p}\right)^{B} - \left(\frac{q}{p}\right)^{A}}, \qquad \beta = \frac{1 - \left(\frac{q}{p}\right)^{A}}{\left(\frac{q}{p}\right)^{B} - \left(\frac{q}{p}\right)^{A}}. \tag{17}$$

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 (17).

EXAMPLE 2. In the example considered above, let $p = q = \frac{1}{2}$. Let us show that for τ defined in (16) and every λ in $0 < \lambda < \pi/(B+|A|)$

$$\mathsf{E}(\cos\lambda)^{-\tau} = \frac{\cos\left(\lambda\frac{B+A}{2}\right)}{\cos\left(\lambda\frac{B+|A|}{2}\right)}. \tag{18}$$

For this purpose we consider the martingale $X = (X_n, \mathscr{F}_n^\xi)_{n\geq 0}$ with

$$X_n = (\cos\lambda)^{-n}\cos\left(\lambda\left(S_n - \frac{B+A}{2}\right)\right) \tag{19}$$

and $S_0 = 0$. It is clear that

$$\mathsf{E}\, X_n = \mathsf{E}\, X_0 = \cos\left(\lambda\frac{B+A}{2}\right). \tag{20}$$

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 $0 < \lambda < \pi/(B+|A|)$,

$$\begin{aligned}\mathsf{E}\, X_0 = \mathsf{E}\, X_{n\wedge\tau} &= \mathsf{E}(\cos\lambda)^{-(n\wedge\tau)}\cos\left(\lambda\left(S_{n\wedge\tau} - \frac{B+A}{2}\right)\right)\\ &\geq \mathsf{E}(\cos\lambda)^{-(n\wedge\tau)}\cos\left(\lambda\frac{B-A}{2}\right).\end{aligned}$$

Therefore, by (20),

$$\mathsf{E}(\cos\lambda)^{-(n\wedge\tau)} \leq \frac{\cos\left(\lambda\frac{B+A}{2}\right)}{\cos\left(\lambda\frac{B+|A|}{2}\right)},$$

and consequently, by Fatou's lemma,

$$\mathsf{E}(\cos\lambda)^{-\tau} \leq \frac{\cos\left(\lambda\frac{B+A}{2}\right)}{\cos\left(\lambda\frac{B+|A|}{2}\right)}. \tag{21}$$

Consequently, by (19),

$$|X_{n\wedge\tau}| \leq (\cos\lambda)^{-\tau}.$$

With (21), this establishes the uniform integrability of the family $\{X_{n\wedge\tau}\}$. Then, by Corollary 2 to Theorem 1,

$$\cos\left(\lambda\frac{B+A}{2}\right) = \mathsf{E}\,X_0 = \mathsf{E}\,X_\tau = \mathsf{E}(\cos\lambda)^{-\tau}\cos\left(\lambda\frac{B-A}{2}\right),$$

from which the required equality (18) follows.

4. As an application of Wald's identity (13), we will give the proof of the "elementary theorem" of renewal theory: If $N = (N_t)_{t\geq 0}$ is a renewal process ($N_t = \sum_{n=1}^{\infty} I(T_n \leq t)$, $T_n = \sigma_1 + \cdots + \sigma_n$, where $\sigma_1, \sigma_2, \ldots$ is a sequence of independent identically distributed random variables (Subsection 4, Sect. 9, Chap. 2, Vol. 1)) and $\mu = \mathsf{E}\,\sigma_1 < \infty$, then the renewal function $m(t) = \mathsf{E}\,N_t$ satisfies

$$\frac{m(t)}{t} \to \frac{1}{\mu}, \quad t \to \infty. \tag{22}$$

(Recall that the process $N_t = (N_t)_{t\geq 0}$ itself obeys the *strong law of large numbers*:

$$\frac{N_t}{t} \to \frac{1}{\mu} \quad (\mathsf{P}\text{-a.s.}), \quad t \to \infty;$$

see Example 4 in Sect. 3, Chap. 4.)

To prove (22), we will show that

$$\liminf_{t\to\infty} \frac{m(t)}{t} \geq \frac{1}{\mu} \quad \text{and} \quad \limsup_{t\to\infty} \frac{m(t)}{t} \leq \frac{1}{\mu}. \tag{23}$$

To this end we notice that

$$T_{N_t} \leq t < T_{N_t+1}, \quad t > 0. \tag{24}$$

Since for any $n \geq 1$

$$\{N_t + 1 \leq n\} = \{N_t \leq n-1\} = \{N_t < n\} = \{T_n > t\} = \left\{\sum_{k=1}^{n} \sigma_k > t\right\} \in \mathscr{F}_n,$$

where $\mathscr{F}_n$ is the σ-algebra generated by $\sigma_1, \ldots, \sigma_n$, we have that $N_t + 1$ (but not N_t) for any fixed $t > 0$ is a Markov time. Then Wald's identity (13) implies that

$$\mathsf{E}\,T_{N_t+1} = \mu[m(t) + 1]. \tag{25}$$

Hence we see from the right inequality in (24) that $t < \mu[m(t) + 1]$, i.e.,

$$\frac{m(t)}{t} > \frac{1}{\mu} - \frac{1}{t}, \tag{26}$$

whence, letting $t \to \infty$, we obtain the first inequality in (23).

Next, the left inequality in (24) implies that $t \geq \mathsf{E}\,T_{N_t}$. Since $T_{N_t+1} = T_{N_t} + \sigma_{N_t+1}$, we have

$$t \geq \mathsf{E}\,T_{N_t} = \mathsf{E}(T_{N_t+1} - \sigma_{N_t+1}) = \mu[m(t) + 1] - \mathsf{E}\,\sigma_{N_t+1}. \tag{27}$$

If we assume that the variables σ_i are bounded from above ($\sigma_i \le c$), then (27) implies that $t \ge \mu[m(t)+1] - c$, and hence

$$\frac{m(t)}{t} \le \frac{1}{\mu} + \frac{1}{t} \cdot \frac{c-\mu}{\mu}. \tag{28}$$

Then the second inequality in (23) would follow.

To discard the restriction $\sigma_i \le c$, $i \ge 1$, we introduce, for some $c > 0$, the variables

$$\sigma_i^c = \sigma_i I(\sigma_i < c) + cI(\sigma_i \ge c)$$

and define the related renewal process $N^c = (N_t^c)_{t\ge 0}$ with $N_t^c = \sum_{n=1}^{\infty} I(T_n^c \le t)$, $T_n^c = \sigma_1^c + \cdots + \sigma_n^c$. Since $\sigma_i^c \le \sigma_i$, $i \ge 1$, we have $N_t^c \ge N_t$; hence

$$m^c(t) = \mathsf{E}\, N_t^c \ge \mathsf{E}\, N_t = m(t).$$

Then we see from (28) that

$$\frac{m(t)}{t} \le \frac{m^c(t)}{t} \le \frac{1}{\mu^c} + \frac{1}{t} \cdot \frac{c-\mu^c}{\mu^c},$$

where $\mu^c = \mathsf{E}\, \sigma_1^c$.

Therefore

$$\limsup_{t\to\infty} \frac{m(t)}{t} \le \frac{1}{\mu^c}.$$

Letting now $c \to \infty$ and using that $\mu^c \to \mu$, we obtain the required second inequality in (23).

Thus (22) is established.

Remark. For more general results of renewal theory see, for example, [10, Chap. 9], [25, Chap. 13].

5. Problems

1. Show that
$$\mathsf{E}\,|X_\tau| \le \varliminf_{N\to\infty} \mathsf{E}\,|X_N|$$
for any martingale or nonnegative submartingale $X = (X_n, \mathscr{F}_n)_{n\ge 0}$ and any finite (P-a.s.) stopping time τ. (Compare with inequality $\mathsf{E}\,|X_\tau| \le 3 \sup_N \mathsf{E}\,|X_N|$ in Corollary 2 to Theorem 1.)
2. Let $X = (X_n, \mathscr{F}_n)_{n\ge 0}$ be a square-integrable martingale, $\mathsf{E}\, X_0 = 0$, τ a stopping time, and
$$\varliminf_{n\to\infty} \int_{\{\tau > n\}} X_n^2 \, d\mathsf{P} = 0.$$
Show that
$$\mathsf{E}\, X_\tau^2 = \mathsf{E}\langle X\rangle_\tau \quad \left(= \mathsf{E} \sum_{j=0}^{\tau} (\Delta X_j)^2 \right),$$
where $\Delta X_0 = X_0$, $\Delta X_j = X_j - X_{j-1}$, $j \ge 1$.

3. Let $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ be a supermartingale such that $X_n \geq \mathsf{E}(\xi \,|\, \mathscr{F}_n)$ (P-a.s.), $n \geq 0$, where $\mathsf{E}\,|\xi| < \infty$. Show that for stopping times σ and τ with $\mathsf{P}\{\sigma \leq \tau\} = 1$ the following relation holds:

$$X_\sigma \geq \mathsf{E}(X_\tau \,|\, \mathscr{F}_\sigma) \quad (\mathsf{P}\text{-a.s.}).$$

4. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables with $\mathsf{P}(\xi_1 = 1) = \mathsf{P}(\xi_1 = -1) = \frac{1}{2}$, 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

$$\tau = \min\{n \geq 1 \colon X_n \leq -r\}, \quad r > 0.$$

Show that $\mathsf{E}\, e^{\lambda\tau} < \infty$ for $\lambda \leq \alpha_0$ and $\mathsf{E}\, e^{\lambda\tau} = \infty$ for $\lambda > \alpha_0$, where

$$\alpha_0 = \frac{b}{a+b}\log\frac{2b}{a+b} + \frac{a}{a+b}\log\frac{2a}{a+b}.$$

5. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with $\mathsf{E}\,\xi_i = 0$, $\operatorname{Var}\xi_i = \sigma_i^2$, $S_n = \xi_1 + \cdots + \xi_n$, $\mathscr{F}_n^\xi = \sigma\{\xi_1, \ldots, \xi_n\}$. Prove the following generalizations of Wald's identities (13) and (14): 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

$$\mathsf{E}\, S_\tau^2 = \mathsf{E}\sum_{j=1}^{\tau}\xi_j^2 = \mathsf{E}\sum_{j=1}^{\tau}\sigma_j^2. \tag{29}$$

6. Let $X = (X_n, \mathscr{F})_{n\geq 1}$ be a square-integrable martingale and τ a stopping time. Establish the inequality

$$\mathsf{E}\, X_\tau^2 \leq \mathsf{E}\sum_{n=1}^{\tau}(\Delta X_n)^2.$$

Show that if

$$\liminf_{n\to\infty}\mathsf{E}(X_n^2\, I(\tau > n)) < \infty \quad \text{or} \quad \liminf_{n\to\infty}\mathsf{E}(|X_n|\, I(\tau > n)) = 0,$$

then $\mathsf{E}\, X_\tau^2 = \mathsf{E}\sum_{n=1}^{\tau}(\Delta X_n)^2$.

7. Let $X = (X_n, \mathscr{F}_n)_{n\geq 1}$ be a submartingale and $\tau_1 \leq \tau_2 \leq \ldots$ stopping times such that $\mathsf{E}\, X_{\tau_m}$ are defined and

$$\liminf_{n\to\infty}\mathsf{E}(X_n^+ I(\tau_m > n)) = 0, \quad m \geq 1.$$

Prove that the sequence $(X_{\tau_m}, \mathscr{F}_{\tau_m})_{m\geq 1}$ is a submartingale. (As usual, $\mathscr{F}_{\tau_m} = \{A \in \mathscr{F} \colon A \cap \{\tau_m = j\} \in \mathscr{F}_j,\ j \geq 1\}$.)

3. Fundamental Inequalities

1. Let $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ be a stochastic sequence,

$$X_n^* = \max_{0\leq j\leq n} |X_j|, \qquad \|X_n\|_p = (\mathsf{E}\,|X_n|^p)^{1/p}, \quad p > 0.$$

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 $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ be a submartingale. Then for all $\lambda > 0$*

$$\lambda\,\mathsf{P}\Big\{\max_{k\leq n} X_k \geq \lambda\Big\} \leq \mathsf{E}\Big[X_n^+ I\Big(\max_{k\leq n} X_k \geq \lambda\Big)\Big] \leq \mathsf{E}\,X_n^+, \tag{1}$$

$$\lambda\,\mathsf{P}\Big\{\min_{k\leq n} X_k \leq -\lambda\Big\} \leq \mathsf{E}\Big[X_n I\Big(\min_{k\leq n} X_k > -\lambda\Big)\Big] - \mathsf{E}\,X_0 \leq \mathsf{E}\,X_n^+ - \mathsf{E}\,X_0, \tag{2}$$

$$\lambda\,\mathsf{P}\Big\{\max_{k\leq n} |X_k| \geq \lambda\Big\} \leq 3\max_{k\leq n} \mathsf{E}\,|X_k|. \tag{3}$$

II. *Let $Y = (Y_n, \mathscr{F}_n)_{n\geq 0}$ be a supermartingale. Then for all $\lambda > 0$*

$$\lambda\,\mathsf{P}\Big\{\max_{k\leq n} Y_k \geq \lambda\Big\} \leq \mathsf{E}\,Y_0 - \mathsf{E}\Big[Y_n I\Big(\max_{k\leq n} Y_k < \lambda\Big)\Big] \leq \mathsf{E}\,Y_0 + \mathsf{E}\,Y_n^-, \tag{4}$$

$$\lambda\,\mathsf{P}\Big\{\min_{k\leq n} Y_k \leq -\lambda\Big\} \leq -\mathsf{E}\Big[Y_n I\Big(\min_{k\leq n} Y_k \leq -\lambda\Big)\Big] \leq \mathsf{E}\,Y_n^-, \tag{5}$$

$$\lambda\,\mathsf{P}\Big\{\max_{k\leq n} |Y_k| \geq \lambda\Big\} \leq 3\max_{k\leq n} \mathsf{E}\,|Y_k|. \tag{6}$$

III. *Let $Y = (Y_n, \mathscr{F}_n)_{n\geq 0}$ be a nonnegative supermartingale. Then for all $\lambda > 0$*

$$\lambda\,\mathsf{P}\Big\{\max_{k\leq n} Y_k \geq \lambda\Big\} \leq \mathsf{E}\,Y_0, \tag{7}$$

$$\lambda\,\mathsf{P}\Big\{\sup_{k\geq n} Y_k \geq \lambda\Big\} \leq \mathsf{E}\,Y_n. \tag{8}$$

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 $p = 1$,

$$\|X_n\|_1 \leq \|X_n^*\|_1 \leq \frac{e}{e-1}\{1 + \|X_n \log^+ X_n\|_1\}. \tag{10}$$

Theorem 3. *Let $X = (X_n, \mathscr{F}_n)_{n\geq 0}$ be a martingale, $\lambda > 0$ and $p \geq 1$. Then*

$$\mathsf{P}\Big\{\max_{k\leq n} |X_k| \geq \lambda\Big\} \leq \frac{\mathsf{E}\,|X_n|^p}{\lambda^p} \tag{11}$$

and if $p > 1$,

$$\|X_n\|_p \le \|X_n^*\|_p \le \frac{p}{p-1}\|X_n\|_p. \tag{12}$$

In particular, if $p = 2$,

$$\mathsf{P}\left\{\max_{k\le n}|X_k| \ge \lambda\right\} \le \frac{\mathsf{E}\,|X_n|^2}{\lambda^2}, \tag{13}$$

$$\mathsf{E}\left[\max_{k\le n} X_k^2\right] \le 4\,\mathsf{E}\,X_n^2. \tag{14}$$

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 $\tau = \min\{k \le n\colon Y_k \ge \lambda\}$ with $\tau = n$ if $\max_{k\le n} Y_k < \lambda$. Then, by (6), Sect. 2,

$$\begin{aligned}\mathsf{E}\,Y_0 \ge \mathsf{E}\,Y_\tau &= \mathsf{E}\left[Y_\tau;\, \max_{k\le n} Y_k \ge \lambda\right] + \mathsf{E}\left[Y_\tau;\, \max_{k\le n} Y_k < \lambda\right]\\ &\ge \lambda\,\mathsf{P}\left\{\max_{k\le n} Y_k \ge \lambda\right\} + \mathsf{E}\left[Y_n;\, \max_{k\le n} Y_k < \lambda\right],\end{aligned}$$

which proves (4).

Now let us set $\sigma = \min\{k \le n\colon Y_k \le -\lambda\}$ and take $\sigma = n$ if $\min_{k\le n} Y_k > -\lambda$. Again, by (6), Sect. 2,

$$\begin{aligned}\mathsf{E}\,Y_n \le \mathsf{E}\,Y_\tau &= \mathsf{E}\left[Y_\tau;\, \min_{k\le n} Y_k \le -\lambda\right] + \mathsf{E}\left[Y_\tau;\, \min_{k\le n} Y_k > -\lambda\right]\\ &\le -\lambda\,\mathsf{P}\left\{\min_{k\le n} Y_k \le -\lambda\right\} + \mathsf{E}\left[Y_n;\, \min_{k\le n} Y_k > -\lambda\right].\end{aligned}$$

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),

$$\begin{aligned}\lambda\,\mathsf{P}\left\{\max_{k\le n}|Y_k| \ge \lambda\right\} &\le \lambda\,\mathsf{P}\left\{\max_{k\le n} Y_k^+ \ge \lambda\right\} + \lambda\,\mathsf{P}\{\max_{k\le n} Y_k^- \ge \lambda\}\\ &= \lambda\,\mathsf{P}\left\{\max_{k\le n} Y_k \ge \lambda\right\} + \lambda\,\mathsf{P}\{\max_{k\le n} Y_k^- \ge \lambda\}\\ &\le \mathsf{E}\,Y_0 + 2\,\mathsf{E}\,Y_n^- \le 3\max_{k\le n}\mathsf{E}\,|Y_k|.\end{aligned}$$

Inequality (7) follows from (4).

To prove (8), we set $\gamma = \min\{k \ge n\colon Y_k \ge \lambda\}$, taking $\gamma = \infty$ if $Y_k < \lambda$ for all $k \ge n$. Now let $n < N < \infty$. Then, by (6), Sect. 2,

$$\mathsf{E}\, Y_n \geq \mathsf{E}\, Y_{\gamma\wedge N} \geq \mathsf{E}[Y_{\gamma\wedge N} I(\gamma \leq N)] \geq \lambda\, \mathsf{P}\{\gamma \leq N\},$$

from which, as $N \to \infty$,

$$\mathsf{E}\, Y_n \geq \lambda\, \mathsf{P}\{\gamma < \infty\} = \lambda\, \mathsf{P}\left\{ \sup_{k\geq n} Y_k \geq \lambda \right\}.$$

□

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 $r > 0$,

$$\mathsf{E}\, \xi^r = r \int_0^\infty t^{r-1}\, \mathsf{P}(\xi \geq t)\, dt \tag{16}$$

(see (69) in Sect. 6, Chap. 2, Vol. 1). Then we obtain, by (1) and Fubini's theorem, that for $p > 1$

$$\begin{aligned} \mathsf{E}(X_n^*)^p &= p \int_0^\infty t^{p-1}\, \mathsf{P}\{X_n^* \geq t\}\, dt \leq p \int_0^\infty t^{p-2} \left(\int_{\{X_n^* \geq t\}} X_n\, d\,\mathsf{P} \right) dt \\ &= p \int_0^\infty t^{p-2} \left[\int_\Omega X_n\, I\{X_n^* \geq t\}\, d\,\mathsf{P} \right] dt \\ &= p \int_\Omega X_n \left[\int_0^{X_n^*} t^{p-2}\, dt \right] d\,\mathsf{P} = \frac{p}{p-1}\, \mathsf{E}\left[X_n (X_n^*)^{p-1} \right]. \end{aligned} \tag{17}$$

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 $\mathsf{E}(X_n^* \wedge L)^p \leq L^p < \infty$, 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

$$\mathsf{E}\,X_n^* - 1 \le \mathsf{E}(X_n^* - 1)^+ = \int_0^\infty \mathsf{P}\{X_n^* - 1 \ge t\}\,dt$$
$$\le \int_0^\infty \frac{1}{1+t}\left[\int_{\{X_n^* \ge 1+t\}} X_n\,d\mathsf{P}\right] dt = \mathsf{E}\,X_n \int_0^{X_n^*-1} \frac{dt}{1+t} = \mathsf{E}\,X_n \log X_n^*.$$

Since, for arbitrary $a \ge 0$ and $b > 0$,

$$a \log b \le a \log^+ a + be^{-1}, \tag{19}$$

we have

$$\mathsf{E}\,X_n^* - 1 \le \mathsf{E}\,X_n \log X_n^* \le \mathsf{E}\,X_n \log^+ X_n + e^{-1}\,\mathsf{E}\,X_n^*.$$

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^* with $X_n^* \wedge L$.

This proves the theorem. □

PROOF OF THEOREM 3. The proof follows from the remark that $|X|^p$, $p \ge 1$, is a nonnegative submartingale (if $\mathsf{E}\,|X_n|^p < \infty$, $n \ge 0$), and from inequalities (1) and (9).

□

Corollary of Theorem 3. *Let $X_n = \xi_0 + \cdots + \xi_n$, $n \ge 0$, where $(\xi_k)_{k\ge 0}$ 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 (Sect. 2, Chap. 4).*

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

$$\mathsf{P}\{X_n^* \ge \varepsilon\} \le \frac{\mathsf{E}\,A_n}{\varepsilon}.$$

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 sequence A if

$$\mathsf{E}\,X_\tau \le \mathsf{E}\,A_\tau \tag{20}$$

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,*

$$\mathsf{P}\{X_\tau^* \geq \lambda\} \leq \frac{\mathsf{E}\, A_\tau}{\lambda}, \tag{21}$$

$$\mathsf{P}\{X_\tau^* \geq \lambda\} \leq \frac{1}{\lambda}\,\mathsf{E}(A_\tau \wedge a) + \mathsf{P}(A_\tau \geq a), \tag{22}$$

$$\|X_\tau^*\|_p \leq \left(\frac{2-p}{1-p}\right)^{1/p} \|A_\tau\|_p, \quad 0 < p < 1. \tag{23}$$

PROOF. We set

$$\sigma_n = \min\{j \leq \tau \wedge n \colon X_j \geq \lambda\},$$

taking $\sigma_n = \tau \wedge n$, if $\{\cdot\} = \varnothing$. Then

$$\mathsf{E}\, A_\tau \geq \mathsf{E}\, A_{\sigma_n} \geq \mathsf{E}\, X_{\sigma_n} \geq \int_{\{X^*_{\tau\wedge n} > \lambda\}} X_{\sigma_n} d\,\mathsf{P} \geq \lambda\,\mathsf{P}\{X^*_{\tau\wedge n} > \lambda\},$$

from which

$$\mathsf{P}\{X^*_{\tau\wedge n} > \lambda\} \leq \frac{1}{\lambda}\,\mathsf{E}\, A_\tau,$$

and we obtain (21) by Fatou's lemma.

For the proof of (22), we introduce the time

$$\gamma = \min\{j \colon A_{j+1} \geq a\},$$

setting $\gamma = \infty$ if $\{\cdot\} = \varnothing$. Then

$$\begin{aligned}
\mathsf{P}\{X_\tau^* \geq \lambda\} &= \mathsf{P}\{X_\tau^* \geq \lambda,\ A_\tau < a\} + \mathsf{P}\{X_\tau^* \geq \lambda,\ A_\tau \geq a\} \\
&\leq \mathsf{P}\{I_{\{A_\tau < a\}} X_\tau^* \geq \lambda\} + \mathsf{P}\{A_\tau \geq a\} \\
&\leq \mathsf{P}\{X^*_{\tau\wedge\gamma} \geq \lambda\} + \mathsf{P}\{A_\tau \geq a\} \leq \frac{1}{\lambda}\,\mathsf{E}\, A_{\tau\wedge\gamma} + \mathsf{P}\{A_\tau \geq a\} \\
&\leq \frac{1}{\lambda}\,\mathsf{E}(A_\tau \wedge a) + \mathsf{P}(A_\tau \geq a),
\end{aligned}$$

where we used (21) and the inequality $I_{\{A_\tau < a\}} X_\tau^* \leq X^*_{\tau\wedge\gamma}$. Finally, by (22),

$$\begin{aligned}
\|X_\tau^*\|_p^p = \mathsf{E}(X_\tau^*)^p &= \int_0^\infty \mathsf{P}\{(X_\tau^*)^p \geq t\}\, dt = \int_0^\infty \mathsf{P}\{X_\tau^* \geq t^{1/p}\}\, dt \\
&\leq \int_0^\infty t^{-1/p}\,\mathsf{E}[A_\tau \wedge t^{1/p}]\, dt + \int_0^\infty \mathsf{P}\{A_\tau^p \geq t\}\, dt \\
&= \mathsf{E}\int_0^{A_\tau^p} dt + \mathsf{E}\int_{A_\tau^p}^\infty (A_\tau t^{-1/p})\, dt + \mathsf{E}\, A_\tau^p = \frac{2-p}{1-p}\,\mathsf{E}\, A_\tau^p.
\end{aligned}$$

This completes the proof.

□

Remark. Let us suppose that the hypotheses of Theorem 4 are satisfied, except that the sequence $A = (A_n, \mathscr{F}_n)_{n\geq 0}$ is not necessarily predictable but has the property that for some positive constant c

$$\mathsf{P}\left\{\sup_{k\geq 1}|\Delta A_k|\leq c\right\}=1,$$

where $\Delta A_k = A_k - A_{k-1}$. Then the following inequality is satisfied (cf. (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 $\gamma = \min\{j\colon A_{j+1}\geq a\}$ with $\gamma=\min\{j\colon A_j\geq a\}$ 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 stated below.

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 $n\geq 1$.

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*

$$A_p\left\|\left(\sum_{i=1}^n\xi_j^2\right)^{1/2}\right\|_p\leq\left\|\sum_{j=1}^n\xi_j\right\|_p\leq B_p\left\|\left(\sum_{j=1}^n\xi_j^2\right)^{1/2}\right\|_p \tag{26}$$

for every $n\geq 1$.

The sequences $X=(X_n)$ with $X_n=\sum_{j=1}^n c_j\xi_j$ and $X_n=\sum_{j=1}^n\xi_j$ in (25) and (26) are martingales involving independent ξ_j. It is natural to ask whether these 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:

$$[X]_n = \sum_{j=1}^{n} (\Delta X_j)^2, \quad X_0 = 0. \tag{28}$$

The constants A_p and B_p can be taken to have the values

$$A_p = [18p^{3/2}/(p-1)]^{-1}, \quad B_p = 18p^{3/2}/(p-1)^{1/2}.$$

It follows from (27), using (12), 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

$$\tau = \min\left\{ n \ge 1 \colon \sum_{i=1}^{n} \xi_j = 1 \right\}.$$

The sequence $X = (X_n,\ \mathscr{F}_n^{\xi})$ is a martingale, with

$$\|X_n\|_1 = \mathsf{E}\,|X_n| = 2\,\mathsf{E}\,X_n^+ \to 2, \quad n \to \infty.$$

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 (29) rather than (27) (which is equivalent when $p > 1$).

Davis' 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.,

$$A\,\mathsf{E}\sqrt{\sum_{j=1}^{n}(\Delta X_j)^2} \le \mathsf{E}\left[\max_{1\le j\le n}|X_j|\right] \le B\,\mathsf{E}\sqrt{\sum_{j=1}^{n}(\Delta X_j)^2}.$$

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* (13) (Sect. 2)*, we have*

$$\mathsf{E}\,S_\tau = 0 \tag{31}$$

for every stopping time τ (with respect to $(\mathscr{F}_n^\xi)$) for which $\mathsf{E}\,\tau < \infty$.

If we assume additionally that $\mathsf{E}\,|\xi_1|^r < \infty$, where $1 < r \le 2$, then the condition $\mathsf{E}\,\tau^{1/r} < \infty$ is sufficient for (31).

For the proof, we set $\tau_n = \tau \wedge n$, $Y = \sup_n |S_{\tau_n}|$ and let $m = [t^r]$ (integral part of t^r) for $t > 0$. By Corollary 1 to Theorem 1 (Sect. 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

$$\begin{aligned}
\mathsf{P}(Y \ge t) &= \mathsf{P}(\tau \ge t^r,\ Y \ge t) + \mathsf{P}(\tau < t^r,\ Y \ge t)\\
&\le \mathsf{P}(\tau \ge t^r) + \mathsf{P}\left\{\max_{1\le j\le m}|S_{\tau_j}| \ge t\right\}\\
&\le \mathsf{P}(\tau \ge t^r) + t^{-r}\,\mathsf{E}\,|S_{\tau_m}|^r\\
&\le \mathsf{P}(\tau \ge t^r) + t^{-r}B_r^r\,\mathsf{E}\left(\sum_{j=1}^{\tau_m}\xi_j^2\right)^{r/2}\\
&\le \mathsf{P}(\tau \ge t^r) + t^{-r}B_r^r\,\mathsf{E}\sum_{j=1}^{\tau_m}|\xi_j|^r.
\end{aligned}$$

Notice that (with $\mathscr{F}_0^\xi = \{\varnothing,\ \Omega\}$)

$$\begin{aligned}
\mathsf{E}\sum_{j=1}^{\tau_m}|\xi_j|^r &= \mathsf{E}\sum_{j=1}^{\infty}I(j \le \tau_m)|\xi_j|^r\\
&= \sum_{j=1}^{\infty}\mathsf{E}\,\mathsf{E}[I(j \le \tau_m)|\xi_j|^r \mid \mathscr{F}_{j-1}^\xi]\\
&= \mathsf{E}\sum_{j=1}^{\infty}I(j \le \tau_m)\,\mathsf{E}[|\xi|^r \mid \mathscr{F}_{j-1}^\xi] = \mathsf{E}\sum_{j=1}^{\tau_m}\mathsf{E}\,|\xi_j|^r = \mu_r\,\mathsf{E}\,\tau_m,
\end{aligned}$$

where $\mu_r = \mathsf{E}\,|\xi_1|^r$. Consequently,

$$\begin{aligned}\mathsf{P}(Y \geq t) &\leq \mathsf{P}(\tau \geq t^r) + t^{-r} B_r^r \mu_r \,\mathsf{E}\,\tau_m \\ &= \mathsf{P}(\tau \geq t^r) + B_r^r \mu_r t^{-r} \left[m\,\mathsf{P}(\tau \geq t^r) + \int_{\{\tau < t^r\}} \tau \, d\,\mathsf{P} \right] \\ &\leq (1 + B_r^r \mu_r)\,\mathsf{P}(\tau \geq t^r) + B_r^r \mu_r t^{-r} \int_{\{\tau < t^r\}} \tau \, d\,\mathsf{P}\end{aligned}$$

and therefore

$$\begin{aligned}\mathsf{E}\,Y = \int_0^\infty \mathsf{P}(Y \geq t)\,dt &\leq (1 + B_r^r \mu_r)\,\mathsf{E}\,\tau^{1/r} + B_r^r \mu_r \int_0^\infty t^{-r} \left[\int_{\{\tau < t^r\}} \tau \, d\,\mathsf{P} \right] dt \\ &= (1 + B_r^r \mu_r)\,\mathsf{E}\,\tau^{1/r} + B_r^r \mu_r \int_\Omega \tau \left[\int_{\tau^{1/r}}^\infty t^{-r} dt \right] d\,\mathsf{P} \\ &= \left(1 + B_r^r \mu_r + \frac{B_r^r \mu_r}{r-1} \right) \mathsf{E}\,\tau^{1/r} < \infty.\end{aligned}$$

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 (cf. Theorem 2 in Sect. 3, Chap. 4) we have the strong law of large numbers:

$$\frac{M_n}{n} \to 0 \quad (\mathsf{P}\text{-}a.s.), \quad n \to \infty. \tag{33}$$

When $r = 1$, the proof follows the same lines as the proof of Theorem 2 in Sect. 3, Chap. 4. 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 (Sect. 3, Chap. 4), a sufficient condition for the limit relation (P-a.s.)

$$\frac{1}{n} \sum_{k=1}^n k \Delta m_k \to 0, \quad n \to \infty,$$

is that the limit $\lim_n m_n$ exists and is finite (P-a.s.), which in turn (Theorems 1 and 4 in Sect. 10, Chap. 2, Vol. 1) is true if and only if

$$\mathsf{P}\left\{ \sup_{k \geq 1} |m_{n+k} - m_n| \geq \varepsilon \right\} \to 0, \quad n \to \infty. \tag{34}$$

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 statement (33) is equivalent (Theorem 1 of Sect. 10, Chap. 2, Vol. 1) to the statement that

$$\varepsilon^{2r}\,\mathsf{P}\left\{\sup_{j\geq n}\frac{|M_j|}{j}\geq\varepsilon\right\}\to 0,\quad n\to\infty, \tag{35}$$

for every $\varepsilon>0$. By inequality (52) of Problem 1,

$$\begin{aligned}\varepsilon^{2r}\,\mathsf{P}\left\{\sup_{j\geq n}\frac{|M_j|}{j}\geq\varepsilon\right\}&=\varepsilon^{2r}\lim_{m\to\infty}\mathsf{P}\left\{\max_{n\leq j\leq m}\frac{|M_j|^{2r}}{j^{2r}}\geq\varepsilon^{2r}\right\}\\&\leq\frac{1}{n^{2r}}\,\mathsf{E}\,|M_n|^{2r}+\sum_{j\geq n+1}\frac{1}{j^{2r}}\,\mathsf{E}(|M_j|^{2r}-|M_{j-1}|^{2r}).\end{aligned}$$

It follows from Kronecker's lemma 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

$$\begin{aligned}I_N&=\sum_{j=2}^{N}\frac{1}{j^{2r}}\left[\mathsf{E}\,|M_j|^{2r}-\mathsf{E}\,|M_{j-1}|^{2r}\right]\\&\leq\sum_{j=2}^{N}\left[\frac{1}{(j-1)^{2r}}-\frac{1}{j^{2r}}\right]\mathsf{E}\,|M_{j-1}|^{2r}+\frac{\mathsf{E}\,|M_N|^{2r}}{N^{2r}}.\end{aligned}$$

By Burkholder's inequality (27) and Hölder's inequality,

$$\mathsf{E}\,|M_j|^{2r}\leq B_{2r}^{2r}\,\mathsf{E}\left[\sum_{i=1}^{j}(\Delta M_i)^2\right]^r\leq B_{2r}^{2r}\sum_{i=1}^{j}|\Delta M_i|^{2r}.$$

Hence

$$I_N\leq\sum_{j=2}^{N-1}B_{2r}^{2r}\left[\frac{1}{j^{2r}}-\frac{1}{(j+1)^{2r}}\right]j^{r-1}\sum_{i=1}^{j}\mathsf{E}\,|\Delta M_i|^{2r}\,\frac{\mathsf{E}\,|M_N|^{2r}}{N^{2r}}$$

$$\leq C_1 \sum_{j=2}^{N-1} \frac{1}{j^{r+2}} \sum_{i=1}^{j} \mathsf{E}\,|\Delta M_i|^{2r}\, \frac{\mathsf{E}\,|M_N|^{2r}}{N^{2r}}$$

$$\leq C_2 \sum_{j=2}^{N} \frac{\mathsf{E}\,|\Delta M_j|^{2r}}{j^{r+1}} + C_3$$

(C_k 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. In what follows, Theorem 5 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 $X = (X_n, \mathscr{F}_n)$:

$$\begin{aligned}
\tau_0 &= 0,\\
\tau_1 &= \min\{n > 0\colon X_n \leq a\},\\
\tau_2 &= \min\{n > \tau_1\colon X_n \geq b\},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
\tau_{2m-1} &= \min\{n > \tau_{2m-2}\colon X_n \leq a\},\\
\tau_{2m} &= \min\{n > \tau_{2m-1}\colon X_n \geq b\},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots
\end{aligned}$$

taking $\tau_k = \infty$ if the corresponding set $\{\cdot\}$ is empty.

In addition, for each $n \geq 1$ we define the random variables

$$\beta_n(a,b) = \begin{cases} 0, & \text{if } \tau_2 > n,\\ \max\{m\colon \tau_{2m} \leq n\} & \text{if } \tau_2 \leq n. \end{cases}$$

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

$$\mathsf{E}\,\beta_n(0,b) \leq \frac{\mathsf{E}\,X_n}{b}. \tag{38}$$

Set $X_0 = 0, \mathscr{F}_0 = \{\varnothing, \Omega\}$, and for $i = 1, 2, \ldots$, let

$$\varphi_i = \begin{cases} 1 & \text{if } \tau_m < i \le \tau_{m+1} \text{ for some odd } m, \\ 0 & \text{if } \tau_m < i \le \tau_{m+1} \text{ for some even } m. \end{cases}$$

It is easily seen that

$$b\beta_n(0,b) \le \sum_{i=1}^{n} \varphi_i [X_i - X_{i-1}]$$

and

$$\{\varphi_i = 1\} = \bigcup_{\text{odd } m} [\{\tau_m < i\} \backslash \{\tau_{m+1} < i\}] \in \mathscr{F}_{i-1}.$$

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 probabilities of *large deviations* for square-integrable martingales.

Let $M = (M_n, \mathscr{F}_n)_{n \ge 0}$ be a square-integrable martingale with quadratic characteristic $\langle M \rangle = (\langle M \rangle_n, \mathscr{F}_{n-1})$, setting $M_0 = 0$. If we apply inequality (22) to $X_n = M_n^2$, $A_n = \langle M \rangle_n$, 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 where $|\Delta M_n| \le C$ for all n and $\omega \in \Omega$, this inequality can be substantially improved using the ideas explained in Sect. 5 of Chap. 4 for estimating the probabilities of large deviations for sums of independent identically distributed random variables.

Let us recall that in Sect. 5, Chap. 4, when we introduced the corresponding inequalities, the essential point was to use the property that the sequence

$$(e^{\lambda S_n}/[\varphi(\lambda)]^n,\ \mathscr{F}_n)_{n \ge 1}, \quad \mathscr{F}_n = \sigma\{\xi_1, \ldots, \xi_n\}, \tag{40}$$

formed a nonnegative martingale, to which we could apply inequality (8). If we now take M_n instead of S_n, by analogy with (40), then

$$(e^{\lambda M_n}/\mathscr{E}_n(\lambda),\ \mathscr{F}_n)_{n\geq 1}$$

will be a nonnegative martingale, 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* (see also Subsection 13, Sect. 6, Chap. 2, Vol. 1).

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 simply depends on M_n and $\langle M\rangle_n$, and to which we can apply the method used in Sect. 5, Chap. 4.

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 $\lambda > 0$,*

$$\psi_c(\lambda) = \begin{cases} (e^{\lambda c} - 1 - \lambda^c)/c^2, & c > 0, \\ \frac{1}{2}\lambda^2, & c = 0, \end{cases} \tag{42}$$

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 nonnegative 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

$$\begin{aligned} &\mathsf{E}(\Delta Z_n \mid \mathscr{F}_{n-1}) \\ &\quad = Z_{n-1}[\mathsf{E}(e^{\lambda \Delta M_n} - 1 \mid \mathscr{F}_{n-1})\, e^{-\Delta\langle M\rangle_n \psi_c(\lambda)} + (e^{-\Delta\langle M\rangle_n \psi_c(\lambda)} - 1)] \\ &\quad = Z_{n-1}[\mathsf{E}(e^{\lambda \Delta M_n} - 1 - \lambda \Delta M_n \mid \mathscr{F}_{n-1})\, e^{-\Delta\langle M\rangle_n \psi_c(\lambda)} + (e^{-\Delta\langle M\rangle_n \psi_c(\lambda)} - 1)] \\ &\quad \leq Z_{n-1}[\psi_c(\lambda)\, \mathsf{E}((\Delta M_n)^2 \mid \mathscr{F}_{n-1})\, e^{-\Delta\langle M\rangle_n \psi_c(\lambda)} + (e^{-\Delta\langle M\rangle_n \psi_c(\lambda)} - 1)] \\ &\quad = Z_{n-1}[\psi_c(\lambda)\Delta\langle M\rangle_n\, e^{-\Delta\langle M\rangle_n \psi_c(\lambda)} + (e^{-\Delta\langle M\rangle_n \psi_c(\lambda)} - 1)] \leq 0, \end{aligned} \tag{44}$$

where we have also used the fact that $xe^{-x} + (e^{-x} - 1) \leq 0$ for $x \geq 0$.

We see from (44) that

$$\mathsf{E}(Z_n \mid \mathscr{F}_{n-1}) \le Z_{n-1},$$

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

$$\begin{aligned}
\mathsf{P}\left\{\max_{k\le n} M_k \ge an\right\} &= \mathsf{P}\left\{\max_{k\le n} e^{\lambda M_k} \ge e^{\lambda an}\right\} \\
&\le \mathsf{P}\left\{\max_{k\le n} e^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} \ge e^{\lambda an - \psi_c(\lambda)\langle M\rangle_n}\right\} \\
&= \mathsf{P}\left\{\max_{k\le n} e^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} \ge e^{\lambda an - \psi_c(\lambda)\langle M\rangle_n},\ \langle M\rangle_n \le bn\right\} \\
&\quad + \mathsf{P}\left\{\max_{k\le n} e^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} \ge e^{\lambda an - \psi_c(\lambda)\langle M\rangle_n},\ \langle M\rangle_n > bn\right\} \\
&\le \mathsf{P}\left\{\max_{k\le n} e^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} \ge e^{\lambda an - \psi_c(\lambda) bn}\right\} + \mathsf{P}\{\langle M\rangle_n > bn\} \\
&\le e^{-n(\lambda a - b\psi_c(\lambda))} + \mathsf{P}\{\langle M\rangle_n > bn\}, \qquad (45)
\end{aligned}$$

where the last inequality follows from (7).

Let us write (compare with $H(a)$ in Sect. 5, Chap. 4)

$$H_c(a,b) = \sup_{\lambda>0}[a\lambda - b\psi_c(\lambda)].$$

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 inequalities (46) and (47).*

Remark 2.

$$H_c(a,b) = \frac{1}{c}\left(a + \frac{b}{c}\right)\log\left(1 + \frac{ac}{b}\right) - \frac{a}{c}. \qquad (48)$$

6. Under the hypotheses of Theorem 6, we now consider the question of estimating 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 rate of convergence in the strong law of large numbers for martingales (also see Theorem 4 in Sect. 5).

Proceeding as in Sect. 5, Chap. 4, we find that for every $a > 0$ there is a $\lambda > 0$ for which $a\lambda - \psi_c(\lambda) > 0$. Then, for every $b > 0$,

$$\begin{aligned}\mathsf{P}\left\{\sup_{k\ge n}\frac{M_k}{\langle M\rangle_k} > a\right\} &\le \mathsf{P}\left\{\sup_{k\ge n} e^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} > e^{[\alpha\lambda-\psi_c(\lambda)]\langle M\rangle_n}\right\}\\ &\le \mathsf{P}\left\{\sup_{k\ge n} e^{\lambda M_k - \psi_c(\lambda)\langle M\rangle_k} > e^{[a\lambda-\psi_c(\lambda)]bn}\right\} + \mathsf{P}\{\langle M\rangle_n < bn\}\\ &\le e^{-bn[a\lambda-\psi_c(\lambda)]} + \mathsf{P}\{\langle M\rangle_n < bn\},\end{aligned} \tag{49}$$

from which

$$\mathsf{P}\left\{\sup_{k\ge n}\frac{M_k}{\langle M\rangle_k} > a\right\} \le \mathsf{P}\{\langle M\rangle_n < bn\} + e^{-nH_c(ab,b)}, \tag{50}$$

$$\mathsf{P}\left\{\sup_{k\ge n}\left|\frac{M_k}{\langle M\rangle_k}\right| > a\right\} \le 2\,\mathsf{P}\{\langle M\rangle_n < bn\} + 2e^{-nH_c(ab,b)}. \tag{51}$$

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. A comparison of (51) with estimate (21) in Sect. 5, Chap. 4, 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 a similar result:

$$\mathsf{P}\left\{\sup_{k\ge n}\left|\frac{M_k}{\langle M\rangle_k}\right| > \varepsilon\right\} = \mathsf{P}\left\{\sup_{k\ge n}\left|\frac{S_k - (k/2)}{k}\right| > \frac{\varepsilon}{4}\right\} \le 2e^{-4\varepsilon^2 n}.$$

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 $0 \le V_{n+1} \le V_n \le C$ (P-a.s.), where C is a constant. Establish the following generalization of (1):

$$\varepsilon\,\mathsf{P}\left\{\max_{i\le j\le n} V_j X_j \ge \varepsilon\right\} + \int_{\{\max_{1\le j\le n} V_j X_j < \varepsilon\}} V_n X_n\, d\mathsf{P} \le \sum_{j=1}^{n} \mathsf{E}\, V_j \Delta X_j. \tag{52}$$

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*:

$$\mathsf{P}\left\{\max_{1\le j\le n} |S_j| > 2\varepsilon\right\} \le \frac{\mathsf{P}\{|S_n| > \varepsilon\}}{\min_{1\le j\le n} \mathsf{P}\{|S_{j,n}| \le \varepsilon\}}$$

and deduce (assuming $\mathsf{E}\,\xi_i = 0,\ i \ge 1$) 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-algebras $\mathscr{F}_0, \ldots, \mathscr{F}_n$ 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 $\varepsilon > 0$,

$$\mathsf{P}\left[\bigcup_{k=1}^n A_k\right] \le \varepsilon + \mathsf{P}\left[\sum_{k=1}^n \mathsf{P}(A_k \mid \mathscr{F}_{k-1}) > \varepsilon\right].$$

8. Let $X = (X_n)_{n\ge1}$ be a square-integrable martingale and $(b_n)_{n\ge1}$ a nondecreasing sequence of positive real numbers. Prove the following *Hájek–Rényi inequality*:

$$\mathsf{P}\left\{\max_{1\le k\le n}\left|\frac{X_k}{b_k}\right| \ge \lambda\right\} \le \frac{1}{\lambda^2}\sum_{k=1}^n \frac{\mathsf{E}(\Delta X_k)^2}{b_n^2}, \quad \Delta X_k = X_k - X_{k-1},\ X_0 = 0.$$

9. Let $X = (X_n)_{n\ge1}$ be a submartingale and $g(x)$ a nonnegative increasing convex function. Then, for any $t > 0$ and real x,

$$\mathsf{P}\left\{\max_{1\le k\le n} X_k \ge x\right\} \le \frac{\mathsf{E}\,g(tX_n)}{g(tx)}.$$

In particular,

$$\mathsf{P}\left\{\max_{1\le k\le n} X_k \ge x\right\} \le e^{-tx}\,\mathsf{E}\,e^{tX_n}.$$

10. Let $\xi_1, \xi_2, \ldots$ be independent random variables with $\mathsf{E}\,\xi_n = 0, \mathsf{E}\,\xi_n^2 = 1, n \ge 1$. Let

$$\tau = \min\left\{n \ge 1 \colon \sum_{i=1}^n \xi_i > 0\right\}.$$

Prove that $\mathsf{E}\,\tau^{1/2} < \infty$.

11. Let $\xi = (\xi_n)_{n\geq 1}$ be a martingale difference and $1 < p \leq 2$. Show that

$$\mathsf{E}\sup_{n\geq 1}\left|\sum_{j=1}^{n}\xi_j\right|^p \leq C_p\sum_{j=1}^{\infty}\mathsf{E}\,|\xi_j|^p$$

for a constant C_p.

12. Let $X = (X_n)_{n\geq 1}$ be a martingale with $\mathsf{E}\,X_n = 0$ and $\mathsf{E}\,X_n^2 < \infty$. As a generalization of Problem 5 of Sect. 2, Chap. 4, show that for any $n \geq 1$ and $\varepsilon > 0$

$$\mathsf{P}\Big\{\max_{1\leq k\leq n} X_k > \varepsilon\Big\} \leq \frac{\mathsf{E}\,X_n^2}{\varepsilon^2 + \mathsf{E}\,X_n^2}.$$

4. General Theorems on 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

$$\mathsf{P}(\limsup X_n > \liminf X_n) > 0. \tag{2}$$

Then, since

$$\{\limsup X_n > \liminf X_n\} = \bigcup_{a<b}\{\limsup X_n > b > a > \liminf X_n\}$$

(here a and b are rational numbers), there are values a and b such that

$$\mathsf{P}\{\limsup X_n > b > a > \liminf X_n\} > 0. \tag{3}$$

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 (37), Sect. 3,

$$\mathsf{E}\,\beta_n(a,b) \leq \frac{\mathsf{E}[X_n - a]^+}{b-a} \leq \frac{\mathsf{E}\,X_n^+ + |a|}{b-a}$$

and therefore

$$\mathsf{E}\,\beta_\infty(a,b) = \lim_n \mathsf{E}\,\beta_n(a,b) \le \frac{\sup_n \mathsf{E}\,X_n^+ + |a|}{b-a} < \infty,$$

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 $\mathsf{E}\,X_n^+ \le \mathsf{E}\,|X_n| = 2\,\mathsf{E}\,X_n^+ - \mathsf{E}\,X_n \le 2\,\mathsf{E}\,X_n^+ - \mathsf{E}\,X_1$). 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,

$$\mathsf{E}\,|X_\infty| \le \sup_n \mathsf{E}\,|X_n| < \infty.$$

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 $X = (X_n, \mathscr{F}_n)_{n\ge 1}$ is a nonpositive submartingale, then the sequence $\overline{X} = (X_n, \mathscr{F}_n)$ with $1 \le n \le \infty$, $X_\infty = \lim X_n$ and $\mathscr{F}_\infty = \sigma\{\bigcup \mathscr{F}_n\}$ is a (nonpositive) submartingale.*

In fact, by Fatou's lemma,

$$\mathsf{E}\,X_\infty = \mathsf{E}\,\lim X_n \ge \limsup \mathsf{E}\,X_n \ge \mathsf{E}\,X_1 > -\infty$$

and (P-a.s.)

$$\mathsf{E}(X_\infty \mid \mathscr{F}_m) = \mathsf{E}(\lim X_n \mid \mathscr{F}_m) \ge \limsup \mathsf{E}(X_n \mid \mathscr{F}_m) \ge X_m.$$

Corollary 3. *If $X = (X_n, \mathscr{F}_n)$ is a nonnegative supermartingale (or, in particular, a nonnegative martingale), then* $\lim X_n$ *exists with probability* 1.

In fact, in that case,

$$\sup_n \mathsf{E}\,|X_n| = \sup_n \mathsf{E}\,X_n = \mathsf{E}\,X_1 < \infty,$$

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 $\mathscr{F}_n^\xi = \sigma\{\xi_1, \ldots, \xi_n\}$ 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 (16) of Subsection 5, Sect. 6, Chap. 2, Vol. 1), 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 $\overline{X} = (X_n, \mathscr{F}_n)$, $1 \le n \le \infty$, 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 (Sect. 6, Chap. 2, Vol. 1).

Moreover, if $A \in \mathscr{F}_n$ and $m \ge n$, then

$$\mathsf{E}\, I_A |X_m - X_\infty| \to 0, \quad m \to \infty,$$

and therefore

$$\lim_{m\to\infty} \int_A X_m \, d\mathsf{P} = \int_A X_\infty \, d\mathsf{P}.$$

The sequence $\left(\int_A X_m \, d\mathsf{P} \right)_{m \ge n}$ 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 Sect. 6, Chap. 2, Vol. 1, 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 $(\mathscr{F}_n)_{n\ge1}$ 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,*

$$\mathsf{E}(\xi \,|\, \mathscr{F}_n) \to \mathsf{E}(\xi \,|\, \mathscr{F}_\infty), \quad n \to \infty. \tag{7}$$

PROOF. Let $X_n = \mathsf{E}(\xi \mid \mathscr{F}_n)$, $n \geq 1$. Then, with $a > 0$ and $b > 0$,

$$\begin{aligned}\int_{\{|X_n| \geq a\}} |X_n| \, d\mathsf{P} &\leq \int_{\{|X_n| \geq a\}} \mathsf{E}(|\xi| \mid \mathscr{F}_n) \, d\mathsf{P} = \int_{\{|X_n| \geq a\}} |\xi| \, d\mathsf{P} \\ &= \int_{\{|X_n| \geq a, |\xi| \leq b\}} |\xi| \, d\mathsf{P} + \int_{\{|X_n| \geq a, |\xi| > b\}} |\xi| \, d\mathsf{P} \\ &\leq b \, \mathsf{P}\{|X_n| \geq a\} + \int_{\{|\xi| > b\}} |\xi| \, d\mathsf{P} \\ &\leq \frac{b}{a} \, \mathsf{E}\,|\xi| + \int_{\{|\xi| > b\}} |\xi| \, d\mathsf{P}.\end{aligned}$$

Letting $a \to \infty$ and then $b \to \infty$, we obtain

$$\lim_{a \to \infty} \sup_n \int_{\{|X_n| \geq a\}} |X_n| \, d\mathsf{P} = 0,$$

i.e., the family $\{X_n\}$ is uniformly integrable. Therefore, by Theorem 2, there is a random variable X_∞ such that $X_n = \mathsf{E}(\xi \mid \mathscr{F}_n) \to X_\infty$ (P-a.s. and in the L^1 sense). Hence we only have to show that

$$X_\infty = \mathsf{E}(\xi \mid \mathscr{F}_\infty) \quad (\mathsf{P}\text{-a.s.}).$$

Let $m \geq n$ and $A \in \mathscr{F}_n$. Then

$$\int_A X_m \, d\mathsf{P} = \int_A X_n \, d\mathsf{P} = \int_A \mathsf{E}(\xi \mid \mathscr{F}_n) \, d\mathsf{P} = \int_A \xi \, d\mathsf{P}.$$

Since the family $\{X_n\}$ is uniformly integrable and since, by Theorem 5, Sect. 6, Chap. 2, Vol. 1, 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 from an algebra to the smallest σ-algebra containing it (Carathéodory's theorem, Sect. 3, Chap. 2, Vol. 1), Eq. (8) remains valid for sets $A \in \mathscr{F}_\infty = \sigma(\bigcup \mathscr{F}_n)$. Thus

$$\int_A X_\infty \, d\mathsf{P} = \int_A \xi \, d\mathsf{P} = \int_A \mathsf{E}(\xi \mid \mathscr{F}_\infty) \, d\mathsf{P}, \quad A \in \mathscr{F}_\infty. \tag{9}$$

Since X_∞ and $\mathsf{E}(\xi \mid \mathscr{F}_\infty)$ are $\mathscr{F}_\infty$-measurable, it follows from Property I of Subsection 3, Sect. 6, Chap. 2, Vol. 1, and from (9) that $X_\infty = \mathsf{E}(\xi \mid \mathscr{F}_\infty)$ (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 $X_n = \mathsf{E}(\xi \mid \mathscr{F}_n)$ for all $n \geq 1$. Here $X_n \to \mathsf{E}(\xi \mid \mathscr{F}_\infty)$ (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 $X_n = \mathsf{E}(X_\infty \mid \mathscr{F}_n)$. 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–one law.* Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables, $\mathscr{F}_n^\xi = \sigma\{\xi_1, \ldots, \xi_n\}$, let $\mathscr{X}$ be the σ-algebra of the "tail" events, and $A \in \mathscr{X}$. By Theorem 3, we have $\mathsf{E}(I_A \mid \mathscr{F}_n^\xi) \to \mathsf{E}(I_A \mid \mathscr{F}_\infty^\xi) = I_A$ (P-a.s.). But I_A and $(\xi_1, \ldots, \xi_n)$ are independent. Since $\mathsf{E}(I_A \mid \mathscr{F}_n^\xi) = \mathsf{E}\, I_A$, 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

$$\xi_n(x) = \sum_{k=1}^{2^n} \frac{k-1}{2^n} I\left\{\frac{k-1}{2^n} \leq x < \frac{k}{2^n}\right\},$$

$\mathscr{F}_n = \sigma\{\xi_1, \ldots, \xi_n\} = \sigma\{\xi_n\}$, 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

$$\begin{aligned}\mathsf{E}[X_{n+1} \mid \mathscr{F}_n] &= \mathsf{E}[X_{n+1} \mid \xi_n] = 2^{n+1}\, \mathsf{E}[f(\xi_{n+1} + 2^{-(n+1)}) - f(\xi_{n+1}) \mid \xi_n] \\ &= 2^{n+1}\{\tfrac{1}{2}[f(\xi_n + 2^{-(n+1)}) - f(\xi_n)] + \tfrac{1}{2}[f(\xi_n + 2^{-n}) - f(\xi_n + 2^{-n+1})]\} \\ &= 2^n\{f(\xi_n + 2^{-n}) - f(\xi_n)\} = X_n.\end{aligned}$$

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

$$X_n = \mathsf{E}[g \mid \mathscr{F}_n]. \tag{11}$$

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 $\{H_n(x)\}_{n\ge1}$, as defined in Example 3 of Sect. 11, Chap. 2, Vol. 1. Put $\mathscr{F}_n = \sigma\{H_1, \dots, H_n\}$, 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

$$\mathsf{E}[f(x) \mid \mathscr{F}_n] = \sum_{k=1}^{n} a_k H_k(x) \quad (\mathsf{P}\text{-a.s.}) \tag{12}$$

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 $\mathsf{E}[f(x) \mid \mathscr{F}_n]$ is a partial sum of the Fourier series of $f(x)$ in the Haar system. Then, if we apply Theorem 3 to the martingale $(\mathsf{E}(f \mid \mathscr{F}_n), \mathscr{F}_n)$, 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 $(\xi_n)_{n\ge1}$ be a sequence of random variables. By Theorem 2 of Sect. 10, Chap. 2, Vol. 1, the P-a.s. 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 1.

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 $|\mathsf{E}\, e^{it_0 S_n}| \geq c > 0$ for all $n \geq n_0$, where c is a constant.

For $n \geq n_0$, we form the sequence $X = (X_n, \mathscr{F}_n)$ with

$$X_n = \frac{e^{it_0 S_n}}{\mathsf{E}\, e^{it_0 S_n}}, \quad \mathscr{F}_n = \sigma\{\xi_1, \ldots, \xi_n\}.$$

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 1 the limit $\lim_n X_n$ exists and is finite. Therefore the limit $\lim_{n\to\infty} e^{it_0 S_n}$ also exists with probability 1. Consequently, we can assert that there is a $\delta > 0$ such that for each t in the set $T = \{t\colon |t| < \delta\}$ the limit $\lim_n e^{itS_n}$ exists with probability 1.

Let $T \times \Omega = \{(t, \omega)\colon t \in T, \omega \in \Omega\}$, let $\overline{\mathscr{B}}(T)$ be the σ-algebra of Lebesgue sets on T, and let λ be Lebesgue measure on $(T, \overline{\mathscr{B}}(T))$. Also, let

$$C = \left\{(t,\ \omega) \in T \times \Omega\colon\ \lim_n e^{itS_n(\omega)} \text{ exists}\right\}.$$

It is clear that $C \in \overline{\mathscr{B}}(T) \otimes \mathscr{F}$.

It was shown earlier that $\mathsf{P}(C_t) = 1$ for every $t \in T$, where $C_t = \{\omega \in \Omega\colon (t, \omega) \in C\}$ is the section of C at point t. By Fubini's theorem (Theorem 8 of Sect. 6, Chap. 2, Vol. 1),

$$\begin{aligned}\int_{T\times\Omega} I_C(t,\omega)\, d(\lambda \times \mathsf{P}) &= \int_T \left(\int_\Omega I_C(t,\omega)\, d\mathsf{P}\right) d\lambda \\ &= \int_T \mathsf{P}(C_t)\, d\lambda = \lambda(T) = 2\delta > 0.\end{aligned}$$

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 $C_\omega = \{t\colon (t,\omega) \in C\}$.

Hence it follows that there is a set $\tilde{\Omega}$ with $\mathsf{P}(\tilde{\Omega}) = 1$ 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 $\mathsf{P}(\tilde{\Omega}) = 1$, the limit $\lim_n S_n(\omega)$ exists and is finite with probability 1.

5. Problems

1. Let $\{\mathscr{G}_n\}$ be a *nonincreasing* family of σ-algebras, $\mathscr{G}_1 \supseteq \mathscr{G}_2 \supseteq \cdots$, let $\mathscr{G}_\infty = \bigcap \mathscr{G}_n$, and let η be an integrable random variable. Establish the following analog

of Theorem 3: As $n \to \infty$,

$$\mathsf{E}(\eta \mid \mathscr{G}_n) \to \mathsf{E}(\eta \mid \mathscr{G}_\infty) \quad (\mathsf{P}\text{-a.s. and in the } L^1 \text{ sense}).$$

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 (Problem 2, Sect. 7, Chap. 2, Vol. 1) that

$$\mathsf{E}(\xi_1 \mid S_n, S_{n+1}, \ldots) = \mathsf{E}(\xi_1 \mid S_n) = \frac{S_n}{n} \quad (\mathsf{P}\text{-a.s.}),$$

deduce from Problem 1 a stronger form of the law of large numbers: As $n \to \infty$,

$$\frac{S_n}{n} \to m \quad (\mathsf{P}\text{-a.s. and in the } L^1 \text{ sense}).$$

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 let $\{\mathscr{F}_m\}_{m\geq 1}$ be a nondecreasing family of σ-algebras with $\mathscr{F}_\infty = \sigma(\bigcup \mathscr{F}_n)$. Then

$$\lim_{\substack{m\to\infty \\ n\to\infty}} \mathsf{E}(\xi_n \mid \mathscr{F}_m) = \mathsf{E}(\xi \mid \mathscr{F}_\infty) \quad (\mathsf{P}\text{-a.s.}).$$

4. Establish formula (12).
5. Let $\Omega = [0,1], \mathscr{F} = \mathscr{B}([0,1))$, let P denote Lebesgue measure, and let $f = f(x) \in L^1$. Set

$$f_n(x) = 2^n \int_{k2^{-n}}^{(k+1)2^{-n}} f(y)\,dy, \quad k2^{-n} \leq x < (k+1)2^{-n}.$$

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 set

$$f_n(x) = \sum_{j=1}^{2^n} 2^{-n} f(x + j2^{-n}).$$

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 $\inf_m \sup_{n\geq m} \mathsf{E}(X_n^+ \mid \mathscr{F}_m) < \infty$ (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.
9. Let $F = F(x), x \in R$, be a distribution function, and let $\alpha \in (0,1)$. Suppose that there exists $\theta \in R$ such that $F(\theta) = \alpha$. Let us construct the sequence $X_1, X_2, \ldots$ so that

$$X_{n+1} = X_n - n^{-1}(Y_n - \alpha),$$

where $Y_1, Y_2, \ldots$ are random variables such that

$$\mathsf{P}(Y_n = y \,|\, X_1, \ldots, X_n; Y_1, \ldots, Y_{n-1}) = \begin{cases} F(X_n) & \text{if } y = 1, \\ 1 - F(X_n) & \text{if } y = 0 \end{cases}$$

(the Robbins–Monro procedure). Prove the following result of the stochastic approximation theory: $\mathsf{E}\,|X_n - \theta|^2 \to 0$, $n \to \infty$.

10. Let $X = (X_n, \mathscr{F}_n)_{n\geq 1}$ be a submartingale such that $\mathsf{E}(X_\tau I(\tau < \infty)) \neq \infty$ for any stopping time τ. Show that with probability 1 there exists the limit $\lim_n X_n$.
11. Let $X = (X_n, \mathscr{F}_n)_{n\geq 1}$ be a martingale and $\mathscr{F}_\infty = \sigma\left(\bigcup_{n=1}^\infty \mathscr{F}_n\right)$. Prove that if the sequence $(X_n)_{n\geq 1}$ is uniformly integrable, then the limit $X_\infty = \lim_n X_n$ exists (P-a.s.) and the "closed" sequence $\overline{X} = (X_n, \mathscr{F}_n)_{1\leq n\leq \infty}$ is a martingale.
12. Assume that $X = (X_n, \mathscr{F}_n)_{n\geq 1}$ is a submartingale, and let $\mathscr{F}_\infty = \sigma\left(\bigcup_{n=1}^\infty \mathscr{F}_n\right)$. Prove that if $(X_n^+)_{n\geq 1}$ is uniformly integrable, then the limit $X_\infty = \lim_n X_n$ exists (P-a.s.) and the "closed" sequence $\overline{X} = (X_n, \mathscr{F}_n)_{1\leq n\leq \infty}$ is a submartingale.

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 $\mathsf{P}(I_A \leq I_B) = 1$. We will also write $\{X_n \nrightarrow\}$ for $\Omega \setminus \{X_n \to\}$ and $A = B$ a.s. if $\mathsf{P}(A \triangle B) = 0$.

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 Sect. 4, we have

$$\{X_n \to\} = \Omega \quad (\mathsf{P}\text{-a.s.}), \quad \text{i.e. } \mathsf{P}\{X_n \nrightarrow\} = 0.$$

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 $\tau_a = \min\{n \geq 1 \colon X_n > a\}$ with $\tau_a = \infty$ if $\{\cdot\} = \varnothing$.

Definition. A stochastic sequence $X = (X_n, \mathscr{F}_n)$ belongs to *class* $\mathbb{C}^+$ ($X \in \mathbb{C}^+$) 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 $X \in \mathbb{C}^+$ if

$$\mathsf{E} \sup_n |\Delta X_n| < \infty \tag{2}$$

or, all the more so, if

$$|\Delta X_n| \leq C < \infty \quad (\mathsf{P}\text{-a.s.}) \tag{3}$$

for all $n \geq 1$.

Theorem 1. *If the submartingale $X \in \mathbb{C}^+$, 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 opposite inclusion, 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 from Sect. 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* (P-*a.s.*)

$$\{X_n \to\} \cup \{\liminf X_n = -\infty,\ \limsup X_n = +\infty\} = \Omega. \tag{6}$$

In fact, if we apply Theorem 1 to X and to $-X$, we find that (P-a.s.)

$$\begin{aligned} \{\limsup X_n < \infty\} &= \{\sup X_n < \infty\} = \{X_n \to\}, \\ \{\liminf X_n > -\infty\} &= \{\inf X_n > -\infty\} = \{X_n \to\}. \end{aligned}$$

Therefore (P-a.s.)

$$\{\limsup X_n < \infty\} \cup \{\liminf X_n > -\infty\} = \{X_n \to\},$$

which establishes (6).

Statement (6) means that, provided that $\mathsf{E}\sup|\Delta X_n| < \infty$, either almost all trajectories of the martingale X have finite limits or all behave very badly, in the sense that $\limsup X_n = +\infty$ and $\liminf X_n = -\infty$.

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 from Sect. 2, Chap. 4, 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 $\mathscr{F}_n = \sigma\{\xi_1, \ldots, \xi_n\}$ 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:

$$\{\langle X\rangle_\infty < \infty\} = \{X_n \to\} = \Omega \quad (\mathsf{P}\text{-a.s.}),$$

where $\langle X\rangle_\infty = \lim_n \langle X\rangle_n$.

The following propositions extend 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 $X \in \mathbb{C}^+$, 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 $X \in \mathbb{C}^+$, 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

$$\sigma_a = \min\{n \geq 1 \colon A_{n+1} > a\}, \quad a > 0,$$

taking $\sigma_a = +\infty$ if $\{\cdot\} = \varnothing$. Then $A_{\sigma_a} \leq a$, and, by Corollary 1 to Theorem 1, Sect. 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 in Sect. 4 that

$$\{A_\infty \leq a\} = \{\sigma_a = \infty\} \subseteq \{X_n \to\} \quad (\mathsf{P}\text{-a.s.}).$$

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.*)

$$\left\{ \sum_{n=1}^{\infty} \mathsf{E}(\xi_n \mid \mathscr{F}_{n-1}) < \infty \right\} \subseteq \{X_n \to\}, \tag{10}$$

and if, in addition, $\mathsf{E} \sup_n \xi_n < \infty$, *then* (P-*a.s.*)

$$\left\{ \sum_{n=1}^{\infty} \mathsf{E}(\xi_n \mid \mathscr{F}_{n-1}) < \infty \right\} = \{X_n \to\}. \tag{11}$$

Corollary 2 (Borel–Cantelli–Lévy Lemma). *If the events* $B_n \in \mathscr{F}_n$, *then, if we set* $\xi_n = I_{B_n}$ *in* (11), *we find that* (P-*a.s.*)

$$\left\{ \sum_{n=1}^{\infty} \mathsf{P}(B_n \mid \mathscr{F}_{n-1}) < \infty \right\} = \left\{ \sum_{n=1}^{\infty} I_{B_n} < \infty \right\}. \tag{12}$$

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

$$\langle M \rangle_\infty = \sum_{n=1}^{\infty} \mathsf{E}((\Delta M_n)^2 \mid \mathscr{F}_{n-1}) \tag{15}$$

with $M_0 = 0, \mathscr{F}_0 = \{\varnothing, \Omega\}$.

PROOF. Consider the two submartingales $M^2 = (M_n^2, \mathscr{F}_n)$ and $(M+1)^2 = ((M_n + 1)^2, \mathscr{F}_n)$. Let their Doob decompositions be

$$M_n^2 = m_n' + A_n', \quad (M_n + 1)^2 = m_n'' + A_n''.$$

Then A_n' and A_n'' are the same, since

$$A_n'' = \sum_{k=1}^{n} \mathsf{E}(\Delta(M_k + 1)^2 \mid \mathscr{F}_{k-1}) = \sum_{k=1}^{n} \mathsf{E}(\Delta M_k^2 \mid \mathscr{F}_{k-1}) = A_n'$$

because the linear term in $\mathsf{E}(\Delta(M_k+1)^2 \mid \mathscr{F}_{k-1})$ vanishes. 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), Eq. (14) will be established if we show that the condition $\mathsf{E}\sup|\Delta M_n|^2 < \infty$ guarantees that M^2 belongs to $\mathbb{C}^+$.

Let $\tau_a = \min\{n \geq 1\colon M_n^2 > a\},\ a > 0$. Then, on the set $\{\tau_a < \infty\}$,

$$\begin{aligned}|\Delta M^2_{\tau_a}| = |M^2_{\tau_a} - M^2_{\tau_a-1}| &\leq |M_{\tau_a} - M_{\tau_a-1}|^2 \\ &+2|M_{\tau_a-1}| \cdot |M_{\tau_a} - M_{\tau_a-1}| \leq (\Delta M_{\tau_a})^2 + 2a^{1/2}|\Delta M_{\tau_a}|,\end{aligned}$$

whence

$$\begin{aligned}\mathsf{E}\,|\Delta M^2_{\tau_a}|\, I\{\tau_a < \infty\} &\leq \mathsf{E}(\Delta M_{\tau_a})^2\, I\{\tau_a < \infty\} + 2a^{1/2}\sqrt{\mathsf{E}(\Delta M_{\tau_a})^2\, I\{\tau_a < \infty\}} \\ &\leq \mathsf{E}\sup|\Delta M_n|^2 + 2a^{1/2}\sqrt{\mathsf{E}\sup|\Delta M_n|^2} < \infty.\end{aligned}$$

This completes the proof of the theorem.

□

As an illustration of this theorem, we present the following result, which can be considered as a kind of the *strong law of large numbers* for square-integrable martingales (cf. Theorem 2 in Sect. 3, Chap. 4).

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.*)

$$\sum_{i=1}^{\infty} \frac{\mathsf{E}[(\Delta M_i)^2 \mid \mathscr{F}_{i-1}]}{A_i^2} < \infty, \tag{16}$$

then

$$M_n/A_n \to 0, \quad n \to \infty, \tag{17}$$

with probability 1.

In particular, if $\langle M\rangle = (M_n, \mathscr{F}_{n-1})$ 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

$$\frac{M_n}{\langle M\rangle_n} \to 0, \quad n \to \infty. \tag{18}$$

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

$$\langle m\rangle_n = \sum_{i=1}^{n} \frac{\mathsf{E}[(\Delta M_i)^2 \mid \mathscr{F}_{i-1}]}{A_i^2}. \tag{19}$$

Since

$$\frac{M_n}{A_n} = \frac{\sum_{k=1}^{n} A_k \Delta m_k}{A_n},$$

we have, by Kronecker's lemma (Sect. 3, Chap. 4), $M_n/A_n \to 0$ (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 (Problem 6).

This completes the proof of the theorem.

□

EXAMPLE. Consider a sequence $\xi_1, \xi_2, \ldots$ of independent random variables with $\mathsf{E}\,\xi_i = 0$, $\operatorname{Var} \xi_i = V_i > 0$, 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

$$\hat{\theta}_n = \frac{\sum_{k=0}^{n-1} (X_k X_{k+1})/V_{k+1}}{\sum_{k=0}^{n-1} X_k^2/V_{k+1}}, \tag{22}$$

taking this to be 0 if the denominator is 0. (The quantity $\hat{\theta}_n$ is the *least-squares* estimator of θ.)

It is clear from (21) and (22) that

$$\hat{\theta} = \theta + \frac{M_n}{A_n},$$

where

$$M_n = \sum_{k=0}^{n-1} \frac{X_k \xi_{k+1}}{V_{k+1}}, \quad A_n = \langle M \rangle_n = \sum_{k=0}^{n-1} \frac{X_k^2}{V_{k+1}}.$$

Therefore, if the true value of the unknown parameter is θ, then

$$\mathsf{P}(\hat{\theta}_n \to \theta) = 1 \tag{23}$$

if and only if (P-a.s.)

$$\frac{M_n}{A_n} \to 0, \quad n \to \infty. \tag{24}$$

Let us show that the conditions

$$\sup_n \frac{V_{n+1}}{V_n} < \infty, \quad \sum_{n=1}^{\infty} \mathsf{E}\left(\frac{\xi_n^2}{V_n} \wedge 1\right) = \infty \tag{25}$$

are sufficient for (24), and therefore sufficient for (23). We have

$$\sum_{n=1}^{\infty}\left(\frac{\xi_n^2}{V_n}\wedge 1\right) \le \sum_{n=1}^{\infty}\frac{\xi_n^2}{V_n} = \sum_{n=1}^{\infty}\frac{(X_n-\theta X_{n-1})^2}{V_n}$$
$$\le 2\left[\sum_{n=1}^{\infty}\frac{X_n^2}{V_n}+\theta^2\sum_{n=1}^{\infty}\frac{X_{n-1}^2}{V_n}\right] \le 2\left[\sup\frac{V_{n+1}}{V_n}+\theta^2\right]\langle M\rangle_\infty,$$

which follows because

$$\sum_{n=1}^{\infty}\frac{X_n^2}{V_n} = \sum_{n=1}^{\infty}\frac{X_n^2}{V_{n+1}}\frac{V_{n+1}}{V_n} \le \sup\frac{V_{n+1}}{V_n}\sum_{n=1}^{\infty}\frac{X_n^2}{V_{n+1}} = \sup\frac{V_{n+1}}{V_n}\langle M\rangle_\infty,$$

where $\langle M\rangle_n = \sum_{k=0}^{n-1}\frac{X_k^2}{V_{k+1}}$ by definition.

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, Sect. 2, Chap. 4) 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 $\mathsf{P}\{\langle M\rangle_\infty=\infty\}=1$, hence (24) follows directly from Theorem 4.

Estimators $\hat{\theta}_n$, $n\ge 1$, with property (23) are said to be *strongly consistent*; compare the notion of consistency in Sect. 4, Chap. 1, Vol. 1.

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),

$$\{\langle m\rangle_\infty + A_\infty < \infty\} \subseteq \{X_n \to\} \quad (\mathsf{P}\text{-a.s.}) \tag{30}$$

and, according to (8),

$$\{X_n \to\} \subseteq \{A_\infty < \infty\} \quad (\mathsf{P}\text{-a.s.}).$$

Therefore, by (14) and (30),

$$\begin{aligned}\{X_n \to\} &= \{X_n \to\} \cap \{A_\infty < \infty\} = \{X_n \to\} \cap \{A_\infty < \infty\} \cap \{m_n \to\} \\ &= \{X_n \to\} \cap \{A_\infty < \infty\} \cap \{\langle m\rangle_\infty < \infty\} \\ &= \{X_n \to\} \cap \{A_\infty + \langle m\rangle_\infty < \infty\} = \{A_\infty + \langle m\rangle_\infty < \infty\}.\end{aligned}$$

Finally, the equivalence of (26) and (27) follows because, by (29),

$$\langle m\rangle_n = \sum \{\mathsf{E}[(\Delta X_k)^2 \mid \mathscr{F}_{k-1}] - [\mathsf{E}(\Delta X_k \mid \mathscr{F}_{k-1})]^2\},$$

and the convergence of the series $\sum_{k=1}^\infty \mathsf{E}(\Delta X_k \mid \mathscr{F}_{k-1})$ of nonnegative terms implies the convergence of $\sum_{k=1}^\infty [\mathsf{E}(\Delta X_k \mid \mathscr{F}_{k-1})]^2$. This completes the proof.

□

4. Kolmogorov's three-series theorem (Theorem 3, Sect. 2, Chap. 4) gives a necessary and sufficient condition for the convergence, with probability 1, of a series $\sum \xi_n$ of independent random variables. The following theorem, whose proof is based on Theorems 2 and 3, describes sets of convergence of $\sum \xi_n$ without the assumption that the random variables $\xi_1, \xi_2, \ldots$ are *independent*.

Theorem 6. *Let $\xi = (\xi_n, \mathscr{F}_n)$, $n \ge 1$, 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*

$$\sum \mathsf{P}(|\xi_n| \ge c \mid \mathscr{F}_{n-1}), \quad \sum \mathsf{E}(\xi_n^c \mid \mathscr{F}_{n-1}), \quad \sum \mathrm{Var}(\xi_n^c \mid \mathscr{F}_{n-1})$$

converge, where $\xi_n^c = \xi_n I(|\xi_n| \le c)$.

PROOF. Let $X_n = \sum_{k=1}^n \xi_k$. Since (on the set A) the series $\sum \mathsf{P}(|\xi_n| \ge c \mid \mathscr{F}_{n-1})$ converges, by Corollary 2 of Theorem 2, and by the convergence of the series $\sum \mathsf{E}(\xi_n^c \mid \mathscr{F}_{n-1})$, we have

$$\begin{aligned}A \cap \{X_n \to\} &= A \cap \left\{\sum_{k=1}^n \xi_k I(|\xi_k| \le c) \to\right\} \\ &= A \cap \left\{\sum_{k=1}^n [\xi_k I(|\xi_k| \le c) - \mathsf{E}(\xi_k I(|\xi_k| \le c) \mid \mathscr{F}_{k-1})] \to\right\}. \end{aligned} \tag{31}$$

Let $\eta_k = \xi_k^c - \mathsf{E}(\xi_k^c \mid \mathscr{F}_{k-1})$, and let $Y_n = \sum_{k=1}^n \eta_k$. Then $Y = (Y_n, \mathscr{F}_n)$ is a square-integrable martingale with $|\eta_k| \le 2c$. By Theorem 5 we have

$$A \subseteq \Big\{ \sum \mathrm{Var}(\xi_n^c \mid \mathscr{F}_{n-1}) < \infty \Big\} = \{\langle Y \rangle_\infty < \infty\} = \{Y_n \to\}. \tag{32}$$

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 $\mathbb{C}^+$.
2. Show that Theorems 1 and 2 remain valid for generalized submartingales.
3. Show that generalized submartingales satisfy (P-a.s.) the inclusion

$$\Big\{ \inf_m \sup_{n \ge m} \mathsf{E}(X_n^+ \mid \mathscr{F}_m) < \infty \Big\} \subseteq \{X_n \to\}.$$

4. Show that the corollary of Theorem 1 remains valid for generalized martingales.
5. Show that every generalized submartingale of class $\mathbb{C}^+$ is a local submartingale.
6. Let $a_n > 0,\ n \ge 1$, 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 on a Measurable Space with Filtration

1. Let $(\Omega, \mathscr{F})$ be a measurable space on which there is defined a family $(\mathscr{F}_n)_{n \ge 1}$ of σ-algebras such that $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \cdots \subseteq \mathscr{F}$ and

$$\mathscr{F} = \sigma\Big(\bigcup_{n=1}^\infty \mathscr{F}_n \Big). \tag{1}$$

Let us suppose that two probability measures P and $\tilde{\mathsf{P}}$ are given on $(\Omega, \mathscr{F})$. Let us write

$$\mathsf{P}_n = \mathsf{P} \mid \mathscr{F}_n, \quad \tilde{\mathsf{P}}_n = \tilde{\mathsf{P}} \mid \mathscr{F}_n$$

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

$$\mathsf{P}_n(B) = \mathsf{P}(B), \quad \tilde{\mathsf{P}}_n(B) = \tilde{\mathsf{P}}(B).$$

Recall that 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 $\mathsf{P}(\overline{A}) = 1$ (notation, $\tilde{\mathsf{P}} \perp \mathsf{P}$).

Definition 1. 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.

Recall that the problems of absolute continuity and singularity were considered in Sect. 9, Chap. 3, Vol. 1, for *arbitrary* probability measures. It was shown that the corresponding tests could be stated in terms of the Hellinger integrals (Theorems 2 and 3 therein). The results about absolute continuity and singularity for locally absolutely continuous measures to be stated below could be obtained using those tests. This approach is revealed in the monographs [34, 43]. Here we prefer another presentation, which enables us to better illustrate the possibilities of using the results on the sets of convergence of submartingales obtained in Sect. 5. (Note that throughout this section we assume the property of local absolute continuity. This is done only to simplify the presentation. The reader is referred to [34, 43] for the general case.)

Let us then suppose that $\tilde{\mathsf{P}} \overset{\text{loc}}{\ll} \mathsf{P}$. Denote by

$$z_n = \frac{d\tilde{\mathsf{P}}_n}{d\,\mathsf{P}_n}$$

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*.

The following theorem is the key to problems on absolute continuity and singularity.

Theorem 1. *Let* $\tilde{\mathsf{P}} \overset{\text{loc}}{\ll} \mathsf{P}$.

(a) *Then with* $\frac{1}{2}(\tilde{\mathsf{P}} + \mathsf{P})$*-probability* 1 *there exists the limit* $\lim_n z_n$*, to be denoted by* z_∞*, such that*

$$\mathsf{P}(z_\infty = \infty) = 0.$$

(b) *The Lebesgue decomposition*

$$\tilde{\mathsf{P}}(A) = \int_A z_\infty \, d\mathsf{P} + \tilde{\mathsf{P}}(A \cap \{z_\infty = \infty\}), \quad A \in \mathscr{F}, \tag{3}$$

holds, and the measures $\tilde{\mathsf{P}}(A \cap \{z_\infty = \infty\})$ *and* $\mathsf{P}(A)$, $A \in \mathscr{F}$, *are singular.*

PROOF. Let us notice first that, according to the classical *Lebesgue decomposition* (see (29) in Sect. 9, Chap. 3, Vol. 1) of an arbitrary probability measure P with respect to a probability measure $\tilde{\mathsf{P}}$, the following representation holds:

$$\widetilde{\mathsf{P}}(A) = \int_A \frac{\tilde{\mathfrak{z}}}{\mathfrak{z}} \, d\mathsf{P} + \widetilde{\mathsf{P}}(A \cap \{\mathfrak{z} = 0\}), \quad A \in \mathscr{F}, \tag{4}$$

where

$$\mathfrak{z} = \frac{d\mathsf{P}}{d\mathsf{Q}}, \quad \tilde{\mathfrak{z}} = \frac{d\widetilde{\mathsf{P}}}{d\mathsf{Q}}$$

and the measure Q can be taken, for example, to be $\mathsf{Q} = \frac{1}{2}(\mathsf{P} + \widetilde{\mathsf{P}})$. Conclusion (3) can be thought of as a specialization of decomposition (4) under the assumption that $\tilde{\mathsf{P}} \overset{\text{loc}}{\ll} \mathsf{P}$, i.e., $\tilde{\mathsf{P}}_n \ll \mathsf{P}_n$.

Let

$$\mathfrak{z}_n = \frac{d\mathsf{P}_n}{d\mathsf{Q}_n}, \quad \tilde{\mathfrak{z}}_n = \frac{d\widetilde{\mathsf{P}}_n}{d\mathsf{Q}_n}, \quad \mathsf{Q}_n = \frac{1}{2}(\mathsf{P}_n + \widetilde{\mathsf{P}}_n).$$

The sequences $(\mathfrak{z}_n, \mathscr{F}_n)$ and $(\tilde{\mathfrak{z}}_n, \mathscr{F}_n)$ are martingales with respect to Q such that $0 \le \mathfrak{z}_n \le 2$, $0 \le \tilde{\mathfrak{z}}_n \le 2$. Therefore, by Theorem 2, Sect. 4, there exist the limits

$$\mathfrak{z}_\infty \equiv \lim_n \mathfrak{z}_n, \quad \tilde{\mathfrak{z}}_\infty \equiv \lim_n \tilde{\mathfrak{z}}_n \tag{5}$$

both Q-a.s. and in the sense of convergence in $L^1(\Omega, \mathscr{F}, \mathsf{Q})$.

The convergence in $L^1(\Omega, \mathscr{F}, \mathsf{Q})$ implies, in particular, that for any $A \in \mathscr{F}_m$

$$\int_A \tilde{\mathfrak{z}}_\infty \, d\mathsf{Q} = \lim_{n \uparrow \infty} \int_A \tilde{\mathfrak{z}}_n \, d\mathsf{Q} = \int_A \tilde{\mathfrak{z}}_m \, d\mathsf{Q} = \widetilde{\mathsf{P}}_m(A) = \widetilde{\mathsf{P}}(A).$$

Then we obtain by Carathéodory's theorem (Sect. 3, Chap. 2, Vol. 1) that for any $A \in \mathscr{F} = \sigma(\bigcup_n \mathscr{F}_n)$

$$\int_A \tilde{\mathfrak{z}}_\infty \, d\mathsf{Q} = \widetilde{\mathsf{P}}(A),$$

i.e., $d\widetilde{\mathsf{P}}/d\,\mathsf{Q} = \tilde{\mathfrak{z}}_\infty$, and, similarly,

$$\int_A \mathfrak{z}_\infty \, d\,\mathsf{Q} = \mathsf{P}(A),$$

i.e., $d\,\mathsf{P}/d\,\mathsf{Q} = \mathfrak{z}_\infty$.

Thus, we have established the result that was to be expected: If the measures P and Q are defined on $\mathscr{F} = \sigma(\bigcup \mathscr{F}_n)$ and $\mathsf{P}_n, \mathsf{Q}_n$ are the restrictions of these measures to $\mathscr{F}_n$, then

$$\lim_n \frac{d\,\mathsf{P}_n}{d\,\mathsf{Q}_n} = \frac{d\,\mathsf{P}}{d\,\mathsf{Q}}$$

(Q-a.s. and in $L^1(\Omega, \mathscr{F}, \mathsf{Q})$). Similarly,

$$\lim_n \frac{d\widetilde{\mathsf{P}}_n}{d\,\mathsf{Q}_n} = \frac{d\widetilde{\mathsf{P}}}{d\,\mathsf{Q}}.$$

In the special case under consideration, where $\widetilde{\mathsf{P}}_n \ll \mathsf{P}_n$, $n \geq 1$, it is not hard to show that (Q-a.s.)

$$z_n = \frac{\tilde{\mathfrak{z}}_n}{\mathfrak{z}_n}, \tag{6}$$

and $\mathsf{Q}\{\mathfrak{z}_n = 0, \tilde{\mathfrak{z}}_n = 0\} \leq \frac{1}{2}[\mathsf{P}\{\mathfrak{z}_n = 0\} + \widetilde{\mathsf{P}}\{\tilde{\mathfrak{z}}_n = 0\}] = 0$, so that (6) Q-a.s. does not involve an indeterminacy of the form $\frac{0}{0}$.

The expression of the form $\frac{2}{0}$, as usual, is set at $+\infty$. It is useful to note that, since $(\mathfrak{z}_n, \mathscr{F}_n)$ is a nonnegative martingale, relation (5) of Sect. 2 implies that if $\mathfrak{z}_\tau = 0$, then $\mathfrak{z}_n = 0$ for all $n \geq \tau$ (Q-a.s.). Of course, the same holds also for $(\tilde{\mathfrak{z}}_n, \mathscr{F}_n)$. Therefore the points 0 and $+\infty$ are "absorbing states" for the sequence $(z_n)_{n\geq 1}$.

It follows from (5) and (6) that the limit

$$z_\infty \equiv \lim_n z_n = \frac{\lim_n \tilde{\mathfrak{z}}_n}{\lim_n \mathfrak{z}_n} = \frac{\tilde{\mathfrak{z}}_\infty}{\mathfrak{z}_\infty} \tag{7}$$

exists Q-a.s.

Since $\mathsf{P}\{\mathfrak{z}_\infty = 0\} = \int_{\{\mathfrak{z}_\infty = 0\}} \mathfrak{z}_\infty \, d\,\mathsf{Q} = 0$, we have $\mathsf{P}\{z_\infty = \infty\} = 0$, which proves conclusion (a).

For the proof of (3) we use the general decomposition (4). In our setup, by what has been proved, we have $\mathfrak{z} = \frac{d\,\mathsf{P}}{d\,\mathsf{Q}} = \mathfrak{z}_\infty$, $\tilde{\mathfrak{z}} = \frac{d\widetilde{\mathsf{P}}}{d\,\mathsf{Q}} = \tilde{\mathfrak{z}}_\infty$ (Q-a.s.), hence (4) yields

$$\widetilde{\mathsf{P}}(A) = \int_A \frac{\tilde{\mathfrak{z}}_\infty}{\mathfrak{z}_\infty} \, d\,\mathsf{P} + \widetilde{\mathsf{P}}(A \cap \{\mathfrak{z}_\infty = 0\}).$$

In view of (7) and the fact that $\widetilde{\mathsf{P}}\{\tilde{\mathfrak{z}}_\infty = 0\} = 0$, we obtain the required decomposition (3). Note that due to $\mathsf{P}\{z_\infty < \infty\} = 1$, the measures

$$\mathsf{P}(A) \equiv \mathsf{P}(A \cap \{z_\infty < \infty\}) \quad \text{and} \quad \widetilde{\mathsf{P}}(A \cap \{z_\infty = \infty\}), \quad A \in \mathscr{F},$$

are *singular*.

□

The Lebesgue decomposition (3) implies the following useful tests for absolute continuity or singularity of 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. Setting $A = \Omega$ 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 in terms of either P (verify the equation $\mathsf{E} z_\infty = 1$ or $\mathsf{E} z_\infty = 0$) or $\tilde{\mathsf{P}}$ (verify that $\tilde{\mathsf{P}}(z_\infty < \infty) = 1$ or that $\tilde{\mathsf{P}}(z_\infty = \infty) = 1$).

By Theorem 5 in Sect. 6, Chap. 2, Vol. 1, 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

$$\sup_n \mathsf{E}[z_n \log^+ z_n] < \infty \tag{12}$$

or, if

$$\sup_n \mathsf{E} z_n^{1+\varepsilon} < \infty, \quad \varepsilon > 0, \tag{13}$$

then, by Lemma 3 in Sect. 6, Chap. 2, Vol. 1, 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 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 $\xi = (\xi_1, \xi_2, \ldots)$ and $\tilde{\xi} = (\tilde{\xi}_1, \tilde{\xi}_2, \ldots)$ be sequences of independent random variables defined on a probability space $(\Omega, \mathscr{F}, \mathsf{P})$.

Let $(R^\infty, \mathscr{B}_\infty)$ be the measurable space of sequences $x = (x_1, x_2, \ldots)$ of real numbers with $\mathscr{B}_\infty = \mathscr{B}(R^\infty)$, and let $\mathscr{B}_n = \sigma\{x_1, \ldots, x_n\}$.

Let P and $\tilde{P}$ be the probability distributions on $(R^\infty, \mathscr{B}_\infty)$ for ξ and $\tilde{\xi}$, respectively, i.e.,

$$P(B) = \mathsf{P}\{\xi \in B\}, \quad \tilde{P}(B) = \mathsf{P}\{\xi \in B\}, \qquad B \in \mathscr{B}_\infty.$$

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

$$P_{\xi_n}(A) = \mathsf{P}(\xi_n \in A), \quad P_{\tilde{\xi}_n}(A) = \mathsf{P}(\tilde{\xi}_n \in A), \quad A \in \mathscr{B}(R^1).$$

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*

$$P_{\tilde{\xi}_n} \ll P_{\xi_n}, \quad n \geq 1. \tag{14}$$

Then either $\tilde{\mathsf{P}} \ll \mathsf{P}$ or $\tilde{\mathsf{P}} \perp \mathsf{P}$.

PROOF. Condition (14) is evidently equivalent to $\tilde{P}_n \ll P_n$, $n \geq 1$, i.e., $\tilde{P} \overset{\text{loc}}{\ll} P$. 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,

$$\{x\colon z_\infty < \infty\} = \{x\colon \log z_\infty < \infty\} = \left\{x\colon \sum_{i=1}^{\infty} \log q_i(x_i) < \infty\right\}.$$

The event $\{x\colon \sum_{i=1}^{\infty} \log\, q_i(x_i) < \infty\}$ is a tail event. Therefore, by the Kolmogorov zero–one law (Theorem 1, Sect. 1, Chap. 4) the probability $\tilde{P}\{x\colon z_\infty < \infty\}$ has only two values (0 or 1), and therefore, by Theorem 2, either $\tilde{P} \perp P$ or $\tilde{P} \ll P$.

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 $\tilde{\mathsf{P}} \overset{loc}{\ll} \mathsf{P}$, and let*

$$\alpha_n = z_n z_{n-1}^{\oplus}, \quad n \geq 1,$$

with $z_0 = 1$. Then (with $\mathscr{F}_0 = \{\varnothing, \Omega\}$)

$$\tilde{\mathsf{P}} \ll \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\Big\{\sum_{n=1}^{\infty}[1-\mathsf{E}(\sqrt{\alpha_n}\,|\,\mathscr{F}_{n-1})]<\infty\Big\}=1, \tag{16}$$

$$\tilde{\mathsf{P}} \perp \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\Big\{\sum_{n=1}^{\infty}[1-\mathsf{E}(\sqrt{\alpha_n}\,|\,\mathscr{F}_{n-1})]=\infty\Big\}=1. \tag{17}$$

PROOF. Since

$$\tilde{\mathsf{P}}_n\{z_n=0\}=\int_{\{z_n=0\}} z_n\,d\,\mathsf{P}=0,$$

we have (P-a.s.)

$$z_n=\prod_{k=1}^{n}\alpha_k=\exp\Big\{\sum_{k=1}^{n}\log\alpha_k\Big\}. \tag{18}$$

Setting $A=\{z_\infty=0\}$ in (3), we find that $\tilde{\mathsf{P}}\{z_\infty=0\}=0$. Therefore, by (18), we have (P-a.e.)

$$\begin{aligned}\{z_\infty<\infty\}&=\{0<z_\infty<\infty\}=\{0<\lim z_n<\infty\}\\&=\Big\{-\infty<\lim\sum_{k=1}^{n}\log\alpha_k<\infty\Big\}.\end{aligned} \tag{19}$$

Let us introduce the function

$$u(x)=\begin{cases}x, & |x|\le 1,\\ \operatorname{sign}x, & |x|>1.\end{cases}$$

Then

$$\Big\{-\infty<\lim\sum_{k=1}^{n}\log\alpha_k<\infty\Big\}=\Big\{-\infty<\lim\sum_{k=1}^{n}u(\log\alpha_k)<\infty\Big\}. \tag{20}$$

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 for "recalculation of conditional expectations" (see (44) in Sect. 7, Chap.2, Vol. 1):

$$\tilde{\mathsf{E}}(\eta\,|\,\mathscr{F}_{n-1})=\mathsf{E}(\alpha_n\eta\,|\,\mathscr{F}_{n-1})\quad(\tilde{\mathsf{P}}\text{-a.s.}). \tag{23}$$

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),

$$\tilde{\mathsf{E}}[u(\log \alpha_n) \,|\, \mathscr{F}_{n-1}] = \mathsf{E}[\alpha_n u(\log \alpha_n) \,|\, \mathscr{F}_{n-1}] \quad (\tilde{\mathsf{P}}\text{-a.s.}).$$

Since $xu(\log x) \geq x - 1$ for $x \geq 0$, we have, by (24),

$$\tilde{\mathsf{E}}[u(\log \alpha_n) \,|\, \mathscr{F}_{n-1}] \geq 0 \quad (\tilde{\mathsf{P}}\text{-a.s.}).$$

It follows that the stochastic sequence $X = (X_n, \mathscr{F}_n)$ with

$$X_n = \sum_{k=1}^{n} u(\log \alpha_k),$$

is a submartingale with respect to $\tilde{\mathsf{P}}$ and $|\Delta X_n| = |u(\log \alpha_n)| \leq 1$.

Then, by Theorem 5 in Sect. 5, we have ($\tilde{\mathsf{P}}$-a.e.)

$$\left\{ -\infty < \lim \sum_{k=1}^{n} u(\log \alpha_k) < \infty \right\} = \left\{ \sum_{k=1}^{\infty} \tilde{\mathsf{E}}[u(\log \alpha_k) + u^2(\log \alpha_k) \,|\, \mathscr{F}_{k-1}] < \infty \right\}. \quad (25)$$

Hence we find, by combining (19), (20), (22), and (25), that (P-a.s.)

$$\begin{aligned} \{z_\infty < \infty\} &= \left\{ \sum_{k=1}^{\infty} \tilde{\mathsf{E}}[u(\log \alpha_k) + u^2(\log \alpha_k) \,|\, \mathscr{F}_{k-1}] < \infty \right\} \\ &= \left\{ \sum_{k=1}^{\infty} \mathsf{E}[\alpha_k u(\log \alpha_k) + \alpha_k u^2(\log \alpha_k) \,|\, \mathscr{F}_{k-1}] < \infty \right\} \end{aligned}$$

and consequently, by Theorem 2,

$$\tilde{\mathsf{P}} \ll \mathsf{P} \Leftrightarrow \quad \tilde{\mathsf{P}}\left\{ \sum_{k=1}^{\infty} \mathsf{E}[\alpha_k u(\log \alpha_k) + \alpha_k u^2(\log \alpha_k) \,|\, \mathscr{F}_{k-1}] < \infty \right\} = 1, \quad (26)$$

$$\tilde{\mathsf{P}} \perp \mathsf{P} \Leftrightarrow \quad \tilde{\mathsf{P}}\left\{ \sum_{k=1}^{\infty} \mathsf{E}[\alpha_k u(\log \alpha_k) + \alpha_k u^2(\log \alpha_k) \,|\, \mathscr{F}_{k-1}] = \infty \right\} = 1. \quad (27)$$

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

$$A(1 - \sqrt{x})^2 \leq xu(\log x) + xu^2(\log x) + 1 - x \leq B(1 - \sqrt{x})^2. \quad (28)$$

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 following 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*

$$\tilde{\mathsf{P}}\left\{\sum_{n=1}^{\infty} \mathsf{E}(\alpha_n \log \alpha_n \,|\, \mathscr{F}_{n-1}) < \infty\right\} = 1 \Rightarrow \tilde{\mathsf{P}} \ll \mathsf{P}\,.$$

For the proof, it is enough to notice that

$$x \log x + \tfrac{3}{2}(1 - x) \geq 1 - x^{1/2}, \tag{29}$$

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 $\sum |\log \mathsf{E}(\sqrt{\alpha_n} \,|\, \mathscr{F}_{n-1})|$, conclusions (16) and (17) of Theorem 4 can be put in the form*

$$\tilde{\mathsf{P}} \ll \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\left\{\sum_{n=1}^{\infty} |\log \mathsf{E}(\sqrt{\alpha_n} \,|\, \mathscr{F}_{n-1})| < \infty\right\} = 1, \tag{30}$$

$$\tilde{\mathsf{P}} \perp \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\left\{\sum_{n=1}^{\infty} |\log \mathsf{E}(\sqrt{\alpha_n} \,|\, \mathscr{F}_{n-1})| = \infty\right\} = 1. \tag{31}$$

Corollary 4. *Let there exist constants A and B such that $0 \leq A < 1$, $B \geq 0$, and*

$$\mathsf{P}\{1 - A \leq \alpha_n \leq 1 + B\} = 1, \quad n \geq 1.$$

Then, if $\tilde{\mathsf{P}} \overset{\text{loc}}{\ll} \mathsf{P}$, *we have*

$$\tilde{\mathsf{P}} \ll \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\left\{\sum_{n=1}^{\infty} \mathsf{E}[(1-\alpha_n)^2 \mid \mathscr{F}_{n-1}] < \infty\right\} = 1,$$

$$\tilde{\mathsf{P}} \perp \mathsf{P} \Leftrightarrow \tilde{\mathsf{P}}\left\{\sum_{n=1}^{\infty} \mathsf{E}[(1-\alpha_n)^2 \mid \mathscr{F}_{n-1}] = \infty\right\} = 1.$$

For the proof it is enough to notice that if $x \in [1-A,\ 1+B]$, where $0 \le A < 1$, $B \ge 0$, there are constants c and C $(0 < c < C < \infty)$ such that

$$c(1-x)^2 \le (1-\sqrt{x})^2 \le C(1-x)^2. \tag{32}$$

4. Using 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 and $\tilde{P}_n \sim P_n,\ n \ge 1$. Let us show that, for such sequences, the "predictable" test given above implies the *Hájek–Feldman dichotomy*: either $\tilde{P} \sim P$ or $\tilde{P} \perp P$.

By the theorem on normal correlation (Theorem 2 of Sect. 13, Chap. 2, Vol. 1) the conditional expectations $E(x_n \mid \mathscr{B}_{n-1})$ and $\tilde{E}(x_n \mid \mathscr{B}_{n-1})$, where E and $\tilde{E}$ are expectations 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 $\tilde{a}_{n-1}(x)$ (where $a_0(x) = a_0$, $\tilde{a}_0(x) = \tilde{a}_0$ are constants) and put

$$b_{n-1} = (E[x_n - a_{n-1}(x)]^2)^{1/2},$$
$$\tilde{b}_{n-1} = (\tilde{E}[x_n - \tilde{a}_{n-1}(x)]^2)^{1/2}.$$

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 (P-a.s.)

$$\begin{aligned} \xi_n &= a_{n-1}(\xi) + b_{n-1}\varepsilon_n, \\ \tilde{\xi}_n &= \tilde{a}_{n-1}(\xi) + \tilde{b}_{n-1}\tilde{\varepsilon}_n. \end{aligned} \tag{33}$$

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 distribution of the vector $(x_1, \ldots, x_n)$ will be concentrated (P-a.s.) on the linear manifold $x_n = a_{n-1}(x)$, and since by hypothesis $\tilde{P}_n \sim P_n$, we have $\tilde{b}_{n-1} = 0$, $a_{n-1} = \tilde{a}_{n-1}(x)$, and $\alpha_n(x) = 1$ (P- and $\tilde{P}$-a.s.). Hence we may suppose without loss of generality that $b_n^2 > 0$, $\tilde{b}_n^2 > 0$ for all $n \ge 1$, since otherwise the contribution of the corresponding terms of the sum $\sum_{n=1}^{\infty}[1 - E(\sqrt{\alpha_n} \mid \mathscr{B}_{n-1})]$ (see (16) and (17)) is zero.

Using the Gaussian hypothesis, we find from (33) that, for $n \ge 1$,

$$\alpha_n = d_{n-1}^{-1} \exp\left\{-\frac{(x_n - a_{n-1}(x))^2}{2b_{n-1}^2} + \frac{(x_n - \tilde{a}_{n-1}(x))^2}{2\tilde{b}_{n-1}^2}\right\}, \tag{34}$$

where $d_n = |b_n/\tilde{b}_n|$ and

$$a_0 = E\xi_1, \qquad \tilde{a}_0 = E\tilde{\xi}_1,$$
$$b_0^2 = \mathrm{Var}\xi_1, \quad \tilde{b}_0^2 = \mathrm{Var}\tilde{\xi}_1.$$

From (34),

$$\log E(\alpha_n^{1/2} \mid \mathscr{B}_{n-1}) = \frac{1}{2}\log\frac{2d_{n-1}}{1+d_{n-1}^2} - \frac{d_{n-1}^2}{1+d_{n-1}^2}\left(\frac{a_{n-1}(x)-\tilde{a}_{n-1}(x)}{b_{n-1}}\right)^2.$$

Since $\log[2d_{n-1}/(1+d_{n-1}^2)] \le 0$, statement (30) can be written in the form

$$\tilde{P} \ll P \Leftrightarrow \tilde{P}\Bigg\{\sum_{n=1}^{\infty}\Bigg[\frac{1}{2}\log\frac{1+d_{n-1}^2}{2d_{n-1}} + \frac{d_{n-1}^2}{1+d_{n-1}^2}\left(\frac{a_{n-1}(x)-\tilde{a}_{n-1}(x)}{b_{n-1}}\right)^2\Bigg] < \infty\Bigg\} = 1. \quad (35)$$

The series

$$\sum_{n=1}^{\infty}\log\frac{1+d_{n-1}^2}{2d_{n-1}} \quad \text{and} \quad \sum_{n=1}^{\infty}(d_{n-1}^2-1)$$

converge or diverge together; hence it follows from (35) that

$$\tilde{P} \ll P \Leftrightarrow \tilde{P}\Bigg\{\sum_{n=0}^{\infty}\Bigg[\left(\frac{\Delta_n^2(x)}{b_n^2}\right) + \left(\frac{\tilde{b}_n^2}{b_n^2}-1\right)^2\Bigg] < \infty\Bigg\} = 1, \quad (36)$$

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 $\{\Delta_n(x)/b_n\}_{n\ge 0}$ is a Gaussian system (with respect to both $\tilde{P}$ and P). As follows from the lemma that will be proved below,

$$\tilde{P}\Bigg\{\sum\left(\frac{\Delta_n(x)}{b_n}\right)^2 < \infty\Bigg\} = 1 \Leftrightarrow \sum\tilde{E}\left(\frac{\Delta_n(x)}{b_n}\right)^2 < \infty. \quad (37)$$

Hence it follows from (36) that

$$\tilde{P} \ll P \Leftrightarrow \sum_{n=0}^{\infty}\left[\tilde{E}\left(\frac{\Delta_n(x)}{b_n}\right)^2 + \left(\frac{\tilde{b}_n^2}{b_n^2}-1\right)^2\right] < \infty$$

and in a similar way

$$\tilde{P} \perp P \Leftrightarrow \tilde{P}\Bigg\{\sum_{n=0}^{\infty}\Bigg[\left(\frac{\Delta_n^2(x)}{b_n^2}\right) + \left(\frac{\tilde{b}_n^2}{b_n^2}-1\right)^2\Bigg] < \infty\Bigg\} = 0$$
$$\Leftrightarrow \sum_{n=0}^{\infty}\left[\tilde{E}\left(\frac{\Delta_n(x)}{b_n}\right)^2 + \left(\frac{\tilde{b}_n^2}{b_n^2}-1\right)^2\right] = \infty.$$

Then it is clear that if $\tilde{P}$ and P are not singular measures, we have $\tilde{P} \ll P$. But by hypothesis, $\tilde{P}_n \sim P_n,\ n \geq 1$; hence by symmetry, we have $P \ll \tilde{P}$. 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*,

$$\begin{aligned} \tilde{P} \sim P &\Leftrightarrow \sum_{n=0}^{\infty} \left[\tilde{E} \left(\frac{\Delta_n(x)}{b_n} \right)^2 + \left(\frac{\tilde{b}_n^2}{b_n^2} - 1 \right)^2 \right] < \infty, \\ \tilde{P} \perp P &\Leftrightarrow \sum_{n=0}^{\infty} \left[\tilde{E} \left(\frac{\Delta_n(x)}{b_n} \right)^2 + \left(\frac{\tilde{b}_n^2}{b_n^2} - 1 \right)^2 \right] = \infty. \end{aligned} \tag{38}$$

Lemma. *Let* $\beta = (\beta_n)_{n \geq 1}$ *be a Gaussian sequence defined on* $(\Omega, \mathscr{F}, \mathsf{P})$. *Then*

$$\mathsf{P}\left\{ \sum_{n=1}^{\infty} \beta_n^2 < \infty \right\} > 0 \Leftrightarrow \mathsf{P}\left\{ \sum_{n=1}^{\infty} \beta_n^2 < \infty \right\} = 1 \Leftrightarrow \sum_{n=1}^{\infty} \mathsf{E}\beta_n^2 < \infty. \tag{39}$$

PROOF. The implications $(\Leftarrow)$ are obvious. To establish the implications $(\Rightarrow)$, 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. Therefore then $\sum_{n=1}^{\infty} \mathsf{E}\beta_n^2 < \infty$, and hence $\mathsf{P}\left\{\sum_{n=1}^{\infty} \beta_n^2 < \infty\right\} = 1$ by the implication $(\Leftarrow)$.

Select an $n \geq 1$. Then it follows from Sects. 11 and 13, Chap. 2, Vol. 1, 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 another 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 $\mathsf{P}\{\sum_{n=1}^{\infty}\beta_n^2 < \infty\} > 0$, then $\mathsf{P}\{\sum_{n=1}^{\infty}(\beta_n - \tilde{\beta}_n)^2 < \infty\} > 0$, 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

$$\sum_{n=1}^{\infty}\mathsf{E}\beta_n^2 = \sum_{n=1}^{\infty}(\mathsf{E}\beta_n)^2 + \sum_{n=1}^{\infty}\mathsf{E}(\beta_n - \mathsf{E}\,\beta_n)^2 < \infty.$$

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 $\mathsf{E}\xi_i = 0$, $\operatorname{Var}\xi_i = V_i > 0$.

Again we let

$$X_{n+1} = \theta X_n + \xi_{n+1}$$

for $n \geq 0$, 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.

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

$$M_n = \sum_{k=0}^{n-1}\frac{X_k\xi_{k+1}}{V_{k+1}}, \quad \langle M\rangle_n = \sum_{k=0}^{n-1}\frac{X_k^2}{V_{k+1}}.$$

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 $P_\theta(\hat\theta_n \to \theta) = 1$. Let us show 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, 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 (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 $\mathsf{P}_\tau = \mathsf{P} \mid \mathscr{F}_\tau$ 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 the "recalculation formulas" (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

$$\tau = \min\{n \geq 1 \colon S_n < g(n)\}$$

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 $\xi_i \sim \mathscr{N}(0,1)$. 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

$$\log n = o\Bigg(\sum_{k=2}^{n}[\Delta g(k)]^2\Bigg), \quad n \to \infty. \tag{2}$$

Then

$$\mathsf{P}(\tau > n) = \exp\left\{-\frac{1}{2}\sum_{k=2}^{n}[\Delta g(k)]^2(1+o(1))\right\}, \quad n \to \infty. \tag{3}$$

Before starting the proof, let us observe that (1) and (2) are satisfied if, for example,

$$g(n) = an^{\nu} + b, \quad \tfrac{1}{2} < \nu \le 1, \quad a + b < 0, \quad a > 0,$$

or (for sufficiently large n)

$$g(n) = n^{\nu}L(n), \quad \tfrac{1}{2} \le \nu \le 1,$$

where $L(n)$ is a slowly varying function (e.g., $L(n) = C(\log n)^{\beta}$, $C > 0$, 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 $\mathscr{F}_0 = \{\varnothing, \Omega\}$, $\mathscr{F}_n = \sigma\{\xi_1, \ldots, \xi_n\}$, and let $\alpha = (\alpha_n, \mathscr{F}_{n-1})$ be a predictable sequence with $\mathsf{P}(|\alpha_n| \le C) = 1$, $n \ge 1$, where C is a constant. Form the sequence $z = (z_n, \mathscr{F}_n)$ with

$$z_n = \exp\left\{\sum_{k=1}^{n}\alpha_k\xi_k - \frac{1}{2}\sum_{k=1}^{n}\alpha_k^2\right\}, \quad n \ge 1. \tag{4}$$

It is easily verified that (with respect to P) the sequence $z = (z_n, \mathscr{F}_n)$ is a martingale with $\mathsf{E}\, z_n = 1$, $n \ge 1$.

Choose a value $n \ge 1$ and introduce a probability measure $\tilde{\mathsf{P}}_n$ on the measurable space $(\Omega, \mathscr{F}_n)$ by putting

$$\tilde{\mathsf{P}}_n(A) = \mathsf{E}\, I(A) z_n, \quad A \in \mathscr{F}_n. \tag{5}$$

Lemma 1 (Discrete version of Girsanov's theorem). *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 the expectation with respect to $\tilde{\mathsf{P}}_n$. Then for $\lambda_k \in R$, $1 \le k \le n$,

$$\begin{aligned}\tilde{\mathsf{E}}_n \exp\left\{i\sum_{k=1}^{n}\lambda_k\tilde{\xi}_k\right\} &= \mathsf{E}\exp\left\{i\sum_{k=1}^{n}\lambda_k\tilde{\xi}_k\right\}z_n \\ &= \mathsf{E}\left[\exp\left\{i\sum_{k=1}^{n-1}\lambda_k\tilde{\xi}_k\right\}z_{n-1}\cdot\mathsf{E}\left\{\exp\left(i\lambda_n(\xi_n-\alpha_n)+\alpha_n\xi_n-\frac{\alpha_n^2}{2}\right)\Big|\mathscr{F}_{n-1}\right\}\right]\end{aligned}$$

$$= \mathsf{E}\left[\exp\left\{i\sum_{k=1}^{n-1}\lambda_k\xi_k\right\}z_{n-1}\right]\exp\{-\tfrac{1}{2}\lambda_n^2\} = \cdots = \exp\left\{-\frac{1}{2}\sum_{k=1}^{n}\lambda_k^2\right\}.$$

Now the desired conclusion follows from Theorem 4, Sect. 12, Chap. 2, Vol. 1. □

Lemma 2. *Let $X = (X_n, \mathscr{F}_n)_{n\geq 1}$ be a square-integrable martingale with mean zero and*

$$\sigma = \min\{n \geq 1 : X_n \leq -b\},$$

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 $n \geq 1$,

$$\mathsf{P}(\sigma > n) \geq \frac{C}{\mathsf{E}\,X_n^2}. \tag{6}$$

PROOF. By Corollary 1 to Theorem 1 in Sect. 2, we have $\mathsf{E}\,X_{\sigma\wedge n} = 0$, whence

$$-\mathsf{E}\,I(\sigma \leq n)X_\sigma = \mathsf{E}\,I(\sigma > n)X_n. \tag{7}$$

On the set $\{\sigma \leq n\}$

$$-X_\sigma \geq b > 0.$$

Therefore, for $n \geq 1$,

$$-\mathsf{E}\,I(\sigma \leq n)\,X_\sigma \geq b\,\mathsf{P}(\sigma \leq n) \geq b\,\mathsf{P}(\sigma = 1) = b\,\mathsf{P}(X_1 < -b) > 0. \tag{8}$$

On the other hand, by the Cauchy–Schwarz inequality,

$$\mathsf{E}\,I(\sigma > n)\,X_n \leq [\mathsf{P}(\sigma > n)\cdot \mathsf{E}\,X_n^2]^{1/2}, \tag{9}$$

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

$$\liminf_{n\to\infty} \log \mathsf{P}(\tau > n)\Big/\sum_{k=2}^{n}[\Delta g(k)]^2 \geq -\tfrac{1}{2} \tag{10}$$

and

$$\limsup_{n\to\infty} \log \mathsf{P}(\tau > n)\Big/\sum_{k=2}^{n}[\Delta g(k)]^2 \leq -\tfrac{1}{2}. \tag{11}$$

For this purpose we consider the (nonrandom) sequence $(\alpha_n)_{n\geq 1}$ with

$$\alpha_1 = 0, \quad \alpha_n = \Delta g(n), \quad n \geq 2,$$

and the probability measures $(\tilde{\mathsf{P}}_n)_{n\geq 1}$ 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 $\tilde{\xi}_1, \ldots, \tilde{\xi}_n$ 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 $p > 1$,

$$\mathsf{P}(\tau > n) \geq C_p \exp\left\{ -\frac{p}{2} \sum_{k=2}^{n} [\Delta g(k)]^2 - \frac{p}{p-1} \log n \right\}, \tag{15}$$

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 in Sect. 3, Chap. 4)

$$\sum_{k=2}^{n} \Delta g(k) \cdot \xi_k = \Delta g(n) \cdot S_n - \sum_{k=2}^{n} S_{k-1} \Delta(\Delta g(k)).$$

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),

$$\begin{aligned}\mathsf{P}(\tau > n) &\leq \exp\left\{ -\frac{1}{2} \sum_{k=2}^{n} [\Delta g(k)]^2 - g(1) \Delta g(2) \right\} \tilde{\mathsf{E}}_n I(\tau > n) e^{-\xi_1 \Delta g(2)} \\ &= \exp\{-g(1) \Delta g(2)\} \exp\left\{ -\frac{1}{2} \sum_{k=2}^{n} [\Delta g(k)]^2 \right\} \tilde{\mathsf{E}}_n I(\tau > n) e^{-\xi_1 \Delta g(2)},\end{aligned}$$

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*

$$f(n) \to \infty, \quad n \to \infty,$$

and

$$\sum_{k=2}^{n} [\Delta f(k)]^2 = o\left(\sum_{k=1}^{n} f^{-2}(k) \right), \quad n \to \infty.$$

Then for

$$\sigma = \min\{n \geq 1 \colon |S_n| \geq f(n)\}$$

we have

$$\mathsf{P}(\sigma > n) = \exp\left\{ -\frac{\pi^2}{8}\sum_{k=1}^{n} f^{-2}(k)(1+o(1))\right\}, \quad n \to \infty. \tag{17}$$

4. PROBLEMS

1. Show that the sequence defined in (4) is a martingale. Is it still true without the condition $|\alpha_n| \le c$ (P-a.s.), $n \ge 1$?
2. Establish (13).
3. Prove (17).

8. Central Limit Theorem for Sums of Dependent Random Variables

1. In Sect. 4, Chap. 3, Vol. 1, the central limit theorem for sums $S_n = \xi_{n1} + \cdots + \xi_{nn}$, $n \ge 1$, of random variables $\xi_{n1}, \ldots, \xi_{nn}$ was established under the assumptions of their *independence*, *finiteness of second moments*, and *asymptotic negligibility* of their terms. In this section, we give up both the assumption of independence and even that of the finiteness of the absolute first-order moments. However, the asymptotic negligibility of the terms will be retained.

Thus, we suppose that on the probability space $(\Omega, \mathscr{F}, \mathsf{P})$ there are given stochastic sequences

$$\xi^n = (\xi_{nk}, \mathscr{F}_k^n), \quad 0 \le k \le n, \quad n \ge 1,$$

with $\xi_{n0} = 0$, $\mathscr{F}_0^n = \{\varnothing, \Omega\}$, $\mathscr{F}_k^n \subseteq \mathscr{F}_{k+1}^n \subseteq \mathscr{F}$ $(k+1 \le n)$. We set

$$X_t^n = \sum_{k=0}^{[nt]} \xi_{nk}, \quad 0 \le t \le 1.$$

Theorem 1. *For a given t, $0 < t \le 1$, let the following conditions be satisfied: for each $\varepsilon \in (0, 1)$, as $n \to \infty$,*

(A) $\sum_{k=1}^{[nt]} \mathsf{P}(|\xi_{nk}| > \varepsilon \,|\, \mathscr{F}_{k-1}^n) \xrightarrow{\mathsf{P}} 0$,

(B) $\sum_{k=1}^{[nt]} \mathsf{E}[\xi_{nk} I(|\xi_{nk}| \le \varepsilon) \,|\, \mathscr{F}_{k-1}^n] \xrightarrow{\mathsf{P}} 0$,

(C) $\sum_{k=1}^{[nt]} \mathrm{Var}[\xi_{nk} I(|\xi_{nk}| \le \varepsilon) \,|\, \mathscr{F}_{k-1}^n] \xrightarrow{\mathsf{P}} \sigma_t^2$, *where* $\sigma_t^2 \ge 0$.

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 $Z_t^n \xrightarrow{\mathsf{P}} 0$ 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 where 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) $\displaystyle\sum_{k=1}^{n} \mathsf{P}(|\xi_{nk}| > \varepsilon) \to 0,$

(b) $\displaystyle\sum_{k=1}^{n} \mathsf{E}[\xi_{nk} I(|\xi_{nk}| \le \varepsilon)] \to 0,$

(c) $\displaystyle\sum_{k=1}^{n} \operatorname{Var}[\xi_{nk} I(|\xi_{nk}| \le \varepsilon)] \to \sigma^2.$

These are well known; see the book by Gnedenko and Kolmogorov [33]. Hence we have the following corollary to Theorem 1.

Corollary. *If $\xi_{n1}, \ldots, \xi_{nn}$ are independent random variables, $n \ge 1$, then*

$$(\mathrm{a}), (\mathrm{b}), (\mathrm{c}) \Rightarrow X_1^n = \sum_{k=1}^{n} \xi_{nk} \xrightarrow{d} \mathscr{N}(0, \sigma^2).$$

Remark 2. In hypothesis (C), the case $\sigma_t^2 = 0$ is *not excluded*. Hence, in particular, Theorem 1 yields a convergence condition to the degenerate distribution ($X_t^n \xrightarrow{d} 0$).

Remark 3. The method used to prove Theorem 1 lets us state and prove the following more general proposition.

Let $0 = t_0 < t_1 < t_2 < \cdots < t_j \le 1$, $0 = \sigma_{t_0}^2 \le \sigma_{t_1}^2 \le \sigma_{t_2}^2 \le \cdots \le \sigma_{t_j}^2$, $\sigma_0^2 = 0$, 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 vector $(W_{t_1}, \ldots, W_{t_j})$ 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 $(\mathsf{P}^n_{t_1,\ldots,t_j})$ of the random variables $(X^n_{t_1}, \ldots, X^n_{t_j})$ converges weakly to the Gaussian distribution $\mathsf{P}_{t_1,\ldots,t_j}$ of the variables $(W_{t_1}, \ldots, W_{t_j})$:

$$\mathsf{P}^n_{t_1,\ldots,t_j} \xrightarrow{w} \mathsf{P}_{t_1,\ldots,t_j}.$$

Remark 4. Let $(\sigma_t^2)_{0\le t\le 1}$ be a continuous nondecreasing function, $\sigma_0^2 = 0$. Let $W = (W_t)_{0\le t\le 1}$ denote the *Brownian motion process* (the *Wiener process*) with $\mathsf{E}\,W_t = 0$ and $\mathsf{E}\,W_t^2 = \sigma_t^2$. This process was defined in Sect. 13, Chap. 2, Vol. 1, for $\sigma_t^2 = t$. In the general case, this process is defined in a similar way as the Gaussian process $W = (W_t)_{0\le t\le 1}$ with independent increments, $W_0 = 0$, and covariance function $r(s,t) = \min(\sigma_s^2, \sigma_t^2)$. It is shown in the general theory of stochastic processes that there always exists such a process with continuous paths. (In the case $\sigma_t^2 = t$, this process is called *standard* Brownian motion.)

If we denote by P^n and P the distributions of the processes X^n and W in the functional space $(D, \mathscr{B}(D))$ (Subsection 7, Sect. 2, Chap. 2, Vol. 1), then we can say that conditions (A), (B), and (C), fulfilled for all $0 < t \le 1$, ensure not only the convergence of finite-dimensional distributions ($P^n_{t_1,\ldots,t_j} \xrightarrow{w} P_{t_1,\ldots,t_j}$, $t_1 < t_2 < \cdots < t_j \le t$, $j = 1, 2, \ldots$) stated earlier, but also the *functional* convergence, i.e., the weak convergence of the distributions P^n of the processes X^n to the distribution

of the process W. (For details, see [4, 55, 43].) This result is usually called the *functional central limit theorem* or the *invariance principle* (when $\xi_{n1}, \dots, \xi_{nn}$ are independent, the latter is referred to as the *Donsker–Prohorov* invariance principle).

2. Theorem 2. 1. *Condition (A) is equivalent to the uniform asymptotic negligibility condition*

$$(\mathrm{A}^*) \qquad \max_{1\le k\le [nt]} |\xi_{nk}| \xrightarrow{\mathsf{P}} 0.$$

2. *Assuming* (A) *or* (A^*), *condition* (C) *is equivalent to*

$$(\mathrm{C}^*) \qquad \sum_{k=0}^{[nt]} [\xi_{nk} - \mathsf{E}(\xi_{nk} I(|\xi_{nk}| \le 1) \,|\, \mathscr{F}_{k-1}^n)]^2 \xrightarrow{\mathsf{P}} \sigma_t^2.$$

(The value of t in (A^*) and (C^*) is the same as in (A) and (C).)

Theorem 3. *For each $n \ge 1$ let the sequence*

$$\xi^n = (\xi_{nk}, \mathscr{F}_k^n), \quad 1 \le k \le n,$$

be a square-integrable martingale difference:

$$\mathsf{E}\,\xi_{nk}^2 < \infty, \quad \mathsf{E}(\xi_{nk} \,|\, \mathscr{F}_{k-1}^n) = 0.$$

Suppose that the Lindeberg condition is satisfied: *for any $\varepsilon > 0$,*

$$(\mathrm{L}) \qquad \sum_{k=0}^{[nt]} \mathsf{E}[\xi_{nk}^2 I(|\xi_{nk}| > \varepsilon) \,|\, \mathscr{F}_{k-1}^n] \xrightarrow{\mathsf{P}} 0.$$

Then (C) *is equivalent to*

$$\langle X^n \rangle_t \xrightarrow{\mathsf{P}} \sigma_t^2, \tag{1}$$

where (*quadratic characteristic*)

$$\langle X^n \rangle_t = \sum_{k=0}^{[nt]} \mathsf{E}(\xi_{nk}^2 \,|\, \mathscr{F}_{k-1}^n), \tag{2}$$

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

$$\begin{aligned} X_t^n &= \sum_{k=0}^{[nt]} \xi_{nk} I(|\xi_{nk}| \le 1) + \sum_{k=0}^{[nt]} \xi_{nk} I(|\xi_{nk}| > 1) \\ &= \sum_{k=0}^{[nt]} \mathsf{E}[\xi_{nk} I(|\xi_{nk}| \le 1) \,|\, \mathscr{F}_{k-1}^n] + \sum_{k=0}^{[nt]} \xi_{nk} I(|\xi_{nk}| > 1) \\ &\quad + \sum_{k=0}^{[nt]} \{\xi_{nk} I(|\xi_{nk}| \le 1) - \mathsf{E}[\xi_{nk} I(|\xi_{nk}| \le 1) \,|\, \mathscr{F}_{k-1}^n]\}. \end{aligned} \tag{7}$$

We define

$$\begin{aligned} B_t^n &= \sum_{k=0}^{[nt]} \mathsf{E}[\xi_{nk} I(|\xi_{nk}| \le 1) \,|\, \mathscr{F}_{k-1}^n], \\ \mu_k^n(\Gamma) &= I(\xi_{nk} \in \Gamma), \\ \nu_k^n(\Gamma) &= \mathsf{P}(\xi_{nk} \in \Gamma \,|\, \mathscr{F}_{k-1}^n), \end{aligned} \tag{8}$$

where Γ is a set from the smallest σ-algebra $\mathscr{B}_0 = \sigma(\mathscr{A}_0)$ generated by the system of sets $\mathscr{A}_0$ in $R_0 = R \setminus \{0\}$, which consists of finite sums of disjoint intervals $(a, b]$ not containing the point $\{0\}$, and $\mathsf{P}(\xi_{nk} \in \Gamma \,|\, \mathscr{F}_{k-1}^n)$ is a regular conditional distribution of ξ_{nk} given the σ-algebra $\mathscr{F}_{k-1}^n$.

Then (7) can be rewritten in the following form:

$$X_t^n = B_t^n + \sum_{k=1}^{[nt]} \int_{\{|x|>1\}} x \, d\mu_k^n + \sum_{k=1}^{[nt]} \int_{\{|x|\le 1\}} x \, d(\mu_k^n - \nu_k^n), \tag{9}$$

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

$$\sum_{k=1}^{[nt]} \int_{\{|x|>1\}} |x| \, d\mu_k^n \xrightarrow{\mathsf{P}} 0. \tag{10}$$

We have

$$\sum_{k=1}^{[nt]} \int_{\{|x|>1\}} |x| \, d\mu_k^n = \sum_{k=1}^{[nt]} |\xi_{nk}| \, I(|\xi_{nk}| > 1). \tag{11}$$

For every $\delta \in (0, 1)$,

$$\left\{\sum_{k=1}^{[nt]} |\xi_{nk}|\, I(|\xi_{nk}| > 1) > \delta\right\} = \left\{\sum_{k=1}^{[nt]} I(|\xi_{nk}| > 1) > \delta\right\}, \tag{12}$$

since each sum is greater than δ if $|\xi_{nk}| > 1$ for at least one k. It is clear that

$$\sum_{k=1}^{[nt]} I(|\xi_{nk}| > 1) \;= \sum_{k=1}^{[nt]} \int_{\{|x|>1\}} d\mu_k^n \quad (\equiv U_{[nt]}^n).$$

By (A),

$$V_{[nt]}^n \equiv \sum_{k=1}^{[nt]} \int_{\{|x|>1\}} d\nu_k^n \xrightarrow{\mathsf{P}} 0, \tag{13}$$

and V_k^n is $\mathscr{F}_{k-1}^n$-measurable.

Then, by the corollary to Theorem 4 in Sect. 3,

$$V_{[nt]}^n \xrightarrow{\mathsf{P}} 0 \Rightarrow U_{[nt]}^n \xrightarrow{\mathsf{P}} 0. \tag{14}$$

Note that by the same corollary and the inequality $\Delta U_{[nt]}^n \le 1$, we also have the converse implication:

$$U_{[nt]}^n \xrightarrow{\mathsf{P}} 0 \Rightarrow V_{[nt]}^n \xrightarrow{\mathsf{P}} 0, \tag{15}$$

which will be needed in the proof of Theorem 2.

The required proposition (10) now follows from (11)–(14).

Thus

$$X_t^n = Y_t^n + Z_t^n, \tag{16}$$

where

$$Y_t^n = \sum_{k=1}^{[nt]} \int_{\{|x|\le 1\}} x\, d(\mu_k^n - \nu_k^n), \tag{17}$$

and

$$Z_t^n = B_t^n + \sum_{k=1}^{[nt]} \int_{\{|x|>1\}} x\, d\mu_k^n \xrightarrow{\mathsf{P}} 0. \tag{18}$$

It then follows by Problem 1 that to establish that

$$X_t^n \xrightarrow{d} \mathscr{N}(0, \sigma_t^2),$$

we need only show that

$$Y_t^n \xrightarrow{d} \mathscr{N}(0, \sigma_t^2). \tag{19}$$

Let us represent Y_t^n in the form

$$Y_t^n = \gamma_{[nt]}^n(\varepsilon) + \Delta_{[nt]}^n(\varepsilon), \quad \varepsilon \in (0,1],$$

where

$$\gamma_{[nt]}^n(\varepsilon) = \sum_{k=1}^{[nt]} \int_{\{\varepsilon<|x|\le 1\}} x\, d(\mu_k^n - \nu_k^n), \tag{20}$$

$$\Delta_{[nt]}^n(\varepsilon) = \sum_{k=1}^{[nt]} \int_{\{|x|\le \varepsilon\}} x\, d(\mu_k^n - \nu_k^n). \tag{21}$$

As in the proof of (10), it is easily verified that, because of (A), we have $\gamma_{[nt]}^n(\varepsilon) \xrightarrow{\mathsf{P}} 0$, $n \to \infty$.

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 characteristic

$$\begin{aligned}\langle \Delta^n(\varepsilon)\rangle_k &= \sum_{i=1}^k \left[\int_{\{|x|\le\varepsilon\}} x^2 d\nu_i^n - \left(\int_{\{|x|\le\varepsilon\}} x\, d\nu_i^n \right)^2 \right] \\ &= \sum_{i=1}^k \operatorname{Var}[\xi_{ni} I(|\xi_{ni}| \le \varepsilon) \,|\, \mathscr{F}_{i-1}^n].\end{aligned}$$

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

$$\gamma_{[nt]}^n(\varepsilon_n) \xrightarrow{\mathsf{P}} 0, \quad \langle \Delta^n(\varepsilon_n)\rangle_{[nt]} \xrightarrow{\mathsf{P}} \sigma_t^2.$$

Therefore, again by Problem 1, it is enough to prove that

$$M_{[nt]}^n \xrightarrow{d} \mathscr{N}(0, \sigma_t^2), \tag{22}$$

where

$$M_k^n = \Delta_k^n(\varepsilon_n) = \sum_{i=1}^k \int_{\{|x|\le\varepsilon_n\}} x\, d(\mu_i^n - \nu_i^n). \tag{23}$$

For $\Gamma \in \mathscr{B}_0$, let

$$\tilde{\mu}_k^n(\Gamma) = I(\Delta M_k^n \in \Gamma), \quad \tilde{\nu}_k^n(\Gamma) = \mathsf{P}(\Delta M_k^n \in \Gamma \,|\, \mathscr{F}_{k-1}^n)$$

be a regular conditional probability, $\Delta M_k^n = M_k^n - M_{k-1}^n$, $k \geq 1$, $M_0^n = 0$. Then the square-integrable martingale $M^n = (M_k^n, \mathscr{F}_k^n)$, $1 \leq k \leq n$, can evidently be written in the form

$$M_k^n = \sum_{i=1}^{k} \Delta M_i^n = \sum_{i=1}^{k} \int_{\{|x| \leq 2\varepsilon_n\}} x \, d\tilde{\mu}_k^n.$$

(Notice that $|\Delta M_i^n| \leq 2\varepsilon_n$ by (23).)

To establish (22), we have, by Theorem 1 (Sect. 3, Chap. 3, Vol. 1), to show that, for every real λ,

$$\mathsf{E} \exp\{i\lambda M_{[nt]}^n\} \to \exp(-\tfrac{1}{2}\lambda^2 \sigma_t^2). \tag{24}$$

Set

$$G_k^n = \sum_{j=1}^{k} \int_{\{|x| \leq 2\varepsilon_n\}} (e^{i\lambda x} - 1) \, d\tilde{\nu}_j^n$$

and

$$\mathscr{E}_k^n(G^n) = \prod_{j=1}^{k} (1 + \Delta G_j^n).$$

Observe that

$$1 + \Delta G_k^n = 1 + \int_{\{|x| \leq 2\varepsilon_n\}} (e^{i\lambda x} - 1) \, d\tilde{\nu}_k^n = \mathsf{E}[\exp(i\lambda \Delta M_k^n) \mid \mathscr{F}_{k-1}^n],$$

and consequently,

$$\mathscr{E}_k^n(G^n) = \prod_{j=1}^{k} \mathsf{E}[\exp(i\lambda \Delta M_j^n) \mid \mathscr{F}_{j-1}^n].$$

By the lemma to be proved in Subsection 4, (24) will follow if, for every real λ,

$$|\mathscr{E}_{[nt]}^n(G^n)| = \left| \prod_{j=1}^{[nt]} \mathsf{E}[\exp(i\lambda \Delta M_j^n) \mid \mathscr{F}_{j-1}^n] \right| \geq c(\lambda) > 0 \tag{25}$$

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 $\mathscr{E}_t(A)$ defined by (76) of Sect. 6, Chap. 2, Vol. 1.)

Since

$$\int_{\{|x|\le 2\varepsilon_n\}} x\,d\tilde{\nu}_j^n = \mathsf{E}(\Delta M_j^n \mid \mathscr{F}_{j-1}^n) = 0,$$

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

$$\begin{aligned} |\Delta G_k^n| &\le \int_{\{|x|\le 2\varepsilon_n\}} |e^{i\lambda x} - 1 - i\lambda x|\,d\tilde{\nu}_k^n \le \frac{1}{2}\lambda^2 \int_{\{|x|\le 2\varepsilon_n\}} x^2\,d\tilde{\nu}_k^n \\ &\le \frac{1}{2}\lambda^2 (2\varepsilon_n)^2 \to 0 \end{aligned} \tag{28}$$

and

$$\sum_{j=1}^{k} |\Delta G_j^n| \le \frac{1}{2}\lambda^2 \sum_{j=1}^{k} \int_{\{|x|\le 2\varepsilon_n\}} x^2 d\tilde{\nu}_j^n = \frac{1}{2}\lambda^2 \langle M^n\rangle_k. \tag{29}$$

By (C),

$$\langle M^n\rangle_{[nt]} \xrightarrow{\mathsf{P}} \sigma_t^2. \tag{30}$$

Suppose first that $\langle M^n\rangle_{[nt]} \le a$ (P-a.s.). Then, by (28), (29), and Problem 3,

$$\prod_{k=1}^{[nt]} (1 + \Delta G_k^n)\exp(-\Delta G_k^n) \xrightarrow{\mathsf{P}} 1, \quad n \to \infty,$$

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

$$\begin{aligned} &\sum_{k=1}^{[nt]} \int_{\{|x|\le 2\varepsilon_n\}} |e^{i\lambda x} - 1 - i\lambda x + \tfrac{1}{2}\lambda^2 x^2|\,d\tilde{\nu}_k^n \le \tfrac{1}{6}|\lambda|^3 (2\varepsilon_n) \sum_{k=1}^{[nt]} \int_{\{|x|\le 2\varepsilon_n\}} x^2\,d\tilde{\nu}_k^n \\ &\quad = \tfrac{1}{3}\varepsilon_n|\lambda|^3 \langle M_n\rangle_{[nt]} \le \tfrac{1}{3}\varepsilon_n|\lambda|^3 a \to 0, \quad n \to \infty. \end{aligned}$$

Therefore, if $\langle M^n\rangle_{[nt]} \le a$ (P-a.s.), (31) is established and, consequently, so is (26).

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,

$$|\mathscr{E}_k^n(G^n)| = \left|\prod_{j=1}^{k}(1+\Delta G_i^n)\right| \ge \prod_{j=1}^{k}(1-\tfrac{1}{2}\lambda^2\Delta\langle M^n\rangle_j)$$
$$= \exp\left\{\sum_{j=1}^{k}\log(1-\tfrac{1}{2}\lambda^2\Delta\langle M^n\rangle_j)\right\}.$$

But

$$\log(1-\tfrac{1}{2}\lambda^2\Delta\langle M^n\rangle_j) \ge -\frac{\frac{1}{2}\lambda^2\Delta\langle M^n\rangle_j}{1-\frac{1}{2}\lambda^2\Delta\langle M^n\rangle_j}$$

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

$$\tau^n = \min\{k \le [nt]\colon \langle M^n\rangle_k \ge \sigma_t^2 + 1\},$$

taking $\tau^n = \infty$ if $\langle M^n\rangle_{[nt]} \le \sigma_t^2 + 1$.

Then, for $\overline{M} = M_{k\wedge\tau^n}$, 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 to show (using the Cramér–Wold method [4]) that for all real numbers $\lambda_1, \dots, \lambda_j$

$$\mathsf{E}\exp\left\{i\left[\lambda_1 M^n_{[nt_1]}+\sum_{k=2}^{j}\lambda_k(M^n_{[nt_k]}-M^n_{[nt_{k-1}]})\right]\right\}$$
$$\to\exp\left\{-\frac{1}{2}\lambda_1^2\sigma^2_{t_1}-\frac{1}{2}\sum_{k=2}^{j}\lambda_k^2(\sigma^2_{t_k}-\sigma^2_{t_{k-1}})\right\}.$$

The proof of this is similar to that of (24), replacing $(M^n_k,\ \mathscr{F}^n_k)$ by the square-integrable martingales $(\hat{M}^n_k,\ \mathscr{F}^n_k)$,

$$\hat{M}^n_k=\sum_{i=1}^{k}\nu_i\Delta M^n_i,$$

where $\nu_i=\lambda_1$ for $i\le[nt_1]$ and $\nu_i=\lambda_j$ for $[nt_{j-1}]<i\le[nt_j]$.

4. In this subsection we prove a simple lemma that lets us reduce the verification of (24) to the verification of (25) and (26).

Let $\eta^n=(\eta_{nk},\mathscr{F}^n_k),\ 1\le k\le n,\ n\ge 1$, be stochastic sequences, let

$$Y^n=\sum_{k=1}^{n}\eta_{nk},$$

let

$$\mathscr{E}^n(\lambda)=\prod_{k=1}^{n}\mathsf{E}\left[\exp(i\lambda\eta_{nk})\mid\mathscr{F}^n_{k-1}\right],\quad\lambda\in R,$$

and let Y be a random variable with

$$\mathscr{E}(\lambda)=\mathsf{E}\,e^{i\lambda Y},\quad\lambda\in R.$$

Lemma. *If (for a given λ) $|\mathscr{E}^n(\lambda)|\ge c(\lambda)>0,\ n\ge\mathbf{1}$, 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 $|m^n(\lambda)|\le c^{-1}(\lambda)<\infty$, and it is easily verified that

$$\mathsf{E}\,m^n(\lambda)=1.$$

Hence, by (34) and the Lebesgue dominated convergence theorem,

$$\begin{aligned}|\,\mathsf{E}\,e^{i\lambda Y^n} - \mathsf{E}\,e^{i\lambda Y}| &= |\,\mathsf{E}(e^{i\lambda Y^n} - \mathscr{E}(\lambda))| \le |\,\mathsf{E}(m^n(\lambda)[\mathscr{E}^n(\lambda) - \mathscr{E}(\lambda)])| \\ &\le c^{-1}(\lambda)\,\mathsf{E}\,|\mathscr{E}^n(\lambda) - \mathscr{E}(\lambda)| \to 0, \quad n \to \infty.\end{aligned}$$

□

Remark 5. It follows from (33) and the hypothesis that $\mathscr{E}^n(\lambda) \ge c(\lambda) > 0$ that $\mathscr{E}(\lambda) \ne 0$. In fact, the conclusion of the lemma remains valid without the assumption that $|\mathscr{E}^n(\lambda)| \ge c(\lambda) > 0$, if restated in the following form: *If* $\mathscr{E}^n(\lambda) \xrightarrow{\mathsf{P}} \mathscr{E}(\lambda)$ *and* $\mathscr{E}(\lambda) \ne 0$, *then* (33) *holds* (Problem 5).

5. PROOF OF THEOREM 2. 1. Let $0 < \varepsilon < 1$, $\delta \in (0, \varepsilon)$, and for simplicity let $t = 1$. Since

$$\max_{1\le k\le n} |\xi_{nk}| \le \varepsilon + \sum_{k=1}^{n} |\xi_{nk}| I(|\xi_{nk}| > \varepsilon)$$

and

$$\left\{ \sum_{k=1}^{n} |\xi_{nk}| I(|\xi_{nk}| > \varepsilon) > \delta \right\} = \left\{ \sum_{k=1}^{n} I(|\xi_{nk}| > \varepsilon) > \delta \right\},$$

we have

$$\begin{aligned}\mathsf{P}\left\{ \max_{1\le k\le n} |\xi_{nk}| > \varepsilon + \delta \right\} &\le \mathsf{P}\left\{ \sum_{k=1}^{n} I(|\xi_{nk}| > \varepsilon) > \delta \right\} \\ &= \mathsf{P}\left\{ \sum_{k=1}^{n} \int_{\{|x|>\varepsilon\}} d\mu_k^n > \delta \right\}.\end{aligned}$$

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 (cf. (10)) we also have

$$\mathsf{P}\left\{ \sum_{k=1}^{n} \int_{\{|x|>\varepsilon\}} d\mu_k^n > \delta \right\} \to 0.$$

Therefore $(\mathrm{A}) \Rightarrow (\mathrm{A}^*)$.

Conversely, let

$$\sigma_n = \min\{k \le n\colon |\xi_{nk}| \ge \varepsilon/2\}$$

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^*),

$$\sum_{k=1}^{n\wedge\sigma_n} I(|\xi_{nk}| \geq \varepsilon/2) = \sum_{k=1}^{n\wedge\sigma_n} \int_{\{|x|\geq\varepsilon/2\}} d\mu_k^n \xrightarrow{\mathsf{P}} 0.$$

Therefore, by (15),

$$\sum_{k=1}^{n\wedge\sigma_n} \int_{\{|x|\geq\varepsilon\}} d\nu_k^n \leq \sum_{k=1}^{n\wedge\sigma_n} \int_{\{|x|\geq\varepsilon/2\}} d\nu_k^n \xrightarrow{\mathsf{P}} 0,$$

which, together with the property $\lim_n \mathsf{P}(\sigma_n < \infty) = 0$, proves that $(\mathrm{A}^*) \Rightarrow (\mathrm{A})$.

2. Again suppose that $t = 1$. Choose an $\varepsilon \in (0, 1]$ and consider the square-integrable martingales (see (21))

$$\Delta^n(\delta) = (\Delta_k^n(\delta), \mathscr{F}_k^n) \quad (1 \leq k \leq n)$$

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 from (C^*) and (A) or, equivalently, from (C^*) and (A^*), it follows that, for every $\delta \in (0, \varepsilon]$,

$$[\Delta^n(\delta)]_n \xrightarrow{\mathsf{P}} \sigma_1^2, \tag{36}$$

where

$$[\Delta^n(\delta)]_n = \sum_{k=1}^n \left[\xi_{nk} I(|\xi_{nk}| \leq \delta) - \int_{\{|x|\leq\delta\}} x\, d\nu_k^n \right]^2.$$

In fact, it is easily verified that, by (A),

$$[\Delta^n(\delta)]_n - [\Delta^n(1)]_n \xrightarrow{\mathsf{P}} 0. \tag{37}$$

But

$$\left| \sum_{k=1}^n \left[\xi_{nk} - \int_{\{|x|\leq 1\}} x\, d\nu_k^n \right]^2 - \sum_{k=1}^n \left[\xi_{nk} I(|\xi_{nk}| \leq 1) - \int_{\{|x|\leq 1\}} x\, d\nu_k^n \right]^2 \right|$$
$$\leq \sum_{k=1}^n I(|\xi_{nk}| > 1) \left[\xi_{nk}^2 + 2|\xi_{nk}| \left| \int_{\{|x|\leq 1\}} x\, d(\mu_k^n - \nu_k^n) \right| \right]$$

$$\leq 5\sum_{k=1}^{n} I(|\xi_{nk}| > 1)\xi_{nk}^2$$

$$\leq 5 \max_{1\leq k\leq n} \xi_{nk}^2 \cdot \sum_{k=1}^{n} \int_{\{|x|>1\}} d\mu_k^n \to 0. \tag{38}$$

Hence (36) follows from (37) and (38).

Consequently, to establish the equivalence of (C) and (C*), it is enough to establish that both (C) (for a given $\varepsilon \in (0, 1]$) and (C*) imply that, for every $a > 0$,

$$\lim_{\delta\to 0} \limsup_n \mathsf{P}\{|[\Delta^n(\sigma)]_n - \langle\Delta^n(\delta)\rangle_n| > a\} = 0. \tag{39}$$

Let

$$m_k^n(\delta) = [\Delta^n(\delta)]_k - \langle\Delta^n(\delta)\rangle_k, \quad 1 \leq k \leq n.$$

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 from Sect. 3) by the sequences $[m^n(\delta)]$ and $\langle m^n(\delta)\rangle$.

It is clear that

$$[m^n(\delta)]_n = \sum_{k=1}^{n} (\Delta m_k^n(\delta))^2 \leq \max_{1\leq k\leq n} |\Delta m_k^n(\delta)| \cdot \{[\Delta^n(\delta)]_n + \langle\Delta^n(\delta)\rangle_n\}$$

$$\leq 3\delta^2\{[\Delta^n(\delta)]_n + \langle\Delta^n(\delta)\rangle_n\}. \tag{40}$$

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 (e.g., for $\delta < \min\left(\varepsilon, \frac{1}{6}b(\sigma_1^2 + 1)\right)$)

$$\lim_n \mathsf{P}(6\delta^2\langle\Delta^n(\delta)\rangle_n > b) = 0,$$

and hence, by the corollary to Theorem 4 (Sect. 3), we have (39).

On the other hand, if (C*) is satisfied, then for the same values of δ,

$$\lim_n \mathsf{P}(6\delta^2[\Delta^n(\delta)]_n > b) = 0. \tag{41}$$

Since $|\Delta[\Delta^n(\delta)]_k| \leq (2\delta)^2$, the validity of (39) follows from (41) and another appeal to the corollary to Theorem 4 (Sect. 3).

This completes the proof of Theorem 8. □

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 $B_t^n \xrightarrow{\mathsf{P}} 0$ 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 of the central limit theorem without such a hypothesis. For independent random variables, an example of such a theorem was Theorem 1 in Sect. 5, Chap. 3, Vol. 1 (assuming finite second moments).

We quote (without proof) an analog of this theorem, restricting ourselves to sequences $\xi^n = (\xi_{nk}, \mathscr{F}_k^n)$, $1 \le k \le n$, that are square-integrable martingale differences ($\mathsf{E}\,\xi_{nk}^2 < \infty$, $\mathsf{E}(\xi_{nk} \,|\, \mathscr{F}_{k-1}^n) = 0$).

Let $F_{nk}(x) = \mathsf{P}(\xi_{nk} \le x \,|\, \mathscr{F}_{k-1}^n)$ be a regular distribution function of ξ_{nk} given $\mathscr{F}_{k-1}^n$, and let $\Delta_{nk} = \mathsf{E}(\xi_{nk}^2 \,|\, \mathscr{F}_{k-1}^n)$.

Theorem 5. *If a square-integrable martingale difference $\xi_n = (\xi_{nk}, \mathscr{F}_k^n)$, $0 \le k \le n$, $n \ge 1$, satisfies the conditions*

$$\sum_{k=0}^{[nt]} \Delta_{nk} \xrightarrow{\mathsf{P}} \sigma_t^2, \quad 0 \le \sigma_t^2 < \infty, \ 0 \le t \le 1,$$

and for every $\varepsilon > 0$

$$\sum_{k=0}^{[nt]} \int_{\{|x|>\varepsilon\}} |x| \left| F_{nk}(x) - \Phi\left(\frac{x}{\sqrt{\Delta_{nk}}}\right) \right| dx \xrightarrow{\mathsf{P}} 0,$$

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 $\xi_n(\varepsilon) \xrightarrow{\mathsf{P}} 0$ for each $\varepsilon > 0$ as $n \to \infty$. Using, for example, Problem 11 from Sect. 10, Chap. 2, Vol. 1, prove that there is a sequence $\varepsilon_n \downarrow 0$ such that $\xi_n(\varepsilon_n) \xrightarrow{\mathsf{P}} 0$.
3. Let (α_k^n), $1 \le k \le n$, $n \ge 1$, be complex-valued random variables such that (P-a.s.)

$$\sum_{k=1}^{n} |\alpha_k^n| \le C, \quad |\alpha_k^n| \le a_n \downarrow 0.$$

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 5 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 (e.g., martingales, local martingales, semimartingales) *Itô's change-of-variables formula* plays a key role. In this section, we present a *discrete* (in time) version of this formula and show briefly how *Itô's formula for Brownian motion* could be derived from it 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 $(\Omega, \mathscr{F}, \mathsf{P})$, $X_0 = Y_0 = 0$, 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 *quadratic covariation* of X and Y (Sect. 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 $f = f(y)$, $y \in R$, is a Borel function such that

$$\int_{|y| \le c} |f(y)|\, dy < \infty, \quad c > 0.$$

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" functionals of the sequence $X = (X_n)_{0 \le n \le N}$.

Given the function $f = f(x)$ as in (2), consider the quadratic covariation $[X, f(X)]$ of the sequences X and $f(X) = (f(X_n))_{0 \le n \le N}$. By (1),

$$\begin{aligned}[X, f(X)]_n &= \sum_{k=1}^{n} \Delta f(X_k) \Delta X_k \\ &= \sum_{k=1}^{n} (f(X_k) - f(X_{k-1}))(X_k - X_{k-1}). \end{aligned} \tag{4}$$

We introduce two "discrete integrals" (cf. Definition 5 in Sect. 1):

$$I_n(X, f(X)) = \sum_{k=1}^{n} f(X_{k-1}) \Delta X_k, \quad 1 \le n \le N, \tag{5}$$

$$\tilde{I}_n(X, f(X)) = \sum_{k=1}^{n} f(X_k) \Delta X_k, \quad 1 \le n \le N. \tag{6}$$

Then

$$[X, f(X)]_n = \tilde{I}_n(X, f(X)) - I_n(X, f(X)). \tag{7}$$

(For $n = 0$, we set $I_0 = \tilde{I}_0 = 0$.)

For a *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,

$$\tilde{I}_n(X, f(X)) = -\{I_N(\tilde{X}, f(\tilde{X})) - I_{N-n}(\tilde{X}, f(\tilde{X}))\}.$$

From this and (7) we obtain

$$[X, f(X)]_N = -\{I_N(\tilde{X}, f(\tilde{X})) + I_N(X, f(X))\}$$

and for $0 < n < N$ we have

$$\begin{aligned}[X, f(X)]_n &= -\{I_N(\tilde{X}, f(\tilde{X})) - I_{N-n}(\tilde{X}, f(\tilde{X}))\} - I_n(X, f(X)) \\ &= -\left\{ \sum_{k=N-n+1}^{N} f(\tilde{X}_{k-1}) \Delta \tilde{X}_k + \sum_{k=1}^{n} f(X_{k-1}) \Delta X_k \right\}. \end{aligned} \tag{9}$$

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" in the sense that the value $f(X_{k-1})$ of f at the left end of the interval $[k-1, k]$ is multiplied by the increment $\Delta X_k = X_k - X_{k-1}$ on this interval, whereas in $\tilde{I}_n(X, f(X))$ the increment ΔX_k is multiplied by the value $f(X_k)$ at the right end of $[k-1, k]$.

Thus, (7) contains both the "forward integral" $I_n(X, f(X))$ and the "backward integral" $\tilde{I}_n(X, f(X))$, while in (9), both integrals are "forward integrals," over two *different* 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

$$\begin{aligned} F(X_n) &= F(X_0) + \sum_{k=1}^{n} g(X_{k-1})\Delta X_k + \tfrac{1}{2}[X,\ g(X)]_n \\ &\quad + \sum_{k=1}^{n} \left\{ (F(X_k) - F(X_{k-1})) - \frac{g(X_{k-1}) + g(X_k)}{2} \Delta X_k \right\}. \end{aligned} \tag{10}$$

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 (12),

$$R_n(X, f(X)) = -\frac{1}{12} \sum_{k=1}^{n} f''(\eta_k)(\Delta X_k)^3,$$

where $X_{k-1} \le \eta_k \le X_k$, whence

$$|R_n(X, f(X))| \le \frac{1}{12} \sup f''(\eta) \sum_{k=1}^{n} |\Delta X_k|^3, \tag{13}$$

where the supremum is taken over all η such that

$$\min(X_0, X_1, \dots, X_n) \le \eta \le \max(X_0, X_1, \dots, X_n).$$

We shall refer to formula (11) as the *discrete analog 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 quadratic covariation $[X, f(X)]_n$, and the "remainder" term $R_n(X, f(X))$, which is so termed because it goes to zero in the limit transition to the continuous time (see Subsection 5 for details).

4.

EXAMPLE 1. If $f(x) = a + bx$, then $R_n(X, f(X)) = 0$, and formula (11) takes the following form:

$$F(X_n) = F(X_0) + I_n(X, f(X)) + \tfrac{1}{2}[X, f(X)]_n. \tag{14}$$

(Compare with formula (19) below.)

EXAMPLE 2. Let

$$f(x) = \operatorname{sign} x = \begin{cases} 1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0, \end{cases}$$

and let $F(x) = |x|$.

Let $X_k = S_k$, where

$$S_k = \xi_1 + \xi_2 + \cdots + \xi_k$$

with $\xi_1, \xi_2, \dots$ independent Bernoulli random variables taking values ± 1 with probability 1/2.

If we also set $S_0 = 0$, we obtain from (11) that

$$|S_n| = \sum_{k=1}^{n} (\operatorname{sign} S_{k-1}) \Delta S_k + N_n, \tag{15}$$

where

$$N_n = \#\{0 \le k < n, S_k = 0\}$$

is the number of zeroes in the sequence $S_0, S_1, \dots, S_{n-1}$.

We note that the sequence of discrete integrals $(\sum_{k=1}^{n} (\operatorname{sign} S_{k-1}) \Delta S_k)_{n \ge 1}$ involved in (14) forms a martingale, and therefore

$$\mathsf{E}\,|S_n| = \mathsf{E}\,N_n. \tag{16}$$

Since (Problem 2)

$$\mathsf{E}\,|S_n| \sim \sqrt{\frac{2}{\pi} n}, \quad n \to \infty, \tag{17}$$

(16) yields

$$\mathsf{E}\,N_n \sim \sqrt{\frac{2}{\pi} n}, \quad n \to \infty. \tag{18}$$

In other words, the average number of "draws" in the random walk $S_0, S_1, \ldots, S_n$ has order of growth $\sqrt{n}$ rather than n, which could seem more natural at first glance. Note that the property (18) is closely related to the *arcsine law* (Sect. 10, Chap. 1, Vol. 1) since it is actually its consequence.

5. Let $B = (B_t)_{0\le t\le 1}$ be a standard ($B_0 = 0$, $\mathsf{E}\, B_t = 0$, $\mathsf{E}\, B_t^2 = t$) Brownian motion (Sect. 13, Chap. 2, Vol. 1), and let $X_k = B_{k/n}$, $k = 0, 1, \ldots, n$. Then application of formula (11) leads to the following result:

$$F(B_1) = F(B_0) + \sum_{k=1}^{n} f(B_{(k-1)/n})\Delta B_{k/n} + \frac{1}{2}[f(B_{\cdot/n}),\, B_{\cdot/n}]_n + R_n(B_{\cdot/n}, f(B_{\cdot/n})). \tag{19}$$

It is known from the stochastic calculus of Brownian motion (e.g., [75, 32]) that

$$\sum_{k=1}^{n} |B_{k/n} - B_{(k-1)/n}|^3 \xrightarrow{\mathsf{P}} 0, \quad n \to \infty. \tag{20}$$

Therefore, if $f = f(x)$ is twice differentiable and $|f''(x)| \le C$, $x \in R$, for some $C > 0$, then we obtain from (13) that $R_n(B_{\cdot/n}, f(B_{\cdot/n})) \xrightarrow{\mathsf{P}} 0$.

Appealing again to Brownian motion theory, we obtain that for any Borel function $f = f(x) \in L^2_{\mathrm{loc}}$ (i.e., such that $\int_{|x|\le C} f^2(x)\, dx < \infty$ for any $C > 0$) there exists the limit (in probability) of "discrete integrals" $\sum_{k=1}^{n} f(B_{(k-1)/n})\Delta B_{k/n}$. This limit is denoted by $\int_0^1 f(B_s)\, dB_s$ and called *Itô's stochastic integral with respect to Brownian motion*.

Therefore, turning to (19), we see that $R_n(B_{\cdot/n}, f(B_{\cdot/n})) \xrightarrow{\mathsf{P}} 0$, the "discrete integrals" $\sum_{k=1}^{n} f(B_{(k-1)/n})\Delta B_{k/n}$ converge (in probability) to the "stochastic integral" $\int_0^1 f(B_s)\, dB_s$, and hence there exists the limit in probability of quadratic covariations

$$[B_{\cdot/n}, f(B_{\cdot/n})] \quad (= [f(B_{\cdot/n}),\, B_{\cdot/n}]),$$

which can naturally be denoted by

$$[B, f(B)]_1.$$

Thus, if $f = f(x)$ is twice differentiable, $|f''(x)| \le C$, $x \in R$, and $f \in L^2_{\mathrm{loc}}$, then

$$F(B_1) = F(0) + \int_0^1 f(B_s)\, dB_s + \frac{1}{2}[B, f(B)]_1. \tag{21}$$

We have here

$$[B, f(B)]_1 = \int_0^1 f'(B_s)\, ds, \tag{22}$$

and therefore

$$F(B_1) = F(0) + \int_0^1 f(B_s)\,dB_s + \frac{1}{2}\int_0^1 f'(B_s)\,ds, \tag{23}$$

or, in a more standard form,

$$F(B_1) = F(0) + \int_0^1 F'(B_s)\,dB_s + \frac{1}{2}\int_0^1 F''(B_s)\,ds. \tag{24}$$

This formula (for $F \in C^2$) is referred to as *Itô's change-of-variables formula for Brownian motion*.

6. PROBLEMS

1. Prove formula (15).
2. Establish that property (17) is true.
3. Prove formula (22).
4. Try to prove that (24) holds for any $F \in C^2$.

10. Application of Martingale Methods to Calculation of Probability of Ruin in Insurance

1. The material studied in this section is a good illustration of the fact that the *theory of martingales* provides a simple way of estimating the *risk* faced by an insurance company.

We shall assume that the evolution of the capital of a certain insurance company is described by a random process $X = (X_t)_{t\geq 0}$. 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 $T_1, T_2, \ldots$ $(0 < T_1 < T_2 < \cdots)$, where the amounts to be paid out at these times are described by nonnegative random variables $\xi_1, \xi_2, \ldots$.

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 by

$$T = \inf\{t \geq 0 \colon X_t \leq 0\}$$

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 set equal to $+\infty$.

One of the main questions relating to the operation of an insurance company is the calculation (or estimation) of the *probability of ruin*, $\mathsf{P}(T < \infty)$, and the *probability of ruin before time* t, $\mathsf{P}(T \le t)$ (inclusively).

2. This is a rather complicated problem. However, it can be solved (partially) in the framework of the classical *Cramér–Lundberg model* characterized by the following assumptions.

A. The times $T_1, T_2, \dots$ at which claims are received are such that the variables $(T_0 \equiv 0)$

$$\sigma_i = T_i - T_{i-1}, \quad i \ge 1,$$

are independent identically distributed random variables having an exponential distribution with density $\lambda e^{-\lambda t}$, $t \ge 0$ (see Table 2.3 in Sect. 3, Chap. 2, Vol. 1).

B. The random variables $\xi_1, \xi_2, \dots$ are independent identically distributed with distribution function $F(x) = \mathsf{P}(\xi_1 \le x)$ 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 of Sect. 5, Chap. 2, Vol. 1).

Denote by

$$N_t = \sum_{i \ge 1} I(T_i \le t) \tag{3}$$

the process describing the number of claims before time t (inclusively), $N_0 = 0$.

Since

$$\{T_k > t\} = \{\sigma_1 + \cdots + \sigma_k > t\} = \{N_t < k\}, \quad k \ge 1,$$

under assumption **A**, we find that, according to Problem 6 in Sect. 8, Chap. 2, Vol. 1,

$$\mathsf{P}(N_t < k) = \mathsf{P}(\sigma_1 + \cdots + \sigma_k > t) = \sum_{i=0}^{k-1} e^{-\lambda t} \frac{(\lambda t)^i}{i!},$$

whence

$$\mathsf{P}(N_t = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}, \quad k = 0, 1, \dots, \tag{4}$$

i.e., the random variable N_t has the Poisson distribution (see Table 2.2 in Sect. 3, Chap. 2, Vol. 1) with parameter λt. Here, $\mathsf{E}\, N_t = \lambda t$.

The *Poisson process* $N = (N_t)_{t \ge 0}$ constructed in this way is a special case of a renewal process (Subsection 4, Sect. 9, Chap. 2, Vol. 1). The trajectories of this process are discontinuous (specifically, piecewise-constant, continuous on the right, and with unit jumps). Like Brownian motion (Sect. 13, Chap. 2, Vol. 1) having continuous trajectories, this process plays a fundamental role in the theory of random processes. From these two processes can be built random processes of rather complicated probabilistic structure. (We mention processes with independent increments as a typical example of these; see, e.g., [31, 75, 68].)

3. From assumption **C** we find that

$$\begin{aligned}\mathsf{E}(X_t - X_0) &= ct - \mathsf{E}\,S_t = ct - \mathsf{E}\sum_i \xi_i I(T_i \le t) = ct - \sum_i \mathsf{E}\,\xi_i I(T_i \le t)\\ &= ct - \sum_i \mathsf{E}\,\xi_i\,\mathsf{E}\,I(T_i \le t) = ct - \mu\sum_i \mathsf{P}(T_i \le t)\\ &= ct - \mu\sum_i \mathsf{P}(N_t \ge i) = ct - \mu\,\mathsf{E}\,N_t = t(c - \lambda\mu).\end{aligned}$$

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

$$h(z) = \int_0^\infty (e^{zx} - 1)\,dF(x), \quad z \ge 0, \tag{6}$$

which is equal to $\hat{F}(-z) - 1$, where

$$\hat{F}(s) = \int_0^\infty e^{-sx} dF(x)$$

is the Laplace–Stieltjes transform of F (with s a complex number).

Using the notation

$$g(z) = \lambda h(z) - cz, \quad \xi_0 = 0,$$

we find that for any $r > 0$ 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 $\mathscr{F}_t^X = \sigma(X_s,\ s \le t)$. Since the process $X = (X_t)_{t\ge 0}$ is a process with independent increments (Problem 2), we have (P-a.s.)

$$\mathsf{E}(e^{-r(X_t - X_s)} \mid \mathscr{F}_s^X) = \mathsf{E}\, e^{-r(X_t - X_s)} = e^{(t-s)g(r)},$$

hence (P-a.s.)

$$\mathsf{E}(e^{-rX_t - tg(r)} \mid \mathscr{F}_s^X) = e^{-rX_s - sg(r)}. \tag{8}$$

Using the notation

$$Z_t = e^{-rX_t - tg(r)}, \quad t \ge 0, \tag{9}$$

we see that property (8) can be rewritten in the form

$$\mathsf{E}(Z_t \mid \mathscr{F}_s^X) = Z_s, \quad s \le t \quad (\mathsf{P}\text{-a.s.}). \tag{10}$$

It is natural to say, by analogy with Definition 1 in Sect. 1, that the process $Z = (Z_t)_{t\ge 0}$ is a *martingale*(with respect to the "flow" $(\mathscr{F}_t^X)_{t\ge 0}$ of σ-algebras). Notice that in this case $\mathsf{E}\,|Z_t| < \infty$, $t \ge 0$ (cf. (1) in Sect. 1).

By analogy with Definition 3 in Sect. 1, we shall say that the random variable $\tau = \tau(\omega)$ with values in $[0, +\infty]$ is a *Markov time*, or a *random variable independent of the future* (relative to the "flow" of σ-algebras $(\mathscr{F}_t^X)_{t\ge 0}$) if for each $t \ge 0$ the set

$$\{\tau(\omega) \le t\} \in \mathscr{F}_t^X.$$

It turns out that for martingales with continuous time, which are considered now, Theorem 1 from Sect. 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.

Let $\tau = T$. Then, by virtue of (9), we find from (11) that for any $t > 0$

$$\begin{aligned} e^{-ru} &= \mathsf{E}\, e^{-rX_{t\wedge T} - (t\wedge T)g(r)} \\ &\ge \mathsf{E}[\, e^{-rX_{t\wedge T} - (t\wedge T)g(r)} \mid T \le t\,]\, \mathsf{P}(T \le t) \\ &= \mathsf{E}[\, e^{-rX_T - Tg(r)} \mid T \le t\,]\, \mathsf{P}(T \le t) \\ &\ge \mathsf{E}[\, e^{-Tg(r)} \mid T \le t\,]\, \mathsf{P}(T \le t) \ge \min_{0\le s\le t} e^{-sg(r)}\, \mathsf{P}(T \le t). \end{aligned}$$

Therefore

$$\mathsf{P}(T \le t) \le \frac{e^{-ru}}{\min_{0\le s\le t} e^{-sg(r)}} = e^{-ru} \max_{0\le s\le t} e^{sg(r)}. \tag{12}$$

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 $g(R) = 0$.

Note that for $r > 0$

$$\begin{aligned}\int_0^\infty e^{rx}(1-F(x))\,dx &= \int_0^\infty \int_x^\infty e^{rx}\,dF(y)\,dx \\ &= \int_0^\infty \left(\int_0^y e^{rx}\,dx\right) dF(y) \\ &= \frac{1}{r}\int_0^\infty (e^{ry}-1)\,dF(y) = \frac{1}{r}h(r).\end{aligned}$$

From this and $\lambda H(R) - cR = 0$ we conclude that R is 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 $t > 0$,

$$\mathsf{P}(T \le t) \le e^{-Ru}, \tag{14}$$

whence

$$\mathsf{P}(T < \infty) \le e^{-Ru}. \tag{15}$$

Hence we have proved the following theorem.

Theorem. *Suppose that in the Cramér–Lundberg model assumptions* $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *and property* (5) *are satisfied* (*i.e.*, $\lambda\mu < c$). *Then the ruin probabilities* $\mathsf{P}(T \le t)$ *and* $\mathsf{P}(T < \infty)$ *satisfy* (14) *and* (15), *where* R *is the positive* (*and unique*) *root of Eq.* (13).

4. In the foregoing proof, we used relation (11), which, as we said, follows from a continuous-time analog of Theorem 1, Sect. 2 (on preservation of the martingale property under random time change). The proof of this continuous-time result can be found, for example, in [54, Sect. 3.2]. However, if we assumed that σ_i, $i = 1, 2, \ldots$, had a (discrete) geometric (rather than an exponential) distribution ($\mathsf{P}\{\sigma_i = k\} = q^{k-1}p,\ k \ge 1$), then Theorem 1 of Sect. 2 would suffice.

The derivations in this section, which appeal to the theory of random processes with *continuous time*, demonstrate, in particular, how mathematical models with continuous time arise in applied problems.

5. Problems

1. Prove that the process $N = (N_t)_{t\ge 0}$ (under assumption $\mathbf{A}$) is a process with independent increments.
2. Prove that $X = (X_t)_{t\ge 0}$ is also a process with independent increments.
3. Consider the Cramér–Lundberg model and obtain an analog of the foregoing theorem assuming that the variables σ_i, $i = 1, 2, \ldots$, have a geometric (rather than exponential) distribution ($\mathsf{P}(\sigma_i = k) = q^{k-1}p,\ k = 1, 2, \ldots$).

11. Fundamental Theorems of Stochastic Financial Mathematics: The Martingale Characterization of the Absence of Arbitrage

1. In the previous section we applied the martingale theory to the proof of the *Cramér–Lundberg theorem*, which is a basic result of the mathematical theory of insurance. In this section the martingale theory will be applied to the problem of *absence of arbitrage* in a financial market in the situation of stochastic indeterminacy. In what follows, Theorems 1 and 2, which are called the *fundamental theorems* of arbitrage theory in stochastic financial mathematics, are of particular interest because they state conditions for the absence of arbitrage in *martingale terms* (in a sense to be explained later) in the markets under consideration as well as conditions that guarantee the possibility of meeting financial obligations. (For a more detailed exposition of the financial mathematics, see [71].)

2. Let us give some definitions. It will be assumed throughout that we are given a *filtered probability space* $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geq 0}, \mathsf{P})$, which describes the stochastic indeterminacy of the evolution of prices, financial indexes, and other financial indicators. The totality of events in $\mathscr{F}_n$ will be interpreted as the information available at time n (inclusive). For example, $\mathscr{F}_n$ may comprise information about the particular values of some financial assets or financial indexes, for example.

The main object of the fundamental theorems will be the concept of a (B, S)-market, defined as follows.

Let $B = (B_n)_{n\geq 0}$ and $S = (S_n)_{n\geq 0}$ be positive random sequences. It is assumed that B_n for every $n \geq 0$ is $\mathscr{F}_{n-1}$-measurable, whereas S_n is $\mathscr{F}_n$-measurable. For simplicity, we assume that the initial σ-algebra $\mathscr{F}_0$ is trivial, i.e., $\mathscr{F}_0 = \{\varnothing, \Omega\}$ (Sect. 2, Chap. 2, Vol. 1). Therefore B_0 and S_0 are constants. In the terminology of Sect. 1, $B = (B_n)_{n\geq 0}$ and $S = (S_n)_{n\geq 0}$ are *stochastic sequences*, and moreover, the sequence $B = (B_n)_{n\geq 0}$ is *predictable* (since B_n are $\mathscr{F}_{n-1}$-measurable).

The financial meaning of $B = (B_n)_{n\geq 0}$ is that it describes the evolution of a bank account with initial value B_0. The fact that B_n is $\mathscr{F}_{n-1}$-measurable means that the state of the bank account at time n (say, "today") becomes already known at time $n-1$ ("yesterday").

If we let

$$r_n = \frac{\Delta B_n}{B_{n-1}}, \quad n \geq 1, \tag{1}$$

with $\Delta B_n = B_n - B_{n-1}$, then we obviously get

$$B_n = (1 + r_n)B_{n-1}, \quad n \geq 1, \tag{2}$$

where r_n are $\mathscr{F}_{n-1}$-measurable and satisfy $r_n > -1$ (since $B_n > 0$ by assumption). In the financial literature r_n are called the (*bank*) *interest rates*.

The sequence $S = (S_n)_{n\geq 0}$ differs from $B = (B_n)_{n\geq 0}$ in that S_n is $\mathscr{F}_n$-measurable, in contrast to the $\mathscr{F}_{n-1}$-measurability of B_n. This reflects the situation with *stock prices*,whose actual value at time n becomes known only when it is announced (i.e., "today" rather than "yesterday" as for a bank account).

Similarly to the bank interest rate, we can define the *market interest rate*

$$\rho_n = \frac{\Delta S_n}{S_{n-1}}, \quad n \geq 1, \tag{3}$$

for stock $S = (S_n)_{n\geq 0}$.

Clearly, then

$$S_n = (1 + \rho_n)S_{n-1}, \tag{4}$$

with $\rho_n > -1$, since all $S_n > 0$ (by assumption).

It follows from (2) and (4) that

$$B_n = B_0 \prod_{k=1}^{n} (1 + r_k), \tag{5}$$

$$S_n = S_0 \prod_{k=1}^{n} (1 + \rho_k). \tag{6}$$

By definition, the pair of processes $B = (B_n)_{n\geq 0}$ and $S = (S_n)_{n\geq 0}$ introduced in the foregoing form a *financial* (B, S)*-market* consisting of two assets, the bank account B and the stock S.

Remark. It is clear that this (B, S)-market is merely a *simple* model of real financial markets, which usually consist of many assets of a diverse nature (e.g., [71]). Nevertheless, even this simple example demonstrates that the methods of *martingale theory* are very efficient in the treatment of many issues of a financial and economic nature (including, for example, the question about the absence of arbitrage in a (B, S)-market, which will be solved by the *first fundamental theorem*.)

3. Now we provide a definition of an *investment portfolio* and its *value* and define the important notion of a *self-financing portfolio*.

Let $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geq 0}, \mathsf{P})$ be a basic filtered probability space with $\mathscr{F}_0 = \{\varnothing, \Omega\}$, and let $\pi = (\beta, \gamma)$ be a pair of *predictable* sequences $\beta = (\beta_n)_{n\geq 0}, \gamma = (\gamma_n)_{n\geq 0}$. We impose no other restrictions on β_n and γ_n, $n \geq 0$, except that they are predictable, i.e., $\mathscr{F}_{n-1}$-measurable ($\mathscr{F}_{-1} = \mathscr{F}_0$). In particular, they can take fractional and negative values.

The meaning of β_n is the amount of "units" in a bank account, and that of γ_n is the amount of shares in an investor's possession at time n.

We will call $\pi = (\beta, \gamma)$ the *investment portfolio* in the (B, S)-market under consideration.

We associate with each portfolio $\pi = (\beta, \gamma)$ the corresponding value $X^\pi = (X_n^\pi)_{n\geq 0}$ by setting

$$X_n^\pi = \beta_n B_n + \gamma_n S_n \tag{7}$$

and interpreting $\beta_n B_n$ as the amount of money in the bank account and $\gamma_n S_n$ as the total price of the stock at time n. The intuitive meaning of the predictability of β and γ is also clear: the investment portfolio "for tomorrow" must be composed "today."

The following important notion of a self-financing portfolio expresses the idea of considering the (B,S)-markets that admit neither outflow nor inflow of capital. The formal definition is as follows.

Using the formula of discrete differentiation ($\Delta(a_n b_n) = a_n \Delta b_n + b_{n-1}\Delta a_n$), we find that the increment ΔX_n^π ($= X_n^\pi - X_{n-1}^\pi$) of the value is representable as

$$\Delta X_n^\pi = [\beta_n \Delta B_n + \gamma_n \Delta S_n] + [B_{n-1}\Delta\beta_n + S_{n-1}\Delta\gamma_n]. \tag{8}$$

The *real* change of the value may be caused only by market-based changes in the bank account and the stock price, related to the quantity $\beta_n \Delta B_n + \gamma_n \Delta S_n$. The second expression on the right-hand side of (8), i.e., $B_{n-1}\Delta\beta_n + S_{n-1}\Delta\gamma_n$, is $\mathscr{F}_{n-1}$-measurable and cannot affect X_{n-1}^π at time n. Therefore it must be equal to zero.

In general, the value can vary not only because of market-based changes in interest rates (r_n and ρ_n, $n \geq 1$) but also due to, say, inflow of capital from outside or outflow of capital for operating expenditures, and so on. We will not take into account such possibilities; in addition, we will consider (in accordance with the foregoing discussion) only portfolios $\pi = (\beta, \gamma)$ satisfying the condition

$$\Delta X_n^\pi = \beta_n \Delta B_n + \gamma_n \Delta S_n \tag{9}$$

for all $n \geq 1$.

In stochastic financial mathematics such portfolios are called *self-financing*.

4. It follows from (9) that a self-financing portfolio $\pi = (\beta, \gamma)$ satisfies

$$X_n^\pi = X_0^\pi + \sum_{k=1}^{n} (\beta_k \Delta B_k + \gamma_k \Delta S_k), \tag{10}$$

and since

$$\Delta\Big(\frac{X_n^\pi}{B_n}\Big) = \gamma_n \Delta\Big(\frac{S_n}{B_n}\Big), \tag{11}$$

we have

$$\frac{X_n^\pi}{B_n} = \frac{X_0^\pi}{B_0} + \sum_{k=1}^{n} \gamma_k \Delta\Big(\frac{S_k}{B_k}\Big). \tag{12}$$

Let us fix an $N \geq 1$ and consider the evolution of the (B,S)-market at times $n = 0, 1, \ldots, N$.

Definition 1. We say that a self-financing portfolio $\pi = (\beta, \gamma)$ provides an *arbitrage opportunity* at time N if $X_0^\pi = 0$, $X_N^\pi \geq 0$ (P-a.s.), and $X_N^\pi > 0$ with a positive P-probability, i.e., $\mathsf{P}\{X_N^\pi > 0\} > 0$.

Definition 2. We say that there is *no arbitrage* on the (B,S)-market (at time N), or that this market is *arbitrage-free* if, for any portfolio $\pi = (\beta, \gamma)$ with $X_0^\pi = 0$ and

$\mathsf{P}\{X_N^\pi \geq 0\} = 1$, it holds that $\mathsf{P}\{X_N^\pi = 0\} = 1$, i.e., the event $X_N^\pi > 0$ may occur only with zero P-probability.

The financial meaning of these definitions is that it is impossible to obtain any *risk-free* income in an *arbitrage-free* market.

Clearly, the property of a (B,S)-market to be arbitrage-free, and hence to be in a certain sense "fair" or "rational," depends on the probabilistic properties of the sequences $B = (B_n)_{n\leq N}$ and $S = (S_n)_{n\leq N}$, as well as on the assumptions regarding the structure of the filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\leq N}, \mathsf{P})$.

Remarkably, the theory of martingales enables us to effectively state conditions that guarantee the absence of arbitrage opportunities.

Theorem 1 (First Fundamental Theorem). *Assume that stochastic indeterminacy is described by a filtered probability space* $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\leq N}, \mathsf{P})$ *with* $\mathscr{F}_0 = \{\varnothing, \Omega\}$, $\mathscr{F}_N = \mathscr{F}$.

A (B,S)*-market defined on* $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\leq N}, \mathsf{P})$ *is arbitrage-free if and only if there exists a measure* $\widetilde{\mathsf{P}}$ *on* $(\Omega, \mathscr{F})$ *equivalent to* P ($\widetilde{\mathsf{P}} \sim \mathsf{P}$) *such that the discounted sequence* $\frac{S}{B} = \left(\frac{S_n}{B_n}\right)_{n\leq N}$ *is a martingale with respect to this measure, i.e.,*

$$\widetilde{\mathsf{E}}\left|\frac{S_n}{B_n}\right| < \infty, \quad n \leq N,$$

and

$$\widetilde{\mathsf{E}}\left(\frac{S_n}{B_n} \mid \mathscr{F}_{n-1}\right) = \frac{S_{n-1}}{B_{n-1}}, \quad n \leq N,$$

where $\widetilde{\mathsf{E}}$ *is the expectation with respect to* $\widetilde{\mathsf{P}}$.

Remark 1. The statement of the theorem remains valid also for vector processes $S = (S^1, \ldots, S^d)$ with $d < \infty$ [71, Chap. V, Sect. 2b].

Remark 2. For obvious reasons, the measure $\widetilde{\mathsf{P}}$ involved in the theorem is called the *martingale measure.*

Denote by $\mathsf{M}(\mathsf{P}) = \left\{\widetilde{\mathsf{P}} \sim \mathsf{P}\colon \frac{S}{B} \text{ is a } \widetilde{\mathsf{P}}\text{-martingale}\right\}$ the class of measures $\widetilde{\mathsf{P}}$, which are equivalent to P and such that the sequence $\frac{S}{B} = \left(\frac{S_n}{B_n}\right)_{n\leq N}$ is a *martingale* with respect to $\widetilde{\mathsf{P}}$.

We will write NA for the *absence of arbitrage* (no arbitrage). Using this notation the conclusion of Theorem 1 can be written

$$\mathsf{NA} \iff \mathsf{M}(\mathsf{P}) \neq \varnothing. \tag{13}$$

Proof of Theorem 1. *Sufficiency.* Let $\widetilde{\mathsf{P}} \in \mathsf{M}(\mathsf{P})$ be a martingale measure and $\pi = (\beta, \gamma)$ a portfolio with $X_0^\pi = \beta_0 B_0 + \gamma_0 S_0 = 0$. Then (12) implies

$$\frac{X_n^\pi}{B_n} = \sum_{k=1}^{n} \gamma_k \Delta\left(\frac{S_k}{B_k}\right), \quad 1 \leq n \leq N. \tag{14}$$

The sequence $\frac{S}{B} = \left(\frac{S_k}{B_k}\right)_{k\le N}$ is a $\widetilde{\mathsf{P}}$-martingale; therefore the sequence $G = (G_n^\pi)_{0\le n\le N}$ with $G_0^\pi = 0$ and $G_n^\pi = \sum_{k=1}^n \gamma_k \Delta\left(\frac{S_k}{B_k}\right)$, $1 \le n \le N$, is a *martingale transform*. Hence the sequence $\left(\frac{X_n^\pi}{B_n}\right)_{0\le n\le N}$ is also a martingale transform.

When testing for arbitrage or its absence, we must consider portfolios π such that not only $X_0^\pi = 0$, but also $X_N^\pi \ge 0$ (P-a.s.). Since $\widetilde{\mathsf{P}} \sim \mathsf{P}$ and $B_N > 0$ (P- and $\widetilde{\mathsf{P}}$-a.s.), we obtain that $\widetilde{\mathsf{P}}\left\{\frac{X_N^\pi}{B_N} \ge 0\right\} = 1$.

Then, applying Theorem 3 in Sect. 1 to the martingale transform $\left(\frac{X_n^\pi}{B_n}\right)_{0\le n\le N}$, we obtain that this sequence is in fact a $\widetilde{\mathsf{P}}$-martingale. Thus, $\widetilde{\mathsf{E}}\frac{X_N^\pi}{B_N} = \widetilde{\mathsf{E}}\frac{X_0^\pi}{B_0} = 0$, and since $\widetilde{\mathsf{P}}\left\{\frac{X_N^\pi}{B_N} \ge 0\right\} = 1$, we have $\widetilde{\mathsf{P}}\left\{\frac{X_N^\pi}{B_N} = 0\right\} = 1$.

Hence we see that $X_N^\pi = 0$ ($\widetilde{\mathsf{P}}$- and P-a.s.), and therefore $X_N^\pi = 0$ (P-a.s.) for any self-financing portfolio π with $X_0^\pi = 0$ and $X_N^\pi \ge 0$ (P-a.s.), which by definition means the absence of arbitrage opportunities.

Necessity. We will give the proof only for the one-step model of a (B, S)-market, i.e., for $N = 1$. But even this simple case will enable us to demonstrate the idea of the proof, which consists in an explicit construction of a martingale measure using the absence of arbitrage. We will construct this measure using the *Esscher transform* (see subsequent discussion). (For the proof in the general case $N \ge 1$ see [71, Chapter V, Sect. 2d].)

Without loss of generality we can assume that $B_0 = B_1 = 1$. In the current setup, the *absence of arbitrage opportunities* reduces (Problem 1) to the condition

$$\mathsf{P}\{\Delta S_1 > 0\} > 0 \quad \text{and} \quad \mathsf{P}\{\Delta S_1 < 0\} > 0. \tag{15}$$

(We exclude the trivial case $\mathsf{P}\{\Delta S_1 = 0\} = 1$.)

We must derive from this that *there exists* an equivalent martingale measure $\widetilde{\mathsf{P}}$, i.e., such that $\widetilde{\mathsf{P}} \sim \mathsf{P}$ and $\widetilde{\mathsf{E}}|\Delta S_1| < \infty$, $\widetilde{\mathsf{E}}\Delta S_1 = 0$.

This immediately follows from the following lemma, which is also of interest in its own right for probability theory.

Lemma 1. *Let $(\Omega, \mathscr{F}) = (R, \mathscr{B}(R))$, and let $X = X(\omega)$ be the coordinate random variable $(X(\omega) = \omega)$. Let P be a probability measure on $(\Omega, \mathscr{F})$ such that*

$$\mathsf{P}\{X > 0\} > 0 \quad \textit{and} \quad \mathsf{P}\{X < 0\} > 0. \tag{16}$$

Then for any real a there exists a probability measure $\widetilde{\mathsf{P}} \sim \mathsf{P}$ on $(\Omega, \mathscr{F})$ such that

$$\widetilde{\mathsf{E}} e^{aX} < \infty. \tag{17}$$

In particular, $\widetilde{\mathsf{E}}|X| < \infty$ and, moreover,

$$\widetilde{\mathsf{E}} X = 0. \tag{18}$$

PROOF. Define the measure $\mathsf{Q} = \mathsf{Q}(dx)$ with $\mathsf{Q}(dx) = ce^{-x^2}\,\mathsf{P}(dx)$ and normalizing constant $c = (\mathsf{E}\,e^{-X^2})^{-1}$.

For any real a, set

$$\varphi(a) = \mathsf{E}_{\mathsf{Q}}\,e^{aX}, \tag{19}$$

where E_{Q} is the expectation related to Q.

Let

$$Z_a(x) = \frac{e^{ax}}{\varphi(a)}. \tag{20}$$

Since $Z_a(x) > 0$ and $\mathsf{E}_{\mathsf{Q}}\,Z_a(X) = 1$, the measure $\widetilde{\mathsf{P}}_a$ with

$$\widetilde{\mathsf{P}}_a(dx) = Z_a(x)\,\mathsf{Q}(dx) \tag{21}$$

is a probability measure for any real a. Clearly, $\widetilde{\mathsf{P}}_a \sim \mathsf{Q} \sim \mathsf{P}$.

□

Remark 3. The transformation $x \rightsquigarrow \frac{e^{ax}}{\varphi(a)}$ is known as the *Esscher transform*. As we will see later, the measure $\widetilde{\mathsf{P}} = \widetilde{\mathsf{P}}_{a_*}$ for a certain value a_* possesses the martingale property (18). This measure is referred to as the *Esscher measure* or the *martingale Esscher measure*.

Now we return to the proof of Theorem 1. The function $\varphi = \varphi(a)$ defined for all real a is strictly convex, since $\varphi''(a) > 0$. Let $\varphi_* = \inf\{\varphi(a)\colon a \in R\}$. The following two cases are possible: (i) there exists a_* such that $\varphi(a_*) = \varphi_*$, and (ii) there is no such (finite) a_*.

In the first case, $\varphi'(a_*) = 0$. Therefore

$$\mathsf{E}_{\widetilde{\mathsf{P}}_{a_*}} X = \mathsf{E}_{\mathsf{Q}}\,\frac{Xe^{a_*X}}{\varphi(a_*)} = \frac{\varphi'(a_*)}{\varphi(a_*)} = 0,$$

and we can take the measure $\widetilde{\mathsf{P}}_{a_*}$ for the required measure $\widetilde{\mathsf{P}}$.

So far we have not used the no-arbitrage assumption (16). It is not hard to show (Problem 2) that this assumption excludes possibility (ii). Therefore there remains only the first possibility, which has already been considered.

Thus, we have proved the necessity part (which consists in the existence of a martingale measure) for $N = 1$. For the general case $N \geq 1$ the reader is referred, as stated earlier, to [71, Chap. V, Sect. 2d].

□

5. Now we give some examples of arbitrage-free (B, S)-markets.

EXAMPLE 1. Suppose that the (B, S)-market is described by (5) and (6) with $1 \leq k \leq N$, where $r_k = r$ (a constant) for all $1 \leq k \leq N$ and $\rho = (\rho_1, \rho_2, \ldots, \rho_N)$ is a sequence of independent identically distributed Bernoulli random variables taking

values a and b ($a < b$) with probabilities $\mathsf{P}\{\rho_1 = a\} = q$, $\mathsf{P}\{\rho_1 = b\} = p$, $p + q = 1$, $0 < p < 1$. Moreover, assume that

$$-1 < a < r < b. \tag{22}$$

This model of a (B, S)-market is known as the *CRR model*, after the names of its authors J. C. Cox, R. A. Ross, and M. Rubinstein; for more details see [71].

Since in this model

$$\frac{S_n}{B_n} = \Big(\frac{1+\rho_n}{1+r}\Big)\frac{S_{n-1}}{B_{n-1}},$$

it is clear that the martingale measure $\widetilde{\mathsf{P}}$ must satisfy

$$\widetilde{\mathsf{E}}\frac{1+\rho_n}{1+r} = 1,$$

i.e., $\widetilde{\mathsf{E}}\rho_n = r$.

Use the notation $\tilde{p} = \widetilde{\mathsf{P}}\{\rho_n = b\}$, $\tilde{q} = \widetilde{\mathsf{P}}\{\rho_n = a\}$; then for any $n \geq 1$

$$\tilde{p} + \tilde{q} = 1, \quad b\tilde{p} + a\tilde{q} = r.$$

Hence

$$\tilde{p} = \frac{r-a}{b-a}, \quad \tilde{q} = \frac{b-r}{b-a}. \tag{23}$$

In this case the whole "randomness" is determined by the Bernoulli sequence $\rho = (\rho_1, \rho_2, \ldots, \rho_N)$. We let $\Omega = \{a, b\}^N$, i.e., we assume that the space of elementary outcomes consists of sequences $(x_1, \ldots, x_N)$ with $x_i = a$ or b. (Assuming this specific "coordinate" structure of Ω does not restrict generality; in this connection, see the end of the proof of sufficiency in Theorem 2, Subsection 6.)

As an exercise (Problem 2) we suggest showing that the measure $\widetilde{\mathsf{P}}$ defined by

$$\widetilde{\mathsf{P}}(x_1, \ldots, x_N) = \tilde{p}^{\nu_b(x_1,\ldots,x_N)}\tilde{q}^{N-\nu_b(x_1,\ldots,x_N)}, \tag{24}$$

where $\nu_b(x_1, \ldots, x_N) = \sum_{i=1}^N I_b(x_i)$ (the number of x_i's equal to b) is a *martingale measure*, and this measure is *unique*. It is clear from (24) that $\widetilde{\mathsf{P}}\{\rho_n = b\} = \tilde{p}$ and $\widetilde{\mathsf{P}}\{\rho_n = a\} = \tilde{q}$.

Thus, by Theorem 1, the *CRR model* is an example of an *arbitrage-free* market.

EXAMPLE 2. Let the (B, S)-market have the structure $B_n = 1$ for all $n = 0, 1, \ldots, N$ and

$$S_n = S_0 \exp\left(\sum_{k=1}^n \hat{\rho}_k\right), \quad 1 \leq n \leq N. \tag{25}$$

Let $\hat{\rho}_k = \mu_k + \sigma_k\varepsilon_k$, where μ_k and $\sigma_k > 0$ are $\mathscr{F}_{k-1}$-measurable and $(\varepsilon_1, \ldots, \varepsilon_N)$ are independent standard Gaussian random variables, $\varepsilon_k \sim \mathscr{N}(0, 1)$.

We will construct the required Esscher measure $\widetilde{\mathsf{P}}$ (on $(\Omega, \mathscr{F}_N)$) by means of the *conditional Esscher transform*. That is, let $\widetilde{\mathsf{P}}(d\omega) = Z_N(\omega)\,\mathsf{P}(d\omega)$, where $Z_N(\omega) = \prod_{1\le k\le N} z_k(\omega)$ with

$$z_k(\omega) = \frac{e^{a_k\hat{\rho}_k}}{\mathsf{E}(e^{a_k\hat{\rho}_k} \mid \mathscr{F}_{k-1})} \tag{26}$$

($\mathscr{F}_0 = \{\varnothing, \Omega\}$) and where the $\mathscr{F}_{k-1}$-measurable random variables $a_k = a_k(\omega)$ are to be chosen so that the sequence $(S_n)_{0\le n\le N}$ is a $\widetilde{\mathsf{P}}$-martingale.

In view of (25), $\widetilde{\mathsf{P}}$ is a martingale measure if and only if

$$\mathsf{E}[e^{(a_n+1)\hat{\rho}_n} \mid \mathscr{F}_{n-1}] = \mathsf{E}[e^{a_n\hat{\rho}_n} \mid \mathscr{F}_{n-1}], \quad 1 \le n \le N \tag{27}$$

(with respect to the *initial* measure P).

Since $\hat{\rho}_n = \mu_n + \sigma_n\varepsilon_n$, we find from (27) that a_n must be chosen so that

$$\mu_n + \frac{\sigma_n^2}{2} = -a_n\sigma_n^2,$$

i.e.,

$$a_n = -\frac{\mu_n}{\sigma_n^2} - \frac{1}{2}.$$

With this choice of a_n, $1 \le n \le N$, the density $Z_N(\omega)$ is given by the formula

$$Z_N(\omega) = \exp\Big\{-\sum_{n=1}^{N}\Big[\Big(\frac{\mu_n}{\sigma_n} + \frac{\sigma_n}{2}\Big)\varepsilon_n + \frac{1}{2}\Big(\frac{\mu_n}{\sigma_n} + \frac{\sigma_n}{2}\Big)^2\Big]\Big\}. \tag{28}$$

If $\mu_n = -\frac{\sigma_n^2}{2}$ for all $1 \le n \le N$ from the outset, then $\widetilde{\mathsf{P}} = \mathsf{P}$. In other words, in this case the initial measure P itself is a martingale measure.

Thus the (B,S)-market with $B = (B_n)_{0\le n\le N}$ such that $B_n \equiv 1$ and $S = (S_n)_{0\le n\le N}$ as specified by (25), is, as in Example 1, arbitrage-free. In Problem 4, we propose to examine whether the martingale measure $\widetilde{\mathsf{P}}$ constructed earlier is *unique*.

6. The notion of a *complete* (B,S)-market to be introduced in what follows is of great interest to stochastic financial mathematics because (irrespective of whether or not the market is arbitrage-free) it is related to the natural question of whether, for a given $\mathscr{F}_N$-measurable contingent claim f_N, there is a self-financing portfolio π such that the corresponding capital X_N^π "*offsets*" (or is at least no less than) f_N.

Definition 3. A (B,S)-market is said to be *complete* (relative to time instant N) or *N-complete* if any *bounded* $\mathscr{F}_N$-measurable contingent claim f_N is *replicable*, i.e., there exists a self-financing portfolio π such that $X_N^\pi = f_N$ (P-a.s.).

Theorem 2 (Second Fundamental Theorem). *Similarly to Theorem* 1, *we assume that* $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{0\le n\le N}, \mathsf{P})$ *is a filtered probability space*, $\mathscr{F}_0 = \{\varnothing, \Omega\}$, $\mathscr{F}_N = \mathscr{F}$,

and the (B,S)-market defined on this space is arbitrage-free ($\mathsf{M}(\mathsf{P}) \neq \varnothing$). *Then this market is complete if and only if there exists only a unique equivalent martingale measure* ($|\mathsf{M}(\mathsf{P})| = 1$).

PROOF. *Necessity.* Let the market at hand be complete. This means that for any $\mathscr{F}_N$-measurable contingent claim f_N there is a self-financing portfolio $\pi = (\beta,\gamma)$ such that $X_N^\pi = f_N$ (P-a.s.). Without loss of generality we may assume that $B_n = 1$, $0 \leq n \leq N$. Hence we see from (10) that

$$f_N = X_N^\pi = X_0^\pi + \sum_{k=1}^{N} \gamma_k \Delta S_k. \tag{29}$$

Since the market is arbitrage-free by assumption, the set of martingale measures is nonempty, $\mathsf{M}(\mathsf{P}) \neq \varnothing$. We will show that the completeness assumption implies the *uniqueness* of the martingale measure ($|\mathsf{M}(\mathsf{P})| = 1$).

Let P^1 and P^2 be two martingale measures. Then $\left(\sum_{k=1}^n \gamma_k \Delta S_k\right)_{1\leq n\leq N}$ is a martingale transform with respect to either of these measures.

Take a set $A \in \mathscr{F}_N$, and let $f_N(\omega) = I_A(\omega)$. Since for some π

$$I_A(\omega) = X_N^\pi = X_0^\pi + \sum_{k=1}^{N} \gamma_k \Delta S_k \quad (\mathsf{P}\text{-a.s.}),$$

we conclude from Theorem 3 in Sect. 1 that the sequence $\left(\sum_{k=1}^n \gamma_k \Delta S_k\right)_{1\leq n\leq N}$ is a martingale with respect to either of the measures P^1 and P^2. Therefore

$$\mathsf{E}_{\mathsf{P}^i} I_A(\omega) = x, \quad i = 1,2, \tag{30}$$

where $\mathsf{E}_{\mathsf{P}^i}$ is the expectation with respect to P^i and $x = X_0^\pi$, which is a constant since $\mathscr{F}_0 = \{\varnothing,\Omega\}$. Now (30) implies that $\mathsf{P}^1(A) = \mathsf{P}^2(A)$ for any set $A \in \mathscr{F}_N$. Hence the uniqueness of the martingale measure is established.

The proof of *sufficiency* is more complicated and will be carried out in several steps. We consider an arbitrage-free (B,S)-market ($\mathsf{M}(\mathsf{P}) \neq \varnothing$) such that the martingale measure is unique ($|\mathsf{M}(\mathsf{P})| = 1$).

It is worth noting that both assumptions of the *uniqueness* of the martingale measure and the *completeness* of the market are *strong restrictions*. What is more, it turns out that these assumptions imply that the trajectories $S = (S_n)_{0\leq n\leq N}$ are "conditionally two-pointed," which will be explained subsequently. (This may be exemplified by the CRR model $\Delta S_n = \rho_n S_{n-1}$, where ρ_n takes only two values, so that the conditional probabilities $\mathsf{P}(\Delta S_n \in \cdot \mid \mathscr{F}_{n-1})$ "sit" on two points, aS_{n-1} and bS_{n-1}.)

The uniqueness of the martingale measure ($|\mathsf{M}(\mathsf{P})| = 1$) also imposes restrictions on the *structure* of the filtration $(\mathscr{F}_n)_{n\leq N}$. Under this condition the σ-algebras $\mathscr{F}_n$ must be generated by the prices $S_0, S_1, \ldots, S_n$ (assuming that $B_k \equiv 1$, $k \leq n$). In this regard, see the diagram on p. 610 of [71] and Chap. V, Sect. 4e, therein.

As an intermediate result for establishing the implication "$|\mathsf{M}(\mathsf{P})| = 1 \Rightarrow$ completeness" we will prove the following useful lemma, which provides an equivalent characterization of *completeness* of an *arbitrage-free* market.

Lemma 2. *An arbitrage-free (B,S)-market is complete if and only if there exists a measure $\widetilde{\mathsf{P}}$ in the set $\mathsf{M}(\mathsf{P})$ of all martingale measures such that any bounded martingale $m = (m_n, \mathscr{F}_n, \widetilde{\mathsf{P}})_{0\le n\le N}$ admits an "$\frac{S}{B}$-representation":*

$$m_n = m_0 + \sum_{k=1}^{n} \gamma_k^* \Delta\Big(\frac{S_k}{B_k}\Big) \tag{31}$$

with predictable γ_k^*, $1 \le k \le n$.

PROOF. We consider an arbitrage-free complete (B,S)-market. (Without loss of generality, assume that $B_n = 1$, $0 \le n \le N$.)

Take an arbitrary measure $\widetilde{\mathsf{P}} \in \mathsf{M}(\mathsf{P})$, and let $m = (m_n, \mathscr{F}_n, \widetilde{\mathsf{P}})_{0\le n\le N}$ be a bounded martingale ($|m_n| \le c$, $0 \le n \le N$). Set $f_N = m_N$. Then, by the definition of completeness (Definition 3), there is a portfolio $\pi^* = (\beta^*, \gamma^*)$ such that $X_N^{\pi^*} = f_N$ and for all $0 \le n \le N$

$$X_n^{\pi^*} = x + \sum_{k=1}^{n} \gamma_k^* \Delta S_k, \tag{32}$$

with $x = X_0^{\pi^*}$.

Since $X_N^{\pi^*} = f_N \le c$, the sequence $X^{\pi^*} = (X_n^{\pi^*}, \mathscr{F}_n, \widetilde{\mathsf{P}})_{0\le n\le N}$ is a martingale (Theorem 3, Sect. 1). Thus, we have two martingales, m and X^{π^*}, with the same *terminal* value f_N ($X_N^{\pi^*} = m_N = f_N$). But by the definition of the martingale property, $m_n = \mathsf{E}(m_N \,|\, \mathscr{F}_n)$ and $X_n^{\pi^*} = \mathsf{E}(X_N^{\pi^*} \,|\, \mathscr{F}_n)$, $0 \le n \le N$. Therefore the *Lévy martingales* m and X^{π^*} are the same, and by (32) the martingale $m = (m_n, \mathscr{F}_n, \widetilde{\mathsf{P}})_{0\le n\le N}$ admits the "S-representation"

$$m_n = x + \sum_{k=1}^{n} \gamma_k^* \Delta S_k, \quad 1 \le n \le N, \tag{33}$$

with $x = m_0$.

Let us now prove the reverse statememt (S-representation $\Rightarrow$ completeness).

By assumption, there exists a measure $\widetilde{\mathsf{P}} \in \mathsf{M}(\mathsf{P})$ such that any bounded $\widetilde{\mathsf{P}}$-martingale admits an S-representation.

Take for such a martingale $X = (X_n, \mathscr{F}_n, \widetilde{\mathsf{P}})_{0\le n\le N}$ a martingale with $X_n = \widetilde{\mathsf{E}}(f_N \,|\, \mathscr{F}_n)$, where $\widetilde{\mathsf{E}}$ is the expectation with respect to $\widetilde{\mathsf{P}}$ and f_N is the contingent claim involved in Definition 3, for which we must find a self-financing portfolio π such that $X_N^{\pi} = f_N$ ($\widetilde{\mathsf{P}}$- and P-a.s.).

For a (bounded) martingale $X = (X_n, \mathscr{F}_n, \mathsf{P})_{0\le n\le N}$ consider its S-representation

$$X_n = X_0 + \sum_{k=1}^{n} \gamma_k \Delta S_k \tag{34}$$

with some $\mathscr{F}_{k-1}$-measurable variables γ_k.

Let us show that this implies the existence of a self-financing portfolio $\tilde\pi = (\tilde\beta, \tilde\gamma)$ such that $X_n^{\tilde\pi} = X_n$ for all $0 \le n \le N$ and, in particular, $f_N = X_N = X_N^{\tilde\pi}$ admits the representation

$$f_N = X_0^{\tilde\pi} + \sum_{k=1}^{N} \tilde\gamma_k \Delta S_k, \tag{35}$$

as required in Definition 3.

Using representation (34), set $\tilde\gamma_n = \gamma_n$ and define

$$\tilde\beta_n = X_n - \gamma_n S_n. \tag{36}$$

Then (34) implies that $\tilde\beta_n$ are $\mathscr{F}_{n-1}$-measurable. Moreover,

$$\begin{aligned} S_{n-1}\Delta\tilde\gamma_n + \Delta\tilde\beta_n &= S_{n-1}\Delta\gamma_n + \Delta X_n - \Delta(\gamma_n S_n) \\ &= S_{n-1}\Delta\gamma_n + \gamma_n \Delta S_n - \Delta(\gamma_n S_n) = 0. \end{aligned}$$

Thus, according to Subsection 3, the portfolio $\tilde\pi = (\tilde\beta, \tilde\gamma)$ so constructed is *self-financing* and $X_N^{\tilde\pi} = f_N$, i.e., the *completeness* property is fulfilled.

□

With this lemma, we see that to complete the proof of the theorem, we must establish the implication $\{3\}$ in the following chain of implications:

$$\boxed{|\,\mathsf{M}(\mathsf{P})| = 1} \overset{\{3\}}{\Longrightarrow} \boxed{S\text{-representation}} \overset{\{2\}}{\Longleftrightarrow} \boxed{\text{completeness}} \overset{\{1\}}{\Longrightarrow} \boxed{|\,\mathsf{M}(\mathsf{P})| = 1}.$$

(Implication $\{1\}$ was established in the proof of necessity and implication $\{2\}$ in the foregoing lemma.)

To make the proof of $\{3\}$ more transparent, we will consider the particular case of a (B, S)-market described by the *CRR model*.

As was pointed out earlier (Example 1), in this model the martingale measure $\widetilde{\mathsf{P}}$ is unique ($|\,\mathsf{M}(\mathsf{P})| = 1$). So we need to understand why in this case the S-representation (with respect to the martingale measure $\widetilde{\mathsf{P}}$) holds. We have already indicated that the key reason for that is the fact that the ρ_n in (4) take only two values, a and b, and therefore the conditional distributions $\mathsf{P}(\Delta S_n \in \cdot \mid \mathscr{F}_{n-1})$ are two-pointed.

Thus we will consider the CRR model introduced in Example 1 and assume additionally that $\mathscr{F}_n = \sigma(\rho_1, \dots, \rho_n)$ for $1 \le n \le N$ and $\mathscr{F}_0 = \{\varnothing, \Omega\}$. Let $\widetilde{\mathsf{P}}$ denote the martingale measure on $(\Omega, \mathscr{F}_N)$ defined by (24).

Let $X = (X_n, \mathscr{F}_n, \widetilde{\mathsf{P}})_{0\le n\le N}$ be a bounded martingale. Then there are functions $g_n = g_n(x_1, \dots, x_n)$ such that $X_n(\omega) = g_n(\rho_1(\omega), \dots, \rho_n(\omega))$, so that

$$\Delta X_n = g_n(\rho_1, \dots, \rho_n) - g_{n-1}(\rho_1, \dots, \rho_{n-1}).$$

Since $\widetilde{\mathsf{E}}(\Delta X_n \mid \mathscr{F}_{n-1}) = 0$, we have

$$\tilde{p}\, g_n(\rho_1, \dots, \rho_{n-1}, b) + \tilde{q}\, g_n(\rho_1, \dots, \rho_{n-1}, a) = g_{n-1}(\rho_1, \dots, \rho_{n-1}),$$

i.e.,

$$\frac{g_n(\rho_1, \dots, \rho_{n-1}, b) - g_{n-1}(\rho_1, \dots, \rho_{n-1})}{\tilde{q}} = \frac{g_{n-1}(\rho_1, \dots, \rho_{n-1}) - g_n(\rho_1, \dots, \rho_{n-1}, a)}{\tilde{p}}. \quad (37)$$

Since $\tilde{p} = \frac{r-a}{b-a}$, $\tilde{q} = \frac{b-r}{b-a}$, we find from (37) that

$$\frac{g_n(\rho_1, \dots, \rho_{n-1}, b) - g_{n-1}(\rho_1, \dots, \rho_{n-1})}{b-r} = \frac{g_n(\rho_1, \dots, \rho_{n-1}, a) - g_{n-1}(\rho_1, \dots, \rho_{n-1})}{a-r}. \quad (38)$$

Let $\mu_n(\{a\}; \omega) = I(\rho_n(\omega) = a)$, $\mu_n(\{b\}; \omega) = I(\rho_n(\omega) = b)$, and let

$$W_n(\omega, x) = g_n(\rho_1(\omega), \dots, \rho_{n-1}(\omega), x) - g_{n-1}(\rho_1(\omega), \dots, \rho_{n-1}(\omega)),$$
$$W_n^*(\omega, x) = \frac{W_n(\omega, x)}{x - r}.$$

Using this notation we obtain

$$\Delta X_n(\omega) = W_n(\omega, \rho_n(\omega)) = \int W_n(\omega, x)\, \mu_n(dx; \omega) = \int (x-r) W_n^*(\omega, x)\, \mu_n(dx; \omega).$$

By (38) the functions $W_n^*(\omega, x)$ *do not depend* on x. Therefore denoting the expression in the left-hand side (or, equivalently, in the right-hand side) of (38) by $\gamma_n^*(\omega)$ we find that

$$\Delta X_n(\omega) = \gamma_n^*(\omega)\, (\rho_n(\omega) - r). \quad (39)$$

Therefore

$$X_n(\omega) = X_0(\omega) + \sum_{k=1}^{n} \gamma_k^*(\omega)\, (\rho_k(\omega) - r). \quad (40)$$

It is easily seen that

$$\Delta\Big(\frac{S_n}{B_n}\Big) = \frac{S_{n-1}}{B_{n-1}} \cdot \frac{\rho_n - r}{1 + r}.$$

Hence

$$\rho_n - r = (1+r)\frac{B_{n-1}}{S_{n-1}}\Delta\Big(\frac{S_n}{B_n}\Big),$$

and consequently we see from (40) that

$$X_n(\omega) = X_0(\omega) + \sum_{k=1}^{n}\gamma_k(\omega)\Delta\Big(\frac{S_k(\omega)}{B_k}\Big), \tag{41}$$

where

$$\gamma_k(\omega) = \gamma_k^*(\omega)\,(1+r)\frac{B_{k-1}}{S_{k-1}}.$$

The sequence $\frac{S}{B} = \left(\frac{S_n}{B_n}\right)_{0\le n\le N}$ is a martingale with respect to $\widetilde{\mathsf{P}}$. Thus, (41) is simply the "$\frac{S}{B}$-representation" for X with respect to the (basic) $\widetilde{\mathsf{P}}$-martingale $\frac{S}{B}$.

The key argument in the proof of $\{3\}$ for the CRR model (where $|\mathsf{M}(\mathsf{P})| = 1$) was the fact that the ρ_n take on only *two* values. However, it turns out that the uniqueness assumption of the martingale measure $\widetilde{\mathsf{P}}$ is so strong that in the general case it also implies that the variables $\rho_n = \frac{\Delta S_n}{S_{n-1}}$ are "two-pointed," i.e., there exist predictable $a_n = a_n(\omega)$ and $b_n = b_n(\omega)$ such that

$$\widetilde{\mathsf{P}}(\rho_n = a_n \,|\, \mathscr{F}_{n-1}) + \widetilde{\mathsf{P}}(\rho_n = b_n \,|\, \mathscr{F}_{n-1}) = 1 \quad (\mathsf{P}\text{-a.s.}). \tag{42}$$

Taking for granted this property, the foregoing proof of the $\frac{S}{B}$-representation in the CRR model will "work" also in the general case. Thus, all that remains is to establish (42). We leave obtaining this result to the reader (Problem 5). Nevertheless, we give some heuristic arguments showing how the *uniqueness* of the martingale measure leads to two-pointed conditional distributions.

Let $\mathsf{Q} = \mathsf{Q}(dx)$ be a probability distribution on $(R, \mathscr{B}(R))$ and $\xi = \xi(x)$ the coordinate random variable ($\xi(x) = x$). Let $\mathsf{E}_{\mathsf{Q}}|\xi| < \infty$, $\mathsf{E}_{\mathsf{Q}}\,\xi = 0$ ("martingale property"), and the measure Q has the property that for any other measure $\widetilde{\mathsf{Q}}$ such that $\mathsf{E}_{\widetilde{\mathsf{Q}}}|\xi| < \infty$ and $\mathsf{E}_{\widetilde{\mathsf{Q}}}\,\xi = 0$, it holds that $\widetilde{\mathsf{Q}} = \mathsf{Q}$ ("uniqueness of the martingale measure").

We assert that in this case Q is supported on at most two points ($a \le 0$ and $b \ge 0$) that may stick together as one zero point ($a = b = 0$).

The aforementioned heuristic arguments, which make this assertion very likely, are as follows.

Suppose that Q is supported on three points, $x_- \le x_0 \le x_+$, with masses q_-, q_0, q_+, respectively. The condition $\mathsf{E}_{\mathsf{Q}}\,\xi = 0$ means that

$$q_- x_- + q_0 x_0 + q_+ x_+ = 0.$$

If $x_0 = 0$, then $q_- x_- + q_+ x_+ = 0$.

Let

$$\tilde{q}_- = \frac{q_-}{2}, \quad \tilde{q}_0 = \frac{1}{2} + \frac{q_0}{2}, \quad \tilde{q}_+ = \frac{q_+}{2}, \tag{43}$$

i.e., we move some parts of masses q_- and q_+ from the points x_- and x_+ to x_0.

It is seen from (43) that the corresponding measure $\widetilde{\mathsf{Q}} \sim \mathsf{Q}$ and $\mathsf{E}_{\widetilde{\mathsf{Q}}}\,\xi = 0$, although $\widetilde{\mathsf{Q}} \neq \mathsf{Q}$. But this contradicts the *uniqueness* assumption of measure Q such that $\mathsf{E}_{\mathsf{Q}}\,\xi = 0$.

Therefore the measure Q cannot be supported at three points (x_-, x_0, x_+) with $x_0 = 0$. In a similar way, utilizing the same idea of "moving masses," the case $x_0 \neq 0$ is treated. (For more details, see [71, Chap. 5, Sect. 4e].)

7. Problems.

1. Show that for $N = 1$ the no-arbitrage condition is equivalent to inequalities (15). (It is assumed that $\mathsf{P}\{\Delta S_1 = 0\} < 1$.)
2. Show that possibility (ii) in the proof of Lemma 1 (Subsection 4) is excluded by conditions (16).
3. Prove that the measure $\widetilde{\mathsf{P}}$ in Example 1 (Subsection 5) is a martingale measure, which is unique in the class $\mathsf{M}(\mathsf{P})$.
4. Explore the problem of uniqueness of the martingale measure constructed in Example 2 (Subsection 5).
5. Prove that the assumption $|\,\mathsf{M}(\mathsf{P})| = 1$ in the (B, S)-model implies the "conditional two-pointedness" for the distribution of $\frac{S_n}{B_n}$, $1 \leq n \leq N$.

12. Hedging in Arbitrage-Free Models

1. *Hedging* is one of the basic methods of the dynamic control of investment portfolios. We will set out some basic notions and results related to this method considering as an example the pricing of so-called *option contracts* (or simply *options*).

Options (as instruments of financial engineering), being derivative securities, are fairly *risky*. But at the same time they (along with other securities, e.g., *forwards*) are successfully used not only for earning profit due to market price fluctuations but also for protection (*hedging*) against unexpected changes in stock prices.

An *option* is a security (contract) issued by a financial institution that gives its holder the *right* to buy or sell something valuable (e.g., a share, a bond, currency) at a certain period or instant of time on specified terms.

Whereas an option gives the *right* to buy or sell something, the other financial instrument, a *forward contract* (or a forward), is a commitment to buy or sell something of value at a certain time in the future at a price fixed at the moment of signing the deal.

One of the main questions regarding option pricing concerns the *price* at which the options are to be sold. Clearly, the seller wants to charge as much as possible, while the buyer wants to pay as little as possible. What is the fair, rational price, acceptable to both buyer and seller?

Naturally, this fair price must be "reasonable." That is, the buyer must realize that a lower price for the option may put the seller in a position where he is unable to meet the obligations fixed by the agreement because of insufficient payment.

At the same time, the amount of this payment should not give the seller arbitrage possibilities of a "free lunch" type, i.e., the chance to earn a risk-free profit.

Before defining what the fair price of an option should mean, we give the commonly accepted classification of options.

2. We will consider a (B, S)-market, $B = (B_n)_{0\le n\le N}$, $S = (S_n)_{0\le n\le N}$, operating at time instants $n = 0, 1, \dots, N$ and defined on a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{0\le n\le N}, \mathsf{P})$ with $\mathscr{F}_0 = \{\varnothing, \Omega\}$ and $\mathscr{F}_N = \mathscr{F}$.

We will consider options written on stock with prices described by the random sequence $S = (S_n)_{0\le n\le N}$.

With regard to the time of their *exercise*, options are of two types: *European* and *American*.

If an option can be exercised only at the time instant N fixed in the contract, then N is called the time of its exercise, and this option is said to be of the *European type*.

Alternatively, if an option can be exercised at any Markov time (or stopping time; see Definition 3 in Sect. 1) $\tau = \tau(\omega)$, taking values in the range $\{0, 1, \dots, N\}$ specified by the contract, then this option is of the *American type*.

According to the generally adopted terminology, there are two types of options:

1. *buyer's options* (*call options*) and
2. *seller's options* (*put options*).

The difference between these types is that call options grant the *right to buy*, while put options grant the *right to sell*.

For definiteness, we consider examples of *standard* options of the *European type*.

These options are characterized by two constants: N, the time of exercise, and K, the price (fixed by the contract) at which a certain asset (say, a share) can be bought (buyer's options) or for which it can be sold (seller's options).

In the case of a buyer's option, the buyer buys at time 0 from the seller an option at price C. This option stipulates that the buyer can buy from the seller at time N the share for price K. Let S_0 and S_N be the market prices of the share at times 0 and N. If $S_N > K$, then the buyer can sell it right away for price S_N, thereby earning a profit $S_N - K$. Otherwise, if $S_N < K$, there is no sense in exercising the right to buy for price K since the buyer can buy the share at the lower market price S_N.

Thus, combining these two cases, we find that for buyer's options, the buyer at time N earns the profit

$$f_N = (S_N - K)^+, \tag{1}$$

where $a^+ = \max(a, 0)$. The buyer's net return is equal to this quantity minus the amount C that he has paid to the seller at time 0 (the negative value $-C$ in the case $S_N < K$ indicates a loss C).

In a similar way, the profit of the buyer of a put option is given by the formula

$$f_N = (K - S_N)^+. \tag{2}$$

3. When defining the fair price in an arbitrage-free (B, S)-market we must distinguish between two cases, *complete* and *incomplete* markets.

Definition 1. Let a (B,S)-market be *arbitrage-free* and *complete*. The fair price of an option of the European type with $\mathscr{F}_N$-measurable bounded (nonnegative) contingent claim f_N is the price of *perfect hedging*,

$$\mathbb{C}(f_N;\mathsf{P}) = \inf\{x\colon \exists\pi \text{ with } X_0^\pi = x \text{ and } X_N^\pi = f_N\ (\mathsf{P}\text{-a.s.})\}. \tag{3}$$

A portfolio π is called a *hedge* of the contingent claim f_N if $X_N^\pi \geq f_N$ with probability 1.

It follows from the results of Sect. 11 that in the case of complete arbitrage-free markets, for any bounded contingent claim there exists a perfect hedging π, i.e., such that $X_N^\pi = f_N$ (P-a.s.). This is why in definition (3) we consider a (nonempty) class of portfolios with the property $X_N^\pi = f_N$ (P-a.s.).

The following definition is natural for *incomplete* arbitrage-free markets.

Definition 2. Let a (B,S)-market be *arbitrage-free*. The fair price of an option of the European type with $\mathscr{F}_N$-measurable bounded (nonnegative) contingent claim f_N is the *superhedging* price

$$\mathbb{C}(f_N;\mathsf{P}) = \inf\{x\colon \exists\pi \text{ with } X_0^\pi = x \text{ and } X_N^\pi \geq f_N\ (\mathsf{P}\text{-a.s.})\}. \tag{4}$$

Note that this definition is correct, i.e., for any bounded function f_N there always exists a portfolio π with some initial capital x such that $X_N^\pi \geq f_N$ (P-a.s.).

4. Now we give a formula for the price $\mathbb{C}(f_N;\mathsf{P})$. We will prove it for complete markets and refer the reader to specialized literature for incomplete markets (e.g., [71, Chap. VI, Sect. 1c]).

Theorem 1. (i) *For a complete arbitrage-free (B,S)-market, the fair price of a European-type option with a contingent claim f_N is*

$$\mathbb{C}(f_N;\mathsf{P}) = B_0\,\mathsf{E}_{\widetilde{\mathsf{P}}}\frac{f_N}{B_N}, \tag{5}$$

where $\mathsf{E}_{\widetilde{\mathsf{P}}}$ is the expectation with respect to the (unique) martingale measure $\widetilde{\mathsf{P}}$.

(ii) *For a general arbitrage-free (B,S)-market, the fair price of a European-type option with a contingent claim f_N is*

$$\mathbb{C}(f_N;\mathsf{P}) = \sup_{\widetilde{\mathsf{P}}\in\mathsf{M}(\mathsf{P})} B_0\,\mathsf{E}_{\widetilde{\mathsf{P}}}\frac{f_N}{B_N}, \tag{6}$$

where the sup *is taken over the set of all martingale measures $\mathsf{M}(\mathsf{P})$.*

PROOF. (i) Let π be a perfect hedge with $X_0^\pi = x$ and $X_N^\pi = f_N$ (P-a.s.). Then (see (15) in Sect. 11)

$$\frac{f_N}{B_N} = \frac{X_N^\pi}{B_N} = \frac{x}{B_0} + \sum_{k=1}^{N}\gamma_k\Delta\Big(\frac{S_k}{B_k}\Big). \tag{7}$$

Hence, by Theorem 3 from Sect. 1,

$$\mathsf{E}_{\widetilde{\mathsf{P}}} \frac{f_N}{B_N} = \frac{x}{B_0}, \tag{8}$$

since the martingale transform $\left(\frac{x}{B_0} + \sum_{k=1}^{n} \gamma_k \Delta\left(\frac{S_k}{B_k}\right)\right)_{1 \le n \le N}$ is such that at the terminal time N

$$\frac{x}{B_0} + \sum_{k=1}^{N} \gamma_k \Delta\left(\frac{S_k}{B_k}\right) = \frac{f_N}{B_N} \ge 0. \tag{9}$$

Note that the left-hand side of (8) *does not depend* on the structure of a particular hedge π with initial value $X_0^\pi = x$. If we take another hedge π' with initial value $X_0^{\pi'}$, then, according to (8), this initial value is again equal to $B_0 \mathsf{E}_{\widetilde{\mathsf{P}}} \frac{f_N}{B_N}$. Hence it is clear that the initial value x is *the same* for *all* perfect hedges, which proves (5).

(ii) Here we only prove the inequality

$$\sup_{\widetilde{\mathsf{P}} \in \mathsf{M}(\mathsf{P})} B_0 \mathsf{E}_{\widetilde{\mathsf{P}}} \frac{f_N}{B_N} \le \mathbb{C}(f_N; \mathsf{P}). \tag{10}$$

(The proof of the reverse inequality relies on the so-called optional decomposition, which goes beyond the scope of this book; see, e.g., [71, Chap. VI, Sects. 1c and 2d].)

Suppose that the hedge π is such that $X_0^\pi = x$ and $X_N^\pi \ge f_N$ (P-a.s.).

Then (7) implies that

$$\frac{x}{B_0} + \sum_{k=1}^{N} \gamma_k \Delta\left(\frac{S_k}{B_k}\right) \ge \frac{f_N}{B_N} \ge 0.$$

Therefore, for *any* measure $\widetilde{\mathsf{P}} \in \mathsf{M}(\mathsf{P})$,

$$B_0 \mathsf{E}_{\widetilde{\mathsf{P}}} \frac{f_N}{B_N} \le x$$

(cf. (8) and (9)). Hence, taking the supremum on the left-hand side over *all* measures $\widetilde{\mathsf{P}} \in \mathsf{M}(\mathsf{P})$, we arrive at the required inequality (10).

□

5. Now we consider some definitions and results related to options of the *American type*. For these options we must assume that we are given not a single contingent claim f_N related to time N, but a *collection* of claims $f_0, f_1, \ldots, f_N$ whose meaning is that once the buyer exercises the option at time n, the payoff (by the option seller to the buyer) is determined by the ($\mathscr{F}_n$-mesurable) function $f_n = f_n(\omega)$.

If the buyer of an option decides to exercise the option at time $\tau = \tau(\omega)$, which is a Markov time with values in $\{0, 1, \ldots, N\}$, then the payoff is $f_{\tau(\omega)}(\omega)$. Therefore the seller of the option when composing his portfolio π must envisage that for any τ the following *hedging* condition must hold: $X_\tau^\pi \ge f_\tau$ (P-a.s.).

This explains the expedience of the following definition.

Definition 3. Let a (B,S)-market be arbitrage-free. The fair price of an option of the American type with the system $f=(f_n)_{0\le n\le N}$ of $\mathscr{F}_n$-measurable nonnegative payoff functions f_n is the *upper superhedging price*, i.e., the price

$$\overline{\mathbb{C}}(f;\mathsf{P})=\inf\{x\colon \exists\pi \text{ with } X_0^\pi=x \text{ and } X_n^\pi\ge f_n\ (\mathsf{P}\text{-a.s.}),\ 0\le n\le N\}. \tag{11}$$

We state (without proof) an analog of Theorem 1 for American-type options.

Theorem 2. (i) *For a complete arbitrage-free (B,S)-market, the fair price of an American-type option with a system of payoff functions $f=(f_n)_{0\le n\le N}$ is given by*

$$\overline{\mathbb{C}}(f;\mathsf{P})=\sup_{\tau\in\mathfrak{M}_0^N} B_0\,\mathsf{E}_{\widetilde{\mathsf{P}}}\frac{f_\tau}{B_\tau}, \tag{12}$$

where $\mathfrak{M}_0^N=\{\tau\colon \tau\le N\}$ is the class of stopping times (with respect to $(\mathscr{F}_n)_{0\le n\le N}$) and $\widetilde{\mathsf{P}}$ is the unique martingale measure.

(ii) *In the general case of an (incomplete) arbitrage-free (B,S)-market, the fair price of an American-type option with a system of payoff functions $f=(f_n)_{0\le n\le N}$ is given by*

$$\overline{\mathbb{C}}(f;\mathsf{P})=\sup_{\tau\in\mathfrak{M}_0^N,\ \widetilde{\mathsf{P}}\in\mathsf{M}(\mathsf{P})} B_0\,\mathsf{E}_{\widetilde{\mathsf{P}}}\frac{f_\tau}{B_\tau}, \tag{13}$$

where $\mathsf{M}(\mathsf{P})$ is the set of martingale measures $\widetilde{\mathsf{P}}$.

For the proof, see [71, Chap. VI, Sect. 2c].

6. The foregoing theorems enable one to determine the fair price of an option. Another important question is how the seller of an option should compose the hedging portfolio π^* having received the premium $\mathbb{C}(f_N;\mathsf{P})$ or $\overline{\mathbb{C}}(f;\mathsf{P})$.

For simplicity, we restrict ourselves to the case of a *complete* (B,S)-market of European-type options.

Theorem 3. *Consider an arbitrage-free complete (B,S)-market. There exists a self-financing portfolio $\pi^*=(\beta^*,\gamma^*)$ with initial capital $X_0^{\pi^*}=\mathbb{C}(f_N;\mathsf{P})$ implementing the perfect hedging of the terminal payoff f_N:*

$$X_N^{\pi^*}=f_N\quad(\mathsf{P}\text{-}a.s.).$$

The dynamics of capital $X_n^{\pi^*}=\beta_n^*B_n+\gamma_n^*S_n$, $0\le n\le N$, is determined by

$$X_n^{\pi^*}=B_n\,\mathsf{E}_{\widetilde{\mathsf{P}}}\Big(\frac{f_N}{B_N}\mid\mathscr{F}_n\Big). \tag{14}$$

The component $\gamma^*=(\gamma_n^*)_{0\le n\le N}$ of the hedge $\pi^*=(\beta^*,\gamma^*)$ is obtained from $X^{\pi^*}=(X_n^{\pi^*})_{0\le n\le N}$ by the formula

$$\Delta\Big(\frac{X_n^{\pi^*}}{B_n}\Big) = \gamma_n^* \Delta\Big(\frac{S_n}{B_n}\Big) \tag{15}$$

and the component $\beta^* = (\beta_n^*)_{0\le n\le N}$ *by the formula*

$$X_n^{\pi^*} = \beta_n^* B_n + \gamma_n^* S_n. \tag{16}$$

The *proof* follows directly from the proof of the implication "completeness" $\Rightarrow$ "$\frac{S}{B}$-represenatation" in Lemma 2, Sect. 11, applied to the martingale $m = (m_n)_{0\le n\le N}$ with $m_n = \mathsf{E}_{\widetilde{\mathsf{P}}}\Big(\frac{f_N}{B_N} \mid \mathscr{F}_n\Big)$.

7. As an example of actual option pricing consider a (B, S)-market described by the CRR *model*,

$$\begin{aligned} B_n &= B_{n-1}(1+r), \\ S_n &= S_{n-1}(1+\rho_n), \end{aligned} \tag{17}$$

where $\rho_1, \dots, \rho_N$ are independent identically distributed random variables taking two values, a and b, $-1 < a < r < b$.

This market is arbitrage-free and complete (Problem 3, Sect. 11) with martingale measure $\widetilde{\mathsf{P}}$ such that $\widetilde{\mathsf{P}}\{\rho_n = b\} = \tilde{p}$, $\widetilde{\mathsf{P}}\{\rho_n = a\} = \tilde{q}$, where

$$\tilde{p} = \frac{r-a}{b-a}, \quad \tilde{q} = \frac{b-r}{b-a}. \tag{18}$$

(See Example 1 in Sect. 11, Subsection 5.)

By formula (5) of Theorem 1, the fair price for this (B, S)-market is

$$\mathbb{C}(f_N; \mathsf{P}) = \mathsf{E}_{\widetilde{\mathsf{P}}} \frac{f_N}{(1+r)^N}. \tag{19}$$

And, according to Theorem 3, to compute the perfect hedging portfolio $\pi^* = (\beta^*, \gamma^*)$, we must first calculate

$$X_n^{\pi^*} = \mathsf{E}_{\widetilde{\mathsf{P}}}\Big(\frac{f_N}{(1+r)^N} \mid \mathscr{F}_n\Big) \tag{20}$$

(with $\mathscr{F}_n = \sigma(\rho_1, \dots, \rho_n)$, $1 \le n \le N$, and $\mathscr{F}_0 = \{\varnothing, \Omega\}$) and then to find γ_n^* and β_n^* by (15) and (16).

Since $X_0^{\pi^*} = \mathbb{C}(f_N; \mathsf{P})$, the problem amounts to finding the conditional expectations on the right-hand side of (20) for $n = 0, 1, \dots, N$.

We will assume that the $\mathscr{F}_N$-measurable function f_N has a "Markov" structure, i.e., $f_N = f(S_N)$, where $f = f(x)$ is a nonnegative function of $x \ge 0$.

Use the notation

$$F_n(x; p) = \sum_{k=0}^{n} f\big(x(1+b)^k(1+a)^{n-k}\big)\, C_n^k p^k (1-p)^{n-k}. \tag{21}$$

Taking into account that

$$\prod_{n<k\le N}(1+\rho_k) = (1+b)^{\Delta_N-\Delta_n}(1+a)^{(N-n)-(\Delta_N-\Delta_n)},$$

where $\Delta_n = \delta_1 + \cdots + \delta_n$, $\delta_k = (\rho_k - a)/(b-a)$, we obtain

$$\mathsf{E}_{\tilde{\mathsf{P}}} f\left(x\prod_{n<k\le N}(1+\rho_k)\right) = F_{N-n}(x;\tilde{p}), \tag{22}$$

with $\tilde{p} = (r-a)/(b-a)$.

Using that $S_N = S_n\prod_{n<k\le N}(1+\rho_k)$, (21) and (20) imply that

$$X_n^{\pi^*} = \mathsf{E}_{\tilde{\mathsf{P}}}\Big(\frac{f_N}{(1+r)^N}\,|\,\mathscr{F}_n\Big) = (1+r)^{-N}F_{N-n}(S_n;\tilde{p}). \tag{23}$$

In particular,

$$\mathbb{C}(f_N;\mathsf{P}) = X_0^{\pi^*} = (1+r)^{-N}F_N(S_0;\tilde{p}). \tag{24}$$

Finally, taking into account (23), we obtain from (15) that

$$\gamma_n^* = \Delta\Big(\frac{X_n^{\pi^*}}{B_n}\Big)\Big/\Delta\Big(\frac{S_n}{B_n}\Big)$$

is given by

$$\gamma_n^* = (1+r)^{-(N-n)}\frac{F_{N-n}(S_{n-1}(1+b);\tilde{p}) - F_{N-n}(S_{n-1}(1+a);\tilde{p})}{S_{n-1}(b-a)}. \tag{25}$$

To find β_n^*, note that $B_{n-1}\Delta\beta_n^* + S_{n-1}\Delta\gamma_n^* = 0$ by the self-financing condition. Therefore

$$X_{n-1}^{\pi^*} = \beta_n^* B_{n-1} + \gamma_n^* S_{n-1}, \tag{26}$$

and consequently,

$$\beta_n^* = \frac{X_{n-1}^{\pi^*} - \gamma_n^* S_{n-1}}{B_{n-1}}. \tag{27}$$

Using this formula along with (23) and (25) we obtain

$$\beta_n^* = \frac{1}{B_N}\Big\{F_{N-n+1}(S_{n-1};\tilde{p}) - \frac{1+r}{1+b}[F_{N-n}(S_{n-1}(1+b);\tilde{p}) - F_{N-n}(S_{n-1}(1+a);\tilde{p})]\Big\}. \tag{28}$$

Let us see, finally, what the fair price $\mathbb{C}(f_N;\mathsf{P})$ is in the case of a *standard* buyer's (call) option when $f_N = (S_N - K)^+$.

Let $K_0 = K_0(a, b, N; s_0/K)$ be the smallest integer for which

$$S_0(1+a)^N\Big(\frac{1+b}{1+a}\Big)^{K_0} > K, \tag{29}$$

i.e., let

$$K_0 = 1 + \Big[\log\frac{K}{S_0(1+a)^N} \Big/ \log\frac{1+b}{1+a}\Big], \tag{30}$$

where $[x]$ is the integral part of x.

Using the notation

$$p^* = \frac{1+b}{1+r}\tilde{p},$$

where $\tilde{p} = (r-a)/(b-a)$, and

$$\mathbb{B}(K_0, N; p) = \sum_{k=K_0}^{N} C_N^k p^k (1-p)^{N-k}, \tag{31}$$

it is not hard to derive from (24) the following formula (Cox–Ross–Rubinstein) for the fair price (denoted presently by $\mathbb{C}_N$) of the *standard call option*:

$$\mathbb{C}_N = S_0\mathbb{B}(K_0, N; p^*) - K(1+r)^{-N}\mathbb{B}(K_0, N; \tilde{p}). \tag{32}$$

If $K_0 > N$, then $\mathbb{C}_N = 0$.

Remark. Since

$$(K - S_N)^+ = (S_N - K)^+ - S_N + K,$$

the fair price of a *standard* seller's (put) option denoted by $\mathbb{P}_N$ $(= \mathbb{C}(f_N; \mathsf{P})$ with $f_N = (K - S_N)^+)$ is given by

$$\mathbb{P}_N = \widetilde{\mathsf{E}}(1+r)^{-N}(K - S_N)^+ = \mathbb{C}_N - \widetilde{\mathsf{E}}(1+r)^{-N}S_N + K(1+r)^{-N}.$$

Since $\widetilde{\mathsf{E}}(1+r)^{-N}S_N = S_0$, we obviously have the following *identity* (the *call–put parity*):

$$\mathbb{P}_N = \mathbb{C}_N - S_0 + K(1+r)^{-N}. \tag{33}$$

8. Problems

1. Find the price $\mathbb{C}(f_N; \mathsf{P})$ of a standard call option with $f_N = (S_N - K)^+$ for the model of the (B, S)-market considered in Example 2, Subsection 5, Sect. 11.
2. Try to prove the reverse inequality in (10).
3. Prove (12), and try to prove (13).
4. Give a detailed derivation of (23).
5. Prove (25) and (28).
6. Give a detailed derivation of (32).

13. Optimal Stopping Problems: Martingale Approach

1. We have already encountered an optimal stopping problem when we dealt with the fair price of an American-type option. That is, formula (12) in Sect. 12 shows that, to find this price, we must (under the simplified conditions $B_n = 1, 0 \le n \le N$, and $\widetilde{\mathsf{P}} = \mathsf{P}$) determine the quantity (also called a "price")

$$V_0^N = \sup_{\tau \in \mathfrak{M}_0^N} \mathsf{E} f_\tau, \tag{1}$$

where $f = (f_0, f_1, \dots, f_N)$ is a sequence of $\mathscr{F}_n$-measurable nonnegative functions and $\tau = \tau(\omega)$ are Markov times (or stopping times) of class $\mathfrak{M}_0^N$ consisting of random variables $\tau = \tau(\omega)$ taking values $\{0, 1, \dots, N\}$ and such that for any n in this set

$$\{\omega \colon \tau(\omega) = n\} \in \mathscr{F}_n. \tag{2}$$

(In this section we assume given a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n \ge 0}, \mathsf{P})$ with $\mathscr{F}_0 = \{\varnothing, \Omega\}$.)

Along with problem (1), where the times $\tau = \tau(\omega)$ belong to $\mathfrak{M}_0^N$, the problem of finding the price

$$V_0^\infty = \sup_{\tau \in \mathfrak{M}_0^\infty} \mathsf{E} f_\tau \tag{3}$$

is also of interest, where $\mathfrak{M}_0^\infty = \{\tau \colon \tau < \infty\}$ and $f = (f_0, f_1, \dots)$ is a stochastic sequence of $\mathscr{F}_n$-measurable random variables f_n, $n \ge 0$, with $\mathsf{E}\,|f_\tau| < \infty$.

In both cases (1) and (3), the problem is not only in finding the prices V_0^N and V_0^∞, but also in determining the *optimal times* (provided they exist) when the supremum is attained.

In many problems it makes sense to consider also infinite Markov times (taking also the value $+\infty$). In this case when dealing with $\mathsf{E} f_\tau$ we should agree what we mean by f_∞. One natural way is to take $\limsup_n f_n$ for f_∞. Another convention when admitting infinite values for τ is to define the price as

$$\overline{V}_0^\infty = \sup_{\tau \in \overline{\mathfrak{M}}_0^\infty} \mathsf{E} f_\tau I(\tau < \infty), \tag{4}$$

where $\overline{\mathfrak{M}}_0^\infty = \{\tau \colon \tau \le \infty\}$ is the class of all Markov times. Obviously, $\overline{V}_0^\infty = \sup_{\tau \in \overline{\mathfrak{M}}_0^\infty} \mathsf{E} f_\tau$ when letting $f_\infty = 0$ (cf. Sect. 1, Subsection 3).

In what follows we will only treat problem (1). (Regarding the case $N = \infty$, see Sect. 9, Chap. 8.) If the probabilistic structure of the sequence $f = (f_0, f_1, \dots, f_N)$ is not specified, the most efficient method of solving problems (1) and (3) is the "martingale" method described in what follows. (We will always assume without mention that $\mathsf{E}\,|f_n| < \infty$ for all $n \le N$.)

2. Thus, let $N < \infty$. This case may be treated by what is known as backward induction, which is carried out here as follows.

Along with V_0^N, define the "prices"

$$V_n^N = \sup_{\tau \in \mathfrak{M}_n^N} \mathsf{E} f_\tau, \tag{5}$$

where $\mathfrak{M}_n^N = \{\tau : n \le \tau \le N\}$ is the class of stopping times such that $n \le \tau(\omega) \le N$ for all $\omega \in \Omega$.

Moreover, define inductively the *stochastic sequence* $v^N = (v_n^N)_{0 \le n \le N}$ as follows:

$$v_N^N = f_N, \quad v_n^N = \max(f_n, \mathsf{E}(v_{n+1}^N \mid \mathscr{F}_n)) \tag{6}$$

for $n = N-1, \dots, 0$.

For $0 \le n \le N$, define

$$\tau_n^N = \min\{n \le k \le N : f_k = v_k^N\}. \tag{7}$$

Using this notation, the following theorem completely describes the solution of the optimal stopping problems (1) and (5).

Theorem 1. *Let $f = (f_0, f_1, \dots, f_N)$ be such that every f_n is $\mathscr{F}_n$-measurable.*

(i) *For any n, $0 \le n \le N$, the stopping time*

$$\tau_n^N = \min\{n \le k \le N : v_k^N = f_k\} \tag{8}$$

is optimal within the class $\mathfrak{M}_n^N$:

$$\mathsf{E} f_{\tau_n^N} = \sup_{\tau \in \mathfrak{M}_n^N} \mathsf{E} f_\tau \quad (= V_n^N). \tag{9}$$

(ii) *The stopping times τ_n^N, $0 \le n \le N$, are optimal also in the following "conditional" sense*:

$$\mathsf{E}(f_{\tau_n^N} \mid \mathscr{F}_n) = \operatorname*{ess\,sup}_{\tau \in \mathfrak{M}_n^N} \mathsf{E}(f_\tau \mid \mathscr{F}_n) \quad (\mathsf{P}\text{-}a.s). \tag{10}$$

The "stochastic prices" $\operatorname{ess\,sup}_{\tau \in \mathfrak{M}_n^N} \mathsf{E}(f_\tau \mid \mathscr{F}_n)$ *are equal to* v_n^N:

$$\operatorname*{ess\,sup}_{\tau \in \mathfrak{M}_n^N} \mathsf{E}(f_\tau \mid \mathscr{F}_n) = v_n^N \quad (\mathsf{P}\text{-}a.s.) \tag{11}$$

and

$$V_n^N = \mathsf{E}\, v_n^N. \tag{12}$$

If $n = 0$, then

$$V_0^N = v_0^N. \tag{13}$$

For $n = N$,

$$V_N^N = \mathsf{E} f_N. \tag{14}$$

3. Before we proceed to the proof, let us recall the definition of the *essential supremum* $\operatorname{ess\,sup}_{\alpha \in \mathfrak{A}} \xi_\alpha(\omega)$ of a family of $\mathscr{F}$-measurable random variables $\{\xi_\alpha(\omega), \alpha \in \mathfrak{A}\}$ involved in (10).

We need this concept because in the case of an *uncountable* set $\mathfrak{A}$ the use of the ordinary $\sup_{\alpha\in\mathfrak{A}}\xi_\alpha(\omega)$ may, in general, give rise to functions (of $\omega\in\Omega$) that are not $\mathscr{F}$-measurable.

Indeed, for any $c\in R$

$$\Big\{\omega\colon \sup_{\alpha\in\mathfrak{A}}\xi_\alpha(\omega)\le c\Big\}=\bigcap_{\alpha\in\mathfrak{A}}\{\omega\colon \xi_\alpha(\omega)\le c\}.$$

Here, the sets $A_\alpha=\{\omega\colon \xi_\alpha(\omega)\le c\}$ belong to $\mathscr{F}$ (i.e., they are *events*). However, since the set $\mathfrak{A}$ is uncountable, we are not guaranteed that $\bigcap_{\alpha\in\mathfrak{A}}A_\alpha\in\mathscr{F}$.

Definition. Let $\{\xi_\alpha(\omega),\alpha\in\mathfrak{A}\}$ be a family of random variables (i.e., of $\mathscr{F}$-measurable functions taking values in $(-\infty,+\infty)$). An extended *random variable* $\xi(\omega)$ (an $\mathscr{F}$-measurable function with values in $(-\infty,+\infty]$) is said to be the essential supremum of the family $\{\xi_\alpha(\omega),\alpha\in\mathfrak{A}\}$ (denoted $\xi(\omega)=\operatorname{ess\,sup}_{\alpha\in\mathfrak{A}}\xi_\alpha(\omega)$) if

(a) $\xi(\omega)\ge\xi_\alpha(\omega)$ (P-a.s.) for all $\alpha\in\mathfrak{A}$;

(b) For any (extended) random variable $\eta(\omega)$ such that $\eta(\omega)\ge\xi_\alpha(\omega)$ (P-a.s.) for all $\alpha\in\mathfrak{A}$ we have $\xi(\omega)\le\eta(\omega)$ (P-a.s.).

In other words, $\xi(\omega)$ is the *smallest* among all (extended) random variables majorizing $\xi_\alpha(\omega)$ for all $\alpha\in\mathfrak{A}$.

Of course, we must prove first of all that this definition is meaningful. This is done by the following statement.

Lemma. *For any family* $\{\xi_\alpha(\omega),\alpha\in\mathfrak{A}\}$ *of random variables there exists a random variable* (*in general, extended*) $\xi(\omega)$ (*denoted by* $\operatorname{ess\,sup}_{\alpha\in\mathfrak{A}}\xi_\alpha(\omega)$) *with properties* (a) *and* (b) *as in the definition.*

There is a countable subset $\mathfrak{A}_0\subseteq\mathfrak{A}$ *with the property that this variable can be taken to be*

$$\xi(\omega)=\sup_{\alpha\in\mathfrak{A}_0}\xi_\alpha(\omega).$$

PROOF. First, assume that all $\xi_\alpha(\omega)$, $\alpha\in\mathfrak{A}$, are uniformly bounded ($|\xi_\alpha(\omega)|\le c$, $\omega\in\Omega$, $\alpha\in\mathfrak{A}$).

Let A be a *finite* set of indices $\alpha\in\mathfrak{A}$. Set $S(A)=\mathsf{E}\big(\max_{\alpha\in A}\xi_\alpha(\omega)\big)$. Let, further, $S=\sup S(A)$, where the supremum is taken over all finite subsets $A\subseteq\mathfrak{A}$.

Denote by A_n, $n\ge1$, a finite set such that

$$\mathsf{E}\Big(\max_{\alpha\in A_n}\xi_\alpha(\omega)\Big)\ge S-\frac{1}{n}.$$

Let $\mathfrak{A}_0=\bigcup_{n\ge1}A_n$. This set is countable, hence

$$\xi(\omega)=\sup_{\alpha\in\mathfrak{A}_0}\xi_\alpha(\omega)$$

is $\mathscr{F}$-measurable, i.e., this is a random variable. (Note that $|\xi(\omega)|\le c$, hence $\xi(\omega)$ is an *ordinary* rather than extended random variable.)

This construction of $\xi(\omega)$ implies (Problem 1) that this random variable has properties (a) and (b) of the foregoing definition.

Therefore we have established the existence of the essential supremum for a uniformly bounded family $\{\xi_\alpha(\omega), \alpha \in \mathfrak{A}\}$.

In the general case we first go from $\xi_\alpha(\omega)$ to the bounded random variables $\tilde{\xi}_\alpha(\omega) = \arctan \xi_\alpha(\omega)$, for which $|\tilde{\xi}_\alpha(\omega)| \le \pi/2$, $\alpha \in \mathfrak{A}$, $\omega \in \Omega$, and then we let $\tilde{\xi}(\omega) = \operatorname{ess\,sup}_{\alpha\in\mathfrak{A}} \tilde{\xi}_\alpha(\omega)$.

Then the random variable $\xi(\omega) = \tan \tilde{\xi}(\omega)$ will satisfy requirements (a) and (b) of the definition of the essential supremum (Problem 3).

4. PROOF OF THEOREM 1. Let us fix an index N. To simplify writing, it will often be omitted.

If $n = N$, then $v_N = f_N$ and $\tau_N = N$, and properties (9)–(12) and (14) are obvious. Now we will argue inductively.

Suppose that the theorem is established for $n = N, N-1, \dots, k$. Let us show that it holds then for $n = k-1$.

Let $\tau \in \mathfrak{M}_{k-1}$ $(= \mathfrak{M}_{k-1}^N)$ and $A \in \mathscr{F}_{k-1}$. Define the stopping time $\overline{\tau} \in \mathfrak{M}_k$ by letting $\overline{\tau} = \max(\tau, k)$. Since $\overline{\tau} \in \mathfrak{M}_k$ and the event $\{\tau \ge k\} \in \mathscr{F}_{k-1}$, we find that

$$\begin{aligned}
\mathsf{E}[I_A f_\tau] &= \mathsf{E}[I_{A\cap\{\tau=k-1\}} f_\tau] + \mathsf{E}[I_{A\cap\{\tau\ge k\}} f_\tau] \\
&= \mathsf{E}[I_{A\cap\{\tau=k-1\}} f_\tau] + \mathsf{E}[I_{A\cap\{\tau\ge k\}}\, \mathsf{E}(f_\tau \mid \mathscr{F}_{k-1})] \\
&= \mathsf{E}[I_{A\cap\{\tau=k-1\}} f_\tau] + \mathsf{E}[I_{A\cap\{\tau\ge k\}}\, \mathsf{E}(\mathsf{E}(f_{\overline{\tau}} \mid \mathscr{F}_k) \mid \mathscr{F}_{k-1})] \\
&\le \mathsf{E}[I_{A\cap\{\tau=k-1\}} f_{k-1}] + \mathsf{E}[I_{A\cap\{\tau\ge k\}}\, \mathsf{E}(v_k \mid \mathscr{F}_{k-1})] \le \mathsf{E}[I_A v_{k-1}]. \qquad (15)
\end{aligned}$$

In view of the $\mathscr{F}_{k-1}$-measurability of the set A, this implies that for any $\tau \in \mathfrak{M}_{k-1}$

$$\mathsf{E}(f_\tau \mid \mathscr{F}_{k-1}) \le v_{k-1} \quad (\mathsf{P}\text{-a.s.}). \qquad (16)$$

We will show now that for the Markov time τ_{k-1}

$$\mathsf{E}(f_{\tau_{k-1}} \mid \mathscr{F}_{k-1}) = v_{k-1}, \qquad (17)$$

with P-probability 1. (If this equality is established, we obtain by (16) that (10) and (11) hold also for $n = k-1$.)

For that purpose it suffices to show that (15) holds for $\tau = \tau_{k-1}$ with equality rather than inequality signs throughout.

Beginning as in (15) and using then that on the set $\{\tau_{k-1} \ge k\}$ we have $\tau = \tau_k$ by definition (5) and that (by the induction assumption) $\mathsf{E}(f_{\tau_k} \mid \mathscr{F}_k) = v_k$ (P-a.s.), we obtain

$$\begin{aligned}
\mathsf{E}[I_A f_{\tau_{k-1}}] &= \mathsf{E}[I_{A\cap\{\tau_{k-1}=k-1\}} f_{k-1}] + \mathsf{E}[I_{A\cap\{\tau_{k-1}\ge k\}}\, \mathsf{E}(f_{\tau_{k-1}} \mid \mathscr{F}_{k-1})] \\
&= \mathsf{E}[I_{A\cap\{\tau_{k-1}=k-1\}} f_{k-1}] + \mathsf{E}[I_{A\cap\{\tau_{k-1}\ge k\}}\, \mathsf{E}(f_{\tau_k} \mid \mathscr{F}_{k-1})] \\
&= \mathsf{E}[I_{A\cap\{\tau_{k-1}=k-1\}} f_{k-1}] + \mathsf{E}[I_{A\cap\{\tau_{k-1}\ge k\}}\, \mathsf{E}(v_k \mid \mathscr{F}_{k-1})] = \mathsf{E}[I_A v_{k-1}],
\end{aligned}$$

where the last equality holds because $v_{k-1} = \max(f_{k-1}, \mathsf{E}(v_k \,|\, \mathscr{F}_{k-1}))$ by definition, hence $v_{k-1} = f_{k-1}$ on the set $\{\tau_{k-1} = k-1\}$ and $v_{k-1} > f_{k-1}$ on the set $\{\tau_{k-1} > k-1\} = \{\tau_{k-1} \geq k\}$ (so that on this set $v_{k-1} = \mathsf{E}(v_k \,|\, \mathscr{F}_{k-1})$).

Thus (17) is established. As was pointed out earlier, this property, together with (16), implies that (10) and (11) hold.

It follows from these relations that (P-a.s.)

$$v_n = \mathsf{E}(f_{\tau_n} \,|\, \mathscr{F}_n) \geq \mathsf{E}(f_\tau \,|\, \mathscr{F}_n) \tag{18}$$

for any $\tau \in \mathfrak{M}_n$ $(= \mathfrak{M}_n^N)$. Therefore, taking into account the convention $v_n^N = v_n$, we find that

$$\mathsf{E}\, v_n^N = \mathsf{E} f_{\tau_n} \geq \sup_{\tau \in \mathfrak{M}_n^N} \mathsf{E} f_\tau = V_n^N, \tag{19}$$

which proves (9) and (12).

Property (13) is a particular case of (12) (for $n = 0$) due to the fact that v_0^N is a constant by (11) and since the σ-algebra $\mathscr{F}_0$ $(= \{\varnothing, \Omega\})$ is trivial. Finally, (14) follows from definition (5) (for $n = N$). □

5. To clarify the "martingale" nature of the optimal problem at hand, consider the recurrence relations (6) for the sequence $v^N = (v_0^N, v_1^N, \ldots, v_N^N)$ with the "boundary" condition $v_N^N = f_N$.

We see from (6) that for every $n = 0, 1, \ldots, N-1$ (P-a.s.)

$$v_n^N \geq f_n, \tag{20}$$

$$v_n^N \geq \mathsf{E}(v_{n+1}^N \,|\, \mathscr{F}_n). \tag{21}$$

The first inequality here means that the sequence v^N *majorizes* the sequence $f = (f_0, f_1, \ldots, f_N)$. The second inequality shows that v^N is a *supermartingale* with "terminal" value $v_N^N = f_N$. Thus, we can say that $v^N = (v_0^N, v_1^N, \ldots, v_N^N)$ with v_n^N's defined by (6) or by (11), is a *supermartingale majorant* for the sequence $f = (f_0, f_1, \ldots, f_N)$.

In other words, this means that the sequence v^N belongs to the class of sequences $\gamma^N = (\gamma_0^N, \gamma_1^N, \ldots, \gamma_N^N)$ with $\gamma_N^N \geq f_N$ satisfying (P-a.s.) the "variational inequalities"

$$\gamma_n^N \geq \max(f_n, \mathsf{E}(\gamma_{n+1}^N \,|\, \mathscr{F}_n)) \tag{22}$$

for all $n = 0, 1, \ldots, N-1$.

But the sequence v^N possesses additionally the property that (22) holds for v^N not only with *nonstrict inequality* "$\geq$" but with *equality* "$=$" (see (6)). This property singles out the sequence v^N among sequences γ^N (with $\gamma_N^N \geq f_N$) as follows.

Theorem 2. *The sequence v^N is the least supermartingale majorant for the sequence $f = (f_0, f_1, \ldots, f_N)$.*

PROOF. Since $v_N^N = f_N$ and $\gamma_N^N \geq f_N$, we have $\gamma_N^N \geq v_N^N$. Together with (22) and (6) this implies that (P-a.s.)

$$\gamma_{N-1}^N \geq \max(f_{N-1}, \mathsf{E}(\gamma_N^N \,|\, \mathscr{F}_{N-1})) \geq \max(f_{N-1}, \mathsf{E}(v_N^N \,|\, \mathscr{F}_{N-1})) = v_{N-1}^N.$$

In a similar way we find that $\gamma_n^N \geq v_n^N$ (P-a.s.) also for all $n < N-1$.
□

Remark. The result of this theorem can be restated as follows: The solution $v^N = (v_0^N, v_1^N, \ldots, v_N^N)$ of the recurrence system

$$v_n^N = \max(f_n, \mathsf{E}(v_{n+1}^N \mid \mathscr{F}_n)), \quad n < N,$$

with $v_N^N = f_N$, is the smallest among solutions $\gamma^N = (\gamma_0^N, \gamma_1^N, \ldots, \gamma_N^N)$ of the recurrence system of inequalities

$$\gamma_n^N \geq \max(f_n, \mathsf{E}(\gamma_{n+1}^N \mid \mathscr{F}_n)), \quad n < N, \tag{23}$$

with $\gamma_N^N \geq f_N$.

6. Theorems 1 and 2 not only describe the method of finding the price $V_0^N = \sup \mathsf{E} f_\tau$, where sup is taken over the class of Markov times $\mathfrak{M}_0^N$, but also enable us to determine the optimal time τ_0^N, i.e., the time for which $\mathsf{E} f_{\tau_0^N} = V_0^N$.

According to (8),

$$\tau_0^N = \min\{0 \leq k \leq N \colon v_k^N = f_k\}. \tag{24}$$

When solving specific optimal stopping problems, the following equivalent description of this stopping time τ_0^N is usable.

Let

$$D_n^N = \{\omega \colon v_n^N(\omega) = f_n(\omega)\} \tag{25}$$

and

$$C_n^N = \Omega \setminus D_n^N = \{\omega \colon v_n^N(\omega) = \mathsf{E}(v_{n+1}^N \mid \mathscr{F}_n)(\omega)\}.$$

Clearly, $D_N^N = \Omega$, $C_N^N = \varnothing$, and

$$D_0^N \subseteq D_1^N \subseteq \cdots \subseteq D_N^N = \Omega,$$
$$C_0^N \supseteq C_1^N \supseteq \cdots \supseteq C_N^N = \varnothing.$$

It follows from (24) and (25) that the stopping time τ_0^N can also be defined as

$$\tau_0^N = \min\{0 \leq k \leq N \colon \omega \in D_k^N\}. \tag{26}$$

It is natural to call D_k^N the "*stopping sets*" and C_k^N the "*continuation of observation sets*." This terminology can be justified as follows.

Consider the time instant $n = 0$, and divide the set Ω into sets D_0^N and C_0^N ($\Omega = D_0^N \cup C_0^N$, $D_0^N \cap C_0^N = \varnothing$). If $\omega \in D_0^N$, then $\tau_0^N(\omega) = 0$. In other words, "stopping" is done then at time $n = 0$. But if $\omega \in C_0^N$, then $\tau_0^N(\omega) \geq 1$. In the case where $\omega \in D_1^N \cap C_0^N$, we have $\tau_0^N(\omega) = 1$. The subsequent steps are considered in a similar manner. At time N the observations are certainly terminated.

7. Consider some examples.

EXAMPLE 1. Let $f = (f_0, f_1, \ldots, f_N)$ be a *martingale* with $f_0 = 1$. Then, according to Corollary 1 to Theorem 1, Sect. 2, $\mathsf{E} f_\tau = 1$ for any Markov time $\tau \in \mathfrak{M}_0^N$. Therefore, in this case, $V_0^N = \sup_{\tau \in \mathfrak{M}_0^N} \mathsf{E} f_\tau = 1$.

The functions $v_n^N = f_n$ for all $1 \le n \le N$, and $v_0^N = 1$. Then it is clear that $\tau_0^N = \min\{0 \le k \le N : f_k = v_k^N\} = 0$ and $\tau_n^N = n$ for any $1 \le n \le N$.

Thus the optimal stopping problem for martingale sequences is solved actually in a trivial manner: the optimal stopping time is $\tau_0^N(\omega) = 0$, $\omega \in \Omega$ (as well as, by the way, any other stopping time $\tau_n^N(\omega) = n$, $\omega \in \Omega$, $1 \le n \le N$).

EXAMPLE 2. If $f = (f_0, f_1, \ldots, f_N)$ is a *submartingale*, then $\mathsf{E} f_\tau \le \mathsf{E} f_N$ for any $\tau \in \mathfrak{M}_0^N$ (Theorem 1, Sect. 2). Thus the optimal stopping time here is $\tau^* \equiv N$. Since $v_k^N = \mathsf{E}(f_N \mid \mathscr{F}_k) \ge f_k$ (P-a.s.), it is possible that $\tau_0^N(\omega)$ may be less than N for some ω. But in any case both stopping times τ_0^N and $\tau^* \equiv N$ are optimal. Although $\tau^* \equiv N$ has a simple structure, τ_0^N nevertheless has a certain advantage, namely, it is the smallest among possible stopping times, i.e., if $\tilde{\tau}$ is also a stopping time in the class $\mathfrak{M}_0^N$, then $\mathsf{P}\{\tau_0^N \le \tilde{\tau}\} = 1$.

EXAMPLE 3. Let $f = (f_0, f_1, \ldots, f_N)$ be a *supermartingale*. Then $v_n^N = f_n$ for all $0 \le n \le N$. Therefore the optimal stopping time is (as in the martingale case) $\tau_0^N = 0$.

The preceding examples are fairly simple, and the problem of finding optimal stopping times is solved in them actually without invoking the theory given by Theorems 1 and 2. Their solution relies on the results on preservation of martingale, submartingale, and supermartingale properties under time change by a Markov time (Sect. 2). But in general finding the price V_0^N and the optimal stopping time τ_0^N may be a very difficult problem.

Of great interest are the cases where the functions f_n have the form

$$f_n(\omega) = f(X_n(\omega)),$$

where $X = (X_n)_{n \ge 0}$ is a Markov chain. As will be shown in Sect. 9 of Chap. 8, the solution of optimal stopping problems reduces in fact to the solution of *variational inequalities* and *Wald–Bellman equatons of dynamic programming*.

We also provide therein (nontrivial) examples of complete solutions to some optimal stopping problems for Markov sequences.

8. Problems

1. Show that the random variable $\xi(\omega) = \sup_{\alpha \in \mathfrak{A}_0} \xi_\alpha(\omega)$ constructed in the proof of the lemma (Subsection 3) satisfies requirements (a) and (b) in the definition of essential supremum. (*Hint*: In the case $\alpha \notin \mathfrak{A}_0$, consider $\mathsf{E} \max(\xi(\omega), \xi_\alpha(\omega))$.)
2. Show that $\xi(\omega) = \tan \tilde{\xi}(\omega)$ (see the end of the proof of the lemma in Subsection 3) also satisfies requirements (a) and (b).

3. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent identically distributed random variables with $\mathsf{E}|\xi_1| < \infty$. Consider the optimal stopping problem (in the class $\mathfrak{M}_1^\infty = \{\tau\colon 1 \le \tau < \infty\}$):

$$V^* = \sup_{\tau \in \mathfrak{M}_1^\infty} \mathsf{E}\Big(\max_{i \le \tau} \xi_i - c\tau\Big).$$

Let $\tau^* = \min\{n \ge 1\colon \xi_n \ge A^*\}$, where A^* is the unique root of the equation $\mathsf{E}(\xi_1 - A^*) = c$. Show that whenever $\mathsf{P}\{\tau^* < \infty\} = 1$, the stopping time τ^* is optimal in the class of all finite stopping times τ for which $\mathsf{E}\Big(\max_{i \le \tau} \xi_i - c\tau\Big)$ exists. Show also that $V^* = A^*$.

4. In this and the following problems, let

$$\begin{aligned}
\mathfrak{M}_n^\infty &= \{\tau\colon n \le \tau < \infty\},\\
V_n^\infty &= \sup_{\tau \in \mathfrak{M}_n^\infty} \mathsf{E} f_\tau,\\
v_n^\infty &= \operatorname*{ess\,sup}_{\tau \in \mathfrak{M}_n^\infty} \mathsf{E}(f_\tau \mid \mathscr{F}_n),\\
\tau_n^\infty &= \min\{k \ge n\colon v_n^\infty = f_n\}.
\end{aligned}$$

Assuming that $\mathsf{E}\sup f_n^- < \infty$, show that the limiting random variables

$$\tilde{v}_n = \lim_{N \to \infty} v_n^N$$

have the following properties:

(a) For any $\tau \in \mathfrak{M}_n^\infty$

$$\tilde{v}_n \ge \mathsf{E}(f_\tau \mid \mathscr{F}_n);$$

(b) If the stopping time $\tau_n^\infty \in \mathfrak{M}_n^\infty$, then

$$\begin{aligned}
&\tilde{v}_n = \mathsf{E}(f_{\tau_n^\infty} \mid \mathscr{F}_n),\\
&\tilde{v}_n = v_n^\infty \quad \Big(= \operatorname*{ess\,sup}_{\tau \in \mathfrak{M}_n^\infty} \mathsf{E}(f_\tau \mid \mathscr{F}_n)\Big).
\end{aligned}$$

5. Let $\tau_n^\infty \in \mathfrak{M}_n^\infty$. Deduce from (a) and (b) of the previous problem that τ_n^∞ is the optimal stopping time in the sense that

$$\operatorname*{ess\,sup}_{\tau \in \mathfrak{M}_n^\infty} \mathsf{E}(f_\tau \mid \mathscr{F}_n) = \mathsf{E}(f_{\tau_n^\infty} \mid \mathscr{F}_n) \quad (\mathsf{P}\text{-a.s.})$$

and

$$\sup_{\tau \in \mathfrak{M}_n^\infty} \mathsf{E} f_\tau = \mathsf{E} f_{\tau_n^\infty},$$

i.e., $V_n^\infty = \mathsf{E} f_{\tau_n^\infty}$.

Chapter 8
Markov Chains

The modern theory of Markov processes has its origins in the studies of A. A. Markov (1906–1907) on sequences of experiments "connected in a chain" and in attempts to describe mathematically the physical phenomenon known as Brownian motion (L. Bachelier 1900, A. Einstein 1905).

E. B. Dynkin "Markov processes," [21, Vol. 1]

1. Definitions and Basic Properties

1. In Sect. 12 of Chap. 1 we set out, for the case of *finite* probability spaces, the fundamental ideas and principles behind the concept of *Markov dependence* (see property (7) therein) of random variables, which is designed to describe the evolution of *memoryless systems*. In this chapter we extend this treatment to more general probability spaces.

One of the main problems of the theory of Markov processes is the study of the *asymptotic behavior* (as time goes to infinity) of memoryless systems. Remarkably, under very broad assumptions, such a system evolves as if it "forgot" the initial state, its behavior "stabilizes," and the system reaches a "steady-state" regime. We will analyze in detail the asymptotic behavior of systems described as *Markov chains with countable many states*. To this end we will provide a classification of the states of Markov chains according to the algebraic and asymptotic properties of their transition probabilities.

2. Let $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geq 0}, \mathsf{P})$ be a filtered probability space, i.e., a probability space $(\Omega, \mathscr{F}, \mathsf{P})$ with a specified filtration (flow) $(\mathscr{F}_n)_{n\geq 0}$ of σ-algebras $\mathscr{F}_n$, $n \geq 0$, such that $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \ldots \subseteq \mathscr{F}$. Intuitively, $\mathscr{F}_n$ describes the "information" available by the time n (inclusive).

Let, further, $(E, \mathscr{E})$ be a measurable space representing the "state space," where systems under consideration take their values. For "technical reasons" (e.g., so that, for any random element $X_0(\omega)$ and $x \in E$, the set $\{\omega\colon X_0(\omega) = x\} \in \mathscr{F}$) it will be

A. N. Shiryaev, *Probability-2*, Graduate Texts in Mathematics 95,
https://doi.org/10.1007/978-0-387-72208-5_5

assumed that the σ-algebra $\mathscr{E}$ contains all singletons in E, i.e., sets consisting of one point. (Regarding this assumption, see Subsection 6 below.)

The measurable spaces $(E, \mathscr{E})$ subject to this assumption are called the *phase spaces* or the *state spaces* of the systems under consideration.

Definition 1 (Markov chain in wide sense). Let $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geq 0}, \mathsf{P})$ be a filtered probability space and $(E, \mathscr{E})$ a phase space. A sequence $X = (X_n)_{n\geq 0}$ of E-valued $\mathscr{F}_n/\mathscr{E}$-measurable random elements $X_n = X_n(\omega)$, $n \geq 0$, defined on $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geq 0}, \mathsf{P})$, is a *sequence of random variables with Markov dependence*(or a *Markov chain*) *in the wide sense* if for any $n \geq 0$ and $B \in \mathscr{E}$ the following *wide-sense Markov property holds*:

$$\mathsf{P}(X_{n+1} \in B \,|\, \mathscr{F}_n)\,(\omega) = \mathsf{P}(X_{n+1} \in B \,|\, X_n(\omega)) \quad (\mathsf{P}\text{-a.s.}). \tag{1}$$

Let $\mathscr{F}_n^X = \sigma(X_0, X_1, \ldots, X_n)$ be the σ-algebra generated by $X_0, X_1, \ldots, X_n$. Since $\mathscr{F}_n^X \subseteq \mathscr{F}_n$ and X_n are $\mathscr{F}_n^X$-measurable, (1) implies the *Markov property in the strict sense* (or simply *Markov property*):

$$\mathsf{P}(X_{n+1} \in B \,|\, \mathscr{F}_n^X)\,(\omega) = \mathsf{P}(X_{n+1} \in B \,|\, X_n(\omega)) \quad (\mathsf{P}\text{-a.s.}). \tag{2}$$

For clarity (cf. Sect. 12 in Chap. 1, Vol. 1), this property is often written

$$\mathsf{P}(X_{n+1} \in B \,|\, X_0(\omega), \ldots, X_n(\omega)) = \mathsf{P}(X_{n+1} \in B \,|\, X_n(\omega)) \quad (\mathsf{P}\text{-a.s.}). \tag{3}$$

The strict-sense Markov property (2) deduced from (1) suggests the definition of the Markov dependence in the case where the flow $(\mathscr{F}_n)_{n\geq 0}$ is not specified a priori.

Definition 2 (Markov chain). Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space and $(E, \mathscr{E})$ a phase space. A sequence $X = (X_n)_{n\geq 0}$ of $\mathscr{F}/\mathscr{E}$-measurable E-valued random elements $X_n = X_n(\omega)$ is a *sequence of random variables with Markov dependence*, or a *Markov chain*, if for any $n \geq 0$ and $B \in \mathscr{E}$ the *strict-sense Markov property* (2) holds.

Remark. The introduction from the outset of a filtered probability space on which a Markov chain in the wide sense is defined is useful in many problems where the behavior of systems depends on a "flow of information" $(\mathscr{F}_n)_{n\geq 0}$. For example, it may happen that the first component $X = (X_n)_{n\geq 0}$ of a "two-dimensional" process $(X, Y) = (X_n, Y_n)_{n\geq 0}$ is not a Markov chain in the sense of (2), but nevertheless it is a Markov chain in the sense of (1) with $\mathscr{F}_n = \mathscr{F}_n^{X,Y}$, $n \geq 0$.

However, in the elementary exposition of the theory of Markov chains to be set out in this chapter, the flow $(\mathscr{F}_n)_{n\geq 0}$ is not usually introduced and the presentation is based on Definition 2.

3. The Markov property characterizes the "lack of aftereffect" (lack of memory) in the evolution of a system whose states are described by the sequence $X = (X_n)_{n\geq 0}$. In the case of a finite space Ω, this was stated in Sect. 12, Chap. 1 (Vol. 1) as the property

$$\mathsf{P}(\mathrm{F}\,|\,\mathrm{PN}) = \mathsf{P}(\mathrm{F}\,|\,\mathrm{N}), \tag{4}$$

where F stands for "future", P for "past", and N for "present" ("now"). It was pointed out there that Markov systems also possess the property

$$\mathsf{P}(\mathrm{PF}\,|\,\mathrm{N}) = \mathsf{P}(\mathrm{P}\,|\,\mathrm{N})\ \mathsf{P}(\mathrm{F}\,|\,\mathrm{N}), \tag{5}$$

interpretable as independence of past and future for a given present.

In the general case, the analogs of (4) and (5) are stated as properties (6) and (7) in the following theorem, which gives various equivalent formulations of the Markov property (in the sense of Definition 2). In this theorem the following notation is used:

$$\mathscr{F}^X_{[0,n]} = \sigma(X_0, X_1, \dots, X_n),$$
$$\mathscr{F}^X_{[n,\infty)} = \sigma(X_n, X_{n+1}, \dots),$$
$$\mathscr{F}^X_{(n,\infty)} = \sigma(X_{n+1}, X_{n+2}, \dots).$$

Theorem 1. *The Markov property* (2) *is equivalent to either of the following two properties*: *for* $n \geq 0$,

$$\mathsf{P}(\mathrm{F}\,|\,\mathscr{F}^X_{[0,n]})(\omega) = \mathsf{P}(\mathrm{F}\,|\,X_n(\omega)) \quad (\mathsf{P}\text{-a.s.}) \tag{6}$$

for any future event $\mathrm{F} \in \mathscr{F}^X_{(n,\infty)}$, *or for* $n \geq 1$

$$\mathsf{P}(\mathrm{PF}\,|\,X_n(\omega)) = \mathsf{P}(\mathrm{P}\,|\,X_n(\omega))\ \mathsf{P}(\mathrm{F}\,|\,X_n(\omega)) \quad (\mathsf{P}\text{-a.s.}) \tag{7}$$

for any future event $\mathrm{F} \in \mathscr{F}^X_{(n,\infty)}$ *and past event* $\mathrm{P} \in \mathscr{F}^X_{[0,n-1]}$.

PROOF. First of all, we prove the equivalence of (6) and (7).

(6) $\Rightarrow$ (7). We have (P-a.s.)

$$\begin{aligned}\mathsf{P}(\mathrm{P}\,|\,X_n(\omega))\,\mathsf{P}(\mathrm{F}\,|\,X_n(\omega)) &= \mathsf{E}(I_{\mathrm{P}}\,|\,X_n(\omega))\ \mathsf{E}(I_{\mathrm{F}}\,|\,X_n(\omega))\\ &= \mathsf{E}\{I_{\mathrm{P}}\,\mathsf{E}(I_{\mathrm{F}}\,|\,X_n(\omega))\,|\,X_n(\omega)\} = \mathsf{E}\{I_{\mathrm{P}}\,\mathsf{E}(I_{\mathrm{F}}\,|\,\mathscr{F}^X_{[0,n]})(\omega)\,|\,X_n(\omega)\}\\ &= \mathsf{E}\{\mathsf{E}(I_{\mathrm{P}}I_{\mathrm{F}}\,|\,\mathscr{F}^X_{[0,n]})(\omega)\,|\,X_n(\omega)\} = \mathsf{E}\{I_{\mathrm{P}}I_{\mathrm{F}}\,|\,X_n(\omega)\} = \mathsf{P}(\mathrm{PF}\,|\,X_n(\omega)).\end{aligned}$$

(7) $\Rightarrow$ (6). We must show that for any set C in $\mathscr{F}^X_{[0,n]}$

$$\mathsf{E}(I_C\,\mathsf{P}(\mathrm{F}\,|\,X_n)) = \mathsf{E}(I_C\,\mathsf{P}(\mathrm{F}\,|\,\mathscr{F}^X_{[0,n]})). \tag{6$'$}$$

To this end, consider first a particular case of such a set, namely, a set PN, where $\mathrm{P} \in \mathscr{F}^X_{[0,n-1]}$ and $\mathrm{N} \in \sigma(X_n)$, and show that in this case (6$'$) follows from (7).

Indeed,

$$\begin{aligned}&\mathsf{E}(I_{\mathrm{PN}}\,\mathsf{P}(\mathrm{F}\,|\,X_n)) = \mathsf{E}(I_{\mathrm{P}}I_{\mathrm{N}}\,\mathsf{E}(\mathrm{F}\,|\,X_n)) = \mathsf{E}(I_{\mathrm{N}}\,\mathsf{E}(I_{\mathrm{P}}\,\mathsf{E}(I_{\mathrm{F}}\,|\,X_n)\,|\,X_n))\\ &= \mathsf{E}(I_{\mathrm{N}}\,\mathsf{E}(I_{\mathrm{P}}\,|\,X_n)\ \mathsf{E}(I_{\mathrm{F}}\,|\,X_n)) = \mathsf{E}(I_{\mathrm{N}}\,\mathsf{P}(\mathrm{P}\,|\,X_n)\ \mathsf{P}(\mathrm{F}\,|\,X_n)) \overset{(7)}{=} \mathsf{E}(I_{\mathrm{N}}\,\mathsf{P}(\mathrm{PF}\,|\,X_n))\\ &\qquad = \mathsf{P}(\mathrm{PNF}) = \mathsf{E}(I_{\mathrm{PN}}\,\mathsf{P}(\mathrm{F}\,|\,\mathscr{F}^X_{[0,n]})), \quad (8)\end{aligned}$$

i.e., the property (6′) holds for sets C of the form PN, where $\mathrm{P} \in \mathscr{F}^X_{[0,n-1]}$ and $\mathrm{N} \in \sigma(X_n)$. By means of monotone classes arguments (Sect. 2, Chap. 2, Vol. 1) we deduce that property (6′) is valid for any sets C in $\mathscr{F}^X_{[0,n]}$. Since the function $\mathsf{P}(\mathrm{F} \,|\, X_n)$ is $\mathscr{F}^X_{[0,n]}$-measurable, (6′) implies that $\mathsf{P}(\mathrm{F} \,|\, X_n)$ is a version of the conditional probability $\mathsf{P}(\mathrm{F} \,|\, \mathscr{F}^X_{[0,n]})$, i.e., (6) holds.

Let us turn to the proof of equivalence of (2) and (6), or, in view of the foregoing proof, that of (2) and (7). The implication (6) $\Rightarrow$ (2) is obvious. Let us prove the implication (2) $\Rightarrow$ (6), invoking again the monotone classes arguments.

The sets F in (6) belong to the σ-algebra $\mathscr{F}^X_{(n,\infty)} = \mathscr{F}^X_{[n+1,\infty)}$, the σ-algebra generated by the algebra $\bigcup_{k=1}^{\infty} \mathscr{F}^X_{[n+1,n+k]}$, where $\mathscr{F}^X_{[n+1,n+k]} = \sigma(X_{n+1}, \ldots, X_{n+k})$. Therefore it is natural to start with the proof of (6) for sets F in the σ-algebras $\mathscr{F}^X_{[n+1,n+k]}$.

We will prove this by induction. If $k = 1$, then $\mathscr{F}^X_{[n+1,n+1]} = \sigma(X_{n+1})$, and (6) is the same as (2), which is assumed to hold.

Now let (6) hold for some $k \geq 1$. Let us prove its validity for $k+1$.

To this end, let us take a set $\mathrm{F} \in \mathscr{F}^X_{[n+1,n+k+1]}$ of the form $\mathrm{F} = \mathrm{F}^1 \cap \mathrm{F}^2$, where $\mathrm{F}^1 \in \mathscr{F}^X_{[n+1,n+k]}$ and $\mathrm{F}^2 \in \sigma(X_{n+k+1})$. Then, using the induction assumption, we find that (P-a.s.)

$$
\begin{aligned}
\mathsf{P}(\mathrm{F} \,|\, \mathscr{F}^X_{[0,n]}) &= \mathsf{E}(I_{\mathrm{F}} \,|\, \mathscr{F}^X_{[0,n]}) = \mathsf{E}[I_{\mathrm{F}^1 \cap \mathrm{F}^2} \,|\, \mathscr{F}^X_{[0,n]}] \\
&= \mathsf{E}[I_{\mathrm{F}^1} \, \mathsf{E}(I_{\mathrm{F}^2} \,|\, \mathscr{F}^X_{[0,n+k]}) \,|\, \mathscr{F}^X_{[0,n]}] \\
&= \mathsf{E}[I_{\mathrm{F}^1} \, \mathsf{E}(I_{\mathrm{F}^2} \,|\, X_{n+k}) \,|\, \mathscr{F}^X_{[0,n]}] = \mathsf{E}[I_{\mathrm{F}^1} \, \mathsf{E}(I_{\mathrm{F}^2} \,|\, X_{n+k}) \,|\, X_n] \\
&= \mathsf{E}[I_{\mathrm{F}^1} \, \mathsf{E}(I_{\mathrm{F}^2} \,|\, \mathscr{F}_{[n,n+k]}) \,|\, X_n] = \mathsf{E}[\mathsf{E}(I_{\mathrm{F}^1} I_{\mathrm{F}^2} \,|\, \mathscr{F}_{[n,n+k]}) \,|\, X_n] \\
&= \mathsf{E}[I_{\mathrm{F}^1} I_{\mathrm{F}^2} \,|\, X_n] = \mathsf{P}(\mathrm{F}^1 \cap \mathrm{F}^2 \,|\, X_n) = \mathsf{P}(\mathrm{F} \,|\, X_n). \qquad (9)
\end{aligned}
$$

The fact that property (9) holds, as we proved, for the sets $\mathrm{F} \in \mathscr{F}^X_{[n+1,n+k+1]}$ of the form $\mathrm{F} = \mathrm{F}^1 \cap \mathrm{F}^2$ with $\mathrm{F}^1 \in \mathscr{F}^X_{[n+1,n+k]}$ and $\mathrm{F}^2 \in \sigma(X_{n+k+1})$ implies (Problem 1a) that this property holds for *any* sets $\mathrm{F} \in \mathscr{F}^X_{[n+1,n+k+1]}$. Hence we conclude (Problem 1b) that (9) is valid also for F in the algebra $\bigcup_{k=1}^{\infty} \mathscr{F}^X_{[n+1,n+k]}$, which implies in turn (Problem 1c) that this property is satisfied also for the σ-algebra $\sigma\left(\bigcup_{k=1}^{\infty} \mathscr{F}^X_{[n+1,n+k]}\right) = \mathscr{F}^X_{(n,\infty)}$.

□

Remark. The reasoning in this proof is based on the *principle of appropriate sets* (starting the proof with sets of a "simple" structure) by applying subsequently the results on *monotone classes* (Sect. 2, Chap. 2, Vol. 1). In what follows this method will be repeatedly used (e.g., proofs of Theorems 2 and 3, which, in particular, enable one to recover the parts of the foregoing proof of Theorem 1 that were stated as Problems 1a, 1b, and 1c).

4. As a classical example of the Markov chain, consider the random walk $X = (X_n)_{n \geq 0}$ with

$$
X_n = X_0 + S_n, \quad n \geq 1, \qquad (10)
$$

where $S_n = \xi_1 + \cdots + \xi_n$ and $X_0, \xi_1, \xi_2, \ldots$ are independent random variables defined on a probability space $(\Omega, \mathscr{F}, \mathsf{P})$.

Theorem 2. *Let $\mathscr{F}_0 = \sigma(X_0)$, $\mathscr{F}_n = \sigma(X_0, \xi_1, \ldots, \xi_n)$, $n \geq 1$. The sequence $X = (X_n)_{n\geq 0}$ considered on the filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geq 0}, \mathsf{P})$ is a Markov chain (in the wide as well in the strict sense), i.e.,*

$$\mathsf{P}(X_{n+1} \in B \,|\, \mathscr{F}_n)(\omega) = \mathsf{P}(X_{n+1} \in B \,|\, X_n(\omega)) \quad (\mathsf{P}\text{-a.s.}) \tag{11}$$

for $n \geq 0$ and $B \in \mathscr{B}(R)$, and

$$\mathsf{P}(X_{n+1} \in B \,|\, X_n(\omega)) = P_{n+1}(B - X_n(\omega)) \quad (\mathsf{P}\text{-a.s.}), \tag{12}$$

where

$$P_{n+1}(A) = \mathsf{P}\{\xi_{n+1} \in A\} \tag{13}$$

and

$$B - X_n(\omega) = \{y\colon y + X_n(\omega) \in B\}, \quad B \in \mathscr{B}(R).$$

PROOF. We will prove (11) and (12) simultaneously.

For discrete probability spaces, similar results were proved in Sect. 12, Chap. 1, Vol. 1, and it may appear that the proof here should be rather simple, too. But, as will be seen from the subsequent proof, the present situation is more complicated.

Let A be a set such that $A \in \{X_0 \in B_0, \xi_1 \in B_1, \ldots, \xi_n \in B_n\}$, where $B_i \in \mathscr{B}(R)$, $i = 0, 1, \ldots, n$. By the definition of conditional probability $\mathsf{P}(X_{n+1} \in B \,|\, \mathscr{F}_n)\,(\omega)$ (Sect. 7, Chap. 2, Vol. 1),

$$\begin{aligned}
\int_A \mathsf{P}(X_{n+1} \in B \,|\, \mathscr{F}_n)(\omega)\ \mathsf{P}(d\omega) &= \int_A I_{\{X_{n+1}\in B\}}(\omega)\ \mathsf{P}(d\omega) \\
&= \mathsf{P}\{X_0 \in B_0, \xi_1 \in B_1, \ldots, \xi_n \in B_n, X_{n+1} \in B\} \\
&= \int_{B_0\times\cdots\times B_n} P_{n+1}(B - (x_0 + x_1 + \cdots + x_n))\, P_0(dx_0) \ldots P_n(dx_n) \\
&= \int_A P_{n+1}(B - X_n(\omega))\ \mathsf{P}(d\omega).
\end{aligned} \tag{14}$$

Thus we have proved the equality

$$\int_A \mathsf{P}(X_{n+1} \in B \,|\, \mathscr{F}_n)(\omega)\ \mathsf{P}(d\omega) = \int_A P_{n+1}(B - X_n(\omega))\ \mathsf{P}(d\omega) \tag{15}$$

for sets $A \in \mathscr{F}_n$ of the form $A = \{X_0 \in B_0, \xi_1 \in B_1, \ldots, \xi_n \in B_n\}$.

Obviously, the system $\mathscr{A}_n$ of such sets is a π-system ($\Omega \in \mathscr{A}_n$ and if $A_1 \in \mathscr{A}_n$ and $A_2 \in \mathscr{A}_n$, then also $A_1 \cap A_2 \in \mathscr{A}_n$; see Definition 2 in Sect. 2, Chap. 2, Vol. 1). Further, let $\mathscr{L}$ be the class of sets $A \in \mathscr{F}_n$ that satisfy (15).

Let us show that $\mathscr{L}$ is a λ-system (Definition 2, Sect. 2, Chap. 2, Vol. 1). It is clear that $\Omega \in \mathscr{L}$, i.e., the property (λ_a) of this definition is satisfied. The property (λ_b) of the same definition holds because of the additivity of Lebesgue integral. Finally, the property (λ_c) of that definition follows from the theorem on monotone convergence of Lebesgue integrals (Sect. 6, Chap. 2, Vol. 1).

Thus, $\mathscr{L}$ is a λ-system. Applying statement (c) of Theorem 2 in Sect. 2, Chap. 2, Vol. 1, we obtain that $\sigma(\mathscr{A}_n) \subseteq \mathscr{L}$. But $\sigma(\mathscr{A}_n) = \mathscr{F}_n$, so property (15) is valid also for sets A in $\mathscr{F}_n$.

Consequently, taking into account that $P_{n+1}(B - X_n(\omega))$ (as a function of ω) is $\mathscr{F}_n$-measurable (Problem 2) we obtain from (15) (by the definition of conditional probability) that $P_{n+1}(B - X_n(\omega))$ is a version of the conditional probability $\mathsf{P}(X_{n+1} \in B \,|\, \mathscr{F}_n)(\omega)$. Finally, using the "telescopic property" of conditional expectations (see property **H*** in Sect. 7, Chap. 2, Vol. 1) we find that (P-a.s.)

$$\mathsf{P}(X_{n+1} \in B \,|\, X_n)(\omega) = \mathsf{E}[\, I_{\{X_{n+1} \in B\}} \,|\, X_n](\omega) = \mathsf{E}[\, \mathsf{E}(I_{\{X_{n+1} \in B\}} \,|\, \mathscr{F}_n) \,|\, X_n](\omega) =$$
$$= \mathsf{E}[P_{n+1}(B - X_n) \,|\, X_n(\omega)] = P_{n+1}(B - X_n(\omega)). \quad (16)$$

Thus, both properties (11) and (12) are proved.

□

Remark. Properties (11) and (12) could also be deduced (Problem 3) directly from Lemma 3 in Sect. 2, Chap. 2, Vol. 1. We carried out a detailed proof of these "almost obvious" properties to demonstrate once more the technique of the proof of such assertions based on the principle of *appropriate sets* and results on *monotone classes*.

5. Consider the Markov property (1). If $(E, \mathscr{E})$ is a *Borel* space, then, by Theorem 3 in Sect. 7, Chap. 2, Vol. 1, for any $n \geq 0$ there exists a regular conditional distribution $P_{n+1}(x; B)$ such that (P-a.s.)

$$\mathsf{P}(X_{n+1} \in B \,|\, X_n(\omega)) = P_{n+1}(X_n(\omega); B), \quad (17)$$

where the function $P_{n+1}(x; B)$, $B \in \mathscr{E}$, $x \in E$, has the following properties (Definition 7, Sect. 7, Chap. 2, Vol. 1):

(a) For any x the set function $P_{n+1}(x, \cdot)$ is a *measure* on $(E, \mathscr{E})$;
(b) For any $B \in \mathscr{E}$ the function $P_{n+1}(\cdot\,; B)$ is *$\mathscr{E}$-measurable*.

The functions $P_n = P_n(x; B)$, $n \geq 1$, are called *transition functions* (or *Markov kernels*).

The case of special interest to us will be the one where all these transition functions are the same, $P_1 = P_2 = \ldots$, or, more precisely, when the conditional probabilities $\mathsf{P}(X_{n+1} \in B \,|\, X_n(\omega))$, $n \geq 0$, have a common version of regular conditional distribution $P(x; B)$ such that (P-a.s.)

$$\mathsf{P}(X_{n+1} \in B \,|\, X_n(\omega)) = P(X_n(\omega); B) \quad (18)$$

for all $n \geq 0$ and $B \in \mathscr{E}$.

If such a version $P = P(x;B)$ exists (in which case we can set all $P_n = P$, $n \geq 0$), then the Markov chain is called *homogeneous* (in time) with transition function $P = P(x;B)$, $x \in E$, $B \in \mathscr{E}$.

The intuitive meaning of the homogeneity property of Markov chains is clear: the corresponding system evolves *homogeneously* in the sense that the probabilistic mechanisms governing the transitions of the system remain the same for all time instants $n \geq 0$. (In the theory of dynamical systems, systems with this property are said to be *conservative*.)

Besides the transition probabilities $P_1, P_2, \dots$, or the transition probability P for *homogeneous* chains, the important characteristic of Markov chains is the *initial* distribution $\pi = \pi(B)$, $B \in \mathscr{E}$, i.e., the probability distribution $\pi(B) = \mathsf{P}\{X_0 \in B\}$, $B \in \mathscr{E}$.

The set of objects $(\pi, P_1, P_2, \dots)$ completely determines the *probabilistic properties* of the sequence $X = (X_n)_{n\geq 0}$, since all the *finite-dimensional* distributions of this sequence are given by the formulas

$$\mathsf{P}\{X_0 \in B\} = \pi(B), \quad B \in \mathscr{E},$$

and

$$\mathsf{P}\{(X_0, X_1, \dots, X_n) \in B\} = \int\limits_{E\times\cdots\times E} I_B(x_0, x_1, \dots, x_n)\, \pi(dx_0) P_1(x_0; dx_1) \cdots P_n(x_{n-1}; dx_n) \quad (19)$$

for any $n \geq 1$ and $B \in \mathscr{B}(E^{n+1})$ $(= \mathscr{E}^{n+1} = \mathscr{E} \otimes \cdots \otimes \mathscr{E}$ $(n+1)$ times).

Indeed, consider first the set B of the form $B = B_0 \times \cdots \times B_n$. Then for $n = 1$ we have, by the formula for total probability (see (5) in Sect. 7, Chap. 2, Vol. 2),

$$\mathsf{P}\{X_0 \in B_0, X_1 \in B_1\} = \int\limits_{\Omega} I_{\{X_0\in B_0\}}(\omega)\ \mathsf{P}(X_1 \in B_1 \,|\, X_0(\omega))\ \mathsf{P}(d\omega)$$

$$= \int\limits_{\Omega} I_{\{X_0\in B_0\}}(\omega)\, P_1(B_1; X_0(\omega))\ \mathsf{P}(d\omega)$$

$$= \int\limits_{E} I_{B_0}(x_0)\, P_1(B_1; x_0)\, \pi(dx_0) = \int\limits_{E\times E} I_{B_0\times B_1}(x_0, x_1)\, P_1(dx_1; x_0)\, \pi(dx_0).$$

The further proof proceeds by induction:

$$\mathsf{P}\{X_0 \in B_0, X_1 \in B_1, \dots, X_n \in B_n\}$$

$$= \int_{\Omega} I_{\{X_0 \in B_0, \dots, X_{n-1} \in B_{n-1}\}}(\omega)\ \mathsf{P}(X_n \in B_n \,|\, X_0(\omega), \dots, X_{n-1}(\omega))\ \mathsf{P}(d\omega)$$

$$= \int_{\Omega} I_{\{X_0 \in B_0, \dots, X_{n-1} \in B_{n-1}\}}(\omega)\ \mathsf{P}(X_n \in B_n \,|\, X_{n-1}(\omega))\ \mathsf{P}(d\omega)$$

$$= \int_{\Omega} I_{\{X_0 \in B_0, \dots, X_{n-1} \in B_{n-1}\}}(\omega)\, P_n(B_n; X_{n-1}(\omega))\ \mathsf{P}(d\omega)$$

$$= \int_{E \times \dots \times E} I_{B_0 \times B_1 \times \dots \times B_{n-1}}(x_0, x_1, \dots, x_{n-1}) \times P_n(B_n; x_{n-1})\ \mathsf{P}\{X_0 \in dx_1, \dots, X_{n-1} \in dx_n\}$$

$$= \int_{E \times \dots \times E} I_{B_0 \times B_1 \times \dots \times B_{n-1} \times B_n}(x_0, x_1, \dots, x_{n-1}, x_n) \times P_n(dx_n; x_{n-1})\, P_{n-1}(dx_{n-1}; x_{n-2}) \dots P_1(dx_1; x_0)\, \pi(dx_0),$$

which coincides with (19) for sets B of the form $B = B_0 \times B_1 \times \dots \times B_n$. The general case of the sets $B \in \mathscr{B}(E^{n+1})$ is treated in the same way as in the proof of the similar point in Theorem 2.

Using the results on *monotone classes* (Sect. 2, Chap. 2, Vol. 1), one can deduce from (19) (Problem 4) that for any bounded $\mathscr{B}(E^{n+1})$-measurable function $h = h(x_0, x_1, \dots, x_n)$

$$\mathsf{E}\, h(X_0, X_1, \dots, X_n) = \int_{E^{n+1}} h(x_0, x_1, \dots, x_n)\, \pi(dx_0)\, P_1(dx_1; x_0) \cdots P_n(dx_n; x_{n-1}). \quad (20)$$

6. Thus, if we have a Markov chain (in the wide or strict sense), then, by means of formula (19), we can recover the distribution $Law(X_0, X_1, \dots, X_n)$ of any collection of random variables $X_0, X_1, \dots, X_n$, $n \geq 1$, from its initial distribution $\pi = \pi(B) = \mathsf{P}\{X_0 \in B\}$, $B \in \mathscr{E}$, and its transition probabilities $P_n(x; B)$, $n \geq 1$, $x \in E$, $B \in \mathscr{E}$.

Now we take another look at defining Markov chains. We will require that they must be completely determined by a *given collection* of distributions $(\pi, P_1, P_2, \dots)$, where the meaning of π is the probability distribution of the *initial* state of the system and the functions $P_{n+1} = P_{n+1}(x; B)$, $n \geq 0$, satisfying (a) and (b) of Subsection 5, play the role of transition probabilities, i.e., the probabilities that the system in state x at time n will get into the set $B \in \mathscr{E}$ at time $n + 1$. Naturally, when our initial object is the collection $(\pi, P_1, P_2, \dots)$, the question arises as to whether there is *any* Markov chain with initial distribution π and transition probabilities $P_1, P_2, \dots$.

An (affirmative) answer to this question is virtually given by Kolmogorov's theorem (Theorem 1 and Corollary 3 in Sect. 9, Chap. 2, Vol. 1), at least for $E = R^d$, and by Ionescu Tulcea's theorem (Theorem 2 in Sect. 9, Chap. 2, Vol. 1) for arbitrary measurable spaces $(E, \mathscr{E})$.

Following the proofs of these theorems, define first of all the measurable space $(\Omega, \mathscr{F})$ by setting $(\Omega, \mathscr{F}) = (E^\infty, \mathscr{B}(E^\infty))$, where $E^\infty = E \times E \times \cdots$, $\mathscr{B}(E^\infty) = \mathscr{E} \otimes \mathscr{E} \otimes \cdots$; in other words, we take the elementary events to be the "points" $\omega = (x_0, x_1, \dots)$, where $x_i \in E$.

Define the *flow* $(\mathscr{F}_n)_{n\geq 0}$ by setting $\mathscr{F}_n = \sigma(x_0, x_1, \dots, x_n)$. The random variables $X_n = X_n(\omega)$ will be defined "*canonically*" by setting $X_n(\omega) = x_n$ if $\omega = (x_0, x_1, \dots)$.

Ionescu Tulcea's theorem states that for *arbitrary measurable* spaces $(E, \mathscr{E})$ (and in particular for phase spaces under consideration) *there exists a probability measure* P_π on $(\Omega, \mathscr{F})$ *such that*

$$\mathsf{P}_\pi\{X_0 \in B\} = \pi(B), \quad B \in \mathscr{E}, \tag{21}$$

and the *finite-dimensional distributions* for all $n \geq 1$ are given by

$$\mathsf{P}_\pi\{(X_0, X_1, \dots, X_n) \in B\} = \int_E \pi(dx_0) \int_E P_1(x_0; dx_1) \dots \int_E I_B(x_0, \dots, x_n)\, P_n(x_{n-1}; dx_n). \tag{22}$$

Theorem 3. *The canonically defined sequence $X = (X_n)_{n\geq 0}$ is a Markov chain (in the sense of Definition 2) with respect to the measure P_π specified by Ionescu Tulcea's theorem.*

PROOF. We must prove that for $n \geq 0$ and $B \in \mathscr{E}$

$$\mathsf{P}_\pi(X_{n+1} \in B \,|\, \mathscr{F}_n)(\omega) = \mathsf{P}_\pi(X_{n+1} \in B \,|\, X_n(\omega)) \quad (\mathsf{P}_\pi\text{-a.s.}) \tag{23}$$

and, moreover, for $n \geq 0$

$$\mathsf{P}_\pi(X_{n+1} \in B \,|\, X_n(\omega)) = P_n(X_n(\omega); B) \quad (\mathsf{P}_\pi\text{-a.s.}). \tag{24}$$

We will prove this by using the principle of appropriate sets and the results on monotone classes (Sect. 2, Chap. 2, Vol. 1).

As before, we take for appropriate sets the sets of a "simple" structure $A \in \mathscr{F}_n$ of the form

$$A = \{\omega\colon X_0(\omega) \in B_0, \dots, X_n(\omega) \in B_n\},$$

where $B_i \in \mathscr{E}$, $i = 0, 1, \dots, n$, and let $B \in \mathscr{E}$.

Then the construction of the measure P_π (see (22)) implies

$$\begin{aligned}\int_A I_{\{X_{n+1}\in B\}}(\omega)\ \mathsf{P}_\pi(d\omega) &= \mathsf{P}_\pi\{X_0 \in B_0, \dots, X_n \in B_n, X_{n+1} \in B\} \\ &= \int_{B_0} \pi(dx_0) \int_{B_1} P_1(x_0; dx_1) \dots \int_{B_n} P_n(x_{n-1}; dx_n) \int_B P_{n+1}(x_n; dx_{n+1}) \\ &= \int_A P_{n+1}(X_n(\omega); B)\ \mathsf{P}_\pi(d\omega). \end{aligned} \tag{25}$$

By the same arguments as in the proof of Theorem 2 (see the proof of (15) for sets $A \in \mathscr{F}_n$) we find that (25) is also fulfilled for $A \in \mathscr{F}_n$, i.e., for sets of the form $A = \{\omega\colon (X_0(\omega), \dots, X_n(\omega)) \in C\}$, where $C \in \mathscr{B}(E^{n+1})$.

Since, by the definition of conditional probabilities (Sect. 7, Chap. 2, Vol. 2),

$$\int\limits_A I_{\{X_{n+1}\in B\}}(\omega)\ \mathsf{P}_\pi(d\omega) = \int\limits_A \mathsf{P}_\pi(X_{n+1} \in B \,|\, \mathscr{F}_n)\,(\omega)\ \mathsf{P}_\pi(d\omega), \tag{26}$$

and the functions $P_{n+1}(X_n(\omega); B)$ are $\mathscr{F}_n$-measurable, we obtain the required properties (23) and (24) using (25) and the "telescopic" property of conditional expectations (see **H*** in Subsection 4, Sect. 7, Chap. 2, Vol. 1).

□

7. Thus, with any given collection of distributions $(\pi, P_1, P_2, \dots)$, we can associate a Markov chain (to be denoted by $X^\pi = (X_n, \mathsf{P}_\pi)_{n\geq 0}$) with *initial distribution* π and *transition probabilities* $P_1, P_2, \dots$ (i.e., a chain with properties (21), (23), and (24)). This chain proceeds as follows.

At the time instant $n = 0$, the initial state is randomly chosen according to the distribution π. If the initial value X_0 is equal to x, then at the first step the system moves from this state to a state x_1 with distribution $P_1(\cdot\,; x)$, and so on.

Therefore the initial distribution π acts only at time $n = 0$, while the subsequent evolution of the system is determined by the transition probabilities $P_1, P_2, \dots$. Consequently, if the random choice of two initial distributions π_1 and π_2 results in the same state x, then the behavior of the system will be the same (in probabilistic terms), being determined only by the transition probabilities $P_1, P_2, \dots$. This can also be expressed as follows.

Let P_x denote the distribution P_π corresponding to the case where π is supported at a single point x: $\pi(dy) = \delta_x(dy)$, i.e., $\pi(\{x\}) = 1$, where $\{x\}$ is a singleton, which belongs to $\mathscr{E}$ by the assumption about the phase space $(E, \mathscr{E})$ (Subsection 2).

Then (22) implies (Problem 4) that for any $A \in \mathscr{B}(E^\infty)$ and $x \in E$ the probability $\mathsf{P}_x(A)$ is (for each π) a version of the conditional probability $\mathsf{P}_\pi(A \,|\, X_0 = x)$, i.e.,

$$\mathsf{P}_\pi(A \,|\, X_0 = x) = \mathsf{P}_x(A) \quad (\mathsf{P}_\pi\text{-a.s.}). \tag{27}$$

For any $x \in E$ the probabilities $\mathsf{P}_x(\cdot)$ are completely determined by the collection of transition probabilities $(P_1, P_2, \dots)$.

Therefore, if we are primarily interested in knowing how the behavior of the system depends on the transition probabilities $(P_1, P_2, \dots)$, we can restrict ourselves to the probabilities $\mathsf{P}_x(\cdot)$, $x \in E$, obtaining, if needed, the probabilities $\mathsf{P}_\pi(\cdot)$ simply by the integration

$$\mathsf{P}_\pi(A) = \int\limits_E \mathsf{P}_x(A)\,\pi(dx), \quad A \in \mathscr{B}(E^\infty). \tag{28}$$

These arguments gave rise to an approach according to which the *main object* in the "general theory of Markov processes" (see [21]) (with discrete time in the present case) is not a particular Markov chain $X^\pi = (X_n, \mathsf{P}_\pi)_{n\geq 0}$, but rather a *family*

of Markov chains $X^x = (X_n, \mathsf{P}_x)_{n\geq 0}$ with $x \in E$. (Nevertheless, instead of the words "a family of Markov chains" one often says simply "a Markov chain" and writes "$X = (X_n, \mathscr{F}_n, \mathsf{P}_x)$" instead of "$X^x = (X_n, \mathsf{P}_x)_{n\geq 0}$ with $x \in E$.")

Let us emphasize that all these considerations presume that the chains are defined "canonically": the space $(\Omega, \mathscr{F})$ is taken to be $(E^\infty, \mathscr{E}^\infty)$, $\mathscr{E}^\infty = \mathscr{E} \otimes \mathscr{E} \otimes \cdots$, the random variables $X_n(\omega)$ are defined so that $X_n(\omega) = x_n$ if $\omega = (x_0, x_1, \ldots)$. Therefore in $X^x = (X_n, \mathsf{P}_x)$, only the probability P_x depends on x, whereas no dependence of X_n on x is assumed. This implies that, according to the measure P_x, all the trajectories $(X_n)_{n\geq 0}$ "start" at the point x, i.e., $\mathsf{P}_x\{X_0 = x\} = 1$.

8. In the case of *finite* Markov chains (Sect. 12, Chap. 1, Vol. 1), their behavior was analyzed by exploring the transition probabilities $p_{ij}^{(n)} = \mathsf{P}(X_n = j \,|\, X_0 = i)$, which were shown to satisfy the *Kolmogorov–Chapman equation* (see (13) therein) from which, in turn, the *forward and backward Kolmogorov equations* ((16) and (15) therein) were derived.

Now we turn to the Kolmogorov–Chapman equation for Markov chains with *arbitrary* phase space $(E, \mathscr{E})$. We will restrict ourselves to *homogeneous* chains for which $P_1 = P_2 = \cdots = P$.

In this case, in view of (22),

$$\mathsf{P}_\pi\{(X_0, X_1, \ldots, X_n) \in B\} = \int_E \pi(dx_0) \int_E P(x_0; dx_1) \ldots \int_E I_B(x_0, x_1, \ldots, x_n) P(x_{n-1}; dx_n). \quad (29)$$

In particular, for $n = 2$ we have

$$\mathsf{P}_\pi\{X_0 \in B_0, X_2 \in B_2\} = \int_{B_0} \int_E P(x_1; B_2) P(x_0; dx_1) \pi(dx_0). \quad (30)$$

Hence, by the Radon–Nikodym theorem (Sect. 6, Chap. 2, Vol. 1) and the definition of the conditional probabilities, we find that (π-a.s.)

$$\mathsf{P}_\pi(X_2 \in B_2 \,|\, X_0 = x) = \int_E P(x; dx_1) P(x_1; B_2). \quad (31)$$

Let us notice now that, by (27), $\mathsf{P}_\pi(X_2 \in B_2 \,|\, X_0 = x) = \mathsf{P}_x\{X_2 \in B_2\}$ (π-a.s.), where the probability $\mathsf{P}_x\{X_2 \in B_2\}$ has a simple meaning: this is the probability of the transition of the system from state x at time $n = 0$ into set B_2 at time $n = 2$, i.e., this is the transition probability for *two* steps.

Let $P^{(n)}(x; B_n) = \mathsf{P}_x\{X_n \in B_n\}$ denote the transition probability for n steps. Then, in view of the homogeneity of the chains at hand, $P^{(1)}(x; B_1) = P(x; B_1)$, hence we find from (31) that (π-a.s.)

$$P^{(2)}(x; B) = \int_E P^{(1)}(x; dx_1) P^{(1)}(x_1; B), \quad (32)$$

where $B \in \mathscr{E}$.

In a similar manner, one can establish (Problem 5) that for any $n \geq 0$, $m \geq 0$ (π-a.s.)

$$P^{(n+m)}(x;B) = \int_E P^{(n)}(x;dy)P^{(m)}(y;B). \tag{33}$$

This is the well-known **Kolmogorov–Chapman equation**, whose intuitive meaning is quite clear: to compute the probability $P^{(m+n)}(x;B)$ of a transition from the point $x \in E$ to the set $B \in \mathscr{E}$ for $n+m$ steps we must multiply the probability $P^{(n)}(x;dy)$ of transition from x into an "infinitesimal" neighborhood dy of $y \in E$ for n steps by the probability of transition from y to B for m steps (with subsequent integration over all "intermediate" points y).

Regarding the Kolmogorov–Chapman equation, which relates the transition probabilities for a varying number of steps, we should point out that it is established only up to "π-almost sure." In particular, this implies that this is not a relation that holds *for all* $x \in E$. This should not come as a surprise because, as on many previous occasions, where we had to choose *versions* of conditional probabilities, we are not guaranteed that these versions are such that the properties of interest are fulfilled identically in x rather than π-almost sure.

Nevertheless, it is possible to explicitly specify the versions for which the Kolmogorov–Chapman equation (33) is fulfilled *for all* $x \in E$.

This follows from the following assertions (Problem 6). Let the "transition probabilities" $P^{(n)}(x;B)$ be defined as follows:

$$P^{(1)}(x;B) = P(x;B)$$

and for $n > 1$

$$P^{(n)}(x;B) = \int_E P(x;dy)P^{(n-1)}(y;B).$$

Then

(i) $P^{(n)}(x;B)$, $n \geq 1$, are regular conditional probabilities on $\mathscr{E}$ for a fixed x;
(ii) $P^{(n)}(x;B)$ is equal to $\mathsf{P}_x\{X_n \in B\}$, hence it is a version of $\mathsf{P}_\pi(X_n \in B \,|\, X_0 = x)$ (π-a.s.);
(iii) For the functions $P^{(n)}(x;B)$, $n \geq 1$, the Kolmogorov–Chapman equations hold identically in $x \in E$.

9. Problems

1. Prove Problems 1a, 1b, and 1c stated in the proof of Theorem 1.
2. Prove that the function $P_{n+1}(B - X_n(\omega))$ in Theorem 2 is $\mathscr{F}_n$-measurable in ω.
3. Deduce the properties (11) and (12) from Lemma 3 in Sect. 2, Chap. 2, Vol. 1.
4. Prove (20) and (27).
5. Establish the validity of (33).
6. Prove statements (i), (ii), and (iii) given at the end of Subsection 8.
7. Establish whether the Markov property (3) implies that

$$\mathsf{P}(X_{n+1} \in B \,|\, X_0 \in B_0, X_1 \in B_1, \ldots, X_n \in B_n) = \mathsf{P}(X_{n+1} \in B \,|\, X_n \in B_n),$$

where $B, B_0, B_1, \ldots, B_n$ are subsets of $\mathscr{E}$ and $\mathsf{P}\{X_0 \in B_0, X_1 \in B_1, \ldots, X_n \in B_n\} > 0$.

2. Generalized Markov and Strong Markov Properties

1. In this section we mostly consider families $X^x = (X_n, P_x)_{n\geq 0}$, $x \in E$, of homogeneous Markov chains defined "canonically" on the coordinate space $(\Omega, \mathscr{F}) = (E^\infty, \mathscr{E}^\infty)$ and specified by a transition function $P = P(x; B)$, $x \in E$, $B \in \mathscr{E}$.

Let us define the shift operators $\theta_n \colon \Omega \to \Omega$ on $(\Omega, \mathscr{F})$ (cf. Sect. 1, Chap. 5) by setting

$$\theta_n(\omega) = (x_n, x_{n+1}, \dots)$$

for $\omega = (x_0, x_1, \dots)$.

If $H = H(\omega)$ is a $\mathscr{F}$-measurable function, then $H \circ \theta_n$ will denote the function $(H \circ \theta_n)(\omega)$ defined by

$$(H \circ \theta_n)(\omega) = H(\theta_n(\omega)). \tag{1}$$

Thus, if $\omega = (x_0, x_1, \dots)$ and $H = H(x_0, x_1, \dots)$, then $(H \circ \theta_n)(x_0, x_1, \dots) = H(x_n, x_{n+1}, \dots)$.

The following theorem is virtually property (6) in Sect. 1 restated in the context of the present case of a *family of homogeneous Markov chains*.

Theorem 1. *Let $X^x = (X_n, \mathsf{P}_x)_{n\geq 0}$, $x \in E$, be a family of homogeneous Markov chains determined by a transition function $P = P(x; B)$, $x \in E$, $B \in \mathscr{E}$. Assume that the probabilities $\mathsf{P}_x\{(X_0, X_1, \dots, X_n) \in B\}$ for $B \in \mathscr{B}(E^{n+1})$ and $n \geq 0$ are determined by (22) of Sect. 1 with $\pi(dy) = \delta_{\{x\}}(dy)$ and $P_1 = P_2 = \dots = P$.*

Then for any initial distribution π, any $n \geq 0$, and any bounded (or nonnegative) $\mathscr{F}$-measurable function $H = H(\omega)$ the following generalized Markov property holds:

$$\mathsf{E}_\pi(H \circ \theta_n \mid \mathscr{F}_n^X)(\omega) = \mathsf{E}_{X_n(\omega)} H \quad (\mathsf{P}_\pi\text{-}a.s.). \tag{2}$$

Remark. Although the notation in the theorem is self-explanatory, let us note nevertheless that E_π is the expectation with respect to $\mathsf{P}_\pi(\cdot) = \int_E \mathsf{P}_x(\cdot)\, \pi(dx)$, and $\mathsf{E}_{X_n(\omega)} H$ is to be understood as follows. Take the expectation $\mathsf{E}_x H$, i.e., the averaging of H with respect to P_x (denote it by $\psi(x)$), and then plug $X_n(\omega)$ for x into the expression thus obtained, so that $\mathsf{E}_{X_n(\omega)} H = \psi(X_n(\omega))$. (Note that $\mathsf{E}_x H$ is an $\mathscr{E}$-measurable function of x (Problem 1), so $\mathsf{E}_{X_n(\omega)} H$ is a random variable, i.e., an $\mathscr{F}/\mathscr{E}$-measurable function.)

PROOF. The proof of Theorem 1 again uses the principle of *appropriate sets and functions* with subsequent application of results on *monotone classes*.

To prove (2), we must show that for any $A \in \mathscr{F}_n^X = \sigma(x_0, x_1, \dots, x_n)$

$$\int_A (H \circ \theta_n)(\omega)\ \mathsf{P}_\pi(d\omega) = \int_A (\mathsf{E}_{X_n(\omega)} H)\ \mathsf{P}_\pi(d\omega), \tag{3}$$

or, in a more concise form,

$$\mathsf{E}_\pi(H \circ \theta_n; A) = \mathsf{E}_\pi(\mathsf{E}_{X_n} H; A), \tag{4}$$

where $\mathsf{E}_\pi(\xi; A)$ denotes $\mathsf{E}_\pi(\xi I_A)$ (Subsection 2, Sect. 6, Chap. 2, Vol. 1).

According to the principle of appropriate sets, consider sets A of a "simple" structure, that is, the sets $A = \{\omega\colon x_0 \in B_0, \dots, x_n \in B_n\}$, $B_i \in \mathscr{E}_i$, and a function $H = H(x_0, x_1, \dots, x_m)$, $m \geq 0$ (more precisely, an $\mathscr{F}_m^X$-measurable function H). Then (4) becomes

$$\mathsf{E}_\pi(H(X_n, X_{n+1}, \dots, X_{n+m}); A) = \mathsf{E}_\pi(\mathsf{E}_{X_n} H(X_0, X_1, \dots, X_m); A). \tag{5}$$

Using (22) of Sect. 1 we find that

$$\begin{aligned}
\mathsf{E}_\pi(H(X_n, X_{n+1}, \dots, X_{n+m}); A) &= \mathsf{E}_\pi(I_A(X_0, \dots, X_n)H(X_n, \dots, X_{n+m})) \\
&= \int\limits_{E^{n+m+1}} I_A(x_0, \dots, x_n)H(x_n, \dots, x_{n+m}) \\
&\qquad \times \pi(dx_0)P(x_0; dx_1) \cdots P(x_{n+m-1}; dx_{n+m}) \\
&= \int\limits_{E^{n+1}} I_A(x_0, \dots, x_n)\, \pi(dx_0)P(x_0; dx_1) \cdots P(x_{n-1}; dx_n) \\
&\qquad \times \int\limits_{E^m} H(x_n, \dots, x_{n+m})P(x_n; dx_{n+1}) \cdots P(x_{n+m-1}; dx_{n+m}) \\
&= \int\limits_{E^{n+1}} I_A(x_0, \dots, x_n)\pi(dx_0)P(x_0; dx_1) \cdots P(x_{n-1}; dx_n) \\
&\qquad \times \int\limits_{E^m} H(x_0, \dots, x_m)\ \mathsf{P}_x(dx_1, \dots, dx_m) = \mathsf{E}_\pi(\mathsf{E}_{X_n} H(X_0, \dots, X_m); A),
\end{aligned}$$

where $\mathsf{P}_x(dx_1, \dots, dx_m) = P(x; dx_1)P(x_1; dx_2) \cdots P(x_{m-1}; dx_m)$.

Thus, (5) for the sets $A = \{\omega\colon x_0 \in B_0, x_1 \in B_1, \dots, x_n \in B_n\}$ and functons H of the form $H = H(x_0, x_1, \dots, x_m)$ is established. The case of general $A \in \mathscr{F}_n^X$ is treated (for a fixed m) in the same way as in Theorem 2 of Sect. 1.

It remains to show that the properties just proved remain true also for all $\mathscr{F}$ $(= \mathscr{E}^\infty)$-measurable bounded (or nonnegative) functions $H = H(x_0, x_1, \dots)$.

For that it suffices to prove that if $A \in \mathscr{F}_n^X$, then (5) holds true for such functions, i.e., that

$$\mathsf{E}_\pi(H(X_n, X_{n+1}, \dots); A) = \mathsf{E}_\pi(\mathsf{E}_{X_n} H(X_0, X_1, \dots); A). \tag{6}$$

Having in mind an application of the principle of appropriate sets (Sect. 2, Chap. 2, Vol. 1), denote by $\mathscr{H}$ the set of all bounded (or nonnegative) $\mathscr{F}$-measurable functions $H = H(x_0, x_1, \dots)$ for which (5) is true.

Denote by J the set of (cylindrical) sets of the form $I_m = \{\omega\colon x_0 \in B_0, \dots, x_m \in B_m\}$ with some $B_i \in \mathscr{E}$, $i = 0, 1, \dots, m$, $m \geq 0$. Clearly, J is a π-system of sets in $\mathscr{F}$ $(= \mathscr{E}^\infty)$.

To prove that it is also a λ-system, we turn to the conditions of Theorem 3 of Sect. 2, Chap. 2, Vol. 1.

Condition (h_1) is fulfilled because $I_A \in \mathscr{H}$ for $A \in J$ by what was proved earlier (take $H(x_0, \dots, x_m) = I_A(x_0, \dots, x_m)$ in (5)). Condition (h_2) follows from the additivity of the Lebesgue integral, and (h_3) from the monotone convergence theorem for Lebesgue integrals.

According to Theorem 5 mentioned earlier, $\mathscr{H}$ contains, then, all functions measurable with respect to $\sigma(J)$, which by definition is the σ-algebra $\mathscr{E}^\infty = \mathscr{B}(E^\infty)$ (Subsections 4 and 8, Sect. 2, Chap. 2, Vol. 1).

□

2. Now we proceed to another generalization of the Markov property, the so-called *strong Markov* property related to the change from "time n" to "random time τ." (The general setup will be the same as at the start of this section: $(\Omega, \mathscr{F}) = (E^\infty, \mathscr{E}^\infty)$ and so on.)

We will denote by $\tau = \tau(\omega)$ *finite* random variables $\tau(\omega)$ such that for any $n \geq 0$

$$\{\omega\colon \tau(\omega) = n\} \in \mathscr{F}_n^X.$$

According to the terminology used in Sect. 1, Chap. 7 (Definition 3), such a random variable is called a (finite) *Markov* or *stopping time*.

We will associate with the flow $(\mathscr{F}_n^X)_{n\geq 0}$ and the stopping time τ the σ-algebra

$$\mathscr{F}_\tau^X = \{A \in \mathscr{F}^X\colon A \cap \{\tau = n\} \in \mathscr{F}_n^X \text{ for all } n \geq 0\},$$

where the σ-algebra $\mathscr{F}^X = \sigma(\bigcup \mathscr{F}_n^X)$ is interpreted as the σ-algebra of events observed on the "random interval" $[0, \tau]$.

Theorem 2. *Suppose that the conditions of Theorem* 1 *are fulfilled, and let* $\tau = \tau(\omega)$ *be a finite Markov time. Then the following strong Markov property holds*:

$$\mathsf{E}_\pi(H \circ \theta_\tau \mid \mathscr{F}_\tau^X) = \mathsf{E}_{X_\tau} H \quad (\mathsf{P}_\pi\text{-}a.s.). \tag{7}$$

Before we proceed to the proof, let us comment on how $\mathsf{E}_{X_\tau} H$ and $H \circ \theta_\tau$ must be understood.

Let $\psi(x) = \mathsf{E}_x H$. (We pointed out in Subsection 1 that $\psi(x)$ is a $\mathscr{E}$-measurable function of x.) By $\mathsf{E}_{X_\tau} H$ we mean $\psi(X_\tau) = \psi(X_{\tau(\omega)}(\omega))$. As concerns $(H \circ \theta_\tau)(\omega)$, this is the random variable $(H \circ \theta_{\tau(\omega)})(\omega) = H(\theta_{\tau(\omega)}(\omega))$.

PROOF. Take a set $A \in \mathscr{F}_\tau$. As in Theorem 1, for the proof of (7) we must show that

$$\mathsf{E}_\pi(H \circ \theta_\tau; A) = \mathsf{E}_\pi(\mathsf{E}_{X_\tau} H; A). \tag{8}$$

Consider the left-hand side. We have

$$\begin{aligned} \mathsf{E}_\pi(H \circ \theta_\tau; A) &= \sum_{n=0}^{\infty} \mathsf{E}_\pi(H \circ \theta_\tau; A \cap \{\tau = n\}) \\ &= \sum_{n=0}^{\infty} \mathsf{E}_\pi(H \circ \theta_n; A \cap \{\tau = n\}). \end{aligned} \tag{9}$$

The right-hand side of (8) is

$$\mathsf{E}_\pi(\mathsf{E}_{X_\tau} H; A) = \sum_{n=0}^{\infty} \mathsf{E}_\pi(\mathsf{E}_{X_\pi} H; A \cap \{\tau = n\}). \tag{10}$$

Obviously, $A \cap \{\tau = n\} \in \mathscr{F}_n^X$. Therefore, in view of (4), the right-hand sides in (9) and (10) are the same, which proves the strong Markov property (7).

□

Corollary. If we let $H(x_0, x_1, \dots) = I_A(x_0, x_1, \dots)$, where $A = \{\omega\colon (x_0, x_1, \dots) \in B\}$, $B \in \mathscr{E}^\infty = \mathscr{B}(E^\infty)$, we obtain from (7) the following widely used form of the strong Markov property:

$$\begin{aligned} &\mathsf{P}_\pi((X_\tau, X_{\tau+1}, \dots) \in B \,|\, X_0, X_1, \dots, X_\tau) \\ &\quad = \mathsf{P}_{X_\tau}\{(X_0, X_1, \dots) \in B\} \quad (\mathsf{P}_\pi\text{-a.s.}). \end{aligned} \tag{11}$$

Remark 1. If we analyze the proof of the strong Markov property (7), we can see that in fact the following property also holds.

Let for any $n \geq 0$ the real-valued functions $H_n = H_n(\omega)$ defined on $\Omega = E^\infty$ be $\mathscr{F}$-measurable ($\mathscr{F} = \mathscr{E}^\infty$) and uniformly bounded (i.e., $|H_n(\omega)| \leq c$, $n \geq 0$, $\omega \in \Omega$). Then for any finite Markov time $\tau = \tau(\omega)$ ($\tau(\omega) < \infty$, $\omega \in \Omega$) the following form of the strong Markov property holds (Problem 2):

$$\mathsf{E}_\pi[\Psi_\tau \,|\, \mathscr{F}_\tau^X] = \psi(\tau, X_\tau) \quad (\mathsf{P}_\pi\text{-a.s.}), \tag{12}$$

where $\Psi(\omega) = H_n(\theta_n(\omega))$, $\psi(n, x) = \mathsf{E}_x H_n$ (see [21]).

Remark 2. We assumed earlier that $\tau = \tau(\omega)$ is a *finite* Markov time. If this is not the case, i.e., $\tau(\omega) \leq \infty$, $\omega \in \Omega$, then (12) must be changed as follows (Problem 3):

$$\mathsf{E}_\pi[\Psi_\tau \,|\, \mathscr{F}_\tau^X] = \psi(\tau, X_\tau) \quad (\{\tau < \infty\}; \mathsf{P}_\pi\text{-a.s.}). \tag{13}$$

In other words, in this case, (12) holds P_π-a.s. on the set $\{\tau < \infty\}$.

3. Example (Related to the strong Markov property). When dealing with the law of iterated logarithm we used an *inequality* (Lemma 1 in Sect. 4, Chap. 4, see also (14) in what follows) whose counterpart for the Brownian motion $B = (B_t)_{t \leq T}$ is the *equality* $\mathsf{P}\{\max_{0 \leq t \leq T} B_t > a\} = 2\,\mathsf{P}\{|B_T| > a\}$ ([12, Chap. 3]).

Let $\xi_1, \xi_2, \dots$ be a sequence of independent identically distributed random variables with symmetric (about zero) distribution. Let $X_0 = x \in R$, $X_m = X_0 + (\xi_1 + \cdots + \xi_m)$, $m \geq 1$. As before, we denote by P_x the probability distribution of the sequence $X = (X_m)_{m \geq 0}$ with $X_0 = x$. (The space Ω is assumed to be specified coordinate-wise, $\omega = (x_0, x_1, \dots)$ and $X_m(\omega) = x_m$.)

According to (slightly modified) inequality (9) of Sect. 4, Chap. 4,

$$\mathsf{P}_0\Big\{\max_{0 \leq m \leq n} X_m > a\Big\} \leq 2\,\mathsf{P}_0\{X_n > a\} \tag{14}$$

for any $a > 0$.

Define the Markov time $\tau = \tau(\omega)$ by

$$\tau(\omega) = \inf\{0 \le m \le n\colon X_m(\omega) > a\}. \tag{15}$$

(As usual, we set $\inf \varnothing = \infty$.) Let us demonstrate an "easy proof" of (14) using this Markov time, which would be valid if such a (random) time could be treated in the same manner as if it were nonrandom. We have (cf. proof of Lemma 1 in Sect. 4, Chap. 4)

$$\begin{aligned}\mathsf{P}_0\{X_n > a\} &= \mathsf{P}_0\{(X_n - X_{\tau\wedge n}) + X_{\tau\wedge n} > a\}\\ &\ge \mathsf{P}_0\{X_n - X_{\tau\wedge n} \ge 0, X_{\tau\wedge n} > a\} = \mathsf{P}_0\{X_n - X_{\tau\wedge n} \ge 0\}\ \mathsf{P}_0\{X_{\tau\wedge n} > a\}\\ &\ge \tfrac{1}{2}\,\mathsf{P}_0\{X_{\tau\wedge n} > a\} = \tfrac{1}{2}\,\mathsf{P}_0\{\tau \le n\} = \tfrac{1}{2}\,\mathsf{P}_0\Big\{\max_{0\le m\le n} X_m > a\Big\},\end{aligned} \tag{16}$$

where we have used the seemingly "almost obvious" property that $X_n - X_{\tau\wedge n}$ and $X_{\tau\wedge n}$ are *independent*, which is true for a *deterministic* time τ but is, in general, false for a random τ (Problem 4). (This means that our "easy proof" is incorrect.)

Now we give a correct proof of (14) based on the strong Markov property (13).

Since $\{X_n > a\} \subseteq \{\tau \le n\}$, we have

$$\mathsf{P}_0\{X_n > a\} = \mathsf{E}_0(I_{\{X_n > a\}}; \tau \le n). \tag{17}$$

Define the functions $H_m = H_m(x_0, x_1, \dots)$ by setting

$$H_m(x_0, x_1, \dots) = \begin{cases} 1, & \text{if } m \le n \text{ and } x_{n-m} > a,\\ 0 & \text{otherwise.}\end{cases}$$

It follows from this definition that on the set $\{\tau \le n\}$

$$(H_\tau \circ \theta_\tau)\,(x_0, x_1, \dots) = \begin{cases} 1, & \text{if } x_n > a,\\ 0 & \text{otherwise.}\end{cases} \tag{18}$$

Since $\{X_n > a\} \subseteq \{\tau \le n\}$ and $\{\tau \le n\} \in \mathscr{F}_\tau$, we obtain from (17)

$$\mathsf{P}_0\{X_n > a\} = \mathsf{E}_0(H_\tau \circ \theta_\tau; \tau \le n) = \mathsf{E}_0(\mathsf{E}_0(H_\tau \circ \theta_\tau \mid \mathscr{F}_\tau); \tau \le n). \tag{19}$$

According to the strong Markov property (13), we have

$$\mathsf{E}_0(H_\tau \circ \theta_\tau \mid \mathscr{F}_\tau) = \psi(\tau, X_\tau) \quad (\mathsf{P}_0\text{-a.s.}) \tag{20}$$

on the set $\{\tau \le n\}$. By definition, $\psi(m, x) = \mathsf{E}_x H_m$, and we obtain for $x \ge a$

$$\mathsf{E}_x H_m = \mathsf{P}_x\{X_{n-m} > a\} \ge \mathsf{P}_x\{X_{n-m} > x\} \ge \tfrac{1}{2}$$

(the last inequality follows from the symmetry of the distributions of $\xi_1, \xi_2, \dots$).

Hence

$$\mathsf{E}_0(H_\tau \circ \theta_\tau \,|\, \mathscr{F}_\tau) \geq \frac{1}{2} \quad (\mathsf{P}\text{-a.s.}) \tag{21}$$

on the set $\{\tau \leq n\}$. Together with (19) and (20), this implies the required inequality (14).

4. If we compare the Kolmogorov–Chapman Eq. (13) with Eq. (38), both in Sect. 12, Chap. 1, Vol. 1, we can observe that they are very similar. Therefore it is of interest to analyze the common points and the differences in their statements and proofs. (We restrict ourselves to homogeneous Markov chains with discrete state space E.)

Using (1) and (2), we obtain for $n \geq 1$, $1 \leq k \leq n$, and $i, j \in E$, that

$$\begin{aligned}
\mathsf{P}_i\{X_n = j\} &= \sum_{\alpha \in E} \mathsf{P}_i\{X_n = j, X_k = \alpha\} = \sum_{\alpha \in E} \mathsf{E}_i\, I(X_n = j) I(X_k = \alpha) \\
&= \sum_{\alpha \in E} \mathsf{E}_i[\mathsf{E}_i(I(X_n = j) I(X_k = \alpha) \,|\, \mathscr{F}_k)] \\
&= \sum_{\alpha \in E} \mathsf{E}_i[I(X_k = \alpha)\ \mathsf{E}_i(I(X_n = j) \,|\, \mathscr{F}_k)] \\
&\overset{(1)}{=} \sum_{\alpha \in E} \mathsf{E}_i[I(X_k = \alpha)\ \mathsf{E}_i(I(X_{n-k} = j) \circ \theta_k \,|\, \mathscr{F}_k)] \\
&\overset{(2)}{=} \sum_{\alpha \in E} \mathsf{E}_i[I(X_k = \alpha)\ \mathsf{E}_{X_k}\, I(X_{n-k} = j)] \\
&= \sum_{\alpha \in E} \mathsf{E}_i[I(X_k = \alpha)\ \mathsf{E}_\alpha\, I(X_{n-k} = j)] \\
= \sum_{\alpha \in E} \mathsf{E}_i\, I(X_k = \alpha)\ \mathsf{E}_\alpha\, I(X_{n-k} = j) &= \sum_{\alpha \in E} \mathsf{P}_i\{X_k = \alpha\}\ \mathsf{P}_\alpha\{X_{n-k} = j\},
\end{aligned} \tag{22}$$

which is exactly the Kolmogorov–Chapman Eq. (13) as in Sect. 12, Chap. 1, Vol. 1, written there as

$$p_{ij}^{(n)} = \sum_{\alpha \in E} p_{i\alpha}^{(k)} p_{\alpha j}^{(n-k)}.$$

If we replace the time k in (22) with a Markov time τ (taking values $1, 2, \ldots, n$) and use the strong Markov property (7) instead of Markov property (2), we obtain (Problem 5) the following natural (generalized) form of the Kolmogorov–Chapman equation:

$$\mathsf{P}_i\{X_n = j\} = \sum_{\alpha \in E} \mathsf{P}_i\{X_\tau = \alpha\}\ \mathsf{P}_\alpha\{X_{n-\tau} = j\}. \tag{23}$$

Both in (22) and (23) the summation is done over the *phase* variable $\alpha \in E$, whereas in (38) of Sect. 12, Chap. 1, Vol. 1, the summation is over the *time* variable.

Having noticed this, assume that τ is a Markov time with values in $\{1, 2, \ldots\}$. Starting as in the derivation of (38) given earlier, we find that

$$\begin{aligned}
\mathsf{P}_i\{X_n = j\} &= \sum_{k=1}^{n} \mathsf{P}_i\{X_n = j, \tau = k\} + \mathsf{P}_i\{X_n = j, \tau \geq n+1\} \\
&= \sum_{k=1}^{n} \mathsf{E}_i\, I(X_n = j) I(\tau = k) + \mathsf{P}_i\{X_n = j, \tau \geq n+1\} \\
&= \sum_{k=1}^{n} \mathsf{E}_i[\mathsf{E}_i(I(X_n = j) I(\tau = k) \mid \mathscr{F}_k)] + \mathsf{P}_i\{X_n = j, \tau \geq n+1\} \\
&= \sum_{k=1}^{n} \mathsf{E}_i[I(\tau = k)\ \mathsf{E}_i(I(X_n = j) \mid \mathscr{F}_k)] + \mathsf{P}_i\{X_n = j, \tau \geq n+1\} \\
&= \sum_{k=1}^{n} \mathsf{E}_i[I(\tau = k)\ \mathsf{E}_i(I(X_{n-k} = j) \circ \theta_k \mid \mathscr{F}_k)] + \mathsf{P}_i\{X_n = j, \tau \geq n+1\} \\
&= \sum_{k=1}^{n} \mathsf{E}_i[I(\tau = k)\ \mathsf{E}_{X_k} I(X_{n-k} = j)] + \mathsf{P}_i\{X_n = j, \tau \geq n+1\}. \qquad (24)
\end{aligned}$$

In Subsection 7 of Sect. 12, Chap. 1, Vol. 1, the role of τ was

$$\tau_j = \min\{1 \leq k \leq n\colon X_k = j\}$$

with the condition that $\tau_j = n+1$ if the set $\{\cdot\} = \varnothing$. In this case, (24) simplifies to

$$\begin{aligned}
\mathsf{P}_i\{X_n = j\} &= \sum_{k=1}^{n} \mathsf{E}_i(I(\tau_j = k)\ \mathsf{E}_{X_{\tau_j}} I(X_{n-k} = j)) \\
&= \sum_{k=1}^{n} \mathsf{E}_i(I(\tau_j = k)\ \mathsf{E}_j I(X_{n-k} = j)) = \sum_{k=1}^{n} \mathsf{E}_i I(\tau_j = k)\ \mathsf{E}_j I(X_{n-k} = j) \\
&= \sum_{k=1}^{n} \mathsf{P}_i\{\tau_j = k\}\ \mathsf{P}_j\{X_{n-k} = j\}
\end{aligned}$$

to become Eq. (38) in Sect. 12, Chap. 1, Vol. 1:

$$p_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)}. \qquad (25)$$

Equation (24) makes it possible to also derive other useful formulas involving summation with respect to the time variable (in contrast to the Kolmogorov–Chapman equation). For example, consider the Markov time

$$\tau(\alpha) = \min\{1 \leq k \leq n\colon X_k = \alpha(k)\},$$

where the (deterministic) function $\alpha = \alpha(k)$, $1 \leq k \leq n$, and the Markov chain are such that $\mathsf{P}_i\{\tau(\alpha) \leq n\} = 1$ (for fixed i and n). Then (24) implies that

$$\mathsf{P}_i\{X_n = j\} = \sum_{k=1}^{n} \mathsf{E}_i[I(\tau(\alpha) = k)\ \mathsf{E}_{X_{\tau(\alpha)}}\, I(X_{n-k} = j)]$$
$$= \sum_{k=1}^{n} \mathsf{E}_i\, I(\tau(\alpha) = k)\ \mathsf{E}_{\alpha(k)}\, I(X_{n-k} = j),$$

i.e.,

$$\mathsf{P}_i\{X_n = j\} = \sum_{k=1}^{n} \mathsf{P}_i\{\tau(\alpha) = k\}\ \mathsf{P}_{\alpha(k)}\{X_{n-k} = j\}$$

(cf. (23)).

5. Problems

1. Prove that the function $\psi(x) = \mathsf{E}_x H$ introduced in the remark in Subsection 1 is $\mathscr{E}$-measurable.
2. Prove (12).
3. Prove (13).
4. Is the independence property of $X_n - X_{\tau\wedge n}$ and $X_{\tau\wedge n}$ in the example in Subsection 3 true?
5. Prove (23).

3. Limiting, Ergodic, and Stationary Probability Distributions for Markov Chains

1. As mentioned in Sect. 1, the problem of the *asymptotic* behavior of *memoryless* stochastic systems described by Markov chains is of great importance for the theory of Markov random processes. One of the reasons for that is the fact that, under very broad assumptions, the behavior of a Markov system "stabilizes" and the system reaches a "steady-state" regime.

The *limiting* behavior of homogeneous Markov chains $X = (X_n)_{n\geq 0}$ may be studied in different aspects. For example, one can explore the P_π-almost sure convergence for functionals of the form $\frac{1}{n}\sum_{m=0}^{n-1} f(X_m)$ as $n \to \infty$ for various functions $f = f(x)$, as was done in the ergodic theorem for strict-sense stationary random sequences (Theorem 3 in Sect. 3, Chap. 5). It is also of interest to investigate the conditions for the law of large numbers as in Sect. 12, Chap. 1, Vol. 1.

Instead of these types of questions concerning convergence *almost sure* or *in probability*, in our exposition we will be mostly interested in the *asymptotic behavior of the transition probabilities* $P^{(n)}(x; A)$ for n steps as $n \to \infty$ (see (10) in Sect. 1) and in the existence of *nontrivial stationary* (*invariant*) *measures* $q = q(A)$, i.e., measures such that $q(E) > 0$ and

$$q(A) = \int P(x; A)\, q(dx), \tag{1}$$

where $P(x; A)$ is the transition function (for one step).

Let us emphasize that definition (1) does not, in general, presume that $q = q(A)$ is a probability measure ($q(E) = 1$).

If this is a probability measure, it is said to be a*stationary* or *invariant distribution*. The meaning of this terminology is clear: If we take q for the initial distribution π, i.e., assume that $\mathsf{P}_q\{X_0 \in A\} = q(A)$, then (1) will imply that $\mathsf{P}_q\{X_n \in A\} = q(A)$ for any $n \geq 1$, i.e., this distribution remains *invariant in time*.

It is easy to come up with an example where there is *no* stationary *distribution* $q = q(A)$, but there *are* stationary measures.

EXAMPLE. Let $X = (X_n)_{n\geq 0}$ be the Markov chain generated by Bernoulli trials, i.e., $X_{n+1} = X_n + \xi_{n+1}$, where $\xi_1, \xi_2, \dots$ are a sequence of independent identically distributed random variables with $\mathsf{P}\{\xi_n = +1\} = p$, $\mathsf{P}\{\xi_n = -1\} = q$. Let $X_0 = x$, $x \in \{0, \pm 1, \dots\}$. It is clear that the transition function here is

$$P(x; \{x+1\}) = p, \quad P(x; \{x-1\}) = q.$$

It is not hard to verify that one of the solutions to (1) is the measure $q(A)$ such that $q(\{x\}) = 1$ for any $x \in \{0, \pm 1, \dots\}$. If $p \neq q > 0$, then $q(A)$ with $q(\{x\}) = (p/q)^x$ is *another* invariant measure. It is obvious that *neither* of them is a *probability measure*, and there is no invariant probability measure here.

This simple example shows that the *existence* of a stationary (invariant) distribution requires certain assumptions about Markov chains.

The main interest in the problem of *convergence* of transition probabilities $P^{(n)}(x; A)$ as $n \to \infty$ lies in the *existence* of a limit that is *independent of the initial state* x. We must bear in mind that there may exist no limiting distribution at all, for example, it may happen that $\lim P^{(n)}(x; A) = 0$ for any $A \in \mathscr{E}$ and any initial state $x \in E$. For example, take $p = 1$ in the preceding example, i.e., consider the deterministic motion to the right. (See also Examples 4 and 5 in Sect. 8; cf. Problem 6 in Sect. 5.)

Establishing the conditions for the existence of stationary (invariant) distributions and the convergence of transition probabilities (and obtaining their properties) for *arbitrary* phase spaces $(E, \mathscr{E})$ is a very difficult problem (e.g., [9]). However, in the case of a *countable* state space (for "countable Markov chains"), interesting results in this area admit fairly transparent formulations. They will be presented in Sects. 6 and 7. But before that we will give a detailed classification of the states of countable Markov chains according to the algebraic and asymptotic properties of transition probabilities.

Let us point out that the questions concerning stationary distributions and the existence of the limits $\lim_n P^{(n)}(x; A)$ are closely related. Indeed, if the limit $\lim_n P^{(n)}(x; A)$ $(= \nu(A))$ *exists, does not depend on x, and is a measure* (in $A \in \mathscr{E}$), then we find from the *Kolmogorov–Chapman equation*

$$P^{(n+1)}(x; A) = \int P^{(n)}(x; dy)\, P(y; A)$$

by (formally) taking the limit as $n \to \infty$ that

$$\nu(A) = \int P(y; A)\, \nu(dy).$$

Thus $\nu = \nu(A)$ is then a *stationary* (*invariant*) *measure*.

2. Throughout the sequel, we assume that the Markov chains $X = (X_n)_{n \geq 0}$ under consideration take values in a countable phase space $E = \{1, 2, \dots\}$. For simplicity of notation, we will denote the transition functions $P(i, \{j\})$ by p_{ij} $(i, j \in E)$. The transition probabilities (of a randomly moving "particle") from state i to state j for n steps will be denoted by $p_{ij}^{(n)}$.

We will be interested in obtaining conditions under which the following statements hold true:

A. For all $j \in E$ *there exist* the limits

$$\pi_j = \lim_n p_{ij}^{(n)},$$

independent of the initial states $i \in E$;

B. These limiting values $\Pi = (\pi_1, \pi_2, \dots)$ form a probability *distribution*, i.e., $\pi_j \geq 0$ and $\sum_{j \in E} \pi_j = 1$;

C. The Markov chain is *ergodic*, i.e., the limiting values $\Pi = (\pi_1, \pi_2, \dots)$ are such that *all* $\pi_j > 0$ and $\sum_{j \in E} \pi_j = 1$;

D. There exists a unique *stationary* (*invariant*) probability distribution $\mathbb{Q} = (q_1, q_2, \dots)$, i.e., such that $q_j \geq 0$, $\sum_{i \in E} q_i = 1$, and

$$q_j = \sum_{i \in E} q_i p_{ij}$$

for all $j \in E$.

Remark. The term "ergodicity" used here appeared already in Chap. 5 (*ergodicity* as a metric transitivity property, the Birkhoff–Khinchin *ergodic theorem*). Formally, these terms are related to different objects, but their common feature is that they reflect the *asymptotic* behavior of various probabilistic characteristics as the time parameter goes to infinity.

3. Problems.

1. Give examples of Markov chains for which the limits $\pi_j = \lim_b p_{ij}^{(n)}$ exist and (a) are independent of the initial state j and (b) depend on j.
2. Give examples of ergodic and nonergodic Markov chains.
3. Give examples where the stationary distribution is not ergodic.

4. Classification of States of Markov Chains in Terms of Algebraic Properties of Matrices of Transition Probabilities

1. We will assume that the Markov chain under consideration has a countable set of states $E = \{1, 2, \dots\}$ and transition probabilities p_{ij}, $i, j \in E$. The matrix of these probabilities will be denoted by $\mathbb{P} = \|p_{ij}\|$ or, in expanded form,

$$\mathbb{P} = \begin{Vmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ p_{i1} & p_{i2} & p_{i3} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{Vmatrix}.$$

(Sometimes we will write $(\cdot)$ instead of $\|\cdot\|$ to denote matrices.)

In what follows we give the classification of the states of a Markov chain in terms of the *algebraic* properties of the matrices of transition probabilities $\mathbb{P}$ and $\mathbb{P}^{(n)}$, $n \geq 1$.

The matrix of transition probabilities $\mathbb{P}$ completely determines transitions for *one step* from one state to another, while transitions for *n steps* are determined (due to the Markov property) by the matrices $\mathbb{P}^{(n)} = \|p_{ij}^{(n)}\|$.

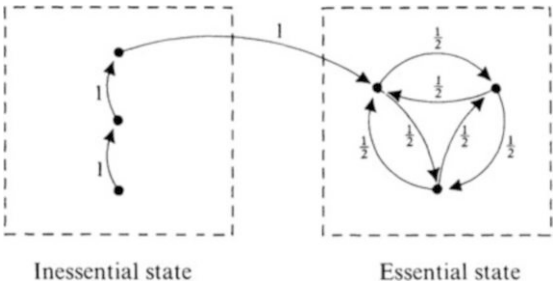


Fig. 36 Inessential and essential states

For example, the matrix

$$\mathbb{P} = \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \end{pmatrix}$$

and the corresponding graph (Sect. 12, Chap. 1, Vol. 1) show that in the random walk over states 0 and 1 driven by this matrix, the move $0 \to 1$ for one step is possible (with probability $1/2$), whereas the move $1 \to 0$ is impossible. Clearly, the transition $1 \to 0$ is impossible for any number of steps, which can be seen from the structure of the matrices

$$\mathbb{P}^{(n)} = \begin{Vmatrix} 2^{-n} & 1 - 2^{-n} \\ 0 & 1 \end{Vmatrix}$$

showing that $p_{10}^{(n)} = 0$ for every $n \geq 1$.

State 1 in this example is such that the particle can *enter* into it (from state 0), but cannot *leave* it.

Consider the graph in Fig. 36, from which one can easily recover the transition matrix $\mathbb{P}$. It is clear from this graph that there are three states (the left-hand part of the figure) such that leaving any of them, there is no way to return back.

With regard of the future behavior of the "particle" wandering in accordance with this graph, these three states are *inessential* because the particle can *leave* them but *cannot return* anymore.

These "inessential" states are of little interest, and we can discard them to focus our attention on the classification of the remaining "essential" states. (This descriptive definition of "inessential" and "essential" states can be formulated precisely in terms of the transition probabilities $p_{ij}^{(n)}$, $i,j \in E$, $n \geq 1$, Problem 1.)

2. To classify essential states or groups of such states, we need some definitions.

Definition 1. We say that state j is *accessible* from point i (notation: $i \to j$) if there is $n \geq 0$ such that $p_{ij}^{(n)} > 0$ (with $p_{ij}^{(0)} = 1$ if $i = j$ and 0 if $i \neq j$).

States i and j *communicate* (notation: $i \leftrightarrow j$) if $i \to j$ and $j \to i$, i.e., if they are *mutually accessible*.

Lemma 1. *The property of states to communicate* ($\leftrightarrow$) *is an equivalence relation between states of the Markov chain with transition probabilities matrix* $\mathbb{P}$.

PROOF. By the definition of the *equivalence relation* (in this case "$\leftrightarrow$"), we must verify that it is *reflexive* ($i \leftrightarrow i$), *symmetric* (if $i \leftrightarrow j$, then $j \leftrightarrow i$), and *transitive* (if $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$).

The first two properties follow directly from the definition of *communicating* states. Transitivity follows from the Kolmogorov–Chapman equation: If $p_{ij}^{(n)} > 0$ and $p_{jk}^{(m)} > 0$, then

$$p_{ik}^{(n+m)} = \sum_{l \in E} p_{il}^{(n)} p_{lk}^{(m)} \geq p_{ij}^{(n)} p_{jk}^{(m)} > 0,$$

i.e., $i \to k$. In a similar way, $k \to i$, hence $i \leftrightarrow k$.

□

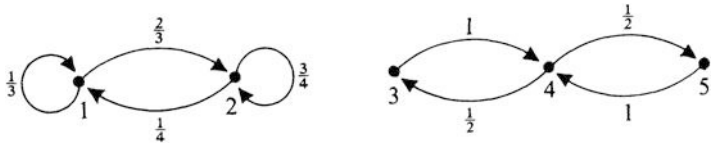
We will gather the states $i, j, k, \ldots$ that communicate with each other ($i \leftrightarrow j, j \leftrightarrow k, k \leftrightarrow i, \ldots$) into the same *class*. Then any two such classes of states either coincide or are disjoint. Thus, the relation that two states may communicate induces a *partition* of the set of (essential) states E into a finite or countable set of disjoint classes $E_1, E_2, \ldots$ ($E = E_1 + E_2 + \ldots$).

These classes will be called *indecomposable classes* (of essential communicating) states. A Markov chain whose states form a *single* indecomposable class will be said to be *indecomposable*.

As an illustration we consider the chain with state space $E = \{1, 2, 3, 4, 5\}$ and the matrix of transition probabilities

$$\mathbb{P} = \left(\begin{array}{cc|ccc} 1/3 & 2/3 & 0 & 0 & 0 \\ 1/4 & 3/4 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{array}\right) = \left(\begin{array}{c|c} \mathbb{P}_1 & 0 \\ \hline 0 & \mathbb{P}_2 \end{array}\right).$$

The graph of this chain 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 its properties reduces to the investigation of the two separate chains with state spaces E_1 and E_2 and transition matrices $\mathbb{P}_1$ and $\mathbb{P}_2$.

Now let us consider any indecomposable class E, for example, the one sketched in Fig. 37.

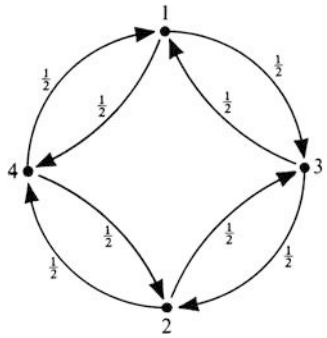


Fig. 37 Example of a Markov chain with period $d = 2$

Observe that in this case a return to each state is possible only after an *even* number of steps, and a transition to an adjacent state after an *odd* number. The transition matrix has a block structure,

$$\mathbb{P} = \left(\begin{array}{cc|cc} 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ \hline 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{array}\right).$$

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.

3. This example suggests that, in general, it is possible to give a classification of *indecomposable* classes into *cyclic subclasses*.

To this end we will need some definitions and a fact from number theory.

Definition 2. Let $\varphi = (\varphi_1, \varphi_2, \dots)$ be a sequence of nonnegative numbers $\varphi_n \geq 0$, $n \geq 1$. The *period* of the sequence φ (notation: $d(\varphi)$) is the number

$$d(\varphi) = \mathrm{GCD}(M_\varphi) = \mathrm{GCD}\{\, n \geq 1 \colon \varphi_n > 0 \,\},$$

where $\mathrm{GCD}(M_\varphi)$ is the *greatest common divisor* of the set M_φ of indices $n \geq 1$ for which $\varphi_n > 0$; if $\varphi_n = 0$, $n \geq 1$, then $M_\varphi = \varnothing$, and $\mathrm{GCD}(M_\varphi)$ is set to be zero.

In other words, the period of a sequence φ is $d(\varphi)$ if n is divisible by $d(\varphi)$ whenever $\varphi_n > 0$ (i.e., n equals $d(\varphi)k$ with some $k \geq 1$) and $d(\varphi)$ is the greatest number among d with this property, i.e., such that $n = dl$ for some integer $l \geq 1$.

For example, the sequence $\varphi = (\varphi_1, \varphi_2, \dots)$ such that $\varphi_{4k} > 0$ for $k = 1, 2, \dots$ and $\varphi_n = 0$ for $n \neq 4k$, has period $d(\varphi) = 4$ rather than 2, although $\varphi_{2l} > 0$ for $l = 2, 4, 8$.

Definition 3. A sequence $\varphi = (\varphi_1, \varphi_2, \dots)$ is *aperiodic* if its period $d(\varphi) = 1$.

The following elementary result of number theory will be useful in the sequel for the classification of states in terms of the cyclicity property.

Lemma 2. *Let M be a set of nonnegative integers ($M \subseteq E$) closed with respect to addition and such that* $\mathrm{GCD}(M) = 1$.

Then there is an n_0 such that M contains all numbers $n \geq n_0$.

We will apply this lemma to the set $M = M_\varphi$ taking as the sequence $\varphi = (\varphi_1, \varphi_2, \dots)$ the sequence $(p_{jj}^{(1)}, p_{jj}^{(2)}, \dots)$ or the sequence $(p_{jj}^{(d)}, p_{jj}^{(2d)}, \dots)$, $d \geq 1$, where j is a state of the Markov chain with a matrix of transition probabilities $\mathbb{P} = \|p_{ij}\|$, and $p_{jj}^{(n)}$ is an element of the matrix $\mathbb{P}^{(n)}$, $n \geq 1$, $\mathbb{P}^{(1)} = \mathbb{P}$. (We will say that a *state j* has period $d(j)$ if $d(j)$ is the period of the sequence $(p_{jj}^{(1)}, p_{jj}^{(2)}, \dots)$.) Then we obtain the following result.

Theorem 1. *Let a state j have period $d = d(j)$.*

If $d = 1$, then there is $n_0 = n_0(j)$ such that the transition probabilities $p_{jj}^{(n)} > 0$ for all $n \geq n_0$.

If $d > 1$, then there is $n_0 = n_0(j, d)$ such that $p_{jj}^{(nd)} > 0$ for all $n \geq n_0$.

If $d \geq 1$ and $p_{ij}^{(m)} > 0$ for some $i \in E$ and $m \geq 1$, then there is $n_0 = n_0(j, d, m)$ such that $p_{ij}^{(m+nd)} > 0$ for all $n \geq n_0$.

Now we state a theorem showing that the *periods* of the states of an indecomposable class are of the same "type."

Theorem 2. *Let $E_* = \{i, j, \dots\}$ be an indecomposable class of (communicating) states of set E.*

Then all the states of this class are "of the same type" in the sense that they have the same period (denoted by $d(E_*)$) called the period of class E_*.

PROOF. Let $i, j \in E_*$. Then there are k and l such that $p_{ij}^{(k)} > 0$ and $p_{ji}^{(l)} > 0$. But, by the Kolmogorov–Chapman equation, we have then

$$p_{ii}^{(k+l)} = \sum_{a \in E} p_{ia}^{(k)} p_{ai}^{(l)} \geq p_{ij}^{(k)} p_{ji}^{(l)} > 0,$$

hence $k + l$ must be divisible by $d(i)$, the period of the state $i \in E_*$.

Let $d(j)$ be the period of the state $j \in E_*$, and let n be such that $p_{jj}^{(n)} > 0$. Then n must be divisible by $d(j)$, and since

$$p_{ii}^{(n+k+l)} \geq p_{ij}^{(k)} p_{jj}^{(n)} p_{ji}^{(l)} > 0,$$

we obtain that $n + k + l$ is divisible by $d(i)$. But $k + l$ is divisible by $d(i)$, hence n is divisible by $d(i)$, and since $d(j) = \text{GCD}\{ n \colon p_{jj}^{(n)} > 0 \}$, we have $d(i) \leq d(j)$.

By symmetry, $d(j) \leq d(i)$, hence $d(i) = d(j)$.

□

4. If a set $E_* \subseteq E$ forms an indecomposable class of (communicating) states and $d(E_*) = 1$, then it is said to be an *aperiodic class* of states.

Now we consider the case $d(E_*) > 1$. The transitions within such a class may be quite freakish (as in the preceding example of a Markov chain with period $d(E_*) = 2$, see Fig. 37). However, there is a *cyclic* character of the transitions from one group of states to another.

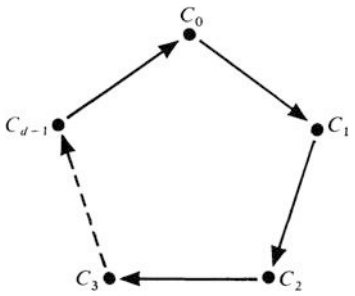


Fig. 38 Motion among cyclic subclasses

Theorem 3. *Let E_* be an indecomposable class of states, $E_* \subseteq E$, with period $d = d(E_*) > 1$.*

Then there are d groups of states $C_0, C_1, \ldots, C_{d-1}$, called cyclic subclasses $(E_* = C_0 + C_1 + \cdots + C_{d-1})$, such that at the time instants $n = p + kd$, with $p = 0, 1, \ldots, d-1$ and $k = 0, 1, \ldots$, the "particle" is in the subclass C_p with a transition at the next time to C_{p+1}, then to $C_{p+2}, \ldots, C_{d-1}$, then from C_{d-1} to C_0 and so on.

PROOF. Let us fix a state $i_0 \in E_*$ and define the following *subclasses*:

$$C_0 = \{j \in E_* : p_{i_0 j}^{(n)} > 0, n = kd,\ k = 0, 1, \dots\},$$
$$C_1 = \{j \in E_* : p_{i_0 j}^{(n)} > 0, n = kd + 1,\ k = 0, 1, \dots\},$$
$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$
$$C_{d-1} = \{j \in E_* : p_{i_0 j}^{(n)} > 0, n = kd + (d-1),\ k = 0, 1, \dots\}.$$

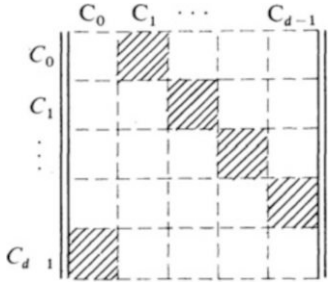
It is clear that $E_* = C_0 + C_1 + \dots + C_{d-1}$. Let us show that the "particle" moves from one subclass to another following the rule described in the theorem (Fig. 38).

In fact, consider a state $i \in C_p$, and let the state $j \in E_*$ be such that $p_{ij} > 0$. We will show that then $j \in C_{(p+1) \pmod d}$.

Let n be such that $p_{i_0 j}^{(n)} > 0$. Then n can be written as $n = p + kd$ with some $p = 0, 1, \dots, d-1$ and $k = 0, 1, \dots$. Hence $n \equiv p \pmod d$ and $n + 1 \equiv (p+1) \pmod d$. This implies that $p_{i_0 j}^{(n+1)} > 0$ (by the definition of the period $d = d(E_*)$), so that $j \in C_{(p+1) \pmod d}$, which was to be proved.

□

Let us observe that it now follows that the transition matrix $\mathbb{P}$ of an indecomposable chain has the following block structure:



Suppose now that the wandering particle whose evolution is driven by matrix $\mathbb{P}$ starts from a state in the subclass C_0. Then at each time $n = p + kd$ this particle will be (by the definition of subclasses $C_0, C_1, \dots, C_{d-1}$) in the set C_p.

Therefore with each set C_p of states we can associate a *new* Markov chain with transition matrix $\|p_{ij}^{(d)}\|$, where $i, j \in C_p$. This new chain will be *indecomposable* and *aperiodic*.

Thus, taking into account the foregoing classification (into inessential and essential states, indecomposable classes and cyclic subclasses; see Fig. 39), we can draw the following conclusion:

> To investigate the limiting behavior of transition probabilites $p_{ij}^{(n)}$, $n \geq 1$, $i, j \in E$, which determine the motion of the "Markov particle," we can restrict our attention to the case where the phase space E itself is a *unique indecomposable and aperiodic class* of states.

In this case, the Markov chain $X = (X_n)_{n \geq 0}$ itself with such a phase space and the matrix of transition probabilities $\mathbb{P}$ is called *indecomposable and aperiodic*.

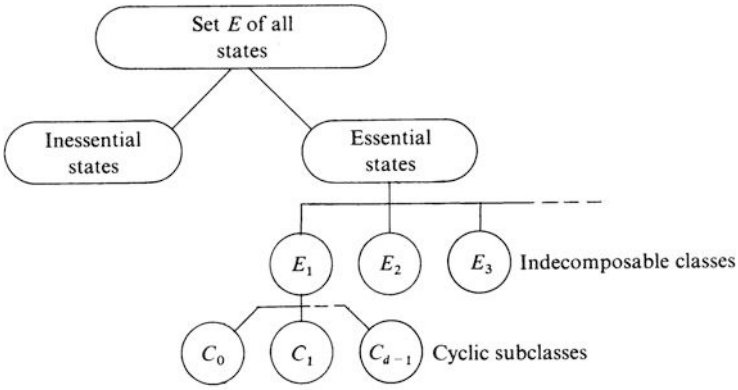


Fig. 39 Classification of states of a Markov chain in terms of arithmetic properties of probabilities $p_{ij}^{(n)}$

5. Problems

1. Give an accurate formulation in terms of transition probabilities $p_{ij}^{(n)}$, $i,j \in E$, $n \geq 1$, to the descriptive definition of inessential and essential states stated at the end of Subsection 1.
2. Let $\mathbb{P}$ be the matrix of transition probabilities of an indecomposable Markov chain with finitely many states. Let $\mathbb{P}^2 = \mathbb{P}$. Explore the structure of $\mathbb{P}$.
3. Let $\mathbb{P}$ be the matrix of transition probabilities of a finite Markov chain $X = (X_n)_{n\geq 0}$. Let $\sigma_1, \sigma_2, \dots$ be a sequence of independent identically distributed nonnegative integer-valued random variables independent of X, and let $\tau_0 = 0$, $\tau_n = \sigma_1 + \cdots + \sigma_n$, $n \geq 1$. Show that the sequence $\widetilde{X} = (\widetilde{X}_n)_{n\geq 0}$ with $\widetilde{X}_n = X_{\tau_n}$ is a Markov chain. Find the matrix $\widetilde{\mathbb{P}}$ of transition probabilities for this chain. Show that if states i and j communicate for the chain X, they do so for $\widetilde{X}$.
4. Consider a Markov chain with two states, $E = \{0, 1\}$, and the matrix of transition probabilities

$$\mathbb{P} = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix}, \quad 0 < \alpha < 1,\ 0 < \beta < 1.$$

Describe the structure of the matrices $\mathbb{P}^{(n)}$, $n \geq 2$.

5. Classification of States of Markov Chains in Terms of Asymptotic Properties of Transition Probabilities

1. Let $X = (X_n)_{n\geq 0}$ be a homogeneous Markov chain with countable state space $E = \{1, 2, \dots\}$ and transition probabilities $p_{ij} = \mathsf{P}_i\{X_1 = j\}$, $i,j \in E$.

Let

$$f_{ii}^{(n)} = \mathsf{P}_i\{X_n = i, X_k \neq i, 1 \leq k \leq n-1\} \tag{1}$$

and (for $i \neq j$)

$$f_{ij}^{(n)} = \mathsf{P}_i\{X_n = j, X_k \neq j, 1 \leq k \leq n-1\}. \tag{2}$$

It is clear that $f_{ii}^{(n)}$ is the probability of *first return* to state i at time n, while $f_{ij}^{(n)}$ is the probability of *first arrival* at state j at time n, provided that $X_0 = i$.

If we set

$$\sigma_i(\omega) = \min\{\, n \geq 1 \colon X_n(\omega) = i \,\} \tag{3}$$

with $\sigma_i(\omega) = \infty$ when the set in (3) is empty, then the probabilities $f_{ii}^{(n)}$ and $f_{ij}^{(n)}$ can be represented as

$$f_{ii}^{(n)} = \mathsf{P}_i\{\sigma_i = n\}, \quad f_{ij}^{(n)} = \mathsf{P}_i\{\sigma_j = n\}. \tag{4}$$

For $i, j \in E$ define the quantities

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}. \tag{5}$$

It is seen from (4) that

$$f_{ij} = \mathsf{P}_i\{\sigma_j < \infty\}. \tag{6}$$

In other words, f_{ij} is the probability that the "particle" leaving state i will ultimately arrive at state j.

In the sequel, of special importance is the probability f_{ii} that the "particle" leaving state i will ultimately *return* to this state. These probabilities are used in the following definitions.

Definition 1. A state $i \in E$ is *recurrent* if $f_{ii} = 1$.

Definition 2. A state $i \in E$ is *transient* if $f_{ii} < 1$.

There are the following conditions for recurrence and transience.

Theorem 1. (a) *The state $i \in E$ is recurrent if and only if either of the following two conditions is satisfied*:

$$\mathsf{P}_i\{X_n = i \ \ i.o.\} = 1 \quad or \quad \sum_n p_{ii}^{(n)} = \infty.$$

(b) *The state $i \in E$ is transient if and only if either of the following two conditions is satisfied*:

$$\mathsf{P}_i\{X_n = i \ \ i.o.\} = 0 \quad or \quad \sum_n p_{ii}^{(n)} < \infty.$$

Therefore, according to this theorem,

$$f_{ii} = 1 \iff \mathsf{P}_i\{X_n = i \ \text{ i.o.}\} = 1 \iff \sum_n p_{ii}^{(n)} = \infty, \tag{7}$$

$$f_{ii} < 1 \iff \mathsf{P}_i\{X_n = i \ \text{ i.o.}\} = 0 \iff \sum_n p_{ii}^{(n)} < \infty. \tag{8}$$

Remark. Recall that, according to Table 2.1 in Sect. 1, Chap. 2, Vol. 1, the event $\{X_n = i \text{ i.o.}\}$ is the set of outcomes ω for which $X_n(\omega) = i$ for infinitely many indices n. If we use the notation $A_n = \{\omega\colon X_n(\omega) = i\}$, then $\{X_n = i \text{ i.o.}\} = \bigcap_{n=1}^{\infty} \bigcup_{k=n} A_k$; see the table mentioned earlier.

PROOF. We can observe immediately that the implication

$$\sum_n p_{ii}^{(n)} < \infty \implies \mathsf{P}_i\{X_n = i \text{ i.o.}\} = 0 \tag{9}$$

follows from the Borel–Cantelli lemma since $p_{ii}^{(n)} = \mathsf{P}_i\{X_n = i\}$ (see statememt (a) of this lemma, Sect. 10, Chap. 2, Vol. 1).

Let us show that

$$f_{ii} = 1 \iff \sum_n p_{ii}^{(n)} = \infty. \tag{10}$$

The Markov property and homogeneity imply that for any collections $(i_1, \dots, i_k)$ and $(j_1, \dots, j_n)$

$$\begin{aligned}\mathsf{P}_i\big\{(X_1, \dots, X_k) = (i_1, \dots, i_k), (X_{k+1}, \dots, X_{k+n}) = (j_1, \dots, j_n)\big\} \\ = \mathsf{P}_i\big\{(X_1, \dots, X_k) = (i_1, \dots, i_k)\big\}\ \mathsf{P}_{i_k}\big\{(X_1, \dots, X_n) = (j_1, \dots, j_n)\big\}.\end{aligned}$$

This implies at once that (cf. derivation of (38) in Sect. 12, Chap. 1, Vol. 1 and (25) in Sect. 2 of this chapter)

$$\begin{aligned}p_{ij}^{(n)} = \mathsf{P}_i\{X_n = j\} &= \sum_{k=0}^{n-1} \mathsf{P}_i\{X_1 \neq j, \dots, X_{n-k-1} \neq j, X_{n-k} = j, X_n = j\} \\ &= \sum_{k=0}^{n-1} \mathsf{P}_i\{X_1 \neq j, \dots, X_{n-k-1} \neq j, X_{n-k} = j\}\ \mathsf{P}_j\{X_k = j\} \\ &= \sum_{k=0}^{n-1} f_{ij}^{(n-k)} p_{jj}^{(k)} = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)}.\end{aligned}$$

Thus

$$p_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)}. \tag{11}$$

Letting $j = i$ we find that (with $p_{ii}^{(0)} = 1$)

$$\begin{aligned}\sum_{n=1}^{\infty} p_{ii}^{(n)} = \sum_{n=1}^{\infty} \sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)} = \sum_{k=1}^{\infty} f_{ii}^{(k)} \sum_{n=k}^{\infty} p_{ii}^{(n-k)} &= f_{ii} \sum_{n=0}^{\infty} p_{ii}^{(n)} \\ &= f_{ii}\left(1 + \sum_{n=1}^{\infty} p_{ii}^{(n)}\right).\end{aligned} \tag{12}$$

Hence we see that

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty \implies f_{ii} = \frac{\sum\limits_{n=1}^{\infty} p_{ii}^{(n)}}{1 + \sum\limits_{n=1}^{\infty} p_{ii}^{(n)}}. \tag{13}$$

Let now $\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$. Then

$$\sum_{n=1}^{N} p_{ii}^{(n)} = \sum_{n=1}^{N} \sum_{k=1}^{n} f_{ii}^{(k)} p_{ii}^{(n-k)} = \sum_{k=1}^{N} f_{ii}^{(k)} \sum_{n=k}^{N} p_{ii}^{(n-k)} \le \sum_{k=1}^{N} f_{ii}^{(k)} \sum_{l=0}^{N} p_{ii}^{(l)},$$

hence

$$f_{ii} = \sum_{k=1}^{\infty} f_{ii}^{(k)} \ge \sum_{k=1}^{N} f_{ii}^{(k)} \ge \frac{\sum\limits_{n=1}^{N} p_{ii}^{(n)}}{\sum\limits_{l=0}^{N} p_{ii}^{(l)}} \to 1, \quad N \to \infty.$$

Thus,

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty \implies f_{ii} = 1. \tag{14}$$

The implications (13) and (14) imply the following equivalences:

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} < \infty \iff f_{ii} < 1, \tag{15}$$

$$\sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty \iff f_{ii} = 1. \tag{16}$$

To complete the proof, it remains to show that

$$f_{ii} < 1 \iff \mathsf{P}_i\{X_n = i \text{ i.o.}\} = 0, \tag{17}$$

$$f_{ii} = 1 \iff \mathsf{P}_i\{X_n = i \text{ i.o.}\} = 1. \tag{18}$$

These properties are easily comprehensible from an intuitive point of view. For example, if $f_{ii} = 1$, then $\mathsf{P}_i\{\sigma_i < \infty\} = 1$, i.e., the "particle" sooner or later will return to the same state i from where it started its motion. But then, by the strong Markov property, the "life of the particle" starts anew from this (random) time. Continuing this reasoning, we obtain that the events $\{X_n = i\}$ will occur for infinitely many indices n, i.e., $\mathsf{P}_i\{X_n = i \text{ i.o.}\} = 1$.

Let us give a formal proof of (17) and (18). For a given state $i \in E$, consider the probability that *the number of returns to i* is *greater than or equal to m*. We claim that this probability is equal to $(f_{ii})^m$.

Indeed, for $m = 1$ this follows from the definition of f_{ii}. Suppose that our claim is true for $m - 1$. We will show that the probability of interest is then equal to $(f_{ii})^m$.

By the strong Markov property (see (8) in Sect. 2) and since $\{\sigma_i = k\} \in \mathscr{F}_{\sigma_i}$, we find

$$
\begin{aligned}
&\mathsf{P}_i(\text{the number of returns to } i \text{ is at least } m)\\
&= \sum_{k=1}^{\infty} \mathsf{P}_i(\sigma_i = k \text{ and there are at least } m-1 \text{ returns to } i \text{ after time } k)\\
&= \sum_{k=1}^{\infty} \mathsf{P}_i\{\sigma_i = k\}\ \mathsf{P}_i(\text{at least } m-1 \text{ of } X_{\sigma_i+1}, X_{\sigma_i+2}, \dots \text{ are equal to } i \,|\, \sigma_i = k)\\
&= \sum_{k=1}^{\infty} \mathsf{P}_i\{\sigma_i = k\}\ \mathsf{P}_i(\text{at least } m-1 \text{ of } X_1, X_2, \dots \text{ are equal to } i)\\
&= \sum_{k=1}^{\infty} f_{ii}^{(k)} (f_{ii})^{m-1} = f_{ii}(f_{ii})^{m-1} = (f_{ii})^m.
\end{aligned}
$$

This implies that

$$
\mathsf{P}_i\{X_n = i \text{ i.o.}\} = \lim_{m\to\infty} (f_{ii})^m = \begin{cases} 1, & \text{if } f_{ii} = 1,\\ 0, & \text{if } f_{ii} < 1. \end{cases} \tag{19}
$$

Using the notation $A = \{A_n \text{ i.o.}\}$ $(= \limsup A_n)$, where $A_n = \{X_n = i\}$, we see from (19) that $\mathsf{P}_i(A)$ obeys the "0 or 1 law," i.e., $\mathsf{P}_i(A)$ can take only *two* values 0 or 1. (Note that this property *does not follow* directly from statements (a) and (b) of the Borel–Cantelli lemma (Sect. 10, Chap. 2, Vol. 1) since the events A_n, $n \geq 1$, are, in general, *dependent*.)

Equation (19) and the property that $\mathsf{P}_i(A)$ can take only the values 0 and 1 imply the required implications in (17) and (18).

□

2. The theorem just proved implies the following simple, but important, property of *transient* states.

Theorem 2. *If a state j is transient, then for any $i \in E$*

$$
\sum_{n=1}^{\infty} p_{ij}^{(n)} < \infty, \tag{20}
$$

hence, for any $i \in E$,

$$
p_{ij}^{(n)} \to 0, \quad n \to \infty. \tag{21}
$$

PROOF. We have from (11) (with $p_{jj}^{(0)} = 1$)

$$
\sum_{n=1}^{\infty} p_{ij}^{(n)} = \sum_{n=1}^{\infty} \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)} = \sum_{k=1}^{\infty} f_{ij}^{(k)} \sum_{n=0}^{\infty} p_{jj}^{(n)} = f_{ij} \sum_{n=0}^{\infty} p_{jj}^{(n)} \leq \sum_{n=0}^{\infty} p_{jj}^{(n)} < \infty,
$$

where we have used that $f_{ij} = \sum_{k=1}^{\infty} f_{ij}^{(k)} \leq 1$ (being the probability that the particle leaving state i will ultimately arrive at state j).

Property (21) obviously follows from (20).

□

3. Now we proceed to *recurrent* states.

Every recurrent state $i \in E$ can be classified according to whether the *average time of return* to this state

$$\mu_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)} \quad (= \mathsf{E}_i \, \sigma_i) \tag{22}$$

is finite or infinite. (Recall that, according to (1), $f_{ii}^{(n)}$ is the probability of the first return for exactly n steps.)

Definition 3. Let us say that a recurrent state $i \in E$ is *positive* if

$$\mu_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)} < \infty \tag{23}$$

and *null* if

$$\mu_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)} = \infty. \tag{24}$$

Hence, according to this definition, the first return to a *null* (recurrent) state requires (in average) infinite time. Alternatively, the average time of first return to a *positive* (recurrent) state is *finite*.

4. The following figure illustrates the classification of states of a Markov chain in terms of *recurrence* and *transience*, and *positive* and *null* recurrence (Fig. 40).

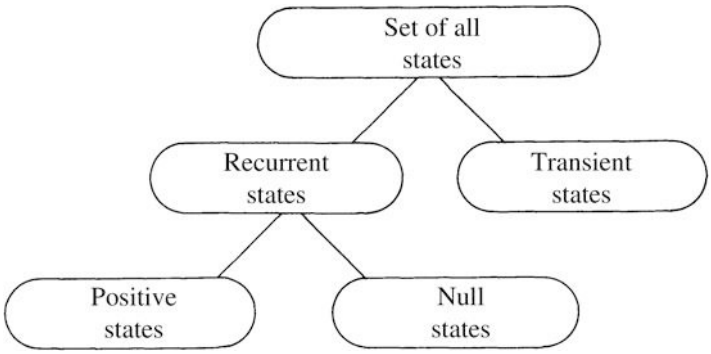


Fig. 40 Classification of states of Markov chain in terms of asymptotic properties of probabilities $p_{ii}^{(n)}$

5. Theorem 3. *Let a state $j \in E$ of a Markov chain be recurrent and aperiodic* $(d(j) = 1)$.

Then for any $i \in E$

$$p_{ij}^{(n)} \to \frac{f_{ij}}{\mu_j}, \quad n \to \infty. \tag{25}$$

If, moreover, states i and j communicate ($i \leftrightarrow j$), i.e., belong to the same indecomposable class, then

$$p_{ij}^{(n)} \to \frac{1}{\mu_j}, \quad n \to \infty. \tag{26}$$

The proof given below will rely on Lemma 1, which is one of the key results of "discrete renewal theory." For another proof of Theorem 5, based on the concept of coupling (Sect. 8, Chap. 3, Vol. 1), see, for example, [9, 35].

Lemma 1. (From "discrete renewal theory.") *Let $\varphi = (\varphi_1, \varphi_2, \dots)$ be an aperiodic sequence ($d(\varphi) = 1$) of nonnegative numbers and $u = (u_0, u_1, \dots)$ a sequence constructed by the following recurrence rule: $u_0 = 1$ and for every $n \geq 1$*

$$u_n = \varphi_1 u_{n-1} + \varphi_2 u_{n-2} + \cdots + \varphi_n u_0. \tag{27}$$

Then

$$u_n \to \mu^{-1}$$

as $n \to \infty$, where $\mu = \sum_{n=1}^{\infty} n\varphi_n$.

For the proof, see, for example, [25, XIII.10].

PROOF OF THEOREM 3. First, let us show for $i = j$ that

$$p_{jj}^{(n)} \to \frac{1}{\mu_j}, \quad n \to \infty. \tag{28}$$

To this end, we rewrite (11) (for $i = j$) as

$$p_{jj}^{(n)} = f_{jj}^{(1)} p_{jj}^{(n-1)} + f_{jj}^{(2)} p_{jj}^{(n-2)} + \cdots + f_{jj}^{(n)} p_{jj}^{(0)}, \tag{29}$$

where we set $p_{jj}^{(0)} = 1$ and, obviously, $f_{jj}^{(1)} = p_{jj}^{(1)}$. Letting

$$u_k = p_{jj}^{(k)}, \quad \varphi_k = f_{jj}^{(k)}, \tag{30}$$

(29) becomes

$$u_n = \varphi_1 u_{n-1} + \varphi_2 u_{n-2} + \cdots + \varphi_n u_0,$$

which is *exactly* the recurrence formula (27).

The required result (28) will follow directly from Lemma 1 if we show that the period $d_f(j)$ of the sequence $(f_{jj}^{(1)}, f_{jj}^{(2)}, \dots)$ is equal to 1, provided that the period of $(p_{jj}^{(1)}, p_{jj}^{(2)}, \dots)$ is 1.

This, in turn, follows from the following general result.

Lemma 2. *For any $j \in E$*

$$\mathrm{GCD}(n \geq 1 \colon p_{jj}^{(n)} > 0) = \mathrm{GCD}(n \geq 1 \colon f_{jj}^{(n)} > 0), \tag{31}$$

i.e., the periods $d_f(j)$ and $d(j)$ are the same.

PROOF. Let

$$M = \{\, n \colon p_{jj}^{(n)} > 0 \,\} \quad \text{and} \quad M_f = \{\, n \colon f_{jj}^{(n)} > 0 \,\}.$$

Since $M_f \subseteq M$, we have

$$\mathrm{GCD}(M) \leq \mathrm{GCD}(M_f),$$

i.e., $d(j) \leq d_f(j)$.

The reverse inequality follows from the following probabilistic meaning of $p_{jj}^{(n)}$ and $f_{jj}^{(n)}$, $n \geq 1$.

If the "particle" leaving state j arrives in this state in n steps again ($p_{jj}^{(n)} > 0$), this means that it returned to j for the *first time* in k_1 steps ($f_{jj}^{(k_1)} > 0$), then in k_2 steps ($f_{jj}^{(k_2)} > 0$), …, and finally in k_l steps ($f_{jj}^{(k_l)} > 0$). Therefore $n = k_1 + k_2 + \cdots + k_l$. The number $d_f(j)$ is divisible by $k_1, k_2, \ldots, k_l$; hence it is a divisor of n. But $d(j)$ is the largest among the divisors of n for which $p_{jj}^{(n)} > 0$. Hence $d(j) \geq d_f(j)$.

Thus $d(j) = d_f(j)$. This, by the way, means that instead of defining the period $d(j)$ of a state j by the formula $d(j) = \mathrm{GCD}(n \geq 1 \colon p_{jj}^{(n)} > 0)$, we could also define it by $d(j) = \mathrm{GCD}(n \geq 1 \colon f_{jj}^{(n)} > 0)$.

The proof of Lemma 2 is completed.

□

Now we proceed to the proof of (25) for $i \neq j$. Rewrite (11) in the form

$$p_{ij}^{(n)} = \sum_{k=1}^{\infty} f_{ij}^{(k)} p_{jj}^{(n-k)}, \tag{32}$$

where we set $p_{jj}^{(l)} = 0$, $l < 0$.

Since $p_{jj}^{(n)} \to \frac{1}{\mu_j}$ here and $\sum_{k=1}^{\infty} f_{ij}^{(k)} \leq 1$, we have, by the dominated convergence theorem (Theorem 3 in Sect. 6, Chap. 2, Vol. 1),

$$\lim_n \sum_{k=1}^{\infty} f_{ij}^{(k)} p_{jj}^{(n-k)} = \sum_{k=1}^{\infty} f_{ij}^{(k)} \lim_n p_{jj}^{(n-k)} = \frac{1}{\mu_j} \sum_{k=1}^{\infty} f_{ij}^{(k)} = \frac{1}{\mu_j} f_{ij}. \tag{33}$$

Now (32) and (33) imply that

$$\lim_n p_{ij}^{(n)} = \frac{f_{ij}}{\mu_j}, \tag{34}$$

i.e., (25) holds true.

Finally, we will show that under the additional assumption $i \leftrightarrow j$ (i.e. i, j belong to the *same* indecomposable class of communicating states), we have $f_{ij} = 1$. Then (34) will imply property (26).

State j is recursive by assumption. Therefore $\mathsf{P}_j\{X_n = j \text{ i.o.}\} = 1$ by statement (a) of Theorem 1. Hence for any m

$$\begin{aligned} p_{ji}^{(m)} &= \mathsf{P}_j(\{X_m = i\} \cap \{X_n = j \text{ i.o.}\}) \\ &\le \sum_{n>m} \mathsf{P}_j\{X_m = i, X_{m+1} \ne j, \dots, X_{n-1} \ne j, X_n = j\} \\ &= \sum_{n>m} p_{ji}^{(m)} f_{ij}^{(n-m)} = p_{ji}^{(m)} f_{ij}, \end{aligned} \tag{35}$$

where the next-to-last equality follows from the generalized Markov property (see (2) in Sect. 2).

Since E is a class of communicating states, there is m such that $p_{ji}^{(m)} > 0$. Therefore (35) implies that $f_{ij} = 1$.

The proof of Theorem 5 is completed. □

6. It is natural to state an analog of Theorem 5 for an arbitrary period d of state j of interest ($d = d(j) \ge 1$).

Theorem 4. *Let the state $j \in E$ of a Markov chain be recurrent with period $d = d(j) \ge 1$, and let i be a state in E (possibly coinciding with j).*

(a) *Suppose that i and j are in the same indecomposable class $C \subseteq E$ with (cyclic) subclasses $C_0, C_1, \dots, C_{d-1}$ numbered so that $j \in C_0$, $i \in C_a$, where $a \in \{0, 1, \dots, d-1\}$, and the motion over them goes in cyclic order, $C_0 \to C_1 \to \cdots \to C_a \to \cdots \to C_{d-1} \to C_0$. Then*

$$p_{ij}^{(nd+a)} \to \frac{d}{\mu_j} \quad \text{as} \quad n \to \infty. \tag{36}$$

(b) *In the general case, when i and j may belong to different indecomposable classes,*

$$p_{ij}^{(nd+a)} \to \frac{d}{\mu_j} \left[\sum_{k=0}^{n} f_{jj}^{(kd+a)} \right] \quad \text{as} \quad n \to \infty \tag{37}$$

for any $a = 0, 1, \dots, d-1$.

PROOF. (a) At first, let $a = 0$, i.e., i and j belong to the same indecomposable class C and, moreover, to the same cyclic subclass C_0.

Consider the transition probabilities $p_{ij}^{(d)}$, $i, j \in C$, and arrange from them a *new* Markov chain (according to the constructions from Sect. 1).

For this new chain state j will be recurrent and aperiodic, and states i and j remain communicating ($i \leftrightarrow j$). Therefore, by property (26) of Theorem 5,

$$p_{ij}^{(nd)} \to \frac{1}{\sum_{k=1}^{\infty} k f_{jj}^{(kd)}} = \frac{d}{\sum_{k=1}^{\infty} (kd) f_{jj}^{(kd)}} = \frac{d}{\mu_j},$$

where the last equality holds because $f_{jj}^{(l)} = 0$ for all l not divisible by d and $\mu_j = \sum_{l=1}^{\infty} l f_{jj}^{(l)}$ by definition.

Assume now that (36) has been proved for $a = 0, 1, \ldots, r$ $(\leq d-2)$. By the dominated convergence theorem (Theorem 3 in Sect. 4, Chap. 2, Vol. 1),

$$p_{ij}^{(nd+r+1)} = \sum_{k=1}^{\infty} p_{ik} p_{kj}^{(nd+r)} \to \sum_{k=1}^{\infty} p_{ik} \frac{d}{\mu_j} = \frac{d}{\mu_j}.$$

Therefore (36) is true for $a = r+1$ $(\leq d-1)$, hence it is established by induction for all $a = 0, 1, \ldots, d-1$.

(b) For all i and j in E we have (see (11))

$$p_{ij}^{(nd+a)} = \sum_{k=1}^{nd+a} f_{ij}^{(k)} p_{jj}^{(nd+a-k)}, \quad a = 0, 1, \ldots, d-1.$$

By assumption, the period of j is d. Hence $p_{jj}^{(nd+a-k)} = 0$, unless $k-a$ has the form rd. Therefore

$$p_{ij}^{(nd+a)} = \sum_{r=0}^{n} f_{ij}^{(rd+a)} p_{jj}^{((n-r)d)}.$$

Using this equality and (36) and applying again the dominated convergence theorem, we arrive at the required relation (37).

□

7. As was pointed out at the end of Sect. 4, in the problem of classifying Markov chains in terms of asymptotic properties of transition probabilities, we can restrict ourselves to *aperiodic indecomposable chains*.

The results of Theorems 1–5 actually contain all that we need for the complete *classification* of such chains.

The following lemma is one of the results saying that for an *indecomposable* chain all states are of the same (recurrent or transient) type. (Compare with the property that the states are "of the same type" in Theorem 2, Sect. 4.)

Lemma 3. *Let E be an indecomposable class (of communicating states). Then all its states are either recurrent or transient.*

PROOF. Let the chain have at least one transient state, say, state i. By Theorem 1, $\sum_n p_{ii}^{(n)} < \infty$.

Now let j be another state. Since E is an indecomposable class of communicating states ($i \leftrightarrow j$), there are k and l such that $p_{ij}^{(k)} > 0$ and $p_{ji}^{(l)} > 0$. The obvious inequality

$$p_{ii}^{(n+k+l)} \geq p_{ij}^{(k)} p_{jj}^{(n)} p_{ji}^{(l)}$$

implies now that

$$\sum_n p_{ii}^{(n+k+l)} \geq p_{ij}^{(k)} p_{ji}^{(l)} \sum_n p_{jj}^{(n)}.$$

By assumption, $\sum_n p_{ii}^{(n)} < \infty$ and k, l satisfy $p_{ij}^{(k)} p_{ji}^{(l)} > 0$. Hence $\sum_n p_{jj}^{(n)} < \infty$.

By statement (b) of Theorem 1, this implies that j is also a transient state. In other words, if at least one state of an indecomposable chain is transient, then so are all other states.

Now let i be a recurrent state. We will show that all the other states are then recurrent. Suppose that (along with the recurrent state i) there is at least one transient state. Then, by what has been proved, all other states must be transient, which contradicts the assumption that i is a recurrent state. Thus the presence of at least one recurrent state implies that all other states (of an indecomposable chain) are also recurrent.

□

This lemma justifies the commonly used terminology of saying about an indecomposable chain (rather than about a single state) that it is recurrent or transient.

Theorem 5. *Let a Markov chain consist of a single indecomposable class E of aperiodic states. Then only one of three possibilities may occur.*

(i) *The chain is transient. In this case*

$$\lim_n p_{ij}^{(n)} = 0$$

for any $i, j \in E$ with convergence to zero rather "fast" in the sense that

$$\sum_n p_{ij}^{(n)} < \infty.$$

(ii) *The chain is recurrent and null. In this case, again*

$$\lim_n p_{ij}^{(n)} = 0$$

for any $i, j \in E$, but the convergence is "slow" in the sense that

$$\sum_n p_{ij}^{(n)} = \infty$$

and the average time μ_j of first return from j to j is infinite.

(iii) *The chain is recurrent and positive. In this case*

$$\lim_n p_{ij}^{(n)} = \frac{1}{\mu_j} > 0$$

for all $i, j \in E$, where μ_j, the average time of return from j to j, is finite.

PROOF. Statement (i) has been proved in Theorems 1 (b) and 2. Statements (ii) and (iii) follow directly from Theorems 1 (a) and 3.

□

Consider the case of *finite Markov chains*, i.e., the case where the state set E consists of *finitely many* elements.

It turns out that in this case only the third of the three options (i), (ii), (iii) in Theorem 5 is possible.

Theorem 6. *Let a finite Markov chain be indecomposable and aperiodic. Then this chain is recurrent and positive, and* $\lim_n p_{ij}^{(n)} = \frac{1}{\mu_j} > 0$.

PROOF. Suppose that the chain is transient. If the state space consists of r states ($E = \{1, 2, \ldots, r\}$), then

$$\lim_n \sum_{j=1}^{r} p_{ij}^{(n)} = \sum_{j=1}^{r} \lim_n p_{ij}^{(n)}. \tag{38}$$

Obviously, the left-hand side is equal to 1. But the assumption that the chain is transient implies (by Theorem 1 (i)) that the right-hand side is zero.

Suppose now that the states of the chain are recurrent.

Since by Theorem 5 there remain only two options, (ii) and (iii), we must exclude (ii). But since $\lim_n p_{ij}^{(n)} = 0$ for all $i, j \in E$ in this case, we arrive at a contradiction using (38) in the same way as in the case of transient states.

Thus only (iii) is possible.

□

9. Problems

1. Consider an indecomposable chain with state space $0, 1, 2, \ldots$. This chain is transient if and only if 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 $0, 1, \ldots$ 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 $0, 1, \ldots$ 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 $0, 1, 2, \ldots$ and transition probabilities

$$p_{00} = r_0, \quad p_{01} = p_0 > 0,$$
$$p_{ij} = \begin{cases} p_i > 0, & j = i + 1, \\ r_i \geq 0, & j = i, \\ q_i > 0, & j = i - 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\rho_0 = 1$, $\rho_m = (q_1 \ldots q_m)/(p_1 \ldots p_m)$. Prove the following propositions.

$$\begin{aligned}
\text{A chain is recursive} &\iff \sum \rho_m = \infty,\\
\text{A chain is transient} &\iff \sum \rho_m < \infty,\\
\text{A chain is positive} &\iff \sum \rho_m = \infty, \quad \sum \frac{1}{p_m \rho_m} < \infty,\\
\text{A chain is null} &\iff \sum \rho_m = \infty, \quad \sum \frac{1}{p_m \rho_m} = \infty.
\end{aligned}$$

5. Show that
$$f_{ik} \geq f_{ij} f_{jk}, \qquad \sup_n p_{ij}^{(n)} \leq f_{ij} \leq \sum_{n=1}^{\infty} p_{ij}^{(n)}.$$

6. Show that for any Markov chain with countable state space the limits of $p_{ij}^{(n)}$ always exist 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 $\xi_0, \xi_1, \dots$ with $\xi_{k+1} = (\xi_k)^+ + \eta_{k+1}$, $k \geq 0$, where $\eta_1, \eta_2, \dots$ is a sequence of independent identically distributed random variables with $\mathsf{P}(\eta_k = j) = p_j, j = 0, 1, \dots$. Write the transition matrix and show that if $p_0 > 0$, $p_0 + p_1 < 1$, the chain is recurrent if and only if $\sum_k k p_k \leq 1$.

6. Limiting, Stationary, and Ergodic Distributions for Countable Markov Chains

1. We begin with a general result clarifying the relationship between the *limits* $\Pi = (\pi_1, \pi_2, \dots)$, where $\pi_j = \lim_n p_{ij}^{(n)}$, $j = 1, 2, \dots$, and *stationary distributions* $\mathbb{Q} = (q_1, q_2, \dots)$.

Theorem 1. *Consider a Markov chain with a countable state space $E = \{1, 2, \dots\}$ and transition probabilities p_{ij}, $i, j \in E$, such that the limits*

$$\pi_j = \lim_n p_{ij}^{(n)}, \quad j \in E,$$

exist and are independent of the initial states $i \in E$. Then

(a) $\sum_{j=1}^{\infty} \pi_j \leq 1$, $\sum_{i=1}^{\infty} \pi_i p_{ij} = \pi_j$, $j \in E$;

(b) *Either* $\sum_{j=1}^{\infty} \pi_j = 0$ *(hence all* $\pi_j = 0$, $j \in E$*) or* $\sum_{j=1}^{\infty} \pi_j = 1$;

(c) *If* $\sum_{j=1}^{\infty} \pi_j = 0$, *then the Markov chain has no stationary distributions, and if* $\sum_{j=1}^{\infty} \pi_j = 1$, *then the vector of limiting values* $\Pi = (\pi_1, \pi_2, \dots)$ *is a stationary distribution for this chain, and the chain has no other stationary distribution.*

PROOF. We have

$$\sum_{j=1}^{\infty} \pi_j = \sum_{j=1}^{\infty} \lim_n p_{ij}^{(n)} \leq \liminf_n \sum_{j=1}^{\infty} p_{ij}^{(n)} = 1, \tag{1}$$

and, for any $j \in E$, $k \in E$,

$$\sum_{i=1}^{\infty} \pi_i p_{ij} = \sum_{i=1}^{\infty} \lim_n p_{ki}^{(n)} p_{ij} \le \liminf_n \sum_{i=1}^{\infty} p_{ki}^{(n)} p_{ij} = \liminf_n p_{kj}^{(n+1)} = \pi_j. \quad (2)$$

Remark. Note that the inequalities and lower limits appear here, of course, due to Fatou's lemma, which is applied to Lebesgue's integral over a *σ-finite* (nonnegative) measure rather than a *probability* measure as in Sect. 6, Chap. 2, Vol. 1.

Thus the vector $\Pi = (\pi_1, \pi_2, \dots)$ satisfies

$$\sum_{j=1}^{\infty} \pi_j \le 1 \quad \text{and} \quad \sum_{i=1}^{\infty} \pi_i p_{ij} \le \pi_j, \quad j \in E. \quad (3)$$

Let us show that the latter inequality is in fact the equality.

Let for some $j_0 \in E$

$$\sum_{i=1}^{\infty} \pi_i p_{ij_0} < \pi_{j_0}. \quad (4)$$

Then

$$\sum_{j=1}^{\infty} \pi_j > \sum_{j=1}^{\infty} \Big(\sum_{i=1}^{\infty} \pi_i p_{ij} \Big) = \sum_{i=1}^{\infty} \pi_i \sum_{j=1}^{\infty} p_{ij} = \sum_{i=1}^{\infty} \pi_i.$$

The contradiction thus obtained shows that $\sum_{i=1}^{\infty} \pi_i p_{ij} = \pi_j$. Together with inequality $\sum_{j=1}^{\infty} \pi_j \le 1$ this proves conclusion (a).

For the proof of (b), we iterate the equality $\sum_{i=1}^{\infty} \pi_i p_{ij} = \pi_j$ to obtain

$$\sum_{i=1}^{\infty} \pi_i p_{ij}^{(n)} = \pi_j$$

for any $n \ge 1$ and any $j \in E$. Hence, by the dominated convergence theorem (Theorem 3, Sect. 6, Chap. 2, Vol. 1),

$$\pi_j = \lim_n \sum_{i=1}^{\infty} \pi_i p_{ij}^{(n)} = \sum_{i=1}^{\infty} \pi_i \lim_n p_{ij}^{(n)} = \Big(\sum_{i=1}^{\infty} \pi_i \Big) \pi_j,$$

i.e.,

$$\pi_j \Big(1 - \sum_{i=1}^{\infty} \pi_i \Big) = 0, \quad j \in E,$$

so that $\big(\sum_{j=1}^{\infty} \pi_j\big)\big(1 - \sum_{i=1}^{\infty} \pi_i\big) = 0$. Thus $a(1-a) = 0$ with $a = \sum_{i=1}^{\infty} \pi_i$, implying that either $a = 1$ or $a = 0$, which proves conclusion (b).

For the proof of (c), assume that $\mathbb{Q} = (q_1, q_2, \dots)$ is a stationary distribution. Then $\sum_{i=1}^{\infty} q_i p_{ij}^{(n)} = q_j$ and we obtain $\big(\sum_{i=1}^{\infty} q_i\big)\pi_j = q_j$, $j \in E$, by the dominated convergence theorem.

Therefore, if $\mathbb{Q}$ is a stationary distribution, then $\sum_{i=1}^{\infty} q_i = 1$, and hence this stationary distribution must satisfy $q_j = \pi_j$ for all $j \in E$. Thus in the case where $\sum_{j=1}^{\infty} \pi_j = 0$, it is impossible to have $\sum_{i=1}^{\infty} q_i = 1$, so that there is no stationary distribution in this case.

According to (b), there remains the possibility that $\sum_{j=1}^{\infty} \pi_j = 1$. In this case, by (a), $\Pi = (\pi_1, \pi_2, \dots)$ is itself a stationary distribution, and the foregoing proof implies that if $\mathbb{Q}$ is also a stationary distribution, then it must coincide with Π, which proves the uniqueness of the stationary distribution when $\sum_{j=1}^{\infty} \pi_j = 1$.

□

2. Theorem 1 provides a *sufficient* condition for the existence of a unique stationary distribution. This condition requires that for all $j \in E$ *there exist* the limiting values $\pi_j = \lim_n p_{ij}^{(n)}$ independent of $i \in E$ such that $\pi_j > 0$ for at least one $j \in E$.

At the same time, the more general problem of the *existence* of the limits $\lim_n p_{ij}^{(n)}$ was thoroughly explored in Sect. 5 in terms of the "intrinsic" properties of the chains such as indecomposability, periodicity, recurrence and transience, and positive and null recurrence. Therefore it would be natural to formulate the conditions for the existence of the stationary distribution in terms of these intrinsic properties determined by the structure of the matrix of transition probabilities p_{ij}, $i, j \in E$. It is seen also that if conditions stated in these terms imply that *all* the limiting values are positive, $\pi_j > 0$, $j \in E$, then by definition (see property **C** in Sect. 3) the vector $\Pi = (\pi_1, \pi_2, \dots)$ will be an *ergodic* limit distribution.

The answers to these questions are given in the following two theorems.

Theorem 2 ("Basic theorem on stationary distributions"). *Consider a Markov chain with a countable state space E. A necessary and sufficient condition for the existence of a unique stationary distribution is that*

(a) *There exists a unique indecomposable subclass and*
(b) *All the states are positive recurrent.*

Theorem 3 ("Basic theorem on ergodic distributions"). *Consider a Markov chain with a countable state space E. A necessary and sufficient condition for the existence of an ergodic distribution is that the chain is*

(a) *Indecomposable*,
(b) *Positive recurrent, and*
(c) *Aperiodic*.

3. PROOF OF THEOREM 2. *Necessity*. Let the chain at hand have a unique stationary distribution, to be denoted by $\widetilde{\mathbb{Q}}$. We will show that in this case there is a unique positive recurrent subclass in state space E.

Let N denote the conceivable number of such subclasses ($0 \le N \le \infty$).

Suppose $N = 0$, and let j be a state in E. Since there are no positive recurrent classes, state j may be either transitive or null recurrent.

In the former case, the limits $\lim_n p_{ij}^{(n)}$ exist and are equal to zero for all $i \in E$ by Theorem 2 in Sect. 5.

In the latter case these limits also exist and are equal to zero, which follows from (37) in Sect. 5 and the fact that $\mu_j = \infty$, since state j is null recurrent.

Thus, if $N = 0$, then the limits $\pi_j = \lim_n p_{ij}^{(n)}$ exist and are equal to zero for all $i, j \in E$. Therefore, by Theorem 1 (c), in this case there is no stationary distribution, so the case $N = 0$ is excluded by the assumption of the *existence* of a stationary distribution $\widetilde{\mathbb{Q}}$.

Suppose now that $N = 1$. Denote the only positive recurrent class by C. If the period of this class $d(C) = 1$, then by (26) of Theorem 5, Sect. 5,

$$p_{ij}^{(n)} \to \mu_j^{-1}, \quad n \to \infty,$$

for all $i, j \in C$. If $j \notin C$, then this state is transient and by property (21) of Theorem 2, Sect. 5,

$$p_{ij}^{(n)} \to 0, \quad n \to \infty,$$

for all $i \in E$.

Let

$$q_j = \begin{cases} \mu_j^{-1}\ (>0), & \text{if } j \in C, \\ 0, & \text{if } j \notin C. \end{cases} \tag{5}$$

Then, since $C \neq \varnothing$, the collection $\mathbb{Q} = (q_1, q_2, \dots)$ is (by Theorem 1 (a)) a *unique stationary distribution*, therefore $\mathbb{Q} = \widetilde{\mathbb{Q}}$.

Suppose now that the period $d(C) > 1$. Let $C_0, C_1, \dots, C_{d-1}$ be the cyclic subclasses of the (positive recurrent) class C.

Every C_k, $k = 0, 1, \dots, d-1$, is a recurrent and aperiodic subclass of the matrix of transition probabilities $p_{ij}^{(d)}$, $i, j \in C$. Hence, for $i, j \in C_k$,

$$p_{ij}^{(nd)} \to \frac{d}{\mu_j} > 0$$

by (36) from Sect. 5. Therefore, for each set C_k, the collection $\{d/\mu_j,\ j \in C_k\}$ is (by Theorem 1 (b)) a unique stationary distribution (with regard to the matrix $p_{ij}^{(d)}$, $i, j \in C$). This implies, in particular, that $\sum_{j \in C_k} \frac{d}{\mu_j} = 1$, i.e., $\sum_{j \in C_k} \frac{1}{\mu_j} = \frac{1}{d}$.

Let us set

$$q_j = \begin{cases} \mu_j^{-1}, & j \in C = C_0 + \dots + C_{d-1}, \\ 0, & j \notin C, \end{cases} \tag{6}$$

and show that the collection $\mathbb{Q} = (q_1, q_2, \dots)$ is a unique stationary distribution.

Indeed, if $i \in C$, then

$$p_{ii}^{(nd)} = \sum_{j \in C} p_{ij}^{(nd-1)} p_{ji}.$$

Then we find in the same way as in (1) that

$$\frac{d}{\mu_i} = \lim_n p_{ii}^{(nd)} \geq \sum_{j \in C} \liminf_n p_{ij}^{(nd-1)} p_{ji} = \sum_{j \in C} \frac{d}{\mu_j} p_{ji},$$

hence

$$\frac{1}{\mu_i} \geq \sum_{j \in C} \frac{1}{\mu_j} p_{ji}. \tag{7}$$

But

$$\sum_{i \in C} \frac{1}{\mu_i} = \sum_{k=0}^{d-1} \left(\sum_{i \in C_k} \frac{1}{\mu_i} \right) = \sum_{k=0}^{d-1} \frac{1}{d} = 1. \tag{8}$$

As in the proof of Theorem 1 (see (3) and (4)), we obtain from (7) and (8) that (7) holds in fact with an equality sign:

$$\frac{1}{\mu_i} = \sum_{j \in C} \frac{1}{\mu_j} p_{ji}. \tag{9}$$

Since $q_i = \mu_i^{-1} > 0$, we see from (9) that the collection $\mathbb{Q} = (q_1, q_2, \dots)$ is a stationary distribution, which is unique by Theorem 1. Therefore $\mathbb{Q} = \widetilde{\mathbb{Q}}$.

Finally, let $2 \leq N < \infty$ or $N = \infty$. Denote the positive recurrent subclasses by $C^1, \dots, C^N$ if $N < \infty$, and by $C^1, C^2, \dots$ if $N = \infty$.

Let $\mathbb{Q}^k = (q_1^k, q_2^k, \dots)$ be a stationary distribution for a class C^k, given by the formula (compare with (5), (6))

$$q_j^k = \begin{cases} \mu_j^{-1} > 0, & j \in C^k, \\ 0, & j \notin C^k. \end{cases}$$

Then for any nonnegative numbers $a_1, a_2, \dots$ with $\sum_{k=1}^{\infty} a_k = 1$ ($a_{N+1} = \dots = 0$ if $N < \infty$), the collection $a_1 \mathbb{Q}^1 + \dots + a_N \mathbb{Q}^N + \dots$ is, obviously, a stationary distribution. Hence the assumption $2 \leq N \leq \infty$ leads us to the existence of a *continuum* of stationary distributions, which contradicts the assumption of its uniqueness.

Thus, the foregoing proof shows that only the case $N = 1$ is possible. In other words, the *existence of a unique stationary distribution* implies that the chain has *only one indecomposable class, which consists of positive recurrent states*.

Sufficiency. If the chain has an indecomposable subclass of positive recurrent states, i.e., the case $N = 1$ takes place, then the preceding arguments imply (by Theorem 1 (c)) the existence and uniquence of the stationary distribution.

This completes the proof of Theorem 2.

□

4. PROOF OF THEOREM 3. Actually, all we need is contained in Theorem 2 and its proof.

Sufficiency. Using the notation in the proof of Theorem 2, we have by the conditions of the present theorem that $N = 1$, $C = E$, and $d(E) = 1$ (aperiodicity). Then the reasoning in the case $N = 1$ of the proof of Theorem 2 implies that $\mathbb{Q} = (q_1, q_2, \dots)$ with $q_j = \mu_j^{-1}$, $j \in E$, is a stationary and ergodic distribution, since all $\mu_j^{-1} < \infty$, $j \in E$.

Thus, the existence of an ergodic distribution $\Pi = (\pi_1, \pi_2, \dots)$ is established ($\Pi = \mathbb{Q}$).

Necessity. If there exists an ergodic distribution $\Pi = (\pi_1, \pi_2, \dots)$, then by Theorem 1 there exists a unique stationary distibution $\mathbb{Q}$ coinciding with Π.

It follows from Theorem 2 (and its proof) that the cases $N = 0$ and $2 \le N \le \infty$ cannot occur, so that $N = 1$, and there is only one indecomposable class C consisting of positive recurrent states. It remains to show that $C = E$ and $d(E) = 1$.

Assume that $C \ne E$ and $d(C) = 1$. Then the same reasoning as for $N = 1$ in the proof of Theorem 2 shows that there is a state $j \notin C$ such that $p_{ij}^{(n)} \to 0$ for all $i \in E$. This, however, contradicts the property that $\pi_j = \lim_n p_{ij}^{(n)} > 0$ for all $i \in E$.

Therefore, if $d(C) = 1$, then $C = E$ and $d(E) = 1$ (aperiodicity).

Finally, if $C \ne E$ and $d(C) > 1$, the arguments in the proof of Theorem 2 (case $N = 1$) again imply that there is a stationary distribution $\mathbb{Q} = (q_1, q_2, \dots)$ with some $q_j = 0$, which is in contradiction with $\mathbb{Q} = \Pi$, where $\Pi = (\pi_1, \pi_2, \dots)$ is an ergodic distribution whose probabilities are positive by definition, $\pi_j > 0$, $j \in E$.

□

5. By the definition of the stationary (invariant) distribution $\mathbb{Q} = (q_1, q_2, \dots)$, these probabilities are subject to the conditions

$$q_j \ge 0,\ j \in E = \{1, 2, \dots\}, \qquad \sum_{j=1}^{\infty} q_j = 1 \tag{10}$$

and satisfy the equations

$$q_j = \sum_{i=1}^{\infty} q_i p_{ij}, \quad j \in E. \tag{11}$$

In other words, the stationary distribution $\mathbb{Q} = (q_1, q_2, \dots)$ is *one of the solutions* to the system of equations

$$x_j = \sum_{i=1}^{\infty} x_i p_{ij}, \quad j \in E, \tag{12}$$

with components of these solutions being *nonnegative* $(x_j \ge 0, j \in E)$ and *normalized* $\left(\sum_{j=1}^{\infty} x_j = 1\right)$.

Under the conditions of Theorem 5 there exists a stationary solution that at the same time is ergodic. Hence, by Theorem 1 (c), there is a unique solution to system (12) *within the class* of sequences $x = (x_1, x_2, \dots)$ with $x_j \ge 0$, $j \in E$, and $\sum_{j=1}^{\infty} x_j = 1$.

But in fact we can make a stronger assertion. Since we assume that the conditions of Theorem 5 are fulfilled, there exists an ergodic distribution $\Pi = (\pi_1, \pi_2, \dots)$.

Consider under this assumption the problem of existence of a solution to (12) in a *wider* class of sequences $x = (x_1, x_2, \dots)$ such that $x_j \in R$, $j \in E$, $\sum_{j=1}^{\infty} |x_j| < \infty$ and $\sum_{j=1}^{\infty} x_j = 1$. We will show that there is a unique solution in this class given by the ergodic distribution Π.

Indeed, if $x = (x_1, x_2, \dots)$ is a solution, then, using that $\sum_{j=1}^{\infty} |x_j| < \infty$, we obtain the following chain of inequalities:

$$x_j = \sum_{i=1}^{\infty} x_i p_{ij} = \sum_{i=1}^{\infty} \left(\sum_{k=1}^{\infty} x_k p_{ki} \right) p_{ij}$$
$$= \sum_{k=1}^{\infty} x_k \left(\sum_{i=1}^{\infty} p_{ki} p_{ij} \right) = \sum_{k=1}^{\infty} x_k p_{kj}^{(2)} = \cdots = \sum_{k=1}^{\infty} x_k p_{kj}^{(n)}$$

for any $n \geq 1$. Taking the limit as $n \to \infty$, we obtain (by the dominated convergence theorem) that $x_j = \left(\sum_{k=1}^{\infty} x_k\right) \pi_j$, where $\pi_j = \lim_n p_{kj}^{(n)}$ for any $k \in E$. By assumption, $\sum_{k=1}^{\infty} x_k = 1$. Hence $x_j = \pi_j$, $j \in E$, which was to be proved.

6. Problems

1. Investigate the problem of stationary, limiting, and ergodic distributions for a Markov chain with the transition probabilities matrix

$$\mathbb{P} = \begin{pmatrix} 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \\ 1/4 & 1/2 & 1/4 & 0 \\ 0 & 1/2 & 1/2 & 0 \end{pmatrix}.$$

2. Let $\mathbb{P} = \|p_{ij}\|$ be a finite doubly stochastic matrix (i.e., $\sum_{j=1}^{m} p_{ij} = 1$ for $i = 1, \ldots, m$ and $\sum_{i=1}^{m} p_{ij} = 1$ for $j = 1, \ldots, m$). Show that $\mathbb{Q} = (1/m, \ldots, 1/m)$ is a stationary distribution of the corresponding Markov chain.
3. Let X be a Markov chain with two states, $E = \{0, 1\}$, and the transition probabilities matrix

$$\mathbb{P} = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix}, \quad 0 < \alpha < 1, \quad 0 < \beta < 1.$$

Explore the limiting, ergodic, and stationary distributions for this chain.

7. Limiting, Stationary, and Ergodic Distributions for Finite Markov Chains

1. According to Theorem 6 in Sect. 5, every indecomposable aperiodic Markov chain with a finite state space is positive recurrent. This conclusion allows us to state Theorem 3 from Sect. 6 in the following form. (Compare with questions **A**, **B**, **C**, and **D** in Sect. 3.)

Theorem 1. *Consider an indecomposable aperiodic Markov chain $X = (X_n)_{n \geq 0}$ with finite state space $E = \{1, 2, \ldots, r\}$. Then*

(a) *For all $j \in E$ there exist limits $\pi_j = \lim_n p_{ij}^{(n)}$ independent of the initial state $i \in E$.*
(b) *The limits $\Pi = (\pi_1, \pi_2, \ldots, \pi_r)$ form a probability distribution, i.e., $\pi_j \geq 0$ and $\sum_{i=1}^{r} \pi_i = 1$, $j \in E$.*

(c) *Moreover, these limits* π_j *are equal to* $\mu_j^{-1} > 0$ *for all* $j \in E$, *where* $\mu_j = \sum_{n=1}^{\infty} n f_{jj}^{(n)}$ *is the mean time of return to state* j (*i.e.*, $\mu_j = \mathsf{E}_j \tau(j)$ *with* $\tau(j) = \min\{ n \geq 1 \colon X_n = j\}$), *so that* $\Pi = (\pi_1, \pi_2, \dots, \pi_r)$ *is an ergodic distribution.*
(d) *The stationary distribution* $\mathbb{Q} = (q_1, q_2, \dots, q_r)$ *exists, is unique, and is equal to* $\Pi = (\pi_1, \pi_2, \dots, \pi_r)$.

2. In addition to Theorem 1, we state the following result clarifying the role of the properties of a chain being indecomposable and aperiodic.

Theorem 2. *Consider a Markov chain with a finite state space* $E = \{1, 2, \dots, r\}$. *The following statements are equivalent*:

(a) *The chain is indecomposable and aperiodic* $(d = 1)$.
(b) *The chain is indecomposable, aperiodic* $(d = 1)$, *and positive recurrent*.
(c) *The chain is ergodic*.
(d) *There is an* n_0 *such that for all* $n \geq n_0$

$$\min_{i,j \in E} p_{ij}^{(n)} > 0.$$

PROOF. The implication (d) $\Rightarrow$ (c) was proved in Theorem 1 of Sect. 12, Chap. 1 (Vol. 1). The converse implication (c) $\Rightarrow$ (d) is obvious. The implication (a) $\Rightarrow$ (b) follows from Theorem 6, Sect. 5, while (b) $\Rightarrow$ (a) is obvious. Finally, the equivalence of (b) and (c) is contained in Theorem 5 from Sect. 6.

□

8. Simple Random Walk as a Markov Chain

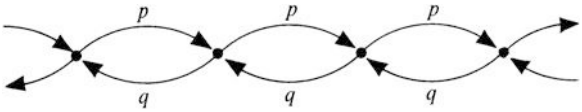
1. A *simple d-dimensional random walk* is a homogeneous Markov chain $X = (X_n)_{n \geq 0}$ describing the motion of a random "particle" over the nodes of the *lattice* $\mathbb{Z}^d = \{0, \pm 1, \pm 2, \dots\}^d$ when this particle at every step either stays in the same state or passes to one of the *adjacent* states with some probabilities.

EXAMPLE 1. Let $d = 1$ and the state space of the chain be $E = \mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$. Let the transition probabilities matrix be

$$p_{ij} = \begin{cases} p, & j = i + 1, \\ q, & j = i - 1, \\ 0 & \text{otherwise}, \end{cases}$$

where $p + q = 1$.

The following graph demonstrates the possible transitions of this chain.



If $p = 0$ or 1, the motion is *deterministic*, and the particle moves to the left or the right, respectively.

These deterministic cases are of little interest; all the states here are inessential. Hence we will assume that $0 < p < 1$.

Under this assumption the states form a single class of *essential communicating* states. In other words, when $0 < p < 1$, the chain is *indecomposable* (Sect. 4).

By the formula for the binomial distribution (Sect. 2, Chap. 1, Vol. 1),

$$p_{jj}^{(2n)} = C_{2n}^{n}(pq)^n = \frac{(2n)!}{(n!)^2}(pq)^n \tag{1}$$

for any $j \in E$. By Stirling's formula (see (6) in Sect. 2, Chap. 1, Vol. 1); see also Problem 1),

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

Therefore we find from (1) that

$$p_{jj}^{(2n)} \sim \frac{(4pq)^n}{\sqrt{\pi n}}, \tag{2}$$

hence

$$\sum_{n=1}^{\infty} p_{jj}^{(2n)} = \infty, \quad \text{if } p = q, \tag{3}$$

$$\sum_{n=1}^{\infty} p_{jj}^{(2n)} < \infty, \quad \text{if } p \neq q. \tag{4}$$

These formulas, together with Theorem 1 in Sect. 5, yield the following result.

The simple one-dimensional random walk over the set $E = \mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$ is recurrent in the symmetric case when $p = q = \frac{1}{2}$ and transient when $p \neq q$.

If $p = q = 1/2$, then, as was shown in Sect. 10, Chap. 1, Vol. 1, for large n

$$f_{jj}^{(2n)} \sim \frac{1}{2\sqrt{\pi}\, n^{3/2}}. \tag{5}$$

Therefore

$$\mu_j = \sum_{n=1}^{\infty} (2n) f_{jj}^{(2n)} = \infty, \quad j \in E. \tag{6}$$

Therefore all the states in this case are *null recurrent*. Hence, by Theorem 5, Sect. 5, we obtain that for any i and j, $p_{ij}^{(n)} \to 0$ as $n \to \infty$ for all $0 < p < 1$. This implies (Theorem 1, Sect. 6) that there are no limit distributions and no stationary or ergodic distributions.

EXAMPLE 2. Let $d = 2$. Consider the symmetric motion in the plain (corresponding to the case $p = q = 1/2$ in the previous example), when the particle can move one step in either direction (to the left or right, up or down) with probability $1/4$ (Fig. 41).

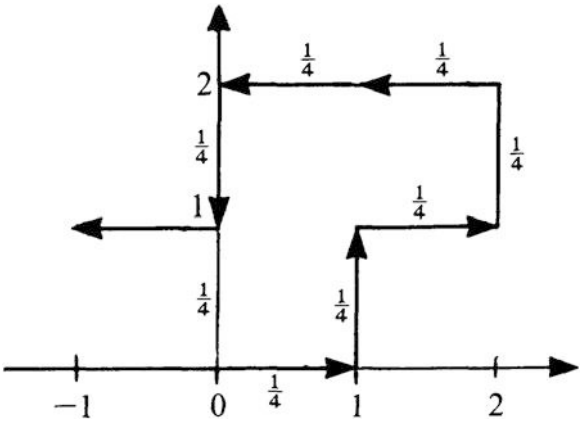


Fig. 41 A walk in the plane

Assuming for definiteness that the particle was at the origin $\mathbf{0} = (0, 0)$ at the initial time instant, we will investigate the problem of its *return* or *nonreturn* to this zero state.

To this end, consider the paths in which a particle makes i steps to the right and i steps to the left and j steps up and j steps down. If $2i + 2j = 2n$, this means that the particle starting from the origin returns to this state in $2n$ steps. It is also clear that the particle cannot return to the origin after an odd number of steps.

This implies that the probabilities of transition from state $\mathbf{0}$ to the same state $\mathbf{0}$ are given by the following formulas:

$$p_{\mathbf{00}}^{(2n+1)} = 0, \quad n = 0, 1, 2, \ldots,$$

and (by the formula for total probability)

$$p_{\mathbf{00}}^{(2n)} = \sum_{(i,j):\ i+j=n} \frac{(2n)!}{(i!)^2 (j!)^2} \left(\frac{1}{4}\right)^{2n}, \quad n = 1, 2, \ldots \tag{7}$$

(see also Subsection 2, "Multinomial Distribution," in Sect. 2, Chap. 1, Vol. 1.)

Multiplying the numerator and denominator in (7) by $(n!)^2$ yields

$$p_{\mathbf{00}}^{(2n)} = \left(\frac{1}{4}\right)^{2n} C_{2n}^{n} \sum_{i=0}^{n} C_n^i C_n^{n-i} = \left(\frac{1}{4}\right)^{2n} (C_{2n}^{n})^2, \tag{8}$$

where we have used the formula (Problem 4 in Sect. 2, Chap. 1, Vol. 1).

$$\sum_{i=0}^{n} C_n^i C_n^{n-i} = C_{2n}^{n}.$$

By Stirling's formula we obtain from (8) that $p_{00}^{(2n)} \sim \frac{1}{\pi n}$, hence

$$\sum_{n=0}^{\infty} p_{00}^{(2n)} = \infty. \tag{9}$$

Of course, a similar assertion, by symmetry, holds not only for $(0,0)$ but also for any state (i,j).

As in the case $d = 1$, we obtain from (9) and Theorem 1 of Sect. 5 the following statement.

The simple two-dimensional symmetric random walk over the set $E = \mathbb{Z}^2 = \{0, \pm 1, \pm 2, \dots\}^2$ *is recurrent.*

EXAMPLE 3. It turns out that in the case $d \geq 3$, the behavior of the symmetric random walk over the states $E = \mathbb{Z}^d = \{0, \pm 1, \pm 2, \dots\}^d$ is quite different from the cases $d = 1$ and $d = 2$ considered above.

That is:

The simple d-dimensional symmetric random walk over the set $E = \mathbb{Z}^d = \{0, \pm 1, \pm 2, \dots\}^d$ *for every* $d \geq 3$ *is transitive.*

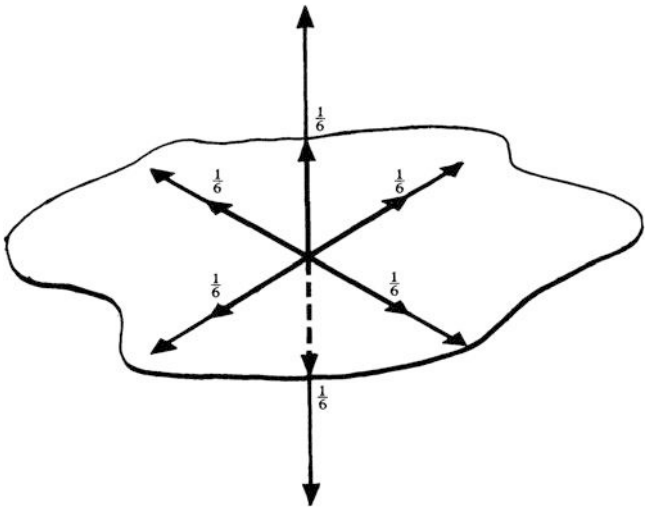
The proof relies on the fact that the asymptotic behavior of the probabilities $p_{jj}^{(2n)}$ as $n \to \infty$ is

$$p_{jj}^{(2n)} \sim \frac{c(d)}{n^{d/2}} \tag{10}$$

with a positive constant $c(d)$ depending only on dimension d.

We will give the proof for $d = 3$ leaving the case $d > 3$ as a problem.

The symmetry of the random walk means that at every step the particle moves by one unit in one of the six coordinate directions with probabilities $1/6$.



Let the particle start from the state $\mathbf{0} = (0,0,0)$. Then, as for $d = 2$, we find from the formulas for the multinomial distribution (Sect. 2, Chap. 1, Vol. 1) that

$$\begin{aligned} p_{\mathbf{00}}^{(2n)} &= \sum_{(i,j):\,0\le i+j\le n} \frac{(2n)!}{(i!)^2(j!)^2((n-i-j)!)^2}\Big(\frac{1}{6}\Big)^{2n} \\ &= 2^{-2n}C_{2n}^n \sum_{(i,j):\,0\le i+j\le n} \Big[\frac{n!}{i!\,j!\,(n-i-j)!}\Big]^2\Big(\frac{1}{3}\Big)^{2n} \\ &\le C_n\,2^{-2n}C_{2n}^n\,3^{-n} \sum_{(i,j):\,0\le i+j\le n} \frac{n!}{i!\,j!\,(n-i-j)!}\Big(\frac{1}{3}\Big)^{2n} \\ &= C_n\,2^{-2n}C_{2n}^n\,3^{-n}, \end{aligned} \tag{11}$$

where

$$C_n = \max_{(i,j):\,0\le i+j\le n}\Big(\frac{n!}{i!\,j!\,(n-i-j)!}\Big) \tag{12}$$

and where we have used the obvious fact that

$$\sum_{(i,j):\,0\le i+j\le n} \frac{n!}{i!\,j!\,(n-i-j)!}\Big(\frac{1}{3}\Big)^{2n} = 1.$$

It will be established subsequently that

$$C_n \sim \frac{n!}{[(n/3)!]^3}. \tag{13}$$

Then, by Stirling's formula, (13) implies that

$$C_n 2^{-2n}C_{2n}^n 3^{-n} \sim \frac{3\sqrt{3}}{2\pi^{3/2}n^{3/2}}. \tag{14}$$

Hence (11) yields

$$\sum_{n=1}^{\infty} p_{\mathbf{00}}^{(2n)} < \infty, \tag{15}$$

and therefore, by Theorem 1, Sect. 5, the state $\mathbf{0} = (0,0,0)$ is *transient*. By symmetry, the same holds for any state in $E = \mathbb{Z}^3$.

It remains to establish (13). Let

$$m_n(i,j) = \frac{n!}{i!\,j!\,(n-i-j)!},$$

and let $i_0 = i_0(n), j_0 = j_0(n)$ be the values of i, j for which

$$\max_{(i,j):\,0\le i+j\le n} m_n(i,j) = m_n(i_0,j_0).$$

Taking four points (i_0-1,j_0), (i_0+1,j_0), (i_0,j_0-1), (i_0,j_0+1) and using that the corresponding values $m_n(i_0-1,j_0)$, $m_n(i_0+1,j_0)$, $m_n(i_0,j_0-1)$, and $m_n(i_0,j_0+1)$ are less than or equal to $m_n(i_0,j_0)$, we obtain the inequalities

$$n - i_0 - 1 \le 2j_0 \le n - i_0 + 1,$$
$$n - j_0 - 1 \le 2i_0 \le n - j_0 + 1.$$

One can easily deduce from these inequalities that

$$i_0(n) \sim \frac{n}{3}, \quad j_0(n) \sim \frac{n}{3},$$

which implies the required formula (13).

Summarizing these cases, we can state the following theorem due to G. Pólya.

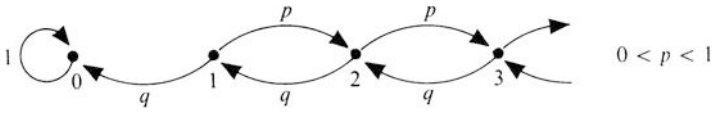
Theorem. *The simple symmetric random walk over the set*

$$E = \mathbb{Z}^d = \{0, \pm 1, \pm 2, \dots\}^d$$

is recurrent when $d = 1$ *or* $d = 2$ *and transitive when* $d \ge 3$.

2. The previous examples dealt with a simple random walk in the *entire* space $\mathbb{Z}^d$. In this subsection we will consider examples of simple random walks with state space E strictly less than $\mathbb{Z}^d$. We will restrict ourselves to the case $d = 1$.

EXAMPLE 4. Consider a simple random walk with state space $E = \{0, 1, 2, \dots\}$ and *absorbing* zero state 0. Its graph of transitions is as follows.



State 0 is here the only positive recurrent state that forms a unique indecomposable subclass. (All the other states are transient.) By Theorem 2 in Sect. 6 there exists a unique stationary distribution $\mathbb{Q} = (q_0, q_1, \dots)$ with $q_0 = 1$ and $q_i = 0$, $i = 1, 2, \dots$.

This walk provides an example, where (for some i and j) the limits $\lim_n p_{ij}^{(n)}$ exist but *depend* on the initial state, which means, in particular, that this random walk possesses no ergodic distribution.

It is clear that $p_{00}^{(n)} = 1$ and $p_{0j}^{(n)} = 0$ for $j = 1, 2, \dots$, and an easy calculation shows that $p_{ij}^{(n)} \to 0$ for all $i, j = 1, 2, \dots$.

Let us show that the limits $\alpha(i) = \lim_n p_{i0}^{(n)}$ exist for all $i = 1, 2, \dots$ and are given by the formula

$$\alpha(i) = \begin{cases} (q/p)^i, & p > q, \\ 1, & p \le q. \end{cases} \tag{16}$$

This formula demonstrates that when $p > q$ (trend to the right), the limiting probability $\lim_n p_{i0}^{(n)}$ of transition from state i ($i = 1, 2, \dots$) to state 0 depends on i decreasing geometrically as i grows.

For the proof of (16) notice that $p_{i0}^{(n)} = \sum_{k \le n} f_{i0}^{(k)}$ since the null state is absorbing, hence the limit $\lim_n p_{i0}^{(n)}$ $(= \alpha(i))$ exists and equals f_{i0}, i.e., the probability of interest is the probability that the particle leaving state i eventually reaches the null state. By the same method as in Sect. 12, Chap. 1, Vol. 1 (see also Sect. 2, Chap. 7), we obtain recursive relations for these probabilities:

$$\alpha(i) = p\alpha(i+1) + q\alpha(i-1), \tag{17}$$

with $\alpha(0) = 1$. A general solution to this equation is

$$\alpha(i) = a + b(q/p)^i, \tag{18}$$

and condition $\alpha(0) = 1$ provides a condition $a + b = 1$ on a and b.

When $q > p$, we immediately obtain that $b = 0$, hence $\alpha(i) = 1$, because the $\alpha(i)$ are bounded. This result is easily understandable since in the case $q > p$ a particle has a tendency to move toward the null state.

In contrast, if $p > q$, we have the reverse situation: a particle has a tendency to move to the right, and it is natural to expect that

$$\alpha(i) \to 0, \quad i \to \infty, \tag{19}$$

so that $a = 0$ and

$$\alpha(i) = (q/p)^i. \tag{20}$$

Instead of establishing (19) first, we will prove this equality in another way.

Along with the absorbing barrier at point 0, consider one more absorbing barrier at point N. Denote by $\alpha_N(i)$ the probability that a particle leaving point i reaches the null state before getting to state N. The probabilities $\alpha_N(i)$ satisfy equations (17) with boundary conditions

$$\alpha_N(0) = 1, \quad \alpha_N(N) = 0,$$

and, as was shown in Sect. 9, Chap. 1, Vol. 1,

$$\alpha_N(i) = \frac{(q/p)^i - (q/p)^N}{1 - (q/p)^N}, \quad 0 \le i \le N. \tag{21}$$

Therefore $\lim_N \alpha_N(i) = (q/p)^i$, so that for the proof of (20) we must show that

$$\alpha(i) = \lim_N \alpha_N(i). \tag{22}$$

This is easily seen intuitively. The proof can be carried out as follows.

Assume that the particle starts from a given state i. Then

$$\alpha(i) = \mathsf{P}_i(A), \tag{23}$$

where A is the event that there is an N such that the particle leaving state i reaches the null state before state N. If

$$A_N = \{\text{the particle reaches } 0 \text{ before } N\},$$

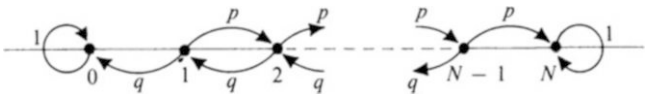
then $A = \bigcup_{N=i+1}^{\infty} A_N$. Clearly, $A_N \subseteq A_{N+1}$ and

$$\mathsf{P}_i\left(\bigcup_{N=i+1}^{\infty} A_N\right) = \lim_{N\to\infty} \mathsf{P}_i(A_N). \tag{24}$$

But $\alpha_N(i) = \mathsf{P}_i(A_N)$, so that (22) follows directly from (23) and (24).

Thus, when $p > q$, the limits $\lim_n p_{i0}$ *depend* on i. If $p \le q$, then $\lim_n p_{i0}^{(n)} = 1$ for any i and $\lim_n p_{ij}^{(n)} = 0, j \ge 1$. Hence in this case there exists a limit distribution $\Pi = (\pi_0, \pi_1, \dots)$ with $\pi_j = \lim_n p_{ij}^{(n)}$ independent of i, and $\Pi = (1, 0, 0, \dots)$.

EXAMPLE 5. Consider a simple random walk with state space $E = \{0, 1, \dots, N\}$ and *absorbing* boundary states 0 and N:



In this case there are two indecomposable positive recurrent classes $\{0\}$ and $\{N\}$. All the other states $1, 2, \dots, N-1$ are transient. We can see from the proof of Theorem 2, Sect. 6, that there exists a *continuum* of stationary distributions $\mathbb{Q} = (q_0, q_1, \dots, q_N)$, all of which have the form $q_1 = \cdots = q_{N-1} = 0$ and $q_0 = a$, $q_N = b$ with $a \ge 0$, $b \ge 0$, and $a + b = 1$.

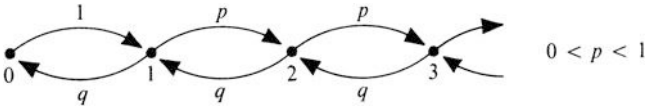
According to the results of Subsection 2 in Sect. 9, Chap. 1, Vol. 1,

$$\lim_n p_{i0}^{(n)} = \begin{cases} \dfrac{(q/p)^i - (q/p)^N}{1 - (q/p)^N}, & p \ne q, \\ 1 - (i/N), & p = q = 1/2, \end{cases} \tag{25}$$

$\lim_n p_{iN}^{(n)} = 1 - \lim_n p_{i0}^{(n)}$ and $\lim_n p_{ij}^{(n)} = 0$, $1 \le j \le N-1$.

Let us emphasize that, as in the previous example, the limiting values $\lim_n p_{ij}^{(n)}$ of the transition probabilities depend on the initial state.

EXAMPLE 6. Consider a simple random walk with state space $E = \{0, 1, \dots\}$ and a *reflecting barrier at* 0:



The behavior of this chain essentially depends on p and q.

If $p > q$, then a wandering particle has a trend to the right and the reflecting barrier enhances this trend, unlike the chain in Example 4, where a particle may become "stuck" in the zero state. All the states are transitive: $p_{ij}^{(n)} \to 0$, $n \to \infty$, for all $i, j \in E$; there are no stationary or ergodic distributions.

If $p < q$, there is a leftward trend, and the chain is recurrent, and so is the chain for $p = q$.

Let us write down the system of equations (cf. (12) in Sect. 6) for the stationary distribution $\mathbb{Q} = (q_0, q_1, \dots)$:

$$\begin{aligned} q_0 &= q_1 q, \\ q_1 &= q_0 + q_2 q, \\ q_2 &= q_1 p + q_3 q, \\ &\dots\dots\dots\dots\dots \end{aligned}$$

Hence

$$\begin{aligned} q_1 &= q(q_1 + q_2), \\ q_2 &= q(q_2 + q_3), \\ &\dots\dots\dots\dots\dots \end{aligned}$$

Therefore

$$q_j = \left(\frac{p}{q}\right) q_{j-1}, \quad j = 2, 3, \dots .$$

If $p = q$, then $q_1 = q_2 = \dots$, and hence there is no nonnegative solution to this system satisfying the conditions $\sum_{j=0}^{\infty} q_j = 1$ and $q_0 = q_1 q$. Therefore, when $p = q = 1/2$, there is *no* stationary distribution. All the states in this case are *recurrent*.

Finally, let $p < q$. The condition $\sum_{j=0}^{\infty} q_j = 1$ yields

$$q_1 \left[q + 1 + \frac{p}{q} + \left(\frac{p}{q}\right)^2 + \dots \right] = 1.$$

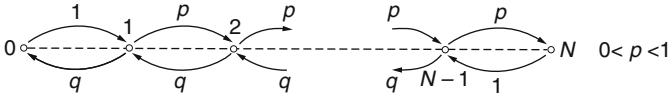
Hence

$$q_1 = \frac{q-p}{2q}, \quad q_0 = q_1 q = \frac{q-p}{2},$$

and

$$q_j = \frac{q-p}{2q} \left(\frac{p}{q}\right)^{j-1}, \quad j \geq 2.$$

EXAMPLE 7. The state space of the simple random walk in this example is $E = \{0, 1, \dots, N\}$, with *reflecting* barriers 0 and N:



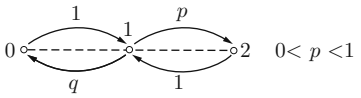
The states of this chain constitute one indecomposable class. They are positive recurrent with period $d = 2$. By Theorem 2 in Sect. 6, *there is a unique* stationary dis-

tribution $\mathbb{Q} = (q_0, q_1, \dots, q_N)$. On solving the system of equations $q_j = \sum_{i=0}^{N} q_i p_{ij}$ subject to the conditions $\sum_{i=0}^{N} q_i = 1$, $q_j \geq 0$, $j \in E$, we find

$$q_j = \frac{(p/q)^{j-1}}{1 + \sum_{i=1}^{N-1} (p/q)^{i-1}}, \quad 1 \leq j \leq N-1, \tag{26}$$

and $q_0 = q_1 q$, $q_N = q_{N-1} q$.

There is no ergodic distribution, which follows from Theorem 3, Sect. 6, and the fact that this chain has period $d = 2$. The lack of an ergodic distribution can also be seen directly. For example, let $N = 2$:



Then we see that $p_{11}^{(2n)} = 1$, but $p_{11}^{(2n+1)} = 0$, so the limit $\lim_n p_{11}^{(n)}$ does not exist. At the same time, the stationary distribution $\mathbb{Q} = (q_0, q_1, q_2)$ exists and, by (26), has the form

$$q_0 = \frac{1}{2} q, \quad q_1 = \frac{1}{2}, \quad q_2 = \frac{1}{2} p.$$

3. The material set out in the book shows that the *simple random walk* is a classical model that makes it possible to develop probabilistic ideology, elaborate probabilistic techniques, and discover many probabilistic laws. In a similar way, the study of the sums $X_n = \xi_1 + \cdots + \xi_n$, $n \geq 1$, of independent Bernoulli random variables $\xi_1, \xi_2, \dots$ taking only two values, and hence giving rise to a *simple random walk* $X = (X_n)_{n \geq 1}$ (which is a Markov chain), led to the discovery of the law of large numbers (Sect. 5, Chap. 1, Vol. 1), the de Moivre–Laplace theorem (Sect. 6, Chap. 1, Vol. 1), the arcsine law (Sect. 10, Chap. 1, Vol. 1), and many other probabilistic regularities.

In this subsection we will consider two *discrete diffusion models* that provide a good illustration of how a simple random walk can describe real physical processes.

A. Ehrenfest Model. As in Example 7, consider the simple random walk with phase space $E = \{0, 1, \dots, N\}$ and reflecting barriers at 0 and N.

The transition probabilities from these states are $p_{01} = 1$, $p_{N,N-1} = 1$. At other states $i = 1, \dots, N-1$ only transitions by one step to the right or to the left are possible with probabilities

$$p_{ij} = \begin{cases} 1 - \frac{i}{N}, & j = i + 1, \\ \frac{i}{N}, & j = i - 1. \end{cases} \tag{27}$$

In 1907, Paul and Tatiana Ehrenfest [23] proposed this Markov chain as the model of *statistical mechanics* describing the motion of gas molecules from one container (A or B) to the other (B or A) through the membrane between them.

It is assumed that the total number of molecules in the two containers is N, and at each step one of them is randomly chosen (with probability $1/N$) and placed into the other container. The choice of the molecule at each step is made independently of its prehistory.

Let X_n be the number of molecules in container A at time n. The random mechanism of the motion of molecules fulfills the Markov property (Problem 2):

$$\mathsf{P}(X_{n+1} = j \,|\, X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i) = \mathsf{P}(X_{n+1} = j \,|\, X_n = i) \quad (28)$$

and

$$\mathsf{P}(X_{n+1} = j \,|\, X_n = i) = p_{ij}, \quad (29)$$

with p_{ij} defined by (27).

For this model there exists a stationary distribution $\mathbb{Q} = (q_0, q_1, \ldots, q_N)$ given by the following binomial formula (Problem 3):

$$q_j = C_N^j \left(\frac{1}{2}\right)^N, \quad j = 0, 1, \ldots, N. \quad (30)$$

All the states of this Markov chain are recurrent (Problem 4).

It is of interest to note that the maximum of q_j, $j = 0, 1, \ldots, N$, is attained for, say, even N, at the central value $j = N/2$, which corresponds to the most probable "equilibrium" state, when the number of molecules in both containers is the same.

Clearly, this equilibrium, which is established in the course of time, is of a probabilistic nature (specified by the stationary distribution $\mathbb{Q}$).

The possibility of "stabilization" of the number of molecules in the containers is quite understandable intuitively: the farther state i is from the central value, the larger the probability (by (27)) that the molecule will move *toward* this value.

B. D. Bernoulli–Laplace Model. This model, which is akin to the Ehrenfest model, was proposed by Daniel Bernoulli in 1769 and analyzed by Laplace in 1812 in the context of describing the exchange of particles between two ideal liquids.

Specifically, there are two containers, A and B, containing $2N$ particles, of which N particles are white and N particles are black.

The system is said to be in state i, where $i \in E = \{0, 1, \ldots, N\}$, if there are i white particles and $N - i$ black particles in container A. The assumption of ideal liquids means that in state i there are $N - i$ white particles and i black particles in container B, i.e., the number of particles in each container remains equal to N.

At each step n one particle in each container is randomly chosen (with probability $1/N$), and these particles interchange their containers. The two choices are independent, and each choice is independent of the choices in the previous steps.

Let X_n be the number of white particles in container A. Then the aforementioned mechanism of particle interchange obeys the Markov property (28) with transition probabilities p_{ij} in (29) given by the formula (Problem 5)

$$p_{ij} = \begin{cases} (i/N)^2, & j = i - 1, \\ (1 - (i/N))^2, & j = i + 1, \\ 2\,(i/N)(1 - (i/N)), & j = i, \end{cases} \quad (31)$$

and $p_{ij} = 0$ if $|i - j| > 1$, $i = 0, 1, \ldots, N$.

As in the Ehrenfest model, all the states are recurrent. There exists a unique stationary distribution $\mathbb{Q} = (q_0, q_1, \dots, q_N)$ determined by (Problem 5)

$$q_j = \frac{(C_N^j)^2}{(C_{2N}^N)^2}, \quad j = 0, 1, \dots, N. \tag{32}$$

4. At the beginning of this chapter we wrote that the issue of major interest here is the asymptotic behavior (as $n \to \infty$) of memoryless systems. In the previous sections we considered Markov chains with countable state space $E = \{i, j, \dots\}$ as a specific class of these systems, and we studied in them the behavior of transition probabilities $p_{ij}^{(n)}$ as $n \to \infty$. In particular, we investigated the asymptotic behavior of a simple random walk in which transitions are possible only to adjacent states.

Of great interest is the study of similar problems for Markov chains with more complicated state spaces. In this regard, see, for example, [14, 65].

5. The two models considered above (Ehrenfest and Bernoulli–Laplace) are said to be *discrete diffusion models*.

We will give an explanation to this expression in terms of asymptotic behavior of a simple random walk in R. Let $S_n = \xi_1 + \dots + \xi_n$, $n \geq 1$, $S_0 = 0$, where $\xi_1, \xi_2, \dots$ is a sequence of independent identically distributed random variables with $\mathsf{E}\,\xi_i = 0$, $\operatorname{Var} \xi_i = 1$. Let $X_0^n = 0$ and

$$X_t^n = \frac{S_{[nt]}}{\sqrt{n}} \quad \left(= \frac{1}{\sqrt{n}} \sum_{k=1}^{[nt]} \xi_k \right), \quad 0 < t \leq 1.$$

Clearly, the sequence $(0, X_{1/n}^N, X_{2/n}^N, \dots, X_1^N)$ may be regarded as a simple random walk in times $\Delta, 2\Delta, \dots, 1$ with $\Delta = 1/n$ and jumps of order $\sqrt{\Delta}$ ($\Delta X_{k\Delta}^n \equiv X_{k\Delta}^n - X_{(k-1)\Delta}^n = \xi_k \sqrt{\Delta}$).

As was pointed out in Remark 4, Sect. 8, Chap. 7, the finite-dimensional distributions of the random walk $X^n = (X_t^n)_{0 \leq t \leq 1}$ weakly converge to those of the Wiener process (Brownian motion) $W = (W_t)_{0 \leq t \leq 1}$. Moreover, we stated there that a functional convergence also holds, i.e., the weak convergence of the distributions of X^n to the distribution of W (in the same sense as the convergence of empirical processes to the Brownian bridge, see Sect. 13, Chap. 3, Vol. 1). The Wiener process is a typical (and the most important) example of a *diffusion process*; see [26, 21, 12]. This explains why the processes like X^n and those arising in the Ehrenfest and Bernoulli–Laplace models are naturally called *discrete diffusion models*.

6. Problems.

1. Prove Stirling's formula ($n! \sim \sqrt{2\pi}\, n^{n+1/2} e^{-n}$) using the following probabilistic arguments [5, Problem 27.18]. Let $S_n = X_1 + \dots + X_n$, $n \geq 1$, where $X_1, X_2, \dots$ are independent identically distributed random variables distributed according to Poisson's law with parameter $\lambda = 1$. Prove successively that

(a) $$\mathsf{E}\left(\frac{S_n - n}{\sqrt{n}}\right)^- = e^{-n} \sum_{k=0}^{n} \left(\frac{n-k}{\sqrt{n}}\right) \frac{n^k}{k!} = \frac{n^{n+1/2} e^{-n}}{n!};$$

(b) $\operatorname{Law}\left[\left(\frac{S_n - n}{\sqrt{n}}\right)^-\right] \to \operatorname{Law}[N^-]$,
where N is a standard normal random variable;

(c) $\mathsf{E}\left[\left(\frac{S_n - n}{\sqrt{n}}\right)^-\right] \to \mathsf{E}\,N^- = \frac{1}{\sqrt{2\pi}}$;

(d) $n! \sim \sqrt{2\pi}\, n^{n+1/2} e^{-n}$.

2. Establish the Markov property (28).
3. Prove (30).
4. Prove that all the states in the Ehrenfest model are recurrent.
5. Verify that (31) and (32) hold true.

9. Optimal Stopping Problems for Markov Chains

1. The subject of this section is closely related to Sect. 13, Chap. 7, which dealt with the *martingale approach* to optimal stopping problems for *arbitrary* stochastic sequences. In this section we focus on the case where stochastic sequences are generated by functions of states of Markov chains, which enables us to present and interpret the general results of Sect. 13, Chap. 7, in a simple and conceivable way.

2. Let $X = (X_n, \mathscr{F}_n, \mathsf{P}_x)$ be a homogeneous Markov chain with discrete time and phase space $(E, \mathscr{E})$.

We will assume that the space $(\Omega, \mathscr{F})$ on which the variables $X_n = X_n(\omega)$, $n \geq 0$, are defined is a coordinate space (as in Subsection 6, Sect. 1) and that the $X_n(\omega)$ are specified coordinate-wise, i.e., $X_n(\omega) = x_n$ if $\omega = (x_0, x_1, \dots) \in \Omega$. The σ-algebra $\mathscr{F}$ is defined as $\sigma(\bigcup \mathscr{F}_n)$, where $\mathscr{F}_n = \sigma(x_0, \dots, x_n)$, $n \geq 0$.

Remark. In the "general theory of optimal stopping rules" there is no need to require that Ω be a *coordinate* space. Nevertheless in the "general theory" one also must assume that the space is sufficiently "rich." (For details, see [69].)

In the present exposition, the assumption of coordinate space simplifies the presentation, in particular, regarding the generalized Markov property (Theorem 1 in Sect. 2), which was defined in this very framework.

As before, $P(x; B)$ will denote the transition function of our chain ($P(x; B) = \mathsf{P}_x\{X_1 \in B\}$), $x \in E$, $B \in \mathscr{E}$).

Let T be the one-step *transition operator* that acts on $\mathscr{E}$-measurable functions $f = f(x)$ satisfying $\mathsf{E}_x |f(X_1)| < \infty$, $x \in E$, in the following way:

$$(Tf)(x) = \mathsf{E}_x f(X_1) \quad \Big(= \int_E f(y)\, P(x; dy) \Big). \tag{1}$$

(For notational simplicity, we will write $Tf(x)$ instead of $(Tf)(x)$. A similar convention will also be used in other cases.)

3. To state the optimal stopping problem for the Markov chain X, let $g = g(x)$ be a given $\mathscr{E}$-measurable real-valued function such that $\mathsf{E}_x\,|g(X_n)| < \infty$, $x \in E$, for all $n \geq 0$ (or for $0 \leq n \leq N$ if we are to take the "optimal decision" before a time N specified a priori).

Let $\mathfrak{M}_0^n$ be the class of Markov times $\tau = \tau(\omega)$ (with respect to the filtration $(\mathscr{F}_k)_{0\leq k\leq N}$) taking values in the set $\{0, 1, \ldots, n\}$.

The following theorem is a "Markov" version of Theorems 1 and 2 in Sect. 13, Chap. 7.

Theorem 1. *For any $0 \leq n \leq N$ and $x \in E$, define the "price"*

$$s_n(x) = \sup_{\tau\in\mathfrak{M}_0^n} \mathsf{E}_x\, g(X_\tau), \tag{2}$$

where E_x is the expectation with respect to P_x.

Let

$$\tau_0^n = \min\{0 \leq k \leq n\colon s_{n-k}(X_k) = g(X_k)\} \tag{3}$$

and

$$Qg(x) = \max(g(x), Tg(x)). \tag{4}$$

Then the following statements hold true:

(1) *The Markov time τ_0^n is an optimal stopping time in the class $\mathfrak{M}_0^n$:*

$$\mathsf{E}_x\, g(X_{\tau_0^n}) = s_n(x) \tag{5}$$

for all $x \in E$.

(2) *The functions $s_n(x)$ are determined by the formula*

$$s_n(x) = Q^n g(x), \quad x \in E, \tag{6}$$

where $Q^0 g(x) = g(x)$ for $n = 0$.

(3) *The functions $s_n(x)$, $n \leq N$, satisfy the recurrence relations*

$$s_n(x) = \max(g(x), Ts_{n-1}(x)), \quad x \in E,\ 1 \leq n \leq N \tag{7}$$

(with $s_0(x) = g(x)$).

PROOF. Let us apply Theorems 1 and 2 of Sect. 13, Chap. 7, to the functions $f_n = g(X_n)$, $0 \leq n \leq N$. To this end, fix an initial state $x \in E$ and consider the functions V_n^N and v_n^N introduced therein. To highlight the dependence on the initial state, we will write $V_n^N = V_n^N(x)$. Thus,

$$V_n^N(x) = \sup_{\tau\in\mathfrak{M}_n^N} \mathsf{E}_x\, g(X_\tau), \tag{8}$$

where $\mathfrak{M}_n^N$ is the class of Markov times (with respect to the filtration $(\mathscr{F}_k)_{k\leq N}$) taking values in the set $\{n, n+1, \ldots, N\}$.

In accordance with (6) of Sect. 13, Chap. 7, the functions v_n^N are defined recursively:

$$v_N^N = g(X_N), \quad v_n^N = \max(g(X_n), \mathsf{E}_x(v_{n+1}^N \mid \mathscr{F}_n)). \tag{9}$$

By the generalized Markov property (Theorem 1 in Sect. 2),

$$\mathsf{E}_x(v_N^N \mid \mathscr{F}_{N-1}) = \mathsf{E}_x(g(X_N) \mid \mathscr{F}_{N-1}) = \mathsf{E}_{X_{N-1}}\, g(X_1) \quad (\mathsf{P}_x\text{-a.s.}), \tag{10}$$

where $\mathsf{E}_{X_{N-1}}\, g(X_1)$ is to be understood as follows (Sect. 2): for the function $\psi(x) = \mathsf{E}_x\, g(X_1)$, i.e., $\psi(x) = (Tg)(x)$, we define $\mathsf{E}_{X_{N-1}}\, g(X_1) \equiv \psi(X_{N-1}) = (Tg)(X_{N-1})$.

Hence $v_N^N = g(X_N)$ and

$$v_{N-1}^N = \max(g(X_{N-1}), (Tg)(X_{N-1})) = (Qg)(X_{N-1}). \tag{11}$$

Proceeding in a similar manner, we find that

$$v_n^N = (Q^{N-n}g)(X_n) \tag{12}$$

for all $0 \le n \le N-1$ and, in particular,

$$v_0^N = (Q^N g)(X_0) = (Q^N g)(x) \quad (\mathsf{P}_x\text{-a.s.}).$$

By (13) of Sect. 13, Chap. 7, we have $v_0^N = V_0^N$. Since $V_0^N = V_0^N(x) = s_N(x)$, we have $s_N(x) = (Q^N g)(x)$, which proves (6) for $n = N$ (and similarly for any $n < N$).

The recurrence formulas (7) follow from (6) and the definition of Q.

We show now that the stopping time defined by (3) is *optimal* (for $n = N$) in the class $\mathfrak{M}_0^N$ (and similarly in the classes $\mathfrak{M}_0^n$ for $n < N$).

By Theorem 1 of Sect. 13, Chap. 7, the optimal stopping time is

$$\tau_0^N = \min\{0 \le k \le N\colon v_k^N = g(X_k)\}.$$

Now (12) and the fact established above that $s_n(x) = (Q^n g)(x)$ for any $n \ge 0$ imply that

$$v_k^N = (Q^{N-k}g)(X_k) = s_{N-k}(X_k). \tag{13}$$

Therefore

$$\tau_0^N = \min\{0 \le k \le N\colon s_{N-k}(X_k) = g(X_k)\}, \tag{14}$$

which proves the optimality of this stopping time in the class $\mathfrak{M}_0^N$.

□

4. Use the notation

$$\mathbb{D}_k^N = \{x \in E\colon s_{N-k}(x) = g(x)\}, \tag{15}$$

$$\mathbb{C}_k^N = E \setminus \mathbb{D}_k^N = \{x \in E\colon s_{N-k}(x) > g(x)\}. \tag{16}$$

Then we see from (14) that

$$\tau_0^N(\omega) = \min\{0 \le k \le N\colon X_k(\omega) \in \mathbb{D}_k^N\}, \tag{17}$$

and, by analogy with the sets D_k^N and C_k^N (in Ω) introduced in Subsection 6, Sect. 13, Chap. 7, the sets

$$\mathbb{D}_0^N \subseteq \mathbb{D}_1^N \subseteq \cdots \subseteq \mathbb{D}_N^N = E, \tag{18}$$

$$\mathbb{C}_0^N \supseteq \mathbb{C}_1^N \supseteq \cdots \supseteq \mathbb{C}_N^N = \varnothing \tag{19}$$

can be called the *stopping sets* and *continuation of observation sets* (in E), respectively.

Let us point out the specific features of the stopping problems in the case of Markov chains. Unlike the general case, the answer to the question of whether observations are to be stopped or continued is given in the Markov case in terms of the states of the Markov chain itself ($\tau_0^N = \min\{0 \le k \le N\colon X_k \in \mathbb{D}_k^N\}$), in other words, depending on the position of the wandering particle. And the *complete* solution of the optimal stopping problems (i.e., the description of the "price" $s_N(x)$ and the optimal stopping time τ_0^N) is obtained from the recurrence "dynamic programming equations" (7) by finding successively the functions $s_0(x) = g(x)$, $s_1(x), \ldots, s_N(x)$.

5. Consider now the optimal stopping problem assuming that $\tau \in \mathfrak{M}_0^\infty$, where $\mathfrak{M}_0^\infty$ is the class of all *finite* Markov times. (The assumption $\tau \in \mathfrak{M}_0^N$ means that $\tau \le N$, while the assumption $\tau \in \mathfrak{M}_0^\infty$ means only that $\tau = \tau(\omega) < \infty$ for all $\omega \in \Omega$.)

Thus we consider the price

$$s(x) = \sup_{\tau \in \mathfrak{M}_0^\infty} \mathsf{E}_x\, g(X_\tau). \tag{20}$$

To avoid any questions about the existence of expectations $\mathsf{E}_x\, g(X_\tau)$, we can assume, for example, that

$$\mathsf{E}_x \left(\sup_n g^-(X_n) \right) < \infty, \quad x \in E. \tag{21}$$

Clearly, this assumption is satisfied whenever $g = g(x)$ is *bounded* ($|g(x)| \le C$, $x \in E$). In particular, (21) holds if the state space E is *finite*.

The definitions of the prices $s_N(x)$ and $s(x)$ imply that

$$s_N(x) \le s_{N+1}(x) \le \cdots \le s(x) \tag{22}$$

for all $x \in E$. Of course, it is natural to expect that $\lim_{N\to\infty} s_N(x)$ equals $s(x)$. If this is the case, then, passing to the limit in (7), we find that $s(x)$ satisfies the equation

$$s(x) = \max(g(x), Ts(x)), \quad x \in E. \tag{23}$$

This equation implies that $s(x)$, $x \in E$, fulfills the following "variational inequalities":

$$s(x) \ge g(x), \tag{24}$$

$$s(x) \ge Ts(x). \tag{25}$$

Inequality (24) says that $s(x)$ is a *majorant* of $g(x)$. Inequality (25), according to the terminology of the general theory of Markov processes, means that $s(x)$ is an *excessive* or a *superharmonic* function.

Therefore, if we could establish that $s(x)$ satisfies (23), then the price $s(x)$ would be an *excessive majorant* for $g(x)$.

Note now that if a function $v(x)$ is an excessive majorant for $g(x)$, then, obviously, the following variational inequalities hold:

$$v(x) \geq \max(g(x), Tv(x)), \quad x \in E. \tag{26}$$

It turns out, however, that if we assume additionally that $v(x)$ is the *least excessive majorant* then (26) becomes an equality, i.e., $v(x)$ satisfies the equation

$$v(x) = \max(g(x), Tv(x)), \quad x \in E. \tag{27}$$

Lemma 1. *Any least excessive majorant $v(x)$ of $g(x)$ satisfies equation* (27).

PROOF. The proof is fairly simple. Clearly, $v(x)$ satisfies inequality (26). Let $v_1(x) = \max(g(x), Tv(x))$. Since $v_1(x) \geq g(x)$ and $v_1(x) \leq v(x)$, $x \in E$, we have

$$Tv_1(x) \leq Tv(x) \leq \max(g(x), Tv(x)) = v_1(x).$$

Therefore $v_1(x)$ is an excessive majorant for $g(x)$. But $v(x)$ is the *least* excessive majorant. Hence $v(x) \leq v_1(x)$, i.e., $v(x) \leq \max(g(x), Tv(x))$. Together with (26) this implies (27).

□

The preliminary discussions presented earlier, which were based on the assumption $s(x) = \lim_{N\to\infty} s_N(x)$ and led to (23), as well as the statement of Lemma 1 suggest a characterization of the the price $s(x)$, namely, that it is likely to be the least excessive majorant of $g(x)$.

Indeed, the following theorem is true.

Theorem 2. *Suppose a function $g = g(x)$ satisfies $\mathsf{E}_x\left[\sup_n g^-(X_n)\right] < \infty$, $x \in E$. Then the following statements are valid.*

(a) *The price $s = s(x)$ is the least excessive majorant of $g = g(x)$.*

(b) *The price $s(x)$ is equal to $\lim_{N\to\infty} s_N(x) = \lim_{N\to\infty} Q^N g(x)$ and satisfies the Wald–Bellman dynamic programming equation*

$$s(x) = \max(g(x), Ts(x)), \quad x \in E.$$

(c) *If $\mathsf{E}_x\left[\sup_n |g(X_n)|\right] < \infty$, $x \in E$, then for any $\varepsilon > 0$ the stopping time*

$$\tau_\varepsilon^* = \min\{n \geq 0 \colon s(X_n) \leq g(X_n) + \varepsilon\}$$

is ε-optimal in the class $\mathfrak{M}_0^\infty$, i.e.,

$$s(x) - \varepsilon \leq \mathsf{E}_x\, g(X_{\tau_\varepsilon^*}), \quad x \in E.$$

If $\mathsf{P}_x\{\tau_0^* < \infty\} = 1$, $x \in E$, *then the stopping time* τ_0^* *is optimal* (0-optimal), *i.e.*,

$$s(x) = \mathsf{E}_x\, g(X_{\tau_0^*}), \quad x \in E. \tag{28}$$

(d) *If the set E is finite, then* τ_0^* *belongs to* $\mathfrak{M}_0^\infty$ *and is optimal.*

Remark. The stopping time $\tau_0^* = \min\{n \geq 0\colon s(X_n) = g(X_n)\}$ may happen to be infinite for some $x \in E$ with positive probability, $\mathsf{P}_x\{\tau_0^* = \infty\} > 0$. (This can occur even in the case of a countable set of states, Problem 1.) In view of this we should agree about the meaning of $\mathsf{E}_x\, g(X_\tau)$ if τ can equal $+\infty$ since the value X_∞ has not been defined.

The value of $g(X_\infty)$ is often defined to be $\limsup_n g(X_n)$ (Subsection 1, Sect. 13, Chap. 7, and [69]). Another option is to consider $g(X_\tau)\, I(\tau < \infty)$ instead of $g(X_\tau)$. Then denoting by $\overline{\mathfrak{M}}_0^\infty$ the class of *all* Markov times, possibly equal to $+\infty$, the price

$$\bar{s}(x) = \sup_{\tau \in \overline{\mathfrak{M}}_0^\infty} \mathsf{E}_x\, g(X_\tau)\, I(\tau < \infty) \tag{29}$$

is well defined, so that we can consider the optimal stopping problem in the class $\overline{\mathfrak{M}}_0^\infty$ also.

PROOF OF THEOREM 2. We will give the proof only for the case of a finite set E. In this case the proof is rather simple and clarifies quite well the appearance of excessive functions in optimal stopping problems. For the proof in the general case, see [69, 22].

Proof of (a). Let us show that $s(x)$ is excessive, i.e., $s(x) \geq Ts(x)$, $x \in E$.

It is obvious that for any state $y \in E$ and any $\varepsilon > 0$ there is a finite (P_y-a.s.) Markov time $\tau_y \in \mathfrak{M}_0^\infty$ (depending in general on $\varepsilon > 0$) such that

$$\mathsf{E}_y\, g(X_{\tau_y}) \geq s(y) - \varepsilon. \tag{30}$$

Using these times τ_y, $y \in E$, we will construct one more time $\hat{\tau}$, which will determine the following strategy of the choice of the stopping time.

Let the particle be in the state $x \in E$ at the initial time instant. The observation process is surely not stopped at this time, and one observation is produced. Let at $n = 1$ the particle occur in the state $y \in E$. Then the strategy determined by $\hat{\tau}$ consists, informally speaking, in treating the "life" of the particle as if it started anew at this time subject further to the stopping rule governed by τ_y.

The formal definition of $\hat{\tau}$ is as follows.

Let $y \in E$. Consider the event $\{\omega\colon \tau_y(\omega) = n\}$, $n \geq 0$. Since τ_y is a Markov time, this event belongs to $\mathscr{F}_n$. We assume that Ω is a *coordinate* space generated by sequences $\omega = (x_0, x_1, \dots)$ with $x_i \in E$, and $\mathscr{F}_n = \sigma(x_0, \dots, x_n)$. This implies that the set $\{\omega\colon \tau_y(\omega) = n\}$ can be written $\{\omega\colon (X_0(\omega), \dots, X_n(\omega)) \in B_y(n)\}$, where $B_y(n)$ is a set in $\mathscr{E}^{n+1} = \mathscr{E} \otimes \cdots \otimes \mathscr{E}$ ($n+1$ times). (See also Theorem 4 in Sect. 2, Chap. 2, Vol. 1).

By definition, the Markov time $\hat{\tau} = \hat{\tau}(\omega)$ equals $n+1$ with $n \geq 0$ on the set

$$\hat{A}_n = \sum_{y \in E} \{\omega\colon X_1(\omega) = y, (X_1(\omega), \dots, X_{n+1}(\omega)) \in B_y(n)\}.$$

(The time $\hat{\tau}$ can be described heuristically as a rule for making an observation at time $n = 0$, whatever the state x, and using subsequently the Markov time τ_y if $X_1 = y$.)

Since $\sum_{n\geq 0} \hat{A}_n = \Omega$, $\hat{\tau} = \hat{\tau}(\omega)$ is well defined for all $\omega \in \Omega$ and is a Markov time (Problem 2).

Using this construction, the generalized Markov property, and (30), we find that, for any $x \in E$,

$$\begin{aligned}
\mathsf{E}_x\, g(X_{\hat{\tau}}) &= \sum_{n\geq 0}\sum_{y\in E}\sum_{z\in E} \mathsf{P}_x\{X_1 = y, (X_1,\dots,X_{n+1}) \in B_y(n), X_{n+1} = z\}\, g(z) \\
&= \sum_{n\geq 0}\sum_{y\in E}\sum_{z\in E} p_{xy}\, \mathsf{P}_y\{X_0 = y, (X_0,\dots,X_n) \in B_y(n), X_n = z\}\, g(z) \\
&= \sum_{n\geq 0}\sum_{y\in E}\sum_{z\in E} p_{xy}\, \mathsf{P}_y\{(X_0,\dots,X_n) \in B_y(n), X_n = z\}\, g(z) \\
&= \sum_{y\in E} p_{xy}\, \mathsf{E}_y\, g(X_{\tau_y}) \geq \sum_{y\in E} p_{xy}(s(y) - \varepsilon) = Ts(x) - \varepsilon.
\end{aligned}$$

Thus,

$$s(x) = \mathsf{E}_x\, g(X_{\hat{\tau}}) \geq Ts(x) - \varepsilon, \quad x \in E,$$

and, since $\varepsilon > 0$ is arbitrary,

$$s(x) \geq Ts(x), \quad x \in E,$$

which proves that $s = s(x)$, $x \in E$, is an excessive function.

The property just obtained, that $s(x)$ is excessive (superharmonic), immediately provides the following important result.

Corollary 1. *For any $x \in E$ the process* (*sequence*)

$$s = (s(X_n))_{n\geq 0} \tag{31}$$

is a supermartingale (*with respect to the P_x-probability*).

Theorem 1 of Sect. 2, Chap. 7, applied to this supermartingale implies that for any stopping time $\tau \in \mathfrak{M}_0^\infty$ we have

$$s(x) \geq \mathsf{E}_x\, s(X_\tau), \quad x \in E, \tag{32}$$

and if σ and τ are Markov times in $\mathfrak{M}_0^\infty$ such that $\sigma \leq \tau$ (P_x-a.s., $x \in E$), then

$$\mathsf{E}_x\, s(X_\sigma) \geq \mathsf{E}_x\, s(X_\tau),\ x \in E. \tag{33}$$

(Note that the conditions of Theorem 1, Sect. 2, Chap. 7, mentioned earlier are fulfilled in this case since space E is finite.)

Now we deduce from (32) the following corollary.

Corollary 2. *Let the function $g = g(x)$, $x \in E$, in the optimal stopping problem* (20) *be excessive* (*superharmonic*). *Then $\tau_0^* \equiv 0$ is an optimal stopping time.*

Proof of (b). Let us show that $s(x) = \lim_N s_N(x)$, $x \in E$.

Since $s_N(x) \le s_{N+1}(x)$, the limit $\lim_N s_N(x)$ exists. Denote it by $\bar{s}(x)$. Since E is finite and the $s_N(x)$, $N \ge 0$, satisfy the recurrence relations

$$s_N(x) = \max(g(x), Ts_{N-1}(x)),$$

we can pass to the limit as $N \to \infty$ in them to obtain that

$$\bar{s}(x) = \max(g(x), T\bar{s}(x)).$$

This implies that $\bar{s}(x)$ is an excessive majorant for $g(x)$. But $s(x)$ is the least excessive majorant. Hence $s(x) \le \bar{s}(x)$. On the other hand, since $s_N(x) \le s(x)$ for any $N \ge 0$, we have $\bar{s}(x) \le s(x)$.

Therefore $\bar{s}(x) = s(x)$, which proves statement (b).

Proof of (c, d). Finally, we will show that the stopping time

$$\tau_0^* = \min\{n \ge 0 \colon s(X_n) = g(X_n)\}, \tag{34}$$

i.e., the time

$$\tau_0^* = \min\{n \ge 0 \colon X_n \in \mathbb{D}^*\} \tag{35}$$

of the *first* entering the (stopping) set

$$\mathbb{D}^* = \{x \in E \colon s(x) = g(x)\} \tag{36}$$

is (for finite E) optimal in the class $\mathfrak{M}_0^\infty$.

To this end, note that the set $\mathbb{D}^*$ is *not empty* because it certainly contains those $\tilde{x}$ for which $g(\tilde{x}) = \max_{x \in E} g(x)$. In these states $s(\tilde{x}) = g(\tilde{x})$, and it is obvious that the optimal strategy with regard to these states is to stop when getting into $\tilde{x}$. This is exactly what the stopping time τ_0^* does.

To discuss τ_0^* from the point of view of optimality in the class $\mathfrak{M}_0^\infty$, we must first of all establish that this stopping time belongs to this class, i.e., that

$$\mathsf{P}_x\{\tau_0^* < \infty\} = 1, \quad x \in E. \tag{37}$$

This is true indeed under our assumption that the state space E is *finite*. (For an *infinite* E this is, in general, not the case; see Problem 1).

For the proof of this, note that the event $\{\tau_0^* = \infty\}$ is the same as $A = \bigcap_{n \ge 0}\{X_n \notin \mathbb{D}^*\}$. Thus, we are to show that $\mathsf{P}_x(A) = 0$ for all $x \in E$.

Obviously, this is the case if $\mathbb{D}^* = E$.

Let $\mathbb{D}^* \ne E$. Since E is *finite*, there is $\alpha > 0$ such that $g(y) \le s(y) - \alpha$ for all $y \in E \setminus \mathbb{D}^*$. Then, for any $\tau \in \mathfrak{M}_0^\infty$,

$$\mathsf{E}_x\, g(X_\tau) = \sum_{n=0}^{\infty} \sum_{y\in E} \mathsf{P}_x\{\tau = n, X_n = y\}\, g(y)$$

$$= \sum_{n=0}^{\infty} \sum_{y\in \mathbb{D}^*} \mathsf{P}_x\{\tau = n, X_n = y\}\, g(y) + \sum_{n=0}^{\infty} \sum_{y\in E\setminus \mathbb{D}^*} \mathsf{P}_x\{\tau = n, X_n = y\}\, g(y)$$

$$\le \sum_{n=0}^{\infty} \sum_{y\in \mathbb{D}^*} \mathsf{P}_x\{\tau = n, X_n = y\}\, s(y) + \sum_{n=0}^{\infty} \sum_{y\in E\setminus \mathbb{D}^*} \mathsf{P}_x\{\tau = n, X_n = y\}\, (s(y) - \alpha)$$

$$\le \mathsf{E}_x\, s(X_\tau) - \alpha\, \mathsf{P}_x(A) \le s(x) - \alpha\, \mathsf{P}_x(A), \quad (38)$$

where the last inequality follows because $s(x)$ is excessive (superharmonic) and satisfies Eq. (32).

Taking the supremum over all $\tau \in \mathfrak{M}_0^\infty$ on the left-hand side of (38), we obtain

$$s(x) \le s(x) - \alpha\, \mathsf{P}_x(A), \quad x \in E.$$

But $|s(x)| < \infty$ and $\alpha > 0$. Therefore $\mathsf{P}_x(A) = 0$, $x \in E$, which proves the finiteness of the stopping time τ_0^*.

We will show now that this stopping time is optimal in the class $\mathfrak{M}_0^\infty$. By the definition of τ_0^*,

$$s(X_{\tau_0^*}) = g(X_{\tau_0^*}). \quad (39)$$

With this property in mind, consider the function $\gamma(x) = \mathsf{E}_x\, g(X_{\tau_0^*}) = \mathsf{E}_x\, s(X_{\tau_0^*})$. In what follows, we show that $\gamma(x)$ has the properties:

(i) $\gamma(x)$ is excessive;
(ii) $\gamma(x)$ majorizes $g(x)$, i.e., $\gamma(x) \ge g(x)$, $x \in E$;
(iii) Obviously, $\gamma(x) \le s(x)$.

Properties (i) and (ii) imply that $\gamma(x)$ is an excessive majorant for $s(x)$, which, in turn, is the *least* excessive majorant for $g(x)$. Hence, by (iii), $\gamma(x) = s(x)$, $x \in E$, which yields

$$s(x) = \mathsf{E}_x\, g(X_{\tau_0^*}), \quad x \in E,$$

thereby proving the required optimality of τ_0^* within $\mathfrak{M}_0^\infty$.

Let us prove (i). Use the notation $\bar\tau = \min\{n \ge 1 \colon X_n \in \mathbb{D}^*\}$. This is a Markov time, $\tau_0^* \le \bar\tau$, $\bar\tau \in \mathfrak{M}_1^\infty$, and since $s(x)$ is excessive, we have, by (33),

$$\mathsf{E}_x\, s(X_{\bar\tau}) \le \mathsf{E}_x\, s(X_{\tau_0^*}), \quad x \in E. \quad (40)$$

Next, using the generalized Markov property (see (2) in Theorem 1, Sect. 2) we obtain

$$\mathsf{E}_x\, s(X_{\bar\tau}) = \sum_{n=1}^{\infty} \sum_{y\in \mathbb{D}^*} \mathsf{P}_x\{X_1 \notin \mathbb{D}^*, \dots, X_{n-1} \notin \mathbb{D}^*, X_n = y\}\, s(y)$$

$$= \sum_{n=1}^{\infty} \sum_{y \in \mathbb{D}^*} \sum_{z \in E} p_{xz} \mathsf{P}_z\{X_0 \notin \mathbb{D}^*, \ldots, X_{n-2} \notin \mathbb{D}^*, X_{n-1} = y\}\, s(y)$$

$$= \sum_{z \in E} p_{xz} \mathsf{E}_z\, s(X_{\tau_0^*}). \quad (41)$$

Hence, by (40), we find that

$$\mathsf{E}_x\, s(X_{\tau_0^*}) \geq \sum_{z \in E} p_{xz} \mathsf{E}_z\, s(X_{\tau_0^*}),$$

i.e.,

$$\gamma(x) \geq \sum_{z \in E} p_{xz} \gamma(z), \quad x \in E,$$

which shows that the function $\gamma(x)$ is excessive.

It remains to show that $\gamma(x)$ majorizes $g(x)$. If $x \in \mathbb{D}^*$, then $\tau_0^* = 0$ and, obviously, $\gamma(x) = \mathsf{E}_x\, g(X_{\tau_0^*}) = g(x)$.

Consider the set $E \setminus \mathbb{D}^*$, and let $E_0^* = \{x \in E \setminus \mathbb{D}^* \colon \gamma(x) < g(x)\}$. Let x_0^* be the point in the *finite* set E_0^*, where the maximum of $g(x) - \gamma(x)$ is attained:

$$g(x_0^*) - \gamma(x_0^*) = \max_{x \in E_0^*} (g(x) - \gamma(x)).$$

Define the function

$$\tilde{\gamma}(x) = \gamma(x) + [g(x_0^*) - \gamma(x_0^*)], \quad x \in E. \quad (42)$$

Clearly, this function is excessive (being the sum of an excessive function and a constant) and

$$\tilde{\gamma}(x) - g(x) = [g(x_0^*) - \gamma(x_0^*)] - [g(x) - \gamma(x)] \geq 0$$

for all $x \in E$. Thus, $\tilde{\gamma}(x)$ is an excessive majorant for $g(x)$, and hence $\tilde{\gamma}(x) \geq s(x)$, since $s(x)$ is the least excessive majorant for $g(x)$.

This implies that

$$\tilde{\gamma}(x_0^*) \geq s(x_0^*).$$

But $\tilde{\gamma}(x_0^*) = g(x_0^*)$ by (42); therefore $g(x_0^*) \geq s(x_0^*)$. Since $s(x) \geq g(x)$ for all $x \in E$, we obtain $g(x_0^*) = s(x_0^*)$, i.e., x_0^* is in $\mathbb{D}^*$, while $x^* \in E \setminus \mathbb{D}^*$ by assumption.

This contradiction shows that $E \setminus \mathbb{D}^* = \varnothing$, so $\gamma(x) \geq g(x)$ for all $x \in E$.

□

6. Let us give some examples.

EXAMPLE 1. Consider the simple random walk with two absorbing barriers described in Example 5, Sect. 8, assuming that $p = q = 1/2$ (symmetric random walk). If a function $\gamma(x)$, $x \in E = \{0, 1, \ldots, N\}$, is excessive for this random walk, then

$$\gamma(x) \geq \frac{1}{2}\gamma(x-1) + \frac{1}{2}\gamma(x+1) \tag{43}$$

for all $x = 1, \ldots, N-1$.

Suppose we are given a function $g = g(x)$, $x \in \{0, 1, \ldots, N\}$. Since states 0 and N are absorbing, the function $s(x)$ must be sought among the functions $\gamma(x)$ satisfying condition (43) and boundary conditons $\gamma(0) = g(0)$, $\gamma(N) = g(N)$.

Condition (43) means that $\gamma(x)$ is *convex* on the set $\{1, 2, \ldots, N-1\}$. Hence we can conclude that the price $s(x)$ in the problem $s(x) = \sup_{\tau \in \mathfrak{M}_0^\infty} \mathsf{E}_x\, g(X_\tau)$ is the least *convex* function subject to the boundary conditions $s(0) = g(0)$, $s(N) = g(N)$. A visual description of the rule for determinimg $s(x)$ is as follows. Let us cover the values of $g(x)$ by a stretched thread. In Fig. 42 this thread passes through the points $(0, a)$, $(1, b)$, $(4, c)$, $(6, d)$, where the points $0, 1, 4, 6$ form the set $\mathbb{D}^*$ of stopping states. In these points we have $s(x) = g(x)$. The values of $s(x)$ at other points $x = 2, 3, 5$ are determined by linear interpolation. In the general case, the convex hull $s(x)$ for all $x \in E$ is constructed in a similar manner.

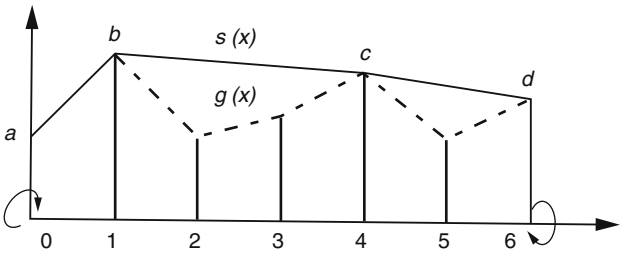


Fig. 42 The function $g(x)$ (dashed line) and its convex hull $s(x)$, $x = 0, 1, \ldots, 6$

EXAMPLE 2. Consider, as in Example 7, Sect. 8, a simple symmetric ($p = q = 1/2$) random walk over the set $E = \{0, 1, \ldots, N\}$ with reflecting barriers at 0 and N. This random walk is positive recurrent, which implies that the optimal rule in the optimal stopping problem $s(x) = \sup_{\tau \in \mathfrak{M}_0^\infty} \mathsf{E}_x\, g(X_\tau)$ has a very simple and natural structure: wait until the particle reaches a point where $g(x)$ attains its maximum and then stop the observations.

EXAMPLE 3. Suppose that in a simple symmetric random walk over the set $E = \{0, 1, \ldots, N\}$ state 0 is absorbing and N is reflecting. Let x_0 be the state where $g(x)$ attains its maximum and that is closest to N if there are several maxima. Then the optimal stopping rule is as follows: If $x_0 \leq x \leq N$, then the walk stops when (with P_x-probability 1) the state x_0 is achieved, while for x between 0 and x_0 the decision rule is the same as in Example 1 taking $E = \{0, 1, \ldots, x_0\}$ with absorbing barriers 0 and x_0.

7. Finally, consider the widely known "best choice problem" also known as "the fiancée problem," "the secretary problem," and so on (see [66], [69, 22, 5]). We will interpret it as "the fiancée problem."

Suppose that a fiancée is going to choose the best of N candidates. It is assumed that N is known and the candidates are ranked. Let, for definiteness, the best candidate be ranked number N, the second $N-1$, and so on to 1.

The candidates are presented to the fiancée in random order, which is formalized as follows. Let $(a_1, a_2, \dots, a_N)$ be a random permutation taking on any of $N!$ possible permutations of $(1, 2, \dots, N)$ with probability $1/N!$. Then the fiancée meets first the candidate of rank a_1, then a_2, and so on till the last a_Nth.

The fiancée can choose one of the candidates according to the following rules. She does not know the ranks of the candidates and can only compare the current candidate with those she has seen before. Once rejected, a candidate *cannot return* anymore (even though he might have been the best one).

Based on a consecutive assessment of candidates (keeping in mind the outcomes of their pairwise comparison and the "quality" of rejected candidates), the fiancée must choose a stopping time τ^* such that

$$\mathsf{P}\{a_{\tau^*} = N\} = \sup_{\tau} \mathsf{P}\{a_\tau = N\}, \tag{44}$$

where τ runs over a class of stopping times $\mathfrak{M}_1^N$ determined by information accessible to the fiancée.

To describe the class $\mathfrak{M}_1^N$ more precisely, let us construct the rank sequence $X = (X_1, X_2, \dots)$ depending on $\omega = (a_1, a_2, \dots, a_N)$, which will determine the action of the fiancée.

That is, let $X_1 = 1$, and let X_2 be the order number of the candidate that dominates all the preceding ones. If, for example, $X_2 = 3$, this means that in $\omega = (a_1, a_2, \dots, a_N)$ we have $a_1 > a_2$, but $a_3 > a_1$ $(> a_2)$. If, say, $X_3 = 5$, then $a_3 > a_4$, but $a_5 > a_3$ $(> a_4)$.

There might be at most N dominants (when $(a_1, a_2, \dots, a_N) = (1, 2, \dots, N)$). If $\omega = (a_1, a_2, \dots, a_N)$ contains m dominants, we set $X_{m+1} = X_{m+2} = \cdots = N + 1$.

The class $\mathfrak{M}_1^N$ of admissible stopping times will consist of those $\tau = \tau(\omega)$ for which

$$\{\omega\colon \tau(\omega) = n\} \in \mathscr{F}_n^X,$$

where $\mathscr{F}_n^X = \sigma(X_1, \dots, X_n)$, $1 \le n \le N$.

Consider the structure of the rank sequence $X = (X_1, X_2, \dots)$ in more detail. It is not hard to see (Problem 3) that this sequence is a homogeneous Markov chain (with phase space $E = \{1, 2, \dots, N+1\}$). The transition probabilities of this chain are given by the following formulas:

$$p_{ij} = \frac{i}{j(j-1)}, \quad 1 \le i < j \le N, \tag{45}$$

$$p_{i,N+1} = \frac{i}{N}, \quad 1 \le i \le N, \tag{46}$$

$$p_{N+1,N+1} = 1. \tag{47}$$

It is seen from these formulas that the state $N+1$ is *absorbing* and all the transitions on set E are upward, i.e., the only possible transitions are $i \to j$ with $j > i$.

Remark. Formula (45) follows from the following simple arguments taking into account that the probability of every sequence $\omega = (a_1, \dots, a_N)$ is $1/N!$.

For $1 \le i < j \le N$ the transition probability is equal to

$$p_{ij} = \mathsf{P}(X_{n+1} = j \,|\, X_n = i) = \frac{\mathsf{P}\{X_n = i, X_{n+1} = j\}}{\mathsf{P}\{X_n = i\}}. \tag{48}$$

The event $\{X_n = i, X_{n+1} = j\}$ means that a_j dominates among $a_1, \dots, a_j$ and $a_j > a_i$. The probability of this event is $\frac{(j-2)!}{j!} = \frac{1}{j(j-1)}$. In the same way, the event $\{X_n = i\}$ means that a_i dominates among $a_1, \dots, a_i$ and the probability of this event is $\frac{(i-1)!}{i!} = \frac{1}{i}$. These considerations and (48) imply (45).

For the proof of (46) it suffices to note that if $X_n = i$, then $X_{n+1} = N+1$ implies that a_i dominates both $a_{i+1}, \dots, a_N$ and $a_1, \dots, a_{i-1}$. Formula (47) is obvious.

Suppose now that the fiancée adopted a stopping time τ (with respect to the system of σ-algebras $(\mathscr{F}_n^X)$) and $X_\tau = i$. Then the conditional probability that this stopping time is successful (i.e., $a_\tau = N$) is, according to (46), equal to $\frac{X_\tau}{N}$ $\left(= \frac{i}{N}\right)$. Therefore

$$\mathsf{P}\{a_\tau = N\} = \mathsf{E}\,\frac{X_\tau}{N},$$

and hence seeking the optimal stopping time τ^* (i.e., the stopping time for which $\mathsf{P}\{a_{\tau^*} = N\} = \sup_\tau \mathsf{P}\{a_\tau = N\}$) reduces to the optimal stopping problem

$$V^* = \sup_\tau \mathsf{E}\,\frac{X_\tau}{N}, \tag{49}$$

where τ is a Markov time with respect to $(\mathscr{F}_n^X)$.

It is assumed in (49) that $X_1 = 1$. In accordance with the general method of solving optimal stopping problems for Markov chains, use the notation

$$v(i) = \sup_\tau \mathsf{E}_i\, g(X_\tau),$$

where E_i is the expectation given that $X_1 = i$, and

$$g(i) = \frac{i}{N}, \quad i \le N, \quad g(N+1) = 0.$$

As we know (Theorem 2), the function $v(i)$, $1 \le i \le N+1$, is an excessive majorant for $g(i)$, $1 \le i \le N+1$:

$$v(i) \ge Tv(i) = \sum_{j=i+1}^{N} \frac{i}{j(j-1)}\, v(j), \tag{50}$$

$$v(i) \ge g(i), \tag{51}$$

and moreover it is the *least* excessive majorant. The same Theorem 2 implies that $v(i)$, $1 \le i \le N+1$, satisfies the equation

$$v(i) = \max(g(i), Tv(i)), \quad 1 \le i \le N+1, \tag{52}$$

and, as is easy to see, $v(i)$ must fulfill

$$v(N+1) = 0, \quad v(N) = g(N) = 1.$$

Denote by $\mathbb{D}^*$ the set of the states $i \in E$, where observations are stopped. By Theorem 1, this set has the form

$$\mathbb{D}^* = \{i \in E\colon v(i) = g(i)\}.$$

Accordingly, the set where observations are continued is

$$\mathbb{C}^* = \{i \in E\colon v(i) > g(i)\}.$$

Therefore, if $i \in \mathbb{D}^*$, then

$$\begin{aligned} g(i) = v(i) \ge Tv(i) &= \sum_{j=i+1}^{N} \frac{i}{j} \cdot \frac{1}{j-1}\, v(j) \ge \sum_{j=i+1}^{N} \frac{i}{j} \cdot \frac{1}{j-1}\, g(j) \\ &= \sum_{j=i+1}^{N} \frac{i}{j} \cdot \frac{1}{j-1} \cdot \frac{j}{N} = g(i) \sum_{j=i+1}^{N} \frac{1}{j-1}. \end{aligned}$$

Hence, for $i \in \mathbb{D}^*$, we must have

$$\sum_{j=i+1}^{N} \frac{1}{j-1} \le 1.$$

Further, if this inequality is fulfilled and the values $i+1, \dots, N$ belong to $\mathbb{D}^*$, then

$$Tv(i) = \sum_{j=i+1}^{N} \frac{i}{j} \cdot \frac{1}{j-1}\, g(j) = g(i) \sum_{j=i+1}^{N} \frac{1}{j-1} \le g(i),$$

so that i also belongs to $\mathbb{D}^*$.

The preceding arguments (together with $N \in \mathbb{D}^*$, since $v(N) = g(N)$) show that the set $\mathbb{D}^*$ has the form

$$\mathbb{D}^* = \{i^*, i^*+1, \dots, N, N+1\},$$

where $i^* = i^*(N)$ is determined by the inequalities

$$\frac{1}{i^*} + \frac{1}{i^*+1} + \cdots + \frac{1}{N-1} \le 1 < \frac{1}{i^*-1} + \frac{1}{i^*} + \cdots + \frac{1}{N-1}, \tag{53}$$

which imply that for large N

$$i^*(N) \sim \frac{N}{e}. \tag{54}$$

Indeed, for any $n \geq 2$ we have

$$\log(n+1) - \log n < \frac{1}{n} < \log n - \log(n-1).$$

Hence

$$\log \frac{N}{n} < \frac{1}{n} + \cdots + \frac{1}{N-1} < \log \frac{N-1}{n-1}.$$

Together with (53), this yields

$$\log \frac{N}{i^*(N)} < 1 < \log \frac{N-1}{i^*(N)-2},$$

which implies (54).

Let us find now $v = v(i)$ for $i \in E = \{1, 2, \ldots, N+1\}$.

If $i \in \mathbb{D}^* = \{i^*, i^*+1, \ldots, N, N+1\}$, then $v(i) = g(i) = \frac{i}{N}$. Let $i = i^* - 1$. Then

$$v(i^*-1) = Tv(i^*-1) = \sum_{j=i^*}^{N} \frac{i^*-1}{j(j-1)}\, g(j) = \frac{i^*-1}{N}\Big(\frac{1}{i^*-1} + \cdots + \frac{1}{N-1}\Big).$$

Now let $i = i^* - 2$. Then

$$\begin{aligned} v(i^*-2) = Tv(i^*-2) &= \frac{i^*-2}{(i^*-1)(i^*-2)}\, v(i^*-1) + \sum_{j=i^*}^{N} \frac{i^*-2}{j(j-1)}\, g(j) \\ &= \frac{1}{N}\Big(\frac{1}{i^*-1} + \cdots + \frac{1}{N-1}\Big) + \frac{i^*-2}{N} \sum_{j=i^*}^{N} \frac{1}{j-1} \\ &= \frac{i^*-1}{N}\Big(\frac{1}{i^*-1} + \cdots + \frac{1}{N-1}\Big). \end{aligned}$$

By induction we establish that

$$v(i) = v^*(N) = \frac{i^*-1}{N}\Big(\frac{1}{i^*-1} + \cdots + \frac{1}{N-1}\Big) \tag{55}$$

for $1 \leq i < i^*$. Therefore

$$v(i) = \begin{cases} v^*(N), & 1 \leq i < i^*(N), \\ g(i) = \frac{i}{N}, & i^*(N) \leq i \leq N, \end{cases} \tag{56}$$

for $i \in \{1, 2, \ldots, N\}$.

By (53) we have

$$\lim_{N\to\infty} \Big(\frac{1}{i^*(N)-1} + \cdots + \frac{1}{N-1}\Big) = 1, \tag{57}$$

and hence (55) implies that

$$\lim_{N\to\infty} v^*(N) = \lim_{N\to\infty} \frac{i^*(N)-1}{N} = \frac{1}{e} \approx 0.368. \tag{58}$$

This result may appear somewhat surprising since it implies that for a large number N of candidates the fiancée can choose the best one with a fairly high probability, $V^* = \sup_\tau \mathsf{P}\{a_\tau = N\} = v^*(N) \approx 0.368$. The optimal stopping time for that is

$$\tau^* = \min\{n\colon X_n \in \mathbb{D}^*\},$$

where $\mathbb{D}^* = \{i^*, i^*+1, \dots, N, N+1\}$.

Thus the optimal strategy of the fiancée is to see i^*-1 candidates without making any choice (where $i^* = i^*(N) \sim N/e$, $n \to \infty$) and then to choose the first one who is better than those she has already seen.

When $N = 10$, a more detailed analysis shows (e.g., [22, Section 1, Chap. III]) that $i^*(10) = 4$. In other words, in this case the fiancée should see the first three candidates and then choose the first one who dominates over all the preceding ones. The probability of choosing the best fiancé in this case (i.e., $v^*(10)$) is approximately 0.399.

8. Problems

1. Give an example showing that the optimal stopping time (in the class $\mathfrak{M}_0^\infty$) may not exist for Markov chains with a *countable* state space.
2. Check that the time τ_y introduced in the proof of Theorem 2 is a Markov time.
3. Show that the sequence $X = (X_1, X_2, \dots)$ in the fiancée problem forms a homogeneous Markov chain.
4. Let $X = (X_n)_{n\ge 0}$ be a real-valued homogeneous Markov chain with transition function $P = P(x; B)$, $x \in R$, $B \in \mathscr{B}(R)$. An $\overline{R}$-valued function $f = f(x)$, $x \in R$, is *P-harmonic* (or harmonic with respect to P) if

$$\mathsf{E}_x\,|f(X_1)| = \int_R |f(y)|\,P(x; dy) < \infty, \quad x \in R,$$

and

$$f(x) = \int_R f(y)\,P(x; dy), \quad x \in R. \tag{59}$$

(If the equality sign in (59) is replaced by the greater than or equal to symbol, then f is *superharmonic*.) Prove that if f is superharmonic, then for any $x \in R$ the sequence $(f(X_n))_{n\ge 0}$ with $X_0 = x$ is a supermartingale (with respect to P_x).
5. Show that the stopping time $\bar\tau$ involved in (38) belongs to the class $\mathfrak{M}_1^\infty$.
6. Similarly to Example 1 in Subsection 6, consider the optimal stopping problems

$$s_N(x) = \sup_{\tau\in\mathfrak{M}_0^N} \mathsf{E}_x\, g(X_\tau)$$

and

$$s(x) = \sup_{\tau \in \mathfrak{M}_0^\infty} \mathsf{E}_x\, g(X_\tau)$$

for simple random walks treated in the examples in Sect. 8.

Development of Mathematical Theory of Probability: Historical Review

In the history of probability theory, we can distinguish the following periods of its development (cf. [34, 46]*):

1. Prehistory,
2. First period (the seventeenth century–early eighteenth century),
3. Second period (the eighteenth century–early nineteenth century),
4. Third period (second half of the nineteenth century),
5. Fourth period (the twentieth century).

Prehistory. Intuitive notions of *randomness* and the beginning of reflection about *chances* (in ritual practice, deciding controversies, fortune telling, and so on) appeared far back in the past. In the prescientific ages, these phenomena were regarded as incomprehensible for human intelligence and rational investigation. It was only several centuries ago that their understanding and logically formalized study began.

Archeological findings tell us about ancient "randomization instruments", which were used for gambling games. They were made from the ankle bone (latin: *astragalus*) of hoofed animals and had four faces on which they could fall. Such dice were definitely used for gambling during the First Dynasty in Egypt (around 3500 BC), and then in ancient Greece and ancient Rome. It is known ([14]) that the Roman Emperors August (63 BC–14 AC) and Claudius (10 BC–54 AC) were passionate dicers.

In addition to gambling, which even at that time raised issues of favorable and unfavorable outcomes, similar questions appeared in insurance and commerce. The oldest forms of insurance were contracts for maritime transportation, which were found in Babylonian records of the 4th to 3rd Millennium BC. Afterwards the practice of similar contracts was taken over by the Phoenicians and then came to Greece, Rome, and India. Its traces can be found in early Roman legal codes and in legislation of the Byzantine Empire. In connection with life insurance, the Roman jurist Ulpian compiled (220 BC) the first mortality tables.

* The citations here refer to the list of References following this section.

A. N. Shiryaev, *Probability-2*, Graduate Texts in Mathematics 95,
https://doi.org/10.1007/978-0-387-72208-5

In the time of flourishing Italian city-states (Rome, Venice, Genoa, Pisa, Florence), the practice of insurance caused the necessity of statistics and actuarial calculations. It is known that the first dated life insurance contract was concluded in Genoa in 1347.

The city-states gave rise to the Renaissance (fourteenth to early seventeenth centuries), the period of social and cultural upheaval in Western Europe. In the Italian Renaissance, there appeared the first discussions, mostly of a philosophical nature, regarding the "probabilistic" arguments, attributed to Luca Pacioli (1447–1517), Celio Calcagnini (1479–1541), and Niccòlo Fontana Tartaglia (1500–1557) (see [46, 14]).

Apparently, one of the first people to *mathematically* analyse gambling chances was Gerolamo Cardano (1501–1576), who was widely known for inventing the Cardan gear and solving the cubic equation (although this was apparently solved by Tartaglia, whose solution Cardano published). His manuscript (written around 1525 but not published until 1663) "Liber de ludo aleae" ("Book on Games of Chance") was more than a kind of practical manual for gamblers. Cardano was first to state the idea of *combinations* by which one could describe the set of all possible outcomes (in throwing dice of various kinds and numbers). He observed also that for true dice "the ratio of the number of favorable outcomes to the total number of possible outcomes is in good agreement with gambling practice" ([14]).

1. First period (the seventeenth century–early eighteenth century). Many mathematicians and historians, such as Laplace [44] (see also [64]), related the beginning of the "calculus of probabilities" with correspondence between Blaise Pascal (1623–1662) and Pierre de Fermat (1601–1665). This correspondence arose from certain questions that Antoine Gombaud (alias Chevalier de Méré, a writer and moralist, 1607–1684) asked Fermat.

One of the questions was how to divide the stake in an interrupted game. Namely, suppose two gamblers, A and B, agreed to play a certain number of games, say a best-of-five series, but were interrupted by an external cause when A has won 4 games and B has won 3 games. A seemingly natural answer is to divide it in the proportion $2:1$. Indeed, the game certainly finishes in two steps, of which A may win 1 time, while B has to win both. This apparently implies the proportion $2:1$.

However, A has won 4 games against 3 won by B, so that the proportion $4:3$ also looks natural. In fact, the correct answer found by Pascal and Fermat was $3:1$.

Another question was: what is more likely, to have at least one 6 in 4 throwings of a dice or to have at least one pair $(6,6)$ in 24 simultaneous throwings of two dice? In this problem, Pascal and Fermat also gave a correct answer: the former combination is slightly more probable than the latter ($1-(5/6)^2 = 0.516$ against $1-(35/36)^2 = 0.491$).

In solving these problems, Pascal and Fermat (as well as Cardano) applied combinatorial arguments that became one of the basic tools in "calculus of probabilities" for the calculation of various chances. Among these tools, Pascal's triangle also found its place (although it was known before).

In 1657, the book by Christianus Huygens (1629–1695) "De Ratiociniis in Ludo Aleæ" ("On Reasoning in Games of Chance") appeared, which is regarded as the

first systematic presentation of the "calculus of probabilities". In this book, Huygens formulates many basic notions and principles, states the rules of addition and multiplication of probabilities, and discusses the concept of expectation. This book became for long the main textbook in elementary probability theory.

A prominent figure of this formative stage of probability theory was Jacob (James, Jacques) Bernoulli (1654–1705), who is credited with introducing the classical concept of the "probability of an event" as the ratio of the number of outcomes favorable for this event to the total number of possible outcomes.

The main result of J. Bernoulli, with which his name is associated, is, of course, the law of large numbers, which is fundamental for all applications of probability theory.

This law, stated as a limit theorem, is dated from 1713 when Bernoulli's treatise "Ars Conjectandi" ("The Art of Guessing") was published (posthumously) with involvement of his nephew Nikolaus Bernoulli (see [3]). As was indicated by A. A. Markov in his speech on the occasion of the 200th anniversary of the law of large numbers (see [56]), J. Bernoulli wrote in his letters (of October 3, 1703, and April 20, 1704) that this theorem was known to him "already twelve years ago". (The very term "law of large numbers" was proposed by Poisson in 1835.)

Another member of the Bernoulli family, Daniel Bernoulli (1667–1748), is known for probability theory in connection with discussion regarding the so-called "St. Petersburg paradox", where he proposed to use the notion of "moral expectation".

The first period of development of probability theory coincided in time with the formation of mathematical natural science. This was the time when the concepts of continuity, infinity, and infinitesimally small quantities prevailed. This was when Issaac Newton (1642–1727) and Gottfried Wilhelm Leibniz (1646–1716) developed differential and integral calculus. As A. N. Kolmogorov [34] wrote, the problem of that epoch was to "comprehend the extraordinary breadth and flexibility (and omnipotence, as appeared then) of the mathematical method of study of causality. The idea of a differential equation as a law which determines uniquely the evolution of a system from its present state on took then even more prominent position in the mathematical natural science then now. The probability theory is needed in the mathematical natural science when this deterministic approach based on differential equations fails. But at that time there was no concrete numerical material for application of probability theory."

Nevertheless, it became clear that the description of real data by deterministic models like differential equations was inevitably only a rough approximation. It was also understood that, in the chaos of large masses of unrelated events, there may appear in average certain regularities. This envisaged the fundamental natural-philosophic role of the probability theory, which was revealed by J. Bernoulli's law of large numbers.

It should be noted that J. Bernoulli realized the importance of dealing with *infinite* sequences of repeated trials, which was a radically new idea in probabilistic considerations restricted at that time to elementary arithmetic and combinatorial tools. The statement of the question that led to the law of large numbers revealed

the difference between the notions of the *probability* of an event and the frequency of its appearance in a finite number of trials, as well as the possibility of determination of this probability (with certain accuracy) from its frequency in large numbers of trials.

2. Second period (the eighteenth century–early nineteenth century). This period is associated, essentially, with the names of Pierre-Rémond de Montmort (1678–1719), Abraham de Moivre (1667–1754), Thomas Bayes (1702–1761), Pierre Simon de Laplace (1749–1827), Carl Friedrich Gauss (1777–1855), and Siméon Denis Poisson (1781–1840).

While the first period was largely of a *philosophical* nature, in the second one the *analytic* methods were developed and perfected, computations became necessary in various applications, and probabilistic and statistical approaches were introduced in the theory of observation errors and shooting theory.

Both Montmort and de Moivre were greatly influenced by Bernoulli's work in the calculus of probability. In his book "Essai d'Analyse sur les Jeux de Hasard" ("Essay on the analysis of gambling"), Montmort pays major attention to the methods of computations in diverse gambling games.

In the books "Doctrine of Chances" (1718) and "Miscellanea Analytica" ("Analytical Miscellany", 1730), de Moivre carefully defines such concepts as *independence*, *expectation*, and *conditional probability*.

De Moivre's name is best known in connection with the normal approximation for the binomial distribution. While J. Bernoulli's law of large numbers showed that the relative frequencies obey a certain regularity, namely, they converge to the corresponding probabilities, the normal approximation discovered by de Moivre revealed another universal regularity in the behavior of *deviations* from the mean value. This de Moivre's result and its subsequent generalizations played such a significant role in the probability theory that the corresponding "integral limit theorem" became known as the Central Limit Theorem. (This term was introduced by G. Pólya (1887–1985) in 1920, see [60].)

The main figure of this period was, of course, Laplace (Pierre-Simon de Laplace, 1749–1827). His treatise "Théorie analytique des probabilités" ("Analytic Theory of Probability") published in 1812 was the main manual on the probability theory in the nineteenth century. He also wrote several memoirs on foundations, philosophical issues, and particular problems of the probability theory, in addition to his works on astronomy and calculus. He made a significant contribution to the theory of errors. The idea that the measurement errors are normally distributed as a result of the summation of many independent elementary random errors is due to Laplace and Gauss. Laplace not only restated Moivre's integral limit theorem in a more general form (the "de Moivre–Laplace theorem"), but also gave it a new analytic proof.

Following Bernoulli, Laplace maintained the equiprobability principle implying the classical definition of probability (in the case of finitely many possible outcomes).

However, already at that time there appeared "nonclassical" probability distributions that did not conform to the classical concepts. So were, for example, the

normal and Poisson laws, which for long time were considered merely as certain approximations rather than probability distributions per se (in the modern sense of the term).

Other problems, where "nonclassical" probabilities arose, were the ones related to "geometric probabilities" (treated, e.g., by Newton 1665, see [55, p. 60]). An example of such a problem is the "Buffon needle". Moreover, unequal probabilities arose from the Bayes formula (presented in "An Essay towards Solving a Problem in the Doctrine of Chances" which was read to the Royal Society in 1763 after Bayes' death). This formula gives the rule for recalculation of prior probabilities (assumed equal by Bayes) into posterior ones given the occurrence of a certain event. This formula gave rise to the statistical approach called nowadays "the Bayes approach".

It can be seen from all that has been said that the framework of the "classical" (finite) probability theory limited the possibilities of its development and application, and the interpretation of the normal, Poisson, and other distributions merely as limiting objects was giving rise to the feeling of incompleteness. During this period, there were no abstract mathematical concepts in the probability theory and it was regarded as nothing but a branch of applied mathematics. Moreover, its methods were confined to the needs of specific applications (such as gambling, theory of observation errors, theory of shooting, insurance, demography, and so on).

3. Third period (second half of the nineteenth century). During the third period, the general problems of probability theory developed primarily in St. Petersburg. The Russian mathematicians P. L. Chebyshev (1821–1894), A. A. Markov (1856–1922), and A. M. Lyapunov (1857–1918) made an essential contribution to the broadening and in-depth study of probability theory. It was due to them that the limitation to "classical" probability was abandoned. Chebyshev clearly realized the role of the notions of a random variable and expectation and demonstrated their usability, which have now become a matter of course.

Bernoulli's law of large numbers and the de Moivre–Laplace theorem dealt with random variables taking only two values. Chebyshev extended the scope of these theorems to much more general random variables. Already his first result established the law of large numbers for sums of arbitrary independent random variables bounded by a constant. (The next step was done by Markov who used in the proof the "Chebyshev–Markov inequality".)

After the law of large numbers, Chebyshev turned to establishing the de Moivre–Laplace theorem for sums of independent random variables, for which he worked out a new tool, the method of moments, which was later elaborated by Markov.

The next unexpected step in finding general conditions for the validity of the de Moivre–Laplace theorem was done by Lyapunov, who used the method of characteristic functions taking its origin from Laplace. He proved this theorem assuming only that the random variables involved in the sum have moments of order $2 + \delta$, $\delta > 0$ (rather then the moments of all orders required by the method of moments) and satisfy *Lyapunov's condition.*

Moreover, Markov introduced a principally new concept, namely, that of a sequence of *dependent* random variables possessing the memoryless property known nowadays as a Markov chain, for which he rigorously proved the first "ergodic theorem".

Thus, we can definitely state that the works of Chebyshev, Markov, and Lyapunov ("Petersbourg school") laid the foundation for all subsequent development of the probability theory.

In Western Europe, the interest to probability theory in the late nineteenth century was rapidly increasing due to its deep connections discovered at that time with pure mathematics, statistical physics, and flourished mathematical statistics.

It became clear that the development of probability theory was restrained by its classical framework (finitely many equiprobable outcomes) and its extension had to be sought in the models of pure mathematics. (Recall that at that time the set theory only began to be developed and measure theory was on the threshold of its creation.)

At the same time, pure mathematics, particularly number theory, which is an area apparently very remote from probability theory, began to use concepts and obtain results of a probabilistic nature with the help of probabilistic intuition.

For example, Jules Henri Poincaré (1854–1912), in his paper [58] of 1890 dealing with the three-body problem, stated a result on return of the motion of a dynamical system described by a transformation T preserving the "volume". This result asserted that if A is the set of initial states ω, then for "typical" states $\omega \in A$ the trajectories $T^n\omega$ would return into the set A infinitely often (in the modern language, the system returns for *almost all* [rather then for all] initial states of the system).

In considerations of that time, expressions like "random choice", "typical case", "special case" are often used. In the handbook *Calcul des Probabilités* [59], 1896, H. Poincaré asks the question about the probability that a randomly chosen point of $[0, 1]$ happens to be a rational number.

In 1888, the astronomer Johan August Hugo Guldén (1841–1896) published a paper [24] that dealt (like Poincaré's [58]) with planetary stability and which nowadays would fall within the domain of number theory.

Let $\omega \in [0, 1]$ be a number chosen "at random" and let $\omega = (a_1, a_2, \dots)$ be its continued fraction representation, where $a_n = a_n(\omega)$ are integers. (For a rational number ω there are only finitely many nonvanishing a_n's; the numbers $\omega^k = (a_1, a_2, \dots, a_k, 0, 0, \dots)$ formed from the representation $\omega = (a_1, a_2, \dots)$ are in a sense "best possible" rational approximations of ω.) The question is how the numbers $a_n(\omega)$ behave for large n in "typical" cases.

Guldén established (though nonrigorously) that the "probability" to have $a_n = k$ in the representation $\omega = (a_1, a_2, \dots)$ is "more or less" inversely proportional to k^2 for large h. Somewhat later, T. Brodén [9] and A. Wiman [69] showed by dealing with geometric probabilities that if the "random" choice of $\omega \in [0, 1]$ is determined by the uniform distribution of ω on $[0, 1]$, then the probability that $a_n(\omega) = k$ tends to

$$(\log 2)^{-1} \log \left[\left(1 + \frac{1}{k}\right) \Big/ \left(1 + \frac{1}{k+1}\right) \right]$$

as $n \to \infty$. This expression is inversely proportional to k^2 for large k, which Guldén essentially meant.

In the second half of the nineteenth century, the probabilistic concepts and arguments found their way to the classical physics and statistical mechanics. Let us mention, for example, the *Maxwell distribution* (James Clerk Maxwell, 1831–1879)

for molecular velocities, see [51], and Boltzmann's *temporal averages* and *ergodic hypothesis* (Ludwig Boltzmann, 1844–1906), see [6, 7].

With their names, the concept of a *statistical ensemble* is connected, which was further elaborated by Josiah Willard Gibbs (1839–1903), see [23].

An important role in development of probability theory and understanding its concepts and approaches was played by the discovery in 1827 by Robert Brown (1773–1858) of the phenomenon now known as *Brownian motion*, which he described in the paper "A Brief Account of Microscopical Observations . . ." published in 1828 (see [11]). Another phenomenon of this kind was the *radioactive decay* discovered in 1896 by Antoine Henri Becquerel (1852–1908), who studied the properties of uranium. In 1900, Louis Bachelier (1870–1946) used the Brownian motion for mathematical description of stock value, see [2].

A qualitative explanation and quantitative description of the Brownian motion was given later by Albert Einstein (1879–1955) [15] and Marian Smoluchowski (1872–1917) [63]. The phenomenon of radioactivity was explained in the framework of quantum mechanics, which was created in 1920s.

From all that has been said, it becomes apparent that the appearance of new probabilistic models and the use of probabilistic methodology were far beyond the scope of the "classical probability" and required new concepts that would enable one to give a precise mathematical meaning to expressions such as "randomly chosen point from the interval $[0, 1]$", let alone the probabilistic description of Brownian motion. From this perspective, very well-timed were measure theory and the notion of the "Borel measure" introduced by Émile Borel (1871–1956) in 1898 [8], and the theory of integration by Henri Lebesgue (1875–1941) exposed in his book [45] of 1904. (Borel introduced the measure on the Euclidean space as a generalization of the notion of length. The modern presentation of measure theory on *abstract* measurable spaces follows Maurice Fréchet (1878–1973), see [22] of 1915. The history of measure theory and integration can be found, e.g., in [25].)

It was immediately recognized that Borel's measure theory along with Lebesgue's theory of integration form the conceptual basis that may justify many probabilistic considerations and give a precise meaning to intuitive formulations like the "random choice of a point from $[0, 1]$". Soon afterwards (1905), Borel himself produced an application of the measure-theoretic approach to the probability theory by proving the first limit theorem, viz. *strong law of large numbers*, regarding certain properties of real numbers that hold "with probability one".

This theorem, giving a certain idea of "how many" real numbers with exceptional (in the sense to be specified) properties are there, consists of the following.

Let $\omega = 0.\alpha_1, \alpha_2, \dots$ be the binary representation with $\alpha_n = 0$ or 1 of a real number $\omega \in [0, 1]$ (compare with continued fraction representation $\omega = (a_1, a_2, \dots)$ considered above). Let $\nu_n(\omega)$ be the (relative) frequency of ones among the first n digits $\alpha_1, \dots, \alpha_n$, then the set of those numbers ω for which $\nu_n(\omega) \to 1/2$ as $n \to \infty$ ("normal" numbers according to Borel) has Borel measure 1, while the ("exceptional") numbers for which this convergence fails form a set of zero measure.

This result (Borel's law of large numbers) bears a superficial resemblance to Bernoulli's law of large numbers. However, there is a great formally mathematical and conceptually philosophical difference between them. In fact, the law of large numbers says that for any $\varepsilon > 0$ the probability of the event $\{\omega\colon |\nu_n(\omega) - \frac{1}{2}| \geq \epsilon\}$ tends to 0 as $n \to \infty$. But the strong law of large numbers says more, namely, it states that the probability of the event $\left\{\omega\colon \sup_{m\geq n} |\nu_m(\omega) - \frac{1}{2}| \geq \varepsilon\right\}$ tends to 0. Further, in the former case, the assertion concerns the probabilities related to *finite* sequences $(\alpha_1, \alpha_2, \ldots, \alpha_n)$, $n \geq 1$, and the limits of these probabilities. The latter case, however, deals with *infinite* sequences $(\alpha_1, \alpha_2, \ldots \alpha_n, \ldots)$ and probabilities related to them.[†] (A detailed presentation of a wide variety of mathematical and philosophical issues connected with application of probabilistic methods in the number theory, as well as a comprehensive information about the development of the modern probability theory, can be found in Jan von Plato's "Creating Modern Probability" [57].)

4. Fourth period (the twentieth century). The interplay between the probability theory and pure mathematics, which became apparent by the end of the nineteenth century, made David Hilbert (1862–1943) pose the problem of mathematization of the probability theory in his lecture on the 2nd Mathematical Congress in Paris on August 8, 1900. Among his renowned problems (the first of which pertained to the continuum-hypothesis), the sixth was formulated as the one of axiomatization of physical disciplines where mathematics plays a dominant role. Hilbert associated with this disciplines the probability theory and mechanics, having pointed out the necessity of the rigorous development of the method of mean values in physics and, in particular, in kinetic theory of gases. (Hilbert pointed out that the axiomatization of the probability theory was initiated by Georg Bohlmann (1869–1928), a privatdozent in Göttingen, who spoke on this matter on the Actuarial Congress in Paris, 1900, see [5, 62]. The probability introduced by Bohlmann was defined as a (finite-additive) function of events, but without clear definition of the system of events, which he well recognized himself.)

The fourth period in the development of probability theory is the time of its logical justification and becoming a mathematical discipline.

Soon after Gilbert's lecture, several attempts of building a mathematical theory of probability involving elements of set and measure theory were made. In 1904, R. Lämmel [43], see also [62], used set theory to describe possible outcomes. However, the notion of probability (termed "content" and associated with volume, area, length, etc.) remained on the intuitive level of the previous period.

In the thesis [10] (see also [62]), produced in 1907 under the guidance of Hilbert, Ugo Broggi (1880–1965) exploited Borel's and Lebesgue's measure (using its presentation in Lebesgue's book [45] of 1904), but the notion of the (finite-additive)

[†] Bernoulli's LLN dealt with a problem quite different from the one solved by the SLLN, namely, obtaining an approximation for the distribution of the sum of n independent variables. Of course, it was only the first step in this direction; the next one was the de Moivre–Laplace theorem. This problem does not concern infinite sequences of random variables and is of current interest for probability theory and mathematical statistics. For the modern form of the LLN, see, e.g., the *degenerate convergence criterion* in Loéve's "Probability Theory" (*Translator*).

probability required (in the simplest cases) the concepts of "relative measures", "relative frequences" and (in the general cases) some artificial limiting procedures.

Among the authors of subsequent work on logical justification of probability theory, we mention first of all S. N. Bernstein (1880–1968) and Richard von Mises (1883–1953).

Bernstein's system of axioms ([4], 1917) was based on the notion of *qualitative* comparison of events according to their greater or smaller likelihood. But the numerical value of probability was defined as a subordinate notion.

Afterwards, a very similar approach based on subjective qualitative statements ("system of knowledge of the subject") was extensively developed by Bruno de Finetti (1906–1985) in the late 1920s–early 1930s (see, e.g., [16–21]).

The ideas of de Finetti found support from some statisticians following the Bayes approach, e.g., Leonard Savage (1917–1971), see [61], and were adopted in game and decision theory, where subjectivity plays a significant role.

In 1919, Mises proposed ([52, 53]) the *frequentist* (or in other terms, *statistical* or *empirical*) approach to foundation of probability theory. His basic idea was that the probabilistic concepts are applicable only to so called "collectives", i.e., *individual infinite ordered sequences* possessing a certain property of their "random" formation. The general Mises' scheme may be outlined as follows.

We have a space of outcomes of an "experiment" and assume that we can produce infinitely many trials resulting in a sequence $x = (x_1, x_2, \ldots)$ of outcomes. Let A be a subset in the set of outcomes and $\nu_n(A; x) = n^{-1} \sum_{i=1}^{n} I_A(x_i)$ the relative frequency of occurrence of the "event" A in the first n trials.

The sequence $x = (x_1, x_2, \ldots)$ is said to be a "collective" if it satisfies the following two postulates (which Mises calls *alternative conditions*, see ([52–54]):

(i) (The existence of the limiting frequencies for the *sequence*). For all "admissible" sets A, there exists the limit

$$\lim_n \nu_n(A; x) \quad (= P(A; x)).$$

(ii) (The existence of the limiting frequencies for *subsequences*). For all subsequences $x' = (x'_1, x'_2, \ldots)$ obtained from the sequence $x = (x_1, x_2, \ldots)$ by means of a certain preconditioned system of ("admissible") rules (termed by Mises as "place-selection functions"), the limits of frequencies $\lim_n \nu_n(A; x')$ must be the same as for the initial sequence $x = (x_1, x_2, \ldots)$, i.e. must be equal to $\lim_n \nu_n(A; x)$.

According to Mises, one can speak of the "probability of A" only in connection with a certain "collective", and this probability $P(A; x)$ is defined (by (i)) as the limit $\lim_n \nu_n(A; x)$. It should be emphasized that if this limit does not exist (so that x by definition is not a "collective"), this probability is not defined. The second postulate was intended to set forth the concept of "randomness" in the formation of the "collective" $x = (x_1, x_2, \ldots)$ (which is the cornerstone of the probabilistic reasoning and must be in accordance with intuition). It had to express the idea of "irregularity" of this sequence and "unpredictability" of its "future values" $(x_n, x_{n+1}, \ldots)$ from the "past" $(x_1, x_2, \ldots, x_{n-1})$ for any $n \geq 1$. (In probability theory based on

Kolmogorov's axioms exposed in Sect. 1, Chap. 2, Vol. 1, such sequences are "typical" sequences of *independent identically distributed* random variables, see Subsection 4 of Sect. 5, Chap. 1).

The postulates used by Mises in construction of "a mathematical theory of repetitive events" (as he wrote in [54]) caused much discussion and criticism, especially in the 1930s. The objections concerned mainly the fact that in practice we deal only with *finite* rather than infinite sequences. Therefore, in reality, it is impossible to determine whether the limit $\lim_n \nu_n(A;x)$ does exist and how sensitive this limit is to taking it along a subsequence x' instead of the sequence x. The issues, which were also severely criticised, were the manner of defining by Mises "admissible" rules of selecting subsequences as well as vagueness in defining the set of those ("test") rules that can be considered in the alternative condition (ii).

If we consider a sequence $x = (x_1, x_2, \ldots)$ of zeroes and ones such that the limit $\lim_n \nu_n(\{1\};x)$ lies in $(0,1)$, then this sequence must contain infinitely many both zeroes and ones. Therefore, if *any* rules of forming subsequences are admitted, then we always can take a subsequence of x consisting, e.g., only of ones for which $\lim_n \nu_n(\{1\};x') = 1$. Hence nontrivial collectives invariant with respect to *all* rules of taking subsequences *do not exist*.

The first step towards the proof that the class of collectives is not empty was taken in 1937 by Abraham Wald (1902–1950), see [68]. In his construction, the rules of selecting subsequences $x' = (x'_1, x'_2, \ldots)$ from $x = (x_1, x_2, \ldots)$ were determined by a countable collection of functions $f_i = f_i(x_1, \ldots, x_i)$, $i = 1, 2, \ldots$, taking two values 0 and 1 so that x_{i+1} is included into x' if $f_i(x_1, \ldots, x_i) = 1$ and not included otherwise. In 1940, Alonzo Church (1903–1995) proposed [12] another approach to forming subsequences based on the idea that every rule must be "effectively computable" in practice. This idea led Church to the concept of *algorithmically computable* functions (i.e., computable by means of, say, *Turing machine*). (Let, for example, x_j take two values, $\omega_1 = 0$ and $\omega_2 = 1$. Let us make correspond to $(x_1, \ldots, x_n)$ the integer

$$\lambda = \sum_{k=1}^{n} i_k 2^{k-1},$$

where i_k is defined by $x_k = \omega_{i_k}$. Let $\varphi = \varphi(\lambda)$ be a $\{0,1\}$-valued function defined on the set of nonnegative integers. Then x_{n+1} is included into x' if $\varphi(\lambda_n) = 1$ and not included otherwise.)

For explanation and justification of his concept of a "collective" as a sequence with the "randomness" property, Mises brought forward a heuristic argument that it is impossible for such sequences to construct a "winning system of a game".

These arguments were critically analyzed in a 1939 monograph [65] by Jean Ville (1910–1988), where he put Mises' reasoning into a rigorous mathematical form. It is interesting to note that this is the paper where the term "martingale" (in the *mathematical* sense) was first used.

The above description of various approaches to the axiomatics of probability theory (e.g., Bernstein, de Finetti, Mises) shows that they were complicated and overburdened with concepts stemming from the intention of their authors to make

probability theory closer to applications. As Kolmogorov pointed out in his *Foundations of the Theory of Probability* [33], this could not lead to a simple system of axioms.

The first publication by Kolmogorov demonstrating his interest in the logical justification of probability theory was his paper (regretfully, not widely known) *General measure theory and calculus of probability* [27], see also [37]. Both the title of the paper and its content show that Kolmogorov envisaged the way of the logical justification of probability theory in the framework of measure and set theory. As follows from the above exposition, this was not a novelty and was quite natural for the Moscow mathematical school, where set theory and the metric theory of functions were prevailing directions of research.

In the time between this paper (1929) and appearance of *Foundations* ([31], 1933) Kolmogorov published one of his most renowned probabilistic papers, "On Analytic Methods in Probability Theory" [28]. P. S. Aleksandrov and A. Ya. Khinchin wrote [1] about this paper: "In the entire probability theory of twentieth century it is difficult to find a research so essential to development of science."

The fundamentality of this paper consisted not only in that it laid the basis for the theory of *Markov random processes*, but also in that it demonstrated close relations of this theory, and probability theory in whole, with calculus (in particular, with the theory of ordinary and partial differential equations) as well as with classical mechanics and physics.

In connection with the problem of justification of mathematical probability theory, note that Kolmogorov's paper "On Analytic Methods" provided in a sense a physical motivation to the necessity of the logical construction of the fundamentals of random processes, which, apart from axiomatics, was one of the aims of his *Foundations*.

Kolmogorov's axiomatization of probability theory is based on the concept of the probability space

$$(\Omega, \mathcal{F}, \mathsf{P}),$$

where $(\Omega, \mathcal{F})$ is an (abstract) measurable space (of "elementary events" or outcomes) and P is a nonnegative countably additive set function on $\mathcal{F}$ normalized by $\mathsf{P}(\Omega) = 1$ (to be a "probability", see Sect. 1, Chap. 2, Vol. 1). The random variables are defined as $\mathcal{F}$-measurable functions $\xi = \xi(\omega)$; the expectation of ξ is the Lebesgue integral of $\xi(\omega)$ with respect to P.

A novel concept was that of the conditional expectation $\mathsf{E}(\xi \mid \mathcal{G})$ with respect to a σ-algebra $\mathcal{G} \in \mathcal{F}$ (see in this connection Kolmogorov's preface to the 2nd edition [32] of *Foundations*).

There is a theorem (on the existence of a process with specified finite-dimensional distributions) in *Foundations* that Kolmogorov called *basic*, thereby emphasizing its particular importance. The matter is as follows.

In *Analytic Methods*, the Markov processes were designed to describe the evolution of "stochastically determined systems", and this description was given in terms of differential properties of functions $P(s, x, t, A)$ satisfying the "Kolmogorov–

Chapman equation". These functions were called "transition probabilities" due to their interpretation as the probabilities that the system being in the state x at time s will occur in the set A of states at time t.

In a similar way, in the papers [29, 30, 17, 18] of that time, which dealt with "homogeneous stochastic processes with independent increments", these processes were treated in terms of functions $P_t(x)$ satisfying the equation

$$P_{s+t}(x) = \int P_s(x-y)\, dP_t(y),$$

which is naturally derived from the interpretation of $P_t(x)$ as the probability that the increment of the process for the time t is no greater than x.

However, from the formally-logical point of view the *existence* of an object, which could be called a "process" with transition probabilities $P(s,x,t,A)$ or with increments distributed according to $P_t(x)$, remained an open question.

This was the question solved by the *basic* theorem stating that for any system of *consistent* finite-dimensional probability distributions

$$F_{t_1,t_2,\ldots,t_n}(x_1,x_2,\ldots,x_n), \quad 0 \le t_1 < t_2 < \cdots < t_n, \quad x_i \in R,$$

one can construct a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ and a system of random variables $X = (X_t)_{t\ge 0}$, $X_t = X_t(\omega)$, such that

$$\mathsf{P}\{X_{t_1} \le x_1, X_{t_2} \le X_2, \ldots, X_{t_n} \le x_n\} = F_{t_1,t_2,\ldots,t_n}(x_1,x_2,\ldots,x_n).$$

Here, Ω is taken to be the space $R^{[0,\infty)}$ of real-valued functions $\omega = \{\omega_t\}_{t\ge 0}$, $\mathcal{F}$ is the σ-algebra generated by cylinder sets, and the measure P is defined as the extension of the measure from the algebra of cylinder sets (on which this measure is naturally defined by the finite-dimensional distributions) to the smallest σ-algebra generated by this algebra. The random variables $X_i(\omega)$ are defined coordinate-wise: if $\omega = \{\omega_t\}_{t\ge 0}$, then $X_t(\omega) = \omega_t$. (This construction explains why the notion of a "random process" is often identified with the corresponding *measure* on $R^{[0,\infty)}$).

There is a little section in the *Foundations* dealing with *applicability* of probability theory.

Describing the *scheme of conditions* according to which this theory is applied to the "real world of experiment", Kolmogorov largely follows Mises, demonstrating thereby that Mises' frequentist approach to interpretation and applicability of probability theory was not alien to him.

This scheme of conditions is essentially as follows.

It is assumed that there is a *complex* of conditions which presumes the possibility to run infinitely many repeated experiments. Let $(x_1,\ldots,x_n)$ be the outcomes of n experiments taking their values in a set X and let A be a subset of X in which we are interested. If $x_i \in A$, we say that the event A occurred in the ith experiment. (Note that no assumptions of probabilistic nature are made *a priori*, e.g., that the experiments are carried out randomly and independently or anything about the "chances" of A to occur, etc.)

Further, it is assumed that to the event A a certain number (to be denoted by $\mathsf{P}(A)$) is assigned such that we may be *practically certain* that the relative frequency $\nu_n(A)$ of occurrence of A in n trials will differ very slightly from $\mathsf{P}(A)$ for large n. Moreover, if $\mathsf{P}(A)$ is very small, we may be certain that A cannot occur in a single experiment.

In his *Foundations*, Kolmogorov does not discuss in detail the conditions for applicability of probability theory to the "real world", saying that we "disregard the deep philosophical dissertations on the concept of probability in the experimental world". However, he points out in the Introduction to Chap. 1 that there are domains of applicability of probability theory "which have no relation to the concepts of random event and of probability in the precise meaning of these words".

Thirty years after, Kolmogorov turned again (see [35, 36, 38, 39, 40, 41]) to the issue of applicability of probability theory and proposed two approaches to resolve it, which are based on the concepts of "approximative randomness" and "algorithmic complexity". In this regard, he emphasized [39] that in contrast to Mises and Church who operated with infinite sequences $x_1, x_2, \ldots$ his approaches to defining randomness are of *strictly finite* nature, i.e. they are related to *finite* sequences $x_1, x_2, \ldots, x_N$ (called subsequently *chains* according to [42]), which are so in real problems.

The concept of "approximative randomness" is introduced as follows. Let $x_1, x_2, \ldots, x_N$ be a binary ($x_i = 0, 1$) sequence of length N and let $n \leq N$. This chain is said to be (n, ε)-random with respect to a finite collection Φ of *admissible algorithms* [13] if there exists a number p ($= \mathsf{P}(\{1\})$) such that for any chain $(x'_1, x'_2, \ldots, x'_m)$ with $n \leq m \leq N$ obtained from $x_1, x_2, \ldots, x_N$ by means of an algorithm $A \in \Phi$ the relative frequency $\nu_m(\{1\}, x')$ differs from p no more than by ε. (The algorithms in Φ producing chains of length $m < n$ are neglected.)

Kolmogorov shows in [35] that if for a given n and $0 < \varepsilon < 1$ the number of admissible algorithms is no greater than

$$\frac{1}{2}\exp\{2n\varepsilon^2(1-\varepsilon)\},$$

then for any $0 < p < 1$ and any $N \geq n$ there is a chain $(x_1, x_2, \ldots, x_N)$, which is (n, ε)-random (having the property of "approximative randomness").

This approach to the identification of "random" chains involves (as well as Mises' one) a certain arbitrariness connected with indeterminacy in description and selection of admissible algorithms. Obviously, this class of algorithms cannot be too large because otherwise the set of "approximatively random" chains would be empty. On the other hand, it is desirable that the admissible algorithms would be sufficiently simple (e.g., were presented in a tabular form).

In probability theory, the idea that typical random realizations have a very complicated, irregular form has been established on the basis of various probabilistic statements.

Therefore, if we want the algorithmic definition of randomness of a chain to be as close as possible to the probabilistic conception of the structure of a random real-

ization, the algorithms in Φ must reject atypical chains of simple structure, selecting as random those sufficiently complicated.

This consideration led Kolmogorov to the "second" approach to the concept of randomness. The emphasis in this approach is made on the "complexity" of the chains rather then on "simplicity" of the related algorithms. Kolmogorov introduces a certain numerical characteristic of complexity, which is designed to show the degree of "irregularity" in formation of these chains.

This characteristic is known as "algorithmic" (or "Kolmogorov's") complexity $K_A(x)$ of an individual chain x with respect to the algorithm A, which can be heuristically described as the shortest length of a binary chain at the input of the algorithm A for which this algorithm can recover this chain at the output.

The formal definitions are as follows.

Let Σ be a collection of all finite binary chains $x = (x_1, x_2, \ldots, x_n)$, let $|x|$ $(= n)$ denote the length of a chain, and let Φ be a certain class of algorithms. The complexity of a chain $x \in \Sigma$ with respect to an algorithm $A \in \Phi$ is the number

$$K_A(x) = \min\{|p| : A(p) = x\},$$

i.e., the minimal length $|p|$ of a binary chain p at the input of the algorithm A from which x can be recovered at the output of A $(A(p) = x)$.

In [36], Kolmogorov establishes that (for some important classes of algorithms Φ) the following statement holds: there exists a *universal* algorithm $U \in \Phi$ such that for any $A \in \Phi$ there is a constant $C(A)$ satisfying

$$K_U(x) \leq K_A(x) + C(A)$$

for any chain $x \in \Sigma$ and for two universal algorithms U' and U''

$$|K_{U'}(x) - K_{U''}(x)| \leq C, \quad x \in \Sigma,$$

where C does not depend on $x \in \Sigma$. (Kolmogorov points out in [36] that a similar result was simultaneously obtained by R. Solomonov.)

Taking into account the fact that $K_U(x)$ grows to infinity with $|x|$ for "typical" chains x, this result justifies the following definition: the *complexity* of a chain $x \in \Sigma$ with respect to a class Φ of algorithms is $K(x) = K_U(x)$, where U is a universal algorithm in Φ.

The quantity $K(x)$ is customarily referred to as *algorithmic* or *Kolmogorov's* complexity of an "object" x. Kolmogorov regarded this quantity as measuring the amount of algorithmic information contained in the "finite object" x. He believed that this concept is even more fundamental than the *probabilistic* notion of information, which requires knowledge of a probability distribution on objects x for its definition.

The quantity $K(x)$ may be considered also as a measure of compression of a "text" x. If the class Φ includes algorithms like simple enumeration of elements, then (up to a constant factor) the complexity $K(x)$ is no greater than the length $|x|$. On the other hand, it is easy to show that the number of (binary) chains x of

complexity less than K is no greater than $2^K - 1$, which is the number of possible binary chains of length less than x ($1 + 2 + \cdots + 2^{K-1} = 2^K - 1$) at the input.

Further, it can be shown by simple arguments (see, e.g., [66]) that *there exist* chains x whose complexity is equal (up to a constant factor) to the length $|x|$ and that there are not many chains that admit high compression (the fraction of chains of complexity $n - a$ does not exceed 2^{-a}). These arguments naturally lead to the following definition: "algorithmically random chains" (with respect to a class of algorithms Φ) are those chains x whose algorithmic complexity $K(x)$ is close to $|x|$.

In other words, the algorithmic approach regards as "random" the chains x of maximal complexity ($K(x) \sim x$).

Kolmogorov's concepts of complexity and algorithmic randomness gave rise to a new direction called "Kolmogorov's complexity," which is applicable in diverse fields of mathematics and its applications (see, e.g. [42, 47, 48, 49, 50, 26] for details).

With regard to probability theory, these new concepts initiated a field of research aiming to determine for what algorithmically random chains and sequences probabilistic laws (such as the law of large numbers or the law of the iterated logarithm, see, e.g., [67]) are valid. Results of this kind provide the opportunity to apply the methods and results of probability theory in the areas which, as was pointed out with reference to [31] (or [32, 33]), "have no direct relation to the concepts of random event and probability in the precise meaning of these words".

References

[1] P. S. Aleksandrov and A. Ya. Khinchin. Andrey Nikolaevich Kolmogorov (for his 50-ies anniversary) (in Russian), *Uspekhi Matem. Nauk*, **8**, 3 (1953), 177–200.

[2] L. Bachelier. Théorie de la speculation. *Annales de l'Ecole Normale Supérieure*, **17** (1900), 21–86.

[3] Translations from James Bernoulli, transl. by Bing Sung, Dept. Statist., Harvard Univ., Preprint No. 2 (1966); Chs. 1–4 also available on: http://cerebro.xu.edu/math/Sources/JakobBernoulli/ars_sung.pdf.

[4] S. N. Bernstein. Axiomatic Justification of Probability Theory. [Opyt axiomaticheskogo obosnovaniya teorii veroyatnostey]. (In Russian.) *Soobshcheniya Khar'kovskogo Matematichskogo Obshchestva*, Ser. 2, **15** (1917), 209–274.

[5] G. Bohlmann. Lebensversicherungsmathematik. *Encyklopaedie der mathematischen Wissenschaften*. Bd. 1, Heft 2. Artikel ID4b. Teubner, Leipzig, 1903.

[6] L. Boltzmann. *Wissenschaftliche Abhandlungen*. Bd. 1–3. Barth, Leipzig, 1909.

[7] L. Boltzmann, J. Nabl. Kinetische Theorie der Materie. *Encyklopaedie der mathematischen Wissenschaften*. Bd. V, Heft 4. Teubner, Leipzig, 1907. 493–557.

[8] É. Borel. *Leçons sur la théorie des fonctions*. Gauthier-Villars, Paris, 1898; Éd. 2. Gauthier-Villars, Paris, 1914.
[9] T. Brodén. Wahrscheinlichkeitsbestimmungen bei der gewöhnlichen Kettenbruchentwicklung reeller Zahlen. *Akad. Förh. Stockholm* **57** (1900), 239–266.
[10] U. Broggi. Die Axiome der Wahrscheinlichkeitsrechnung. Dissertation. Göttingen, 1907.
[11] R. Brown. 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.* **4** (1828), 161–173.
[12] A. Church. On the concept of a random sequence. *Bull. Amer. Math. Soc*. **46**, 2 (1940), 130–135.
[13] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to Algorithms*. 3rd ed. MIT Press, 2009.
[14] F. N. David. *Games, Gods and Gambling. The Origin and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era.* Griffin, London, 1962.
[15] A. Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, **17** (1905), 549–560.
[16] B. de Finetti. Sulle probabilità numerabili e geometriche. *Istituto Lombardo. Accademia di Scienze e Lettere*. Rendiconti (2), **61** (1928), 817–824.
[17] B. de Finetti. Sulle funzioni a incremento aleatorio. *Accademia Nazionale dei Lincei*. Rendiconti (6), **10** (1929), 163–168.
[18] B. de Finetti. Integrazione delle funzioni a incremento aleatorio. *Accademia Nazionale dei Lincei*. Rendiconti (6), **10** (1929), 548–553.
[19] B. de Finetti. Probabilismo: saggio critico sulla teoria delle probabilità e sul valore della scienza. Perrella, Napoli, 1931; Logos. **14** (1931), 163–219. English transl.: *Erkenntnis. The International Journal of Analytic Philosophy* **31** (1989), 169–223.
[20] B. de Finetti. *Probability, Induction and Statistics. The Art of Guessing*. Wiley, New York etc., 1972.
[21] B. de Finetti. *Teoria delle probabilità: sintesi introduttiva con appendice critica*. Vol. 1, 2. Einaudi, Torino, 1970. English transl.: *Theory of Probability: A Critical Introductory Treatment*. Vol. 1, 2. Wiley, New York etc., 1974, 1975.
[22] M. Fréchet. Sur l'intégrale d'une fonctionnelle étendue à un ensemble abstrait. *Bulletin de la Société Mathématique de France*. **43** (1915), 248–265.
[23] J. W. Gibbs. *Elementary Principles in Statistical Mechanics. Developed with especial reference to the rational foundation of thermodynamics*. Yale Univ. Press, New Haven, 1902; Dover, New York, 1960.
[24] H. Gyldén. Quelques remarques relativement à la représentation de nombres irrationnels au moyen des fractions continues. *Comptes Rendus, Paris*, **107** (1888), 1584–1587.
[25] T. Hawkins. *Lebesgue's Theory of Integration. Its Origin and Development*. Univ. Wisconsin Press, Madison, Wis. – London, 1970.

[26] W. Kirchherr, M. Li, P. Vitányi. The Miraculous Universal Distribution. *Mathematical Intelligencer*, **19**, 4 (1997), 7–15.
[27] A. N. Kolmogorov. General Measure Theory and Calculus of Probabilities (in Russian), in: *Communist Academy. Section of natural and exact sciences. Mathematical papers.* Moscow, 1929, Vol. 1, 8–21.
[28] A. Kolmogoroff. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Mathematische Annalen*, **104** (1931), 415–458.
[29] A. Kolmogoroff. Sulla forma generale di un processo stochastico omogeneo. (Un problema di Bruno de Finetti.) *Atti della Accademia Nazionale dei Lincei*, **15** (1932), 805–808.
[30] A. Kolmogoroff. Ancora sulla forma generale di un processo omogeneo. *Atti della Accademia Nazionale dei Lincei*, **15** (1932), 866–869.
[31] A. Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933; Springer, Berlin–New York, 1973.
[32] A. N. Kolmogorov. Foundations of the Theory of Probability (in Russian). Moscow–Leningrad, ONTI, 1936; Moscow, Nauka, 1974 (2nd ed.); Moscow, PHASIS Publishing House, 1998 (3rd ed.).
[33] A. N. Kolmogorov. Foundations of the Theory of Probability. Chelsea, New York, 1950; 2nd ed. Chelsea, New York, 1956.
[34] A. N. Kolmogorov. The Contribution of Russian Science to the Development of Probability Theory. *Uchen. Zap. Moskov. Univ*. 1947, no. 91, 56ff. (in Russian).
[35] A. N. Kolmogorov. On Tables of Random Numbers, *Sankhyā* A, **25**, 4 (1963), 369–376.
[36] A. N. Kolmogorov. Three Approaches to the Definition of the Notion of "Amount of Information" (in Russian). *Problems of Infromation Transmission*, **1**, 1 (1965), 3–11.
[37] A. N. Kolmogorov. *Probability Theory and Mathematical Statistics* (in Russian). Nauka, Moscow, 1986.
[38] A. N. Kolmogorov. Logical Basis for Information Theory and Probability Theory. *IEEE Transactions on Information Theory*, **14**, 5 (1968), 662–664.
[39] A. N. Kolmogorov. On Logical Foundations of Probability Theory. *Probability Theory and Mathematical Statistics* (Tbilisi, 1982). Springer-Verlag, Berlin etc., 1983, 1–5 (Lecture Notes in Mathematics, Vol. 1021).
[40] A. N. Kolmogorov. Combinatorial Foundations of Information Theory and the Calculus of Probabilities (in Russian). Uspekhi Matem. Nauk (1983), 27–36; Russian Math. Surveys, **38**, 4 (1983), 29–40.
[41] A. N. Kolmogorov. *Information Theory and Theory of Algorithms* (in Russian). Nauka, Moscow, 1987.
[42] A. N. Kolmogorov and V. A. Uspensky. Algorithms and Randomness. *Theory Probab. Appl.*, **32**, 3 (1988), 389–412.
[43] R. Lämmel. Untersuchungen über die Ermittlung der Wahrscheinlichkeiten. Dissertation. Zürich, 1904. (See also [62].)
[44] P. S. Laplace, de. *Essai philosophique sur les probabilités*. Paris, 1814. English transl.: *A Philosophical Essay on Probabilities*. Dover, New York, 1951.

[45] H. Lebesgue. *Leçons sur l'intégration et la recherche des fonctions primitives*. Gauthier-Villars, Paris, 1904.
[46] D. E. Maistrov. *Probability Theory*: *A Historical Sketch*. Academic Press, New York, 1974.
[47] P. Martin-Löf, On the concept of a random sequence. *Theory Probab. Appl.*, **11**, 1 (1966), 413–425.
[48] P. Martin-Löf. The Definition of Random Sequences. *Information and Control*, **9**, 6 (1966), 602–619.
[49] P. Martin-Löf. On the Notion of Randomness. *Intuitionism and Proof Theory*, Proc. Conf. at Buffalo, NY, 1968, Ed. A. Kino et al.; North-Holland, Amsterdam, 1970, 73–78.
[50] P. Martin-Löf. Complexity Oscillations in Infinite Binary Sequences. *Z. Wahrsch. verw. Gebiete*, **19** (1971), 225–230.
[51] J. C. Maxwell. *The Scientific Letters and Papers of James Clerk Maxwell*. Vol. 1: 1846–1862, Vol. 2: 1862–1873, Vol. 3: 1874–1879. Ed. P. M. Harman. Cambridge, Cambridge Univ. Press, 1990, 1995, 2002.
[52] R. von Mises. Fundamentalsätze der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, **4** (1919), 1–97.
[53] R. von Mises. Grundlagen der Wahrscheinlichkeitsrechnung, *Mathematische Zeitschrift*. **5** (1919), 52–99; **7** (1920), 323.
[54] R. von Mises. *Mathematical Theory of Probability and Statistics*. Academic Press, New York–London, 1964.
[55] J. Newton. *The Mathematical Works of Isaac Newton*. D. T. Whiteside ed., vol. 1, Johnson, New York, 1967.
[56] *On Probability Theory and Mathematical Statistics* (correspondence between A. A. Markov and A. A. Chuprov). [O teorii veroyatnostey i matematicheskoy statistike (perepiska A. A. Markova i A. A. Chuprova)] (in Russian) Moscow, Nauka, 1977.
[57] J. Plato, von. *Creating Modern Probability. Its Mathematics, Physics and Philosophy in Historical Perspective*. Cambridge Univ. Press, Cambridge, 1994.
[58] H. Poincaré. Sur le problème des trois corps et les équations de la dynamique. I, II. *Acta Mathematica*, **13** (1890), 1–270.
[59] H. Poincaré. *Calcul des probabilités*. G. Carré, Paris, 1896.
[60] G. Pólya. Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem. *Mathematische Zeitschrift*, (1920), 171–181.
[61] L. J. Savage. *The Foundations of Statistics*. Wiley, New York; Chapman–Hall, London, 1954.
[62] I. Schneider, Ed. *Die Entwicklung der Wahrscheinlichkeitstheorie von den Anfängen bis* 1933. Akademie-Verlag, Berlin, 1989.
[63] M. R. Smoluchowski, von. Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen. *Annalen der Physik*, **21** (1906), 756–780.
[64] I. Todhunter. *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace*. Chelsea, New York, 1949. 1st Edition: Macmillan, Cambridge, 1865.

[65] J. A. Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.
[66] P. Vitányi and M. Li. Two Decades of Applied Kolmogorov Complexity. Uspekhi Matem. Nauk, **43**, 6 (1988), 129–166.
[67] V. G. Vovk. The Law of the Iterated Logarithm for Random Kolmogorov, or Chaotic, Sequences. *Theory Probab. Appl.*, **32**, 3 (1988), 413–425.
[68] A. Wald. Die Widerspruchsfreiheit des Kollektivbegriffes der Wahrscheinlichkeitsrechnung. *Ergebnisse eines mathematischen Kolloquiums*, **8** (1937), 38–72.
[69] A. Wiman. Über eine Wahrscheinlichkeitsaufgabe bei Kettenbruchentwicklungen, *Akad. Förh. Stockholm*, **57** (1900), 829–841.

Historical and Bibliographical Notes (Chaps. 4–8)

Chapter 4

Section 1. Kolmogorov's zero–one law appears in his book [50]. For the Hewitt–Savage zero–one law see also Borovkov [10], Breiman [11], and Ash [2].
Sections 2–4. Here the fundamental results were obtained by Kolmogorov and Khinchin (see [50] and references therein). See also Petrov [62], Stout [74], and Durret [20]. For probabilistic methods in number theory see Kubilius [51].

It is appropriate to recall here the historical background of the strong law of large numbers and the law of the iterated logarithm for the Bernoulli scheme.

The first paper where the strong law of large numbers appeared was Borel's paper [7] on normal numbers in $[0, 1)$. Using the notation of Example 3 in Sect. 3, let

$$S_n = \sum_{k=1}^{n} \Big(I(\xi_k = 1) - \frac{1}{2} \Big).$$

Then Borel's result stated that for almost all (Lebesgue) $\omega \in [0, 1)$ there exists $N = N(\omega)$ such that

$$\Big| \frac{S_n(\omega)}{n} \Big| \leq \frac{\log(n/2)}{\sqrt{2n}}$$

for all $n \geq N(\omega)$. This implies, in particular, that $S_n = o(n)$ almost surely.

The next step was done by Hausdorff [41], who established that $S_n = o(n^{1/2+\varepsilon})$ almost surely for any $\varepsilon > 0$. In 1914 Hardy and Littlewood [39] showed that $S_n = O((n \log n)^{1/2})$ almost surely. In 1922 Steinhaus [73] improved their result by showing that

$$\limsup_n \frac{S_n}{\sqrt{2n \log n}} \leq 1$$

almost surely.

A. N. Shiryaev, *Probability-2*, Graduate Texts in Mathematics 95,
https://doi.org/10.1007/978-0-387-72208-5

In 1923 Khinchin [45] showed that $S_n = O(\sqrt{n \log\log n})$ almost surely. Finally, in a year Khinchin obtained [46] the final result (the "law of the iterated logarithm"):

$$\limsup_n \frac{S_n}{\sqrt{(n/2)\log\log n}} = 1$$

almost surely. (Note that in this case $\sigma^2 = \mathsf{E}[I(\xi_k = 1) - 1/2]^2 = 1/4$, which explains the appearance of the factor $n/2$ rather than the usual $2n$; cf. Theorem 4 in Sect. 4.)

As was mentioned in Sect. 4, the next step in establishing the law of the iterated logarithm for a broad class of independent random variables was taken in 1929 by Kolmogorov [48].

Section 5. Regarding these questions, see Petrov [62], Borovkov [8–10], and Dacunha-Castelle and Duflo [16].

Chapter 5

Sections 1–3. Our exposition of the theory of (strict sense) stationary random processes is based on Breiman [11], Sinai [72], and Lamperti [52]. The simple proof of the maximal ergodic theorem was given by Garsia [28].

Chapter 6

Section 1. The books by Rozanov [67] and Gihman and Skorohod [30, 31] are devoted to the theory of (wide sense) stationary random processes. Example 6 was frequently presented in Kolmogorov's lectures.

Section 2. For orthogonal stochastic measures and stochastic integrals see also Doob [18], Gihman and Skorohod [31], Rozanov [67], and Ash and Gardner [3].

Section 3. The spectral representation (2) was obtained by Cramér and Loève (e.g., [56]). Also see Doob [18], Rozanov [67], and Ash and Gardner [3].

Section 4. There is a detailed exposition of problems of statistical estimation of the covariance function and spectral density in Hannan [37, 38].

Sections 5–6. See also Rozanov [67], Lamperti [52], and Gihman and Skorohod [30, 31].

Section 7. The presentation follows Liptser and Shiryaev [54].

Chapter 7

Section 1. Most of the fundamental results of the theory of martingales were obtained by Doob [18]. Theorem 1 is taken from Meyer [57]. Also see Meyer [58], Liptser and Shiryaev [54], Gihman and Skorohod [31], and Jacod and Shiryaev [43].

Theorem 1 is often called the theorem "on transformation under a system of optional stopping" (Doob [18]). For identities (13) and (14) and Wald's fundamental identity see Wald [76].

Section 3. The right inequality in (25) was established by Khinchin [45] (1923) in the course of proving the law of the iterated logarithm. To explain what led Khinchin to obtain this inequality, let us recall the line of the proof of the strong law of large numbers by Borel and Hausdorff (see also the earlier comment to Sects. 2–4, Chap. 4).

Let $\xi_1, \xi_2, \dots$ be a sequence of independent identically distributed random variables with $\mathsf{P}\{\xi_1 = 1\} = \mathsf{P}\{\xi_1 = -1\} = 1/2$ (Bernoulli scheme), and let $S_n = \xi_1 + \cdots + \xi_n$.

Borel's proof that $S_n = o(n)$ almost surely was essentially as follows. Since

$$\mathsf{P}\left\{\left|\frac{S_n}{n}\right| \geq \delta\right\} \leq \frac{\mathsf{E}\, S_n^4}{n^4\delta^4} \leq \frac{3n^2}{n^4\delta^4} = \frac{3}{n^2\delta^4} \quad \text{for any} \quad \delta > 0,$$

we have

$$\mathsf{P}\left\{\sup_{k\geq n}\left|\frac{S_k}{k}\right| \geq \delta\right\} \leq \sum_{k\geq n} \mathsf{P}\left\{\left|\frac{S_k}{k}\right| \geq \delta\right\} \leq \frac{3}{\delta^4}\sum_{k\geq n}\frac{1}{k^2} \to 0$$

as $n \to \infty$; therefore $S_n/n \to 0$ almost surely by the Borel–Cantelli lemma (Chap. 2, Sect. 10).

Hausdorff's proof that $S_n = o(n^{1/2+\varepsilon})$ almost surely for any $\varepsilon > 0$ proceeded in a similar way: since $\mathsf{E}\, S_n^{2r} = O(n^r)$ for any integer $r > 1/(2\varepsilon)$, we have

$$\begin{aligned}\mathsf{P}\left\{\sup_{k\geq n}\left|\frac{S_k}{k^{1/2+\varepsilon}}\right| \geq \delta\right\} &\leq \sum_{k\geq n}\mathsf{P}\left\{\left|\frac{S_k}{k^{1/2+\varepsilon}}\right| \geq \delta\right\} \\ &\leq \frac{1}{\delta^{2r}}\sum_{k\geq n}\mathsf{E}\left|\frac{S_k}{k^{1/2+\varepsilon}}\right|^{2r} \leq \frac{c}{\delta^{2r}}\sum_{k\geq n}\frac{k^r}{k^{r+2\varepsilon r}} \to 0\end{aligned}$$

as $n \to \infty$, where c is a positive constant. This implies (again by the Borel–Cantelli lemma) that

$$\frac{S_n}{n^{1/2+\varepsilon}} \to 0$$

almost surely.

The foregoing considerations show that the key element of the proofs was obtaining a "good" bound for the probabilities $\mathsf{P}\{|S_n| \geq t(n)\}$, where $t(n) = n$ in

Borel's proof and $t(n) = n^{1/2+\varepsilon}$ in Hausdorff's (while Hardy and Littlewood needed $t(n) = (n \log n)^{1/2}$).

Analogously, Khinchin needed inequalities (25) (in fact, the right one) to obtain a "good" bound for the probabilities $\mathsf{P}\{|S_n| \geq t(n)\}$.

Regarding the derivation of Khinchin's inequalities (both right and left) for any $p > 0$ and optimality of the constants A_p and B_p in (25) see the survey paper by Peškir and Shiryaev [61].

Khinchin derives from the right inequality in (25) for $p = 2m$ that for any $t > 0$

$$\mathsf{P}\{|X_n| > t\} \leq t^{-2m}\,\mathsf{E}\,|X_n|^{2m} \leq \frac{(2m)!}{2^m\,m!}\,t^{-2m}[X]_n^{2m}.$$

By Stirling's formula

$$\frac{(2m)!}{2^m m!} \leq D\Big(\frac{2}{e}\Big)^m m^m,$$

where $D = \sqrt{2}$. Therefore, setting $m = \Big[\frac{t^2}{2[X]_n^2}\Big]$, we obtain

$$\begin{aligned}\mathsf{P}\{|X_n| > t\} &\leq D\Big(\frac{2m[X]_n^2}{et^2}\Big)^m \leq D\,e^{-m} \\ &\leq D\exp\Big\{1 - \frac{t^2}{2[X]_n^2}\Big\} = D\,e\exp\Big\{-\frac{t^2}{2[X]_n^2}\Big\} \\ &= c\exp\Big\{-\frac{t^2}{2[X]_n^2}\Big\}\end{aligned}$$

with $c = De = \sqrt{2}e$.

This inequality implies the bound

$$\mathsf{P}\{|S_n| > t\} \leq e^{-\frac{t^2}{2n^2}},$$

which was used by Khinchin for the proof that $S_n = O(\sqrt{n \log\log n})$ almost surely.

Chow and Teicher [13] contains an illuminating study of the inequalities presented here. Theorem 2 is due to Lenglart [53].

Section 4. See Doob [18].

Section 5. Here we follow Kabanov, Liptser, and Shiryaev [44], Engelbert and Shiryaev [24], and Neveu [59]. Theorem 4 and the examples were given by Liptser.

Section 6. This approach to problems of absolute continuity and singularity, and the results given here, can be found in Kabanov, Liptser, and Shiryaev [44].

Section 7. Theorems 1 and 2 were given by Novikov [60]. Lemma 1 is a discrete analog of Girsanov's lemma (see [54]).

Section 8. See also Liptser and Shiryaev [55] and Jacod and Shiryaev [43], which discuss limit theorems for random processes of a rather general nature (e.g., martingales, semimartingales).

Section 9. The presentation follows Shiryaev [70, 71]. For the development of the approach given here to the generalization of Ito's formula see [27].

Section 10. Martingale methods in insurance are treated in [29]. The proofs presented here are close to those in [70].
Section 11–12. For more detailed exposition of the topics related to application of martingale methods in financial mathematics and engineering see [71].
Section 13. The basic monographs on the theory and problems of optimal stopping rules are Dynkin and Yushkevich [22], Robbins, Chow, and Siegmund [66], and Shiryaev [69].

Chapter 8

Sections 1–2. For the definitions and basic properties of Markov chains see also Dynkin and Yushkevich [22], Dynkin [21], Ventzel [75], Doob [18], Gihman and Skorohod [31], Breiman [11], Chung [14, 15], and Revuz [65].
Sections 3–7. For problems related to limiting, ergodic, and stationary distributions for Markov chains see Kolmogorov's paper [49] and the books by Feller [25, 26], Borovkov [10, 9], Ash [1], Chung [15], Revuz [65], and Dynkin and Yushkevich [22].
Section 8. The simple random walk is a textbook example of the simplest Markov chain for which many regularities were discovered (e.g., the properties of recurrence, transience, and ergodicity). These issues are treated in many of the books cited earlier; see, for example, [1, 10, 15, 65].
Section 9. The interest in optimal stopping was due to the development of statistical sequential analysis (Wald [76], De Groot [17], Zacks [77], Shiryaev [69]). The theory of optimal stopping rules is treated in Dynkin and Yushkevich [22], Shiryaev [69], and Billingsley [5]. The martingale approach to the optimal stopping problems is presented in Robbins, Chow, and Siegmund [66].

DEVELOPMENT OF MATHEMATICAL THEORY OF PROBABILITY: HISTORICAL REVIEW. This historical review was written by the author as a supplement to the third edition of Kolmogorov's *Foundations of the Theory of Probability* [50].

References

[1] R. B. Ash. *Basic Probability Theory*. Wiley, New York, 1970.
[2] R. B. Ash. *Real Analysis and Probability*. Academic Press, New York, 1972.
[3] R. B. Ash and M. F. Gardner. *Topics in Stochastic Processes*. Academic Press, New York, 1975.
[4] P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, 1968.
[5] P. Billingsley. *Probability and Measure*. 3rd ed. New York, Wiley, 1995.
[6] G. D. Birkhoff. Proof of the ergodic theorem. *Proc. Nat. Acad. Sci USA*, **17**, 650–660.
[7] É. Borel. Les probabilités dénombrables et leurs applications arithmétiques. *Rendiconti del Circolo Matematico di Palermo*. **27**, (1909), 247–271.
[8] A. A. Borovkov. *Mathematical Statistics* [*Matematicheskaya Statistika*] (in Russian). Nauka, Moscow, 1984.
[9] A. A. Borovkov. *Ergodicity and Stability of Random Processes* [Ergodichnost' i ustoichivost' Sluchaĭnykh Processov] (in Russian). Moscow, URSS, 1999.
[10] A. A. Borovkov. *Wahrscheinlichkeitstheorie*: *eine Einführung*, 1st edition Birkhäuser, Basel–Stuttgart, 1976; *Theory of Probability* (in Russian), 3rd edition [*Teoriya veroyatnosteĭ*]. Moscow, URSS, 1999.
[11] L. Breiman. *Probability*. Addison-Wesley, Reading, MA, 1968.
[12] A. V. Bulinsky and A. N. Shiryayev. *Theory of Random Processes* [Teoriya Sluchaĭnykh Processov] (in Russian). Fizmatlit, Moscow, 2005.
[13] Y. S. Chow and H. Teicher. *Probability Theory*: *Independence, Interchangeability, Martingales*. Springer-Verlag, New York, 1978.
[14] K. L. Chung. *Markov Chains with Stationary Transition Probabilities*. Springer-Verlag, New York, 1967.
[15] K. L. Chung. *Elementary Probability Theory with Stochastic Processes*. 3rd ed. Springer-Verlag, Berlin, 1979.
[16] 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).
[17] M. H. De Groot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

A. N. Shiryaev, *Probability-2*, Graduate Texts in Mathematics 95,
https://doi.org/10.1007/978-0-387-72208-5

[18] J. L. Doob. *Stochastic Processes*. Wiley, New York, 1953.
[19] J. L. Doob. What is a martingale? *American Mathematical Monthly* **78** (1971), 451–463.
[20] R. Durrett. *Probability: Theory and Examples*. Wadsworth & Brooks/Cole, Pacific Grove, CA, 1991.
[21] E. B. Dynkin. *Markov Processes*, Vol. 1, 2, Academic Press, New York; Springer-Verlag, Berlin, 1965.
[22] E. B. Dynkin and A. A. Yushkevich. *Markov Processes*: *Theorems and Problems*, Plenum, New York, 1969.
[23] P. Ehrenfest and T. Ehrenfest. Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem. *Physikalische Zeitschrift*, **8** (1907), 311–314.
[24] H. J. Engelbert and A. N. Shiryaev. On the sets of convergence of generalized submartingales. *Stochastics* **2** (1979), 155–166.
[25] W. Feller. *An Introduction to Probability Theory and Its Applications*, vol. 1, 3rd ed. Wiley, New York, 1968.
[26] W. Feller. *An Introduction to Probability Theory and Its Applications*, vol. 2, 3rd ed. Wiley, New York, 1971.
[27] H. Föllmer, Ph. Protter, and A. N. Shiryaev. Quadratic covariation and an extension of Itô's formula. *Bernoulli*, **1**, 1/2 (1995), 149–170.
[28] A. Garcia. A simple proof of E. Hopf's maximal ergodic theorem. *J. Math. Mech*. **14** (1965), 381–382.
[29] H. U. Gerber, *Life Insurance Mathematics*, Springer, Zürich, 1997.
[30] I. I. Gihman [Gikhman] and A. V. Skorohod [Skorokhod]. *Introduction to the Theory of Random Processes*, 1st ed. Saunders, Philadelphia, 1969; 2nd ed. [*Vvedenie v teoriyu sluchaĭnyh protsessov*]. Nauka, Moscow, 1977.
[31] I. I. Gihman and A. V. Skorohod. *Theory of Stochastic Processes*, 3 vols. Springer-Verlag, New York–Berlin, 1974–1979.
[32] B. V. Gnedenko and A. Ya. Khinchin. *An Elementary Introduction to the Theory of Probability*. Freeman, San Francisco, 1961; 9th ed. [*Elementarnoe vvedenie v teoriyu veroyatnosteĭ*]. Nauka, Moscow, 1982.
[33] B. V. Gnedenko and A. N. Kolmogorov. *Limit Distributions for Sums of Independent Random Variables*, revised edition. Addison-Wesley, Reading, MA, 1968.
[34] P. E. Greenwood and A. N. Shiryaev. *Contiguity and the Statistical Invariance Principle*. Gordon and Breach, New York, 1985.
[35] G. R. Grimmet and D. R. Stirzaker. *Probability and Random Processes*. 3rd ed. Oxford University Press, Oxford, 2001.
[36] J. B. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, NJ, 1994.
[37] E. J. Hannan. *Time Series Analysis*. Methuen, London, 1960.
[38] E. J. Hannan. *Multiple Time Series*. Wiley, New York, 1970.
[39] G. H. Hardy and J. E. Littlewood. Some problems of Diophantine approximation. *Acta Mathematica*. **37** (1914), 155–239.
[40] P. Hartman and A. Wintner. On the law of iterated logarithm. *Amer. J. Math.* **63**, 1 (1941), 169–176.

[41] F. Hausdorff. *Grundzüge der Mengenlehre*. Veit, Leipzig, 1914.
[42] M. Hazewinkel, editor. *Encyclopaedia of Mathematics*, Vols. 1–10 + Supplement I–III. Kluwer, 1987–2002. [Engl. transl. (extended) of: I. M. Vinogradov, editor. *Matematicheskaya Entsiklopediya*, in 5 Vols.], Moscow, Soviet Entsiklopediya, 1977–1985.
[43] J. Jacod and A. N. Shiryaev, *Limit Theorems for Stochastic Processes*, 2nd ed. Springer-Verlag, Berlin, 2003.
[44] 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.
[45] A. Khintchine. Über dyadische Brüche. *Mathematische Zeitschrift*. **18** (1923), 109–116.
[46] A. Khintchine. Über einen Satz der Wahrscheinlichkeitsrechnung. *Fundamenta Mathematicae*. **6** (1924), 9–20.
[47] A. Khintchine. Zu Birkhoffs Lösung des Ergodenproblems. *Mathematische Annalen*, **107** (1932), 485–488.
[48] A. Kolmogoroff. Über das Gesetz des iterierten Logarithmus. *Mathematische Annalen*. **101** (1929), 126–135.
[49] A. N. Kolmogorov. Markov Chains with finitely many states (in Russian). *Bull. Moscow State Univ.* [Bulleten' MGU], **1**, 3 (1937), 1–16.
[50] A. N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea, New York, 1956; 2nd ed. [*Osnovnye poniatiya Teorii Veroyatnosteĭ*]. Nauka, Moscow, 1974.
[51] J. Kubilius. *Probabilistic Methods in the Theory of Numbers*. American Mathematical Society, Providence, RI, 1964.
[52] J. Lamperti. *Stochastic Processes*. Springer-Verlag, New York, 1977.
[53] E. Lenglart. Relation de domination entre deux processus. *Ann. Inst. H. Poincaré*. Sect. B (N.S.), **13** (1977), 171–179.
[54] R. S. Liptser and A. N. Shiryaev. *Statistics of Random Processes*. Springer-Verlag, New York, 1977.
[55] R. Sh. Liptser and A. N. Shiryaev. *Theory of Martingales*. Kluwer, Dordrecht, Boston, 1989.
[56] M. Loève. *Probability Theory*. Springer-Verlag, New York, 1977–78.
[57] P.-A. Meyer, Martingales and stochastic integrals. I. *Lecture Notes in Mathematics*, no. 284. Springer-Verlag, Berlin, 1972.
[58] P.-A. Meyer. *Probability and Potentials*. Blaisdell, Waltham, MA, 1966.
[59] J. Neveu. *Discrete Parameter Martingales*. North-Holland, Amsterdam, 1975.
[60] 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.
[61] G. Peškir and A. N. Shiryaev, The Khintchine inequalities and martingale expanding sphere of their action, *Russian Math. Surveys*, **50**, 5 (1995), 849–904.
[62] V. V. Petrov. *Sums of Independent Random Variables*. Springer-Verlag, Berlin, 1975.
[63] H. Poincaré. *Calcul des probabilitiés*, 2nd ed. Gauthier Villars, Paris, 1912.

[64] I. I. Privalov. *Randeigenschaften analytischer Functionen.* Deutscher Verlag der Vissenschaft, 1956.
[65] D. Revuz. *Markov Chains*, 2nd ed. North-Holland, Amsterdam, 1984.
[66] H. Robbins, Y. S. Chow, and D. Siegmund. *Great Expectations*: *The Theory of Optimal Stopping*. Houghton Mifflin, Boston, 1971.
[67] Yu. A. Rozanov. *Stationary Random Processes*. Holden-Day, San Francisco, 1967.
[68] A. N. Shiryaev. *Random Processes* [*Sluchainye processy*] (in Russian). Moscow State University Press, 1972.
[69] A. N. Shiryayev. *Optimal Stopping Rules*. Applications of Mathematics, Vol. 8. Springer-Verlag, New York-Heidelberg, 1978.
[70] A. N. Shiryayev. *Probability*, 2nd ed. Springer-Verlag, Berlin, 1995.
[71] A. N. Shiryaev. *Essentials of Stochastic Finance: Facts, Models, Theory*. World Scientific, Singapore, 1999.
[72] Ya. G. Sinai. *Introduction to Ergodic Theory*. Princeton University Press, Princeton, NJ, 1976.
[73] H. Steinhaus. Les probabilités dénombrables et leur rapport à la théorie de la mesure. *Fundamenta Mathematicae*. **4** (1923), 286–310.
[74] W. F. Stout. *Almost Sure Convergence*. Academic Press, New York, 1974.
[75] A. D. Ventsel. *A Course in the Theory of Stochastic Processes*. McGraw-Hill, New York, 1981.
[76] A. Wald. *Sequential Analysis*. Wiley, New York, 1947.
[77] S. Zacks. *The Theory of Statistical Inference*. Wiley, New York, 1971.

Index

A. N. Shiryaev, *Probability-2*, Graduate Texts in Mathematics 95,
https://doi.org/10.1007/978-0-387-72208-5