A. V. Skorokhod

Basic Principles and Applications of Probability Theory

Springer

Basic Principles and Applications
of Probability Theory

A.V. Skorokhod

Basic Principles and Applications of Probability Theory

Edited by Yu.V. Prokhorov
Translated by B. D. Seckler

Springer

A.V. Skorokhod
Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824, USA

Yu.V. Prokhorov *(Editor)*
Russian Academy of Science
Steklov Mathematical Institute
ul. Gubkina 8
117966 Moscow, Russia

B. D. Seckler *(Translator)*
19 Ramsey Road
Great Neck, NY 11023-1611, USA
e-mail: bersec@aol.com

Original Russian edition published by Viniti, Moscow 1989

Title of the Russian edition: Teoriya Veroyatnostej 1

Published in the series: Itogi Nauki i Tekhniki. Sovremennye Problemy Matematiki. Fundamental'nye Napravleniya, Tom 43

Library of Congress Control Number: 2004110444

Mathematics Subject Classification (2000):
60Axx, 60Dxx, 60Fxx, 60Gxx, 60Jxx, 62Cxx, 94Axx

ISBN 3-540-54686-3 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springeronline.com

Printed in Germany

Typeset by Steingraeber Satztechnik GmbH, Heidelberg
using a Springer TEX macro package

Cover design: Erich Kirchner, Heidelberg
Printed on acid-free paper 46/3142LK - 5 4 3 2 1 0

Contents

Probability. Basic Notions. Structure. Methods

Contents

1

Introduction

Probability theory arose originally in connection with games of chance and then for a long time it was used primarily to investigate the credibility of testimony of witnesses in the "ethical" sciences. Nevertheless, probability has become a very powerful mathematical tool in understanding those aspects of the world that cannot be described by deterministic laws. Probability has succeeded in finding strict determinate relationships where chance seemed to reign and so terming them "laws of chance" combining such contrasting notions in the nomenclature appears to be quite justified. This introductory chapter discusses such notions as determinism, chaos and randomness, predictibility and unpredictibility, some initial approaches to formalizing randomness and it surveys certain problems that can be solved by probability theory. This will perhaps give one an idea to what extent the theory can answer questions arising in specific random occurrences and the character of the answers provided by the theory.

1.1 The Nature of Randomness

The phrase "by chance" has no single meaning in ordinary language. For instance, it may mean unpremeditated, nonobligatory, unexpected, and so on. Its opposite sense is simpler: "not by chance" signifies obliged to or bound to (happen). In philosophy, necessity counteracts randomness. Necessity signifies conforming to law – it can be expressed by an exact law. The basic laws of mechanics, physics and astronomy can be formulated in terms of precise quantitative relations which must hold with ironclad necessity. True, this state of affairs existed in the classical period when science did not delve into the microworld. But even before, chance had been encountered in everyday life at practicaily every step. Birth and death and even the entire life of a person is a chain of chance occurrences that cannot be computed or foreseen with the aid of determinate laws. What then can be studied and how studied and what sort of answers may be obtained in a world of chance? Science can merely treat the

intrinsic in occurrences and so it is important to extract the essential features of a chance occurrence that we shall take into account in what follows.

1.1.1 Determinism and Chaos

In a deterministic world, randomness must be absent – it is absolutely subject to laws that specify its state uniquely at each moment of time. This idea of the world (setting aside philosophical and theological considerations) existed among mathematicians and physicists in the 18th and 19th centuries (Newton, Laplace, etc.). However, such a world was all the same unpredictable because of its complex arrangement. In order to determine a future state, it is necessary to know its present state absolutely precisely and that is impossible. It is more promising to apply determinism to individual phenomena or aggregates of them. There is a determinate relationship between occurrences if one entails the other necessarily. The heating of water to 100°C under standard atmospheric pressure, let us say, implies that the water will boil. Thus, in a determinate situation, there is complete order in a system of phenomena or the objects to which these phenomena pertain. People have observed that kind of order in the motion of the planets (and also the Moon and Sun) and this order has made it possible to predict celestial occurrences like lunar and solar eclipses. Such order can be observed in the disposition of molecules in a crystal (it is easy to give other examples of complete order). The most precise idea of complete order is expressed by a collection of absolutely indistinguishable objects.

In contrast to a deterministic world would be a chaotic world in which no relationships are present. The ancient Greeks had some notion of such a chaotic world. According to their conception, the existing world arose out of a primary chaos. Again, if we confine ourselves just to some group of objects, then we may regard this system to be completely chaotic if the things are entirely distinct. We are excluding the possibility of comparing the objects and ascertaining relationships among them (including even causal relationships). Both of these cases are similar: the selection of one (or several objects) from the collection yields no information. In the first case, we know right away that all of the objects are identical and in the second, the heterogeneity of the objects makes it impossible to draw any conclusions about the remaining ones. Observe that this is not the only way in which these two contrasting situations resemble one another. As might be expected, according to Hegel's laws of logic, these totally contrasting situations describe the exact same situation. If the objects in a chaotic system are impossible to compare, then one cannot distinguish between them so that instead of complete disorder, we have complete order.

1.1.2 Unpredictability and Randomness

A large number of phenomena exist that are neither completely determinate nor completely chaotic. To describe them, one may use a system of noniden-

tical but mutually comparable objects and then classify them into several groups. Of interest to us might be to what group a given object belongs. We shall illustrate how the existence of differences relates to the absence of complete determinism. Suppose that we are interested in the sex of newborn children. It is known that roughly half of births are boys and half are girls. In other words, the "things" being considered split into two groups. If a strictly valid law existed for the birth of a boy or girl, then it would still be impossible to produce the mechanism which would continually equalize the sexes of babies being born in the requisite proportion (without assuming the effect of the results of prior births on succeeding births, such a premise is meaningless). One may give numerous examples of valid statements like "such a thing happens in such and such fraction of the cases", for instance, "1% of males are color-blind." As in the case of the sex of babies, the phenomenon cannot be explained on the basis of determinate laws. It is advantageous to view a set-up of things as a sequence of events proceeding in time.

The absence of determinism means that future events are unpredictable. Since events can be classified in some sort of way, one may ask to what class will a future event belong? But once again (determinism not being present), one cannot furnish an answer in advance. The question is ill posed in the given situation. The examples cited suggest a proper way to state the question: how often will a phenomenon of a given class occur in the sequence? We shall speak about chance in precisely such situations and it will be natural to raise such questions and to find answers for them.

1.1.3 Sources of Randomness.

We shall now point out a few of the most important existing physical sources of randomness in the real world. In so doing, we view the world to be sufficiently organized (unchaotic) and randomness will be understood as in Sect. 1.1.2.

(a) *Quantum-mechanical laws.* The laws of quantum mechanics are statements about the wave functions of micro-objects. According to these laws, we can specify, for instance, just the wave function of an electron in a field of force. Based on the wave function, only the probability of detecting the electron in some particular region of space may be found – to predict its position is impossible. In exactly the same way, one cannot ascertain the energy of an electron and it is only possible to determine a discrete number of possible energy levels and the probability that the energy of the electron has a specified value. We perceive that the fundamental laws of the microworld make use of the language of probability and thus phenomena in the microworld are random. An important example of a random phenomenon in the microworld is the emission of a quantum of light by an excited atom. Another important example are nuclear reactions.

(b) *Thermal motion of molecules.* The molecules of any substance are in constant thermal motion. If the substance is a solid, then the molecules range

close to positions of equilibrium in a crystal lattice. But in fluids and gases, the molecules perform rather complex movements changing their directions of motion frequently as they interact with one another. The presence of such a motion may be ascertained by watching the movement of microscopic particles suspended in a fluid or gas (this is so-called Brownian motion). This motion is of a random nature and the energies of the individual molecules are also random, that is, the energies of the molecules can assume different values and so one talks about the fraction of molecules having an energy within narrow specified bounds. This is the familiar Maxwell distribution in physics. A simple experiment will convince one that the energies of the molecules are different. Take the phenomenon of boiling water: if all of the molecules had the same energy, then the water would become steam all at once, that is, with an explosion, and this does not happen.

(c) *Discreteness of matter.* The discreteness of matter leads to the occurrence of randomness in another way. Items (a) and (b) also considered material particles. The following fact should now be noted: the laws of classical physics have been formulated for macrobodies just as if matter filled up space continuously. The discreteness of matter leads to the occurrence of deviations of the actual values of physical quantities from those predicted by the laws. These deviations or "fluctuations" are of a random nature and they affect the course of a process substantially. Thus, the discreteness of the carriers of electricity in metallic conductors – the electrons – is the source of fluctuation currents which are the reason for internal noise in radios. The discreteness of matter results in the mutual permeation of substances. Furthermore the absence of pure substances, that is, the existence of impurities, also results in random deviations from the calculated flow of phenomena.

(d) *Cosmic radiation.* Experimentation shows that it is irregular (aperiodic and unpredictable) but it conforms to laws that can be studied by probability theory.

1.1.4 The Role of Chance

It is hard to overestimate the role played in our lives by those phenomena that are of a chance nature. The nuclear reactions occurring in the depths of the Sun are the source of the energy sustaining all life on Earth. We are surrounded by the medium of light and the electromagnetic field which are composed of the quanta emitted by the individual atoms of the Sun's corona. Fluctuations in this emission – the solar flares – affect meteorological processes in a substantial way. Random mechanisms also lead to explosions of supernova stars and to sources of cosmic radiation. Brownian motion results in diffusion and in the mutual permeation of substances and due to it, there are reactions possible and hence even life. Chance mechanisms are responsible for the transmission of hereditary characteristics from parents to children. Cosmic radiation, which is also of a random nature, is one of the sources of mutation of genes due to

which we have biological evolution. Many phenomena conform strictly to laws only due to chance and this proves to be the case whenever a phenomenon is dependent upon a large number of independent random microphenomena (for instance, in gases, where there are a huge number of molecules moving randomly and one has the exact Clapeyron law).

1.2 Formalization of Randomness

In order to make chance a subject of mathematical research, it is necessary to construct a formal system which can be interpreted by real phenomena in which chance is observed. This section is devoted to a first discussion.

1.2.1 Selection from Among Several Possibilities. Random Experiments. Events

A most simple scheme in which unpredictable phenomena occur is in the selection of one element from a finite collection. To describe this situation, probability theory makes use of urn models. Let there be an urn containing balls that differ from one another. A ball is drawn from the urn at random. The phrase "at random" means that each ball in the urn can be withdrawn. Later, we shall make at random still more precise. This single selection can be described strictly speaking as being the enumeration of possibilities and furnishes little for discussion. The matter changes substantially when there are a large number of selections. After drawing a ball from the urn and observing what it was, we return it and we again remove one ball from the urn (at random). Observing what the second ball was, we return it to the urn and we repeat the operation again and so on. Let the balls be numbered $1, 2, \ldots, s$ and repeat the selection n times. The results of our operations (termed an experiment in what follows) can be described by the sequence of numbers of the balls drawn: $\alpha_1, \alpha_2, \ldots, \alpha_n$ with $\alpha_n \in \{1, 2, \ldots, s\}$. Questions of interest in probability include this one. How often is the exact same number encountered in such a sequence? At first glance, the question is meaningless: it can still be anything. Nevertheless, although there are certain restrictions, they are based on the following fact. If n_i is the number of times that ball numbered i is drawn, then $n_1 + n_2 + \ldots + n_s = n$. This is of course a trivial remark but, as explained later on, it will serve as a starting point for building a satisfactorily developed mathematical theory. However, there is another nontrivial fact demonstrated by the simplest case $s = 2$. We write out all of the possible results of the n extractions of which there are 2^n. These are all of the possible sequences of digits 1 and 2 of length $n_1 + n_2 = n$, where n_1 is the number of ones in the sequence and n_2 the number of twos. Let N_ε be the amount of those sequences for which $|n_1/n - 1/2| > \varepsilon$. Then $\lim_{n\to\infty} 2^{-n} N_\varepsilon = 0$ for all positive ε. This is an important assertion and it indicates that for large n the fraction of ones in an overwhelming majority of the sequences is close to

1/2. If the same computation is done for s balls, then it can be shown that the fraction of ones is $1/s$ in an overwhelming majority of the sequences. This holds for any $i \leq s$. That the "encounterability" of different numbers in the sequences must be the same can be discerned directly without computation by way of the following symmetry property. If the places of two numbers are interchanged, there are again the same 2^n sequences. Probability theory treats this property as the "*equal likelihood*" of occurrence of each of the numbers in the sequence. Assertions about the relative number of sequences for which n_i/n deviates from $1/s$ by less than ε are examples of the "*law of large numbers*", the class of probability theorems most generally used in applications.

We now consider the notion of "*random experiment*", which is a generalization of the selection scheme discussed above. Suppose that a certain complex of conditions is realized resulting in one of several possible *events*, where generally a different event can occur on iterating the conditions. We then say that we have a random experiment. It is determined by the set of conditions and the set of possible outcomes (observed events). The conditions of the experiment may or may not depend on the will of an experimenter (created artificially) and the presence or absence of an experimenter also plays no role. It is also inessential whether it is possible in principle to observe the outcome of the experiment. Any sufficiently complicated event can generally be placed under the concept of random experiment if one chooses as conditions those that do not determine its course completely. The pattern of its course is then a result of the experiment. The main thing for us in a random experiment is the possibility of repeating it indefinitely. Only for large series of iterated experiments is it possible to obtain meaningful assertions. Examples of physical phenomena have already been given above in which randomness enters. If we consider radioactive decay, for example, then each individual atom of a radioactive element undergoes radioactive conversion in a random fashion. Although we cannot follow each atom, a conceptual experiment can be performed which can help establish which of the atoms have already undergone a nuclear reaction and which still have not. In the same way, by considering a volume of gas, we can conceive an experiment which can ascertain the energies of all of the molecules in the gas. If the possible outcomes of an experiment are known, then we can imagine the experiment as choosing from among several possibilities. Again considering an urn containing balls, we can assume that each ball has one of the possible outcomes of the pertinent experiment written on it and any possibility has been written on one of the balls. On drawing one of the balls, we ascertain which one of the possibilities has been realized. Such a description of an experiment is advantageous because of its uniformness. We point out two difficulties arising in associating an urn model with an experiment. First, it is easy to imagine an experiment which in principle has infinitely many different outcomes. This will always be the case whenever an experiment is measuring a continuously varying quantity (position, energy, etc.). However, in practical situations a continuously varying quantity is measured with a certain accuracy. Second, there is a definite symmetry

among the possibilities in the urn model, which was discussed above. It would be unnatural to expect every experiment to have this property. However, the symmetry can be broken by increasing the number of balls and viewing some of them as identical. The indistinguishable balls correspond to one and the same outcome of the experiment but the number of such balls varies from outcome to outcome. Say that an experiment has two outcomes and one ball corresponds to outcome 1 and two balls to outcome 2. Then in a long run of trials, outcome 2 should be encountered twice as often as outcome 1.

In discussing the outcomes of an experiment above, we meant all possible mutually exclusive outcomes. They are usually called "*elementary events*" or "*sample points*". They can be used to construct an "algebra of events" that are observable in an experiment. Events that are observable in an experiment will be denoted by $A, B, C, \ldots$. We now define operations on events. The sum or union of two events A and B is the event that occurs if and only if at least one of A or B occurs and it is denoted by $A \cup B$ or $A + B$. The product or intersection of two events A and B is the event that both A and B occur (simultaneously) and it is denoted by $A \cap B$ or AB. An event is said to be impossible if it can never occur in an experiment (we denote it by $\emptyset$) and to be sure if it always occurs (we denote it by U). The event $\bar{A}$ is the complement of A and corresponds to A not happening. The event $A \cap \bar{B}$ is the difference of A and B and is denoted by $A \setminus B$.

A collection $\mathcal{A}$ of events observable in an experiment is called an *algebra of events* if together with each A it contains $\bar{A}$ and together with each pair A and B it contains $A \cup B$ (the collection $\mathcal{A}$ is nonempty). Since $A \cup \bar{A} = U$, $U \in \mathcal{A}$ and $\emptyset = \bar{U} \in \mathcal{A}$. If A and $B \in \mathcal{A}$, then $A \cap B = \overline{(\bar{A} \cup \bar{B})} \in \mathcal{A}$ and $A \cap \bar{B} \in \mathcal{A}$. Thus the operations on events introduced above do not lead out of the algebra. Let $A_1, A_2, \ldots, A_m$ be a set of events. A smallest algebra of events exists containing these events. We introduce the natural assumption that the events that are observable in an experiment form an algebra. If $A_1, A_2, \ldots, A_m$ are all elementary events of a given experiment, then the algebra of events observable in the experiment comprises events of the form

$$A = \bigcup_{k \in \Lambda} A_k, \quad \Lambda \subset \{1, 2, \ldots, m\}\,, \tag{1.2.1}$$

where Λ is any subset of the segment of integers $\overline{1, m}$; if $\Lambda = \emptyset$, then A is considered to be the impossible event. Let Ω denote the set of elementary events or *sample space*. Every event may be viewed as a subset of Ω. More precisely, one can associate with each event A the set of elementary events A_k occurring in the union on the right of (1.2.1).

As a result there is a one-to-one correspondence between the events in an experiment and the subsets of Ω in which a sum of events corresponds to a union of sets, a product of events to an intersection of sets and the opposite event to the complement of a set in Ω. The relation $A \subset B$ for subsets of Ω has the probabilistic meaning that the event A implies event B because B occurs

whenever A occurs. The interpretation of events as subsets of a set enables us to make set theory the basis of our probability-theoretic development and to avoid in what follows such indefinite terminology as "event", "occurs in an experiment" and so on.

1.2.2 Relative Frequencies. Probability as an Ideal Relative Frequency

Consider some experiment and let Ω be the set of elementary events that can occur in the experiment. Let $\mathcal{A}$ be an algebra of observable events in the experiment. $\mathcal{A}$ is a collection of subsets of which together with A contains $\Omega \setminus A$ and together with each pair of sets A and B contains $A \cup B$. The elements of Ω will be denoted by $\omega, \omega_1, \omega'$, etc. Suppose that the experiment is repeated n times. Let ω_k denote the outcome in the k-th experiment; the n-fold repetition of the experiment determines a sequence $(\omega_1, \ldots, \omega_n)$, or in other words, a point of the space Ω^n (the n-th Cartesian power of Ω). An event A occurred in the k-th experiment if $\omega_k \in A$. Let $n(A)$ denote the number of occurrences of A in these n experiments. The quantity

$$\nu_n(A) = \frac{n(A)}{n} \tag{1.2.2}$$

is the *relative frequency* of A (in the stated series of experiments). The relative frequency of A characterizes a connection between A and the conditions of the experiment. Thus, if the conditions of the experiment always imply the occurrence of A, that is, the connection between the conditions of the experiment and A is determinate, then $\nu_n(A) = 1$. If A is impossible under the conditions of the experiment, then $\nu_n(A) = 0$. The closer $\nu_n(A)$ is to 1 or 0, the more "strictly" is the occurrence (nonoccurrence) of A tied to the conditions of the experiment.

We now indicate the basic properties of a relative frequency.

1. $0 \le \nu_n(A) \le 1$ with $\nu_n(\emptyset) = 0$ and $\nu_n(U) = 1$. Two events A and B are said to be disjoint or mutually exclusive if $A \cap B = \emptyset$, that is, they cannot occur simultaneously.
2. If A and B are mutually exclusive events, then $\nu_n(A \cup B) = \nu_n(A) + \nu_n(B)$. Thus the relative frequency is a non-negative additive set-function defined on $\mathcal{A}$ and it is normalized: $\nu_n(\Omega) = \nu_n(U) = 1$.

Relative frequency is a function of the sequence of outcomes of an experiment:

$$\nu_n(A) = n^{-1} \sum_{k=1}^{n} I_A(\omega_k), \tag{1.2.3}$$

where I_A is the indicator function of A. If another sequence of outcomes is considered, the relative frequency can change. In the discussion of the urn

model, it was said that for a large number n of observations, the fraction of sequences $(\omega_1, \ldots, \omega_n)$ for which a relative frequency differs little from a certain number approaches 1. Therefore the variability of relative frequency does not preclude some "ideal" value around which it fluctuates and which it approaches in some sense. This ideal value of the relative frequency of an event is then its *probability*. Our discussion has a very vague meaning and it may be viewed as a heuristic argument. Just as actual cats are imperfect "copies" of an ideal cat (the idea of a cat) according to Plato, relative frequencies are likewise realizations of an absolute (ideal) relative frequency – the probability. The sole pithy conclusion that can be drawn from the above heuristic discussion is that probability must preserve the essential properties of relative frequency, that is, it should be a non-negative additive function of events and the probability of the sure event should be 1.

1.2.3 The Definition of Probability

The preceding considerations can be used in different ways to define probability. The initial naive view of the matter was that probabilities of events exist objectively and therefore probability needs no defining. The question was how to calculate a probability.

(a) *The classical definition of probability.* Games of chance and the analysis of testimony of witnesses were originally the basic areas of application of probability theory. Games of chance involving cards, dice and flipping coins naturally permitted the creation of appropriate random experiments (this terminology first appeared in the twentieth century) so that their outcomes had symmetry in relation to the conditions of the experiment. These outcomes were treated as "*equally likely*" and they were assigned the same probabilities. Thus, if there are s outcomes in the experiment, each elementary event was assigned a probability of $1/s$ (it is easy to see that an elementary event has that probability using the additivity of probability and the fact that the sure event has probability one). If an event is expressed as the union of r elementary events ($r \leq s$), then the probability of A is r/s by virtue of the additivity. Thus we arrive at the definition of probability that has been in use for about two centuries.

The probability of an event A is the quotient of the number of outcomes favorable to A and the number of all possible outcomes. The outcomes favorable to A are understood to be those that imply A.

This is the classical definition of probability. With this definition as a starting point, it is possible to establish that probability has the properties indicated in Sect. 1.2.2. The definition is convenient, consistent and allows results obtained by the theory to have a simple interpretation. A deficiency is the impossiblity of extending it to experiments with infinitely many outcomes or to any case in which the outcomes are asymmetric in relation to the conditions of the experiment. In particular, the classical set-up has no events with irrational probabilities.

(b) *The axioms of von Mises.* The German mathematician R. von Mises proposed as the definition of probability the second of the properties mentioned for urn models – the convergence of a relative frequency to some limiting value in the sense indicated there. Von Mises gave a system of probability axioms whose first one postulates the existence of the limit of a relative frequency and this limit is called the probability of an event. Such a system of axioms results in considerable mathematical difficulties. On the one hand, there is the possibility of varying the sequence of experiments and on the other hand, the definition is too empirical and so it hardly accommodates mathematical study. The ideas of von Mises can be used in some interpretations of the results of probability but they are untenable for constructing a mathematical theory.

(c) *The axioms of Kolmogorov.* The set of axioms of A.N. Kolmogorov has been universally recognized as the starting point for the development of probability theory. He proposed them in his book "Fundamental Concepts of Probability Theory." These axioms employ only the most general properties which are inherent to probability about which we spoke above. First of all, Kolmogorov considered the set-theoretic treatment already discussed above and also the notion of random experiment. He postulated the existence of the probability of each event occurring in a random experiment. Probability was assumed to be a nonnegative additive function on the algebra of events with the probability of the sure event equal to 1. Thus a random experiment is formally specified by a triple of things: 1. a sample space Ω of elementary events; 2. an algebra $\mathcal{A}$ of its subsets, the members of $\mathcal{A}$ being the random events; 3. a nonnegative additive function $\mathbf{P}(A)$ defined on $\mathcal{A}$ for which $\mathbf{P}(\Omega) = 1$; $\mathbf{P}(A)$ is termed the probability of A. If random experiments with infinitely many outcomes are considered, then it is natural to require that $\mathcal{A}$ be a *σ-algebra* (or *σ-field*). In other words, together with each sequence of events A_n, $\mathcal{A}$ also contains the countable union $\bigcup_n A_n$ and $\mathbf{P}(A)$ must be a countably-additive function on $\mathcal{A}$: if $A_n \cap A_m = \emptyset$ for $n \neq m$, then $\mathbf{P}(\bigcup_n A_n) = \sum_n \mathbf{P}(A_n)$. This means that $\mathbf{P}$ is a measure on $\mathcal{A}$ and since $\mathbf{P}(\Omega) = 1$, the measure is normalized.

1.3 Problems of Probability Theory

Initially, probability theory was the study of ways of computing probabilities of events knowing the probabilities of other given events. The techniques developed for computing the probabilities of certain classes of events now form a constituent unit of probability but only partly and far from the main part. However, as before, probability theory only deals with the probabilities of events independently of what meaningful sense can be invested in the words "the probability of event A is p". This means that probability theory itself does interpret its results meaningfully but in so doing it does not exclude the term "probability". There is no statement like "A always occurs" but rather the statement "A occurs with probability one".

1.3.1 Probability and Measure Theory

Kolmogorov's axioms make probability theory a special part of measure theory namely finite measure theory (being finite and being normalized are clearly essentially equivalent since any finite measure may be converted into a normalized measure by multiplication by a constant). If this is so, is probability theory unnecessary? The answer to this question has already been given by the development of probability theory following the introduction of Kolmogorov's axioms. Probability theory does employ measure theory in an essential way but classical measure theory really involves the construction of a measure by extension and the development of the integral and its properties including the Radon-Nikodym theorem. Probability theory has inspired new problems in measure theory: the convergence of measures and construction of a measure fibre ("conditional" measure); these now belong traditionally to probability theory. A completely new area of measure theory is the analysis of absolute continuity and singularity of measures. The Radon-Nikodym theorem of measure theory serves merely as a starting point for the development of the very important theory of absolute continuity and singularity of probability measures (also of consequence in applications). Its meaningfulness lies in the broad class of special probability measures that it examines. Finally, the specific classes of measures in probability theory, say, product measures or fibre bundles of measures, establish the nature of its position in relation to general measure theory. This manifests itself in the concepts utilized such as independence, weak dependence and conditional dependence, which are more associated with certain physical ideas at the basis of our probabilistic intuition. These same concepts lead to problems whose reformulations in the language of measure theory prove to be cumbersome, unclear and perplexing making one wonder where these problems arose. (For individuals familiar with probability theory, as an example, it is suggested that one formulate the degeneracy problem for the simplest branching process in terms of measure theory.) Nonetheless, there are a number of sections of probability that can relate immediately to measure theory, for instance, measure theory in infinite-dimensional linear spaces. Having originated in probability problems, they remain traditionally within the framework of probability theory.

1.3.2 Independence

Independence is one of the basic concepts of probability theory. According to Kolmogorov, it is exactly this that distinguishes probability theory from measure theory. Independence will be discussed more precisely later on. For the moment, we merely point out that stochastic independence and physical independence of events (one event having no effect on another) are identical in content. Stochastic independence is a precisely-defined mathematical concept to be given below. At this point, we note that independence was already used in latent form in the definition of random experiment. One of the requirements

imposed on an experiment is the possibility of iterating it indefinitely. To iterate it assumes that the conditions of the experiment can be reconstructed after which the one just performed and all of the prior ones have no affect on the outcome of the next experiment. This means that the events occurring in different experiments must be independent.

Probability theory also studies laws of large numbers for independent experiments. One such law has already been stated on an intuitive level. An example is Bernoulli's form of the law of large numbers: "Given a series of independent trials in each of which an event A can occur with probability p and $\nu_n(A)$ the relative frequency of A in the first n trials. Then the probability that $|\nu_n(A) - p| > \varepsilon$ tends to zero as $n \to \infty$ for any positive ε." Observe that the value of $\nu_n(A)$ is random and so the fulfillment of the inequality in this theorem is a random event. The theorem is a precise statement of the fact that the relative frequency of an event approaches its probability. As will be seen below, the proof of this assertion is strictly mathematical. It may seem paradoxical that it is possible to use mathematics to obtain precise knowledge about randomly-occurring events (that it is possible to do so in a determinate world, say, to calculate the dates of lunar eclipses, is quite natural). In fact, the choice of p is supposedly arbitrary and only the fulfillment of Kolmogorov's axioms is required. However, something interesting can be extracted from Bernoulli's theorem only if events of small probability actually rarely occur in practice. It is precisely these kinds of events (or events whose probability is close to 1) that interest us primarily in probability. If one comes to the point of view that events of probability 0 practically never occur and events of probability 1 practically always occur, then the kind of conclusions that may be drawn from random premises will be of interest.

1.3.3 Asymptotic Behavior of Stochastic Systems

Many physical, engineering and biological objects may be viewed as randomly evolving systems. Such a system is in one of its possible states (frequently viewable as finitely many) and with the passage of time the system changes its state at random. One of the major problems of probability is to study the asymptotic behavior of these systems over unbounded time intervals. We give one of the possible results in order to demonstrate the problems arising here. Let $T_t(E)$ be the total time that a system spends in the state E on the time interval $[0, t]$. Then the nonrandom

$$\lim_{t\to\infty} \frac{1}{t} T_t(E) = \pi(E)$$

exists with probability 1; $\pi(E)$ is the probability that the system will be found in the state E after a sufficiently long time. More precisely, the probability that the system is in the state E at time t tends to $\pi(E)$ as $t \to \infty$. This assertion holds of course under certain assumptions on the system in question. We cannot state them at this point since the needed concepts still have

not been introduced. Assertions of this kind are lumped together under the generic name of *ergodic theorems*. Just as for the laws of large numbers, they provide reliable conclusions from random premises. One may be interested in a more exact behavior of the sojourn time in a given state, for instance, in studying the behavior of the difference $[t^{-1}T_t(E) - \pi(E)]$ multiplied by a suitable increasing function of t (the difference itself tends to zero). Under very broad assumptions, this difference multiplied by $\sqrt{t}$ behaves primarily the same way for all systems. We have now the second most important probability law (after the law of large numbers), which may be called the *law of normal fluctuations*. It holds also for relative frequencies and says that the deviation of a relative frequency from a probability after multiplication by a suitable constant behaves the same way in all cases (this is expressed precisely by the phrase "has a normal distribution"; what this means will be explained later on). Among the practically important problems involving stochastic systems is "predicting" their behavior from observations of their past behavior.

1.3.4 Stochastic Analysis

Moving on from the concept of random event, one could "randomize" any mathematical object. Such randomization is widely employed and studied in probability. The new objects do not result in idle philosophizing. They come about in an essential way and nontrivial important theorems are associated with them that find extensive application in the natural sciences and engineering. The first thing of this kind is the random number (or random variable in the accepted terminology). Such variables appear in experiments in which one or more characteristics of the experimental results are being measured. Following this, it is natural to consider the arithmetic of these variables and then to extend the concepts of mathematical analysis to them: limit, functional dependence and so on. Thus we arrive at the notions of random function, random operator, random mapping, stochastic integral, stochastic differential equation, etc. This is a comparatively new rather intensively developing area of probability theory. Despite their stochastic coloration, the problems that arise here are often analogous to problems of ordinary analysis.

2

Probability Space

The probability space is the basic object of study in probability theory and formalizes the notion of random experiment. A *probability space* is defined by three things: the space Ω of elementary events or sample space, a σ-algebra $\mathcal{A}$ of subsets of Ω called events, and a countably-additive nonnegative normalized set function $\mathbf{P}(A)$ defined on $\mathcal{A}$, which is called probability. A probability space defined by this triple is denoted by $(\Omega, \mathcal{A}, \mathbf{P})$.

2.1 Finite Probability Space

A finite probability space is one whose sample space is a finite set and $\mathcal{A}$ comprises all of the subsets of Ω. The probability is defined by its values on the elementary events.

2.1.1 Combinatorial Analysis

Suppose that the probabilities of all of the elementary events are the same (they are equally likely). To find the probability of an event A, it is necessary to know the overall number of elementary events and the number of those elementary events which imply A. The number of elements in a finite set can be calculated using direct methods that sort out all of the possibilities or combinatorial methods. Only the latter are of mathematical interest. We consider some examples applying them.

(a) *Allocation of particles in cells.* Problems of this kind arise in statistical physics. Given n cells in which N particles are distributed at random. What is the distribution of the particles in the cells? The answer depends on what are considered to be the elementary events.

Maxwell-Boltzmann statistics. We assume that all of the particles are distinct and all allocations of particles are equally likely. An elementary event is given by the sequence $(k_1, k_2, \dots k_N)$, where k_i is the number of the cell into which the particle numbered i has fallen. Since each k_i assumes n distinct values, the number of such sequences is n^N. The probability of an elementary event is n^{-N}.

Bose-Einstein statistics. The particles are indistinguishable. Again all of the allocations are equally likely. An elementary event is given by the sequence $(\ell_1, \dots, \ell_n)$, where $(\ell_1 + \dots + \ell_n = N$ and ℓ_i is the number of particles in the i-th cell, $i \le n$. The number of such sequences can be calculated as follows. With each $(\ell_1, \dots, \ell_n)$ associate a sequence of zeroes and ones $(i_1, \dots, i_{k+n-1})$ with zeroes in the positions numbered $\ell_1+1, \ell_1+\ell_2+2, \dots, \ell_1+\ell_2+\dots+\ell_{n-1}+n-1$ (there are $n-1$ of them) and ones in the remaining positions. The number of such sequences is equal to the number of combinations of $N+n-1$ things taken $n-1$ at a time. The probability of an elementary event is $\binom{N+n-1}{n-1}^{-1}$.

Fermi-Dirac statistics. In this case $N < n$ and each cell contains at most one particle. Then the number of elementary events is $\binom{n}{N}^{-1}$.

For each of the three statistics, we find the probability that a given cell (say, number 1) has no particle. Each time the number of favorable elementary events equals the number of allocations of the particles into $n-1$ cells. Therefore if we let p_1, p_2, and p_3 be the probabilities of the specified event for each statistics (in order of discussion), we have

$$
\begin{aligned}
p_1 &= (n-1)^N / n^N = \left(1 - \frac{1}{n}\right)^N , \\
p_2 &= \binom{N+n-2}{n-2} \Big/ \binom{N+n-1}{n-1} = \frac{n-1}{N+n-1} , \\
p_3 &= \binom{n-1}{N} \Big/ \binom{n}{N} = 1 - \frac{N}{n} .
\end{aligned}
$$

If $N/n = \alpha$ and $n \to \infty$, then

$$
p_1 = e^{-\alpha}, \quad p_2 = \frac{1}{1+\alpha}, \quad p_3 = 1 - \alpha .
$$

For small α, these probabilities coincide up to $O(\alpha^2)$. α characterizes the "average density" of the particles. If α is small, then the three probabilities are primarily equal.

(b) *Samples.* A sample may be defined in general as follows. There are m finite sets $A_1, A_2, \dots, A_m$. From each set, we choose an element $a_i \in A_i$ one by one. The collection $(a_1, \dots, a_m)$ is then the sample. Samples are distinguished

by identification rules (let us say, we are not interested in the order of the elements in a sample). Each sample is regarded as an elementary event and the elementary events are considered to be equally likely.

1. *Sampling with replacement.* In this instance, the A_i coincide: $A_i = A$ and the number of samples is n^m, where n is the number of elements in A.
2. *Sampling without replacement.* A sample is constructed as follows. $A_1 = A$, $A_2 = A\backslash\{a_1\}, \ldots, A_k = A\backslash\{a_1, \ldots, a_{k-1}\}$. In other words, only samples $(a_1, \ldots, a_m)$, $a_i \in A$, are considered in which all of the elements are distinct. If A has n elements, then the number of samples without replacement is $n(n-1)\ldots(n-m+1)/m! = \binom{n}{m}$.
3. *Sampling without replacement from intersecting sets.* In this instance, the A_i have points in common but we are considering samples in which all of the elements are distinct. The number of such samples may be computed as follows. Consider the set $A = \bigcup_{k=1}^{m} A_k$ and the algebra $\mathcal{A}$ of subsets of it generated by $A_1, \ldots, A_m$. This is a finite algebra. Let $B_1, B_2, \ldots, B_N$ be atoms of the algebra, that is, they each have no subsets belonging to the algebra other than the *empty set* and themselves. Let $n(B_{i_1}, \ldots, B_{i_m})$ denote the number of samples without replacement from $B_{i_1}, \ldots, B_{i_m}$, where each B_{i_k} may be any atom. The value of $n(B_{i_1}, \ldots, B_{i_m})$ depends on the distinct sets encountered in the sequence and on the number of times these sets are repeated. Let $n(\ell_1, \ell_2, \ldots, \ell_N)$ be the number of samples from such a sequence, where B_1 occurs ℓ_1 times, B_2 occurs ℓ_2 times and so on, $\ell_i \geq 0, \ell_1 + \ldots + \ell_N = m$. If B_i has n_i elements, then

$$n(\ell_1, \ldots, \ell_N) = \prod_{i=1}^{N} \frac{n_i!}{(n_i - \ell_i)!} .$$

The number of samples of interest to us equals

$$\sum_{B_{i_1} \subset A_1, \ldots, B_{i_m} \subset A_m} n(B_{i_1}, \ldots, B_{i_m}) .$$

2.1.2 Conditional Probability

The *conditional probability* of an event A given event B having positive probability has occurred is the quantity

$$P(A|B) = \frac{P(A \cap B)}{P(B)} . \tag{2.1.1}$$

As a function of A, $\mathbf{P}(A|B)$ possesses all of the properties of a probability. The meaning of conditional probability may be explained as follows. Together with the original experiment, consider a conditional probability experiment which is performed if event B has happened in the original experiment. Thus if

the original experiment has been done n times and B has happened n_B times, then this sequence contains n_B conditional experiments. The event A will have occurred in the conditional experiment if A and B occur simultaneously, i.e., if $A \cap B$ occurs. If $n_{A\cap B}$ is the number of experiments in which the event $A \cap B$ is observed (of the n carried out), then the relative frequency of occurrence in the n_B conditional experiments is $n_{A\cap B}/n_B = \nu_n(A\cap B)/\nu_n(B)$. If we replace the relative frequencies by the probabilities, then we have the right-hand side of (1.2.1).

(a) *Formula of total probability. Bayes's theorem.* A finite collection of events $H_1, H_2, \ldots, H_r$ is said to form a complete group of events if they are pairwise disjoint and their union is the sure event: 1. $H_i \cap H_j = \emptyset$ if $i \neq j$; 2. $\bigcup_i H_i = \Omega$. One can consider a supplementary experiment in which the H_i are the elementary events and the original experiment is viewed as a compound experiment: first one clarifies which H_i has occurred and then knowing H_i, one performs a conditional experiment under the assumption that H_i has occurred. An event A occurs in the conditional experiment with probability $P(A|H_i)$, the conditional probability of A given H_i. In many problems, the H_i are called the *causes* or *hypotheses* and the conditional probabilities given the causes are prescribed. The following relation expressing the probability of an event in terms of these conditional probabilities and the probabilities of causes is called the *formula of total probability*:

$$\mathbf{P}(A) = \sum_{i=1}^{r} \mathbf{P}(A|H_i)\mathbf{P}(H_i) \; . \tag{2.1.2}$$

On the basis of (2.1.1) the right-hand side becomes $\sum_{i=1}^{r} \mathbf{P}(A\cap H_i)$ and since the events $A \cap H_i$ are mutually exclusive and $\cup H_i = \Omega$, it follows that

$$\sum_{i=1}^{r} \mathbf{P}(A \cap H_i) = \mathbf{P}\left(\bigcup_{i=1}^{r}(A \cap H_i)\right) = \mathbf{P}\left(A \cap \bigcup_{i=1}^{r} H_i\right) = \mathbf{P}(A) \; .$$

Formula (2.1.2) is really useful when considering a compound experiment.

Example. There are r urns containing black and white balls. The probability of drawing a white ball from the urn numbered i is p_i. One of the urns is chosen at random and then a ball is drawn from it. By formula (2.1.2), we determine the probability of drawing a white ball. In our case, $\mathbf{P}(H_i) = 1/r$, $\mathbf{P}(A|H_i) = p_i$ and hence $\mathbf{P}(A) = r^{-1} \sum_{i=1}^{r} p_i$.

The formula of total probability leads to an important result called *Bayes's theorem*. It enables one to find the conditional probabilities of the causes given that an event A has occurred:

$$\mathbf{P}(H_k|A) = \mathbf{P}(A|H_k)\mathbf{P}(H_k) \Big/ \sum_{i=1}^{r} \mathbf{P}(A|H_i)\mathbf{P}(H_i) \; . \tag{2.1.3}$$

This formula is commonly interpreted as follows. The conditional probabilities of an event given each of the causes $H_1, \ldots, H_r$ and the probabilities of the causes are assumed to be known. If the experiment has resulted in the occurrence of event A, then the probabilities of the causes have changed: once we know that A has already occurred, then it is natural to treat the probabilities of the causes as their conditional probabilities given A. The $\mathbf{P}(H_i)$ are called the apriori probabilities of the causes and the $\mathbf{P}(H_i|A)$ are their aposteriori probabilities. Bayes's theorem expresses the aposteriori probabilities of the causes in terms of their apriori probabilities and the conditional probabilities of an event given the various causes.

Example. There are two urns of which the first contains 2 white and 8 black balls and the second 8 white and 2 black balls. An urn is selected at random and a ball is drawn from it. It is white. What is the probability that the first urn was chosen? Here we have $\mathbf{P}(H_1) = \mathbf{P}(H_2) = 1/2$, $\mathbf{P}(A|H_1) = 1/5$ and $\mathbf{P}(A|H_2) = 4/5$. By (2.1.3),

$$\mathbf{P}(H_1|A) = 1/2 \cdot 1/5/(1/2 \cdot 1/5 + 1/2 \cdot 4/5) = 1/5 \, .$$

(b) *Independence.* An event A does not depend on an event B if the conditional probability $\mathbf{P}(A|B)$ equals the unconditional probability $\mathbf{P}(A)$. In that case,

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) \, , \tag{2.1.4}$$

which shows that the property of *independence* is symmetric. Formula (2.1.4) could serve as a definition of independence of two events A and B. The first definition is more meaningful: the fact that B has occurred has no affect on the probability of A and it is reasonable to assume that A does not depend on B. It follows from (2.1.4) that the independence of A and B implies the independence of A and $\bar{B}$, $\bar{A}$ and B, and $\bar{A}$ and $\bar{B}$ ($\bar{A}$ is the negation of the event A). Independence is defined for several events as follows. $A_1, A_2, \ldots, A_m$ are said to be mutually independent if

$$\mathbf{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_K}) = \mathbf{P}(A_{i_1}) \ldots \mathbf{P}(A_{i_k}) \tag{2.1.5}$$

for any $k \leq m$ and $i_1 < i_2 \ldots < i_k \leq m$. Thus for three events A, B and C their independence means that the following four equalities hold: $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$, $\mathbf{P}(A \cap C) = \mathbf{P}(A)\mathbf{P}(C)$, $\mathbf{P}(B \cap C) = \mathbf{P}(B)\mathbf{P}(C)$ and $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)$.

Bernstein's example. The sample space consists of four elements E_1, E_2, E_3, and E_4 with $\mathbf{P}(E_k) = 1/4$, $k = 1, 2, 3, 4$. Let $A_i = E_i \cup E_4, i = 1, 2, 3$. Then $A_1 \cap A_2 = A_1 \cap A_3 = A_2 \cap A_3 = A_1 \cap A_2 \cap A_3 = E_4$. Therefore $\mathbf{P}(A_1 \cap A_2) = \mathbf{P}(A_1)\mathbf{P}(A_2)$, $\mathbf{P}(A_1 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_3)$ and $\mathbf{P}(A_2 \cap A_3) = \mathbf{P}(A_2)\mathbf{P}(A_3)$. But $\mathbf{P}(A_1 \cap A_2 \cap A_3) \neq \mathbf{P}(A_1)\mathbf{P}(A_2)\mathbf{P}(A_3)$. The events are pairwise independent but they are not mutually independent.

2.1.3 Bernoulli's Scheme. Limit Theorems

Let $A_1, A_2, \ldots, A_r$ be a complete group of events. An event B is independent of this group if it does not depend on any of the events $A_k, k = 1, \ldots, r$. Let $\mathcal{A}$ be the algebra generated by the events $A_1, \ldots, A_r$; it comprises the impossible event and all unions of the form $\bigcup_k A_{i_k}$, $i_k \leq r$. Then B is independent of the algebra $\mathcal{A}$, that is, it does not depend on any event $A \in \mathcal{A}$. Two *algebras* of events $\mathcal{A}_1$ and $\mathcal{A}_2$ are said to be *independent* if A_1 and A_2 are independent for each pair of events $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$. Algebras of events $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m$ are independent if $A_1, A_2, \ldots, A_m$ are mutually independent, where $A_i \in \mathcal{A}_i$, $i \leq m$. To this end, it suffices that

$$\mathbf{P}\left(\bigcap_{i=1}^{m} A_i\right) = \prod_{i=1}^{m} \mathbf{P}(A_i) \tag{2.1.6}$$

for any choice of $A_i \in \mathcal{A}_i$. (This definition simplifies as compared to that of independent events in (2.1.5) because some A_i may be chosen to be Ω.)

Consider several experiments specified by the probability spaces $(\Omega_k, \mathcal{A}_k, \mathbf{P}_k)$, $k = 1, 2, \ldots, n$. We now form a new probability space $(\Omega, \mathcal{A}, \mathbf{P})$. Ω is taken to be the Cartesian product $\Omega_1 \times \Omega_2 \times \ldots \times \Omega_n$. The algebra $\mathcal{A}$ is the *product of algebras* $\mathcal{A}_1 \otimes \mathcal{A}_2 \otimes \ldots \otimes \mathcal{A}_n$ of subsets of Ω generated by sets of the form $A_1 \times A_2 \times \ldots \times A_n$ with $A_k \in \mathcal{A}_k$, $k = 1, 2, \ldots, n$ (an algebra is said to be generated by a collection of sets if it is the smallest algebra containing that collection). Finally, the measure $\mathbf{P}$ is the product of measures $\mathbf{P}_k$: $\mathbf{P} = \prod_{k=1}^{n} \mathbf{P}_k$, that is, $\mathbf{P}(A_1 \times A_2 \ldots \times A_n) = \mathbf{P}(A_1)\mathbf{P}(A_2) \ldots \mathbf{P}(A_n)$. The probability space $(\Omega, \mathcal{A}, \mathbf{P})$ corresponds to a compound experiment in which each of the n experiments specified above is performed independently.

(a) *Bernoulli's scheme* involves a series of independent and identical experiments (trials). This just means that a probability space $(\Omega_1 \times \ldots \times \Omega_n, \mathcal{A}_1 \otimes \ldots \otimes \mathcal{A}_n, \prod_{i=1}^{n} \mathbf{P}_i)$ is defined for every n in which each probability space $(\Omega_k, \mathcal{A}_k, \mathbf{P}_k)$ coincides with the exact same space $(\Omega, \mathcal{A}, \mathbf{P})$. (As we shall see below, it is possible to consider an infinite product of such probability spaces right away; it will not be finite if the given space is nontrivial, that is, Ω contains more than one element.) Let $A \in \mathcal{A}$. The event $\Omega \times \ldots \times A \times \ldots \times \Omega$, where A is in the k-th position and the remaining factors are Ω, is interpreted as the event "A occurred in the k-th experiment." Let $p_n(m)$ denote the probability that A happens exactly m times in n independent trials. Then

$$p_n(m) = \binom{n}{m} p^m (1-p)^{n-m}, \quad p = \mathbf{P}(A)\ . \tag{2.1.7}$$

Indeed, our event of interest is the union of events of the form $A \times \bar{A} \times \ldots \times A \times \ldots \times \bar{A} \times \ldots \times A$, where A occurs in the product m times and $\bar{A}$ occurs $n - m$ times. There are $\binom{n}{m}$ such distinct products and the probability of one such event is $p^m(1-p)^{n-m}$.

Let $A_1, A_2, \dots A_r$ be a complete group of events in an algebra $\mathcal{A}$. Let $p_n(k_1, \dots, k_r)$ be the probability that in n independent trials A_i occurs k_i times, $i = 1, \dots, r$ and $k_1 +, \dots, +k_r = n$. Similarly to the preceding, one can establish that

$$p_n(k_1, \dots, k_r) = \frac{n!}{k_1! \dots k_r!} p_1^{k_1} \dots p_r^{k_r}, \quad p_i = \mathbf{P}(A_i) , \quad i = 1, \dots, r . \quad (2.1.8)$$

The probabilities (2.1.7) are called the *binomial probabilities* and (2.1.8) the *multinomial probabilities.*

(b) *The law of large numbers.* This law has been mentioned several times in the introductory chapter. We are now in a position to prove it.

Bernoulli's Theorem. *Let ν_n be the number of occurences of an event A in n independent trials having probability p in each trial, $0 < p < 1$. Then for any positive ε,*

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{1}{n}\nu_n - p\right| > \varepsilon\right\} = 0 . \quad (2.1.9)$$

Proof. For fixed n, the event $\{\nu_n = k\}$ has probability $p_n(k)$. For different values of k, these events are mutually exclusive. Therefore

$$\mathbf{P}\left\{\left|\frac{1}{n}\nu_n - p\right| > \varepsilon\right\} = \sum_{k<n(p-\varepsilon)} p_n(k) + \sum_{k>n(p+\varepsilon)} p_n(k) .$$

Starting with (2.1.7), we find that

$$\frac{p_n(k+1)}{p_n(k)} = \frac{n-k}{k+1}\frac{p}{1-p} .$$

Therefore for $k > n(p+\varepsilon)$,

$$\frac{p_n(k+1)}{p_n(k)} < \frac{n-n(p+\varepsilon)}{np}\frac{p}{1-p} = 1 - \frac{\varepsilon}{1-p} .$$

Let $k*$ denote the smallest value of k satisfying $k > n(p+\varepsilon)$ and let k_* be the smallest value of k for which $(n-k)p/[(k+1)(1-p)] < 1$. Then $p_n(k+1) < p_n(k)$ for $k \geq k_*$. Therefore

$$\sum_{k>n(p+\varepsilon)} p_n(k) = \sum_{k\geq k^*} p_n(k) < p_n(k^*) \sum_{m=0}^{\infty} \left(1 - \frac{\varepsilon}{1-p}\right)^m = \frac{(1-p)}{\varepsilon} p_n(k^*) .$$

Next,

$$1 \geq \sum_{k=k_*}^{k^*} p_n(x) \geq (k^* - k_*) p_n(k^*)$$

and so

$$p_n(k^*) \le (k^* - k_*)^{-1} .$$

Since k_* is the smallest value of k such that $k > np+p-1$, we have $k^* - k_* \ge \varepsilon n - 2$ and thus $\sum_{k>n(p+\varepsilon)} p_n(k) = O\left(\frac{1}{\varepsilon^2 n}\right)$ as $n \to \infty$. A similar estimate also holds for $\sum_{k<n(p-\varepsilon)} p_n(k)$. □

(c) *Law of rare events.* Consider now the asymptotic behavior of the binomial probabilities assuming that the probability of A tends to zero. A nontrivial result is obtained if the number of trials increases in such a way that the product $np = a$ remains bounded and nonvanishing.

Poisson's Theorem. *If the specified assumptions hold, then*

$$p_n(m) \sim (m!)^{-1} a^m e^{-a} . \tag{2.1.10}$$

The proof is a consequence of the following operations:

$$\begin{aligned} p_n(m) &= \frac{n(n-1)\dots(n-m+1)}{m!} \left(\frac{a}{n}\right)^m \left(1 - \frac{a}{n}\right)^{n-m} \\ &\sim \frac{1}{m!} \left(1 - \frac{1}{n}\right) \dots \left(1 - \frac{m-1}{n}\right) a^m \exp\left\{(m-n)\frac{a}{n}\right\} \sim (m!)^{-1} a^m e^{-a} . \end{aligned}$$

The collection of quantities $p_m(a) = (m!)^{-1} a^m e^{-a}$, $m = 0, 1, 2, \dots$, defines the *Poisson distribution* with parameter a; the sum of all of the probabilities $p_m(a)$ is 1. Many practical situations involve "rare" random events of the same kind that obey a Poisson distribution: the probability that m events occur in a certain time interval equals $p_m(a)$, where the parameter a is proportional to the length of the interval. Examples of such rare events are: 1. the number of cosmic particles registered by a Geiger counter; 2. the number of calls received at a telephone exchange; 3. the number of accidents; 4. the number of spontaneous catastrophes, and so on.

(d) *Normal approximation.* We now find an asymptotic approximation to $p_n(m)$ for large n and for p bounded away from 0 and 1.

DeMoivre-Laplace Theorem. *Let δ be any positive quantity. Then uniformly for $p(1-p) \ge \delta$ and $|x| \le 1/\delta$,*

$$p_n(m) \sim (2\pi np(1-p))^{-1/2} \exp\left\{-\frac{x^2}{2}\right\} , \tag{2.1.11}$$

where $x = (m - np)/\sqrt{np(1-p)}$.

To show that (2.1.11) holds, we express $p_n(m)$ with the help of *Stirling's formula* ($n! \sim \sqrt{2\pi n}\, n^n e^{-n}$) obtaining

$$p_n(m) = \frac{n!}{(n-m)!m!}p^m(1-p)^{n-m} \sim n^n e^{-n}\sqrt{2\pi n}(n-m)^{-n+m}$$
$$\times \left[e^{-n+m}\sqrt{2\pi(n-m)}\right]^{-1} m^{-m}\left[e^{-m}\sqrt{2\pi m}\right]^{-1} p^m(1-p)^{n-m}$$
$$= \left(2\pi n\left(1-\frac{m}{n}\right)\frac{m}{n}\right)^{-1/2}\left(\frac{np}{m}\right)^m\left(\frac{n(1-p)}{n-m}\right)^{n-m}.$$

Solving for m in terms of x so that $m = np + x\sqrt{np(1-p)}$ and $n - m = n(1-p) - x\sqrt{np(1-p)}$, and using the boundedness of x, we obtain

$$p_n(m) \sim (2\pi np(1-p))^{-1/2}\left(1 + x\sqrt{\frac{1-p}{np}}\right)^{-np-x\sqrt{np(1-p)}}$$
$$\times \left(1 - x\sqrt{\frac{p}{n(1-p)}}\right)^{-n(1-p)+x\sqrt{np(1-p)}}.$$

Taking the logarithm of the product of the two power terms involving x and using the expansions of $\ln(1+x\sqrt{(1-p)/np})$ and $\ln\left(1-x\sqrt{p/n(1-p)}\right)$, one can show that this product equals $-\frac{1}{2}x^2 + O(1/\sqrt{n})$.

2.2 Definition of Probability Space

We now discard the finiteness of Ω. It is then natural to replace an algebra of events by a σ-algebra and to define probability as a countably-additive function of the events.

2.2.1 σ-algebras. Probability

A *σ-algebra* $\mathcal{A}$ of subsets of Ω is an algebra which together with each sequence $A_n \in \mathcal{A}$ contains $\bigcup_n A_n$. Then $\mathcal{A}$ also contains $\bigcap_n A_n = \Omega \setminus \bigcup_n (\Omega \setminus A_n)$ and it is therefore closed under countable unions and countable intersections of events. Each algebra of events $\mathcal{A}_0$ can be extended to a σ-algebra by considering all of the sets that can be obtained from those in $\mathcal{A}_0$ by the operations $\cap, \cup$ and $\setminus$ applied at most countably many times. To express such sets, one would have to use transfinite numbers. It is more convenient to employ the following construction which allows one to extend a measure to the σ-algebra also. A collection $\mathcal{M}$ of subsets is said to be monotone if together with every increasing sequence of sets A_n it contains $\bigcup_n A_n$ and with every decreasing sequence B_n it contains $\bigcap_n B_n$.

Theorem on a Monotone Collection. *The smallest monotone collection $\mathcal{M}$ containing an algebra $\mathcal{A}_0$ is the same as the smallest σ-algebra containing $\mathcal{A}_0$.*

This σ-algebra is said to be the σ-algebra generated by $\mathcal{A}_0$. If S is a collection of subsets of Ω, then the smallest σ-algebra containing all of the sets in S is the σ-algebra generated by S and is denoted by $\sigma(S)$.

(a) *Definition of probability.* If Ω is infinite, then a σ-algebra is nondenumerable since the elementary events belong to the σ-algebra as singletons and its power set is nondenumerable. Therefore it is impossible in general to give an effective definition of probability for all events. It is possible to do this in the simplest infinite case where Ω is denumerable and $\mathcal{A}$ is the σ-algebra of subsets of Ω. Every subset is representable as a union of at most countably many elementary events and so a probability can be defined by its values on elementary events. Let $\Omega = \{\omega_k, k = 1, 2, \ldots\}$ and $p_k = \mathbf{P}(\{\omega_k\})$. Then $\mathbf{P}(A) = \sum I_A(\omega_k)p_k$.

If Ω is nondenumerable, one customarily defines probability as an extension from finite algebras. Let $\mathcal{A}_n$ be an increasing sequence of finite algebras and let $\mathcal{A}_0 = \bigcup_n \mathcal{A}_n$ be a denumerable algebra. Let $\{E_n^i, i = 1, \ldots, k_n\}$ be atoms of $\mathcal{A}_n$. To specify probability on the algebra $\mathcal{A}_n$, it suffices to specify the values of $\mathbf{P}(E_n^i), i = 1, \ldots, k_n$. They must satisfy the condition

$$\mathbf{P}(E_n^i) = \sum \mathbf{P}(E_{n+1}^j) I_{\{E_{n+1}^j \subset E_n^i\}}. \tag{2.2.1}$$

This determines the probability on $\mathcal{A}_0$. The probabilities on $\mathcal{A}_0$ uniquely determine the probabilities on $\sigma(\mathcal{A}_0)$. Indeed, the countable additivity of probability is equivalent to its continuity: if $A_n \uparrow$ or $A_n \downarrow$, then $\mathbf{P}(\cup A_n) = \lim_{n\to\infty} \mathbf{P}(A_n)$ or $\mathbf{P}(\cap A_n) = \lim_{n\to\infty} \mathbf{P}(A_n)$. Therefore if two probabilities $\mathbf{P}$ and $\mathbf{P}^*$ coincide on $\mathcal{A}_0$, they also coincide on some monotone collection containing $\mathcal{A}_0$ and consequently also on $\sigma(\mathcal{A}_0)$.

Relation (2.2.1) is not the sole restriction on $\mathbf{P}(E_n^i)$; it ensures additivity on $\mathcal{A}_0$ (the nonnegativity and normalization of $\mathbf{P}$ are understood; the normalization is ensured by the condition $\sum \mathbf{P}(E_1^i) = 1$). In order for $\mathbf{P}$ to be extendable to a σ-additive function on $\sigma(\mathcal{A}_0)$, it is necessary and sufficient that $\mathbf{P}$ be σ-additive on $\mathcal{A}_0$. A necessary condition for this is the following: if $E_n^{i_n}$ is a decreasing sequence for which $\bigcap_n E_n^{i_n} = \emptyset$, then $\lim \mathbf{P}(E_n^{i_n}) = 0$. The fulfillment of this condition makes it possible to extend a measure. It suffices to prove this for the case where $\lim_{n\to\infty} \mathbf{P}(E_n^{i_n}) = 0$ for every decreasing sequence $(E_n^{i_n})$ (a "continuous" measure). This follows because there exist only at most countably many decreasing sequences $E_n^{i_n(k)}, k = 1, 2, \ldots$, such that $\lim_{n\to\infty} \mathbf{P}(E_n^{i_n(k)}) = q_k > 0$ and for which $\bigcap_n E_n^{i_n(k)} = F_k \in \sigma(\mathcal{A}_0)$ is not empty. Each F_k is an atom of $\sigma(\mathcal{A}_0)$ and if $Q_1(A) = \sum I_{\{F_k \subset A\}} q_k$, then clearly Q_1 is a countably-additive measure. Putting $Q_2(A) = \mathbf{P}(A) - Q_1(A), A \in \mathcal{A}_0$, we then obtain a continuous measure. For this measure, one can make use of the interpretation of the E_n^i as intervals in $[0, 1]$ of length $Q_2(E_n^i)$ and the intervals are chosen so that the inclusion relations for the intervals and sets E_n^i coincide.

(b) *Geometrical probabilities.* Geometrical probabilities arose in the attempt to generalize the notion of equal likelihood on which the classical definition of probability is based. They involve choosing a point at random in some geometric figure. If the figure is planar, then it is assumed that the probability

of choosing the point in a given part of the figure equals the quotient of the area of that part of the figure and the area of the entire figure. Very simple illustrations of geometrical probability are the following.

The encounter problem. Two persons agree to meet between 12:00 P.M. and 1:00 P.M. The first to arrive at the meeting place waits 20 minutes. What is the probability of an encounter if the time of arrival of each is chosen at random and independently of the time of arrival of the other person? If x is the fraction of the hour after 12:00 P.M. when the first person arrives and y is the fraction of the second, then a meeting will take place if $|x-y| < 1/3$. We take the sample space to be the square of side 1 with one vertex at the origin and the two sides going from that vertex on the coordinate axes. We identify the pair (x, y) with the point of the square with these coordinates. For the σ-algebra of events, we take the Borel subsets of the square and the probability is Lebesgue measure. The points satisfying the condition $|x-y| < 1/3$ lie between the lines $x - y < 1/3$ and $x - y = -1/3$. The complement of this set consists of the two triangles $x > y + 1/3$ and $y > x + 1/3$ which together form a square of side $2/3$ and area $4/9$. Hence, the probability of a meeting is $5/9$.

Buffon's problem. A needle of length 2ℓ is tossed onto a plane on which parallel lines have been ruled lying a distance $2a$ apart. What is the probability that the needle intersects a line?

We locate the needle by means of the distance x of its midpoint to the closest line and the acute angle φ between that line and the needle, $0 \le x \le a$ and $0 \le \varphi \le \pi/2$. The rectangle determined by these inequalities is the sample space. The needle intersects a line if $x \le l \sin\varphi$. The required probability is the quotient of the area of the figure determined by the three inequalities $0 \le x \le \ell$, $0 \le \varphi \le \pi/2$ and $x \le l \sin\varphi$ and the area of the rectangle $\pi a/2$. Assume that $\ell < a$. Then the area of the figure is $\int_0^{\pi/2} d\varphi \int_0^{l \sin\varphi} dx = \ell$ and so the probability of intersection is $2\ell/\pi a$.

Geometrical probability now also encompasses probability spaces in which some subset of finite-dimensional Euclidean space plays the role of the sample space and Lebesgue measure (with appropriate normalization) is the probability. It should be pointed out that the applications of geometrical probability show that the expression "at random" is meaningless for spaces with infinitely many outcomes. By using different sets for the sample space, one can deduce different values for probabilities. Thus, if we locate a chord in a circle by the position of its midpoint, then the probability that its length exceeds the radius equals $3/4$. But if we specify the position of the chord by one point on the circumference and the angle between the chord and the tangent, then the probability is $2/3$.

2.2.2 Random Variables. Expectation

Random variables are quantities which can be measured in random experiments. This means that the value of the quantity is determined once an experiment has been performed or, in other words, an elementary event has been

chosen. Therefore a random variable is a function of the elementary events (we are considering numerical variables here).

The possibility of making a measurement means that for each interval we can observe an event: the measured quantity assumes a value in that interval. Thus a random variable ξ is a measurable function of the elementary events: $\xi = \xi(\omega)$ and $\{\omega : \xi(\omega) < x\} \in \mathcal{A}$ for all $x \in R$ (the reals). The mapping $\xi : \Omega \to R$ sends the measure $\mathbf{P}$ on $\mathcal{A}$ into some measure μ_ξ defined on the σ-algebra $\mathcal{B}$ of Borel sets of R (the *Borel algebra*); μ_ξ is also a probability measure and it is called the *distribution of* ξ. It is given by its values on the intervals $[a, b[$ and hence it is determined just by specifying the *distribution function* $F_\xi(x) = \mu_\xi(]-\infty, x[) = \mathbf{P}(\{\omega : \xi(\omega) < x)$. A random variable is *discrete* if a countable set S can be specified such that $\mu_\xi(S) = 1$. If $S = \{x_1, x_2, \ldots\}$, then the distribution of ξ is the set of probabilities $p_k = \mathbf{P}(\{\omega : \xi(\omega) = x_k\})$; and $\mu_\xi(B) = \sum p_k I_B(x_k)$ for any $B \in \mathcal{B}$. A *distribution is continuous* if $\mu_\xi(\{x\}) = 0$ for all x. It is called *absolutely continuous* if there exists a measurable function $f_\xi(x) : R \to R$ such that

$$\mu_\xi(B) = \int_B f_\xi(x)dx \ ;$$

$f_\xi(x)$ is termed the (*distribution*) *density* of the random variable ξ.

Let ξ assume finitely many values $x_1, \ldots, x_r$. Let A_k be the event that ξ takes the value x_k. Suppose that n experiments have been performed in which ξ takes the values $\xi_1, \xi_2, \ldots, \xi_n$. Consider the average value of the resulting observations

$$\begin{aligned}\bar{\xi} = \frac{1}{n}(\xi_1 + \ldots + \xi_n) = \frac{1}{n}\Big(x_1 \sum I_{A_1} + x_2 \sum I_{A_2} + \ldots \\ + x_r \sum I_{A_r}\Big) = \frac{m_1}{n}x_1 + \frac{m_2}{n}x_2 + \ldots \\ + \frac{m_r}{n}x_r = \sum_{k=1}^{r} x_k \nu_n(A_k) \ .\end{aligned}$$

Here m_i is the number of occurrences of A_i in the n experiments and $\nu_n(A_i)$ is the relative frequency of A_i. If we replace the relative frequencies on the right-hand side by probabilities, we obtain $\sum x_k \mathbf{P}\{\xi = x_k\}$. It is natural to view this as the stochastic average of the random variable. Then clearly

$$\sum x_k \mathbf{P}\{\xi = x_k\} = \int \xi(\omega)\mathbf{P}(d\omega) = \int \xi d\mathbf{P} \ . \tag{2.2.2}$$

If the integral on the right-hand side of (2.2.2) is defined for a random variable ξ (with arbitrary distribution), then ξ is said to have a (*mathematical*) *expectation*, *mean* or *expected value*. It is denoted by $\mathbf{E}\xi$:

$$\mathbf{E}\xi = \int \xi(\omega)\mathbf{P}(d\omega)\ . \tag{2.2.3}$$

A change of variables in the integral results in the following formula for $\mathbf{E}\xi$ in terms of the distribution function of ξ:

$$\mathbf{E}\xi = \int x\mu_\xi(dx) = \int x d_x F_\xi(x) \tag{2.2.4}$$

(the existence of the expectation implies the existence of the indicated integrals). If ξ is non-negative, then $\mathbf{E}\xi$ is always regarded as defined but it may have the value $+\infty$. Therefore one can talk about random variables with finite expectation. From (2.2.3) it follows that $\mathbf{E}\xi$ is a linear function of a random variable; $\mathbf{E}\xi \geq 0$ if $\xi \geq 0$ and if $\xi \geq 0$ and $\mathbf{E}\xi = 0$, then $\mathbf{P}\{\xi = 0\} = 1$.

(a) *Expectation of a function of a random variable.* Let $g(x)$ be a Borel function from R to R. If $\xi(\omega)$ is a random variable, then so is $\mu(\omega) = g(\xi(\omega))$. On the basis of the formula for a change of variables in an integral,

$$\mathbf{E}g(\xi) = \int g(x)\mu_\xi(dx) = \int g(x) d_x F_\xi(x)$$

if these integrals exist.

To characterize a random variable, use is made of its *moments*

$$\mathbf{E}\xi^k = \int x^k \mu_\xi(dx),$$

where k is positive integer. This is the k-th order moment. If $\mathbf{E}\xi = a$ exists, then the expression

$$\mathbf{E}(\xi - a)^k = \int (x - a)^k \mu_\xi(dx)$$

is the k-th order *central moment*. In particular,

$$\mathbf{V}\xi = \mathbf{E}(\xi - a)^2 = \int (x - a)^2 \mu_\xi(dx)$$

is the *variance* of the random variable ξ.[1] If $\mathbf{V}\xi = 0$, then $\mathbf{P}\{\xi = a\} = 1$.

2.2.3 Conditional Expectation

Let $E_1, E_2, \ldots, E_r$ be a complete group of events. The conditional expectation of a random variable ξ given E_k is defined by

$$\frac{1}{P(E_k)} \int_{E_k} \xi(\omega)\mathbf{P}(d\omega) = \mathbf{E}(\xi | E_k)\ . \tag{2.2.5}$$

[1] It is easy to show that $\mathbf{V}(\xi) = \mathbf{E}\xi^2 - (\mathbf{E}\xi)^2$.

The formula (2.2.5) can be obtained from (2.2.3) if the measure $\mathbf{P}(A)$ is replaced in it by the conditional probability $\mathbf{P}(A|E_k)$. We now consider conditional expectation with respect to $\{E_1, \dots, E_r\}$. Introduce the algebra $\mathcal{E}$, generated by $E_1, \dots, E_r$. We define

$$\mathbf{E}(\xi|\mathcal{E}) = \sum_{k=1}^{r} I_{E_k} \mathbf{E}(\xi|E_k) \, . \tag{2.2.6}$$

We point out two properties of $\mathbf{E}(\xi|\mathcal{E})$.

1. $\mathbf{E}(\xi|\mathcal{E})$ is a random variable which is measurable with respect to $\mathcal{E}$.
2. For any set $B \in \mathcal{E}$,
$$\int_B \mathbf{E}(\xi|\mathcal{E}) d\mathbf{P} = \int_B \xi d\mathbf{P} \, . \tag{2.2.7}$$

The first assertion follows since $\mathbf{E}(\xi|\mathcal{E})$ is constant on the atoms of $\mathcal{E}$. The second assertion holds for the sets E_k, which results from (2.2.5), and so holds also for their unions which comprise $\mathcal{E}$. These two properties determine $\mathbf{E}(\xi|\mathcal{E})$ uniquely. By the first property, this variable is constant on the atoms of $\mathcal{E}$, that is, on each E_k. This constant is determined by (2.2.7): if $c_k = \mathbf{E}(\xi|\mathcal{E})$ for $\omega \in E_k$, then $c_k \mathbf{P}(E_k) = \int_{E_k} \xi d\mathbf{P}$, where we replaced B by E_k in (2.2.7). This makes it possible to extend *conditional expectation* to the case where $\mathcal{E}$ is an arbitrary σ-algebra.

Definition. Let $\mathcal{E} \subset \mathcal{A}$ be a σ-algebra and ξ a random variable for which $\mathbf{E}\xi \leq \infty$. Then $\mathbf{E}(\xi|\mathcal{E})$ is called the conditional expectation of ξ with respect to the σ-algebra $\mathcal{E}$ if conditions 1 and 2 hold. If $A \in \mathcal{A}$, then $\mathbf{E}(I_A|\mathcal{E})$ is the conditional probability of A with respect to $\mathcal{E}$ and it is denoted by $\mathbf{P}(A|\mathcal{E})$.

Properties 1 and 2 determine a conditional expectation uniquely up to sets of measure 0, that is, if $\eta_1 = \mathbf{E}(\xi|\mathcal{E})$ and $\eta_2 = \mathbf{E}(\xi|\mathcal{E})$, then $\mathbf{P}\{\eta_1 = \eta_2\} = 1$. Indeed, $\{\omega : \eta_1 - \eta_2 > 0\}$ belongs to $\mathcal{E}$ and hence by 2,

$$\int (\eta_1 - \eta_2) I_{\{\eta_1 - \eta_2 > 0\}} d\mathbf{P} = \int \xi I_{\{\eta_1 - \eta_2 > 0\}} d\mathbf{P} - \int \xi I_{\{\eta_1 - \eta_2 > 0\}} d\mathbf{P} = 0 \, .$$

In similar fashion, $\int (\eta_1 - \eta_2) I_{\{\eta_2 - \eta_1 > 0\}} d\mathbf{P} = 0$ implies $\int |\eta_1 - \eta_2| d\mathbf{P} = 0$.

We now show that a conditional expectation exists. Consider the countably-additive set function $\mathbf{Q}(E) = \int_E \xi d\mathbf{P}$ on $\mathcal{E}$. It is clearly absolutely continuous with respect to the measure $\mathbf{P}$ considered on $\mathcal{E}$. Therefore by the Radon-Nikodym theorem, the density of $\mathbf{Q}$ with respect to $\mathbf{P}$ exists, that is, an $\mathcal{E}$-measurable function $q(\omega)$ exists such that $\mathbf{Q}(B) = \int_B q(\omega) \mathbf{P}(d\omega)$. It is easy to see that this last relation is simply (2.2.7).

We now state the main properties of conditional expectation. Since it is a random variable and is determined up to sets of measure 0, we emphasize that all of the equalities (and inequalities) below are understood to hold with probability 1.

I. Conditional expectation is an additive function of random variables. This means the following. Let $\xi_n = \xi_n(\omega)$ be a finite sequence of random variables. Then

$$\mathbf{E}\left(\sum \xi_n|\mathcal{E}\right) = \sum \mathbf{E}(\xi_n|\mathcal{E}) \ . \tag{2.2.8}$$

Proving this amounts to showing that (2.2.7) holds for $\eta = \sum \xi_n$ if $\mathbf{E}(\eta|\mathcal{E})$ is replaced by the quantity on the right-hand side of (2.2.8).

II. Let η be $\mathcal{E}$-measurable and let ξ be such that $\mathbf{E}|\xi\eta| < \infty$ and $\mathbf{E}|\xi| < \infty$. Then

$$\mathbf{E}(\xi\eta \mid \mathcal{E}) = \eta\mathbf{E}(\xi \mid \mathcal{E}) \ .$$

It is necessary to show that

$$\int_B \xi\eta d\mathbf{P} = \int_B \eta\mathbf{E}(\xi|\mathcal{E})d\mathbf{P} \tag{2.2.9}$$

for every $B \in \mathcal{E}$. Let $\eta = I_C$ with $C \in \mathcal{E}$. Then the preceding relation may be written as

$$\int_{B\cap C} \xi d\mathbf{P} = \int_{B\cap C} \mathbf{E}(\xi|\mathcal{E})d\mathbf{P} \ .$$

Thus (2.2.2) holds for η assuming finitely many values. From this it is easy to deduce this relation for all η for which one of the sides of the equality is meaningful.

III. Formula for iterated expectations. Given two σ-algebras $\mathcal{E} \subset \mathcal{F} \subset \mathcal{A}$. Then

$$\mathbf{E}(\xi|\mathcal{E}) = \mathbf{E}(\mathbf{E}(\xi|\mathcal{F})|\mathcal{E}) \ . \tag{2.2.10}$$

Let $E \in \mathcal{E}$. Then

$$\int_E \mathbf{E}(\mathbf{E}(\xi|\mathcal{F})|\mathcal{E})d\mathbf{P} = \int_E \mathbf{E}(\xi|\mathcal{F})d\mathbf{P} = \int_E \xi d\mathbf{P}$$

(the fact that $E \in \mathcal{E} \subset \mathcal{F}$ was used in the last relation). This shows that the right-hand side of (2.3.10) satisfies property 2.

Let ζ be a random variable. Let $\mathcal{B}_\zeta$ be the σ-algebra of sets of the form $\{\omega : \zeta \in B\}$ with $B \in \mathcal{B}$ ($\mathcal{B}$ is the σ-algebra of Borel sets on the line). The conditional expectation with respect to $\mathcal{B}_\zeta$ must be measurable with respect to $\mathcal{B}_\zeta$ and this signifies that it is a Borel function of ζ. We shall denote it by $\mathbf{E}(\xi|\zeta)$. Similarly, if $\{\zeta_\lambda, \lambda \in \Lambda\}$ is a family of random variables and $\sigma\{\zeta_\lambda, \lambda \in \Lambda\}$ is the smallest σ-algebra with respect to which the variables ζ_λ are measurable, then the conditional expectation with respect to this σ-algebra can be denoted by $\mathbf{E}(\xi|\zeta_\lambda, \lambda \in \Lambda)$. For the conditional probabilities, we shall use the notation $\mathbf{P}(A|\zeta)$ and $\mathbf{P}(A|\zeta_\lambda, \lambda \in \Lambda)$.

2.2.4 Regular Conditional Distributions

Let $\mathcal{A}_0 \subset \mathcal{A}$ be a countable algebra. Since $\mathbf{P}(A_1 \cup A_2|\mathcal{E}) = \mathbf{P}(A_1|\mathcal{E}) + \mathbf{P}(A_2|\mathcal{E})$ for all A_1 and $A_2 \in \mathcal{A}_0$ if $A_1 \cap A_2 = \emptyset$ (the equality holds with probability 1), a $C \in \mathcal{A}$ can be specified such that $\mathbf{P}(C) = 0$ so that this equality holds for all $\omega \notin C$. Therefore using the countability of $\mathcal{A}_0$ we can say that there is a subset $U \in \mathcal{A}$ such that $\mathbf{P}(U) = 1$ and for all $\omega \in U$ the function $\mathbf{P}(A|\mathcal{E})$ is additive with respect to A on $\mathcal{A}_0$ (for each A it is a function of ω).

Suppose that there exists a function $p(A, \omega)$ satisfying: 1. $p(A, \omega)$ is a measure in A on $\mathcal{A}$; 2. $p(A, \omega)$ is $\mathcal{E}$-measurable for all $A \in \mathcal{A}$; 3. $\mathbf{P}(A|\mathcal{E}) = p(A, \omega)$ (with probability 1). Then $p(A, \omega)$ is called *regular conditional probability.* Examples show that the regular conditional probability does not exist in general. At the same time, it is possible to form a function $p(A, \omega)$ on $\mathcal{A}_0$ which is additive for each ω and coincides with a conditional probability ($p(A, \omega)$ may be specified arbitrarily for $\omega \in \Omega \setminus U$, provided there is additivity and $\mathcal{E}$-measurability). Therefore it is apparently impossible to construct a regular conditional probability due to $\mathcal{A}$ containing too many sets. But if $\mathcal{A}_0$ is such that each additive function on $\mathcal{A}_0$ can be extended to a countably-additive one on $\sigma(\mathcal{A}_0)$, then $p(A, \omega)$ will be countably-additive on $\sigma(\mathcal{A}_0)$. The simplest example of such an $\mathcal{A}_0$ is the algebra generated by the union of countably many finite covers of a compact set by spheres of radius ε_k, $k = 1, 2, \ldots$, with $\varepsilon_k \to 0$. If ν is a given additive function on such an $\mathcal{A}_0$, then $\int \varphi d\nu$ is defined for every continuous function φ and due to the form of the linear functional, this integral must be an integral with respect to a countably-additive function.

Let X be a complete separable metric space. Every measure μ on the σ-algebra $\mathcal{B}_X$ of Borel sets has the following property: $\forall \varepsilon > 0$, there exists a compact set $\mathcal{K}_\varepsilon$ such that $\mu(X \setminus \mathcal{K}_\varepsilon) < \varepsilon$. Let $x(\omega)$ be the meaurable mapping of Ω into X : $\{\omega : x(\omega) \in B\} \in \mathcal{A}$ for $B \subset \mathcal{B}_X$. Let $\mu_x(B)$ denote the measure into which $x(\omega)$ maps the measure $\mathbf{P}$: $\mu_x(B) = \mathbf{P}(\{\omega : x(\omega) \in B\})$. The mapping $x(\omega)$ is called a *random element* in X (a random element in R is merely a random variable) and $\mu_x(B)$ is its distribution. Let $\mathcal{B}^x$ be the σ-algebra of subsets of $\mathcal{A}$ of the form $\{\omega : x(\omega) \in B\}$ with $B \in \mathcal{B}_X$. The conditional probability $\mathbf{P}(C|\mathcal{E})$ considered on $\mathcal{B}^x$ is mapped by $x(\omega)$ into a function $\mu_x(B|\mathcal{E})$, $B \in \mathcal{B}_X$, which will be called the conditional distribution of $x(\omega)$. It clearly determines a regular conditional distribution.

Theorem. *A random element in a complete separable metric space has a regular conditional distribution.*

The proof of this rests on the following assertions. 1. An increasing sequence of compact sets $\mathbf{K}_n$ can be specified so that $\mu_x(\mathbf{K}_n) \uparrow 1$; then there is a $U \in \mathcal{A}$ with $\mathbf{P}(U) = 1$ such that $\mu_x(\mathbf{K}_n|\mathcal{E}) \uparrow 1$ for all $\omega \in U$. 2. For each $\mathbf{K}_n$, it is possible to specify a countable algebra of its subsets $\mathcal{A}_n$ and a $U_n \in \mathcal{A}$ such that $\mathbf{P}(U_n) = 1$ and $\mu_x(B|\mathcal{E})$ is additive on $\mathcal{A}_n$ for $\omega \in U_n$. 3. $\mathcal{A}_n$ may be chosen so that every additive function on it can be extended in a unique way to a countably-additive function on $\mathcal{B}_{\mathbf{K}_n}$, the σ-algebra of Borel subsets

of $\mathbf{K}_n$. 4. $\mathcal{A}_n$ may always be chosen to be increasing with n. Then $\mu_x(B|\mathcal{E})$ will be countably-additive on $\mathcal{B}_X$ for $\omega \in (\bigcap_n U_n) \cap U$, $\mathbf{P}((\bigcap_n U_n) \cap U) = 1$. For the remaining $\omega, \mu_x(B|\mathcal{E})$ may be chosen to coincide with μ_x.

2.2.5 Spaces of Random Variables. Convergence

Consider the space of real random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$, which we denote by $R(\Omega)$. This is a linear space. A sequence of random variables ξ_n is said to *converge to a random variable* ξ *in probability* if $\lim_{n\to\infty} \mathbf{P}\{|\xi_n - \xi| > \varepsilon\} = 0$ for any positive ε.

Convergence in probability is equivalent to the convergence of $1 - \mathbf{E}e^{-|\xi_n - \xi|}$ to zero. To show this, we need an important inequality.

(a) *Chebyshev's inequality.* If $\xi \geq 0$ and $\mathbf{E}\xi < \infty$, then $\mathbf{P}\{\xi > a\} \leq \frac{1}{a}\mathbf{E}\xi$ *for* $a > 0$.

Proof. Since $\xi \geq aI_{\{\xi > a\}}$, we have $\mathbf{E}\xi \geq \mathbf{E}aI_{\{\xi>a\}} = a\mathbf{P}\{\xi > a\}$. □

Put $r_{\mathbf{P}}(\xi, \eta) = 1 - \mathbf{E}\exp\{-|\xi - \eta|\}$. Then by Chebyshev's inequality,

$$\begin{aligned}\mathbf{P}\{|\xi - \eta| > a\} &= \mathbf{P}\{1 - e^{-|\xi-\eta|} > 1 - e^{-a}\} \\ &\leq r_{\mathbf{P}}(\xi, \eta)(1 - e^{-a})^{-1}\end{aligned}$$

On the other hand, if $0 < \varepsilon < 1$,

$$r_{\mathbf{P}}(\xi, \eta) \leq \varepsilon + \mathbf{E}I_{\{1-\exp\{-|\xi-\eta|\}>\varepsilon\}} \leq \varepsilon + \mathbf{P}\{|\xi - \eta| > \ln\frac{1}{1-\varepsilon}\}.$$

Therefore if ξ_n converges to ξ in probability, $\overline{\lim}_{n\to\infty} r_{\mathbf{P}}(\xi_n, \xi) \leq \varepsilon$ for any positive ε. But if $r_{\mathbf{P}}(\xi_n, \xi) \to 0$, then $\overline{\lim}_{n\to\infty}\mathbf{P}\{|\xi_n - \xi| > \varepsilon\} \leq \overline{\lim}_{n\to\infty}$ $r_{\mathbf{P}}(\xi_n, \xi)(1 - e^{-\varepsilon})^{-1}$ for any positive ξ.

The quantity $r_{\mathbf{P}}(\xi, \eta)$ satisfies the triangle inequality. If ξ, η, and ζ are any three random variables, then

$$r_{\mathbf{P}}(\xi, \eta) \leq r_{\mathbf{P}}(\xi, \zeta) + r_{\mathbf{P}}(\zeta, \eta).$$

Random variables that coincide almost everywhere with respect to a measure $\mathbf{P}$ are identified (properties that hold almost everywhere with respect to $\mathbf{P}$ are said to hold almost surely). Since this identification will be assumed throughout the sequel, we shall retain the old notation for random variables so determined and for the set of all random variables. Therefore $r_{\mathbf{P}}(\xi, \eta)$ is a metric on $R(\Omega)$ and $R(\Omega)$ is complete in this metric. We shall discuss completeness a little bit below.

A sequence of random variables ξ_n *converges almost surely or with probability* 1 to a random variable ξ if there exists a $U \in \mathcal{A}$ such that $\xi_n \to \xi$ for all $\omega \in U$ with $\mathbf{P}(U) = 1$. The set of ω for which $\xi_n \to \xi$ can be written as

$$\bigcap_k \bigcup_m \bigcap_{n\geq m} \{|\xi_n - \xi| \leq 1/k\} \,. \tag{2.2.11}$$

This set belongs to $\mathcal{A}$ and if the probability of this set is 1, then $\xi_n \to \xi$ with probability 1. If $\xi_n \to \xi$ with probability 1, then $\lim_{m\to\infty} \mathbf{P}\left(\bigcap_{n\geq m}\{|\xi_n - \xi| \leq 1/k\}\right) = 1$ for all k and hence $\lim_{m\to\infty} \mathbf{P}\{|\xi_m - \xi| > 1/k\} = 0$, that is, $\xi_n \to \xi$ in probability. If ξ_n is a sequence such that $\sum r_{\mathbf{P}}(\xi_n, \xi) < \infty$, then $\xi_n \to \xi$ with probability 1. To prove that the probability of (2.2.11) is 1, it suffices to show that the probability of the complement of this set is zero or that for all k,

$$\mathbf{P}\left(\bigcap_m \bigcup_{n\geq m} \left\{|\xi_n - \xi| > \frac{1}{k}\right\}\right) = \lim_{m\to\infty} \mathbf{P}\left(\bigcup_{n\geq m} \left\{|\xi_n - \xi| > \frac{1}{k}\right\}\right) = 0 \,. \tag{2.2.12}$$

But

$$\begin{aligned} \mathbf{P}\left(\bigcup_{n\geq m} \{|\xi_n - \xi| > 1/k\}\right) &\leq \sum_{n\geq m} \mathbf{P}\{|\xi_n - \xi| > 1/k\} \\ &\leq \sum_{n\geq m} r_{\mathbf{P}}(\xi_n, \xi) \left(1 - e^{-\frac{1}{k}}\right)^{-1} \end{aligned}$$

and the right-hand side tends to zero as $m \to \infty$. Thus we have the following.

Theorem 2.2.1. *If $r_{\mathbf{P}}(\xi_n, \xi) \to 0$, then there is a sequence n_k such that $\xi_{n_k} \to \xi$ with probability 1.*

The next theorem establishes a connection between convergence in probability and convergence with probability 1.

Theorem 2.2.2. *A sequence ξ_n converges to ξ in probability if and only if every subsequence ξ_{n_k} contains a subsequence $\xi_{n_{k_m}}$ converging to ξ with probability* 1*.*

If ξ_n converges to ξ in probability, then the necessity follows by Theorem 2.2.1. If ξ_n did not converge to ξ in probability, there would be a positive ε and a subsequence ξ_{n_k} such that $r_{\mathbf{P}}(\xi_{n_k}, \xi) \geq \varepsilon$. No sequence could be extracted from the subsequence ξ_{n_k} converging to ξ in probability (and hence with probability 1). □

Let $\xi_n^i, i = 1, \ldots, r$, be sequences of random variables converging to ξ^i respectively in probability and let $\Phi(x_1, x_2, \ldots, x_r)$ be a continuous function from R^r to R. Then $\Phi(\xi_n^1, \ldots, \xi_n^r)$ converges to $\Phi(\xi^1, \ldots, \xi^r)$ in probability. This assertion follows from Theorem 2.2.2 and the fact that $\xi_{n_k}^i \to \xi, i = 1, \ldots, r$, implies that $\Phi(\xi_{n_k}^1, \ldots, \xi_{n_k}^r) \to \Phi(\xi^1, \ldots, \xi^r)$.

A sequence ξ_n is *fundamental* with probability 1 if

$$\mathbf{P}\left(\bigcap_k \bigcup_l \bigcap_{n,m\geq l} \left\{|\xi_n - \xi_m| \leq \frac{1}{k}\right\}\right) = 1\,. \tag{2.2.13}$$

The sequence $\xi_n(\omega)$ is fundamental for all ω belonging to the set under the probability sign and hence it has a limit. Therefore $\lim_{n\to\infty} \xi_n(\omega)$ exists for almost all ω and the limit is a random variable.

Now let $r_{\mathbf{P}}(\xi_n, \xi_m) \to 0$. Choose a sequence n_k so that $r_{\mathbf{P}}(\xi_{n_k}, \xi_{n_{k+1}})(1 - e^{-1/k^2})^{-1} \leq 1/k^2$. Then $\mathbf{P}\{|\xi_{n_k} - \xi_{n_{k+1}}| \leq 1/k^2\} \leq 1/k^2$. Write $U = \bigcup_m \bigcap_{k\geq m}\{|\xi_{n_k} - \xi_{n_{k+1}}| \leq k^{-2}\}$. Then $\mathbf{P}(U) = 1$ since

$$\begin{aligned}\mathbf{P}(\Omega \setminus U) &= \lim_{m\to\infty} \mathbf{P}\left(\bigcup_{k\geq m} \left\{|\xi_{n_k} - \xi_{n_{k+1}}| > \frac{1}{k^2}\right\}\right)\\ &\leq \lim_{m\to\infty} \sum_{k=m}^{\infty} \mathbf{P}\left\{|\xi_{n_k} - \xi_{n_{k+1}}| > \frac{1}{k^2}\right\} = 0\,.\end{aligned}$$

For $\omega \in U$,

$$\overline{\lim}_{k\to\infty} |\xi_{n_k} - \xi_{n_{k+1}}| k^2 \leq 1$$

and so the series $\xi_{n_1} + \xi_{n_2} - \xi_{n_1} + \ldots + \xi_{n_{k+1}} - \xi_{n_k} + ..$ converges absolutely, that is, $\lim_{k\to\infty} \xi_{n_k} = \xi$ exists. But then

$$\overline{\lim}_{n\to\infty} r_{\mathbf{P}}(\xi_n, \xi) \leq \overline{\lim}_{n\to\infty} r_{\mathbf{P}}(\xi_n, \xi_{n_k}) + r_{\mathbf{P}}(\xi_{n_k}, \xi).$$

Letting $k \to \infty$, one can see that $r_{\mathbf{P}}(\xi_n, \xi) \to 0$.

(b) *Passage to the limit under the expectation sign.* A sequence ξ_n is *uniformly integrable* if

$$\lim_{\alpha\to\infty} \sup_n \mathbf{E}|\xi_n| I_{\{|\xi_n|>\alpha\}} = 0\,. \tag{2.2.14}$$

Theorem 2.2.3. *Let ξ_n converge to ξ in probability.* 1. *If ξ_n is uniformly integrable, then $\mathbf{E}|\xi| < \infty$ and $\lim_{n\to\infty} \mathbf{E}\xi_n = \mathbf{E}\xi$.* 2. *If $\xi_n \geq 0, \mathbf{E}\xi < \infty$ and $\lim_{n\to\infty} \mathbf{E}\xi_n = \mathbf{E}\xi$, then ξ_n is uniformly integrable.*

Proof. 1. Let $g_a(x) = -a$ for $x < -a$, $g_a(x) = x$ for $|x| \leq a$ and $g_a(x) = a$ for $x > a$. By Lebesgue's dominated convergence theorem, $\lim_{n\to\infty} \mathbf{E}_{g_a}(\xi_n) = \mathbf{E}_{g_a}(\xi)$. By the uniform integrability, it follows that

$$\sup_n \mathbf{E}|g_a(\xi_n) - \xi_n| \leq \sup_n \mathbf{E}|\xi_n| I_{\{|\xi_n|>a\}}\ ,$$

and this quantity may be made arbitrarily small by the choice of a. 2. Let us show that a may be chosen so that $\mathbf{E}\xi_n I_{\{\xi_n>a\}} < \varepsilon$ for all n and any

positive ε. To this end, it suffices that $\mathbf{E}f_a(\xi_n) < \varepsilon$, where $f_a(x) = 0$ for $x < a/2$, $f_a(x) = 2x - a$ for $a/2 \le x \le a$ and $f_a(x) = x$ for $x > a$. Since f_a is continuous, $f_a(\xi_n) \to f_a(\xi)$ in probability and $\xi_n - f_a(\xi_n) \to \xi - f_a(\xi)$ in probability. But $x - f_a(x)$ is a bounded function and by Lebesgue's theorem $\lim_{n\to\infty} \mathbf{E}(\xi_n - f_a(\xi_n)) = \mathbf{E}(\xi - f_a(\xi))$. Now $fa(\xi) \to 0$ as $a \to \infty$ and $f_a(\xi) \le \xi$ and so again by the same theorem,

$$\lim_{a\to\infty} \mathbf{E}f_a(\xi) = 0 \,.$$

Now choose a_1 so that $\mathbf{E}f_{a_1}(\xi) < \varepsilon/2$. Since $\mathbf{E}\xi_n \to \mathbf{E}\xi$, we have $\lim_{n\to\infty} \mathbf{E}f_a(\xi_n) = \mathbf{E}f_a(\xi)$ and there is an n_1 such that $\mathbf{E}f_{a_1}(\xi_n) < \varepsilon$ for $n > n_1$. Letting a increase, one may also deduce for a finite collection $\xi_1, \ldots, \xi_{n_1}$, that $\mathbf{E}\xi_k I_{\{\xi_k > a\}} < \varepsilon$ for $k \le n_1$. □

In addition to the space with convergence in probability discussed above, we shall also utilize the space $\mathbf{L}_p(\Omega, \mathbf{P})$ of functions $\xi(\omega)$ for which $\int |\xi|^p d\mathbf{P} < \infty$ with norm $||\xi||_p = (\int |\xi|^p d\mathbf{P})^{1/p}$.

Of special interest is $\mathbf{L}_2(\Omega, \mathbf{P})$, which is a Hilbert space with the inner product $< \xi, \eta >= \mathbf{E}\xi\eta$.

2.3 Random Mappings

2.3.1 Random Elements

Let $(X, \mathcal{B})$ be a measurable space. Consider a *measurable mapping* $x = f(\omega)$ from $(\Omega, \mathcal{A})$ to $(X, \mathcal{B})$ or, in other words, a mapping such that $f^{-1}(B) = \{\omega : f(\omega) \in B\} \in \mathcal{A}$ for all $B \in \mathcal{B}$. It is called a *random element* in X; X is the phase space of the element, $\{x(\omega) \in B\}$, $B \in \mathcal{B}$, is the σ-algebra generated by $x(\omega)$, which we shall denote by $\sigma(x(\omega))$, and the measure $\mu_x(B) = \mathbf{P}(\{\omega : x(\omega) \in B\})$ on $\mathcal{B}$ is the distribution of $x(\omega)$.

If we are just examining a random element $x(\omega)$, then it is natural to deal with the image of the original probability space under the mapping $x(\omega)$. It will also be a probability space $(X, \mathcal{B}, \mu_x)$, where X serves as the sample space, $\mathcal{B}$ is the σ-algebra of events and μ_x is the probability. This new probability space describes an experiment involving the "measurement of $x(\omega)$."

(a) *Random variables.* If X is the real line R and $\mathcal{B} = \mathcal{B}_R$ is the Borel σ-algebra of R, then $x(\omega)$ is a numerical random variable. If we consider this variable by itself, we can view the random experiment as measuring the variable. It is then described by the space $\{R, \mathcal{B}_R, \mu_x\}$, where μ_x is some probability measure on the real line. It is customary to specify it by a *distribution function*

$$F_x(t) = \mu_x(]-\infty, t[) = \mathbf{P}(\{x(\omega) < t\}).$$

This function determines the distribution μ_x uniquely. If two measures coincide on intervals of the form $]-\infty, t[$, then they coincide on the algebra generated by these intervals and thus on the smallest σ-algebra containing all such intervals, that is, on $\mathcal{B}_R$. We now give examples of distributions that occur frequently in probability theory and its applications.

A *discrete distribution* is a measure concentrated on at most a countable set.

1. *The binomial distribution.* Let $0 < p < 1$. Then

$$\mu_x(A) = \sum_{k=0}^{n} I_A(k) \binom{n}{k} p^k (1-p)^{n-k}.$$

2. *The Poisson distribution*: $a > 0, \mu_x(A) = \sum_{k=0}^{n} I_A(k) a^k e^{-a}/k!$.

These distributions were encountered earlier in independent trials. The first is the distribution of the number of occurrences of an event in n independent trials having probability p in each trial. The second occurred in the Poisson limit theorem.

We give one further example involving independent trials. Assume that the trials are continued as long as the event A does not occur. The probability that A happens for the first time in the n-th trial is $(1-p)^{n-1}p$.

3. *The geometric distribution.* Let $0 < p < 1$. Then

$$\mu_x(A) = \sum_{k=1}^{\infty} p(1-p)^k I_A(k) \ .$$

Continuous distributions. These are the distributions with no atoms. In other words, no $A \in \mathcal{B}$ exists with $\mu_x(A) > 0$ such that either $\mu_X(C) = 0$ or $\mu_x(C) = \mu_x(A)$ for all $C \in \mathcal{B}, C \subset A$. The corresponding distribution functions are continuous. If a distribution is absolutely continuous with respect to Lebesgue measure, that is,

$$\mu_x(A) = \int_A f(y)dy,$$

then $f(y)$ is called the density of the distribution.

4. *The uniform distribution* on a set $C \in \mathcal{B}_R$ has the density $I_C(y)/m(C)$ where m is Lebesgue measure. The distribution itself is $\mu_x(A) = m(C \cap A)/m(C)$ and it is assumed that $0 < m(C) < \infty$. If $C = [a, b]$, then the uniform distribution on $[a, b]$ has the density $I_{[a,b]}(x)/(b-a)$.

5. *The exponential distribution.* This is the distribution with density $f(t) = \lambda e^{-\lambda t} I_{t>0}$; $\lambda > 0$ is the parameter of the exponential distribution. The exponential distribution occurs in experiments involving the observation of rare events to which Poisson's theorem is applicable. The parameter a in Poisson's theorem (Sect. 2.1.3(c)) is generally proportional to time: $a = \lambda t$. The probability that a rare event has not occurred up to time t is $e^{-a} = e^{-\lambda t}$. Therefore

if τ is the time when the required event first happens, $\mathbf{P}\{\tau \geq t\} = e^{-\lambda t}$ (the left-hand side is the probability that the event has not occurred up to time t). The distribution function of τ is $F_\tau(t) = (1 - e^{-\lambda t})I_{\{t>0\}}$.

The exponential distribution is always encountered when one considers the "*delay*" in some event. Suppose that a system spontaneously changes its state abruptly (say, a neutron is converted into a proton). Let $g(t)$ be the probability that the state of the system did not change during the time t. Let A_t be the event that the state of the system did not change during the time t. Then the conditional probability of A_{t+s} (that the state is unchanged a further time s after it was unchanged up to time t) given A_t must equal simply the probability that the system did not change state during time s. That is,

$$g(s) = \mathbf{P}(A_{t+s}|A_t) = \frac{P(A_{t+s} \cap A_t)}{P(A_t)} = \frac{P(A_{t+s})}{P(A_t)} = \frac{g(t+s)}{g(t)},$$

$$g(t+s) = g(t)g(s).$$

Since $0 \leq g \leq 1$, it follows that g is montone nonincreasing; $g(t+) = g(t)g(0+)$ and so $g(0+)$ is 0 or 1. In the first case, $g(t) = 0$ for all positive t. In the second case, $g(t) > 0$ for all positive t since $g(2t) = g^2(t)$. Consequently, $\ln g(t)$ is a monotone additive function; $\ln g(t) = -\lambda t, \lambda \geq 0$, or $g(t) = e^{-\lambda t}$.

6. *The normal(Gaussian) distribution.* The DeMoivre-Laplace theorem involved the function $\varphi(x) = (2\pi)^{-1/2}e^{-x^2/2}$. It is the density of a distribution: $\varphi(x) > 0$ and $\int \varphi(x)dx = 1$. Similarly, the function

$$g(a, b, x) = (2\pi b)^{-1/2} \exp\left\{-\frac{(x-a)^2}{2b}\right\} \tag{2.3.1}$$

is also the density of a distribution. The distribution with density (2.3.1) is called the *normal (Gaussian)* distribution and its parameters a and b are expressed in terms of the density by the formulas

$$a = \int xg(a, b, x)dx, \quad b = \int (x-a)^2 g(a, b, x)dx\,. \tag{2.3.2}$$

The formulas (2.3.2) have the following interpretation. Let ξ be a random variable having a distribution with density $g(a, b, x)$. Then $a = \mathbf{E}\xi$ and $b = \mathbf{E}(\xi - a)^2 = \mathbf{V}\xi$. Thus a is the expectation of a random variable with density $g(a, b, x)$ and b is its variance.

(b) *Random vectors.* Now let $X = R^n$ (that is, the space of vectors $(x^1, x^2, \ldots x^n)$). Then a mapping $x(\omega)$ defines a *random vector* $(\xi^1(\omega), \ldots, \xi^n(\omega))$. Giving a random vector is equivalent to specifying its n component real random variables. The distribution of a random vector is a probability measure on the σ-algebra of Borel sets of R^n. It is also called the

joint distribution of the random variables $\xi^1(\omega), \ldots, \xi^n(\omega)$ and it is completely determined by the joint distribution function of $\xi^1(\omega), \ldots, \xi^n(\omega)$,

$$F_{\xi^1,\ldots,\xi^n}(t_1, \ldots, t_n) = \mathbf{P}\left(\bigcap_{k=1}^{n} \{\omega : \xi^k(\omega) < t_k\}\right) . \tag{2.3.3}$$

Sometimes it is more advantageous to view X without a fixed basis (simply as n-dimensional Euclidean space). A random vector can be prescribed by giving its coordinates in some basis and the coordinates in any other basis may be found by the familiar formulas. Distributions in R^n are also called n-dimensional distributions. If a distribution is absolutely continuous with respect to Lebesgue measure in R^n, then its density is called the density of the distribution.

1. *Uniform distribution in C.* Let m be Lebesgue measure in R^n. If C is a Borel set in R^n with $0 < m(C) < \infty$, then

$$\mu(B) = \frac{m(B \cap C)}{m(C)}$$

is the uniform distribution in C. It has the density $\varphi(x) = I_C(x)/m(C), x \in R^n$.

2. *Normal (Gaussian) distribution in R^n.* Let $a \in R^n$, B be an n-th order positive symmetric matrix, $\det B$ be its determinant and B^{-1} be its inverse. The distribution with density

$$g_n(a, B, x) = ((2\pi)^n \det B)^{-1/2} \exp\left\{-\frac{1}{2}((x-a)B^{-1}, x-a)\right\} \tag{2.3.4}$$

is the n-dimensional normal or Gaussian distribution. If $\tilde{b}_{ij}$ are the elements of B^{-1} and $a = (a^1, \ldots, a^n)$, then the argument of the exponential function in (2.3.4) is $-\frac{1}{2}\sum_{i,j=1}^{n} \tilde{b}_{ij}(x^i - a^i)(x^j - a^j)$. In addition,

$$a^i = \int x^i g_n(a, B, x)dx ,$$

while the elements b_{ij}, of B are given by

$$b_{ij} = \int (x^i - a^i)(x^j - a^j) g_n(a, B, x)dx .$$

If $(\xi^1, \xi^2, \ldots, \xi^n)$ is a random vector having density (2.3.4), then $a^i = \mathbf{E}\xi^i$ and

$$b_{ij} = \mathbf{E}(\xi^i - \mathbf{E}\xi^i)(\xi^j - \mathbf{E}\xi^j) . \tag{2.3.5}$$

Let $(\xi^1, \xi^2, \ldots, \xi^n)$ be any vector satisfying $\sum_1^n \mathbf{E}(\xi^i)^2 < \infty$. The vector $\mathbf{E}\xi = \mathbf{a} = (\mathbf{E}\xi^1, \ldots, \mathbf{E}\xi^n)$ is its expectation (mean) and the linear operator

acting in R^n by way of the matrix (b_{ij}), where b_{ij} is given by (2.3.5), is its *covariance*.

Let $\xi(\omega)$ be a random vector in n-dimensional Euclidean space R (no basis has been selected in it) with $\mathbf{E}|\xi(\omega)|^2 < \infty$. Then $\mathbf{E}\xi(\omega)$ can be found from the relation $\mathbf{E}(\xi(\omega), z) = (\mathbf{E}\xi(\omega), z)$, which must hold for all $z \in X$. Similarly, the symmetric operator B can be found using the relation $(Bz, u) = \mathbf{E}(\xi(\omega) - \mathbf{E}\xi(\omega), z)(\xi(\omega) - \mathbf{E}\xi(\omega), u)$ with u and $z \in X$ (the right-hand side is a bilinear form and so this operator exists).

2.3.2 Random Functions

Let Θ be a parameter set (space) and let $(X, \mathcal{B})$ be a measurable space (phase space). A *random function* with domain Θ and phase space $(X, \mathcal{B})$ is a family of mappings $x(\theta, \omega)$ of the probability space $(\Omega, \mathcal{A}, \mathbf{P})$ into $(X, \mathcal{B})$ defined for all $\theta \in \Theta$. In other words, it is a function $x(\theta, \omega) : \Theta \times \Omega \to X$, with $x(\theta, \omega) : (\Omega, A) \to (X, B)$ a measurable mapping for all $\theta \in \Theta$. If Θ is a set on the real line, then the parameter θ is treated as time and the random function is then called a random or stochastic process. Commonly considered are real-valued functions ($X = R$), complex-valued functions (X is the complex plane Z) and vector spaces ($X = R^n$ or $X = Z^n$). If $\Theta = R^n$, then the further term random field is used.

(a) *Finite-dimensional distributions.* One of the basic probability characteristics of a random function are its *finite-dimensional distribution functions*

$$F_{\theta_1,\dots\theta_n}(B_1, \dots B_n) = \mathbf{P}\left(\bigcap_{k=1}^{n} \{\omega : x(\theta_k, \omega) \in B_k\}\right) \tag{2.3.6}$$

defined for all $n \geq l$, $\theta_k \in \Theta$, $k \leq n$ and $B_k \in \mathcal{B}$, $k \leq n$. If we know the finite-dimensional distribution functions, we can make judgments about the joint behavior of the values of the random function on any finite subset of the parameter space (and so also any denumerable one). In principle, this makes it possible to know "everything" about the random function. A more precise meaning of this statement will be discussed in Sect. 2.4 and later in Chap. 4. For given n and $\theta_1, \dots, \theta_n$, the functions (2.3.6) determine a measure on $(X^n, \mathcal{B}^n)$ – the joint distribution of the values of the random function at the points $\theta_1, \dots, \theta_n$. For given n, the functions (2.3.6) are called the n-dimensional distributions of the random function. Finite-dimensional distributions satisfy the following *consistency conditions*:

1. If $i_1, \dots, i_n$ is a permutation of $1, \dots, n$, then

$$F_{\theta_1,\dots,\theta_n}(B_1, \dots, B_n) = F_{\theta_{i_1},\dots,\theta_{i_n}}(B_{i_1}, \dots, B_{i_n});$$

2. $F_{\theta_1,\dots,\theta_{n-1},\theta_n}(B_1, \dots, B_{n-1}, X) = F_{\theta_1,\dots,\theta_{n-1}}(B_1, \dots, B_{n-1})$

(b) *Moment functions.* To specify all of the finite-dimensional distribution functions is constructively impossible. Therefore other parameters of random functions are used particularly the moment functions. Let $X = R^1$. Assume that $\mathbf{E}|x(\theta,\omega)|^m < \infty$ for all $\theta \in \Theta$ with m a positive integer. Then

$$M_k(\theta_1,\dots,\theta_k) = \mathbf{E}x(\theta_1,\omega)x(\theta_2,\omega)\dots x(\theta_k,\omega), \quad k \le m\ , \tag{2.3.7}$$

is the k-th *moment function* of $x(\theta,\omega)$. The first moment function $M_1(\theta) = \mathbf{E}_x(\theta,\omega)$ is again called the mean of $x(\theta,\omega)$ and

$$\begin{aligned} R(\theta_1,\theta_2) &= M_2(\theta_1,\theta_2) - M_1(\theta_1)M_1(\theta_2) \\ &= \mathbf{E}(x(\theta_1,\omega) - \mathbf{E}x(\theta_1,\omega))(x(\theta_2,\omega) - \mathbf{E}x(\theta_2,\omega)) \end{aligned} \tag{2.3.8}$$

is its *covariance function.* In many applications, one is satisfied in knowing just these two parameters of a random function. It should be pointed out that they assist in solving an entire class of problems in the theory of stochastic processes (the linear problems). Any function $M_1(\theta)$ may clearly play the role of a mean. But $M_2(\theta_1,\theta_2)$ and $R(\theta_1,\theta_2)$ are *positive-definite* functions:

$$\sum_{i,j=1}^{k} M_2(\theta_i,\theta_j)z_i\bar{z}_j \ge 0, \quad \sum_{i,j=1}^{k} R(\theta_i,\theta_j)z_i\bar{z}_j \ge 0 \tag{2.3.9}$$

for any $k, \theta_1,\dots,\theta_k \in \Theta$ and complex $z_1,\dots,z_k$ ($\bar{z}$ is the conjugate of z). The first inequality in (2.3.9) follows from the relation

$$\sum_{i,j=1}^{k} M_2(\theta_i,\theta_j)z_i\bar{z}_j = \mathbf{E}\left|\sum_{i=1}^{k} z_i x(\theta_i,\omega)\right|^2 .$$

If $X = R^n$, $y_i \in X$ and $\mathbf{E}|x(\theta,\omega|^m < \infty$ for all $\theta \in \Theta$, then the k-th moment function can be expressed as

$$M_k(\theta_1,\dots,\theta_k,y_1,\dots y_k) = \mathbf{E}(x(\theta_1,\omega),y_1)\dots(x(\theta,\omega),y_k)\ . \tag{2.3.10}$$

This is a k-linear function of $y_1,\dots,y_k$. The first two moment functions are $M_1(\theta,y) = \mathbf{E}(x(\theta,\omega),y)$ and

$$M_2(\theta_1,\theta_2,y_1,y_2) = (B(\theta_1,\theta_2)y_1,y_2) - M_1(\theta_1,y_1)M_1(\theta_2,y_2)\ ,$$

in which $B(\theta_1,\theta_2)$ is a function defined on Θ^2 whose values are symmetric operators in R^n. It is termed the operator covariance function of $x(\theta,\omega)$. $\mathbf{E}x(\theta,\omega)$ may again clearly be any function. The covariance function is positive-definite:

$$\sum_{i,j=1}^{k} (B(\theta_i,\theta_j)y_i,y_j) \ge 0$$

for arbitrary $k, \theta_1,\dots,\theta_k$ and $y_i \in X, i = 1,\dots,k$.

(c) *Gaussian random functions.* Let $X = R^1$. A random function $x(\theta, \omega)$ is said to be *Gaussian* if $\sum \lambda_k x(\theta_k, \omega)$ (a scalar random variable) has a normal distribution for any n, $\theta_1, \ldots, \theta_n \in \Theta$ and numbers $\lambda_1, \ldots, \lambda_n$. In other words, it has a density of the form (2.3.1) with certain a and positive b (or it is a constant with probability 1 in which event we take $b = 0$). This is equivalent to $x(\theta_1, \omega), \ldots, x(\theta_n, \omega)$ having a joint n-dimensional normal distribution for all $n, \theta_1, \ldots, \theta_n$. (The density (2.3.4) determines a nondegenerate distribution. A more general distribution may be obtained from a nondegenerate one by means of a passage to the limit.) To specify the joint distribution of n Gaussian random variables, it suffices to assign their expectations and covariances (2.3.5). Therefore to specify the finite-dimensional distributions of a Gaussian random function, it suffices to give its mean $a(\theta) = \mathbf{E}x(\theta, \omega)$ and covariance function

$$R(\theta_1, \theta_2) = \mathbf{E}(x(\theta_1, \omega) - a(\theta_1))(x(\theta_2, \omega) - a(\theta_2)) \ . \tag{2.3.11}$$

2.3.3 Random Elements in Linear Spaces

Let X be a linear space and let L be a linear set of linear functionals on X (that is, linear mappings from X to R). The functionals in L will be assumed to separate the points in X. Denote by $\mathcal{B}^L$ the smallest σ-algebra of subsets of X with respect to which all the functionals in L are measurable. We shall view $(X, \mathcal{B}^L)$ as a measurable space. We are interested in the random elements in this space and their distributions. If X is a locally convex linear topological space, we take L to be the space X^* of all continuous linear functionals and if $X$is separable, then $\mathcal{B}^L$ coincides with the σ-algebra $\mathcal{B}_X$ of Borel sets of X. Most interesting is the case where X is a separable Banach space and particularly a Hilbert space. The main question is the assignment of the probability distributions on $(X, \mathcal{B}^L)$. A procedure will be discussed below for specifying a probability distribution by way of its Fourier transform or, in other words, its characteristic functional. This procedure is used extensively also for a finite-dimensional space and we shall explain it in detail for that case.

(a) *The characteristic function of a random variable and a random vector.* Let ξ be a real random variable. Then $f_\xi(t) = \mathbf{E}\exp\{i\xi t\}, t \in R$, is called the *characteristic function* of ξ (or of its distribution). Let us mention a few properties of a characteristic function.

1. It is uniformly continuous in t and $f_\xi(0) = 1$.
2. $f_\xi(t)$ is positive-definite:

$$\sum_{k,j=1}^{m} f_\xi(t_k - t_j) z_k \bar{z}_j \geq 0 \tag{2.3.12}$$

for any choice of m, real $t_1, \ldots, t_m$ and complex $z_1, \ldots, z_m$.

Bochner's Theorem. *If $f(t)$ is a continuous positive-definite function with $f(0) = 1$, then $f(t) = \int e^{itx}\mu(dx)$, where μ is a probability measure on R. In other words, f is a characteristic function.*

3. A characteristic function determines the distribution of a random variable. To see this, we introduce the notion of *complete set* of functions. F is said to be a complete set of continuous functions if for any choice of distinct probability measures μ_1 and μ_2 on R, there is an $f \in F$ such that $\int f\mu_1 \neq \int f d\mu_2$. To show that property 3 is valid, it suffices to establish that $\{e^{itx}, t \in R\}$ is a total family of functions. However, this is a consequence of being able to specify for every bounded and continuous function $f(x)$ a sequence of trigonometric polynomials $g_n(x)$ such that

$$\sup_{n,x} |g_n(x)| < \infty, \quad \lim_{n\to\infty} \sup_{|x|\leq n} |g_n(x) - f(x)| = 0 .$$

Now let $\xi = (\xi^1, \ldots, \xi^n)$ be a random vector. Its characteristic function is $f_\xi(t) = \mathbf{E}\exp\{i(\xi, t)\}, t \in R^n$ and $(\xi, t) = \sum_k \xi^k t^k$, where $t = (t^1, \ldots, t^n)$. It is also called the joint characteristic function of $\xi^1, \ldots, \xi^n$ or n-dimensional characteristic function. Statements 1-3 hold for multivariate characteristic functions. (Positive-definiteness is also expressed here by (2.3.12) except $t_l, \ldots, t_m \in R^n$).

Bochner's Theorem remains valid also in the multivariate case.

(b) *The characteristic functional.* Let μ be a probability measure on $(X, \mathcal{B}^L)$. The characteristic functional of μ is the function

$$\varphi_\mu(l) = \int \exp\{il(x)\}\mu(dx) . \tag{2.3.13}$$

If μ is the distribution of a random element $x(\omega)$, then $\varphi_\mu(l) = \mathbf{E}\exp\{il(x(\omega))\}$ and $\varphi_\mu(l)$ is also called the characteristic functional of $x(\omega)$. We now note the main properties of $\varphi_\mu(l)$.

1. $\varphi_\mu(l)$ is weakly continuous in l: if $l_n(x) \to l(x)$ for all x, then $\varphi_\mu(l_n) \to \varphi_\mu(l)$; $\varphi_\mu(0) = 1$.
2. $\varphi_\mu(l)$ is a positive-definite function:

$$\sum_{k,j=1}^{m} \varphi_\mu(l_k - l_j) z_k \bar{z}_j \geq 0$$

for any choice of m, $l_1, \ldots, l_m \in L$ and complex $z_1, \ldots, z_m$.

3. $\varphi_\mu(l)$ determines the measure μ uniquely. This statement follows because every bounded $\mathcal{B}^L$-measurable function $g(x)$ is the limit of a sequence of collectively bounded functions of the form

$$g_n(x) = \sum_k c_{nk} \exp\{il_{nk}(x))\},$$

where $c_{nk} \in Z$, $\sum_k |c_{nk}| < \infty$ and $l_{nk} \in L$.

Bochner's Theorem is generally false when X is infinite-dimensional. However, it is possible to indicate an X for which it is preserved. Let $X = R^\infty$, the space of X-sequences $x = (x^1, x^2, \ldots, x^n, \ldots)$, $x^i \in R$. Take L to be the collection of all functionals of the form $l(x) = \sum \alpha_k x^k$, in which only finitely many of the numbers $\alpha_k \in R$ are nonvanishing. In this case, the σ-algebra $\mathcal{B}^L$ is the smallest σ-algebra containing the sets of the form $\{x : x^k \in \Delta\}$, where Δ is an interval of the line. It is natural to denote this σ-algebra by $\mathcal{B}_{R^\infty}$.

Theorem. *Let $\varphi(l)$ be a function on L satisfying:* 1. *it is continuous and $\varphi(0) = 1$;* 2. *it is positive-definite. Then $\varphi(l)$ is the characteristic functional of some distribution on $\mathcal{B}_{R^\infty}$.*

An example of an infinite-dimensional linear space X in which Bochner's theorem is false is a separable Hilbert space with L taken to be $X^* = X$. This fact will be discussed in the next section.

2.4 Construction of Probability Spaces

We now turn to probability spaces with specific sample spaces Ω. We shall describe the σ-algebras in these spaces and ways of assigning measures on these σ-algebras.

2.4.1 Finite-dimensional Space

Such a space has already been discussed when defining a random vector. The σ-algebra was chosen to be the Borel σ-algebra. The measure was determined by the distribution function. A nontrivial assertion is that every distribution function determines some measure. This fact will now be established here. We first define distribution function.

Definition. A function $F(x^1, x^2, \ldots, x^n)$ defined on R^n and assuming values in $[0, 1]$ is an n-dimensional distribution function if

1. $\lim\limits_{\substack{x^1 \to -\infty, \\ \ldots, x^n \to -\infty}} F(x^1, \ldots, x^n) = 0, \qquad \lim\limits_{\substack{x^1 \to \infty, \\ \ldots, x^n \to \infty}} F(x^1, \ldots, x^n) = 1,$
2. $\Delta_{h_1}^{(1)} \ldots \Delta_{h_n}^{(n)} F(x^1, \ldots x^n) \geq 0$ for all $h_1 > 0, \ldots, h_n > 0$, where for any $G(x^1, \ldots, x^n)$ define on R^n,

$$\Delta_h^{(k)} G(x^1, \ldots, x^n) = G(x^1, \ldots, x^{k-1}, x^k + h, x^{k+1}, \ldots, x^n) - G(x^1, \ldots, x^n),$$

3. $F(x^1, \ldots, x^n)$ is left-continuous jointly in its arguments.
   If there exists a measure μ such that $F(x^1, \ldots, x^n) = \mu(\{y : y^1 < x^1, \ldots, y^n < x^n\})$, then F is the distribution function for μ. Observe that then

$$\Delta_{h_1}^{(1)} \dots \Delta_{h_n}^{(n)} F(x^1, \dots, x^n) = \mu([x^1, x^1 + h_1[\times \dots \times [x^n, x^n + h_n[) \, . \tag{2.4.1}$$

Theorem. *To every distribution function, there exists a probability measure μ satisfying (2.4.1).*

Proof. Consider the sets in R^n that are representable as a finite union of half-open intervals in R^n of the form $[a_1, b_1[\times [a_2, b_2[\times \dots \times [a_n, b_n[$ (a_i may be $-\infty$ and b_i may be ∞). These sets form an algebra $\mathcal{A}_0$. Every set in $\mathcal{A}_0$ is expressible as the union of disjoint half-open intervals. Let I be a half-open interval of the above form. Put

$$\tilde{\mu}(I) = \Delta_{b_1 - a_1}^{(1)} \dots \Delta_{b_n - a_n}^{(n)} F(a_1, \dots, a_n)$$

and extend $\tilde{\mu}(I)$ to $\mathcal{A}_0$ as an additive function. Thus one can form a finitely additive function for which (2.4.1) holds.

It remains to show that it can be extended to a countably-additive function. To this end, it has to be continuous on $\mathcal{A}_0$. Let C_m be a sequence of sets in $\mathcal{A}_0$, $C_m \supset C_{m+1}$, with $\mu(C_m) \geq \delta > 0$. It is necessary to show that $\cap C_m$ is nonempty. By property 3, if $I = [a_1, b_1[\times, \dots \times [a_n, b_n[$ is any interval, it is possible to find an $I' = [a_1, b_1'[\times \dots \times [a_n, b_n'[$ such that $a_i < b_i' < b_i$ and $\tilde{\mu}(I) - \tilde{\mu}(I')$ is arbitrarily small. Let $C_m = \bigcup_{k=1}^{N_m} I_{mk}$ and suppose that the I_{mk} are disjoint. As was indicated above, for each m and k we can form an $I'_{mk} \subset I_{mk}$ so that $[I'_{mk}] \subset I_{mk}$ and $\tilde{\mu}(I_{mk}) - \tilde{\mu}(I'_{mk}) < \delta / 2^{m+k+1}$ ($[\cdot]$ is the closure of a set). Put $C'_m = \bigcup_{k=1}^{N_m} I'_{mk}$ and $\tilde{C}_m = \bigcap_{i=1}^{m} C'_i$. Then

$$\begin{aligned} \tilde{\mu}(\tilde{C}_m) &\geq \mu(C_m) - \sum_{i=1}^{m} \tilde{\mu}(C_i \setminus C'_i) \geq \delta - \sum_{i=1}^{m} \sum_{k} \tilde{\mu}(I \setminus I'_{mk}) \\ &\geq \delta - \delta \sum_{i=1}^{m} \sum_{k=1}^{\infty} \frac{1}{2^{m+k+1}} \geq \frac{\delta}{2} \, . \end{aligned}$$

Clearly $\tilde{C}_m \supset \tilde{C}_{m+1}$, $\tilde{C}_m \in \mathcal{A}_0$ and $[\tilde{C}_m] \subset [C'_m] \subset C_m$. Since $\tilde{C}_m$ is nonempty, we have $\bigcap_m [\tilde{C}_m] \neq \emptyset$ and hence so is $\cap C_m$ nonempty. □

2.4.2 Function Spaces

Let $(X, \mathcal{B})$ be a measurable space and let Θ be any (parameter) space. Let X^θ denote the space of all functions $\theta : \Theta \to X$. We now examine how to construct a σ-algebra in X^θ and a measure on this σ-algebra so that a random function can be defined on the probability space having X^θ as sample space.

(a) *Cylinder sets and σ-algebras.* Let Λ be a finite subset of Θ and n_λ the number of elements in Λ. Let P_Λ denote a mapping from X^θ to X^{n_Λ} : $P_\Lambda(x(\cdot)) = (x(\theta_1^\Lambda), \dots, x(\theta_{n_\Lambda}^\Lambda))$, where θ_i^Λ, $i = 1, \dots, n_\Lambda$, are all the points of Λ. Sets in

X^θ of the form $P_\Lambda^{-1}(B_{n_\Lambda})$ with $B_{n_\Lambda} \in \mathcal{B}^{n_\Lambda}$ are called *cylinders* and Λ is said to be the base of the cylinder. The collection of all cylinder sets with a given base forms a σ-algebra, which we shall denote by $\mathcal{C}^\Lambda$. If $\Lambda_1 \subset \Lambda_2$, then clearly $\mathcal{C}^{\Lambda_1} \subset \mathcal{C}^{\Lambda_2}$. The collection $\bigcup_{\Lambda \subset \Theta} \mathcal{C}^\Lambda$, where the union is taken over all finite subsets Λ, is the algebra of cylinder sets. The smallest σ-algebra including it is the *cylinder σ-algebra*. We denote it by $\mathcal{C}(X, \Theta)$. If $\Theta_1 \subset \Theta$, then $C_{\theta_1}(X, \Theta)$ denotes the smallest σ-algebra containing $\bigcup_{\Lambda \subset \Theta_1} \mathcal{C}^\Lambda$ with the Λ finite sets.

Every set in $\mathcal{C}(X, \Theta)$ is generated by at most countable unions and intersections of cylinder sets. Therefore if A is any set in $\mathcal{C}(X, \Theta)$, one can find a sequence $\{C_k, k \geq 1\}$ of cylinder sets producing A and if Θ_1 is a countable subset of Θ such that $C_k \in \bigcup_{\Lambda \subset \Theta_1} \mathcal{C}^\Lambda$, then $A \in \mathcal{C}_{\Theta_1}(X, \Theta)$. Thus, $\mathcal{C}(X, \Theta) = \bigcup_{\Theta_1 \subset \Theta} \mathcal{C}_{\Theta_1}(X, \Theta)$, the union being over all countable subsets Θ_1. To assign a measure on $\mathcal{C}(X, \Theta)$, it suffices to assign it on each σ-algebra $\mathcal{C}_{\Theta_1}(X, \Theta)$.

(b) *Consistent finite-dimensional distributions.* Suppose that to each finite set $(\theta_1, \theta_2, \ldots, \theta_n)$ there corresponds a probability measure $\mu_{\theta_1,\ldots,\theta_n}$ on $\mathcal{B}^n$. All such measures are *consistent finite-dimensional distributions* if the following conditions hold:

1. Let T be a mapping of X^n into itself: $T(x_1, \ldots, x_n) = (x_{i_1}, \ldots, x_{i_n})$, where $i_1, \ldots, i_n$ is a permutation of $1, 2, \ldots, n$. Then $\mu_{\theta_1,\ldots,\theta_n}(B) = \mu_{\theta_{i_1},\ldots,\theta_{i_n}}(T^{-1}B)$ for $B \in \mathcal{B}^n$;
2. If $Q(x_1, \ldots, x_n) = (x_1, \ldots, x_{n-1})$ is a mapping from X^n to X^{n-1}, then

$$\mu_{\theta_1,\ldots,\theta_{n-1}}(B) = \mu_{\theta_1,\ldots,\theta_n}(Q^{-1}B), \quad B \in \mathcal{B}^n.$$

If $(\theta_1, \ldots, \theta_n) = \Lambda$, then a measure μ_Λ, is defined on $\mathcal{C}^\Lambda$ by

$$\mu_\Lambda(P_\Lambda^{-1}B) = \mu_{\theta_1,\ldots,\theta_n}(B), \quad B \in \mathcal{B}^n.$$

Property 1 guarantees that the values of the measure are independent of the way the elements of Λ are numbered. Further, since $\mathcal{C}^{\Lambda_1} \subset \mathcal{C}^{\Lambda_2}$ if $\Lambda_1 \subset \Lambda_2$, the natural question arises of how μ_{Λ_1} and μ_{Λ_2} are related. Property 2 ensures that μ_{Λ_2} and μ_{Λ_1} coincide on $\mathcal{C}^{\Lambda_1}$, that is, μ_{Λ_2} is an extension of μ_{Λ_1} (property 2 establishes this if $\Lambda_2 \setminus \Lambda_1$ is a singleton set). This determines a non-negative finitely-additive set function on the algebra $\bigcup_{\Lambda \subset \Theta} \mathcal{C}^\Lambda$. It can be extended to a countably-additive measure on $\mathcal{C}(X, \Theta)$ if and only if it is countably additive (or continuous) on the algebra of cylinder sets.

(c) *Kolmogorov's theorem.* This theorem gives sufficient conditions for the existence of a measure on $\mathcal{C}(X, \Theta)$ with given finite-dimensional distribution functions. These conditions are formulated in terms of a measurable space.

Condition K. For all $n \geq 1$, there is a class of sets $\mathcal{K}_n \subset \mathcal{B}^n$, satisfying:

K1. $QS \in \mathcal{K}_{n-1}$ for $n > 1$ and $S \in \mathcal{K}_n$ in which Q is the projection of X^n on X^{n-1}.

K2. $\cap S_m \in \mathcal{K}_n$ if $S_m \in \mathcal{K}_n$, $m = 1, 2 \ldots$
K3. For all $B \in \mathcal{B}^n$ and any measure μ_n on $\mathcal{B}^n$,

$$\mu_n(B) = \sup[\mu_n(S) : S \in \mathcal{K}_n, S \subset B].$$

If X is a Borel space (that is, it is a Borel subset of a complete separable metric space), $\mathcal{K}_n$ may be taken to be the class of compact sets in X^n.

Theorem. *If $(X, \mathcal{B})$ satisfies Condition* K, *consistent finite-dimensional distribution functions determine a probability measure on $\mathcal{C}(X, \Theta)$.*

Proof. It suffices to consider denumerable Θ. We take $\Theta = \{1, 2, \ldots\}$. Let μ_n be the probability measure on $(X^n, \mathcal{B}^n)$ corresponding to $\Lambda_n = \{1, 2, \ldots, n\}$. Since any finite subset belongs to Λ_n for n sufficiently large, the finite-dimensional distributions determine μ_n and the consistency conditions for them are the same. In the situation in question, the theorem says that there exists a sequence of random elements $x_n(\omega)$ in $(X, \mathcal{B})$ such that $x_1(\omega), \ldots, x_k(\omega)$ have the given joint distributions for all k. To prove the theorem, it suffices to show that if $\{C_n\}$ is a sequence of cylinder sets with $C_n \supset C_{n+1}$ and $\mu(C_n) \geq \delta > 0$, then $\cap C_n$ is nonempty. Here μ is a finitely-additive function on $\bigcup_n \mathcal{C}^{\Lambda_n}$ which coincides with μ_{Λ_n} on $\mathcal{C}^{\Lambda_n}$. Without loss of generality, we may assume that $C_n \in \mathcal{C}^{\Lambda_n}$. Let $C_n = \{P_{\Lambda_n} x \in \hat{C}_n\}$, where $\hat{C}_n \in \mathcal{B}^n$. By K3, we may assume that $\hat{C}_n \in \mathcal{K}_n$. Consider the operator Q_n on $\bigcup_{m \geq n} X^m$ given by $Q_n(x_1, \ldots, x_m) = (x_1, \ldots, x_n), m \geq n$. Define the following sets in X^n:

$$\hat{C}_n^{(1)} = \bigcap_{m \geq n} Q_n(\hat{C}_m), \quad \hat{C}_n^{(k+1)} = \bigcap_{m \geq n} Q_n(\hat{C}_m^{(k)}), \quad k \geq n, \quad \bar{C}_n = \cap \hat{C}_n^{(k)} .$$

If $\hat{C}_n \in \mathcal{K}_n$ and $\hat{C}_n^{(k)} \in \mathcal{K}_n$, then $\bar{C}_n \in \mathcal{K}_n$. Furthermore,

$$\begin{aligned} \mu_n(\hat{C}_n^{(k+1)}) &= \lim_{m \to \infty} \mu_n(Q_N(\hat{C}_m^{(k)})) = \lim_{m \to \infty} \mu_m(\hat{C}_m^{(k)}) \\ &= \lim_{m \to \infty} \mu_m(\hat{C}_m^{(k-1)}) = \lim_{m \to \infty} \mu_m(\hat{C}_m) \geq \delta . \end{aligned}$$

Hence $\mu_n(\bar{C}_n) > \delta$ also and so $\bar{C}_n \neq \emptyset$. It can be shown that $Q_n(\bar{C}_{n+1}) = \bar{C}_n$. We say that $(x_1, \ldots x_m) \in X^m$ can be extended if for all $l > m$ there are $x_{m+1}^\ell, \ldots, x_\ell^\ell$ such that $(x_1, \ldots, x_m, x_{m+1}^\ell, \ldots, x_\ell^\ell) \in \hat{C}_\ell$. These new points will be called extensions of $(x_1, \ldots x_m)$. It is easy to see that $\hat{C}_n^{(1)}$ comprises the points $(x_1, \ldots x_n)$ that can be extended, $\hat{C}_n^{(2)}$ the points whose extensions can be extended (we then say that $(x_1, \ldots x_n)$ admits a 2-fold extension), $\hat{C}_n^{(k)}$ the points that admit k-fold extensions and $\bar{C}_n$ comprises the points admitting extensions of any frequency. If $(x_1, \ldots, x_{n+1})$ can be extended any number of times, then clearly $(x_1, \ldots, x_n)$ possesses this property: $Q_n(\bar{C}_{n+1}) = \bar{C}_n$. Let $\bar{x}_1 \in \bar{C}_1 \subset \hat{C}_1$. There is a point $(\bar{x}_1, \bar{x}_2) \in \bar{C} \subset \hat{C}_2$ and so on. Thus for all n, there is an $\bar{x}_n$ such that $(\bar{x}_1, \ldots, \bar{x}_n) \in \bar{C}_n \subset \hat{C}_n$. But then $\bar{x} = (\bar{x}_1, \bar{x}_2, \ldots) \in \bigcap_n C_n$. □

2.4.3 Linear Topological Spaces. Weak Distributions

Let X be a locally convex linear topological space and X^* the space of linear functionals on X. To each finite-dimensional linear subset $\Lambda \subset X^*$, introduce the σ-algebra $\mathcal{B}^\Lambda$ generated by the sets $\{x : \varphi(x) < \alpha\}, \alpha \in R$ and $\varphi \in \Lambda$. If $\mathcal{B}$ is the σ-algebra generated by all linear functionals in X^*, then $B \supset B^\Lambda$. On the other hand, if $\mathcal{B}_0 = \bigcup_{\Lambda \subset X^*} \mathcal{B}^\Lambda$ taken over all finite-dimensional subspaces Λ, then $\mathcal{B}_0$ is an algebra and $\mathcal{B}$ is the smallest σ-algebra containing $\mathcal{B}_0$. The members of $\mathcal{B}_0$ are called cylinder sets and those of B^Λ cylinder sets with base Λ. To assign a measure on $\mathcal{B}$, it suffices to know it on $\mathcal{B}_0$. Let $\tilde{\mu}$ be a given additive set-function on $\mathcal{B}_0$, which is a probability measure on each σ-algebra B^Λ with Λ finite-dimensional. Then $\tilde{\mu}$, is called a *weak distribution*. Every probability measure on $\mathcal{B}$ clearly generates a weak distribution on $\mathcal{B}_0$ and the weak distribution determines this measure uniquely. It is therefore natural to assign probability measures on $(X, \mathcal{B})$ with the help of weak distributions. The characteristic functional of a probability distribution on a linear space has already been discussed. Since for all $\varphi \in X^*$ the function $\varphi(x)$ is $\mathcal{B}^\Lambda$-measurable if $\varphi \in \Lambda$, a weak distribution $\tilde{\mu}$ also determines the characteristic functional

$$f(\varphi) = \int e^{i\varphi(x)} \tilde{\mu}(dx) \,. \tag{2.4.2}$$

It is easy to show that $f(\varphi)$ has the following properties:

1. $f(\varphi)$ is positive-definite and $f(0) = 1$;
2. $f(\varphi)$ is continuous in φ on each finite-dimensional subspace Λ of X^*.

The converse is also true. If $f(\varphi)$ has properties 1 and 2 , then there exists a weak distribution $\tilde{\mu}$ for which (2.4.2) holds. $\tilde{\mu}$ may be constructed as follows. Let Λ be a finite-dimensional subspace of X^* and $\varphi_1, \ldots, \varphi_n$ a basis for Λ. Define $f_\Lambda(t_1, \ldots, t_n) = f\left(\sum_{k=1}^n t_k \varphi_k\right)$. The function $f_\Lambda(t_1, \ldots, t_n)$ is continuous and positive-definite. Thus by Bochner's theorem, a probability measure $\mu_\Lambda(dx)$ exists in R^n such that

$$f_\Lambda(t_1, \ldots, t_n) = \int \exp\left\{ i \sum_{k=1}^n t_k y^k \right\} \mu_\Lambda(dy),$$

$y = (y^1, \ldots, y^n) \in R^n$. Let $\Phi_\Lambda : X \to R^n$, $\Phi_\Lambda(x) = (\varphi_1(x), \ldots, \varphi_n(x))$, and put $\tilde{\mu}(\Phi_\Lambda^{-1}(B)) = \mu_\Lambda(B)$ for $B \in \mathcal{B}_{R^n}$. Thus a measure has been defined on $\mathcal{B}^\Lambda$. We now use the fact that if $\Lambda \subset \Lambda_1$ and $\varphi_1(x), \ldots, \varphi_n(x)$, $\varphi_{n+1}(x), \ldots, \varphi_m(x)$ is a basis for Λ_1, then

$$f_\Lambda(t_1, \ldots, t_n) = f_{\Lambda_1}(t_1, \ldots, t_n, 0, \ldots, 0).$$

This helps to show that the values of $\tilde{\mu}$ on B^Λ and B^{Λ_1} are consistent and do not depend on the choice of basis. That (2.4.2) holds for $\tilde{\mu}$ follows from our construction (μ_Λ was found instantly with the help of Bochner's theorem).

The question of the extendability of $\tilde{\mu}$ to $\mathcal{B}$ as a countably-additive function reduces again to showing the continuity of $\tilde{\mu}$ on the algebra $\mathcal{B}_0$. The following is a general result on extendability.

Theorem. *Suppose that to any positive ε there is a sequence of weakly closed cylinder sets K_n such that $\tilde{\mu}(K_n) \geq 1 - \varepsilon, K_n \supset K_{n+1}$ and $\cap K_n$ is weakly compact. Then $\tilde{\mu}$ can be extended to a measure.*

Proof. Let $B_n \in \mathcal{B}_0, B_n \supset B_{n+1}$, and $\tilde{\mu}(B_n) \geq \delta$. If $B_n = \Phi_{\Lambda_n}^{-1}(\tilde{B}_n)$ with $\tilde{B}_n \in \mathcal{B}_{R^n}$, we can choose a closed set $\tilde{F}_n \subset \tilde{B}_n$ so that $\mu_{\Lambda_n}(\tilde{B}_n \setminus \tilde{F}_n) < \delta/2^{n+1}$. Put $F_n = \Phi_{\Lambda_n}^{-1}(\tilde{F}_n)$ and $\hat{F}_n = \bigcap_{k=1}^{n} F_k$. Then each $\hat{F}_n$ is weakly closed, $\hat{F}_{n+1} \subset \tilde{F}_n$ and $\tilde{\mu}(\hat{F}_n) \geq \delta/2$. Let K_n be a sequence for which the hypotheses of the theorem hold when $\varepsilon < \delta/2$. Then $K_n \cap \hat{F}_n$ is a nonempty closed cylinder set with base Λ_n. Without loss of generality, we may assume $\hat{F}_n$ and K_n to be cylinder sets with the exact same base Λ_n. From the condition on K_n, it follows that $\bigcap_n (K_n \cap \hat{F}_n)$ is nonempty. □

2.4.4 The Minlos-Sazonov Theorem

Let X be a separable Hilbert space. We identify X^* with X by expressing the linear functionals in terms of the inner product putting $\varphi(x) = (\varphi, x), \varphi \in X$. Introduce a topology S in X as follows. X is a linear topological space with neighborhoods of zero $V_A = \{x : (Ax, x) < 1\}$, where A is a nonnegative symmetric operator with a finite trace: if $\{e_k\}$ is a basis in X, then $\operatorname{Tr} A = \sum(Ae_k, e_k) < \infty$. This system of neighborhoods clearly gives a separable topology.

Theorem. *Suppose that $f(\varphi)$, $\varphi \in X$, is a continuous positive-definite functional with $f(0) = 1$. It is the characteristic functional of a probability distribution in X if and only if* $\operatorname{Re} f(\varphi)$ *is continuous at $\varphi = 0$ in the topology S.*

Necessity. Let μ be a probability measure on X. Then

$$1 - \operatorname{Re} f(\varphi) = \int (1 - \cos(\varphi, x))\mu(dx) \leq 2 \int_{|x|>\rho} \mu(dx) + \frac{1}{2} \int_{|x|\leq\rho} (\varphi, x)^2 \mu(dx).$$

Put $(A_\rho \varphi, \varphi) = \int_{|x|\leq\rho} (\varphi, x)^2 \mu(dx)$; A_ρ is a non negative symmetric operator. Then

$$\operatorname{Tr} A_\rho = \int_{|x|\leq\rho} \sum_{k=1}^{\infty} (e_k, x)^2 \mu(dx) \leq \rho^2 \ .$$

Take a positive ε, choose ρ so that $\int_{|x|>\rho} \mu(dx) < \varepsilon/4$ and let $B = \frac{1}{\varepsilon} A_\rho$. Then $(A_\rho \varphi, \varphi) < \varepsilon/2$ when $(B\varphi, \varphi) < 1$ and so

$$1 - \operatorname{Re} f(\varphi) < \varepsilon/2 + 1/2(A_\rho \varphi, \varphi) < \varepsilon \ .$$

Sufficiency. Let $\{e_k\}$ be a fixed basis. Put $K_n(\rho) = \{x : \sum_{k=1}^n (x, e_k)^2 < \rho^2\}$. $\bigcap_n K_n(\rho)$ is clearly a weakly compact ball in X of radius ρ. On the basis of the theorem in the preceding section, it suffices to show that to any positive ε there is a ρ such that $\tilde{\mu}(K_n(\rho)) > 1-\varepsilon$ for all n. Here $\tilde{\mu}$ is the weak distribution constructed from f. We now make use of the relation

$$\exp\left\{-\frac{\lambda}{2}\sum_{k=1}^n (x, e_k)^2\right\}$$
$$= (2\pi\lambda)^{-n/2} \int \exp\left\{i\sum_{k=1}^n t_k (x, e_k\right\} \exp\left\{-\frac{1}{2\lambda}\sum_{k=1}^n t_k^2\right\} dt_1 \ldots dt_n \ .$$

Integration with respect to $\tilde{\mu}$ yields

$$\int \left(1 - \exp\left\{-\frac{\lambda}{2}\sum_{k=1}^n (x, e_k)^2\right\}\right) \tilde{\mu}(dx)$$
$$= (2\pi\lambda)^{-n/2} \int \left(1 - \operatorname{Re} f\left(\sum_{k=1}^n t_k, e_k\right)\right) \exp\left\{-\frac{1}{2\lambda}\sum_{k=1}^n t_k^2\right\} dt_1 \ldots dt_n \ .$$

If A is a kernel operator such that $1 - \operatorname{Re} f(\varphi) < \varepsilon/4$ when $(A\varphi, \varphi) < 1$, then the right-hand side is majorized by the quantity

$$(2\pi\lambda)^{-n/2} \int \left(\varepsilon/4 + 2\left(A\sum_{k=1}^n t_k e_k, \sum_{k=1}^n t_k e_k\right)\right) \exp\left\{-\frac{1}{2\lambda}\sum_{k=1}^n t_k^2\right\} dt_1$$
$$\ldots dt_n = \varepsilon/4 + 2\lambda \operatorname{Tr} A \ .$$

Thus,

$$\tilde{\mu}(K_n(\rho)) \left(1 - \exp\{-\frac{\lambda}{2}\rho^2\}\right)$$
$$\leq \int \left(1 - \exp\left\{-\sum_{k=1}^n (x, e_k)^2\right\}\right) \tilde{\mu}(dx) \leq \varepsilon/4 + 2\lambda \operatorname{Tr} A \ .$$

Let $\lambda = \varepsilon/(8 \operatorname{Tr} A)$. Then $\tilde{\mu}(K_n(\rho)) < \frac{1}{2}\varepsilon(1 - \exp\{-\frac{1}{2}\lambda\rho^2\})^{-1} < \varepsilon$ if $\lambda\rho^2 < 1$.

□

3

Independence

Independence is one of the basic concepts of probability theory. The study of independent events, random variables, random elements and σ-algebras comprises to a considerable extent the content of probability theory. This chapter presents the main concepts and facts concerning independence and it examines sequences of independent events and variables and their related random processes.

3.1 Independence of σ-Algebras

3.1.1 Independent Algebras

It will be recalled that algebras $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are independent if

$$\mathbf{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathbf{P}(A_i) \tag{3.1.1}$$

for any choice of $A_1 \in \mathcal{A}_1, \ldots, A_n \in \mathcal{A}_n$ (see p. 24).

Let $\mathcal{A}_i, i = 1, \ldots, l$, be finite algebras and $A_1^{(i)}, \ldots, A_{n_i}^{(i)}$ be the atoms of $\mathcal{A}_i$.

Theorem 3.1.1. *Algebras $\mathcal{A}_i$ are independent if and only if for any $k_i \le n_i$,*

$$\mathbf{P}\left(\bigcap_{i=1}^{l} A_{k_i}^{(i)}\right) = \prod_{i=1}^{l} \mathbf{P}(A_{k_i}^{(i)}). \tag{3.1.2}$$

Proof. The necessity is obvious. It is easy to prove the sufficiency by induction making use of the following trivial assertion: if A and B are independent, A and C are independent and $B \cap C = \emptyset$, then A and $B \cup C$ are also independent. □

Observe that there are only $n_1 n_2 \dots n_l - n_1 - \dots - n_l + l - 1$ independent relations among those in (3.1.2). This is by virtue of the relations $\sum_{k=1}^{n_i} \mathbf{P}(A_k^{(i)}) = 1$ and $\sum_{k=1}^{n_i} \mathbf{P}(A_k^{(i)} \cap (A_j^{(1)}) = \mathbf{P}(A_j^{(1)})$ and others similar to the latter which are based on the fact that $\bigcup_{k=1}^{n_i} A_k^{(i)} = \Omega$. Since $\mathcal{A}_i$ has 2^{n_i} elements, (3.1.2) involves $2^{n_1+\dots+n_l}$ equalities. But the theorem shows that only $n_1 n_2 \dots n_l - n_1 - \dots - n_l + l - 1$ of them are independent.

Remark. Each algebra $\mathcal{A}$ is a union of its finite subalgebras; if $A \in \mathcal{A}$, then A belongs to the four-element algebra $\{\emptyset, A, \bar{A}, \Omega\}$ (if $A = \emptyset$ or $A = \Omega$ the algebra consists of two elements $\{\emptyset, \Omega\}$).

Let $\mathcal{A}$ and $\mathcal{B}$ be algebras. $\mathcal{A} \vee \mathcal{B}$ denotes the smallest algebra containing $\mathcal{A}$ and $\mathcal{B}$. The algebra $\mathcal{A} \vee \mathcal{B}$ consists of sets of the form

$$\bigcup_{k=1}^{n} (A_k \cap B_k) \,,$$

where $A_k \in \mathcal{A}, B_k \in \mathcal{B}$ and n is arbitrary. If $\mathcal{A}$ and $\mathcal{B}$ are finite and A_k and B_l are atoms of them, then $A_k \cap B_l$ is an atom of $\mathcal{A} \vee \mathcal{B}$. Similarly, if $A_k, k = 1, \dots, n$, are algebras, then $\bigvee_{k=1}^{n} \mathcal{A}_k$ is the smallest algebra containing $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n$. It involves sets of the form

$$\bigcup_{k=1}^{m} \left(\bigcap_{i=1}^{n} A_{ki} \right), \quad A_{ki} \in \mathcal{A}_i, \quad m \text{ arbitrary} \,.$$

Let I be an index set. Let $\mathcal{A}_i^\alpha \subset \mathcal{A}_i$ and $\bigcup_\alpha \mathcal{A}_i^\alpha = \mathcal{A}_i$, where the $\mathcal{A}_i^\alpha$'s are finite algebras, $\alpha \in I$. Then

$$\bigvee_{i=1}^{n} \mathcal{A}_i = \bigcup_{\alpha_1 \in I, \dots, \alpha_n \in I} \bigvee_{i=1}^{n} \mathcal{A}_i^{a_i} \,. \tag{3.1.3}$$

This makes it possible to prove the next two theorems just for finite algebras.

Theorem 3.1.2. *Let $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_m$ and $\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_n$ be independent algebras and let $\mathcal{A} = \bigvee_{k=1}^{m} \mathcal{A}_k$. Then the algebras $\mathcal{A}$ and $\mathcal{B}_1, \dots, \mathcal{B}_n$ are also independent.*

Proof. All algebras may be considered finite. $\bigvee_{k=1}^{m} \mathcal{A}_k$ is also a finite algebra and its atoms are of the form $A_1 \cap A_2 \cap \dots \cap A_m$, where$A_i$ is an atom of $\mathcal{A}_i$. If $B_k \in \mathcal{B}_k$, then

$$\begin{aligned} \mathbf{P}\left(\left(\bigcap_{k=1}^{m} A_k \right) \cap \left(\bigcap_{i=1}^{n} B_i \right) \right) &= \prod_{k=1}^{m} \mathbf{P}(A_k) \cdot \prod_{i=1}^{n} \mathbf{P}(B_i) \\ &= \mathbf{P}\left(\bigcap_{k=1}^{m} A_k \right) \cdot \mathbf{P}\left(\bigcap_{i=1}^{n} B_i \right) \,. \end{aligned}$$

Therefore the independence of the algebras follows by Theorem 3.1.1. □

Theorem 3.1.3. *Let $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ be algebras of events. They are independent if and only if $\mathcal{A}_i$ and $\bigvee_{k<i} \mathcal{A}_k$ are independent for all i.*

Proof. That the latter algebras are independent if $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are independent follows from Theorem 3.1.2. Let the hypothesis of the theorem hold with $A_i \in \mathcal{A}_i$. Since $\bigcap_{k=1}^{n-1} A_k \in \bigvee_{k=1}^{n-1} \mathcal{A}_k$, we have

$$\mathbf{P}\left(\left(\bigcap_{k=1}^{n-1} A_k\right) \cap A_n\right) = \mathbf{P}\left(\bigcap_{k=1}^{n-1} A_k\right) \mathbf{P}(A_n) \, .$$

Furthermore, $\bigcap_{k=1}^{n-2} A_k \in \bigvee_{i<n-1} \mathcal{A}_i$ which implies that

$$\mathbf{P}\left(\bigcap_{k=1}^{n-1} A_k\right) = \mathbf{P}\left(\bigcap_{k=1}^{n-2} A_k\right) \mathbf{P}(A_{n-1}) \, .$$

In the exact same way,

$$\mathbf{P}\left(\left(\bigcap_{k=1}^{m} A_k\right) \cap A_{m+1}\right) = \mathbf{P}\left(\bigcap_{k=1}^{m} A_k\right) \mathbf{P}(A_{m+1})$$

for all $m < n$ since $\bigcap_{k=1}^{m} A_k \in \bigvee_{k<m+1} \mathcal{A}_k$. Thus, $\mathbf{P}\left(\bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} \mathbf{P}(A_k)$. □

3.1.2 Conditions for the Independence of σ-Algebras

The preceding section gave some conditions for the independence of algebras of events. The next theorem makes it possible to reduce the question of independence of σ-algebras to the independence of the algebras generating them. (Since a σ-algebra is an algebra, the definition of independence given above is applicable to σ-algebras.)

Theorem 3.1.4. *Let $\mathcal{A}_1^0, \ldots, \mathcal{A}_n^0$ be algebras of events and let $\mathcal{A}_i = \sigma(\mathcal{A}_i^0)$. If $\mathcal{A}_1^0, \ldots, \mathcal{A}_n^0$ are independent, then the σ-algebras $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are independent.*

Proof. We shall apply Theorem 3.1.3. Let us show that $\bigvee_{i<k} \mathcal{A}_i$, and $\mathcal{A}_k$ are independent. Observe that $\sigma(\bigvee_{i<k} \mathcal{A}_i^0)$ is a σ-algebra containing $\mathcal{A}_i^0, i < k$. This implies that $\sigma(\bigvee_{i<k} \mathcal{A}_i^0) \supset \bigvee_{i<k} \mathcal{A}_i = \bigvee_{i<k} \sigma(\mathcal{A}_i^0)$ and so $\bigvee_{i<k} \mathcal{A}_i \subset \sigma(\bigvee_{i<k} \mathcal{A}_i^0)$. But $\bigvee_{i<k} \mathcal{A}_i^0$ and $\mathcal{A}_k^0$ are independent. Therefore to prove the theorem, it suffices to show that it is true for $n = 2$. For fixed $A_1 \in \mathcal{A}_i^0$,

$$\mathbf{P}(A_1 \cap A_2) = \mathbf{P}(A_1)\mathbf{P}(A_2) \tag{3.1.4}$$

on a monotone class of sets A_2. This holds for $A_2 \in \mathcal{A}_2^0$ and thus on $\sigma(\mathcal{A}_2^0) = \mathcal{A}_2$. Let A_2 be a fixed member of $\mathcal{A}_2$. Then (3.1.4) holds on a monotone class of sets $\mathcal{A}_1$ which contains $\mathcal{A}_1^0$ and thus for all $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$. □

Corollary 3.1.1. *Let $\mathcal{A}_1, \ldots, \mathcal{A}_n$ and $\mathcal{B}_1, \ldots, \mathcal{B}_m$ be independent σ-algebras. Then $\sigma(\bigvee_{i=1}^{n} \mathcal{A}_i)$ and $\sigma(\bigvee_{j=1}^{m} \mathcal{B}_j)$ are independent.*

This is a consequence of the independence of the algebras $\bigvee_{i=1}^{n} \mathcal{A}_i$ and $\bigvee_{j=1}^{m} \mathcal{B}_j$ and Theorem 3.1.4.

3.1.3 Infinite Sequences of Independent σ-Algebras

Let $\mathcal{A}_n, n = 1, 2, \ldots,$ be σ-algebras of events. They are independent if $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are independent for all n. If $\mathcal{B}_n$ is a sequence of algebras , then $\bigcup_n \bigvee_{k=1}^n \mathcal{B}_k$ will be denoted by $\bigvee_n \mathcal{B}_n$. It is clearly an algebra and it is the smallest algebra containing all $\mathcal{B}_n$. It consists of finite unions of sets of the form $\bigcap_{k=1}^n B_k$, where $B_k \in \mathcal{B}_k$ and n is arbitrary.

Theorem 3.1.5. *If $\mathcal{A}_n$ is a sequence of independent σ-algebras, then $\mathcal{A}_1, \ldots, \mathcal{A}_n$ and $\sigma(\bigvee_{k=n+1}^\infty \mathcal{A}_k)$ are independent for any n.*

Proof. By Theorem 3.1.2, $\mathcal{A}_1, \ldots, \mathcal{A}_n$ and $\bigvee_{k=n+1}^m \mathcal{A}_k$ are independent for all $m > n$. Therefore the algebras $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ and $\bigvee_{k=n+1}^\infty \mathcal{A}_k = \bigcup_m \bigvee_{k=n+1}^m \mathcal{A}_k$ are independent. It remains to apply Theorem 3.1.4. □

(a) *Kolmogorov's zero-one law.* Let $\mathcal{A}_n$ be a sequence of σ-algebras. $\bigcap_n \sigma(\bigvee_{k=n}^\infty \mathcal{A}_k)$ is called the *tail σ-algebra*. It characterizes the events that are expressible in terms of the events in $\mathcal{A}_k$ with arbitarily large indices. If we interpret the index k of $\mathcal{A}_k$ as time and $\mathcal{A}_k$ as the σ-algebra of observable events at time k, then the tail σ-algebra comprises events that occur "later" than any finite moment of time.

Kolmogorov's Theorem (Zero-One Law). *If the σ-algebras $\mathcal{A}_n$ are independent and A is any event in the tail σ-algebra, then $\mathbf{P}(A)$ is either 0 or 1 (this means that the tail σ-algebra is trivial).*

Proof. By Theorem 3.1.5, the σ-algebras $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ and $\bigcap_m \sigma(\bigvee_{k=m}^\infty \mathcal{A}_k)$ are independent for every n (the last σ-algebra is a subset of $\sigma(\bigvee_{k=n+1}^\infty A_k)$). Thus for all n, the algebras $\bigvee_{i=1}^n \mathcal{A}_i, \bigcap_m \sigma(\bigvee_{k=m}^\infty \mathcal{A}_k)$ and $\bigcup_n \bigvee_{i=1}^n \mathcal{A}_i = \bigvee_{i=1}^\infty \mathcal{A}_i, \bigcap_m \sigma(\bigvee_{k=m}^\infty \mathcal{A}_k)$ are independent. This is a consequence of Theorem 3.1.2. Therefore, by Theorem 3.1.4, $\sigma(\bigvee_{i=1}^\infty \mathcal{A}_i)$ and $\bigcap_m \sigma(\bigvee_{k=m}^\infty \mathcal{A}_k)$ are independent. If $A \in \bigcap_m \sigma(\bigvee_{k=m}^\infty \mathcal{A}_k)$, then $A \in \sigma(\bigvee_{k=1}^\infty \mathcal{A}_i)$ and by virtue of the independence of these σ-algebras, $\mathbf{P}(A \cap A) = \mathbf{P}(A)\mathbf{P}(A)$ or $\mathbf{P}(A) = \mathbf{P}^2(A)$. The theorem is proved. □

(b) *The Borel-Cantelli Lemma.* The next assertion concerns an infinite sequence of events and for independent events, furnishes necessary and sufficient conditions for finitely many of these events to happen with probability 1. Let A_k be a sequence of events. The event that A_k's with arbitrarily large indices have occurred (which is equivalent to infinitely many of the A_k's having occurred) is representable as $\bigcap_n \bigcup_{k=n}^\infty A_k$. Let $\mathcal{A}_k$ be the σ-algebra comprising the events $\Omega, A_k, \Omega \setminus A_k, \emptyset$. Then $\bigcap_n \bigcup_{k=n}^\infty A_k \in \bigcap_n \sigma(\bigvee_{k=n}^\infty \mathcal{A}_k)$. If the A_k's are independent (that is, $A_1, \ldots, A_n$ are independent for all n), then the σ-algebras $\mathcal{A}_k$ are independent. In that case, $\mathbf{P}(\bigcap_n \bigcup_{k=n}^\infty A_k)$ equals 0 or 1 by the zero-one law.

Borel-Cantelli Lemma.

1. *If* $\sum_{k=1}^{\infty} \mathbf{P}(A_k) < \infty$, *then* $\mathbf{P}(\bigcap_n \bigcup_{k=n}^{\infty} A_k) = 0$;
2. *If the events* A_k *are independent and* $\sum_{k=1}^{\infty} \mathbf{P}(A_k) = \infty$, *then* $\mathbf{P}(\bigcap_n \bigcup_{k=n}^{\infty} A_k) = 1$.

Proof. 1. For any choice of m,

$$\mathbf{P}\left(\bigcap_n \bigcup_{k=n}^{\infty} A_k\right) \le \mathbf{P}\left(\bigcup_{k=m}^{\infty} A_k\right) \le \sum_{k=m}^{\infty} \mathbf{P}(A_k)$$

and the right-hand side tends to zero.

2. The events $\Omega \setminus A_k$ are also independent. We have

$$\mathbf{P}\left(\bigcap_n \bigcup_{k=n}^{\infty} A_k\right) = 1 - \mathbf{P}\left(\bigcup_n \bigcap_{k=n}^{\infty} (\Omega \setminus A_k)\right) .$$

Then

$$\begin{aligned}
&\mathbf{P}\left(\bigcup_n \bigcap_{k=n}^{\infty} (\Omega \setminus A_k\right) \\
&= \lim_{n\to\infty} \mathbf{P}\left(\bigcap_{k=n}^{\infty} (\Omega \setminus A_k\right) = \lim_{n\to\infty} \lim_{m\to\infty} \mathbf{P}\left(\bigcap_{k=n}^{m} (\Omega \setminus A_k)\right) \\
&= \lim_{n\to\infty} \lim_{m\to\infty} \prod_{k=n}^{m} (1 - \mathbf{P}(A_k)) = \lim_{n\to\infty} \prod_{k=n}^{\infty} (1 - \mathbf{P}(A_k)) = 0
\end{aligned}$$

since the divergence of $\sum_{k=1}^{\infty} \mathbf{P}(A_k)$ implies that $\prod_{k=n}^{\infty}(1 - \mathbf{P}(A_k)) = 0$ for all n. □

3.1.4 Independent Random Variables

Let $(X_n, \mathcal{B}_n)$ be a (finite or infinite) sequence of measurable spaces and let $\xi_n(\omega)$ be a random element in $(X_n, \mathcal{B}_n)$ defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Let $\mathcal{A}_{\xi_n}$ be the σ-algebra of events $\{\omega : \xi_n(\omega) \in B_n\}$ with $B_n \in \mathcal{B}_n$ (this is the σ-algebra generated by ξ_n). Then $\{\xi_n\}$ is said to be a sequence of *independent random elements* if $\{\mathcal{A}_{\xi_n}\}$ is a sequence of independent σ-algebras.

Let $\mathcal{F}_{(X_n,\mathcal{B})_n}$ be the space of bounded $\mathcal{B}_n$-measurable scalar functions on X_n. A subset T_n of $\mathcal{F}_{(X_n,\mathcal{B}_n)}$ is called *complete* if the relation

$$\int f(x)\mu_1(dx) = \int f(x)\mu_2(dx), \quad f \in T_n ,$$

with μ_1 and μ_2 finite measures on $\mathcal{B}_n$, entails that $\mu_1 = \mu_2$.

Theorem 3.1.6. 1. *If $\xi_1(\omega), \dots, \xi_n(\omega)$ are independent random elements, then*

$$\mathbf{E} \prod_{k=1}^{n} g_k(\xi_k) = \prod_{k=1}^{n} \mathbf{E} g_k(\xi_k) \tag{3.1.5}$$

for any choice of $g_k \in \mathcal{F}_{(X_k, \mathcal{B}_k)}, k = 1, \dots, n$.
2. *If (3.1.5) holds for all $g_k \in T_k$, with T_k a complete subset of $\mathcal{F}_{(X_k, \mathcal{B}_k)}$ then $\xi_1(\omega), \dots, \xi_n(\omega)$ are independent.*

Proof. 1. The relation (3.1.5) holds if the g_k are indicator functions. Both sides of (3.1.5) are linear and continuous in g_k under uniform convergence and every function in $\mathcal{F}_{(X_k, \mathcal{B}_k)}$ can be expressed as the uniform limit of finite linear combinations of indicator functions.

2. Let us show that (3.1.5) is satisfied by all $g_k \in \mathcal{F}_{(X_k, \mathcal{B}_k)}$. The left-hand side of (3.1.5) is clearly representable as $\int g_1(x) \mu_1(dx)$, where $\mu_1(B_1) = \mathbf{E} I_{B_1}(\xi_1) g_2(\xi_2) \dots g_n(\xi_n)$. The right-hand side of (3.1.5) is representable as $\int g_1(x) \tilde{\mu}_1(dx)$, where $\tilde{\mu}_1(B) = \mathbf{P}\{\xi_1 \in B_1\} \mathbf{E} g_2(\xi_2) \dots g_n(\xi_n)$. This is true for all $g_1 \in \mathcal{F}_{(X_1, \mathcal{B}_1)}$. If $g_1 \in T_1$, then $\int g_1(x) \mu_1(dx) = \int g_1(x) \tilde{\mu}_1(dx)$ and hence $\mu_1 = \tilde{\mu}_1$. Therefore (3.1.5) holds for all $g_1 \in \mathcal{F}_{(X_1, \mathcal{B}_1)}$, $g_2 \in T_2, \dots, g_n \in T_n$. Similar reasoning shows that (3.1.5) holds for all $g_1 \in \mathcal{F}_{(X_1, \mathcal{B}_1)}; g_2 \in \mathcal{F}_{(X_2, \mathcal{B}_2)}, g_3 \in T_3, \dots g_n \in T_n$ and so on. For $\xi_1, \dots \xi_n$ to be independent, it suffices to prove that (3.1.5) is satisfied by indicator functions.

□

Remark. If ξ_k is an infinite sequence of independent elements, then statement 1 is valid for all n; if (3.1.5) holds for all n and $g_k \in T_k$, then the ξ_k are independent.

Corollary 3.1.2. *Let $X_n = R^d$ and $\mathcal{B}_n = \mathcal{B}_{R^d}$. Vectors $\xi_k \in R^d, k = 1, \dots, n$, are independent if*
1. *for any continuous functions $f_1, \dots, f_n \in C_{R^d}$,*

$$\mathbf{E} \prod_{k=1}^{n} f_k(\xi_k) = \prod_{k=1}^{n} \mathbf{E} f_k(\xi_k) \ ;$$

2. *for any $z_1, \dots, z_n \in R^d$*

$$\mathbf{E} \exp\left\{ i \sum_{k=1}^{n} (\xi_k, z_k) \right\} = \prod_{k=1}^{n} \mathbf{E} \exp\{ i(\xi_k, z_k) \} \ .$$

Statements 1 and 2 result because C_{R^d}, and $\{\exp\{i(z, x)\} : z \in R^d\}$ are respectively complete sets.

Corollary 3.1.3. *If the R^1-variables $\xi_1, \xi_2, \dots, \xi_n$ are independent, then*
1. *their joint distribution function is the product of the separate (marginal) distribution functions of the ξ_k:*

$$F_{\xi_1,\ldots,\xi_n}(x_1,\ldots,x_n) = \mathbf{P}\{\xi_1 < x_1,\ldots,\xi_n < x_n\} = \prod_{k=1}^{n} F_{\xi_k}(x_k)\ ;$$

2. *their joint characteristic function is the product of the characteristic functions of the* ξ_k*:*

$$\varphi_{\xi_1\ldots\xi_n}(t_1,\ldots,t_n) = \mathbf{E}\exp\left\{i\sum_{k=1}^{n} t_k\xi_k\right\} = \prod_{k=1}^{n}\varphi_{\xi_k}(t_k)$$

and each of these conditions implies the independence of $\xi_1,\ldots,\xi_n$*.*

Corollary 3.1.4. *If* $\xi_1,\ldots,\xi_n$ *are independent random variables and* $\mathbf{E}|\xi_k| < \infty$*, then* $\mathbf{E}\xi_1\ldots\xi_n = \prod_{k=1}^{n}\mathbf{E}\xi_k$*.*

This assertion is easily proved by passing to the limit from bounded variables.

3.2 Sequences of Independent Random Variables

3.2.1 Sums of Independent Random Variables

Let $\xi_1,\xi_2,\ldots$ be a sequence of independent random variables in R^1. Consider the successive sums $\zeta_k = \xi_1 + \ldots + \xi_k, k = 1,2,\ldots$. The distribution function of the sum of two independent variables ξ and η is easily seen to be expressible as the convolution of their distribution functions:

$$F_{\xi+\eta}(x) = F_\xi(x) * F_\eta(x) = \int F_\xi(x-y)dF_\eta(y)\ .$$

If F_i is the distribution function of ξ_i, then the distribution function F_{ζ_k} of ζ_k is $F_1 * F_2 * \ldots * F_k$ (it is easy to show that convolution is commutative and associative). The characteristic function of a sum is even more simple to express in terms of the characteristic functions of the variables, namely,

$$\mathbf{E}\exp\left\{it\sum_{1}^{n}\xi_k\right\} = \mathbf{E}\prod_{k=1}^{n}\exp\{it\xi_k\} = \prod_{k=1}^{n}\mathbf{E}\exp\{it\xi_k\}\ .$$

That is, the characteristic function of a sum of independent random variables is the product of the characteristic functions of the terms. In exactly the same way, if the ξ_k's are nonnegative random variables and $\lambda > 0$, then

$$\mathbf{E}\exp\left\{-\lambda\sum_{1}^{n}\xi_k\right\} = \prod_{k=1}^{n}\mathbf{E}\exp\{-\lambda\xi_k\}\ .$$

The Laplace transform of a sum of independent random variables is the product of their Laplace transforms. If the ξ_k's are nonnegative integer-valued independent random variables and $|z| \le 1$, then

$$\mathbf{E}z^{\xi_1+\ldots+\xi_n} = \prod_{k=1}^{n} \mathbf{E}z^{\xi_k} \ .$$

The function $\mathbf{E}z^{\xi}$, with ξ integer-valued and nonnegative, is called the *generating function* of ξ. Thus, the generating function of a sum of independent random variables equals the product of the generating functions of the terms.

(a) *Chebyshev's inequality. The law of large numbers.* Let ξ be a random variable for which $\mathbf{E}\xi^2 < \infty$. Then for any positive c,

$$\mathbf{P}\{|\xi - \mathbf{E}\xi| > c\} \le \mathbf{V}\xi/c^2 \ . \tag{3.2.1}$$

This inequality is proved as follows. If $\eta > 0$, $\mathbf{E}\eta < \infty$ and $a > 0$, then

$$\mathbf{P}\{\eta > a\} \le a^{-1}\mathbf{E}\eta \ . \tag{3.2.2}$$

Clearly, $\eta \ge aI_{\{\eta \ge a\}}$. Taking the expectation, we arrive at (3.2.2). To deduce (3.2.1), one need only apply (3.2.2) with $\eta = (\xi - \mathbf{E}\xi)^2$ and $a = c^2$. The inequalities (3.2.1) and (3.2.2) are known as Chebyshev's inequality. Chebyshev obtained (3.2.1) for sums of independent random variables and used it to prove the next theorem known as the law of large numbers.

We need the following basic property. If ξ and η are independent, then $\mathbf{V}(\xi + \eta) = \mathbf{V}(\xi) + \mathbf{V}(\eta)$ provided the right-hand side is finite. Indeed,

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

since $\mathbf{E}\xi\eta = \mathbf{E}\xi\mathbf{E}\eta$ by the independence of ξ and η. The variance of a sum of any number of independent random variables equals the sum of their variances. This is simple to prove.

Chebyshev's Theorem. *Let $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ be independent random variables for which $\mathbf{E}\xi_n$ and $\mathbf{V}\xi_n$ exist and let $\sup_n \mathbf{V}\xi_n < \infty$. Then for any positive ϵ,*

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{1}{n}\sum_{k=1}^{n}\xi_k - \frac{1}{n}\sum_{k=1}^{n}\mathbf{E}\xi_k\right| > \varepsilon\right\} = 0 \ . \tag{3.2.3}$$

Proof. On the basis of (3.2.1), the probability in (3.2.3) can be majorized above by

$$\frac{1}{\varepsilon^2}\mathbf{V}\left(\frac{1}{n}\sum_{k=1}^{n}\xi_k\right) = \frac{1}{n^2\varepsilon^2}\sum_{k=1}^{n}\mathbf{V}\xi_k = O\left(\frac{1}{n\varepsilon^2}\right) \ .$$

□

Chebyshev's Theorem establishes that the difference between the average of random variables and the average of their expectations converges to zero in probability. Theorems that give conditions for this property to be true are called laws of large numbers. The law of large numbers asserts that the average of a large number of random variables is "practically" nonrandom. A physical illustration of the law is the constancy of the pressure of a gas on the walls of a container although this pressure is determined by the overall momentum of the molecules of gas colliding with the walls. The law of large numbers is used in precision measurements with instruments that yield random errors. By taking the average of n independent measurements, one will obtain as precise a value as desired for the quantity being measured if n is sufficiently large.

Remark. Bernoulli's theorem (p. 25) is a special case of Chebyshev's theorem. For, if the A_n are independent random events with $\mathbf{P}(A_n) = p$, and if $\xi_n = I_{A_n}$, then $\nu_n = \sum_1^n \xi_k$, $\mathbf{E}\xi_n = p$ and $\mathbf{V}\xi_n = p - p^2$. Since the variables ξ_n are independent,

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{1}{n}\nu_n - p\right| > \varepsilon\right\} = 0$$

for all positive ε.

3.2.2 Kolmogorov's Inequality

Kolmogorov's Theorem. *Let $\xi_1, \ldots, \xi_n$ be independent random variables with $\mathbf{E}\xi_k = 0$ and $\mathbf{V}\xi_k < \infty$ and let $\zeta_k = \xi_1 + \ldots + \xi_k$. Then for every positive a*

1. $\mathbf{P}\left\{\sup_{k\le n} |\zeta_k| > a\right\} \le a^{-2}\mathbf{V}\zeta_n$*; if in addition, $|\xi_k| \le c, k = 1, \ldots, n$, then*
2. $\mathbf{V}\zeta_n \le \dfrac{(a+c)^2}{\mathbf{P}\{\sup_{k\le n} |\zeta_k| \le a\}}$.

Proof. Let $\chi_1 = I_{\{|\zeta_1|>a\}}$ and $\chi_k = I_{\{|\zeta_1|\le a,\ldots,|\zeta_{k-1}|\le a,|\zeta_k|>a\}}, k = 2, \ldots, n$. Then $\sum_{k=1}^n \chi_k = I_{\{\sup_k |\zeta_k|>a\}}$. χ_k is independent of $\xi_{k+1}, \ldots, \xi_n$. Therefore

$$\mathbf{E}(\zeta_n - \zeta_k)\chi_k\zeta_k = \mathbf{E}(\zeta_n - \zeta_k)\mathbf{E}\chi_k\zeta_k = 0\ ,$$
$$\mathbf{E}(\zeta_n - \zeta_k)^2\chi_k = \mathbf{E}(\zeta_n - \zeta_k)^2\mathbf{E}\chi_k \le \mathbf{E}\chi_k \cdot \mathbf{E}\zeta_n^2\ .$$

It is also evident that

$$a\chi_k \le |\zeta_k|\chi_k \le (a + \sup_k |\zeta_k|)\chi_k\ .$$

Thus

$$\mathbf{E}\zeta_n^2 \geq \mathbf{E}\zeta_n^2 \sum_{k=1}^{n} \chi_k = \mathbf{E} \sum_{k=1}^{n} [(\zeta_n - \zeta_k)^2 \chi_k + 2(\zeta_n - \zeta_k)\chi_k \zeta_k + \zeta_k^2 \chi_k]$$
$$\geq a^2 \mathbf{E} \sum_{k=1}^{n} \chi_k = a^2 \mathbf{P} \left\{ \sup_{k \leq n} |\zeta_k| > a \right\} .$$

This yields statement 1. For 2,

$$\mathbf{E}\zeta_n^2 = \mathbf{E}\zeta_n^2 \sum_{k=1}^{n} \chi_k + \mathbf{E}\zeta_n^2 \left(1 - \sum_{k=1}^{n} \chi_k\right) \leq \sum_{k=1}^{n} (\mathbf{E}\zeta_k^2 \chi_k + \mathbf{E}\chi_k \cdot \mathbf{E}\zeta_n^2)$$
$$+ a^2 \mathbf{E} \left(1 - \sum_{k=1}^{n} \chi_k\right) \leq \mathbf{P} \left\{ \sup_{k \leq n} |\zeta_k| > a \right\} \mathbf{E}\zeta_n^2 + (a + c)^2 \sum_{k=1}^{n} \mathbf{E}\chi_k$$
$$+ a^2 \left(1 - \sum_{k=1}^{n} \mathbf{E}\chi_k\right) \leq \mathbf{E}\zeta_n^2 \mathbf{P} \left\{ \sup_{k \leq n} |\zeta_k| > a \right\} + (a + c)^2 .$$

From this we obtain statement 2. □

We now give one further inequality in which $\mathbf{P}\left\{\sup_{k \leq n} |\zeta_k| > a\right\}$ is estimated. First, there is this result.

Theorem 3.2.1. *For some α and all $k = 1, 2, \ldots, n$, let $\mathbf{P}\{\zeta_n - \zeta_k \geq \alpha\} \geq \beta > 0$. Then*

$$\mathbf{P} \left\{ \sup_{k \leq n} \zeta_k > a \right\} \leq \frac{1}{\beta} \mathbf{P}\{\zeta_n > a + \alpha\} . \tag{3.2.4}$$

Proof.

$$\mathbf{P}\{\chi_k = 1\} \leq \frac{1}{\beta} \mathbf{P}\{\chi_k = 1\} \mathbf{P}\{\zeta_n \zeta_k \geq \alpha\} = \frac{1}{\beta} \mathbf{P}\{\chi_k = 1, \zeta_n - \zeta_k \geq \alpha\}$$
$$\leq \frac{1}{\beta} \mathbf{P}\{\chi_k = 1, \zeta_n > a + \alpha\}$$

(we have made use of the independence of χ_k and $\zeta_n - \zeta_k$ in this). Summing these inequalities, we obtain (3.2.4) since

$$\sum \mathbf{P}\{\chi_k = 1, \zeta_n > a + \alpha\} \leq \mathbf{P}\{\zeta_n > a + \alpha\}$$

because the events $\{\chi_k = 1\}$ are mutually exclusive for different values of k. □

Corollaries.

1. *If the random variables ξ_k have symmetric distributions (which means that ξ_k and $-\xi_k$ have the exact same distribution), then*

$$\mathbf{P} \left\{ \sup_{k \leq n} \zeta_k > a \right\} \leq 2\mathbf{P}\{\zeta_n > a\}, \quad \mathbf{P} \left\{ \sup_{k \leq n} |\zeta_k| > a \right\} \leq 2\mathbf{P}\{|\zeta_n| > a\} .$$

2. *If* $\mathbf{P}\{|\zeta_n - \zeta_k| \le c\} \ge \beta$ *for some positive* c, *all* $k = 1,2,\dots,n$ *and positive* β, *then for* $a > c$,

$$\mathbf{P}\left\{\sup_{k\le n}|\zeta_k| > a\right\} \le \frac{1}{\beta}\mathbf{P}\{|\zeta_n| > a - c\}\ .$$

In the first case, ξ_k and $-\xi_k$ satisfy the hypotheses of the theorem with $\alpha = 0$ and $\beta = \frac{1}{2}$. In the second case, ξ_k and $-\xi_k$ satisfy the hypotheses of the theorem with $\alpha = -c$. Writing out the further relation

$$\mathbf{P}\left\{\sup_{k\le n}(-\zeta_k) > a\right\} \le \frac{1}{\beta}\mathbf{P}\{-\zeta_n > a + \alpha\}$$

and combining it with (3.2.4), we arrive at the inequality for the absolute value. □

3.2.3 Convergence of Series of Independent Random Variables

Suppose that a_n decreases monotonely to zero. Then the series $\sum(-1)^n a_n$ always converges. One could pose the following question. What can be said about the convergence of a series if the sign of the n-th term is selected at random? More precisely, this means that we want to investigate $\sum \varepsilon_k a_k$ where ε_k is a sequence of independent random variables each assuming the values 1 or -1 with probabilities $1/2$. (This is an example of a series of independent random variables.) It is natural to examine the convergence of such a series. Kolmogorv's zero-one law implies that the series either converges with probability 1 or else it diverges with probability 1. Henceforth, we shall consider $\sum_{k=1}^{\infty}\xi_k$ whose terms ξ_k are independent random variables; $\zeta_n = \sum_{k=1}^{n}\xi_k$ are its partial sums.

Theorem 3.2.2. *Suppose that* $\mathbf{E}\xi_k$ *and* $\mathbf{V}\xi_k$ *exist.*
1. *The series* $\sum_{k=1}^{\infty}\xi_k$ *is convergent with probability* 1 *if* $\sum_{k=1}^{\infty}\mathbf{E}\xi_k$ *and* $\sum_{k=1}^{\infty}\mathbf{V}\xi_k$ *are convergent.*
2. *If* $\mathbf{P}\{|\xi_k| > c\} = 0$ *for some positive* c, *then the conditions in part* 1 *are necessary for* $\sum_{k=1}^{\infty}\xi_k$ *to converge in probability (and thus also with probability* 1*).*

Proof. 1. With no loss of generality, we may assume that $\mathbf{E}\xi_k = 0$. Then for $a > 0$, we have by Kolmogorov's inequality that

$$\mathbf{P}\left\{\sup_{n<k\le m}|\zeta_k - \zeta_n| > a\right\} \le a^{-2}\sum_{k=n+1}^{m}\mathbf{V}\xi_k \le a^{-2}\sum_{k=n+1}^{\infty}\mathbf{V}\xi_k\ .$$

Letting $m \to \infty$, we find that

$$\mathbf{P}\left\{\sup_{k>n}|\zeta_k - \zeta_n| > a\right\} \le a^{-2}\sum_{k=n+1}^{\infty}\mathbf{V}\xi_k\ .$$

Thus

$$\mathbf{P}\left\{\sup_{k,l>n}|\zeta_k-\zeta_l|>2a\right\}\le\mathbf{P}\left\{\sup_{k>n}|\zeta_k-\zeta_n|>a\right\}$$
$$+\mathbf{P}\left\{\sup_{l>n}|\zeta_l-\zeta_n|>a\right\}\le 2a^{-2}\sum_{k=n+1}^{\infty}\mathbf{V}\xi_k\ .$$

Let $\eta_n=\sup_{k,l>n}|\zeta_k-\zeta_l|$. The random variables η_n decrease with n and so $\lim_{n\to\infty}\eta_n=\eta$ exists. Clearly,

$$\mathbf{P}\{\eta>2a\}\le\lim_{n\to\infty}\mathbf{P}\{\eta_n>a\}=0\ .$$

Therefore the series $\sum\xi_k$ satisfies the Cauchy convergence criterion with probability 1.

2. We now apply the *symmetrization process* which is often used in the study of sums of independent random variables. Consider along with the random variables ξ_k another sequence ξ_k' of independent random variables. The collections of variables $\{\xi_k,k=1,2,\ldots\}$ and $\{\xi_k',k=1,2,\ldots\}$ are also independent and ξ_k and ξ_k' have the exact same distribution. Such variables may be formed by taking a second copy of the original probability space $(\Omega,\mathcal{F},\mathbf{P})$. If we denote it by $(\Omega',\mathcal{F}',\mathbf{P}')$, we can investigate the product of the probability spaces $(\Omega\times\Omega',\mathcal{F}\times\mathcal{F}',\mathbf{P}\times\mathbf{P}')$. If the $\xi_k(\omega)$ are the original random variables, then they may be viewed as variables on the product space: $\xi_k(\omega,\omega')=\xi_k(\omega)$. The variables ξ_k' are then $\xi_k(\omega')$. The random variables $\tilde{\xi}_k=\xi_k-\xi_k'$ are symmetric. Clearly, $|\tilde{\xi}_k|\le 2c$, $\mathbf{E}\tilde{\xi}_k=0$ and $\mathbf{V}\tilde{\xi}_k=2\mathbf{V}\xi_k$.

By hypothesis, there is a positive r such that for all n,

$$\mathbf{P}\{|\tilde{\zeta}_n|>r\}\le\frac{1}{4}\ .$$

By Corollary 1 of Sect. 3.2.2,

$$\mathbf{P}\left\{\sup_{k\le n}|\tilde{\zeta}_k|>r\right\}\le\frac{1}{2}\ .$$

From part 2 of Kolmogorov's inequality, it follows that for all n,

$$\mathbf{V}\tilde{\zeta}_n\le 2(r+2c)^2\quad\text{and}\quad 2\sum_{k=1}^{n}\mathbf{V}\xi_k\le 2(r+2c)^2\ .$$

Thus, $\sum_{k=1}^{\infty}\mathbf{V}\xi_k<\infty$. The sequence $\xi_k-\mathbf{E}\xi_k$ satisfies the hypotheses of part 1 of the theorem and so $\sum_{k=1}^{\infty}(\xi_k-\mathbf{E}\xi_k)$ converges with probability 1 and in probability. By hypothesis, $\sum_{k=1}^{\infty}\xi_k$ also converges in probability and so the difference $\sum_{k=1}^{\infty}\mathbf{E}\xi_k$ of these two series also converges (the latter is a series of nonrandom terms; its convergence in probability is the same as ordinary convergence). □

Corollary. *The series* $\sum_{k=1}^{\infty} \varepsilon_k a_k$, *in which the* ε_k *are independent and identically distributed random variables,* $\mathbf{P}\{\varepsilon_k = \pm 1\} = \frac{1}{2}$, *and* a_k *is a bounded sequence, converges with probability* 1 *if and and only if* $\sum a_k^2 < \infty$.

Kolmogorov's Three-Series Theorem. *The series* $\sum \xi_k$ *of independent random variables is convergent if and only if for some positive* c

$$\text{(a)} \quad \sum \mathbf{P}\{|\xi_k| > c\}, \quad \text{(b)} \quad \sum \mathbf{E}\xi_k I_{\{|\xi_k| \le c\}}, \quad \text{(c)} \quad \sum \mathbf{V}\xi_k I_{\{|\xi_k| \le c\}}$$

are convergent.

Proof. If the series (a) converges, then so does

$$\sum \xi_k I_{\{|\xi_k| > c\}} \tag{3.2.5}$$

since by the Borel-Cantelli lemma it has only finitely many nonzero terms with probability 1. If the series (b) and (c) converge, then so does $\sum \xi_k I_{\{|\xi_k| \le c\}}$ by virtue of part 1 of Theorem 3.2.2. This establishes the sufficiency of the hypotheses of the theorem.

Now let $\sum \xi_k$ converge with probability 1. Then $\mathbf{P}\{\lim_{n\to\infty} \xi_n = 0\} = 1$. Thus,for any choice of positive c, only finitely many of the events $\{|\xi_k| > c\}$ occur with probability 1 and so the series (a) is convergent. But then the series (3.2.5) converges with probability 1 and hence so does the series

$$\sum \xi_k I_{\{|\xi_k| \le c\}} \ .$$

It remains to make use of part 2 of Theorem 3.2.2. □

3.2.4 The Strong Law of Large Numbers

The law of large numbers establishes the convergence in probability of the difference between the average of n random variables and the average of their expectations. If convergence in probability is replaced by convergence with probability one then the corresponding theorems are known as strong laws of large numbers.

Theorem 3.2.3. *Suppose that* ξ_k *is a sequence of independent random variables for which* $\mathbf{E}\xi_k$ *and* $\mathbf{V}\xi_k$ *exist. If* $\sum (1/k^2)\mathbf{V}\xi_k < \infty$, *then*

$$\mathbf{P}\left\{\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (\xi_k - \mathbf{E}\xi_k) = 0\right\} = 1 \ .$$

Proof. We may assume that $\mathbf{E}\xi_k = 0$. Put $\eta_n = \sup_{m \le 2^n} |\sum_{k=1}^{m} \xi_k|$. Then

$$\left| \frac{1}{m} \sum_{k=1}^{m} \xi_k \right| \le 2^{-n} \eta_{n+1}$$

for $2^n \le m \le 2^{n+1}$. We have to prove that $\lim_{n\to\infty} 2^{-n}\eta_n = 0$ with probability 1. To this end, it suffices to show that $\sum_n \mathbf{P}\{2^{-n}\eta_n > \varepsilon\}$ converges for any positive ε since by the Borel-Cantelli lemma, there then exists an N such that $2^{-n}\eta_n \le \varepsilon$ for $n > N$. On the basis of Kolmogorov's inequality,

$$\sum_n \mathbf{P}\{\eta_n > 2^n\varepsilon\} \le \sum_n \varepsilon^{-2}2^{-2n} \sum_{k\le 2^n}^{n} \mathbf{V}\xi_k$$
$$= \varepsilon^{-2}\sum_{k=1}^{\infty} \mathbf{V}\xi_k \sum_{n\ge \log_2 k} 2^{-2n} = O(\varepsilon^{-2})\sum_{k=1}^{\infty} \frac{\mathbf{V}\xi_k}{k^2} < \infty\,.$$

□

Theorem 3.2.4. *Suppose that ξ_k is a sequence of independent and identically distributed random variables. Then there exists a constant a such that*

$$\mathbf{P}\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\xi_k = a\right\} = 1 \tag{3.2.6}$$

if and only if $\mathbf{E}\xi_1$ exists and $a = \mathbf{E}\xi_1$.

Proof. Let (3.2.6) hold for some a. Then $\mathbf{P}\{\lim_{n\to\infty}\frac{1}{n}\xi_n = 0\} = 1$. By the Borel-Cantelli lemma, $\sum \mathbf{P}\{|\frac{1}{n}\xi_n| > c\} < \infty$, $c > 0$, since the events $\{|\frac{1}{n}\xi_n| > c\}$ are independent and only finitely many of them occur. Therefore,

$$\mathbf{E}|\xi_1| \le \sum_{k=1}^{\infty} ck\mathbf{E}I_{\{(k-1)c<|\xi_1|\le kc\}} = c\sum_{k=1}^{\infty} k\mathbf{P}\{(k-1)c < |\xi_1| \le kc\}$$
$$\le c\sum_{k=1}^{\infty}\mathbf{P}\{|\xi_1| > kc\} < \infty\,.$$

Now let $\mathbf{E}|\xi_1| < \infty$ We may assume that $\mathbf{E}\xi_1 = 0$. Put $\xi_n' = \xi_n I_{\{|\xi_n|\le n\}}$ and $\xi_n'' = \xi_n - \xi_n'$. Then

$$\sum \mathbf{P}\{\xi_n'' \ne 0\} = \sum \mathbf{P}\{\xi_n| > n\} < \infty$$

and beginning with some index, $\xi_n'' = 0$ and

$$\mathbf{P}\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\xi_k'' = 0\right\} = 1\,. \tag{3.2.7}$$

The hypotheses of Theorem 3.2.2 hold for ξ_n' because

$$\sum_{n=1}^{\infty}\frac{1}{n^2}\mathbf{V}\xi_n' \le \sum_{n=1}^{\infty}\frac{1}{n^2}\int_{|x|\le n} x^2\mathbf{P}\{\xi_1 \in dx\} = \int\sum_{n=1}^{\infty}\frac{x^2}{n^2}I_{\{|x|\le n\}}\mathbf{P}\{\xi_1 \in dx\}$$
$$\le c_1\int |x|\mathbf{P}\{\xi_1 \in dx\} = c_1\mathbf{E}|\xi_1|\,,$$

in which $c_1 = \sup_x |x| \sum_{n>|x|} n^{-2}$. Hence,

$$\mathbf{P}\left\{\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}(\xi'_k - \mathbf{E}\xi'_k) = 0\right\} = 1\ .$$

Since

$$\lim_{n\to\infty} \frac{1}{n}\sum_{1}^{n} \mathbf{E}\xi'_k = \lim_{n\to\infty} \int_{-n}^{n} x\mathbf{P}\{\xi_1 \in dx\} = \mathbf{E}\xi_1 = 0\ ,$$

it follows that also

$$\mathbf{P}\left\{\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \xi'_k = 0\right\} = 1\ . \tag{3.2.8}$$

The relations (3.2.7) and (3.2.8) show that (3.2.6) holds with $a = 0$. □

A most important consequence of Theorem 3.2.4, is the following refinement of Bernoulli's theorem.

Corollary. *If A_n is a series of independent events with $\mathbf{P}(A_n) = p$ and ν_n is the number of events that occur among $A_1, A_2, \ldots, A_n$, then*

$$\mathbf{P}\left\{\lim \frac{1}{n}\nu_n = p\right\} = 1,$$

that is, the relative frequency tends to the probability almost surely.

Nevertheless, the attempt to define probability as the limit of the relative frequency (the axiom of von Mises) proves to be logically untenable.

3.3 Random Walks

Let $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ be a sequence of independent and identically distributed random variables. The sequence with $\zeta_0 = x$ and $\zeta_n = x + \sum_{k=1}^{n} \xi_k, n = 1, 2, \ldots$, is called a *random walk*, x is its initial position, ξ_k is the k-th *step of the walk* and ζ_n is its position at time n (after the n-th step). A random walk is customarily interpreted as the motion of a particle taking independent random steps. It is one of the simplest stochastic processes with discrete time. One is usually interested in the behavior of this process on an infinite time interval.

3.3.1 The Renewal Scheme

Let $x = 0$ and $\mathbf{P}\{\xi_1 > 0\} = 1$. The corresponding random walk is often interpreted as follows. Let there exist a device running for a random time. As soon as it goes out of commission, it is replaced by an identical one and so

on. If 0 is the moment that the first device is switched on, ξ_1 is its running time and ξ_n is the running time of the n-th device, then it is reasonable to assume that the ξ_k's are independent and identically distributed. Thus, $\zeta_n = \xi_1 + \ldots + \xi_n$, the moment that the n-th device stops running (and the $(n+1)$st is switched on), is called a *renewal or regeneration time.*

Let $\nu(t)$ be the number of the device running at time t; $\nu(t) = n$ if $\zeta_{n-1} \le t < \zeta_n$. $N(t) = \mathbf{E}\nu(t)$ is called the *renewal function.* The basic results of renewal theory (this is the name given to the section of probability that studies the process $\nu(t)$) concern the asymptotic behavior of $N(t)$. Let $g(\lambda) = \mathbf{E}e^{-\lambda\xi_1}$ be the Laplace transform of ξ_1. It is easy to see that

$$N(t) = \sum_{n=1}^{\infty} n\mathbf{P}\{\nu(t) = n\} = \sum_{n=1}^{\infty} n\mathbf{P}\{\zeta_{n-1} \le t < \zeta_n\}$$
$$= \sum_{n=1}^{\infty} n(\mathbf{P}\{\zeta_n > t\} - \mathbf{P}\{\zeta_{n-1} > t\}) .$$

Therefore

$$\int_0^{\infty} e^{-\lambda t} N(t)dt = \sum_{n=1}^{\infty} n \int_0^{\infty} e^{-\lambda t}(\mathbf{P}\{\zeta_n > t\} - \mathbf{P}\{\zeta_{n-1} > t\})dt$$
$$= \sum_{n=1}^{\infty} n(g^{n-1}(\lambda) - g^n(\lambda))/\lambda = \frac{1}{\lambda(1-g(\lambda))} .$$

It follows from this relation that $N(t)$ is finite for all t.

Theorem 3.3.1. *Let* $\mathbf{E}\xi_1 < \infty$. *Then*

$$\lim_{t\to\infty} \frac{1}{t} N(t) = \frac{1}{\mathbf{E}\xi_1} .$$

Proof. It is easy to see that $\nu(\zeta_n + h) - \nu(\zeta_n)$ is independent of $\xi_1, \ldots, \xi_n$ and is distributed the same as $\nu(h) - \nu(0) = \nu(h) - 1$ ($\nu(\zeta_n + h) - \nu(\zeta_n)$ is the number of renewals on $]\zeta_n, \zeta_n + h]$; since the left-hand endpoint is a renewal time, this number has the same distribution as the number of renewals on $]0, h]$). If $\zeta_{k-1} \le t < \zeta_k$, then $\nu(\zeta_k) - \nu(t) = 1$ and so

$$\nu(t+h) - \nu(t) = \sum_k I_{\{\zeta_{k-1} < t \le \zeta_k\}} I_{\{\zeta_k < t+h\}} (\nu(t+h) - \nu(t))$$
$$\le \sum_k I_{\{\zeta_{k-1} < t \le \zeta_k\}} I_{\{\zeta_k < t+h\}} (\nu(t+h) - \nu(\zeta_k) + 1)$$
$$\le \nu(\zeta_k + h) - \nu(\zeta_k) + 1 .$$

Taking the expectation, we obtain $N(t+h) \le N(t) + N(h)$. From this inequality it follows that $\limsup_{t\to\infty} N(t)/t < \infty$. Let $c = \liminf_{t\to\infty} N(t)/t$.

If t_n is a sequence such that $\lim_{n\to\infty} \frac{N(t_n)}{t_n} = c$, then $\lim_{n,m\to\infty} \frac{N(mt_n)}{mt_n} \le c$ and so

$$\lim_{t\to\infty} \frac{N(t)}{t} = c\,.$$

To determine c, we use the relation

$$\int_0^\infty e^{-\lambda t} N(t)dt = \int_0^\infty e^{-u} \frac{1}{\lambda} N\left(\frac{u}{\lambda}\right) du = \frac{1}{\lambda(1-g(\lambda))}$$

or

$$\frac{\lambda}{1-g(\lambda)} = \int_0^\infty e^{-u} \lambda N\left(\frac{u}{\lambda}\right) du\,.$$

Letting $\lambda \to 0$ and noting that $\lim_{\lambda\to 0}(1-\mathbf{E}e^{-\lambda\xi_1})/\lambda = \mathbf{E}\xi_1$, we complete the proof of the theorem. □

This theorem shows that $N(t) \sim \frac{1}{\mathbf{E}\xi_1} \cdot t$, that is, the expected number of renewals in time t is asymptotically proportional to the time and inversely proportional to the mean running time of a device. A more precise study of the asymptotic behavior of $N(t)$ as $t \to \infty$ necessitates distinguishing two cases.

(a) *Arithmetic distributions.* A random variable ξ in R has an *arithmetic distribution* if $h^{-1}\xi$ is integer-valued for some h. The largest such h is the *distribution span*. If $f(z) = \mathbf{E}e^{iz\xi}$ is the characteristic function of an arithmetically distributed ξ with span h, then $2\pi h^{-1}$ is the smallest positive root of the equation $f(z) = 1$. This follows because h is the smallest positive number belonging to the minimal group G that contains all those real x for which $\mathbf{P}\{\xi = x\} > 0$.

Theorem 3.3.2. *Suppose that ξ_1 is arithmetically distributed with span h and $\mathbf{E}\xi_1 < \infty$. Then*

$$\lim_{t\to\infty} [N(t+h) - N(t)] = \frac{h}{\mathbf{E}\xi_1}$$

Proof. We shall assume that $h = 1$. Let $q_n = \sum_k \mathbf{P}\{\zeta_k = n\}, n \ge 1, q_0 = 1$. Then $N(t) = q_0 + \ldots + q_t$ for positive integral t. To prove the theorem, it suffices to show that

$$\lim_{t\to\infty} q_t = \frac{1}{\mathbf{E}\xi_1}\,.$$

Let $f(z) = \mathbf{E}e^{iz\xi_1} = \sum \mathbf{P}\{\xi_1 = n\}e^{izn}$. Then $f^k(z) = \mathbf{E}e^{iz\zeta_k} = \sum \mathbf{P}\{\zeta_k = n\}e^{izn}$. Therefore

$$q_n = \sum_{k=0}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-izn} f^k(z) dz = \sum_{k=0}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} (e^{-izn} - e^{izn}) f^k(z) dz$$
$$= \frac{1}{\pi i} \sum_{k=0}^{\infty} \int_{-\pi}^{\pi} \sin nz f^k(z) dz \, .$$

Noting that $f(z) \neq 1$ for $0 \leq |z| \leq \pi, f'(0) = i\mathbf{E}\xi_1$ and thus that $\left| \frac{\sin nz}{1-f(z)} \right|$ is bounded, we can interchange the summation and integration. Therefore

$$q_n = \frac{1}{\pi i} \int_{-\pi}^{\pi} \frac{\sin nz}{z} \cdot \frac{z}{1-f(z)} dz = \frac{1}{\mathbf{E}\xi_1} - \frac{1}{\pi i f'(0)} \int_{|z|>n\pi} \frac{\sin z}{z} dz$$
$$+ \frac{1}{\pi i} \int_{-\pi}^{\pi} \frac{\sin nz}{z} \left(\frac{z}{1-f(z)} - \frac{1}{f'(0)} \right) dz \, .$$

The last integral tends to zero if $f(z)$ is continuously differentiable and $f(z) \neq f(0)$. This is a purely analytic and easily proved fact. □

Remark. If $\mathbf{E}\xi_1 = \infty$, the theorem remains true if we put $\frac{1}{\infty} = 0$.

(b) *Nonarithmetic Distributions.* Now let ξ_1 have a *nonarithmetic distribution*, that is, it is not arithmetically distributed. If $f(z) = \mathbf{E}e^{iz\xi_1}$, then $\operatorname{Re} f(z) < 1$ for $z \neq 0$.

Theorem 3.3.3. *If ξ_1 has a nonarithmetic distribution and $\mathbf{E}\xi_1 < \infty$, then for all $u > 0$,*

$$\lim_{t\to\infty} (N(t+u) - N(t)) = \frac{u}{\mathbf{E}\xi_1} \, .$$

If $\mathbf{E}\xi_1 = \infty$, the right-hand side of this last relation is zero.

Proof. It suffices to prove that if $\psi(u)$ is any sufficiently smooth function with compact support, then

$$\lim_{t\to\infty} \int \psi(u) dN(t+u) = \lim_{t\to\infty} \int \psi(u-t) dN(u) = \frac{1}{\mathbf{E}\xi_1} \int \psi(u) du \, .$$

Now $\int \psi(u-t) dN(u) = \sum_{n=0}^{\infty} \mathbf{E}\psi(\zeta_n - t)$. If $\psi(u) = \int e^{iuv} \tilde{\psi}(v) dv$, then $\mathbf{E}\psi(\zeta_n - t) = \int e^{-ivt} \sum_{n=0}^{\infty} f^n(v) \tilde{\psi}(v) dv$. Hence,

$$\int \psi(u-t) dN(u) = \int e^{-ivt} \sum_{n=0}^{\infty} f^n(v) \tilde{\psi}(v) dv$$
$$= \int \frac{e^{-ivt} - e^{ivt}}{1-f(v)} \tilde{\psi}(v) dv = 2 \int \frac{\sin vt}{v} \cdot \frac{-iv}{1-f(v)} \tilde{\psi}(v) dv \, .$$

Again we use the fact that $\lim_{v\to 0} \frac{-iv}{1-f(v)} = \frac{1}{\mathbf{E}\xi_1}$ and that it is permissible to take the limit (which is a purely analytic problem causing no difficulties). We find that the limit of the right-hand side is

$$\frac{2\tilde{\psi}(0)}{\mathbf{E}\xi_1}\lim_{t\to\infty}\int\frac{\sin vt}{v}dv=\frac{2\pi\tilde{\psi}(0)}{\mathbf{E}\xi_1}=\frac{1}{\mathbf{E}\xi_1}\int\psi(u)du$$

since by the inversion formula,

$$\tilde{\psi}(v)=\frac{1}{2\pi}\int e^{-ivu}\psi(u)du,\tilde{\psi}(0)=\frac{1}{2\pi}\int\psi(u)du\ .$$

□

3.3.2 Recurrency

We now examine a random walk that starts at $x=0$. A random walk is called arithmetic if the step ξ_1 is arithmetically distributed. Otherwise, it is nonarithmetic. The distribution span is then the step of the random walk.

(a) *Arithmetic walks.* Consider an integer-valued walk of step 1. Define the random variable ν to be the smallest positive n for which $\zeta_n=0$. If $\zeta_n\neq 0$ for all $n>0$, then set $\nu=\infty$. A random walk is called *recurrent* if $\mathbf{P}\{\nu<\infty\}=1$. If $\mathbf{P}\{\nu=\infty\}>0$, then it is called *nonrecurrent.*

Theorem 3.3.4. *In order for a walk to be recurrent, it is necessary and sufficient that* $\sum_{n=1}^{\infty}\mathbf{P}\{\zeta_n=0\}=\infty$.

Proof. If $\zeta_n=0$, then $\nu<n$. Thus, these events satisfy the relation

$$\{\zeta_n=0\}=\bigcup_{k=1}^{n}\{\nu=k\}\cap\{\zeta_n=0\}=\bigcup_{k=1}^{n}\{\nu=k\}\cap\{\zeta_n-\zeta_k=0\}\ .$$

The events $\{\nu=k\}$ and $\{\zeta_n-\zeta_k=0\}$ are independent. Consequently,

$$\begin{aligned}\mathbf{P}\{\nu=k,\zeta_n-\zeta_k=0\}&=\mathbf{P}\{\nu=k\}\mathbf{P}\{\zeta_n-\zeta_k=0\}\\&=\mathbf{P}\{\nu=k\}\mathbf{P}\{\zeta_{n-k}=0\}\ ,\end{aligned}$$

and so

$$\mathbf{P}\{\zeta_n=0\}=\sum_{k=1}^{n}\mathbf{P}\{\nu=k\}\mathbf{P}\{\zeta_{n-k}=0\}\ .$$

Multiply this equation by λ^n with $|\lambda|<1$ and sum over n obtaining

$$\begin{aligned}\sum_{n=1}^{\infty}\lambda^n\mathbf{P}\{\zeta_n=0\}&=\sum_{n=1}^{\infty}\sum_{k=1}^{n}\lambda^k\mathbf{P}\{\nu=k\}\lambda^{n-k}\mathbf{P}\{\zeta_{n-k}=0\}\\&=\sum_{n=1}^{\infty}\lambda^n\mathbf{P}\{\nu=n\}\left(1+\sum_{n=1}^{\infty}\lambda^n\mathbf{P}\{\zeta_n=0\}\right)\end{aligned}$$

or

$$\sum_{n=1}^{\infty} \mathbf{P}\{\nu = n\}\lambda^n = \sum_{n=1}^{\infty} \mathbf{P}\{\zeta_n = 0\}\lambda^n \left(1 + \sum_{n=1}^{\infty} \lambda^n \mathbf{P}\{\zeta_n = 0\}\right)^{-1} .$$

Thus

$$\mathbf{P}\{\nu < \infty\} = \lim_{\lambda \uparrow 1} \sum_{n=1}^{\infty} \mathbf{P}\{\zeta_n = 0\}\lambda^n \left(1 + \sum_{n=1}^{\infty} \lambda^n \mathbf{P}\{\zeta_n = 0\}\right)^{-1}$$

and the limit of the right-hand side is 1 if and only if $\sum_{n=1}^{\infty} \mathbf{P}\{\zeta_n = 0\} = \infty$. □

Remark. Let $f(z) = \mathbf{E}e^{iz\xi_1}$. Then

$$\mathbf{P}\{\zeta_n = 0\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f^n(z)dz = \frac{1}{2\pi} \int_{-\pi}^{\pi} \operatorname{Re} f^n(z)dz$$

and so

$$\sum_{n=0}^{\infty} \mathbf{P}\{\zeta_n = 0\} = \lim_{\lambda \uparrow 1} \sum_{n=0}^{\infty} \lambda^n \mathbf{P}\{\zeta_n = 0\} = \lim_{\lambda \uparrow 1} \frac{1}{2\pi} \int_{-\pi}^{\pi} \operatorname{Re} \frac{1}{1 - \lambda f(z)} dz .$$

The limit may be taken under the integral sign and so a random walk is recurrent if and only if $\int_{-\pi}^{\pi} \operatorname{Re} \frac{1}{1-f(z)} dz = \infty$ (the integrand is nonnegative and finite if $z \neq 0$). A probabilistic proof exists but it is quite complicated and there is no sense in reproducing it.

(b) *Nonarithmetic walk.* In this case, a walk is said to be recurrent if $\sum_{n=1}^{\infty} I_{\{\zeta_n \in U\}} \geq 1$ almost surely for U any open set containing 0. Let ν_U be the smallest subscript for which $\zeta_{\nu_U} \in U$. $\mathbf{P}\{\nu_U < \infty\} = 1$ for a recurrent walk. Notice that

$$\sum_{n=1}^{\infty} I_{\{\zeta_n \in U\}} = \infty$$

for a recurrent walk. For, by taking V so that $x + y \in U$ if $x \in V$ and $y \in V$, one can write

$$\sum_{n=1}^{\infty} I_{\{\zeta_n \in U\}} \geq \sum_{n > \nu}^{\infty} I_{\{\zeta_n - \zeta_{\nu_V} \in V\}} + 1 ,$$

and the right-hand sum is distributed the same as $\sum_{n=1}^{\infty} I_{\{\zeta_n \in V\}}$.[1] Therefore if $\mathbf{E} \sum_{n=1}^{\infty} I_{\{\zeta_n \in U\}} < \infty$ for some open set U containing 0, the random walk is not recurrent.

[1] This is true because the event $\{\nu_n = m\}$ is independent of $\xi_{m+1}, \xi_{m+2}, \ldots$ for all m.

Assume now that a random walk is nonrecurrent. This means that there exists a neighborhood V (it may be taken to be an interval of the form $(-2\delta, 2\delta)$) for which $\mathbf{P}\left\{\sum_{n=1}^{\infty} I_{\{\zeta_n \in V\}} < \infty\right\} > 0$. Thus, $\mathbf{P}\left\{\sum_{n=1}^{\infty} I_{\{\zeta_n \in V\}} \geq r\right\} < \alpha$ for some positive r and $\alpha < 1$. If $U = (-\delta, \delta)$, then for all x

$$\mathbf{P}\left\{\sum_{n=1}^{\infty} I_{\{\zeta_n - x \in U\}} \geq r+1\right\} \leq \sum_{k=1}^{\infty} \mathbf{P}\Big\{\zeta_1 - x \notin U, \ldots, \zeta_k - x \in U,$$

$$\sum_{n=1}^{\infty} I_{\{\zeta_n - \zeta_k \in U\}} \geq r\Big\} \leq \sum_{k=1}^{\infty} \mathbf{P}\left\{\zeta - x \notin U, \ldots, \zeta_{k-1} - x \notin U, \zeta_k - x \in U\right\} \leq \alpha\ .$$

Next,

$$\mathbf{P}\left\{\sum_{n=1}^{\infty} I_{\{\zeta_n - x \in U\}} \geq 2r+2\right\} = \sum_{k=1}^{\infty} \mathbf{P}\left\{\sum_{n=1}^{k-1} I_{\{\zeta_n - x \in U\}} < r+1,\right.$$

$$\left.\sum_{n=1}^{k} I_{\{\zeta_n - x \in U\}} = r+1,\ \sum_{n=k+1}^{\infty} I_{\{\zeta_n - x \in U\}} \geq r+1\right\}$$

$$= \sum_{k=1}^{\infty} \mathbf{E} I_{\left\{\sum_{n=1}^{k-1} I_{\{\zeta_n - x \in U\}} < r+1 = \sum_{n=1}^{k} I_{\{\zeta_n - x \in U\}}\right\}}$$

$$\times \mathbf{P}\left\{\sum_{n=k+1}^{\infty} I_{\{\zeta_n - \zeta_k + \zeta_k - x \in U\}} \geq r+1 | \zeta_k\right\}$$

since $\zeta_n - \zeta_k$ is independent of ζ_k. The last probability does not exceed α and so

$$\mathbf{P}\left\{\sum_{n=1}^{\infty} I_{\{\zeta_n - x \in U\}} \geq 2r+2\right\}$$

$$\leq \alpha \sum_{k=1}^{\infty} \mathbf{P}\left\{\sum_{n=1}^{k-1} I_{\{\zeta_n - x \in U\}} \leq r+1 = \sum_{n=1}^{k} I_{\{\zeta_n - x \in U\}}\right\} \leq \alpha^2\ .$$

In similar fashion, we can establish that

$$\mathbf{P}\left\{\sum_{n=1}^{\infty} I_{\{\zeta_n - x \in U\}} \geq mr + m\right\} \leq \alpha^m\ .$$

Therefore

$$\mathbf{E} \sum_{n=1}^{\infty} I_{\{\zeta_n \in U\}} \leq (r+1) \sum_{m} m \alpha^{m-1} < \infty\ .$$

We have proved the following statement.

Theorem 3.3.5. *A random walk is not recurrent if and only if* $\sum_n \mathbf{P}\{\zeta_n \in U\} < \infty$ *for some open set* U *containing* 0.

It is easy to see that this also holds for any bounded open set U.

Remark. As in the arithmetic case, it is possible to express this condition in terms of the characteristic function of the step. For recurrency, it is necessary and sufficient that

$$\int_{-\delta}^{\delta} \operatorname{Re} \frac{1}{1-f(z)} dz = \infty$$

for all positive δ.

3.3.3 Ladder Functionals

Let $\{\zeta_n\}$ with $\zeta_0 = 0$ be a random walk and let $a \geq 0$. Define $\tau_a = \inf\{n > 0 : \zeta_n > a\}$ (if the set is empty, we take inf $= \infty$). If $\tau_a < \infty$, let $\gamma_a = \zeta_{\tau_a} - a$. The variable τ_a is called the *overshoot time* of level a and γ_a is the *overshoot* of level a. They are also called *ladder functionals*. We shall study their joint distribution. Observe that $\mathbf{P}\{\tau_a = k\} = \mathbf{P}\{\zeta_1 \leq a, \ldots, \zeta_{k-1} \leq a, \zeta_k > a\}$. Let $F(y)$ be the distribution function of ξ_1. Then

$$\begin{aligned}\mathbf{P}\{\tau_a = k\} &= \int_{-\infty}^{a} dF(y)\mathbf{P}\{\zeta_2 \leq a, \ldots, \zeta_{k-1} \leq a, \zeta_k > a | \zeta_1 = y\} \\ &= \int_{-\infty}^{a} dF(y)\mathbf{P}\{\zeta_2 - \zeta_1 \leq a - y, \ldots, \zeta_{k-1} - \zeta_1 \leq a - y, \\ &\zeta_k - \zeta_1 > a - y\} = \int_{-\infty}^{a} dF(y)\mathbf{P}\{\tau_{a-y} = k-1\}\end{aligned}$$

(we have used the fact that $\{\zeta_{n-1} - \zeta_1, n \geq 2\}$ is also a random walk with steps $\xi_2, \xi_3, \ldots$). Let $|\lambda| < 1$ and $Q(\lambda, a) = \sum_{k=1}^{\infty} \lambda^k \mathbf{P}\{\tau_a = k\}$. Then

$$Q(\lambda, a) = \lambda \mathbf{P}\{\xi_1 > a\} + \lambda \int_{-\infty}^{a} Q(\lambda, a-y) dF(y), \quad a \geq 0\,.$$

For $a < 0$, define $Q(\lambda, a) = 0$. This leads to the system of equations

$$\left.\begin{array}{l} Q(\lambda, a) = \lambda(1 - F(a+)) + \lambda \int Q(\lambda, a-y) dF(y), \quad a \geq 0 \\ Q(\lambda, a) = 0, \quad a < 0 \end{array}\right\}\,. \tag{3.3.1}$$

The first is a *convolution equation on the half-line*. Below we describe Wiener's method for solving such an equation.

Let $\varepsilon(z) = 1$ if $z \geq 0$ and $\varepsilon(z) = 0$ if $z < 0$. We can rewrite (3.3.1) in the form $\varepsilon(z)\lambda(1 - F(z+)) = \varepsilon(z) \int Q(\lambda, z-y) d(\varepsilon(y) - \lambda F(y))$. Consider the convolution of this equation with a function $v_1(t)$ of bounded variation such that $v_1(t) = v_1(0)$ for $t > 0$. We obtain

$$\lambda \int \varepsilon(z-t)[1-F((z-t)+)]dv_1(t)$$
$$= \int \varepsilon(z-t) \int Q(\lambda, z-t-y)d(\varepsilon(y)-\lambda F(y))dv_1(t)\ . \quad (3.3.2)$$

Since $\varepsilon(z-t)dv_1(t) = dv_1(t)$ for positive z, (3.3.2) leads to the equation

$$\lambda\Phi_1(z) = \int Q(\lambda, z-y)dv_2(y)\ , \quad (3.3.3)$$

in which $\lambda\Phi_1(z)$ is the left-hand side of (3.3.2) and $v_2(y) = \int[\varepsilon(y-t)-\lambda F(y-t)]dv_1(t)$. If $v_2(t)$ is constant for negative t, then equation (3.3.3) can be solved by using Fourier transforms. Let

$$\tilde{\Phi}_1(\mu) = \int_0^\infty e^{i\mu z}d\Phi_1(z),\ \tilde{Q}(\lambda,\mu) = \int_0^\infty e^{i\mu z}dQ(\lambda,z),\ \tilde{v}_2(\mu) = \int_0^\infty e^{i\mu z}dv_2(z)$$

(since $Q(\lambda, a) = \mathbf{E}\lambda^{\tau_a}$ and τ_a increases with $a, Q(\lambda, a)$ is monotone in a for $0 < \lambda < 1$). From (3.3.3) we obtain

$$\lambda\tilde{\Phi}_1(\mu) = \tilde{Q}(\lambda,\mu)\tilde{v}_2(\mu), \tilde{Q}(\lambda,\mu) = \lambda\tilde{\Phi}_1(\mu)/\tilde{v}_2(\mu)\ . \quad (3.3.4)$$

To determine the functions $\tilde{v}_1(\mu)$ and $\tilde{v}_2(\mu)$, we make use of the equation

$$\tilde{v}_2(\mu) = (1-\lambda f(\mu))\tilde{v}_1(\mu)\ , \quad (3.3.5)$$

in which $f(\mu) = \int e^{i\mu y}dF(y)$ is the characteristic function of the step. Let $F_n(y)$ be the distribution function of ζ_n. Then

$$(1-\lambda f(\mu))^{-1} = \exp\left\{\sum_{n=1}^\infty \frac{\lambda^n}{n} f^n(\mu)\right\} = \exp\left\{\sum_{n=1}^\infty \frac{\lambda^n}{n}\int e^{i\mu y}dF_n(y)\right\}$$
$$= \exp\left\{\sum_{n=1}^\infty \frac{\lambda^n}{n}\int_{-\infty}^0 e^{i\mu y}dF_n(y)\right\} \Big/ \exp\left\{-\sum_{n=1}^\infty \frac{\lambda^n}{n}\int_0^\infty e^{i\mu y}dF_n(y)\right\}\ .$$

Thus (3.3.5) will hold if we put

$$\tilde{v}_1(\mu) = \exp\left\{\sum_{n=1}^\infty \frac{\lambda^n}{n}\int_{-\infty}^0 e^{i\mu y}dF_n(y)\right\}$$

and

$$\tilde{v}_2(\mu) = \exp\left\{-\sum_{n=1}^\infty \frac{\lambda^n}{n}\int_0^\infty e^{i\mu y}dF_n(y)\right\}\ .$$

These are functions of bounded variation with the required properties since $\tilde{v}_k(\mu) = \int e^{i\mu t}dv_k(t)$ with[2]

[2] $a \wedge b = \min(a,b)$; $a \vee b = \max(a,b)$.

$$v_k(t) = \varepsilon(t) + \sum \frac{1}{n} w_k^{*n}(t),$$

$$w_1(t) = \sum_{n=1}^{\infty} \frac{\lambda^n}{n} F_n(t \wedge 0), \quad w_2(t) = \sum_{n=1}^{\infty} \frac{\lambda^n}{n} [F_n(t) - F_n(0)] I_{\{t \geq 0\}} ,$$

and w^{*n} is the n-fold convolution of w with itself. Let $v_+(\lambda, t)$ be defined by its Fourier transform

$$\int e^{i\mu t} dv_+(\lambda, t) = \frac{1}{\tilde{v}_2(\mu)} = \exp\left\{ \sum_{n=1}^{\infty} \frac{\lambda^n}{n} \int_0^{\infty} e^{i\mu t} dF_n(t) \right\} . \tag{3.3.6}$$

Then from (3.3.4) we find that

$$Q(\lambda, x) = \lambda \int_0^x \Phi_1(x - t) dv_+(\lambda, t) . \tag{3.3.7}$$

Or if we substitute the value of $\Phi_1(x)$ and recall that $\varepsilon(z - t) dv_1(t) = dv_1(t)$ for $z \geq 0$,

$$Q(\lambda, x) = \lambda \int_0^x \int_{-\infty}^{\infty} [1 - F((x - t - z)+) dv_-(\lambda, z) dv_+(\lambda, t) , \tag{3.3.8}$$

where $v_-(\lambda, t) = v_1(t)$.

The expression on the right-hand side of (3.3.8) may be transformed by using

$$\lambda \int (1 - F(t - z)) dv_1(z)$$

$$= (\lambda - 1) v_1(0) + \int (\varepsilon(t - z) - \lambda F(t - z)) dv_1(z) = v_2(z),$$

$$(1 - \lambda) v_1(0) = (1 - \lambda f(0)) \tilde{v}_1(0) = \tilde{v}_2(0),$$

$$\int_0^x v_2(x - t) dv_+(\lambda, t) = \varepsilon(x)$$

for positive t to obtain

$$Q(\lambda, x) = \varepsilon(x) - \tilde{v}_2(0) v_+(\lambda, x) .$$

This formula leads to the following expression for $\tilde{Q}(\lambda, \mu)$:

$$\tilde{Q}(\lambda, \mu) = 1 - \exp\left\{ \sum_{n=1}^{\infty} \frac{\lambda^n}{n} \int_0^{\infty} (e^{i\mu x} - 1) dF_n(x) \right\} . \tag{3.3.9}$$

Now consider the function

$$Q_1(\lambda, x, y) = \sum_{k=1}^{\infty} \lambda^n \mathbf{P}\{\tau_x = n, \gamma_x > y\}, \quad x \geq 0 .$$

Extending its definition so that $Q_1(\lambda, x, y) = 0$ for $x < 0$, we find similarly to Eq. (3.3.1) that

$$Q_1(\lambda, x, y) = \lambda(1-F((x+y)+))+\lambda \int Q_1(\lambda, x-z, y)dF(z), \quad x \geq 0 \,. \quad (3.3.10)$$

This equation differs from (3.3.1) in its free term. Its solution is given by (3.3.7) with $\Phi_1(x-t)$ replaced by $\Phi_1(x+y-t)$. Therefore

$$Q_1(\lambda, x, y) = \lambda \int_0^x \int [1 - F((x+y-t-z)+)]dv_-(\lambda, z)dv_+(\lambda, z) \,. \quad (3.3.11)$$

(a) *Semibounded walks.* The zero-one law implies that the $\sup_n \zeta_n$ is finite either with probability 0 or probability 1. In the latter instance, the random walk is bounded from above. If $\mathbf{P}\{\inf_n \zeta_n > -\infty\} = 1$, then the walk is bounded from below.

Theorem 3.3.6. *A random walk is bounded from above if and only if*

$$\sum_{n=1}^{\infty} \frac{1}{n}\mathbf{P}\{\zeta_n > 0\} < \infty \,. \quad (3.3.12)$$

Proof. Formula (3.3.9) is true for complex μ with $\operatorname{Im} \mu \geq 0$. We replace μ in it by $i\mu$ with $\mu > 0$. This gives

$$\int_0^{\infty} e^{-\mu x}dQ(\lambda, x) = 1 - \exp\left\{\sum_{n=1}^{\infty} \frac{\lambda^n}{n} \int_0^{\infty} (e^{-\mu x} - 1)dF_n(x)\right\}.$$

Since $Q(\lambda, 0-) = 0$ by definition and the integrals are over regions that include zero, the left-hand integral is equal to $Q(\lambda, 0)+\int_{0+}^{\infty} e^{-\mu x}dQ(\lambda, x)$. This integral approaches zero as $\mu \to \infty$. Therefore

$$\lim_{\lambda \uparrow 1} \lambda^{\tau_0} = I_{\{\tau_0 < \infty\}}, \ \mathbf{P}\{\tau_0 < \infty\} = 1 - \exp\left\{-\sum_{n=1}^{\infty} \frac{1}{n}\mathbf{P}\{\zeta_n > 0\}\right\} \,. \quad (3.3.13)$$

Consider the pairs $(\tau_0, \gamma_0), (\tau_0^{(1)}, \gamma_0^{(1)}), \ldots$ defined sequentially as follows: if $\tau_0 < \infty$, then $\tau_0^{(1)}$ and $\gamma_0^{(1)}$ are the time and size of the first overshoot of zero by the random walk $\zeta_n^{(2)} = \zeta^{(1)}_{\tau_0^{(1)}+n} - \zeta^{(1)}_{\tau_0^{(1)}}$ and so on. It is easy to show that the distribution of the pair $\tau_0^{(k)}$, $\gamma_0^{(k)}$, if it is determined under the condition that $(\tau_0, \gamma_0), \ldots, (\tau_0^{(k-1)}, \gamma_0^{(k-1)})$ have been specified, coincides with the distribution of (τ_0, γ_0). Let condition (3.3.12) hold. Then on the basis of (3.3.13),

$$\mathbf{P}\left\{\sup_n \zeta_n < \infty\right\} > \mathbf{P}\{\tau_0 = \infty\} = \exp\left\{-\sum_{n=1}^{\infty} \frac{1}{n}\mathbf{P}\{\zeta_n > 0\}\right\} > 0$$

and thus $\mathbf{P}\{\sup_n \zeta_n < \infty\} = 1$.

Now let the series in (3.3.12) be divergent. Then $\mathbf{P}\{\tau_0 < \infty\} = 1$ and all $\tau^{(k)}$ and $\gamma^{(k)}$ are well defined, independent and identically distributed and $\mathbf{P}\{\gamma_0^{(k)} > 0\} = 1$. Therefore

$$\mathbf{P}\left\{\sup_n \zeta_n = \infty\right\} = \mathbf{P}\left\{\sum_{k=0}^{\infty} \gamma_0^{(k)} = \infty\right\} = 1, \quad \gamma_0^{(0)} = \gamma_0 \, .$$

This follows by the three-series theorem since $\mathbf{P}\{\gamma_0^{(k)} > c\} = \mathbf{P}\{\gamma_0 > c\} > 0$ for some positive c and hence $\sum \mathbf{P}\{\gamma_0^{(k)} > c\} = \infty$. □

3.4 Processes with Independent Increments

3.4.1 Definition

Let $T \subset R$. A process $\xi(t)$, $t \in T$, with values in a measurable linear space $(X, \mathcal{B})$ is called a *process with independent increments* if for any choice of $t_0 < t_1 < \ldots < t_n$ in T, the random variables $\xi(t_0), \xi(t_1) - \xi(t_0), \ldots, \xi(t_n) - \xi(t_{n-1})$ with values in X are mutually independent. A linear space X with σ-algebra $\mathcal{B}$ is said to be measurable if the mapping of $X \times X$ into X, defined by the sum $x + y$, is measurable. The difference of two random variables will, also be a random variable. A random walk with $T = \{0, 1, \ldots\}$ exemplifies a process with independent increments. We shall concentrate on real-valued processes.

Introduce the characteristic functions $f(t, z) = \mathbf{E} \exp\{iz\xi(t)\}$ and $g(s, t, z) = \mathbf{E} \exp\{iz(\xi(t) - \xi(s))\}$, $t, s \in T$ and $s < t$. They satisfy the following: 1. $f(t, z) = g(s, t, z) f(s, z)$ for $s < t$; 2. $g(s, u, z) = g(s, t, z) g(t, u, z)$ for $s < t < u$. The characteristic functions $f(t, z)$ and $g(s, t, z)$ determine the finite-dimensional distributions of the process. If $t_1 < t_2 < \ldots < t_n$, then

$$\begin{aligned}
\mathbf{E} \exp\left\{i \sum_{k=1}^{n} z_k \xi(t_k)\right\} &= \mathbf{E} \exp\Big\{iz_n(\xi(t_n) - \xi(t_{n-1})) \\
&\quad + i(z_n + z_{n-1})(\xi(t_{n-1} - \xi(t_{n-2})) + \ldots + i(z_1 + \ldots + z_n)\xi(t_1)\Big\} \\
&= f(t_1, z_1 + \ldots + z_n) g(t_1, t_2, z_2 + \ldots + z_n) \ldots g(t_{n-1}, t_n, z_n) \, .
\end{aligned}$$

Thus describing processes with independent increments reduces to describing the above characteristic functions. We shall assume that T is R_+. Evidently, $\xi_{(0)}$ may be anything. Therefore the problem becomes one of finding $g(s, t, z)$.

(a) *Discrete processes with independent increments.* Let $\{t_k, k = 1, 2, \ldots\}$ be a sequence in R_+. Let ξ_k^+ and ξ_k^- be two independent sequences of independent random variables such that the series

$$\sum I_{\{t_k \le t\}} \xi_k^- \quad \text{and} \quad \sum I_{\{t_k < t\}} \xi_k^+$$

converge for all $t \in R_+$ and their sums are independent of the order of the terms (by analyzing the proof of the three-series theorem, one can convince oneself that the latter will hold if

$$\sum I_{\{t_k \le t\}} (|\mathbf{E}\xi_k^- I_{\{|\xi_k^-| \le c\}}| + |\mathbf{E}\xi_k^+ I_{\{|\xi_k^+| \le c\}}|) < \infty \Big) .$$

The process

$$\xi'(t) = \sum I_{\{t_k \le t\}} \xi_k^- + \sum I_{\{t_k < t\}} \xi_k^+$$

has independent increments. A sequence $n_m \to \infty$ may be selected so that the process

$$\xi'_m(t) = \sum_{k \le n_m} I_{\{t_k \le t\}} \xi_k^- + \sum_{k \le n_m} I_{\{t_k < t\}} \xi_k^+$$

converges uniformly with probability 1 as $m \to \infty$. To this end, it suffices that

$$\sum_m \mathbf{P} \left\{ \sup_{t \le a_m} |\xi'_m(t) - \xi'_{m+1}(t)| > \frac{1}{m^2} \right\} < \infty$$

for some sequence $a_m \uparrow \infty$. The probability being summed can be estimated using the inequalities in Sect. 3.2.3. If $\xi'(t)$ is understood to be the limit of $\xi'_m(t)$ ($\xi'(t)$ is determined up to modification; see Chap. 4, Sect. 4.1), then $\xi'(t)$ is continuous everywhere except at the points $t_k, k = 1, 2, \ldots$. In this connection, $\xi_k^+ = \xi'(t_k+) - \xi'(t_k)$ and $\xi_k^- = \xi'(t_k) - \xi'(t_k-)$.

(b) *Stochastically continuous processes.* A stochastic process $\xi(t)$ is *stochastically continuous* at t_0 if $\xi(t_n) \to \xi(t_0)$ in probability as $t_n \to t_0$. Let $\xi(t)$ be defined for $t \in [a, b]$ and stochastically continuous at each point. Then given any $\varepsilon > 0$ and $\rho > 0$, there exists a $\delta > 0$ such that $\mathbf{P}\{|\xi(t_1) - \xi(t_2)| > \varepsilon\} < \rho$ for $|t_1 - t_2| < \delta$. This property is termed uniform stochastic continuity. If this were not so, then there would exist two sequences t'_n and t''_n with $t'_n - t''_n \to 0$ such that $\mathbf{P}\{|\xi(t'_n) - \xi(t''_n)| > \varepsilon\} \ge \rho$, for some $\varepsilon > 0$ and $\rho > 0$. With no loss of generality, we could assume that $t'_n \to t_0$ and $t''_n \to t_0$ with $t_0 \in [a, b]$. But then $\xi(t'_n) \to \xi(t_0)$, $\xi(t''_n) \to \xi(t_0)$ and so $\xi(t''_n) - \xi(t'_n) \to 0$ in probability contradicting the assumption.

(c) *Lévy's decomposition.*

Theorem 3.4.1. *Let $\xi(t)$ be any process with independent increments. There exists a nonrandom function $a(t)$ such that $\xi(t) = a(t) + \xi'(t) + \xi''(t)$, where $\xi'(t)$ and $\xi''(t)$ are independent processes with independent increments, $\xi'(t)$ being discrete and $\xi''(t)$ being stochastically continuous.*

Proof. Let $a(t) = \tan(\mathbf{E} \arctan \xi(t))$. Then one can show that the function

$$g_1(s, t, z) = \exp\{-i(a(t) - a(s))z\} g(s, t, z)$$

has right-hand and left-hand limits in s and in t. In other words, it has at most jump discontinuities. If $\{t_k\}$ are all the discontinuities, then $g(t_k-, t_k, z)$ and $g(t_k, t_k+, z)$ will be characteristic functions of certain variables. These variables can be defined to be the limits of $\xi(t_k) - \xi(t_k - h)$ and $\xi(t_k + h) - \xi(t_k)$ in probability as $h \to 0$. Denoting the respective limits by ξ_k^- and ξ_k^+, one can show that they are independent random variables. If $\xi'(t)$ is the discrete process formed from these variables, then $\xi''(t) = \xi(t) - \xi'(t)$ will be a stochastically continuous process with independent increments that is independent of $\xi'(t)$. □

3.4.2 Stochastically Continuous Processes

Let $\xi(t)$ be a stochastically continuous process on R_+. Let D_+ be the set of all nonnegative binary rational numbers. It can be shown that $\lim \xi(t_n)$ exists with probability 1 for every monotone decreasing sequence $t_n \in D_+$. Since the process is stochastically continuous, this limit equals $\xi(\lim t_n)$ with probability 1. Extending $\xi(t)$ from D_+ in the stated way, we arrive at a right-continuous modification $\xi(t)$. It has a left-hand limit at each point so that it will be a process with at most jump discontinuities.

Let $x(t)$ be a numerical-valued function defined on the set T. We shall say that $x(t)$ has k ε-oscillations if there exist $t_0 < t_1 < \ldots < t_k$ in T such that $|x(t_i) - x(t_{i-1})| \geq \varepsilon$, $i = 1, \ldots, k$, and no $k+2$ points exist with the same property. The function $x(t)$ has at most jump discontinuities if and only if it has finitely many ε-oscillations for all positive ε.

Lemma 3.4.1. *Let $\xi(t)$ be a process with independent increments defined on $\Lambda_n = \{t_1, \ldots, t_n\}$ with $t_l < \ldots < t_n$, and let $\mathbf{P}\{|\xi(t_n) - \xi(t_k)| \geq \frac{\varepsilon}{4}\} \leq \alpha < \frac{1}{2}$ for all k. If ν_ε is the number of ε-oscillations of $\xi(t)$ on Λ_n, then $\mathbf{E}\nu_\varepsilon \leq \alpha/(1-2\alpha)$.*

Proof. The event $\{\nu_\varepsilon \geq m\}$ implies one of the events $\bigcap_{j=1}^{k-1}\{|\xi(t_j)-\xi(t_1)| < \frac{\varepsilon}{2}\} \cap \{|\xi(t_k)-\xi(t_1)| \geq \frac{\varepsilon}{2}\} \cap$ {the number of ε-oscillations of $\xi(t)$ on $\Lambda_n \cap [t_k, \infty[\geq m-1\}$. These events are mutually exclusive, the last event in the intersection is independent of the first k events and its probability does not exceed $\mathbf{P}\{\nu_\varepsilon \geq m-1\}$. Thus

$$\mathbf{P}\{\nu_\varepsilon \geq m\} \leq \mathbf{P}\{\nu_\varepsilon \geq m-1\}\mathbf{P}\left\{\sup_k |\xi(t_k) - \xi(t_1)| \geq \frac{\varepsilon}{2}\right\}$$
$$\leq \mathbf{P}\Big\{\nu_\varepsilon \geq m-1\Big\}\frac{\alpha}{1-\alpha}$$

(we have made use of Corollary 2 of Sect. 3.2.2). Hence

$$\mathbf{P}\{\nu_\varepsilon \geq m\} \leq \left(\frac{\alpha}{1-\alpha}\right)^m, \quad \mathbf{E}\nu_\varepsilon \leq \sum_{m=1}^{\infty}\left(\frac{\alpha}{1-\alpha}\right)^m \leq \frac{\alpha}{1-2\alpha}\,.$$

□

Applying this lemma, we can say the following. If h is chosen so that $\mathbf{P}\{|\xi(s)-\xi(t)|\geq\frac{\varepsilon}{4}\}<\frac{1}{3}$ for $|s-t|<h$ and $s,t\leq T$, then $\xi(t)$ has finitely many ε-oscillations on $D_+\cap[0,T]$ with probability 1 (since this will hold for any finite subset of $[kh,(k+1)h]$ and the expected number of ε-oscillations is uniformly bounded).

Lemma 3.4.2. *Let* $\xi_\varepsilon(t)=\sum_{s\leq t}(\xi(s)-\xi(s-))I_{\{|\xi(s)-\xi(s-)|>\varepsilon\}}$ *and* $\bar{\xi}_\varepsilon(t)=\xi(t)-\xi_\varepsilon(t)$. *Then* $\xi_\varepsilon(t)$ *and* $\bar{\xi}_\varepsilon(t)$ *are mutually independent processes with independent increments.*

Proof. Let $\mathcal{F}_t^s$ be the σ-algebra generated by $\xi(u)-\xi(s)$ with $s\leq u\leq t$. For $t_0<t_1<\ldots<t_n$, the σ-algebras $\mathcal{F}_{t_1}^{t_0},\mathcal{F}_{t_2}^{t_1},\ldots,\mathcal{F}_{t_n}^{t_{n-1}}$ are independent. Clearly, $\xi_\varepsilon(t)-\xi_\varepsilon(s)$ and $\bar{\xi}_\varepsilon(t)-\bar{\xi}_\varepsilon(s)$ are $\mathcal{F}_t^s$-measurable. Therefore to complete the proof, it suffices to show that $\xi_\varepsilon(t)-\xi_\varepsilon(s)$ and $\bar{\xi}_\varepsilon(t)-\bar{\xi}_\varepsilon(s)$ are independent. With no loss of generality, we may assume that $s=0$ and $\xi(0)=0$. Let us prove that $\xi_\varepsilon(t)$ and $\bar{\xi}_\varepsilon(t)$ are independent.

Let $t_{nk}=\frac{k}{n}t$. For all $\varepsilon>0$, with the possible exception of a countable set, $\xi_\varepsilon(t)=\lim_{n\to\infty}\sum_{k=1}^n\xi_{nk}$, where $\xi_{nk}=(\xi(t_{nk})-\xi(t_{nk-1}))I_{\{|\xi(t_{nk})-\xi(t_{nk-1})|>\varepsilon\}}$ and $\bar{\xi}_\varepsilon(t)=\lim_{n\to\infty}\sum_{k=1}^n\eta_{nk}$, where $\eta_{nk}=\xi(t_{nk})-\xi(t_{nk-1})-\xi_{nk}$. Therefore

$$\begin{aligned}
&\left|\mathbf{E}e^{iu\xi_\varepsilon(t)+iv\bar{\xi}_\varepsilon(t)}-\mathbf{E}e^{iu\xi_\varepsilon(t)}\mathbf{E}e^{iv\bar{\xi}_\varepsilon(t)}\right|\\
&\quad=\left|\prod_{k=1}^n\mathbf{E}e^{iu\xi_{nk}+iv\eta_{nk}}-\prod_{k=1}^n\mathbf{E}e^{iu\xi_{nk}}\right|\\
&\quad\leq\sum_{k=1}^n|\mathbf{E}e^{iu\xi nk+iv\eta_{nk}}-\mathbf{E}e^{iu\xi_{nk}}e^{iv\eta_{nk}}|\\
&\quad=\sum_{k=1}^n|\mathbf{E}e^{iu\xi_{nk}}+\mathbf{E}e^{iv\eta_{nk}}-1-\mathbf{E}e^{iu\xi_{nk}}\mathbf{E}e^{iv\eta_{nk}}|\\
&\leq\sum_{k=1}^n|\mathbf{E}e^{iu\xi_{nk}}-1|\cdot|\mathbf{E}e^{iv\eta_{nk}}-1|
\end{aligned}$$

($e^{iu\xi_{nk}+iv\eta_{nk}}=e^{iu\xi_{nk}}+e^{iv\eta_{nk}}-1$ since $\xi_{nk}\eta_{nk}=0$). The last sum is majorized by

$$\begin{aligned}
&\sup_k|\mathbf{E}e^{iv\eta_{nk}}-1|\sum_{k=1}^n|\mathbf{E}e^{iu\xi_{nk}}-1|\\
&\leq 2\sup_k|\mathbf{E}\exp\{iv(\xi(t_{nk})-\xi(t_{nk-1})\}-1|\sum_{k=1}^n\mathbf{P}\{|\xi(t_{nk})-\xi(t_{nk-1})|>\varepsilon\}.
\end{aligned}$$

Lemma 3.4.1 can be applied to obtain a bound for the sum that is uniform in n and the factor in front of the sum approaches zero by the uniform stochastic continuity of $\xi(t)$. □

Corollary. *Let* $0 < \varepsilon_n < \varepsilon_{n-1} < \ldots < \varepsilon_1$. *Then* $\bar{\xi}_{\varepsilon_n}(t), \xi_{\varepsilon_n}(t) - \xi_{\varepsilon_{n-1}}(t), \ldots, \xi_{\varepsilon_2}(t) - \xi_{\varepsilon_1}(t)$ *and* $\xi_{\varepsilon_1}(t)$ *are independent processes with independent increments.*

Proof. $\bar{\xi}_{\varepsilon_n}(t)$ is independent of $\xi_{\varepsilon_n}(t)$ and thus also of $\xi_{\varepsilon_n}(t) - \xi_{\varepsilon_{n-1}}(t), \ldots, \xi_{\varepsilon_1}(t)$, which are expressible in terms of it. This is because $\xi_{\varepsilon_k}(t) - \xi_{\varepsilon_{k-1}}(t)$ is the sum of the jumps in $\xi(t)$ that occur up to time t inclusively and their absolute values lie in the interval $]\varepsilon_k, \varepsilon_{k-1}]$. If $\bar{\xi}(t) = \xi_{\varepsilon_k}(t)$, then $\bar{\xi}_{\varepsilon_{k-1}}(t) = \xi_{\varepsilon_k}(t) - \xi_{\varepsilon_{k-1}}(t)$. This fact can be used to show that $\xi_{\varepsilon_k}(t) - \xi_{\varepsilon_{k-1}}(t), \xi_{\varepsilon_{k-1}}(t) - \xi_{\varepsilon_{k-2}}(t), \ldots, \xi_{\varepsilon_1}(t)$ are independent processes for all values of k. □

Lemma 3.4.3. $\mathbf{E}(\bar{\xi}_\varepsilon(t) - \bar{\xi}_\varepsilon(0))$ *and* $\mathbf{V}(\bar{\xi}_\varepsilon(t) - \bar{\xi}_\varepsilon(0))$ *exist for any* $\varepsilon > 0$. *A sequence* $\varepsilon_n \downarrow 0$ *may be chosen so that for all* t *the processes* $\bar{\xi}_{\varepsilon_n}(t) - \bar{\xi}_{\varepsilon_n}(0) - \mathbf{E}(\bar{\xi}_{\varepsilon_n}(t) - \bar{\xi}_{\varepsilon_n}(0))$ *converge uniformly to a process* $\xi_0(t)$ *with probability* 1. *This process is continuous and has independent increments.*

Proof. Assume that $\xi(0) = 0$. Let the η_{nk} be the variables introduced in the proof of Lemma 3.4.2. The symmetrization method and part 2 of Kolmogorov's theorem imply that $\mathbf{E}\bar{\xi}_\varepsilon(t)$ and $\mathbf{V}\bar{\xi}_\varepsilon(t)$ exist. Now, $\bar{\xi}_{\varepsilon_1}(t) = \bar{\xi}_{\varepsilon_2}(t) - (\xi_{\varepsilon_2}(t) - \xi_{\varepsilon_1}(t))$ if $\varepsilon_2 < \varepsilon_1$ and the terms on the right-hand side are independent. Thus $\mathbf{V}\bar{\xi}_{\varepsilon_1}(t) = \mathbf{V}\bar{\xi}_{\varepsilon_2}(t) + \mathbf{V}(\xi_{\varepsilon_2}(t) - \xi_{\varepsilon_1}(t))$ and hence $\mathbf{V}\bar{\xi}_\varepsilon(t)$ is decreasing together with ε. Choose the sequence ε_n so that

$$\sum n^4 [\mathbf{V}\bar{\xi}_{\varepsilon_n}(n) - \mathbf{V}\bar{\xi}_{\varepsilon_{n+1}}(n)] < \infty$$

(this is possible because $\mathbf{V}\bar{\xi}_\varepsilon(t) - \mathbf{V}\xi_{\varepsilon'}(t) \to 0$ for all t as ε and $\varepsilon' \to 0$). Then by Kolmogorov's inequality,

$$\begin{aligned} &\mathbf{P}\left\{ \sup_{t \le n} |\bar{\xi}_{\varepsilon_{n+1}}(t) - \mathbf{E}\bar{\xi}_{\varepsilon_{n+1}}(t) + \mathbf{E}\bar{\xi}_{\varepsilon - n}(t)| > \frac{1}{n^2} \right\} \\ &\quad \le n^4 \mathbf{V}(\bar{\xi}_{\varepsilon_{n+1}}(t) - \bar{\xi}_{\varepsilon_n}(t)) = n^4 \mathbf{V}\bar{\xi}_{\varepsilon_n}(n) - \mathbf{V}\bar{\xi}_{\varepsilon_{n+1}}(n)) \,, \end{aligned}$$

Kolmogorov's inequality is employed first for values of $t = \frac{k}{2^m}, k \le n2^m$, and then let $m \to \infty$ (the processes $\bar{\xi}_\varepsilon(t)$ are right-continuous and the supremum over binary rationals t equals the supremum over all t).

Now apply the Borel-Cantelli lemma: starting with some subscript n, $\sup_{t \le n} |\bar{\xi}_{\varepsilon_{n+1}}(t) - \mathbf{E}\bar{\xi}_{\varepsilon_{n+1}} + \mathbf{E}\bar{\xi}_{\varepsilon_n}(t) - \bar{\xi}_{\varepsilon_n}(t)| \le 1/n^2$ and so the series $\sum_n (\bar{\xi}_{\varepsilon_{n+1}}(t) - \mathbf{E}\bar{\xi}_{\varepsilon_{n+1}}(t) + \mathbf{E}\bar{\xi}_{\varepsilon_n}(t) - \bar{\xi}_{\varepsilon_n}(t))$ is uniformly convergent. This implies the uniform convergence of the sequence $\bar{\xi}_{\varepsilon_n}(t) - \mathbf{E}\bar{\xi}_{\varepsilon_n}(t)$ on each finite interval. Consequently, the limiting process $\xi_0(t)$ has no jumps greater than ε_n no matter what n is. Thus, the process $\xi_0(t)$ is continuous. It has independent increments, being the limit of such processes, and since $\bar{\xi}_{\varepsilon_n}(t)$ is independent of $\xi_\varepsilon(t)$ if $\varepsilon_n < \varepsilon$ the limit $\xi_0(t)$ will also not depend on $\xi_\varepsilon(t)$. □

Corollary. *With probability* 1, $\lim_{n\to\infty}[\xi_{\varepsilon_n}(t)+\mathbf{E}(\xi_{\varepsilon_1}(t)-\xi_{\varepsilon_n}(t))]$ *exists uniformly on each bounded interval. If this limit is* $\xi'(t)$, *then* $\xi(t)=\xi_0(t)+\xi'(t)+\mathbf{E}\bar{\xi}_{\varepsilon_1}(t)$. *The process* $\xi'(t)$ *is referred to as the discontinuous part of* $\xi(t)$.

Since $\xi_0(t)$ is the limit of random variables with zero expectations and bounded variances, it follows that $\mathbf{E}\xi_0(t)=0$. The stochastic continuity of $\bar{\xi}_{\varepsilon_1}(t)$ and boundedness of $\mathbf{E}|\bar{\xi}_{\varepsilon_1}(t)|^2$ imply that $\mathbf{E}\bar{\xi}_{\varepsilon_1}(t)$ is continuous.

We have obtained a decomposition into a sum of a continuous process with zero mean, a discontinuous part and a nonrandom continuous function.

3.4.3 Lévy's Formula

We now find an expression for the characteristic function of a stochastically continuous process with independent increments. We shall assume that $\xi(0)=0$.

(a) *Poisson process.* Let $\xi(t)$ be a process having a finite number of jumps on each bounded interval. It is constant between discontinuities and each jump is of size 1. $\xi(t)$ assumes only nonnegative integer values. Certainly, $\xi(t)$ has a Poisson distribution: there exists a nondecreasing continuous function $\lambda(t)$ with $\lambda(0)=0$ such that

$$\mathbf{P}\{\xi(t)=m\}=\frac{(\lambda(t))^m}{m!}e^{-\lambda(t)}\,. \tag{3.4.1}$$

Let us prove this.

The stochastic continuity implies that $p_0(t)=\mathbf{P}\{\xi(t)=0\}$ is a continuous function. Since $p_0(s+t)=p_0(s)\times\mathbf{P}\{\xi(s+t)-\xi(s)=0\}$ and the probability on the right is nonvanishing for t sufficiently small, the set $\{t: p_0(t)=0\}$ is both open and closed. But $p_0(0)=1$ and so this set is empty. The function $\lambda(t)=-\ln p_0(t)$ is defined for all t, it is nonnegative and continuous and it is nondecreasing since $p_0(t)$ is nondecreasing. With this choice of $\lambda(t)$, formula (3.4.1) holds for $m=0$. Observe that

$$\begin{aligned}\mathbf{P}\{\xi(t)-\xi(s)>0\}&=1-e^{-(\lambda(t)-\lambda(s))}\\&=\lambda(t)-\lambda(s)+O((\lambda(t)-\lambda(s))^2),\quad t>s\,.\end{aligned}$$

If $t_{nk}=\frac{k}{n}t$, then $\mathbf{P}\{\sup_{k\le n}(\xi(t_{nk})-\xi(t_{nk-1}))\le 1\}\to 1$ as $n\to\infty$ and hence $\prod_{k=1}^n(1-\mathbf{P}\{\xi(t_{nk})-\xi(t_{nk-1})>1\})\to 1$.

This last relation implies that

$$\lim_{n\to\infty}\sum_{k=1}^n\mathbf{P}\{\xi(t_{nk})-\xi(t_{nk-1})>1\}=0\,. \tag{3.4.2}$$

For $m>0$, it now follows that

$$\mathbf{P}\{\xi(t)=m\}=\lim_{n\to\infty}\sum_{k=1}^{n}\mathbf{P}\{\xi(t_{nk-1})=m-1\}\mathbf{P}\{\xi(t_{nk})-\xi(t_{nk-1})=1\}$$
$$+O\left(\sum_{k=1}^{n}\mathbf{P}\{\xi(t_{nk})-\xi(t_{nk-1})>1\}\right)=\lim_{n\to\infty}\sum_{k=1}^{n}\mathbf{P}\{\xi(t_{nk-1})=m-1\}$$
$$\times\mathbf{P}\{\xi(t_{nk+1})-\xi(t_{nk})\geq 1\}=\lim_{n\to\infty}\left(\sum_{k=1}^{n}\mathbf{P}\{\xi(t_{nk-1})=m-1\}\right)$$
$$\times[\lambda(t_{nk+1})-\lambda(t_{nk})]+O\left(\sum_{k=1}^{n}(\lambda(t_{nk+1})-(t_{nk})\right)^2$$
$$=\int_0^t\mathbf{P}\{\xi(t)=m-1\}d\lambda(t)\ .$$

Formula (3.4.1) is obtained from this by induction.

(b) *Brownian motion (Wiener process).* We now consider a continuous process $\xi_0(t)$ with independent increments. It is known by the two different names. Using the fact that $\mathbf{P}\{\sup_k|\xi_0(t_{nk})-\xi_0(t_{nk-1})|<\varepsilon\}\to 1$ for every positive ε, one can show that

$$\lim_{n\to\infty}\sum_{k=1}^{n}\mathbf{P}\{|\xi_0(t_{nk})-\xi_0(t_{nk-1})|>\varepsilon\}=0 \tag{3.4.3}$$

(the derivation of (3.4.3) is analogous to that of (3.4.2)). Let $\varepsilon_n\to 0$ be chosen so that (3.4.3) holds with ε replaced by ε_n. Then writing

$$\eta_{nk}=[\xi_0(t_{nk}-\xi_0(t_{nk-1})]I_{\{|\xi_0(t_{nk}-\xi_0(t_{nk-1}|\leq\varepsilon_n\}}\ ,$$

we have

$$\mathbf{P}\left\{\xi_0(t)-\sum_{k=1}^{n}\eta_{nk}\neq 0\right\}\leq\sum_{k=1}^{n}\mathbf{P}\{|\xi_0(t_{nk})-\xi_0(t_{nk-1})|>\varepsilon_n\}\to 0$$

and hence

$$\mathbf{E}\ \ e^{iz\xi_0(t)}=\lim_{n\to\infty}\prod_{k=1}^{n}\mathbf{E}e^{iz\eta_k}=\lim_{n\to\infty}e^{iz\sum\mathbf{E}\eta_{nk}}\prod_{k=1}^{n}\mathbf{E}e^{iz(\eta_{nk}-\mathbf{E}\eta_{nk})}$$
$$=\lim_{n\to\infty}\exp\left\{iz\sum_{k=1}^{n}\mathbf{E}\eta_{nk}\right\}\prod_{k=1}^{n}\left(1-\frac{z^2}{2}\mathbf{V}\eta_{nk}(1+O(\varepsilon_n)\right)$$
$$=\lim_{n\to\infty}\exp\left\{iz\sum_{k=1}^{n}\mathbf{E}\eta_{nk}-\frac{z^2}{2}\sum_{k=1}^{n}\mathbf{V}\eta_{nk}\right\}(1+O(\varepsilon_n)+O(\max_k\mathbf{V}\eta_{nk}))\ .$$

From the existence of the limit, it follows that it has the form $\exp\{iza(t)-\frac{z^2}{2}b(t)\}$. In other words $\xi_0(t)$ is normally distributed for all $t>0$ with mean

$a(t)$ and variance $b(t)$. Now $\xi_0(t) - \xi_0(s)$ also is normally distributed with mean $a(t) - a(s)$ and variance $b(t) - b(s)$, $t > s$. Thus $b(t)$ is nondecreasing and the stochastic continuity of $\xi_0(t)$ implies the continuity of $a(t)$ and $b(t)$.

The standard Wiener process (or simply Wiener process) is the process $w(t)$ for which $a(t) = 0$ and $b(t) = t$. The process $w(t)$ is continuous with independent increments and $w(0) = 0$. For $t \geq 0$ and $h > 0$, the increment $w(t+h) - w(t)$ is normally distributed with mean 0 and variance h.

(c) *Jump component.* Let A be a Borel set in R lying at a positive distance from the origin. Let $\xi(A,t)$ denote the sum of the jumps of a process which happen up to time t inclusively with values lying in A. Let $\nu(A,t)$ be the number of such jumps. Then as functions of t, both processes are stochastically continuous with independent increments. The process $\nu(A,t)$ is Poisson. Put $\Pi(t,A) = \mathbf{E}\nu(A,t)$. $\Pi(t,A)$ is a nondecreasing continuous function of t and a measure with respect to A. If $A_1, A_2, \ldots, A_n$ are pairwise disjoint, then the processes $\xi(A_1,t), \xi(A_2,t), \ldots, \xi(A_n,t)$ are independent (this is proved in exactly the same way as Lemma 3.4.2 and its Corollary). Hence, $\nu(A_1,t), \nu(A_2,t), \ldots, \nu(A_n,t)$ are also independent. Let Δ be a bounded closed interval not containing the origin. Let $\Delta = \bigcup_{k=1}^{n} \Delta_{nk}$, where the Δ_{nk} are pairwise disjoint and $\max_k \operatorname{diam} \Delta_{nk} \to 0$ as $n \to \infty$, and $x_{nk} \in \Delta_{nk}$. Then

$$\xi(\Delta,t) = \lim_{n\to\infty} \sum x_{nk}\nu(\Delta_{nk},t) \ ,$$

and

$$\begin{aligned} \mathbf{E}e^{iz\xi(\Delta,t)} &= \lim_{n\to\infty} \prod_{k=1}^{n} \exp\{(e^{izx_{nk}} - 1)\Pi(t,\Delta_{nk})\} \\ &= \exp\left\{\int_{\Delta} (e^{izx} - 1)\Pi(t,dx)\right\} \ . \end{aligned} \tag{3.4.4}$$

(We have made use of the fact that the characteristic function of a Poisson distribution with parameter a is $\exp\{a(e^{iz} - 1)\}$.) Differentiating (3.4.4) and setting $z = 0$, we find that

$$\mathbf{E}\xi(\Delta,t) = \int_{\Delta} x\Pi(t,dx)$$

and

$$\mathbf{V}\xi(\Delta,t) = \int_{\Delta} x^2\Pi(t,dx)$$

Therefore

$$\mathbf{E}[\xi_{\varepsilon_n}(t) - \xi_{\varepsilon_{n-1}}(t)] = \int_{\varepsilon_n < |x| \leq \varepsilon_{n-1}} x\Pi(t,dx),$$

and

$$\mathbf{V}[\xi_{\varepsilon_n}(t) - \xi_{\varepsilon_{n-1}}(t)] = \int_{\varepsilon_n < |x| \leq \varepsilon_{n-1}} x^2\Pi(t,dx) \ ,$$

(d) *General form of a characteristic function.* By the uniform boundedness of $\mathbf{V}[\xi_{\varepsilon_1}(t) - \xi_{\varepsilon_n}(t)]$,

$$\int_{0<|x|<\varepsilon_1} x^2 \Pi(t, dx) < \infty$$

for all positive ε_1. Let $\varepsilon_1 = 1$. Lemma 3.4.3 and its Corollary may be utilized to obtain the following formula for the characteristic function of $\xi(t)$:

$$\mathbf{E}e^{iz\xi(t)} = \exp\left\{iza(t) - \frac{1}{2}z^2 b(t)\right\} + \int_{|x|>1} (e^{izx} - 1)\Pi(t, dx) + \int_{|x|\le 1} (e^{izx} - 1 - izx)\Pi(t; dx)\,. \tag{3.4.5}$$

This is *Lévy's formula.* The characteristic function for an increment in the process can be deduced from (3.4.5) by way of division. Thus, we have proved the following.

Theorem 3.4.2. *Let $\xi(t)$ be a stochastically continuous process with independent increments and $\xi(0) = 0$. There exist (i) a continuous function $a(t)$, (ii) a nondecreasing continuous function $b(t)$, (iii) a function $\Pi(t, A)$, which is a σ-finite measure on A satisfying $\int (x^2/(1+x^2))\Pi(t, dx) < \infty$ and nondecreasing and continuous in t for A bounded away from 0, such that* (3.4.5) *holds.*

Remark. A process $\xi(t), t \in R_+$, is said to have stationary independent increments if $\xi(0) = 0$ and the distribution of $\xi(t+h) - \xi(t)$ is independent of t for all positive h. For this process, the functions $a(t), b(t)$ and $\Pi(t, A)$ in (3.4.5) are proportional to t. Therefore there exist positive a and b and a measure $\Pi(dx)$ satisfying $\int x^2/(1+x^2)\Pi(dx) < \infty$ such that the characteristic function of $\xi(t)$ is expressible in the form

$$\mathbf{E}e^{iz\xi(t)} = \exp\left\{t\left[iaz - \frac{b}{2}z^2 + \int (e^{izx} - 1 - izxI_{\{|x|\le 1\}})\Pi(dx)\right]\right\}\,. \tag{3.4.6}$$

3.5 Product Measures

3.5.1 Definition

Let $(X_n, \mathcal{B}_n), n = 1, 2, \ldots$, be a sequence of measurable spaces and let μ_n be a probability measure on $\mathcal{B}_n$. The product of the measures μ_n, denoted by $\mu = \times_{n=1}^{\infty}\mu_n$, is called the *product measure.* It is defined on the product $\prod_{n=1}^{\infty} X_n$ of the spaces with the σ-algebra $\otimes_{n=1}^{\infty}\mathcal{B}_n$. The latter is the smallest σ-algebra containing the cylinder sets of the form $C = \{(x_1, x_2, \ldots) : x_1 \in A_1, \ldots, x_k \in A_k\}$ with $A_i \in \mathcal{B}_i$. In addition,

$$\mu(C) = \mu_1(A_1) \dots \mu_k(A_k) \,. \tag{3.5.1}$$

The existence of a product measure, that is, a measure satisfying (3.5.1) for all k and $A_i \in \mathcal{B}_i$ with identical $(X_n, \mathcal{B}_n)$, follows by Kolmogorov's theorem, which carries over to this case with slight changes. To consider a product measure is equivalent to considering a sequence of independent random elements $\xi_n(\omega)$ assuming values in the respective spaces $(X_n, \mathcal{B}_n)$. Of interest is a probability measure in a measurable space $(Y, \mathcal{C})$ for which there exists a measurable mapping of $(Y, \mathcal{C})$ into $(\prod_{n=1}^{\infty} X_n, \otimes_{n=1}^{\infty} \mathcal{B}_n)$ which carries that measure into a product measure.

(a) *Gaussian measures.* Consider a Gaussian measure in a separable Hilbert space $(H, \mathcal{B}_H)$. This is a measure which is specified by the characteristic functional $\varphi(z) = \exp\{i(a,z) - \frac{1}{2}(Bz,z)\}$, where $a \in H$ and B is a kernel operator from H to H. Let a_k be a sequence in H and consider the mapping from H to R^{∞} defined by $x \to ((x,a_1),(x,a_2),\dots)$. If x is a random variable with the distribution μ, then the joint distribution of the variables $(x,a_1),\dots,(x,a_n)$ is given by their joint characteristic function

$$\int \exp\left\{ i \sum_{k=1}^{n} s_k(x,a_k) \right\} \mu(dx)$$

$$= \exp\left\{ i \sum_{k=1}^{n} s_k(a,a_k) - \frac{1}{2} \sum_{k,l=1}^{n} s_k s_l (Ba_k, a_l) \right\}, s_k \in R \,.$$

From this formula, it follows that $(x,a_1),(x,a_2),\dots$ are independent if $B(a_k,a_l) = 0$ for $k \neq l$. This is possible to have, for instance, by choosing the a_k's to be eigenvectors of the operator B. This choice is not unique since any sequence may be orthogonalized by the Gram- Schmidt process with (Bx,y) viewed as the inner product. With such a choice of a_k, the specified mapping (it is linear) sends μ into a product measure on R^{∞}. It is expressible in the form $\times_{n=1}^{\infty} \mu_n$, where μ_n is a Gaussian measure on R with mean (a_n,a_n) and variance (Ba_n,a_n).

3.5.2 Absolute Continuity and Singularity of Measures

Let μ and ν be two given measures on a measurable space $(X, \mathcal{B})$ (only the case of finite measures will be of interest to us). The measure ν is said to be absolutely continuous with respect to μ if $\nu(A) = 0$ for all $A \in \mathcal{B}$ for which $\mu(A) = 0$. A measure ν is singular with respect to μ (in which case μ is also singular with respect to ν) if a set $S \in \mathcal{B}$ exists such that $\mu(S) = 0$ and $\nu(X \setminus S) = 0$. The following is a well-known result of measure theory.

Radon-Nikodym Theorem. *If ν is absolutely continuous with respect to μ, then there exists a $\mathcal{B}$-measurable μ-integrable function $f(x)$ such that for all $A \in \mathcal{B}$,*

$$\nu(A) = \int_A f(x)\mu(dx) \; ; \tag{3.5.2}$$

the representation (3.5.2) *is sufficient for* ν *to be absolutely continuous with respect to* μ.

The function f is called the density of ν or derivative of ν with respect to μ and is denoted by $\frac{d\nu}{d\mu}(x)$. If $f > 0$ almost everywhere with respect to μ, then μ is also absolutely continuous with respect to ν and $\frac{d\mu}{d\nu} = f^{-1}$. The measures ν and μ are then said to be equivalent. The relation (3.5.2) is equivalent to the following:

$$\int g(x)\nu(dx) = \int g(x)f(x)\mu(dx) \tag{3.5.3}$$

with $g(x)$ any bounded measurable function. The notation $\nu \ll \mu$ is used if ν is absolutely continuous with respect to μ and $\nu \perp \mu$ if the measures are mutually singular (they are then also called orthogonal).

If μ, and ν are any finite measures on $\mathcal{B}$, then the following representation always holds:

$$\nu(A) = \int_A f(x)\mu(dx) + \nu(A \cap S) \; , \tag{3.5.4}$$

where f is integrable with respect to μ and $\mu(S) = 0$. The function f is again called the derivative of ν with respect to μ and is denoted by $\frac{d\nu}{d\mu}$. The representation (3.5.4) gives a decomposition of ν into an absolutely continuous component and a singular component with respect to μ (the Jordan decomposition).

3.5.3 Kakutani's Theorem

Let $X = X_1 \times X_2, \mathcal{B} = \mathcal{B}_1 \otimes \mathcal{B}_2$, $\mu = \mu_1 \times \mu_2$, and $\nu = \nu_1 \times \nu_2$, where μ_i and ν_i are measures on $\mathcal{B}_i$. In order for $\nu \ll \mu$, it is necessary and sufficient that $\nu_i \ll \mu_i$ with

$$\frac{d\nu}{d\mu}(x_1, x_2) = \frac{d_{\nu_1}}{d_{\mu_1}}(x_1)\frac{d_{\nu_2}}{d_{\mu_2}}(x_2), \quad x_i \in X_i \; .$$

This statement is a simple consequence of Fubini's theorem and formula (3.5.3) with $g(x)$ the function $g(x_1, x_2) = g(x_1)g(x_2)$ and the fact that such functions form a complete set. The above statement clearly carries over to any finite number of factors. Kakutani's theorem concerns two infinite products of measures, that is, product measures.

Kakutani's Theorem. *Given measures* $\mu = \times_{n=1}^{\infty}\mu_n$ *and* $\nu = \times_{n=1}^{\infty}\nu_n$ *on* $(\prod_{n=1}^{\infty} X_n, \otimes_{n=1}^{\infty}\mathcal{B}_n)$ *with* μ_n *and* ν_n *probability measures on* $\mathcal{B}_n$. *In order for* $\nu \ll \mu$, *it is necessary and sufficient that the following two conditions hold:*
(a) $\nu_n \ll \mu_n$ *for all* n *and*
(b) *the numerical infinite product*

$$\prod_{n=1}^{\infty} \int \left(\frac{d\nu_n}{d\mu_n} \right)^{1/2} d\mu_n \tag{3.5.5}$$

should converge.

Proof. The necessity of (a) follows because for every n, the measure μ may be expressed as the product of $\mu_1 \times \ldots \times \mu_n$ and $\bigcap_{k=n+1}^{\infty} \mu_k$. The same is true for ν. If $\nu \ll \mu$, then so is $\nu_1 \times \nu_2 \times \ldots \times \nu_n \ll \mu_1 \times \mu_2 \ldots \times \mu_n$. (b) Let $\nu \ll \mu$. Then for all n,

$$\frac{d\nu}{d\mu}(x_1, x_2, \ldots) = \frac{d\nu_1}{d\mu_1}(x_1) \ldots \frac{d\nu_n}{d\mu_(x_n)} \rho_{n+1}(x_{n+1}, \ldots) \ ,$$

where

$$\rho_{n+1}(x_{n+1}, \ldots) = \frac{d \overset{\infty}{\underset{k=n+1}{\times}} \nu_k}{d \overset{\infty}{\underset{k=n+1}{\times}} \mu_k}(x_{n+1}, \ldots) \ .$$

It is easy to show that $\int \rho_{n+1}(x_{n+1}, \ldots) d\mu = 1$ and that $\lim_{n\to\infty} \rho_{n+1}(x_{n+1}, \ldots)$ exists almost everywhere with respect to μ and is a constant by virtue of the zero-one law. Therefore

$$\frac{d\nu}{d\mu}(x_1, \ldots) = \prod_{n=1}^{\infty} \frac{d\nu_n}{d\mu_n}(x_n) \ . \tag{3.5.6}$$

The integral $\int \left(\frac{d\nu}{d\mu}(x_1, \ldots) \right)^{1/2} d\mu$ coincides with $\lim_{n\to\infty} \int \left(\frac{d\nu_1}{d\mu_1}(x_1) \ldots \frac{d\nu_n}{d\mu_n}(x_n) \right)^{1/2} d\mu, \ldots, d\mu_n$ and so with (3.5.5) (taking the limit under the integral sign is permissible because the integral of the limiting integrand squared equals 1). The necessity of the hypotheses of the theorem has been proved.

Sufficiency. Suppose that we can establish the μ-convergence of the infinite product on the right-hand side of (3.5.6) and the relation

$$\lim_{n\to\infty} \int \prod_{k=n+1}^{\infty} \frac{d\nu_k}{d\mu_k}(x_k) d\mu = 1 \ . \tag{3.5.7}$$

Then if $g(x_1, \ldots, x_m)$ is any bounded measurable function, we obtain

$$\begin{aligned}
&\int g(x_1,\dots,x_m)\prod_{k=1}^{\infty}\frac{d\nu_k}{d\mu_k}(x_k)d\mu \\
&\quad=\int g(x_1,\dots,x_m)\prod_{k=1}^{m}\frac{d\nu_k}{d\mu_k}(x_k)d\mu_1\dots d\mu_m \\
&\quad\times\int\prod_{k=m+1}^{n}\frac{d\nu_k}{d\mu_k}(x_k)d\mu_{m+1}\dots d\mu_n\int\prod_{k=n+1}^{\infty}\frac{d\nu_k}{d\mu_k}(x_k)d\left(\times_{k=n+1}^{\infty}\mu_k\right) \\
&\quad=\int g(x_1,\dots,x_m)d\nu_1\dots d\nu_m\int\prod_{k=n+1}^{\infty}\frac{d\nu_k}{d\mu_k}(x_k)d\mu\ .
\end{aligned}$$

Letting $n\to\infty$, we find that

$$\begin{aligned}
\int g(x_1,\dots,x_m)\prod_{k=1}^{\infty}\frac{d\nu_k}{d\mu_k}(x_k)d\mu &= \int g(x_1,\dots,x_m)d\nu_1\dots d\nu_m \\
&= \int g(x_1,\dots,x_m)d\nu\ ,
\end{aligned}$$

that is, the right-hand side of (3.5.6) equals $\frac{d\nu}{d\mu}$.

We now show that $\prod_{n=1}^{\infty}\left(\frac{d\nu_n}{d\mu_n}(x_n)\right)^{\frac{1}{2}}$ converges in the mean-square. Let $m<n$. Then

$$\begin{aligned}
&\int\left(\prod_{k=1}^{m}\left(\frac{d\nu_k}{d\mu_k}(x_k)\right)^{1/2}-\prod_{k=1}^{n}\left(\frac{d\nu_k}{d\mu_k}(x_k)\right)^{1/2}\right)^2 d\mu \\
&\quad=\int\prod_{k=1}^{m}\frac{d\nu_k}{d\mu_k}(x_k)\left(1+\prod_{k=m+1}^{n}\frac{d\nu_k}{d\mu_k}(x_k)-2\prod_{k=m+1}^{n}\left(\frac{d\nu_k}{d\mu_k}(x_k)\right)^{1/2}\right)d\mu \\
&\quad=2-2\prod_{k=m+1}^{n}\int\left(\frac{d\nu_k}{d\mu_k}(x_k)\right)^{1/2}d\mu\to 0
\end{aligned}$$

as n and $m\to\infty$. By Fatou's lemma,

$$\int\prod_{k=m+1}^{\infty}\frac{d\nu_k}{d\mu_k}(x_k)d\mu\le\lim_{n\to\infty}\int\prod_{n=m+1}^{n}\frac{d\nu_k}{d\mu_k}(x_k)d\mu\le 1\ ,$$

and by Cauchy's inequality,

$$\begin{aligned}
\int\prod_{k=m+1}^{\infty}\frac{d\nu_k}{d\mu_k}(x_k)d\mu &\ge\left(\int\prod_{k=m+1}^{\infty}\left(\frac{d\nu_k}{d\mu_k}(x_k)\right)^{1/2}d\mu\right)^2 \\
&=\prod_{k=m+1}^{\infty}\left(\int\left(\frac{d\nu_k}{d\mu_k}(x_k)\right)^{1/2}d\mu_k\right)^2\ ,
\end{aligned}$$

and the expression on the right approaches 1 as $m \to \infty$. The last equality follows because the integral of the functions $\prod_{k=m+1}^{\infty} \left(\frac{d\nu_k}{d\mu_k}(x_k)\right)^{1/2}$ squared with respect to μ equals 1. Therefore these functions are uniformly integrable and so taking the limit under the integral sign is permissible. The sufficiency of (3.5.7) has been established and thus the right-hand side of (3.5.6) equals $\frac{d\nu}{d\mu}$. □

3.5.4 Absolute Continuity of Gaussian Product Measures

For all n, let $X_n = R$ and $\mathcal{B}_n = \mathcal{B}_R$ and let μ_n and ν_n be Gaussian measures with means a_n and $\tilde{a}_n$ and variances b_n and $\tilde{b}_n$. Then

$$\frac{d\nu_n}{d\mu_n}(x) = \sqrt{\frac{b_n}{\tilde{b}_n}} \exp\left\{\frac{x^2}{2}\left(\frac{1}{b_n} - \frac{1}{\tilde{b}_n}\right) + \left(\frac{\tilde{a}_n}{\tilde{b}_n} - \frac{a_n}{b_n}\right)x + \frac{a_n^2}{2b_n} - \frac{\tilde{a}_n^2}{2\tilde{b}_n}\right\} .$$

Integration yields

$$\int \left(\frac{d\nu_n}{d\mu_n}(x)\right)^{1/2} d\mu_n = \left(\frac{2b_n^{1/2}\tilde{b}_n^{1/2}}{b_n + \tilde{b}_n}\right) \exp\left\{-\frac{(a_n - \tilde{a}_n)^2}{4(b_n + \tilde{b}_n)}\right\} ,$$

The product $\prod_{n=1}^{\infty} \int \left(\frac{d\nu_n}{d\mu_n}\right)^{1/2} d\mu_n$ is convergent if and only if

$$\sum_{n=1}^{\infty} \left(\frac{b_n - \tilde{b}_n}{b_n}\right)^2 < \infty, \quad \sum_{n=1}^{\infty} \frac{(a_n - \tilde{a}_n)^2}{b_n} < \infty .$$

Let $\tilde{\mu}$ and $\tilde{\nu}$ be Gaussian measures in H with charactistic functionals

$$\varphi(x) = \exp\left\{i(a, z) - \frac{1}{2}(Bz, z)\right\}, \quad \tilde{\varphi}(z) = \exp\left\{i(\tilde{a}, z - \frac{1}{2}(\tilde{B}z, z)\right\} .$$

Choose a sequence of vectors $c_k \in H$ such that $(Bc_k, c_j) = 0$ and $(\tilde{B}c_k, c_j) = 0$ for $k \neq j$. Then the mapping $x \to ((x, c_1), (x, c_2), \ldots)$ sends H into R^{∞} and $\tilde{\mu}$ and $\tilde{\nu}$ into Gaussian product measures $\mu = \times_{n=1}^{\infty}\mu_n$ and $\nu = \times_{n=1}^{\infty}\nu_n$; μ_n has mean (c_n, a) and variance (Bc_n, c_n), ν_n has mean $(c_n, \tilde{a})$ and variance $(\tilde{B}c_n, c_n)$ and $\nu \ll \mu$ if

$$\sum_{n=1}^{\infty} \left(\frac{(Bc_n, c_n) - (\tilde{B}c_n, c_n)}{(Bc_n, c_n)}\right)^2 < \infty, \quad \sum_{n=1}^{\infty} \frac{(a - \tilde{a}, c_n)^2}{(Bc_n, c_n)} < \infty . \quad (3.5.8)$$

The next result is a consequence of these two conditions and the fact that $\nu \ll \mu$ if and only if $\tilde{\nu} \ll \tilde{\mu}$.

Theorem. *$\tilde{\nu} \ll \tilde{\mu}$ if and only if there exists a symmetric Hilbert-Schmidt operator D such that $\tilde{B} - B = B^{\frac{1}{2}}DB^{\frac{1}{2}}$ and $a - \tilde{a} = B^{\frac{1}{2}}d$ with $d \in H$.*

Deriving these conditions from (3.5.8) is a purely technical matter.

4

General Theory of Stochastic Processes and Random Functions

4.1 Regular Modifications

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, Θ a parameter set and $(X, \mathcal{B})$ a measurable space. Consider a random function $x(\theta, \omega)$ defined on Θ with values in $(X, \mathcal{B})$. If $x^*(\theta, \omega)$ is another such function such that

$$\mathbf{P}\{x^*(\theta, \omega) = x(\theta, \omega)\} = 1, \quad \theta \in \Theta ,$$

then $x^*(\theta, \omega)$ and $x(\theta, \omega)$ are said to be stochastically equivalent. One also says that $x^*(\theta, \omega)$ and $x(\theta, \omega)$ are modifications of each other (or modifications of one and the same process). From the second standpoint, a random variable is a class of functions $\{\xi(\omega)\}$ with any two of its representatives $\xi_1(\omega)$ and $\xi_2(\omega)$ satisfying $\mathbf{P}\{\xi_1(\omega) = \xi_2(\omega)\} = 1$. When viewed as functions of θ, modifications may be essentially different.

Example. Suppose that $\Theta = R$, τ is a random variable in R with absolutely continuous distribution and A is a Borel set in R of Lebesgue measure 0. Define $\xi_1(t) = I_A(t - \tau)$ and $\xi_2(t) = 0$. Then $\mathbf{P}\{\xi_1(t) = 0\} = 1$ and thus $\xi_1(t)$ and $\xi_2(t)$ are stochastically equivalent. But $\xi_2(t)$ is a continuous process with $\mathbf{P}\{\sup_t \xi_2(t) = 0\} = 1$, while $\xi_1(t)$ has points of discontinuity, the boundary A' of A (possibly R), and $\mathbf{P}\{\sup_t \xi_1(t) = 1\} = 1$ providing A is not empty.

The second process is obviously a more natural modification (for instance, it is continuous). The question arises at to whether a given random function has a modification that possesses pre-assigned regularity properties (continuity, monotonicity, at most jump discontinuities, differentiability and so on). A reasonable way to answer the question is in terms of the finite-dimensional distributions of the process. We now make the statement of the problem more precise.

Let X^θ be the space of all functions from Θ to X, $\mathcal{C}(X, \Theta)$ the cylinder σ-algebra of subsets of this set, and F^r some subset of regular functions of X^θ. Given a consistent family of finite-dimensional distribution functions,

construct with respect to them a probability measure μ on $\mathcal{C}(X, \Theta)$ (as in the proof of Kolmogorov's theorem; see p. 49). If $F^r \in \mathcal{C}(X, \Theta)$, then one may speak of the probability that a random function is regular. This is $\mu(F^r)$. However, for the interesting regularity properties (continuity, at most jump discontinuities, boundedness and so on), the corresponding set F^r does not belong to $\mathcal{C}(X, \Theta)$. This is because $\mathcal{C}(X, \Theta)$ contains those sets for which the values of the functions on a countable subset of Θ determine if they belong to $\mathcal{C}(X, \Theta)$ (continuity of a function, for example, cannot be determined from its values on a countable set). What is of interest is this: When does there exist a modification $x^*(\theta, \omega)$ of a random function such that $x^*(\theta, \omega) \in F^r$ for all ω? Such a modification will be called an F^r-modification. The answer to the question must only utilize the measure μ and the set F^r.

Let μ^* be the outer measure formed from μ: For all $F \subset X^\theta$,

$$\mu^*(F) = \inf_{\bigcup C_k \supset F} \sum \mu(C_k) \, ,$$

where the C_k's are cylinder sets in $\mathcal{C}(X, \Theta)$.

Theorem 4.1.1. *There exists an F^r-modification with finite-dimensional distributions generating μ if and only if $\mu^*(F^r) = 1$.*

Proof. Assume that such a modification $x^*(\theta, \omega)$ exists. Then

$$\mathbf{P}(x^*(\theta, \omega) \in C) = 1$$

for every $C \in \mathcal{C}(X, \Theta)$ such that $C \supset F^r$. Thus $\mu^*(F^r) = 1$. Now assume that this condition holds. Let C^r be the σ-algebra of subsets of F^r of the form $F = C \cap F^r$ with $C \in \mathcal{C}(X, \Theta)$. If $C_1 \cap F^r = C_2 \cap F^r$, then $[(C_1 \setminus C_2) \cup (C_2 \setminus C_1)] \cap F^r = \emptyset$ and therefore $\mu[(C_1 \setminus C_2) \cup (C_2 \setminus C_1)] = 0$ (the condition $\mu^*(F^r) = 1$ entails that $\mu(C) = 0$ for all C such that $C \cap F^r = \emptyset$.) Therefore $\mu(C_1) = \mu(C_2)$. Consequently, a measure can be introduced on $\mathcal{C}^r$: $\bar{\mu}(F) = \mu(C)$ if $F = C \cap F^r$. If we view $(F^r, \mathcal{C}^r, \bar{\mu})$ as the probability space and we put $x^*(\theta, \omega) = \bar{\omega}(\theta)$ with $\bar{\omega}(\theta) \in F^r$, then we obtain an F^r-modification on this specific probability space. □

4.1.1 Separable Random Functions

Doob proposed a procedure for selecting regular modifications. Let Θ be a separable topological space, Λ a denumerable everywhere dense subset of Θ and X a topological space. Consider any open $U \subset \Theta$ and closed $F \subset X$. A function $x(\theta)$ from Θ to X is said to be *Λ-separable* if $x(\theta) \in F$ for $\theta \in U$ when $x(\theta) \in F$ for $\theta \in U \cap \Lambda$.

A random function $x(\theta, \omega)$ with phase space X defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is said to be Λ-separable if there exists an $S \in F$ with $\mathbf{P}(S) = 0$ such that

$$\bigcap_{\theta\in U\cap A}\{\omega : x(\theta,\omega)\in F\}\setminus\bigcap_{\theta\in U}\{\omega : x(\theta,\omega)\in F\}\subset S \tag{4.1.1}$$

for all open $U\subset\Theta$ and closed $F\subset X$. In that case, if $\omega\in\Omega\setminus S$, the function $x(\cdot,\omega)$ will be Λ-separable.

If X is a complete separable metric space with metric $r(\cdot,\cdot)$ and $x(\theta)$ is a Λ-separable function, then

1. $x(\theta)$ is uniformly continuous on a closed subset Θ_1 of Θ if and only if $x(\theta)$ is uniformly continuous on $\Theta_1\cap\Lambda$;
2. $x(\theta)$ is bounded if and only if it is bounded on Λ;
3. if Θ is an interval of the line, then $x(\theta)$ has at most jump discontinuities.

If $x(\theta)$ has finitely many ε-oscillations on Λ for each $\varepsilon>0$, this means that to any $\varepsilon>0$ there exists a k such that for $n>k$,

$$\inf[r(x(\theta_i),x(\theta_{i+1})), i=0,1,\ldots,n-1]<\varepsilon$$

for any choice of $\theta_0<\theta_1<\ldots<\theta_n$ in Λ.

The following theorem was proved by Doob.

Theorem 4.1.2. *If Θ is a separable metric space and X is a compact space, then to every random function there exists a separable modification. That is, there is a denumerable dense subset Λ of Θ for which a Λ-separable modification exists.*

The proof relies on the following considerations. It is sufficient that relation (4.1.1) hold for a denumerable family of open sets $U_k\subset\Theta$ (forming an open base in Θ) and a denumerable family of closed sets $F_k\subset X$ (whose complements form an open base in X).

If to each pair k,m, we can construct a set $\Lambda_{k,m}$ such that

$$\mathbf{P}\left(\bigcap_{\theta\in U_k\cap\Lambda_{k,m}}\{\omega : x^*(\theta,\omega)\in F_m\}\setminus\bigcap_{\theta\in U_k}\{\omega : x^*(\theta,\omega)\in F_m\right)=0\ , \tag{4.1.2}$$

where x^* is some modification of x, and if we take $\Lambda=\bigcup_{k,m}\Lambda_{k,m}$ and

$$S=\bigcup_{k,m}\left(\bigcap_{\theta\in U_k\cap\Lambda_{k,m}}\{\omega : x^*(\theta,\omega)\in F_m\}\setminus\bigcap_{\theta\in U_k}\{\omega : x^*(\theta,\omega)\in F_m\}\right)\ , \tag{4.1.3}$$

then we can show that $\mathbf{P}(S)=0$.

The problem now reduces to constructing sets $\Lambda_{k,m}$ so that (4.1.2) holds and finding the modification x^* itself.

Lemma. *To any U_k and F_m, it is possible to form sets $\Lambda_{k,m}$ so that for all $\theta\in U_k$,*

$$\mathbf{P}\{x(\theta,\omega)\in F_m,\ \theta\in\Lambda_{k,m},\ x(\theta)\notin F_m\}=0\ .$$

Proof. Assuming that $\theta_1, \theta_2, \ldots, \theta_n$ have already been selected, choose $\theta_{n+1} \in U_k$ so that

$$\begin{aligned} &\mathbf{P}\{x(\theta_i, \omega) \in F_m, i = 1, 2, \ldots, n, x(\theta_{n+1}, \omega) \notin F_m\} \\ &\geq \frac{1}{2} \sup_{\theta \in U_k} \mathbf{P}\{x(\theta_i, \omega) \in F_m, i = \overline{1, n}, x(\theta, \omega) \notin F_m\} . \end{aligned}$$

The sum of the probabilities on the left converges since it involves the probabilities of mutually exclusive events. Therefore its general term approaches zero. On putting $\Lambda_{k,m} = \{\theta_1, \theta_2, \ldots\}$, we can complete the proof of the lemma.

We now form the required modification. Let $\Lambda_k = \bigcup_m \Lambda_{k,m}$ and $D_k(\omega)$ be the closure of $\{x(\theta, \omega), \theta \in \Lambda_k\}$ (it depends on ω). Put $D^\theta(\omega) = \bigcap_{U_k \ni \theta} D_k(\omega)$. This set is nonempty since X is a compact space. Let $x^*(\theta, \omega) = x(\theta, \omega)$ if $x(\theta, \omega) \in D^\theta(\omega)$ and let $x^*(\theta, \omega) \in D^\theta(\omega)$ be chosen arbitrarily if $x(\theta, \omega) \notin D^\theta$. This then is the required modification. □

4.1.2 Continuous Stochastic Processes

Now consider the case of a stochastic process $x(t, \omega)$ defined on a subset T of R with values in a complete separable metric space X. If it has a continuous modification, then we shall simply say that it is continuous. To prove the continuity of a process on T, it suffices to show that it is uniformly continuous on a denumerable dense subset T_0 of T, that is,

$$\mathbf{P}\left\{\lim_{\delta \to 0} \sup_{t_1, t_2 \in T_0,\ |t_1 - t_2| \leq \delta} r(x(t_1, \omega), x(t_2, \omega)) = 0\right\} = 1 . \tag{4.1.4}$$

In that case,

$$\lim_{t' \in T_0,\ t' \to t} x(t', \omega) = x^*(t, \omega)$$

exists with probability 1 for all $t \in T$ and $x^*(t, \omega)$ is a continuous modification.

Theorem 4.1.3 (Kolmogorov). *Let $T = [0, 1]$ and suppose that there exist positive α, β and k such that*

$$\mathbf{E}(r(x(t_1, \omega), x(t_2, \omega))^\beta \leq k|t_2 \cdot t_1|^{1+\alpha} \tag{4.1.5}$$

Then $x(t, \omega)$ is a continuous process.

Proof. Let T_0 be the set of binary rational numbers. Let us show that $x(t, \omega)$ is uniformly continuous on T_0. Put

$$\eta_n = \sup_{1 \leq i \leq 2^n} r\left(x\left(\frac{i}{2^n}, \omega\right), x\left(\frac{i-1}{2^n}, \omega\right)\right) .$$

Then for $0 < a < 1$,

$$\begin{aligned}\mathbf{P}\{\eta_n > a^n\} &\le \sum_{i=1}^{2^n} \mathbf{P}\left\{r\left(x\left(\frac{i}{2^n},\omega\right), x\left(\frac{i-1}{2^n},\omega\right)\right) > a^n\right\} \\ &\le \sum_{i=1}^{2^n} \frac{1}{a^{n\beta}} \mathbf{E} r^\beta \left(x\left(\frac{i}{2^n},\omega\right), x\left(\frac{i-1}{2^n},\omega\right)\right) \\ &\le 2^n \cdot \frac{1}{a^{n\beta}} 2^{-n(1+\alpha)} = \left(\frac{1}{a^\beta 2^\alpha}\right)^n .\end{aligned}$$

Now choose $a > 2^{-\alpha/\beta}$. Then $\sum \mathbf{P}\{\eta_n > a^n\} < \infty$ and so $\eta_n \le a^n$ for n sufficiently large by the Borel-Cantelli lemma. It remains to observe that if $t_1 = i/2^m$, $|t_1 - t_2| < 1/2^m$, $t_2 = k/2^n$, $n > m$, then $r(x(t_1,\omega), x(t_2,\omega)) \le \eta_{m+1} + \ldots + \eta_n$. Therefore $r(x(t_1,\omega), x(t_2,\omega)) \le 2\sum_{n=m+1}^{\infty} \eta_k$ for any t_1 and t_2 in T_0 where m is such that $1/2^m > |t_2 - t_1| \ge 1/2^{m+1}$. The theorem follows from this last statement. □

4.1.3 Processes With at Most Jump Discontinuities

Consider again a stochastic process on $[0,1]$ with phase space X which is a complete separable metric space. A stochastic process $x(t,\omega)$ has at most jump discontinuities if

$$\lim_{t' \uparrow t,\ t' \in T_0} x(t',\omega), \qquad \lim_{t' \downarrow t,\ t' \in T_0} x(t',\omega) ,$$

exist on a dense set T_0 for almost all ω. These limits will be denoted hereafter by $x(t-,\omega)$ and $x(t+,\omega)$. In that case, if we put $x^*(t,\omega) = x(t-,\omega)$ when the process is stochastically left-continuous at t, $x^*(t,\omega) = x(t+,\omega)$ when stochastically right-continuous at t (but not from the left) and $x^*(t,\omega) = x(t,\omega)$ if there is no stochastic continuity from the left or right, we arrive at a modification of $x(t,\omega)$ with at most jump discontinuities. It exists if and only if $x(t,\omega)$ has finitely many ε-oscillations on T_0 with probability 1.

Theorem 4.1.4. *The process $x(t,\omega)$ has at most jump discontinuities with probability 1 if*

$$\mathbf{E}[r(x(t_1,\omega), x(t_2,\omega)) r(x(t_2,\omega), x(t_3,\omega))]^\beta \le k|t_3 - t_1|^{1+\alpha}$$

for some $\alpha > 0, \beta > 0, k > 0$ and $t_1 < t_2 < t_3$.

Proof. We have

$$\mathbf{P}\left\{\max_t\left[r\left(x\left(\frac{i-1}{2^n},\omega\right),x\left(\frac{i}{2^n},\omega\right)\right)\wedge r\left(x\left(\frac{i}{2^n},\omega\right),x\left(\frac{i+1}{2^n},\omega\right)\right)\right]>a^n\right\}$$

$$\leq\sum_i\mathbf{P}\left\{r\left(x\left(\frac{i-1}{2^n},\omega\right),x\left(\frac{i}{2^n},\omega\right)\right)\wedge r\left(x\left(\frac{i}{2^n},\omega\right),x\left(\frac{i+1}{2^n},\omega\right)\right)>a^n\right\}$$

$$\leq\sum_i\mathbf{P}\left\{r\left(x\left(\frac{i-1}{2^n},\omega\right),x\left(\frac{i}{2^n},\omega\right)\right)r\left(x\left(\frac{i}{2^n},\omega\right),x\left(\frac{i+1}{2^n},\omega\right)\right)\geq a^{2n}\right\}$$

$$\leq ka^{-2n\beta}\left(\frac{1}{2^{n-1}}\right)^{1+\alpha}2^n\leq k_1(2^\alpha a^{2\beta})^{-n},\quad k_1=k\cdot 2^{1+\alpha}\,.$$

If $2^\alpha a^{2\beta}>1$ with $a<1$, then

$$r\left(x\left(\frac{i-1}{2^n},\omega\right),x\left(\frac{i}{2^n},\omega\right)\right)\wedge r\left(x\left(\frac{i}{2^n},\omega\right),x\left(\frac{i+1}{2^n},\omega\right)\right)\leq a^n$$

for all n sufficiently large by virtue of the Borel-Cantelli lemma. It can be shown that if there is a function $x(t)$ satisfying

$$r\left(x\left(\frac{i-1}{2^n}\right),x\left(\frac{i}{2^n}\right)\right)\wedge r\left(x\left(\frac{i}{2^n}\right),x\left(\frac{i+1}{2^n}\right)\right)\leq a^n$$

for $n\geq n_0$ and $a<1$, then it has finitely many ε-oscillations for all positive ε. The theorem follows from this. □

A second theorem makes use of conditional distributions.

Let $x(t,\omega)$ be a process on $[0,1]$ with values in X. Let $\mathcal{F}_t$ be the σ-algebra generated by $x(s,\omega)$, $s\leq t$.

Theorem 4.1.5. *Suppose that there exists a function $\varphi_\varepsilon(h)\downarrow 0$ as $h\downarrow 0$ such that*

$$\mathbf{P}\{r(x(t+h,\omega),x(t,\omega))>\varepsilon|\mathcal{F}_t\}\leq\varphi_\varepsilon(h)$$

with probability 1 for any $\varepsilon>0$ and all $t\in[0,1]$. Then $x(t,\omega)$ has at most jump discontinuities.

Proof. Similarly to the proof of Lemma 3.4.1 on p. 80, an upper bound for the expected number of ε-oscillations ν_ε of the sequence $x(t_1,\omega),\ldots,x(t_n,\omega)$ with $t_n-t_1\leq h$ and $\varphi_{\varepsilon/4}(h)<1/2$ is $\mathbf{E}\nu_\varepsilon\leq\varphi_{\varepsilon/4}(h)/(1-2\varphi_{\varepsilon/4}(h))$. This implies that the number of ε-oscillations of $x(t,\omega)$ is finite on any countable subset T_0 of $[0,1]$.

4.1.4 Markov Processes

Let $(X,\mathcal{B})$ be a measurable space and $T\subset R$. A stochastic process $x(t,\omega)$ with phase space $(X,\mathcal{B})$ defined on T is said to be a *Markov process* if and there exists a function $P(s,x,t,E)$ defined for $x\in X$, $E\in\mathcal{B}$, and $s<t$ with s and $t\in T$ such that

1. it is a probability measure with respect to E and $\mathcal{B}$-measurable in x,
2. the *Chapman-Kolmogorov equation* holds: for $s < t < u$, with s, t, and $u \in T$,

$$P(s,x,u,E) = \int P(t,y,u,E)P(s,x,t,dy)\ , \tag{4.1.6}$$

3. if $\mathcal{F}_t^*$ is the σ-algebra generated by $x(s,\omega)$, $s \in T, s \le t$, then

$$\mathbf{P}(x(u,\omega) \in E|\mathcal{F}_t^*) = P(t,x(t,\omega),u,E) \tag{4.1.7}$$

for t and $u \in T$, $t < u$, with probability 1.

$P(s,x,t,E)$ is called the *transition probability* of the process. It determines the conditional finite-dimensional distributions of the process. If $t_0 < t_1 < \ldots < t_n$, with $t_i \in T$, and $E_1, \ldots, E_n \in \mathcal{B}$, then

$$\begin{aligned}&\mathbf{P}\Big\{x(t_1) \in E_1, \ldots, x(t_n) \in E_n | x(t_0)\Big\}\\ &= \int_{E_1} \cdots \int_{E_n} P(t_0, x(t_0), t_1, dx_1) \ldots P(t_{n-1}, x_{n-1}, t_n, dx_n)\ .\end{aligned} \tag{4.1.8}$$

This formula results by applying (4.1.7) repeatedly. If T has a smallest (initial) point t_0, then to specify the finite-dimensional distributions of a Markov process, it suffices to give the distribution of $x(t_0)$ (the initial distribution) and the transition probability. Markov processes describe the evolution of dynamic systems undergoing independent random perturbations at different moments of time.

Theorem 4.1.5 leads to a condition for a Markov process to have at most jump discontinuities.

Theorem 4.1.6. *Suppose that X is a complete separable metric space and $T = [0,1]$. Let $S_\rho(x)$ be the ball of radius ρ centered at x. Put*

$$\varphi_\rho(h) = \sup_{x \in X} \sup_{\substack{0 \le t-s \le h \\ t,s \in [0,1]}} P(s,x,t,X \setminus S_\rho(x))\ .$$

If $\varphi_\rho(0+) = 0$ for all $\rho > 0$, then the Markov process $x(t,\omega)$ has at most jump discontinuities.

This is a consequence of Theorem 4.1.5 and formula (4.1.7).

Corollary. *A stochastically continuous process with independent increments in a separable Banach space X has at most jump discontinuities.*

For, if $x(t,\omega)$ is such a process, then it is a Markov process with transition probability $P(s,x,t,E) = \mathbf{P}\{x(t,\omega) - x(s,\omega) + x \in E\}$. Hence, $P(s,x,t,X \setminus S_\rho(x)) = \mathbf{P}\{|x(t,\omega) - x(s,\omega)| > \rho\}$ ($|\cdot|$ is the norm in X) and so

$$\varphi_\rho(h) = \sup_{\substack{0 \le t-s \le h \\ t,s \in [0,1]}} \mathbf{P}\Big\{|x(t,\omega) - x(s,\omega)| > \rho\Big\}\ .$$

It remains to observe that stochastic continuity implies uniform stochastic continuity.

4.2 Measurability

Let $(\Theta, \mathcal{C})$ and $(X, \mathcal{B})$ be two measurable spaces and let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A random function $x(\theta, \omega)$ defined on Θ with phase space X is said to be *measurable* if the mapping $x(\theta, \omega) : \Theta \times \Omega \to X$ is measurable with respect to $\mathcal{C} \otimes \mathcal{F}$. In other words, $\{(\theta, \omega) \in \Theta \times \Omega : x(\theta, \omega) \in B\} \in \mathcal{C} \otimes \mathcal{F}$ for all $B \in \mathcal{B}$. If random function $x(\theta, \omega)$ is measurable, then $x(\theta, \omega)$ is measurable with respect to the σ-algebra $\mathcal{C}$ for all $\omega \in \Omega$. If $g(x)$ is a bounded measurable scalar function and ν is a measure on $\mathcal{C}$, then $\int g(x(\theta, \omega))\nu(d\theta)$ is defined for all ω. It is an $\mathcal{F}$-measurable (random) variable and

$$\mathbf{E} \int g(x(\theta, \omega))\nu(d\theta) = \int \mathbf{E} g(x(\theta, \omega))\nu(d\theta) \, . \tag{4.2.1}$$

All of these statements are consequences of Fubini's theorem.

4.2.1 Existence of a Measurable Modification

We shall assume that the σ-algebras $\mathcal{B}$ in X and $\mathcal{C}$ in Θ are countably generated. Let $x(\theta, \omega)$ be a random function and let $g(x)$ be a $\mathcal{B}$-measurable function from X to $[0, 1]$. We shall say that $x(\theta, \omega)$ is countably generated if the space of random variables of the form $\eta = g(x(\theta, \omega))$, $\theta \in \Theta$, with the metric

$$r(\eta_1, \eta_2) = \mathbf{E}|\eta_1 - \eta_2| \, , \tag{4.2.2}$$

is separable.

A measurable space $(X, \mathcal{B})$ is called a Borel space if it is separable and metric, it is a Borel subset of its completion and $\mathcal{B}$ is its Borel σ-algebra.

Theorem 4.2.1. *In order for a random function $x(\theta, \omega)$ to have a measurable modification, it is necessary and, if $(X, \mathcal{B})$ is a Borel space, also sufficient that the following conditions hold:* (a) *$x(\theta, \omega)$ is countably generated and* (b) *for any choice of B_1 and $B_2 \in \mathcal{B}$ and $\bar{\theta}$ and $\theta \in \Theta$, the numerical function*

$$\mathbf{P}\{x(\bar{\theta}, \omega) \in B_1, x(\theta, \omega) \in B_2\} = \mathbf{E} I_{B_1}(x(\bar{\theta}, \omega)) I_{B_2}(x(\theta, \omega))$$

is $\mathcal{C}$-measurable with respect to θ.

Necessity. That (b) is necessary follows from Fubini's theorem and the measurability of the function $I_B(x(\theta, \omega))$. To prove the necessity of (a), it suffices to show that if $x(\theta, \omega)$ is measurable, then $I_B(x(\theta, \omega))$ is countably generated for any choice of $B \in \mathcal{B}$. Since $\{(\theta, \omega) : x(\theta, \omega) \in B\} \in \mathcal{C} \otimes \mathcal{B}$, there are sequences of sets $C_k \in \mathcal{C}$ and $A_k \in \mathcal{F}$ for which that set belongs to the σ-algebra generated by the rectangles $C_k \times A_k$, $k \geq 1$. Then the closure of the set of random variables $\{I_{A_k}(\omega)\}$ in the metric r contains all of the variables $\{I_B(x(\theta, \omega)), \theta \in \Theta\}$, which means $I_B(x(\theta, \omega))$ is countably generated.

Sufficiency. A Borel space may be mapped into $[0,1]$ in one-one fashion by a Borel function. Therefore $x(\theta,\omega)$ may be assumed to take values in $[0,1]$ (with $\mathcal{B}_{[0,1]}$ the σ-algebra of Borel sets). Let H be the closure in $L_2(\Omega,\mathbf{P})$ of the linear span of the variables $I_B(x(\theta,\omega))$, with $B\in\mathcal{B}_{[0,1]}$. H is a separable Hilbert space and $x(\theta,\omega)$ is a measurable function with values in H. To such a function, there exists a sequence of simple H-valued functions

$$x_n(\theta,\omega)=\sum_k \xi_{nk}(\omega)I_{C_{nk}}(\theta),\quad C_{nk}\in\mathcal{C}$$

such that $x_n(\theta,\omega)$ converges to $x(\theta,\omega)$ in H for all θ, that is,

$$\mathbf{E}|x_n(\theta,\omega)-x(\theta,\omega)|^2\to 0\ .$$

The function $\lambda_n(\theta)=\mathbf{E}|x_n(\theta,\omega)-x(\theta,\omega)|^2$ is measurable.

Define the measurable functions

$$n_m(\theta)=\inf\{k:\lambda_k(\theta)\le 2^{-m}\},$$
$$\tilde{x}_m(\theta,\omega)=x_{n_m(\theta)}(\theta,\omega)=\textstyle\sum_n\sum_k \xi_{nk}(\omega)I_{C_{nk}}(\theta)I_{\{n_m(\theta)=n\}}\ .$$

Since $\mathbf{E}|\tilde{x}_m(\theta,\omega)-x(\theta,\omega)|^2\le 2^{-m}$, it follows that

$$\sum \mathbf{P}\{|\tilde{x}_m(\theta,\omega)-x(\theta,\omega)|>\varepsilon\}\le\sum\varepsilon^2 2^{-m}<\infty \tag{4.2.3}$$

for all θ and positive ε and so $\tilde{x}_m(\theta,\omega)-x(\theta,\omega)\to 0$ with probability 1. Putting $x^*(\theta,\omega)=\lim\tilde{x}_m(\theta,\omega)$ if the limit exists, and $x^*(\theta,\omega)=0$ otherwise, we obtain our required modification. □

4.2.2 Mean-Square Integration

Let Θ be a compact metric space and m a finite measure on $\mathcal{C}$. Assume further that $x(\theta,\omega)$ is a random function defined on Θ taking values in R and that $\mathbf{E}|x(\theta,\omega)|^2<\infty$ for all $\theta\in\Theta$. Write $a(\theta)=\mathbf{E}x(\theta,\omega)$ and $R(\theta_1,\theta_2)=\mathbf{E}x(\theta_1,\omega)x(\theta_2,\omega)-a(\theta_1)a(\theta_2)$. They are the mean and covariance function of $x(\theta,\omega)$. We are interested in conditions under which the Riemann sums

$$S_n=\sum_{k=1}^n x(\theta_k,\omega)m(C_k) \tag{4.2.4}$$

have a limit in the mean square as $\max_k \operatorname{diam} C_k\to 0$, where $C_i\cap C_j=\emptyset$ for $i\ne j$, $\bigcup_{i=1}^n C_i=\Theta$, $\theta_k\in C_k$ and $C_i\in\mathcal{C}$, and it must be independent of the choice of θ_k and C_k. If this limit exists, then it is natural to call it the Riemann mean-square integral of $x(\theta,\omega)$ with respect to the measure $m(d\theta)$. The integral will also be denoted by $\int_\Theta x(\theta,\omega)m(d\theta)$ (like an ordinary integral). Observe that the integral sums (4.2.4) coincide with probability 1 for different modifications. Therefore the values of the integrals of different modifications are the same with probability 1. It is shown below that the Riemann mean-square integral of a measurable modification coincides with its ordinary one.

Theorem 4.2.2. *A random function is mean-square Riemann integrable if and only if $a(\theta)$ is Riemann integrable with respect to the measure $m(d\theta)$ and $R(\theta_1, \theta_2)$ is Riemann integrable with respect to the measure $m(d\theta_1)m(d\theta_2)$.*

Proof. This theorem is a consequence of the fact that the partial sums S_n converge in the mean-square if and only if the following limits exist:

$$\lim_{n\to\infty} \mathbf{E}S_n, \lim_{\substack{n\to\infty \\ m\to\infty}} \mathbf{E}S_nS_m \ ,$$

$\mathbf{E}S_n$ is the Riemann sum for $a(\theta)$ with respect to $m(d\theta)$ and $\mathbf{E}S_nS_m$ is the Riemann sum for $R(\theta_1, \theta_2) + a(\theta_1)a(\theta_2)$ with respect to $m(d\theta_1)m(d\theta_2)$. □

Corollary. *If $x(\theta, \omega)$ is Riemann mean-square integrable, then the functions $a(\theta)$ and $R(\theta_1, \theta_2)$ are continuous almost everywhere with respect to the respective measures $m(d\theta)$ and $m(d\theta_1)m(d\theta_2)$.*

Theorem 4.2.3. *Suppose that $x(\theta, \omega)$ is a measurable function that is Riemann mean-square integrable. Then the Riemann integral $\int x(\theta, \omega)m(d\theta)$ coincides with the ordinary integral.*

Proof. The Riemann integrability of $a(\theta)$ and $R(\theta_1, \theta_2)$ implies that they are bounded and thus

$$\mathbf{E}\int |x(\theta,\omega)|m(d\theta) = \int \mathbf{E}|x(\theta,\omega)|m(d\theta)$$

(the integral here is Lebesgue and we have made use of Fubini's theorem). Therefore $\int x(\theta, \omega)m(d\theta)$ exists with probability 1. Using the Riemann integrability and that Riemann integrable (nonrandom) functions have coinciding Riemann and Lebesgue integrals, we can show that

$$\lim_{n\to\infty} \mathbf{E}S_n^2 = \left(\int a(\theta)m(d\theta)\right)^2 + \iint R(\theta_1,\theta_2)m(d\theta_1)m(d\theta_2),$$

$$\begin{aligned}
&\lim_{n\to\infty} \mathbf{E}S_n \int x(\theta,\omega)m(d\theta) \\
&\quad = \lim_{n\to\infty} \left[\int a(\theta)m(d\theta)\sum_{k=1}^{n} a(\theta_k)m(C_k) + \int \sum_{k=1}^{n} R(\theta,\theta_k)m(C_k)m(d\theta)\right] \\
&\quad = \left(\int a(\theta)m(d\theta)\right)^2 + \iint R(\theta',\theta'')m(d\theta')m(d\theta'') \ .
\end{aligned}$$

Therefore

$$\begin{aligned}
&\lim_{n\to\infty} \mathbf{E}\left(S_n - \int x(\theta,\omega)m(d\theta)\right)^2 \\
&\quad = \lim_{n\to\infty} \mathbf{E}S_n^2 - 2\lim_{n\to\infty} \mathbf{E}S_n \int x(\theta,\omega)m(d\theta) + \mathbf{E}\left(\int x(\theta,\omega)m(d\theta)\right)^2 = 0 \ .
\end{aligned}$$

□

4.2.3 Expansion of a Random Function in an Orthogonal Series

Let $(\Theta, \mathcal{C})$ be the same as in the preceding section and let the random function $x(\theta, \omega)$ be continuous in the mean-square:

$$\lim_{\theta \to \theta_0} \mathbf{E}|x(\theta, \omega) - x(\theta_0, \omega)|^2 = 0, \quad \forall \theta_0 \in \Theta \, .$$

Then the mean $a(\theta)$ and covariance function $R(\theta_1, \theta_2)$ are continuous. Therefore $x(\theta, \omega)$ is Riemann mean-square integrable. If $f_1(x)$ and $f_2(x)$ are any bounded continuous functions, then $\mathbf{E}(f_1(x(\theta_1, \omega) \times f_2(x(\theta, \omega))$ is continuous in θ_1 and θ_2 and so it is measurable in θ_2 for all θ_1. The set F of functions for which such measurability holds is closed under bounded convergence (a sequence g_n converges boundedly to g if $\sup_{x,n} |g_n(x)| < \infty$ and $g_n(x) \to g(x)$ for all x). Thus F contains all bounded Borel functions and particularly the indicator functions of Borel sets. Thus, by virtue of Theorem 4.2.1, $x(\theta, \omega)$ has a measurable modification. We shall view it as $x(\theta, \omega)$ itself .

Let $L_2(\Theta, m)$ be the space of numerical $\mathcal{C}$-measurable functions $g(x)$ such that $\int g^2(\theta) m(d\theta) < \infty$. It is a Hilbert space with the inner product $(g_1, g_2) = \int g_1(\theta) g_2(\theta) m(d\theta)$. Since

$$\int \mathbf{E} x^2(\theta, \omega) m(d\theta) = \mathbf{E} \int x^2(\theta, \omega) m(d\theta) < \infty \, ,$$

$x(\cdot, \omega) \in L_2(\Theta, m)$ for almost all ω. If $\{g_k(\theta)\}$ is an orthonormal basis in $L_2(\Theta, m)$, then every $h(\theta) \in L_2(\Theta, m)$ can be represented as

$$h(\theta) = \sum \alpha_k g_k(\theta), \quad \alpha_k = \int h(\theta) g_k(\theta) m(d\theta) \, . \tag{4.2.5}$$

Therefore

$$x(\theta, \omega) = \sum \xi_k(\omega) g_k(\theta), \quad \xi_k(\omega) = \int x(\theta, \omega) g_k(\theta) m(d\theta) \, . \tag{4.2.6}$$

The series (4.2.5) and (4.2.6) converge in $L_2(\Theta, m)$ (the second one for almost all ω). Since by Parseval's equality,

$$\sum_{k=1}^{\infty} \xi_k^2(\omega) = \int x^2(\theta, \omega) m(d\theta) \quad \text{and} \quad \mathbf{E} \int x^2(\theta, \omega) m(d\theta) < \infty \, ,$$

it follows that

$$\sum_{k=1}^{\infty} \mathbf{E} \xi_k^2(\omega) < \infty \, .$$

Consequently

$$\lim_{n\to\infty}\int \mathbf{E}\left|x(\theta,\omega)-\sum_{k=1}^{n}\xi_k(\omega)g_k(\theta)\right|^2 m(d\theta)=\lim_{n\to\infty}\sum_{k=n+1}^{\infty}\mathbf{E}\xi_n^2(\omega)=0\ .$$

Thus the series (4.2.6) converges in $L_2(\Theta,m)$ in the mean-square.

Let $a(\theta)=0$. The integral operator $Rg(\theta)=\int R(\theta,\theta')g(\theta')m(d\theta')$ is completely continuous, symmetric and nonnegative in $L_2(\Theta,m)$. Let $\{\varphi_k\}$ be a complete orthogonal system of its eigenfunctions. Then

$$x(\theta,\omega)=\sum\eta_k\varphi_k(\theta),\quad \eta_k=\int x(\theta,\omega)\varphi_k(\theta)m(d\theta)$$

and the η_k's are uncorrelated. Namely, for $k\neq l$,

$$\begin{aligned}\mathbf{E}\eta_k\eta_l&=\mathbf{E}\iint x(\theta,\omega)x(\theta',\omega)\varphi_k(\theta)\varphi_l(\theta')m(d\theta)m(d\theta')\\&=\iint R(\theta,\theta')\varphi_k(\theta)\varphi_l(\theta')m(d\theta)m(d\theta')\\&=\lambda_k\int\varphi_k(\theta')\varphi_l(\theta)m(d\theta')=0\ ;\end{aligned}$$

λ_k in this is the eigenvalue of R corresponding to to the eigenfunction φ_k. In addition, $\lambda_k\geq 0$ and $\sum\lambda_k=\int R(\theta,\theta)m(d\theta)$. In particular, if $x(\theta,\omega)$ is a Gaussian random function, the $\{\eta_k\}$ also have a joint Gaussian distribution. Therefore $\eta_k=\sqrt{\lambda_k}\xi_k$, where $\{\xi_k\}$ is a sequence of independent Gaussian random variables with $\mathbf{E}\xi_k=0$ and $\mathbf{V}\xi_k=1$.

Expansion of a Wiener process on $[0,1]$. Let $w(t)$, $t\in[0,1]$, be a Wiener process, that is, a Gaussian process with independent increments such that $\mathbf{E}w(t)=0$ and $\mathbf{V}w(t)=t$. Then $\mathbf{E}w(t)w(s)=t\wedge s$.

The eigenfunctions of the operator R are determined from the equation

$$\lambda\varphi(t)=\int_0^1(t\wedge s)\varphi(s)ds=\int_0^t s\varphi(s)ds+t\int_t^1\varphi(s)ds\ .$$

This gives $\varphi(0)=0$, $\varphi'(1)=0$ and $\lambda\varphi''(t)+\varphi(t)=0$. Therefore

$$w(t)=\sum_{k=0}^{\infty}\frac{2\sqrt{2}}{\pi(2k+1)}\xi_k\sin\frac{\pi}{2}(2k+1)t\ ,\tag{4.2.7}$$

where the ξ_k's are independent Gaussian variables with $\mathbf{E}\xi_k=0$ and $\mathbf{V}\xi_k=1$.

4.3 Adapted Processes

The number of theorems on measurability of stochastic processes can be expanded considerably if use is made of the ordering in the parameter space.

We shall consider stochastic processes defined on R_+. With each R_+, we associate a σ-algebra of events $\mathcal{F}_t$ which is a subalgebra of the algebra $\mathcal{F}$ of the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. This $\mathcal{F}_t$ will be interpreted as the σ-algebra of events observed up to time t inclusively. Thus $\mathcal{F}_{t_1} \subset \mathcal{F}_{t_2}$ when $t_1 < t_2$. We now impose some additional conditions on the collection $\{\mathcal{F}_t\}$. First, $\mathcal{F}$ is complete with respect to the measure $\mathbf{P}$ (that is, if $A \in \mathcal{F}$ and $\mathbf{P}(A) = 0$, then $B \in \mathcal{F}$ for all $B \subset A$) and the σ-algebra $\mathcal{F}_0$ contains all sets in $\mathcal{F}$ of $\mathbf{P}$-measure zero. This is the completeness condition. Second is the following condition of right-continuity: For all $t \in R_+$

$$\mathcal{F}_{t+} = \bigcap_{s>t} \mathcal{F}_s = \mathcal{F}_t \ .$$

This collection of σ-algebras is called a *flow*. In what follows, the flow of σ-algebras is viewed as fixed. Processes $x(t, \omega)$ will be studied for which for each fixed t the random variable $x(t, \omega)$ is $\mathcal{F}_t$-measurable. Such processes are said to be *adapted* to the flow $\{\mathcal{F}_t\}$.

4.3.1 Stopping Times

A random variable τ taking nonnegative values including ∞ is called a *stopping time* if $\{\tau \le t\} \in \mathcal{F}_t$ for all $t \in R_+$ where $\mathcal{F}_\infty = \bigvee_t \mathcal{F}_t$. Associated with τ is the σ-algebra $\mathcal{F}_\tau$ of events $A \in \mathcal{F}_\infty$ such that $A \cap \{\tau \le t\} \in \mathcal{F}_t$ for all $t \in R_+$. The properties of $\mathcal{F}_t$ are preserved in a sense if t is replaced by stopping times. Let us establish some of these properties.

Theorem 4.3.1. 1. *If τ_1 and τ_2 are stopping times, then $\tau_1 \vee \tau_2$ and $\tau_1 \wedge \tau_2$ are also stopping times.*
2. *If $\tau_1 \le \tau_2$ and τ_1 and τ_2 are stopping times, then $\mathcal{F}_{\tau_1} \subset \mathcal{F}_{\tau_2}$.*
3. *If $\tau_n \downarrow \tau$ and the τ_n's are stopping times, then τ is a stopping time and $\mathcal{F}_\tau = \bigcap \mathcal{F}_{\tau_n}$.*
4. *If τ_n is a stopping time and $\tau_n \uparrow \tau$, then τ is a stopping time. If $\tau_n < \tau$, then $\bigvee_n \mathcal{F}_{\tau_n} = \mathcal{F}_{\tau-}$, where $\mathcal{F}_{\tau-}$ is the σ-algebra generated by the events $A_t \cap \{\tau > t\}$, with $t \in R_+$ and $A_t \in \mathcal{F}_t$.*

Proof. Statement 1 is trivial.

2. If $A \in \mathcal{F}_{\tau_1}$, then $A \cap \{\tau_2 \le t\} = A \cap \{\tau_1 \le t\} \cap \{\tau_2 \le t\} \in \mathcal{F}_t$ since $A \cap \{\tau_1 \le t\} \in \mathcal{F}_t$ and τ_2 is a stopping time.
3. $\{\tau < s\} = \bigcup_n \{\tau_n < s\} \in \mathcal{F}_s$. But $\{\tau \le s\} = \bigcap_{k \ge N} \left\{\tau < s + \frac{1}{k}\right\} \in \mathcal{F}_{s+\frac{1}{N}}$. Therefore $\{\tau \le s\} \in \bigcap_n \mathcal{F}_{s+\frac{1}{n}} = \mathcal{F}_{s+} + \mathcal{F}_s$. Since $\tau \le \tau_n$, it follows that $\mathcal{F}_\tau \subset \bigcap_n \mathcal{F}_{\tau_n}$. Let $A \in \bigcap_n \mathcal{F}_{\tau_n}$. Then $A \cap \{\tau_n \le t\} \in \mathcal{F}_t$ and $A \cap \{\tau < t\} = A \cap (\bigcup_n \{\tau_n < t\}) = \bigcup_n A \cap \{\tau_n < t\} \in \mathcal{F}_t$. In other words, $A \cap \{\tau \le t\} \in \mathcal{F}_{t+} = \mathcal{F}_t$ with $A \in \mathcal{F}_\tau$.
4. $\{\tau \le t\} = \bigcap_n \{\tau_n \le t\} \in \mathcal{F}_t$. Let $A \in \mathcal{F}_{\tau_n}$. Then $A = \bigcup_n A \cap \{\tau_n \le k/m\}$ for all m. Hence,

$$A = \bigcup_{m>0} \bigcup_{k>0} \left\{\tau > \frac{k+1}{m}\right\} \cap \left\{\left\{\tau_n \le \frac{k}{m}\right\} \cap A\right\}$$

($\bigcup_m \bigcup_n \left\{\tau > \frac{k+1}{m}\right\} \cap \left\{\tau_n \le \frac{k}{m}\right\} = \Omega$ because $\tau_n < \tau$). Since $\left\{\tau_n \le \frac{k}{m}\right\} \cap A \in \mathcal{F}_{k+m^{-1}}$, we have

$$\left\{\tau > \frac{k+1}{m}\right\} \cap \left(\left\{\tau_n \le \frac{k}{m}\right\} \cap A\right) \in \mathcal{F}_{\tau-} \ .$$

We have shown that $\bigvee_n \mathcal{F}_{\tau_n} \subset \mathcal{F}_{\tau-}$.

Now let $A \in \mathcal{F}_{\tau-}$ and consider an A of the form $A_t \cap \{\tau > t\}$. Then

$$A = \bigcup_k \bigcap_n \left\{\tau_n > t + \frac{1}{k}\right\} \cap A_t \ .$$

Observe that $\left\{\tau_n > t + \frac{1}{k}\right\} \cap A_t \in \mathcal{F}_{\tau_n}$ since $\left\{\tau_n > t + \frac{1}{k}\right\} \cap A_t \cap \{\tau_n \le s\}$ is empty for $s \le t + \frac{1}{k}$ and equals $A_t \cap \left\{t + \frac{1}{k} < \tau_n \le s\right\} \in \mathcal{F}_s$ for $s > t + \frac{1}{k}$. Thus for all k,

$$\bigcup_k \bigcap_n \left\{\tau_n \le t + \frac{1}{k}\right\} \cap A_t \in \bigvee_n \mathcal{F}_{\tau_n}, \quad A \in \bigvee_n \mathcal{F}_{\tau_n} \ .$$

□

A stopping time τ is said to be *predictable* if there exists a sequence of stopping times τ_n such that $\mathbf{P}\{\tau_n < \tau\} = 1$ and $\tau_n \to \tau$. A stopping time ζ is said to be *completely unpredictable* if $\mathbf{P}\{\tau = \zeta\} = 0$ for every predictable stopping time τ.

4.3.2 Progressive Measurability

This concept relates measurability and adaptedness. A process $x(t, \omega)$ with values in a measurable phase space $(X, \mathcal{B})$ is *progressivly measurable* if $x(s, \omega)$ is a $\mathcal{B}_{[0,1]} \otimes \mathcal{F}_t$-measurable function on $[0, t] \times \Omega$ for all $t \in R_+$, where $\mathcal{B}_{[0,1]}$ is the σ-algebra of Borel sets in $[0, t]$. An important consequence of progressive measurability is the following property. For any stopping time τ, if $x(s, \omega)$ is a progressively measurable process, then $x(\tau, \omega)$ is $\mathcal{F}_\tau$-measurable. In fact, $\{x(\tau, \omega) \in B\} \cap \{\tau \le t\}$ belongs to $\mathcal{F}_t$ for all $t \in R_+$ because of the measurability of a composition of measurable functions and the $\mathcal{B}_{[0,1]} \otimes \mathcal{F}_t$-measurability of $x(s, \omega)$ for $s \le t$.

The next result gives sufficient conditions for progressive measurability.

Theorem 4.3.2. *A right- (left-) continuous adapted process in a topological phase space is progressively measurable.*

Proof. Let t be fixed. We have to show that the process $x(s,\omega)$, $s \in [0,t]$, is $\mathcal{B}_{[0,t]} \otimes \mathcal{F}_t$-measurable. A process of the form $x_n(s,\omega) = x_{n,k}(\omega)$ with $\frac{k}{n}t < s < \frac{k+1}{n}t$, which is left- or right-continuous, will be $\mathcal{B}_{[0,t]} \otimes \mathcal{F}_t$-measurable if the $x_{n,k}(\omega)$'s are $\mathcal{F}_t$-measurable. Choosing $x_{n,k}(\omega) = x\left(\frac{k}{n}t, \omega\right)$ if $x(s,\omega)$ is left-continuous and $x_{n,k}(\omega) = x\left(\frac{k+1}{n}t, \omega\right)$ if $x(s,\omega)$ is right-continuous, we arrive at a sequence of $\mathcal{B}_{[0,t]} \otimes \mathcal{F}_t$-measurable processes that converges to $x(s,\omega)$.

4.3.3 Completely Measurable and Predictable σ-Algebras

Let τ_1 and τ_2 be stopping times. The set

$$\{(t,\omega) : \tau_1(\omega) \le t < \tau_2(\omega)\} \in \mathcal{B}_{R_+} \otimes \mathcal{F}$$

is called a stochastic interval and will be denoted by $[\![\tau_1, \tau_2[\![$. The stochastic intervals $]\!]\tau_1, \tau_2]\!]$, $]\!]\tau_1, \tau_2[\![$ and $[\![\tau_1, \tau_2]\!]$ are defined in similar fashion.

Definition. The σ-algebra of completely measurable sets is the σ-algebra $\mathcal{W}$ in $R_+ \times \Omega$ generated by the stochastic intervals $[\![\tau_1, \tau_2[\![$.

Theorem 4.3.3. 1. *$\mathcal{W}$ is the smallest σ-algebra relative to which all adapted right-continuous numerical processes are measurable.*
2. *$\mathcal{W}$ is generated by intervals of the form $[\![t_A, \infty[\![$, where $t \in R_+$, $A \in \mathcal{F}_t$, $t_A = t$, $\omega \in A$, and $t_A = \infty, \omega \notin A$ (it is a stopping time).*

Proof. 1. Let $\mathcal{W}_0$ be the σ-algebra relative to which all adapted right-continuous processes are measurable. $I_{[\![\tau_1,\tau_2[\![}$ is such a process and hence it is $\mathcal{W}_0$-measurable, $[\![\tau_1, \tau_2[\![\in \mathcal{W}_0$ and $\mathcal{W} \subset \mathcal{W}_0$. Let us show that an adapted right-continuous process $x(t,\omega)$ is $\mathcal{W}$-measurable. Choose $\varepsilon > 0$ and form a (generally speaking, transfinite) sequence of stopping times

$$\tau_1^\varepsilon = \inf\{t : |x(0,\omega) - x(t,\omega)| \ge \varepsilon\}$$

(we consider the $\inf \emptyset = \infty$). If $\tau_1^\varepsilon < \infty$, put

$$\tau_2^\varepsilon = \inf\{t > \tau_1^\varepsilon : |x(\tau_1^\varepsilon, \omega) - x(t,\omega)| \ge \varepsilon\},$$

and if $\tau_1^\varepsilon = \infty$, then also $\tau_2^\varepsilon = \infty$ and so on. It is easy to see that τ_k^ε is a stopping time. Put $x^\varepsilon(t) = x(\tau_k^\varepsilon)$ for $t \in [\![\tau_k^\varepsilon, \tau_{k+1}^\varepsilon[\![$ $(\tau_0^\varepsilon = 0)$. This process is $\mathcal{W}$-measurable:

$$\{(t,\omega) : x^\varepsilon(t) < a\} = \bigcup [\![\tau_k^\varepsilon, \tau_{k+1}^\varepsilon[\![\cap \{\omega : x(\tau_k^\varepsilon) < a\} \times R_+ \, .$$

Since $x(t,\omega)$ is right-continuous, it is progressively measurable and so $x(t,\omega)$ is $\mathcal{F}_\tau$-measurable if τ is a stopping time. Put $\tau_k^* = \tau_k^\varepsilon$ for $x(\tau_k^\varepsilon) < a$ and $\tau_k^* = \infty$ otherwise. Then τ_k^* is a stopping time because $\{\omega : x(\tau_k^\varepsilon) < a\} \in \mathcal{F}_{\tau_k^\varepsilon}$ and so $\{(t,\omega) : x^\varepsilon(t) < a\} = \bigcup_k [\![\tau_k^*, \tau_{k+1}^\varepsilon[\![\in \mathcal{W}$. Thus, $x^\varepsilon(t,\omega)$ is $\mathcal{W}$-measurable.

By construction, $|x^\varepsilon(t,\omega) - x(t,\omega)| \le \varepsilon$. Being the limit of $\mathcal{W}$-measurable processes, $x(t,\omega)$ is also $\mathcal{W}$-measurable.

2. Let $\mathcal{W}_1$ denote the σ-algebra generated by the intervals $[\![t_A, \infty[\![$ with $t \in R_+$. Clearly, $\mathcal{W}_1 \subset \mathcal{W}$. Observe that $\{\tau > t\} \in \mathcal{F}_t$ if τ is a stopping time. Therefore

$$[\![t_{\{r>t\}}, \infty[\![\in \mathcal{W}_1, \quad [\![\tau, \infty[\![= \bigcap_{k,n} \left[\!\!\left[\left(\frac{k}{n}\right)_{\left\{\tau > \frac{k}{n}\right\}}, \infty \right[\!\!\right[\in \mathcal{W}_1 \ ,$$

and hence $\mathcal{W} \subset \mathcal{W}_1$. □

Let $\mathcal{P}$ be the σ-algebra generated by the stochastic intervals $[\![\tau_1, \tau_2[\![$, where the τ_i's are predictable stopping times. This is the σ-algebra of *predictable sets*.

Theorem 4.3.4. 1. *$\mathcal{P}$ is generated also by stochastic intervals $]\!]\tau_1, \tau_2]\!] = \{(t,\omega) : \tau_1(\omega) < t \le \tau_2(\omega)\}$ with τ_1 and τ_2 stopping times.*
2. *$\mathcal{P}$ is generated by adapted left-continuous scalar processes.*
3. *$\mathcal{P}$ is generated by adapted continuous scalar processes.*

Proof. 1. Let τ be a stopping time. Then $\tau + s$ is also a stopping time for positive s. Therefore it is a predictable stopping time: $\tau + s = \lim(\tau + s - n^{-1})$. Hence, $[\![0, \tau + s[\![\in \mathcal{P}$, $[\![0, \tau]\!] = \bigcap_n [\![0, \tau + \frac{1}{n}]\!] \in \mathcal{P}$, $]\!]\tau_1, \tau_2]\!] = [\![0, \tau_2]\!] \setminus [\![0, \tau_1]\!] \in \mathcal{P}$. On the other hand, if τ is a predictable stopping time, $\tau_n < \tau$ for positive τ and $\tau_n \to \tau$, then $[\![0, \tau[\![= \bigcup_n [\![0, \tau_n]\!]$.

2. If $\mathcal{P}_1$ is a σ-algebra generated by adapted left-continuous processes, then the $\mathcal{P}$-measurability of $I_{]\!]\tau_1, \tau_2]\!]}$ and left-continuity of this process imply that $\mathcal{P} \subset \mathcal{P}_1$. Let us show that an adapted left-continuous process $x(t,\omega)$ is $\mathcal{P}$-measurable. Define $x_n(t,\omega) = x(k/n, \omega)$ for $k/n < s \le (k+1)/n$ and $x_n(0,\omega) = x(0,\omega)$. Since $x(t,\omega) = \lim x_n(t,\omega)$, the $\mathcal{P}$-measurability of $x(t,\omega)$ follows from the $\mathcal{P}$-measurability of $x_n(t,\omega)$. The latter is a consequence of the relation

$$\{(\omega,t) : x_n(t,\omega) < a\} = \bigcup_k \left[\!\!\left[\left(\frac{k}{n}\right)_{\left\{x\left(\frac{k}{n},\omega\right) < a\right\}}, \frac{k+1}{n} \right]\!\!\right] \ ,$$

since $\left(\frac{k}{n}\right)_{\left\{x\left(\frac{k}{n},\omega\right)<a\right\}}$ is a stopping time; by what has been proved, the stochastic intervals on the right belong to $\mathcal{P}$.
This results from part 1 and the following fact. If τ is a stopping time, then

$$]\!]0, \tau]\!] = \{(t,\omega) : \exp\{0 \vee (t - \tau)\} = 1\} \ .$$

□

4.3.4 Completely Measurable and Predictable Processes

A stochastic process $x(t,\omega)$ defined on R_+ taking values in a measurable space $(X, \mathcal{B})$ and adapted to the flow $\{\mathcal{F}_t\}$ is said to be *completely measurable* if it is $\mathcal{W}$-measurable and it is *predictable* if it is $\mathcal{P}$-measurable. In what follows, we shall consider only countably-generated σ-algebras $\mathcal{B}$.

(a) *Indistiguishability.* We shall say that two processes $x_1(t,\omega)$ and $x_2(t,\omega)$ are *indistiguishable* if $x_1(t,\omega) = x_2(t,\omega)$ for almost all ω and for all t. If $A \subset \Omega \times R_+$, then $\pi(A)$ is the projection of A on Ω: $\omega \in \pi(A)$ if $\{\omega\} \times R_+ \cap A$ is not empty, with $\{\omega\}$ a singleton. In measure theory, it is proved that if $\mathbf{P}$ is a complete measure, then $\pi(A) \in \mathcal{F}$ if $A \in \mathcal{F} \otimes \mathcal{B}_{R_+}$. This fact may be used to show that $x_1(t,\omega)$ and $x_2(t,\omega)$ are indistinguishable if

$$\mathbf{P}(\pi(\{(\omega;t) : x_1(t,\omega) \neq x_2(t,\omega)\})) = 0 .$$

Theorem 4.3.5. *Let $x_1(t,\omega)$ and $x_2(t,\omega)$ be two $\mathcal{W}$-measurable (or $\mathcal{P}$-measurable) processes. If for every stopping time τ (or predictable stopping time τ) $x_1(t,\omega) = x_2(t,\omega)$ almost everywhere on $\{\tau < \infty\}$, then $x_1(t,\omega)$ and $x_2(t,\omega)$ are indistinguishable.*

The proof of this statement rests on the following fact.

Lemma. *If $A \in \mathcal{W}$ (or $\mathcal{P}$) and $\mathbf{P}(\pi(A)) > 0$, then to each positive ε there corresponds a stopping time τ (predictable stopping time τ) such that $\mathbf{P}\{(\omega : \tau(\omega)) \in A\} > \mathbf{P}(\pi(A)) - \varepsilon$.*

Proof. Consider the case $A \in \mathcal{W}$. Let $\mathcal{K}$ be the class of sets $B \in \mathcal{W}$ for which $\{t : (\omega;t) \in B\}$ contains its infimum. Denote it by $\operatorname{deb} B$ and notice that it is a stopping time. Next let $\tilde{\mathcal{K}}$ be the collection of sets $A \in \mathcal{W}$ for which to every positive ε, there corresponds a $B_\varepsilon \in \mathcal{K}$ such that $B_\varepsilon \subset A$ and $\mathbf{P}(\pi(A)) < P(\pi(B_\varepsilon)) + \varepsilon$. One can show that $\tilde{\mathcal{K}}$ is a monotone collection and since $\tilde{\mathcal{K}}$ contains an algebra generated by stochastic intervals, $\tilde{\mathcal{K}} = \mathcal{W}$. If $A \in \mathcal{W}$, $B_\varepsilon \subset A$, $B_\varepsilon \in \mathcal{K}$ and $\mathbf{P}(\pi(A) < \mathbf{P}(\pi(B_\varepsilon)) + \varepsilon$, then $\operatorname{deb} B_\varepsilon$ is the required stopping time. □

Proof of the theorem. If $\mathbf{P}(\pi(\{(\omega;t) : x_1(t,\omega) \neq x_2(t,\omega)\})) > 0$, there would be a stopping time τ (predictable stopping time τ) such that $\mathbf{P}((\omega;\tau) \in \{(\omega,t) : x_1(t,\omega) \neq x_2(t,\omega)\}) > 0$. In other words, $\mathbf{P}\{x_1(t,\omega) \neq x_2(t,\omega)\} > 0$ which contradicts the hypothesis. □

(b) *Existence of measurable modifications.*

Theorem 4.3.6. 1. *Let $x(t,\omega)$ be an adapted measurable process in a Borel space $(X,\mathcal{B})$. Then it has a $\mathcal{W}$-measurable modification.* 2. *If in addition $x(t,\omega)$ is $\mathcal{F}_{t-}$-measurable for all $t \in R_+$, where $\mathcal{F}_{0-} = \mathcal{F}_0$ and $\mathcal{F}_{t-} = \bigvee_{s<t} \mathcal{F}_s$, $t > 0$, then it has a $\mathcal{P}$-measurable modification.*

Proof. Just as in Theorem 4.2.1 of Sect. 4.2.1, we can confine ourselves to a process $x(t,\omega)$ with values in $[0,1]$. Let H be a Hilbert space as in that theorem, and let H_t be the subspace of H generated by the variables $I_B(x(s(\omega))$, $s \leq t$, Let P_t denote projection on H_t. If $\xi(t,\omega)$ as a function with values in H is measurable, then the same thing will be true of the function $P_t(\xi(t,\omega))$. Let $x_n(t,\omega)$ be just as in Theorem 4.2.1 of Sect. 4.2.1. It suffices to prove the existence of a $\mathcal{W}$-measurable modification for the processes

$$P_t x_n(t,\omega) = \sum I_{C_{nk}}(t) P_t \xi_{nk} \ ,$$

and to this end, for the processes $P_t\xi_{nk}$.

The process $P_t\xi_{nk} = \mathbf{E}(\xi_{nk}|\mathcal{F}_t)$ is a martingale (see Sect. 4.4 below); and so on the basis of Sect. 4.4.3, it has a right-continuous modification. It is $\mathcal{W}$-measurable by virtue of Theorem 4.3.3.

2. Write $x^{(h)}(t,\omega) = P_{t-h\vee 0}x(t,\omega)$. This is also measurable in H. Since by hypothesis

$$\lim_{h\to 0} \mathbf{E}|x(t,\omega) - x^{(h)}(t,\omega)|^2 = 0 \ ,$$

it suffices to prove the existence of a $\mathcal{P}$-measurable modification for $x^h(t,\omega)$ and to this end for $P_{t-h}\xi_{nk}$. This is a martingale adapted to the flow $\mathcal{F}_{t-h\vee 0}$ and its right-continuous modification is predictable since every $\mathcal{F}_{t-h\vee 0}$-adapted step-process has that property. □

4.4 Martingales

4.4.1 Definition and Simplest Properties

Let $T \subset R$ and to each $t \in T$, let there correspond a σ-algebra $\mathcal{F}_t$ so that $\mathcal{F}_t \subset \mathcal{F}_s$ when $t < s$. A family of numerical random variables $\{\xi_t, t \in T\}$ is a *martingale* with respect to $\{\mathcal{F}_t\}$ if: (i) ξ_t is $\mathcal{F}_t$-measurable for all $t \in T$; (ii) $\mathbf{E}\xi_t$ exists; (iii) $\mathbf{E}(\xi_t|\mathcal{F}_s) = \xi_s$ for $s < t$ (equality of random variables is understood everywhere to be with probability 1) Sometimes, one says that $\{\xi_t, \mathcal{F}_t, t \in T\}$ is a martingale. It is possible to speak about martingales without mentioning σ-algebras. It is then kept in mind that the $\mathcal{F}_t$'s are the σ-algebras generated by $\{\xi_s, s \le t, s \in T\}$. We shall be primarily interested in three cases: T is a finite set, T is the set Z_+ of nonnegative integers and $T = R_+$.

A very simple example of a martingale is a process with independent increments in which the expectation of an increment is zero. Less trivial are the following examples.

Example 4.4.1. Let $\eta(t), t \in R_+$ be a process with stationary independent increments for which $\mathbf{E}\exp\{\lambda\eta(t)\}$ exists for some λ. Then $\mathbf{E}\exp\{\lambda\eta(t)\} = \exp\{ta(\lambda)\}$, where $a(\lambda)$ is some number. The process

$$\xi(t) = \exp\{\lambda\eta(t) - ta(\lambda)\}$$

is a martingale with respect to the flow $\{\mathcal{F}_t\}$ generated by $\eta(s)$, $s \le t$.

Example 4.4.2. Let T be arbitrary and let $\{\mathcal{F}_t\}$ satisfy the monotonicity condition. Suppose that η is an arbitrary random variable in R for which $\mathbf{E}|\eta| < \infty$. If $\xi_t = \mathbf{E}(\eta|\mathcal{F}_t)$, then $\{\xi_t, \mathcal{F}_t, t \in T\}$ is a martingale.

If $\{\xi_t\}$ obeys (i), (ii) and (iii′) $\mathbf{E}(\xi_t|\mathcal{F}_s) \leq \xi_s$ for $s < t$, then it is called a *supermartingale*. And if it obeys (i), (ii) and (iii″) $\mathbf{E}(\xi_t|\mathcal{F}_s) \geq \xi_s$ for $s < t$, it is called a *submartingale*. Both are also termed *semimartingales*. It is easy to verify the following properties.

I. If $\{\xi_t\}$ is a martingale and $g(x)$ is convex down, then $g(\xi_t)$ is a submartingale (for instance, $|\xi_t|$ and ξ_t^2).
II. If $\{\xi_t\}$ is a supermartingale and $g(x)$ is convex up and increasing, then $g(\xi_t)$ is also a supermartingale.

More important is that (iii), (iii′) and (iii″) continue to hold for stopping times under certain additional restrictions. These restrictions disappear if T is a finite set. Stopping times are understood here to be random variables τ taking values in $T \cup \{\infty\}$ for which $\{\tau \leq t\} \in \mathcal{F}_t$. When T is at most countable, it means that $\{\tau = t\} \in \mathcal{F}_t$ for all $t \in T$. We shall say that τ is a stopping time in T.

Theorem 4.4.1. *Suppose that T is a finite set and τ_1 and τ_2 are stopping times in T. Then on the set $\{\tau_1 \leq \tau_2\}$, $\mathbf{E}(\xi_{\tau_2}|\mathcal{F}_{\tau_1}) \leq \xi_{\tau_1}$ if $\{\xi_t\}$ is a supermartingale, $\mathbf{E}(\xi_{\tau_2}|\mathcal{F}_{\tau_1}) = \xi_{\tau_1}$ if $\{\xi_t\}$ is a martingale, and $\mathbf{E}(\xi_{\tau_2}|\mathcal{F}_{\tau_1}) \geq \xi_{\tau_1}$ if $\{\xi_t\}$ is a submartingale. $\mathcal{F}_{\tau_1}$ is the σ-algebra of sets A such that $A \cap \{\tau_1 = t\} \in \mathcal{F}_t$ for all $t \in T$.*

Proof. All three relations are proved in similar fashion. Let $\{\xi_t\}$ be a martingale. Notice that $\{\tau_1 < \tau_2\} \in \mathcal{F}_{\tau_1}$ since $\{\tau_1 < \tau_2\} \cap \{\tau_1 = t\} = \{\tau_1 = t\} \cap \{\tau_2 > t\}$. We have to show that

$$\mathbf{E} I_A \xi_{\tau_2} = \mathbf{E} I_A \xi_{\tau_1}$$

for all $A \in \mathcal{F}_{\tau_1}$, $A \subset \{\tau_1 < \tau_2\}$. There is no loss of generality in assuming that $T = \{0, 1, \ldots, n\}$ and $A \subset \{\tau_1 = k\} \cap \{\tau_2 > k\}$. Then

$$I_A \xi_{\tau_2} = I_A \sum_{m=k}^{n-1} (\xi_{m+1} - \xi_m) I_{\{\tau_2 > m\}} + I_A \xi_{\tau_1} \ .$$

Since $I_A I_{\{\tau_2 > m\}}$ is $\mathcal{F}_m$-measurable if $m \geq k$, it follows that $\mathbf{E} I_A I_{\{\tau_2 > m\}} \xi_{m+1} = \mathbf{E} I_A I_{\{\tau_2 > m\}} \xi_m$. Taking the expectation, we arrive at our required result. □

4.4.2 Inequalities. Existence of the Limit

We concentrate first on finite sequences $\xi_1, \ldots, \xi_n$ of random variables forming a martingale, a submartingale or a supermartingale. $\{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$ are the corresponding σ-algebras.

(a) *Inequalities for the maximum.*

Theorem 4.4.2. 1. *If* $\{\xi_k, \mathcal{F}_k, k = x_1, \ldots, x_n\}$ *is a submartingale, then*

$$a\mathbf{P}\{\max_k \xi_k \geq a\} \leq \mathbf{E}(\xi_n \vee 0) \,. \tag{4.4.1}$$

2. *If* $\{\xi_k, \mathcal{F}_k, k = x_1, \ldots, x_n\}$ *is a supermartingale, then*

$$a\mathbf{P}\{\max_k \xi_k \geq a\} \leq \mathbf{E}\xi_0 - \mathbf{E}(\xi_n \wedge 0) \,. \tag{4.4.2}$$

Proof. 1. Let $\tau = k < n$ if $\xi_k \geq a, \xi_{k-1} < a, \ldots, \xi_1 < a$ and let $\tau = n$ if $\xi_{n-1} < a, \ldots, \xi_1 < a$. Then τ is a stopping time relative to $\{\mathcal{F}_k\}$ since $\{\tau = k\} \in \mathcal{F}_k$. Let A be the event $\{\max_k \xi_k \geq a\}$. $A \in \mathcal{F}_\tau$ because $A \cap \{\tau = k\} \in \mathcal{F}_k$. Hence on the basis of Theorem 4.4.1,

$$\mathbf{E}\xi_n I_A \geq \mathbf{E}\xi_\tau I_A \geq a\mathbf{P}(A) \,.$$

It remains to observe that $\mathbf{E}\xi_n I_A \leq \mathbf{E}(\xi_n \vee 0) I_A \leq \mathbf{E}(\xi_n \vee 0)$.

2. Retaining the preceding notation, we have $\mathbf{E}\xi_\tau \leq \mathbf{E}\xi_0$ and

$$\mathbf{E}\xi_\tau = \mathbf{E}\xi_\tau I_A + \mathbf{E}\xi_\tau(1 - I_A) \geq a\mathbf{P}(A) + \mathbf{E}\xi_n(1 - I_A) \,.$$

Therefore

$$a\mathbf{P}(A) \leq \mathbf{E}\xi_\tau - \mathbf{E}\xi_n(1 - I_A) \leq \mathbf{E}\xi_\tau - \mathbf{E}(\xi_n \wedge 0)(1 - I_A) \leq \mathbf{E}\xi_0 - \mathbf{E}(\xi_n \wedge 0).$$

□

Corollary. *If* $\{\xi_k, \mathcal{F}_k, k = 1, \ldots, n\}$ *is a supermartingale, then*

$$a\mathbf{P}\{\sup_k |\xi_k \geq a\} \leq \mathbf{E}|\xi_0| + \mathbf{E}|\xi_n|.$$

(b) *Inequalities for the number of crossings.* A numerical sequence x_1, $x_2, \ldots, x_n$ is said to cross the band $[a, b] (a < b)$ at least k times downward from above if there are numbers $i_1 < i_2 < \ldots < i_{2k-1} < i_{2k}$ such that $x_{i_1} \leq a, x_{i_3} \leq a, \ldots, x_{i_{2k-1}} \leq a, x_{i_2} \geq b, x_{i_4} \geq b, \ldots, x_{i_{2k}} \geq b$. The number of crossings upward from below is defined in similar fashion and their sum is the number of crossings of $[a, b]$ by the sequence $x_1, \ldots, x_n$.

Theorem 4.4.3. *Let* $\{\xi_k, \mathcal{F}_k, k = 1, \ldots, n\}$ *be a supermartingale. If* $\nu_+[a, b]$ *is the number of times the sequence* $\xi_1, \ldots, \xi_n$ *crosses* $[a, b]$ *upward from below, then*

$$(b - a)\mathbf{E}\nu_+[a, b] \leq \mathbf{E}(a - \xi_n) \vee 0 \tag{4.4.3}$$

Proof. Consider stopping times in $\{1, 2, \dots, n\}$ defined by $\tau_1 = \inf(\{i : \xi_i \le a\} \cup \{n\})$, $\tau_2 = \inf(\{i \ge \tau_1 : \xi_1 \ge b\} \cup \{n\})$ and so on. Let $A_m = \{\nu_+[a,b] \ge m\} = \{\tau_{2m-1} < n\} \cap \{\xi_{2m} \ge b\}, \{\tau_{2m-1} < n\} \in \mathcal{F}_{\tau_{2m-1}}$. Therefore

$$\begin{aligned} 0 &\ge \mathbf{E}(\xi_{\tau_{2m}} - \xi_{\tau_{2m-1}}) I_{\{\tau_{2m-1} < n\}} \\ &\ge (b-a)\mathbf{P}(A_m) + \mathbf{E}(\xi_{\tau_{2m}} - a) I_{\{\tau_{2m-1} < n, \xi_{\tau_{2m}} < b\}} \\ &\ge (b-a)\mathbf{P}(A_m) + \mathbf{E}(\xi_n - a) I_{\{\tau_{2m-1} < n, \xi_{\tau_{2m}} < b\}}. \end{aligned}$$

Since the events $\{\tau_{2m-1} < n, \xi_{\tau_{2m}} < b\} \subset \{\tau_{2m-1} < n, \tau_{2m} = n\}$ are mutually exclusive, it follows that

$$\begin{aligned} (b-a) \sum \mathbf{P}(A_m) &\le \mathbf{E}[(a - \xi_n) \vee 0] \\ &\times \sum I_{\{\tau_{2m-1} < n, \tau_{2m} = n\}} \le \mathbf{E}[(a - \xi_n) \vee 0]. \end{aligned}$$

But

$$\sum \mathbf{P}(A_m) = \mathbf{E}\nu_+[a,b]. \qquad \square$$

(c) *Limit theorem.* We next consider an infinite sequence $\{\xi_k, \mathcal{F}_k, k = 1, 2, \dots\}$ which is either a super- or submartingale. We wish to find conditions under which the sequence has a limit with probability 1.

Theorem 4.4.4. *Let $\{\xi_k, \mathcal{F}_k, k = 1, 2, \dots\}$ be a supermartingale for which $\inf_n \mathbf{E}(\xi_n \wedge 0) > -\infty$. Then $\lim\limits_{n\to\infty} \xi_n$ exists with probability* 1.

Proof. If $\{x_n\}$ is a numerical sequence, then it has a limit if (a) it is bounded and (b) it crosses any $[r_1, r_2]$, with $r_1 < r_2$ rational numbers, finitely many times. Write $\eta_n^+ = \max\{\xi_1, \dots, \xi_n\}, \eta_n^- = \max\{-\xi_1, \dots, -\xi_n\}, \eta^+ = \lim\limits_{n\to\infty} \eta_n^+$, and $\eta^- = \lim\limits_{n\to\infty} \eta_n^-$. Let $\nu^n[r_1, r_2]$ be the number of times the sequence $\xi_1, \dots, \xi_n$ crosses $[r_1, r_2]$ and $\nu[r_1, r_2] = \lim\limits_{n\to\infty} \nu^n[r_1, r_2]$ the number of times that the infinite sequence crosses $[r_1, r_2]$. Since $\{-\xi_k\}$ is a submartingale, Theorem 4.4.2 implies for positive a that

$$\begin{aligned} &a\mathbf{P}\{\eta_n^+ \ge a\} \le \mathbf{E}\xi_0 - \mathbf{E}(\xi_n \wedge 0), \quad a\mathbf{P}\{\eta^+ \ge a\} \le \mathbf{E}\xi_0 - \inf_n \mathbf{E}(\xi_n \wedge 0), \\ &a\mathbf{P}\{\eta_n^- \ge a\} \le \mathbf{E}(-\xi_n) \vee 0 = -\mathbf{E}(\xi_n \wedge 0), \\ &a\mathbf{P}\{\eta^- \ge a\} \le -\inf_n \mathbf{E}(\xi_n \wedge 0). \end{aligned}$$

Hence, $\mathbf{P}\{\sup\limits_k |\xi_k| \ge a\} \le \frac{1}{a}(\mathbf{E}\xi_0 - 2\inf\limits_n \mathbf{E}(\xi_n \wedge 0))$. The sequence $\{\xi_k\}$ is bounded almost surely. Since $\nu^n[r_1, r_2] \le 2\nu_+[r_1, r_2] + 1$,

$$\mathbf{E}\nu[r_1, r_2] \le \frac{1}{(b-a)} \left(1 + 2(|a| - \inf_n \mathbf{E}(\xi_n \wedge 0))\right) < \infty$$

on the basis of Theorem 4.4.3. $\square$

Corollary. *A nonnegative supermartingale has a limit with probability* 1.

Remark. Let $\{\xi_n, \mathcal{F}_n, n = 1, 2, \ldots\}$ be a uniformly integrable martingale. Then $\sup_n \mathbf{E}|\xi_n| < \infty$ and by Theorem 4.4.4, $\lim\limits_{n\to\infty} \xi_n = \xi_\infty$ exists. If $A_m \in \mathcal{F}_m$,

$$\mathbf{E}\xi_\infty I_{A_m} = \lim_{n\to\infty} \mathbf{E}\xi_n I_{A_m} = \mathbf{E}\xi_m I_{A_m}$$

(taking the limit under the integral sign is permissible in view of the uniform integrability). Thus

$$\xi_m = \mathbf{E}(\xi_\infty|\mathcal{F}_m). \tag{4.4.4}$$

Conversely, if $\{\xi_n\}$ is representable as (4.4.4), then it is a uniformly integrable martingale. That it is a martingale is obvious. The uniform integrability is a consequence of the next assertion.

Lemma. *Let $\mathbf{E}|\xi| < \infty$ and let $\{\mathcal{F}_\theta, \theta \in \Theta\}$ be a collection of σ-algebras with $\mathcal{F}_\theta \subset \mathcal{F}$. Then the family of random variables $\{\eta_\theta = \mathbf{E}(\xi|\mathcal{F}_\theta), \theta \in \Theta\}$ is uniformly integrable.*

Proof. Clearly $\mathbf{E}|\eta_\theta| \le \mathbf{E}|\xi|$. $\mathbf{P}\{|\eta_\theta| > c^2\} \le c^{-2}\mathbf{E}|\xi|$. Therefore

$$\begin{aligned}\mathbf{E}|\eta_\theta| I_{\{|\eta_\theta|>c^2\}} &= \mathbf{E}|\mathbf{E}(\xi|\mathcal{F}_\theta)| I_{\{|\eta_\theta|>c^2\}} \le \mathbf{E}\mathbf{E}(|\xi|\big|\mathcal{F}_\theta) I_{\{|\eta_\theta|>c^2\}} \\ &= \mathbf{E}|\xi| I_{\{|\eta_\theta|>c^2\}} = \mathbf{E}|\xi| I_{\{|\xi|\le c\}} I_{\{|\eta_\theta|>c^2\}} \\ &+ \mathbf{E}|\xi| I_{\{|\xi|>c\}} I_{\{|\eta_\theta|>c^2\}} \le c\frac{\mathbf{E}|\xi|}{c^2} + \mathbf{E}|\xi| I_{\{|\xi|>c\}}\end{aligned}$$

The right-hand side does not depend on θ and approaches zero as $c \to \infty$. □

Corollary. *Let $\mathcal{F}_n \subset \mathcal{F}_{n+1}$, $\mathcal{F}_\infty = \bigvee_n \mathcal{F}_n$ and $\mathbf{E}|\xi| < \infty$. Then*

$$\mathbf{E}(\xi|\mathcal{F}_\infty) = \lim_{n\to\infty} \mathbf{E}(\xi|\mathcal{F}_n)$$

with probability 1.

4.4.3 Continuous Parameter

Let $T \subset R_+$. Assume that $\{F_t, t \in R_+\}$ is a flow of σ-algebras.

Theorem 4.4.5. *Suppose that $\{\xi_t, \mathcal{F}_t, t \in R_+\}$ is a supermartingale and that the family $\{\xi_t, t \le s\}$ is uniformly integrable for any s. Then ξ_t has a right-continuous modification if $\mathbf{E}\xi_t$ is right-continuous.*

Proof. Let D_+ be the set of nonnegative rationals. Using Theorem 4.4.2 and 4.4.3, one can show that $\sup\limits_{t\in D_+\cap[0,s]} |\xi_t|$ is finite with probability 1 for every s. The same is true for $\nu(D_+ \cap [0, s], r_1, r_2)$, the number of crossings of the band $[r_1, r_2]$ by the collection $\{\xi_t, t \in D_+ \cap [0, s]\}$; it is defined as the supremum of the number of crossings of the band $[r_1, r_2]$ by $\{\xi_1, \ldots, \xi_n\}$ over all n and all $t_1 < t_2 < \ldots t_n$ in $D_+ \cap [0, s]$. Therefore

$$\lim_{u\in D^+, u\downarrow t} \xi_u = \xi_t^*$$

exists for all t. Let us show that $\xi_t^* = \xi_t$ with probability 1. Let $u_n \downarrow t$ with $u_n \in D_+$. Then $\mathbf{E}I_A\xi_{u_n} \le \mathbf{E}I_A\xi_t$ for all $A \in \mathcal{F}_t$. Utilizing the uniform integrability, we see that $\mathbf{E}I_A\xi_t^* \le \mathbf{E}I_A\xi_t$ so that $\xi_t^* \le \xi_t$ with probability 1. But $\mathbf{E}\xi_t^* = \lim \mathbf{E}\xi_{u_n} = \mathbf{E}\xi_t$ and hence $\mathbf{P}\{\xi_t = \xi_t^*\} = 1$. □

Corollary. *If $\{\xi_t, \mathcal{F}_t, t \in R_+\}$ is a martingale, then ξ_t has a right-continuous modification.*

4.5 Stochastic Integrals and Integral Representations of Random Functions

We shall consider complex-valued random variables on a given probability space $(\Omega, \mathcal{F}, \mathbf{P})$ that belong to $L_2(\Omega, \mathbf{P})$, as well as random functions with such values. $L_2(\Omega, \mathbf{P})$ is now a complex Hilbert space with the inner product $\langle \xi, \eta \rangle = \mathbf{E}\xi\bar{\eta}$.

4.5.1 Random Measures

Let $(X, \mathcal{B})$ be a measurable space. Consider a complex-valued function $\mu(B)$ defined on $\mathcal{B}$ that satisfies the following:

A. There exists a finite measure m on $\mathcal{B}$ such that

$$\mathbf{E}\mu(B_1)\overline{\mu(B_2)} = m(B_1 \cap B_2), \quad B_1, B_2 \in \mathcal{B}\,. \tag{4.5.1}$$

Then $\mu(B)$ is called a *random measure*. This term is warranted for the following reasons.

B. If B_1 and $B_2 \in \mathcal{B}$, $B_1 \cap B_2 = \emptyset$, then $\mu(B_1 \cup B_2) = \mu(B_1) + \mu(B_2)$. To see this, consider

$$\begin{aligned}&\mathbf{E}|\mu(B_1 \cup B_2) - \mu(B_1) - \mu(B_2)|^2\\ &\quad = m(B_1 \cup B_2) - 2m(B_1) - 2m(B_2) + m(B_1) + m(B_2) = 0\,.\end{aligned}$$

C. If $\{B_n, n \ge 1\} \subset \mathcal{B}$, $B_i \cap B_j = \emptyset$ for $i \ne j$, then

$$\mu\left(\bigcup_n B_n\right) = \sum_n \mu(B_n)\,. \tag{4.5.2}$$

This is a consequence of the relation

$$\mathbf{E}\left|\mu\left(\bigcup_n B_n\right) - \sum_{n=1}^{l} \mu(B_n)\right|^2 = m\left(\bigcup_n B_n\right) - \sum_{n=1}^{l} m(B_n)\,.$$

We point out one further important property which can be used to extend stochastic measures.

D. Let $\mathcal{B}_0$ be a subalgebra of $\mathcal{B}$ generating $\mathcal{B}$ and let $\mu(B)$ be given on $\mathcal{B}_0$ and satisfy (4.5.1) for B_1 and $B_2 \in \mathcal{B}_0$. Finally, let m be a finite measure on $\mathcal{B}$. Then there exists an extension of $\mu(B)$ to a stochastic measure on $\mathcal{B}$. This extension can be determined by taking the limit over a monotone sequence of sets: $\lim \mu(B_n)$ exists for each monotone sequence B_n. For example, for an increasing sequence, with $n < k$, $\mathbf{E}|\mu(B_k) - \mu(B_n)|^2 = m(B_k - B_n) \to 0$.

(a) *Stochastic integrals.* Let $L_2(m)$ be the space of complex-valued $\mathcal{B}$-measurable functions $\varphi(x)$ defined on X for which $\int |\varphi(x)|^2 m(x)dx < \infty$ It is also a complex Hilbert space. Let $H_2(\mu)$ be the subspace of $L_2(\Omega, \mathbf{P})$ generated by $\{\mu(B), B \in \mathcal{B}\}$ and let $H^0(\mu)$ be the linear span of this set. Finally, let $B_0(X)$ be the space of simple functions in $L_2(m)$, that is, the linear span of $\{I_B(x), B \in \mathcal{B}\}$. Define a mapping I from $B_0(X)$ to $H^0(\mu)$ by

$$I\left(\sum c_k I_{B_k}\right) = \sum c_k \mu(B_k) \; .$$

I is a linear isometry. If the B_k's are disjoint, then

$$\int \left|\sum c_k I_{B_k}(x)\right|^2 m(dx) = \sum |c_k|^2 m(B_k) = \mathbf{E}\left|\sum c_k \mu(B_k)\right|^2 \; .$$

Therefore I can be extended to $L_2(m)$ as an isometry and it is a bijection from $L_2(m)$ to $H_2(\mu)$.

The random variable $I(\varphi)$ defines the *stochastic integral* of a function φ with respect to the measure μ and it is denoted by $\int \varphi d\mu = \int \varphi(x)\mu(dx)$.

A stochastic integral is uniquely determined by its following properties: 1. $\int I_B d\mu = \mu(B)$; 2. a stochastic integral is a homogenous linear function; 3. $\mathbf{E} \int \varphi du \overline{\int \psi d\mu} = \int \varphi \bar{\psi} dm$.

4.5.2 Karhunen's Theorem

A stochastic integral can be used to express a random function defined on Θ with values in $L_2(\Omega, P)$ as follows:

$$\xi(\theta) = \int k(\theta, y)\mu(dy) \; . \tag{4.5.3}$$

where $k(\theta, y)$ as a function of y belongs to $L_2(m)$ for all $\theta \in \Theta$. Consider the function

$$r(\theta_1, \theta_2) = \mathbf{E}\xi(\theta_1)\overline{\xi(\theta_2)} \; . \tag{4.5.4}$$

If $\xi(\theta)$ has the representation (4.5.3), then by property 3, $r(\theta_1, \theta_2)$ can be expressed as

$$r(\theta_1.\theta_2) = \int k(\theta_1, y)k(\theta_2, y)m(dy) \; . \tag{4.5.5}$$

A more profound result is due to Karhunen.

Theorem 4.5.1. *Suppose that $\xi(\theta)$ is a random function for which $\mathbf{E}|\xi(\theta)|^2 < \infty$ and that the function $r(\theta_1, \theta_2)$ defined by* (4.5.4) *has the representation* (4.5.5) *with m a finite measure and $k(\theta, \cdot) \in L_2(m)$ for all $\theta \in \Theta$. Then there exists a random measure μ such that $\xi(\theta)$ is representable as* (4.5.3).

Proof. Let $\mathcal{K}_0 \subset L_2(m)$ be the linear span of the set of functions $\{k(\theta, \cdot), \theta \in \Theta\}$ and $\mathcal{K}$ its closure. Let $H^0(\xi)$ be the linear span of the random variables $\{\xi(\theta), \theta \in \Theta\}$ and $H_2(\xi)$ its closure in $L_2(\Omega, \mathbf{P})$. There is no loss of generality in assuming that the Hilbert space $L_2(\Omega, \mathbf{P}) \ominus H_2(\xi)$ is of dimensionality no less than that of $L_2(m) \ominus \mathcal{K}$. This is because the original probability space may always be expanded by replacing it by $(\Omega \times \Omega', \mathcal{F} \otimes \mathcal{F}', \mathbf{P} \times \mathbf{P}')$, where $(\Omega', \mathcal{F}', \mathbf{P}')$ is arbitrary. Now define a mapping $S : L_2(m) \to L_2(\Omega, \mathbf{P})$ as follows: $Sk(\theta, \cdot) = \xi(\theta), \theta \in \Theta$, and $S\varphi_\alpha = \eta_\alpha$, where $\{\varphi_\alpha\}$ is an orthonormal system in $L_2(m) \ominus \mathcal{K}$ and $\{\eta_\alpha\}$ is an orthonormal system in $L_2(\Omega, \mathbf{P}) \ominus H(\xi)$. Then S can be extended linearly. From (4.5.4) and (4.5.5), it follows that S is an isometry. Define $SI_B = \mu(B)$. Then

$$\mathbf{E}\mu(B_1)\overline{\mu(B_2)} = \int I_{B_1}(x) I_{B_2}(x) m(dx) = m(B_1 \cap B_2) \, .$$

Hence, $\mu(B)$ is a random measure. Let $H_2(\mu)$ be defined as in Sect. 4.5.1. Then S maps $L_2(m)$ onto $H_2(\mu)$ and $S_\varphi = \int \varphi(y)\mu(dy)$. Thus

$$\xi(x) = S(k, \cdot) = \int k(x, y)\mu(dy) \, .$$

□

Remark. Let Π be orthogonal projection of $H_2(\mu)$ on $H_2(\xi)$ and $\tilde{\mu}(B) = \Pi\mu(B)$. Since $\Pi\xi(\theta) = \xi(\theta)$, it follows that $\xi(\theta) = \int k(\theta, y)\tilde{\mu}(dy)$; $\tilde{\mu}$ is also a random measure and satisfies $\mathbf{E}\tilde{\mu}(B_1)\overline{\tilde{\mu}(B_2)} = \tilde{m}(B_1 \cap B_2) = \mathbf{E}|\Pi\mu(B_1 \cap B_2)|^2 \leq m(B_1 \cap B_2)$. Using this fact, we can confirm the truth of the following statement:

If (4.5.5) holds for $R(\theta_1, \theta_2)$ with some finite measure m, then a representation (4.5.5) exists with $\tilde{m}$ the smallest measure of all those for which this representation is possible; the random measure $\tilde{\mu}$ is the one taking values in $H_2(\xi)$ for which $\mathbf{E}\tilde{\mu}(B_1)\overline{\tilde{\mu}(B_2)} = \tilde{m}(B_1 \cap B_2)$, and

$$\xi(\theta) = \int k(\theta, y)\tilde{m}(dy) \, .$$

4.5.3 Spectral Representation of Some Random Functions

(a) *Stationary sequences.* Consider a complex-valued random sequence $\{\xi_n, n = 0, \pm 1, \pm 2, \ldots\}$. It is (wide-sense) *stationary* if $\mathbf{E}|\xi_n|^2 < \infty$, $\mathbf{E}\xi_n$ does not depend on n and $\mathbf{E}\xi_n\bar{\xi}_k$ depends only on $n - k$. Let $r_k = \mathbf{E}\xi_k\bar{\xi}_0 - |a|^2$, where $a = \mathbf{E}\xi_n$.

Then r_k is a positive-definite sequence (this means that the quadratic form $\sum_{k,j=1}^{n} r_{k-j} z_k \bar{z}_j$ is nonnegative-definite for all n). By Bochner's theorem below, r_k can be expressed as $r_k = \int_{-\pi}^{\pi} e^{ik\lambda} d\sigma(\lambda)$, where $\sigma(\lambda)$ is a nondecreasing bounded function on $[-\pi, \pi[$.

Hence,

$$\mathbf{E}\xi_k \bar{\xi}_1 = |a|^2 + \int_{-\pi}^{\pi} e^{ik\lambda} e^{\overline{il\lambda}} d\sigma(\lambda)$$

and by Karhunen's theorem, there exists a random measure $\mu(d\lambda)$ on $[-\pi, \pi[$ for which $\mathbf{E}|\mu(d\lambda)|^2 = d\sigma(\lambda)$ and $\xi_k = a + \int_{-\pi}^{\pi} e^{ik\lambda} \mu(d\lambda)$.

(b) *Stationary processes.* Consider a complex-valued random process $\xi(t)$, $t \in R$. It is (wide-sense) *stationary* if: 1. $\mathbf{E}|\xi(t)|^2 < \infty$, $t \in R_+$; 2. $\mathbf{E}\xi(t)$ does not depend on t; 3. $\mathbf{E}\xi(t)\overline{\xi(s)}$ depends on the difference $t - s$. Let $a = \mathbf{E}\xi(t)$ and let $r(t) = \mathbf{E}\xi(t)\xi(0)$ be the covariance function; it is positive-definite.

Bochner's Theorem. *If $r(t)$ is continuous at 0, then it can be represented in the form*

$$r(t) = \int e^{i\lambda t} d\sigma(\lambda) \, ,$$

where $\sigma(\lambda)$ is a nondecreasing bounded function.

By Karhunen's theorem, there exists a stochastic measure $\mu(d\lambda)$ such that

$$\xi(t) = \int e^{i\lambda t} \mu(d\lambda) + a \, .$$

5

Limit Theorems

In Chapter 1 it was pointed out that probability laws manifest themselves when considering a large number of random objects. The main instrument for exposing these laws are the limit theorems of probability theory. They comprise a considerable part of the theory and have remained a principal area of its development. Examples of limit theorems may be found in Chap. 2, Sect. 2.1.3 and in Chap. 3, Sects. 3.1–3.3. The present chapter furnishes some general results on the convergence of distributions as well as on two important classes: ergodic theorems and central limit theorems.

5.1 Weak Convergence of Distributions

5.1.1 Weak Convergence of Measures in Metric Spaces

Let X be a complete separable metric space and $\mathcal{B}$ a Borel σ-algebra. Let $\mathcal{M}$ denote the set of finite measures on $\mathcal{B}$ and $\mathbf{C}$ the space of bounded continuous functions from X to R.

Definition. A sequence of measures $\mu_n \in \mathcal{M}$ *converges weakly* to a measure μ if $\lim_{n\to\infty} \int f d\mu_n = \int f d\mu$ for all $f \in \mathbf{C}$ (this fact will be denoted by $\mu_n \Rightarrow \mu$).

We now give a characterization of weak convergence in terms of convergence of measures on sets.

Theorem 5.1.1. *In order that $\mu_n \Rightarrow \mu_0$ it is necessary and sufficient that*

$$\lim_{n\to\infty} \mu_n(X) = \mu_0(X) \quad \text{and} \quad \overline{\lim_{n\to\infty}}\, \mu_n(F) \le \mu_0(F) \tag{5.1.1}$$

for every closed set F.

Necessity. Let $f \in \mathbf{C}$ and let $f \ge I_F$. Then

$$\overline{\lim_{n\to\infty}}\, \mu_n(F) \le \lim_{n\to\infty} \int f d\mu_n = \int f d\mu_0.$$

But the infimum of the right-hand side over all $f \geq I_F$ is $\mu_0(F)$. The necessity of the first part of (5.1.1) is obvious.

Sufficiency. Let G be an open set. It follows from (5.1.1) that $\underline{\lim}_{n\to\infty}\mu_n(G) \geq \mu_0(G)$. Let $A \in \mathcal{B}$ and let int A, $[A]$, and A' be the interior, closure and boundary of A, respectively. Then

$$\begin{aligned}\mu_0(\text{int } A) &\leq \underline{\lim_{n\to\infty}}\, \mu_n(A) \leq \overline{\lim_{n\to\infty}}\, \mu_n(A) \leq \mu_0([A]) \\ &= \mu_0(\text{int } A) + \mu_0(A').\end{aligned}$$

Hence, when $\mu_0(A') = 0$,

$$\lim_{n\to\infty} \mu_n(A) = \mu_0(A).$$

To any $f \in \mathbf{C}$ and positive ε, there are $a_0 < a_1 < \ldots < a_m$ such that $a_0 < f < a_m$, $\mu_0(\{x : f(x) = a_i\}) = 0, i = 0, 1, \ldots, m$ and $a_i - a_{i-1} < \varepsilon$. Thus for $n = 0, 1, 2, \ldots$

$$\left| \int f d\mu_n - \sum a_i \mu_n(\{x : a_i \leq f(x) < a_{i+1}\}) \right| < \varepsilon \mu_n(X),$$

$$\begin{aligned}\lim_{n\to\infty} &\sum a_i \mu_n(\{x : a_i \leq f(x) < a_{i+1}\}) \\ &= \sum a_i \mu_0(\{x : a_i \leq f(x) < a_{i+1}\}),\end{aligned}$$

and so

$$\overline{\lim_{n\to\infty}} \left| \int f d\mu_n - \int f d\mu_0 \right| \leq 2\varepsilon \mu_0(X).$$

□

Remark 5.1.1. A is called a continuity set for measure μ_0 if $\mu_0(A') = 0$. The proof of the theorem establishes that when $\mu_n \Rightarrow \mu_0$, then $\mu_n(A) \to \mu_0(A)$ for all continuity sets of the measure μ_0.

Remark 5.1.2. Let $x_n(\omega), \quad n = 0, 1, 2, \ldots$, be a sequence of X-valued random elements such that $x_n(\omega) \to x(\omega)$ in probability and let μ_n be the distribution of $x_n(\omega)$. Then

$$\mu_n \Rightarrow \mu_0 : \int f d\mu_n = \mathbf{E} f(x_n(\omega)) \to \mathbf{E} f(x_0(\omega)) = \int f d\mu_0.$$

The converse is also true in the following sense.

Theorem 5.1.2. *If $\mu_n \Rightarrow \mu_0$ and the μ_n are probability distributions on $\mathcal{B}$, then there exists a sequence of X-valued random elements $x_n(\omega)$ such that $x_n(\omega) \to x_0(\omega)$ in probability and μ_n is the distribution of $x_n(\omega)$.*

Proof. To each $\varepsilon > 0$, it is possible to partition X into countably many subsets $\{U_k^\varepsilon, k = 1, 2, \ldots\}$ so that diam $(U_k^\varepsilon) < \varepsilon$ and $\mu_n(U_k^\varepsilon) \to \mu_0(U_k^\varepsilon)$. Choose a point x_k^ε in each U_k^ε. Take $([0,1], \mathcal{B}_{[0,1]}, m)$ to be the probability space, where

m is Lebesgue measure. Let $\Delta_{nk}^{\varepsilon}$ be an interval in [0,1] of length $\mu_n(U_k^{\varepsilon})$, where $\Delta_{n1}^{\varepsilon} = (0, \mu_n(U_1^{\varepsilon}))$ and $\Delta_{nk}^{\varepsilon}$ and $\Delta_{nk+1}^{\varepsilon}$ have a common boundary point. The intervals are uniquely determined by these properties. Put $x_n^{\varepsilon}(\omega) = x_k^{\varepsilon}$ if $\omega \in \Delta_{nk}^{\varepsilon}$. Then $x_n^{\varepsilon}(\omega) \to x_0^{\varepsilon}(\omega)$ for all ω not lying at the endpoints of the intervals $\Delta_{0k}^{\varepsilon}$. For $f \in \mathbf{C}$

$$\mathbf{E}f(x_n^{\varepsilon}(\omega)) = \sum f(x_k^{\varepsilon})\mu_0(U_k^{\varepsilon}), \tag{5.1.2}$$

and the right-hand side converges to $\int f d\mu_n$ as $\varepsilon \to 0$. Consider a sequence of such partitions of X corresponding to $\varepsilon \to 0$. One can determine $x_n^{\varepsilon_r}(\omega)$ to satisfy (5.1.2) with $\varepsilon = \varepsilon_r$ and so that $\rho(x_n^{\varepsilon_r}(\omega), x_n^{\varepsilon_s}(\omega)) \le \varepsilon_s$ for $\varepsilon_r < \varepsilon_s$ and all ω (ρ is the metric in X). Taking $x_n(\omega) = \lim_{r\to\infty} x_n^{\varepsilon_r}(\omega)$, we arrive at the required random elements. □

The finite measures on $\mathcal{B}$ are Radon measures. This means that $\mu \in \mathcal{M}$ possesses the property that to every $\varepsilon > 0$, there is a compact set $K \subset X$ such that $\mu(X \setminus K) < \varepsilon$. Measures having this property are said to be *tight*.

A subset $\mathcal{M}_1$ of measures of $\mathcal{M}$ is *uniformly tight* if to every $\varepsilon > 0$, there corresponds a compact subset K of X such that $\mu(X \setminus K) < \varepsilon$ for all $\mu \in \mathcal{M}_1$.

Theorem 5.1.3. *If $\mu_n \Rightarrow \mu_0$, then $\{\mu_n, n \ge 1\}$ is uniformly tight.*

Proof. Assume the contrary. Let $K^{\delta} = \{y : \rho(y, K) < \delta\}$. It is easy to see that if to all $\varepsilon > 0$ and $\delta > 0$, there were a compact K such that $\sup_n \mu_n(X \setminus K^{\delta}) < \varepsilon$, then $\{\mu_n\}$ would be uniformly tight. Thus, by assumption, there exist $\varepsilon > 0$ and $\delta > 0$ such that $\sup_n \mu_n(X \setminus K^{\delta}) > \varepsilon$ for all compact sets K. Choose K_1 so that $\mu_1(X \setminus K_1) < \varepsilon$. One can find an n_2 such that $\mu_{n_2}(X \setminus K_1^{\delta}) \ge \varepsilon$. Then using the tightness of μ_{n_2} on $X \setminus K_1^{\delta}$, one can find a compact set $K_2 \subset X \setminus K_1^{\delta}$ such that $\mu_{n_2}(K_2) \ge \varepsilon/2$. Since $\sup_n \mu_n(X \setminus (K_1 \cup K_2)^{\delta}) \ge \varepsilon$, one can find an n_3 and a compact set $K_3 \subset X \setminus (K_1 \cup K_2)^{\delta})$ such that $\mu_{n_3}(K_3) \ge \varepsilon/2$. Thus we can form a sequence μ_{n_i} and compact sets K_i such that $\mu_{n_i}(K_i) \ge \varepsilon/2$ and the distance between each two compact sets is at least δ. Let $\chi_i \in \mathbf{C}$ with $\chi_i(x) = 1$ if $x \in K_i$ and $\chi_i(x) = 0$ if $\rho(x, K_i) \ge \delta/2$. Then $\Sigma\chi_i(x) \in \mathbf{C}$. For all m,

$$\lim_{i\to\infty} \int \sum_{k=m}^{\infty} \chi_k(x)\mu_{n_i}(dx) = \int \sum_{k=m}^{\infty} \chi_k(x)\mu_0(dx).$$

The right-hand side approaches zero as $m \to \infty$ but

$$\int \sum_{k=m}^{\infty} \chi_k(x)\mu_{n_i}(dx) \ge \frac{\varepsilon}{2},$$

a contradiction. □

Remark. Suppose that $\lim \int f d\mu_n$ exists for all $f \in \mathbf{C}$. Then $\{\mu_n, n \ge 1\}$ is also uniformly tight. For, if this were not so, one could find a sequence n_i and

form functions $\chi_k(x)$ as indicated in the proof of the theorem. Define $\psi_p(x) = \sum_{k=1}^{\infty} \chi_{p^k}(x)$, where p is a prime. $\psi_p(x) \in \mathbf{C}$ and thus $\lim_{n\to\infty} \int \psi_p(x)\mu_n(dx)$ exists. It is easy to see that $\int \psi_p d\mu_{n_{p^i}} \geq \varepsilon/2$ for all i which implies that

$$\lim_{n\to\infty} \int \psi_p(\cdot) d\mu_n \geq \varepsilon/2.$$

Therefore

$$\lim_{n\to\infty} \int \sum_p \psi_p d\mu_n = \infty, \quad \sum \psi_p \in \mathbf{C}$$

and again we have a contradiction.

5.1.2 Weak Compactness

Definition. A subset $\mathcal{M}_1$ of $\mathcal{M}$ is *weakly compact* if every sequence $\mu_n \in M_1$ contains a weakly convergent subsequence.

Every weakly compact sequence having a single limit point is weakly convergent. In particular, μ_n converges weakly to some limit if $\{\mu_n, n \geq 1\}$ is a weakly compact set and $\lim_{n\to\infty} \int f d\mu_n$ exists for $f \in T \subset \mathbf{C}$, where $\mathbf{C}$ is a complete set of functions. We now give conditions for the weak compactness of a set of measures.

Theorem 5.1.4. *A set of measures $\mathcal{M}_1 \subset \mathcal{M}$ is weakly compact if and only if* (i) *$\{\mu(X), \mu \in \mathcal{M}_1\}$ is bounded and* (ii) *$\mathcal{M}_1$ is uniformly tight.*

Proof. The necessity of (i) is obvious and (ii) follows from Theorem 5.1.3. □

Sufficiency. Let K_n be a sequence of compact subsets of X for which $\mu(X \setminus K_n) \leq 2^{-n}$ with $\mu \in \mathcal{M}_1$ and let $c = \sup\{\mu(X), \mu \in \mathcal{M}_1\}$. Let $\mathbf{C}(K_n)$ be the space of continuous functions on K_n. It is a separable space. Therefore, it is possible to find a sequence $f_m \in \mathbf{C}$ such that to all $f \in \mathbf{C}$, there exists a subsequence f_{m_k} satisfying $\sup_k \|f_{m_k}\| < \infty$ ($\|\cdot\|$ is the norm in $\mathbf{C}$) and

$$\lim_{k\to\infty} \sup_{x\in K_n} |f_{m_k}(x) - f(x)| = 0$$

for all n.

Let $\mu_i \in \mathcal{M}$ and $\lim_{i\to\infty} \int f_m d\mu_i$ exist for all m. If $f \in \mathbf{C}$ and f_{m_k} has been determined as indicated above, then

$$\begin{aligned}\overline{\lim_{\substack{i\to\infty\\ j\to\infty}}} \left| \int f d\mu_i - \int f d\mu_j \right| &\leq \overline{\lim_{\substack{i\to\infty\\ j\to\infty}}} \left| \int_{K_n} f d\mu_i - \int_{K_n} f d\mu_j \right| \\ &+ 2\|f\| 2^{-n} \leq \overline{\lim_{\substack{i\to\infty\\ j\to\infty}}} \left| \int f_{m_k} d\mu_i - \int f_{m_k} d\mu_j \right| \\ &+ 2c \sup_{x\in K_n} \left| f_{m_k}(x) - f(x) \right| + 2\|f\| 2^{-n} .\end{aligned}$$

The first term on the right equals zero and the remaining two may be made arbitrarily small by the choice of n and k. Thus, $\lim \int f d\mu_i$ exists for all $f \in \mathbf{C}$.

Given any sequence $\mu_i \in \mathcal{M}_1$, one can extract a subsequence μ_{i_n} for which $\lim_{n\to\infty} \int f_m d\mu_{i_n}$ exists and hence $\lim_{n\to\infty} \int f d\mu_{i_n}$ exists for all $f \in \mathbf{C}$. Write $l(f) = \lim_{n\to\infty} \int f d\mu_{i_n}$. Using the form of a linear functional on $\mathbf{C}(K)$, where K is a compact set, one can see that $l(f) = \int f d\mu_0$ with $\mu_0 \in \mathcal{M}$.

5.1.3 Weak Convergence of Measures in R^d

Since compact sets in R^d are bounded closed sets, Theorem 5.1.4 leads to the following result. A subset of measures $\mathcal{M}_1$ of $\mathcal{M}$ in R^d is weakly compact if and only if (i) $\{\mu(R^d), \mu \in \mathcal{M}_1\}$ is a bounded set and (ii) to every $\varepsilon > 0$, there is an r such that $\mu(R^d \setminus S_r) < \varepsilon$ where S_r is a ball of radius r. Notice that $\{e^{i(z,x)}, z \in R^d\}$ is a complete set of functions. Starting with these two facts, one can establish the next assertion.

Theorem 5.1.5. *Let $\{\mu_n, n = 0, 1, 2, \ldots\}$ be a sequence of probability distributions in R^d and let $\varphi_n(z) = \int \exp i(z,x)\mu_n(dx)$ be their characteristic functions.*

1. *If $\mu_n \Rightarrow \mu_0$, then $\varphi_n(z) \to \varphi_0(z)$ uniformly on each bounded set;*
2. *If $\varphi_n(z) \to \psi(z)$ and $\psi(z)$ is a continuous function, then $\mu_n \Rightarrow \mu_0$, where μ_0 is some probability measure.*

Proof. Statement 1 follows because $\varphi_n(z) \to \varphi_0(x)$ for all z and because

$$\begin{aligned} |\varphi_n(z) - \varphi_n(z_1)| &\le \int_{S_r} |e^{i(z,x)} - e^{i(z_1,x)}|\mu_n(dx) + 2\varepsilon \\ &\le r|z_1 - z_2| + 2\varepsilon \end{aligned}$$

if r is chosen so that $\mu_n(S_r) \ge 1 - \varepsilon$ for all n.

Let us show that $\{\mu_n, n \ge 1\}$ is weakly compact under hypothesis 2. Let $z^1, \ldots, z^d$ and $x^1, \ldots, x^d$ be the respective coordinates of z and x. Then

$$\left(\frac{1}{2\delta}\right)^d \int_{-\delta}^{\delta} \cdots \int_{-\delta}^{\delta} \exp\{i\Sigma x^k z^k\} dz^1 \ldots dz^d = \prod_{k=1}^{d} \frac{\sin \delta x^k}{\delta x^k} ,$$

$$\left(\frac{1}{2\delta}\right)^d \int_{-\delta}^{\delta} \cdots \int_{-\delta}^{\delta} \varphi_n(z) dz^1 \ldots dz^d = \int \prod_{k=1}^{d} \frac{\sin \delta x^k}{\delta x^k} \mu_n(dx) ,$$

$$\left(\frac{1}{2\delta}\right)^d \int_{-\delta}^{\delta} \cdots \int_{-\delta}^{\delta} (1 - \varphi_n(z)) dz^1 \ldots dz^d = \int \left(1 - \prod_{k=1}^{d} \frac{\sin \delta x^k}{\delta x^k}\right) \mu_n(dx) .$$

Using the convergence of $\varphi_n(z)$ to a continuous function $\psi(z)$ and the fact that $\varphi_n(0) = \psi(0) = 1$, we can choose a $\delta > 0$ so that

$$\sup_n \int \left(1 - \prod_{k=1}^{d} \frac{\sin \delta x_k}{\delta x_k}\right) \mu_n(dx) < \varepsilon \, .$$

Then

$$\sup_n \mu_n(X \setminus S_r) \leq \varepsilon \left(1 - \sup_{|x|>r} \prod_{k=1}^{d} \frac{\sin \delta x^k}{\delta x^k}\right)^{-1} .$$

From the convergence of the characteristic functions and the completeness of the set of functions $\{\exp i(z,x),\ z \in R^d\}$, it follows that μ_n converges weakly to a measure μ_0. □

5.2 Ergodic Theorems

The simplest illustration of an ergodic theorem is the strong law of large numbers for independent and identically distributed random variables (Chap. 3, Sect. 3.2.4). Such theorems assert the existence of the limits of the means along the path of a stochastic process. We shall only consider processes with discrete time and we shall discuss the existence of the limit of the means of random sequences.

5.2.1 Measure-Preserving Transformations

Let there be given a σ-finite measure μ on a measurable space $(X, \mathcal{B})$ and a measurable mapping T from $(X, \mathcal{B})$ to $(X, \mathcal{B})$. Denote the image of x under T by Tx (we are not assuming that X has a linear structure and so this notation should cause no confusion). The transformation T preserves μ if $\mu(T^{-1}B) = \mu(B)$ for $B \in \mathcal{B}$, where $T^{-1}B$ is the pre-image of B under T. Transformations that preserve a measure usually occur when considering dynamical systems. They also arise in probability theory in a natural way.

(a) *Strict-sense stationary sequences.* Let $(X, \mathcal{B})$ be a measurable space and let $\{\xi_n,\ n = 0, 1, 2, \dots\}$ be a sequence of random elements with values in $(X, \mathcal{B})$. It is called *stationary in the strict sense* if for all n the joint distributions of $\xi_0, \dots, \xi_n$ and $\xi_1, \dots, \xi_{n+1}$ are the same. It easily follows from this that the joint distribution of $\xi_0, \dots, \xi_n$ and $\xi_m, \dots, \xi_{n+m}$ are the same for any m. In other words, the distribution of a segment of the sequence does not depend on the shifting of time (or the starting reference point). Many systems in the absence of external effects become almost stationary after the passage of a long enough time. This section will examine stationarity in the strict sense.

Let X^∞ be the space of sequences $(x_0, x_1, \dots)$ and let $\mathcal{B}^\infty$ be the cylinder σ-algebra in X^∞. The finite-dimensional distribution functions of the sequence

$\{\xi_n,\ n = 0, 1, 2, \ldots\}$ generate a probability measure μ on $\mathcal{B}^\infty$. Let T be the shifting operator on X^∞ defined by

$$T(x_0, x_1, \ldots) = (x_1, x_2, \ldots) .$$

The stationarity of a sequence is equivalent to T preserving the measure μ. Observe that: 1. a sequence of independent and identically distributed random variables $\{\xi_n,\ n = 0, 1, 2, \ldots\}$ is stationary; 2. if $f(x_0, x_1, \ldots)$ is $\mathcal{B}^\infty$-measurable, then $\eta_k = f(\xi_k, \xi_{k+1}, \ldots)$ is stationary when $\{\xi_n,\ n = 0, 1, 2, \ldots\}$ is stationary.

(b) *Homogeneous Markov chains.* A function $Q(t, B)$ defined on $X \times \mathcal{B}$ and assuming values in $[0, 1]$ is called a transition probability if: 1. it is $\mathcal{B}$-measurable for fixed $B \in \mathcal{B}$; 2. it is a probability distribution on $\mathcal{B}$ for fixed $x \in X$. If Q_1 and Q_2 are two transition probabilities, then their composition $Q_1 * Q_2(x, B) = \int Q_1(x, dy) Q_2(y, B)$ is well defined; it is also a transition probability. A random sequence $\{\xi_n,\ n = 0, 1, 2, \ldots\}$ is said to be a *homogeneous* Markov chain (or a Markov chain with stationary transition probabilities) if

$$\mathbf{P}\{\xi_{n+1} \in B \mid \xi_0, \ldots, \xi_n\} = Q(\xi_n, B) \quad \text{with probability 1}$$

for all n, where Q is a transition probability. If ξ_k is viewed as the state of a system at time k, then for a Markov sequence the probability of being in some state at the nth step depends only on the state at the preceding moment of time (this is the Markov property) and it does not depend on the moment of time (this is the stationarity in time).

Let ν_0 be the distribution of ξ_0. The stated definition leads to the following formula for the finite-dimensional distributions of a Markov sequence:

$$\begin{aligned} &\mathbf{P}\{\xi_0 \in B_0,\ \xi_1 \in B_1, \ldots, \xi_n \in B_n\} \\ &\quad = \int_{B_0} \nu(dx_0) \int_{B_1} Q(x_0, dx_1) \ldots \int_{B_n} Q(x_{n-1}, dx_n) . \end{aligned} \tag{5.2.1}$$

The transition probability Q appearing in (5.2.1) is the transition probability of the chain $\{\xi_n,\ n = 0, 1, 2, \ldots\}$. When is this chain a stationary sequence?

A σ-finite measure ν is said to be *invariant* under a transition probability Q if

$$\nu(B) = \int Q(x, B) \nu(dx)$$

for all $B \in \mathcal{B}$. It is easy to see that a Markov chain with transition probablility Q is stationary if and only the distribution ν_0 of ξ_0 is invariant under Q.

We now state a simple and widely-used condition for the existence of invariant measures.

Lemma. *Let X be a compact space and $\mathcal{B}$ the σ-algebra of its Borel sets. If $Q(x, B)$ has the property that $\int f(y) Q(x, dy) \in \mathbf{C}$ for all $f \in \mathbf{C}$, then there exist invariant probability measures under Q.*

Proof. Define

$$Q_1 = Q, \quad Q_{n+1}(x, B) = \int Q_n(x, dy)Q(y, B) \; .$$

For all n and $f \in \mathbf{C}$,

$$\int f(y)Q_n(x, dy) \in \mathbf{C} \; .$$

Put

$$R_n(x, B) = \frac{1}{n} \sum_{k=1}^{n} Q_k(x, B) \; .$$

The sequence of measures $R_n(x, \cdot)$ is weakly compact for any x. Let n_i be a sequence such that $R_{n_i}(x, \cdot) \Rightarrow \nu$. Since R_n is a transition probability, ν is a probability measure. For $f \in \mathbf{C}$ it follows that

$$\begin{aligned}
&\int \nu(dy) \int Q(y, dz)f(z) \\
&\quad = \lim_{i\to\infty} \int \frac{1}{n_i} \sum_{k=1}^{n_i} Q_k(x, dy) \int Q(y, dz)f(z) \\
&\quad = \lim_{i\to\infty} \frac{1}{n_i} \sum_{k=1}^{n_i} \int Q_{k+1}(x, dz)f(z) \\
&\quad = \lim_{i\to\infty} \left(\frac{1}{n_i} \sum_{k=1}^{n_i} \int Q_k(x, dz)f(z) - \frac{1}{n_i} \int Q_1(x, dz)f(z) \right. \\
&\qquad \left. + \frac{1}{n_i} \int Q_{n_i+1}(x, dz)f(z) \right) .
\end{aligned}$$

If $\tilde{\nu}(B) = \int Q(y, B)\nu(dy)$, then $\int f(z)\tilde{\nu}(dz) = \int f(z)\nu(dz)$, $\tilde{\nu} = \nu$ and so ν is invariant under Q. □

5.2.2 Birkhoff's Theorem

Let $(X, \mathcal{B})$ be a measurable space and let T be a transformation preserving σ-additive measure μ. Denote by $T^k x$ the k-fold application of T to x; $T^0 x = x$. The sequence $\{x, Tx, T^2x, \ldots\}$ is the path (or orbit) of the point x. We are interested in the behavior of

$$\sum_{k=0}^{n} f(T^k x) = S_n(f, x)$$

as $n \to \infty$ where f is a $\mathcal{B}$-measurable scalar function. Let $L_1(X, \mu)$ be the space of $\mathcal{B}$-measurable and μ-integrable scalar functions.

Theorem 5.2.1. *Suppose that*

$$U_n = \{x : S_n(f,x) \geq 0\}, \quad V_n = \bigcup_{k=0}^{n} U_k \quad \text{and} \quad V = \bigcup_n V_n \ .$$

If $f \in L_1(X,\mu)$, *then* $\int_V f(x)\mu(dx) \geq 0$.

This result is known as the *maximal ergodic theorem.*

Proof. Let

$$\hat{S}_n(f,x) = \max_{k \leq n} S_k(f,x) \ .$$

Clearly, $V_n = \{x : \hat{S}_n(f,x) \geq 0\}$. If we write $T^* f(x) = f(Tx)$, then

$$\hat{S}_{n+1}(f,x) = f(x) + \hat{S}_n(T^* f, x) I_{\{\hat{S}_n(T^* f,x) \geq 0\}} \ .$$

The second term is nonnegative and so

$$\begin{aligned} &\int_{V_{n+1}} \hat{S}_{n+1}(f,x)\mu(dx) \\ &\quad \leq \int_{V_{n+1}} f(x)\mu(dx) + \int \hat{S}_n(T^* f,x) I_{\{\hat{S}_n(T^* f,x) \geq 0\}} \mu(dx) \\ &\quad = \int_{V_{n+1}} f(x)\mu(dx) + \int \hat{S}_n(f,x) I_{\{\hat{S}_n f, x \geq 0\}} \mu(dx) \end{aligned} \tag{5.2.2}$$

(we have made use of the invariance of μ by virtue of which the integrals in the last equation coincide).

Since $\hat{S}_n(f,x) \leq \hat{S}_{n+1}(f,x)$,

$$\begin{aligned} \int_{V_{n+1}} \hat{S}_{n+1}(f,x)\mu(dx) &= \int \hat{S}_{n+1}(f,x) I_{\{\hat{S}_{n+1}(f,x) \geq 0\}} \mu(dx) \\ &\geq \int \hat{S}_n(f,x) I_{\{\hat{S}_n(f,x) \geq 0\}} \mu(dx) \ . \end{aligned} \tag{5.2.3}$$

From (5.2.2) and (5.2.3), it follows that

$$\int_{V_{n+1}} f(x)\mu(dx) \geq 0 \ .$$

Letting $n \to \infty$, we complete the proof.

The maximal ergodic theorem leads readily to the following *ratio theorem.*

Theorem 5.2.2. *Let* f *and* $g \in L_1(X,\mu)$ *with* g *positive and let* $\sum_{k=1}^{\infty} g(T^k x) = \infty$ μ*-almost everywhere. Then*

1. $\lim_{n\to\infty}(S_n(f,x)/S_n(g,x))$ *exists* μ*-almost everywhere;*
2. *if this limit is denoted by* $f^*(x)$*, then* $f^*(Tx) = f^*(x)$ μ*-almost everywhere and*

$$\int f(x)\mu(dx) = \int g(x)f^*(x)\mu(dx) \; . \tag{5.2.4}$$

Proof. Let Z^β be the set of x for which

$$\sum_{k=1}^{\infty} g(T^k x) = \infty \; , \quad \overline{\lim_{n\to\infty}}\,(S_n(f,x)/S_n(g,x)) > \beta \; .$$

This set is invariant under T (A is invariant under T if $T(A) \subset A$). Let Z_α be the set of x for which

$$\sum_{k=1}^{\infty} g(T^k x) = \infty \; , \quad \underline{\lim_{n\to\infty}}\,(S_n(f,x)/S_n(g,x)) < \alpha \; .$$

This set is also invariant under T and thus so is $Z_\alpha \cap Z^\beta$. Clearly T may be applied to any invariant A and it will preserve the restriction of the measure μ to A. Consider T on $Z_\alpha \cap Z^\beta$. Then

$$\bigcup_k \{x : f(T^k x) - \beta g(T^k x) \geq 0\} \cap Z_\alpha \cap Z^\beta = Z_\alpha \cap Z^\beta \; ,$$

and by Theorem 5.2.1

$$\int_{Z_\alpha \cap Z^\beta} (f(x) - \beta g(x))\mu(dx) \geq 0 \; ,$$

or

$$\int_{Z_\alpha \cap Z^\beta} f(x)\mu(dx) \geq \beta \int_{Z_\alpha \cap Z^\beta} g(x)\mu(dx) \; . \tag{5.2.5}$$

Applying Theorem 5.2.1 to $\alpha g(x) - f(x)$ one can show that

$$\int_{Z_\alpha \cap Z^\beta} f(x)\mu(dx) \leq \alpha \int_{Z_\alpha \cap Z^\beta} g(x)\mu(dx) \; . \tag{5.2.6}$$

Therefore

$$\beta \int_{Z_\alpha \cap Z^\beta} g(x)\mu(dx) \leq \alpha \int_{Z_\alpha \cap Z^\beta} g(x)\mu(dx) \; .$$

With $\beta < \alpha$, we find from this that

$$\int_{Z_\alpha \cap Z^\beta} g(x)\mu(dx) = 0 \; ,$$

in other words, $\mu(Z_\alpha \cap Z^\beta) = 0$ since $g > 0$. From (5.2.5) and (5.2.6), it also follows that

$$\begin{aligned} \int_{Z_\alpha} g(x)\mu(dx) &\le \frac{1}{|\alpha|}\int |f(x)|\mu(dx)\ , \quad \alpha < 0\ , \\ \int_{Z^\beta} g(x)\mu(dx) &\le \frac{1}{\beta}\int |f(x)|\mu(dx)\ , \quad \beta > 0\ , \end{aligned} \tag{5.2.7}$$

Since $g(x) > 0$,

$$\mu\left(\bigcap_{\alpha<0} Z_\alpha\right) = \mu\left(\bigcap_{\beta>0} Z^\beta\right) = 0\ .$$

Hence, $\overline{\lim}(S_n(f,x)/S_n(g,x))$ and $\underline{\lim}(S_n(f,x)/S_n(g,x))$ are finite and are equal μ-almost everywhere. Statement 1 has been proved.

To deduce statement 2, observe that if $D_\alpha^\beta = \{x : f^*(x) \in [\alpha, \beta[\}$, then as in the proof of (5.2.6) and (5.2.7),

$$\begin{aligned} &\beta \int_{D_\alpha^\beta} g(x)\mu(dx) \ge \int_{D_\alpha^\beta} f(x)\mu(dx) \ge \alpha \int_{D_\alpha^\beta} g(x)\mu(dx)\ , \\ &\left|\int_{D_\alpha^\beta} f(x)\mu(dx) - \int_{D_\alpha^\beta} f^*(x)g(x)\mu(dx)\right| \le (\beta - \alpha)\int_{D_\alpha^\beta} g(x)\mu(dx)\ . \end{aligned}$$

Taking h to be arbitrary and positive, we have

$$\begin{aligned} &\left|\int f(x)\mu(dx) - \int f^*(x)g(x)\mu(dx)\right| \\ &\quad \le \sum_{k=-\infty}^{\infty} \left|\int_{D_{kh}^{(k+1)h}} f(x)\mu(dx) - \int_{D_{kh}^{(k+1)h}} f^*(x)g(x)\mu(dx)\right| \\ &\quad \le h\sum \int_{D_{kh}^{(k+1)h}} g(x)\mu(dx) = h\int g(x)\mu(dx)\ . \end{aligned}$$

This implies (5.2.4) and so $f^*(Tx) = f^*(x)$ if $\sum g(T^k x) = \infty$. □

Denote by $\mathcal{T}$ the σ-algebra generated by the invariant sets and the sets of μ-measure zero. A function $\psi(x)$ is $\mathcal{T}$-measurable if $\mu(\{x : \psi(Tx) \ne \psi(x)\}) = 0$. The limit of the ratio occurring in Theorem 5.2.2 is $\mathcal{T}$-measurable.

In probability theory, the most interesting case is that of a probability measure μ.

Theorem 5.2.3. *Suppose that $\mu(X) = 1$ and $f \in L_1(x, \mu)$. Then*

$$\lim_{n\to\infty} \frac{1}{n} S_n(f,x) = f^*(x) \tag{5.2.8}$$

exists μ-almost everywhere, where $f^*(x) = \mathbf{E}_\mu(f(x) \mid \mathcal{T})$, $\mathbf{E}_\mu$ is expectation on the probability space $(X, \mathcal{B}, \mu)$ and $\mathbf{E}_\mu(\cdot \mid \mathcal{T})$ is the conditional expectation given the σ-algebra $\mathcal{T}$.

Proof. Relation (5.2.8) follows from Theorem 5.2.2 by taking $g = 1$ ($\int g d\mu = \mu(X) < \infty$).

The function f^* is $\mathcal{T}$-measurable. Let $A \in \mathcal{T}$. Then $I_A(T^k x) = I_A(x)$ μ-almost everywhere. Thus

$$\lim_{n\to\infty} \frac{1}{n} S_n(fI_A, x) = I_A f^*(x) \, .$$

Applying (5.2.3), we obtain (with $g = 1$)

$$\int_A f(x)\mu(dx) = \int_A f^*(x)\mu(dx) \, , \quad A \in \mathcal{T} \, .$$

□

Corollary. *Let* $\{\xi_n, \ n = 0, 1, 2, \ldots\}$ *be a scalar stationary sequence with* $\mathbf{E}|\xi_0| < \infty$. *Then with probability* 1

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \xi_k = \mathbf{E}(\xi_0 \mid \mathcal{T}) \, .$$

($(\Omega, \mathcal{F}, \mathbf{P})$ is viewed as the space $(X^\infty, \mathcal{B}^\infty, \mu)$ formed in Sect. 5.2.1a with T the shifting transformation).

5.2.3 Metric Transitivity

When does the law of large numbers hold in the sense that the limits in Theorem 5.2.2 and 5.2.3 are constants? Observe that the limit $f^*(x)$ in Theorem 5.2.2 (and hence in Theorem 5.2.3) may turn out to be any bounded $\mathcal{T}$-measurable function. To this end, we need only take $f(x) = f^*(x)g(x)$. $\mathcal{T}$-measurable functions are μ-almost constant if and only if σ-algebra $\mathcal{T}$ is trivial under the measure μ, that is, it is generated by sets of measure zero.

A transformation T is *metrically transitive* under a measure μ, which it preserves, if $\mathcal{T}$ is trivial for μ. In that event, μ is said to be *ergodic* for T. A stationary process is *ergodic* if the probability measure generated by it on the space of sequences is ergodic for the sequence shift transformation.

If μ is ergodic for T, then $f^*(x)$ in Theorem 5.2.2 is a constant f^* and so

$$f^* = \int f d\mu \Big/ \int g d\mu \, .$$

This result may be stated as follows for a stationary process.

Theorem 5.2.4. *Suppose that* $\{\xi_n, \ n = 0, 1, \ldots\}$ *is a scalar stationary ergodic process for which* $\mathbf{E}|\xi_0| < \infty$. *Then with probability* 1

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \xi_k = \mathbf{E}\xi_0 \, .$$

Remark. If μ is a probability measure that is ergodic for T and $f \in L_1(X,\mu)$, then the sequence $\{f(x),\ f(Tx),\dots\}$ is stationary and ergodic on the probability space $(X,\mathcal{B},\mu)$ and Theorem 5.2.4 is applicable to it. If it is ergodic for all $f \in L_1(X,\mu)$, then μ is also ergodic for T.

Theorem 5.2.5. *Suppose that $\{\xi_n,\ n=0,1,2,\dots\}$ is a stationary sequence in the measurable space $(X,\mathcal{B})$. It is ergodic if and only if*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\big(\mathbf{P}\{\xi_0\in A_0,\ \xi_1\in A_1,\dots,\xi_m\in A_m,\ \xi_k\in A_0,\\ \xi_{k+1}\in A_1,\dots,\xi_{k+m}\in A_m\}-(\mathbf{P}\{\xi_0\in A_0,\dots,\xi_m\in A_m\})^2\big)=0 \tag{5.2.9}$$

for any m and $A_0,\dots,A_m\in\mathcal{B}$.

Proof. Let $\{\eta_k\}$ be a bounded scalar stationary sequence. For $n^{-1}\sum_{k=1}^{n}\eta_k$ to converge to a constant with probability 1 (we know that the limit exists), it is necessary and sufficient that $\mathbf{V}(n^{-1}\sum_{k=1}^{n}\eta_k)\to 0$. Using the stationarity ($\mathbf{E}\eta_k\eta_j=\mathbf{E}\eta_0\eta_{j-k}$ for $j>k$), we find that

$$\mathbf{V}\left(\frac{1}{n}\sum_{k=1}^{n}\eta_k\right)=\frac{1}{n}[\mathbf{E}\eta_0^2-(\mathbf{E}\eta_0)^2]+\frac{2}{n}\sum_{k=1}^{n-1}(\mathbf{E}\eta_0\eta_k-(\mathbf{E}\eta_0)^2)\\ -\frac{2}{n^2}\sum_{k=1}^{n-1}k(\mathbf{E}\eta_0\eta_k-(\mathbf{E}\eta_0)^2)\ .$$

The condition

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}k(\mathbf{E}\eta_0\eta_k-(\mathbf{E}\eta_0)^2)=0$$

is necessary and sufficient for $\mathbf{V}(n^{-1}\sum_{k=1}^{n}\eta_k)\to 0$. Taking $\eta_n=I_{A_0}(\xi_k)\dots I_{A_m}(\xi_m)$, we conclude that the condition (5.2.9) is necessary.

Now let (5.2.9) hold and let $(Y,\mathcal{C},\mu)$ be the space X^∞ with cylinder σ-algebra and measure generated by the stationary sequence. Denote by $L_1^0(Y,\mu)$ the subset of functions $f(y)\in L_1(Y,\mu)$ of the form $f(y)=f_m(x_0,\dots,x_m)$, $y=(x_0,x_1,\dots,)$. It is dense in $L_1(Y,\mu)$. Furthermore, let F_0 be the set of functions in the linear span of the indicator functions of the cylinder sets. For $f\in F_0$ it follows from (5.2.9) that

$$\lim_{n\to\infty}\int\left|\frac{1}{n}\sum_{k=1}^{n}f(T^ky)-\int f(y)\mu(dy)\right|\mu(dy)=0\ . \tag{5.2.10}$$

Since F_0 is dense in $L_1^0(Y,\mu)$, (5.2.10) holds for all $f\in L_1(Y,\mu)$. Hence it follows that every $\mathcal{T}$-measurable function is a constant μ-almost everywhere. □

5.3 Central Limit Theorem and Invariance Principle

The *central limit theorem* says that the distribution of a sum of independent random variables approaches the normal distribution as the number of variables is increased indefinitely. This statement has to be made more precise. If the sums of independent random variables are bounded in probability, then the convergence of the distributions entails the convergence of the series. The limiting distribution equals the distribution of the sum of the series. It will be normal only if the terms have normal distributions. Therefore the problem is posed differently. Let $\{\xi_n\}$ be a sequence of independent scalar random variables and let $\xi_n = \xi_1 + \ldots + \xi_n$. Of interest is the case where the ξ_n are unbounded in probability. We shall study conditions under which there exist constants a_n and b_n such that the distribution of $b_n^{-1}(\xi_n - a_n)$ converges weakly to the normal distribution.

5.3.1 Identically Distributed Terms

Let the ξ_k have identical distributions with $\mathbf{E}\xi_k = a$ and $\mathbf{V}\xi_k = b < \infty$. Then ξ_n has expectation na and variance nb. It is reasonable to take $a_n = na$ and $b_n = \sqrt{nb}$ so that $\eta_n = (nb)^{-1/2}(\xi_n - na)$ has expectation 0 and variance 1.

Theorem 5.3.1. *As $n \to \infty$, the distribution of η_n converges to the normal distribution with mean 0 and variance 1.*

Proof. Denote the characteristic function of η_n by $f_n(z)$. Then

$$f_n(z) = \mathbf{E}\exp\{iz\eta_n\} = \exp\left\{in\frac{az}{\sqrt{ab}}\right\}\left[\varphi\left(z(nb)^{-1/2}\right)\right]^n ,$$

where $\varphi(z) = \mathbf{E}e^{iz\xi_1}$. By Theorem 5.1.5 of Sect. 5.1.3, it suffices to prove that $f_n(z) \to e^{-z^2/2}$ (this is the characteristic function of a normal distribution with mean 0 and variance 1). The existence of $\mathbf{V}\xi_1$ implies that $\varphi(z)$ is twice continuously differentiable and

$$e^{-iaz}\varphi(z) = 1 - \frac{b}{2}z^2\alpha(z) ,$$

where $\alpha(z) \to 1$ as $z \to 0$. Therefore

$$f_n(z) = \left(1 - \frac{1}{2n}z^2\alpha\left(\frac{z}{\sqrt{nb}}\right)\right)^n \to e^{-\frac{z^2}{2}} .$$

□

Corollary. *Suppose that the ξ_k's assume the value 1 with probability p and 0 with probability $1-p$, $0 < p < 1$. Then for all x*

$$\lim_{n\to\infty} P\left\{\sum_{k=1}^{n}\xi_k < \sqrt{np(1-p)}x + np\right\} = \Phi(x), \tag{5.3.1}$$

where $\Phi(x) = (2\pi)^{-\frac{1}{2}}\int_{-\infty}^{x} e^{-\frac{1}{2}u^2}du$.

We next make use of Remark 5.1.2, Theorem 5.1.1 of Sect. 5.1.1. Consider a sequence of independent trials in each of which an event may occur with probability p. By defining $\xi_k = I_{A_k}$, where A_k is the event in question in the k-th trial, we have a sequence of variables for which (5.3.1) holds. Furthermore, $n^{-1}\sum_{k=1}^{n}\xi_k = \nu_n$ is the relative frequency of occurrence of the event in the n trials. From (5.3.1) we find for $a < b$ that

$$\lim_{n\to\infty} \mathbf{P}\left\{a\sqrt{\frac{p(1-p)}{n}} \le \nu_n - p < b\sqrt{\frac{p(1-p)}{n}}\right\} = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-u^2/2}du\,. \tag{5.3.2}$$

5.3.2 Lindeberg's Theorem

Let $\{\xi_n\}$ be a sequence of independent random variables with $\mathbf{E}\xi_k = a_k$ and and $\mathbf{V}\xi_k = b < \infty$. Define

$$\eta_n = \left(\sum_{k=1}^{n}\xi_k - \sum_{k=1}^{n}a_k\right)\left(\sum_{k=1}^{n}b_k\right)^{-1/2}\,.$$

We are interested in conditions under which the distribution of η_n will converge to the normal distribution with mean 0 and variance 1. One such condition was found by Lindeberg and it bears his name. Write $c_n = \sum_{k=1}^{n} b_n$.

Lindeberg's Condition. For any $\varepsilon > 0$

$$\lim_{n\to\infty} c_n^{-1}\sum_{k=1}^{n}\mathbf{E}(\xi_k - a_k)^2 I_{\{|\xi_k - a_k| > \varepsilon\sqrt{c_n}\}} = 0. \tag{5.3.3}$$

Theorem 5.3.2. *If Lindeberg's condition holds, then*

$$\lim_{n\to\infty}\mathbf{P}\{\eta_n < x\} = \Phi(x) \tag{5.3.4}$$

for all x.

Proof. With no loss of generality, we may assume that $a_k = 0$. Write $\varphi_k(z) = \mathbf{E}e^{iz\xi_k}$ and $f_n(z) = \mathbf{E}e^{iz\eta_n}$. Then

$$f_n(z) = \prod_{k=1}^{n}\varphi_k(c_n^{-1/2}z).$$

Let

$$\alpha_{nk}(\varepsilon) = c_n^{-1}\mathbf{E}\xi_k^2 I_{\{|\xi_k| > \varepsilon\sqrt{c_n}\}}. \tag{5.3.5}$$

By condition (5.3.3), $\sum_{k=1}^{n}\alpha_{nk}(\varepsilon) \to 0$ for every $\varepsilon > 0$. Since $b_k/c_n \le \varepsilon + \alpha_{nk}(\varepsilon)$, it follows that

$$\lim_{n\to\infty} \max_{1\le k\le n} \frac{b_k}{c_n} = 0$$

and hence $c_n \to \infty$.

Condition (5.3.3) implies that

$$\varphi_k(z) = 1 - \frac{z^2}{2} b_k + \beta_k(z),$$

where $\lim_{n\to\infty} \sum_{k=1}^n |\beta_k(z c_n^{-\frac{1}{2}})| = 0$. But

$$\left| \prod_{k=1}^n u_k - \prod_{k=1}^n v_k \right| \le \sum_{k=1}^n |u_k - v_k|$$

for $|u_k| \le 1$ and $|v_k| \le 1$. Thus

$$\lim_{n\to\infty} \left| \prod_{k=1}^n \varphi_k(z c_n^{-1/2}) - \prod_{k=1}^n \left(1 - \frac{b_k z^2}{2c_n}\right) \right| \le \lim \sum_{k=1}^n |\beta_{nk}(z)| = 0.$$

It remains to observe that

$$\lim_{n\to\infty} \sum_{k=1}^n \left(1 - \frac{b_k}{2c_n} z^2\right) = \exp\left\{-\frac{z^2}{2}\right\}.$$

□

Remark. Lindeberg's condition assures that the contribution of each separate term in the overall sum approaches zero in probability:

$$\begin{aligned} \mathbf{P}\left\{ \max_{1\le k\le n} \frac{|\xi_k - a_k|}{\sqrt{c_n}} > \varepsilon \right\} &\le \sum_{k=1}^n \mathbf{P}\{|\xi_k - a_k| > \varepsilon\sqrt{c_n}\} \\ &\le \frac{1}{\varepsilon^2 c_n} \sum_{k=1}^n \mathbf{E}|\xi_k - a_k|^2 I_{\{|\xi_k - a_k| > \varepsilon\sqrt{c_n}\}} \to 0. \end{aligned}$$

A consequence of Lindeberg's theorem is Lyapunov's theorem (it was proved earlier and is more useful in applications).

Theorem 5.3.3. *Suppose that*

$$\lim_{n\to\infty} c^{-1-\frac{\alpha}{2}} \sum_{k=1}^n \mathbf{E}|\xi_k - a_k|^{2+\alpha} = 0 \tag{5.3.6}$$

for some positive α. *Then* (5.3.4) *holds.*

Proof. Again assume that $a_k = 0$. If $\alpha_{nk}(\varepsilon)$ is given by (5.3.5), then

$$\alpha_{nk}(\varepsilon) \le \varepsilon^{-\alpha} c^{-1-\frac{\alpha}{2}} \mathbf{E}|\xi_k|^{2+\alpha}.$$

Therefore (5.3.3) results from (5.3.6). □

5.3.3 Donsker-Prokhorov Theorem

(a) *Random polygonal paths.* Let $\{\xi_k\}$ be a sequence of independent random variables with $\mathbf{E}\xi_k = 0$ and $\mathbf{V}\xi_k = b_k$, and let Lindeberg's condition be satisfied. Put

$$t_{nk} = \frac{1}{c_n}\sum_{i=1}^{k} b_k, \quad t_{n_0} = 0, \quad \zeta_{nk} = c_n^{-1/2}\sum_{i=1}^{k}\xi_i, \quad \zeta_{n_0} = 0\ .$$

Denote by $\zeta_n(t)$ the random polygonal line with vertices at (t_{nk}, ζ_{nk}):

$$\zeta_n(t) = \zeta_{nk}\frac{t_{nk+1} - t}{t_{nk+1} - t_{nk}} + \zeta_{nk+1}\frac{t - t_{nk}}{t_{nk+1} - t_{nk}}, \quad t_{nk} \le t \le t_{nk+1}\ .$$

$\zeta_n(t)$ is a continuous random process. Associated with it is a probability measure on $\mathbf{C}_{[0,1]}$, the space of scalar continuous functions. Denote it by μ_n. Let μ_w be the measure on $\mathbf{C}_{[0,1]}$ corresponding to the standard Wiener process $w(t)$ (or more precisely, its continuous modification) (see pp. 84–85).

Theorem 5.3.4. (Yu.V. Prokhorov) $\mu_n \Longrightarrow \mu_w$ *in* $\mathbf{C}_{[0,1]}$.

A special but most important case in practical applications is that of identically distributed ξ_k considered by Donsker.

(b) *Invariance principle.* Assume that the ξ_k's are identically distributed with $\mathbf{V}\xi_k = 1$. Then $c_n = n$, Lindeberg's condition holds automatically, $t_{nk} = k/n$, $\zeta_{nk} = n^{-\frac{1}{2}}\sum_{i=1}^{k}\xi_i$ and $\zeta_{n0} = 0$. Various functionals of the sequential sums may be expressed as functions of $\zeta_n(t)$. Thus

$$\max_{k\le n}\zeta_k = \sqrt{n}\max_t \zeta_n(t),$$
$$\max_{k\le n}|\zeta_k| = \sqrt{n}\max_t |\zeta_n(t)|,$$
$$\sum_{k=1}^{n}|\zeta_k|^\alpha \sim n^{1+\frac{\alpha}{2}}\int_0^1 |\zeta_n(t)|^\alpha dt.$$

The *invariance principle* says that under suitable normalization functionals of this kind do not depend on the distributions of the variables ξ_k. Therefore by selecting the distribution of ξ_k (usually, $\xi_k = \pm 1$ with probabilities $1/2$), the limiting distribution of the functional is determined. The distribution turns out to coincide with the distribution of the same functional of the Wiener process. The next result serves as an example of this kind of theorem.

Theorem 5.3.5.

$$\lim_{n\to\infty}\mathbf{P}\{\max_{k\le n}\zeta_k \le x \le \sqrt{n}\} = \sqrt{\frac{2}{\pi}}\int_0^x e^{-u^2/2}du.$$

The right-hand side coincides with the distribution of $\sup_{0\le t\le 1} w(t)$.

(c) *Weak compactness of a family of probability measures in* $\mathbf{C}_{[0,1]}$. We shall make use of Theorem 5.1.4 of Sect. 5.1.2. To derive effective conditions for weak compactness, we must be able to give an efficacious description of a sufficiently broad class of compact sets in $\mathbf{C}_{[0,1]}$.

Let $\delta(t)$ be defined for $t \geq 0$ with $\delta(t) > 0$ for $t \neq 0$ and $\delta(t) \downarrow 0$ as $t \downarrow 0$. Let $K(\delta, c)$ be the subset of functions $x(t)$ of $\mathbf{C}_{[0,1]}$ such that $|x(t)| \leq c$ and $|x(t_1) - x(t_2)| \leq \delta(|t_1 - t_2|)$ for all t, t_1 and t_2.

Lemma 5.3.1. *$K(\delta, c)$ is a compact set in $\mathbf{C}_{[0,1]}$ and to every compact K in $\mathbf{C}_{[0,1]}$ there exist a $\delta(\cdot)$ and c such that $K \subset K(\delta, c)$.*

Proof. The first statement follows by Arzelà's theorem. If K is a compact set, take $\delta(t) = \sup\{|x(t_1) - x(t_2)|, |t_1 - t_2| \leq t, x(\cdot) \in K\}$. Then by Dini's theorem, $\delta(t) \downarrow 0$ as $t \downarrow 0$. If $c = \sup\{|x(t)|, t \in [0,1], x(\cdot) \in K\}$, then $K \subset K(\delta, c)$. □

Theorem 5.1.4 of Sect. 5.1.2 and the lemma yield the following result.

Theorem 5.3.6. *Let $\{\xi_n(t)\}$ be a sequence of continuous numerical processes defined on $[0,1]$. The sequence of measures $\{\nu_n\}$ on $\mathbf{C}_{[0,1]}$ corresponding to these processes is weakly compact if:*

1. *$\xi_n(0)$ is bounded in probability.*
2. *for every $\varepsilon > 0$*

$$\lim_{h \to 0} \overline{\lim_{n \to \infty}} \mathbf{P}\left\{ \sup_{|t_1 - t_2| \leq h} |\xi_n(t_1) - \xi_n(t_2)| > \varepsilon \right\} = 0 \,. \tag{5.3.7}$$

Proof. From (5.3.7) it follows that

$$\lim_{h \to 0} \sup_n \mathbf{P}\left\{ \sup_{|t_1 - t_2| \leq h} |\xi_n(t_1) - \xi_n(t_2)| > \rho \right\} = 0$$

for all positive ρ. Choose sequences $h_k \to 0$ and $\rho_k \to 0$ so that

$$\sum_k \sup_n \mathbf{P}\left\{ \sup_{|t_1 - t_2| \leq h_k} |\xi_n(t_1) - \xi_n(t_2)| > \rho_k \right\} \leq \varepsilon/2 \,.$$

If $\delta(t) = \rho_k$ for $h_{k+1} \leq t < h_k$ and $\delta(t) = (1 + h_1)\rho_1/h_1$ for $t \geq h_1$, then for all n

$$\mathbf{P}\left\{ \sup_{|t_1 - t_2| \leq s} |\xi_n(t_1) - \xi_n(t_2)| \leq \delta(s), 0 \leq s \leq 1 \right\} \geq 1 - \frac{\varepsilon}{2}.$$

If now $\mathbf{P}\{|\xi_n(0)| > c_1\} \leq \varepsilon/2$, then

$$\mathbf{P}\Bigg\{ \sup_t |\xi_n(t)| \leq c_1 + \frac{1 + h_1}{h_1}\rho_1, \\ \sup_{|t_1 - t_2| \leq s} |\xi_n(t_1) - \xi_n(t_2)| \leq \delta(s), 0 \leq s \leq 1 \Bigg\} \geq 1 - \varepsilon \,.$$

Thus $\mu_n(K(\delta, c)) \geq 1 - \varepsilon$ if $c = c_1 + (1 + h_1)\rho_1/h_1$. □

(d) *Proof of Prokhorov's theorem.* Let us verify that the processes $\zeta_n(t)$ satisfy the hypotheses of Theorem 5.3.6. Since $\zeta_n(0) = 0$, it is only necessary to check part 2.

The proof of Theorem 5.3.2 shows that $\delta_n = \max_k(t_{nk} - t_{nk-1}) = \max_k(b_k/c_k) \to 0$. Since

$$\sup_{|t_1-t_2|\le h} |\zeta_n(t_1) - \zeta_n(t_2)| \le \sup_{|t_{nk}-t_{nj}|\le h+2\delta_n} |\zeta_n(t_{nk}) - \zeta_n(t_{nj})| \,,$$

it suffices to prove that

$$\lim_{h\to 0} \overline{\lim_{n\to\infty}} \mathbf{P}\left\{ \sup_{|t_{nk}-t_{nj}|\le h} |\zeta_n(t_{nk}) - \zeta_n(t_{nj})| > \varepsilon \right\} = 0 \,.$$

Let $s_n(t) = \inf\{t_{nk} : t_{nk} \ge t\}$. Then $|s_n(t) - t| \le \delta_n$ and

$$\sup_{|t_{nk}-t_{nj}|\le h} |\zeta_n(t_{nk}) - \zeta_n(t_{nj})| \le 2 \sup_{k<h^{-1}} \eta_n(k,h),$$

where $\eta_n(k,h) = \sup\{|\zeta_n(t_{nj}) - \zeta_n(s_n(kh))| : t_{nj} \in [s_n(kh), s_n(kh) + 2h]\}$. To prove that the μ_n's are weakly compact, it suffices to show that

$$\lim_{h\to 0} \overline{\lim_{n\to\infty}} \mathbf{P}\left\{ \sup_{k<h^{-1}} \eta_n(k,h) > \varepsilon \right\} = 0 \,,$$

for every positive ε, or that

$$\lim_{h\to 0} \overline{\lim_{n\to\infty}} \sum_{k<h^{-1}} \mathbf{P}\{\eta_n(k,h) > \varepsilon\} = 0 \,.$$

Since

$$\mathbf{P}\{|\zeta_n(t_{nk}) - \zeta_n(t_{nj})| > \rho\} \le \frac{1}{\rho^2}\mathbf{E}(\zeta_n(t_{nk}) - \zeta_n(t_{nj}))^2 \le \frac{|t_{nk} - t_{nj}|}{\rho^2} \,,$$

it follows that

$$\mathbf{P}\left\{|\zeta_n(t_{nj}) - \zeta_n(t_{nj})| > \frac{\varepsilon}{2}\right\} \le \frac{1}{2}$$

for $16h < \varepsilon^2$ providing $|t_{nk} - t_{nj}| \le 2h$. Hence, on the basis of Sect. 3.2 of Chapter 3,

$$\mathbf{P}\{\eta_n(k,x) > \varepsilon\} \le 2\mathbf{P}\left\{|\zeta_n((s_n(kh) + 2h) \wedge 1) - \zeta_n(s_n(kh))| > \frac{\varepsilon}{2}\right\}.$$

Using the independence of $\zeta_n(t_{nk_1}), \zeta_n(t_{nk_2}) - \zeta_n(t_{nk_1}), \ldots, \zeta_n(t_{nk_l}) - \zeta_n(t_{nk_{l-1}})$ for $t_{nk_1} < t_{nk_2} < \ldots < t_{nk_l}$, Lindeberg's theorem and

$$\sup_t |\zeta_n(t) - \zeta_n(s_n(t))| \le \max_k \frac{|\xi_k|}{c_n} \to 0$$

in probability, one can demonstrate that the following statement is true.

Lemma 5.3.2. 1. *The distribution of*

$$\zeta_n((s_n(kh)+2h)\wedge 1)-\zeta_n(s_n(kh))$$

converges to the distribution of $w((k+2)h\wedge 1)-w(kh)$.

2. *The joint distribution of* $\zeta_n(t_1),\zeta_n(t_2),\dots,\zeta_n(t_l)$ *with* $t_i\in[0,1]$ *converges to the joint distribution of* $w(t_1),w(t_2),\dots,w(t_l)$.

Proof. Applying part 1, we find that

$$\begin{aligned}\overline{\lim_{n\to\infty}}\sum_{k<h^{-1}}\mathbf{P}\{\eta_n(k,h)>\varepsilon\} &\le 2\sum_{k<h^{-1}}\mathbf{P}\left\{|w((k+2)h\wedge 1)-w(kh)|>\frac{\varepsilon}{2}\right\}\\ &\le 2\sum_{k<h^{-1}}\frac{16}{\varepsilon^2}\mathbf{E}|w((k+2)h\wedge 1)-w(kh)|^4\\ &\le \frac{32}{\varepsilon^2}h^{-1}(3(2h)^2)=O(h)\ .\end{aligned}$$

This expression approaches zero. Hence, the set $\{\mu_n\}$ is compact. By statement 2,

$$\int f d\mu_n\to\int f\mu_w$$

for any function $f(x(\cdot))$ of the form

$$f(x(\cdot))=\Phi(x(t_1),\dots,x(t_l))\ ,$$

where $l,t_1,\dots,t_l\in[0,1]$ and $\Phi\in C_R^l$ are arbitrary. The collection of f's forms a complete set in $\mathbf{C}_{[0,1]}$. Therefore $\mu_n\Longrightarrow\mu_w$. □

Historic and Bibliographic Comments

The philosphical foundations of probability theory are discussed in Borel (1963). It contains a popular account of the author's views on the essence of probability theory and its possible applications (translation of a book published in French in 1956).

The now-accepted axiomatic treatment of probability theory was first proposed in an article by Kolmogorov in 1933 (see Kolmogorov (1974)).

A presentation of the basic concepts of probability may be found in Borovkov (1976), Feller (1968, 1971), Halmos (1956), Loèv (1963), Neveu (1965), and Prokhorov (1973). The first volume of Feller (1968) involves elementary probability theory with very many interesting problems that are helpful in understanding the sources of concepts and methods. The second volume (1971) gives an account of the main results on limit theorems, random walks, Markov processes and stationary processes. Loèv (1963) contains a fundamental presentation of probability theory, independent and dependent random variables as well as elements of the theory of stochastic processes. Attention is given in Neveu (1965) mainly to rigorous definitions of the basic notions of probability. It gives a fairly complete development of ergodic theory. Shiryaev (1980) gives a contemporary treatment of the basic notions of probability theory, limit theorems and ergodic theory, stationary sequences, martingales and Markov chains. Borovkov (1976) presents basic information about probability theory (limit theorems, random walks, Markov chains and the simplest stochastic processes). Prokhorov (1973) is more of a handbook and pays attention mainly to limit theorems and stochastic processes.

Dellachérie (1972), Doob (1953), Gikhman (1977), Meyer (1966), Rozanov (1963) and Skorokhod (1986) consider various aspects of the theory of stochastic processes. The first most complete presentation of stochastic processes is found in Doob (1953). Its treatment of ergodic theory for Markov chains should especially be noted. Gikhman (1977) gives the foundations of stochastic processes including an axiomatic treatment, conditional probabilities, independence, general theory of random functions, information about the basic classes of stochastic processes and limit theorems. Meyer (1966) and the

closely-related Doob (1953) are devoted to studying currents of σ-algebras on probability spaces and the related stopping times, measurability of stochastic processes, martingales and supermartingales. Rozanov (1967) presents the spectral theory of wide-sense stationary processes with applications to extrapolation and filtering, as well as the ergodic theory of strict-sense stationary processes. Skorokhod (1986) gives basic facts about processes with independent increments (including those with discrete time, that is, random walks).

The main subject of Ibragimov (1965) is the extension of limit theorems for independent random variables to dependent variables.

Halmos (1956) contains the basic notions and theorems of ergodic theory in a form that is accessible to a wide range of mathematicians.

Finally, Skorokhod (1975) studies ways of constructing measures, transformations of measures by smooth mappings, conditions for absolute continuity and quasi-invariant measures.

References

Borel, E. (1963) Probability and Certainty. Walker, New York

Borovkov, A.A. (1976) Probability Theory, Nauka, Moscow. German transl.: Basel-Birkhäuser, 1976

Dellachérie, C. (1972) Capacités et processus stochastiques. Springer-Verlag, New York

Doob, J.L. (1953) Stochastic Processes. John Wiley, New York

Feller, W. (1968) An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd ed., John Wiley, New York
(1971) An Introduction to Probability Theory and Its Applications, Vol. 2, 2nd. ed., John Wiley, New York

Gikhman, I.I. and Skorokhod, A.V. (1977) Introduction to the Theory of Random Processes, Rev. and Suppl., 2nd. ed. Nauka, Moscow. English transl.: 1st ed., W.B. Saunders, Philadelphia, 1969

Halmos, P. (1956) Lectures on Ergodic Theory, Mathematics Society of Japan, Tokyo

Ibragimov, I.A. and Linnik, Yu.V. (1965) Independent and Stationary Sequences of Random Variables, Nauka, Moscow. English transl.: Wolters-Noordhoff, Groningen, 1971

Kolmogorov, A.N. (1974) Foundations of the Theory of Probability. Nauka, Moscow. English transl. of 1933 German ed.: Chelsea Publishing CO., New York, 2nd ed., 1956

Loev, M. (1963) Probability Theory, 3rd,ed.. Van Nostrand, Princeton

Meyer, P.A. (1966) Probability and Potentials. Blaisdell, Walthem (MA)

Neveu, P.A. (1965) Mathematical Foundations of the Calculus of Probabilities. Holden-Day, San Francisco

Prokhorov, Yu.V. and Rozanov, Yu.A. (1973) Probability Theory. Nauka, Moscow (Russian)

Rozanov, Yu.A. (1963) Stationary Random Processes. Fizmatgiz, Moscow. English transl.: Holden-Day, San Francisco

Shiryaev, A.N. (1980) Probability, Nauka, Moscow. English transl.: Springer-Verlag, New York etc., 1985

Skorokhod, A.V. (1975) Integration in Hilbert Space. Nauka, Moscow. English transl.-earlier ed.: Springer-Verlag, New York, 1974
(1986) Stochastic Processes with Independent Increments. Nauka, Moscow. English transi. earlier ed.: U.S. Government Publ., 1966

Markov Processes and Probability Applications in Analysis

Contents

Introduction

In 1931, Kolmogorov published his article "On analytic methods in probability theory". In that paper, he introduced a class of stochastic processes which have since been called Markov processes. To study their probabilistic parameters (namely, the transition probabilities), he proposed using differential equations. For processes with a finite or denumerable phase space, the transition probabilities satisfy finite or denumerable systems of ordinary differential equations and for processes with a finite-dimensional phase space, second-order parabolic partial differential equations. The creation of this new powerful method made it possible to solve other problems involving stochastic processes and to deduce new kinds of limit theorems (for instance, diffusion problems for random walks considered by Kolmogorov, Petrovsky and Khinchin). Although certain analytic techniques had been used to prove limit theorems (the Fourier transform), it was precisely due to Kolmogorov's papers that analysis came to be applied universally in probability theory. Thus (at least formally), probability theory could now be viewed as a section of analysis.

The connection that Kolmogorov established between differential equations and Markov processes made possible their "reverse" application to differential equations. To do this, it was necessary to have ways of studying Markov processes that were independent of analysis. Such purely probabilistic methods were initially developed for the Wiener process. The formation of Wiener measures led to the solution of numerous problems in analysis. The most siginficant step here was Kac's representation of the solution to the Cauchy problem for the equation

$$\frac{\partial U}{\partial t} = \Delta U + vU$$

as an integral with respect to a Wiener measure. It was precisely this result that demonstrated how probability-theoretic techniques could help to overcome essential difficulties that are encountered in analysis.

The theory of stochastic differential equations developed by Itô and Gikhman has led to the construction of a broad class of Markov processes by probability-theoretic methods which can then be used to solve partial differential equations.

1

Markov Processes

Markov processes describe the evolution of systems undergoing independent random perturbations at different moments of time. In this case, the state of a system at a given moment of time completely determines the probability parameters of the process that describes the further evolution of the system. This means that the past behavior of a system has no effect on its future behavior except for its present state. Discrete sequences of events possessing this property were introduced by Markov and have come to be called Markov chains. Processes with continuous time were introduced by Kolmogorov and he referred to them as stochastically determinate processes. Subsequently they came to be called Markov processes (by analogy with discrete-time processes).

1.1 Definition and General Properties

1.1.1 Definition of a Markov Process

Let $(X, \mathcal{B})$ be a measurable space. It will play the role of the *phase* or *state* space of a process being considered. We shall examine processes defined on R_+. To define a Markov process, it is natural to consider a whole family of processes $\xi_{s,x}(t)$, $t \geq s$, rather than a single process; $\xi_{s,x}(t)$ describes the evolution of a system which is in the state x at the initial time s. Let $(\Omega, \mathcal{F})$ be a sample space with a σ-algebra $\mathcal{F}$. It is convenient to assume that each $\xi_{s,x}(t)$ has its own probability measure $\mathbf{P}_{s,x}(\cdot)$ on $(\Omega, \mathcal{F})$. Thus for a Markov process there will be a family of probability spaces $\{(\Omega, \mathcal{F}, \mathbf{P}_{s,x}),\ s \in R_+,\ x \in X\}$ whose measures $\mathbf{P}_{s,x}$ are consistent in a certain way. Let $\mathcal{F}_t^s$, $s < t$, be the σ-algebra of events observed on $[s, t]$ (in particular, $\xi_{s,x}(u)$ with $u \in [s, t]$, is measurable relative to $\mathcal{F}_t^s$). From the above intuitive definition, it follows for $s < u$ that the conditional distribution of $\xi_{s,x}(t)$ on $[t, \infty)$, given $\xi_{s,x}(v)$ on $[s, u]$, must be the same as the distribution of the process that begins at time u at the point $\xi_{s,x}(u)$. This property may be stated precisely as follows. With probability $\mathbf{P}_{s,x} = 1$,

$$\begin{aligned}&\mathbf{P}_{s,x}\{\xi(t_1)\in A_1,\dots,\xi(t_n)\in A_n|\mathcal{F}_n^s\}\\&\quad=\mathbf{P}_{u,\xi(u)}\{\xi(t_1)\in A_1,\dots,\xi(t_n)\in A_n\}\,,\\&0\le s<u\le t_1<\ldots<t_n\,,\quad A_k\in\mathcal{B}\end{aligned}\tag{1.1.1}$$

($\xi(\cdot)$ in this denotes a process whose distributions are specified by the measure operating on $\xi(\cdot):\mathbf{P}_{u,\xi(u)}(C)=\mathbf{P}_{u,x}(C)|_{x=\xi(u)}$). Observe that it is precisely having different measures that permits one to denote the process by the exact same symbol. Since we are going to consider processes later on that begin at different moments of time, we shall utilize the notation $\xi_s(t)$ for a process defined for $t\ge s$.

Then (1.1.1) can be rewritten as follows:

$$\begin{aligned}&\mathbf{P}_{s,x}\{\xi_s(t_1)\in A_1,\dots,\xi_s(t_n)\in A_n\}|\mathcal{F}_u^s\}\\&\quad=\mathbf{P}_{u,\xi_s(u)}\{\xi_u(t_1)\in A_1,\dots,\xi_u(t_n)\in A_n\}.\end{aligned}\tag{1.1.2}$$

Thus a *Markov process* is defined by

1. a phase space – a measurable space $(X,\mathcal{B})$,
2. a measurable space $(\Omega,\mathcal{F})$ and family of σ-algebras $\mathcal{F}_t^s\subset\mathcal{F}$, $0\le s<t<\infty$, such that $\mathcal{F}_t^s\subset\mathcal{F}_v^u$ when $u\le s<t\le v$,
3. a family of functions $\xi_s(t)=\xi_s(t,\omega)$, $s\in R_+$, $t>s$, with $\xi_s(t,\omega)$ an $\mathcal{F}_t^s$-measurable mapping from Ω to X for all $0\le s<t$, and
4. a family of probability measures $\mathbf{P}_{s,x}(\cdot)$ on $\mathcal{F}$ satisfying
   i. $\mathbf{P}_{s,x}\{\xi_s(s)=x\}=1$ for all $s\in R_+$ and $x\in X$,
   ii. relation (1.1.2) with probability $\mathbf{P}_{s,x}=1$ for all n, $0\le s<u\le t_1<\ldots<t_n$ and $A_1,\dots,A_n\in\mathcal{B}$,
   iii. the function $\mathbf{P}_{s,x}\{\xi_s(t)\in A\}$ is measurable in x for $s<t$ and $A\in\mathcal{B}$.

The expectation with respect to $\mathbf{P}_{s,x}$ will be denoted by $\mathbf{E}_{s,x}$. Relation (1.1.2) may be rewritten as follows:

$$\mathbf{E}_{s,x}\left(\prod_{k=1}^n I_{A_k}(\xi_s(t_k))\big|\mathcal{F}_u^s\right)=\mathbf{E}_{u,\xi_s(u)}\prod_{k=1}^n I_{A_k}(\xi_s(t_k))\,.\tag{1.1.3}$$

In this $s<u\le t_1<\ldots<t_n$ and $A_k\in\mathcal{B}$, $k=1,\dots,n$. The relation holds with probability $\mathbf{P}_{s,x}=1$.

1.1.2 Transition Probability

The function $P(s,x,t,A)=\mathbf{P}_{s,x}\{\xi_s(t)\in A\}$ is called the *transition probability*. It has these obvious properties.

I. $P(s,x,t,A)$ is defined for $0\le s\le t$, $x\in X$, and is measurable in x.
II. $P(s,x,t,A)$ is a probability measure with respect to $A\in\mathcal{B}$.

The next property establishes the Markov nature of the process.

III. $P(s,x,t,A)$ satisfies the equation

$$P(s,x,t,A) = \int P(s,x,u,dz)P(u,z,t,A) \tag{1.1.4}$$

for $0 \le s < u < t$, and $A \in \mathcal{B}$ (the *Chapman-Kolmogorov equation*).

For, on making use of part ii of property 4, we find that

$$\mathbf{P}_{s,x}\{\xi_s(t) \in A|\mathcal{F}_u^s\} = \mathbf{P}_{u,\xi_s(u)}(\xi_u(t) \in A\} = P(u,\xi_s(u),t,A) \ .$$

Taking the expectation of both sides of this equation with respect to $\mathbf{P}_{s,x}$, we obtain

$$\begin{aligned}\int \mathbf{P}_{s,x}\{\xi_s(u) \in dz\}P(u,z,t,A) &= \mathbf{E}_{s,x}\mathbf{P}_{s,x}\{\xi_s(t) \in A|\mathcal{F}_u^s\} \\ &= \mathbf{P}_{s,x}\{\xi_s(t) \in A\} \ .\end{aligned}$$

This proves (1.1.4).

Using (1.1.3), one can derive the following formula for the finite-dimensional distributions of $\xi_s(t)$ in terms of $\mathbf{P}_{s,x}$: For $s < t_1 < \ldots < t_n$,

$$\begin{aligned}&\mathbf{P}_{s,x}\{\xi_s(t_1) \in A_1, \ldots, \xi_s(t_n) \in A_n\} \\ &\quad = \mathbf{E}_{s,x}I_{A_1}(\xi_s(t_1))\mathbf{E}_{s,x}\left(\prod_{k=2}^{n} I_{A_k}(\xi_s(t_k))\big|\mathcal{F}_t^s\right) \\ &\quad = \mathbf{E}_{s,x}I_{A_1}(\xi_s(t_1))\mathbf{E}_{t_1,\xi_s(t_1)}\prod_{k=2}^{n} I_{A_k}\{\xi_{t_1}(t_k)) \\ &\quad = \int_{A_1} P(s,x,t_1,dx_1)\mathbf{E}_{t_1,x_1}\prod_{k=2}^{n} I_{A_k}(\xi_{t_1}(t_k)) \ .\end{aligned}$$

By induction, this yields

$$\begin{aligned}&\mathbf{P}_{s,x}\{\xi_s(t_1) \in A_1, \ldots, \xi_s(t_n) \in A_n\} \\ &\quad = \int_{A_1} P(s,x,t_1,dx_1)\int_{A_2} P(t_1,x_1,t_2,dx_2) \\ &\qquad \ldots \int_{A_n} P(t_{n-1},x_{n-1},t_n,dx_n) \ .\end{aligned} \tag{1.1.5}$$

Thus, the transition probability gives the values of $\mathbf{P}_{s,x}$ on the σ-algebras $\mathcal{C}_u^s$ determined by the events $\{\xi_s(t) \in A\}$, $t \in [s,u]$ and $A \in \mathcal{B}$. This shows that the transition probability permits one to construct a Markov process on a specific $(\Omega,\mathcal{B})$; Ω is taken to be the space X^{R_+} of all X-valued functions $\omega = \omega(t)$ on R_+, $\mathcal{F}$ to be a cylinder σ-algebra in X^{R_+} and $\xi_s(\omega) = \omega(t)$ for $t \ge s$. We regard $\xi_s(\omega) \in C$, with $C \in \mathcal{F}$, if there exists an $\bar{\omega}$ such that $\omega(t) = \bar{\omega}(t)$ for $t \ge s$ and $\bar{\omega} \in C$. $\mathcal{F}_t^s$, where $s < t$, will denote the σ-algebra

of cylinder sets in $\mathcal{F}$ with bases in $[s,t]$. Conditions 1–3 are clearly satisfied. We first define the measures $\mathbf{P}_{s,x}$ on the σ-algebra $\mathcal{F}_\infty^s$ with the help of the finite-dimensional distribution functions given by (1.1.5). For any $C \in \mathcal{B}$, we regard $\xi_s(\omega)$ as belonging to C if $\xi_s(t,\omega) = \xi_s(t,\bar{\omega})$ for some $\bar{\omega} \in C$. Parts i–iii of property 4 follow from formula (1.1.5) and properties I–III of a transition probability.

Therefore, in a sense, the study of Markov processes reduces to studying their transition probabilities $P(s,x,t,A)$, functions satisfying conditions I–III. Equation (1.1.4) is fundamental there.

Two families of linear transformations can be associated with a transition probability. One acts in the space $\mathbf{B}_x$ of bounded $\mathcal{B}$-measurable functions from X to R. It is a Banach space with norm $\|f\| = \sup_x |f(x)|$. The other acts in the space $\mathcal{M}_X$ of charges (countably-additive functions of bounded variation) on $\mathcal{B}$. If ν is such a charge, then

$$\|\nu\| = \operatorname{var} \nu = \sup_{\|f\| \le 1} \left| \int f d\nu \right| .$$

These families of transformations $T_{s,t}$ are given by

$$[T_{s,t} f](x) = \int f(y) P(s,x,t,dy) , \quad f \in \mathbf{B}_X , \tag{1.1.6}$$

and

$$[\nu T_{s,t}](A) = \int P(s,x,t,A) \nu(dx) , \quad \nu \in \mathcal{M}_X . \tag{1.1.7}$$

The first maps $\mathbf{B}_X$ into $\mathbf{B}_X$ while the second maps $\mathcal{M}_X$ into $\mathcal{M}_X$. We shall denote a transformation by the same letter but applied on the right for measures and on the left for functions (this is similar to matrix multiplication on the right for row vectors and on the left for column vectors).

It follows from condition II that $\|T_{s,t}\| = 1$. Finally, condition III can be rewritten as follows. For $0 \le s < t < u$,

$$T_{s,t} T_{t,u} = T_{s,u} . \tag{1.1.8}$$

Since $P(t,x,t,A) = I_A(x)$, we have $T_{s,s} = I$, where I is the identity operator. The operators $T_{s,t}$ possess one further property. They map nonnegative functions in $\mathbf{B}_X$ into nonnegative functions and measures in $\mathcal{M}_X$ into measures.

We give now a relation between a transition probability and a class of martingales generated by a Markov process. Let $f \in \mathbf{B}_X$ and for $0 \le t \le a$, let

$$u(t,x) = \int P(t,x,a,dy) f(y) . \tag{1.1.9}$$

Then when $t \in [s,a]$, the numerical process $\zeta_t = u(t, \xi_s(t))$ is a martingale with respect to the flow $\mathcal{F}_t^s$ on the probability space $(\Omega, \mathcal{F}, \mathbf{P}_{s,x})$ for any $x \in X$. In fact, if $s < v < t$, then

$$\mathbf{E}_{s,x}(u(t,\xi_s(t))|\mathcal{F}_v^s) = \mathbf{E}_{v,\xi_s(v)}u(t,\xi_v(t))$$
$$= \int u(t,z)P(v,\xi_s(v),t,dz) = \int \left(\int P(t,z,a,dy)f(y)\right)$$
$$\times P(v,\xi_s(v),t,dz) = \int f(y)\int P(v,\xi_s(v),t,dz)P(t,z,a,dy)$$
$$= \int f(y)P(v,\xi_s(v),a,dy) = u(v,\xi_s(v))\ .$$

Conversely, if $\Phi(t,x)$ is a bounded function such that $\Phi(t,\xi_s(t))$ is a martingale when $t\in[s,a]$, then

$$\Phi(t,\xi_s(t)) = \mathbf{E}_{s,x}(\Phi(a,\xi_s(a))|\mathcal{F}_t^s)$$
$$= \mathbf{E}_{t,\xi_s(t)}\Phi(t,\xi_t(a)) = \int \Phi(t,z)P(t,\xi_s(t),a,dz)\ .$$

Thus formula (1.1.9) gives all of the bounded functions $u(t,x)$ of two variables that are measurable in x for which $u(t,\xi_s(t))$ is a martingale with respect to the measures $\mathbf{P}_{s,x}$.

(a) *Homogeneous Markov Processes.* A Markov process is *homogeneous* or *has stationary transition probabilities* if $P(s,x,t,A)$ depends only on the difference $t-s$: $P(s,x,t,A) = P(0,x,t-s,A)$. When a process is homogeneous, $P(0,x,t,A)$ will be denoted by $P(t,x,A)$ and we shall say that the transition probability is temporally homogeneous. For a homogeneous process, the Chapman-Kolmogorov equation takes the form

$$P(t+s,x,A) = \int P(t,x,dz)P(s,z,A)\ . \tag{1.1.10}$$

The operators $T_{s,t}$ depend only on the difference $t-s$. If we set $T_{0,t} = T_t$, then the family $\{T_t, t>0\}$ is a one-parameter semigroup of operators: $T_{t+s} = T_tT_s$ and the operators T_t commute with one another. For a homogeneous process, formula (1.1.5) becomes

$$\mathbf{P}_{s,x}\{\xi_s(t_1)\in A_1,\dots,\xi_s(t_n)\in A_n\}$$
$$= \int_{A_1} P(t_1-s,x,dx_1)\int_{A_2} P(t_2-t_1,x_1,dx_1)$$
$$\dots\int_{A_n} P(t_n-t_{n-1},x_{n-1}dx_n)\ . \tag{1.1.11}$$

This formula shows that the finite-dimensional distributions of $\xi_s(t+s)$ only depend on x (the initial state) and not on s. Therefore it is possible to concentrate on one process $\xi(t)$ and a family of measures $\mathbf{P}_x$, $x\in X$, which correspond to a process beginning at time 0 in the state x. For a homogeneous Markov process, condition (1.1.2) assumes this form: For $0\le u<t_1\dots<t_n$,

$$\mathbf{P}_x\{\xi(t_1) \in A_1, \ldots, \xi(t_n) \in A_n | \mathcal{F}_u^0\} = \mathbf{P}_{\xi(u)}\{\xi(t_1 - u) \in A_1, \ldots, \xi(t_n - u) \in A_n\} . \tag{1.1.12}$$

Let $u(t,x) = T_t f(x)$. Then the numerical process $u(t, \xi(t-s))$ is a martingale with respect to measure $\mathbf{P}_x$ for all $x \in X$.

1.1.3 Regularity

We are interested in when a Markov process can be considered continuous or right-continuous, assuming that it has at most jump discontinuities. The following general result is a consequence of Theorem 4.1.6 on p. 99. Let X be a complete metric space with metric $\varrho(x,y)$ and let $\mathcal{B}$ be the σ-algebra generated by its balls. Denote by $S_r(y)$ the ball with center at y of radius r.

Theorem 1.1.1. *Suppose that the transition probability $P(s,x,t,A)$ is uniformly stochastically continuous: For all T*

$$\lim_{h \to 0} \sup_{x \in X} \sup_{t \le T, s \le T, |t-s| \le h} P(s,x,t, X \setminus S_r(x)) = 0 . \tag{1.1.13}$$

Then there exist functions $\tilde{\xi}_s(t,\omega)$ such that: 1. *$\tilde{\xi}_s(t,\omega)$ has at most jump discontinuities as a function of $t \ge s$;* 2. *$\mathbf{P}_{s,x}\{\xi_s(t,\omega) = \tilde{\xi}_s(t,\omega)\} = 1$ for all $s \in R_+$ and $x \in X$.*

Proof. Consider the value of the process $\xi_s(t,\omega)$ on the set $\Lambda \cap [s,\infty)$, where Λ are the rationals. As shown in the above-mentioned theorem, $\xi_s(t,\omega)$ has a limit with probability $\mathbf{P}_{s,x} = 1$ on $\Lambda \cap [s,\infty)$ over any decreasing or bounded increasing sequence of arguments. The process $\xi_s(t+,\omega) = \xi_s(t,\omega)$ will then have the necessary requirements. □

Remark. It is easy to derive sufficient conditions for the processes $\tilde{\xi}_s(t)$ to be continuous. To this end, it is sufficient for any $r > 0$, $T > 0$ and $x \in X$ that

$$\lim_{\max \Delta t_k \to 0} \sum_{k=0}^{n-1} \int P(s,x,t_k,dy) P(t_k,y,t_{k+1}, X \setminus S_r(y)) = 0 ,$$
$$s = t_0 < t_1 < \ldots < t_n = T , \quad \Delta t_k = t_{k+1} - t_k .$$

The sum in this limit equals

$$\mathbf{E}_{s,x} \sum_{k=0}^{n-1} I_{\{\varrho(\tilde{\xi}_s(t_k), \tilde{\xi}_s(t_{k+1})) > r\}} ,$$

and the quantity under the expectation sign tends (for almost all r) to the number of jumps of the process exceeding r. In particular, the stated condition will hold if

$$P(t,x,t+h, X \setminus S_r(x)) = o(h)$$

uniformly in $t \le T$ and $x \in X$.

(a) *Conditions associated with martingales.* Let X be a complete separable metric space and let $\mathbf{C}_X$ be the space of bounded continuous functions with the usual norm. Assume that the following conditions hold:

Φ 1. The operators $T_{s,t}$ map $\mathbf{C}_X$ into $\mathbf{C}_X$.
Φ 2. $T_{s,t}f(x) \to f(x)$ as $s \to t$ uniformly on each ball.

Let $u_f(s,x,t) = T_{s,t}f(x)$. Then when $u < s < t$, the process $u_f(s, \xi_u(s), t)$ is a martingale. It is possible to choose a modification of this martingale so that it has at most jump discontinuities and is continuous from the right. Using this fact, one can select a countable set of continuous functions $f_n(x)$ (for instance, $r(x_n, x)$, where $\{x_n\}$ is dense in X) which separate the points in X. One can show that $\xi_u(s)$ also has a modification with at most jump discontinuities which is continuous from the right.

1.2 Purely Discontinuous Processes

We now assume that a process has a phase space $(X, \mathcal{B})$ such that $\mathcal{B}$ contains all singletons $\{x\}$, $x \in X$.

Definition 1.2.1. A Markov process is uniformly purely discontinuous if for any positive T

$$\lim_{h \downarrow 0} \frac{1}{h}(P(t,x,t+h,A) - I_A(x)) = q(t,x,A) \tag{1.2.1}$$

exists uniformly in $x \in X$, $t \leq T$ and $A \in \mathcal{B}$.

Being the uniform limit of countably additive functions (in A) of bounded variation, $q(t,x,A)$ will be the same kind of function. Furthermore, $q(t,x,\{x\})$ and $q(t,x,X \setminus \{x\})$ will be bounded. It follows from (1.2.1) that $q(t,x,X) = 0$. Define

$$\lambda(t,x) = -q(t,x,\{x\}) \; . \tag{1.2.2}$$

Then $q(t,x,X \setminus \{x\}) = \lambda(t,x)$ and

$$\pi(t,x,A) = \frac{1}{\lambda(t,x)} q(t,x,A \setminus \{x\})$$

is a probability measure with respect to A. (It is defined for $\lambda(t,x) > 0$; if $\lambda(t,x) = 0$, we take $\pi(t,x,A) = I_A(x)$.) As already indicated, $\lambda(t,x)$ is bounded. If $\lambda(t,x) < C$, then

$$1 - P\{t,x,t+h,\{x\}) < Ch$$

or

$$P(t,x,t+h,\{x)\} > e^{-Ch}$$

for positive h small enough. Since

$$P(t,x,u,\{x\}) \geq P(t,x,s,\{x\})P(s,x,u,\{x\})$$

when $t < s < u$ (this is a consequence of the Chapman-Kolmogorov equation), for all x we have

$$P(s,x,t,\{x\}) \geq e^{-C(t-s)}$$

or

$$P(s,x,t,X \setminus \{x\}) \leq 1 - e^{-C(t-s)} \ .$$

Viewing X as a discrete metric space, one can see that the process has a modification with at most jump discontinuities that is right-continuous.

Theorem 1.2.1. *Suppose that a Markov process is purely discontinuous and that $\xi_s(t)$ has at most jump discontinuities and is right-continuous. Denote by τ the first exit time of the process from its initial state. Then*

$$\mathbf{P}_{s,x}\{\tau < t, \xi_s(\tau) \in A\} = \int_s^t \lambda(u,x) e^{-\int_s^u \lambda(v,x)dv} \pi(u,x,A)du \ . \tag{1.2.3}$$

Proof. Noting that

$$\xi_s(\tau) = \lim_{n\to\infty} \sum_{k=0}^{\infty} \prod_{i=1}^{k} I_{\{\xi_s(s+i/2^n)=x\}} I_{\{\xi_s(s+(k+1)/2^n)\in X\setminus\{x\}\}} \xi_s\left(\frac{k+1}{2^n}\right)$$

(although X is not necessarily linear, we take $0 \cdot x = 0$ and $0 + x = x$ for $x \in X$), we obtain

$$I_A(\xi_s(\tau)) = \lim_{n\to\infty} \sum_{k=0}^{\infty} \left(\prod_{i=1}^{k} I_{\{\xi_s(s+i/2^n)=x\}} \right) I_{\{\xi_s(s+(k+1)/2^n)\in A\setminus\{x\}\}}.$$

In exactly the same way

$$\tau - s = \lim_{n\to\infty} \sum_{k=1}^{\infty} \frac{k}{2^n} \left(\prod_{i=1}^{k} I_{\{\xi_s(s+i/2^n)=x\}} \right) I_{\{\xi_s(s+(k+1)/2^n)\neq x\}}$$

with probability $\mathbf{P}_{s,x} = 1$. For $z > 0$,

$$\begin{aligned}
\mathbf{E}_{s,x} e^{-z(\tau-s)} &= \mathbf{E}_{s,x} \lim_{n\to\infty} \sum_{k=1}^{\infty} e^{-kz/2^n} \left(\prod_{i=1}^{k} I_{\{\xi_s(s+i/2^n)=x\}} \right) I_{\{\xi_s(s+(k+1)/2^n)\neq x\}} \\
&= \lim_{n\to\infty} \sum_{k=1}^{\infty} e^{-kz/2^n} \mathbf{E}_{s,x} \left(\prod_{i=1}^{k} I_{\{\xi_s(s+i/2^n)=x\}} \right) I_{\{\xi_s(s+(k+1)/2^n)\neq x\}}.
\end{aligned}$$

(It is permissible to take $\mathbf{E}_{s,x}$ under the limit sign because the sum does not exceed 1. Since

$$\mathbf{E}_{s,x}\left(\prod_{i=1}^{k} I_{\{\xi_s(s+i/2^n)=x\}}\right) I_{\{\xi_s(s+(k+1)/2^n)\neq x\}}$$
$$= \left(\prod_{i=1}^{k} P\left\{s+(i-1)/2^n, x, s+i/2^n, \{x\}\right\}\right)$$
$$\times P(s+k/2^n, x, s+(k+1)/2^n, X\setminus\{x\})$$
$$= \left(\prod_{i=1}^{k}\left(1-\lambda\left(s+(i-1)/2^n, x\right)\frac{1}{2^n}+o\left(\frac{1}{2^n}\right)\right)\right.$$
$$\left.\times\,\lambda\left(s+k/2^n, x\right)\left(\frac{1}{2^n}+o\left(\frac{1}{2^n}\right)\right)\right),$$

it follows that

$$\mathbf{E}_{s,x}e^{-z(\tau-s)}$$
$$= \lim_{n\to\infty}\frac{1}{2^n}\sum_{k=1}^{\infty}\lambda\left(s+k/2^n, x\right)\exp\left\{-\frac{1}{2^n}\sum_{i=1}^{k}\lambda\left(s+(i-1)/2^n, x\right)-\frac{zk}{2^n}\right\}$$
$$= \int_0^{\infty}\exp\left\{-\int_0^{t-s}\lambda(s+u,x)du-(t-s)z\right\}\lambda(t,x)dt\,.$$

This implies that τ has a continuous distribution with respect to the measure $\mathbf{P}_{s,x}$ with density

$$\frac{d}{dt}\mathbf{P}_{s,x}\{\tau<t\} = I_{(t>s)}\exp\left\{-\int_s^t\lambda(u,x)du\right\}\lambda(t,x)\,. \tag{1.2.4}$$

By the continuity of the distribution of τ,

$$I_{\{\tau<t\}}I_{\{\xi_s(\tau)\in A\}}\lim_{n\to\infty}\sum_{k+1<2^n(t-s)}\left(\prod_{i=1}^{k}I_{\{\xi_s(s+i/2^n)=x\}}\right)I_{\{\xi_s(s+(k+1)/2^n)\in A\setminus\{x\}\}}$$

with probability $\mathbf{P}_{s,x}=1$. Thus

$$\mathbf{E}_{s,x}I_{\{\tau<t\}}I_{\{\xi_s(\tau)\in A\}}$$
$$= \lim_{n\to\infty}\sum_{k+1<2^n(t-s)}\mathbf{E}_{s,x}\left(\prod_{i=1}^{k}I_{\{\xi_s(s+i/2^n)=x\}}\right)I_{\{\xi_s(s+(k+1)/2^n)\in A\setminus\{x\}\}}$$
$$= \lim_{n\to\infty}\sum_{k+1<2^n(t-s)}\left(\prod_{i=1}^{k}P\left(s+(i-1)/2^n, x, s+i/2^n, \{x\}\right)\right)$$
$$\times P\big(s+k/2^n, x, s+(k+1)/2^n, A\setminus\{x\}\big)$$
$$= \lim_{n\to\infty}\sum_{k+1<2^n(t-s)}\left(\prod_{i=1}^{k}\left(1-\frac{1}{2^n}\lambda\left(s+(i-1)/2^n, x\right)+o\left(\frac{1}{2^n}\right)\right)\right)$$
$$\times\left(\frac{1}{2^n}\lambda\left(s+\frac{k}{2^n}, x\right)+o\left(\frac{1}{2^n}\right)\right)\pi\left(s+k/2^n, x, A\right).$$

From this we obtain (1.2.3). □

Remark. For $s < t$, formulas (1.2.3) and (1.2.4) give

$$\pi(t, x, A) = \mathbf{P}_{s,x}\{\xi_s(\tau) \in A | \tau = t\} .$$

1.2.1 Kolmogorov's Equations

Theorem 1.2.2. *The transition probability of a purely discontinuous process obeying* (1.2.1) *satisfies each of the following equations:*

$$-\frac{\partial}{\partial s}P(s, x, t, A) = \int q(s, x, dy)P(s, y, t, A), \quad s < t , \tag{1.2.5}$$

$$\frac{\partial}{\partial t}P(s, x, t, A) = \int P(s, x, dy)q(t, y, A), \quad t > s , \tag{1.2.6}$$

Equation (1.2.5) is called the *backward* (or *first*) *Kolmogorov equation* and (1.2.6) is the *forward* (or *second*) *Kolmogorov equation.*

Proof. We have

$$\begin{aligned} &P(s - h, x, t, A) - P(s, x, t, A) \\ &\quad = \int [P(s - h, x, s, dz) - I_{\{x \in dz\}}]P(s, z, t, A) . \end{aligned}$$

Therefore

$$\begin{aligned} &|P(s - h, x, t, A) - P(s, x, t, A)| \\ &\quad \le \operatorname{var} |P(s - h, x, s, \cdot) - I_{\{x\}}| . \end{aligned}$$

The right-hand side involves the variation of a difference of measures, $I_{\{x\}}$ being a measure concentrated at x and equaling 1 there. Clearly,

$$\begin{aligned} \operatorname{var}|P(s - h, x, s, \cdot) - I_{\{x\}}| &\le (1 - P(s - h, x, s, \{x\}) \\ &+P(s - h, x, s, X \setminus \{x\}) = 2|1 - P(s - h, x, s, \{x\})| \\ &\le 2(1 - e^{-Ch}) . \end{aligned}$$

Thus $P(s, x, t, A)$ is continuous in s. The continuity of $P(s, x, t, A)$ in t and jointly in s and t may be established in similar fashion. The continuity in t of the pre-limiting expression on the left-hand side of (1.2.1) and the uniform convergence entail the continuity of $q(t, x, A)$ in t. We then have

$$\begin{aligned} &\frac{1}{h}(P(s - h, x, t, A) - P(s, x, t, A)) \\ &\quad = \int \frac{1}{h}(P(s - h, x, s, dz) - I_{\{x \in dz\}})P(s, z, t, A) . \end{aligned}$$

By (1.2.1), the right-hand limit exists and equals the right-hand side of (1.2.5). Therefore the limit of the left-hand side also exists and it is the left-hand partial derivative of $P(s,x,t,A)$ with respect to s. But this derivative is continuous in s and so it is the same as the regular partial derivative. Thus (1.2.5) holds. In similar fashion, starting from the relation

$$\begin{aligned}&\frac{1}{h}(P(s,x,t+h,A)-P(s,x,t,A))\\&\quad=\int P(s,x,t,dy)\frac{1}{h}(P(t,y,t+h,A)-I_A(y))\end{aligned}$$

one can derive (1.2.6). □

Remark. The functions $\lambda(s,x)$ and $\pi(s,x,A)$ can be used to rewrite (1.2.5) as follows:

$$\begin{aligned}-\frac{\partial}{\partial s}P(s,x,t,A)&=-\lambda(s,x)P(s,x,t,A)\\&\quad+\lambda(s,x)\int\pi(s,x,dy)P(s,y,t,A)\ .\end{aligned}$$

Since

$$\begin{aligned}&-\frac{\partial}{\partial s}P(s,x,t,A)+\lambda(s,x)P(s,x,t,A)\\&\quad=-\exp\left\{-\int_s^t\lambda(u,x)du\right\}\frac{\partial}{\partial s}\left(\exp\left\{\int_s^t\lambda(u,x)du\right\}P(s,x,t,A)\right)\end{aligned}$$

and

$$\begin{aligned}&\int_s^t\exp\left\{-\int_v^t\lambda(u,x)du\right\}P(v,x,t,A)dv\\&\quad=I_A(x)-\exp\left\{\int_s^t\lambda(u,x)du\right\}P(s,x,t,A)\ ,\end{aligned}$$

we arrive at the following integral equation for $P(s,x,t,A)$:

$$\begin{aligned}P(s,x,t,A)&=I_A(x)\exp\left\{-\int_s^t\lambda(u,x)du\right\}\\&\quad+\int_s^t\int\lambda(v,x)\exp\left\{-\int_v^t\lambda(u,x)du\right\}\pi(v,x,dy)P(v,y,t,A)dv\ .\end{aligned}\tag{1.2.7}$$

Equation (1.2.7) can be solved by the method of successive approximations. If we let

$$\left.\begin{aligned}Q_0(s,x,t,A)&=\exp\left\{-\int_s^t\lambda(u,x)du\right\}I_A(x),\\Q_n(s,x,t,A)&\\=\int_s^t\int\lambda(v,x)\exp\left\{-\int_v^t\lambda(u,x)du\right\}\pi(v,x,dy)&\\\times Q_{n-1}(v,y,A)dv\ ,\quad n\geq 1\ ,&\end{aligned}\right\}\tag{1.2.8}$$

then

$$P(s,x,t,A) = \sum_{n=0}^{\infty} Q_n(s,x,t,A) \ . \tag{1.2.9}$$

Equation (1.2.7) is meaningful and has a unique solution determined by (1.2.8), (1.2.9) even under the broader conditions:

(i) $\lambda(t,x)$ is measurable jointly in its arguments, it is nonnegative and it is bounded,

(ii) $\pi(t,x,A)$ is a probability measure with respect to A; $\pi(t,x,\{x\}) = 0$ when $\lambda(t,x) > 0$ and $\pi(t,x,\{x\}) = 1$ when $\lambda(t,x) = 0$; $\pi(t,x,A)$ is measurable in t,x for all $A \in \mathcal{B}$.

If the transition probability of a Markov process satisfies (1.2.7) in which λ and π obey (i) and (ii), then the process is said to be *purely discontinuous*.

(a) *Processes with a finite set of states.* Let X be a finite set and $\mathcal{B}$ the σ-algebra of all its subsets. The transition probability is determined upon giving the function

$$p(s,x,t,y) = P(s,x,t,\{y\}),\ x,\ y \in X,\ 0 \le s \le t \ ,$$

since

$$P(s,x,t,A) = \sum_{y \in A} p(s,x,t,y) \ .$$

Condition I of Sect. 1.1.2 holds automatically and condition II reduces to

$$p(s,x,t,y) > 0 \ , \quad \sum_{y \in X} p(s,x,t,y) = 1 \ .$$

The Chapman-Kolmogorov equation assumes this form: For $s < t < u$,

$$p(s,x,u,y) = \sum_{z \in X} p(s,x,t,z) p(t,z,u,y) \ .$$

The process is regular if to every $\varepsilon > 0$ and $T > 0$, there exists a $\delta > 0$ such that $1 - p(s,x,t,y) < \delta$ for $s < T$, $0 < t - s < \delta$ and all $x \in X$. Operators $T_{s,t}$ may be defined as follows. Enumerate the elements in X, to wit, $x_1, x_2, \ldots, x_m$. The space $\mathbf{B}_X$ is determined by the vectors $(f_1, f_2, \ldots, f_m)$, where f_k is the value of function f at x_k, and the space $\mathcal{M}_X$, by the vectors $(p_1, \ldots, p_m)$, where p_i is the value of a (signed) measure at x_i. Then

$$T_{s,t} f(x_k) = \sum_{j=1}^{m} p(s,x_k,t,x_j) f_j$$

and

$$\mu T_{s,t}(\{x_j\}) = \sum_k \mu(x_k) p(s,x_k,t,x_j) \ .$$

Let $\prod_{s,t}$ be the matrix with elements $p(s, x_k, t, x_j)$ (in the k-th row and j-th column). Then $T_{s,t}$ acts on functions like this matrix does on columns and it acts on measures like this matrix on rows.

The process will be purely discontinuous if

$$\lim_{h\downarrow 0} \frac{1}{h} p(t, x, t+h, y) = a(t, x, y), \quad x \neq y \ ,$$

exists uniformly in $t \leq T$ for any positive T. Write $a(t, x, x) = -\sum_{y \neq x} a(t, x, y)$.

Then Kolmogorov's backward and forward equations become

$$-\frac{\partial}{\partial s} p(s, x, t, y) = \sum_{z \in X} a(s, x, z) p(s, z, t, y) \tag{1.2.10}$$

and

$$\frac{\partial}{\partial t} p(s, x, t, y) = \sum_{z \in X} p(s, x, t, z) a(t, z, y) \ . \tag{1.2.11}$$

Equation (1.2.10) converts to the following integral equation:

$$p(s, x, t, y) = I_{\{y\}}(x) \exp\left\{\int_s^t a(u, x, x) du\right\} + \sum_{z \in X} \int_s^t \exp\left\{\int_v^t a(u, x, x) du\right\} a(v, x, z) p(v, z, t, y) dv \ . \tag{1.2.12}$$

(b) *Homogeneous processes.* A homogeneous process with transition probability $P(t, x, A)$ is called purely discontinuous if

1. $\lim_{h\to 0} \frac{1}{h} P(h, x, A) = q(x, A)$ exists uniformly in $A \subset X \setminus \{x\}$ for every $x \in X$, where $q(x, A)$ is a measure with respect to A;

2. the function $\lambda(x) = q(x, X \setminus \{x\})$ is bounded.

(These conditions are weaker in form than those given on p. 153. However they are equivalent to them in the homogeneous case.) From the relation $P(h, x, \{x\}) \geq e^{-hC}$, holding for h sufficiently small, it follows that

$$P(t, x, \{x\}) \geq e^{-tC}$$

for all t; C does not depend on x. This inequality implies that the variation of the set-function

$$\frac{1}{h}(P(h, x, A) - I_A(x))$$

satisfies

$$\frac{2}{h}(1 - P(h, x, \{x\})) \leq \frac{2}{h}(1 - e^{-hC}) \leq 2C \ .$$

Just as in Theorem 1.2.2, this last relation leads to the derivation of the backward and forward Kolmogorov equations:

$$\frac{\partial}{\partial t}P(t,x,A) = -\lambda(x)P(t,x,A) + \int q(x,dy)P(t,x,A)\ , \tag{1.2.13}$$

$$\frac{\partial}{\partial t}P(t,x,A) = -\int_A P(t,x,dy)\lambda(y) + \int P(t,x,dy)q(y,A)\ . \tag{1.2.14}$$

From (1.2.13), we obtain the following analogue of Eq. (1.2.7):

$$P(t,x,A) = e^{-t\lambda(x)}I_A(x) + \int_0^t \int e^{-(t-s)\lambda(x)}P(s,y,A)q(x,dy)\ . \tag{1.2.15}$$

With $x \notin A$, this gives

$$\begin{aligned} P(t,x,A) &= \int_0^t e^{-(t-s)\lambda(x)} \int_A q(x,dy)e^{-s\lambda(y)}ds + O(t^2) \\ &= tq(x,A) + O(t^2)\ ; \\ P(t,x,\{x\}) &= e^{-t\lambda(x)} + O(t^2)\ . \end{aligned}$$

From these relations, we find that

$$\lim_{h\downarrow 0}\frac{1}{h}(P(h,x,A) - I_A(x)) = q(x,A) - \lambda(x)I_A(x)$$

exists uniformly in x and A.

(c) *Denumerable homogeneous processes.* In this instance, $X = (1,2,\ldots)$ and $\mathcal{B}$ is the σ-algebra of all subsets of X. The transition probability is given by the collection of functions $p_{ij}(t) = P(t,i,\{j\})$. The Chapman-Kolmogorov equation assumes the form

$$p_{ij}(t+s) = \sum_k p_{ik}(t)p_{kj}(s)\ .$$

The process will be purely discontinuous if

$$(\alpha) \qquad a_{ij} = \lim_{h\to 0}\frac{1}{h}p_{ij}(h) \quad \text{exists for} \quad i \neq j\ ;$$

$$(\beta) \qquad -a_{ii} = \lim_{h\to 0}\frac{1}{h}(1 - p_{ii}(h)) \quad \text{exists for all} \quad i\ ;$$

$$(\gamma) \qquad \sup_i(-a_{ii}) < \infty \quad \text{and} \quad \sum_k a_{ik} = 0\ .$$

Kolmogorov's backward and forward equations are

$$\frac{dp_{ij}(t)}{dt} = \sum_k a_{ik}p_{kj}(t) \tag{1.2.16}$$

and

$$\frac{dp_{ij}(t)}{dt} = \sum_k p_{ik}(t)a_{kj}\ .$$

Let P_t be an operator in the space of bounded sequences $x = \{x_k\}$ in R acting according to the formula

$$(P_t x)_k = \sum p_{ki}(t) x_i \ .$$

Equation (1.2.16) becomes

$$\frac{d}{dt} P_t = A P_t \ , \tag{1.2.17}$$

where A is a bounded operator defined by

$$(Ax)_k = \sum a_{ki} x_i \ .$$

From (1.2.17) and the condition $P_0 = I$, we obtain

$$P_t = \exp\{tA\} = I + \sum_{n=1}^{\infty} \frac{t^n}{n!} A^n \ . \tag{1.2.18}$$

1.3 Diffusion Processes

We next consider Markov processes with phase space R^d. $\mathcal{B}$ is the σ-algebra of Borel sets in R^d. A transition probability will be assumed to satisfy the following condition: For all $T > 0$ and $\varepsilon > 0$,

$$\lim_{\delta \to 0} \sup_{t \le T} \sup_{x \in X} \sup_{0 < t-s < \delta} \frac{1}{\delta} P(s, x, t, V_\varepsilon(x)) = 0 \ , \tag{1.3.1}$$

where $V_\varepsilon(x) = \{y : |x - y| > \varepsilon\}$. Section 1.1.3 shows that a continuous modification exists for the process under this assumption. A continuous Markov process is capable of serving as model for the motion of a particle colliding with the chaotically moving molecules of a medium, that is, as a model of the diffusion or dispersal of alien particles in the medium. This explains the name given to the class of Markov processes being considered.

Definition 1.3.1. A continuous Markov process with a transition probability $P(s, x, t, A)$ satisfying (1.3.1) is called a *diffusion process* if it fulfills the following two conditions:

I. There exists a continuous function $a(t, x)$ defined on $R_+ \times R^d$ with values in R^d such that for all $\varepsilon > 0$ and $T > 0$,

$$\lim_{h \to 0} \frac{1}{h} \int_{|y-x| < \varepsilon} (y - x) P(t, x, t + h, dy) = a(t, x) \tag{1.3.2}$$

uniformly in $x \in R^d$ and $t \le T$.

II. There exists a continuous function $B(t, x)$ defined on $R_+ \times R^d$ with values in $L(R^d)$ such that for all $\varepsilon > 0$, $T > 0$ and $v, z \in R^d$,

$$\lim_{h\to 0}\frac{1}{h}\int_{|y-x|\le\varepsilon}(y-x,z)(y-x,\upsilon)P(t,x,t+h,dy)$$
$$=(B(t,x)z,\upsilon)\ . \tag{1.3.3}$$

The vector function $a(t,x)$ is called the *transport coefficient* of the diffusion process. It is the rate of flow in the medium in which the diffusing particle is located. The function $B(t,x)$ is called the *diffusion operator*. It characterizes the variance of the particle deviation from its initial position. $B(t,x)$ is a nonnegative symmetric operator. $B(t,x)$ and $a(t,x)$ together are the *diffusion coefficients* of the process. More precisely, if some basis $\{e_1,\dots,e_d\}$ has been chosen in R^d, then the diffusion coefficients are the coordinates $a^k(t,x)=(a(t,x),e_k)$, $k=1,\dots,d$, of $a(t,x)$ in this basis and the elements of the matrix of the operator $B(t,x)$ in this basis,

$$b^{ik}(t,x)=(B(t,x)e_i,e_k)\ ,\quad i,k\le d\ .$$

A very simple example of a diffusion process in R^d is the Wiener process. It is a homogeneous process $\xi(t)$ with independent increments for which $\xi(t+h)-\xi(t)$ has a normal distribution with zero mean and identity covariance matrix. In other words, the distribution of $\xi(t+h)-\xi(t)$ has the density

$$(2\pi h)^{-d/2}\exp\left\{-\frac{1}{2h}(x,x)\right\}\ .$$

The transition probability of such a process is given by

$$P(s,x,t,A)=(2\pi(t-s))^{-d/2}\int_A\exp\left\{-\frac{1}{2(t-s)}(y-x,y-x)\right\}dy\ .$$

A simple computation shows that for all positive m,

$$\int|y-x|^m P(s,x,t,dy)=O((t-s)^{m/2})\ ,$$
$$\int(y-x)P(s,x,t,dy)=0\ ,$$
$$\int(y-x,z)(y-x,\upsilon)P(s,x,t,dy)=(z,\upsilon)(t-s)\ .$$

Therefore $\xi(t)$ is a diffusion process with $a=0$ and $B=I$, the identity matrix.

1.3.1 Kolmogorov's Equations

Our goal is to derive differential equations by means of which the transition probability may be determined. These equations resemble the diffusion equations of physics, excessive evidence that the term "diffusion process" reflects the essential aspects of the process. We first prove a preliminary result.

Lemma 1.3.1. *If $\varphi(x)$ is a scalar function on R^d which is bounded, continuous and twice differentiable in a neighborhood of x_0, then*

$$\lim_{h\to 0}\frac{1}{h}\left[\int \varphi(y)P(t,x_0,t+h,dy)-\varphi(x_0)\right]$$
$$=(a(t,x_0),\varphi'(x_0))+\frac{1}{2}\operatorname{Tr} B(t,x_0)\varphi''(x_0) \tag{1.3.4}$$

($\varphi''(x_0)$ is an operator in $L(R^d)$ such that

$$(\varphi''(x_0)u,v)=\frac{\partial^2}{\partial t\partial s}\varphi(x_0+tu+sv)|_{\substack{t=0\\ s=0}}\ ,$$

and $\operatorname{Tr} A$ is the trace of operator A).

Proof. For any positive ε,

$$\begin{aligned}
&\int(\varphi(y)-\varphi(x_0))P(t,x_0,t+h,dy)\\
&\quad=\int_{|x-y|\le\varepsilon}(\varphi(y)-\varphi(x_0))P(t,x_0,t+h,dy)+o(h)\\
&\quad=\int_{|x-y|\le\varepsilon}(\varphi'(x_0),y-x_0)P(t,x_0,t+h,dy)\\
&\qquad+\int_{|x-y|\le\varepsilon}\frac{1}{2}(\varphi''(x_0)(y-x_0),y-x_0)P(t,x_0,t+h,dy)\\
&\qquad+\int\alpha_\varepsilon|y-x_0|^2P(t,x_0,t+h,dy)+o(h)\ ,
\end{aligned}$$

where

$$\begin{aligned}
\alpha_\varepsilon|y-x_0|^2&=\varphi(y)-\varphi(x_0)\\
&\quad-(\varphi'(x_0),y-x_0)-\frac{1}{2}(\varphi''(x_0)(y-x_0),y-x_0)\ ,
\end{aligned}$$

and $\alpha_\varepsilon\to 0$ as $\varepsilon\to 0$. Applying relations (1.3.2) and (1.3.3), we arrive at (1.3.4). □

Theorem 1.3.1. *Suppose that $\varphi\in\mathbf{C}_{R^d}$, the function*

$$u(s,x)=\int P(s,x,t,dy)\varphi(y)$$

is twice differentiable in x and $\partial u/\partial x$ and $\partial^2u/\partial x^2$ are continuous in s. Then for $s<t$, it satisfies the differential equation

$$-\frac{\partial}{\partial s}u(s,x)=(a(s,x),u_x(s,x))+\frac{1}{2}\operatorname{Tr} B(s,x)u_{xx}(s,x) \tag{1.3.5}$$

and boundary condition

$$\lim_{s\to t}u(s,x)=\varphi(x)\ .$$

Proof. On the basis of Lemma 1.3.1, we have

$$u(s-h,x)-u(s,x)=\int u(s,y)P(s-h,x,s,dy)-u(s,x)$$
$$=h\left[(a(s,x),u_x(s,x))+\frac{1}{2}\mathrm{Tr}\,B(s,x)u_{xx}(s,x)\right]+o(h)\,.$$

This implies that $u(s,x)$ has a left-hand partial derivative with respect to s and its value equals the left-hand side of (1.3.5) taken with a minus. Thus, it is continuous and so it equals the regular derivative. □

Equation (1.3.5) can be related further to a diffusion process as follows.

Theorem 1.3.2. *For $s \le t$, suppose that $u(s,x)$ is continuous, bounded and has bounded continuous derivatives $\partial u(s,x)/\partial s$, $\partial u(s,x)/\partial x$ and $\partial^2 u(s,x)/\partial x^2$. If $u(s,x)$ satisfies* (1.3.5)*, then*

$$u(s,x)=\int u(t,y)P(s,x,t,dy)\,. \tag{1.3.6}$$

Proof. To show that (1.3.6) holds, it is sufficient to demonstrate that $u(s,\xi(s))$ is a martingale. We have

$$\mathbf{E}(u(s+h,\xi(s+h))|\xi(s))=\int P(s,\xi(s),s+h,dy)u(s+h,y)$$
$$=u(s,\xi(s))+u_s(s,\xi(s))h+o(h)+\int u(s,y)P(s,\xi(s),s+h,dy)$$
$$=u(s,\xi(s))+h(u_s(s,\xi(s)))+(u_x(s,\xi(s)),a(s,\xi(s)))$$
$$+\frac{1}{2}\mathrm{Tr}\,B(s,\xi(s))u_{xx}(s,\xi(s))+o(h)=u(s,\xi(s))+o(h)\,.$$

We have made use of Lemma 1.3.1 and the fact that $u(s,x)$ satisfies (1.3.5). The last relation easily leads to

$$\mathbf{E}(u(s+nh,\xi(s+nh))|\xi(s))=u(s,\xi(s))+no(h)\,.$$

If $s+nh=v$, then $h=(v-s)/n$, $n=(v-s)/h$ and

$$\mathbf{E}(u(v,\xi(v))|\xi(s))=u(s,\xi(s))+\frac{1}{h}o(h)\,.$$

The left-hand side does not depend on h. Letting $h\to 0$, we obtain

$$\mathbf{E}(u(v,\xi(v))|\xi(s))=u(s,\xi(s))=\int u(v,y)P(s,\xi(s),v,dy)\,.$$ □

2

Probability Representation of Solutions of Partial Differential Equations

Kolmogorov's equations for diffusion processes establish a connection between Markov processes and second-order partial differential operators with diffusion coefficients. If $a^k(t,x)$, $k=1,\dots,d$, are the coordinates of the transport vector in some basis and $b^{ik}(t,x)$ are the elements of the matrix of the diffusion operator in the same basis, then the differential operator has the form

$$L_t u(t,x) = \frac{1}{2}\sum_{i,k=1}^{d} b^{ik}(t,x)\frac{\partial^2 u}{\partial x^i \partial x^k}(t,x) + \sum_{k=1}^{d} a^k(t,x)\frac{\partial u}{\partial x_k}(t,x) \qquad (2.0.1)$$

(the x^k's are the coordinates of x in the indicated basis). It turns out that if one is able to construct a diffusion process with the diffusion coefficients occurring on the right-hand side of (2.0.1) (or in other words, a family of measures $\mathbf{P}_{s,x}$), then the solution of many problems involving the differential operator L may be written in the form

$$\mathbf{E}_{t,x} F(t, \xi_t(\cdot)) \ .$$

For fixed t, $F(t,x(\cdot))$ is a function defined on $\mathbf{C}_{R^d}^{[t,\infty)}$, the space of continuous functions $x(\cdot)$ on $[t,\infty)$ taking values in R^d. This is the probability representation of a solution. Such a representation and the properties of a stochastic process can be used to obtain results about the solutions. In addition, probability theory has procedures for constructing diffusion processes. Therefore the formulas of this kind can be used to compute the values of the solution itself.

2.1 Problems for a Parabolic Equation

In this section, L_t is the differential operator given by (2.0.1). The subscript t tells us that the coefficients of the operator are time-dependent. The functions $b^{ik}(t,x)$ and $a^k(t,x)$ will be assumed to be bounded and sufficiently smooth so that solutions of the considered problems exist and are unique.

2.1.1 Cauchy Problem

(a) *Backward Cauchy problem for a nonhomogeneous equation.* Let $T > 0$. Consider the solution of the problem

$$\begin{gathered}\frac{\partial}{\partial t}u(t,x) + L_t u(t,x) = f(t,x),\ t < T,\ x \in R^d\ , \\ u(T,x) = \varphi(x)\ .\end{gathered} \tag{2.1.1}$$

The boundary condition is given on the hyperplane $t = T$. The functions $f(t,x)$ and $\varphi(x)$ are bounded and continuous.

Theorem 2.1.1. *Let $\xi_t(\cdot)$ be a Markov diffusion process having diffusion coefficients $a(t,x)$ and $B(t,x)$ with coordinates a^k and matrix (b^{ik}) in the standard basis in R^d. If $\mathbf{P}_{s,x}$ denotes the probability corresponding to the process, then*

$$u(t,x) = \mathbf{E}_{t,x}\varphi(\xi_t(T)) - \mathbf{E}_{t,x}\int_t^T f(s,\xi_t(s))ds\ . \tag{2.1.2}$$

Proof. Let us show that the numerical process

$$\zeta(t) = u(t,\xi_s(t)) - \int_s^t f(u,\xi_s(u))du$$

is a martingale with respect to measure $\mathbf{P}_{s,x}$ for all $0 \le s < t < T$ and $x \in X$. If $t + h \le T$, we have (in the notation of the preceding chapter)

$$\begin{aligned}&\mathbf{E}_{s,x}(\zeta(t+h) - \zeta(t)|\mathcal{F}_t^s) \\ &\quad = \mathbf{E}_{s,x}(u(t+h,\xi_s(t+h)) - u(t,\xi_s(t))|\mathcal{F}_t^s) \\ &\qquad - E_{s,x}\left(\int_t^{t+h} f\left(u,\xi_s(u)\right)du|\mathcal{F}_t^s\right) \\ &\quad = \mathbf{E}_{s,x}\Bigg(u_t(t,\xi_s(t))h + (u_x(t,\xi_s(t)),a(t,\xi_s(t)))h \\ &\qquad + \frac{1}{2}\mathrm{Tr}\,u_{xx}(t,\xi_s(t))B(t,\xi_s(t))h - hf(t,\xi_s(t))|\mathcal{F}_t^s\Bigg) + o(h) = o(h)\ .\end{aligned}$$

We have made use of (2.1.1) and Lemma 1.3.1 on p. 163.

If $0 \le s < t < u \le T$ and $h = (u-t)/n$, then summing the equalities

$$\mathbf{E}_{s,x}(\zeta(t+kh) - \zeta(t+(k-1)h)|\mathcal{F}^s_{t+(k-1)h}) = o(h)\ ,\quad k = 1,\dots,n\ ,$$

and then taking the conditional expectation given $\mathcal{F}_t^s$, we obtain

$$\mathbf{E}_{s,x}(\zeta(u) - \zeta(t)|\mathcal{F}_t^s) = no(h) = \frac{u-t}{h}o(h)\ .$$

After letting $h \to 0$, we observe that $\zeta(t)$ is a martingale. Formula (2.1.2) follows from the relation $\mathbf{E}_{s,x}\zeta(T) = \mathbf{E}_{s,x}\zeta(s) = u(s,x)$. □

(b) *Forward Cauchy problem for a nonhomgeneous equation.* Let $T > 0$. We consider solving the problem

$$\begin{gathered}\frac{\partial}{\partial t}u(t,x) - L_t u(t,x) = f(t,x)\ , \quad 0 \le t \le T\ , \quad x \in R^d\ , \\ u(0,x) = \varphi(x)\ .\end{gathered} \tag{2.1.3}$$

The function $v_T(t,x) = u(T-t,x)$ satisfies the equation

$$\begin{aligned}&\frac{\partial v_T(t,x)}{\partial t} + \sum_{k=1}^{d} a^k(T-t,x)\frac{\partial v_T(t,x)}{\partial x^k} \\ &\quad + \frac{1}{2}\sum_{i,k=1}^{d} b^{ik}(T-t,x)\frac{\partial^2 v_T(t,x)}{\partial x^i \partial x^k} = -f(T-t,x)\ .\end{aligned} \tag{2.1.4}$$

Let $\xi^T(t)$ be a diffusion process defined on $[0,T]$ with diffusion coefficients occurring in (2.1.4). Then applying Theorem 2.1.1, we find that

$$\begin{aligned}u(t,x) = v_T(T-t,x) &= \mathbf{E}^T_{T-t,x}\varphi(\xi^T_{T-t}(T)) \\ &\quad + \mathbf{E}^T_{T-t,x}\int_{T-t}^{T} f(T-u,\xi^T_{T-t}(u))du\ .\end{aligned} \tag{2.1.5}$$

Here $\mathbf{E}^T_{s,x}$ is the measure corresponding to the process $\xi^T_s(t)$ that starts at time s in state x.

Remark. Let $L_t = L$ so that the coefficients of L no longer depend on t. Then the distribution of $\xi^T_s(s+t)$ depends on neither s nor T. The solution to problem (2.1.3) can be expressed for all T by means of the following formula:

$$u(t,x) = \mathbf{E}_x\varphi(\xi(t)) = \mathbf{E}_x\int_0^t f(u,\xi(u))du\ , \tag{2.1.6}$$

in which $\xi(u)$ is a homogeneous Markov diffusion process with diffusion coefficients $a(x)$ and $B(x)$ of the operator L.

2.1.2 Kac's Formula

Let $v(t,x)$ be a bounded continuous function. Consider the backward Cauchy problem

$$\begin{gathered}\frac{\partial}{\partial t}u(t,x) + L_t u(t,x) + v(t,x)u(t,x) = 0\ , \quad t \in [0,T]\ , \\ u(T,x) = \varphi(x)\ ;\end{gathered} \tag{2.1.7}$$

the function φ is bounded and continuous.

Theorem 2.1.2. *Let $\xi_s(t)$ be the same Markov diffusion process as in Theorem* 2.1.1. *Then*

$$u(t,x) = \mathbf{E}_{t,x}\varphi(\xi_t(T)) \exp\left\{\int_t^T v(u,\xi_t(u))du\right\} . \tag{2.1.8}$$

This result is known as *Kac's formula*.

Proof. Let $0 \le s < t \le T$. Consider the process

$$\zeta(t) = u(t,\xi_s(t)) \exp\left\{\int_s^t v(u,\xi_s(u))du\right\} . \tag{2.1.9}$$

Let us show that it is a martingale under the measure $\mathbf{P}_{s,x}$. As in Theorem 2.1.1, it suffices to demonstrate that

$$\mathbf{E}_{s,x}(\zeta(t+h) - \zeta(t)|\mathcal{F}_t^s) = o(h) .$$

We have

$$\begin{aligned}
\mathbf{E}_{s,x}(\zeta(t+h) - \zeta(t)|\mathcal{F}_t^s) &= \exp\left\{\int_s^t v(u,\xi_s(u))du\right\} \\
&\times \mathbf{E}_{s,x}\left(u(t+h,\xi_s(t+h))e^{\int_t^{t+h} v(u,\xi_s(u))du} - u(t,\xi_s(t))|\mathcal{F}_t^s\right) \\
&= \exp\left\{\int_s^t v(u,\xi_s(u))du\right\} \mathbf{E}_{s,x}(u(t+h,\xi_s(t+h)) \\
&+ hu(t+h,\xi_s(t+h))v(t,\xi_s(t)) - u(t,\xi_s(t))|\mathcal{F}_t^s) + o(h) \\
&= \exp\left\{\int_s^t v(u,\xi_s(u))du\right\} [u_t((t,\xi_s(t)) + (u_x(t,\xi_s(t)), a(t,\xi_s(t))) \\
&+ \frac{1}{2}\operatorname{Tr} u_{xx}(t,\xi_s(t))B(t,\xi_s(t)) + u(t,\xi_s(t))v(t,\xi_s(t))]h \\
&+ o(h) = o(h)
\end{aligned}$$

(we have again applied Lemma 2.1.1 on p. 163 and (2.1.7)). Hence, $\zeta(t)$ is a martingale and

$$\mathbf{E}_{s,x}\zeta(T) = \mathbf{E}_{s,x}\zeta(s) .$$

But $\zeta(s) = u(s,\xi_s(s)) = u(s,x)$ with probability $\mathbf{P}_{s,x} = 1$. From this we obtain (2.1.8) with s instead of t. □

Remark 2.1.1. Let $f(t,x)$ be a bounded continuous function. Then the solution to the problem

$$\begin{aligned}
&\frac{\partial}{\partial t}u(t,x) + L_t u(t,x) + v(t,x)u(t,x) = f(t,x), \quad t \in [0,T], \\
&u(T,x) = \varphi(x) ,
\end{aligned} \tag{2.1.10}$$

is expressible as

$$u(t,x) = \mathbf{E}_{t,x}\varphi(\xi_t(T))\exp\left\{\int_t^T v(u,\xi_t(u))du\right\}$$
$$- \mathbf{E}_{t,x}\int_t^T f(s,\xi_t(s))\exp\left\{\int_t^s v(u,\xi_t(u))du\right\}ds\ . \quad (2.1.11)$$

The proof of this formula amounts to showing for $0 \le s < t \le T$ that the process

$$\zeta(t) = u(t,\xi_s(t))\exp\left\{\int_s^t v(u,\xi_s(u))du\right\}$$
$$- \int_s^t f(u,\xi_s(u))\exp\left\{\int_s^u v(t,\xi_s(\tau))d\tau\right\}du \quad (2.1.12)$$

is a martingale. The proof of this is exactly the same as in Theorems 2.1.1 and 2.1.2.

Remark 2.1.2. Let $L_t = L$ (that is, the coefficients are independent of t). Then the solution to the problem

$$\frac{\partial u(t,x)}{\partial t} - Lu(t,x) - v(t,x)u(t,x) = f(t,x),$$
$$u(0,x) = \varphi(x)\ ,$$

is expressible as

$$u(t,x) = \mathbf{E}_x\varphi(\xi(t))\exp\left\{\int_0^t v(s,\xi(s))ds\right\}$$
$$+ \mathbf{E}_x\int_0^t f(s,\xi(s))\exp\left\{\int_0^s v(u,\xi(u))du\right\}ds\ , \quad (2.1.13)$$

where $\xi(u)$ is a homogeneous Markov diffusion process like the one in (2.1.6). Formula (2.1.13) is derived from (2.1.12) just as (2.1.6) was derived from (2.1.5).

2.1.3 Mixed Backward Problem for a Parabolic Equation

Let $T > 0$. Let $G(t,x)$ be a sufficiently smooth function defined an $[0,T]\times R^d$ for which the equation $G(t,x) = 0$ defines a smooth surface in $[0,T]\times R^d$. Its sections S_t by hyperplanes perpendicular to the t-axis are smooth surfaces in R^d each being the boundary of a bounded simply-connected region V_t, Let $V = \{(t,x) : t \in [0,T], x \in V_t\}$. We shall examine the following boundary-value problem in V:

$$\frac{\partial u(t,x)}{\partial t} + L_t u(t,x) = f(t,x), \quad (2.1.14)$$
$$u(t,x)\Big|_{x\in S_t} = \psi(t,x), \quad 0 \le t < T, \quad u(T,x) = \varphi(x)\ .$$

The function $\psi(t,x)$ is sufficiently smooth, $\varphi(x)$ and $f(t,x)$ are continuous, and $\psi(t,x)=\varphi(x)$ for $x\in S_T$.

Let V' be the boundary of V (it comprises the surfaces $\bigcup_t\{t\}\times S_t$ and V_T). Let $\tau_{V'}$ denote the *first exit time* of the process $\xi_s(t)$ from the region V. $\tau_{V'}=T$ if $\xi_s(t)\in V_t$ for $t\in[s,T]$. If $G(t,x)>0$ for $(t,x)\in V$ and if $\tau_{V'}<T$, then $\tau_{V'}$ is the first time for which $G(t,\xi_s(t))=0$. Clearly, $\tau_{V'}$ is a stopping time with respect to the flow $\mathcal{F}_t^s$ under the probability measure $\mathbf{P}_{s,x}$ for any $x\in X$.

Theorem 2.1.3. *Let $u(t,x)$ be the solution to problem* (2.1.14). *Then*

$$u(t,x)=\mathbf{E}_{t,x}\psi(\tau_{V'},\xi_t(\tau_{V'}))I_{\{\tau_{V'}<T\}}+\mathbf{E}_{t,x}\varphi(\xi_t(T))I_{\{\tau_{V'}=T\}}-\mathbf{E}_{t,x}\int_t^{\tau_{V'}}f(u,\xi_t(u))du\ . \quad (2.1.15)$$

Proof. Under the assumptions made, $u(t,x)$ can be extended to $[0,T]\times R^d$ with bounded continuous derivatives $\partial u/\partial t, \partial u/\partial x, \partial^2 u/\partial x^2$ (we are using the same symbol for the extension). The extended function will now satisfy (2.1.1) in $[0,T]\times R^d$; in $[0,T]\times R^d\times V$, the function $f(t,x)$ is simply set equal to the left-hand side of (2.1.1) in which the chosen extension of $u(t,x)$ has been substituted. As was shown in Theorem 2.1.1, the process

$$\zeta(t)=u(t,\xi_s(t))-\int_s^t f(u,\xi_s(u))du$$

is a martingale. Since $s\le\tau_{V'}$, it follows with probability $\mathbf{P}_{s,x}=1$ that

$$\mathbf{E}_{s,x}\zeta(\tau_{V'})=\mathbf{E}_{s,x}\zeta(s)\ .$$

But $\zeta(s)=u(s,x)$ with probability $\mathbf{P}_{s,x}=1$ and so

$$u(s,x)=\mathbf{E}_{s,x}\zeta(s)\ . \quad (2.1.16)$$

We have

$$\zeta(\tau_{V'})=u(\tau_{V'},\xi_s(\tau_{V'}))-\int_s^{\tau_{V'}}f(u,\xi_s(u))du \quad (2.1.17)$$

in which $f(u,\xi_s(u))$ has the same values as f on V, namely, those values of f that occur an the right-hand side of (2.1.14). Furthermore if $\tau_{V'}<T$, then $u(\tau_{V'},\xi_s(\tau_{V'}))=\psi(\tau_{V'},\xi_s(\tau_{V'}))$. If $\tau_{V'}=T$, then $u(\tau_{V'},\xi_s(\tau_{V'}))=\varphi(\xi_s(T))$. On substituting these values in (2.1.17), we can obtain (2.1.15) from (2.1.16). □

The next statement can be proved similarly to Theorem (2.1.3).

Theorem 2.1.4. *Suppose that the hypotheses of Theorem* (2.1.3) *hold and that the function $v(t,x)$ is continuous together with its derivatives $\partial v/\partial t$, $\partial v/\partial x$ and $\partial^2 v/\partial x^2$ in the closure of V. If $u(t,x)$ satisfies the equation*

$$\frac{\partial u(t,x)}{\partial t} + L_t(u) + v(t,x)u(t,x) = f(t,x) \tag{2.1.18}$$

in V and the boundary conditions $u(t,x)\,\big|_{x\in S_t} = \psi(t,x)$, $0 \le t < T$, and $u(T,x) = \varphi(x)$, $x \in V_T$, then $u(t,x)$ is expressible as

$$\begin{aligned} u(t,x) = {} & \mathbf{E}_{t,x}\psi(\tau_{V'}, \xi_s(\tau_{V'}))\exp\left\{\int_t^{\tau_{V'}} v(u,\xi(u))du\right\} I_{\{\tau_{V'}<T\}} \\ & + \mathbf{E}_{t,x}\varphi(\xi_t(T))\exp\left\{\int_t^{T} v(u,\xi_t(u))du\right\} I_{\{\tau_{V'}=T\}} \\ & - \mathbf{E}_{t,x}\int_t^{\tau_{V'}} f(u,\xi_t(u))\exp\left\{\int_t^{u} v(\tau,\xi_t(\tau))d\tau\right\} du\ . \end{aligned} \tag{2.1.19}$$

2.2 Boundary-Value Problems for Elliptic Operators

Let $a(x)$ and $B(x)$ be bounded and sufficiently smooth functions with values in R^d and $L(R^d)$, respectively; $a^k(x), k = 1,2,\dots,d$, are the coordinates of $a(x)$ in the standard basis and $b^{ik}(x)$ are the elements of the matrix of the operator $B(x)$ in this basis. A Markov diffusion process is assumed to exist having these diffusion coefficients. Let G be a bounded region in R^d with smooth boundary G'. We shall study the differential operator

$$Lu = \frac{1}{2}\sum_{i,k=1}^{d} b^{ik}(x)\frac{\partial^2 u}{\partial x^i \partial x^k}(x) + \sum_{k=1}^{d} a^k(x)\frac{\partial u}{\partial x^k}(x)\ . \tag{2.2.1}$$

The fact that the $b^{ik}(x)$ are diffusion coefficients means that the matrix $\big(b^{ik}(x)\big)$ is nonnegative-definite. We shall assume that the matrix is positive-definite. Therefore L is an elliptic operator.

2.2.1 Exit Times from a Bounded Region

Let $\xi(t)$ be a homogeneous Markov process with diffusion coefficients $a(x)$ and $B(x)$. Let τ_G be the first exit time from the bounded region G:

$$\tau_G = \inf\{s : \rho(\xi(s), X \setminus G) = 0\}\ .$$

Here $\rho(x,F)$ is the distance from point x to the set F. If $\rho(\xi(s), X\setminus G) > 0$ for all s, then we take $\tau_G = \infty$. Our aim in this subsection is to show that $\mathbf{E}_x\tau_G < \infty$ and hence that

$$\mathbf{P}_x\{\tau_G < \infty\} = 1\ .$$

We first prove a statement which is of interest in its own right.

Lemma 2.2.1. *Let $f(x)$ be a twice continuously differentiable function with compact support. Then the process*

$$\zeta(t) = f(\xi(t)) - \int_0^t Lf(\xi(s))ds$$

is a martingale with respect to the flow $\mathcal{F}_t^0$ and measure $\mathbf{P}_x$ for any value of x.

Proof. We have

$$\mathbf{E}_x(\zeta(t+h) - \zeta(t)|\mathcal{F}_t^0) = \mathbf{E}_{\xi(t)}f(\xi(h)) - f(\xi(t)) - \int_0^h \mathbf{E}_{\xi(t)}Lf(\xi(s))ds\ .$$

By Lemma 1.3.1 on p. 163, it follows that

$$\mathbf{E}_x f(\xi(h)) = hLf(x) + h\alpha_h(x)\ ,$$

where $\alpha_h(x)$ is bounded and approaches 0 as $h \to 0$ uniformly in x. Next

$$\int_0^h \mathbf{E}_x Lf(\xi(s)ds = hLf(x) + \int_0^h [\mathbf{E}_x Lf(\xi(s)) - Lf(x)]ds\ .$$

It is easy to see that $\sup_x |\mathbf{E}_x g(\xi(s)) - g(x)| \to 0$ as $s \to 0$ for every continuous g having compact support. Therefore

$$\lim_{h\to 0} \frac{1}{h} \int_0^h [\mathbf{E}_x Lf(\xi(s)) - Lf(x)]ds = 0$$

uniformly in x. Hence

$$\mathbf{E}_x(\zeta(t+h) - \zeta(t)|\mathcal{F}_t^0) = o(h)\ , \tag{2.2.2}$$

in which $o(h)$ depends on $\xi(t)$ but $o(h)/h \to 0$ uniformly in $\xi(t)$. Just as in Theorem 2.1.1 of Sect. 2.1.1, relation (2.2.2) implies the statement in the lemma. □

Since G is bounded, it is possible to find a ball S such that $G \subset S$. Then for $x \in G$

$$\mathbf{P}_x\{\tau_G < \tau_S\} = 1\ .$$

Thus to study the boundedness of τ_G, it suffices to concentrate on the case $G = S$.

Theorem 2.2.1. $\mathbf{E}_x\tau_S$ *is bounded uniformly in $x \in S$.*

Proof. There is no loss of generality in assuming that $S = \{y : |y| < r\}$. Let $c \in R^d$ with $|c| = 1$ and let

$$\varphi(x) = \exp\{k(c,x)\} \ .$$

Then

$$L\varphi(x) = \exp\{k(c,x)\}\left[k(c,a(x)) + \frac{k^2}{2}(B(x)c,c)\right] \ .$$

Write $\alpha = \sup_{x\in S}|a(x)|$ and $\beta = \inf_{x\in S}(B(x)c,c)$. By the assumed positive-definiteness of $B(x)$, β is positive. Therefore

$$L\varphi(x) \geq \exp\{-kr\}\left[-\alpha k + \frac{k^2}{2}\beta\right]$$

and choosing $k = 3\alpha/\beta$, we find for $x \in S$ that

$$L\varphi(x) \geq \exp\left\{-\frac{3\alpha r}{\beta}\right\}\frac{3\alpha^2}{2\beta} = c_1 > 0 \ .$$

Let $s < \tau_S$. Then $\xi(s) \in S$ and $L\varphi(\xi(s)) > c_1$. Noting that $\varphi(\xi(t)) - \int_0^t L\varphi(\xi(s))ds$ is a martingale, we obtain

$$\mathbf{E}_x\left[\varphi(\xi(\tau_S \wedge t)) - \int_0^{\tau_S\wedge t} L\varphi(\xi(s))ds\right] = \varphi(x) \ ,$$

or

$$\mathbf{E}_x\int_0^{\tau_S\wedge t} L\varphi(\xi(s))ds = \mathbf{E}_x\varphi(\xi(\tau_S \wedge t)) - \varphi(x) \ ,$$

and so

$$c_1\mathbf{E}_x\tau_S \wedge t \leq \sup_{x\in S} e^{k(c,x)} - \inf_{x\in S} e^{k(c,x)} < e^{kr} \ .$$

This is a uniform upper bound for $\mathbf{E}_x\tau_S \wedge t$. Letting $t \to \infty$, we observe that a bound also holds for τ_S:

$$\mathbf{E}_x\tau_S \leq \frac{2\beta}{3\alpha^2}\exp\left\{\frac{6\alpha r}{\beta}\right\} \ .$$

□

2.2.2 Solution of the Interior Boundary-Value Problem

We first concentrate on solving the *Dirichlet problem* for the equation

$$Lu(x) = 0, \quad x \in G \ , \tag{2.2.3}$$

with $u(x) = \varphi(x)$ on the boundary G' of G. Since $u(t,x) = u(x)$ is simultaneously a solution to the problem

$$\frac{\partial u}{\partial t} - Lu = 0, \quad u(t,x) = \varphi(x) \quad \text{on} \quad G', \quad u(T,x) = u(x)$$

in the region $[0,T] \times G$, Theorem 2.1.3 on p. 170 gives

$$u(x) = \mathbf{E}_x \varphi(\xi(\tau_G)) I_{\{\tau_G < T\}} + \mathbf{E}_x u(\xi(T)) I_{\{\tau_G \geq T\}} .$$

Letting $T \to \infty$ in this equation, we obtain

$$u(x) = \mathbf{E}_x \varphi(\xi(\tau_G)) .$$

A similar approach may be used to solve the more complicated nonhomogenous boundary-value problem for L.

Theorem 2.2.2. *Let $V(x) \leq 0$ and $f(x)$ be bounded continuous functions and let $u(x)$ be a solution of the equation*

$$Lu(x) + V(x)u(x) = f(x) \tag{2.2.4}$$

with the boundary condition

$$u(x) = \varphi(x) \ \text{on} \ G' .$$

Then

$$u(x) = \mathbf{E}_x \left[\varphi(\xi(\tau_G)) \exp\left\{ \int_0^{\tau_G} V(\xi(s)) ds \right\} - \int_0^{\tau_G} f(\xi(s)) \exp\left\{ \int_0^s V(\xi(u)) du \right\} ds \right] . \tag{2.2.5}$$

Proof. The proof of Theorem 2.1.2, p. 168 (see also Remark 2.1.1 to that theorem) established that the process

$$\zeta(t) = u(\xi(t)) \exp\left\{ \int_0^t V(\xi(u)) du \right\} - \int_0^t f(\xi(s)) \exp\left\{ \int_0^t V(\xi(u)) du \right\} ds$$

is a martingale. Therefore for all positive t,

$$\mathbf{E}_x \zeta(t \wedge \tau_G) = \mathbf{E}_x(\zeta(0)) = u(x) ,$$

$$u(x) = \mathbf{E}_x u(\xi(t \wedge \tau_G)) \exp\left\{ \int_0^{t \wedge \tau_G} V(\xi(s)) ds \right\} - \mathbf{E}_x \int_0^{t \wedge \tau_G} f(\xi(s)) \exp\left\{ \int_0^s V(\xi(u)) du \right\} ds . \tag{2.2.6}$$

Formula (2.2.5) is derived from (2.2.6) by letting $t \to \infty$. It is permissible to take the limit under the expectation sign because $V \leq 0$ and so

$$\exp\left\{\int_0^t V(\xi(s))ds\right\} \leq 1 .$$

Hence, for $t < \tau_G$

$$\left|u(\xi(t))\exp\left\{\int_0^t V(\xi(s))ds\right\} - \int_0^t f(\xi(s))\exp\left\{\int_0^s V(\xi(u))du\right\}ds\right| \leq c_1 + c_2\tau_G,$$

where $c_1 = \sup_{x\in G} u(x)$, $c_2 = \sup_{x\in G} f(x)$ and $\mathbf{E}\tau_G < \infty$ by virtue of Theorem 2.2.1 of Sect. 2.2.1. □

(a) *Equation with V of alternating sign.* If $V(x)$ changes sign in equation (2.2.4) (perhaps it is simply positive), formula (2.2.5) need not hold. For example, the equation (2.2.4) with $f = 0$ and $\varphi = 0$ can have a non-zero solution but the right-hand side of (2.2.5) vanishes. This happens because the passage to the limit in (2.2.6) is unjustified. To justify it, we shall need the finiteness of $\mathbf{E}_x e^{\lambda\tau_G}$ for $\lambda > 0$.

Lemma 2.2.2. *If $\sup_{x\in G} \mathbf{E}_x\tau_G \leq q$, then $\mathbf{E}_x e^{\lambda\tau_G} < \infty$ for $\lambda < (eq)^{-1}$.*

Proof. For $c > q$ and all $x \in G$,

$$\mathbf{P}_x\{\tau_G > c\} \leq \frac{q}{c} .$$

Observe that

$$\begin{aligned}\mathbf{P}_x\{\tau_G > nc\} &= \mathbf{E}_x I_{\{\tau_G > nc\}} = \mathbf{E}_x I_{\{\tau_G > (n-1)c\}}\mathbf{E}_x(I_{\{\tau_G > nc\}}|\mathcal{F}^0_{(n-1)c}\}\\ &= \mathbf{E}_x I_{\{\tau_G > (n-1)c\}}\mathbf{E}_{\xi((n-1)c)} I_{\{\tau_G > c\}}\\ &\leq \frac{q}{c}\mathbf{P}_x\{\tau_G > (n-1)c\} \leq \left(\frac{q}{c}\right)^n .\end{aligned}$$

Therefore

$$\mathbf{E}_x e^{\lambda\tau_G} \leq \sum_{n=1}^{\infty} e^{\lambda nc}\mathbf{P}_x\{\tau_G \geq (n-1)c\} \leq e^{\lambda c}\sum_{n=0}^{\infty}\left(e^{\lambda c}\cdot\frac{q}{c}\right)^n$$

and the series on the right converges when $\lambda < c^{-1}\ln(c/q)$. The right-hand side has a maximum when $c = qe$. □

Lemma 2.2.3. *Suppose that a strictly positive solution exists to the equation*

$$Lu(x) + \lambda u(x) = 0, \quad x \in G, \quad \inf_{x\in G} u(x) = c > 0 ,$$

for some $\lambda > 0$. Then $\mathbf{E}_x e^{\lambda\tau_G} < \infty$.

Proof. Formula (2.2.6) can be applied with $V = \lambda$ and $f = 0$. We obtain

$$u(x) = \mathbf{E}_x u(\xi(t \wedge \tau_G)) \exp\{\lambda(t \wedge \tau_G)\} .$$

Therefore

$$\mathbf{E}_x \exp\{\lambda(t \wedge \tau_G)\} \leq c^{-1} u(x) .$$

On letting $t \to \infty$, we complete the proof of the lemma. □

Theorem 2.2.3. *Suppose that* $\mathbf{E}_x e^{\lambda \tau_G} < \infty$ *for some* $\lambda > 0$. *Then if* $V(x) \leq \lambda$, *the solution to equation* (2.2.4) *with boundary condition* $u(x) = \varphi(x)$ *on* G', *is representable by formula* (2.2.5).

Proof. Under the assumption made,

$$\exp\left\{\int_0^s V(\xi(s))ds\right\} \leq e^{\lambda s}$$

and

$$\int_0^{t \wedge \tau_G} \exp\left\{\int_0^s V(\xi(u))du\right\} ds \leq \frac{\exp\{\lambda t \wedge \tau_G\} - 1}{\lambda}$$

for $s \leq \tau_G$. Therefore

$$\left| u(\xi(t \wedge \tau_G)) \exp\left\{\int_0^{t \wedge \tau_G} V(\xi(u))du\right\} \right.$$
$$\left. - \int_0^{t \wedge \tau_G} f(\xi(s)) \exp\left\{\int_0^s V(\xi(u))du\right\} ds \right| \leq c_1 + c_2 e^{\lambda \tau_G} ,$$

where c_1 and c_2 are constants. The passage to the limit in (2.2.6). under the expectation sign is thus permissible. □

2.3 Wiener Measure and the Solution of Equations Involving the Laplace Operator

2.3.1 Wiener Process in R^d

A *Wiener process* in R^d is a homogeneous process $w(t)$ with independent increments for which $w(t+h) - w(t)$ is normally distributed with mean zero and covariance operator hI (I is the identity operator in R^d). In other words, the increment has the density

$$g_h(x) = (2\pi h)^{-d/2} \exp\left\{-\frac{1}{2h}|x|^2\right\} . \tag{2.3.1}$$

(a) *A Markov process related to a Wiener process.* Since a Wiener process has independent increments, then for $0 \le t_1 < \ldots < t_n \le t < t+h$,

$$\begin{aligned}
&\mathbf{P}\{w(t+h) \in A | w(t_1), \ldots, w(t_n), w(t)\} \\
&= \mathbf{P}\{w(t+h) - w(t) \in A - w(t) | w(t_1), \ldots, w(t_n), w(t)\} \\
&= \mathbf{P}\{w(t+h) - w(t) \in A - w(t) | w(t)\} \\
&= \mathbf{P}(w(h) - w(0) \in A - x\} \Big|_{x=w(t)} ;
\end{aligned}$$

in this, $A - x = \{y : y + x \in A\}$ and we have used not only the independence of the increments but also the homogeneity. Putting

$$\begin{aligned}
P^w(h, x, A) &= \mathbf{P}\{w(h) - w(0) \in A - x\} \\
&= (2\pi h)^{-d/2} \int_A \exp\left\{-\frac{1}{2h}|y-x|^2\right\} ds, \qquad (2.3.2)
\end{aligned}$$

one can verify that

$$\mathbf{P}\{w(t+h) \in A | w(t_1), \ldots, w(t_n), w(t)\} = P^w(h, w(t), A) .$$

This means that the family of processes $w(t) - w(0) + x$ may be viewed as a homogeneous Markov process with transition probability (2.3.2). This process has spatial homogeneity: if it starts at x, the process can be determined from a process starting at 0 by shifting the latter's path by amount x. In what follows, we take $w(0) = 0$. Then $w(t)$ is the path of a Markov process starting at 0. If as previously, $\mathbf{P}_x$ is the distribution in $\mathbf{C}_{R^d}$ corresponding to the initial value x, then spatial homogenity means that

$$\mathbf{E}_x f(\xi(\cdot)) = \mathbf{E}_0 f(\xi(\cdot) + x)$$

for every bounded measurable function $f(x(\cdot))$ in $\mathbf{C}_{R^d}$ (here $\xi(\cdot)$ denotes the path of a process). If $f(x(\cdot))$ is measurable, then so is the function $f_a(x(\cdot)) = f(x(\cdot) + a)$. Thus, when a process is spatially homogeneous, it is possible to consider a single measure instead of a family of measures $\mathbf{P}_x$.

Noting the form of the transition probability, one can show that

$$\begin{aligned}
P^w(h, x, V_\varepsilon(x)) &= \mathbf{P}\{|w(h)| > \varepsilon\} \le \frac{1}{\varepsilon^m}\mathbf{E}|w(h)|^m \\
&= \frac{h^{m/2}}{\varepsilon^m}\mathbf{E}|w(1)|^m = o(h), \quad m > 2 , \qquad (2.3.3)
\end{aligned}$$

and

$$\begin{aligned}
&\int_{|y-x|\le\varepsilon} (y^i - x^i) P^w(h, x, dy) = \int_{|y|\le\varepsilon} y^i P^w(h, 0, dy) \\
&= \mathbf{E}w^i(h) I_{\{|w(h)|\le\varepsilon\}} = \mathbf{E}w^i(h) + \mathbf{E}w^i(h) I_{\{|w(h)|>\varepsilon\}} \\
&= O((\mathbf{E}|w(h)|^2)^{1/2} (\mathbf{P}\{|w(h)| > \varepsilon\})^{1/2}) = o(h) \qquad (2.3.4)
\end{aligned}$$

(we have made use of the estimate (2.3.3)). Finally, in similar fashion,

$$\int_{|y-x|\le\varepsilon}(y^i-x^i)(y^k-x^k)P^w(h,x,dy)=\mathbf{E}(w^i(h)w^k(h))+v(h)$$
$$=\delta_{ik}h+o(h)\ .$$

Thus, $w(t)$ is a diffusion process with constant diffusion coefficients: $a=0$ and $b^{ik}=\delta_{ik}$.

(b) *Wiener measure.* A Wiener process $w(t)$ (as stated above, we are assuming that $w(0)=0$) possesses one further remarkable property, that of self-similarity. This means that there exists a function ψ from R_+ to R_+ for which $w(\lambda t)$ has the same distribution as $\psi(\lambda)w(t)$ for every positive value of λ. Specifically, for the Wiener process, $\psi(\lambda)=\sqrt{\lambda}$. Indeed, $w(\lambda t)$ and $\sqrt{\lambda}w(t)$ are homogeneous processes with independent increments and both are normally distributed with mean 0 and covariance operator λI.

Let $\mathbf{P}_x^T$ be the measure corresponding to a Markov process on the space $\mathbf{C}_{R^d}[0,T]$ of functions defined on $[0,T]$ (it is determined by its values on the cylinder sets with bases in $[0,T]$; see Part I, pp. 47–48). The homogeneity and self-similarity imply that every bounded measurable function $f(x(\cdot))$ on $\mathbf{C}_{R^d}[0,T]$ satisfies the relation

$$\mathbf{E}_x^T f(\xi(\cdot))=\mathbf{E}_0^1 R_{T,x}f(\xi(\cdot))\ ,$$

where $\mathbf{E}_x^T$ is the expectation with respect to $\mathbf{P}_x^T$, $R_{T,x}f(x(\cdot))=f(R_{T,x}x(\cdot))$ and $R_{T,x}(t)=\sqrt{T}x(t/T)+x$ is a measurable mapping from $\mathbf{C}_{R^d}[0,1]$ to $\mathbf{C}_{R^d}[0,T]$. The measure $\mathbf{P}_0^1$ on $\mathbf{C}_{R^d}[0,1]$ is commonly called a *Wiener measure.* It will be denoted by μ_w and an integral with respect to it will be represented as $\int f d\mu_w$. The measure is determined by its integrals of cylindrical functions. If $\Phi(x_1,\dots,x_n)$ is a measurable numerical function on $(R^d)^n$ and

$$f_\Phi(t_1,\dots,t_n,x(\cdot))=\Phi(x(t_1),\dots,x(t_n))$$

is a cylindrical function with base $\{t_1,\dots,t_n\}$ $(0=t_0<t_1<\dots<t_n\le 1)$, then taking $x_0=0$, we have

$$\int f_\Phi(t_1,\dots,t_n,x(\cdot))d\mu_w$$
$$=\prod_{k=1}^n(2\pi(t_k-t_{k-1}))^{-\frac{d}{2}}\int\cdots\int\exp\left\{-\frac{1}{2}\sum_{k=1}^n\frac{|x_k-x_{k-1}|^2}{t_k-t_{k-1}}\right\}$$
$$\times\Phi(x_1,\dots,x_n)dx_1\dots dx_n\ . \qquad (2.3.5)$$

The approximation of continuous functions by cylindrical functions makes it possible to use (2.3.5) to evaluate integrals with respect to a Wiener measure by passing to the limit.

Lemma 2.3.1. *Let $f(x(\cdot))$ be a bounded continuous functional on $\mathbf{C}_{R^d}[0,1]$ and let $l(t_1,\dots,t_n,x_1,\dots,x_n,t)$ be a polygonal line in $[0,1]\times R^d$ with vertices (t_i,x_i), $0=t_0<t_1<\dots<t_n=1$ and $0=x_0,x_1,\dots,x_n\in R^d$. Then*

$$\int f(x(\cdot))d\mu_w = \lim_{\max \Delta t_k\to 0}\prod_{k=1}^{n}(2\pi(t_k-t_{k-1}))^{-d/2}$$
$$\times\int\cdots\int \exp\left\{-\frac{1}{2}\sum_{k=1}^{n}\frac{|x_k-x_{k-1}|^2}{t_k-t_{k-1}}\right\} f(l(t_1,\dots,t_n,x_1,\dots,x_n,\cdot))$$
$$\times dx_1\dots dx_n\ . \tag{2.3.6}$$

Proof. Appearing under the limit sign is

$$\int f(l(t_1,\dots,t_n,x(t_1),\dots,x(t_n),\cdot))d\mu_w$$

and $f(l(t_1,\dots,t_n,x(t_1),\dots,x(t_n),\cdot))\to f(x(\cdot))$ as $\max\Delta t_k\to 0$ for all $x(\cdot)\in \mathbf{C}_{R^d}[0,1]$. Therefore the lemma follows by virtue of Lebesgue's theorem. □

2.3.2 Stochastic Integral

Our further exposition requires an integral of the form

$$\int_0^1 \Big(f(w(t)),dw(t)\Big)\ ,$$

where $f(x)$ is a sufficiently smooth function from R^d to R^d. It is the special case of a *stochastic integral* with respect to a Wiener process. We shall concentrate on the case $d=1$ in detail.

Lemma 2.3.2. *Let $f(x)$ be a continuously differentiable function from R to R and let $f'(x)$ be bounded. If $w(t)$ is a homogeneous Wiener process, then*

$$\lim_{n\to\infty}\sum_{0\le k<2^n} f\left(w\left(\frac{k}{2^n}\right)\right)\left[w\left(\frac{k+1}{2^n}\right)-w\left(\frac{k}{2^n}\right)\right] \tag{2.3.7}$$

exists with probability 1.

Proof. Denote the pre-limiting quantity in (2.3.7) by $S_n(f)$. Then

$$
\begin{aligned}
&S_n(f) - S_{n+1}(f) \\
&= \sum_{0\le k<2^n} \left(f\left(w\left(\frac{k}{2^n}\right)\right)\left[w\left(\frac{k+1}{2^n}\right) - w\left(\frac{k}{2^n}\right)\right]\right. \\
&\qquad - f\left(w\left(\frac{k}{2^n}\right)\right)\left[w\left(\frac{2k+1}{2^{n+1}}\right) - w\left(\frac{k}{2^n}\right)\right] \\
&\qquad \left. - f\left(w\left(\frac{2k+1}{2^{n+1}}\right)\right)\left[w\left(\frac{k+1}{2^n}\right) - w\left(\frac{2k+1}{2^{n+1}}\right)\right]\right) \\
&= \sum_{0\le k<2^n} \left[f\left(w\left(\frac{k}{2^n}\right)\right) - f\left(w\left(\frac{2k+1}{2^{n+1}}\right)\right)\right]\left[w\left(\frac{k+1}{2^n}\right) - w\left(\frac{2k+1}{2^{n+1}}\right)\right] .
\end{aligned}
$$

Therefore

$$
\begin{aligned}
&\mathbf{E}|S_n(f) - S_{n+1}(f)|^2 \\
&= \sum_{0\le k<2^n} \mathbf{E}\left[f\left(w\left(\frac{k}{2^n}\right)\right) - f\left(w\left(\frac{2k+1}{2^{n+1}}\right)\right)\right]^2 \left[w\left(\frac{k+1}{2^n}\right) - w\left(\frac{2k+1}{2^{n+1}}\right)\right]^2 \\
&+ 2\sum_{0\le k<i<2^n} \mathbf{E}\left[f\left(w\left(\frac{k}{2^n}\right)\right) - f\left(w\left(\frac{2k+1}{2^{n+1}}\right)\right)\right]\left[w\left(\frac{k+1}{2^n}\right) - w\left(\frac{k+1}{2^{n+1}}\right)\right] \\
&\qquad \times \left[f\left(w\left(\frac{i}{2^n}\right) - f\left(w\left(\frac{2i+1}{2^{n+1}}\right)\right)\right]\left[w\left(\frac{i+1}{2^n}\right) - w\left(\frac{2i+1}{2^n}\right)\right] .
\end{aligned}
$$

The second sum vanishes since $w((i+1)/2^n) - w((2i+1)/2^{n+1})$ is independent of the remaining factors and $\mathbf{E}[w((i+1)/2^n) - w((2i+1)/2^{n+1})] = 0$. The factors in the first sum are independent and one term has the form

$$
\mathbf{E}\left[f\left(w\left(\frac{k}{2^n}\right)\right) - f\left(w\left(\frac{2k+1}{2^n}\right)\right)\right]^2 \cdot \frac{1}{2^{n+1}} .
$$

For some L, $|f(x) - f(y)| \le L|x-y|$. Thus

$$
\mathbf{E}|S_n(f) - S_{n+1}(f)|^2 \le L^2 \sum_{0\le k<2^n} \left(\frac{1}{2^{n+1}}\right)^2 \le c_1 \cdot 2^{-n} .
$$

Since

$$
\mathbf{P}\{|S_n(f) - S_{n+1}(f)| \ge 2^{-n/4}\} \le c_1 \cdot 2^{-n} \cdot 2^{n/2} = c_1 \cdot 2^{-n/2} ,
$$

the lemma follows by virtue of the Borel-Cantelli lemma. □

The limit (2.3.7) is denoted by

$$
\int_0^1 f(w(t))dw(t) .
$$

Remark 2.3.1. In similar fashion, it can be demonstrated that

$$\int_0^1 (f(w(t)), dw(t)) = \lim_{n\to\infty} \sum_{0\le k<2^n} \left(f\left(w\left(\frac{k}{2^n}\right)\right), w\left(\frac{k+1}{2^n}\right) - w\left(\frac{k}{2^n}\right)\right) \quad (2.3.8)$$

exists with probability 1. In this, $w(t)$ is a Wiener process in R^d, $f(x)$ is a continuously differentiable function from R^d to R^d and $f'(x)$ is bounded.

Remark 2.3.2. One can define the integral $\int_0^t (f(w(s)), dw(s))$ to be

$$\lim_{n\to\infty} \sum_{0\le k<2^n t} \left(f\left(w\left(\frac{k}{2^n}\right)\right), w\left(\frac{k+1}{2^n}\right) - w\left(\frac{k}{2^n}\right)\right) .$$

The proof that this limit exists with probability 1 is the same as that of Lemma 2.3.2. As usual, we define

$$\int_t^{t+h} = \int_0^{t+h} - \int_0^t .$$

Lemma 2.3.3. *If the function $f(x)$ is bounded, then the following relations hold:*

1. $\mathbf{E}\int_t^{t+h} (f(w(s)), dw(s)) = 0$;
2. $\mathbf{E}\left[\int_t^{t+h} (f(w(s)), dw(s))\right]^2 = \mathbf{E}\int_t^{t+h} |f(w(s))|^2 ds$;
3. $\mathbf{E}\left[\int_t^{t+h} (f(w(s)), dw(s))\right]^4 = O(h^2)$;
4. *For $z \in R^d$,*

$$\mathbf{E}\exp\left\{\int_t^{t+h} (f(w(s)), dw(s))\right\} (w(t+h) - w(t), z) = \mathbf{E}\int_t^{t+h} (f(w(s)), z) ds + o(h) ;$$

5. $\mathbf{E}\exp\left\{\int_t^{t+h} (f(w(s)), dw(s))\right\} = 1 + O(h)$.

Proof. We apply the limit process used to define a stochastic integral. The first three statements can be proved in an obvious way . If $|f| \le c$, by employing the inequality

$$1 \le \mathbf{E}\left(\exp\left\{ \left(f\left(w\left(\frac{k}{2^n}\right)\right), w\left(\frac{k+1}{2^n}\right) - w\left(\frac{k}{2^n}\right)\right)\right\} \,\middle|\, w\left(\frac{k}{2^n}\right)\right)$$
$$= \exp\left\{ \frac{1}{2^{n+1}} \left| f\left(w\left(\frac{k}{2^n}\right)\right)\right|^2 \right\} \le \exp\left\{ c^2 \frac{1}{2^{n+1}} \right\} ,$$

one can deduce that

$$1 \le \mathbf{E} \exp\left\{\lambda \int_t^{t+h} (f(w(s)), dw(s))\right\} \le \exp\{h\lambda^2 c^2\} .$$

This implies 5. Furthermore,

$$\begin{aligned}
&\mathbf{E} \exp\left\{\int_t^{t+h} (f(w(s)), dw(s))\right\} (w(t+h) - w(t), z) \\
&\quad = \mathbf{E}\left(1 + \int_t^{t+h} (f(w(s)), dw(s))\right) (w(t+h) - w(t), z) + \mathbf{E}\theta_h \\
&\quad = \mathbf{E} \int_t^{t+h} (f(w(s)), z) ds + \mathbf{E}\theta_h ,
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{E}|\theta_h| \le \mathbf{E} &\left(\int_t^{t+h} (f(w(s)), dw(s))\right)^2 \\
&\times \left(\exp\left\{\int_t^{t+h} (f(w(s)), dw(s))\right\} + \exp\left\{-\int_t^{t+h} (f(w(s)), dw(s))\right\}\right) \\
&\times |w(t+h) - w(t), z)| .
\end{aligned}$$

A double application of Cauchy's inequality and parts 3 and 5 show that $\mathbf{E}|\theta_h| = o(h)$. □

(a) *Martingales related to a Wiener process.*

Theorem 2.3.1. *Suppose that $f(x)$ is a bounded continuous function from R^d to R^d having bounded continuous derivatives $f'(x)$ and $f''(x)$. If $a(x)$ and $v(x)$ are bounded continuous functions from R^d to R^d and R^d to R respectively, with $a'(x)$ bounded, then the process*

$$\begin{aligned}
\zeta(t) = f(w(t) + x) \exp&\left\{\int_0^t (a(x + w(s)), dw(s)) \right.\\
&\left. + \int_0^t \left[v(x + w(s)) - \frac{1}{2}|a(x + w(s))|^2\right] ds\right\} \\
- \int_0^t L_{a,v} f(x + w(s)) \exp&\left\{\int_0^s (a(x + w(u)), dw(u)) \right.\\
&\left. + \int_0^s \left[v(x + w(u)) - \frac{1}{2}|a(x + w(u))|^2\right] du\right\} ds ,
\end{aligned}$$

where

$$L_{a,v} f = \frac{1}{2}\Delta f + (a(x), f'(x)) + v(x) f(x), \quad \Delta f = Tr\, f'' ,$$

is a martingale.

Proof. It suffices to show that $\mathbf{E}(\zeta(t+h)-\zeta(t)|\mathcal{F}_t^0) = o(h)$ where $\mathcal{F}_t^0$ is the σ-algebra generated by $w(s), s \le t$. It is easy to see that

$$\begin{aligned}
&\mathbf{E}\left(\int_t^{t+h} L_{a,v} f(x+w(s)) \exp\left\{\int_0^s (a(x+w(u)), dw(u)) \right.\right. \\
&\qquad \left.\left. + \int_0^s \left[v(x+w(u)) - \frac{1}{2}|a(x+w(u))|^2\right] du\right\} ds | \mathcal{F}_t^0\right) \\
&= hL_{a,v} f(x+w(t)) \exp\left\{\int_0^t (a(x+w(u)), dw(u)) \right. \\
&\qquad \left. + \int_0^t \left[v(x+w(u)) - \frac{1}{2}|a(x+w(u))|^2\right] du\right\} + o(h) \,.
\end{aligned}$$

Then

$$\begin{aligned}
&f(x+w(t+h)) \exp\left\{\int_0^{t+h} (a(x+w(s)), dw(s)) \right. \\
&\qquad \left. + \int_0^{t+h} \left[v(x+w(s)) - \frac{1}{2}|a(x+w(s))|^2\right] ds\right\} \\
&- f(x+w(t)) \exp\left\{\int_0^t (a(x+w(s)), dw(s)) \right. \\
&\qquad \left. + \int_0^t \left[v(x+w(s)) - \frac{1}{2}|a(x+w(s))|^2\right] ds\right\} \\
&= \exp\left\{\int_0^t (a(x+w(s)), dw(s)) + \int_0^t \left[v(x+w(s)) - \frac{1}{2}|a(x+w(s))|^2\right] ds\right\} \\
&\quad \times \left[f(x+w(t+h)) \exp\left\{\int_t^{t+h} (a(x+w(s)), dw(s))\right\}\right. \\
&\qquad \left. \times \left(1 + h\left[v(x+w(t)) - \frac{1}{2}|a(x+w(s))|^2\right] + o(h)\right) - f(x+w(t))\right] \,.
\end{aligned}$$

Therefore to complete the proof, it suffices to show that

$$\begin{aligned}
&\mathbf{E}\left(f(x+w(t+h)) \exp\left\{\int_t^{t+h} (a(x+w(s)), dw(s))\right\} - f(x+w(t)) \right. \\
&\quad \left. + hf(x+w(t))\left[v(x+w(t)) - \frac{1}{2}|a(x+w(t))|^2\right] - hL_{a,v} f(x+w(t)) \,\middle|\, \mathcal{F}_t^0\right) \\
&\qquad = o(h) \,. \qquad (2.3.9)
\end{aligned}$$

We have

$$
\begin{aligned}
&\mathbf{E}\left(f(x+w(t+h))\exp\left\{\int_t^{t+h}(a(x+w(s)),dw(s))\right\}-f(x+w(t))\,\Big|\,\mathcal{F}_t^0\right)\\
&=\mathbf{E}(f(x+w(t+h))-f(x+w(t))|\mathcal{F}_t^0)\\
&\quad+\mathbf{E}\left(f(x+w(t+h))-f(x+w(t))\left(\exp\left\{\int_t^{t+h}(a(x+w(s)),dw(s))\right\}-1\right)\Big|\mathcal{F}_t^0\right)\\
&\quad+\mathbf{E}\left(f(x+w(t))\left(\exp\left\{\int_t^{t+h}(a(x+w(s)),dw(s))\right\}-1\right)\,\Big|\,\mathcal{F}_t^0\right)\\
&=\frac{h}{2}\,Tr\,f''(x+w(t))+hf'(x+w(t))a(x+w(t))\\
&\quad+\frac{h}{2}f(x+w(t))|a(x+w(t))|^2+o(h)\,.
\end{aligned}
$$

We have applied the relation $\mathbf{E}f(x+w(h))=f(x)+\frac{h}{2}\,Tr\,f''(x)+o(h)$ as well as Lemma 2.3.3. Upon substituting this last expression into the left-hand side of (2.3.9) we obtain the required right-hand side. □

2.3.3 Representation of Solutions of Equations

(a) *Cauchy problem for a parabolic equation.*

Theorem 2.3.2. *Suppose that $u(t,x)$ is a solution to the equation*

$$\frac{\partial u(t,x)}{\partial t}=\frac{1}{2}\Delta u(t,x)+(a(x,u_x(t,x))+v(x)u(t,x)$$

with initial condition $u(0,x)=\varphi(x)$ where $a(x)$ is a bounded continuous function from R^d to R^d, $a'(x)$ is bounded, and $v(x)$ is a bounded continuous function from R^d to R. Then

$$
\begin{aligned}
u(t,x)=\mathbf{E}\varphi(x+w(t))\exp\Bigg\{&\int_0^t(a(x+w(s)),dw(s))\\
&+\int_0^t\left[v(x+w(s))-\frac{1}{2}|a(x+w(s))|^3\right]ds\Bigg\}\,. \qquad (2.3.10)
\end{aligned}
$$

Proof. The function

$$
\begin{aligned}
\zeta(s)=u(t-s,x+w(s))\exp&\int_0^s(a(x+w(u)),dw(u))\\
&+\int_0^s\left[v(x+w(u))-\frac{1}{2}|a(x+w(u))|^2\right)du\Bigg\}
\end{aligned}
$$

is a martingale for all positive t by virtue of Theorem 2.3.1. Since $\zeta(0)=u(t,x)$, the theorem follows from the relation $\mathbf{E}\zeta(t)=\mathbf{E}\zeta(0)$.

Remark. Formula (2.3.10) may be expressed in terms of an integral with respect to the measure μ_w:

$$u(t,x) = \int \varphi(x+\sqrt{t}x(1)) \exp\left\{\sqrt{t}\int_0^1 (a(x+\sqrt{t}x(s)), dx(s)) \right.$$
$$\left. + t\int_0^1 \left[v(x+\sqrt{t}x(s)) - \frac{1}{2}|a(x+\sqrt{t}x(s))|^2\right] ds\right\} d\mu_w \ .$$

(b) *Dirchlet problem for an elliptic equation.* Let G be a region in R^d with a simply-connected bounded smooth boundary G'. Consider the problem

$$\tfrac{1}{2}\Delta u(x) + (a(x), u_x(x)) + v(x)u(x) = f(x),$$
$$u(x) = \psi(x) \text{ on } G'$$

Theorem 2.3.3. *Define* $\tau_G(x(\cdot))$ *on* $\mathbf{C}_{R^d}$ *by*

$$\tau_G(x(\cdot)) = \sup\{t : x(s) \in G, s \le t\} \ .$$

If

$$\sup_{x\in G} \mathbf{E}_x \exp\{\lambda\tau_G\} < \infty$$

for some positive λ, *where* $\tau_G = \tau_G(x + w(\cdot))$, *then when*

$$\frac{1}{2}\sup_{x\in G}|a(x)|^2 + \sup_{x\in G}\left[v(x) - \frac{1}{2}|a(x)|^2\right] \le \lambda$$

the following representation holds:

$$u(x) = \mathbf{E}\exp\left\{\int_0^{\tau_G} (a(x+w(s)), dw(s)) \right.$$
$$\left. + \int_0^{\tau_G}\left[v(x+w(s)) - \frac{1}{2}|a(x+w(s))|^2\right] ds\right\} \psi(x+w(\tau_G)) \ .$$

Proof. Consider an extension of $u(x)$ which is twice continuously differentiable and has compact support. On the basis of Theorem 2.3.1, the process

$$\zeta(t) = u(x+w(t))\exp\left\{\int_0^t (a(x+w(s)), dw(s)) \right.$$
$$\left. + \int_0^t \left[v(x+w(s)) - \frac{1}{2}|a(x+w(s))|^2\right] ds\right\}$$
$$- \int_0^t g(x+w(s))\exp\left\{\int_0^s (a(x+w(u)), dw(u)) \right.$$
$$\left. + \int_0^t \left[v(x+w(u)) - \frac{1}{2}|a(x+w(u))|^2\right] du\right\} ds \ ,$$

where $g(x) = 0$ for $x \in G$, is a martingale. Therefore the process

$$\zeta_1(t) = \zeta(t \wedge \tau_G) = u(x + w(t \wedge \tau_G)) \exp\left\{ \int_0^{t \wedge \tau_G} (a(x + w(s)), dw(s)) \right.$$
$$\left. + \frac{1}{2} \int_0^{t \wedge \tau_G} \left[v(x + w(s)) - \frac{1}{2} |a(x + w(s))|^2 \right] ds \right\}$$

is also a martingale. Thus

$$u(x) = \mathbf{E}\zeta_1(0) = \mathbf{E}\zeta_1(t) \, .$$

Applying Lemma 2.3.3, we can show that it is permissible to let $t \to \infty$ under the expectation sign in the last relation. □

Remark. If $f(t, x)$ is continuous and bounded on $R_+ \times R^d$ and has bounded continuous derivatives $\partial f/\partial t$, $\partial f/\partial x$, and $\partial^2 f/\partial x^2$, then the theorem remains true if $f(x)$ is replaced by $f(t, x)$ and $L_{a,v} f$ is replaced by

$$\hat{L}_{a,v} f(t, x) = \frac{1}{2} \Delta f(t, x) + (a(x), f_x(t, x)) + f_t(t, x) + v(x) f(t, x) \, .$$

Historic and Bibliographic Comments

Wiener (1923) gave a rigorous development of the Wiener process by means of a Wiener measure in **C**.

A broad class of Markov processes was introduced by Kolmogorov (1933). This paper obtained the equations for a transition probability now bearing his name.

Petrovsky (1934) developed a method using differential equations to prove the convergence of Markov random walks to continuous Markov processes. Feller (1936) considered certain classes of Markov processes and their transition probabilities starting from Kolmogorov's equations. Ito (1944) gave the first definition of a stochastic integral with respect to a Wiener measure. Kac (1949) derived the first representation of the solution to the heat equation with a potential as an integral with respect to a Wiener measure.

A complete presentation of the theory of homogeneous Markov processes is contained in Dynkin's book (1963). In particular, it gives probability representations of solutions to equations containing the generating and characteristic operators of the Markov processes. Chung's book (1960) contains a complete presentation of the theory of countable homogeneous chains and Markov processes. Among other things, Venttsel' and Freidlin (1979) furnish probability representations of solutions to boundary-value problems. They also use Markov processes to analyze asymptotically solutions to boundary-value problems involving second-order differential operators with small parameter in the leading coefficients. Khinchin (1937) considered in particular diffusion problems and related limit theorems.

References

Chung, K.L. (1960) Markov Chains with Stationary Transition Probabilities. Springer-Verlag, Berlin-Göttingen-Heidelberg

Dynkin, E.B. (1963) Markov Processes. Fizmatgiz, Moscow. English transl.: Academic Press, New York, Vols. 1,2, 1965

Feller, W. (1936) Zur Theorie der stochastischen Prozesse (Existenz und Eindeutigkeitssätze). Math. Ann., 113, No. 1, 113–160

Itô, K. (1944) Stochastic Integral. Proc. Imp. Acad., 20, 519–524

Kac, M. (1949) On distribution of certain Wiener functionals. Trans. Amer. Math. Soc., 65, 1–13

Khinchin, A.Ya. (1937) Asymptotic Methods in Probability Theory. Gostekhizdat, Moscow-Leningrad

Kolmogorov, A.N. (1931) Über die analytische Methoden in der Wahrscheinlichkeitsrechnung. Math. Ann., 104, 415–458

Petrovsky, I.G. (1934) Über das Irrfahrproblem. Math. Ann., 109, 425–444

Venttsel', A.D. and Freidlin, M.I. (1979) Fluctuations in Randomly Perturbed Dynamical Systems. Nauks, Moscow (Russian)

Wiener, N. (1923) Differential space. J. Math. Phys. Mass. Techn., 2, 131–174

Applied Probability

Contents

Introduction

The applied sections of probability theory may be characterized as the science of practical applications under conditions of uncertainty. In building models to solve problems arising in probability, one encounters two kinds of uncertainty: 1. uncertainty caused by random factors and 2. uncertainty due to not knowing the stochastic parameters of the probability experiments describing a model. In principle, in order to make practical use of stochastic models, it is necessary to be able to study the second kind of uncertainty. In other words, on the basis of experimental data, one has to be able to find an unknown probability. How to do this is what mathematical statistics studies and so it lies at the basis of all applications of probability theory.

If we already have a determined stochastic model (that is, all probabilities are known), then its further study can be accomplished by the purely mathematical methods developed in probability theory.

We are going to examine here three of the most typical areas of applied probability theory: controlled stochastic processes, transmission of information and filtering. They are united by the fact that a basic object in their study is a certain class of stochastic processes and the problem is to select an optimum process.

Naturally, in a relatively short article like this, one cannot encompass all of the aspects of applying the theory. Our aim is to show that such applications are feasible and to illustrate the nature of these applications.

1

Statistical Methods

Statistics is an independent science concerned with the mathematical processing of statistcal data. Considered in a broad sense, statistics falls outside the framework of mathematics and encompasses probability theory merely for its mathematical (or theoretical) justification. In a narrow sense, statistics is a branch of probability theory and its job is to determine stochastic parameters from experimental data. We shall examine some of the very simple problems of this kind in order to illustrate how probability theory "works" in practice.

1.1 Processing of Statistical Data

Sample data are the result of making a series of observations. These observations may involve recording the presence or absence of some feature (for example, the sex of a new-born child) or measuring one or several parameters (for example, the weight of corn, the sweetness of fruit and so on). We shall view such a series as resulting from independent probability experiments. Statistical problems arise if the probability parameters of an experiment are unknown.

1.1.1 Relative Frequency and Probability

(a) What is it necessary to know to determine an unknown probability? Suppose that the sample data establish whether or not some event A has occurred in a series of n experiments (or trials). These data may be expressed by way of rows of zeroes and ones, a one marking the occurrence of A and a zero its non-occurrence. The probability of A is unknown. What could one say about this probability on the basis of the existing data?

Let $\mathbf{P}(A) = p$ and $\mathbf{P}(\bar{A}) = 1 - p = q$ ($\bar{A}$ is the event complementary to A). If A has happened m times, then one of the events

$$B_{i_1}, \ldots, i_m, \quad 1 \leq i_1 < i_2 < \ldots < i_m \leq n \,,$$

has happened (which means that A has occurred in the experiments numbered $i_1, i_2, \ldots, i_m$ and has not occurred in the remaining experiments), and

$$\mathbf{P}(B_{i_1}, \ldots, i_m) = p^m q^{n-m} \ .$$

Thus, letting $B^m = \bigcup_{i_1 < \ldots < i_m} B_{i_1}, \ldots, i_m$, we have

$$P\left(B_{i_1}, \ldots, i_m | B^m\right) = 1 \Big/ \binom{n}{m} = \frac{m!(n-m)!}{n!}. \tag{1.1.1}$$

This probability does not depend on p. Thus, if we know the number of occurrences of A in the n experiments, then any additional information telling us in precisely what trials A occurred furnishes nothing for finding p. Thus all of the possible information about the unknown probability p of A is contained in the relative frequency of occurrence of A.

A measurable function of a random sample is called a *statistic*. Relative frequency is a statistic. A statistic is *sufficient* if the conditional distribution of the random sample, given a fixed value of the statistic, is independent of the underlying distribution. The relative frequency is a sufficient statistic.

(b) *Pointwise estimation of an unknown probability.* If we want ro estimate an unknown probability from sample data, then it is natural to do this by means of some function of the random sample, that is, a statistic. It is self-understood that such a function cannot depend on the unknown probability p. A statistic that estimates an unknown stochastic parameter (in the present case the probability of A) is called an *estimator*. To determine the quality of an estimator necessitates some further concepts. Suppose that p_n^* is an estimator of p. p_n^* is a function of the rows of length n consisting of ones and zeros and it takes a value in [0,1]. After n trials have been performed, we have a specific row of experimental data and p_n^* has a specific value. The estimator p_n^* is said to be *unbiased* if $\mathbf{E}p_n^* = p$. It is natural to assess the quality of an unbiased estimator by means of its variance $\mathbf{V}p_n^*$, which depends on p in general. An unbiased estimator is admissible if for any other unbiased estimator $\tilde{p}_n^*$ it cannot be that $\mathbf{V}\tilde{p}_n^* \leq \mathbf{V}p_n^*$ for all p and $\mathbf{V}\tilde{p}_n^* < \mathbf{V}p_n^*$ for at least one p. If an estimator is inadmissible, one can certainly find a better one and so such estimators are avoided.

Theorem 1.1.1. *If p_n^* is an unbiased estimator of p, then there exists an estimator of the form $g(\nu_n)$, where ν_n is relative frequency, which is also unbiased and for which $\mathbf{V}g(\nu_n) \leq \mathbf{V}p_n^*$.*

Proof. Let $g(\nu_n) = \mathbf{E}(p_n^*|\nu_n)$. Then using (1.1.1), we obtain

$$g\left(\frac{m}{n}\right) = \binom{n}{m}^{-1} \sum_{i_1 < \ldots < i_m} \mathbf{E}(p_n^*|B_i, \ldots, i_m) \ . \tag{1.1.2}$$

$\mathbf{E}(p_n^*|B_{i_1,\dots,i_m})$ is the value of p_n^* on the row having ones in the positions numbered $i_1,\dots,i_m$ and zeroes in the remaining places. Therefore the right-hand side of (1.1.2) does not depend on p and $g(\nu_n)$ is a statistic. We have

$$\mathbf{E}g(\nu_n) = \mathbf{E}\mathbf{E}(p_n^*|\nu_n) = \mathbf{E}p_n^* = p\ .$$

Thus $g(\nu_n)$ is an unbiased estimator. Now

$$\begin{aligned}\mathbf{V}g(\nu_n) &= \mathbf{E}g^2(\nu_n) - p^2 = \mathbf{E}(\mathbf{E}(p_n^*|\nu_n))^2 - p^2\\ &\leqslant \mathbf{E}\mathbf{E}(p_n^{*^2}|\nu_n) - p^2 = \mathbf{E}p_n^{*^2} - p^2 = \mathbf{V}p_n^*\ .\end{aligned}$$

□

Theorem 1.1.1 implies that estimators of the form $g(\nu_n)$ cannot be improved by way of estimators from a broader class. Therefore one may restrict oneself to just these estimators.

Theorem 1.1.2. *ν_n is the only unbiased estimator of the form $g(\nu_n)$.*

Proof. Let $g_m = \binom{n}{m} g(m/n)$. If $g(\nu_n)$ is an unbiased estimator, then g_m satisfies the relation

$$\sum_{m=0}^{n} p^m(1-p)^{n-m} g_m = p$$

for all p, or

$$\sum_{m=0}^{n} \left(\frac{p}{1-p}\right)^m g_m = p(1-p)^{-n}\ .$$

Let $0 \leq p_0 < p_1 \dots < p_n \leq 1$, $p_k(1-p_k)^{-1} = x_k$ and $p_k(1-p_k)^{-n} = a_k$. Then g_m satisfies the system of $n+1$ equations

$$\sum_{m=0}^{n} x_k^m g_m = a_k, \quad k = 0,\dots,n\ .$$

Since the x_k are distinct, the determinant of the system does not vanish and so its solution is unique.

(c) *Interval estimation.* Another way to estimate an unknown probability is to construct a *confidence interval*. Let α_n^* and β_n^* be two statistics with $\alpha_n^* < \beta_n^*$. The interval (α_n^*, β_n^*) is called a confidence interval for any p of level $\varepsilon > 0$ if

$$\mathbf{P}\{\alpha_n^* < p < \beta_n^*\} \geqslant 1 - \varepsilon\ . \tag{1.1.3}$$

The simplest confidence interval for p can be formed by using the estimator ν_n and Chebyshev's inequality. Since $\mathbf{E}\nu_n = p$ and $\mathbf{V}\nu_n = p(1-p)/n \leq 1/4n$, it follows that

$$\mathbf{P}\{|\nu_n - p| \geqslant c\} \leqslant \frac{1}{4nc^2} .$$

Consequently,

$$\mathbf{P}\left\{|\nu_n - p| \geqslant \frac{1}{2\sqrt{n\varepsilon}}\right\} \leqslant \varepsilon .$$

Putting $\alpha_n^* = \nu_n - 1/(2\sqrt{n\varepsilon})$ and $\beta_n^* = \nu_n + 1/(2\sqrt{n\varepsilon})$, we arrive at a confidence interval of level ε. Of all the confidence intervals of a given level one would obviously like to find the one of shortest length.

1.1.2 Empirical Distribution Function

Suppose now that the sample data consist of a series of measurements of some variable. The variable is considered to be random and the measurements to be independent observed values. The distribution of the random variable is unknown. In other words, we are viewing the series of observed values $x_1, x_2, \ldots, x_n$ as a collection of independent and identically distributed random variables (or a random sample) whose distribution function $F(x)$ is unknown. Among the statistical problems are those whose answers are expressed in terms of functions of $x_1, \ldots, x_n$ (statistics) and do not depend on $F(x)$. Generally, there is a priori information known about the nature of the distribution function. Here we shall examine the case where it is only known that $F(x)$ is continuous. Then $x_1, \ldots, x_n$ are distinct with probability 1.

Let ξ have the distribution $F(x)$ and let A_x be the event $\{\xi < x\}$. Having n observations of ξ (the x_i may be viewed as such observations), we also have n observed values from an experiment in which A_x does or does not occur. The relative frequency of occurrence of A_x in these n trials is $n^{-1}\sum_{k=1}^{n} I_{\{x_k<x\}}$. This expression depends on x. As was established in Sect. 1.1.1, it is the only unbiased estimator of the probability $F(x)$ and it depends symmetrically on $I_{\{x_k<x\}}$. The function

$$F_n^*(x) = \frac{1}{n}\sum_{k=1}^{n} I_{\{x_k<x\}} \tag{1.1.4}$$

is called the *empirical or sample distribution function*. It is a statistic with values in the space of distribution functions. In addition, it is a sufficient statistic, that is, the conditional joint distribution of $x_1, \ldots, x_n$, given F_n^*, does not depend on $F(x)$. In order to prove this, we consider the *order statistics* or *variational series* $x_1^* < x_2^* < \ldots x_n^*$ of our sample data, where the x_n^* are determined by the fact that the sets $\{x_1^*, \ldots, x_n^*\}$ and $\{x_1, \ldots, x_n\}$ are the same (we are assuming that all of the observations are distinct). To form order statistics, it is necessary to arrange the sample data in order of growth. The x_n^*'s are points of discontinuity of $F_n^*(x)$ with $F_n^*(x_n^*+) - F_n^*(x_n^*) = 1/n$. An empirical distibution function is completely determined by the order statistics.

Consider the conditional joint distribution of $x_1, \ldots, x_n$ given $x_1^*, \ldots, x_n^*$. Obviously, the x_i's assume one of the values of the order statistics. Let $i_1, \ldots, i_n$ be a permutation of $1, \ldots, n$. Then the probabilities

$$\mathbf{P}\{x_1 = x_{i_1}^*, \ldots, x_n = x_{i_n}^* | x_1^*, \ldots, x_n^*\}$$

are all equal since the order statistics are symmetric functions of $x_1, \ldots, x_n$. Hence, these probabilities all equal $(n!)^{-1}$. This then proves our assertion.

(a) *Sampling parameters of a distribution.* $F_n^*(x)$ is a distribution function and it can be used to compute parameters that are commonly studied for nonrandom distributions. They are termed *sampling parameters*. The quantity

$$\bar{x}_n = \frac{1}{n} \sum_{k=1}^{m} x_k = \int x dF_n^*(x) \tag{1.1.5}$$

is the *sample mean*,

$$s_n^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x}_n)^2 = \int x^2 dF_n^*(x) - (\bar{x}_n)^2 \tag{1.1.6}$$

is the *sample variance* and $s_n = \sqrt{s_n^2}$ is the *sample standard deviation*. One can also consider the *sample moments* of order α:

$$m_n^*(\alpha) = \frac{1}{n} \sum_{k=1}^{n} |x_k|^\alpha = \int |x|^\alpha dF_n^*(x) \; . \tag{1.1.7}$$

Another class of parameters, not expressible by integrals with respect to $F_n^*(x)$, are the sample quantiles. The *sample quantile* of order $p, 0 < p < 1$, or *percentile* is the smallest solution $q_n^*(p)$ of the equation

$$F_n^*(x) \leqslant p \leqslant F_n^*(x+) \; .$$

It is easy to see that $q_n^*(k/n) = x_k^*$ and that $q_n^*(p) = x_{[np]+1}^*$, $p \neq k/n$ ($[\cdot]$ is the integral part of a number).

1.1.3 Strong Law of Large Numbers and Limiting Behavior of Sampling Parameters

The strong law of large numbers tells us that the relative frequency ν_n approaches p with probability 1 as $n \to \infty$. We now prove the strong law of large numbers for an empirical distribution function.

Theorem 1.1.3. *With probability* 1,

$$\lim_{n \to \infty} \sup_x |F_n^*(x) - F(x)| = 0 \; .$$

Proof. Recall that $F(x)$ is continuous by assumption. Let $z_1, \ldots, z_m$ be such that $F(z_k) = k/m$, $k = 1, \ldots, m-1$. The strong law of large numbers for the relative frequency implies that

$$\lim_{n\to\infty} |F_n^*(z_k) - F(z_k)| = 0$$

with probability 1. Define

$$\eta_n = \max_{k\leqslant m} |F_n^*(z_k) - F(z_k)| \ .$$

For $x < z_1$

$$|F_n^*(x) - F(x)| \leqslant F_n^*(z_1) \vee \frac{1}{m} \leqslant \frac{1}{m} + \eta_n \ ,$$

for $z_{k-1} \leqslant x \leqslant z_k$,

$$|F_n^*(x) - F_n(x)| \leqslant \left| F_n^*(z_k) - \frac{k-1}{m} \right| \vee \left| \frac{k}{m} - F_n^*(z_{k-1}) \right| \leqslant \frac{1}{m} + \eta_n \ ,$$

and for $x > z_{m-1}$

$$|F_n^*(x) - F_n(x)| \leqslant \frac{1}{m} \vee |1 - F_n^*(z_{k-1})| \leqslant \frac{1}{m} + \eta_n \ .$$

Thus with probability 1,

$$\overline{\lim_{n\to\infty}} \sup_x |F_n^*(x) - F(x)| \leqslant \frac{1}{m} \ .$$

□

Corollary. *Let q_α satisfy $F(q_\alpha) = \alpha, 0 < \alpha < 1$. For all $x < q_\alpha$, let $F(x) < \alpha$ and for all $x > q_\alpha$, let $F(x) > \alpha$. Then $q_n^*(\alpha) \to q_\alpha$ with probability 1 as $n \to \infty$. If $\int |x|^\alpha dF(x) < \infty$, then*

$$\lim_{n\to\infty} m_n^*(\alpha) = \int |x|^\alpha dF(x) \ .$$

with probability 1*, which is a direct consequence of the strong law of large numbers and formula* (1.1.7)*.*

1.1.4 Kolmogorov-Smirnov Goodness-of-Fit Test

Suppose that certain a apriori data (that is, independent of the existing sample data) allow one to expect the distribution function of $x_1, \ldots, x_n$ to be a given distribution function $F_0(x)$ (for instance, the normal distribution with mean 0 and variance 1). How does one establish whether such an assumption agrees

with the existing sample data? To answer this, statistics utilizes goodness-of-fit tests.

We choose some nonnegative statistic $g_n(x_1, \ldots, x_n)$. We can find its distribution function if the distribution of the variables x_k is $F_0(x)$. Let it be $G_n(x)$. Let ε be a fixed positive number and let z_ε^n be such that $G_n(z_\varepsilon^n) \geq 1-\varepsilon$. If the hypothesis that $F_0(x)$ is the true distribution function is valid and $g_n(x_1, \ldots, x_n) \geq z_\varepsilon^n$, then in a single experiment measuring $g_n(x_1, \ldots, x_n)$, an event has occurred whose probability is less than or equal to ε. Therefore in that case, the hypothesis is rejected. If $g_n(x_1, \ldots, x_n) < z_\varepsilon^n$, the sample data is viewed as not contradicting the hypothesis that $F_0(x)$ is the distribution function. The choice of ε depends on the conditions in each specific problem.

The function $g_n(x_1, \ldots, x_n)$ is also termed a test statistic. We now examine two such test statistics,

$$D_n = \sqrt{n} \sup_x |F_0(x) - F_n^*(x)|$$

and

$$D_n^+ = \sqrt{n} \sup_x (F_n^*(x) - F_0(x)) \, .$$

The first one is due to Kolmogorov and the second to Smirnov. These tests are convenient to use because of the following properties.

1. If $F_0(x)$ is the true distribution function, then the distribution of $D_n (D_n^+)$ does not depend on $F_0(x)$.

If x_k has the distribution function $F_0(x)$, then $F_0(x_k)$ is uniformly distributed over $[0, 1]$. Let $\tilde{x}_i = F_0(x_i)$ and let $\tilde{F}_n^*(x)$ be the empirical distribution function of $\tilde{x}_i$, $i = 1, \ldots, n$. Then

$$\tilde{F}_n^*(F_0(x)) = \frac{1}{n} \sum_{i=1}^n I_{\{F_0(x_i) < F_0(x)\}} = \frac{1}{n} \sum_{i=1}^n I_{\{x_i < x\}} = F_n^*(x) \, .$$

Therefore

$$D_n = \sup_x \sqrt{n} |\tilde{F}_n^*(F_0(x)) - F_0(x)| = \sup_{0 \leqslant t \leqslant 1} \sqrt{n} |\tilde{F}_n^*(t) - t| \, ,$$

and

$$D_n^+ = \sup_{0 \leqslant t \leqslant 1} \sqrt{n} (\tilde{F}_n^*(t) - 1) \, .$$

The right-hand side of each of these last equations does not depend on $F_0(x)$.

2. The limiting distribution functions of D_n and D_n^+ exist as $n \to \infty$ and are

$$\lim_{n \to \infty} \mathbf{P}\{D_n < z\} = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}, \quad (z \geqslant 0) \tag{1.1.8}$$

$$\lim_{n \to \infty} \mathbf{P}\{D_n^+ < z\} = 1 - e^{-2z^2} \, . \tag{1.1.9}$$

The proofs of these formulas are very cumbersome. We shall merely indicate the idea behind their derivation. Consider the process $\eta_n(t) = \sqrt{n}(F_n^*(t) - t)$. Applying the central limit theorem, one can establish that for any $t_1, t_2, \ldots, t_m \in [0,1]$ the joint distribution of $\eta_n(t_1), \ldots, \eta_n(t_m)$ converges to the joint distribution of $\eta(t_1), \ldots, \eta(t_m)$, where $\eta(t), t \in [0,1]$, is a Gaussian stochastic process with mean 0 and covariance function $r(t,s) = t(1-s)$ for $0 \leq t \leq s \leq 1$. It is easy to see that $\eta(t) = w(t) - tw(1)$, where $w(t)$ is a Wiener process. One can show (using the compactness of the measures corresponding to the $\eta_n(t)$ in some metric space) that the distribution of the functionals $\sup_t \eta_n(t)$ and $\sup_t |\eta_n(t)|$ converge to the respective distributions of $\sup_t(w(t) - tw(1))$ and $\sup_t |w(t) - tw(1)|$. The right-hand sides of (1.1.8) and (1.1.9) are precisely the distributions of these two random variables.

1.2 Testing of Hypotheses

The preceding section already considered the question of whether existing sample data were in agreement with the hypothesis that a distribution function was of a given form. However this does not preclude the possibility that the sample data may be consistent with other hypotheses about the distribution function. The question is how to select a proper hypothesis from among several.

1.2.1 Statement of the Problem

The set of possible distribution functions is assumed to be parametrizable and we denote it by $\{F_\theta, \theta \in \Theta\}$. One F_{θ_0} is singled out from among these distribution functions and it plays the role of hypothesis to be tested. The remaining distributions are the alternative hypotheses. On the basis of the sample data, one has to accept or reject the hypothesis H_0 (the *null hypothesis*): the distribution function of the observed variable is F_{θ_0}. It is then natural to concentrate on the possible alternatives. Each test that accepts or rejects H_0 is specified by a set $C \subset R^n$. C is a Borel set such that H_0 is rejected if $(x_1, \ldots, x_n) \in C$ and it is called the *critical region* or *set*. If $(x_1, \ldots, x_n) \notin C$, hypothesis H_0 is accepted.

The quality of a test is characterized first of all by the probability F_{θ_0} of rejecting H_0 when it is true. In that case one says that a type I error has been committed. $F_\theta(C)$ is called the *probability of a type I error*. Naturally, one needs to consider rules for which this probability is sufficiently small. Generally speaking, it may be made arbitrarily small by an appropriate choice of C (for example, by always accepting H_0, that is, taking $C =$empty set, we make the probability of a type I error vanish). But then we would start to accept H_0 even when it is untrue (this is a type II error). Therefore the second parameter of the test is introduced: the *probability of a type II error* $F_\theta(R^n \backslash C), \theta \in \Theta \backslash \{\theta_0\}$. The problem is generally posed this way. From among

the tests for which the probability of a type I error does not exceed a given level α, find the one for which the probability of a type II error is in a sense smallest. Of course, the ideal case is where there exists a rule such that one for any other rule (with α fixed) the probability of a type II error is no less than the given one for any $\theta \neq \theta_0$. Instead of the probability of a type II error, statistics customarily considers the function $m(\theta) = 1 - F_\theta(R^n \backslash C) = F_\theta(C)$ which is called the *power* of the test. A rule accepting an hypothesis will be denoted by R (with or without subscripts). Every rule is defined by some critical region. The power of rule R will be denoted by $m_R(\theta)$. Rule R_1 is uniformly more powerful than rule R_2 if $m_{R_1}(\theta) \geq m_{R_2}(\theta)$ for all $\theta \neq \theta_0$ and $m_{R_1}(\theta) > m_{R_2}(\theta)$ for at least one θ. If only the first relation holds, then R_1 is uniformly no less powerful than R_2. A class $\mathcal{R}$ of rules (tests) is *complete* if to every R there exists an $R' \in \mathcal{R}$ which is uniformly no less powerful than R. A rule is uniformly unimprovable if there exists no uniformly more powerful rule. Every uniformly unimprovable rule may be used to accept an hypothesis against certain alternative hypotheses. In this connection, one proceeds in optimum fashion. The problem of testing hypothesis H_0 (against a given set of alternatives) can be viewed as solved if for every $\alpha > 0$, all unimprovable rules have been described or at least a class of such rules. It may happen that no unimpovable rules exist at all. The problem then becomes one of describing the complete classes.

Example. n trials are performed in which an event A is observed. The null hypothesis H_0 is $\mathbf{P}(A) = 1/2$. The alternative hypothesis is all distributions for which $\mathbf{P}(A) > 1/2$. The possible sample data are rows of zeroes and ones. Since the probability of such a row is $1/2^n$ under H_0, $F_{\theta_0}(C) = k/2^n$ for any critical region C containing k points. Let $\alpha = k/2^n$. If C contains the points $z_1, \ldots, z_k$ (with coordinates of ones and zeroes only) and n_i is the number of ones among the coordinates z_i, then

$$m_R(p) = \sum_{i=1}^{k} p^{n_i}(1-p)^{n-n_i} = (1-p)^n \sum_{i=1}^{k} \left(\frac{p}{1-p}\right)^{n_i} .$$

The right-hand side depends on the choice of C by way of n_i. Since $p/(1-p) > 1$, the test will be uniformly most powerful if the numbers n_i are chosen to be largest. Arranging the points with coordinates 0 and 1 in order of decreasing n_i (for example: $(1,1,\ldots,1), (1,1,\ldots,1,0), (1,\ldots,1,0,1), \ldots\ldots, (0,1,\ldots,1)$, $(1,1,\ldots 1,0,0,)$ and so on) and taking C to be the first k points, we arrive at a rule no less powerful than any other one.

1.2.2 Neyman-Pearson Test

We shall concentrate on the simplest case where there is a single alternative. Denote by F_0 the distribution under hypothesis H_0 and by F_1 the distribution under alternative hypothesis H_1. A sample is taken of a random element in

some measureable space $(X, \mathcal{B})$. We can then imagine that there is one observation (n independent observations of a variable in R may be viewed as one observation of a variable in R^n with independent and identically distributed components). It is necessary to accept or reject H_0 on the basis of this observation. The acceptance rule is determined by a critical region $C \in \mathcal{B}$. The probability of a type I error is $\alpha = F_0(C)$ and the probability of a type II error is $\beta = 1 - F_1(C)$. If the measures F_0 and F_1 were mutually singular, there would exist a set $B_0 \in \mathcal{B}$ such that $F_0(B_0) = 1$ and $F_1(B_0) = 0$. If C is taken to be $X \backslash B_0$, then we would have $\alpha = \beta = 0$. Thus there exists a test rule whose possible probabilities of type I and type II errors are smallest. Consider the case where F_1 is absolutely continuous with respect to F_0. Let $f(x)$ be the density of F_1 relative to F_0. The following theorem describes a class of unimprovable rules.

Theorem 1.2.1 (Neyman-Pearson**).** *Let $C \in \mathcal{B}$ be such that*

$$\sup_{x \in X \backslash C} f(x) \leqslant \inf_{x \in C} f(x)$$

and let $\alpha = F_0(C)$. Then $F_1(C_1) \leq F_1(C)$ for any $C_1 \in \mathcal{B}$ for which $F_0(C_1) \leq \alpha$.

Proof. Write $a = \inf_{x \in C} f(x)$. Then

$$\begin{aligned} F_1(C) - F_1(C_1) &= F_1(C \backslash C_1) - F_1(C_1 \backslash C) \\ &= \int_{C \backslash C_1} f(x) F_0(dx) - \int_{C_1 \backslash C} f(x) F_0(dx) \\ &\geqslant a F_0(C \backslash C_1) - a F_0(C_1 \backslash C) = a(F_0(C) - F_0(C_1)) \end{aligned}$$

($f(x) \geq a$ when $x \in C \backslash C_1$ and $f(x) \leq a$ when $x \in C_1 \backslash C$).

Corollary 1.2.1. *Let $\{R_a,\ a \in R_+\}$ be a class of rules with critical regions $C_a = \{x : f(x) > a\}$. It is a complete class of unimprovable rules.*

Corollary 1.2.2. *Let a sample of size n be taken of an R-valued random variable. H_0 is the hypothesis that the random variable has a density $f_0(x)$. The alternative hypothesis H_1 is that its density is $f_1(x)$. Then the rules with critical regions of the form*

$$\left\{ x \in R^n : \prod_{k=1}^{n} \frac{f_1(x_k)}{f_0(x_k)} > a \right\}$$

are unimprovable.

Example. Let

$$f_0(x) = \frac{1}{\sqrt{2\pi b_0}} \exp\left\{ -\frac{x^2}{2b_o} \right\}$$

and

$$f_1(x) = \frac{1}{\sqrt{2\pi b_1}} \exp\left\{-\frac{x^2}{2b_1}\right\}$$

with $b_0 < b_1$. Then the class of rules with critical regions $\{x \in R^n : \sum_{k=1}^n x_k^2 > t\}$, $t \in R_+$, is a complete class of unimprovable rules. This is a consequence of Corollary 1.2.2 and the fact that to every $\alpha \in (0,1)$, there is a t such that

$$F_0^{(n)}\left(\left\{x \in R^n : \sum x_k^2 > t\right\}\right) = a\ ,$$

where $F_0^{(n)}$ is the normal distribution in R^n with the density $\prod_{k=1}^n f_0(x_k)$.

1.2.3 Detection of a Signal with Background Noise

Consider the following problem. A "message" is received, which can be described by a numerical function defined on the interval $[a,b]$. The function is a stochastic process. The randomness is due to "noise" that exists in the surrounding medium. If the message is transmitted, for example, by radio, then the noise is caused by an external electormagnetic field due to solar radiation, atmospheric effects and man's industrial activities. Suppose that the receiver has to bring in an effective signal of a given form. Therefore having received the message, one must decide whether it is a meaningful signal or the message is pure noise. We shall confine ourselves to the simplest case where the noise and signal occur in the message additively. Let $a(t)$ with $t \in [a,b]$ be the function defining the signal and let $\xi(t)$ be the stochastic process defining the noise. Entering the receiver is either pure noise, that is, the process $\xi(t)$, or the signal with noise, that is, $a(t)+\xi(t)$. Let μ_0 be the measure in function space corresponding to $\xi(t)$ and let μ_1 be the measure corresponding to $a(t)+\xi(t)$. From the standpoint of statistics, we must test the hypothesis H_0 that the process $x(t)$ entering the receiver has distribution μ_0 against the alternative H_1 that $x(t)$ has the distribution μ_1. Assume that μ_1 is absolutely continuous with respect to μ_0 and that $f(x) = d\mu_1(x)/d\mu_0$; its argument is any function defined on $[a,b]$. Then by virtue of the Neyman-Pearson theorem, by accepting H_0 if $f(x) \le c$ and rejecting it if $f(x) > c$, we will minimize the type II error $\mu_1(\{x : f(x) \le c\})$ under a type I error $\alpha = \mu_0(\{x : f(x) > c\})$.

Example. Let $\xi(t)$ be a Gaussian process with mean 0 and continuous covariance function $r(t,s), t$ and $s \in [a,b]$. Denote by $\{\varphi_k(t)\}$ the orthonormalized set of eigenfunctions of the kernel $r(t,s)$ and let $\{\lambda_k\}$ be their corresponding eigenvalues:

$$\lambda_k \varphi_k(t) = \int_a^b r(t,s)\varphi_k(s)ds\ .$$

We assume that $a(t)$ is square-integrable. Let $a_k = \int_a^b a(t)\varphi_k(t)dt$. Under the assumptions made, the process $\xi(t)$ has a square-integrable modification.

Therefore we can view μ_0 and μ_1 as being defined on $L_2[a,b]$. Consider the mapping from $L_2[a,b]$ to l_2 given by

$$x(t) \to \left(\int_a^b x(t)\varphi_1(t)dt, \ldots, \int_a^b x(t)\varphi_k(t)dt, \ldots \right) .$$

It associates with $x(t)$ its Fourier coefficients with respect to $\{\varphi_k\}$ and it is an injection (the functions in $L_2[a,b]$ equal almost everywhere are identified with one another). Let $\tilde{\mu}_0$ and $\tilde{\mu}_1$ be the images of μ_0 and μ_1 under this mapping. They will be product measures in the space of sequences. If we denote a sequence by $x = (x_1, x_2, \ldots)$, then x_i has a normal distribution with mean 0 and variance λ_i with respect to the measure $\tilde{\mu}_0$ and a normal distribution with the same variance but mean a_i with respect to the measure $\tilde{\mu}_1$. From Sect. 3.5.4 on page 91, it follows that $\tilde{\mu}_1 \ll \tilde{\mu}_0$ if and only if

$$\sum a_i^2/\lambda_i < \infty , \tag{1.2.1}$$

and then

$$f(x) = \exp\left\{ \sum_{i=1}^{\infty} \frac{a_i}{\lambda_i} x_i - \frac{1}{2} \sum_{i=1}^{\infty} \frac{a_i^2}{\lambda_i} \right\} . \tag{1.2.2}$$

Thus hypothesis H_0 is accepted if

$$\sum_{i=1}^{\infty} \frac{a_i}{\lambda_i} x_i < r , \tag{1.2.3}$$

and rejected if $\sum_{i=1}^{\infty}(a_i/\lambda_i)x_i \geq r$ for some number r. To determine the type I and type II errors, let $d = \sum_{i=1}^{\infty}(a_i^2/\lambda_i)$. Then under H_0, $\sum_{i=1}^{\infty}(a_i/\lambda_i)x_i$ has a normal distribution with mean 0 and variance d. The probability of a type I error is therefore

$$\alpha(r) = \frac{1}{\sqrt{2\pi}} \int_{\frac{r}{\sqrt{d}}}^{\infty} e^{-\frac{u^2}{2}} du . \tag{1.2.4}$$

Under hypothesis H_1, $\sum_{i=1}^{\infty}(a_i/\lambda_i)x_i$ has a normal distribution with the same variance but with mean d. The probability of a type II error is

$$\beta(r) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{r-d}{\sqrt{d}}} e^{-\frac{u^2}{2}} du . \tag{1.2.5}$$

The condition (1.2.1) implies that the function $b(t) = \sum_{i=1}^{\infty}(a_i/\sqrt{\lambda_i})\varphi_i(t)$ is defined and belongs to $L_2[a,b]$. In addition

$$b(t) = \int_a^b r^{1/2}(t,s)a(s)ds ,$$

where

$$r^{1/2}(t,s) = \sum_{k=1}^{\infty} \sqrt{\lambda_k}\varphi_k(t)\varphi_k(s) \ .$$

Condition (1.2.3) can be rewritten in the form

$$\int_a^b b(t)x(t)dt < r \ .$$

The quantity d involved in the error probabilities can be expressed in terms of $b(t)$ as

$$d = \int_a^b b^2(t)dt \ .$$

Suppose, for example, that $\xi(t) = w(t)$, $t \in [0,1]$, and $a(t) = at$. Then $b(t) = a$ and $d = a^2$. H_0 is accepted if $aw(1) < r$ and the error probabilities are given by (1.2.4) and (1.2.5).

1.3 Decision-Making Under Uncertainty

It is easy to imagine a situation where it is necessary to know the distributions of random variables in order to make certain decisions. For instance in designing a seasonal line of clothes and shoes, one has to know the distributions of the parameters that establish the sizes. The design of various devices must take into account possible effects of random perturbations and for that purpose it is necessary to know their distributions. Almost everywhere in all practical activity, individuals have to make decisions under uncertainty. We shall only consider the uncertainty that exists in not knowing some distribution. A statistical approach to this situation is presented below: one assumes that certain statistical information is available that will make it possible to make a judgment about an unknown distribution. Decisions must be made on the basis of this information.

1.3.1 Statement of the Problem

Suppose that we are given the following objects: 1. a measurable space $(X, \mathcal{B})$, the space of observed values; 2. a set $\{\mathbf{P}_\theta, \theta \in \Theta\}$ of distributions on X; $(\Theta, \mathcal{C})$ is a measurable space and the $\mathbf{P}_\theta$ are the possible values of the unknown distribution; $\mathbf{P}_\theta(B)$ is assumed to be $\mathcal{C}$-measurable for all $B \in \mathcal{B}$; 3. a measurable space $(D, \mathcal{D})$, the decision space; the elements $d \in D$ are the decisions that have to be made; 4. a function $R(\theta, d)$ from $\Theta \times D$ to R which is $\mathcal{C} \otimes \mathcal{D}$-measurable; it is called the *risk function* and it determines the losses in making decision d if the unknown (and designated "true") distribution is $\mathbf{P}_\theta$. This collection of objects specifies the conditions of the problem.

Suppose that there is given a sample of size n of a random element $x(\omega)$ in $(X, \mathcal{B})$ with distribution $\mathbf{P}_\theta$. Let the independent observed values be $x_1, \ldots, x_n$. Let $\mathbf{P}_\theta$ denote the probability and $\mathbf{E}_\theta$ the expectation of the variables in question under the assumption that $\mathbf{P}_\theta$ is the true distribution. The problem is to make a decision d on the basis of the existing sample data. Such a decision will of course be a function $d(x_1, \ldots, x_n)$ from X^n to D. Every $\mathcal{B}^n$-measurable function from X^n to D is called a *decision function*. The problem is thus reduced to choosing a decision function. The quality of this decision function is determined by the average loss sustained in using it. Decision functions will be denoted by $d^* = d(x_1, \ldots, x_n)$ and the collection of all decision functions by D^*. The average loss due to d^* is given by

$$R(\theta, d^*) = \mathbf{E}_\theta R(\theta, d(x_1, \ldots, x_n)) \, .$$

It is natural to seek a decision that minimizes the average loss. Since it depends on θ, it is hard to expect it to be minimized for all θ uniformly. Therefore the notions related to minimizing $R(\theta, d^*)$ will be made more precise later on.

Example 1.3.1. Given $(X, \mathcal{B}), X = \{0, 1\}$ and $\mathcal{B}$ are all the subsets of X. The distribution $\mathbf{P}_\theta$ is defined as $\mathbf{P}_\theta(\{1\}) = p_\theta, \Theta = \{0, 1\}$ and D comprises two decisions: $d_0 = (\theta = 0)$ and $d_1 = (\theta = 1)$. The risk function has four values: $R_{0,0} = R(0, d_0), R_{0,1} = R(0, d_1)$, $R_{1,0} = (1, d_0)$ and $R_{1,1} = R(1, d_1)$. Assume that $R_{0,0} < 0, R_{1,1} < 0, R_{0,1} > 0$ and $R_{1,0} > 0$. The decision function is found by dividing all n-dimensional rows of zeroes and ones into two subsets: D_0 where $d(x)$ takes the value 0 and D_1 where it takes the value 1. If d^* is the decision function determined by these two sets, then letting $n(x)$ be the number of ones among the coordinates, we obtain

$$\begin{aligned} R(\theta, d^*) &= R_{\theta,0} \sum_{x \in D_0} p_\theta^{n(x)} (1 - p_\theta)^{n - n(x)} \\ &\quad + R_{\theta,1} \left(1 - \sum_{x \in D_0} p_\theta^{n(x)} (1 - p_\theta)^{n - n(x)} \right) \\ &= R_{\theta,1} + (R_{\theta,0} - R_{\theta,1})(1 - p_\theta)^n \sum_{x \in D_0} \left(\frac{p_\theta}{1 - p_\theta} \right)^{n(x)} \, . \end{aligned}$$

It is now a question of choosing between two hypotheses about the probability of an event. Clearly, D_0 needs to be chosen so that $\sum_{x \in D_0} (p_0/(1-p_0))^{n(x)}$ is largest $(R_{0,0} - R_{0,1} < 0)$ and $\sum_{x \in D_0} (p_1/(1-p_1)^{n(x)}$ is smallest $(R_{1,0} - R_{1,1} > 0)$. If D_0 is defined to be of the form

$$D_0 = \left\{ x : \left[\frac{p_0(1 - p_1)}{p_1(1 - p_0)} \right]^x < \lambda \right\} \, ,$$

one can show that the values of $R(\theta, d^*)$ resulting for such decision functions will be uniformly no greater than for other decision functions.

Example 1.3.2. $X = R, \mathcal{B}$ is a Borel σ-algebra, $\mathbf{P}_\theta$ is a distribution in R depending on a real parameter θ, $D = R$, and the decisions are to choose the value of the parameter. $R(\theta, d) = (\theta - d)^2$. The decision function $d^* = \theta(x_1, \ldots, x_n)$ is an estimator of the parameter. The quality of the estimator is determined by the mean-square deviation

$$R(\theta, d^*) = \mathbf{E}_\theta (d^* - \theta)^2 \, .$$

Suppose, for example, that θ is the unknown mean of a normal distribution with variance 1. If d^* is an estimator of θ, then $\mathbf{E}_\theta(d^*|\bar{x})$, where $\bar{x}$ is the sample mean, does not depend on θ. Therefore it is also an estimator of θ and in addition

$$\begin{aligned} \mathbf{E}_\theta (\mathbf{E}_\theta(d^*|\bar{x}) - \theta)^2 &= \mathbf{E}_\theta (\mathbf{E}_\theta(d^* - \theta|\bar{x}))^2 \\ &\leqslant \mathbf{E}_\theta (\mathbf{E}_\theta((d^* - \theta)^2|\bar{x})) = \mathbf{E}_\theta (d^* - \theta)^2 \, . \end{aligned}$$

Thus it is meaningful to consider just estimators of the form $\mathbf{E}_\theta g(\bar{x})$. If the estimator is required to be unbiased, that is, $\mathbf{E}_\theta g(\bar{x}) = \theta$, then $\bar{x}$ is the only such estimator.

1.3.2 Minimax and Bayesian Decisions

Of all decision functions, it is reasonable to single out the class of unimprovable decision functions. A decision function $\hat{d}^*$ is unimprovable if the relation $R(\theta, \hat{d}^*) \geq R(\theta, d^*), \theta \in \Theta$, implies that $R(\theta, \hat{d}^*) = R(\theta, d^*), \theta \in \Theta$. A decision function d_1^* is no worse than a decision function d_2^* if $R(\theta, d_1^*) \leq R(\theta, d_2^*)$ for all θ. A class of decision functions $\mathcal{K}$ is said to be complete if to every decision function d^*, there exists a d_1^* in $\mathcal{K}$ which is no worse than d^*. Describing the unimprovable decision functions and complete classes is what mathematics can do to help solve the formulated problem. Specific decision functions are used that can also be considered to solve the problem.

(a) *Minimax decisions.* d^* is said to be a minimax decision function if

$$\max_\theta R(\theta, d^*) \leqslant \max R(\theta, d_1^*)$$

with d_1^* any decision function whatsoever. A *minimax decision* allows one to minimize a maximum possible loss. One could readily imagine such a situation where it is natural to make precisely such a decision, for example, when it is necessary to have a guaranteed result in any situation. In general, one cannot assert that a minimax decision function will exist. For it to exist, one needs to impose rather stringent constraints on the family of measures $\{\mathbf{P}_\theta, \theta \in \Theta\}$, the decision space $(D, \mathcal{D})$ and the risk function $R(\theta, d)$. The situations generally of interest are these involving ε-minimax decisions. These are the d_ε^*'s for which

$$\max_\theta R(\theta, d_\varepsilon^*) \leqslant \min_{d_1^*} \max_\theta R(\theta, d_1^*) + \varepsilon \, .$$

To ascertain these decisions, it is necessary to know the number

$$\min_{d_1^*} \max_{\theta} R(\theta, d_1^*) = R_{\min\max} ,$$

which is called the minimax risk. Of interest then are conditions on $R(\theta, d)$ under which

$$R_{\min\max} = \min_{d_1^*} \max_{\theta} R(\theta, d_1^*) = \min_{d} \max_{\theta} R(\theta, d)$$

(the evaluation of the right-hand side is an analytic problem).

(b) *Bayesian decisions.* Next consider the case where there is additional information about the family $\{\mathbf{P}_\theta, \theta \in \Theta\}$. Namely, we are given $\nu(d\theta)$, an a priori distribution of the parameter on $(\Theta, \mathcal{C})$.This a priori distribution might be discovered as the result of repeated recurrences of the exact same situations. Then after a decision function d^* has been chosen, the mean risk (taking $\nu(d\theta)$ into consideration) is

$$\hat{R}_\nu(d^*) = \int R(\theta, d^*)\nu(d\theta) .$$

It is natural to look for decisions that minimize $\hat{R}_\nu(d^*)$. If such a d_ν^* exists, it is called the *Bayesian decision* corresponding to the a priori distribution ν. The existence of a Bayesian decision for any a priori distribution is bound up with the topological properties of Θ and D (in this connection, $\mathcal{C}$ and $\mathcal{D}$ are Borel σ-algebras) and the continuity properties of $R(\theta, d)$. In any case, ε-Bayesian decisions do exist. In other words, there are decision functions $d_{\nu,\varepsilon}^*$ such that

$$\hat{R}_\nu(d_{\nu,\varepsilon}^*) \leqslant \inf_{d^*} \hat{R}_\nu(d^*) + \varepsilon .$$

The following theorem gives a connection between Bayesian decisions and unimprovable decision functions.

Theorem 1.3.1. *Let Θ be a complete separable metric space and let $R(\theta, d^*)$ be continuous in θ for any decision function d^*. If ν is an a priori distribution satisfying $\nu(C) > 0$ for every nonempty open set C, then a Bayesian decision d_ν^* (if it exists) is an unimprovable decision function.*

Proof. Assume the contrary: $R(\theta, d^*) \leq R(\theta, d_\nu^*)$ for some d^* but

$$\sup_{\theta} [R(\theta, d_\nu^*) - R(\theta, d^*)] > 0 .$$

Then $\{\theta : R(\theta, d_\nu^*) - R(\theta, d^*) > 0\}$ is an open set and hence

$$\int R(\theta, d_\nu^*)\nu(d\theta) > \int R(\theta, d^*)\nu(d\theta) ,$$

which contradicts the definition of d_ν^*. □

1.3.3 Sequential Analysis

Sequential analysis is used to solve statistical problems when the observed values of a random variable are not known all at once but arise one after the other (sequentially) and the statistical information builds up. The sampling is terminated after a sufficient amount of information is considered to have been amassed. The time that the sampling is stopped depends on these observations. It is random and has to be a stopping time in relation to the sequence of σ-algebras generated by the sampling data.

The formulation of the problem differs from the one above in that (a) instead of a sequence of n independent observations $x_1, \dots, x_n$ in the space $(X, \mathcal{B})$, we have an infinite sequence $\{x_n, n \geq 1\}$ and (b) the loss function $R(k, \theta, d)$ depends on one additional parameter $k \in R_+$, which is the sample size for which a decision d is made, θ being the true value of the parameter. The decision function is determined by two measurable functions $\tau(x_1, x_2, \dots)$ and $d(x_1, x_2, \dots)$ defined on $(X^\infty, \mathcal{B}^\infty)$ and taking values in R_+ and $\mathcal{D}$, respectively: $\tau(x_1, x_2, \dots)$ is the time that the decision is made and $d(x_1, x_2, \dots)$ is the decision. These two functions satisfy the following conditions:

1. if $\tau(x_1, x_2, \dots) = n$, then $\tau(x'_1, x'_2, \dots) = n$ when $x_1 = x'_1, \dots, x_n = x'_n$.

2. if $(x_1, \dots, x_n, \dots) = n$, then $d(x'_1, x'_2, \dots) = d(x_1, x_2, \dots)$ when $x_1 = x'_1, \dots, x_n = x'_n$.

The pair of functions $(\tau, d^*), \tau = \tau(x_1, x_2, \dots)$ and $d^* = d(x_1, x_2, \dots)$, will be called a sequential decision.

The average loss in utilizing a sequential decision (τ, d^*) is given by

$$R_\theta(\tau^*, d^*) = \mathbf{E}_\theta R(\tau(x_1, \dots), \theta, d(x_1, \dots))$$

(the expectation is taken with respect to the measure $\mu_\theta = \times_{m=1}^{\infty} \mathbf{P}_\theta(dx_n)$ in $(X^\infty, \mathcal{B}^\infty)$).

It is natural to extend the notions of unimprovable, minimax and Bayesian decisions to sequential decisions.

(a) *Sequential testing of two hypotheses.* Let $\Theta = \{0, 1\}$ and $\mathcal{D} = \{0, 1\}$. Decision $i (i = 0, 1)$ means that hypothesis H_i, that i is the true value of the parameter, is considered to be valid. Let a, b and c be given positive numbers which define R; a is the loss if H_0 is true and decision 1 is made; b is the loss if H_1 is true and decision 0 is made; and c is the cost of one observation. Accepting a true hypothesis is viewed as not involving a loss. Then the loss due to sequential decision (τ, d^*) is

$$\begin{aligned} R_0(\tau, d^*) &= c\mathbf{E}_0\tau + a\mathbf{P}_0\{d^* = 1\} , \\ R_1(\tau, d^*) &= c\mathbf{E}_1\tau + b\mathbf{P}_1\{d^* = 0\} . \end{aligned} \tag{1.3.1}$$

Every sequential decision is determined by a sequence of triples of sets $\{(A_n^0, A_n^c, A_n^1), n = 1, 2, \dots\}$, where A_n^0, A_n^c and $A_n^1 \in \mathcal{B}^n, A_n^0 \cup A_n^c \cup A_n^1 = X^n$

and A_n^0, A_n^c and A_1^1 are pairwaise disjoint. If $(x_1, x_2, \dots, x_n) \in A_n^i, i = 0, 1$, then hypothesis H_i is accepted. If $(x_1, x_2, \dots, x_n) \in A_n^c$, then it is necessary to carry out one more observation. For different n, the sets A_n^c are related as follows: if $(x_1, x_2, \dots, x_n) \in A_n^c$, then $(x_1, x_2, \dots, x_k) \in A_k^c$ for all $k < n$. If such a sequence of sets is given, then

$$\mathbf{E}_\theta \tau = 1 + \sum_{n=1}^{\infty} \mathbf{P}_\theta\{\tau > n\} = 1 + \sum_{n=1}^{\infty} \mathbf{P}_\theta\{(x_1, \dots, x_n) \in A_n^c\} ,$$

and

$$\mathbf{P}_\theta\{d^* = i\} = \sum_{n=1}^{\infty} \mathbf{P}_\theta\{(x_1, \dots, x_n) \in A_n^i\}, \quad i = 0, 1 .$$

Let us consider Bayesian sequential decisions. Let π be the a priori probability of H_1 and $1-\pi$ the a priori probability of H_0. The loss due to sequential decision (τ, d^*) is

$$\rho(\pi, \tau, d^*) = c\mathbf{E}_\pi \tau + \pi a \mathbf{P}_1\{d^* = 1\} + (1 - \pi) b \mathbf{P}_0\{d^* = 1\} , \tag{1.3.2}$$

where $\mathbf{E}_\pi = \pi \mathbf{E}_1 + (1 - \pi)\mathbf{E}_0$.

Let $f_i(x)$ be the density of the measure $F_i(dx)$ with respect to $F_0(dx) + F_1(dx)$. The conditional probabilities of H_0 and H_1, given observed value x, are respectively

$$\pi_x = \frac{f_1(x)\pi}{f_1(x)\pi + f_0(x)(1 - \pi)} \quad \text{and} \quad 1 - \pi_x .$$

Denote by $g(\pi)$ the smallest loss sustained on making a decision without carrying out an observation. If H_0 is accepted, the loss is $b\pi$. If H_1 is accepted, the loss is $a(1 - \pi)$ since $g(\pi) = b\pi \wedge a(1 - \pi)$.

Now let $\rho(\pi) = \inf \rho(\pi, \tau, d^*)$ over all sequential decisions. The function $\rho(\pi)$ satisfies the equation

$$\rho(\pi) = g(\pi \wedge (c + \mathbf{E}_\pi \rho(\pi_{x_1})) . \tag{1.3.3}$$

Relation (1.3.3) means that a smallest loss is sustained either on making a decision before observations or on carrying out a single observation x_1 (the loss from this is c); thereupon the probability of the hypothesis has become π_{x_1} and a minimal loss $\rho(\pi_{x_1})$ can be sustained which has to be averaged with respect to the a priori distribution of x_1. Equation (1.3.3) has a unique solution which can be found using successive approximations taking for example,

$$\rho_0(\pi) = 0, \quad \rho_n(\pi) = g(\pi) \wedge (c + \mathbf{E}_\pi \rho_{n-1}(\pi_{x_1})), \quad n \geq 1 ,$$

and

$$\rho(\pi) = \lim_{n \to \infty} \rho_n(\pi) .$$

It is easy to see from (1.3.2) that $\rho(\pi)$ is convex up and from (1.3.3) that $g(\pi) = \rho(\pi)$ when $g(\pi) \le c$. Therefore one can find numbers $0 < q_1 < q_2 < 1$ such that $g(\pi) > \rho(\pi)$ when $\pi \in (q_1, q_2)$ and $g(\pi) = \rho(\pi)$ when $\pi \le q_1$ or $\pi \ge q_2$. In addition, $g(\pi) = b\pi$ for $\pi \le q_1$ and $g(\pi) = b(1-\pi)$ for $\pi \ge q_2$. If $\pi \le q_1$, one accepts H_0 without carrying out observations. If $\pi \ge q_2$, one accepts H_1 without carrying out observations. These considerations permit one to construct a Bayesian decision rule as follows. Assume that a sample $x_1, \dots, x_n$ has been drawn but no decision has been made up to the n-th step. Then the conditional probability of H_1, given the sample, is

$$\pi(x_1, \dots, x_n) = \frac{f_1(x_1) \dots f_1(x_n)\pi}{f_1(x_1) \dots f_1(x_n)\pi + f_0(x_1) \dots f_0(x_n)(1-\pi)} \, .$$

When $\pi(x_1, \dots, x_n) \le q_1$, one accepts H_0, when $\pi(x_1, \dots, x_n) \ge q_2$, one accepts H_1 and when $q_1 < \pi(x_1, \dots x_n) < q_2$, one has to carry out observations. Let

$$A = \frac{q_1(1-\pi)}{(1-q_1)\pi}, \quad B = \frac{q_2(1-\pi)}{(1-q_2)\pi}, \quad \rho_n = \prod_{k=1}^{n} \frac{f_1(x_k)}{f_0(x_k)} \, .$$

Then the Bayesian decision rule is this: τ is the first time when $\rho_n \notin (A, B)$; accept H_0 if $\rho_n \le A$ and accept H_1 if $\rho_n \ge B$.

2

Controlled Stochastic Processes

We constantly encounter the need to control both production processes and processes permeating public life, particularly those of science, nature and society. The theory of controlled processes studies models enabling one to make a quantitative assessment of the effects of controlling a process. In controlled stochastic processes, the course of a process depends on chance. Thus a control is estimated via the averages of its quantitative parameters. The most typical example are automatic control systems, which are studied taking possible random disorders into account. This chapter concentrates on a few problems in the theory of controlled stochastic processes that enable one to obtain an idea of its basic notions and the techniques it employs.

2.1 Controlled Random Sequences

We first consider a *time-discrete controlled stochastic process*. It is a simpler version of the process both in terms of defining it and in formulating and solving the problem.

Let $(X, \mathcal{B})$ and $(U, \mathcal{C})$ be two measurable spaces. The first is the phase space of the process and the second is the phase space of the control. In order to define a *controlled stochastic process*, we first consider the degenerate case in which there is no randomness. The process manifests itself outwardly through two sequences: its states $\{x_n, n \geq 0\}$ in X and its controls $\{u_n, n \geq 0\}$ in U. The controls are chosen arbitrarily and they determine the states of the controlled process after the initial position. To determine the state at time t, it is necessary to know only the controls at the preceding moments of time. Thus a controlled process is described by a sequence of functions

$$x_1 = f_1(x_0, u_0), x_2 = f_2(x_0, u_0, u_1), \ldots, x_n = f_n(x_0, u_0, u_1, \ldots, u_{n-1}) \, .$$

We point out that the more standard definition in which the state at time t is determined by the previous states and controls before t obviously reduces to the one stated above.

In the case of a controlled stochastic process, we shall view the distribution of its state at time t as depending on the previous states and the controls before t. Therefore such a process is specified by a sequence of conditional distributions

$$p_n(B|x_0, x_1, \dots, x_{n-1}, u_0, u_1, \dots, u_{n-1}), \quad n = 1, 2, \dots .$$

This is the probability that a state belongs to $B \in \mathcal{B}$ at time $t = n$, given that the respective previous states and controls were $x_0, \dots, x_{n-1}$ and $u_0, \dots, u_{n-1}$ (the distribution of the state at time $t = 0$ is $p_0(B)$). We assume that the functions p_n depend $B^n \otimes C^n$-measurably on their arguments. The problem generally boils down to selecting an "optimum" control. The next subsection is devoted to making the meaning of this more precise.

2.1.1 Statement of the Problem

Suppose that the controlled process operates over a finite segment of time: $t = 0, 1, 2, \dots, T$. The purpose of the control is usually to make a profit, or achieve a certain least expensive result or to reduce anticipated losses. The profit, losses and expenses each depend on the values of the process and the controls employed. A profit can be viewed as the negative of a loss and a loss can be viewed as an expense in running the process. Therefore a function $F_N(x_0, \dots, x_N, u_0, \dots, u_{N-1})$ is assumed to be given which characterizes these expenses if the controls were $u_0, \dots, u_{N-1}$ and the states were $x_0, \dots, x_N$. An optimum control is one that minimizes the *control cost*. Suppose that controls $u_0, \dots, u_{N-1}$ have been selected. Then the joint distribution of $\xi_0, \xi_1, \dots, \xi_N$, where ξ_k is a random element in X representing the state of the process at time k, is

$$\begin{aligned}
&\mathbf{P}(\xi_0 \in A, \dots, \xi_N \in A_N) \\
&\quad = \int_{A_0} p(dx_0) \int_{A_1} p(dx_1|x_0, u_0) \dots \int_{A_n} p(dx_N|x_0, \dots, x_{N-1}, u_0, \dots, u_{N-1}) \\
&\quad = p_N(A_0, \dots, A_N|u_0, u_1, \dots, u_{N-1}) .
\end{aligned}$$

The average control cost,

$$\begin{aligned}
\bar{F}_N(u_0, \dots, u_{N-1}) = \int & F_N(x_0, \dots, x_N, u_0, \dots, u_{N-1}) \\
&\times p_N(dx_0, \dots, dx_N|u_0, \dots, u_{N-1}) ,
\end{aligned}$$

is a function of the controls employed. If the pre-selected controls are used, then the optimum control will be the one that minimizes $\bar{F}_N$. However, it is possible to improve the control substantially if it is chosen to depend on the states.

Example. Let $X = R, U = R, x_n = x_{n-1} + \eta_n + u_{n-1}, x_0 = \eta_0$, where $\eta_0, \eta_1, \ldots$ is a sequence of independent and identically distributed random variables with $\mathbf{E}\eta_0 = 0, \mathbf{V}\eta_0 = 1$, and $F_N(x_0, \ldots, x_N, u_0, \ldots, u_{N-1}) = x_N^2$. Then

$$\bar{F}_N(u_0, \ldots, u_{N-1}) = \mathbf{E}\left(\sum_{k=0}^{N} \eta_k + \sum_{k=0}^{N-1} u_k\right)^2 = (N+1) + \left(\sum_{k=0}^{N-1} u_k\right)^2 ,$$

and so $\min \bar{F}_N = N + 1$. If the control at the k-th step depends on the states up to time k inclusively, then we can write

$$\mathbf{E}\left(\sum_{k=0}^{N} \eta_k + \sum_{k=0}^{N-1} u_k\right)^2 = \mathbf{E}\eta_N^2 + \mathbf{E}\left(\sum_{k=0}^{N-1} (\eta_k + u_k)\right)^2 .$$

Choosing the control so that $u_k = -\eta_k$, we wind up with a control of average cost one, which cannot be reduced any further.

(a) *Control strategy.* To refine the way a control is chosen, we introduce a *control strategy.* It is reasonable to assume that a control is independent of future states and controls. But it can be random. Denote by $q_n(C|x_0, x_1, \ldots, x_n, u_0, u_1, \ldots, u_{n-1})$ the conditional probability that the control takes a value in the set $C \in \mathcal{C}$ at time n, given that the preceding controls were $u_0, \ldots, u_{n-1}$ and the states up to time n inclusively were $x_0, x_1, \ldots, x_n$. It is viewed as depending $\mathcal{B}^{n+1} \otimes \mathcal{C}^n$-measurably on its arguments. The family of functions $\{q_n(C|\ldots, \ldots), n = 0, 1, 2, \ldots\}$ is then a control strategy. Such a strategy is said to be *randomized.* Assume that for every n the measure $q_n(\cdot|\cdot)$ is concentrated at a single point. Let it be the point $u_n = g_n(x_0, x_1, \ldots, x_n, u_0, u_1, \ldots, u_{n-1})$. Then $u_0 = g_0(x_0), u_1 = g_1(x_0, x_1, g_0(x_0)) = \hat{g}_1(x_0, x_1), \ldots, u_n = \hat{g}_n(x_0, \ldots, x_n)$. Control strategies of this form are said to be *nonrandomized.* The functions $\hat{g}_i(x_0, \ldots, x_i)$ are also $\mathcal{B}^{i+1}$-measurable.

Giving a controlled process and a control strategy determines a sequence (ξ_n, η_n) in the space $X \times U$ with finite-dimensional distributions

$$\begin{aligned} &\mathbf{P}\{\xi_0 \in A_0, \eta_0 \in C_0, \xi_1 \in A_1, \eta_1 \in C_1, \ldots, \xi_n \in A_n, \eta_n \in C_n\} \\ &= \int_{A_0} p_0(dx_0) \int_{C_0} q_0(du_0|x_0) \ldots \int_{A_n} p_n(dx_n|x_0, \ldots, x_{n-1}, \qquad (2.1.1) \\ &u_0, \ldots, u_{n-1}) \int_{C_n} q_n(du_n|x_0, \ldots, x_n, u_0, \ldots, u_{n-1}) . \end{aligned}$$

This sequence will be called the control process corresponding to the controlled process $\{p_n(\cdot|\cdot), n = 0, 1, \ldots\}$ and control strategy $\{q_n(\cdot|\cdot), n = 0, 1, \ldots\}$. It follows from (2.1.1) that the conditional distribution of ξ_n, given $\xi_0, \eta_0, \ldots, \xi_{n-1}, \eta_{n-1}$, is $p_n(\cdot|\xi_0, \ldots, \xi_{n-1}, \eta_0, \ldots, \eta_{n-1})$, and the conditional distribution of η_n, given $\xi_0, \eta_0, \ldots, \xi_{n-1}, \eta_{n-1}$, is $q_n(\cdot|\xi_0, \ldots, \xi_n, \eta_0, \ldots, \eta_{n-1})$.

The sequences $\{\xi_n\}$ and $\{\eta_n\}$ are the respective states and controls of the controlled process. In what follows, we shall consider the controlled process (that is, the collection $\{p_n(\cdot|\cdot)\}$) and the control cost F_N fixed but the control strategy may change by choosing it so that the control cost is as far as possible less. Strategies will be denoted by the same letter (say S), being identified with a collection $\{q_n(\cdot|\cdot)\}$ by the notation $S = \{q_n(\cdot|\cdot)\}$. $\mathbf{P}_S$ and $\mathbf{E}_S$ will denote the probability and expectation involving the control process $\{(\xi_n, \eta_n)\}$ if strategy S has been chosen.

A *strategy* $S = \{q_n\}$ is *optimum* over $[0, N]$ if

$$\mathbf{E}_S F_N(\xi_0, \eta_0, \ldots, \xi_N, \eta_N) \leqslant \mathbf{E}_{S'} F_N(\xi_0, \eta_0, \ldots, \xi_N, \eta_N)$$

for any other strategy $S' = \{q'_n\}$. A strategy S_ε is *ε-optimum* if

$$\mathbf{E}_{S_\varepsilon} F_N(\xi_0, \eta_0, \ldots, \xi_N, \eta_N) \leqslant \mathbf{E}_{S'} F_N(\xi_0, \eta_0, \ldots, \xi_N, \eta_N) + \varepsilon$$

for any other strategy S'. To find optimum and ε-optimum strategies, it is necessary to know the quantity

$$\bar{F}_N = \inf \mathbf{E}_S F_N(\xi_0, \eta_0, \ldots, \xi_N, \eta_N) \,,$$

which is called the control cost. A basic problem involving controlled processes is to look for ε-optimum controls for any positive ε.

2.1.2 Optimum and ε-Optimum Controls

We shall concentrate on the simplest case where $N = 1$ and F_1 does not depend on u_1. Let us show how to find the control cost. The strategy S is given by the distribution $q_0(du_0|x_0)$. We have

$$\begin{aligned}\mathbf{E}_S F_1(\xi_0, \eta_0, \xi_1) &= \int F_1(x_0, u_0, x_1) p_0(dx_0) q_0(du_0|x_0) p_1(dx_1|x_0, u_0) \\ &= \int p_0(dx_0) \int q_0(du_0|x_0) \left[\int F_1(x_0, u_0, x_1) p_1(dx_1|x_0, u_0) \right] .\end{aligned}$$

Put $\int F_1(x_0, u_0, x_1) p_1(dx_1|x_0, u_0) = \bar{F}_1(x_0, u_0)$. This function is independent of the choice of strategy and

$$\int q_0(du_0|x_0) \bar{F}_1(x_0, u_0) \geqslant \inf_{u_0} \bar{F}_1(x_0, u_0) \,.$$

Under very broad assumptions about the nature of $\bar{F}_1(x_0, u_0)$, it is possible to choose a strategy so that

$$\int \left[\int q_0(du_0|x_0) \bar{F}_1(x_0, u_0) - \inf_{u_0} \bar{F}_1(x_0, u_0) \right] p_0(dx_0)$$

is arbitrarily small. Hence

$$\bar{F}_1 = \int q_0(dx_0) \inf_{u_0} \int F_1(x_0, u_0, x_1) p_1(dx_1|x_0, u_0) \ .$$

Observe that if F_1 depends on u_1, then the control cost for this function will be the same as for $\inf_{u_1} F_1(x_0, u_0, x_1, u_1)$ since the control u_1 may be chosen so that the value of $F_1(x_0, u_0, x_1, u_1)$ is arbitrarily close to $\inf_{u_1} F_1(x_0, u_0, x_1, u_1)$. Thus in this case

$$\bar{F}_1 = \int q_0(dx_0) \inf_{u_0} \int p_1(dx_1|x_0, u_0) \inf_{u_1} F_1(x_0, u_0, x_1, u_1) \ .$$

Consider the sequence of functions

$$\begin{aligned}
&F'_N(x_0, u_0, \ldots, x_{N-1}, u_{N-1}, x_N) = \inf_{u_N} F_N(x_0, u_0, \ldots, x_N, u_N) \\
&F_{N-1}(x_0, u_0, \ldots, x_{N-1}, u_{N-1}) \qquad (2.1.2) \\
&\quad = \int F'_N(x_0, u_0, \ldots, x_{N-1}, u_{N-1}, x_N) p_N(dx_N|x_0, \ldots, u_{N-1}) \\
&F'_n(x_0, u_0, \ldots, u_{n-1}, x_n) = \inf_{u_n} F_n(x_0, u_0, \ldots, x_n, u_n) \\
&F_{n-1}(x_0, \ldots, u_{n-1}) = \int F'_n(x_0, u_0, \ldots, u_{n-1}, x_n) p_n(dx_n|x_0, \ldots, u_{n-1})
\end{aligned}$$

for $n < N$. All of these functions are assumed to be well defined. Then it is natural to expect the control cost to be

$$\bar{F}_N = \int F'_0(x_0) p_0(dx_0) \ . \qquad (2.1.3)$$

For this to be so, it is required that each step in the chain (2.1.2) resuIt in a measurable function and that

$$\begin{aligned}
&\inf_S \int F_n(x_0, u_0, \ldots, x_n, u_n) q_n(du_n|x_0, \ldots, x_n, u_0, \ldots, u_n) \\
&\qquad = \inf_{u_n} F_n(x_0, u_0, \ldots, x_n, u_n) \ . \qquad (2.1.4)
\end{aligned}$$

We point out that when $\min_{u_n} F_n(x_0, u_0, \ldots, x_n, u_n)$ exists for all $n \le N$ and there is a measurable function $g_n(x_0, u_0, \ldots, x_n)$ with values in U such that

$$\begin{aligned}
&\min_{u_n} F_n(x_0, u_0, \ldots, x_n, u_n) \\
&\qquad = F_n(x_0, u_0, \ldots, x_n, g_n(x_0, u_0, \ldots, x_n)) \ , \qquad (2.1.5)
\end{aligned}$$

then the sequence $u_n = g_n(x_0, u_0, \ldots, x_n)$ determines a *nonrandomized optimum control*. Two theorems are given below on optimum and ε-optimum controls.

(a) *Existence and form of optimum controls.* X and U are assumed to be complete separable metric spaces. Let Z be a topological space. $\mathbf{C}_Z$ is the space of real bounded continuous functions on Z with norm $\|f\| = \sup_z |f(z|$. A controlled process defined by the collection $\{p_n(\cdot|\cdot)\}$ will be said to satisfy a weak-continuity condition if for all n and $f(x) \in \mathbf{C}_X$

$$\int f(z)p_n(dz|x_0, u_0, \ldots, x_{n-1}, u_{n-1}) \in \mathbf{C}_{X^n \times U^n} \ . \tag{2.1.6}$$

Recall that a real function $f(z)$ defined on a topological space Z is lower semicontinuous if for all $z_0 \in Z$

$$\varliminf_{z \to z_0} f(z) \geqslant f(z_0) \ .$$

Below are some properties of lower semicontinuous functions.

I. If U is compact and $f(u)$ is lower semicontinuous, then $f(u)$ is bounded from below and $\min_u f(u)$ exists.

II. Let $f(z, u)$ be lower semicontinuous on $Z \times U$, with Z a topological space and U a compact space. Define $g(z) = \min_u f(z, u)$ (by property I it is defined everywhere). Then $g(z)$ is lower semicontinuous.

III. Let Z be a complete separable metric space, U a compact space, $f(z, u)$ lower semicontinuous and $g(z) = \min_u f(z, u)$. There exists a Borel function $\varphi(z)$ from Z to U such that $g(z) = f(z, \varphi(z))$ (this is one of the versions of the theorem on measurable selection).

Let us clarify this statement for the case where U is an interval $[a, b]$ of the real line. Then the set $\Delta_z = \{u : f(z, u) = g(z)\}$ is closed, $\Delta_z \subset [a, b]$ and $\varphi(z)$ may be taken to be $\inf \Delta_z$.

IV. Every lower semicontinuous function $g(z)$ bounded from below and defined on a complete separable metric space is the limit of an increasing sequence of continuous functions.

Theorem 2.1.1. *Let X be a complete separable metric space and U a compact space. If a controlled process satisfies the weak-continuity condition and the control cost F_N is lower semicontinuous and bounded from below on $X^{n+1} \times U^{n+1}$, then:* 1. *all of the functions defined in* (2.1.2) *are lower semicontinuous;* 2. *the control cost is given by* (2.1.3)*;* 3. *there exists a sequence of Borel functions*

$$u_n = g_n(x_0, x_1, \ldots, x_n, u_0, u_1, \ldots, u_{n-1})$$

satisfying relation (2.1.5). *These functions determine a nonrandomized optimum control.*

Proof. 1. If F_n is lower semicontinuous, then the lower semicontinuity of F'_n follows from I. If in addition F_n is bounded from below, then so will F'_n be

bounded from below. When the weak-continuity condition holds, it is possible to show that

$$\int f(x_0,\ldots,x_{n-1},u_0,\ldots,u_{n-1})p_n(dx|x_0,\ldots,x_{n-1},u_0,\ldots,u_{n-1}) \in \mathbf{C}_{X^n\times U^n}$$

for all $f \in \mathbf{C}_{X^n\times U^n}$. Applying IV, we find from this that $F_{n-1}(x_0,\ldots,u_{n-1})$ is lower semicontinuous if $F_n(x_0,u_0,\ldots,u_{n-1},x_n)$ is the same. Statements 2 and 3 are consequences of relation (2.1.5) and the latter follows from property III. □

(b) *ε-optimum controls.* We need the notion of analytic set in what follows.

Let Z be a complete separable metric space. A set $A \subset Z$ is *analytic* if there is a compact set U and a Borel set $B \subset Z \times U$ such that A is the projection of B on Z, that is, $A = \{z : \exists u \in U, (z,u) \in B\}$.

We state without proof some facts about analytic sets.

Al. The collection of analytic subsets of Z (which we denote by $\mathcal{A}_Z$) forms a monotone class which is closed under the operations of union and intersection.

A2. If X and Z are complete separable metric spaces and f is a continuous function from X to Z, then $f(A) \in \mathcal{A}_Z$ when $A \in \mathcal{A}_Z$.

A3. The completion of a Borel measure on Z is well defined on $\mathcal{A}_Z$.

Let $\mathcal{A}_Z^-$ denote the set of scalar functions on Z for which $\{z : f(z) < \lambda\} \in \mathcal{A}_Z$ for all $\lambda \in R$.

A4. If $f \in \mathcal{A}_{Z\times U}^-$ and $\inf > -\infty$, then $\inf_u f(z,u) \in \mathcal{A}_Z^-$.

A5. If $f \in \mathcal{A}_{Z\times U}^-$, then $f(z,\bar{u}) \in \mathcal{A}_Z^-$ for all $\bar{u} \in U$.

A6. Let X and Z be complete separable metric spaces and let $p(A,z)$ be a measure on $\mathcal{B}_X$ for all $z \in Z$ ($\mathcal{B}_X$ is a Borel σ-algebra of X). Let $p(A,\cdot) \in \mathcal{A}_Z^-$ for all $A \in \mathcal{B}_X$ and let $f \in \mathcal{A}_{Z\times U}^-$ be non-negative. Then

$$\int f(x,z)p(dx,z) \in \mathcal{A}_Z^- \,. \tag{2.1.7}$$

A7. (Theorem an Measurable Selection). Suppose that $g \in A_{Z\times U}^-$, $\inf g > -\infty$ and $\min_u g(z,u) = g(z)$ exists for all $z \in Z$. Then there exists a Borel function $u = \varphi(z)$ such that

$$g(z) = g(z,\varphi(z)) \,.$$

Theorem 2.1.2. *Suppose that X and U are complete separable metric spaces, $p_n(A|x_0,\ldots,u_{n-1})$ belongs to $\mathcal{A}_{X^n\times U^n}^-$ for all $A \in \mathcal{B}_x$, $F_N \in \mathcal{A}_{X^{N+1}\times U^{N+1}}^-$ and $F_N > 0$. Then:* 1. *the functions F_n' and F_{n-1} defined by* (2.1.2) *belong to $\mathcal{A}_{X^{n+1}\times U^n}^-$ and $\mathcal{A}_{X^n\times U^n}^-$, respectively;* 2. *the control cost is given by* (2.1.3)*;* 3. *to every $\varepsilon > 0$ there exists a sequence of Borel functions*

$$u_n = g_n^\varepsilon(x_0,\ldots,u_{n-1},x_n) \,,$$

which determine a nonrandomized ε-optimal control.

Proof. Statement 1 follows from A4 and A6. The fact that the control cost is given by (2.1.3), provided that the functions in the chain (2.1.2) are measurable, was established above (in our case, they are measurable with respect to the completion of the Borel σ-algebra in any measure). Let us prove statement 3. It suffices to show that to any positive δ there is a Borel function $\varphi_n(x_0, \dots, u_{n-1}, x_n)$ with values in U such that

$$\inf_{u_n} F_n(x_0, u_0, \dots, x_n, u_n) > F_n(x_0, u_0, \dots x_n, \varphi_n(x_0, \dots, u_{n-1}, x_n)) - \delta \,. \tag{2.1.8}$$

Let $F_n^\delta = k\delta$ if $k\delta \leq F_n < (k+1)\delta$. Clearly, $F_n^\delta \in \mathcal{A}^-_{X^{n+1}\times U^{n+1}}$. The $\min_{u^n} F_n^\delta(x_0, u_0, \dots, x_n, u_n)$ exists for all $x_0, \dots, u_{n-1}$, x_n. On the basis of A7, there is a Borel function $u_n = \varphi_n(x_0, \dots, u_{n-l}, x_n)$ for which

$$\min_{u_n} F_n^\delta(x_0, \dots, x_n, u_n) = F_n^\delta(x_0, \dots, x_n, \varphi_n(x_0, \dots, u_{n-1}, x_n)) \,.$$

Consequently,

$$\inf_{u_n} F_n \geqslant \min_{u_n} F_n^\delta = F_n^\delta(x_0, \dots, x_n, \varphi_n) > F_n(x_0, \dots, x_n, \varphi_n) - \delta \,.$$

Observe that

$$F_n(x_0, \dots, x_n, \varphi_n(x_0, \dots, u_{n-1}, x_n)) \in \mathcal{A}^-_{X^{n+1}\times U^n} \,.$$

This is true since the set

$$\{(x_0, \dots, x_n, u_0, \dots, u_{n-1}) \ : \ F_n(x_0, \dots, x_n, u_n) < \lambda\}$$

in $X^{n+1} \times U^{n+1}$ is the pre-image of the analytic set

$$\{(x_0, \dots, x_n, u_0, \dots, u_n) \ : \ F_n(x_0, \dots, x_n, u_n) < \lambda\}$$

under the Borel mapping of $X^{n+1} \times U^n$ into $X^{n+1} \times U^{n+1}$ defined by $x_i = x_i, i \leq n, u_i = u_i, i \leq n-1$, and $u_n = \varphi_n(x_0, \dots, u_{n-1}, x_n)$. At each step choose control $u_n = \varphi_n(x_0, \dots, u_{n-1}, x_n)$ and let

$$\begin{aligned}
&\tilde{F}_N = F_N \\
&\tilde{F}'_n(x_0, \dots, u_{n-1}, x_n) \\
&\quad = \tilde{F}_n(x_0, \dots, u_{n-1}, x_n, \varphi_n(x_0, u_0, \dots, x_n)), \quad n \leqslant N \,, \\
&\tilde{F}_{n-1}(x_0, \dots, x_{n-1}, u_{n-1}) \\
&\quad = \int \tilde{F}'_n(x_0, \dots, u_{n-1}, x_n) p_n(dx_n | x_0, \dots, u_{n-1}) \,.
\end{aligned}$$

Then by induction, one can establish that

$$\begin{aligned}
&\tilde{F}'_n(x_0, \dots, u_{n-1}, x_n) < F'_n(x_0, \dots, u_{n-1}, x_n) + (N-n+1)\delta, \\
&\tilde{F}_n(x_0, \dots, u_{n-1}) < F_n(x_0, \dots, u_{n-1}) + (N-n)\delta \,.
\end{aligned}$$

Therefore the cost of the chosen control satisfies

$$\int \tilde{F}'_0(x_0)p_0(dx_0) < \bar{F}_N + (N+1)\delta \ .$$

Taking $\delta = \varepsilon/(N+1)$, we complete the proof of statement 3. □

2.2 Controlled Markov Chains

As in the previous section, $(X, \mathcal{B})$ is the phase space of the process and $(U, \mathcal{C})$ is the control phase space. A controlled process is Markov (a controlled Markov chain) if the function $p_n(A|x_0, \ldots . x_{n-1}, u_0, \ldots, u_{n-1})$ depends only on x_{n-1} and u_{n-1}. A controlled Markov chain is specified by giving its transition probabilities $\{p_n(A|x_{n-1}, u_{n-1}), n = 1, 2, \ldots\}$. In contrast to the common case, the initial distribution is not kept fixed (this is the usual approach in studying Markov processes). It turns out that the special form of the conditional probabilities defining the controlled process does not simplify the chain of relations (2.1.2) which are used to find the control cost and optimum (or ε-optimum) controls. However, under the assumption that the control expense is the sum of the control expenses at each step which, in turn, depend only on the initial and final states of the process and the control selected, there is a more efficient way to find an optimum (or ε-optimum) control.

2.2.1 Additive Control Cost. Bellman's Equation

We shall assume for all $n \geq 0$ that we are given a function $f_n(x, u, x')$ specifying the expenses due to the control at the n-th step if the process was in state x and wound up in state x' after the control was applied. The total expense over the interval $[0, N]$, if the process was in the states $x_0, x_1, \ldots, x_N$ and the controls were $u_0, u_1, \ldots, u_{N-1}$, is

$$F_N(x_0, u_0, \ldots, x_{N-1}, u_{N-1}, x_N) = \sum_{n=0}^{N-1} f_n(x_n, u_n, x_{n+1}) \ . \tag{2.2.1}$$

Let $V_0^N(x)$ be the control cost for this process (see Sect. 2.1.1(a)) under the assumption that its initial position x_0 coincides with x. Now consider the same controlled Markov chain except beginning at time $k < N$. The control cost for this chain is

$$F_{k,N}(x_k, u_k, \ldots, u_{N-1}, x_N) = \sum_{n=k}^{N-1} f_n(x_n, u_n, x_{n+1}) \ . \tag{2.2.2}$$

Let $V_k^N(x)$ be the control cost for this process under the assumption that the initial position x_k coincides with x. *Bellman's equation* relates the $V_k^N(x)$ for different $k < N$ and makes it possible to determine them recursively.

We assume that X and U are complete separable metric spaces, the functions $f_n(x,u,x') \in \mathcal{A}^-_{X\times U\times X}$ and are bounded from below, and $p_n(A|x_n,u_n) \in \mathcal{A}^-_{X\times U}$ for all closed sets $A \in \mathcal{B}_X$.

Theorem 2.2.1. *For $k < N$,*

$$V_k^N(x) = \inf_u \int [f_k(x,u,x') + V_{k+1}^N(x')]p_k(dx'|x,u)\ , \tag{2.2.3}$$

with $V_N^N(x) = 0$ by assumption.

Proof. Let $\tilde{V}_N^N(x) = 0$, and for $k < N$, let $\tilde{V}_k^N(x)$ be functions determined recursively by (2.2.3). Applying formula (2.1.2) to the functions F_N of the form (2.2.1) and the relation (2.2.3) for $\tilde{V}_k^N(x)$, we find that

$$F'_{N-1}(x_0,u_0,\ldots,x_{N-1}) = \sum_{k=0}^{N-2} f_k(x_k,u_k,x_{k+1}) + \tilde{V}_{N-1}^N(x_{N-1})\ .$$

Furthermore,

$$\begin{aligned} F'_{N-2} &= \sum_{k=0}^{N-3} f_k(x_k,u_k,x_{k+1}) + \inf_{u_{N-2}} \int \Big[f_{N-2}(x_{N-2},u_{N-2},x_{N-1}) \\ &\quad + \tilde{V}_{N-1}^N(x_{N-1})\Big] p_N(dx_{N-1}|x_{N-2},u_{N-2}) \\ &= \sum_{k=0}^{N-3} f_k(x_k,u_k,x_{k+1}) + \tilde{V}_{N-2}^N(x_{N-2})\ . \end{aligned}$$

Continuing in this fashion, we find that $F'_0(x) = \tilde{V}_0^N(x_0)$ and so $V_0^N(x) = \tilde{V}_0^N(x)$. By similarly considering the process on $[k,N]$, we can establish that $V_k^N(x) = \tilde{V}_k^N(x)$. □

Remark 2.2.1. Assume that the infimum in (2.2.3) is attained for all k and x. Then on the basis of statement A7 on p. 221, there exists a Borel function $g_k(x)$ from X to U such that

$$V_k^N(x) = \int \Big[f_k(x,g_k(x),x') + V_{k+1}^N(x')\Big]p_k(dx'|x,g_k(x))\ . \tag{2.2.4}$$

The sequence $u_k = g_k(x_k)$ determines a nonrandomized optimum control.

Remark 2.2.2. If the functions $f_k(x,u,x')$ are lower semicontinuous and bounded from below, U is a compact set and the transition probabilities satisfy the weak-continuity condition, then all of the functions $V_k^N(x)$ and $\int [f(x,u,x') + V_{k+1}^N(x')]p_k(dx'|x,u)$ are lower semicontinuous. Thus the existence of Borel functions $g_k(x)$ satisfying (2.2.4) follows from III on p. 220.

Definition. A nonrandomized control of the form $\{u_k = g_k(x)\}$ is said to be Markov (the corresponding strategy is also said to be Markov).

Thus Remarks 2.2.1 and 2.2.2 give conditions for a Markov optimum control to exist.

Remark 2.2.3. If for all k a function $g_k^{\delta_k}(x)$ is selected so that

$$V_k^N(x) + \delta_k > \int \Big[f_k(x, g_k^{\delta_k}(x), x') + V_{k+1}^N(x') \Big] p_k(dx'|x, g_k^{\delta_k}(x)) \tag{2.2.5}$$

(the existence of such Borel functions follows from statement A7 on p. 221), then the sequence $u_k = g_k^{\delta_k}(x_k), k = 0, 1, \ldots, N-1$, determines an ε-optimum Markov control provided $\delta_0 + \ldots + \delta_{N-1} < \varepsilon$. This is proved in exactly the same way as in Theorem 2.1.2.

2.2.2 Optimum Stopping of a Markov Chain

Consider a Markov chain in the phase space $(X, \mathcal{B})$ with n-step transition probability $p_n(x, A), n = 0, 1, 2, \ldots$. Denote by $\{\xi_n, n = 0, 1, 2, \ldots\}$ a realization of this chain. Control with the chain consists in choosing a stopping time τ. The control expense is given by a $\mathcal{B}$-measurable function $F(x)$ and if the chain is stopped at time τ, then the expense is $F(\xi_\tau)$. As before, it is necessary to determine a control (that is, a stopping time) which makes the average loss $\mathbf{E}F(\xi_\tau)$ as small as possible.

(a) *Equations for the control cost.* First assume that the process is considered only for $t \le N$. Then the stopping times τ also take values in $[0, N]$ (these are the times for which the events $\{\tau = k\}$ are determined by $\{\xi_0, \ldots, \xi_k\}$). Let $U_k^N(x)$ be the cost of the control by the Markov chain $\xi_k, \ldots, \xi_N$, given that $\xi_k = x$. In order that it be well defined, we shall assume that $F(x)$ is bounded from below. Let τ be any control for the chain $(\xi_k, \ldots, \xi_N)$. The cost of this control, given that $\xi_k = x$, is

$$F(x)\mathbf{P}\{\tau = k|\xi_k = x\} + \mathbf{E}(F(\xi_\tau)|\xi_k = x,\ \tau > k)\mathbf{P}\{\tau > k|\xi_k = x\}\ .$$

If $\mathbf{P}\{\tau > k|\xi_k = x\}$ is kept fixed, then $\mathbf{E}(F(\xi_\tau)|\xi_k = x, \tau > k) = \mathbf{E}(\mathbf{E}(F(\xi_{\tau'})|\xi_{k+1})|\xi_k = x)$ where $\tau' = \tau$ is a stopping time on the set $\tau > k$ for the chain $(\xi_{k+1}, \ldots, \xi_N\}$. The infimum of this expression is obviously $\mathbf{E}(U_{k+1}^N(\xi_{k+1})|\xi_k = x) = \int U_{k+1}^N(x')p_k(x, dx')$. It remains to minimize the expression

$$F(x)\mathbf{P}\{\tau = k\}|\xi_k = x\} + (1 - \mathbf{P}\{\tau = k|\xi_k = x\}) \int U_{k+1}^N(x)p_k(x, dx')$$

via a suitable choice of $\mathbf{P}\{\tau = k|\xi_k = x\}$. The minimum will be attained if this probability is either 0 or 1. Hence

$$U_k^N(x) = F(x) \wedge \int p_k(x, dx') U_{k+1}^N(x') \ . \tag{2.2.6}$$

Since $U_N^N(x) = F(x)$, relation (2.2.6) determines the functions $U_k^N(x)$. Let us show how an optimum control may be constructed (and thereby prove its existence). Suppose that all of the functions $U_k^N(x)$ have been formed. Let

$$\bar{U}_k^N(x) = \int U_{k+1}^N(x') p_k(x, dx') \ . \tag{2.2.7}$$

From the derivation of (2.2.6), it follows that if the process is unstopped before moment k, then it must stop at that moment if $F(\xi_k) \leq \tilde{U}_k^N(\xi_k)$ and continue for $F(\xi_k) > \tilde{U}_k^N(\xi_k)$. Thus τ is the first time for which $F(\xi_k) \leq \tilde{U}_k^N(\xi_k)$, $k \leq N$ (we take $\tilde{U}_N^N(x) = U_N^N(x) = F(x)$).

In the general case, we shall assume that the stopping times are only finite, that is, $\mathbf{P}\{\tau < \infty\} = 1$. If $U_k(x)$ is the control cost for the chain $(\xi_k, \xi_{k+1}, \ldots)$, given that $\xi_k = x$, then $U_k(x) = \lim_{N\to\infty} U_k^N(x)$. This limit exists because $U_k^N(x)$ is bounded from below and monotone decreasing in N. Taking the limit in (2.2.6), we obtain

$$U_k(x) = F(x) \wedge \int p_k(x, dx') U_{k+1}(x') \ . \tag{2.2.8}$$

In order to find $U_k(x)$ one must actually utilize (2.2.6). In the present case, one no longer can assert the existence of an optimum control. But ε-optimum controls do exist. We can first determine a measurable function $N(x)$ so that

$$U_0^{N(x)}(x) \leqslant U_0(x) + \varepsilon \ .$$

If $\xi_0 = x$ is the initial value of the process, then on choosing an optimum control for the chain $(\xi_0, \ldots \xi_{N(x)})$, we arrive at an ε-optimum control for the process over an infinite time interval.

(b) *Homogeneous Markov chains.* Now suppose that the transition probability is independent of n: $P_n(x, A) = P(x, A)$. In other words, the chain is homogeneous. Then the distribution of the sequence $\{\xi_{k+n}, n = 0, 1, \ldots\}$, given $\xi_k = x$, coincides with the distribution of the sequence $\{\xi_n, n = 0, 1, \ldots\}$, given $\xi_k = x$, and so $U_k(x)$ is independent of k. Write $U_k(x) = U(x)$. Relation (2.2.8) leads to the following equation for $U(x)$:

$$U(x) = F(x) \wedge \int U(x') P(x, dx') \ . \tag{2.2.9}$$

We now introduce some notions associated with a homogeneous Markov chain. Let Pf be the operator defined by $Pf(x) = \int f(x') P(x, dx')$ for all f for which $\int [0 \wedge (-f(x'))] P(x, dx') < \infty$. A function f is called *harmonic* if $Pf = f$, *superharmonic* if $Pf \leq f$ and *subharmonic* if $Pf \geq f$.

Theorem 2.2.2. *$U(x)$ is the maximal subharmonic function satisfying the inequality $U(x) \leq F(x)$.*

Proof. Corresponding to stopping time τ, denote by τ' the stopping time for the chain $(\xi_0', \xi_1', \ldots, \xi_n')$, with $\xi_n' = \xi_{n+1}$, constructed just like τ for the chain $(\xi_0, \xi_1, \ldots)$: the event $\{\tau' = n\}$ is also expressible exactly in terms of $\xi_1, \ldots, \xi_{n+1}$ as $\{\tau = n\}$ is in terms of $\xi_0, \ldots, \xi_n$. Then $1+\tau'$ will be a stopping time for $\{\xi_n\}$. Therefore

$$\begin{aligned} U(x) &\leqslant \mathbf{E}(F(\xi_{1+\tau'})|\xi_0 = x) = \mathbf{E}(\mathbf{E}(F(\xi_{1+\tau'})|\xi_1)|\xi_0 = x) \\ &= \int (\mathbf{E}F(\xi'_{\tau'})|\xi_0' = x')P(x, dx') = \int \mathbf{E}(F(\xi_\tau)|\xi_0 = x')P(x, dx') \, . \end{aligned}$$

Since for every $\varepsilon > 0$, one may specify a Markov time τ^ε such that $\mathbf{E}(F(\xi_{\tau^\varepsilon})|\xi_0 = x) \leq U(x) + \varepsilon$ for all x, it follows that

$$U(x) \leqslant \int U(x')p(x, dx') + \varepsilon \, .$$

This implies that $U(x)$ is subharmonic.

Now let $\tilde{U}(x)$ be another subharmonic function with $\tilde{U}(x) \leq F(x)$. If τ is a stopping time, $\tau \leq N$, then $\tilde{U}(\xi_\tau) \leq F(\xi_\tau)$ and

$$\mathbf{E}(\tilde{U}(\xi_\tau)|\xi_0 = x) \leqslant \mathbf{E}(F(\xi_\tau)|\xi_0 = x) \, .$$

Observe that

$$\mathbf{E}(\tilde{U}(\xi_{n+1})|\xi_0, \ldots, \xi_n) = \mathbf{E}(\tilde{U}(\xi_{n+1})|\xi_n) = P\tilde{U}(\xi_n) \geqslant \tilde{U}(\xi_n) \, .$$

Therefore $\{\tilde{U}(\xi_n), n = 0, 1, \ldots, N\}$ is a submartingale and hence

$$\mathbf{E}(\tilde{U}(\xi_\tau)|\xi_0 = x) \geqslant \tilde{U}(x)$$

or

$$\tilde{U}(x) \leqslant \mathbf{E}(F(\xi_\tau)|\xi_0 = x) \, .$$

Choosing τ so that $\mathbf{E}(F(\xi_\tau)|\xi_0{=}x) \leq U(x){+}\varepsilon$, we conclude that $\tilde{U}(x){\leq}U(x)$. □

Example. Consider an asymmetric random walk over the integers with steps ± 1. If ξ_n is the step size at time n, then

$$\mathbf{P}\{\xi_n = 1\} = p, \quad \mathbf{P}\{\xi_n = -1\} = q, \quad p + q = 1, \quad p > q \, .$$

Equation (2.2.9) becomes

$$U(x) = F(x) \wedge (pU(x+1) + qU(x-1) \, .$$

Consider the set of x where $U(x) \le F(x)$. If this inequality holds for $a < x < b$ (a, b, and x integers), then for x

$$U(x) = pU(x+1) + qU(x-1) \; .$$

If $U(a) = F(a)$ and $U(b) = F(b)$ (in other words, the interval (a, b) is maximal and cannot be enlarged), then

$$U(x) = \frac{F(b) - F(a)}{r^b - r^a} r^x + \frac{r^b(a) - r^a F(b)}{r^b - r^a} \; , \quad r = \frac{q}{p} \tag{2.2.10}$$

(this is the solution of the above difference equation with the boundary conditions). $U(x)$ may be found by taking the limit:

$$U(x) = \lim_{N \to \infty} U^N(x) \; ,$$

where U^N is the cost for a random walk over $[-N, N]$ stopping at the endpoints of the interval. The representation (2.2.10) also holds for those intervals in which $U^N(x) \le F(x)$. Therefore $U^N(x)$ may be determined as follows. Find function $\tilde{U}_1(x)$ using (2.2.10) with $a = -N$ and $b = N$. Let c_1 be the first point in moving from $-N$ to N for which $\tilde{U}_1(c_1) > F(c_1)$. Determine $\tilde{U}_2(x)$ using (2.2.10) for $x \in [-N, c_1]$ taking $a = -N$ and $b = c_1$ and then for $x \in [c_1, N]$ taking $a = c_1$ and $b = N$. Now let $c_2 > c_1$ be the first point where $\tilde{U}_2(c_2) > F(c_2)$. Find $\tilde{U}_3(x)$ according to (2.2.10) on the intervals $[-N, c_1], [c_1, c_2]$ and $[c_2, N]$. Continuing in this fashion, one can determine $U^N(x)$.

2.3 Time-Continuous Controlled Markov Processes

Defining a controlled process with time-continuous involves certain difficulties due to there not existing a time "preceding" t. Therefore time-continuous controlled processes cannot be specified by giving finite-dimensional distributions as in the case of the common processes. It is then more reasonable to specify a process by giving its infinitesimal parameters and to assume that they themselves depend on the control. A second difficulty arises here. It would be natural to view the control as depending on the process but the process itself is not specified until the control is given. Thus one is caught in circular reasoning. To circumvent it, we shall consider step-controls. A process with such a control can then be defined. We shall only examine Markov jump processes and Markov diffusion processes.

2.3.1 Jump Processes

A Markov jump process is given in phase space $(X, \mathcal{B})$ by a pair of functions $\lambda(t, x)$ and $\pi(t, x, B)$, with $t \in R_+, x \in X$ and $B \in \mathcal{B}$, satisfying

$$\lambda(t,x) = \lim_{h\downarrow 0} \frac{1}{h} P(t,x,t+h, X\setminus\{x\}),$$
$$\pi(t,x,B) = \frac{1}{\lambda(t,x)} \lim_{h\downarrow 0} \frac{1}{h} P(t,x,t+h,B\setminus\{x\})$$

for almost all t, $P(t,x,s,B)$ being the transition probability of the process.

It is assumed that $\lambda(t,x)$ and $\pi(t,x,B)$ are $\mathcal{B}_{R_+}\otimes\mathcal{B}$-measurable and that π is a probability measure with respect to B on $\mathcal{B}$. A controlled jump process is specified by giving functions $\lambda(t,x,u)$ and $\pi(t,x,u,B)$ where $u\in U$ and $(U,\mathcal{C})$ is the measurable phase space of the controls. These functions are assumed to define a Markov jump process for any fixed $u\in U$. It is assumed in addition that they are $\mathcal{B}_{R_+}\otimes\mathcal{B}\otimes\mathcal{C}$-measurable and that $\lambda(t,x,u)$ is bounded.

(a) *Step-controls.* Let the controlled process be considered on $[0,T]$. Let $\mathbf{D}_X[0,T]$ and $\mathbf{D}_U[0,T]$ be the spaces of right-continuous step-functions defined on $[0,T]$ with respective values in X and U (X and U are regarded to be spaces with discrete topology). Every function $x(t)$ in $\mathbf{D}_X[0,T]$ is determined by its discontinuity points $0<t_1<\ldots<t_k\leq T$ and its values there $x(0)$, $x(t_1),\ldots,x(t_k)$. (A similar statement is also true for $\mathbf{D}_U[0,T]$.) Let $\mathbf{D}_X^r[0,T]$ be the subset of functions of $\mathbf{D}_X[0,T]$ having r discontinuities. A set $F\subset\mathbf{D}_X[0,T]$ will be called measurable if the image of $F\cap\mathbf{D}_X^r[0,T]$ under the mapping $x(\cdot)\to(t_1,\ldots,t_r;x(0),x(t_1),\ldots,x(t_r))$ is $\mathcal{B}_{R_+}^r\otimes\mathcal{B}^{r+1}$-measurable (the t_i's are all points of discontinuity of $x(t)$).

A mapping $S:\mathbf{D}_X[0,T]\longrightarrow\mathbf{D}_U[0,T]$, measurable relative to the above-mentioned classes of measurable sets (they are introduced in $\mathbf{D}_U[0,T]$ in similar fashion), is called a control (or strategy) if for any $t\in[0,T]$ the relation $x_1(s)=x_2(s)$ for $s<t$ implies that $S(x_1(\cdot),t)=S(x_2(\cdot),t)$, where $S(x_1(\cdot),t)$ is the result of S acting on $x(\cdot)$ as a function of t (this function belongs to $\mathbf{D}_U[0,T]$). Let us show how to construct the pair of stochastic processes $(\xi(t),\eta(t))$ from the above control and the infinitesimal parameters with $\xi(\cdot)\in D_X[0,T]$, $\eta(\cdot)\in\mathbf{D}_U[0,T]$, where $\eta(t)=S(\xi(\cdot),t)$ and

$$\mathbf{P}\{\xi(t+h)\neq\xi(t)|\xi(s),s\leqslant t\}=\lambda(t,\xi(t),\eta(t))h+o(h)\ ,$$
$$\mathbf{P}\{\xi(t+h)\neq\xi(t),\xi(t+h)\in B|\xi(s),s\leqslant t\}$$
$$=\lambda(t,\xi(t),\eta(t))\pi(t,\xi(t),\eta(t),B)h+o(h)\ .$$

The construction is done sequentially. Let x_0 be the initial state of the process. Put $\xi_0(t)=x_0$ for all $t\in[0,T]$ and $\eta_0(t)=S(\xi_0(\cdot),t)$, $\eta_0(t)$ being a nonrandom function. Form a Markov process using $\lambda(t,x,\eta_0(t))$ and $\pi(t,x,\eta_0(t))$ (more precisely, its finite-dimensional distributions). Call this process $\tilde{\xi}_1(t)$. If τ_1 is the first jump time, put $\xi_1(t)=x_0$ for $t<\tau_1$ and $\xi_1(t)=\tilde{\xi}_1(\tau_1)$ for $t\geq\tau_1$. Let $\eta_1(t)=S(\xi_1(\cdot),t)$. Clearly, $\eta_1(t)=\eta_0(t)$ when $t\leq\tau_1$. Then form the process $\tilde{\xi}_2(t)$ on $[\tau_1,T]$ keeping τ_1 and $\xi_1(\tau_1)$ fixed. If τ_2 is the first jump time for this process, put $\xi_2(t)=x_0$ for $t<\tau_1$, $\xi_2(t)=\xi_1(\tau_1)$ for $t\in[\tau_1,\tau_2[$ and $\xi_2(t)=\tilde{\xi}_2(\tau_2)$ for $t\geq\tau_2$. Define $\eta_2(t)$ on the basis of $\xi_2(t)$; it turns out that $\eta_2(t)=\eta_1(t)$ for $t\leq\tau_2$. We continue this construction in similar fashion

until $\tilde{\xi}_{k+1}(t)$ becomes constant on $[\tau_k, T]$ for some k, that is, $\tilde{\xi}_{k+1}(t) = \xi_k(t)$. Then $\xi_k(t)$ and $\eta_k(t)$ will be the desired pair of processes. It is easy to see that if $\lambda(t,x,u) \le c$, then $\mathbf{P}\{\nu \ge k\} \le A(cT)^k/k!$ is an estimate of the number ν of jumps that $\xi(t)$ has; A is a constant.

(b) *Bellman's equation.* Suppose that the control cost is given by

$$\int_0^T f(s, x(s), u(s))ds \ .$$

Here $u(s)$ is a function with values in U that specifies the control at time s, $x(s)$ is the state of the process at time s and $f(s,x,u)$ is a bounded $\mathcal{B}_{R_+} \otimes \mathcal{B} \otimes \mathcal{C}$-measurable function. Let $U(t,x)$ be the control cost over $[t,T]$, given $\xi(t) = x$:

$$U(t,x) = \inf \mathbf{E}\left(\int_t^T f(s, \xi(s), \eta(s))ds | \xi(t) = x\right) \ . \tag{2.3.1}$$

The infimum is taken over all step-controls described above. Bellman's equation is one involving $U(t,x)$ for $t \in [0,T]$.

Theorem 2.3.1. *For fixed x and A, let $\lambda(t,x,u)$ and $\pi(t,x,u,A)$ be continuous functions in t uniformly in u. Then $U(t,x)$ is differentiable in t for $t \in [0,T]$ and satisfies the equation*

$$-\frac{\partial U(t,x)}{\partial t} = \inf_u \left[f(t,x,u) + \lambda(t,x,u)\int (U(t,x') - U(t,x))\pi(t,x,u,dx')\right] \tag{2.3.2}$$

and boundary condition $U(T,x) = 0$. Under this condition, (2.3.2) *has a unique solution.*

Equation (2.3.2) is known as Bellman's equation.

Proof. The relation $U(T,x) = 0$ is a consequence of (2.3.1). Let $0 < t < t + h < T$. Let $(\xi(s), \eta(s))$ be the pair of processes formed on the basis of a step-control on $[t,T]$ with $\xi(t) = x$. Let τ denote the first jump time of $\xi(s)$ after time t. Then the strategy $\eta(s)$ is nonrandom on $[t,\tau]$; denote it by $u(s)$. By definition of $\lambda(s,x,u)$ and $\pi(s,x,u,A)$, we have

$$\mathbf{P}\{\tau > s\} = \exp\left\{-\int_t^s \lambda(s', x, u(s'))ds'\right\}$$

and

$$\mathbf{P}\{\xi(\tau) \in A | \tau = s\} = \pi(s,x,u(s),A) \ .$$

Therefore

$$\mathbf{E}\int_t^T f(s,\xi(s),\eta(s))ds = \mathbf{E}\int_t^{t+h} \exp\left\{-\int_t^s \lambda(s',x,u(s'))ds'\right\}$$
$$\times\lambda(s,x,u(s))\int \pi(s,x,u(s),dx')\left[\int_t^s f(s',x,u(s'))ds'\right.$$
$$\left.+\mathbf{E}\left(\int_s^T f(s'',\xi(s''),\eta(s''))ds''|\xi(s)=x'\right)\right]ds$$
$$+\mathbf{P}\{\tau > t+h\}\left[\int_t^{t+h} f(s,x,u(s))ds\right.$$
$$\left.+\mathbf{E}\left(\int_{t+h}^T f(s,\xi(s),\eta(s))ds|\xi(t+h)=x\right)\right] .$$

Choosing a suitable control on the intervals $[t,\tau]$ and $[\tau,T]$, we find from this last relation that

$$U(t,x) = \int_t^{t+h} \lambda(s,x,u(s))\int \pi(s,x,u(s),dx')U(s,x')ds$$
$$+\mathbf{P}\{\tau > t+h\}\left[\int_t^{t+h} f(s,x,u(s))ds + U(t+h,x)\right] + o(h)$$

(the quantity $o(h)$ is uniform in t). This implies that $|U(t,x) - U(t+h,x)| \leq c_1 h$ where c_1 is a constant. Using the continuity of λ, f and π in s, we obtain

$$U(t,x) = \int_t^{t+h} \lambda(t,x,u(s))\int \pi(t,x,u(s),dx')U(t,x')ds$$
$$+\left(1-\int_t^{t+h}\lambda(t,x,u(s))ds\right)\left[\int_t^{t+h} f(t,x,u(s))ds + U(t+h,x)\right] + o(h) ,$$

and then

$$U(t,x) - U(t+h,x) = \int_t^{t+h} f(t,x,u(s))ds \tag{2.3.3}$$
$$+\int_t^{t+h}\lambda(t,x,u(s))\int \pi(t,x,u(s),dx')(U(t,x')-U(t,x))ds + o(h) .$$

To minimize $U(t,x)$, it is clearly necessary to minimize the right-hand side; its greatest lower bound is

$$h\inf_u\left[f(t,x,u) + \lambda(t,x,u)\int \pi(t,x,u,dx')\big(U(t,x')-U(t,x)\big)\right] + o(h) .$$

Consequently,

$$\frac{U(t,x)-U(t+h,x)}{h} = \inf_u \Big[f(t,x,u) + \lambda(t,x,u) \int \pi(t,x,u,dx') \times (U(t,x')-U(t,x)\Big] + \frac{o(h)}{h} .$$

This yields equation (2.3.2).

To prove uniqueness, we rewrite (2.3.2) in the form

$$U(t,x) = \int_t^T \inf_u \Big[f(s,x,u) + \lambda(s,x,u)\} \int \pi(s,x,u,dx') \times (U(s,x')-U(s,x)\Big] ds .$$

If $\bar{U}(t,x)$ is a second solution of this equation, then

$$|U(t,x)-\bar{U}(t,x)| \leqslant \int_t^T \sup_u \lambda(s,x,u) \int \pi(s,x,u,dx') \times |U(s,x')-\bar{U}(s,x')-U(s,x)+\bar{U}(s,x)|ds,$$

$$\sup_x |U(t,x)-\bar{U}(t,x)| \leqslant 2\int_t^T \sup_x |U(s,x)-\bar{U}(s,x)|ds .$$

From this it follows that $\sup_x |U(t,x)-\bar{U}(t,x)| = 0$. □

(c) *Optimum and ε-optimum controls.* Suppose that $U(t,x)$ is the solution of (2.3.2) and that a measurable function $u=\varphi(t,x)$ exists for which

$$f(t,x,\varphi(t,x)) + \int [U(t,x')-U(t,x)] \times \lambda(t,x,\varphi(t,x))\pi(t,x,\varphi(t,x),dx') = -\frac{\partial U(t,x)}{\partial t} . \tag{2.3.4}$$

Then $\eta(t) = \varphi(t,\xi(t))$ may be considered a Markov control; $\xi(t)$ is a Markov jump process with infinitesimal parameters, $\lambda(t,x,\varphi(t,x))$ and $\pi(t,x,\varphi(t,x),A)$. Let us show that

$$\mathbf{E}\left(\int_t^T f(s,\xi(s),\varphi(s,\xi(s))ds | \xi(t)=x \right) = U(t,x) .$$

Employing the same reasoning as in the derivation of (2.3.3), we find that the function

$$U^*(t,x) = \mathbf{E}\left(\int_t^T f(s,\xi(s),\varphi(s,\xi(s))ds | \xi(t)=x \right)$$

satisfies the equation

$$U^*(t,x) - U^*(t+h,x) = \int_t^{t+h} f(t,x,\varphi(s,x))ds$$
$$+ \int_t^{t+h} \lambda(t,x,\varphi(s,x)) \int \pi(t,x,\varphi(s,x),dx')$$
$$\times [U^*(t,x') - U^*(t,x)]ds + o(h) \ .$$

This yields

$$-\frac{\partial U^*(t,x)}{\partial t} = f(t,x,\varphi(t,x))$$
$$+\lambda(t,x,\varphi(t,x)) \int \pi(t,x,\varphi(t,x),dx')[U^*(t,x') - U^*(t,x)] \ . \quad (2.3.5)$$

From (2.3.4) we see that $U(t,x)$ is a solution of the same equation with the same boundary condition. Thus

$$U(t,x) = U^*(t,x) \ .$$

In similar fashion, it can be shown that if $\varphi_\varepsilon(t,x)$ is such that

$$f(t,x,\varphi_\varepsilon(t,x))$$
$$+\lambda(t,x,\varphi_\varepsilon(t,x)) \int \pi(t,x,\varphi_\varepsilon(t,x),dx')[U(t,x') - U(t,x)]$$
$$< -\frac{\partial U(t,x)}{\partial t} + \varepsilon \ ,$$

then $\eta(t) = \varphi_\varepsilon(t,\xi(t))$ will be a control for which

$$\mathbf{E}\left(\int_t^T f(s,\xi(s),\varphi(s,\xi(s)))ds | \xi(t) = x\right) \leqslant U(t,x) + \varepsilon(T-t)$$

for all $t \in [0,T]$ and x.

2.3.2 Controlled Diffusion Processes

To simplify formulations, we shall confine ourselves to the one-dimensional case: $X = R$ and $\mathcal{B} = \mathcal{B}_R$. A one-dimensional diffusion process is determined by its transport coefficient $a(t,x)$ and diffusion coefficient $b(t,x)$. In a controlled process, these functions also depend on the control $u \in U$. Thus a *controlled diffusion process* is determined by its diffusion coefficients $a(t,x,u)$ and $b(t,x,u)$. They are assumed to be locally bounded and to be measurable jointly in their arguments.

We shall examine step-controls and Markov controls for such processes. In the case of a step-control, the controlled process is constructed over the intervals of constancy of the control. If the values of the process and control

are given before the time such an interval begins, its conditional distributions on that interval coincide with the distributions of a process with diffusion coefficients $a(t,x,\bar{u})$ and $b(t,x,\bar{u})$, where $\bar{u}$ is fixed by the condition imposed. A process with a Markov control $u=\varphi(t,x)$ is a diffusion process with coefficients $a(t,x,\varphi(t,x))$ and $b(t,x,\varphi(t,x))$.

The results for controlled diffusion processes are similar in form to those obtained for jump processes. But their proofs and especially the clarification of conditions for the existence of solutions to Bellman's equation are very complex. Therefore we shall only state a few results.

Let $U(t,x)$ be the control cost if the expense is $\int_t^T f(s,x(s),u(s))ds$, given that $\xi(t)=x$, with $f(s,x,u)$ a bounded and measurable function of its arguments. Assume that $U(t,x)$ is differentiable in t and twice differentiable in x. Then $U(t,x)$ satisfies the following (Bellman) equation:

$$-\frac{\partial U(t,x)}{\partial t}=\sup_u\Big[f(t,x,u)+a(t,x,u)\frac{\partial U(t,x)}{\partial x}+\frac{1}{2}b(t,x,u)\frac{\partial^2 U(t,x)}{\partial x^2}\Big]$$

and boundary condition $U(T,x)=0$.

A similar equation may be derived for the more general cost function

$$\int_t^T f(s,x(s),u(s))\exp\left\{\int_t^s g(s,x(s'),u(s'))ds'\right\}ds\Phi(x(T))\,. \tag{2.3.6}$$

If $V(t,x)$ is the corresponding control cost and if $\partial V(t,x)/\partial t$, $\partial V(t,x)/\partial x$ and $\partial^2 V(t,x)/\partial x^2$ exist, then $V(t,x)$ satisfies

$$-\frac{\partial V(t,x)}{\partial t}=\sup_u\Big[f(t,x,u)+g(t,x,u)V(t,x)+a(t,x,u)\frac{\partial V(t,x)}{\partial x}+\frac{1}{2}b(t,x,u)\frac{\partial^2 V(t,x)}{\partial x^2}\Big] \tag{2.3.7}$$

and boundary condition $V(T,x)=\Phi(x)$.

If a solution to (2.3.7) exists and $u=\varphi(t,x)$ is a measurable function such that

$$-\frac{\partial V(t,x)}{\partial t}=f(t,x,\varphi(t,x))+g(t,x,\varphi(t,x))V(t,x)+a(t,x,\varphi(t,x))\frac{\partial V(t,x)}{\partial x}+\frac{1}{2}b(t,x,\varphi(t,x))\frac{\partial^2 V(t,x)}{\partial x^2} \tag{2.3.8}$$

then $u=\varphi(t,x)$ is an optimum Markov control.

If $u=\varphi_\varepsilon(t,x)$ is a measurable function and its substitution for φ on the right-hand side of (2.3.8) results in an expression not exceeding $-\partial V(t,x)/\partial t+\varepsilon$, then $u(t)=\varphi_\varepsilon(t,x(t))$ will be an $\varepsilon(T-t)$-optimum control on $[t,T]$.

3

Information

The reception, storage and transmission of information are phenomena encountered constantly. The concept of information arose a long time ago and had a purely qualitative character. The need to measure the amount of information manifested itself during the development of ways of storing and transmitting it. Despite their qualitative differences (verbal information, pictures, music and so on), it became clear that information of diverse content could be transformed into one another. Of simplest form is information written in a binary code and the average number of binary digits (bits) needed to express it is then roughly speaking the amount of information.

Information theory is one of the applied areas of probability and it is among the subjects comprising cybernetics. Its basic function is to investigate the possibilities for transmitting information. It takes into account, on the one hand, the randomness of messages that have to be transmitted and, on the other hand, the errors in transmission that are also of a random nature. It turns out that even when transmission errors are present, it is possible to send information in errorless fashion with probability as as close to one as desired. How to do this is what information theory studies.

3.1 Entropy

Before we can introduce the concept of amount of information, we need a more general concept, that of entropy. In mathematics, we view information as something that reduces existing uncertainty. In order to measure information, we need to measure to what extent uncertainty was reduced due to the information received. Therefore it is necessary to be able to quantify uncertainty. Such a measure of the uncertainty of something is then entropy.

3.1.1 Entropy of a Probability Experiment

The simplest form of uncertainty is in a probability experiment with a finite number of outcomes. Let $E_1, E_2 \ldots, E_r$ be the elementary events in an experiment and let $p_1, p_2, \ldots, p_r$ be their respective probabilities. We denote the experiment by the one letter $\mathcal{E}$ and its *entropy* by $H(\mathcal{E})$. By definition,

$$H(\mathcal{E}) = \sum_{k=1}^{r} p_k \log_2 \frac{1}{p_k} \,. \tag{3.1.1}$$

The choice of the number on the right-hand side of (3.1.1) as a measure of uncertainty may be clarified as follows. Suppose that the experiment is done independently n times with E_{i_k} the event occurring in the k-th experiment. The sequence $E_{i_1}, E_{i_2}, \ldots, E_{i_n}$ is said to be a message of length n. The probability of this message is

$$\prod_{m=1}^{n} p_m^{n_m}, \quad n_m = \sum_{k=1}^{n} I_{\{i_k = m\}} \,. \tag{3.1.2}$$

The total number of messages of length n is clearly $r^n = 2^{n \log_2 r}$. But among these messages are those whose probability is negligible in comparison with the probabilities of other messages. Let the probabilities in (3.1.2) be arranged according to size. Choose a positive δ and let $N_\delta(n)$ be the smallest number of messages whose total probability is at least $1 - \delta$.

Theorem 3.1.1. *For all $\delta \in (0, 1)$,*

$$\lim_{n \to \infty} \frac{1}{n} \log_2 N_\delta(n) = H(\mathcal{E})$$

no matter what δ is.

Proof. Consider the set S_n^ε of events for which

$$-n(H(\mathcal{E}) + \varepsilon) \le \log_2 \mathbf{P}(E_{i_1} \ldots E_{i_n}) \le -n(H(\mathcal{E}) - \varepsilon) \,, \tag{3.1.3}$$

and $\overline{S}_n^\varepsilon$, the set of those for which

$$\log_2 \mathbf{P}(E_{i_1} \ldots E_{i_n}) > -n(H(\mathcal{E}) - \varepsilon) \,. \tag{3.1.4}$$

Then

$$\mathbf{P}(S_n^\varepsilon) = \mathbf{P}\left(H(\mathcal{E}) - \varepsilon \le \sum_{i=1}^{r} \nu_i \log_2 \frac{1}{p_i} \le H(\mathcal{E}) + \varepsilon \right) \,,$$

and

$$\mathbf{P}(\overline{S}_n^\varepsilon) = \mathbf{P}\left(\sum_{i=1}^{r} \nu_i \log_2 \frac{1}{p_i} < H(\mathcal{E}) - \varepsilon \right) \,.$$

Since the relative frequency $\nu_i \to p_i$ in probability by virtue of the law of large numbers, we have $\mathbf{P}(S_n^\varepsilon) \to 1$ and $\mathbf{P}(\overline{S}_n^\varepsilon) \to 0$. Let N_n^ε and $\overline{N}_n^\varepsilon$ be the respective number of members of S_n^ε and $\overline{S}_n^\varepsilon$. Then

$$\mathbf{P}(S_n^\varepsilon)2^{n(H(\mathcal{E})-\varepsilon)} \le N_n^\varepsilon \le \mathbf{P}(S_n^\varepsilon)2^{n(H(\mathcal{E})+\varepsilon)}$$

and

$$\overline{N}_n^\varepsilon \le \mathbf{P}(\overline{S}_n^\varepsilon)2^{n(H(\mathcal{E})-\varepsilon)}$$

(the total probability has been divided into the minimum probability for the upper bound and the maximum probability for the lower bound).

For n sufficiently large (when $\mathbf{P}(S_n^\varepsilon) > 1 - \delta$),

$$(1-\delta)2^{n(H(\mathcal{E})-\varepsilon)} \le N_\delta(n) \le N_n^\varepsilon + \overline{N}_n^\varepsilon$$

or

$$\log_2(1-\delta) + n(H(\mathcal{E}) - \varepsilon) \le \log_2 N_\delta(n) \le n(H(\mathcal{E}) + \varepsilon) .$$

It now remains to use the fact that ε is arbitrary. □

Remark 3.1.1. Suppose that it is necessary to write out each of N messages in binary form. This means that each message is in correspondence with a sequence of zeroes and ones and that different messages have different sequences. What is the shortest length of the sequences of zeroes and ones that are used? If $2^m < N \le 2^{m+1}$, then this length is obviously $m + 1 = -[\log_2 N^{-1}], [\,\cdot\,]$ being the integral part of a number. If we now consider the sequences for the original messages of length r, then there will be N^r of them and the number of bits needed is $-[\log_2 N^{-r}]$. Hence, if these messages are written out repeatedly, the average number of bits needed to be utilized for one such message is

$$-\lim_{r\to\infty} \frac{1}{r}[\log_2 N^{-r}] = \log_2 N .$$

This number is the entropy of an experiment $\mathcal{E}_N$ with N equally likely outcomes:

$$H(\mathcal{E}_N) = \sum_{i=1}^{N} \frac{1}{N} \log_2 N = \log_2 N .$$

Theorem 3.1.1 says that even for general experiments, the entropy is the average number of bits needed to write down one message if independent repetitions of the experiment are performed indefinitely.

3.1.2 Properties of Entropy

Formula (3.1.1) shows that $H(\mathcal{E}) > 0$. The maximum entropy of an experiment with r outcomes is $\log_2 r$, which is the entropy of an experiment with equally likely outcomes. This is easy to see by maximizing the right-hand side of (3.1.1) with respect to $p_1, \ldots, p_r$ subject to the condition $p_1 + \ldots + p_r = 1$.

(a) *Entropy of a compound experiment.* Consider an experiment $\mathcal{E}$ in which two experiments $\mathcal{E}_1$ and $\mathcal{E}_2$ are performed in succession. Write $\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2$. $\mathcal{E}$ is called a compound experiment. Let $A_1, \ldots, A_m$ be the elementary events of $\mathcal{E}_1$ and $B_1, \ldots, B_l$ be the elementary events of $\mathcal{E}_2$. Then the elementary events of $\mathcal{E}_1 \times \mathcal{E}_2$ are $\{A_i \cap B_j, i = 1, \ldots, m;\ \ j = 1, 2, \ldots, l\}$. We have

$$\begin{aligned} H(\mathcal{E}_1 \times \mathcal{E}_2) &= -\sum_{i,j} \mathbf{P}(A_i \cap B_j) \log_2 \mathbf{P}(A_i \cap B_j) \\ &= -\sum_{i,j} \mathbf{P}(A_i)\mathbf{P}(B_j|A_i)[\log_2 \mathbf{P}(A_i) + \log_2 \mathbf{P}(B_j|A_i)] \\ &= -\sum_{i,j} \mathbf{P}(A_i)\mathbf{P}(B_j|A_i) \log_2 \mathbf{P}(A_i) \\ &\quad - \sum_{i,j} \mathbf{P}(A_i)\mathbf{P}(B_j|A_i) \log_2 \mathbf{P}(B_j|A_i) \\ &= H(\mathcal{E}_1) + \sum_i \mathbf{P}(A_i) H(\mathcal{E}_2|A_i)\ . \end{aligned}$$

$H(\mathcal{E}_2|A_i)$ would be the entropy of $\mathcal{E}_2$ if the probability of each elementary event B_j coincided with its conditional probability, given that A_i has happened. The expression

$$\sum_i \mathbf{P}(A_i) H(\mathcal{E}_2|A_i) = H(\mathcal{E}_2|\mathcal{E}_1) \tag{3.1.5}$$

is called the conditional entropy of experiment $\mathcal{E}_2$ with respect to experiment $\mathcal{E}_1$. Thus

$$H(\mathcal{E}_1 \times \mathcal{E}_2) = H(\mathcal{E}_1) + H(\mathcal{E}_2|\mathcal{E}_1)\ . \tag{3.1.6}$$

A compound experiment consisting of n successive experiments $\mathcal{E}_1$, $\mathcal{E}_2, \ldots, \mathcal{E}_n$ is defined similarly. We denote it by $\mathcal{E}_1 \times \ldots \times \mathcal{E}_n$. Since $\mathcal{E}_1 \times \ldots \times \mathcal{E}_n = (\mathcal{E}_1 \times \ldots \times \mathcal{E}_{n-1}) \times \mathcal{E}_n$, formula (3.1.6) yields the following general formula:

E1.

$$\begin{aligned} H(\mathcal{E}_1 \times \ldots \times \mathcal{E}_n) = H(\mathcal{E}_1) + H(\mathcal{E}_2|\mathcal{E}_1) + \ldots \\ + H(\mathcal{E}_n|\mathcal{E}_1 \times \ldots \times \mathcal{E}_{n-1})\ . \end{aligned} \tag{3.1.7}$$

Suppose that $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent experiments. Then $\mathbf{P}(B_j|A_i) = \mathbf{P}(B_j)$ and so $H(\mathcal{E}_2|A_i) = H(\mathcal{E}_2)$. Consequently, $H(\mathcal{E}_1 \times \mathcal{E}_2) = H(\mathcal{E}_1) + H(\mathcal{E}_2)$. This implies

E2. If $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ are independent experiments, then

$$H(\mathcal{E}_1 \times \ldots \times \mathcal{E}_n) = H(\mathcal{E}_1) + \ldots + H(\mathcal{E}_n)\ . \tag{3.1.8}$$

E3. For any $\mathcal{E}_1$ and $\mathcal{E}_2$,

$$H(\mathcal{E}_2|\mathcal{E}_1) \le H(\mathcal{E}_2) \tag{3.1.9}$$

and $H(\mathcal{E}_2|\mathcal{E}_1) = H(\mathcal{E}_2)$ only where $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent.

Indeed,

$$\begin{aligned} H(\mathcal{E}_2|\mathcal{E}_1) &= -\sum_i \mathbf{P}(A_i) \sum_j \mathbf{P}(B_j|A_i) \log_2 \mathbf{P}(B_j|A_i) \\ &= \sum_j \sum_i \mathbf{P}(A_i)\psi(\mathbf{P}(B_j|A_i)) \ , \end{aligned}$$

where $\psi(x) = -x \log_2 \frac{1}{x}$. The function $\psi(x)$ is convex up in (0,1) and so

$$\sum_i \mathbf{P}(A_i)\psi(x_i) \le \psi\left(\sum_i \mathbf{P}(A_i)x_i\right)$$

if $x_i \in (0,1)$. Equality is possible only when $x_i = x_1$ for all i. Thus

$$\sum_i \mathbf{P}(A_i)\psi(\mathbf{P}(B_j|A_i)) \le \psi\left(\sum_i \mathbf{P}(A_i)\mathbf{P}(B_j|A_i)\right) = \psi(\mathbf{P}(B_j)) \ ,$$

with equality possible only when the $\mathbf{P}(B_j|A_i)$ are equal for all i, that is, $\mathbf{P}(B_j|A_i) = \mathbf{P}(B_j)$. It remains to observe that $\sum_j \psi(\mathbf{P}(B_j)) = H(\mathcal{E}_2)$.

E4. If $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_3$ are arbitrary experiments with finitely many outcomes, then

$$H(\mathcal{E}_3|\mathcal{E}_1 \times \mathcal{E}_2) \le H(\mathcal{E}_3|\mathcal{E}_2) \ . \tag{3.1.10}$$

The proof is similar to the derivation of (3.1.9). Equality is attained in (3.1.10) if $\mathcal{E}_3$ and $\mathcal{E}_1$ are conditionally independent given $\mathcal{E}_2$.

(b) *Entropy of a stationary sequence.* We next consider a stationary sequence $\{\zeta_n, n \ge 1\}$ whose terms ζ_n assume finitely many distinct values. Let $\mathcal{E}_n$ be an experiment involving the measurement of ζ_n. The entropy of stationary sequence $\{\zeta_n\}$ is defined to be

$$H(\{\zeta_n\}) = \lim_{n\to\infty} \frac{1}{n} H(\mathcal{E}_1 \times \ldots \times \mathcal{E}_n) \ . \tag{3.1.11}$$

E5. The limit on the right hand side of (3.1.11) exists.

For, we have

$$\begin{aligned} H(\mathcal{E}_1 \times \ldots \times \mathcal{E}_n) = H(\mathcal{E}_1) + H(\mathcal{E}_2|\mathcal{E}_1) + \ldots \\ + H(\mathcal{E}_n|\mathcal{E}_1 \times \ldots \times \mathcal{E}_{n-1}) \ . \end{aligned}$$

Therefore the limit on the right-hand side of (3.1.11) is the same as

$$\begin{aligned} &\lim_{n\to\infty} [H(\mathcal{E}_1 \times \ldots \times \mathcal{E}_n) - H(\mathcal{E}_1 \times \ldots \times \mathcal{E}_{n-1})] \\ &\quad = \lim_{n\to\infty} H(\mathcal{E}_n|\mathcal{E}_1 \times \ldots \times \mathcal{E}_{n-1}) \end{aligned} \tag{3.1.12}$$

provided that the latter exists. But on the basis of E4,

$$H(\mathcal{E}_n|\mathcal{E}_1 \times \ldots \times \mathcal{E}_{n-1}) \le H(\mathcal{E}_n|\mathcal{E}_2 \times \ldots \times \mathcal{E}_{n-1}) \\ = H(\mathcal{E}_{n-1}|\mathcal{E}_1 \times \ldots \times \mathcal{E}_{n-2})$$

(in obtaining this last equality, we made use of the stationarity of $\{\zeta_n\}$). Therefore the pre-limiting quantity on the right-hand side of (3.1.12) is a nonincreasing nonnegative sequence.

E6. If a stationary sequence $\{\zeta_n\}$ is Markov, then

$$H(\{\zeta_n\}) = -\sum_{i,k} p_i p_{ik} \log_2 p_{ik} \ , \tag{3.1.13}$$

with $p_i = \mathbf{P}\{\zeta_1 = a_i\}$, $p_{ik} = \mathbf{P}\{\zeta_2 = a_k|\zeta_1 = a_i\}$ and $\{a_i\}$ the set of values of the variable ζ_n.

To see this, consider the corresponding sequence of experiments $\{\mathcal{E}_n\}$. They form a Markov chain. The probabilities in the experiment $\mathcal{E}_n$, given $\mathcal{E}_1, \ldots, \mathcal{E}_{n-1}$, depend only on $\mathcal{E}_{n-1}$: $\mathbf{P}\{\zeta_n = a_k|\mathcal{E}_1 \times \ldots \times \mathcal{E}_{n-1}\} = \mathbf{P}\{\zeta_n = a_k|\mathcal{E}_{n-1}\}$. Therefore $\mathcal{E}_1, \ldots, \mathcal{E}_{n-2}$ and $\mathcal{E}_n$ are conditionally independent given $\mathcal{E}_{n-1}$. Thus

$$H(\mathcal{E}_n|\mathcal{E}_1 \times \ldots \times \mathcal{E}_{n-1}) = H(\mathcal{E}_n|\mathcal{E}_{n-1}) \\ = -\sum_{i,k} \mathbf{P}\{\zeta_{n-1} = a_i\}\mathbf{P}\{\zeta_n = a_k|\zeta_{n-1} = a_i\} \log_2 \mathbf{P}\{\zeta_n = a_k|\zeta_{n-1} = a_i\} \ ,$$

and this expression is the same as the right-hand side of (3.1.13).

3.1.3 ε-Entropy and Entropy of a Continuous Random Variable

Now consider a random variable ζ with a continuous distribution function $F(x)$. Its *ε-entropy* is defined to be

$$-\sum_k \mathbf{P}\{k\varepsilon \le \zeta < (k+1)\varepsilon\} \log_2 \mathbf{P}\{k\varepsilon \le \zeta < (k+1)\varepsilon\} \ . \tag{3.1.14}$$

Suppose that ζ has a density $f(x)$. Then the relation (3.1.14) is expressible in the form

$$-\int f(x) \log_2 f(x) dx + \log_2 \frac{1}{\varepsilon} + o(1)$$

with $o(1) \to 0$ as $\varepsilon \to 0$ (the integral is assumed to exist). As $\varepsilon \to 0$, this expression becomes infinite. But for various random variables, the difference of their ε-entropies has a finite limit. Write

$$H(\zeta) = -\int f(x) \log_2 f(x) dx \tag{3.1.15}$$

This quantity is defined to be the *entropy of a continuously distributed random variable*. Then the difference of the ε-entropies of two continuous variables ζ_1 and ζ_2 approaches $H(\mathcal{E}_1) - H(\mathcal{E}_2)$ under the assumption that these entropies are finite.

Similarly, if ζ is a random vector in R^d having a density $f(x)$, then its entropy is also defined by (3.1.15) (the integral being over R^d). Its ε-entropy is given by the expression

$$-\sum_{k_1,\dots,k_d} \mathbf{P}\{\zeta \in V^{(\varepsilon)}_{k_1,\dots,k_d}\} \log_2 \mathbf{P}\{\zeta \in V^{(\varepsilon)}_{k_1,\dots,k_d}\}\,, \tag{3.1.16}$$

where $V^{(\varepsilon)}_{k_1,\dots,k_d} = \{x : k_1\varepsilon \le x_1 < (k_1+1)\varepsilon, \dots, k_d\varepsilon \le x_d < (k_d+1)\varepsilon\}$ and $x_1, \dots, x_d$ are the coordinates of x. It differs from the entropy by the amount $d \log_2 \frac{1}{\varepsilon} + o(1)$.

Let ζ and η be two variables with a joint density $f_{\zeta,\eta}(x,y)$. Let $f_\zeta(x)$ and $f_\eta(x)$ be their marginal densities. The expression

$$H(\zeta|\eta) = -\int f_{\zeta,\eta}(x,y) \log_2 \frac{f_{\zeta,\eta}(x,y)}{f_\eta(y)} dxdy$$

is called the conditional entropy of ζ relative to η. If ζ and η are independent, then $H(\zeta|\eta) = H(\zeta)$.

Let ζ and η be random variables in finite-dimensional Euclidean spaces X and Y having densities, and let (ζ,η) have a density in the space $X \times Y$. Then the following formulas hold:

E7. $H((\zeta,\eta)) = H(\zeta) + H(\eta|\zeta)$.
E8. $H(\zeta,\eta) \le H(\zeta)$.
E9. $H((\zeta,\eta)) = H(\zeta) + H(\eta)$ when ζ and η are independent.

E7 and E9 follow from (3.1.16) and E8 may be deduced by passing to the limit in the discrete case.

3.1.4 Information

(a) *Information in one experiment with respect to another.* Suppose that $\mathcal{E}_1$ and $\mathcal{E}_2$ are two experiments each with finitely many outcomes. The quantity

$$I(\mathcal{E}_1, \mathcal{E}_2) = H(\mathcal{E}_1) - H(\mathcal{E}_1|\mathcal{E}_2) \tag{3.1.17}$$

is defined to be the *amount of information* contained in $\mathcal{E}_2$ with respect to $\mathcal{E}_1$. If $\{A_i\}$ are the outcomes of $\mathcal{E}_1$ and $\{B_k\}$ the outcomes of $\mathcal{E}_2$, then

$$\begin{aligned} I(\mathcal{E}_1,\mathcal{E}_2) &= -\sum_i \mathbf{P}(A_i) \log_2 \mathbf{P}(A_i) + \sum_{i,k} \mathbf{P}\{A_i \cap B_k\} \log_2 \frac{\mathbf{P}\{A_i \cap B_k\}}{\mathbf{P}\{B_k\}} \\ &= -\sum_i \mathbf{P}\{A_i\} \log_2 \mathbf{P}\{A_i\} - \sum_k \mathbf{P}\{B_k\} \log_2 \mathbf{P}\{B_k\} \\ &\quad + \sum_{i,k} \mathbf{P}\{A_i \cap B_k\} \log_2 \mathbf{P}\{A_i \cap B_k\}\,. \end{aligned}$$

This furnishes a formula for the amount of information which we take as one of its properties:

I1. $I(\mathcal{E}_1, \mathcal{E}_2) = H(\mathcal{E}_1) + H(\mathcal{E}_2) - H(\mathcal{E}_1 \times \mathcal{E}_2)$.

This formula leads in turn to the following property:

I2. $I(\mathcal{E}_1, \mathcal{E}_2) = I(\mathcal{E}_2, \mathcal{E}_1) \geq 0$; $I(\mathcal{E}_1, \mathcal{E}_2) = 0$ only when $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent (Property E3 and (3.1.17) were made use of).

Let $\mathcal{E}$ and $\mathcal{E}'$ be two experiments with respective outcomes $\{A_i\}$ and $\{A'_j\}$. If to every i, there exist $j_1, \ldots, j_r$ such that $A_i = \cup A'_{j_k}$, then we write $\mathcal{E} \subset \mathcal{E}'$.

I3. If $\mathcal{E}_1 \subset \mathcal{E}'$, then $I(\mathcal{E}_1, \mathcal{E}_2) \leq I(\mathcal{E}'_1, \mathcal{E}_2)$.

To prove this, it suffices to show that $H(\mathcal{E}_2|\mathcal{E}'_1) \leq H(\mathcal{E}_2|\mathcal{E}_1)$. We have

$$\begin{aligned} H(\mathcal{E}_2|\mathcal{E}'_1) &= -\sum_{i,j} \mathbf{P}(B_j \cap A'_i) \log_2 \mathbf{P}(B_j|A'_i) \\ &= -\sum_{i,k} \sum_{A'_j \subset A_k} \mathbf{P}(B_j \cap A'_i) \log_2 \mathbf{P}(B_j|A'_j) \\ &\leq -\sum_{j,k} \mathbf{P}(B_j \cap A_k) \log_2 \mathbf{P}(B_j|A_k) = H(\mathcal{E}_2|\mathcal{E}_1) \, . \end{aligned}$$

The last inequality is derived in exactly the same way as E3.

(b) *Information in one stationary sequence with respect to another.* Let $\{(\zeta_n, \eta_n), n = 0, 1, \ldots\}$ be a stationary sequence of pairs of random variables. Let ζ_n and η_n assume finitely many values. The amount of information in sequence $\{\eta_n\}$ with respect to sequence $\{\zeta_n\}$ is the quantity

$$I(\{\zeta_n\}, \{\eta_n\}) = \lim_{k \to \infty} \frac{1}{k} I(\mathcal{E}_0^{(1)} \times \ldots \times \mathcal{E}_k^{(1)}, \mathcal{E}_0^{(2)} \times \ldots \times \mathcal{E}_k^{(2)}) \, , \quad (3.1.18)$$

where $\mathcal{E}_n^{(1)}$ and $\mathcal{E}_n^{(2)}$ are experiments measuring ζ_n and η_n, respectively.

I4. The limit in (3.1.18) exists and

$$I(\{\zeta_n\}, \{\eta_n\}) = H(\{\zeta_n\}) + H(\{\eta_n\}) - H(\{(\zeta_n, \eta_n)\}) \quad (3.1.19)$$

This statement is a consequence of E5.

I5. Let ζ_n be a stationary ergodic Markov chain and let $\{(\zeta_n, \eta_n)\}$ also be a stationary Markov chain. Then

$$\begin{aligned} I(\{\zeta_n\}, \{\eta_n\}) = \sum_{i,j,k,l} & [p_{a_k b_l} P(a_k, b_l; a_i, b_j) \quad (3.1.20) \\ & - \sum_{r,s} p_{a_k} P(a_k; a_j) p_{a_r b_l} P(a_r, b_l; a_s, b_j)] \log_2 P(a_k, b_l; a_i, b_j) \, . \end{aligned}$$

In this $\{a_i\}$ are the values of ζ_n and $\{b_j\}$ are the values of η_n, and

$$\begin{aligned} p_{a_k} &= \mathbf{P}\{\zeta_1 = a_k\}, \quad P(a_k; a_j) = \mathbf{P}\{\zeta_2 = a_j | \zeta_1 = a_k\}, \\ p_{a_k b_l} &= \mathbf{P}\{\zeta_1 = a_k, \eta_1 = b_l\}, \quad P(a_k, b_l; a_i, b_j) \\ &= \mathbf{P}\{\zeta_2 = a_i, \eta_2 = b_j | \zeta_1 = a_k, \eta_1 = b_l\} \, . \end{aligned}$$

Proof. We shall make use of formula (3.1.19). Since $\{\zeta_n\}$ and $\{(\zeta_n, \eta_n)\}$ are stationary Markov chains, they obey formula (3.1.13). Let us compute the entropy of $\{\eta_n\}$. By definition,

$$\begin{aligned} H(\{\eta_n\}) &= -\lim_{m\to\infty} \frac{1}{m} \mathbf{E} \log_2 \psi_m(\eta_1, \dots, \eta_m), \\ \psi_m(b_{i_1}, \dots, b_{i_m}) &= \mathbf{P}\{\eta_1 = b_{i_1}, \dots, \eta_m = b_{i_m}\} \\ &= \mathbf{E}\mathbf{P}\{\eta_1 = b_{i_1}, \dots, \eta_m = b_i | \zeta_1, \dots, \zeta_m\} \\ &= \mathbf{E}\mathbf{P}\{\eta = b_{i_1} | \zeta_1\} \prod_{k=1}^{m-1} Q(b_{i_k}, b_{i_{k+1}}; \zeta_k, \zeta_{k+1}) \ , \end{aligned}$$

where in turn

$$Q(b_i, b_j; a_k, a_l) = \mathbf{P}\{\eta_2 = b_j | \eta_1 = b_i, \zeta_1 = a_k, \zeta_2 = a_l\} = \frac{P(a_k, b_i; a_l, b_j)}{P(a_k; a_l)} \ .$$

Using the ergodicity of $\{\zeta_k\}$, we can easily show that

$$\begin{aligned} &\mathbf{E}\mathbf{P}\{\eta_1 = b_{i_1} | \zeta_1\} \prod_{k=1}^{m-1} Q(b_{i_k}, b_{i_{k+1}}; \zeta_k, \zeta_{k+1}) \\ &\quad = \mathbf{E} \exp\left\{ \ln \mathbf{P}\{\eta_1 = b_{i_1} | \zeta_1\} + \sum_{k=1}^{m-1} \ln Q(b_{i_k}, b_{i_{k+1}}; \zeta_k, \zeta_{k+1}) \right\} \\ &\quad = \exp\left\{ \sum_{k=1}^{m-1} \mathbf{E} \ln Q(b_{i_k}, b_{i_{k+1}}; \zeta_k, \zeta_{k+1}) + m_{\varepsilon_m} \right\} , \end{aligned}$$

where $\varepsilon_m \to 0$. Thus

$$\begin{aligned} H(\{\eta_n\}) &= -\lim_{m\to\infty} \frac{1}{m} \mathbf{E} \sum_{k=1}^{m-1} \left[\log_2 \sum_{a_k, a_l} Q(\eta_k, \eta_{k+1}; a_i, a_l) p_{a_i} P(a_i; a_l) \right] \\ &= -\sum_{i,j} p_{a_j} P(a_i; a_j) \sum_{k,l} \mathbf{P}\{\eta_1 = b_k, \eta_2 = b_l\} \\ &\quad \times \log_2 \frac{P(a_i, b_k; a_j, b_l)}{P(a_i; a_j)} = -\sum_{i,j,k,l} p_{a_i} P(a_i; a_j) \\ &\quad \times \sum_{r,s} p_{a_r b_k} P(a_r, b_k; a_s, b_l) \log_2 P(a_i, b_k; a_j, b_l) - H(\{\zeta_n\}) \ . \end{aligned}$$

Using (3.1.13) to find $H(\{\zeta_n, \eta_n\})$ and inserting the resulting expression for $H(\{\eta_n\})$ in the right-hand side of (3.1.19), we arrive at (3.1.20). □

(c) *Information in one continuously distributed random variable with respect to another.* Let ζ and η take values in finite-dimensional Euclidean spaces X and Y, respectively. Assume that the pair of variables (ζ, η) has a density

$f_{\zeta,\eta}(x,y)$ in $X \times Y$. Then ζ and η have respective marginal densities $f_\zeta(x)$ and $f_\eta(y)$. The quantity

$$I(\zeta,\eta) = \int\int [f_{\zeta,\eta}(x,y)\log_2 f_{\zeta,\eta}(x,y) - f_\zeta(x)f_\eta(y)\log_2(f_\zeta(x)f_\eta(y))]dxdy \tag{3.1.21}$$

is the amount of information in η with respect to ζ.

I6. Let F_ζ be a set of random variables each assuming finitely many values which are functions of ζ and let F_η be defined similarly. Then

$$I(\zeta,\eta) = \sup_{\zeta_1 \in F_\zeta, \eta_1 \in F_\eta} I(\zeta_1,\eta_1)\ . \tag{3.1.22}$$

Applying I4, one can establish that $I(\zeta_1,\eta_1) \le I(\zeta,\eta)$. The supremum is attained by $I(\zeta,\eta)$ because of formula (3.1.16).

3.2 Transmission of Information

We now examine a mathematical model for the transmission of information. Information subject to transmission is a time-discrete stochastic process with a finite phase space. The process is said to be the source of information. The values of the process are viewed as the message units subject to transmission. For example, if a telegram is sent, messages may be considered to be letters, words, phrases or entire texts. The source of information is characterized quantitatively by the amount of information produced per unit time. We shall assume that the process describing the information is stationary otherwise the concept is meaningless. The amount of information generated by the source of information is the entropy of this stationary process (computed per unit time).

Information being produced must be transmitted via a communication channel. Let us consider a mathematical model for it.

3.2.1 Communication Channels

A communication channel C is specified, first of all, by two finite sets – the input set X and output set Y. The channel converts a sequence of input symbols $(x_1, x_2, \ldots, x_n)$ into some sequence of output symbols $(y_1, y_2, \ldots, y_n)$; in other words, for each n, there is a function φ_n sending X^n into Y^n. For various n, the functions φ_n are mutually consistent in the following sense. Let P_X and P_Y be operators in $\bigcup_{n>1} X^n$ and in $\bigcup_{n>1} Y^n$, respectively, defined by $P_X(x_1, x_2, \ldots, x_{n-1}, x_n) = (x_1, x_2, \ldots, x_{n-1})$ and $P_Y(y_1, y_2, \ldots, y_{n-1}, y_n) = (y_1, y_2, \ldots, y_{n-1})$. Then

$$P_Y\varphi_n(x_1,\ldots,x_n) = \varphi_{n-1}(P_X(x_1,\ldots,x_n)). \tag{3.2.1}$$

In other words, if $\varphi_n(x_1, \ldots, x_n) = (y_1, \ldots, y_n)$, then $\varphi_{n-1}(x_1, \ldots, x_{n-1}) = (y_1, \ldots, y_{n-1})$, that is, the communication channel transmits symbols sequentially.

Thus, a channel is specified by giving sets X and Y and a collection of functions $\{\varphi_n, n = 1, 2, \ldots\}$. Each φ_n was assumed to be nonrandom in this description and such a channel is said to be *determinate* or *noiseless*. A channel is said to have finite memory if for $n > m$, $\varphi_n(x_1, \ldots, x_n) = (\varphi_{11}(x_1), \varphi_{21}(x_1, x_2), \ldots, \varphi_{nn}(x_1, x_2, \ldots, x_n))$ is such that φ_{nn} depends only on $x_n, x_{n-1}, \ldots, x_{n-m+1}$. It is stationary if there is a function $g(x_1, \ldots, x_n)$ from X^m to Y not depending on n such that if $\varphi(x_1, \ldots, x_n) = (y_1, \ldots, y_n)$, then $y_n = g(x_{n-m+1}, \ldots, x_n)$ for $n > m$. Generally speaking, the definition of a communication channel imposes only the condition (3.2.1) on the functions φ_n and this does not preclude the possibility of the φ_n being random. If they are random, then one says that the communication channel specified by these functions is *noisy*.

(a) *Noisy communication channels.* When the functions φ_n carry the sequence $(x_1, x_2, \ldots)$ into a random sequence $(\eta_1, \eta_2, \ldots)$ with values in Y, it is convenient to describe the latter by means of finite-dimensional distributions. Condition (3.2.1) becomes a "nonpredictive" property: the joint distribution of $\eta_1, \ldots, \eta_n$ depends only on $x_1, \ldots, x_n$. Let $\mathbf{x}$ be a point in X^∞. A noisy channel is defined by a family $\mu(C|\mathbf{x})$ of distributions in Y^∞ such that $\mu(C|\mathbf{x})$ depends only on $x_1, x_2, \ldots, x_n$ for all n and C_n of the form $\{\mathbf{y} : y_1 = b_{i_1}, \ldots, y_n = b_{i_n}\}$ with $b_{i_k} \in Y$. Probabilities computed with respect to the measure $\mu(C|\mathbf{x})$ will be denoted by $\mathbf{P}_\mathbf{x}$.

The simplest communication channel is a memoryless one for which

$$\mathbf{P}_\mathbf{x}\{\eta_1 = b_1, \ldots, \eta_n = b_n\} = \prod_{k=1}^{n} \mathbf{P}_\mathbf{x}\{\eta_k = b_k\},$$

and $\mathbf{P}_\mathbf{x}\{\eta_k = b_k\} = g_k(b_k, x_k)$. A channel has finite memory if for some m with $n > m$,

$$\begin{aligned} &\mathbf{P}_\mathbf{x}\{\eta_n = b_n | \eta_{n-1} = b_{n-1}, \ldots, \eta_{n-m} = b_{n-m}\} \\ &= g_n(b_n, b_{n-1}, \ldots, b_{n-m}, x_n, x_{n-1}, \ldots, x_{n-m}). \end{aligned} \tag{3.2.2}$$

Observe that a memoryless channel has finite memory with $m = 0$. A channel with finite memory is called stationary if the function g_n in (3.2.2) is independent of n for $n > m$. Let us examine the distribution of the output symbols if the input sequence is random. Let it be $\{\xi_n\}$ where the ξ_n take on values in X. A noisy channel can be specified by the conditional distributions

$$\mathbf{P}_\mathbf{x}\{\eta_1 = b_1\}, \ldots, \mathbf{P}_\mathbf{x}\{\eta_n = b_n | \eta_1 = b_1, \ldots, \eta_{n-1} = b_{n-1}\}.$$

The joint distribution of $\{\xi_n\}$ and $\{\eta_n\}$ is

$$\begin{aligned}&\mathbf{P}\{\xi_1 = x_1, \dots, \xi_n = x_n, \eta_1 = b_1, \dots, \eta_n = b_n\}\\&= \mathbf{P}\{\xi_1 = x_1, \dots, \xi_n = x_n\}\mathbf{P}_{\mathbf{x}}\{\eta_1 = b_{i_1}\}\\&\quad \times \prod_{k=1}^{n-1} \mathbf{P}_{\mathbf{x}}\{\eta_{k+1} = b_{i_{k+1}} | \eta_1 = b_{i_1}, \dots, \eta_k = b_{i_k}\}.\end{aligned} \tag{3.2.3}$$

A noisy channel is said to be stationary if for any stationary sequence $\{\xi_n\}$, the sequence $\{(\xi_n, \eta_n)\}$ is also stationary in $X \times Y$.

(b) *Channel capacity.* We first concentrate on a determinate channel. The number of signals of length n that may be received at the channel output equals the number l_n of points in the range of φ_n. If the range is denoted by N_n, then $N_n \subset Y^n$ and consequently $l_n \leq l^n$, where l is the number of points in X. Hence, a channel may transmit l_n different messages of length n. These messages are points of the space X^n. The function φ_n maps these points onto N_n in one-to-one fashion. The *channel capacity* is defined to be

$$c = \lim_{n\to\infty} \frac{1}{n} \log_2 l_n$$

if the limit exists. This limit always exists for a stationary channel. The capacity is the average number of binary digits that are transmittable over the channel per unit time.

Let there be given a noisy channel defined by the collection of distributions $\mu(C|\mathbf{x})$. Let $\mathcal{P}(X)$ be the set of all X-valued random sequences $\{\xi_n, n \geq 1\}$. With each such sequence there is associated the sequence $\{(\xi_n, \eta_n), n \geq 1\}$ in $X \times Y$ whose distribution is given by (3.2.3) (only the distributions of the sequences interest us here). The distributions of $\{(\xi_n, \eta_n)\}$ evidently depend only on that of $\{\xi_n\}$ (if $(C|\mathbf{x})$ is held fixed). Put

$$c(\{\xi_n\}) = \lim_{m\to\infty} \frac{1}{m} I((\xi_1, \dots, \xi_m), (\eta_1, \dots, \eta_m)\} \tag{3.2.4}$$

if this limit exists, and

$$c = \sup_{\{\xi_n\}\in\mathcal{P}(X)} c(\{\xi_n\}). \tag{3.2.5}$$

This number c is then the noisy channel capacity. By this definition, the channel capacity is roughly speaking the maximum amount of information that is contained in a message at the channel output about the input message per unit time.

If a communication channel is stationary and has finite memory, then the limit (3.2.4) exists for every stationary sequence $\{\xi_n\}$ and the least upper bound in (3.2.5) remains unchanged if it is taken over $\mathcal{P}_S(X)$, the subset of all stationary sequences $\{\xi_n\}$ of $\mathcal{P}(X)$.

Example. Consider the very simple channel for which $X = Y = \{0, 1\}$. Moreover, it is a memoryless stationary channel. $\mathbf{P}_0\{\{0\}\} = p$, $\mathbf{P}_0\{\{1\}\} = 1 - q$,

$\mathbf{P}_1\{\{0\}\} = q$ and $\mathbf{P}_1\{\{1\}\} = 1 - q$. One can show that the supremum in (3.2.5) is attained here if $\{\xi_n\}$ are independent and identically distributed. Let $\mathbf{P}\{\xi_k = 0\} = s$ and $\mathbf{P}\{\xi_k = 1\} = 1 - s$. Then

$$\begin{aligned}
&\mathbf{P}\{\eta_k = 0\} = sp + (1-s)q = \psi(s),\\
&\mathbf{P}\{\eta_k = 1\} = 1 - \psi(s),\\
&I(s) = I(\xi_1, \eta_1) = H(\eta_1) - H(\eta_1|\xi_1) = \psi(s)\log_2\frac{1}{\psi(s)}\\
&\qquad + (1-\psi(s))\log_2\frac{1}{1-\psi(s)} - sH_0 - (1-s)H_1,
\end{aligned}$$

where

$$\begin{aligned}
H_0 &= p\log_2\frac{1}{p} + (1-p)\log_2\frac{1}{1-p},\\
H_1 &= q\log_2\frac{1}{q} + (1-q)\log_2\frac{1}{1-q}.
\end{aligned}$$

Since $s = [\psi(s) - q]/(p - q)$, it follows that

$$I(s) = \psi(s)\log_2\frac{1}{\psi(s)} + (1-\psi(s))\log_2\frac{1}{1-\psi(s)} + A\psi(s) + B \quad (3.2.6)$$

with

$$A = \frac{H_1 - H_0}{p - q}, \quad B = \frac{qH_0 - pH_1}{p - q}$$

The extreme value of the right-hand side of (3.2.6) occurs when

$$\log_2\frac{1}{\psi} - \log_2\frac{1}{1-\psi} + A = 0$$

or

$$\psi = \frac{2^A}{1 + 2^A}.$$

Consequently, the channel capacity is

$$\begin{aligned}
c &= \frac{2^A}{1+2^A}\log_2(1 + 2^{-A}) + \frac{1}{1+2^A}\log_2(1 + 2^A)\\
&\quad + A\frac{2^A}{1+2^A} + B = \frac{2^A}{1+2^A}\log_2((1 + 2^{-A})\cdot 2^A)\\
&\quad + \frac{1}{1+2^A}\log_2(1 + 2^A) + B = \log_2(2^B + 2^{A+B})\\
&= \log_2\left(2^{\frac{qH_0 - pH_1}{p-q}} + 2^{\frac{(1-p)H_1 - (1-q)H_0}{p-q}}\right).
\end{aligned}$$

3.2.2 Coding and Decoding

For a stream of messages generated by a source of information to be transmittable over a communication channel, it is necessary to encode the messages via the elements of X. Then upon the reception of the sequence of Y-valued symbols at the output, it is necessary to recover a transmitted message via decoding. We introduce two additional finite sets: Z, the message units produced by the source of information and U, the symbols used to write out the information that has to be received. Suppose that one has to receive information about the motion of a sputnik and its interior state. The source of information are the readings of the transmitters installed inside the sputnik (this is the set Z). X, Y and U characterize respectively, the current arriving at the sputnik's antenna, the audio and light signals of the transponder and the required numerical data.

The *coding* operation is specified by a nonpredictive function mapping Z-valued sequences into X-valued ones. Thus coding is via a function ψ from Z^∞ to X^∞ and if $\psi(z_1, z_2, \ldots) = (x_1, x_2, \ldots)$, $\psi(z'_1, z'_2, \ldots) = (x'_1, x'_2, \ldots)$, and $z_1 = z'_1, \ldots, z_n = z'_n$, then $x_1 = x'_1, \ldots, x_n = x'_n$. The simplest code is a memoryless code determined by a function $\psi_1(z)$ from Z to X given by

$$\psi_1(z_1, z_2, \ldots) = (\psi_1(z_1), \psi_1(z_2), \ldots).$$

One way of coding is to split the sequence of messages into segments of length n and to encode each such segment. If $\psi(z_1, \ldots, z_n) = (x_1, \ldots, x_m)$ and $\psi(\bar{z}_1, \ldots, \bar{z}_n) = (\bar{x}_1, \ldots, \bar{x}_m)$, then $\psi(z_1, z_2, \ldots, z_n, \bar{z}_1, \bar{z}_2, \ldots, \bar{z}_n) = (x_1, \ldots, x_m, \bar{x}_1, \ldots, \bar{x}_m)$.

The *decoding* operation is specified by a function θ from Y^∞ to U^∞, which must also be nonpredictive. The most natural case is where U can be identified with Z. If this is so and the transmission of messages is errorless, then the goal is to receive the same sequence of messages at the output as generated by the source of information. Therefore the series of mappings

$$Z^\infty \xrightarrow{\psi} X^\infty \xrightarrow{\varphi} Y^\infty \xrightarrow{\theta} Z^\infty$$

must be such that the composition $\theta \circ \varphi \circ \psi$ is in a sense the identity mapping (this does not mean that we recover the same finite segment $z_1, z_2, \ldots, z_n$ but rather $z_1, \ldots, z_{k_n}$, where $k_n \leq n$ and $n - k_n$ is the size of lag in transmission). Therefore the θ accomplishing the decoding is determined by the ψ accomplishing the encoding (the communication channel is fixed and φ is unaltered).

Sometimes the decoding process is specified by a sequence of functions θ_n on Y_n taking values in Z^{k_n}; k_n is nondecreasing and the θ_n for different n are consistent in the following way. For $n \leq m$, from $\theta_n(y_1, \ldots, y_n) = (z_1, \ldots, z_{k_n})$ it follows that

$$\theta_m(y_1, \ldots, y_n, \ldots, y_m) = (z_1, \ldots, z_{k_n}, \ldots, z_{k_m}).$$

One may also consider randomized decoding rules involving given probabilities

$$\mathbf{P}\big(\theta_m(y_1,\dots,y_n)=(z_1,\dots,z_{k_n})\big)$$

(they are dependent on y_i and z_i).

Example. Let $X = Y = \{0,1\}$. The channel is memoryless and noiseless; 0 goes into 0 and 1 into 1. $Z = \{1,2,3\}$. Messages arise independently and are equally likely. During time n, 3^n equally likely messages can be sent. To give them binary codes, we need no less than $n\log_2 3$ bits. Thus messages of length n cannot be received sooner than a length of time $n\log_2 3$; $k_n \sim n/\log_2 3$ and so the lag is primarily proportional to the length of a message.

3.3 Shannon's Theorem

The example given at the end of the preceding section shows that this communication channel cannot always transmit messages from the source of information so as to prevent the lag from increasing indefinitely. If $Z = \{1,2\}$ in that example, then it evidently would have been possible to transmit messages without lag. Observe that the entropy of the process at the output of the communication channel (per unit time) does not exceed the channel capacity which in the considered example is $1 = \log_2 2$, as one can readily see. Therefore, if it were possible to map the process describing the stream of messages generated by the source of information in one-to-one fashion into the process received at the channel output, then the entropy of the source of messages should not exceed 1 since a one-to-one mapping of experiments into one another does not affect the entropy. This is a general fact. A source of information cannot send messages without indefinitely increasing lag if its entropy is greater than the channel capacity. Shannon's theorem says that when the entropy of a source of information is less than the channel capacity, then such transmission is possible with probability arbitrarily close to one. We shall give a more precise formulation below. We mention also that Shannon himself considered the case where the source of information generates independent and identically distributed messages and the communication channel is stationary and memoryless. We shall refer to this as the simplest case. Its generalization will also be called Shannon's Theorem.

3.3.1 Simplest Transmission of Information

The stream of messages generated by a source of information is a process $\{\zeta_n, n \geq 1\}$ of independent and identically distributed random variables taking values in Z. The entropy of the source of messages is

$$H = -\sum_{z\in Z} \mathbf{P}\{\zeta_1 = z\}\log_2 \mathbf{P}\{\zeta_1 = z\}.$$

A memoryless stationary communication channel is specified by the probabilities

$$\mathbf{P}\{\eta_k = y | \xi_k = x\} = p(x, y), x \in X, y \in Y.$$

Its capacity is c. We shall take $U = Z$. Suppose that coding and decoding rules have been selected. The transmission lag in time N is defined as $\tau_N = \max_{n \le N}(n - k_n)$, where k_n is the number of bits in the decoded sequence after transmission of n bits. The transmission error in time N is defined to be

$$\mathbf{P}\left\{\bigcup_{k=1}^{k_N}\{\zeta_i \neq \zeta_i^*\}\right\} = \varepsilon_N,$$

where $(\zeta_1^*, \ldots, \zeta_{k_n}^*)$ is the sequence resulting on decoding the messages after the transmission of $(\zeta_1, \ldots, \zeta_N)$.

Shannon's Theorem. *Let $H < c$. Then to every positive ε, there is an N_ε such that coding and decoding rules exist for all $N > N_\varepsilon$ for which $\tau_N \le \varepsilon N$ and $\varepsilon_N \le \varepsilon$.*

The proof of the theorem will be carried out separately for deterministic channels and for noisy channels.

(a) *Deterministic channels.* In this case $p(x, y)$ equals either 0 or 1. Let Y_1 be the subset of those $y \in Y$ that can be at the channel output. If $y \in Y_1$, there is an $x \in X$ such that $p(x, y) = 1$. For each $y \in Y_1$, choose one such x and let $X_1 \subset X$ be the set of x chosen. The communication channel carries X_1 into Y_1 in one-to-one fashion. It is easy to show that the channel capacity is $\log_2 m_1$, where m_1 is the number of elements of Y_1 (which is also the number of elements of X_1). This is the amount of information (per unit time) which is contained about the sequence $\{\xi_n\}$ in $\{\eta_n\}$ if ξ_n takes values in X_1 with probabilities $1/m_1$. By assumption, $H < \log_2 m_1$. On the basis of Theorem 3.1.1 on p. 236, to every positive ε, one can find an n_{ε_1} such that for all $n > n_{\varepsilon_1}$ there are 2^{ns} chains of the form $z_1, z_2, \ldots, z_n$ all of whose probabilities $\mathbf{P}\{\zeta_1 = z_1, \zeta_2 = z_2, \ldots, \zeta_n = z_n\}$ add up to at least $1 - \varepsilon_1$ with $s \in (H, \log_2 m_1 - \delta)$. To transmit a series of messages $(z_1, z_2, \ldots, z_n)$ in this collection (denote this subset of Z^n by Z_n), we employ sequences $x_1, x_2, \ldots, x_n$ with $x_i \in X_1$. There are $m_1^n = 2^{n \log_2 m_1} > 2^{ns}$ such sequences. Thus Z_n is carried into X_1^n in one-to-one fashion. X_1^n is carried by the channel into Y_1^n in one-to-one fashion and Y_1^n is carried into Z_n in one-to-one fashion. If $(z_1, \ldots, z_n) \notin Z_n$, we map it into one of the sequences in X^n not belonging to the image of Z_n. We decode the sequence $(y_1, \ldots, y_n)$ corresponding to it in an arbitrary way. The probability that a message of length n will be transmitted without error is greater than $1 - \varepsilon_1$. If there is a message of length $N > n$, we split it into pieces of length n and we transmit each segment of length n in the above way. Since the transmission is done with segments of length n, the lag $\tau_N \le n$. The error probability is less than $l\varepsilon_1$, where l is the integral part of N/n. Now choosing ε_1 and l so that $l\varepsilon_1 \le \varepsilon$ and $n/N < 1/l < \varepsilon$, we arrive at our required assertion. N_ε is any arbitary number satisfying $N_\varepsilon > n_{\varepsilon_1}/\varepsilon$.

(b) *Noisy channels.* As was seen in the proof of the preceding part, it is enough to show the following: To every positive ε, one can find an n_ε such that for $n > n_\varepsilon$ it is possible to transmit one of the $2^{n(H+\delta)}$ messages produced by the source of information in time n with error probability less than ε, δ being an arbitrarily selected positive number. Suppose that the distribution of ξ in X has been chosen so that $H < I(\xi, \eta) < C$ for pairs ξ and η satisfying

$$\mathbf{P}\{\xi = x, \eta = y\} = \mathbf{P}\{\xi = x\}p(x, y).$$

Let $H(\xi)$ be the entropy of the distribution of ξ. Consider the subset X_n of X^n of those $(x_1, x_2, \dots, x_n) \in X$ for which

$$\left| \sum_{k=1}^{n} I_{\{x_k = x\}} - n\mathbf{P}\{\xi = x\} \right| \le na_n, \quad x \in X, \tag{3.3.1}$$

for some sequence a_n such that $a_n \downarrow 0$, $na_n^2 \to \infty$. The number of points in X_n is clearly $2^{n(H(\xi)+\varepsilon_n)}$, where $\varepsilon_n \to 0$. Now let $\{\eta_k(x), x \in X\}$ be independent for different values of k and let $\mathbf{P}\{\eta_k(x) = y\} = p(x, y)$. The entropy of the sequence $\eta_1(x_1), \dots, \eta_n(x_n)$ with $(x_1, x_2, \dots, x_n) \in X$ is

$$\begin{aligned} \sum_{i=1}^{n} \sum_{y} p(x_i, y) \log_2 \frac{1}{p(x_i, y)} &\sim n \sum_{x} \mathbf{P}\{\xi = x\} \sum_{y} p(x, y) \log_2 \frac{1}{p(x, y)} \\ &= nH(\eta|\xi). \end{aligned}$$

(We have made use of (3.3.1).) For n sufficiently large, to each $(x_1, \dots, x_n) \in X_n$ there is a set $S^{(n)}_{(x_1,\dots,x_n)} \subset Y^n$ such that

$$\mathbf{P}\{(\eta_1(x_1), \dots, \eta_n(x_n)) \in S^{(n)}_{(x_1,\dots,x_n)}\} \ge 1 - \varepsilon_1 \tag{3.3.2}$$

with the number of points in $S^{(n)}_{(x_1,\dots,x_n)}$ at most $2^{n(H(\eta|\xi)+\delta)}$. Since $H < I(\xi, \eta) < H(\xi)$, one can choose a subset $\tilde{X}_n$ of X_n whose cardinality equals the number of messages that must be transmitted (there are at most $2^{n(H+\delta)}$ of them). One may assume that $H + 2\delta < I(\xi, \eta)$.

The decoding rule is this. If the sequence $(y_1, y_2, \dots, y_n)$ is observed at the channel output, then put

$$\mathbf{P}\{\theta_{\tilde{X}_n}(y_1, \dots, y_n) = (x_1, \dots, x_n)\} = \left(\sum_{(x_1,\dots,x_n) \in \tilde{X}_n} I_{\left\{(y_1,\dots,y_n) \in S^{(n)}_{(x_1,\dots,x_n)}\right\}} \right)^{-1}$$

if $(y_1, \dots, y_n) \in \bigcup_{(x_1,\dots,x_n) \in \tilde{X}_n} S^{(n)}_{(x_1,\dots,x_n)}$. But if $(y_1, \dots, y_n)$ does not fall in this set, $\theta_{\tilde{X}_n}(y_1, \dots, y_n)$ is specified arbitrarily. Let us estimate $\alpha_{\tilde{X}_n}$, the transmission error probability. We have

$$\alpha_{\tilde{X}_n} \le \sum_{(x_1,\dots,x_n)\in\tilde{X}_n} p(x_1,\dots,x_n) \sum_{(y_1,\dots,y_n)\in S^{(n)}_{(x_1,\dots,x_n)}} \prod_{i=1}^{n} p(x_i,y_i)$$
$$\times \left[1 - \left(\sum_{(x_1,\dots,x_n)\in\tilde{X}_n} I_{\{(y_1,\dots,y_n)\in S^{(n)}_{(x_1,\dots,x_n)}\}}\right)^{-1}\right] + \varepsilon_1 \quad (3.3.3)$$

(we have made use of (3.3.2)). In this, $p(x_1,\dots,x_n)$ is the probability that a message encoded by the sequence $(x_1,\dots,x_n)$ is transmitted. The proof of Theorem 3.1.1 on p. 236 shows that

$$p(x_1,\dots,x_n) \le 2^{-n(H-\delta)}$$

for n sufficiently large. In exactly the same way,

$$\prod_{i=1}^{n} p(x_i,y_i) \le 2^{-n(H(\eta|\xi)-\delta)}.$$

Therefore

$$\alpha_{\tilde{X}_n} \le 2^{-n(H+H(\eta|\xi)-2\delta)} \sum_{(x_1,\dots,x_n)\in\tilde{X}_n} \sum_{(y_1,\dots,y_n)\in S^{(n)}_{(x_1,\dots,x_n)}}$$
$$\times \left[1 - \left(\sum_{(x_1,\dots,x_n)\in\tilde{X}_n} I_{\{(y_1,\dots,y_n)\in S^{(n)}_{(x_1,\dots,x_n)}\}}\right)^{-1}\right].$$

Let $m(A)$ be the number of points in set A. Then we have

$$\sum_{(x_1,\dots,x_n)\in\tilde{X}_n} \sum_{(y_1,\dots,y_n)\in S^{(n)}_{(x_1,\dots,x_n)}} \left[1 - \left(\sum_{(x_1,\dots,x_n)\in\tilde{X}_n} I_{\{(y_1,\dots,y_n)\in S^{(n)}_{(x_1,\dots,x_n)}\}}\right)^{-1}\right]$$
$$= \sum_{(x_1,\dots,x_n)\in\tilde{X}_n} m\left(S^{(n)}_{(x_1,\dots,x_n)}\right) - m\left(\bigcup_{(x_1,\dots,x_n)\in\tilde{X}_n} m\left(S^{(n)}_{(x_1,\dots,x_n)}\right)\right)$$
$$\le \sum_{\substack{(x_1,\dots,x_n)\in\tilde{X}_n \\ (\bar{x}_1,\dots,\bar{x}_n)\in\tilde{X}_n}} m\left(S^{(n)}_{(x_1,\dots,x_n)} \cap S^{(n)}_{(\bar{x}_1,\dots,\bar{x}_n)}\right).$$

Thus

$$\alpha_{\tilde{X}_n} \le 2^{-n(H+H(\eta|\xi)-2\delta)} \sum_{\substack{(x_1,\dots,x_n)\in\tilde{X}_n \\ (\bar{x}_1,\dots,\bar{x}_n)\in\tilde{X}_n}} m\left(S^{(n)}_{(x_1,\dots,x_n)} \cap S^{(n)}_{(\bar{x}_1,\dots,\bar{x}_n)}\right) + \varepsilon_1.$$

We now show that one may choose $\tilde{X}_n$ so that $\alpha_{\tilde{X}_n}$ is less than $2\varepsilon_1$. Let t be the number of points in X_n, s the number in $\tilde{X}_n$, u the number in $S^{(n)}_{(x_1,\dots,x_n)}$

and v the number of $(x_1, \ldots, x_n)$ for which $(y_1, \ldots, y_n) \in S^{(n)}_{(x_1,\ldots,x_n)}$; u may be considered independent of $(x_1, \ldots, x_n) \in X_n$ and v considered independent of $(y_1, \ldots, y_n) \in \bigcup_{(x_1,\ldots,x_n)\in\tilde{X}_n} S^{(n)}_{(x_1,\ldots,x_n)}$.

The average error due to the possible choices of $\tilde{X}_n$ is

$$\begin{aligned}
\bar{\alpha} &= \sum_{\tilde{X}_n \subset X_n} \binom{t}{s}^{-1} \alpha_{\tilde{X}_n} \\
&\le \varepsilon_1 + \binom{t}{s}^{-1} 2^{-n(H+H(\eta|\xi)-2\delta)} \\
&\quad \times \sum_{\tilde{X}_n \subset X_n} \sum_{\substack{(x_1,\ldots,x_n)\in\tilde{X}_n \\ (\bar{x}_1,\ldots,\bar{x}_n)\in\tilde{X}_n}} \sum_{(y_1,\ldots,y_n)} I_{S^{(n)}_{(x_1,\ldots,x_n)}}(y_1, \ldots, y_n) I_{S^{(n)}_{(\bar{x}_1,\ldots,\bar{x}_n)}}(y_1, \ldots, y_n) \\
&\le \varepsilon_1 + \sum_{\substack{(x_1,\ldots,x_n)\in X_n \\ (\bar{x}_1,\ldots,\bar{x}_n)\in X_n}} \sum_{(y_1,\ldots,y_n)} I_{S^{(n)}_{(x_1,\ldots,x_n)}}(y_1, \ldots, y_n) \\
&\quad \times I_{S^{(n)}_{(\bar{x}_1,\ldots,\bar{x}_n)}}(y_1, \ldots, y_n) \binom{t}{s}^{-1} \binom{t-2}{s-2} 2^{-n(H+H(\eta|\xi)-2\delta)} \\
&= \varepsilon_1 + \binom{t}{s}^{-1} \binom{t-2}{s-2} 2^{-n(H+H(\eta|\xi)-2\delta)} \\
&\quad \times \sum_{(y_1,\ldots,y_n)} \left(\sum_{(x_1,\ldots,x_n)\in X_n} I_{S^{(n)}_{(x_1,\ldots,x_n)}}(y_1, \ldots, y_n) \right)^2 \\
&= \varepsilon_1 + \binom{t}{s}^{-1} \binom{t-2}{s-2} 2^{-n(H+H(\eta|\xi)-2\delta)} v^2 m \left(\bigcup S^{(n)}_{(x_1,\ldots,x_n)} \right).
\end{aligned}$$

Notice that $vm\left(\bigcup S^{(n)}_{(x_1,\ldots,x_n)}\right) = tu$. Therefore

$$\bar{\alpha} \le \varepsilon_1 + \frac{s(s-1)}{t(t-1)} 2^{-n(H+H(\eta|\xi)-2\delta)} v \cdot t \cdot u \le \varepsilon_1 + 2\frac{svu}{t} 2^{-n(H+H(\eta|\xi)-2\delta)}.$$

Since $s \le 2^{n(H+\delta)}$ and $u \le 2^{n(H(\eta|\xi)+\delta)}$ for n sufficiently large, it follows that

$$\bar{\alpha} \le \varepsilon_1 + 2^{5\delta n+1} \cdot \frac{v}{t} \cdot 2^{nH}.$$

One can show that $v \le 2^{n(H(\eta|\xi)+\delta)}$ and $t \ge 2^{n(H(\xi)-\delta)}$ and so

$$\bar{\alpha} \le \varepsilon_1 + 2^{7\delta n+1} 2^{n(H-I(\xi,\eta))}.$$

This last expression may be made arbitrarily small if $H + 7\delta < I(\xi, \eta)$. There clearly exist $\tilde{X}_n$ such that $\alpha_{\tilde{X}_n} \le \bar{\alpha}$. The theorem is proved. □

3.3.2 Generalizations

(a) *Ergodic source of messages.* Now suppose that the stream of messages generated by the source of information is a stationary ergodic process $\{\zeta_n, n \geq 1\}$ with values in a finite set Z.

Lemma 3.3.1. *If $H = H(\zeta_n)$ is the entropy of a stationary process $\{\zeta_n\}$, then to every positive ε and δ, there exists a subset Z_n of Z^n for n greater than some n_0 such that*

1. *the number of points $(z_1, \ldots, z_n) \in Z_N$ does not exceed $2^{n(H+\delta)}$,*
2. $\sum_{(z_1,\ldots,z_n)\in Z_n} \mathbf{P}\{\zeta_1 = z_1, \ldots, \zeta_n = z_n\} \geq 1 - \varepsilon$,
3. $H - \delta \leq \frac{1}{n} \log_2 \mathbf{P}\{\zeta_1 = z_1, \ldots, \zeta_n = z_n\} \leq H + \delta$.

Proof. It is convenient to view the process $\{\zeta_n\}$ as being defined for all integral n (the process may be extended in an obvious way). Put

$$\varphi_n(z) = \mathbf{P}\{\zeta_0 = z|\zeta_1, \zeta_2, \ldots, \zeta_n\}, \varphi(z) = \mathbf{P}\{\zeta_0 = z|\zeta_1, \ldots\}.$$

Then $\mathbf{E} \log_2 \varphi(\zeta_0) = H$ and $\mathbf{E}|\log_2 \varphi_n(\zeta_0) - \log_2 \varphi(\zeta_0)| \to 0$. Thus with probability 1

$$\lim_{n\to\infty} \frac{1}{n} \log_2 Q(\zeta_1, \zeta_2, \ldots, \zeta_n) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} q_k(\zeta_k|\zeta_1, \ldots, \zeta_{k-1}) = H,$$

where $Q(z_1, \ldots, z_n) = \mathbf{P}(\zeta_1 = z_1, \ldots, \zeta_n = z_n)$ and $q_k(z_k|z_1, \ldots, z_{k-1}) = \log_2 \mathbf{P}\{\zeta_k = z_k|\zeta_1 = z_1, \ldots, \zeta_{k-1} = z_{k-1}\}$. Given positive ε and δ, suppose that

$$\mathbf{P}\left\{\left|\frac{1}{n} \log_2 Q(\zeta_1, \ldots, \zeta_n) - H\right| < \delta\right\} \geq 1 - \varepsilon.$$

Put $Z_n = \left\{(z_1, \ldots, z_n) : \left|\frac{1}{n} \log_2 Q(z_1, \ldots, z_n) - H\right| < \delta\right\}$. Then conditions 2 and 3 hold. The number $m(Z_n)$ satisfies $m(Z_n)2^{-n(H+\delta)} \leq 1$ so that condition 1 also holds. □

(b) *Ergodic stationary communication channel.* Suppose that the channel is stationary. In other words, if the input is a stationary process $\{\xi_k, k = 0, \pm 1, \ldots\}$ and $\{\eta_k\}$ is the output process, then $\{(\xi_k, \eta_k), k = 0, \pm 1, \ldots\}$ is also stationary. (It is convenient here to consider processes defined for all integral times. When a communication channel has finite memory m, the process $\{(\xi_k, \eta_k), k \geq m\}$ is stationary if $\{\xi_k, 1, 2, \ldots\}$ is stationary.) A channel is said to be ergodic if $\{(\xi_k, \eta_k), k = 0, \pm 1, \pm 2, \ldots\}$ is ergodic when $\{\xi_k, k = 0, \pm 1, \ldots\}$ is any ergodic process. The quantity $\sup I(\{\xi_n\}, \{\eta_n\})$ is defined to be the ergodic channel capacity of an ergodic channel where the supremum is taken over all ergodic X-valued processes $\{\xi_n, n = 0, \pm 1, \pm 2, \ldots\}$.

Lemma 3.3.2. *Suppose that a communication channel and a process* $\{\xi_n, n = 0, \pm 1, \pm 2, \ldots\}$ *are both ergodic. Then for all positive* ε *and* δ*, there exist an* n_0 *and sets* $X_n \subset X^n$*,* $Y_n \subset Y^n$ *and* $W_n \subset X_n \times Y_n$ *for some* $n > n_0$ *such that*

1. $\log_2 m(X_n) \leq n(H(\{\xi_n\}) + \delta)$, $\log_2 m(Y_n) \leq n(H(\{\eta_n\} + \delta))$ *and* $\log_2 m(W_n) \leq n(H(\{(\xi_n, \eta_n)\}) + \delta)$,
2. *if* $W_n(x_1, x_2, \ldots, x_n) = \{(y_1, \ldots, y_n) : ((x_1, \ldots, x_n), (y_1, \ldots, y_n)) \in W_n\}$, *then* $\log_2 m(W_n(x_1, \ldots, x_n)) \leq n(H(\{\eta_n\}|\{\xi_n\}) + \delta)$,
3. $\sum_{((x_1,\ldots,x_n),(y_1,\ldots,y_n))\in W_n} \mathbf{P}\{\xi_1 = x_1, \ldots, \xi_n = x_n, \eta_1 = y_1, \ldots, \eta_n = y_n\} \geq 1 - \varepsilon$,
4. *for all* $(x_1, \ldots, x_n) \in X_n$,

$$\log_2 \mathbf{P}\{\xi_1 = x_1, \ldots, \xi_n = x_n\} \geq -n(H(\{\xi\}) + \delta),$$

and

$$\log_2 \mathbf{P}\{\eta_1 = y_1, \ldots, \eta_n = y_n | \xi_1 = x_1, \ldots, \xi_n = x_n\} \geq -n(H(\{\eta_n\}|\{\xi_n\}) + \delta), (y_1, \ldots, y_n) \in W_n(x_1, \ldots, x_n).$$

Shannon's theorem for ergodic stationary sources and communication channels. If one analyzes the proof of Shannon's theorem in the simplest case for a noisy channel, it can be seen that it merely involves computing the number of members of the sets X_n, $\tilde{X}_n$ and $W_n(x_1, \ldots, x_n)$. Therefore Lemma 3.3.2 can be used to show that Shannon's theorem is true when a source of messages is stationary and ergodic with entropy H and the ergodic channel capacity of a stationary ergodic channel equals c with $c > H$.

4

Filtering

When messages are sent by radio, they are affected by random noise caused by electricity in the atmosphere, solar flares and man's industrial activities. Therefore a message received is a mixture of the signal sent and noise. One of the ways of extracting information from a transmitted signal (in terms of the preceding chapter, one of the ways of decoding) is to "sift out" the noise from the message received. This operation is called filtering. The filtering methods considered here are based on the random nature of the signals and noise. Just as noise is filtered out, it is possible to filter out possible random effects from quantities characterizing the state of a system. By viewing the system as transmitting a "signal" from the "past" to the "future", one can predict its "future" if the system is evolving randomly. We shall therefore also view prediction as a special case of filtering.

4.1 Linear Prediction and Filtering of Stationary Stochastic Processes

4.1.1 General Approach to Finding a Linear Estimator of a Random Variable

Let there be given n random variables $\xi_1, \ldots, \xi_n$ that are observable in some experiment and a random variable η that is not accessible to observation. How can one estimate η from the values of $\xi_1, \ldots, \xi_n$? Here we shall utilize linear estimators of η or, in other words, linear functions of $\xi_1, \ldots, \xi_n$ of the form $\bar{\eta} = c + \sum_{k=1}^{n} a_k \xi_k$ to approximate η. We shall look for a function of this form that estimates η the best. If $\mathbf{E}\xi_k^2 < \infty, k = 1, 2, \ldots, n$, and $\mathbf{E}\eta^2 < \infty$, then it is natural to view the best $\bar{\eta}$ as the one that minimizes $\mathbf{E}|\eta - \bar{\eta}|^2$. There is no loss of generality in assuming that $\mathbf{E}\xi_k = 0, k = 1, 2, \ldots, n$. Since

$$\mathbf{E}|\eta - \bar{\eta}|^2 = (\mathbf{E}\eta - \mathbf{E}\bar{\eta})^2 + \mathbf{V}(\eta - \bar{\eta}) = (\mathbf{E}\eta - c)^2 + \mathbf{V}(\eta - \bar{\eta})$$

and the second term does not depend on c, it follows that $c = \mathbf{E}\eta$. To find the coefficients a_k, it is advantageous first to orthogonalize the sequence $\xi_1, \xi_2, \ldots, \xi_n$. Let

$$\begin{aligned} &\xi_1^* = \xi_1, \xi_2^* = \xi_2 + \alpha_{21}\xi_1^*, \ldots, \\ &\xi_k^* = \xi_k + \alpha_{k1}\xi_1^* + \ldots + \alpha_{kk-1}\xi_{k-1}^*, \ldots, \end{aligned} \tag{4.1.1}$$

where $\alpha_{ki}, 1 \le i < k \le n$, are chosen so that $\mathbf{E}\xi_k^*\xi_i^* = 0$ or, in other words, $\mathbf{E}\xi_k\xi_i^* = -\alpha_{ki}\mathbf{E}|\xi_i^*|^2$. When $\xi_i^* = 0$, α_{ki} may be chosen arbitrarily but we shall assume that $\alpha_{ki} = 0$. By relations (4.1.1), the ξ_k are also expressible linearly in terms of the ξ_k^* by the formulas

$$\xi_k = \xi_k^* - \alpha_{k1}\xi_1^* - \ldots - \alpha_{kk-1}\xi_{k-1}^*. \tag{4.1.2}$$

Therefore every linear combination of the ξ_k's is a linear combination of the ξ_k^*'s and conversely.

Thus, let $\bar{\eta} = \mathbf{E}\eta + \sum_{k=1}^n a_k^*\xi_k^*$. Then

$$\mathbf{E}|\eta - \bar{\eta}|^2 = \mathbf{E}|\eta|^2 - |\mathbf{E}\eta|^2 - 2\sum_{k=1}^n a_k^*\mathbf{E}\eta\xi_k^* + \sum_{k=1}^n (a_k^*)^2\mathbf{E}(\xi_k^*)^2. \tag{4.1.3}$$

This expression has a minimum when

$$a_k^* = \mathbf{E}\eta\xi_k^*/\mathbf{E}(\xi_k^*)^2 \tag{4.1.4}$$

(if $\mathbf{E}(\xi_k^*)^2 = 0$ and $\mathbf{E}\eta\xi_k^* = 0$, so that the right-hand side of (4.1.3) is independent of a_k^*, we set $a_k^* = 0$). If a_k^* is chosen according to (4.1.4), then

$$\mathbf{E}(\eta - \bar{\eta})\xi_k^* = 0 \text{ for } k = 1, 2, \ldots, n,$$

and hence by (4.1.2),

$$\mathbf{E}(\eta - \bar{\eta})\xi_k = 0, k = 1, 2, \ldots, n. \tag{4.1.5}$$

Equations (4.1.5) are a system for determining the coefficients a_k.

Now let there be given a family of random variables $\{\xi_\lambda, \lambda \in \Lambda\}$ for which $\mathbf{E}\xi_\lambda = 0$ and $\mathbf{E}|\xi_\lambda|^2 < \infty$ and a variable η. Again, the ξ_λ's are observable and η has to be estimated linearly from sampling of $\{\xi_\lambda\}$. Consider the linear subspace $L\{\xi_\lambda, \lambda \in \Lambda\}$ of random variables which is the closure under mean-square convergence of the set of linear combinations

$$\left\{\sum_{k=1}^n \alpha_k\xi_{\lambda_k}, n = 1, 2, \ldots, \ \lambda_k \in \Lambda, \ \alpha_k \in R\right\}.$$

There is no loss of generality in assuming that $\mathbf{E}\eta = 0$. Let $L\{\eta, \xi_\lambda, \lambda \in \Lambda\}$ be the subspace of random variables which is the closure under mean-square convergence of the set of linear combinations

$$\left\{\beta\eta + \sum_{k=1}^{n} \alpha_k \xi_{\lambda_k}, n = 1, 2, \ldots, \lambda_k \in \Lambda,\ \beta, \alpha_k \in R\right\}.$$

$L\{\eta, \xi_\lambda, \lambda \in \Lambda\}$ is a Hilbert space with inner product $\langle \zeta_1, \zeta_2 \rangle = \mathbf{E}\zeta_1\zeta_2$, $\zeta_i \in L\{\eta, \xi_\lambda, \lambda \in \Lambda\}$. $L\{\xi_\lambda, \lambda \in \Lambda\}$ is a subspace of this space. A linear estimator of η based on the variables $\{\xi_\lambda, \lambda \in \Lambda\}$ is an element of $L\{\xi_\lambda, \lambda \in \Lambda\}$. Thus it is necessary to find an $\bar{\eta} \in L\{\xi_\lambda, \lambda \in \Lambda\}$ which minimizes $E|\eta - \bar{\eta}|^2$. This will be the case when $\bar{\eta}$ is the orthogonal projection of η on $L\{\xi_\lambda, \lambda \in \Lambda\}$, that is, $\eta - \bar{\eta}$ is orthogonal to $L\{\xi_\lambda, \lambda \in \Lambda\}$. This condition is equivalent to the relation

$$\mathbf{E}\eta\xi_\lambda = \mathbf{E}\bar{\eta}\xi_\lambda. \tag{4.1.6}$$

This relation is the one customarily applied when constructing a filter.

4.1.2 Prediction of Stationary Sequences

Consider a numerical (wide-sense) stationary sequence $\{\xi_n,\ n{=}0, \pm1, \pm2, \ldots\}$. Suppose that it has been observed up to and including the present moment of time and it is necessary to "predict" its value at some future moment of time. The present time is taken to be $t = 0$. Thus, we have a collection of variables $\{\xi_n, n \le 0\}$ and a variable $\xi_m, m > 0$. It is necessary to construct a linear estimator of ξ_m with respect to the variables $\{\xi_n,\ n \le 0\}$. For a stationary sequence, use can be made of its spectral representation (see p. 118)

$$\xi_n = \int_{-\pi}^{\pi} e^{i\lambda n} dy(\lambda),$$

where $y(\lambda)$ is a random function with orthogonal increments on $[-\pi, \pi]$ and $\mathbf{E}|dy(\lambda)|^2 = dF(\lambda)$. $F(\lambda)$ is the *spectral function* of the sequence. If $r_n = \mathbf{E}\xi_0\xi_n$ is the covariance function of the sequence (we are assuming that $\mathbf{E}\xi_k = 0$), then

$$r_n = \int_{-\pi}^{\pi} e^{i\lambda n} dF(\lambda).$$

Let $L_2(F)$ be the space of measurable complex-valued functions $g(\lambda)$ defined on $[-\pi, \pi]$ for which

$$\int_{-\pi}^{\pi} |g(\lambda)|^2 dF(\lambda) < \infty.$$

It is a complex Hilbert space. Let $L_2^-(F)$ be the subspace of it which is the closure of linear combinations of the form

$$\sum_{k \le 0} c_k e^{i\lambda k}.$$

It is easy to see that $L\{\xi_n, n \le 0\}$ coincides with the variables of the form $\int_{-\pi}^{\pi} g(\lambda)dy(\lambda)$ with $g(\lambda) \in L_2^-(F)$. Thus, the projection of ξ_m on $L\{\xi_n, n \le 0\}$ has the form

$$\bar{\xi}_m = \int_{-\pi}^{\pi} g_m(\lambda)dy(\lambda),$$

with $g_m(\lambda) \in L_2^-(F)$ and for all $n \le 0$,

$$\begin{aligned} r_{n-m} = \mathbf{E}\xi_m\xi_n = \mathbf{E}\bar{\xi}_m\xi_n = \mathbf{E}\overline{\int_{-\pi}^{\pi} e^{i\lambda n}dy(\lambda)} \times \int_{-\pi}^{\pi} g_m(\lambda)dy(\lambda) \\ = \int_{-\pi}^{\pi} e^{-i\lambda n} g_m(\lambda)dF(\lambda). \end{aligned} \tag{4.1.7}$$

(a) *Solution of the prediction problem.* Equation (4.1.7) determines the $L_2^-(F)$-function $g_m(\lambda)$ uniquely. However it is difficult to find this function effectively at least because there is no satisfactory description of $L_2^-(F)$. One instance is examined below where the prediction problem can in a sense be solved effectively.

Suppose that the *spectral density* $f(\lambda) = dF(\lambda)/d\lambda$ exists and satisfies the following condition. For some positive c,

$$c \le f(\lambda) \le 1/c.$$

Lemma 4.1.1. *$L_2^-(F)$ coincides with the space of functions $g(\lambda)$ representable by series of the form*

$$\sum_{k\le 0} c_k e^{i\lambda k}, \tag{4.1.8}$$

where $\sum_{k\le 0} |c_k|^2 < \infty$

Proof. Let us show that (4.1.8) is convergent in $L_2^-(F)$. For $n < m < 0$, we have

$$\begin{aligned} \int_{-\pi}^{\pi} \left| \sum_{n\le k\le m} c_k e^{i\lambda k} \right|^2 dF(\lambda) = \int_{-\pi}^{\pi} \left| \sum_{n\le k\le m} c_k e^{i\lambda k} \right|^2 f(\lambda)d\lambda \\ \le \frac{1}{c} \int_{-\pi}^{\pi} \left| \sum_{n\le k\le m} c_k e^{i\lambda k} \right|^2 d\lambda = \frac{2\pi}{c} \sum_{n\le k\le m} |c_k|^2 \le \frac{2\pi}{c} \sum_{k\le m} |c_k|^2, \end{aligned}$$

and this expression approaches zero as $m \to -\infty$. Since the partial sums of series (4.1.8) belong to $L_2^-(F)$, the sum will also belong to $L_2^-(F)$. Now suppose that

$$\lim_{n\to\infty}\int_{-\pi}^{\pi}\left|g(\lambda)-\sum_{k\le 0}c_k^{(n)}e^{i\lambda k}\right|^2 f(\lambda)d\lambda=0.$$

Then

$$\lim_{n\to\infty}\int_{-\pi}^{\pi}\left|g(\lambda)-\sum_{k\le 0}c_k^{(n)}e^{i\lambda k}\right|^2 d\lambda=0. \tag{4.1.9}$$

Since

$$\int_{-\pi}^{\pi}|g(\lambda)|^2d\lambda\le\frac{1}{c}\int_{-\pi}^{\pi}|g(\lambda)|^2f(\lambda)d\lambda<\infty,$$

it follows that

$$g(\lambda)=\sum_k c_k e^{i\lambda k}\ \text{with}\ \sum|c_k|^2<\infty.$$

But (4.1.9) leads to

$$\sum_{k\le 0}|c_k-c_k^{(n)}|^2+\sum_{k>0}|c_k|^2\to 0,$$

that is, $\sum_{k>0}|c_k|^2=0$. The lemma is proved. □

Consider now one-step prediction ($m=1$). From (4.1.7) we obtain

$$\int_{-\pi}^{\pi}e^{-i\lambda n}[g_1(\lambda)-e^{i\lambda}]f(\lambda)d\lambda=0,\quad n\le 0,$$

or

$$\int_{-\pi}^{\pi}e^{i\lambda(n+1)}[g_1(\lambda)e^{-i\lambda}-1]f(\lambda)d\lambda=0,\quad n\ge 0. \tag{4.1.10}$$

The function $h_-(\lambda)=g_1(\lambda)e^{-i\lambda}-1$ belongs to the space L_2^- of functions of the form

$$\sum_{k\le 0}a_k e^{i\lambda k},\quad \sum|a_k|^2<\infty,$$

while

$$[g_1(\lambda)e^{-i\lambda}-1]f(\lambda)=h_+(\lambda)$$

belongs to the space L_2^+ of functions of the form

$$\sum_{k\ge 0}b_k e^{i\lambda k},\quad \sum|b_k|^2<\infty$$

(this is a consequence of (4.1.10)). If $f(\lambda)$ is representable in the form

$$f(\lambda) = \frac{h_+(\lambda)}{h_-(\lambda)}, \tag{4.1.11}$$

where $h_-(\lambda) \in L_2^-$ and $h_+(\lambda) \in L_2^+$, then one is able to find $g_1(\lambda)$: If

$$h_-(\lambda) = \sum_{k\le 0} a_k e^{i\lambda k}$$

and $a_0 \neq 0$, then

$$g_1(\lambda) = -\sum_{k\le -1} \frac{a_k}{a_0} e^{i\lambda(k+1)}.$$

In addition, (4.1.10) will hold, which is equivalent to (4.1.7) in our case.

(b) *Yaglom's method.* We now use (4.1.11) to find $h_-(\lambda)$ for one class of spectral densities which are often encountered in practical applications. They are the spectral densities of the form

$$f(\lambda) = \frac{P(e^{i\lambda})}{Q(e^{i\lambda})},$$

where P and Q are polynomials. P and Q are assumed to be non-vanishing on the unit circle. If $\varphi(z) = P(z)/Q(z)$ is analytic, then since it is real for $|z| = 1$, it follows from $P(z_0) = 0$ and $Q(z_1) = 0$ that $P(1/\bar{z}_0) = 0$ and $Q(1/\bar{z}_1) = 0$. Let $z_1, z_2, \ldots, z_m$ be the zeroes of $P(z)$ such that $|z_k| < 1$, and let $u_1, u_2, \ldots, u_n$ be the zeroes of $Q(z)$ such that $|u_k| < 1$. Then

$$\varphi(z) = A\frac{\prod_{k=1}^m \left[(z - z_k)\left(\frac{1}{z} - \bar{z}_k\right)\right]}{\prod_{k=1}^n \left[(z - u_k)\left(\frac{1}{z} - \bar{u}_k\right)\right]}$$

and so

$$f(\lambda) = A\frac{\prod_{k=1}^m (e^{i\lambda} - z_k)(e^{-i\lambda} - \bar{z}_k)}{\prod_{k=1}^n (e^{i\lambda} - u_k)(e^{-i\lambda} - \bar{u}_k)} = A\frac{\prod_{k=1}^m (1 - z_k e^{-i\lambda})(1 - \bar{z}_k e^{i\lambda})}{\prod_{k=1}^n (1 - u_k e^{-i\lambda})(1 - \bar{u}_k e^{i\lambda})},$$

$$h_+(\lambda) = A\prod_{k=1}^m (1 - \bar{z}_k e^{i\lambda}) \prod_{k=1}^n (1 - \bar{u}_k e^{i\lambda})^{-1},$$

$$h_-(\lambda) = \prod_{k=1}^n (1 - u_k e^{-i\lambda}) \prod_{k=1}^m (1 - z_k e^{-i\lambda})^{-1}.$$

(c) *Case of smooth density.* Suppose that $\ln f(\lambda)$ has an absolutely convergent Fourier series:

$$\ln f(\lambda) = \sum_{n=-\infty}^{\infty} a_n e^{i\lambda n} \text{ and } \sum |a_n| < \infty.$$

This is true, for example, when $f'(\lambda)$ is bounded.

Then

$$f(\lambda) = \frac{h_+(\lambda)}{h_-(\lambda)},$$

where

$$h_+ = -\exp\left\{\sum_{n=0}^{\infty} a_n e^{i\lambda n}\right\}, \quad h_- = -\exp\left\{-\sum_{n=0}^{\infty} \bar{a}_n e^{-i\lambda n}\right\}.$$

$h_+(\lambda) \in L_2^+$ and $h_-(\lambda) \in L_2^-$ because the Fourier series for $h_\pm(\lambda)$ can be derived by expanding the exponential function in a Taylor series and then collecting the coefficients of $e^{i\lambda n}$ for the various values of n. The zero Fourier coefficient for $h_-(\lambda)$ is 1.

Let us look at the prediction error. We obtain

$$\begin{aligned}\mathbf{E}|\xi_1 - \bar{\xi}_1|^2 &= \int_{-\pi}^{\pi} (e^{i\lambda} - g_1(\lambda))\overline{(e^{i\lambda} - g_1(\lambda))} f(\lambda) d\lambda \\ &= \int_{-\pi}^{\pi} (e^{i\lambda} - g_1(\lambda)) e^{-i\lambda} f(\lambda) d\lambda\end{aligned}$$

by using the fact that $g_1(\lambda) \in L_2^-(F)$ and therefore

$$\int_{-\pi}^{\pi} (e^{i\lambda} - g_1(\lambda))\overline{g_1(\lambda)} f(\lambda) d\lambda = 0.$$

Thus

$$\begin{aligned}\mathbf{E}|\xi_1 - \bar{\xi}_1|^2 &= \int_{-\pi}^{\pi} (1 - e^{-i\lambda} g_1(\lambda)) f(\lambda) d\lambda \\ &= -\int_{-\pi}^{\pi} h_+(\lambda) d\lambda = \int_{-\pi}^{\pi} \exp\left\{\sum_{n=0}^{\infty} a_n e^{i\lambda n}\right\} d\lambda \\ &= e^{a_0} \int_{-\pi}^{\pi} \left[1 + \sum_{m=1}^{\infty} \frac{1}{m!} \left(\sum_{n=1}^{\infty} a_n e^{i\lambda n}\right)^m\right] d\lambda = 2\pi e^{a_0}.\end{aligned}$$

But

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln f(\lambda) d\lambda,$$

and so

$$\mathbf{E}|\xi_1 - \bar{\xi}_1|^2 = 2\pi \exp\left\{\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln f(\lambda) d\lambda\right\}.$$

4.1.3 Filtering One Stationary Sequence by Another

We now consider a two-dimensional stationary sequence $\{(\xi_k, \eta_k), k = 0, \pm 1, \pm 2, \ldots\}$ in which $\{\xi_k\}$ is the observable sequence and $\{\eta_k\}$ is to be estimated. Write

$$r_{\xi\xi}(k) = \mathbf{E}\xi_k\xi_0, \quad r_{\xi\eta}(k) = \mathbf{E}\xi_k\eta_0, \quad r_{\eta\xi}(k) = \mathbf{E}\eta_k\xi_0, \quad r_{\eta\eta}(k) = \mathbf{E}\eta_k\eta_0.$$

(we are assuming that $\mathbf{E}\xi_k = \mathbf{E}\eta_k = 0$). The following spectral representations hold:

$$r_{\xi\xi}(k) = \int_{-\pi}^{\pi} e^{i\lambda k} dF_{\xi\xi}(\lambda), \quad r_{\xi\eta}(k) = \int_{-\pi}^{\pi} e^{i\lambda k} dF_{\xi\eta}(\lambda),$$

$$r_{\eta\xi}(k) = \int_{-\pi}^{\pi} e^{i\lambda k} dF_{\eta\xi}(\lambda), \quad r_{\eta\eta}(k) = \int_{-\pi}^{\pi} e^{i\lambda k} dF_{\eta\eta}(\lambda),$$

$$\xi_k = \int_{-\pi}^{\pi} e^{i\lambda k} dy_\xi(\lambda), \qquad \eta_k = \int_{-\pi}^{\pi} e^{i\lambda k} dy_\eta(\lambda).$$

In this, $(y_\xi(\lambda), y_\eta(\lambda))$ is a vector function with orthogonal increments and

$$dF_{\xi\xi}(\lambda) = \mathbf{E}|dy_\xi(\lambda)|^2, \quad dF_{\xi\eta}(\lambda) = \mathbf{E}dy_\xi(\lambda)\overline{dy_\eta}(\lambda),$$

$$F_{\eta\xi}(\lambda) = \overline{F_{\xi\eta}(\lambda)}, \qquad dF_{\eta\eta}(\lambda) = \mathbf{E}|dy_\eta(\lambda)|^2.$$

(a) *Filtering by a whole sequence.* In this instance, $\{\xi_k\}$ is observed completely. It is necessary to estimate, say, η_0. The estimate for η_0 is expressible as

$$\bar{\eta}_0 = \int_{-\pi}^{\pi} g_0(\lambda) dy_\xi(\lambda) \text{ with } g_0(\lambda) \in L_2(F_{\xi\xi}).$$

The condition

$$\mathbf{E}\bar{\eta}_0\xi_k = \mathbf{E}\eta_0\xi_k = r_{\xi\eta}(k)$$

must hold for all k. This relation is equivalent to

$$\int_{-\pi}^{\pi} g_0(\lambda) e^{-i\lambda k} dF_{\xi\xi}(\lambda) = r_{\xi\eta}(k) = \int_{-\pi}^{\pi} e^{i\lambda k} dF_{\xi\eta}(\lambda) = \int_{-\pi}^{\pi} e^{-i\lambda k} dF_{\eta\xi}(\lambda)$$

(the realness of $r_{\xi\eta}(k)$ was used). From the relation

$$\int_{-\pi}^{\pi} g_0(\lambda) e^{-i\lambda k} dF_{\xi\xi}(\lambda) = \int_{-\pi}^{\pi} e^{-i\lambda k} dF_{\eta\xi}(\lambda), k = 0, \pm 1, \pm 2, \ldots,$$

it follows that

$$\int_{-\pi}^{\pi} \psi(\lambda) g_0(\lambda) dF_{\xi\xi}(\lambda) = \int_{-\pi}^{\pi} \psi(\lambda) dF_{\eta\xi}(\lambda)$$

if $\psi(\lambda)$ is any continuous function. This yields

$$g_0(\lambda) = \frac{dF_{\eta\xi}(\lambda)}{dF_{\xi\xi}(\lambda)}.$$

(b) *Nonpredictive filtering.* In this case, the sequence $\{\xi_k\}$ is observed at the present time (taken to be 0) and η_0 is to be estimated. The estimator is then of the form

$$\bar{\eta}_0 = \int_{-\pi}^{\pi} g_0(\lambda) dy_\xi(\lambda), \quad g_0(\lambda) \in L_2^-(F_{\xi\xi}),$$

and for all $k \le 0$,

$$\mathbf{E}\bar{\eta}_0 \xi_k = r_{\xi\eta}(k)$$

or

$$\int_{-\pi}^{\pi} g_0(\lambda) e^{-i\lambda k} dF_{\xi\xi}(\lambda) = \int_{-\pi}^{\pi} e^{-i\lambda k} dF_{\eta\xi}(\lambda), \quad k \le 0. \tag{4.1.12}$$

Assume that the spectral density $f_{\xi\xi}(\lambda) = dF_{\xi\xi}(\lambda)/d\lambda$ exists and that $c < f_{\xi\xi}(\lambda) < 1/c$ for some positive c. Then

$$g_0(\lambda) = \sum_{k \le 0} c_k e^{i\lambda k}.$$

It is easy to see that $F_{\eta\xi}(\lambda)$ is absolutely continuous with respect to $F_{\xi\xi}(\lambda)$. Hence, under the above assumption, $f_{\eta\xi}(\lambda) = dF_{\eta\xi}(\lambda)/d\lambda$ exists.

From (4.1.12) it follows that

$$\int_{-\pi}^{\pi} [g_0(\lambda) f_{\xi\xi}(\lambda) - f_{\eta\xi}(\lambda)] e^{-i\lambda k} d\lambda = 0, \quad k \le 0$$
$$g_0(\lambda) f_{\xi\xi}(\lambda) - f_{\eta\xi}(\lambda) = e^{i\lambda} z_+(\lambda),$$

with $z_+(\lambda) \in L_2^+$. Suppose that

$$f_{\xi\xi}(\lambda) = \frac{f_-(\lambda)}{f_+(\lambda)},$$

where $f_+(\lambda) \in L_2^+$, $f_-(\lambda) \in L_2^-$, and $f_-^{-1}(\lambda) \in L_2^-$. As was shown in Sect. 4.1.2(c), such a representation is possible if $\ln f_{\xi\xi}(\lambda)$ can be expanded in an absolutely convergent Fourier series. Then

$$g_0(\lambda) f_-(\lambda) = f_{\eta\xi}(\lambda) f_+(\lambda) + e^{i\lambda} z_+(\lambda) f_+(\lambda). \tag{4.1.13}$$

Introduce an operator P_- in the space L_2 of square-integrable functions on $[-\pi, \pi]$ by $P_- g(\lambda) = \sum_{n \le 0} a_n e^{i\lambda n}$ if $g(\lambda) = \sum_{n=-\infty}^{\infty} a_n e^{i\lambda n}$. Evidently, $P_-(g_0(\lambda) f_-(\lambda)) = g_0(\lambda) f_-(\lambda)$ and $P_-(e^{i\lambda} z_+(\lambda) f_+(\lambda)) = 0$. Therefore

$$g_0(\lambda) = P_-(f_{\eta\xi}(\lambda) f_+(\lambda)) f_-^{-1}(\lambda). \tag{4.1.14}$$

4.2 Nonlinear Filtering

4.2.1 General Remarks

Nonlinear filtering is similar to the linear problem except that the estimator of a random variable may be any arbitrary function of the observations. Let $\xi_1, \dots, \xi_n$ be the observed variables. It is necessary to estimate the variable η. We assume that $\mathbf{E}|\eta|^2 < \infty$ and we look for a measurable function $g(x_1, \dots, x_n)$ on R^n that minimizes $\mathbf{E}(\eta - g(\xi_1, \dots, \xi_n))^2$. We have

$$\begin{aligned}
&\mathbf{E}(\eta - g(\xi_1, \dots, \xi_n))^2 = \mathbf{E}\mathbf{E}((\eta - g(\xi_1, \dots, \xi_n))^2 | \xi_1, \dots, \xi_n) \\
&= \mathbf{E}(\mathbf{E}(\eta^2 | \xi_1, \dots, \xi_n) - 2\mathbf{E}(\eta | \xi_1, \dots, \xi_n) g(\xi_1, \dots, \xi_n) + g^2(\xi_1, \dots, \xi_n)) \\
&= \mathbf{E}[\eta^2 - (\mathbf{E}(\eta | \xi_1, \dots, \xi_n))^2] + \mathbf{E}(\mathbf{E}(\eta | \xi_1, \dots, \xi_n) - g(\xi_1, \dots, \xi_n))^2.
\end{aligned}$$

The first term on the right-hand side is independent of the choice of g and the second will be smallest if it equals zero, that is, if with probability 1

$$g(\xi_1, \dots, \xi_n) = \mathbf{E}(\eta | \xi_1, \dots, \xi_n).$$

Let $\{\xi_\lambda, \lambda \in \Lambda\}$ be a family of observable variables. Denote by $\mathcal{F}^\xi$ the smallest σ-algebra with respect to which all of the ξ_λ are measurable. Estimators then are $\mathcal{F}^\xi$-measurable variables ζ satisfying $\mathbf{E}\zeta^2 < \infty$. To each such variable there is a sequence of measurable functions $g_n(x_1, \dots, x_n)$ and $\lambda_i \in \Lambda$ such that

$$\lim_{n \to \infty} \mathbf{E}|\zeta - g_n(\xi_{\lambda_1}, \dots, \xi_{\lambda_n})|^2 = 0.$$

Let $L_2(\Omega, \mathbf{P})$ be the Hilbert space of all ζ defined on the initial probability space for which $\mathbf{E}\zeta^2 < \infty$. $L_2(\mathcal{F}^\xi)$ is the subspace of $\mathcal{F}^\xi$-measurable variables. Then the best estimator of η is the $\hat{\eta} \in L_2(\mathcal{F}^\xi)$ that minimizes $\mathbf{E}|\eta - \hat{\eta}|^2$. This means that $\hat{\eta}$ is the projection of η on $L_2(\mathcal{F}^\xi)$. Therefore $\mathbf{E}\eta\xi = \mathbf{E}\hat{\eta}\xi$ for every bounded $\mathcal{F}^\xi$-measurable ζ or, in other words,

$$\hat{\eta} = \mathbf{E}(\eta | \mathcal{F}^\xi). \tag{4.2.1}$$

Nonlinear filtering is reduced to finding the conditional expectation of a random variable given the σ-algebra generated by the observed variables.

4.2.2 Change-Point Problem

Given a process $\theta_t = I_{\{t > \tau\}}$. The time τ is a change-point (or disorder moment) of some device. So long as $\theta_t = 0$, the device is in working condition. When $\theta_t = 1$, the device has broken down. It is required to ascertain the change-point as precisely as possible if θ_t is being observed in the presence of additive noise, that is, the process $x_t = \theta_t + \varepsilon_t$ is observed. It is customary to assume that ε_t does not depend on θ_t and has independent values at different moments of time.

(a) *Discrete time.* Suppose that $p_k = \mathbf{P}\{\tau = k\}$ and the distribution of the ε_k's are given. The ε_k's will be taken to be independent and identically distributed. If $F(x)$ is the distribution function of ε_k, then it is assumed that $F(x-1) = \mathbf{P}\{\varepsilon_k + 1 < x\}$ is absolutely continuous with respect to $F(x)$ and that the function

$$\varphi(x) = \frac{dF(x-1)}{dF(x)}$$

is positive $dF(x)$-almost everywhere. Then $dF(x+1)/dF(x) = 1/\varphi(x+1)$ or, in other words, $F(x)$ is also absolutely continuous with respect to $F(x-1)$. If $F(x-1)$ and $F(x)$ were mutually singular, then we could determine τ expeditiously without delay from the observed values of $x_k = \theta_k + \varepsilon_k$ as follows. Let A be a Borel set such that

$$\mathbf{P}\{\varepsilon_1 \in A\} = 0, \quad \mathbf{P}\{\varepsilon_1 + 1 \in A\} = 1.$$

Then $\tau = k$ if $I_A(x_i) = 0$ when $i < k$ and $I_A(x_k) = 1$.

Therefore assuming the equivalence of $F(x-1)$ and $F(x)$ makes the problem more meaningful. It is natural to imagine that there is a loss $a_{nk} > 0$ by having decided $\tau = k$ and having stopped the process at that time when in fact $\tau = n$. The problem is then to minimize the loss. We shall seek a sequential solution to the problem. At each moment of time, a decision is made to stop the process or continue it. The decision is made on the basis of observations of the process to that time.

Introduce the variables

$$z_k = \prod_{i=1}^{k} \frac{1}{\varphi(x_i)}, \quad z_0 = 1.$$

This is also obviously an observable sequence. It turns out that to make a decision it is sufficient to know this sequence since the conditional distribution of τ, given $x_1, \ldots, x_n$, can be expressed in terms of $z_1, \ldots, z_n$ as follows:

$$\mathbf{P}\{\tau = m | x_1, \ldots, x_n\} = \frac{p_m z_{m-1}}{\sum_{i=1}^{n} p_i z_{i-1} + z_n \sum_{j=n+1}^{\infty} p_j}, \quad m \le n,$$

$$\mathbf{P}\{\tau = m | x_1, \ldots, x_n\} = \frac{p_m z_n}{\sum_{i=1}^{n} p_i z_{i-1} + z_n \sum_{j=n+1}^{\infty} p_j}, \quad m > n. \tag{4.2.2}$$

The expressions (4.2.2) are the filtering formulas for the change-point problem. Knowing these probabilities, we can evaluate the quantity

$$W_n(x_1, \ldots, x_n) = \sum a_{nm} \mathbf{P}\{\tau = m | x_1, \ldots, x_n\},$$

the loss resulting if the stoppage is at time m subject to the observations $x_1, \ldots, x_n$. There now remains the optimum stopping problem for the sequence $\{\eta_n = w_n(x_1, \ldots, x_n), n \ge 1\}$: Find a stopping time ζ that minimizes $\mathbf{E}\eta_\zeta$. This kind of problem was treated in the theory of controlled processes.

(b) *Continuous time.* In the case of continuous time, it is natural to replace θ_t, ε_t and x_t by integrals involving these processes (instead of a process with independent values, ε_t is considered to be a process with independent increments). Therefore the observed process (denoted as before by x_t) is taken to be

$$x_t = \int_0^t I_{\{\tau<s\}} ds + \xi_t, \tag{4.2.3}$$

where ξ_t is a homogeneous process with independent increments that does not depend on τ, the change-point time subject to estimation. We assume that τ has a density $\varphi(s)$:

$$\mathbf{P}\{\tau < t\} = \int_0^t \varphi(s) ds,$$

and that ξ_t is a Wiener process with $\mathbf{E}\xi_t = 0$ and $\mathbf{V}\xi_t = bt$. Our aim is to find filtering formulas analogous to (4.2.2). These formulas should give the conditional density of τ, given $x(s)$ for $s \leq$ t.

Choose a positive h and let

$$\begin{aligned}
&\varepsilon_k^h = \frac{1}{h}[\xi_{kh} - \xi_{(k-1)h}], \qquad x_k^h = \frac{1}{h}[x_{kh} - x_{(k-1)h}],\\
&\theta_k^h = \frac{1}{h}\int_{(k-1)h}^{kh} I_{\{\tau<s\}} ds, \quad p_m^h = \mathbf{P}\{(m-1)h \leq \tau \leq mh\}.
\end{aligned}$$

Applying (4.2.2), one can write

$$\mathbf{P}\{\tau \in [m-1)h, mh] | x_1^h, \ldots, x_n^h\} = \frac{z_{m\wedge n}^h}{\sum_k z_{k\wedge n}^h},$$

where

$$\begin{aligned}
z_m^h &= \prod_{k=1}^{m-1} \left(\frac{2\pi b}{h}\right)^{-1/2} \exp\left\{-\frac{hx_k^2}{2b}\right\} \left(\frac{2\pi b}{h}\right)^{-1/2}\\
&\quad \times \int_{(m-1)h}^{mh} \exp\left\{-\frac{h(x_k - (mh-u))^2}{2b}\right\} \varphi(u) du\\
&\quad \times \prod_{k=m+1}^{n} \left(\frac{2\pi b}{h}\right)^{-1/2} \exp\left\{-\frac{h(x_k^h - 1)^2}{2b}\right\}\\
&\quad \times \prod_{k=1}^{n} \left(\left(\frac{2\pi b}{h}\right)^{-1/2} \exp\left\{-\frac{h(x_k - 1)^2}{2b}\right\}\right)^{-1}\\
&= \exp\left\{-h\sum_{k=1}^{m-1} x_k^h + (m-1)h\right\} \int_{(m-1)h}^{mh} e^{-x_m^h(mh-u) + \frac{(mh-u)^2}{bh}} \varphi(u) du.
\end{aligned}$$

Letting $h \to 0$, $nh \to t$ and $mh \to s$, we arrive at the following expression for the conditional density $\varphi(s|\mathcal{F}_t)$, where $\mathcal{F}_t$ is the σ-algebra generated by x_s for $s \le t$:

$$\varphi(s|\mathcal{F}_t) = \frac{\varphi(s)\exp\left\{-\int_0^{s\wedge t} x(u)du - s\wedge t\right\}}{\int_0^\infty \varphi(u)\exp\left\{-\int_0^{u\wedge t} x(u)du - u\wedge t\right\}du}\,. \tag{4.2.4}$$

4.2.3 Filtering of Markov Chains

This problem is similar to the change-point problem except θ_t is a Markov process with a finite set of states. The observed process is $x_t = \theta_t + \varepsilon_t$, where ε_t has independent increments. It is required to find an estimate for θ_t.

(a) *Discrete time.* Let $\{\theta_t, t = 0, 1, 2, \ldots\}$ be a homogeneous Markov chain with states $\{1, 2, \ldots, r\}$. The one-step transition probability is $p_{ij}, i, j \in \{1, \ldots, r\}$, $p_i(0)$ is the initial distribution of the chain, $p_{ij}(n)$ is the n-step transition probability and $p_i(n)$ is the distribution at the n-th step; $\{\varepsilon_t\}$ is assumed to be a sequence of independent and identically distributed random variables collectively not depending an θ_t. The problem is to find the conditional probabilities

$$\mathbf{P}\{\theta_t = k|x_1, \ldots, x_n\},\ \ k = 1, \ldots, r,\ \ t = 0, 1, 2, \ldots,\ \ n = 0, 1, 2, \ldots. \tag{4.2.5}$$

If we know them, we have complete information about θ_t contained in the observations $x_1, \ldots, x_n$.

Let $F(x)$ be the distribution function of ε_1. Then the distribution function of $\varepsilon_1 + i$, $i = 1, \ldots, r$, is $F(x - i)$. Choose a distribution function $G(x)$ so that $F(x - i)$ is absolutely continuous with respect to $G(x)$ and put

$$\varphi_i(x) = \frac{dF(x-i)}{dG(x)}.$$

Let us find the distribution of $\{x_0, x_1, \ldots, x_n\}$. The conditional distribution of this sequence given $\theta_0, \ldots, \theta_k$ is for $k > n$ equal to

$$\begin{aligned}&\mathbf{P}(x_0 \in A_0, \ldots, x_n \in A_n|\theta = i_0, \ldots, \theta_k = i_k)\\ &\quad = \int_{A_0} dF(y_0 - i_0)\int_{A_1} dF(y_1 - i_1)\ldots\int_{A_n} dF(y_n - i_n).\end{aligned}$$

Thus it depends only on $\theta_0, \ldots, \theta_n$. This conditional probability is expressible in terms of the functions φ_i as

$$\begin{aligned}&\mathbf{P}\{x_0 \in A_0, \ldots, x_n \in A_n|\theta_0, \ldots, \theta_n\}\\ &\quad = \int_{A_0} I_{A_0}(y_0)\ldots I_{A_n}(y_n)\prod_{i=0}^n \varphi_{\theta_i}(y_i)\prod dG(y_k).\end{aligned}$$

Let $\alpha_n(\theta_0, \ldots, \theta_n, y_0, \ldots, y_n) = \prod_{i=0}^n \varphi_{\theta_i}(y_i)$ and

$$\alpha_n(y_0,\dots,y_n) = \mathbf{E}(\theta_0,\dots,\theta_n,y_0,\dots,y_n)$$
$$= \sum_{i_0,\dots,i_n} \alpha_n(i_0,\dots,i_n,y_0,\dots,y_n)p_0(i_0)p(i_0,i_1)\dots p(i_{n-1},i_n). \quad (4.2.6)$$

Then

$$\mathbf{P}\{x_0 \in A_0,\dots,x_n \in A_n\} = \int I_{A_0}(y_0)\dots I_{A_n}(y_n)\alpha_n(y_0,\dots,y_n)\prod_{k=0}^{n} dG(y_k).$$

To evaluate $\alpha_n(y_0,\dots,y_n)$, it is convenient to introduce the functions

$$\hat{\alpha}_n(i_0,y_0,\dots,y_n,i)$$
$$= \sum_{i_1,\dots,i_{n-1}} \alpha_n(i_0,\dots,i_{n-1},i,y_0,\dots,y_n)p(i_0,i_1)\dots p(i_{n-1},i). \quad (4.2.7)$$

Then

$$\alpha_n(y_0,\dots,y_n) = \sum_{i_0,i} \hat{\alpha}_n(i_0,y_0,\dots,y_n,i)p_0(i_0), \quad (4.2.8)$$

and

$$\hat{\alpha}_{n+1}(i_0,y_0,\dots,y_{n+1},i) = \sum_j \hat{\alpha}_n(i,y_0,\dots,y_n,j)p_{ji}(y_{n+1})\varphi_i(y_{n+1}). \quad (4.2.9)$$

Formulas (4.2.9) and (4.2.8) enable one to compute $\alpha_n(y_0,\dots,y_n)$ recursively.

Now consider the conditional distribution of $\theta_0,\dots,\theta_n$ given $x_1,\dots,x_n$. Since

$$\mathbf{P}\{x_0 \in A_0,\dots,x_n \in A_n,\theta_0 = i_0,\dots,\theta_n = i_n\}$$
$$= \mathbf{P}\{x_0 \in A_0,\dots,x_n \in A_n|\theta_0 = i_0,\dots,\theta_n = i_n\}$$
$$\times p_0(i_0)p(i_0,i_1)\dots p(i_{n-1},i_n),$$

it follows that

$$\mathbf{P}(\theta_0 = i_0,\dots,\theta_n = i_n|x_0,\dots,x_n)$$
$$= \alpha_n(i_0,\dots,i_n,x_0,\dots,x_n)p_0(i_0)p(i_0,i_1)\dots p(i_{n-1},i_n)/\alpha_n(x_0,\dots,x_n).$$

Therefore

$$\mathbf{P}\{\theta_n = j|x_0,\dots,x_n\} = \frac{\sum_{i_0}\hat{\alpha}_n(i_0,x_0,\dots,x_n,j)p_0(i_0)}{\alpha_n(x_0,\dots,x_n)}, \quad (4.2.10)$$

for $m < n$

$$\mathbf{P}\{\theta_m = j|x_0,\dots,x_n\} = \frac{1}{\alpha_n(x_0,\dots,x_n)}\sum_{i_0,i} p_0(i_0)\hat{\alpha}_m(i_0,x_0,\dots,x_m,j)$$
$$\times \alpha_{n-m}(j,x_{m+1},\dots,x_n,i), \quad (4.2.11)$$

and for $m > n$

$$\mathbf{P}\{\theta_m = j | x_0, \dots, x_n\} = \frac{1}{\alpha_n(x_0, \dots, x_n)} \sum_{i_0, i} \hat{\alpha}(i_0, x_0, \dots, x_m, j) p_{ij}(m-n). \quad (4.2.12)$$

Thus the requisite conditions for the probabilities are expressed in terms of the functions $\hat{\alpha}_n(i, x_0, \dots, x_n, j)$ which can be computed recursively using (4.2.9), (4.2.8), (4.2.10), (4.2.11) and (4.2.12), with $\hat{\alpha}_0(i, y_0, j) = \varphi_i(y_0)\delta_{ij}$.

Formulas (4.2.8)–(4.2.12) are the filtering equations for a time-discrete Markov chain. The advantage of these equations is that all of the conditional probabilities are expressed in terms of the exact same function $\hat{\alpha}_n(i, x_0, \dots, x_n, j)$ (taking (4.2.8) into account) satisfying recursion equation (4.2.9). This fact makes it possible to carry these formulas over to continuous time.

(b) *Continuous time.* Now suppose that $\{\theta_t, t \geq 0\}$ is a homogeneous Markov process with r states $\{1, 2, \dots, r\}$ and transition probabilities $p_{ij}(t), i, j \in \{1, \dots, r\}$ satisfying the relation

$$\lim_{t \downarrow 0} \frac{p_{ij}(t) - \delta_{ij}}{t} = a_{ij}.$$

Consequently, the probabilities $p_i(t) = \mathbf{P}\{\theta_t = i\}$ satisfy the forward Kolmogorov equation

$$\frac{d}{dt} p_i(t) = \sum_j p_j(t) a_{ji}.$$

Let $c_i, i \in \{1, \dots, r\}$, be a real function. The observed process is

$$x_t = \int_0^t c(\theta_s) ds + \xi_t, \quad (4.2.13)$$

where ξ_t is a Wiener process for which $\mathbf{E}\xi_t = 0$ and $\mathbf{V}\xi_t = bt$. Our aim is to construct the best estimate for $c(\theta_s)$ from observations of the process x_u on the interval $[0, t]$ or, in other words, to find

$$\mathbf{E}(c(\theta_s) | x_u, u \leq t).$$

It is convenient to denote the path of the process θ_u on $[s, t]$ by θ_t^s. Similarly, x_t^s is the path of x_u on $[s, t]$. Let

$$\alpha(t, \theta_t^0, x_t^0) = \lim_{n \to \infty} \alpha_n(\theta_0^h, \dots, \theta_n^h, x_0^h, \dots, x_n^h) = \lim_{n \to \infty} \prod_{k=0}^{n} \varphi_{\theta_k^h}^h(x_k^h), \quad h = t/n, \quad \theta_k^h = \theta_{kh},$$

$$x_k^h = x_{kh} - x_{(k-1)h},$$

$$\varphi_\theta^h(x) = \frac{1}{\sqrt{2\pi bh}} \exp\left\{-\frac{(x - hc(\theta))^2}{2bh}\right\} \left(\frac{1}{2\pi bh}\right)^{-1/2} \exp\left\{\frac{x^2}{2b}\right\}$$
$$= \exp\left\{\frac{c(\theta)x}{b} - \frac{c^2(\theta)h}{b}\right\}.$$

Thus

$$\alpha(t, \theta_t^0, x_t^0) = \exp\left\{\frac{1}{b}\int_0^t c(\theta_s)dx_s - \frac{1}{b}\int_0^t c^2(\theta_s)dx_s\right\}. \tag{4.2.14}$$

The first integral on the right is defined for any continuous function since $c(\theta_s)$ is a step-function.

Let

$$\hat{\alpha}(i, t, y_t^0, j) = \mathbf{E}(\alpha(t, \theta_t^0, y_t^0)|\theta_0 = i, \theta_t = j), \tag{4.2.15}$$

where y_u is an arbitrary continuous random function. Taking the limit in (4.2.10)–(4.2.12) after multiplying by $c(j)$ and summing over j, we find that

$$\mathbf{E}(c(\theta_t)|x_t^0) = \frac{1}{\alpha_t(x_t^0)} \sum_{i,j} c(j)\hat{\alpha}(i, t, x_i^0, j)p_i(0)p_{ij}(t), \tag{4.2.16}$$

$$\mathbf{E}(c(\theta_s)|x_t^0) = \frac{1}{\alpha_t(x_t^0)} \sum_{i,j,k} c(j)\hat{\alpha}(i, s, x_s^0, j)\hat{\alpha}(j, t-s, x_t^s, k)$$
$$\times p_i(0)p_{ij}(s)p_{jk}(t-s), \quad s < t, \tag{4.2.17}$$

and

$$\mathbf{E}(c(\theta_s)|x_t^0) = \frac{1}{\alpha_t(x_t^0)} \sum_{i,j,k} c(j)\hat{\alpha}(i, t, x_t^0, k)p_i(0)p_{ik}(t)$$
$$\times p_{kj}(s-t), \quad s > t, \tag{4.2.18}$$

where

$$\alpha_t(x_t^0) = \sum_{i,k} \hat{\alpha}(i, t, x_t^0, k)p_i(0)p_{ik}(t). \tag{4.2.19}$$

To determine the functions $\hat{\alpha}(i, t, x_t^0, k)$ in terms of which the requisite conditional probabilities are expressed, it is convenient to introduce the functions

$$\beta_{ij}(t) = \hat{\alpha}(i, t, x_t^0, j)p_{ij}(t). \tag{4.2.20}$$

Passage to the limit in (4.2.9) yields a system of stochastic differential equations for $\beta_{ij}(t)$:

$$d\beta_{ij}(t) = \sum_k \beta_{ik}(t)a_{kj}(t)dt + \beta_{ij}(t)c_j dx(t), \tag{4.2.21}$$

which must be solved subject to the initial condition $\beta_{ij}(0) = \delta_{ij}$.

Such equations have been studied in the theory of Markov processes.

Historic and Bibliographic Comments

General questions of mathematical statistics are handled by Cramér (1974), Neyman (1950), Van der Waerden (1969) and Zacks (1971). Jerzy Neyman was one of the eminent statisticians of the twentieth century and his book familiarizes the nonspecialist with the essence of probability and statistical problems both in a simple and profound way. The main thrust of Cramér (1974) is his formulation of statistical problems and presentation of methods for solving them. The book contains many interesting and meaningful examples. Zacks (1971) treats the concepts of statistics, the related problems and also ways of solving them. The book is theoretical and is intended for specialists. Also touching on the content of the first chapter is the book by Wald (1947). It contains a presentation of sequential analytic methods of testing statistical hypotheses.

The books by Bellman (1957), Dynkin (1975), Gikhman (1977), Kalman (1969), Kushner (1967), Krylov (1977) and Shiryaev (1976) consider problems on controlled processes. Bellman derives equations for the control cost of controlled Markov random processes. Gikhman (1977) gives the general theory of time-discrete and time-continuous controlled processes including both Markov chains and Markov processes. Dynkin (1975) presents the general concepts of controlled Markov processes, derives the equations for the control cost and proves the existence of optimum controls. Kalman (1969) gives in particular an introduction to the theory of controlled random processes. Krylov's book (1977) is devoted to controlled processes defined by stochastic differential equations of diffusion type. He studies the nonlinear partial differential equations which are Bellman's equations for an optimum control. The book is intended for specialists. A main theme of Shiryaev's book (1976) is the optimum stopping of a Markov chain or Markov process. In particular, it solves the change-point problem.

Problems in information theory are considered by Feinstein (1958), McMillan (1953) and Shannon (1948). Feinstein discusses the basic concepts and theorems of information theory. McMillan's article is devoted to the capacity

of a Markov-type communication channel and proves a corresponding version of Shannon's theorem.

The books and papers by Bucy (1965), Cramér (1940), Kolmogorov (1941), Liptser (1974), Wiener (1949) and Yaglom (1952) are devoted to filtering. Bucy discusses the theory of nonlinear filtering. Liptser (1974) contains much material on martingale theory and stochastic equations. Its basic aim is to construct a theory of nonlinear filtering. It is intended for specialists. Wiener (1949) presents the theory of extrapolation, interpolation, filtering and smoothing of stationary sequences and methods based on the factorization of analytic functions. The general theory is illustrated by the solution of engineering problems. Kolmogorov (1941) reduces problems involving stationary sequences to problems of analysis in Hilbert spaces. This is a fundamental approach to solving them. Cramér's article (1940) presents some solutions to basic problems involving stationary processes. Yaglom's large review article (1952) contains the basic results with complete proofs including those due to the author on the theory of stationary processes.

Author Index

Subject Index

Printing: Mercedes-Druck, Berlin
Binding: Stein+Lehmann, Berlin