KTH Matematik

Lecture Notes: Probability and Random Processes at KTH

for

sf2940 Probability Theory
Edition: 2017

Timo Koski
Department of Mathematics
KTH Royal Institute of Technology
Stockholm, Sweden

Contents

Foreword

This text corresponds to the material in the course on intermediate probability calculus for masters students that has been lectured at the Department of mathematics at KTH during the last decades. Here this material is organized into one document. There are topics that were not included in the earlier courses and a few previous items have been omitted. The author is obviously indebted to Prof.em. Bengt Rosén and Prof.em. Lars Holst who have built up the course.

Boualem Djehiche, Gunnar Englund, Davit Karagulyan, Gaultier Lambert, Harald Lang, Pierre Nyquist and several others are thanked for having suggested improvements and for having pointed out several errors and mistakes in the earlier editions.

Chapter 1

Probability Spaces and Random Variables

1.1 Introduction

In the first courses on probability given at most universities of technology, see, e.g., [12, 16, 101] for a few excellent items in this educational genre, as well as in courses involving probability and random processes in physics and statistical physics, see [17, 58, 62, 73, 78] or reliability of structures [32, 77] or civil engineering [4], one seemingly considers all subsets, called events, of a space of outcomes. Then one treats a (in practice, finitely additive) probability as a positive total mass = 1 distributed on these events. When the goal is to train students in the use of explicit probability distributions and in statistical modelling for engineering, physics and economics problems, the approach is necessary and has definite didactic advantages, and need not be questioned (and the indicted authors are, of course, well aware of the simplifications imposed).

There is, however, a need to introduce the language[1] and viewpoint of rigorous mathematical analysis, as argued in [43]. The precise (and abstract) mathematical theory requires a more restricted set of events than all the subsets. This leads us to introduce algebras of sets and sigma algebras of sets. The material below has approximately the same level of mathematical completeness as [20, 43, 44, 95] and [103, chapter1].

1.2 Terminology and Notations in Elementary Set Theory

We collect first a list of bullet points recapitulating some definitions, notations and rules of elementary (or naïve) set theory. The bulk of these are assumed to be familiar for the student, due to previous exposure via, e.g., [16] or any other equivalent first course in probability. Therefore the presentation is kept at a concise level, e.g., many of the rules of set theory stated below can be made evident by use of Venn diagrams, but these illustrations are not provided in this summary.

We start by postulating an abstract space consisting of elements, denoted here by ω. The elements are the smallest quantities we deal with. The abstract space is also called the universal set and (in probability calculus)

[1] On the other hand, one widely held opinion is expressed in [62, p.179] to the effect that '*the language favored by mathematicians . . . adds little that is of value to (physicists)*'. That notwithstanding, in [62, chapter 10] the merits of surveying 'the concepts and jargon of modern probability theory' (that is, what corresponds to chapters 1 and 2 in these notes) are recognized. The rationale is that a natural scientist or an engineer will learn how to interpret the basic points of a mathematical discourse in a preferred intuitive idiom.

denoted by Ω. Later on we shall refer to Ω as the outcome space or sample space and ω as an elementary outcome.

Example 1.2.1 The examples of Ω first encountered in courses of probability theory are simple. The outcomes ω of a toss of coin are heads and tails, and we write the universal set as

$$\Omega = \{ \text{ heads }, \text{ tails } \}.$$

■

Let now Ω be an abstract universal set and A, B e.t.c. denote sets, collections of elements in Ω.

- $\omega \in A$ means that an element ω *belongs to a set* A. $\omega \notin A$ means that ω *does not belong* to a set A.

- $\emptyset$ denotes the *empty set*, which has no elements.

- A^c is the *complement* set of A. It consists of all elements ω that do not belong to A. It follows that

  $$(A^c)^c = A.$$

  Since Ω is the universal set, we take

  $$\Omega^c = \emptyset.$$

- $\{\omega \in A \mid S(\omega)\}$ stands for the elements ω belonging to A that satisfy a property S.

- $A \subseteq B$ denotes the *inclusion* of sets. It means that A is a *subset* of B. This means that if $\omega \in A$, then $\omega \in B$. In addition, we have for any set $A \subseteq \Omega$.

  Note that $A \subseteq B$ and $B \subseteq A$ if and only if $A = B$.

  We use also on occasion the notation of strict inclusion $A \subset B$, which means that $A \neq B$.

- If $A \subseteq B$, then $B^c \subseteq A^c$.

- $\mathcal{P}(A)$ denotes the family of all subsets of A and is known as the *power set* of A.

- $A \cup B$ is the *union* of the sets A and B. The union consists of all elements ω such that $\omega \in A$ or $\omega \in B$ or both. We have thus

  $$A \cup \Omega = \Omega$$

  $$A \cup \emptyset = A$$

  $$A \cup A^c = \Omega$$

  $$A \cup B = B \cup A$$

  and

  $$A \cup A = A.$$

  For a sequence of sets $A_1, A_2, \ldots$ the union

  $$\cup_{i=1}^{\infty} A_i = A_1 \cup A_2 \cup \ldots$$

  consists of the elements ω such that there is at least one A_i such that $\omega \in A_i$.

- $A \cap B$ is the *intersection* of the sets A and B. The intersection consists of all elements ω such that $\omega \in A$ and $\omega \in B$. It is seen that

  $$A \cap \Omega = A$$

  $$A \cap \emptyset = \emptyset$$

  $$A \cap A^c = \emptyset$$

  $$A \cap B = B \cap A$$

  and

  $$A \cap A = A.$$

  For a sequence of sets $A_1, A_2, \ldots$ the intersection

  $$\cap_{i=1}^{\infty} A_i = A_1 \cap A_2 \cap \ldots$$

  consists of the elements ω such that $\omega \in A_i$ for all i.

- The sets A and B are said to be *disjoint* if $A \cap B = \emptyset$. The sets $A_1, A_2, \ldots, A_n$ are *pairwise disjoint* if all pairs A_i, A_j are disjoint for $i \neq j$.

- $A \setminus B$ is the *set difference* of the sets A and B. It is the complement of B in A, and thus contains all elements in A that are not in B, or, $\omega \in A$ and $\omega \notin B$. Therefore we get

  $$A \setminus B = A \cap B^c.$$

- *De Morgan's Rules*
  The following rules of computation are frequently useful in probability calculus and are easy to memorize.

  $$(A \cup B)^c = A^c \cap B^c$$

  $$(A \cap B)^c = A^c \cup B^c$$

  These two formulas are known as De Morgan's Rules.

  One can prove the countably infinite versions of De Morgan's Rules, too.

  $$\left(\cup_{i=1}^{\infty} A_i\right)^c = \cap_{i=1}^{\infty} A_i^c$$

  and

  $$\left(\cap_{i=1}^{\infty} A_i\right)^c = \cup_{i=1}^{\infty} A_i^c.$$

- It is also readily proved that we have the *distributive rules*

  $$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

  and

  $$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

- $A \times B$ is the *(Cartesian) product* of the sets A and B. It consists of all pairs (ω_1, ω_2) such that $\omega_1 \in A$ and $\omega_2 \in B$.

  The product $A_1 \times A_2 \times \ldots \times A_n$ consists of ordered n-tuples $(\omega_1, \omega_2, \ldots, \omega_n)$ such that $\omega_i \in A_i$ for each $i = 1, \ldots, n$.

  If $A_i = A$ for all i, then we write
  $$A^n = A \times A \times \ldots \times A$$
  as a product of n copies of A.

- **Intervals**
  If a and b are real numbers, $a < b$, then
  $$(a, b), [a, b), (a, b], [a, b]$$
  are *intervals with endpoints* a and b. These are subsets of the real line $\mathbf{R}$, here taken as a universal set with elements denoted by x (=a real number) such that $(a, b) = \{x \in \mathbf{R} \mid a < x < b\}$, $[a, b) = \{x \in \mathbf{R} \mid a \leq x < b\}$, $(a, b] = \{x \in \mathbf{R} \mid a < x \leq b\}$ and $[a, b] = \{x \in \mathbf{R} \mid a \leq x \leq b\}$. We take $[a, a) = \emptyset$. For (a, b) and $(a, b]$ we can let $a = -\infty$ and for $[a, b)$ and (a, b) we can allow $b = +\infty$. Hence we can write $(-\infty, \infty) = \{x \in \mathbf{R} \mid -\infty < x < \infty\}$. The set operations are, e.g., $(a, b)^c = (-\infty, a] \cup [b, \infty)$.

1.3 Algebras of Sets

Definition 1.3.1 (Algebra) Let Ω denote a universal set. A collection $\mathcal{A}$ of subsets of Ω is called an *algebra*, or *field* if

1. $\Omega \in \mathcal{A}$
2. If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$, where A^c denotes the complement of A.
3. If $A \in \mathcal{A}$ and $B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$.

■

Condition 1. above is known as **non-emptiness**, condition 2. above is known as **closure under complement**, and condition 3. above is known as **closure under union**. Note that if $A \in \mathcal{A}$ and $B \in \mathcal{A}$, then there is closure under intersection, $A \cap B \in \mathcal{A}$, too. This follows since $A^c, B^c \in \mathcal{A}$, hence $A^c \cup B^c \in \mathcal{A}$ and $A \cap B = (A^c \cup B^c)^c$ by De Morgan's rule. Since $\emptyset = \Omega^c$, $\emptyset \in \mathcal{A}$.

Example 1.3.1 $\mathcal{A} = \{\emptyset, \Omega\}$ is an algebra.

■

Example 1.3.2 If Ω is a finite set, then the power set $\mathcal{P}(\Omega)$ is an algebra.

■

Definition 1.3.2 (Sigma - Algebra a.k.a. Sigma- Field a.k.a σ -field) A collection $\mathcal{A}$ of subsets of Ω is called a σ - algebra/field if it satisfies

1. $\Omega \in \mathcal{A}$

2. If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$.

3. If $A_n \in \mathcal{A}$ for each n in a countable collection $(A_n)_{n=1}^{\infty}$, then $\cup_{n=1}^{\infty} A_n \in \mathcal{A}$.

■

Here the condition 3. is referred to as **closure under countable union**. If an algebra is finite, then it is also a Sigma - Algebra. A σ - algebra $\mathcal{A}$ is usually constructed by first choosing an algebra, say $\mathcal{C}$, of subsets of Ω that **generates** $\mathcal{A}$. By this we mean that we augment $\mathcal{C}$ by all possible countable unions of sets in $\mathcal{C}$, their complements, all possible countable unions of these complements ad infinitum. We shall describe this procedure in some more detail in the sequel, when $\Omega =$ the real line, denoted by $\mathbf{R}$.

Example 1.3.3 Let $\Omega = \{\text{heads}, \text{tails}\}$. Then

$$\mathcal{F} = \{\{\text{heads}\}, \{\text{tails}\}, \{\text{heads}, \text{tails}\}, \emptyset\}$$

is a sigma-field, and contains also all possible subsets of Ω.

■

Example 1.3.4 Let $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. Then

$$\mathcal{F}_{\min} = \{\emptyset, \{\omega_1, \omega_2, \omega_3, \omega_4\}\},$$

$$\mathcal{F}_1 = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \{\omega_1, \omega_2, \omega_3, \omega_4\}\}$$

and

$$\mathcal{F}_{\max} = \{\emptyset, \{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\},$$
$$\{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}, \{\omega_1, \omega_4\}, \{\omega_2, \omega_3\}, \{\omega_2, \omega_4\}, \{\omega_3, \omega_4\},$$
$$\{\omega_1, \omega_2, \omega_3\}, \{\omega_1, \omega_2, \omega_4\}, \{\omega_1, \omega_3, \omega_4\}, \{\omega_2, \omega_3, \omega_4\}, \Omega\}$$

are sigma-fields. Clearly

$$\mathcal{F}_{\min} \subset \mathcal{F}_1 \subset \mathcal{F}_{\max},$$

in the sense that, e.g., any set found in $\mathcal{F}_1$ is found also in $\mathcal{F}_{\max}$.

■

Example 1.3.5 Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$. Then

$$\{\emptyset, \{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \Omega\}$$

is NOT a sigma-field.

■

Let $A_n \in \mathcal{A}$ for each n in a countable collection $(A_n)_{n=1}^{\infty}$. Suppose that for all $n \geq 1$

$$A_n \subset A_{n+1}$$

and we say that $(A_n)_{n=1}^{\infty}$ is *increasing*. Then we can define

$$\lim_{n \to \infty} A_n \overset{\text{def}}{=} \cup_{n=1}^{\infty} A_n. \tag{1.1}$$

Then $\lim_{n\to\infty} A_n \in \mathcal{A}$. In words, ω is in the limit of a increasing sequence of events, if ω belongs to some A_n and thereby to infinitely many sets in the collection.
Suppose that for all $n \geq 1$

$$A_{n+1} \subset A_n$$

and we say that $(A_n)_{n=1}^{\infty}$ is *decreasing*. Then we can define

$$\lim_{n\to\infty} A_n \overset{\text{def}}{=} \cap_{n=1}^{\infty} A_n, \tag{1.2}$$

and $\lim_{n\to\infty} A_n \in \mathcal{A}$. In other words, ω is in the limit of a decreasing sequence of events, if ω belongs to all A_n.

Example 1.3.6 Let $\Omega = \mathbf{R}$ and suppose that we have a sigma field $\mathcal{A}$ such that all intervals of the form

$$\left[1, 2 - \frac{1}{n}\right) \in \mathcal{A}.$$

Then the sequence of events is increasing

$$\left[1, 2 - \frac{1}{n}\right) \subset \left[1, 2 - \frac{1}{n+1}\right)$$

and $[1,2) \in \mathcal{A}$, since

$$[1,2) = \lim_{n\to\infty} \left[1, 2 - \frac{1}{n}\right).$$

■

Note that a σ algebra is clearly an algebra, but the converse is not always true, as the following example shows:

Example 1.3.7 Let $\Omega = \mathbf{R}$ and let $\mathcal{A}$ denote the collection of subsets of the form:

$$\cup_{i=1}^{k} (a_i, b_i] \qquad -\infty \leq a_i < b_i < +\infty$$

for some $0 \leq k < +\infty$.
This is clearly an algebra, but it is not a sigma algebra. Consider the collection

$$A_n = \left(0, 2 - \frac{1}{n}\right] \qquad n \geq 1.$$

Then $\cup_{n=1}^{\infty} A_n = (0,2)$, which is not in $\mathcal{A}$. ■

Example 1.3.8 Suppose that we have a sigma field $\mathcal{A}$ such that all intervals of the form

$$(a, b) \in \mathcal{A},$$

where $a < b$ are real numbers. Then

$$\left(a - \frac{1}{n+1}, a + \frac{1}{n+1}\right) \subset \left(a - \frac{1}{n}, a + \frac{1}{n}\right)$$

and thus

$$\{a\} = \lim_{n\to\infty} \left(a - \frac{1}{n}, a + \frac{1}{n}\right),$$

which shows that the singleton set $\{a\}$ is an event, i.e., $\{a\} \in \mathcal{A}$.

■

Theorem 1.3.9 Given any collection $\mathcal{C}$ of subsets of a set Ω, there is a smallest algebra $\mathcal{A}$ containing $\mathcal{C}$. That is, there is an algebra $\mathcal{A}$ containing $\mathcal{C}$ such that if $\mathcal{B}$ is any algebra containing $\mathcal{C}$ then $\mathcal{B}$ contains $\mathcal{A}$.

Proof Let $\mathcal{F}$ denote the family of all algebras of subsets of Ω which contain $\mathcal{C}$. The axioms of set theory are required to justify the existence of this family; it is a subset of $\mathcal{P}(\mathcal{P}(\Omega))$ where $\mathcal{P}$ denotes taking the power set. Let $\mathcal{A} = \cap\{\mathcal{B}|\mathcal{B} \in \mathcal{F}\}$. Then, for any $A \in \mathcal{A}$ and $B \in \mathcal{A}$, $A \cup B \in \mathcal{B}$ for all $\mathcal{B} \in \mathcal{F}$ and hence $A \cup B \in \mathcal{A}$. Similarly, if $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$. It follows that $\mathcal{A}$ is an algebra and that $\mathcal{C} \subseteq \mathcal{A}$. Furthermore, $\mathcal{A} \subseteq \mathcal{B}$ for any algebra $\mathcal{B}$ containing $\mathcal{C}$. ■

Lemma 1.3.10 Let C denote an indexing set. If $(\mathcal{A}_c)_{c\in C}$ is a collection of σ algebras, then $\mathcal{A} = \cap_c \mathcal{A}_c$ (that is the collection of sets that are in $\mathcal{A}_c$ for all $c \in \mathcal{C}$) is a σ algebra.

Proof This follows almost directly from the definition. ■

Corollary 1.3.11 Given a collection of sets $\mathcal{C}$, there exists a smallest σ algebra $\mathcal{B}$ containing each set in $\mathcal{C}$. That is, there exists a sigma algebra $\mathcal{B}$ such that if $\mathcal{A}$ is any other sigma algebra containing each set in $\mathcal{C}$, then $\mathcal{B} \subset \mathcal{A}$.

Proof The proof follows in exactly the same way as the proof of the existence of a smallest algebra containing a given collection of sets. The set of all possible sigma algebras containing $\mathcal{S}$ exists by the power set axiom[2](applied twice). Take the intersection. This exists by De Morgan's laws. It is easy to check the hypotheses to see that the resulting set is a σ-algebra; if A is in all the σ -algebras, then so is A^c. If $(A_j)_{j=1}^{\infty}$ are in all the σ algebras, then so is $\cup_{j=1}^{\infty} A_j$. The resulting collection is a σ algebra and is contained in any other σ algebra containing each set in $\mathcal{C}$. ■.

Referring to corollary 1.3.11 we say again that $\mathcal{B}$ is **generated** by $\mathcal{C}$. In addition, we launch the notation

$$\mathcal{F} \subseteq \mathcal{G},$$

which says that any set in the sigma field $\mathcal{F}$ lies also in the sigma field $\mathcal{G}$.

Example 1.3.12 Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and

$$\mathcal{F} = \{\emptyset, \{\omega_2, \omega_3\}, \{\omega_1\}, \{\omega_1, \omega_2, \omega_3\}\}$$

is a sigma-field generated by the collection of sets $\{\{\omega_1\}\}$, or, generated by the set $\{\omega_1\}$,

■

Example 1.3.13 Let $A \subset \Omega$. Then

$$\mathcal{F} = \{\Omega, A, A^c, \emptyset\}$$

is the sigma-field generated by the collection of sets $\{\{A\}\}$ or, by the set A,

[2] The **power set axiom** is stated as follows: Given any set A, there is a set $\mathcal{P}(\mathcal{A})$ such that, given any set B, B is a member of $\mathcal{P}(\mathcal{A})$ if and only if B is a subset of A. Or, every set has a power set. Here one is stepping outside the realm of naïve set theory and considering axiomatic set theory with the *Zermelo-Fraenkel axioms.*

■

Definition 1.3.3 (The Borel Sigma Algebra) The *Borel σ algebra $\mathcal{B}$* over $\mathbf{R}$ is generated by intervals of the form (a,b).

■

Thereby the Borel σ algebra $\mathcal{B}$ contains all sets of the form $(-n,b)$, n is a positive integer, and

$$(-n,b) \subset (-(n+1),b)$$

and thus

$$(-\infty,b) = \lim_{n\to\infty} (-n,b)$$

and since all sets (a,n) are in $\mathcal{B}$,

$$(a,\infty) = \lim_{n\to\infty} (a,n)$$

is in $\mathcal{B}$. In addition

$$\{a\} = \lim_{n\to\infty} \left(a - \frac{1}{n}, a + \frac{1}{n}\right),$$

and we see that all singleton sets belong to $\mathcal{B}$. Furthermore,

$$(-\infty,a] = (a,\infty)^c,$$

and thus in $(-\infty,a] \in \mathcal{B}$, since there is closure under complements. Furthermore

$$(a,b] = (a,b) \cup \{b\} \quad \text{and} \quad [a,b) = (a,b) \cup \{a\}$$

are events in the Borel σ algebra $\mathcal{B}$, and

$$[a,b] = (a,b) \cup \{a\} \cup \{b\},$$

so that all closed intervals are in $\mathcal{B}$.

In addition $\mathcal{B}$ must contain all finite or countable unions and complements of intervals of any of the preceding forms. We may roughly say that $\mathcal{B}$ contains all subsets of the real line that can be obtained as an approximation of countable combinations of intervals.

It is a deep and difficult mathematical result that there are in fact subsets of $\mathbf{R}$ that are not in the Borel σ algebra. These 'unmeasurable' sets have no importance in engineering practice, as they are very hard to construct. Next we recapitulate some further basics about the Borel σ algebra.

Theorem 1.3.14 The Borel σ algebra $\mathcal{B}$ over $\mathbf{R}$ is generated by each and every of

1. open intervals of the form (a,b)
2. half-open intervals of the form $[a,b)$
3. half-open intervals of the form $(a,b]$
4. closed intervals of the form $[a,b]$
5. left intervals $(-\infty,b)$
6. right intervals (a,∞)

7. open sets of **R**
8. closed subsets of **R**

■

The proof is left to the diligent reader.

Definition 1.3.4 (Borel function) A function $f : \mathbf{R} \mapsto \mathbf{R}$ is called a Borel function, if for every set A in $\mathcal{B}$, the Borel σ algebra, we have that

$$f^{-1}(A) = \{x \in \mathbf{R} \mid f(x) \in A\}$$

belongs to the Borel σ algebra, i.e.,

$$f^{-1}(A) \in \mathcal{B}.$$

■

We call $f^{-1}(A)$ the *inverse image of* A.

Familiar examples of functions, like continuous functions, differentiable functions, sums of such functions and products of such functions, and limits of sequences of Borel functions are all Borel functions. It is difficult to construct a function that would not be a Borel function.

1.4 Probability Space

A *probability space* is given by a triple $(\Omega, \mathcal{F}, \mathbf{P})$, where Ω is a set of 'outcomes', $\mathcal{F}$ is a set of subsets of Ω, the set of possible *events* and $\mathbf{P} : \mathcal{F} \to [0, 1]$ is a function assigning probabilities to events. $\mathcal{F}$ is taken to to be a σ algebra.

Note that the word 'space' has many different usages in mathematics, the triumvirate above is a space in a different sense of the word than, say, when we talk about a Euclidean space or a Hilbert space, which are spaces with a geometric structure. A Euclidean space or a Hilbert space does, of course, serve as Ω of a probability space in many applications.

1.4.1 Probability Measures

Intuitive instances of measures are length on the real line, area in two dimensions, volume in three dimensions, when properly defined. The general definition of measure is

Definition 1.4.1 (Measure) A *measure* over a σ- algebra is a non negative set function $\mu : \mathcal{F} \to \mathbf{R}_+$ satisfying

1. $\mu(A) \geq 0$ for all $A \in \mathcal{F}$ and
2. if $A_i \in \mathcal{F}$ for all A_i in the collection $(A_i)_{i=1}^{\infty}$ of *pairwise disjoint* sets, then

$$\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i).$$

This is known as **countable additivity**.

■

If $\mu(\Omega) = 1$, then μ is said to be a **probability measure** and we use the notation **P** for the generic probability measure.

> The definition above is postulated for further mathematical developments of probability calculus. For real world applications of probability the main problem is the choice of the sample space Ω of events and the assignment of probability on the events.

We quote the following fundamental theorem of probability [81, ch. 1.5]. It tells that it is possible to construct a probability measure on a sigma algebra generated by an algebra by first giving the measure on the generating algebra.

Theorem 1.4.1 Let $\mathcal{A}$ be a set algebra and let $\sigma(\mathcal{A})$ be the (smallest) sigma algebra generated by $\mathcal{A}$. If $\mathbf{P}$ is a probability measure defined on $\mathcal{A}$, then there exists one and only one probability measure $\widetilde{\mathbf{P}}$ defined on $\sigma(\mathcal{A})$ such that if $A \in \mathcal{A}$, then $\widetilde{\mathbf{P}}(A) = \mathbf{P}(A)$.

■

We shall next find a few direct consequences of the axiomatic definition of a probability measure $\mathbf{P}$.

Theorem 1.4.2 For any probability measure $\mathbf{P}$ we have

$$\mathbf{P}(\emptyset) = 0.$$

Proof Consider $\Omega = \Omega \cup (\cup_{i=1}^{\infty} A_i)$, where $A_i = \emptyset$ for $i = 1, 2, \ldots,$. Then $A_i \cap A_j = \emptyset$ and $\Omega \cap A_j = \emptyset$, i.e., the sets in the union $\Omega \cup (\cup_{i=1}^{\infty} A_i)$ are disjoint. We set $a = \mathbf{P}(\emptyset)$. Then countable additivity yields

$$1 = \mathbf{P}(\Omega) = \mathbf{P}\left(\Omega \cup (\cup_{i=1}^{\infty} A_i)\right)$$

$$= \mathbf{P}(\Omega) + \sum_{i=1}^{\infty} \mathbf{P}(A_i) = 1 + \sum_{i=1}^{\infty} \mathbf{P}(A_i)$$

$$= 1 + a + a + a + \ldots,$$

which is possible if and only if $a = 0$. ■

Theorem 1.4.3 (Finite Additivity) Any countably additive probability measure is finitely additive, i.e., for all A_i in the collection $(A_i)_{i=1}^{n}$ of *pairwise disjoint* sets

$$\mathbf{P}(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbf{P}(A_i).$$

Proof Take $A_i = \emptyset$ for $i = n+1, n+2, \ldots,$. Then $A_i \cap \emptyset = \emptyset$ and $\cup_{i=1}^{n} A_i = \cup_{i=1}^{\infty} A_i$. Thus, by countable additivity,

$$\mathbf{P}(\cup_{i=1}^{n} A_i) = \mathbf{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbf{P}(A_i)$$

$$= \sum_{i=1}^{n} \mathbf{P}(A_i) + \mathbf{P}(\emptyset) + \mathbf{P}(\emptyset) + \ldots$$

$$= \sum_{i=1}^{n} \mathbf{P}(A_i)$$

by virtue of Theorem 1.4.2 above. ■

Theorem 1.4.4 For any $A \in \mathcal{F}$

$$\mathbf{P}(A^c) = 1 - \mathbf{P}(A).$$

Proof $\Omega = A \cup A^c$ and $A \cap A^c = \emptyset$. Then finite additivity of Theorem 1.4.3 gives

$$1 = \mathbf{P}(\Omega) = \mathbf{P}(A \cup A^c) = \mathbf{P}(A) + \mathbf{P}(A^c).$$

■

Theorem 1.4.5 (Monotonicity) For any $A \in \mathcal{F}$ and $B \in \mathcal{F}$ such that $A \subseteq B$, we have

$$\mathbf{P}(A) \leq \mathbf{P}(B).$$

Proof $B = A \cup (B \cap A^c)$. Then $A \cap (B \cap A^c) = \emptyset$, and finite additivity gives

$$\mathbf{P}(B) = \mathbf{P}(A \cup (B \cap A^c)) = \mathbf{P}(A) + \mathbf{P}(B \cap A^c) \geq \mathbf{P}(A),$$

as $\mathbf{P}$ is a non negative set function. ■

The identity in the proof above says also the following. If $A \in \mathcal{F}$ and $B \in \mathcal{F}$ and $A \subseteq B$, then we have

$$\mathbf{P}(B \setminus A) = \mathbf{P}(B) - \mathbf{P}(A).$$

In the same manner we can prove the next theorem.

Theorem 1.4.6 For any $A \in \mathcal{F}$ and $B \in \mathcal{F}$ we have

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B).$$

■

Example 1.4.7 (Probability measure on a countable outcome space) We consider the special case $\Omega = \{\omega = (\omega_i)_{i=1}^n \mid \omega_i \in \{0,1\}\}$. In words, the elementary outcomes are finite sequences of digital bits. Ω is countable. The sigma field $\mathcal{F}_o$ is generated by the collection of sets A_k (a.k.a. cylinders) of the form

$$A_k = \{\omega = (\omega_i)_{i=1}^k \mid \omega_1 = x_{l_1}, \omega_2 = x_{l_2}, \ldots, \omega_n = x_{l_k}\}$$

for any integer $k \leq n$ and arbitrary string of bits, $x_{l_1}x_{l_2}\ldots x_{l_k}$. We assign the weight $p(\omega) \geq 0$ to every elementary outcome ω and require that $\sum_\omega p(\omega) = 1$. Then the probability of any set A in $\mathcal{F}_o$ is defined by

$$\mathbf{P}(A) \overset{\text{def}}{=} \sum_{\omega \in A} p(\omega). \tag{1.3}$$

It can be shown (an exercise to this section) that $\mathbf{P}$ is a countably additive probability measure, and therefore $(\Omega, \mathcal{F}_o, \mathbf{P})$ is a probability space. The measure $\mathbf{P}$ can be extended to the σ-field of measurable subsets $\mathcal{F}$ of the uncountable $\{(\omega_i)_{i=1}^\infty \mid \omega_i \in \{0,1\}\}$.

■

1.4.2 Continuity from below and Continuity from above

A probability measure has furthermore the following properties:

1. **subadditivity** If $A \subset \cup_i A_i$ then $\mathbf{P}(A) \leq \sum_i \mathbf{P}(A_i)$

2. **continuity from below** If $A_1 \subset A_2 \subset \ldots$ and $A = \cup_i A_i$, then $\mathbf{P}(A) = \lim_{i \to +\infty} \mathbf{P}(A_i)$.

3. **continuity from above** If $A_1 \supset A_2 \supset \ldots$ and $A = \cap_i A_i$ then $\mathbf{P}(A) = \lim_{i\to+\infty} \mathbf{P}(A_i)$.

The proofs of the continuity properties are given below. One needs to recall (1.1) and (1.2).

Theorem 1.4.8 If $B_n \uparrow \cup_{k=1}^{\infty} B_k$, then $\mathbf{P}(B) = \lim_{n\to\infty} \mathbf{P}\left(B_n\right)$.

Proof: We use the notation for 'set difference',

$$A \setminus B = A \cap B^c,$$

and we can write $\cup_{k=1}^{\infty} B_k = \cup_{k=2}^{\infty}(B_k \setminus B_{k-1}) \cup B_1$, since B_k are increasing.

$$\mathbf{P}(B) = \mathbf{P}\left(\cup_{k=1}^{\infty} B_k\right) = \mathbf{P}\left(\cup_{k=2}^{\infty}(B_k \setminus B_{k-1}) \cup B_1\right)$$

But the sets in the decomposition are seen to be pairwise disjoint, and hence the countable additivity yields

$$\mathbf{P}\left(\cup_{k=2}^{\infty}(B_k \setminus B_{k-1}) \cup B_1\right) = \sum_{k=2}^{\infty} \mathbf{P}\left(B_k \setminus B_{k-1}\right) + \mathbf{P}\left(B_1\right)$$

$$= \lim_{n\to\infty} \sum_{k=2}^{n} \mathbf{P}\left(B_k \setminus B_{k-1}\right) + \mathbf{P}\left(B_1\right)$$

Now we observe that since $B_{k-1} \subset B_k$, we have

$$\mathbf{P}\left(B_k \setminus B_{k-1}\right) = \mathbf{P}\left(B_k\right) - P\left(B_{k-1}\right).$$

Therefore, we get a telescoping series

$$\sum_{k=2}^{n} \mathbf{P}\left(B_k \setminus B_{k-1}\right) + \mathbf{P}\left(B_1\right) = \mathbf{P}\left(B_n\right) - \mathbf{P}\left(B_{n-1}\right) + \mathbf{P}\left(B_{n-1}\right) - \mathbf{P}\left(B_{n-2}\right) + \ldots +$$

$$+\mathbf{P}\left(B_2\right) - \mathbf{P}\left(B_1\right) + \mathbf{P}\left(B_1\right) = \mathbf{P}\left(B_n\right).$$

In other words we have shown that

$$\mathbf{P}(B) = \lim_{n\to\infty} \mathbf{P}\left(B_n\right).$$

■

Theorem 1.4.9 If $B_n \downarrow \cap_{k=1}^{\infty} B_k$, then $\mathbf{P}\left(\cap_{k=1}^{\infty} B_k\right) = \lim_{n\to\infty} \mathbf{P}\left(B_n\right)$.

Proof: We use theorem 1.4.4 in the preceding

$$\mathbf{P}\left(\cap_{k=1}^{\infty} B_k\right) = 1 - \mathbf{P}\left(\left(\cap_{k=1}^{\infty} B_k\right)^c\right). \tag{1.4}$$

When we apply one of De Morgan's rules we get

$$\left(\cap_{k=1}^{\infty} B_k\right)^c = \cup_{k=1}^{\infty} B_k^c.$$

Now we observe that if $B_k \supset B_{k+1}$, then $B_k^c \subset B_{k+1}^c$, i.e., the complement events of a decreasing sequence of events are an increasing sequence of events. Thus the theorem 1.4.8 above implies

$$\mathbf{P}\left(\cup_{k=1}^{\infty} B_k^c\right) = \lim_{n\to\infty} \mathbf{P}\left(B_n^c\right).$$

By one of the De Morgan's rules we have that

$$\mathbf{P}\left((\cap_{k=1}^{\infty} B_k)^c\right) = \mathbf{P}\left(\cup_{k=1}^{\infty} B_k^c\right)$$

$$= \lim_{n\to\infty} \mathbf{P}\left(B_n^c\right).$$

This we shall insert in (1.4) and get

$$\mathbf{P}\left(\cap_{k=1}^{\infty} B_k\right) = 1 - \lim_{n\to\infty} \mathbf{P}\left(B_n^c\right)$$

$$= 1 - \lim_{n\to\infty} \left(1 - \mathbf{P}\left(B_n\right)\right)$$

$$= 1 - 1 + \lim_{n\to\infty} \mathbf{P}\left(B_n\right) = \lim_{n\to\infty} \mathbf{P}\left(B_n\right).$$

This completes the proof. ■

1.4.3 Why Do We Need Sigma-Fields?

This subsection is based on [43, 44, 50] and its contents will NOT be actively examined. There are several statements below and in the next subsection that should be verified (as exercises) but we do not expect the student to do this piece of work, whereas these can be recommended for the seriously interested.

In broad terms, if Ω is finite or countably infinite, we can consider all subsets of Ω to be the family of events. When Ω is uncountably infinite, as in the case of $\Omega =$ the real line, one cannot build a useful theory without confining the allowable subsets to which one will assign probability. Roughly said, all probabilities are obtained by integrating over sets, and some sets are too nasty to be integrated over. It is, however, difficult to show but for such Ω there does not exist a reasonable and consistent means of assigning probabilities to all subsets without contradiction or without violating desirable properties. **The student should be aware of the problem so that the need for specifying $\mathcal{F}$ is understood**.

Let us consider $\Omega = \{\omega \mid 0 \leq \omega \leq 1\}$, i.e., the outcome space is the unit interval in the real line $\mathbf{R}$. Suppose we want the set of events $\mathcal{F}$ to include intervals $[a, b] \subseteq [0, 1]$ and the probability of any interval to be given by the length of the interval:

$$\mathbf{P}\left([a, b]\right) = b - a. \tag{1.5}$$

If we take $a = b$, we need to have the singleton sets $\{a\}$ in $\mathcal{F}$, and their probability is zero. If $\mathcal{F}$ is to be a sigma-field, then the open interval $(a, b) = \cup_{i=1}^{\infty}[a + \frac{1}{i}, b - \frac{1}{i}]$ must be in $\mathcal{F}$, and the probability of such an open interval is by continuity from below (see condition 2. in section 1.4.2 above)

$$\mathbf{P}\left((a, b)\right) = \lim_{i\to\infty} \mathbf{P}\left(\left[a + \frac{1}{i}, b - \frac{1}{i}\right]\right) = \lim_{i\to\infty}\left(b - a - \frac{2}{i}\right) = b - a.$$

Any open subset of Ω is the union of finite or countably infinite set of open intervals, so that $\mathcal{F}$ should contain all open and closed subsets of Ω. Hence $\mathcal{F}$ must contain any set that is the intersection of countably many open sets, and so on.

The specification (1.5) of probability must therefore be extended from all intervals to all of $\mathcal{F}$. We cannot figure out *a priori* how large $\mathcal{F}$ will be. One might think that $\mathcal{F}$ *should be the set of all subsets of* Ω. However, this does not work:

> Suppose that we wish to define a measure μ to called *length*, length(A), for all subsets A of R such that
>
> $$\text{length}\left([a, b]\right) = b - a \quad a < b,$$

and such that the measure satisfies the additional condition of *translation invariance*

$$\text{length}\,(A+y) = \text{length}\,(A)$$

where $A+y = \{x+y \mid x \in A\}$.

This is now shown to lead to a contradiction. Take $Q =$ the set of rational numbers, i.e., $Q = \{p/q \mid p \in Z, q \in Z\}$. For any real $x \in \mathbf{R}$ let $Q_x = Q + x$. One can show that for any $x \in \mathbf{R}$ and $y \in \mathbf{R}$ either $Q_x = Q_y$ or Q_x and Q_y are disjoint. One can also show that $Q_x \cap [0,1] \neq \emptyset$ for all $x \in \mathbf{R}$, or in plain words, each Q_x contains at least one element from $[0,1]$.

Let V be a set obtained by choosing exactly one element from the interval $[0,1]$ from each Q_x. (V is well defined, if we accept the *Axiom of Choice*[3].)

Thus V is a subset of $[0,1]$. Suppose $q_1, q_2, \ldots$ is an enumeration of all the rational numbers in the interval $[-1,1]$, with no number appearing twice in the list. Let for $i \geq 1$

$$V_i = V + q_i.$$

It can be verified that all the sets V_i are disjoint and

$$[0,1] \subset \cup_{i=1}^{\infty} V_i \subset [-1,2].$$

Since V_i are translations of V, they should have the same length as V. If the length of V is defined to be zero, so $[0,1]$ would also have length zero by monotonicity. If the length of V were strictly positive, then the length of $\cup_{i=1}^{\infty} V_i$ would by countable additivity be infinite, and hence the interval $[-1,2]$ would have infinite length. In either way we have a contradiction.

The difficulty will be resolved by taking $\mathcal{F}$ to be the Borel sigma algebra, c.f. definition 1.3.3 above, and by construction of the Lebesgue measure.

For the construction of the **Lebesgue measure** we refer to [36, chapter 1.] or [91, chapter 11.]. We outline a rudiment of this theory. Lebesgue measure over the real line is defined as follows: The length of an interval $[a,b]$, (a,b), $(a,b]$ or $[a,b)$ is given by $b-a$ (c.f. the measure length above). The *outer measure* of a set A is given as the infimum over open intervals $(I_n)_{n=1}^{\infty}$

$$m^*(A) = \inf_{(I_n)_{n=1}^{\infty}: A \subset \cup_n I_n} \sum_{n=1}^{\infty} |I_n|,$$

where $|I_n|$ denotes the length of the interval I_n. A set B is said to be *measurable*, with measure $\lambda(B) = m^*(B)$ if for any set $A \subset \mathbf{R}$ it holds that

$$m^*(A) = m^*(A \cap B) + m^*(A \cap B^c).$$

The *Heine Borel lemma* states that *every* covering by open sets has a finite subcovering.

One then uses the *Carathéodory Extension Theorem* to show that Lebesgue measure is well defined over the *Borel σ algebra* .

Finally, why not be content with probability measures only on set algebras ? The answer is that a good theory of probability needs limits of random variables and infinite sums of random variables, which require events outside a set algebra.

[3] http://en.wikipedia.org/wiki/Axiom_of_choice

1.4.4 P - Negligible Events and P -Almost Sure Properties

An event $A \in \mathcal{F}$ such that $\mathbf{P}(A) = 0$ is called a $\mathbf{P}$ - negligible event, or just negligible event, if there is no possibility of confusion.

A property that holds everywhere except possible for ω in a $\mathbf{P}$ -negligible set is said to hold $\mathbf{P}$ -almost surely or we say that the property holds **almost surely**, and abridge this often to a.s.. Examples of such properties will be in the sequel encountered under the guise of '$X \geq 0$ a.s.' or 'convergence almost surely', or 'continuity of sample paths almost surely', to mention the main ones.

Example 1.4.10 Let $\mathcal{F}$ to be the Borel sigma algebra, c.f. definition 1.3.3, restricted to $[0, 1]$. Let $A =]a, b]$ with $0 \leq a \leq b \leq 1$ and $\mathbf{P}(A) = b - a$. By definition 1.3.3 we know that singleton sets $\{a\}$ belong to $\mathcal{F}$ and thus $\mathbf{P}(\{a\}) = 0$. Hence, e.g., the set of rational numbers in $[0, 1]$, i.e., $\frac{p}{q}$ with p and q positive integers $p \leq q$, is a countable disjoint union of measurable sets, is by countable additivity $\mathbf{P}$ - negligible.

■

1.5 Random Variables and Distribution Functions

1.5.1 Randomness?

As will become evident by scrutiny of this section, random variables of probability calculus are functions with certain properties, and have as such nothing to do with randomness, regardless of how randomness is defined, and regardless of whether such a definition possible at all. Randomness has been aptly described as a negative property [53, p.20], as it is not possible to definitely prove its presence, but it is possible to prove the absence of it.

One makes customarily the interpretation of a random variable as a real valued measurement of the outcomes of a random phenomenon that is governed by a physical probability. One criticism of the notion of physical probability decried as 'mind projection fallacy' has been voiced by Ed Jaynes [65, p.500]:

> ...(statistics) has never produced any definition of the term 'random variable' that could actually be used in practice to decide whether some specific quantity, such as the number of beans in a can, is or is not 'random'.

Random variables and later random processes are in a very useful manner seen as mathematical models of physical **noise**. As examples an engineer might quote thermal noise (a.k.a. Nyquist-Johnson noise, produced by the thermal motion of electrons inside an electrical conductor), quantum noise and shot noise, see [11, 33, 71, **?**]. Does this provide grounds for claiming a physical countably additive probability measure? The foundational question of how to define randomness is, certainly, not resolved by this manœuver, at any rate not, if one in a circular manner describes the physical noise as the result of many random events happening at the microscopic level.

Remark 1.5.1 Physical noise, in particular measurement error (mätfel), is described as follows in [52, p.13]:

> ... felens storlek och tecken (kan) inte individuellt påvisas någon lag och de kan alltså inte i förväg beräknas eller individuellt korrigeras. ... Vanligen antas en viss relation föreligga emellan de oregelbundna felens storlek och deras frekvens.

As an interpretation in English, the quoted Swedish author describes random measurement errors as quantities, whose magnitude and sign do not follow any known law and cannot be compensated for in advance as individual items. One assumes, however, that there is a statistical relation or regularity between the magnitudes and their frequencies.

■

One foundational approach to discussing randomness is due to G. Chaitin [21]. Chaitin argues that randomness has to do with complexity. Or, a random object cannot be compressed at all: since in randomness there is no structure or pattern ('lag', a known law in the Swedish quote above), you cannot give a more concise or less complex description (by a computer program or a rule) other than the object itself. For a statistical modelling theory with complexity as platform we refer to the lectures by J. Rissanen [86].

In the sequel we shall not pursue the foundational topics or the related critical discourses any further, but continue by presenting probabilistic tools for modelling of noise and for **modelling by means of noise**.

1.5.2 Random Variables and Sigma Fields Generated by Random Variables

Definition 1.5.1 *[***Random Variable**] A *real valued random variable* is a *real valued function* $X : \Omega \to \mathbf{R}$ such that for every set $A \in \mathcal{B}$, the Borel σ algebra over $\mathbf{R}$,

$$X^{-1}(A) = \{\omega : X(\omega) \in A\} \in \mathcal{F}. \tag{1.6}$$

■

The condition in (1.6) means in words that the pre-image of any $A \in \mathcal{B}$ is in $\mathcal{F}$, and X is called *measurable* , *or a measurable function* from Ω to $\mathbf{R}$. We can also write

$$X : (\Omega, \mathcal{F}) \mapsto (\mathbf{R}, \mathcal{B}).$$

Example 1.5.1 Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Take $F \in \mathcal{F}$, and introduce the **indicator function** of F, to be denoted by χ_F, as the real valued function defined by

$$\chi_F(\omega) = \begin{cases} 1 & \text{if } \omega \in F \\ 0 & \text{if } \omega \notin F. \end{cases} \tag{1.7}$$

We show now that χ_F is a random variable. We take any $A \in \mathcal{B}$ and find that

$$\chi_F^{-1}(A) = \{\omega : \chi_F(\omega) \in A\} = \begin{cases} \emptyset & \text{if } 0 \notin A,\ 1 \notin A \\ F & \text{if } 0 \notin A,\ 1 \in A \\ F^c & \text{if } 0 \in A,\ 1 \notin A \\ \Omega & \text{if } 0 \in A,\ 1 \in A. \end{cases}$$

Since $\mathcal{F}$ is a sigma field, we see that $\chi_F^{-1}(A) \in \mathcal{F}$.

■

For the next result one needs to recall the definition of a Borel function in 1.3.4.

Theorem 1.5.2 Let $f : \mathbf{R} \mapsto \mathbf{R}$ be a Borel function, and X be a random variable. Then Y defined by

$$Y = f(X)$$

is a random variable.

Proof Let A be a Borel set, i.e., $A \in \mathcal{B}$. We consider

$$Y^{-1}(A) = \{\omega \in \Omega \mid Y(\omega) \in A\}.$$

By construction we have $Y(\omega) = f(X(\omega))$, and thus

$$Y^{-1}(A) = \{\omega \in \Omega \mid f(X(\omega)) \in A\} = \{\omega \in \Omega \mid X(\omega) \in f^{-1}(A)\},$$

where the inverse image is $f^{-1}(A) = \{x \in \mathbf{R} \mid f(x) \in A\}$. Since f is a Borel function, we have by definition that $f^{-1}(A) \in \mathcal{B}$, since $A \in \mathcal{B}$. But then

$$\{\omega \in \Omega \mid X(\omega) \in f^{-1}(A)\} \in \mathcal{F},$$

since X is a random variable. But thereby we have established that $Y^{-1}(A) \in \mathcal{F}$ for any A in $\mathcal{B}$, which by definition means that Y is a random variable. ■

> By this theorem we have, amongst other things, provided a slick mathematical explanation of one basic tenet of statistics, namely that an estimator of a parameter in a probability distribution is a random variable. Of course, for students of a first course in probability and statistics the understanding of this fact may require much more effort and pedagogic ingenuity[4].

Definition 1.5.2 **A sigma field generated by a real valued random variable** X, denoted by $\mathcal{F}_X$ and/or $\sigma(X)$, consists of all events of the form $\{\omega : X(\omega) \in A\} \in \mathcal{F}$, $A \in \mathcal{B}$, where $\mathcal{B}$ is the Borel σ algebra over $\mathbf{R}$.

■

Example 1.5.3 In example 1.5.1 it was verified that χ_F is a random variable for any $F \in \mathcal{F}$. Then it follows by the same example and the definition above that the sigma-field generated by χ_F is

$$\mathcal{F}_{\chi_F} = \{\Omega, F, F^c, \emptyset\}.$$

In view of example 1.3.13 $\mathcal{F}_{\chi_F}$ is the sigma-field generated by the set F, as seems natural.

■

Definition 1.5.3 **A sigma field generated by a family** $\{X_i \mid i \in I\}$ **of real valued random variables** X_i, denoted by $\mathcal{F}_{X_i, i\in I}$, is defined to be the smallest σ algebra containing all events of the form $\{\omega : X_i(\omega) \in A\} \in \mathcal{F}$, $A \in \mathcal{B}$, where $\mathcal{B}$ is the Borel σ algebra over $\mathbf{R}$ and $i \in I$.

■

If it holds for all events in A in a sigma-field $\mathcal{H}$ that $A \in \mathcal{F}$, then we say that $\mathcal{H}$ is a subsigma-field of $\mathcal{F}$ and write

$$\mathcal{H} \subseteq \mathcal{F}.$$

Example 1.5.4 Let $Y = f(X)$, where X is a random variable and f is a Borel function. Then

$$\mathcal{F}_Y \subseteq \mathcal{F}_X.$$

We shall now establish this inclusion. The sigma field generated by a real valued random variable Y, or $\mathcal{F}_Y$, consists of all events of the form $\{\omega : Y(\omega) \in A\} \in \mathcal{F}$, $A \in \mathcal{B}$. Now

$$\{\omega : Y(\omega) \in A\} = \{\omega : f(X(\omega)) \in A\} = \{\omega : X(\omega) \in f^{-1}(A)\}.$$

[4] c.f., K. Vännman: How to Convince a Student that an Estimator is a Random Variable. *Teaching Statistics*, vol. 5, n:o 2, pp. 49−54, 1983.

Since $f^{-1}(A)$ is a Borel set, we have by definition of $\mathcal{F}_X$ that $\{\omega : X(\omega) \in f^{-1}(A)\} \in \mathcal{F}_X$. Therefore we have shown that every event in $\mathcal{F}_Y$ is also in $\mathcal{F}_X$, and this finishes the proof of the inclusion $\mathcal{F}_Y \subseteq \mathcal{F}_X$. ■

The result is natural, as events involving $Y$are in fact events determined by X. If $f(x)$ is invertible in whole of its domain of definition, then clearly $\mathcal{F}_Y = \mathcal{F}_X$.

Theorem 1.5.5 (Doob-Dynkin) Let X be a real valued random variable and let Y be another real valued random variable such that Y is $\sigma(X)$ -measurable, or,

$$Y^{-1}(A) = \{\omega : Y(\omega) \in A\} \in \sigma(X)$$

for all A in the Borel σ algebra over $\mathbf{R}$. Then there is a (Borel) function $H(x)$ (definition 1.3.4) such that

$$Y = H(X).$$

Proof is omitted, and is not trivial. The interested student can find one proof in [63, thm 23.2]. ■

1.5.3 Distribution Functions

The probability distribution of a random variable X may be described by its *distribution function* $F(x) = \mathbf{P}(\{X \leq x\})$. This is a quick and convenient shorthand for the complete expression in the following sense

$$F(x) = \mathbf{P}(\{X \leq x\}) \equiv \mathbf{P}(\{\omega \in \Omega \mid X(\omega) \in (-\infty, x]\}).$$

Note that our preceding efforts pay here a dividend: $(-\infty, x]$ is a Borel event, and as X is a random variable, $\{\omega \in \Omega \mid X(\omega) \in (-\infty, x]\}$ is an event in $\mathcal{F}$ and therefore we may rest assured that $\mathbf{P}(\{\omega \in \Omega \mid X(\omega) \in (-\infty, x]\})$ is defined. In the chapters to follow it will contribute to clarity of thought to indicate the random variable connected to the distribution function, so we shall be writing there

$$F_X(x) = \mathbf{P}(\{X \leq x\}). \tag{1.8}$$

Remark 1.5.2 In statistical physics, see, e.g., [17], a distribution function pertains[5] often to a different concept. For example, the distribution function of the velocities v of molecules in a gas is the fraction, $f(v)dv$, of molecules with velocities between v and $v + dv$, and is shown in [17, p. 48] or [18] to be

$$f(v)dv \propto e^{-mv^2/k_BT}dv, \tag{1.9}$$

where m is the mass, k_B is the Boltzmann constant, and T is temperature. In probability theory's terms $f(v)$ is is obviously the probability density of the velocity. The density above will be re-derived in section 11.3 using an explicit and well defined random process, known as the **Ornstein-Uhlenbeck process**.

■

Theorem 1.5.6 Any distribution function has the following properties:

1. F is non decreasing,

2. $\lim_{x\to+\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$,

[5] [17] is the textbook in SI1161 statistisk fysik för F3 (statistical physics for students of CTFYS at KTH).

3. F is right continuous; $\lim_{y\downarrow x} F(y) = F(x)$

4. If $F(x-) = \lim_{y\uparrow x} F(y)$, then $F(x-) = \Pr\{X < x\}$ and

5. $\mathbf{P}\{X = x\} = F(x) - F(x-)$.

Proof Clear ■

Theorem 1.5.7 If F satisfies 1., 2. and 3. above, then it is the distribution function of a random variable.

■

Proof Consider $\Omega = (0,1)$ and $\mathbf{P}$ the uniform distribution, which means that $\mathbf{P}((a,b]) = b-a$, for $0 \leq a < b \leq 1$. Set

$$X(\omega) = \sup\{y : F(y) < \omega\}.$$

Firstly, notice that if $\omega \leq F(x)$, then $X(\omega) \leq x$, since $x \notin \{y : F(y) < \omega\}$. Next: Suppose $\omega > F(x)$. Since F is right continuous, there is an $\epsilon > 0$ such that $F(x+\epsilon) < \omega$. Therefore, $X(\omega) \geq x + \epsilon > x$. ■

Next we define $\liminf_{n\to+\infty} X_n$ and $\limsup_{n\to+\infty} X_n$ For the definitions of $\liminf_{n\to+\infty} x_n$ and $\limsup_{n\to+\infty} x_n$ for sequences of real numbers $(x_n)_{n=1}^{\infty}$ we refer to Appendices 1.9 and 1.10. Here

$$\liminf_{n\to+\infty} X_n = \sup_n \left(\inf_{m\geq n} X_m \right) \tag{1.10}$$

and

$$\limsup_{n\to+\infty} X_n = \inf_n \left(\sup_{m\geq n} X_m \right). \tag{1.11}$$

Theorem 1.5.8 If $X_1, X_2, \ldots$ are random variables, then so are

$$\inf_n X_n,\ \sup_n X_n,\ \limsup_n X_n \qquad \text{and} \qquad \liminf_n X_n.$$

Proof Provided $\mathcal{F}$ is a σ algebra, it follows that $\{\inf_n X_n < a\} = \cup_{n=1}^{\infty}\{X_n < a\} \in \mathcal{F}$. Now, the sets $(-\infty, a)$ are in the Borel sigma algebra, proving that $\inf_n X_n$ is measurable. Similarly, $\{\sup_n X_n > a\} = \cup_{n=1}^{\infty}\{X_n > a\} \in \mathcal{F}$. For the last two statements, the conclusion is clear in view of (1.10) and (1.11) and by what was just found. ■

1.6 Independence of Random Variables and Sigma Fields, I.I.D. r.v.'s

We know from any first course in probability and statistics that two events A and B are called *independent* if

$$\mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B).$$

We shall now see that we can exploit this to define independence of random variables and of sigma fields.

Definition 1.6.1 Assume that we have a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and random variables X and Y on it.

- Two sigma fields $\mathcal{H} \subseteq \mathcal{F}$ and $\mathcal{G} \subseteq \mathcal{F}$ are **independent** if any two events $A \in \mathcal{H}$ and $B \in \mathcal{G}$ are independent, i.e.,

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

- Two **random variables** X **and** Y **are independent**, if the sigma-algebras generated by them, $\mathcal{F}_X$ and $\mathcal{F}_Y$, respectively, are independent.

■

It follows that two random variables X and Y are independent, if and only if

$$\mathbf{P}\left(X \in A, Y \in B\right) = \mathbf{P}\left(X \in A\right) \cdot \mathbf{P}\left(Y \in B\right)$$

for all Borel sets A and B. In particular, if we take $A = (-\infty, x]$ and $B = (-\infty, y]$, we obtain for all $x \in \mathbf{R}$, $y \in \mathbf{R}$

$$F_{X,Y}(x,y) = \mathbf{P}\left(X \in (-\infty, x], Y \in (-\infty, y]\right) = F_X(x) \cdot F_Y(y), \tag{1.12}$$

which is the familiar definition of independence for X and Y, see, e.g., [15], in terms of distribution functions. ■

Theorem 1.6.1 Let X and Y be independent random variables and f and g be two Borel functions. Then $f(X)$ and $g(Y)$ are independent.

Proof Set $U = f(X)$ and $V = g(Y)$. These are random variables by theorem 1.5.2. Then

$$\mathcal{F}_U \subseteq \mathcal{F}_X, \quad \mathcal{F}_V \subseteq \mathcal{F}_Y,$$

as shown in example 1.5.4. Thus, if we take any $A \in \mathcal{F}_U$ and any $B \in \mathcal{F}_V$, it holds that $A \in \mathcal{F}_X$ and $B \in \mathcal{F}_Y$. But $\mathcal{F}_X$ and $\mathcal{F}_Y$ are independent sigma fields, since X and Y are independent. Therefore it holds for every set $A \in \mathcal{F}_U$ and every $B \in \mathcal{F}_V$, that $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(A)$, and therefore $\mathcal{F}_U$ and $\mathcal{F}_V$ are independent, and this means that $U = f(X)$ and $V = g(Y)$ are independent, as was asserted. ■

If we have two independent random variables X and Y that have the same distribution (i.e., $F_X(x) = F_Y(x)$ for all x), we say that X and Y are **independent, identically distributed** r.v.'s and abridge this with **I.I.D.**. The same terminology can be extended to state $X_1, \ldots, X_n$ as being **I.I.D.** r.v.'s. Sequences of I.I.D. r.v.'s will be a main theme in the sequel.

1.7 The Borel-Cantelli Lemmas

Borel-Cantelli lemmas are indispensable, for example, for proving the law of large numbers in the strong form, section 6.7.4 below, and for proving the almost sure continuity of sample paths of the Wiener process, see section 10.4.1.

We consider a sequence events $A_1, A_2, A_3, \ldots$ and are interested in the question of whether infinitely many events occur or if possibly only a finite number of them occur. We set

$$F_n = \bigcup_{k=n}^{\infty} A_k \text{ and } G_n = \bigcap_{k=n}^{\infty} A_k. \tag{1.13}$$

If G_n in (1.13) occurs, this means that all A_k for $k \geq n$ occur. If there is some such n, this means in other words that from this n on all A_k occur for $k \geq n$. With

$$H = \bigcup_{n=1}^{\infty} G_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$$

we have that if H occurs, then there is an n such that all A_k with $k \geq n$ occur. Sometimes we denote H with $\liminf A_k$.

The fact that F_n occurs implies that there is some A_k for $k \geq n$ which occurs. If F_n in (1.13) occurs for all n this implies that infinitely many of the A_k:s occur. We form therefore

$$E = \bigcap_{n=1}^{\infty} F_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k. \tag{1.14}$$

If E occurs, then infinitely many of A_k:s occur. Sometimes we write this as $E = \{A_n \text{ i.o.}\}$ where i.o. is to be read as "infinitely often", i.e., infinitely many times. E is sometimes denoted with $\limsup A_k$, c.f. Appendix 1.10.

Lemma 1.7.1 Borel-Cantelli lemma If $\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$ then it holds that $P(E) = P(A_n \text{ i.o}) = 0$, i.e., that with probability 1 only finitely many A_n occur.

Proof One notes that F_n is a decreasing set of events. This is simply so because

$$F_n = \bigcup_{k=n}^{\infty} A_k = A_n \bigcup \left(\bigcup_{k=n+1}^{\infty} A_k \right) = A_n \bigcup F_{n+1}$$

and thus

$$F_n \supset F_{n+1}.$$

Thus the theorem 1.4.9 above gives

$$\mathbf{P}(E) = \mathbf{P}(\bigcap_{n=1}^{\infty} F_n) = \lim_{n\to\infty} \mathbf{P}(F_n) = \lim_{n\to\infty} \mathbf{P}(\bigcup_{k=n}^{\infty} A_k).$$

We have, however, by subadditivity that

$$\mathbf{P}(\bigcup_{k=n}^{\infty} A_k) \leq \sum_{k=n}^{\infty} \mathbf{P}(A_k)$$

and this sum $\to 0$ as $n \to \infty$, if the sum $\sum_1^{\infty} \mathbf{P}(A_k)$ converges. Thus we have shown the proposition, as claimed. ■

One can observe that no form of independence is required, but the proposition holds in general, i.e., for any sequence of events.

A counterpart to the Borel-Cantelli lemma is obtained, if we assume that the events $A_1, A_2, \ldots$ are independent.

Lemma 1.7.2 Converse Borel-Cantelli lemma If $A_1, A_2, \ldots$ are independent and

$$\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty,$$

then it holds that $\mathbf{P}(E) = \mathbf{P}(A_n \text{ i.o}) = 1$, i.e., it holds with probability 1 that infinitely many A_n occur.

Proof We have by independence and probability of the complement

$$\mathbf{P}(\bigcap_{k=n}^{\infty} A_k^c) = \prod_{k=n}^{\infty} \mathbf{P}(A_k^c) = \prod_{k=n}^{\infty} (1 - \mathbf{P}(A_k)).$$

Since $1 - x \leq e^{-x}$ we get $1 - \mathbf{P}(A_k) \leq e^{-\mathbf{P}(A_k)}$ and

$$\mathbf{P}(\bigcap_{k=n}^{\infty} A_k^c) \leq \exp(-\sum_{k=n}^{\infty} \mathbf{P}(A_k)).$$

If now $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$, then the sum in the exponent diverges and we obtain

$$\mathbf{P}(\bigcap_{k=n}^{\infty} A_k^c) = 0.$$

Thus it holds also that

$$\mathbf{P}(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c) = 0,$$

which implies by De Morgan's rules that

$$\mathbf{P}(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k) = 1 - \mathbf{P}(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c) = 1 - 0 = 1$$

i.e., that infinitely many A_k:n occur with probability 1. ■

1.8 Expected Value of a Random Variable

1.8.1 A First Definition and Some Developments

Let X be a **simple random variable**. This is nothing but a special case of what will in the sequel be called a discrete random variable. In detail, we think of a set of real numbers, $\{x_1, \ldots, x_m\}$, such that X takes its values in $\{x_1, \ldots, x_m\}$. The expected value $E[X]$ (a.k.a mean or expectation) of a simple random variable is defined to be

$$E[X] \overset{\text{def}}{=} \sum_{i=1}^{m} x_i \mathbf{P}(X = x_i). \tag{1.15}$$

Again we write $\mathbf{P}(X = x_i)$, when we mean

$$\mathbf{P}(X = x_i) \equiv \mathbf{P}(\{\omega \in \Omega \mid X(\omega) = x_i\}).$$

The numbers $\mathbf{P}(X = x_i)$, $i = 1, \ldots, m$ will be later called a **probability mass function**, **p.m.f.**

$$p_X(x_i) = \mathbf{P}(X = x_i).$$

The definition in (1.15) clearly depends only on the probality measure $\mathbf{P}$ for given X.
Now, we want to interpret in $E[X]$ in (1.15) as an integral of X and use this inspirational recipe for non-simple X. For this we must develop a more general or powerful concept of integration, than the one incorporated in the Rieman integral treated in basic integral and differential calculus.

Here is an outline. We consider first an arbitrary nonnegative random variable $X \geq 0$. Then we can find (see below) a infinite sequence of simple random variables $X_1, X_2, \ldots$, such that

- for all $\omega \in \Omega$
$$X_1(\omega) \leq X_2(\omega) \leq \ldots$$
- and for all $\omega \in \Omega$
$$X_n(\omega) \uparrow X(\omega),$$
as $n \to \infty$.

Then $E[X_n]$ is defined for each n and is non-decreasing, and has a limit $E[X_n] \uparrow C \in [0,\infty]$, as $n \to \infty$. The limit C defines $E[X]$. Thus

$$E[X] \stackrel{\text{def}}{=} \lim_{n\to\infty} E[X_n]. \tag{1.16}$$

This is well defined, and it can happen that $E[X] = +\infty$. Let us take a look at the details of this procedure. The discussion of these details in the rest of this section can be skipped as the issues inolved will NOT be actively examined, but are recommended for the specially interested.

For an arbitrary nonnegative random variable $X \geq 0$ we define the simple random variable X_n, $n \geq 1$, as (an electrical engineer might think of this as a digitalized signal)

$$X_n(\omega) = \begin{cases} \frac{k}{2^n} & \text{if } \frac{k}{2^n} \leq X(\omega) \leq \frac{k+1}{2^n},\ k = 0,1,2,\ldots,n2^n - 1 \\ n & \text{else.} \end{cases}$$

This means that we partition for each n the range of X (not its domain !), $\mathbf{R}_+ \cup 0$, the nonnegative real line, so that $[0,n[$ is partitioned into $n2^n$ disjoint intervals of the form

$$E_{n,k} = \left[\frac{k}{2^n}, \frac{k+1}{2^n}\right],$$

and the rest of the range $\mathbf{R}_+ \cup 0$ is in $E_n = [n,\infty]$. Then we see that

$$\mid X_n(\omega) - X(\omega) \mid \leq \frac{1}{2^n} \quad \text{if } X(\omega) < n \tag{1.17}$$

and

$$X_n(\omega) = n, \text{ if } X(\omega) \geq n. \tag{1.18}$$

When we next go over to $n+1$, $[0,n+1[$ is partitioned into intervals of the form $E_{n+1,k} = \left[\frac{k}{2^{n+1}}, \frac{k+1}{2^{n+1}}\right]$. This is smart, because each of the previous intervals is halved, i.e.,

$$E_{n,k} = E_{n+1,2k} \cup E_{n+1,2k+1}. \tag{1.19}$$

But then it is clear that $X_n \leq X_{n+1}$. We show this for each ω. First, if $X_n(\omega) = \frac{k}{2^n}$, then by (1.19) either $X_{n+1}(\omega) = \frac{2k}{2^{n+1}} = \frac{k}{2^n}$ or $X_{n+1}(\omega) = \frac{2k+1}{2^{n+1}} > \frac{2k}{2^{n+1}} = \frac{k}{2^n}$, and thus $X_n(\omega) \leq X_{n+1}(\omega)$. If $X_n(\omega) = n$, then $X_n(\omega) \leq X_{n+1}(\omega)$. By this and by (1.17)- (1.18) we see that for each ω

$$X_n(\omega) \leq X_{n+1}(\omega) \uparrow X(\omega).$$

Then

$$E[X_n] = \sum_{k=0}^{n2^n-1} \frac{k}{2^n}\mathbf{P}(X \in E_{n,k}) + n\mathbf{P}(X \geq n)$$

$$= \sum_{k=0}^{n2^n-1} \frac{k}{2^n-1}\left(F_X\left(\frac{k+1}{2^n}\right) - F_X\left(\frac{k}{2^n}\right)\right) + n\mathbf{P}(X \geq n).$$

We write this (for some omitted details see eq. (1.35) in an exercise of section 1.12.3) as

$$E[X_n] = \sum_{k=0}^{n2^n-1} \int_{\{\omega \mid X_n(\omega) \in E_{n,k}\}} X_n(\omega)\mathbf{P}(d\omega) + \int_{\{\omega \mid X_n(\omega) \in E_n\}} X_n(\omega)\mathbf{P}(d\omega)$$

and since $\Omega = \left(\{\omega \mid X_n(\omega) \in E_{n,k}\}_{k=1}^{n2^n}, \{\omega \mid X_n(\omega) \in E_n\}\right)$ is a partition of Ω, we set

$$= \int_\Omega X_n(\omega) d\mathbf{P}(\omega).$$

Then it is seen by $X_n(\omega) \leq X_{n+1}(\omega)$ for all ω and as $E_{n,k} = E_{n+1,2k} \cup E_{n+1,2k+1}$ that

$$E\left[X_n\right] \leq E\left[X_{n+1}\right].$$

As $E\left[X_n\right]$ is a non-decreasing sequence, it has the limit $E\left[X\right] \leq +\infty$.

Example 1.8.1 For $A \in \mathcal{B}$ the function χ_A defined on $\mathbf{R}$ by

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases} \tag{1.20}$$

is a Borel function. Let X be a random variable. Then $\chi_A(X)$ is a random variable by theorem 1.5.2 and is non negative and simple. We get

$$E\left[\chi_A(X)\right] = 0 \cdot \mathbf{P}\left(X \in A^c\right) + 1 \cdot \mathbf{P}\left(X \in A\right).$$

We write then using (1.8)

$$E\left[\chi_A(X)\right] = \mathbf{P}\left(X \in A\right) = \int_A dF_X(x). \tag{1.21}$$

■

We shall define $E[X]$ for an arbitrary random variable in the next section.

1.8.2 The General Definition

Let $X \geq 0$ denote a random variable on $(\Omega, \mathcal{F}, \mathbf{P})$. Then its *expectation* was above defined as

$$E[X] = \int_\Omega X(\omega) d\mathbf{P}(\omega).$$

Again, we often revert to a useful shorthand, or

$$E[X] = \int_\Omega X d\mathbf{P}.$$

Let $F = F_X$ be the distribution function associated with X, then this may be rewritten as

$$E[X] = \int_0^\infty x dF(x),$$

as follows by the considerations in the preceding example 1.8.1.
For a real valued random variable, its expectation exists if $E[X^+] := \int_0^\infty xF(dx) < +\infty$ and $E[X^-] := -\int_{-\infty}^0 xF(dx) < +\infty$. Then the expectation is given by

$$E[X] = E[X^+] - E[X^-].$$

If we encounter a case where $E[X^+] = \infty$ and $E[X^-] = \infty$, the expected value is not defined.

1.8.3 The Law of the Unconscious Statistician

The following theorem, sometimes known as the law of the unconscious statistician, is extremely useful for computation and will be frequently cited in the sequel.

Theorem 1.8.2 Let X be a random variable and g a Borel function such that $E\left[g(X)\right] < \infty$.

$$E\left[g(X)\right] = \int_\Omega g(X) d\mathbf{P} = \int_{-\infty}^\infty g(x) dF(x) \tag{1.22}$$

Proof We follow [103, p. 317]. We assume that $g(x) \geq 0$, since otherwise we can use decomposition into the negative and positive parts g^+ and g^- as shown above. We assume in addition that g is simple, i.e., there are Borel sets $G_1, G_2, \ldots, G_m$ that are a partition of $\mathbf{R}$ such that

$$g(x) = g_k, \quad \text{if } x \in G_k \quad k = 1, 2, \ldots, m.$$

and $\cup_{k=1}^{m} G_k = \mathbf{R}$. We can use the construction in (1.15) with $Y = g(X)$

$$E[Y] = \sum_{k=1}^{m} g_k \mathbf{P}(Y = g_k). \tag{1.23}$$

Here

$$\{Y = g_k\} = \{\omega \mid g(X(\omega)) = g_k\} = \{\omega \mid X(\omega) \in G_k\} = \{X \in G_k\}.$$

And thus

$$\mathbf{P}(Y = g_k) = \mathbf{P}(X \in G_k).$$

Hence in (1.23)

$$E[Y] = \sum_{k=1}^{m} g_k \mathbf{P}(X \in G_k) = \sum_{k=1}^{m} \int_{X \in G_k} g(X(\omega)) d\mathbf{P}(\omega) = \int_{\Omega} g(X(\omega)) d\mathbf{P}(\omega),$$

where we used the result (1.35) in the exercises of this chapter, since $(\{X \in G_k\})_{k=1}^{m}$ is a partition of Ω, and thus

$$E[Y] = \int_{\Omega} g(X(\omega)) d\mathbf{P}(\omega). \tag{1.24}$$

On the other hand, the discussion in example 1.8.1 and the expression (1.21) tell us that

$$\mathbf{P}(X \in G_k) = \int_{G_k} dF_X(x),$$

and thus

$$\begin{aligned} E[Y] = \sum_{k=1}^{m} g_k \mathbf{P}(X \in G_k) &= \sum_{k=1}^{m} g_k \int_{G_k} dF_X(x) = \sum_{k=1}^{m} \int_{G_k} g(x) dF_X(x) \\ &= \int_{\mathbf{R}} g(x) dF_X(x), \end{aligned}$$

and thus

$$E[Y] = \int_{\mathbf{R}} g(x) dF_X(x).$$

Hence we have established the law of the unconscious statistician for non negative and simple g. The general statement follows by approximating a non negative g by simple functions (see the preceding) and then using g^+ and g^-. ■

1.8.4 Three Inequalities for Expectations

Theorem 1.8.3 (Jensen's Inequality) Suppose that ϕ is a convex function; namely, for any $\lambda \in (0, 1)$

$$\lambda \phi(x) + (1 - \lambda)\phi(y) \geq \phi(\lambda x + (1 - \lambda)y).$$

Then

$$E[\phi(X)] \geq \phi(E[X]).$$

Proof Let $c = E[X]$ and let $l(x) = ax + b$, where a and b are such that $l(c) = \phi(c)$ and $\phi(x) \geq l(x)$. Choose a such that

$$\lim_{h \downarrow 0} \frac{\phi(c) - \phi(c-h)}{h} \leq a \leq \lim_{h \downarrow 0} \frac{\phi(c+h) - \phi(c)}{h}.$$

Then set

$$l(x) = a(x - c) + \phi(c).$$

With this choice of function,

$$E[\phi(X)] \geq E[aX + b] = aE[X] + b = l(E[X]) = \phi(E[X]).$$

■

Theorem 1.8.4 (Hölder's Inequality) If $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$, then

$$E[| \ XY \ |] \leq E[|X|^p]^{1/p} E[|Y|^q]^{1/q}. \tag{1.25}$$

Proof By dividing through by $E[|X|^p]^{1/p} E[|Y|^q]^{1/q}$, one may consider the case of $E[|X|^p]^{1/p} = E[|Y|^q]^{1/q} = 1$. Furthermore, we use the notation

$$E[|X|^p]^{1/p} \overset{\text{def}}{=} \|X\|_p.$$

In chapter 7 we shall be specially interested in $E[|X|^2]^{1/2}$.

For $x \geq 0$ and $y \geq 0$, set

$$\phi(x, y) = \frac{1}{p}x^p + \frac{1}{q}y^q - xy$$

for $x \geq 0$, so that taking derivative with respect to x gives

$$\phi_x(x, y) = x^{p-1} - y$$

and

$$\phi_{xx}(x, y) = (p-1)x^{p-2}.$$

For fixed y, it follows that $\phi(x, y)$ has a minimum (in x) at $x_0 = y^{1/(p-1)}$. Note that $x_0^p = y^{p/(p-1)} = y^q$, so that

$$\phi(x_0) = (\frac{1}{p} + \frac{1}{q})y^p - y^{1/(p-1)}y = 0.$$

Since x_0 is a minimum, it follows that $xy \leq \frac{1}{p}x^p + \frac{1}{q}y^q$. Setting $x = X$, $y = Y$ and taking expectations yields

$$E[|XY|] \leq \frac{1}{p} + \frac{1}{q} = 1 = \|X\|_p \|Y\|_q.$$

■

Let χ_A denote the indicator function of a set A;

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases} \tag{1.26}$$

Theorem 1.8.5 (Chebychev's Inequality) Suppose that $\phi : \mathbf{R} \to \mathbf{R}_+$. Let $i_A = \inf\{\phi(y) : y \in A\}$. Then for any measurable set A,

$$i_A \mathbf{P}(X \in A) \leq E[\phi(X)\chi_A(X)] \leq E[\phi(X)].$$

Proof Exercise ■

One example of Chebychev's inequality as stated above is

$$\mathbf{P}\left(\mid X - E\left[X\right] \mid > a\right) \leq \frac{\mathrm{Var}\left[X\right]}{a^2}. \tag{1.27}$$

1.8.5 Limits and Integrals

There are several results concerning interchange of limits and integrals, for Riemann integrals we refer to [69, chapter 6.6.]. All of them rely crucially on the use of a σ algebra, which is *closed* under countable unions. The proofs require the full machinery of integration theory, c.f. [36, 63], and are therefore beyond the scope of these notes and of this course. For the definitions of lim sup and lim inf we refer to Appendix 1.9 and to (1.10) and (1.11).

Theorem 1.8.6 (Fatou's Lemma) Let $(X_n)_{n=1}^{\infty}$ be a sequence of non negative random variables. It holds that

$$\liminf_{n\to+\infty} E[X_n] \geq E[\liminf_{n\to+\infty} X_n].$$

■

Theorem 1.8.7 (Monotone Convergence Theorem) If $0 \leq X_n \uparrow X$, then $E[X_n] \uparrow E[X]$.

■

We say that a property, described by an event A, for a random variable X holds **almost surely**, if

$$\mathbf{P}\left(X \in A\right) = \mathbf{P}\left(\{\omega \mid X(\omega) \in A\}\right) = 1.$$

Theorem 1.8.8 (Dominated Convergence Theorem) If $X_n \to X$ almost surely, and $|X_n| < Y$ for all n and $E[\mid Y \mid] < +\infty$, then $E[X_n] \to E[X]$.

■

1.9 Appendix: $\limsup x_n$ and $\liminf x_n$

1.9.1 Sequences of real numbers

$(x_n)_{n=1}^{\infty}$ is any sequence of real numbers. For example,

$$x_n = \left(1 + \frac{1}{n}\right)^n.$$

We next define $\liminf_{n\to\infty} x_n$ and $\limsup_{n\to\infty} x_n$.

1.9.2 $\limsup x_n$

Let b be a real number, $-\infty \leq b \leq +\infty$. We say that

$$b = \limsup_{n\to\infty} x_n,$$

if 1) and 2) below hold.

- 1) For every $c > b$ there is an integer N such that

$$n > N \quad \Rightarrow x_n < c.$$

- 2) For every $c^{'} < b$ and for every integer N there is an $n > N$ such that

$$x_n > c^{'}.$$

  In other words, for any $\epsilon > 0$ only finitely many of x_n can be larger than $b+\epsilon$. Also, there are infinitely many x_n larger than $b-\epsilon$.

1.9.3 $\liminf x_n$

Let a be a real number, $-\infty \leq a \leq +\infty$. We say that

$$a = \liminf_{n\to\infty} x_n,$$

if 1) and 2) below hold.

- 1) For every $c > a$ and for every integer N there is an $n > N$ such that

$$x_n < c.$$

- 2) For every $c < a$ there is an integer N such that

$$n > N \quad \Rightarrow x_n > c.$$

  In other words, for any $\epsilon > 0$ there are infinitely many x_n smaller than $a+\epsilon$. Only finitely many of x_n can be smaller than $a-\epsilon$.

1.9.4 Properties, The Limit of a Sequence

$\liminf_{n\to\infty} x_n$ and $\limsup_{n\to\infty} x_n$ always exist, as they can be $\pm\infty$.
We have always that

$$\liminf_{n\to\infty} x_n \leq \limsup_{n\to\infty} x_n.$$

If

$$\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n,$$

then the limit

$$\lim_{n\to\infty} x_n = x$$

exists and

$$\lim_{n\to\infty} x_n = \liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n.$$

This is easy to see, because the properties above imply that for all $n > N$ we have infinitely many x_n with

$$a - \epsilon \leq x_n \leq b + \epsilon$$

But if $a = b$, then this yields the definition of a limit $x = (a = b)$ of $(x_n)_{n=1}^{\infty}$.

Example 1.9.1

$$x_n = (-1)^n \left(1 + \frac{1}{n}\right), \quad n = 1, 2, 3, \ldots, .$$

Then

$$\liminf_{n\to\infty} x_n = -1, \quad \limsup_{n\to\infty} x_n = 1.$$

We show the latter.

1) If $c > 1$, take an integer $N \geq \frac{1}{c-1}$. Then if $n > N$

$$x_n \leq \left(1 + \frac{1}{n}\right) < c.$$

2) If $c^{'} < 1$ and N is an integer, then if $n = 2k$ and $2k > N$, then

$$x_n = x_{2k} = (-1)^{2k}\left(1 + \frac{1}{2k}\right) = 1 + \frac{1}{2k} > c^{'}.$$

Hence

$$\limsup_{n\to\infty} x_n = 1$$

1.10 Appendix: $\limsup A_n$ and $\liminf A_n$

Let $A_n \in \mathcal{A}$ for each n in a countable collection $(A_n)_{n=1}^{\infty}$, where $\mathcal{A}$ is a sigma field. Let us define

$$\limsup_{n\to\infty} A_n \overset{\text{def}}{=} \cap_{n=1}^{\infty} \cup_{m=n}^{\infty} A_m \tag{1.28}$$

and

$$\liminf_{n\to\infty} A_n \overset{\text{def}}{=} \cup_{n=1}^{\infty} \cap_{m=n}^{\infty} A_m. \tag{1.29}$$

Then $\limsup_{n\to\infty} A_n \in \mathcal{A}$ and $\liminf_{n\to\infty} A_n \in \mathcal{A}$ (you should convince yourself of this).

Clearly

$$\liminf_{n\to\infty} A_n \subset \limsup_{n\to\infty} A_n.$$

If $\liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n$, we say that

$$\liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n = \lim_{n\to\infty} A_n.$$

Let χ_A denote the **indicator function** of an event $A \in \mathcal{A}$;

$$\chi_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases} \tag{1.30}$$

Then one can verify that

$$\{\omega \in \Omega \mid \sum_{n=1}^{\infty} \chi_{A_n}(\omega) = \infty\} = \cap_{n=1}^{\infty} \cup_{m=n}^{\infty} A_m,$$

and we say (clearly?) that A_n happens infinitely often. In addition,

$$\{\omega \in \Omega \mid \sum_{n=1}^{\infty} \chi_{A_n^c}(\omega) < \infty\} = \cup_{n=1}^{\infty} \cap_{m=n}^{\infty} A_m,$$

and we say (clearly?) that A_n happens ultimately (i.e, for all but finitely many n). Then (a quiz for self-studies)

$$\chi_{\cap_{n=1}^{\infty} \cup_{m=n}^{\infty} A_m} = \limsup_{n\to\infty} \chi_{A_n}$$

and

$$\chi_{\cup_{n=1}^{\infty} \cap_{m=n}^{\infty} A_m} = \liminf_{n\to\infty} \chi_{A_n}.$$

1.11 Appendix: Combinatorics of Counting and Statistics of Particles in Cells

Combinatorics is connected to probability in a great number of ways [15, 53]. Balls and urns are not idle toys, as often portrayed by, e.g., certain former alumni, but important conceptual models, as shown below by statistics of particles in cells. Here we recapitulate for ease of reference the customary rudiments.

Multiplication principle is a fundamental idea of counting. The principle says that if there are n_1 ways of doing operation 1 and n_2 ways of doing operation 2, then there are $n_1 \cdot n_2$ ways of performing both operations. Therefore n^k equals the number of ways for picking k elements (no restriction on k) with replacement and with regard to order in a collection of n items.

By the multiplication principle the factorial (n is a non negative integer)

$$n! = n \cdot (n-1) \cdot \ldots \cdot 3 \cdot 2 \cdot 1$$

with the convention $0! = 1$, is the number of ways of ordering n items in a collection. Then we define the expression $(n)_k$ for integers $0 \leq k \leq n$ by

$$(n)_k \overset{\text{def}}{=} n \cdot (n-1) \cdot \ldots \cdot (n-k+1).$$

It is seen that $(n)_k$ equals the number of ways to pick k items from the the collection n items without replacement and with the order of the items in the subcollection taken into account.

Let $P(n,k)$ be the number of ways to pick k items from the collection n items without replacement and without the order of the items in the subcollection taken into account. Then by the multiplication principle

$$k!P(n,k) = (n)_k$$

must hold, and we get

$$P(n,k) = \frac{(n)_k}{k!} = \frac{n!}{(n-k)!k!}.$$

The established symbol for $P(n,k)$ is $\begin{pmatrix} n \\ k \end{pmatrix}$, the binomial coefficient (reads as 'n choose k') for non negative integers n and k, $0 \leq k \leq n$, and thus

$$\begin{pmatrix} n \\ k \end{pmatrix} = \frac{n!}{(n-k)!k!}.$$

Then we have found three of the cases in the following table.

	With regard to order	Without regard to order
With replacement	n^k	$\begin{pmatrix} n+k-1 \\ k \end{pmatrix}$
Without replacement, $k \leq n$	$(n)_k$	$\begin{pmatrix} n \\ k \end{pmatrix}$

The derivation of the expression $\begin{pmatrix} n+k-1 \\ k \end{pmatrix}$ is a longer exercise, which we do by changing to an different interpretation of the combinatorial coefficients above.

Sampling with regard to order parallels distributing (k) distinguishable objects into (n) cells and sampling without regard to order parallels distributing indistinguishable objects into cells. Sampling with replacement corresponds to allowing more than one object in a cell, and sampling without replacement corresponds to allowing no more than one object in a cell, hence $k \leq n$. Thus we have the following table:

	Distinguishable Objects	Indistinguishable Objects
No restraints	n^k	$\begin{pmatrix} n+k-1 \\ k \end{pmatrix}$
Not more than one object per cell, $k \leq n$	$(n)_k$	$\begin{pmatrix} n \\ k \end{pmatrix}$

Consider k balls and n cells with $k_i \geq 0$ balls in cell i so that

$$k_1 + k_2 + \ldots + k_n = k$$

We call k_i's *occupation numbers*. We define the occupancy distribution by the n -tuple $(k_1, k_2, \ldots, k_n)$. Two ways of distributing k indistinguishable objects into cells are called indistinguishable, if their occupancy distributions are identical. Let us now consider the following fundamental question: how many distinguishable occupancy distributions can be formed by distributing k indistinguishable objects into n cells ? The answer is $\begin{pmatrix} n+k-1 \\ k \end{pmatrix}$. To prove this, we use a device invented by William Feller. This consists of representing n cells by the space between $n+1$ bars and the balls in the cells by stars between the bars. Thus

$$|||\ast\ast|||\ast\ast\ast||\ast|\ast\ast||\ast\ast\ast\ast|$$

represents $n = 11$, $k = 12$ and the occupancy distribution $(0, 0, 2, 0, 0, 3, 0, 1, 2, 0, 4)$.

Since the first and last symbols in a string of stars and bars must be bars, only $n-1$ bars and k stars can appear in any order. Thus we are back to counting the number of ways on how to pick k objects among $n+k-1$ objects without replacement and without regard to order. This equals by the preceding

$$\begin{pmatrix} n+k-1 \\ k \end{pmatrix}, \tag{1.31}$$

and we have established the last of the arrays in the tables above.

In statistical physics one thinks of a phase space subdivided into a large number, n, regions (cells) and k indistinguishable particles each of which falls into one of the cells. One could guess that each of the possible n^k arrangements of the k particles is equally probable, this is called *Maxwell -Boltzmann statistics*. If on the other hand, each of the possible occupancy distributions for the particles is considered equally probable, and no restrictions are made on the number of particles in each cell, then probability of each distinct arrangement is $1/\begin{pmatrix} n+k-1 \\ k \end{pmatrix}$. This is called *Bose-Einstein statistics* [17, p.354, Exercise 29.6 (b)]. If the k particles are indistinguishable particles and one imposes the restriction that no more than one particle can found in a cell, and the arrangments are equally probable, then the probability of an arragement is $1/\begin{pmatrix} n \\ k \end{pmatrix}$, and one talks about *Fermi-Dirac statistics* [17, p.354, Exercise 29.6 (a)]. The reference [17] loc.cit. shows how one derives Fermi-Dirac and Bose-Einstein distribution functions (which are not distribution functions in the sense defined above) from these expressions. One needs physical experiments to decide, which model of statistics holds for a certain system of particles (e.g., hydrogen atoms, electrons, neutrons, protons). In other words, one cannot argue solely from abstract mathematical principles as to what is to be regarded as equally likely events in reality [53].

1.12 Exercises

1.12.1 Easy Drills

1. $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space. $A \in \mathcal{F}$ and $B \in \mathcal{F}$. $\mathbf{P}((A \cup B)^c) = 0.5$ and $\mathbf{P}(A \cap B) = 0.2$. What is the probability that either A or B but not both will occur. (Answer: 0.3).

2. $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space. $A \in \mathcal{F}$ and $B \in \mathcal{F}$. If the probability that at least one of them occurs is 0.3 and the probability that A occurs but B does not occur is 0.1, what is $\mathbf{P}(B)$? (Answer: 0.2).

1.12.2 Measures, Algebras and Sigma Fields

1. **A measure that is finitely additive but not countably additive** (From [87]) Let Ω be countable. We take
$$\mathcal{A} = \{A \subset \Omega \mid \ A \text{ is finite or } A^c \text{ is finite }\}.$$
   (a) Show that $\mathcal{A}$ is an algebra.

   (b) Set
$$\mathbf{P}(A) = \begin{cases} 0 & \text{if } A \text{ is finite} \\ 1 & \text{if } A^c \text{ is finite.} \end{cases}$$
   Show that $\mathbf{P}$ is finitely additive, but not countably additive measure on $(\Omega, \mathcal{A})$.

2. Let $\Omega = [0, 1)$. For each of the set functions μ defined below, determine whether μ satisfies the axioms of probability. $0 \leq a < b \leq 1$.
   1. $\mu([a, b)) = \frac{b-a}{b+a}$.
   2. $\mu([a, b)) = b^2 - a^2$.
   3. $\mu([a, b)) = b^2 - a^2, \mu((a, b]) = b - a$.

3. $\Omega =$ the non negative integers $= \{0, 1, 2, \ldots\}$. Since Ω is countable, we can take $\mathcal{F} =$ all subsets of Ω. Let $0 < \theta < 1$ be given. For which values of θ is it possible to give a probability measure $\mathbf{P}$ on $(\Omega, \mathcal{F})$ such that $\mathbf{P}(\{i\}) = \theta^i$, $i = 0, 1, 2 \ldots$?

4. (From [43]) Let $\Omega = [0, \infty)$. let $\mathcal{F}$ the sigma field of subsets of Ω generated by sets of the form $(n, n + 1)$ for $n = 1, 2, \ldots$.
   (a) Are the following subsets of Ω in $\mathcal{F}$?
      (i) $[0, \infty)$
      (ii) $\mathbf{Z}_+ = \{0, 1, 2, \ldots\}$
      (iii) $[0, k] \cup [k + 1, \infty)$ for any positive integer k
      (iv) $\{k\}$ for any positive integer k
      (v) $[0, k]$ for any positive integer k
      (vi) $(1/3, 2)$

   (b) Define the following set function $\mathbf{P}$ on subsets A of Ω
$$\mathbf{P}(A) = c \sum_{\{k \in \mathbf{Z}_+ | (k+1/2) \in A\}} 3^{-k}.$$
   If there is no k such that $(k + 1/2) \in A$, then the sum is taken as zero. Is $\mathbf{P}$ a probability measure on $(\Omega, \mathcal{F})$, and if so, what is the value of c?

(c) Repeat part (b) for $\mathcal{F}$ replaced by the Borel sigma field.

(d) Repeat part (b) for $\mathcal{F}$ replaced by the power set of Ω.

5. Show that $\mathbf{P}$ defined in (1.3) in example 1.4.7 is a countably additive probability measure.

6. Assume $(\Omega, \mathcal{F}, \mathbf{P})$ and let $A \in \mathcal{F}$ and $B \in \mathcal{F}$, and $A \subseteq B$. Show that

$$\mathbf{P}\left(B \cap A^c\right) = \mathbf{P}\left(B\right) - \mathbf{P}\left(A\right).$$

7. Show that the probability that one and only one of the events A and B occurs is

$$\mathbf{P}(A) + \mathbf{P}(B) - 2\mathbf{P}(A \cap B).$$

8. Consider $(\Omega, \mathcal{F}, \mathbf{P})$ and let $A \in \mathcal{F}$ and $B \in \mathcal{F}$. Define

$$A \triangle B \overset{\text{def}}{=} \left(A \cap B^c\right) \cup \left(B \cap A^c\right). \tag{1.32}$$

This is known as the symmetric difference of A and B. You should convince yourself of the fact that $A \triangle B \in \mathcal{F}$.

(a) Show that

$$\mid \mathbf{P}(A) - \mathbf{P}(B) \mid \leq \mathbf{P}\left(A \triangle B\right).$$

(b) Show that if $A \in \mathcal{F}$, $B \in \mathcal{F}$ and $C \in \mathcal{F}$, then

$$\mathbf{P}\left(A \triangle B\right) \leq \mathbf{P}\left(A \triangle C\right) + \mathbf{P}\left(C \triangle B\right).$$

The sharp-eyed reader will recognize this as a form of the *triangle inequality.* One can in fact regard $\mathbf{P}\left(A \triangle B\right)$ as a distance or **metric on events**.

9. Show that if A and B are any two events, then

$$\min\left(\mathbf{P}(A), \mathbf{P}(B)\right) \geq \mathbf{P}(A \cap B) \geq \mathbf{P}(A) + \mathbf{P}(B) - 1.$$

10. Show that if $\mathbf{P}\left(A\right) \geq 1 - \delta$ and $\mathbf{P}\left(B\right) \geq 1 - \delta$, then also $\mathbf{P}\left(A \cap B\right) \geq 1 - 2\delta$. In words, if two events have probability near to one, then their intersection has probability nearly one.

11. A_1,A_2, A_3,..., and A_n are events. Show that

$$\mathbf{P}\left(\cap_{j=1}^{n} A_j\right) \geq \sum_{j=1}^{n} \mathbf{P}(A_j) - (n-1).$$

12. **Boole's inequalities**
A_1,A_2, A_3,..., and A_n are events. Prove that

(a)

$$\mathbf{P}\left(\cup_{j=1}^{n} A_j\right) \leq \sum_{j=1}^{n} \mathbf{P}(A_j).$$

(b)

$$\mathbf{P}\left(\cap_{j=1}^{n} A_j\right) \geq 1 - \sum_{j=1}^{n}\left(1 - \mathbf{P}\left(A_j\right)\right).$$

13. A_1,A_2, A_3,..., and A_n are independent events. Prove that their complements are A_1^c,A_2^c, A_3^c,..., and A_n^c are independent events, too.

14. Suppose that $A \cap B \subseteq C$ holds for the events A, B and C. Show that

$$\mathbf{P}\left(C^c\right) \leq \mathbf{P}\left(A^c\right) + \mathbf{P}\left(B^c\right). \tag{1.33}$$

15. [68] Suppose that $\mathbf{P}$ is a finitely additive probability measure defined on a field $\mathcal{G}$ of subsets of a space Ω. Assume that $\mathbf{P}$ is continuous at $\emptyset$, i.e., if $A_n \in \mathcal{G}$ for all n and $A_n \downarrow \emptyset$, then $\mathbf{P}\left(A_n\right) \downarrow 0$. Show that $\mathbf{P}$ is a probability measure on $\mathcal{G}$.

    This result is useful, since in applications one often encounters a finitely additive measure on a field $\mathcal{G}$ rather than a mesure on a σ-field $\mathcal{F}$.

1.12.3 Random Variables and Expectation

1. Let X be a simple function with the range $V_X = \{x_1, \ldots, x_m\}$. Then the sets

$$G_i = \{\omega \in \Omega \mid X(\omega) = x_i\}$$

are a (measurable) partition of Ω in the sense that

$$G_i \cap G_j, i \neq j, \quad \Omega = \cup_{i=1}^m G_i.$$

For any event $A \subset \Omega$ let χ_A be the indicator function of the event A,

$$\chi_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases} \tag{1.34}$$

Then $\chi_A \cdot X$ is a simple random variable with the range $V_{\chi_A \cdot X}$ containing those x_i for which $G_i \cap A \neq \emptyset$. $V_{\chi_A \cdot X}$ is augmented with zero 0, if needed. Then we define $\int_A X d\mathbf{P}$ by

$$\int_A X d\mathbf{P} = \int_\Omega \chi_A \cdot X d\mathbf{P} = E\left[\chi_A \cdot X\right].$$

Show that if $A \cap B = \emptyset$, then

$$\int_{A \cup B} X d\mathbf{P} = \int_A X d\mathbf{P} + \int_B X d\mathbf{P}$$

Hint: Think of how $\chi_{A \cup B}$ can be expressed by means of χ_A and χ_B.

Thus,

$$\int_\Omega X d\mathbf{P} = \sum_{i=1}^m \int_{G_i} X d\mathbf{P}. \tag{1.35}$$

2. Let X be a positive random variable, $\mathbf{P}\left(X > 0\right) = 1$, with $E\left[X\right] < \infty$. Show that

$$E\left[\frac{1}{X}\right] \geq \frac{1}{E\left[X\right]}. \tag{1.36}$$

Note that this inequality is trivially true if $E\left[\frac{1}{X}\right] = +\infty$.

3. Let X and Y be independent random variables. Assume that $E\left[|X|\right] < \infty$, $E\left[|Y|\right] < \infty$, $E\left[|XY|\right] < \infty$. Show that

$$E\left[X \cdot Y\right] = E\left[X\right] \cdot E\left[Y\right]. \tag{1.37}$$

We do this in steps, c.f. [13, p. 403]. Our tools for this are the small pieces of integration theory in section 1.8 and the definition 1.6.1.

(a) Choose arbitrary $A \in \mathcal{F}_X$ and $B \in \mathcal{F}_Y$. Then show that

$$E[\chi_A \cdot \chi_B] = E[\chi_A] \cdot E[\chi_B].$$

(b) By means of the item (a) check in detail that (1.37) holds for all simple random variables X and Y.

(c) Explain how you can obtain (1.37) for $X \geq 0$ and $Y \geq 0$.

4. Let $F_1 \in \mathcal{F}$ and $F_2 \in \mathcal{F}$. Prove that $\chi_{F_1} + \chi_{F_2}$ is a random variable.

5. Show with the aid of Appendix 1.9 that

$$\chi_{\cap_{n=1}^{\infty} \cup_{m=n}^{\infty} A_m} = \limsup_{n \to \infty} \chi_{A_n}$$

and

$$\chi_{\cup_{n=1}^{\infty} \cap_{m=n}^{\infty} A_m} = \liminf_{n \to \infty} \chi_{A_n}.$$

6. **Markov's inequality** Let X be such that $\mathbf{P}(X \geq 0) = 1$, i.e., X is almost surely nonnegative. Show that for any $c > 0$

$$\mathbf{P}(X \geq c) \leq \frac{E[X]}{c}. \tag{1.38}$$

Aid: Let $A = \{\omega \in \Omega \mid X(\omega) \geq c\}$. Let χ_A be the corresponding indicator function, i.e.,

$$\chi_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Then we have clearly $X \geq c\chi_A$.

7. [30] Let for $n = 1, 2, \ldots$

$$A_n = \begin{cases} A & \text{if } n \text{ is even} \\ B & \text{if } n \text{ is odd.} \end{cases}$$

Show that

$$\limsup_{n \to +\infty} A_n = A \cup B, \quad \liminf_{n \to +\infty} A_n = A \cap B.$$

8. [30] Let $\{A_n\}_{n \geq 1}$ be as sequence of pairwise disjoint sets. Show that

$$\limsup_{n \to +\infty} A_n = \liminf_{n \to +\infty} A_n = \emptyset.$$

9. **Monotone class** A class $\mathcal{M}$ of subsets of Ω is called a monotone class, if

$$\lim_{n \to \infty} A_n \in \mathcal{M}$$

for any increasing or decreasing sequence of sets $\{A_n\}_{n \geq 1}$ in $\mathcal{M}$. Show that an algebra $\mathcal{M}$ is a sigma algebra if and only if it is a monotone class. *Aid:* In one direction the assertion is obvious. In the other direction, consider $B_n = \cup_{k=1}^{n} A_k$.

10. $\mathcal{F}_1$ and $\mathcal{F}_2$ are two sigma algebras of subsets of Ω. Show that

$$\mathcal{F}_1 \cap \mathcal{F}_2$$

is a sigma algebra of subsets of Ω.

Chapter 2

Probability Distributions

2.1 Introduction

In this chapter we summarize, for convenience of reference, items of probability calculus that are in the main supposed to be already familiar. Therefore the discourse will partly be sketchy and akin in style to a chapter in a handbook like [92].

We shall first introduce the distinction between continuous and discrete r.v.'s. This is done by specifying the type of the distribution function. Appendix 2.5 provides a theoretical orientation to distribution functions and can be skipped by first reading.

We start by defining the continuous random variables. Let first $f_X(x)$ be function such that

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1, \quad f_X(x) \geq 0, \quad \text{for all } x \text{ in } \mathbf{R}.$$

Then the function

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du \tag{2.1}$$

is the distribution function of a random variable X, as can be checked by theorem 1.5.7, and as was in advance suggested by the notation. We say that X is a **continuous** (univariate) **random variable**. The function $f_X(x)$ is called the **probability density function p.d.f.** (p.d.f.) of X. In fact (c.f. appendix 2.5) we have for any Borel set A that

$$\mathbf{P}\left(X \in A\right) = \int_A f_X(x)dx.$$

In this chapter an array of continuous random variables will be defined by means of families of probability densities $f_X(x)$. By families of probability densities we mean explicit expressions of $f_X(x)$ that depend on a finite set of **parameters**, which assume values in suitable (sub)sets of real numbers. Examples are normal (Gaussian), exponential, Gamma e.t.c. distributions.

> The parameters will be indicated in the symbolical codes for the distributions, e.g., Exp(a) stands for the exponential distribution with the parameter a, $a > 0$. The usage is to write, e.g., $X \in \text{Exp}(a)$, when saying that the r.v. X has an exponential distribution with parameter a.

Next we say that X is a **discrete** (univariate) **random variable**, if there is a countable (finite or infinite) set of real numbers $\{x_k\}$, one frequent example is the non negative integers, such that

$$F_X(x) = \sum_{x_k \leq x} p_X(x_k), \tag{2.2}$$

where

$$p_X(x_k) = \mathbf{P}\left(X = x_k\right).$$

The function $p_X(x_k)$ is called the **probability mass function** (p.m.f.) of X. Then it must hold that

$$\sum_{k=-\infty}^{\infty} p_X(x_k) = 1, p_X(x_k) \geq 0.$$

Again we shall define discrete random variables by parametric families of distributions (Poisson, Binomial, Geometric, Waring e.t.c).

> It is found in appendix 2.5 that there are random variables that are neither continuous or discrete or mixed cases of continuous and discrete. In other words, there are distribution functions that do not have either a p.d.f. or a p.m.f. or a mixture of those. A well known standard instance is the **Cantor distribution**, which is the topic in an exercise to this chapter.
>
> In addition, since the calculus of integrals teaches us that $\mathbf{P}\left(X = x\right) = \int_x f_X(t)dt = 0$, there is a foundational difficulty with continuous random variables likely connected to the nature of real numbers as a description of reality.

If the expectation (or mean) of X, as defined in section 1.8.2, exists, it can be computed by

$$E\left[X\right] = \begin{cases} \sum\limits_{k=-\infty}^{\infty} x_k p_X(x_k) & \text{discrete r.v.,} \\ \int\limits_{-\infty}^{\infty} x f_X(x)\, dx & \text{continuous r.v.} \end{cases} \tag{2.3}$$

The law of the unconscious statistician proved in theorem 1.8.2, see (1.22), can now be written as

$$E\left[H(X)\right] = \begin{cases} \sum\limits_{k=-\infty}^{\infty} H(x_k) p_X(x_k) & \text{discrete r.v.,} \\ \int\limits_{-\infty}^{\infty} H(x) f_X(x)\, dx & \text{continuous r.v..} \end{cases} \tag{2.4}$$

Thereby we have, with $H(x) = \left(x - E\left[X\right]\right)^2$, the variance, when it exists, expressed by

$$\text{Var}\left[X\right] = E\left[H(X)\right] = \begin{cases} \sum\limits_{k=-\infty}^{\infty} \left(x_k - E\left[X\right]\right)^2 p_X(x_k) & \text{discrete r.v.,} \\ \int\limits_{-\infty}^{\infty} \left(x - E\left[X\right]\right)^2 f_X(x)\, dx & \text{continuous r.v.} \end{cases} \tag{2.5}$$

It follows readily that we have **Steiner's formula**

$$\text{Var}\left[X\right] = E\left[X^2\right] - \left(E\left[X\right]\right)^2. \tag{2.6}$$

This formula facilitates computations, and is applicable in both discrete and continuous cases.

Remark 2.1.1 In the sequel we shall frequently come across with $\Gamma(z)$, which is, for z with positive real part, the **Euler gamma function**, see [93, p. 302] for a quick reference, and [54, ch. 6] for a story,

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt. \tag{2.7}$$

Some notable properties of $\Gamma(z)$ are

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \tag{2.8}$$

and

$$\Gamma(n) = (n-1)! \quad n \text{ is a nonnegative integer.} \tag{2.9}$$

Here we recall the convention $0! = 1$. The following integral is useful and inherent in several expressions in the sequel.

$$\int_0^\infty x^t e^{-\lambda x} dx = \frac{\Gamma(t+1)}{\lambda^{t+1}}, \quad \lambda > 0, t > -1. \tag{2.10}$$

2.2 Continuous Distributions

Many continuous and discrete distributions have one or more **generative models**. By a generative model one means in probability theory a mathematical description of the way a particular probability distribution can arise in a physical situation. For example, the logarithm of a product of many positive random variables is a generative model for the log-normal distribution. We shall on occasion try to refer to such models, when introducing a distribution.

2.2.1 Univariate Continuous Distributions

Example 2.2.1 (Uniform Distribution) $X \in U(a,b)$, $a < b$ is a random variable with the p.d.f.

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ \\ 0 & \text{elsewhere.} \end{cases} \tag{2.11}$$

We say that X has the uniform distribution on the interval (a,b). The **parameters** are a and b. We have

$$E\left[X\right] = \frac{a+b}{2}, \mathrm{Var}\left[X\right] = \frac{(b-a)^2}{12}.$$

Frequently encountered special cases are $U(0,1)$ and $U(-1,1)$. The uniform distribution has been discussed in terms of 'complete ignorance'.

■

Example 2.2.2 (Triangular Distribution) $X \in \mathrm{Tri}(-1,1)$ means that the p.d.f. of X is

$$f_X(x) = \begin{cases} 1-|x| & |x| < 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{2.12}$$

This can also be written as

$$f_X(x) = \max\left(0, 1-|x|\right).$$

If one draws a function graph of $\max\left(0, 1-|x|\right)$, one realizes the rationale for the attribute triangular.

$$E\left[X\right] = 0, \mathrm{Var}\left[X\right] = \frac{1}{6}.$$

■

Example 2.2.3 (General Triangular Distribution) $X \in \text{Tri}(a,b)$ means that the p.d.f. of X is

$$f_X(x) = \begin{cases} \frac{2}{b-a}\left(1 - \frac{2}{b-a} \mid x - \left(\frac{a+b}{2}\right) \mid\right) & a < x < b \\ 0 & \text{elsewhere.} \end{cases} \tag{2.13}$$

$$E\left[X\right] = \frac{a+b}{2}, \text{Var}\left[X\right] = \frac{(b-a)^2}{24}.$$

■

Example 2.2.4 (Normal Distribution a.k.a. Gaussian Distribution) $X \in N(\mu, \sigma^2)$, $\mu \in \mathbf{R}$, $\sigma > 0$ means that the p.d.f. of X is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < +\infty. \tag{2.14}$$

We say that X has a normal distribution or a Gaussian distribution with the parameters μ and σ^2, where

$$E\left[X\right] = \mu, \text{Var}\left[X\right] = \sigma^2.$$

The univariate normal or Gaussian distribution will be the platform on which to construct the multivariate Gaussian distribution in chapter 8 and then eventually Gaussian processes in section 9.

■

Example 2.2.5 (Standard Normal Distribution or Standard Gaussian Distribution) The special case $X \in N(0,1)$ of (2.14) is called the standard normal distribution or the standard Gaussian distribution and its p.d.f. is important enough to have a special symbol reserved to it, namely

$$\phi(x) \overset{\text{def}}{=} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < +\infty. \tag{2.15}$$

The corresponding distribution function is designated by $\Phi(x)$, i.e.,

$$\Phi(x) \overset{\text{def}}{=} \int_{-\infty}^{x} \phi(t)dt, \quad -\infty < x < +\infty. \tag{2.16}$$

It follows readily that

$$\Phi(-x) = 1 - \Phi(x), \tag{2.17}$$

$$\Phi\left(0\right) = \frac{1}{2}. \tag{2.18}$$

In the engineering and scientific literature [3] as well as in MATLAB^R, one frequently meets the **error function**[1]

$$\text{erf}(x) \overset{\text{def}}{=} \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt, \quad -\infty < x < \infty, \tag{2.19}$$

and **complementary error function** General

$$\text{erfc}(x) \overset{\text{def}}{=} \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt, \quad -\infty < x < \infty. \tag{2.20}$$

[1] The definition of erf(x) varies in the literature, c.f., for example [32, p.78].

Clearly,

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x).$$

By a change of variable in (2.19) we find that

$$\Phi\left(x\right) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right),$$

and

$$\Phi\left(x\right) = \frac{1}{2}\mathrm{erfc}\left(-\frac{x}{\sqrt{2}}\right).$$

The distribution function of $X \in N(0,1)$, $\Phi(x)$, is often numerically calculated for $x > 0$ by means of the '**Q-function**' or the **error function**. or

$$Q(x) \overset{\mathrm{def}}{=} \int_x^{\infty} \phi(t)dt, \quad \Phi(x) = 1 - Q(x), \tag{2.21}$$

where the following approximation [2] is known to be very accurate

$$Q(x) \approx \left(\frac{1}{\left(1-\frac{1}{\pi}\right)x + \frac{1}{\pi}\sqrt{x^2+2\pi}}\right)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

■

Example 2.2.6 (Skew-Normal Distribution) A random variable X is said to have a skew-normal distribution, if it has the p.d.f.

$$f_X(x) = 2\phi(x)\Phi(\lambda x), \quad -\infty < x < \infty, \tag{2.22}$$

where the parameter $-\infty < \lambda < \infty$ and $\phi(x)$ is the p.d.f. of $N(0,1)$, and $\Phi(x)$ is the distribution function of $N(0,1)$. We write $X \in \mathrm{SN}\left(\lambda\right)$ and note by (2.18) that $\mathrm{SN}\left(0\right) = N(0,1)$. We have two plots of $f_X(x)$ in figure 2.1.

If $\lambda \to \infty$, then $f_X(x)$ converges (pointwise) to

$$f_X(x) = \begin{cases} 2\phi(x) & \text{if } x \geq 0 \\ 0 & \text{if } x < 0, \end{cases}$$

which is a **folded normal distribution**. If $\lambda \to -\infty$, then $f_X(x)$ converges (pointwise) to

$$f_X(x) = \begin{cases} 0 & \text{if } x \geq 0 \\ 2\phi(x) & \text{if } x < 0, \end{cases}$$

which is another folded normal distribution. The mean and variance as well as other properties of $\mathrm{SN}\left(\lambda\right)$ are established in the exercises to this chapter and to a later chapter.

■

Example 2.2.7 (Exponential Distribution) $X \in \mathrm{Exp}\left(\lambda\right)$, $\lambda > 0$, and the p.d.f. is

$$f_X(x) = \begin{cases} \frac{1}{\lambda}e^{-x/\lambda} & 0 \leq x \\ & \\ 0 & x < 0. \end{cases} \tag{2.23}$$

$$E\left[X\right] = \lambda, \mathrm{Var}\left[X\right] = \lambda^2.$$

[2] P.O. Börjesson and C.E.W. Sundberg: Simple Approximations of the Error Function $Q(x)$ for Communication Applications. *IEEE Transactions on Communications*, March 1979, pp. 639−643.

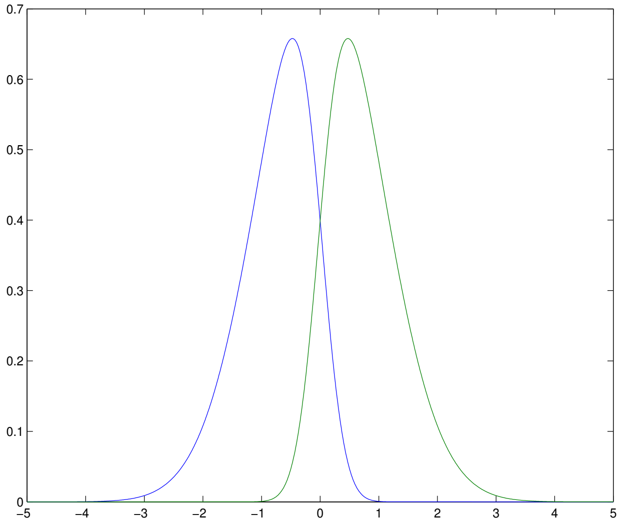

Figure 2.1: The p.d.f.'s of $\mathrm{SN}\,(-3)$ (the left hand function graph) and $\mathrm{SN}\,(3)$ (the right hand function graph).

■

Example 2.2.8 (Laplace Distribution) $X \in L\,(a)$, $a > 0$, means that X is a continuous r.v., and that its p.d.f. is

$$f_X(x) = \frac{1}{2a} e^{-|x|/a}, \quad -\infty < x < +\infty. \tag{2.24}$$

We say that X is Laplace -distributed with parameter a.

$$E\,[X] = 0, \mathrm{Var}\,[X] = 2a^2.$$

The distribution in this example is for obvious reasons also known in the literature as the **Double Exponential** distribution. We shall in the sequel provide exercises generating the Laplace distribution as the distribution of difference between two independent exponential r.v.'s.

■

Example 2.2.9 (Gamma Distribution) Let $X \in \Gamma\,(p, a)$, $p > 0$, $a > 0$. The p.d.f. is

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(p)} \frac{x^{p-1}}{a^p} e^{-x/a} & 0 \leq x \\ \\ 0 & x < 0. \end{cases} \tag{2.25}$$

Note that $\mathrm{Exp}\,(a) = \Gamma\,(1, a)$.

$$E\,[X] = pa, \mathrm{Var}\,[X] = pa^2.$$

Sometimes p is called the shape parameter and a is called the scale parameter. Note that the scale is squared in the expression for variance.

■

Example 2.2.10 (Erlang Distribution) The special case $\Gamma(k,a)$ of the Gamma distribution, where k is a positive integer, is known as the Erlang distribution, say Erlang (k,a) [3].

■

Example 2.2.11 (Weibull Distribution) Let $X \in \mathrm{Wei}(\alpha,\beta)$, $\alpha > 0$, $\beta > 0$. The p.d.f. is

$$f_X(x) = \begin{cases} \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-(x/\beta)^\alpha} & 0 \leq x \\ \\ 0 & x < 0. \end{cases} \tag{2.26}$$

Here α is the **shape parameter**, $\beta > 0$ is the **scale parameter**. Note that $\mathrm{Exp}(a) = \mathrm{Wei}(1,a)$. The exponential distribution is thus a special case of both the Gamma distribution and the Weibull distribution. There are, however, Gamma distributions that are not Weibull distributions and vice versa. The distribution was invented by Waloddi Weibull[4].

$$E[X] = \beta\Gamma\left(1+\frac{1}{\alpha}\right), \mathrm{Var}[X] = \beta^2\left[\Gamma\left(1+\frac{2}{\alpha}\right) - \left(\Gamma\left(1+\frac{1}{\alpha}\right)\right)^2\right].$$

In fracture mechanics one finds the **three parameter Weibull** distribution $\mathrm{Wei}(\alpha,\beta,\theta)$ with the p.d.f.

$$f_X(x) = \begin{cases} \frac{\alpha}{\beta}\left(\frac{x-\theta}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x-\theta}{\beta}\right)^\alpha} & x \geq \theta \\ 0 & x < \theta. \end{cases}$$

α is, as above, the shape parameter, $\beta > 0$ the scale parameter and θ is the **location parameter**. If $\theta = 0$, then $\mathrm{Wei}(\alpha,\beta,0) = \mathrm{Wei}(\alpha,\beta)$.

■

Example 2.2.12 ($\chi^2(f)$- Distribution with f Degrees of Freedom) If the random variable X has the p.d.f. for $f = 1,2,\ldots$

$$f_X(x) = \begin{cases} \dfrac{x^{\frac{f}{2}-1}e^{-x/2}}{\Gamma(f/2)2^{f/2}} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0, \end{cases}$$

then X is said to be $\chi^2(f)$- distributed with f degrees of freedom. We write $X \in \chi^2(f)$. Note that $\chi^2(f) = \Gamma(f/2,2)$.

$$E[X] = f, \mathrm{Var}[X] = 2f^2.$$

The following theorem explains the genesis of $\chi^2(f)$ and is included in section 4.7 as an exercise.

[3] This distribution has been named after A.K. Erlang (1878−1929), Danish mathematician and engineer, a pioneer in the development of statistical models of telephone traffic, see, e.g., [84].

[4] (1887−1979), was an engineer, a commissioned officer of coastal artillery, and a mathematician. He was professor in machine components at KTH. He studied strength of materials, fatigue, bearings, and introduced what we now call the Weibull distribution based on case studies, i.e., not on generative models.

Theorem 2.2.13 $X_1, \ldots, X_n$ are independent and $N(0,1)$ distributed. Then

$$\sum_{i=1}^{n} X_i^2 \in \chi^2(n). \tag{2.27}$$

■

Example 2.2.14 (Student's t-distribution) If the random variable X has the p.d.f. for $n = 1, 2, \ldots$

$$f_X(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \frac{1}{\left(1+\frac{x^2}{n}\right)^{(n+1)/2}}, -\infty < x < \infty,$$

then X is said to be $t(n)$- distributed with n degrees of freedom. We write $X \in t(n)$.

$$E\left[X\right] = 0, \text{Var}\left[X\right] = \frac{n}{n+2}.$$

The following theorem about Student's t-distribution is recognized from courses in statistics. It is in the sequel an exercise on computing the p.d.f. of a ratio of two continuous r.v.'s

Theorem 2.2.15 $X \in N(0,1)$, $Y \in \chi^2(n)$, where X and Y are independent. Then

$$\frac{X}{\sqrt{\frac{Y}{n}}} \in t(n). \tag{2.28}$$

■

Example 2.2.16 (Cauchy Distribution) $X \in C\left(m, a\right)$ has the p.d.f.

$$f_X(x) = \frac{1}{\pi} \frac{a}{a + (x-m)^2}, \quad -\infty < x < +\infty. \tag{2.29}$$

In particle physics, the Cauchy distribution $C\left(m, a\right)$ is known as the (non-relativistic) Wigner distribution [37] or the Breit-Wigner distribution [64, p.85]. An important special case is the standard Caychy distribution $X \in C\left(0, 1\right)$, which has the p.d.f.

$$f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2}, \quad -\infty < x < +\infty. \tag{2.30}$$

If we try to find the expectation of $X \in C\left(0,1\right)$, we start by

$$\int_a^b \frac{x}{1+x^2} dx = \frac{1}{2}\left(\ln\left(1+b^2\right) - \ln\left(1+a^2\right)\right).$$

When $b \to \infty$ and $a \to -\infty$, we see by the above that the integral $\int_{-\infty}^{\infty} \frac{x}{1+x^2} dx$ has no definite meaning[5] . Moments of higher order do not exist for $X \in C\left(0,1\right)$.

■

[5] Unless we define the integral by the Cauchy principal value.

Example 2.2.17 (Rayleigh Distribution) We say that X is Rayleigh distributed, if it has the density, $a > 0$,

$$f_X(x) = \begin{cases} \frac{2x}{a} e^{-x^2/a} & x \geq 0 \\ 0 & \text{elsewhere.} \end{cases}$$

We write $X \in \text{Ra}\,(a)$.

$$E\,[X] = \frac{1}{2}\sqrt{\pi a}, \quad \text{Var}\,[X] = a\left(1 - \frac{1}{4}\pi\right).$$

The parameter in the Rayleigh p.d.f. as recapitulated in [92] is defined in a slightly different manner. The Rayleigh distribution is a special case of the Rice distribution presented an exercise, which therefore is a generative model for $\text{Ra}\,(a)$.

■

Example 2.2.18 (Beta Distribution) The *Beta function* $B\,(x, y)$ (see, e.g., [3, pp. 82−86]) is defined for real $r > 0$ and $s > 0$ as

$$B\,(r, s) = \int_0^1 x^{r-1} \cdot (1 - x)^{s-1} dx = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r + s)}. \tag{2.31}$$

Taking this for granted we have

$$\frac{\Gamma(r + s)}{\Gamma(r)\Gamma(s)} \int_0^1 x^{r-1} \cdot (1 - x)^{s-1} dx = 1. \tag{2.32}$$

Since $\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} x^{r-1} \cdot (1 - x)^{s-1} \geq 0$ for $0 \leq x \leq 1$, we have found that

$$f_X\,(x) = \begin{cases} \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} x^{r-1} \cdot (1 - x)^{s-1} & 0 \leq x \leq 1 \\ 0 & \text{elsewhere,} \end{cases} \tag{2.33}$$

is a p.d.f. to be called the *Beta density.* We write $X \in \beta\,(r, s)$, if X is a random variable that has a Beta density. This p.d.f. plays an important role in Bayesian statistics.

$$E\,[X] = \frac{r}{r + s}, \text{Var}\,[X] = \frac{rs}{(r + s)^2(r + s + 1)}.$$

The function

$$B_x(r, s) = \int_0^x u^{r-1} \cdot (1 - u)^{s-1} du, \tag{2.34}$$

is known as the **incomplete Beta function**.

Example 2.2.19 (Gumbel Distribution) Let us consider the function

$$F(x) = e^{-e^{-x}}, \quad -\infty < x < \infty. \tag{2.35}$$

One should check the sufficient conditions of theorem 1.5.7 ensuring that $F(x)$ is the distribution function of some random variable X. The probability distribution corresponding to (2.35) is known as the **Gumbel** distribution, and the compact notation is $X \in$ Gumbel. The Gumbel distribution belongs to the family of **extreme value distributions**. This indicates that it emerges as a model for the distribution of the maximum (or the minimum) of a number of samples of various distributions. This will be demonstrated for sequences of independent and identically exponentially distributed X in chapter 6 below.

$$E\,[X] = \gamma,$$

where γ is Euler's constant $= 0.577\ldots$[6] , and

$$\mathrm{Var}\left[X\right] = \frac{\pi^2}{6}.$$

The stated expectation and variance of Gumbel distribution will be derived by means of moment generating functions (section 5.7) in an exercise to section 5.8.2 below. The Gumbel distribution and other extreme value distributions are important, e.g., in structural safety analysis [77].

■

Example 2.2.20 (Continuous Pareto Distribution) A continuous random variable X has the p.d.f.

$$f_X(x) = \begin{cases} \frac{\alpha k^{\alpha}}{x^{\alpha+1}} & x > k, \\ 0 & x \leq k, \end{cases} \tag{2.36}$$

where $k > 0$, $\alpha > 0$, which is called a **Pareto density** with parameters k and α. We write $X \in \mathrm{Pa}(k, \alpha)$.

$$E\left[X\right] = \frac{\alpha k}{\alpha - 1}, \mathrm{Var}\left[X\right] = \frac{\alpha k^2}{(\alpha-2)(\alpha-1)^2}, \alpha > 2.$$

This distribution was found by and named after the economist and sociologist Vilfredo Pareto (1848-1923)[7], as a frequency of wealth as a function of income category (above a certain bottom level). In plain words this means: most success seems to migrate to those people or companies, who are already popular.

■

Example 2.2.21 (Inverse Gaussian Distribution) A continuous random variable X with the p.d.f.

$$f_X(x) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} e^{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}}, \quad x > 0, \tag{2.37}$$

is said to have the inverse Gaussian distribution a.k.a. Wald distribution. We write $X \in \mathrm{IG}(\mu, \lambda)$, where

$$E\left[X\right] = \mu > 0, \mathrm{Var}\left[X\right] = \frac{\mu^3}{\lambda}.$$

The inverse Gaussian distribution is the distribution of the time a Wiener process with positive drift takes to reach a fixed positive level.

■

Example 2.2.22 (K-Distribution) A continuous random variable X with the p.d.f.

$$f_X(x) = \frac{2}{x}\left(\frac{L\nu x}{\mu}\right)^{\frac{L+\nu}{2}} \frac{1}{\Gamma(L)\Gamma(\nu)} I_{\nu-L}\left(2\sqrt{\frac{L\nu x}{\mu}}\right), \quad x > 0 \tag{2.38}$$

[6] A review of Euler's constant and related issues is recapitulated very readably in [54].

[7] We quote from the entry on V. Pareto in a classic Swedish Encyclopedia, *Nordisk familjbok, Tjugoförsta bandet*, Uggleupplagan, 1915: "**Pareto** [-tå] Vilfredo, italiensk-schweizisk nationalekonom, född 1848 i Paris, utbildades till ingenjör, men öfvergick så småningom till nationalekonomien, ..., P. har tilldragit sig mycken uppmärksamhet genom sin med matematiska formler demonstrerade och af rikhaltiga statistiska uppgifter belysta teori om inkomstfördelningen mellan de olika samhällsmedlemmarna i skilda länder, en fördelning som mindre motsvara en egentlig pyramid än en sådan med konkava sidor och konvex bas, en toppsnäcka enligt P:s egen beskrivning."

is said to have the K-distribution. Here $I_{\nu-\mu}(z)$ is the modified Bessel function of the second kind. We write $X \in \mathrm{K}(L, \mu, \nu)$. It holds that

$$E[X] = \mu, \quad \mathrm{Var}[X] = \mu^2 \frac{\nu + L + 1}{L\nu}.$$

$X \in \mathrm{K}(L, \mu, \nu)$ is the distribution of the product of two independent random variabels

$$X = X_1 \cdot X_2,$$

where $X_1 \in \Gamma(1/L, L)$, and $X_2 \in \Gamma(\mu/\nu, \nu)$. K-distribution is used as a probabilistic model in Synthetic Aperture Radar (SAR) imagery.

■

Example 2.2.23 (Logistic Distribution) We say that X has a logistic distribution, $X \in \mathrm{logistic}(0, 1)$, if its p.d.f. is

$$f_X(x) = \frac{e^x}{(1+e^x)^2}, -\infty < x < +\infty. \tag{2.39}$$

The corresponding distribution function is

$$F_x = \int_{-\infty}^{x} f_X(t)dt = \sigma(x).$$

The function $\sigma(x) = \frac{1}{1+e^{-x}}$ is known as the *logistic function*, hence the name of the probability distribution. The function $\sigma(x)$ appears also, e.g., in mathematical biology and artificial neural networks.

$$E[X] = 0, \quad \mathrm{Var}[X] = \frac{\pi^2}{3}.$$

■

2.2.2 Continuous Bivariate Distributions

(X, Y) is a bivariate random variable. Let

$$F_{X,Y}(x, y) = \mathbf{P}(X \leq x, Y \leq y), -\infty < x < \infty, -\infty < y < \infty.$$

If

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v)dudv,$$

where

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1, \quad f_{X,Y}(x, y) \geq 0,$$

then (X, Y) is a continuous bivariate random variable. $f_{X,Y}(x, y)$ is called the **joint probability density** for (X, Y). The main explicit case of a continuous bivariate (X, Y) to be treated in the sequel is the bivariate Gaussian in chapter 8. The **marginal distribution function** for Y is

$$F_Y(y) = F_{X,Y}(\infty, y) = \int_{-\infty}^{y} \int_{-\infty}^{\infty} f_{X,Y}(x, v)dxdv$$

and

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx$$

is the **marginal probability density** for Y. Then, of course, the **marginal distribution function** for X is

$$F_X(x) = F_{X,Y}(x, \infty) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(u, y) dy du.$$

and

$$f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$$

is the marginal probability density for X. It follows in view of (1.12) that X and Y are independent random variables, if and only if

$$f_{X,Y}(x, y) = f_X(x) f_Y(y), \text{ for all } (x, y). \tag{2.40}$$

We have even the bivariate version of the **law of the unconscious statistician** for an integrable Borel function $H(x, y)$ as

$$E[H(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x, y) f_{X,Y}(x, y) dy dy. \tag{2.41}$$

This is in the first place applied to $H(x, y) = x \cdot y$, i.e., to computing covariances, which are defined or recalled next. The **covariance** $\text{Cov}(X, Y)$ of the r.v.'s X and Y is

$$\text{Cov}(X, Y) \overset{\text{def}}{=} E[(X - E[X])(Y - E[Y])] \tag{2.42}$$

Here $E[X]$ and $E[Y]$ are computed as in (2.3) using the respective marginal p.d.f.'s. It follows by properties of integrals that

$$\text{Cov}(X, Y) = E[(X \cdot Y)] - E[X] \cdot E[Y]. \tag{2.43}$$

In view of (1.37) it follows that

$$X \text{ and } Y \text{ are independent } \Rightarrow \text{Cov}(X, Y) = 0. \tag{2.44}$$

The converse implication is not true in general, as shown in the next example.

Example 2.2.24 Let $X \in N(0, 1)$ and set $Y = X^2$. Then Y is clearly functionally dependent on X. But we have

$$\text{Cov}(X, Y) = E[(X \cdot Y)] - E[X] \cdot E[Y] = E\left[X^3\right] - 0 \cdot E[Y] = E\left[X^3\right] = 0.$$

The last equality holds, since with (2.15) one has $g(x) = x^3\phi(x)$, so that $g(-x) = -g(x)$. Hence $E\left[X^3\right] = \int_{-\infty}^{+\infty} g(x)dx = 0$, c.f., (4.50) in the sequel, too.

■

It will be shown in chapter 8 that the converse implication holds for bivariate Gaussian (X, Y).

We standardize covariance[8] to get the **coefficient of correlation** between X and Y

$$\rho_{X,Y} \overset{\text{def}}{=} \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X]} \cdot \sqrt{\text{Var}[Y]}}. \tag{2.45}$$

It is shown in an exercise to this chapter that

$$|\rho_{X,Y}| \leq 1. \tag{2.46}$$

The cases $\rho_{X,Y} = \pm 1$ correspond to Y and X being affine functions (e.g., $Y = aX + b$) of each other, the topic of another exercise.

[8] in order to measure dependence in the common unit of the variables.

2.2.3 Mean, Variance and Covariance of Linear Combinations of R.V.s

The following are important tools of computation, especially in the chapters on stochastic processes. The proofs are omitted or left for self study.

$$E[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i E[X_i] \tag{2.47}$$

$$\begin{aligned} \mathrm{Var}[\sum_{i=1}^{n} a_i X_i] &= \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j \mathrm{Cov}(X_i, X_j), \\ \mathrm{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j X_j\right) &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \mathrm{Cov}(X_i, X_j). \end{aligned} \tag{2.48}$$

These expressions are valid for both continuous and discrete distributions.

2.3 Discrete Distributions

2.3.1 Univariate

Example 2.3.1 (Bernoulli Distribution) Let $X \in \mathrm{Be}\,(p)$. X has two values, usually numerically coded as 0 and 1. The p.m.f. is

$$p_X(x) = \begin{cases} p & x = 1 \\ & \\ q = 1 - p & x = 0. \end{cases} \tag{2.49}$$

$$E\,[X] = p, \mathrm{Var}\,[X] = p(1-p).$$

■

Example 2.3.2 (Symmetric Bernoulli Distribution) We say that $X \in \mathrm{SymBe}$, if the p.m.f. is

$$p_X(x) = \begin{cases} \frac{1}{2} & x = -1 \\ & \\ \frac{1}{2} & x = 1. \end{cases} \tag{2.50}$$

Then

$$E\,[X] = 0, \mathrm{Var}\,[X] = 1.$$

■

Example 2.3.3 (Discrete Uniform Distribution) $X \in U\,(1, 2, \ldots, n)$, where $n > 1$. The p.m.f. is

$$p_X(x) = \begin{cases} \frac{1}{n} & x = 1, 2, \ldots, n \\ & \\ 0 & \text{else}. \end{cases} \tag{2.51}$$

I.e., we pick an integer in $1, 2, \ldots, n$ at random.

$$E\,[X] = \frac{n+1}{2}, \mathrm{Var}\,[X] = \frac{n^2-1}{12}.$$

■

Example 2.3.4 (Geometric Distribution) $0 < p < 1$, $q = 1 - p$. The p.m.f. of $X \in \mathrm{Ge}(p)$ is

$$p_X(k) = q^k p, \quad k = 0, 1, 2, \ldots$$

Suppose p is the probability of an event occurring in a trial. Consider the trial of tossing a coin. Let us say that the event of interest is 'heads'. We are interested in the probability of the number of independent trials we perform, before we see the event 'heads' occuring for the first time NOT INCLUDING the successful trial. Let X be this random number. Then we write $X \in \mathrm{Ge}(p)$.

$$E[X] = \frac{q}{p}, \mathrm{Var}[X] = \frac{q}{p^2}.$$

■

Example 2.3.5 (First Success Distribution) $X \in \mathrm{Fs}(p)$, $0 < p < 1$, $q = 1 - p$. The p.m.f. is

$$p_X(k) = q^{k-1} p, \quad k = 1, 2, \ldots.$$

Suppose again p is the probability of an event occurring in a trial. Consider the trial of tossing a coin (modelled as an outcome of a r.v. $\in \mathrm{Be}(p)$). Let us say again that the event of interest is 'heads'. We are interested in the probability of the number of independent trials we perform, before we see the event 'heads' occuring for the first time INCLUDING the successful trial. Let X be this random number. Then we write $X \in \mathrm{Fs}(p)$.

$$E[X] = \frac{1}{p}, \mathrm{Var}[X] = \frac{q}{p^2}.$$

Condsider the measurable space $(\Omega = \{(\omega_i)_{i=1}^{\infty} | \omega_i \in \{0, 1\}\}, \mathcal{F})$, c.f., example 1.4.7 above, $\mathcal{F}_o \subset \mathcal{F}$. Then X above is defined as a map on Ω as

$$X(\omega) = \text{'the first trial at which success occurs in } \omega\text{'}.$$

However, if $\omega_0 = (\omega_i = 0)_{i=1}^{\infty}$, i.e. ω_0 is an infinite sequence of digital zeros (all failures, no successes), we have

$$X(\omega_0) = +\infty.$$

The difficulty is that we have defined r.v.'s as measurable maps from $(\Omega, \mathcal{F})$ to the real numbers, and $+\infty$ is not a real number. Hence X is in principle an extended random variable with values in $\{+\infty\} \cup \mathbf{R}$. However, if we are computing with the probability model of an infinite sequence of independent $\mathrm{Be}(p)$ trials, we have $X \in \mathrm{Fs}(p)$. Then we must have

$$\mathbf{P}(X = +\infty) = 0,$$

since $\sum_{k=1}^{\infty} q^{k-1} p = 1$ and $\{X = +\infty\} = \left(\cup_{k=1}^{\infty} \{X = k\}\right)^c$. Therefore we can define $X(\omega_0)$ in any preferred way, since this choice has no impact whatsoever on the computations of probabilities.

■

The literature in probability calculus is not unanimous about the terminology regarding the geometric distribution. It occurs frequently (mostly?) that $\mathrm{Fs}(p)$ in our sense above is called the geometric distribution, see, e.g., [48, p. 61], [55, p. 62].

Example 2.3.6 (Binomial Distribution) $X \in \text{Bin}(n,p)$, $0 \leq p \leq 1$, $q = 1-p$, and the p.m.f. is

$$p_X(k) = \begin{pmatrix} n \\ k \end{pmatrix} p^k q^{n-k}, k = 0, 1, \ldots, n.$$

This is the probability of an event occurring k times in n independent trials.

$$E[X] = np, \text{Var}[X] = nqp.$$

The distribution function of $X \in \text{Bin}(n,p)$ has been expressed[9] as

$$F_X(x) = \sum_{k=0}^{x} p_X(k) = \frac{B_q(n-x, x+1)}{B(n-x, x+1)},$$

where we used the beta function and the incomplete beta function in (2.31) and (2.34), respectively.

■

Example 2.3.7 (Poisson binomial Distribution) $X \in \text{Pobin}(p_1, p_2, \ldots, p_n)$, $0 \leq p_i \leq 1$, $i = 1, 2, \ldots, n$, and the p.m.f. is

$$p_X(k) = \sum_{A \in F_k} \prod_{i \in A} p_i \prod_{j \in A^c} (1 - p_j), \tag{2.52}$$

where F_k is the collection of all subsets of k integers that can be selected from $\{1, 2, 3, ..., n\}$. F_k has $\begin{pmatrix} n \\ k \end{pmatrix}$ elements. Hence it is not feasible to use (2.52) for computation for large n.

■

Example 2.3.8 (Poisson Distribution) $X \in \text{Po}(\lambda)$, $\lambda > 0$, then its p.m.f. is

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots \tag{2.53}$$

$$E[X] = \lambda, \text{Var}[X] = \lambda.$$

We shall in example 6.6.2 below derive the Poisson distribution as an approximation of $\text{Bin}(n,p)$ for large n and small p.

■

Example 2.3.9 (Compound Poisson Distribution) $X \in \text{ComPo}(\lambda, \mu)$, $\lambda > 0, \mu > 0$, then its p.m.f. is

$$p_X(k) = \sum_{r=0}^{\infty} \frac{(r\mu)^k}{k!} e^{-r\mu} \frac{\lambda^r}{r!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots \tag{2.54}$$

This expression is somewhat unwieldy, and the Compound Poisson distribution is more naturally treated by the methods of probability generating functions developed in chapter 5.

$$E[X] = \lambda\mu, \text{Var}[X] = \lambda\mu(1+\mu).$$

The Compound Poisson distribution has many applications, e.g., in particle physics [37, 64] and queuing theory.

[9] p. xv in H.G. Romig: *50–100 Binomial Tables*, John Wiley & Sons, Inc., New York, 1947.

■

Example 2.3.10 (Pascal Distribution) Suppose p is the probability of an event occurring in a trial. Consider the trial of tossing a coin. Let us say that the event is 'heads'. We are interested in the probability of the number of independent trials we perform, before we see the event 'heads' occuring n times INCLUDING the nth success.

> Texts in engineering statistics suggest the Pascal Distribution as a model of, e.g., the number of days a certain machine works before it breaks down for the nth time. Or, a text can insist upon that 'the number of days a certain machine works before it breaks down for the nth time' **is** Pascal distributed. One can remark that ' applications of probability calculus are based on analogies, which are to a certain degree halting' (J.W. Lindeberg, [75, p.120]). One could once again repeat the words 'mind projection fallacy', too.

Let now X be the number of independent trials we perform, before we have seen the event occurring n times. The random variable X has then said to have the Pascal Distribution, $X \in \text{Pascal}(n, p)$, $n = 1, 2, 3, \ldots$, $0 < p < 1$ and $q = 1-p$. Its p.m.f. can be found, using the same kind of reasoning that underlies the Binomial distribution, [101, p.58], as

$$p_X(k) = \mathbf{P}\left(X = k\right) = \left(\begin{array}{c} k-1 \\ n-1 \end{array}\right) p^n q^{k-n}, \quad k = n, n+1, n+2, \ldots \tag{2.55}$$

Note that we must understand $\left(\begin{array}{c} k-1 \\ n-1 \end{array}\right)$ as $= 0$ for $k = 0, 1, 2, \ldots, n-1$. Note also that $\text{Pascal}(1, p) = \text{Fs}(p)$.

$$E\left[X\right] = \frac{n}{p}, \text{Var}\left[X\right] = n\frac{(1-p)}{p^2}.$$

■

Example 2.3.11 (Negative Binomial Distribution) X is said to follow the Negative Binomial distribution, $X \in \text{NBin}(n, p)$, $0 < p < 1$, $q = 1 - p$, if its p.m.f. is

$$p_X(k) = \left(\begin{array}{c} n+k-1 \\ k \end{array}\right) p^n q^k, \quad k = 0, 1, 2, \ldots \tag{2.56}$$

$$E\left[X\right] = n\frac{q}{p}, \text{Var}\left[X\right] = n\frac{q}{p^2}.$$

Observe that $\text{Ge}(p) = \text{NBin}(1, p)$. The p.m.f. in (2.56) can be established using the same kind of reasoning that underlies the Binomial distribution, where one needs the interpretation of the coefficient (1.31) as given in appendix 1.11.

■

There is a fair deal of confusing variation in the literature w.r.t. the terminology in the two examples above. Sometimes the Pascal distribution defined as above and, e.g., in [101], is called the negative binomial distribution. In some textbooks, e.g., in [49], the negative binomial distribution is as above, but in others it is known as the Pascal distibution. The handbook [92] compromises with (2.55) as 'Negative Binomial or Pascal' (!).

A p.m.f. $p_X(k)$ has a **power-law tail**, or is a **power law**, if it holds that

$$p_X(k) = P\left(X = k\right) \sim k^{-\gamma}, \quad \text{as } k \to \infty. \tag{2.57}$$

A p.d.f. can also have a power-law tail defined in an analogous manner.

Remark 2.3.1 The notation $f(x) \sim g(x)$ (at $x = a$) has the following meaning.

$$\lim_{x \to a} \frac{f(x)}{g(x)} = 1. \tag{2.58}$$

This means that the functions grow at the same rate at a. For example, if

$$f(x) = x^2, g(x) = x^2 + x,$$

then

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = \lim_{x \to \infty} \frac{1}{1 + \frac{1}{x}} = 1,$$

but at the same time $g(x) - f(x) = x$.

Example 2.3.12 (Benford's Law) We say that a random variable X follows Benford's Law if it has the p.m.f.

$$p_X(k) = P(X = k) = \log_{10}\left(1 + \frac{1}{k}\right), \quad k = 1, \ldots, 9. \tag{2.59}$$

This law is, by empirical experience, found as the distribution of the first digit in a large material of numbers. Note that this is not the uniform distribution $p(k) = \frac{1}{9}$, for $k = 1, 2 \ldots 9$ that might have been expected. Benford's Law is known to be valid for many sources of numerical data.

■

Example 2.3.13 (Zipf's Law (rank-frequency form)) We count the frequencies of occurrencies of some N events (e.g., Swedish words in today's issue of some Stockholm daily, digital or paper edition). Then we determine the rank k of each event by the frequency of occurrence (the most frequent is number one and so on). Then, if we consider $p_X(k)$ as the frequency of a word of rank k, this is very likely found to be

$$p_X(k) = c \cdot k^{-\gamma}, \tag{2.60}$$

where γ is close to one, and where c is the normalizing constant

$$c = 1/\sum_{k=1}^{N} k^{-\gamma}.$$

The probability mass function in (2.60) is known as **Zipf's law**, and is an empirical or experimental assertion, which seems to arise in many situations, and is not based on any theoretical generative model. The case with $\gamma = 2$ is known as **Zipf-Lotka's Law**, and was discovered as a bibliometric law on the number of authors making k contributions.

■

Example 2.3.14 (Waring distribution) We write $X \in \text{War}(\rho, \alpha)$ and say that X is **Waring** distributed with parameters $\alpha > 0$ and $\rho > 0$, if X has the p.m.f.

$$p_X(k) = \rho \frac{\alpha_{(k)}}{(\alpha + \rho)_{(k+1)}}, \quad k = 0, 1, 2, \ldots \tag{2.61}$$

Here we invoke the ascending factorials

$$\alpha_{(k)} = \alpha \cdot (\alpha + 1) \cdot \ldots \cdot (\alpha + k - 1) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)},$$

and analogously for $(\alpha+\rho)_{(k+1)}$. If $\rho > 1$, then $E[X]$ exists, and if $\rho > 2$, then $\text{Var}[X]$ exists, too. It can be shown that there is the power-law tail

$$p_X(k) \sim \frac{1}{k^{1+\rho}}.$$

We shall in an exercise to chapter 3 derive $\text{War}(\rho,\alpha)$ under the name **Negative-Binomial Beta** distribution.

$$E[X] = n\frac{\alpha}{\rho-1}, \text{Var}[X] = \frac{\alpha}{\rho-1}\left[\frac{\rho+\alpha}{\rho-2} + \frac{\alpha}{(\rho-1)(\rho-2)}\right].$$

This distribution was invented and named by J.O. Irwin in 1963[10]. It has applications, e.g., in accident theory and in the measurement of scientific productivity.

■

Example 2.3.15 (Skellam distribution) We write $X \in \text{Ske}(\mu_1,\mu_2)$ and say that X is **Skellam** distributed with parameters $\mu_1 > 0$ and $\mu_2 > 0$, if X has the p.m.f. for any integer k

$$p_X(k) = e^{-(\mu_1+\mu_2)}\left(\frac{\mu_1}{\mu_2}\right)^{k/2} I_{|k|}(2\sqrt{\mu_1\mu_2}), \tag{2.62}$$

where $I_k(z)$ is the modified Bessel function of the first kind of order k.

$$E[X] = \mu_1 - \mu_2, \text{Var}[X] = \mu_1 + \mu_2.$$

It can be shown that if $X_1 \in \text{Po}(\mu_1)$ and $X_2 \in \text{Po}(\mu_2)$ and X_1 and X_2 are independent, then $X_1 - X_2 \in \text{Ske}(\mu_1,\mu_2)$.

Skellam distribution is applied to the difference of two images with photon noise. It is also been found useful as a model for the point spread distribution in baseball, hockey and soccer, where all scored points are equal.

■

2.3.2 Bivariate Discrete Distributions

(X,Y) is a bivariate random variable and as earlier

$$F_{X,Y}(x,y) = \mathbf{P}(X \leq x, Y \leq y), -\infty < x < \infty, -\infty < y < \infty.$$

If

$$F_{X,Y}(x,y) = \sum_{-\infty < x_j \leq x}\ \sum_{-\infty < y_k \leq y} p_{X,Y}(x_j,y_k),$$

where

$$\sum_{-\infty < x_j < \infty}\ \sum_{-\infty < y_k < \infty} p_{X,Y}(x_j,y_k) = 1, \quad p_{X,Y}(x_j,y_k) \geq 0,$$

then (X,Y) is a discrete bivariate random variable. The function $p_{X,Y}(x_j,y_k)$ is called the joint probability mass function for (X,Y). Marginal distributions are defined by

$$p_X(x_j) = \sum_{-\infty < y_k < \infty} p_{X,Y}(x_j,y_k), p_Y(y_k) = \sum_{-\infty < x_j < \infty} p_{X,Y}(x_j,y_k).$$

[10] J.O. Irwin: The Place of Mathematics in Medical and Biological Sciences. *Journal of the Royal Statistical Society*, Ser. A, 126, 1963, p. 1−14.

The covariance of (X, Y) is again

$$\mathrm{Cov}(X, Y) = E\left[(X \cdot Y)\right] - E\left[X\right] \cdot E\left[Y\right], \tag{2.63}$$

where we know how to compute with the joint p.m.f. and with the marginal p.m.f.'s and the law of the unconscious statistician.

Example 2.3.16 Bivariate Bernoulli distribution Let (X, Y) be a bivariate random variable, where both X and Y are binary, i.e., their values are 0 or 1. Then we say that (X, Y) has a bivariate Bernoulli distribution, if the p.m.f is

$$p_{X,Y}(x, y) = \theta^x (1-\theta)^{1-x} \lambda^y (1-\lambda)^{1-y}, x \in \{0,1\}, y \in \{0,1\}. \tag{2.64}$$

Here $0 \leq \theta \leq 1$, $0 \leq \lambda \leq 1$.

■

2.4 Transformations of Continuous Distributions

2.4.1 The Probability Density of a Function of Random Variable

Let X be a continuous random variable with the p.d.f. $f_X(x)$. Assume that $H(x)$ is strictly monotonous (= either strictly increasing or strictly decreasing). The p.d.f. of $Y = H(X)$ is ascertained as

$$f_Y(y) = f_X\left(H^{-1}(y)\right) \cdot \mid \frac{d}{dy} H^{-1}(y) \mid. \tag{2.65}$$

Here $H^{-1}(y)$ is the inverse of $H(x)$. In case the domain of definition of the function $H(x)$ can be decomposed into disjoint intervals, where $H(x)$ is strictly monotonous, we have

$$f_Y(y) = \sum_i f_X\left(H_i^{-1}(y)\right) \cdot \mid \frac{d}{dy} H_i^{-1}(y) \mid \chi_{I_i}(y), \tag{2.66}$$

where H_i indicates the function H restricted to the respective domain I_i of strict monotonicity, and χ_{I_i} is the corresponding indicator function.

Example 2.4.1 Let $X \in U\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Set $Y = \sin(X)$. We want to find the p.d.f. $f_Y(y)$. When we recall the graph of $\sin(x)$, we observe that $\sin(x)$ is strictly increasing on $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Then for $-1 \leq y \leq 1$ we have

$$F_Y(y) = \mathbf{P}\left(Y \leq y\right) = \mathbf{P}\left(X \leq \arcsin(y)\right),$$

since $\arcsin(y)$ is the inverse of $\sin(x)$ for $x \in]-\pi/2, \pi/2[$. As $X \in U\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ we have

$$F_Y(y) = \frac{\arcsin(y) - (-\pi/2)}{\pi}, \quad -1 \leq y \leq 1. \tag{2.67}$$

Then

$$f_Y(y) = \frac{1}{\pi\sqrt{1-y^2}}, -1 < y < 1. \tag{2.68}$$

Example 2.4.2 Let $X \in U\left(0, 2\pi\right)$ and $Y = \sin(X)$. We want again to determine the p.d.f. $f_Y(y)$. The function $H(x) = \sin(x)$ is not strictly monotonous in $(0, 2\pi)$, hence we shall find the the p.d.f. $f_Y(y)$ by means of (2.66).

We make the decomposition $(0, 2\pi) = I_1 \cup I_2 \cup I_3$, where $I_1 = (0, \pi/2)$, $I_2 = (\pi/2, 3\pi/2)$ and $I_3 = (3\pi/2, 2\pi)$. Then for $i = 1, 2, 3$, $H_i(x) = \sin(x) \mid I_i$, i.e., the function $\sin(x)$ restricted to H_i, is strictly monotonous. In fact,

$$H_1(x) = \sin(x) \quad 0 \le x \le \frac{\pi}{2},$$

$$H_2(x) = H_2(x - \pi/2), \frac{\pi}{2} \le x \le \frac{3\pi}{2} \Leftrightarrow H_2(t) = \cos(t), \quad 0 \le t \le \pi$$

$$H_3(x) = H_3(x - 2\pi), \frac{\pi}{2} \le x \le \frac{3\pi}{2} \Leftrightarrow H_3(t) = \sin(t), \quad -\frac{\pi}{2} \le t \le 0.$$

Then we have two cases (i)-(ii) to consider:

(i) $0 \le y < 1$. Then (draw a picture)

$$F_Y(y) = \mathbf{P}(Y \le y) = \mathbf{P}\left(0 \le X \le H_1^{-1}(y)\right) + \mathbf{P}\left(H_2^{-1}(y) \le X \le 3\pi/2\right) + \mathbf{P}(3\pi/2 \le X \le 2\pi)$$

$$= \mathbf{P}(0 \le X \le \arcsin(y)) + \mathbf{P}(\arccos(y) \le X \le 3\pi/2) + \frac{1}{4}$$

$$= \frac{\arcsin(y)}{2\pi} + \frac{3\pi/2 - \arccos(y)}{2\pi} + \frac{1}{4}.$$

Then

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{2\pi\sqrt{1-y^2}} - \left(\frac{-1}{2\pi\sqrt{1-y^2}}\right)$$

$$= \frac{1}{\pi\sqrt{1-y^2}}, \quad 0 \le y < 1.$$

(ii) $-1 < y < 0$. Then (draw a picture)

$$F_Y(y) = \mathbf{P}(Y \le y) = \mathbf{P}\left(H_2^{-1}(y) \le X \le 3\pi/2\right) + \mathbf{P}\left(3\pi/2 \le X \le H_3^{-1}(y)\right)$$

$$= \mathbf{P}(\arccos(y) \le X \le 3\pi/2) + \mathbf{P}(3\pi/2 \le X \le \arcsin(y)).$$

$$= \frac{3\pi/2 - \arccos(y)}{2\pi} + \frac{\arcsin(y) - 3\pi/2}{2\pi}.$$

Thus we get again

$$f_Y(y) = \frac{1}{\pi\sqrt{1-y^2}}, \quad -1 < y < 0.$$

In summary, it was found that

$$f_Y(y) = \frac{1}{\pi\sqrt{1-y^2}}, \quad -1 < y < 1.$$

The p.d.f. derived above appears in the introduction to stochastic processess in chapter 9.

The p.d.f.s $f_Y(y)$ derived in examples 2.4.1 and 2.4.2 are identical. Hence, if we were given a sample set of I.I.D. outcomes of $Y = \sin(X)$ for $X \in U\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ or of $Y = \sin(X)$ for $X \in U(0, 2\pi)$, we would have no statistical way of telling from which of the mentioned sources the observations emanate.

■

2.4.2 Change of Variable in a Joint Probability Density

This section consists, for practical purposes, of one single formula, (2.69), and applications of it. The formula follows by the rule for change of variable in multiple integrals. A very elaborate and detailed proof is constructed in [82, pp. 148−168].

Let $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ have the joint p.d.f. $f_{\mathbf{X}}(x_1, x_2, \ldots, x_m)$. Define a new random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$ by

$$Y_i = g_i(X_1, \ldots, X_m), \quad i = 1, 2, \ldots, m,$$

where g_i are continuously differentiable and $(g_1, g_2, \ldots, g_m)$ is invertible (in a domain) with the inverse

$$X_i = h_i(Y_1, \ldots, Y_m), \quad i = 1, 2, \ldots, m,$$

where h_i are continuously differentiable. Then the joint p.d.f. of $\mathbf{Y}$ is (in the domain of invertibility)

$$f_{\mathbf{Y}}(y_1, \ldots, y_m) = f_{\mathbf{X}}(h_1(y_1, y_2, \ldots, y_m), \ldots, h_m(y_1, y_2, \ldots, y_m)) \mid J \mid, \tag{2.69}$$

where J is the Jacobian determinant

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \ldots & \frac{\partial x_1}{\partial y_m} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \ldots & \frac{\partial x_2}{\partial y_m} \\ \vdots & \vdots & \ldots & \vdots \\ \frac{\partial x_m}{\partial y_1} & \frac{\partial x_m}{\partial y_2} & \ldots & \frac{\partial x_m}{\partial y_m} \end{vmatrix}. \tag{2.70}$$

The main point of the proof in loc.cit. may perhaps be said to be the approximation of the domain of invertibility of $(g_1, g_2, \ldots, g_m)$ by intervals I_k in $\mathbf{R}^m$ with volume $V(I_k)$, and then to show that these intervals are mapped by $(g_1, g_2, \ldots, g_m)$ to parallelepipeds P_k with volume $V(P_k)$. The **volume change** incurred by this mapping is then shown to be

$$V(P_k) = \mid J \mid V(I_k).$$

Example 2.4.3 $\mathbf{X}$ has the probability density $f_{\mathbf{X}}(\mathbf{x})$, $Y = A\mathbf{X} + \mathbf{b}$, and A is $m \times m$ and invertible. In this case one finds that the Jacobian is $J = \det(A^{-1})$ and by general properties of determinants $\det(A^{-1}) = \frac{1}{\det A}$. Then $\mathbf{Y}$ has in view of (2.69) the p.d.f.

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\mid \det A \mid} f_{\mathbf{X}}\left(A^{-1}(\mathbf{y} - \mathbf{b})\right). \tag{2.71}$$

■

Example 2.4.4 (Ratio of two random variables) Let X and Y be two independent continuous r.v.'s with p.d.f.'s $f_X(x)$ and $f_Y(y)$, respectively. We are interested in the distribution of $\frac{X}{Y}$. We shall apply (2.69) by the following transformation

$$U = \frac{X}{Y}, \quad V = Y.$$

This is one typical example of the application of the change of variable formula. We are in fact interested in a single r.v., here U, but in to order find its distribution we need an **auxiliary variable**, here V, to use the terminology of [34, p. 68]. Then we determine the joint p.d.f., here $f_{U,V}(u, v)$, and marginalize to U to find the desired p.d.f..

The inverse map is found as

$$X = h_1(U,V) = UV, \quad Y = h_2(U,V) = V.$$

Then the Jacobian is by (2.70)

$$J = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v.$$

By (2.69) and our assumption we get

$$f_{U,V}(u,v) = f_X(uv) f_Y(v) |v|. \tag{2.72}$$

Hence the distribution of the ratio $U = \frac{X}{Y}$ is given by the marginal density

$$f_U(u) = \int_{-\infty}^{\infty} f_{U,V}(u,v)\, dv.$$

In [64, p. 237] this is written as

$$f_U(u) = \int_0^{\infty} f_X(uv) f_Y(v)\, v dv - \int_{-\infty}^{0} f_X(uv) f_Y(v)\, v dv. \tag{2.73}$$

■

Example 2.4.5 (Bivariate Logistic Normal Distribution) (From the exam in sf2940 2010-01-12) X_1,X_2 are two independent standard normal random variables. We introduce two new random variables by

$$\begin{pmatrix} Y_1 \\ \\ Y_2 \end{pmatrix} = \begin{pmatrix} \frac{e^{X_1}}{1+e^{X_1}+e^{X_2}} \\ \\ \frac{e^{X_2}}{1+e^{X_1}+e^{X_2}} \end{pmatrix}.$$

We wish to find the probability density of (Y_1, Y_2). We write first, for clarity of thought,

$$\begin{pmatrix} Y_1 \\ \\ Y_2 \end{pmatrix} = \begin{pmatrix} g_1(X_1,X_2) \\ \\ g_2(X_1,X_2) \end{pmatrix} = \begin{pmatrix} \frac{e^{X_1}}{1+e^{X_1}+e^{X_2}} \\ \\ \frac{e^{X_2}}{1+e^{X_1}+e^{X_2}} \end{pmatrix}.$$

Then we solve to get

$$X_1 = h_1(Y_1,Y_2) = \ln Y_1 + \ln\left(1 + e^{X_1} + e^{X_2}\right) = \ln Y_1 - \ln(1 - (Y_1 + Y_2))$$

and similarly

$$X_2 = h_2(Y_1,Y_2) = \ln Y_2 - \ln(1 - (Y_1 + Y_2)).$$

Then we find the Jacobian, or

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}.$$

Entry by entry we get

$$\frac{\partial x_1}{\partial y_1} = \frac{1}{y_1} + \frac{1}{1-(y_1+y_2)}$$

$$\frac{\partial x_1}{\partial y_2} = \frac{1}{1-(y_1+y_2)}$$

$$\frac{\partial x_2}{\partial y_1} = \frac{1}{1-(y_1+y_2)}$$

$$\frac{\partial x_2}{\partial y_2} = \frac{1}{y_2} + \frac{1}{1-(y_1+y_2)}.$$

Thus, the Jacobian determinant is

$$\begin{aligned}
J &= \frac{\partial x_1}{\partial y_1} \cdot \frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2} \cdot \frac{\partial x_2}{\partial y_1} \\
&= \frac{1}{Y_1}\left(\frac{1}{y_2} + \frac{1}{1-(y_1+y_2)}\right) + \frac{1}{1-(y_1+y_2)}\left(\frac{1}{y_2} + \frac{1}{1-(y_1+y_2)}\right) \\
&\quad - \left(\frac{1}{1-(y_1+y_2)}\right)^2 \\
&= \frac{1}{y_1}\frac{1}{y_2} + \frac{1}{y_1}\frac{1}{1-(y_1+y_2)} + \frac{1}{y_2}\frac{1}{1-(y_1+y_2)} + \left(\frac{1}{1-(y_1+y_2)}\right)^2 - \left(\frac{1}{1-(y_1+y_2)}\right)^2 \\
&= \frac{1}{y_1}\frac{1}{y_2} + \frac{1}{y_1}\frac{1}{1-(y_1+y_2)} + \frac{1}{y_2}\frac{1}{1-(y_1+y_2)} \\
&= \frac{1}{y_1}\frac{1}{y_2} + \frac{1}{1-(y_1+y_2)}\left(\frac{1}{y_1} + \frac{1}{y_2}\right) \\
&= \frac{1}{y_1}\frac{1}{y_2} + \frac{1}{1-(y_1+y_2)}\left(\frac{y_1+y_2}{y_1 y_2}\right) \\
&= \frac{1-(y_1+y_2)+y_1+y_2}{y_1 y_2 (1-(y_1+y_2))} \\
&= \frac{1}{y_1 y_2 (1-(y_1+y_2))}.
\end{aligned}$$

Let us note that by construction $J > 0$. Thus we get by (2.69) that

$$\begin{aligned}
f_{\mathbf{Y}}(y_1, y_2) &= f_{X_1}(h_1(y_1,y_2)) f_{X_2}(h_2(y_1,y_2)) \mid J \mid \\
&= \frac{1}{y_1 y_2 (1-(y_1+y_2))} \phi(\ln y_1 - \ln(1-(y_1+y_2))) \phi(\ln y_2 - \ln(1-(y_1+y_2))).
\end{aligned}$$

This is a case of the bivariate logistic normal distribution. Since $0 < y_1 < 1$ and $0 < y_2 < 1$, $0 < y_1 + y_2 < 1$ and with $y_3 = 1 - (y_1 + y_2)$, the bivariate logistic normal distribution can be taken, e.g., as a prior density on the probability simplex $\{(y_1, y_2, y_3) \mid 0 < y_i < 1, i = 1, 2, 3; 1 = y_1 + y_2 + y_3\}$.

■

Example 2.4.6 Exponential Order Statistics Let $X_1, \ldots, X_n$ be I.I.D. random variables with a continuous distribution. The **order statistic** of $X_1, \ldots, X_n$ is the ordered sample:

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)},$$

Here

$$\begin{aligned}
X_{(1)} &= \min(X_1, \ldots, X_n) \\
X_{(n)} &= \max(X_1, \ldots, X_n)
\end{aligned}$$

and

$$X_{(k)} = \ k\text{th smallest of } X_1, \ldots, X_n \ .$$

The variable $X_{(k)}$ is called the kth order variable. The following theorem has been proved in, e.g., [49, section 4.3., theorem 3.1.].

Theorem 2.4.7 Assume that $X_1, \ldots, X_n$ are I.I.D. random variables with the p.d.f. $f(x)$. The joint p.d.f. of the order statistic is

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}\left(y_1,\ldots,y_n\right) = \begin{cases} n!\prod_{k=1}^{n} f\left(y_k\right) & \text{if } y_1 < y_2 < \ldots < y_n, \\ 0 & \text{elsewhere.} \end{cases} \tag{2.74}$$

■

Let $X_1, \ldots, X_n$ be I.I.D. random variables with the distribution Exp(1). We are interested in the differences of the order variables

$$X_{(1)}, X_{(i)} - X_{(i-1)}, \quad i = 2, \ldots, n.$$

Note that we may consider $X_{(1)} = X_{(1)} - X_{(0)}$, if $X_{(0)} = 0$. The result of interest in this section is the following theorem.

Theorem 2.4.8 Assume that $X_1, \ldots, X_n$ are I.I.D. random variables $X_i \in \text{Exp}(1)$, $i = 1, 2, \ldots, n$. Then

(a)

$$X_{(1)} \in \text{Exp}\left(\frac{1}{n}\right), X_{(i)} - X_{(i-1)} \in \text{Exp}\left(\frac{1}{n+1-i}\right),$$

(b) $X_{(1)}$, $X_{(i)} - X_{(i-1)}$ for $i = 2, \ldots, n$, are n independent random variables.

Proof: We define Y_i for $i = 1, \ldots, n$ by

$$Y_1 = X_{(1)}, \quad Y_i = X_{(i)} - X_{(i-1)}.$$

Then we introduce

$$A = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{pmatrix}. \tag{2.75}$$

so that if

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} X_{(1)} \\ X_{(2)} \\ X_{(3)} \\ \vdots \\ X_{(n)} \end{pmatrix},$$

we have

$$\mathbf{Y} = A\mathbf{X}.$$

It is clear that the inverse matrix A^{-1} exists, because we can uniquely find $\mathbf{X}$ from $\mathbf{Y}$ by

$$X_{(1)} = Y_1, \quad X_{(i)} = Y_i + Y_{i-1} + \ldots + Y_1.$$

We write these last mentioned equalities in matrix form by

$$\mathbf{X} = A^{-1}\mathbf{Y}.$$

Then we have by (2.71)

$$f_{\mathbf{Y}}\left(\mathbf{y}\right) = f_{\mathbf{X}}\left(A^{-1}\mathbf{y}\right)\frac{1}{|\det A|}. \tag{2.76}$$

But now we evoke (2.74) to get

$$f_{\mathbf{X}}\left(A^{-1}\mathbf{y}\right) = n! f\left(y_1\right) f\left(y_1 + y_2\right) \cdots f\left(y_1 + y_2 + \ldots + y_n\right), \tag{2.77}$$

since $y_1 < y_1 + y_2 < \ldots < y_1 + y_2 + \ldots + y_n$. As $f(x) = e^{-x}$, we get

$$f\left(y_1\right) f\left(y_1 + y_2\right) \cdots f\left(y_1 + y_2 + \ldots + y_n\right) = e^{-y_1} e^{-(y_1+y_2)} \ldots e^{-(y_1+y_2+\ldots+y_n)}$$

and rearrange and use $y_1 = x_{(1)}$ and $y_i = x_{(i)} - x_{(i-1)}$,

$$= e^{-ny_1} e^{-(n-1)y_2} \ldots e^{-2y_{n-1}} e^{-y_n}$$

$$= e^{-nx_{(1)}} e^{-(n-1)(x_{(2)}-x_{(1)})} \ldots e^{-(x_{(n)}-x_{(n-1)})}.$$

Hence, if we insert the last result in (2.76) and distribute the factors in $n! = n(n-1)\cdots 3\cdot 2\cdot 1$ into the product of exponentials we get

$$f_{\mathbf{Y}}\left(\mathbf{y}\right) = ne^{-nx_{(1)}}(n-1)e^{-(n-1)(x_{(2)}-x_{(1)})} \ldots e^{-(x_{(n)}-x_{(n-1)})} \frac{1}{|\det A|}. \tag{2.78}$$

Since A in (2.75) is a triangular matrix, its determinant equals the product of its diagonal terms, c.f. [92, p. 93]. Hence from (2.75) we get $\det A = 1$. In other words, we have obtained

$$\begin{aligned} & f_{X_{(1)},X_{(2)}-X_{(1)},\ldots,X_{(n)}-X_{(n-1)}}\left(x_{(1)}, x_{(2)} - x_{(1)}, \ldots, x_{(n)} - x_{(n)}\right) \\ &= ne^{-nx_{(1)}}(n-1)e^{-(n-1)(x_{(2)}-x_{(1)})} \ldots 2e^{-2(x_{(n-1)}-x_{(n-2)})} e^{-(x_{(n)}-x_{(n-1)})}. \end{aligned} \tag{2.79}$$

But, when we check against (2.23), $(n-1)e^{-(n-1)(x_{(2)}-x_{(1)})}$ is nothing but the p.d.f. of $\text{Exp}\left(\frac{1}{n-1}\right)$, and so on, the generic factor in the product in (2.79) being $(n+1-i)e^{-(n+1-i)(x_{(i)}-x_{(i-1)})}$, which is the p.d.f. of $\text{Exp}\left(\frac{1}{n+1-i}\right)$.

Hence we have that the product in (2.79) is a product of the respective p.d.f.'s for the variables $X_{(1)} \in \text{Exp}\left(\frac{1}{n}\right)$ and for $X_{(i)} - X_{(i-1)} \in \text{Exp}\left(\frac{1}{n+1-i}\right)$. Thus we have established the cases (a) and (b) in the theorem as claimed. ■

There exists a more intuitively appealing way of realizing the fact above. First one shows that

$$X_{(1)} = \min\left(X_1, \ldots, X_n\right) \in \text{Exp}\left(\frac{1}{n}\right)$$

(which is also seen above), if $X_1, \ldots, X_n$ are I.I.D. random variables under Exp(1). Then one can argue by independence and the **lack of memory** of the exponential distribution that $X_{(i)} - X_{(i-1)}$ is the minimum of lifetimes of $n+1-i$ independent Exp(1) -distributed random variables.

■

2.5 Appendix: Decompositions of Probability Measures on the Real Line

2.5.1 Introduction

In this appendix we shall make a specialized investigation of probability measures on the real line and the Borel σ field, $(\mathbf{R}, \mathcal{B})$. The goal is to give a brief account of that these measures can be additively decomposed into a sum of an absolutely continuous part, a discrete part and a singular part, in the sense to be made clear in the sequel. Then we check how such dispositions are related to continuous and discrete r.v.'s. We mention here the lecture notes [87], not because these are the authentic source with priority on the results to be discussed, but because we shall follow the good detail of presentation as loc.cit..

We start by a theorem that shows that a probability measure on the real line and its Borel sets can be '*induced*' (in the sense given in the proof below) by a random variable.

Theorem 2.5.1 If F satisfies 1., 2. and 3. in theorem 1.5.6, then there is a unique probability measure μ on $(\mathbf{R}, \mathcal{B})$ such that $\mu((a,b]) = F(b) - F(a)$ for all a, b.

Proof: The sets $(a,b]$ lie in the Borel σ field. The theorem 1.5.7 gives the existence of a random variable X with distribution F. Consider the measure this X induces on $(\mathbf{R}, \mathcal{B})$, which means that for any $A \in \mathcal{B}$ we define

$$\mu^X (A) \overset{\text{def}}{=} \mathbf{P}(X \in A). \tag{2.80}$$

Then, of course, $\mu((a,b]) \overset{\text{def}}{=} \mu^X((a,b]) = F(b) - F(a)$. The uniqueness follows because the sets (a,b) generate the σ field and we can hence apply theorem 1.4.1. ■

We shall return the result in the theorem above. But we continue first by considering a generic probability measure μ on $(\mathbf{R}, \mathcal{B})$.

2.5.2 Decompositions of μ on $(\mathbf{R}, \mathcal{B})$

We have found in example 1.4.10 that the singleton sets $\{x\} \in \mathcal{B}$. Then we can define the **probability mass function** of μ as

$$p(x) \overset{def}{=} \mu(\{x\}), \quad -\infty < x < \infty. \tag{2.81}$$

It is clear that $p(x) \geq 0$.

Lemma 2.5.2 $p(x) > 0$ only for countably many x.

Proof: Set $B_n = \{x | p(x) \geq \frac{1}{n}\}$. Let a_n= number of points in B_n (=cardinality of B_n). Then

$$1 \geq \mu(B_n) \geq a_n \frac{1}{n},$$

or, $a_n \leq n$. Thus B_n is a set with a finite number of elements. Next we observe that

$$\{x | p(x) > 0\} = B_1 \cup B_2 \cup B_3 \cup \ldots$$

This shows that $\{x | p(x) > 0\}$ is a countable union of finite sets, and such a union is a countable set. ■

The singletons $\{x\}$ such that $p(x) > 0$ are also called **atoms** of μ. In view of this lemma we may define a **discrete part** or **pure point mass part of** μ as measure on $(\mathbf{R}, \mathcal{B})$ by the countable sum

$$\mu_D(A) = \sum_{x \in A | p(x) > 0} p(x), \quad A \in \mathcal{B}.$$

We say that a probability measure μ on $(\mathbf{R}, \mathcal{B},)$ is **continuous**, if its pure point mass measure is identically zero, or

$$p(x) = \mu(\{x\}) = 0, \quad \text{for all } x \in \mathbf{R}$$

Then we define for any $A \in \mathcal{B}$ the measure

$$\mu_C(A) \overset{\text{def}}{=} \mu(A) - \mu_D(A).$$

Note that it must hold $\mu_C(A) \geq 0$ for a measure, so we must and can check that $\mu_C(A)$ is a measure. Clearly, μ_C is a continuous measure.

Theorem 2.5.3 Every measure μ on $(\mathbf{R}, \mathcal{B})$ can be expressed uniquely with an additive deomposition to its continuous part and its discrete part by

$$\mu = \mu_C + \mu_D. \tag{2.82}$$

■

By (2.82) we mean $\mu(A) = \mu_C(A) + \mu_D(A)$ for any $A \in \mathcal{B}$.

We shall next proceed by decomposing additively the continuous part. We need a new definition. A measure on μ on $(\mathbf{R}, \mathcal{B})$ is called **absolutely continuous with the density** $f(x)$, if it holds for every interval $I \subset \mathbf{R}$ that

$$\mu(I) = \int_I f(x)dx, \quad f(x) \geq 0. \tag{2.83}$$

Then it follows that an absolutely continuous measure is a continuous measure. This is plausible, since

$$\mu\left(\{x\}\right) \leq \mu\left([x-h, x+h]\right) = \int_{x-h}^{x+h} f(x)dx \to 0,$$

as $h \to 0$.

Theorem 2.5.4 For every probability measure μ on $(\mathbf{R}, \mathcal{B})$ with density $f(x)$ it holds for almost all x that

$$\lim_{h \to 0} \frac{1}{2h} \mu\left(\{x-h, x+h\}\right) = f(x). \tag{2.84}$$

■

It can be shown that $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$. Conversely, any function with the last two properties defines an absolutely continuous measure with density $f(x)$.

Theorem 2.5.5 For every probability measure μ on $(\mathbf{R}, \mathcal{B})$ it holds for almost all x that

$$\lim_{h \to 0} \frac{1}{2h} \mu\left(\{x-h, x+h\}\right) = g(x). \tag{2.85}$$

Proof: is omitted. ■

Let μ be a probability measure on $(\mathbf{R}, \mathcal{B})$ and the corresponding $g(x)$ be defined as in (2.85). By the **absolutely continuous part** of μ we mean the absolutely continuous measure μ_A with density the $g(x)$. It can be shown that $g(x) \geq 0$ and $\int_{-\infty}^{+\infty} g(x)dx = 1$.

Theorem 2.5.6 Let μ be a continuous measure on $(\mathbf{R}, \mathcal{B})$. Let

$$\mu_S = \mu - \mu_A. \tag{2.86}$$

Then μ_S is a continuous measure that satisfies for almost all x

$$\lim_{h \to 0} \frac{1}{2h} \mu_S\left(\{x-h, x+h\}\right) = 0.$$

■

The proof is omitted. The measure μ_S is called the **singular part** of μ. There are trivial examples of singular measures, like the one that assigns measure zero to every Borel set. One can describe a **purely singular measure** μ as follows:

- μ_S is a continuous measure.

- The whole mass of μ_S is on a set with zero (Lebesgue) measure.

By theorem 2.5.3 we have for any probability measure on $(\mathbf{R}, \mathcal{B})$ that $\mu = \mu_C + \mu_D$. By theorem 2.5.6 we have $\mu_C = \mu_A + \mu_S$ for any continuous measure μ_C on $(\mathbf{R}, \mathcal{B})$. This we summarise in the theorem below.

Theorem 2.5.7 Every probability measure μ on $(\mathbf{R}, \mathcal{B})$ can be expressed uniquely with an additive deomposition to its absolutely continuous part, its singular part and its discrete part

$$\mu = \mu_A + \mu_S + \mu_D. \tag{2.87}$$

■

Now we start a journey backwards to the familiar notions in the bulk of this chapter.

2.5.3 Continuous, Discrete and Singular Random Variables

Let μ be probability measure on $(\mathbf{R}, \mathcal{B})$. The **distribution function of** μ is $F_\mu(x)$ defined by

$$F_\mu(x) = \mu\left(] - \infty, x]\right), \quad -\infty < x < \infty. \tag{2.88}$$

The measure μ is uniquely determined by $F_\mu(x)$. It follows that $F_\mu(x)$ satisfies theorem 1.5.6 and in 5. of theorem 1.5.6 and we find

$$p(x) = F_\mu(x) - F_\mu(x-),$$

where $p(x)$ is the point mass function of μ as defined in (2.81). Indeed, by continuity from above of the probability measure μ we get

$$F_\mu(x-) = \lim_{h \to 0} F_\mu(x - h) = \lim_{h \to 0} \mu\left(] - \infty, x + h]\right) = \mu\left(] - \infty, x)\right)$$

$$= \mu\left(] - \infty, x] \setminus \{x\}\right) = \mu\left(] - \infty, x]\right) - \mu\left(\{x\}\right) = F_\mu(x) - p(x).$$

Let μ be an absolutely continuous measure with the density $f_\mu(x)$. Then in view of (2.83)

$$F_\mu(x) = \int_{-\infty}^{x} f_\mu(u) du, \quad -\infty < x < \infty.$$

Then it can be shown for almost all x that

$$\frac{d}{dx} F_\mu(x) = f_\mu(x).$$

In fact we have also the following theorem with a difficult proof, duely omitted.

Theorem 2.5.8 Let μ be a probability measure on $(\mathbf{R}, \mathcal{B})$ and let $F_\mu(x)$ be its distribution function. Then $\frac{d}{dx} F_\mu(x)$ exists for almost all x and

$$\frac{d}{dx} F_\mu(x) = g(x), \tag{2.89}$$

where $g(x)$ is given in (2.85). In addition, the absolutely continuous part μ_A of μ is the probability measure given by the density $g(x)$.

■

As a consequence of the theorem above we can describe the distribution function $F_\mu(x)$ of a singular measure μ by

(a) $F_\mu(x)$ is continuous.

(b) $\frac{d}{dx} F_\mu(x) = 0$ for almost all x.

Finally,

Theorem 2.5.9 Let μ be a probability measure on $(\mathbf{R}, \mathcal{B})$ and let $F_\mu(x)$ be its distribution function. Then μ lacks a purely singular part μ_S, if $F_\mu(x) = \int_{-\infty}^{x} \frac{d}{dx} F_\mu(u) du$ except for a countable number of points.

■

This preceding narrative has amounted to the following. Let as in the proof of theorem 2.5.1 X be a random variable. The probability measure μ^X, which X induces on $(\mathbf{R}, \mathcal{B})$ is

$$\mu^X (A) \overset{\text{def}}{=} \mathbf{P}(X \in A).$$

Then by (2.87),

$$\mu^X = \mu_A^X + \mu_S^X + \mu_D^X, \tag{2.90}$$

where for any $B \in \mathcal{B}$

$$\mu_A^X(B) = \int_B \frac{d}{dx} F_{\mu^X}(x)dx, \quad \mu_D^X(B) = \sum_{x \in B | p(x) > 0} p(x).$$

If the parts μ_S^X and μ_D^X are missing in (2.90), we have what has been in the preceding called X a **continuous r.v.**. If μ_C^X and μ_S^X are missing in (2.90), we have what has been in the preceding called X a **discrete r.v.**. In addition, if μ_S^X is missing in (2.90), we could call X a **mixed r.v.**, and such r.v.'s are not much in evidence in these notes and other texts at the same level [11]. If μ_C^X and μ_D^X are missing in (2.90), the random variable X is called **singular**. The most famous example of a singular r.v. is the r.v. with a **Cantor distribution**.

2.6 Exercises

2.6.1 Distribution Functions

1. A stochastic variable X is said to follow the **two-parameter Birnbaum-Saunders** distribution, we write $X \in BS(\alpha, \beta)$, if its distribution function is

$$F_X(x) = \begin{cases} \Phi\left(\frac{1}{\alpha}\left(\sqrt{\frac{x}{\beta}} - \sqrt{\frac{\beta}{x}}\right)\right) & \text{if } 0 < x < \infty \\ 0 & \text{elsewhere,} \end{cases}$$

where Φ is the cumulative distribution function of $N(0,1)$, $\alpha > 0$, $\beta > 0$.

(a) Verify by means of theorem 1.5.7 that F_X is in fact a distribution function.

(b) Show that $\frac{1}{X} \in BS\left(\alpha, \beta^{-1}\right)$. This is known as the *reciprocal property* of the two-parameter Birnbaum-Saunders distribution.

The two-parameter Birnbaum-Saunders distribution is a life time distribution and has been derived from basic assumptions as a probabilistic generative model of failure times of material specimen.

2. Let $X \in BS(\alpha, \beta)$ (c.f. the exercise above). Let $Y = \ln(X)$. Show that the distribution function of Y is

$$F_Y(y) = \Phi\left(\frac{2}{\alpha}\sinh\left(\frac{y-\mu}{2}\right)\right), \quad -\infty < y < \infty,$$

where Φ is the cumulative distribution function of $N(0,1)$ and where $\mu = \ln(\beta)$. This is known as the distribution function of the **sinh-normal** distribution with parameters α, μ and 2.

[11] We would need the Lebesgue-Stieltjes theory of integration to compute, e.g., the expectations and variances of such X.

3. Justify for yourself that
$$\mathbf{P}\left(a \leq X \leq b\right) = F_X(b) - F_X(a) + \mathbf{P}(X = a). \tag{2.91}$$
How is this related to (2.90) ?

4. A distribution function $F(x)$ with the properties

   (a) $F(x)$ is continuous,

   (b) $\frac{d}{dx}F(x) = 0$ for almost all x,

   i.e, there is neither p.d.f. nor p.m.f., is the distribution function of a **singular probability measure** on the real line. One example of such a distribution function is the **Cantor function**. We require first the construction of the **Cantor set** or more precisely the Cantor ternary set.

   One starts by deleting the open middle third $E_1 = (1/3, 2/3)$ from the interval $[0, 1]$. This leaves the union of two intervals: $[0, 1/3] \cup [2/3, 1]$. Next, the open middle third of each of these two remaining intervals is deleted. The deleted open intervals are $E_2 = (1/9, 2/9) \cup (7/9, 8/9)$ and the remaining closed ones are: $[0, 1/9] \cup [2/9, 1/3] \cup [2/3, 7/9] \cup [8/9, 1]$. This construction is continued: E_n is the union of the middle intervals after $E_1, E_2, \ldots, E_{n-1}$ have been removed. The Cantor set C contains all points in the interval $[0, 1]$ that are not deleted at any step in this infinite construction, or
$$C \stackrel{\text{def}}{=} [0, 1] \setminus \cup_{i=1}^{\infty} E_i.$$
   It follows that C is uncountable and it has length (=Lebesgue measure) zero, see [91, pp. 41, 81, 138, 168, 309]. Let now $A_1, A_2, \ldots, A_{2^n-1}$ be the subintervals of $\cup_{i=1}^{n} E_i$. For example
$$E_1 \cup E_2 = (1/9, 2/9) \cup (1/3, 2/3) \cup (7/9, 8/9) = A_1 \cup A_2 \cup A_3.$$
   Then we define
$$F_n(x) = \begin{cases} 0 & x \leq 0, \\ \frac{k}{2^n} & x \in A_k \quad k = 1, 2, \ldots, 2^n - 1, \\ 1 & 1 \leq x, \end{cases}$$
   with linear interpolation in between. Draw graphs of $F_n(x)$ for $n = 2$ and for $n = 3$ in the same picture. Show that $F_n(x) \to F(x)$ for every x. The limiting function $F(x)$ is the **Cantor function**. Then $F(x)$ is continuous and increasing and $F(x)$ is a distribution function of some random variable according to theorem 1.5.7. $\frac{d}{dx}F(x) = 0$ for almost every x, and $F(x)$ has no p.d.f.. Discuss this challenge for understanding continuity and distribution functions with your fellow student[12].

2.6.2 Univariate Probability Density Functions

1. $X \in \text{Exp}\,(\lambda)$. Show that for $a > 0$, $\sqrt{\frac{aX}{\lambda}} \in \text{Ra}(a)$.

2. $Y \in \text{Exp}\,(1)$. What is the distribution of X in
$$Y = \left(\frac{X}{\beta}\right)^{\alpha},$$
   where $\alpha > 0$ and $\beta > 0$. *Answer:* $\text{Wei}(\alpha, \beta)$.

3. Let $X \in U\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Set $Y = \tan(X)$. Show that $Y \in C(0, 1)$.

[12] in Swedish: diskutera cantorfördelning med din bänkkamrat.

4. We say that '**the r.v.'s X and Y are equal in distribution**, if $F_X(z) = F_Y(z)$ for all $z \in \mathbf{R}$, and write this as

$$X \overset{d}{=} Y.$$

Note that this is a very special sort of equality, since for the individual outcomes ω, the numbers $X(\omega)$ and $Y(\omega)$ need not ever be equal.

Let $X \in N(0,1)$. Show that

$$X \overset{d}{=} -X. \tag{2.92}$$

In this case $X(\omega) \neq -X(\omega)$, except when $X(\omega) = -X(\omega) = 0$, which has probability zero. In addition $X \overset{d}{=} -X$ means that the distribution of X is **symmetric** (w.r.t. the origin).

5. Let $Z \in N(\mu, \sigma^2)$. Let $X = e^Z$. Show that the p.d.f. of X is

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \quad x > 0. \tag{2.93}$$

This distribution is called the **Log-Normal** distribution.

6. X is a continuous r.v. and has the p.d.f.

$$f_X(x) = \frac{1}{2\cosh\left(\frac{\pi}{2}x\right)}, \quad -\infty < x < \infty. \tag{2.94}$$

We say that X has the **hyperbolic secant** distribution, $X \in$ HypSech.

(a) Verify the claim that $f_X(x)$ in (2.94) is a p.d.f..

(b) Show that $E[X] = 0$, and $\mathrm{Var}[X] = 1$.

(c) The p.d.f. $f_X(x)$ in (2.94) is plotted in figure 2.2. Explain in words the features that make this p.d.f. different from the p.d.f. of $N(0,1)$. *Aid:* Consider figure 2.3 and read next about **skewness** and **kurtosis**.

Skewness of a random variable X is a measure of symmetry, or more precisely, the lack of symmetry of its distribution. A continuous distribution is intuitively stated symmetric with respect to a center point, if its p.d.f. looks the same to the left and right of the center point. The symmetry of $N(0,1)$ w.r.t. origin has been stated in (2.92) above. Clearly, $N(\mu,\sigma^2)$ is symmetric w.r.t. μ. Skewness κ_1 is formally defined as

$$\kappa_1 \overset{\text{def}}{=} E\left[\frac{(X - E[X])^3}{\sigma^3}\right] = \frac{E\left[X^3\right] - 3E[X]\sigma^2 - (E[X])^3}{\sigma^3}. \tag{2.95}$$

The reader should check the second equality. If $X \in N(\mu, \sigma^2)$, then the skewness is computed to be $= 0$, see (4.50).

Kurtosis is a measure of whether the distribution of X is peaked or flat relative to a normal distribution. High kurtosis (a.k.a. *leptokurtosis*) tends to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Distributions with low kurtosis (a.k.a. *platykurtosis*) tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case. Kurtosis κ_2 is formally defined as

$$\kappa_2 \overset{\text{def}}{=} E\left[\frac{(X - E[X])^4}{\sigma^4}\right]. \tag{2.96}$$

If $X \in N(\mu, \sigma^2)$, then the kurtosis is computed to be $= 3$, see (4.50). Kurtosis is used to measure how much a distribution differs from the normal distribution.

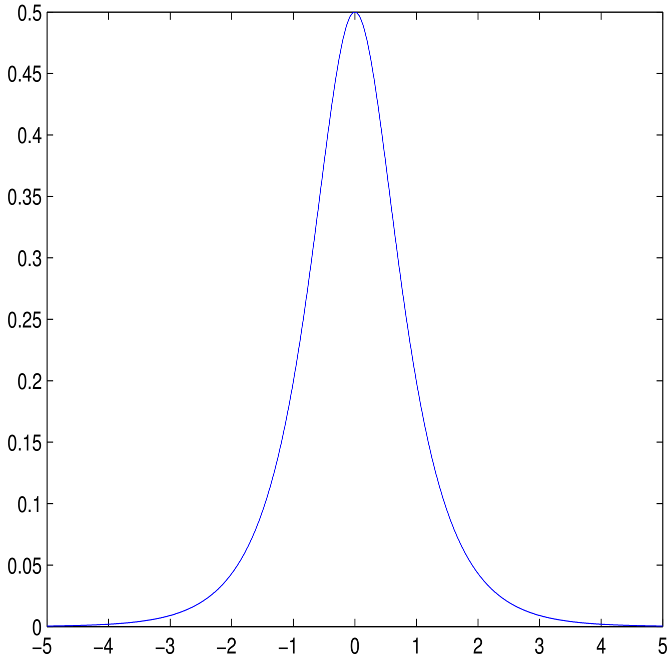

Figure 2.2: The p.d.f. of $X \in$ HypSech.

7. **Hermite Polynomials, Gram-Charlier Expansions, Skewness and Kurtosis** In this exercise we are going to study expansions of 'nearly Gaussian' p.d.f.'s in terms of Hermite polynomials. The resulting expansion of a p.d.f. is known as a Gram-Charlier[13] Expansion [22]. The expansion has recently been in frequent use for **financial analysis**. We need a short summary of the definition and properties of Hermite polynomials.

   The Hermite polynomials $\mathrm{He}_n(x)$, $n = 0, 1, 2, \ldots$, are in [96, p.273], [3, pp.204−209] or [92, pp. 266−267] defined by the **Rodrigues formula**

   $$\mathrm{He}_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}.$$

   This gives $\mathrm{He}_0(x) = 1$, $\mathrm{He}_1(x) = 2x$, $\mathrm{He}_2(x) = 4x^2 - 2$, $\mathrm{He}_3(x) = 8x^3 - 12x$, *ldots* e.t.c.. These polynomials are known as the **physicist's Hermite polynomials**[14]. We shall use another definition to be given next.

   In probability theory [24, p. 133], however, one prefers to work with **probabilist's Hermite polynomials** by

   $$H_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2}. \tag{2.97}$$

[13] Carl Vilhelm Ludwig Charlier (1862−1934) was Professor of Astronomy at Lund University. He is also known for the Charlier-Poisson polynomials.

[14] Indeed, see p. 10 of **Formelsamling i Fysik**, Institutionen för teoretisk fysik, KTH, 2006
`http://courses.theophys.kth.se/SI1161/formelsamling.pdf`

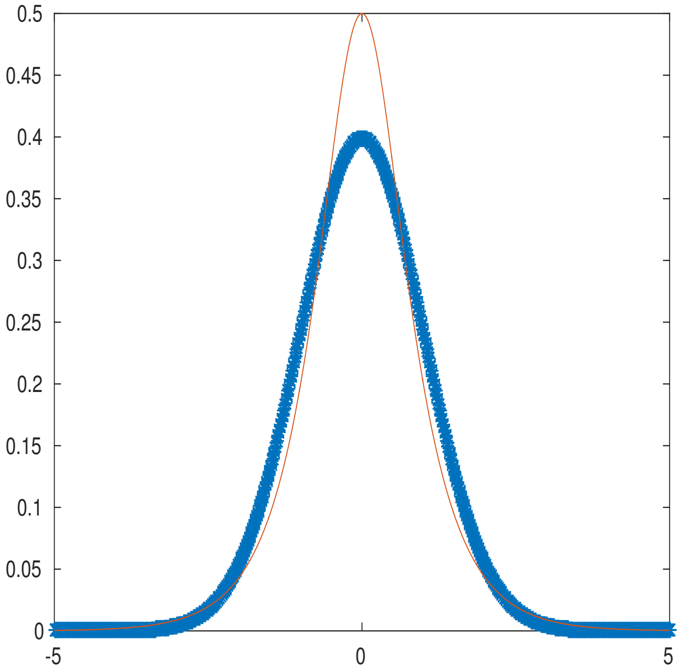

Figure 2.3: The p.d.f. of $X \in \text{HypSech}$ and the p.d.f. of $X \in N(0,1)$ (the thicker function plot).

The first seven are then given by

$$H_0(x) = 1, H_1(x) = x, H_2(x) = x^2 - 1, H_3(x) = x^3 - 3x,$$

$$H_4(x) = x^4 - 6x^2 + 3, H_5(x) = x^5 - 10x^3 + 15x, H_6(x) = x^6 - 15x^4 + 45x^2 - 15.$$

One can in addition define a system of multivariate Hermite polynomials, see [95, p.87]. The Hermite polynomials, as given by (2.97), have the **orthogonality property**

$$\int_{-\infty}^{\infty} e^{-x^2/2} H_n(x) H_m(x) dx = 0, \quad n \neq m, \tag{2.98}$$

and for $n = m$,

$$\int_{-\infty}^{\infty} e^{-x^2/2} \left(H_n(x)\right)^2 dx = n!\sqrt{2\pi}. \tag{2.99}$$

We can now explain the rationale behind the probabilist's definition of Hermite polynomials. Let now $X \in N(0,1)$. Then, by (2.98), if $n \neq m$, and the law of the unconscious statistician (2.4) we have

$$E\left[H_n\left(X\right) H_m\left(X\right)\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} H_n(x) H_m(x) dx = 0, \tag{2.100}$$

and by (2.99)

$$E\left[H_n^2\left(X\right)\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} \left(H_n(x)\right)^2 dx = n!. \tag{2.101}$$

The technical idea of a Gram-Charlier expansion is as follows, [24, pp. 222−223]. Let $f_X(x)$ be a p.d.f.. We consider a symbolic expansion of the form

$$f_X(x) \leftrightarrow c_0\phi(x) + \frac{c_1}{1!}\phi^{(1)}(x) + \frac{c_2}{2!}\phi^{(2)}(x) + \ldots, \tag{2.102}$$

where $\phi(x)$ is the p.d.f. of $N(0,1)$ in (2.15) and $\phi^{(n)}(x) = \frac{d^n}{dx^n}\phi(x)$. The expansion has the attribute 'symbolic', as we are not assured of convergence.

In view of (2.97) we have

$$\phi^{(n)}(x) = (-1)^n\phi(x)H_n(x). \tag{2.103}$$

Thus the right hand side of (2.102) is an expansion in terms of orthogonal polynomials of the type (2.98) and (2.99)[15]. Then we can determine the coefficients c_n by multiplying both sides of (2.102) with $H_n(x)$ and then integrating. The expressions (2.98), (2.99) and (2.103) give

$$c_n = (-1)^n \int_{-\infty}^{\infty} f_X(x)H_n(x)dx. \tag{2.104}$$

We set

$$\widehat{f}_n(x) = \sum_{k=0}^{n} \frac{c_k}{k!}\phi^{(k)}(x) \tag{2.105}$$

Let next X be a standardized r.v., i.e., $E[X] = 0$ and $\text{Var}[X] = 1$.

(a) Show that [22, pp 67−72]

$$\widehat{f}_4(x) = \phi(x) + \frac{c_3}{6}\phi^{(3)}(x) + \frac{c_4}{24}\phi^{(4)}(x). \tag{2.106}$$

Comment: Use the fact that X is standardized, so that, e.g., $\int_{-\infty}^{\infty} x f_X dx = 0$.

(b) Show that in (2.106)

$$c_3 = -\kappa_1,$$

where κ_1 is the **skewness** in (2.95).

(c) Show that in (2.106)

$$c_4 = \kappa_2 - 3,$$

where κ_2 is the **kurtosis** in (2.96).

As stated above, we do not claim in general the convergence of $\widehat{f}_n(x)$ to $f_X(x)$ (or to anything), as $n \to \infty$. In case convergence is there, the speed of convergence can be very slow. But this does not matter for us here. We are interested in an expression like $\widehat{f}_4(x)$ giving us information about the 'near Gaussianness' of $f_X(x)$.

8. $X \in \text{SN}(\lambda)$. Show that

$$X^2 \in \chi^2(1).$$

9. **Exponential Growth Observed at a Random Time** or **a Generative Model for the Pareto Distribution** Let us consider the deterministic (i.e., no random variables involved) exponential growth, or

$$x(t) = e^{\mu t}, \quad t \geq 0, \quad \mu > 0.$$

We stop, or kill, the growth at an exponentially distributed time $T \in \text{Exp}(1/\nu)$. Then we observe the state of the growth at the random time of stopping, or at random age, which is $X = x(T) = e^{\mu T}$. Show

[15] The proper expansion in terms of Hermite polynomials is stated in theorem 9.7.1, but this is not Charlier's concept.

that $X \in \text{Pa}\left(1, \frac{\nu}{\mu}\right)$.
We have here a simple generative model for one of the continuous Pareto distributions in (2.36). *Aid:* Note that since $\mu > 0$ and $T \in \text{Exp}(1/\nu)$, we have $\mathbf{P}(X \leq 1) = 0$.

10. Prove the law of the unconscious statistician (2.4), when $H(x)$ is strictly monotonous, by means of (2.65).

2.6.3 Multivariate P.d.f.'s

1. $\text{Cov}(X, Y) = 0$ **but** X **and** Y **are dependent.** The continuous bivariate random variable (X, Y) has the p.d.f.

$$f_{X,Y}(x, y) = \begin{cases} 1 & -y < x < y, 0 < y < 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{2.107}$$

Show that $\text{Cov}(X, Y) = 0$, but X and Y are not independent.

2. Prove that

$$\mathbf{P}(a < X \leq b, c < Y \leq d) = F_{X,Y}(b, d) - F_{X,Y}(a, d) - F_{X,Y}(b, c) + F_{X,Y}(a, c).$$

Technical Drill

2.1 The four r.v.'s W, X, Y and Z have the joint p.d.f

$$f_{W,X,Y,Z}(w, x, y, z) = 16wxyz, \quad 0 < w < 1, 0 < x < 1, 0 < y < 1, 0 < z < 1.$$

Find $\mathbf{P}\left(0 < W \leq \frac{1}{2}, \frac{1}{2} < X \leq 1\right)$. *Answer:* $\frac{3}{16}$.

3. (From [28]) The continuous bivariate random variable (X, Y) has the p.d.f.

$$f_{X,Y}(x, y) = \begin{cases} xe^{-x(1+y)} & x \geq 0, y \geq 0, \\ \\ 0 & \text{elsewhere.} \end{cases}$$

(a) Find the marginal p.d.f.'s of X and Y. Are X and Y independent ? *Answers:* $f_X(x) = e^{-x}, x \geq 0$, $f_Y(y) = \frac{1}{(1+y)^2}, y \geq 0$. No.

(b) What is the probability that at least one of X and Y exceeds $a > 0$? *Aid:* Consider $\mathbf{P}(\{X \geq a\} \cup \{Y \geq a\})$ and switch over to the complementary probability using De Morgan's rules.
Answer: $e^{-a} + \frac{1}{1+a} - \frac{1}{1+a}e^{-a(1+a)}$.

4. Let $X \in N(0, 1)$ and $Y \in N(0, 1)$ be independent. Set

$$U = \frac{X}{Y}.$$

Show that $U \in C(0, 1)$. *Aid:* The result (2.73) should be useful here.

5. $X \geq 0$ and $Y \geq 0$ are independent continuous random variables with probability densities $f_X(x)$ and $f_Y(y)$, respectively. Show that the p.d.f. of their product XY is

$$f_{XY}(x) = \int_0^\infty \frac{1}{y} f_X\left(\frac{x}{y}\right) f_Y(y) dy = \int_0^\infty \frac{1}{y} f_X(y) f_Y\left(\frac{x}{y}\right) dy. \tag{2.108}$$

Technical Drills

5 $X \in U(0,1)$, and $Y \in U(0,1)$ are independent. Let $W = XY$. Show that

$$f_W(w) = -\ln w, \quad 0 < w \leq 1. \tag{2.109}$$

6. X and Y are independent random variables with p.d.f.'s $f_X(x)$ and $f_Y(y)$, respectively. Show that their sum $Z = X + Y$ has the p.d.f.

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx = \int_{-\infty}^{\infty} f_Y(y) f_X(z-y) dy. \tag{2.110}$$

The integrals in the right hand side are known as **convolutions** of f_X and f_Y. A convolution sum is valid for the probability mass function of a sum of two indepedendent discrete random variables.

7. $X \in \mathrm{Exp}(1/\lambda)$ and $Y \in \mathrm{Exp}(1/\mu)$ are independent, $\lambda > 0$, $\mu > 0$. Let $Z \overset{\text{def}}{=} X - Y$.

(a) Show that $E[Z] = \frac{1}{\lambda} - \frac{1}{\mu}$, and $\mathrm{Var}[Z] = \frac{1}{\lambda^2} + \frac{1}{\mu^2}$.

(b) Show that Z has the p.d.f.

$$f_Z(z) = \begin{cases} \frac{\lambda\mu}{\lambda+\mu} e^{-\lambda z} & z \geq 0 \\ \frac{\lambda\mu}{\lambda+\mu} e^{\mu z} & z < 0. \end{cases} \tag{2.111}$$

(c) In **probabilistic reliability theory of structures**, [32], X would denote the random stress resulting in a bar of constant cross section subjected to an axial random force. Y would denote the resistance, the allowable stress, which is also random. Then R, the reliability of the structure, is

$$R \overset{\text{def}}{=} \mathbf{P}(X \leq Y).$$

Show that

$$R = \frac{\lambda}{\lambda + \mu}.$$

(d) If $\lambda = \mu$, which known distribution is obtained for Z?

8. (From [6]) (X, Y) is a continuous bivariate r.v., and its joint p.d.f is

$$f_{X,Y}(x,y) = \frac{6}{7}x, \quad 0 \leq x, \quad 0 \leq y, \quad 1 \leq x + y \leq 2.$$

Find the marginal p.d.f.'s $f_X(x)$ and $f_Y(y)$. *Answer:*

$$f_X(x) = \begin{cases} \frac{6}{7}x & 0 \leq x \leq 1 \\ \frac{12}{7}x - \frac{6}{7}x^2 & 1 \leq x \leq 2. \end{cases}$$

$$f_Y(y) = \begin{cases} \frac{9}{7} - \frac{6}{7}y & 0 \leq y \leq 1 \\ \frac{3}{7}(2-y)^2 & 1 \leq y \leq 2. \end{cases}$$

9. (X, Y) is a continuous bivariate r.v., and its joint p.d.f is

$$f_{X,Y}(x,y) = \frac{c}{x} e^{-x^2 y}, \quad x \geq 1, y \geq 0.$$

Show that $c = 2$.

10. (X, Y) is a continuous bivariate r.v., and its joint p.d.f is

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{1+c}(xy + c)e^{-(x+y)} & 0 \leq x, 0 \leq y \\ 0 & \text{elsewhere.} \end{cases}$$

(a) Find the marginal p.d.f.'s $f_X(x)$ and $f_Y(y)$. *Answer:*

$$f_X(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x+c}{1+c} e^{-x} & x \geq 0. \end{cases}$$

(b) Show that if $c = 0$, then X and Y are independent, and that if $c > 0$, X and Y are not independent.

11. (X, Y) is a continuous bivariate r.v., and its joint p.d.f is for

$$f_{X,Y}(x, y) = \frac{1}{\pi^2(1 + x^2)(1 + y^2)}, \quad -\infty < x < \infty, -\infty < y < \infty.$$

(a) Find the distribution function $F_{X,Y}(x, y)$. *Aid:* Plain to see.

(b) Find the marginal p.d.f.'s $f_X(x)$ and $f_Y(y)$.

12. The continuous bivariate random variable (X, Y) has the p.d.f.

$$f_{X,Y}(x, y) = \begin{cases} e^{-y} & 0 \leq x \leq y \\ \\ 0 & \text{elsewhere.} \end{cases} \tag{2.112}$$

(a) Find the marginal p.d.f.'s of X and Y. Are X and Y independent ? *Answers:* $f_X(x) = e^{-x}, x > 0$ and $= 0$ elsewhere, $f_Y(y) = ye^{-y}, y > 0$, and $= 0$ elsewhere.

(b) Show that X and $\frac{X}{Y}$ are independent.

(c) Give a generative model for (X, Y). *Aid:* Note that $Y \in \Gamma(2, 1)$.

13. **The t -distribution** $X \in N(0, 1)$, $Y \in \chi^2(n)$. X and Y are independent. Show that

$$\frac{X}{\sqrt{\frac{Y}{n}}} \in t(n). \tag{2.113}$$

14. **The F-distribution** Let $X_1 \in \chi^2(f_1)$, $X_2 \in \chi^2(f_2)$. X_1 and X_2 are independent. Consider the ratio

$$U \stackrel{\text{def}}{=} \frac{\frac{X_1}{f_1}}{\frac{X_2}{f_2}}.$$

Show that the p.d.f. of U is

$$f_U(u) = \frac{\Gamma\left(\frac{f_1+f_2}{2}\right)\left(\frac{f_1}{f_2}\right)^{f_1/2}}{\Gamma\left(\frac{f_1}{2}\right)\Gamma\left(\frac{f_2}{2}\right)} \frac{u^{\frac{f_1}{2}-1}}{\left(1 + \frac{f_1 u}{f_2}\right)^{(f_1+f_2)/2}}, \quad 0 < u < \infty.$$

This is the p.d.f. of what is known as **F -distribution** or **Fisher -Snedecor -distribution**. The distribution is important in the analysis of variance and econometrics (F-test). *Aid:* You need the technique of an **auxiliary variable**, take $V = X_2$. Then consider (U, V) as a transformation of (X_1, X_2). The Jacobian of the transformation is $J = \frac{f_2 V}{f_1}$. Find the joint p.d.f. $f_{(U,V)}(u, v)$, and marginalize to get $f_U(u)$.

15. (From [49]) (X, Y) has the p.d.f.

$$f_{X,Y}(x, y) = \begin{cases} 1 & 0 \leq x \leq 2, \max(0, x - 1) \leq y \leq \min(1, x) \\ 0 & \text{elsewhere.} \end{cases}$$

Show that $X \in \text{Tri}(0, 2)$, $Y \in U(0, 1)$.

16. (From [49]) X_1 and X_2 are independent and have the common the p.d.f.

$$f_X(x) = \begin{cases} 4x^3 & 0 \leq x \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Set $Y_1 = X_1\sqrt{X_2}$, $Y_2 = X_2\sqrt{X_1}$. Find the joint p.d.f. of (Y_1, Y_2). Are Y_1, and Y_2 independent? *Answer:*

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} \frac{64}{3}\left(y_1 y_2\right)^{5/3} & 0 < y_1^2 < y_2 < \sqrt{y_1} < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Y_1 and Y_2 are not independent.

17. (From [49]) (X, Y) has the p.d.f.

$$f_{X,Y}(x, y) = \begin{cases} \frac{2}{(1+x+y)^3} & 0 < x, 0 < y \\ 0 & \text{elsewhere.} \end{cases}$$

Show that

(a) $f_{X+Y}(u) = \frac{2u}{(1+u)^3}$, $0 < u$.

(b) $f_{X-Y}(v) = \frac{1}{2(1+|v|)^2}$, $-\infty < v < \infty$.

18. In this exercise we study the bivariate Bernoulli distribution in example 2.3.16.

(a) Show that the function $p_{X,Y}(x, y)$ in (2.64) is a p.m.f..

(b) Find the marginal p.m.f.s $p_X(x)$ and $p_Y(y)$.

(c) Are X and Y independent ? (Yes)

(d) What is the distribution of X? What is the distribution of Y?

19. (X, Y) is a discrete bivariate r.v., such that their joint p.m.f. is

$$p_{X,Y}(j, k) = c\frac{(j+k)a^{j+k}}{j!k!},$$

where $a > 0$.

(a) Determine c. *Answer:* $c = \frac{e^{-2a}}{2a}$

(b) Find the marginal p.m.f. $p_X(j)$. *Answer:* $p_X(0) = \frac{e^{-a}}{2}$, $p_X(j) = c\frac{a^j}{j!}e^a(j + a)$ for $j \geq 1$.

(c) Find $\mathbf{P}(X + Y = r)$. *Answer:* $\mathbf{P}(X + Y = r) = c\frac{(2a)^r}{(r-1)!}, r \geq 1, \mathbf{P}(X + Y = 0) = 0$.

(d) Find $E[X]$. *Answer:*$\frac{1}{2}\left(e^{-a} + a + 1\right)$.

20. Let $X_1 \in \Gamma(a_1, b)$ and $X_2 \in \Gamma(a_2, b)$ be independent. Show that $\frac{X_1}{X_2}$ and $X_1 + X_2$ are independent.

21. Let $X_1 \in \Gamma(r, 1)$ and $X_2 \in \Gamma(s, 1)$ be independent.

(a) Show that $\frac{X_1}{X_1+X_2}$ and $X_1 + X_2$ are independent.

(b) Show that $\frac{X_1}{X_1+X_2} \in \beta(r, s)$.

22. (See [56, pp.170−171] or [97, pp. 94−95].) $X \in N\left(v\cos(\phi), \sigma^2\right)$, $Y \in N\left(v\sin(\phi), \sigma^2\right)$, where X and Y are independent. Set

$$R = \sqrt{X^2 + Y^2}.$$

(a) Show that the probability density $f_R(r)$ of R is

$$f_R(r) = \frac{r}{\sigma^2} e^{-\frac{\left(r^2+v^2\right)}{2\sigma^2}} I_0\left(\frac{rv}{\sigma^2}\right), \tag{2.114}$$

where $I_0(x)$ is a modified Bessel function of the first kind with order 0. The distribution in this exercise is known as the **Rice distribution**. We write

$$R \in \mathrm{Rice}\,(v, \sigma)\,.$$

The Rice distribution of R is the distibution of the envelope of the narrowband Gaussian noise [3, section 8.3.1.]. The ratio $\frac{v^2}{2\sigma^2}$ is known as the **signal-to-noise ratio** (SNR).

(b) Which distribution is obtained for $v = 0$?

23. The **Marcum Q-function**[16] is a special function important in communications engineering and radar detection and is defined as

$$Q_m(v, b) \stackrel{\text{def}}{=} \frac{1}{v^{m-1}} \int_b^\infty r^m e^{-\frac{\left(r^2+v^2\right)}{2}} I_{m-1}(rv)dr, \tag{2.115}$$

where $I_{m-1}(z)$ is a modified Bessel function of the first kind with order $m-1$.

(a) Show that the Marcum Q-function can be written as

$$Q_m(v, b) = e^{-\frac{\left(b^2+v^2\right)}{2}} \sum_{k=1-m}^{\infty} \left(\frac{v}{b}\right)^k I_k\,(vb)\,. \tag{2.116}$$

(b) Let $F_R(r)$ be the distribution function of $R \in \mathrm{Rice}\,(v, \sigma)$,

$$F_R(r) = \int_0^r f_R(u)du.$$

Show that

$$F_R(r) = 1 - Q_1\left(\frac{v}{\sigma}, \frac{r}{\sigma}\right).$$

This is a useful statement, since there are effective algorithms for numerical computation of the Marcum Q-function.

(c) Let $R_i \in \mathrm{Rice}\,(v_i, \sigma_i)$ for $i = 1, 2$, be independent. Show that

$$\mathbf{P}\,(R_2 > R_1) = Q_1\left(\sqrt{\alpha}, \sqrt{\beta}\right) - \frac{\nu^2}{1+\nu^2} e^{-\frac{\alpha+\beta}{2}} I_0\left(\sqrt{\alpha\beta}\right),$$

where $\alpha = \frac{v_2^2}{\sigma_1^2+\sigma_2^2}$ and $\beta = \frac{v_1^2}{\sigma_1^2+\sigma_2^2}$ and $\nu = \frac{\sigma_1}{\sigma_2}$.

24. **Marcum Q-function and the Poisson distribution** This exercise is found in the technical report in the footnote above. The results are instrumental for computation of $Q_1\left(\frac{v}{\sigma}, \frac{r}{\sigma}\right)$. Let $X \in \mathrm{Po}(\lambda)$, $Y \in \mathrm{Po}(\lambda)$, where X and Y are independent.

(a) Show that

$$\mathbf{P}\,(X = Y) = e^{-2\lambda} I_0(2\lambda),$$

where $I_0(z)$ is a modified Bessel function of the first kind with order 0.

[16] For this topic and the definitions used, see , e.g., G.V. Weinberg: *Stochastic representations of the Marcum Q-function and associated radar detection probabilities.* Australian Goverment. Department of Defence. Defence Science and Technology Organisation. DSTO-RR-0304 (approved for public release), 2005.

(b) Show that

$$\mathbf{P}(X \leq Y) = \frac{1}{2}\left[1 + e^{-2\lambda} I_0(2\lambda)\right].$$

We can perhaps make the surprising link between the Poisson distribution and Marcum Q-function more explicit by the following observation.

By [3, Problem 21 (b). p. 297] we have

$$e^x = I_0(x) + 2\sum_{n=1}^{\infty} I_n(x), \tag{2.117}$$

which can be established be means of the appropriate generating function of modified Bessel functions of the first kind.

Then in view of (2.117) and (2.116) we obtain that

$$Q_1\left(\sqrt{2\lambda}, \sqrt{2\lambda}\right) = \frac{1}{2}\left[1 + e^{-2\lambda} I_0(2\lambda)\right].$$

25. (From [14]) $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with the distribution function $F(x)$ and p.d.f. $f(x)$. We consider the **range** $R = R(X_1, X_2, \ldots, X_n)$ defined by

$$R \overset{\text{def}}{=} \max_{1\leq i\leq n} X_i - \min_{1\leq i\leq n} X_i.$$

This is a function of the n r.v.'s that equals the distance between the largest and the smallest. The text [51, ch. 12] discusses the range as applied in control charts of statistical quality engineering.

The task here is to show that the probability distribution function of R is

$$F_R(x) = n\int_{-\infty}^{\infty} \left[F(t+x) - F(t)\right]^{n-1} f(t)dt. \tag{2.118}$$

In general, $F_R(x)$ cannot be evaluated in a closed form and is computed by numerical quadratures. Next we find the formula in (2.118) by the sequence of steps (a)-(d).

(a) Set $Z \overset{\text{def}}{=} \max_{1\leq i\leq n} X_i$ and $Y \overset{\text{def}}{=} \min_{1\leq i\leq n} X_i$. Let

$$F(y,z) = \mathbf{P}(Y \leq y, Z \leq z), \frac{\partial^2}{\partial y \partial z} F(y,z) = f(y,z).$$

Now show that

$$F_R(x) = \int_{-\infty}^{\infty} \frac{\partial}{\partial y} F(y,z)|_{z=y}^{z=y+x} dy. \tag{2.119}$$

(b) Show next that

$$\mathbf{P}(Y \geq y, Z \leq z) = \left[F(z) - F(y)\right]^n \tag{2.120}$$

(c) Show next using (2.120) that

$$F(y,z) = \mathbf{P}(Y \leq y, Z \leq z) = \mathbf{P}(Z \leq z) - \left[F(z) - F(y)\right]^n \tag{2.121}$$

(d) Next use (2.121) to finally establish (2.118).

26. $X_1 \in \text{Exp}(1/\lambda)$ and $X_2 \in \text{Exp}(1/\mu)$ are independent r.v.'s. We let

$$Y \overset{\text{def}}{=} \min(X_1, X_2), Z \overset{\text{def}}{=} \max(X_1, X_2), R = Z - Y.$$

(a) Show that

$$\mathbf{P}(R \geq a) = \frac{\lambda e^{-\mu a} + \mu e^{-\lambda a}}{\lambda + \mu}.$$

Aid: Draw a picture for $R \geq a$.

(b) Find the distribution of R, when $\lambda = \mu$. *Hint*: E.g., (2.118). *Answer:* $R \in \text{Exp}(1/\lambda)$.

2.6.4 Expectations and Variances

1. Let $X \in \mathrm{Ge}(p)$, see example 2.3.4. Show that

$$E[X] = \frac{q}{p}, \mathrm{Var}[X] = \frac{q}{p^2}.$$

Aid: Let $f(p) = \frac{1}{1-p} = \sum_{k=0}^{\infty} p^k$, $|p| < 1$. Then $f^{'}(p) = \sum_{k=1}^{\infty} kp^{k-1}$ and $f^{''}(p) = \sum_{k=2}^{\infty} k(k-1)p^{k-2}$.

2. Let $X \in \mathrm{Fs}(p)$, see example 2.3.5. Show that

$$E[X] = \frac{1}{p}, \mathrm{Var}[X] = \frac{q}{p^2}.$$

Aid: As above for exercise 1. in this section.

3. **Expectation** and **Variance** of $\mathrm{SN}(\lambda)$
Recall example 2.2.6.

(a) It needs first to be checked that the p.d.f. of $X \in \mathrm{SN}(\lambda)$ as given in (2.22) is in fact a p.d.f.. The serious challenge is to show that

$$\int_{-\infty}^{\infty} f_X(x)dx = 1 \quad \text{for all } \lambda.$$

Note that the chosen notation hides the fact that $f_X(x)$ is also a function of λ. *Aid:* Define $\Psi(\lambda) \overset{\text{def}}{=} \int_{-\infty}^{\infty} f_X(x)dx$. Then we have $\Psi(0) = 1$ and $\frac{d}{d\lambda}\Psi(\lambda) = 0$ for all λ (check this) and thus the claim is proved.

(b) Show that

$$E[X] = \sqrt{\frac{2}{\pi}} \frac{\lambda}{\sqrt{1+\lambda^2}}.$$

Aid: Introduce the auxiliary function $\Psi(\lambda) \overset{\text{def}}{=} \int_{-\infty}^{\infty} x f_X(x)dx$ and find that

$$\frac{d}{d\lambda}\Psi(\lambda) = \sqrt{\frac{2}{\pi}} \frac{1}{(1+\lambda^2)^{3/2}}.$$

Then

$$E[X] = \int \sqrt{\frac{2}{\pi}} \frac{1}{(1+\lambda^2)^{3/2}} d\lambda + C.$$

and the constant of integration C can be determined from $\Psi(0)$.

(c) Show that

$$\mathrm{Var}[X] = 1 - \frac{2}{\pi} \frac{\lambda^2}{1+\lambda^2}.$$

Aid: Use Steiner$^{'}$s formula (2.6) and the fact that $X^2 \in \chi^2(1)$.

4. **Skewness** and **Kurtosis** $\mathrm{SN}(\lambda)$

(a) Check that the skewness (2.95) of $X \in \mathrm{SN}(\lambda)$ is

$$\kappa_1 = \left(\frac{4-\pi}{2}\right) \cdot \frac{(E[X])^3}{(\mathrm{Var}[X])^{3/2}}.$$

Hence $\lambda = 0$ implies $\kappa_1 = 0$, as should be.

(b) Check that the kurtosis (2.96) of $X \in \mathrm{SN}(\lambda)$ is

$$\kappa_2 = 2(\pi - 3) \cdot \frac{(E[X])^4}{(\mathrm{Var}[X])^2}.$$

5. **Chebychev's inequality** Let $X_1, X_2, \ldots, X_n$ be independent r.v.'s, and identically $X_i \in U(-1,1)$. Set $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Use the Chebychev inequality (1.27) to estimate how large n should be so that we have

$$\mathbf{P}\left(\mid \overline{X} \mid > 0.05\right) \leq 0.05.$$

Answer: $n \geq 2667$.

6. |**Coefficient of Correlation**| ≤ 1 The coefficient of correlation is defined in (2.45). The topic of this exercise is to show that (2.46), i.e., $|\rho_{X,Y}| \leq 1$ holds true.

(a) Let now X and Y be two r.v.'s, dependent or not. Assume that $E[X] = E[Y] = 0$ and $\mathrm{Var}[X] = \mathrm{Var}[Y] = 1$. Show that $E[XY] \leq 1$. *Aid:* Since $(X-Y)^2 \geq 0$, we get that $E\left[(X-Y)^2\right] \geq 0$. Expand $(X-Y)^2$ to show the claim.

(b) The r.v.'s are as in (a). Show that $E[XY] \geq -1$. *Aid:* Consider $(X+Y)^2$, and apply steps of argument similar to the one in case (a).

(c) We conclude by (a) and (b) that $|E[XY]| \leq 1$ under the conditions there. Let now X and Y be two r.v.'s, independent or dependent. Assume that $E[X] = \mu_X$ and $E[Y] = \mu_Y$ and $\mathrm{Var}[X] = \sigma_X^2$, $\mathrm{Var}[Y] = \sigma_Y^2$. Set $Z_1 = \frac{X-\mu_X}{\sigma_X}$, and $Z_2 = \frac{Y-\mu_Y}{\sigma_Y}$. Now prove that $|\rho_{X,Y}| \leq 1$ by applying the conclusion above to Z_1 and Z_2.

7. **When is the Coefficient of Correlation** $= \pm 1$ **?** Show for the coefficient of correlation $\rho_{X,Y}$ as defined in (2.45) that

$$\rho_{X,Y} = \pm 1 \Leftrightarrow Y = aX + b.$$

8. $X \in \Gamma(p, 1/\lambda)$. Show that

$$E[X^m] = \frac{(m+p-1)!}{(r-1)!\lambda}.$$

2.6.5 Additional Exercises

1. $X \in \mathrm{Ge}(p)$, $0 < p < 1$. Let m be an integer ≥ 2. The **floor function** or the **integer part** of a real number x is

$$\lfloor x \rfloor \overset{\mathrm{def}}{=} \text{the largest integer smaller than } x. \tag{2.122}$$

We set

$$L_m = \left\lfloor \frac{X}{m} \right\rfloor.$$

and

$$R_m = X - m \cdot L_m.$$

(a) Show that the marginal p.m.f. of L_m is

$$P(L_m = l) = (1 - (1-p)^m)(1-p)^{ml}, \quad l = 0, 1, \ldots,$$

i.e., $L_m \in \mathrm{Ge}((1-p)^m)$ and that the marginal p.m.f. of R_m is

$$P(R_m = r) = \frac{(1-p)^r p}{1-(1-p)^m}, \quad r = 0, 1, \ldots, m-1.$$

(b) Show that L_m and R_m are independent r.v.'s.

2. Let $X \in \mathrm{Exp}\,(1/\lambda)$. Invoking again the integer part (2.122) we set

$$L_m = \left\lfloor \frac{X}{m} \right\rfloor$$

and

$$R_m = X - m \cdot L_m.$$

Show that L_m and R_m are independent r.v.'s. Determine even the marginal distributions of L_m and R_m.

3. $X \in \mathrm{Exp}(1/\lambda)$, and

$$D = X - \lfloor X \rfloor.$$

D is the **fractional part** of X, as $\lfloor X \rfloor$ is the integer part of X. Show that the p.d.f of D is

$$f_D(d) = \begin{cases} \frac{\lambda e^{-\lambda d}}{1-e^{-\lambda}} & 0 < d < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

4. Let $X_1, X_2, \ldots, X_n$ be I.I.D. random variables under a continuous probability distribution with the distribution function $F_X(x)$. Let θ be the median of the distribution, i.e., a number such that

$$\frac{1}{2} = F_X(\theta).$$

Find the probability distribution of the number of the variables $X_1, X_2, \ldots, X_n$ that are less than θ.

5. **Chen's Lemma** $X \in \mathrm{Po}(\lambda)$. $H(x)$ is a locally bounded Borel function. Show that

$$E\left[XH(X)\right] = \lambda E\left[H(X+1)\right]. \tag{2.123}$$

Chen's lemma is found, e.g., in [9]. The cited reference develops a whole theory of Poisson approximation as a consequence of (2.123).

6. $X \in \mathrm{Po}(\lambda)$. Show that

$$E\left[X^n\right] = \lambda \sum_{k=0}^{n-1} \binom{n-1}{k} E\left[X^k\right]. \tag{2.124}$$

Aid: Use Chen's Lemma with a suitable $H(x)$.

7. **Skewness and Kurtosis of Poisson r.v.'s** Recall again (2.95) and (2.96). Show that if $X \in \mathrm{Po}(\lambda)$, $\lambda > 0$, then

(a)

$$\kappa_1 = \frac{1}{\sqrt{\lambda}},$$

(b)

$$\kappa_2 = 3 + \frac{1}{\lambda}.$$

8. $X \in \mathrm{Po}(\lambda)$, $\lambda > 0$. Find

$$E\left[\frac{1}{1+X}\right].$$

Answer: $\frac{1}{1+\lambda}\left(1 - e^{-\lambda}\right)$. Why can we not compute $E\left[\frac{1}{X}\right]$?

9. Let $X_1, X_2, \ldots, X_n$ are I.I.D. and positive r.v.'s. Show that for any $k \leq n$
$$E\left[\frac{X_1 + X_2 + \ldots + X_k}{X_1 + X_2 + \ldots + X_n}\right] = \frac{k}{n}.$$
Aid: Deal first with the case $n = 2$.

10. (From [88]) $X \in \text{Po}(\lambda)$, $\lambda > 0$. Show that
$$\mathbf{P}\left(X \leq k\right) = \frac{1}{k!}\int_{\lambda}^{\infty} e^{-t}t^{k}dt, \quad k = 0, 1, 2, \ldots.$$

11. **Mill's inequality** $X \in N(0,1)$. Show that
$$P(\mid X \mid > t) \leq \sqrt{\frac{2}{\pi}}\frac{e^{-\frac{t^2}{2}}}{t}. \tag{2.125}$$
Aid: Show first that $P(\mid X \mid > t) = 2P(X > t)$. Then find $P(X > t)$ and provide the desired upper bound by observing that if $x > t$, then $\frac{x}{t} > 1$.

12. **Mill's inequality and Chebychev's Inequality** Let $X_1, \ldots, X_n$ are I.I.D. and $\in N(0,1)$. Set $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Use Mill's inequality (2.125) to find and upper bound for $P(\mid \overline{X} \mid > t)$ and make a comparison with the bound by Chebychev's Inequality. *Aid:* The reader is assumed to know that $\overline{X} \in N\left(0, \frac{1}{n}\right)$.

Chapter 3

Conditional Probability and Expectation w.r.t. a Sigma Field

3.1 Introduction

Conditional probability and conditional expectation are fundamental in probability and random processes in the sense that all probabilities referring to the real world are necessarily conditional on the information at hand. The notion of conditioning does not seem to play any significant role in general measure and integration theory, as developed by pure mathematicians, who need not necessarily regard applied processing of information as a concern of their theoretical work.

The conditional probability is in a standard fashion introduced as

$$\mathbf{P}\left(A \mid B\right) \overset{\text{def}}{=} \frac{\mathbf{P}\left(A \cap B\right)}{\mathbf{P}(B)}, \tag{3.1}$$

which is called the conditional probability of the event A given the event B, if $\mathbf{P}(B) > 0$. $\mathbf{P}\left(B \mid A\right)$ is defined analogously. We shall in this chapter expand the mathematical understanding of the concepts inherent in the definition of conditional probability of an event in (3.1).

3.2 Conditional Probability Densities and Conditional Expectations

We are operating with the notations for bivariate random variables in section 2.2.2 above. The **conditional density** for Y given $X = x$ is for $f_X(x) > 0$ defined by

$$f_{Y|X=x}(y) \overset{\text{def}}{=} \frac{f_{X,Y}(x,y)}{f_X(x)}. \tag{3.2}$$

We have that

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{\int_{-\infty}^{\infty} f_{X,Y}(x,t)\,dt}\,. \tag{3.3}$$

Clearly

$$\int_{-\infty}^{\infty} f_{Y|X=x}(y)dy = 1 \quad \text{for all } x.$$

In this setting $P(Y \leq y \mid X = x)$= $F_{Y|X=x}(y) = \int_{-\infty}^{y} f_{Y|X=x}(u)du$ is the conditional distribution function of Y under the condition $X = x$. The conditional density $f_{Y|X=x}(y)$ is thus the derivative of $F_{Y|X=x}(y)$ with respect to y.

In the sequel we shall use a special symbolic notation for conditional distributions $P(Y \leq y \mid X = x)$ using the earlier distribution codes. For example, suppose that for any $x > 0$

$$f_{Y|X=x}(y) = \begin{cases} \frac{1}{x}e^{-y/x} & 0 \leq y \\ \\ 0 & \text{elsewhere.} \end{cases}$$

Then we write this in view of (2.23) as $Y \mid X = x \in \text{Exp}(x)$.

Our calculus should be compatible with (3.1). One difficulty is that the event $\{X = x\}$ has the probability $= 0$, since X is a continuous random variable. We can think heuristically that the conditioning event $\{X = x\}$ is more or less $\{x \leq X \leq x + dx\}$ for an infinitesimal dx. We obtain (c.f., (3.1))

$$\begin{aligned} F_{Y|X=x}(y) = P(Y \leq y \mid X = x) &\approx P(Y \leq y \mid x \leq X \leq x + dx) \\ &= \frac{P(x \leq X \leq x + dx, Y \leq y)}{P(x \leq X \leq x + dx)} = \frac{F_{X,Y}(x + dx, y) - F_{X,Y}(x, y)}{P(x \leq X \leq x + dx)} \\ &\approx \frac{\frac{\partial}{\partial x}F_{X,Y}(x, y)dx}{f_X(x)dx} = \frac{\frac{\partial}{\partial x}F_{X,Y}(x, y)}{f_X(x)}, \end{aligned}$$

and thus

$$f_{Y|X=x}(y) = \frac{\partial}{\partial y}F_{Y|X=x}(x, y) = \frac{\frac{\partial^2}{\partial x \partial y}F_{X,Y}(x, y)}{f_X(x)} = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

In the same way as in (3.2) we can write

$$f_{X,Y}(x, y) = f_X(x)f_{Y|X=x}(y),$$

and this yields

$$f_Y(y) = \int_{-\infty}^{\infty} f_X(x)f_{Y|X=x}(y)\, dx. \tag{3.4}$$

For a bivariate discrete variable (X, Y) we have

$$p_{Y|X=x_k}(y_j) = \frac{p_{X,Y}(x_k, y_j)}{p_X(x_k)} \quad \text{for } j = 0, 1, 2, \ldots, .$$

Example 3.2.1 The operation in (3.4) is used in Bayesian statistics and elsewhere in the following way. Let $f_{Y|\Theta=\theta}(y)$ be a probability density with the parameter θ, which is regarded as an outcome of the r.v. Θ. Then $f_\Theta(\theta)$ is the **prior p.d.f.** of Θ.

To illustrate the idea precisely, think here of, e.g., the r.v. $Y \mid \Theta = \theta \in \text{Exp}(\theta)$, with the p.d.f. $f_{Y|\Theta=\theta}(y) = \frac{1}{\theta}e^{-y/\theta}$, where θ is an outcome Θ, which is a positive r.v. with, e.g., $\Theta \in \text{Exp}(\lambda)$.

Then the **Bayesian integral** or the **mixing integral**

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|\Theta=\theta}(y)f_\Theta(\theta)\, d\theta \tag{3.5}$$

defines a new p.d.f. $f_Y(y)$ (sometimes known as a predictive p.d.f.), which may depend on the so called **hyperparameters** (like λ in the preceding discussion) from $f_\Theta(\theta)$. Following the rules for conditional densities we obtain also

$$f_{\Theta|Y=y}(\theta) = \frac{f_{\Theta,Y}(\theta, y)}{f_Y(y)}$$

and furthermore

$$f_{\Theta|Y=y}(\theta) = \frac{f_{Y|\Theta=\theta}(y) f_{\Theta}(\theta)}{\int_{-\infty}^{\infty} f_{Y|\Theta=\theta}(y) f_{\Theta}(\theta)\, d\theta}. \tag{3.6}$$

This is **Bayes' rule** (with p.d.f.'s), and constitutes an expression for the **posterior p.d.f** of Θ given $Y = y$.

■

In view of the preceding we define quite naturally the **conditional expectation** for Y given $X = x_k$ by

$$E(Y \mid X = x_k) \overset{\text{def}}{=} \sum_{j=-\infty}^{\infty} y_j \cdot p_{Y|X=x_k}(y_j).$$

The **conditional expectation** for Y given $X = x$ is given by

$$E(Y \mid X = x) \overset{\text{def}}{=} \int_{-\infty}^{\infty} y f_{Y|X=x}(y)\, dy. \tag{3.7}$$

Theorem 3.2.2 Double Expectation

$$E(Y) = \begin{cases} \sum\limits_{k=-\infty}^{\infty} E(Y \mid X = x_k) p_X(x_k) & \text{discrete r.v.} \\ \int\limits_{-\infty}^{\infty} E(Y \mid X = x) f_X(x)\, dx & \text{continuous r.v..} \end{cases} \tag{3.8}$$

Proof We deal with the continuous case. The conditional expectation $E(Y \mid X = x)$ is a function of x, which we denote by $H(x)$. Then we have the random variable $H(X) = E(Y \mid X)$ for some (what has to be a Borel) function H. We have by the law of unconscious statistician (2.4) that

$$E\left[H(X)\right] = \int_{-\infty}^{\infty} H(x) f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} E(Y \mid X = x) f_X(x)\, dx,$$

and by the definition in (3.7)

$$= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y f_{Y|X=x}(y)\, dy \right) f_X(x)\, dx = \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx\, dy$$

and from the definition of marginal density

$$= \int_{-\infty}^{\infty} y \underbrace{\int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx}_{=f_Y(y)}\, dy = \int_{-\infty}^{\infty} y f_Y(y)\, dy = E(Y).$$

The proof for the discrete case is now obvious. ■

We write the result (3.7) above as the rule of **double expectation**

$$E(Y) = E(E(Y \mid X)). \tag{3.9}$$

The **conditional variance** of Y given $X = x$ is defined by

$$\text{Var}\,(Y \mid X = x) \overset{\text{def}}{=} E((Y - \mu_{Y|X=x})^2 \mid X = x),$$

where $\mu_{Y|X=x} = E(Y \mid X = x)$. In addition $\text{Var}(Y \mid X = x) = H(x)$ for some Borel function $H(x)$. There is no rule equally simple as (3.9) for variances, but we have the following theorem.

Theorem 3.2.3 (Law of Total Variance)

$$\text{Var}(Y) = E(\text{Var}(Y \mid X)) + \text{Var}(E(Y \mid X)). \tag{3.10}$$

■

The law of total variance above will be proved by means of (3.46) in the exercises to this chapter. We shall next develop a deeper and more abstract (and at the same time more expedient) theory of conditional expectation (and probability) that relieves us from heuristics of the type '$\{x \leq X \leq x + dx\}$ for an infinitesimal dx', and yields the results above as special cases. We start with the simplest case, where we condition w.r.t. an event.

3.3 Conditioning w.r.t. an Event

We have a random variable X on $(\Omega, \mathcal{F}, \mathbf{P})$ and take $A \in \mathcal{F}$. We assume that $\mathbf{P}(A) > 0$. We recall the definition of $\int_A X d\mathbf{P}$ in an exercise in chapter 1, or,

$$\int_A X d\mathbf{P} = \int_\Omega \chi_A \cdot X d\mathbf{P} = E\left[\chi_A \cdot X\right].$$

Definition 3.3.1 For any random variable $E\left[|X|\right] < \infty$ and any $A \in \mathcal{F}$ such that $\mathbf{P}(A) > 0$. The *conditional expectation* of X given A is defined by

$$E\left[X \mid A\right] = \frac{1}{\mathbf{P}(A)} \int_A X d\mathbf{P}. \tag{3.11}$$

■

Example 3.3.1 (Conditional Probability) Let χ_A be the **indicator function** of $A \in \mathcal{F}$

$$\chi_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases} \tag{3.12}$$

Then χ_A is a random variable on $(\Omega, \mathcal{F}, \mathbf{P})$. We take $B \in \mathcal{F}$ with $\mathbf{P}(B) > 0$. Then we have

$$E\left[\chi_A \mid B\right] = \frac{1}{\mathbf{P}(B)} \int_B \chi_A d\mathbf{P}$$

and by an exercise in chapter 1,

$$= \frac{1}{\mathbf{P}(B)} \int_\Omega \chi_A \cdot \chi_B d\mathbf{P}.$$

It holds that

$$\chi_A \cdot \chi_B = \chi_{A \cap B}$$

(check this !). Then

$$\int_\Omega \chi_A \cdot \chi_B d\mathbf{P} = \int_\Omega \chi_{A \cap B} d\mathbf{P}$$

$$= 0 \cdot \mathbf{P}\left((A \cap B)^c\right) + 1 \cdot \mathbf{P}\left(A \cap B\right) = \mathbf{P}\left(A \cap B\right).$$

Thus

$$E\left[\chi_A \mid B\right] = \frac{\mathbf{P}\left(A \cap B\right)}{\mathbf{P}(B)}. \tag{3.13}$$

The alert reader cannot but recognize the expression in the right hand side of (3.13) as the conditional probability of A given B, for which the symbol $\mathbf{P}\left(A \mid B\right)$ has been assigned in (3.1).

■

3.4 Conditioning w.r.t. a Partition

Let $\mathcal{P} = \{A_1, A_2, \ldots, A_k\}$ be a *partition* of Ω, i.e., $A_i \in \mathcal{F}$, $i = 1, 2, \ldots, k$, $A_i \cap A_j = \emptyset$, $j \neq i$ and $\cup_{i=1}^{k} A_i = \Omega$. We can call any set A_i a **cell** of $\mathcal{P}$. We assume that $\mathbf{P}(A_i) > 0$. The sigma-field $\sigma(\mathcal{P})$ generated by $\mathcal{P}$ is such that there are no subsets of any of A_i in $\sigma(\mathcal{P})$. We say that any A_i is an **atom** of $\sigma(\mathcal{P})$.

Consider the following schedule. Somebody chooses randomly (whatever this means in an operational sense) an $\omega \in \Omega$ and informs you about in which cell of $\mathcal{P}$, say A_i, ω lies. For example, the information might be the index of the cell. Thus our information is an outcome of the random variable

$$Y(\omega) = \sum_{i=1}^{k} i\chi_{A_i}(\omega).$$

We can thus say that having access to a partition means having access to a piece of information. The partitions are ordered by inclusion (are a lattice), in the sense that

$$\mathcal{P}_1 \subset \mathcal{P}_2$$

means that all cells in $\mathcal{P}_2$ have been obtained by partitioning of cells in $\mathcal{P}_1$. $\mathcal{P}_1$ is coarser than $\mathcal{P}_2$, and $\mathcal{P}_2$ is finer than $\mathcal{P}_1$, and $\mathcal{P}_2$ contains more information than $\mathcal{P}_1$.

Then we can compute the conditional expectation of X given A_i using (3.11) or

$$E[X \mid A_i] = \frac{1}{\mathbf{P}(A_i)} \int_{A_i} X d\mathbf{P}. \tag{3.14}$$

But hereby we have defined a random variable, by the assignment

$$\Omega \ni \omega \mapsto E[X \mid A_i], \quad \text{if } \omega \in A_i.$$

Then we can define, see [13, p. 495], the conditional expectation w.r.t. to a partition. We remind once more about the definition of the indicator function in (3.12).

Definition 3.4.1 The **conditional expectation given the information in partition** $\mathcal{P}$ is denoted by $E[X \mid \mathcal{P}]$, and is defined by

$$E[X \mid \mathcal{P}](\omega) \stackrel{\text{def}}{=} \sum_{i=1}^{k} \chi_{A_i}(\omega) E[X \mid A_i]. \tag{3.15}$$

■

The point to be harnessed from this is that $E[X \mid \mathcal{P}]$ is not a real number, but a random variable. In fact it is a simple random variable in the sense of section 1.8.1. We shall next pay attention to a few properties of $E[X \mid \mathcal{P}]$, which foreshadow more general conditional expectations.

(a) $E[X \mid \mathcal{P}]$ is measurable w.r.t. the sigma-field $\sigma(\mathcal{P})$ generated by $\mathcal{P}$, as $E[X \mid \mathcal{P}]$ is a constant on each partioning set.

(b) Take one of the partitioning sets A_j in $\sigma(\mathcal{P})$. Then

$$\int_{A_j} E[X \mid \mathcal{P}] d\mathbf{P} = \sum_{i=1}^{k} E[X \mid A_i] \int_{A_j} \chi_{A_i}(\omega) d\mathbf{P}(\omega)$$

$$= \sum_{i=1}^{k} E[X \mid A_i] \int_{\Omega} \chi_{A_j}(\omega) \cdot \chi_{A_i}(\omega) d\mathbf{P}(\omega) = E[X \mid A_j] \int_{\Omega} \chi_{A_j} d\mathbf{P}(\omega),$$

since $\chi_{A_i}(\omega) \cdot \chi_{A_j}(\omega) = 0$ for all ω, unless $i = j$, and $\left(\chi_{A_j}(\omega)\right)^2 = \chi_{A_j}(\omega)$ for all ω, and thus we get

$$= E\left[X \mid A_j\right] \int_{A_j} d\mathbf{P}(\omega) = E\left[X \mid A_j\right] \mathbf{P}\left(A_j\right).$$

By (3.14) this equals

$$= \int_{A_j} X d\mathbf{P}.$$

We summarize; the desired result is

$$\int_{A_j} E\left[X \mid \mathcal{P}\right] d\mathbf{P} = \int_{A_j} X d\mathbf{P}. \tag{3.16}$$

Our strategy is now to define conditional expectation in more general cases by extending the findings (a) and (b) (i.e., (3.16)) about $E\left[X \mid \mathcal{P}\right]$. The way of proceeding in the next section is necessary, because the restriction to $\mathbf{P}(A_i) > 0$ will make it impossible to construct conditional expectation by an approach, where the mesh of cells of the partition gets successively smaller (and the partition becomes finer and finer).

3.5 Conditioning w.r.t. a Random Variable

Definition 3.5.1 Let Y be a r.v. such that $E\left[\mid Y \mid\right] < \infty$, and let X be an arbitrary random variable. Then the **conditional expectation** Y **given** X, $E\left[Y \mid X\right]$, is a random variable such that

1. $E\left[Y \mid X\right]$ is $\mathcal{F}_X$ -measurable.
2. for any event $A \in \mathcal{F}_X$ we have

$$\int_A E\left[Y \mid X\right] d\mathbf{P} = \int_A Y d\mathbf{P}.$$

■

We shall say later a few words about the existence of $E\left[Y \mid X\right]$ as defined here.

We can define conditional probability of an event A given X by

$$\mathbf{P}(A \mid X) \stackrel{\text{def}}{=} E\left[\chi_A \mid X\right],$$

where χ_A is the indicator function (see eq. (3.12)) of the event A in Ω.

We shall need the following lemma that helps us in accepting that $E\left[Y \mid X\right]$ is unique almost surely.

Lemma 3.5.1 Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $\mathcal{G}$ be a sigma field contained in $\mathcal{F}$. If X is a $\mathcal{G}$ -measurable random variable and for any $B \in \mathcal{G}$

$$\int_B X d\mathbf{P} = 0, \tag{3.17}$$

then $X = 0$ almost surely (i.e., $\mathbf{P}\left(X = 0\right) = 1$).

Proof Take any $\varepsilon > 0$. Then $\mathbf{P}\left(X \geq \varepsilon\right) = 0$. This is seen as follows.

$$0 \leq \varepsilon \mathbf{P}\left(X \geq \varepsilon\right) = \int_{\{X \geq \varepsilon\}} \varepsilon d\mathbf{P} \leq \int_{\{X \geq \varepsilon\}} X d\mathbf{P} = 0.$$

The last equality is true by assumption, since $\{X \geq \varepsilon\} \in \mathcal{G}$. In the same way we have that $\mathbf{P}\left(X \leq -\varepsilon\right) = 0$. Therefore

$$\mathbf{P}\left(-\varepsilon \leq X \leq \varepsilon\right) = 1$$

for any $\varepsilon > 0$. Let us set

$$A_n = \left\{ -\frac{1}{n} < X < \frac{1}{n} \right\}.$$

Then $\mathbf{P}(A_n) = 1$, and since A_n is a decreasing sequence of events, $\{X = 0\} = \cap_{n=1}^{\infty} A_n$ and by continuity of probability from above (see theorem 1.4.9 in chapter 1)

$$\mathbf{P}(\{X = 0\}) = \lim_{n \to \infty} \mathbf{P}(A_n) = 1,$$

as was to be proved. ■

Note that the Doob-Dynkin theorem 1.5.5 in Chapter 1 implies that there is a Borel function H such that

$$E[Y \mid X] = H(X).$$

We can every now and then give more or less explicit formulas for H. One such case is investigated in section 3.6 that comes next.

3.6 A Case with an Explicit Rule for Conditional Expectation

The question of existence and uniqueness of $E[Y \mid X]$ may require deep theorems to be proved, but in many practical cases we can find an explicit formula, so that we can verify the conditions 1. and 2. in definition 3.5.1 above directly. We shall next present such a case.

Let us suppose that (X, Y) is a continuous bivariate random variable

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v) du dv.$$

We assume that $E[Y \mid X]$ exists. We shall show that the preceding definition 3.5.1 checks with the formula (3.7).

Theorem 3.6.1 Let Y be a r.v. such that $E[| Y |] < \infty$, and let X be a random variable such that (X, Y) has the joint density $f_{X,Y}$ on all $\mathbf{R} \times \mathbf{R}$. Then

$$E[Y \mid X = x] = \frac{\int_{-\infty}^{+\infty} y f_{X,Y}(x, y) dy}{\int_{-\infty}^{+\infty} f_{X,Y}(x, y) dy}. \tag{3.18}$$

Proof By virtue of definition 3.5.1 we need to find a Borel function, say $H(x)$, such that for any Borel event A we have

$$\int_{X \in A} H(X) d\mathbf{P} = \int_{X \in A} Y d\mathbf{P}. \tag{3.19}$$

Note that $\{X \in A\}$ is an event in $\mathcal{F}_X$. Let us start with the right hand side of (3.19). Since $A \in \mathcal{B}$,

$$\int_{X \in A} Y d\mathbf{P} = \int_{\Omega} \chi_A(X(\omega)) Y(\omega) d\mathbf{P}(\omega),$$

where $\chi_A(x)$ is the indicator of $A \in \mathcal{B}$, see (1.26), and one may compare with the idea in an exercise in Chapter 1. But we can write this in the usual notation

$$= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \chi_A(x) y dF_{X,Y}(x, y)$$

$$= \int_A \left(\int_{-\infty}^{+\infty} y f_{X,Y}(x, y) dy \right) dx. \tag{3.20}$$

Furthermore, in the left hand side of (3.19)

$$\int_{X\in A} H(X)d\mathbf{P} = \int_{\Omega} \chi_A(X(\omega))H(X(\omega))d\mathbf{P}\,(\omega)$$

$$= \int_{-\infty}^{+\infty} \chi_A(x)H(x)dF_X(x)$$

and as $dF_X(x) = f_X(x)dx = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dydx$

$$= \int_{-\infty}^{+\infty} \chi_A(x)H(x)\left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy\right)dx$$

$$= \int_A H(x)\left(\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy\right)dx. \tag{3.21}$$

Now, (3.19) requires that we can choose a Borel function $H(x)$ so that the expressions in (3.20) and (3.21) are equal, i.e.,

$$\int_A \left(\int_{-\infty}^{+\infty} yf_{X,Y}(x,y)dy\right)dx = \int_A H(x)\left(\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy\right)dx.$$

If these integrals are to coincide for each Borel set A, then we must take

$$H(x) = \frac{\int_{-\infty}^{+\infty} yf_{X,Y}(x,y)dy}{\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy},$$

which is the assertion in (3.18), as was to be proved. ■

3.7 Conditioning w.r.t. a σ -Field

Definition 3.7.1 Let Y be a r.v. such that $E\left[\mid Y \mid\right] < \infty$. Let $\mathcal{G}$ be a sub σ field of $\mathcal{F}$, i.e., $\mathcal{G} \subseteq \mathcal{F}$. Then the **conditional expectation** Y **given** $\mathcal{G}$, $E\left[Y \mid \mathcal{G}\right]$, is a random variable such that

1. $E\left[Y \mid \mathcal{G}\right]$ is $\mathcal{G}$-measurable.

2. for any event $A \in \mathcal{G}$ we have

$$\int_A E\left[Y \mid \mathcal{G}\right] d\mathbf{P} = \int_A Y d\mathbf{P}. \tag{3.22}$$

■

We do not prove that the random variable $E\left[Y \mid \mathcal{G}\right]$ exists, as the proof is beyond the scope of these notes/this course. The interested student can check, e.g., [103, p. 27] or [63, p. 200].

We have, when $\mathcal{F}_X$ is the σ field generated by X,

$$E\left[Y \mid \mathcal{F}_X\right] = E\left[Y \mid X\right],$$

hence this definition extends the definition 3.5.1.

In addition we can define the conditional probability

$$\mathbf{P}(A \mid \mathcal{G}) \overset{\text{def}}{=} E\left[\chi_A \mid \mathcal{G}\right]. \tag{3.23}$$

3.7.1 Properties of Conditional Expectation

The statements in the following theorem are the ranking tools for computing with conditional expectations. When a sigma field is generated by a random variable X, and thus $E[Y \mid \mathcal{F}_X] = E[Y \mid X]$, the properties below reduce back to the properties of $E[Y \mid X]$ in section 3.2 above. Thus, for example, (3.9) above is the rule of double expectation below.

In [20] the following basic properties of and useful rules for manipulation with the conditional expectation are given the nice descriptive names recapitulated below.

Theorem 3.7.1 *a and b are real numbers, $E[\mid Y \mid] < \infty$, $E[\mid Z \mid] < \infty$, $E[\mid X \mid] < \infty$ and $\mathcal{H} \subset \mathcal{F}$, $\mathcal{G} \subset \mathcal{F}$,*

1. **Linearity**:
$$E[aX + bY \mid \mathcal{G}] = aE[X \mid \mathcal{G}] + bE[Y \mid \mathcal{G}]$$

2. **Double expectation** :
$$E[E[Y \mid \mathcal{G}]] = E[Y]$$

3. **Taking out what is known**: If Z is $\mathcal{G}$ -measurable, and $E[\mid ZY \mid] < \infty$
$$E[ZY \mid \mathcal{G}] = ZE[Y \mid \mathcal{G}]$$

4. **An independent condition drops out**: If Y is independent of $\mathcal{G}$,
$$E[Y \mid \mathcal{G}] = E[Y]$$

5. **Tower Property** : If $\mathcal{H} \subset \mathcal{G}$,
$$E[E[Y \mid \mathcal{G}] \mid \mathcal{H}] = E[Y \mid \mathcal{H}]$$

6. **Positivity**: If $Y \geq 0$,
$$E[Y \mid \mathcal{G}] \geq 0.$$

All equalities and inequalities hold almost surely.

Proof See, e.g., [103, pp. 29-30].

1. The proof of linearity is more or less straightforward and is left as an exercise.

2. To prove the rule of double expectation we observe that by assumption the condition in (3.22) is to hold for all A in $\mathcal{G}$, hence it must hold for Ω. This means that
$$\int_\Omega E[Y \mid \mathcal{G}]\, d\mathbf{P} = \int_\Omega Y d\mathbf{P} = E[Y],$$
as claimed.

3. We start by verifying the result for $Z = \chi_B$ (see (3.12)), where $B \in \mathcal{G}$. In this special case we get
$$\int_A ZE[Y \mid \mathcal{G}]\, d\mathbf{P} = \int_A \chi_B E[Y \mid \mathcal{G}]\, d\mathbf{P} = \int_{A\cap B} E[Y \mid \mathcal{G}]\, d\mathbf{P} = \int_{A\cap B} Y d\mathbf{P},$$
where we used (3.22), since $A \cap B \in \mathcal{G}$, and
$$= \int_A \chi_B Y d\mathbf{P}.$$

On the other hand, by (3.22), the conditional expectation of $E\left[ZY \mid \mathcal{G}\right]$ satisfies

$$\int_A E\left[ZY \mid \mathcal{G}\right] d\mathbf{P} = \int_A E\left[\chi_B Y \mid \mathcal{G}\right] d\mathbf{P} = \int_A \chi_B Y d\mathbf{P}$$

Hence we have shown that

$$\int_A ZE\left[Y \mid \mathcal{G}\right] d\mathbf{P} = \int_A E\left[ZY \mid \mathcal{G}\right] d\mathbf{P}$$

for all $A \in \mathcal{G}$, and hence the lemma 3.5.1 about uniqueness says that

$$\int_A \left(ZE\left[Y \mid \mathcal{G}\right] - E\left[ZY \mid \mathcal{G}\right]\right) d\mathbf{P} = 0$$

implies

$$ZE\left[Y \mid \mathcal{G}\right] = E\left[ZY \mid \mathcal{G}\right]$$

almost surely.

We can proceed in the same manner to prove that the result holds for *step functions*

$$Z = \sum_{j=1}^{n} a_j \chi_{A_j},$$

where $A_j \in \mathcal{G}$ for $j = 1, 2, \ldots, m$. Finally, we approximate a general Z by a sequence of step functions (recall the operations in section 1.8.1 in Chapter 1).

4. Since we assume that Y is independent of $\mathcal{G}$, Y is independent of the random variable χ_A for all $A \in \mathcal{G}$. Due to (2.44) Y and χ_A have zero covariance, and this means by (2.42)

$$E\left[Y\chi_A\right] = E\left[Y\right] E\left[\chi_A\right].$$

Therefore

$$\int_A Y d\mathbf{P} = \int_\Omega \chi_A Y d\mathbf{P} = E\left[Y\chi_A\right] = E\left[Y\right] E\left[\chi_A\right]$$

$$= E\left[Y\right] \int_\Omega \chi_A d\mathbf{P} = E\left[Y\right] \int_A d\mathbf{P} = \int_A E\left[Y\right] d\mathbf{P},$$

since $E\left[Y\right]$ is a number, whereby, if we read this chain of inequalities from right to left

$$\int_A E\left[Y\right] d\mathbf{P} = \int_A Y d\mathbf{P},$$

for all A in $\mathcal{G}$. Comparison with (3.22) shows that this means

$$E\left[Y\right] = E\left[Y \mid \mathcal{G}\right].$$

5. We shall play a game with the definition. By the condition (3.22) we have again for all $A \in \mathcal{G}$ that

$$\int_A E\left[Y \mid \mathcal{G}\right] d\mathbf{P} = \int_A Y d\mathbf{P}. \tag{3.24}$$

With respect to $\mathcal{H}$, we get for all $A \in \mathcal{H}$

$$\int_A E\left[Y \mid \mathcal{H}\right] d\mathbf{P} = \int_A Y d\mathbf{P}. \tag{3.25}$$

But since $\mathcal{H} \subset \mathcal{F}$, $A \in \mathcal{H}$, we have thus in from (3.24) and (3.25)

$$\int_A E[Y \mid \mathcal{H}]\, d\mathbf{P} = \int_A E[Y \mid \mathcal{G}]\, d\mathbf{P},$$

which holds for all $A \in \mathcal{H}$. But when we check and apply the definition (3.22) once again, we get from this the conclusion that

$$E[E[Y \mid \mathcal{G}] \mid \mathcal{H}] = E[Y \mid \mathcal{H}],$$

as was to be proved.

6. We omit this.

■

Example 3.7.2 Taking out what is known This example is encountered in many situations. Let $H(x)$ be a Borel function. X and Y are random variables. Then the rule of taking out what is known gives

$$E[H(X) \cdot Y \mid \mathcal{F}_X] = H(X) \cdot E[Y \mid \mathcal{F}_X].$$

■

To get a better intuitive feeling for the tower property, which is an enormously versatile tool of computation, we recall example 1.5.4. There X is random variable and for a Borel function $Y = f(X)$. It was shown in loc.cit. that

$$\mathcal{F}_Y \subseteq \mathcal{F}_X.$$

Then the tower property tells us that for a random variable Z

$$E[E[Z \mid \mathcal{F}_X] \mid \mathcal{F}_Y] = E[Z \mid \mathcal{F}_Y].$$

How do we interpret this? We provide an answer to this question in section 3.7.4 below by using the interpretation of a conditional expectation $E[Y \mid X]$ as an estimator of Y by means of X.

3.7.2 An Application of the Properties of Conditional Expectation w.r.t. a σ-Field

Lemma 3.7.3 Le Y be a random variable that has the variance $\mathrm{Var}(Y) < \infty$ and let X be an another random variable (in the same probability space as Y). Set

$$\widehat{Y} = E[Y \mid \mathcal{F}_X]$$

and

$$\widetilde{Y} \stackrel{\text{def}}{=} Y - \widehat{Y}.$$

Then it holds that

$$\mathrm{Var}(Y) = \mathrm{Var}(\widehat{Y}) + \mathrm{Var}(\widetilde{Y}).$$

Proof We recall the well known formula, see, e.g. [15, p. 125],

$$\mathrm{Var}(Y) = \mathrm{Var}(\widehat{Y} + \widetilde{Y}) = \mathrm{Var}(\widehat{Y}) + \mathrm{Var}(\widetilde{Y}) + 2\mathrm{Cov}(\widehat{Y}, \widetilde{Y}).$$

We must investigate the covariance, which obviously must be equal to zero, if the our statement is to be true. We have

$$\mathrm{Cov}(\widehat{Y}, \widetilde{Y}) = E\left[\widehat{Y} \cdot \widetilde{Y}\right] - E\left[\widehat{Y}\right] \cdot E\left[\widetilde{Y}\right]. \tag{3.26}$$

Here we have, since $\widetilde{Y} = Y - \widehat{Y}$, that

$$E\left[\widehat{Y} \cdot \widetilde{Y}\right] = E\left[\widehat{Y}Y\right] - E\left[\widehat{Y}^2\right].$$

We use first the rule of double expectation (property 2.)

$$E\left[\widehat{Y}Y\right] = E\left[E\left[\widehat{Y}Y \mid \mathcal{F}_X\right]\right] =$$

and take out what is known (in $\mathcal{F}_X$) (property 3.)

$$= E\left[\widehat{Y}E\left[Y \mid \mathcal{F}_X\right]\right] = E\left[\widehat{Y}^2\right].$$

Therefore in (3.26)

$$E\left[\widehat{Y} \cdot \widetilde{Y}\right] = 0.$$

Furthermore

$$E\left[\widetilde{Y}\right] = E\left[Y - \widehat{Y}\right] = E\left[Y\right] - E\left[\widehat{Y}\right] =$$

and the rule of double expectation (property 2.)

$$= E\left[Y\right] - E\left[E\left[Y \mid \mathcal{F}_X\right]\right] = E\left[Y\right] - E\left[Y\right] = 0.$$

Thus even the second term in the right hand side of (3.26) is equal to zero. Thereby we have verified the claim as asserted. ■

3.7.3 Estimation Theory

There is an important interpretation of the quantities treated in lemma 3.7.3. We regard

$$\widehat{Y} = E\left[Y \mid \mathcal{F}_X\right] = E\left[Y \mid X\right]$$

as an **estimator of Y based on X**. Then $\widetilde{Y}$ is the **estimation error**

$$\widetilde{Y} = Y - \widehat{Y}.$$

In fact we should pay attention to the result in (3.46) in the exercises. This says that if $E\left[Y^2\right] < \infty$ and $E\left[(g(X))^2\right] < \infty$, where $H(x)$ is a Borel function, then

$$E\left[(Y - H(X))^2\right] = E\left[\mathrm{Var}(Y \mid X)\right] + E\left(\left(E\left[Y \mid X\right] - H(X)\right)^2\right). \tag{3.27}$$

This implies, since both terms in the right hand side are ≥ 0 that for all $H(x)$

$$E\left[\left(Y - E\left[Y \mid X\right]\right)^2\right] \leq E\left[(Y - H(X))^2\right] \tag{3.28}$$

In other words, $H^*(X) = \widehat{Y} = E\left[Y \mid X\right]$ is the optimal estimator of Y based on X, in the sense of minimizing the mean square error. The proof of lemma 3.7.3 above contains the following facts about optimal mean square estimation:

- $$E\left[\widetilde{Y}\right] = 0. \tag{3.29}$$

- the estimation error $\widetilde{Y}$ is uncorrelated with the estimator $\widehat{Y}$

$$\mathrm{Cov}(\widehat{Y}, \widetilde{Y}) = 0. \tag{3.30}$$

- The variance of Y can be decomposed as

$$\mathrm{Var}(Y) = \mathrm{Var}(\widehat{Y}) + E\left[\mathrm{Var}(Y \mid X)\right], \tag{3.31}$$

since by (3.27), (3.28) and by (3.29)

$$\mathrm{Var}(\widetilde{Y}) = E\left[\mathrm{Var}(Y \mid X)\right].$$

This framework yields a particularly effective theory of estimation (prediction, filtering, e.t.c. [90]), when later combined with the properties of Gaussian vectors and Gaussian stochastic processes.

3.7.4 Tower Property and Estimation Theory

Suppose now that X is random variable that we are planning to (or should) use in order to estimate Z, which is not observable to us. Unfortunately, we do not have direct data or observations of X either, but we have merely access to $Y = f(X)$, where f is not invertible. We could think of observing X via an A/D -converter (e.g, a hard limiter) or a clipping or both. The tower property tells us, as stated above, that

$$E\left[E\left[Z \mid \mathcal{F}_X\right] \mid \mathcal{F}_Y\right] = E\left[Z \mid \mathcal{F}_Y\right]. \tag{3.32}$$

Now we recall from the preceding section that $E\left[Z \mid \mathcal{F}_Y\right]$ is our best mean square estimate of Z based on Y. By the same account $\widehat{Z} = E\left[Z \mid \mathcal{F}_X\right]$ is the best mean square estimate of Z based on X. But then, of course, we have in the left hand side of (3.32),

$$\widehat{\widehat{Z}} = E\left[\widehat{Z} \mid \mathcal{F}_Y\right]$$

i.e., $\widehat{\widehat{Z}}$ is our best mean square estimate of $\widehat{Z}$ based on Y. Then we understand that the tower property (3.32) tells us simply that

$$\widehat{\widehat{Z}} = E\left[Z \mid \mathcal{F}_Y\right],$$

or, in other words, that our best mean square estimate of Z based on Y is in fact an estimate of $\widehat{Z}$! This is what is lost, when being forced to estimate Z using Y rather than X. The loss of information is also manifest in the inclusion $\mathcal{F}_Y \subset \mathcal{F}_X$.

3.7.5 Jensen's Inequality for Conditional Expectation

Theorem 3.7.4 Let $\varphi : \mathbf{R} \mapsto \mathbf{R}$ be convex function. Let X be a random variable such that $E\left[|X|\right] < \infty$ and that $E\left[|\varphi(X)|\right] < \infty$. Let $\mathcal{G} \subset \mathcal{F}$. Then

$$\varphi\left(E\left[X \mid \mathcal{G}\right]\right) \leq E\left[\varphi\left(X\right) \mid \mathcal{G}\right]. \tag{3.33}$$

Proof: is omitted, since it can be done as the proof of theorem 1.8.3 in chapter 1.. ■

3.8 Exercises

3.8.1 Easy Drills

1. A and B are two events with $\mathbf{P}(A) > 0$ and $\mathbf{P}(B) > 0$. $A \cap B = \emptyset$. Are A and B independent?

2. $\mathbf{P}(A \cap B) = 0.2$, $\mathbf{P}(A) = 0.6$ and $\mathbf{P}(B) = 0.5$.

   (a) Is $A \cap B = \emptyset$?

   (b) Are A and B independent?

   (c) Find $\mathbf{P}(A^c \cup B^c)$.

3. Given $\mathbf{P}(A \cap B^c) = 0.3$, $\mathbf{P}((A \cup B)^c) = 0.2$ and $\mathbf{P}(A \cap B) = 0.1$, find $\mathbf{P}(A \mid B)$.

4. If $\mathbf{P}(A \mid B) \leq \mathbf{P}(A)$, show that $\mathbf{P}(B \mid A) \leq \mathbf{P}(B)$.

5. A and B are two events with $\mathbf{P}((A \cup B)^c) = 0.6$ and $\mathbf{P}(A \cap B) = 0.1$. Let E be the event that either A or B but not both will occur. Find $\mathbf{P}(E \mid A \cup B)$.

6. A and B are two disjoint events. Show that

$$\mathbf{P}(A \mid A \cup B) = \frac{\mathbf{P}(A)}{\mathbf{P}(A) + \mathbf{P}(B)}.$$

3.8.2 Conditional Probability

1. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be probability space. Let $B \in \mathcal{F}$ and $\mathbf{P}(B) > 0$. Then we define for any $A \in \mathcal{F}$

$$\mathbf{P}^{\dagger}(A) \overset{\text{def}}{=} \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)},$$

or, $\mathbf{P}^{\dagger}(A) = \mathbf{P}(A \mid B)$ in (3.1). Show that $\left(\Omega, \mathcal{F}, \mathbf{P}^{\dagger}\right)$ is a probability space.

2. **The Chain Rule of Probability** Let $A_1, A_2, \ldots, A_n$, $n \geq 2$ be a events such that $\mathbf{P}\left(\cap_{i=1}^{n-1} A_i\right) > 0$. Show that

$$\mathbf{P}\left(\cap_{i=1}^{n} A_i\right) = \mathbf{P}\left(A_n \mid \cap_{i=1}^{n-1} A_i\right) \cdots \mathbf{P}\left(A_3 \mid A_2 \cap A_1\right) \mathbf{P}\left(A_2 \mid A_1\right) \mathbf{P}\left(A_1\right). \tag{3.34}$$

This rule is easy to prove and often omitted from courses in probability, but has its merits, as will be seen.

3. **Law of Total Probability** $\mathcal{P} = \{A_1, A_2, \ldots, A_k\}$ is a partition of Ω. Thus $A_i \in \mathcal{F}$, $i = 1, 2, \ldots, k$, $A_i \cap A_j = \emptyset$, $j \neq i$ and $\cup_{i=1}^{k} A_i = \Omega$. Show that for any event $B \in \mathcal{F}$

$$\mathbf{P}(B) = \sum_{i=1}^{k} \mathbf{P}\left(B \mid A_i\right) \mathbf{P}\left(A_i\right). \tag{3.35}$$

The expression is known as the law of total probability . How is this related to the expression in (3.15) ?

4. **Inverse Probability or Bayes' Formula** $\mathcal{P} = \{A_1, A_2, \ldots, A_k\}$ is a partition of Ω, i.e., $A_i \in \mathcal{F}$, $i = 1, 2, \ldots, k$, $A_i \cap A_j = \emptyset$, $j \neq i$ and $\cup_{i=1}^{k} A_i = \Omega$. Show that for any event $B \in \mathcal{F}$ and any A_l

$$\mathbf{P}\left(A_l \mid B\right) = \frac{\mathbf{P}\left(B \mid A_l\right) \mathbf{P}\left(A_l\right)}{\sum_{i=1}^{k} \mathbf{P}\left(B \mid A_i\right) \mathbf{P}\left(A_i\right)}, \quad l = 1, 2, \ldots, k. \tag{3.36}$$

The expression is nowadays known as Bayes' Formula or Rule, c.f. (3.6), but was in the past centuries called the rule of inverse probability.

5. $X \in \mathrm{Exp}(\lambda)$, $\lambda > 0$. Show that

$$\mathbf{P}(X > t + s \mid X > s) = \mathbf{P}(X > t). \tag{3.37}$$

This is known as the **lack of memory property** of the exponential distribution.

6. $X \in \mathrm{Fs}(p)$, $0 < p < 1$. Show that for every pair (k, m), $k = 0, 1, \ldots, m = 0, 1, 0, 1, \ldots,$

$$\mathbf{P}(X > m + k \mid X > m) = \mathbf{P}(X \geq k). \tag{3.38}$$

This is known as the **lack of memory property** of the first success distribution.

7. Let X_1 and X_2 be two independent r.v.'s with the same p.m.f. $p_X(k)$ on the positive integers, $k = 1, 2, \ldots,$. We know that $p_X(k) \leq c(< 1)$ for every k. Show that $\mathbf{P}(X_1 + X_2 = n) \leq c$.

8. $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of independent and identically distributed r.v.'s $\in \mathrm{Po}(2)$. N is independent of the X_n, and $N \in \mathrm{Po}(1)$. We consider the following **sum of a random number of random variables** $S_N = X_1 + X_2 + \ldots + X_N$, $S_0 = 0$. Find that

$$\mathbf{P}(S_N = 0) = e^{e^{-2}-1}. \tag{3.39}$$

The same formula will be derived using **generating functions** in an exercise of chapter 5.

9. The following is an idea in molecular biotechnology about a p.d.f. of p-values, when testing hypotheses of gene expressions in microarrays:

$$f(p) = \begin{cases} \lambda + (1-\lambda) \cdot a p^{a-1} & 0 < p < 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{3.40}$$

Here $0 < \lambda < 1$, and $0 < a < 1$. This distribution has been called the **BUM** distribution. The acronym **BUM** stands for **Beta-Uniform Mixture**. Find a generative model for the the **BUM** distribution.

10. $X \in U(0,1)$. Find $\mathbf{P}\left(X \leq x \mid X^2 = y\right)$.

11. **Poisson Plus Gauss Distribution** [33, p.327] Let $N \in \mathrm{Po}(\lambda)$, $X \in N\left(0, \sigma^2\right)$, N and X are independent. Set

$$U = N + X.$$

Show that the p.d.f. of U is

$$f_U(u) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}}{\sigma\sqrt{2\pi}} \frac{\lambda^k}{k!} e^{-\frac{(u-k)^2}{2\sigma^2}}. \tag{3.41}$$

12. This exercise is excerpted from [84, p. 145−146][1], and is in loc.cit. a small step in developing methods for treating measurements of real telephone traffic.

Let $N \in \mathrm{Po}(\lambda t)$. Let $T \mid N = n \in \mathrm{Erlang}\left(n, \frac{1}{s}\right)$, see example 2.2.10. Show that

$$E[T] = \lambda s t.$$

The model discussed in loc.cit. is the following. N is the number of phone calls coming to a telephone exchange during $(0, t]$. If $N = n$, then the total length of the n calls is T. Hence $E[T]$ is the expected size of the telephone traffic started during $(0, t]$.

[1] [84] is the Ph.D.-thesis (teknologie doktorsavhandling) from 1943 at KTH by Conrad 'Conny' Palm (1907−1951). Palm was an electrical engineer and statistician, recognized for several pioneering contributions to teletraffic engineering and queueing theory.

13. $X \in \mathrm{Exp}(1)$, $Y \in \mathrm{Exp}(1)$ are independent. Find the distribution of $X \mid X+Y=c$, where $c>0$ is a constant.

14. $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of independent and identically distributed r.v.'s $\in \mathrm{Be}\,(1/2)$. N is independent of the X_i's, and $N \in \mathrm{Po}(\lambda)$. We consider the r.v.'s

$$Y_1 = X_1 + X_2 + \ldots + X_N; Y_1 = 0, N = 0, \quad Y_2 = N - X_1.$$

Show that $Y_1 \in \mathrm{Po}\left(\frac{\lambda}{2}\right)$ and $Y_2 \in \mathrm{Po}\left(\frac{\lambda}{2}\right)$ and that they are independent.

15. (From [49]) Let $N \in \mathrm{Ge}(p)$ and set $X=(-1)^N$. Compute

   (a) $E[X]$ and $\mathrm{Var}[X]$. *Answer:* $\frac{p}{2-p}$, $\frac{4(1-p)}{(2-p)^2}$.

   (b) the p.m.f. of X. *Answer:* $p_X(1)=\frac{1}{2-p}$, $p_X(-1)=\frac{1-p}{2-p}$.

16. (From [30]) Given $\mathbf{P}(A)=a$ and $\mathbf{P}(B)=b$, show that $\frac{a+b-1}{b} \leq \mathbf{P}(A \mid B) \leq \frac{a}{b}$.

3.8.3 Joint Distributions & Conditional Expectations

1. (From [97]) Let (X,Y) is a bivariate random variable, where X is discrete and Y is continuous. (X,Y) has a joint probability mass - and density function given by

$$f_{X,Y}(k,y) = \begin{cases} \frac{\partial P(X=k, Y\leq y)}{\partial y} = \lambda\frac{(\lambda y)^k}{k!}e^{-2\lambda y} & \text{for } k=0,1,2,\ldots, \text{ and } y \in [0,\infty) \\ 0 & \text{elsewhere.} \end{cases}$$

   (a) Check that

$$\sum_{k=0}^{\infty}\int_0^{\infty} f_{X,Y}(k,y)dy = \int_0^{\infty}\sum_{k=0}^{\infty} f_{X,Y}(k,y)dy = 1.$$

   (b) Compute the mixed moment $E[XY]$ defined as

$$E[XY] = \sum_{k=0}^{\infty}\int_0^{\infty} ky f_{X,Y}(k,y)dy.$$

   Answer: $\frac{2}{\lambda}$.

   (c) Find the marginal p.m.f. of X. *Answer*: $X \in \mathrm{Ge}(1/2)$.

   (d) Compute the marginal density of Y here defined as

$$f_Y(y) = \begin{cases} \sum_{k=0}^{\infty} f_{X,Y}(k,y) & y \in [0,\infty) \\ 0 & \text{elsewhere.} \end{cases}$$

   Answer: $Y \in \mathrm{Exp}(1/\lambda)$.

   (e) Find

$$p_{X|Y}(k|y) = P(X=k|Y=y), k=0,1,2,\ldots,.$$

   Answer: $X|Y=y \in \mathrm{Po}(\lambda y)$.

   (f) Compute $E[X|Y=y]$ and then $E[XY]$ using double expectation. Compare your result with (b).

2. (From [35]) Let $X \in \mathrm{Po}(\lambda)$ and $Y \in \mathrm{Po}(\mu)$. X and Y are independent. Set $Z=X+Y$.

   (a) Find the conditional distribution $X \mid Z=z$. *Answer:* $X \mid Z=z \in \mathrm{Bin}\left(z, \frac{\lambda}{\lambda+\mu}\right)$.

(b) Find $E[X \mid Z=z]$, $E[X \mid Z]$, $\text{Var}[X \mid Z=z]$, $\text{Var}[X \mid Z]$. *Answer:* $z\frac{\lambda}{\lambda+\mu}$, $E[X \mid Z] = Z\frac{\lambda}{\lambda+\mu}$, $\text{Var}[X \mid Z=z] = z\frac{\lambda}{\lambda+\mu}\left(1-\frac{\lambda}{\lambda+\mu}\right)$, $\text{Var}[X \mid Z] = Z\frac{\lambda}{\lambda+\mu}\left(1-\frac{\lambda}{\lambda+\mu}\right)$.

(c) Find the coefficient of correlation $\rho_{X,Z}$,

$$\rho_{X,Z} = \frac{\text{Cov}(X,Z)}{\sqrt{\text{Var}[X]}\sqrt{\text{Var}[Z]}}.$$

Answer: $\sqrt{\frac{\lambda}{\lambda+\mu}}$.

3. (From [35]) $X \in \text{Exp}(\lambda)$ and $Y \in \text{U}(0,\theta)$. X and Y are independent. Find $\mathbf{P}(X > Y)$. *Answer:* $\frac{\lambda}{\theta}\left(1-e^{-\frac{\theta}{\lambda}}\right)$.

4. (From [35]) The joint distribution of (X,Y) is for $\beta > -1$ and $\alpha > -1$.

$$f_{X,Y}(x,y) = \begin{cases} c(\alpha,\beta)\, y^{\beta}(1-x)^{\alpha} & 0 \le x \le 1, 0 \le y \le x, \\ 0 & \text{elsewhere.} \end{cases}$$

(a) Determine $c(\alpha,\beta)$. *Aid:* Consider a suitable beta function, c.f., (2.31).

(b) Find the marginal distributions and compute $E[X]$, $\text{Var}[X]$.

(c) Determine $E[X \mid Y=y]$, $\text{Var}[X \mid Y=y]$, $E[Y \mid X=x]$, $\text{Var}[Y \mid X=x]$.

Answers

(a) $c(\alpha,\beta) = \frac{(\beta+1)\Gamma(\alpha+\beta+3)}{\Gamma(\alpha+1)\Gamma(\beta+2)}$.

(b)

$$f_X(x) = \frac{c(\alpha,\beta)}{\beta+1}x^{\beta+1}(1-x)^{\alpha}, \quad 0 \le x \le 1,$$

$$f_Y(y) = \frac{c(\alpha,\beta)}{\alpha+1}y^{\beta}(1-y)^{\alpha+1} \quad 0 \le y \le 1,$$

$$E[X] = \frac{\beta+2}{\alpha+\beta+3},$$

$$\text{Var}[X] = \frac{(\alpha+1)(\beta+2)}{(\alpha+\beta+4)((\alpha+\beta+3)^2}.$$

(c)

$$E[X \mid Y=y] = 1 - \frac{\alpha+1}{\alpha+2}(1-y),$$

$$\text{Var}[X \mid Y=y] = \frac{(\alpha+1)(1-y)^2}{(\alpha+3)(\alpha+2)^2}.$$

You obtain $E[Y \mid X=x]$, $\text{Var}[Y \mid X=x]$ from this by replacing y with $1-x$ and α with β and β with α.

5. (From [35]) Let $X_1, X_2, \ldots, X_n$ be independent and $\text{Po}(\lambda_i)$ -distributed random variables, respectively. Let the r.v. $I \in U(1,2,\ldots,n)$, c.f., Example 2.3.3. Find $E[X_I]$ and $\text{Var}[X_I]$. *Answer:* Let $\overline{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\lambda_i$. Then

$$E[X_I] = \overline{\lambda},$$

and

$$\text{Var}[X_I] = \overline{\lambda} - \overline{\lambda}^2 + \frac{1}{n}\sum_{i=1}^{n}\lambda_i^2.$$

6. (From [35]) Let $X_1, X_2, \ldots, X_n$ be independent and identically $\mathrm{Exp}(1/\lambda)$ -distributed random variables. Let in addition $S_0 = 0$ and $S_n = X_1 + X_2 + \ldots + X_n$. Set

$$N = \max\{n \mid S_n \leq x\}.$$

N is a **random time**, equal to the number of that random sample, when S_n for the last time stays under x. Then show that $N \in \mathrm{Po}\,(\lambda x)$.

7. Show using the properties of conditional expectation, that if X and Y are independent and the expectations exist, then

$$E\left[X \cdot Y\right] = E\left[X\right] \cdot E\left[Y\right]. \tag{3.42}$$

8. Let (X, Y) be a continuous bivariate r.v. with the joint p.d.f.

$$f_{X,Y}(x,y) = \begin{cases} c(x+y) & 0 < x < y < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

   (a) Find c.

   (b) Find $f_X(x)$ and $f_Y(y)$.

   (c) Find $E\left[X\right]$.

   (d) Find $E\left[X \mid Y = y\right]$.

   Answers: (a) $c = 2$, (b) $f_X(x) = 1 + 2x - 3x^2, 0 < x < 1$, $f_Y(y) = 3y^2, 0 < y < 1$. (c) $E\left[X\right] = \frac{5}{12}$, (d) $E\left[X \mid Y = y\right] = \frac{5}{9}y$.

9. Let (X, Y) be a continuous bivariate r.v. with the joint p.d.f. in (2.112). Find $f_{Y|X=x}(y)$. *Answer:*

$$f_{Y|X=x}(y) = \begin{cases} e^{(x-y)} & x < y \\ 0 & \text{elsewhere.} \end{cases}$$

10. Let $X \in \mathrm{Exp}\,(1/a)$, $Y \in \mathrm{Exp}\,(1/a)$ are independent. Show that $X \mid X + Y = z \in U(0, z)$.

11. Let $X \in \mathrm{Exp}\,(1)$, $Y \in \mathrm{Exp}\,(1)$ are independent. Show that $\frac{X}{X+Y} \in U(0, 1)$.

12. **Rosenblatt Transformation, PIT**[2] Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a continuous random vector with the joint distribution $F_{\mathbf{X}}\left(x_1, \ldots, x_n\right)$. We transform $(X_1, \ldots, X_n)$ to $(Y_1, \ldots, Y_n)$ by

$$Y_i = g_i\left(X_i\right),$$

where the transformations $y_i = g_i\left(x_i\right)$ are given by

$$\begin{aligned} y_1 &= g_1\left(x_1\right) = F_{X_1}\left(x_1\right) \\ y_2 &= g_2\left(x_2\right) = F_{X_2|X_1=x_1}\left(x_2\right) \\ &\vdots \\ y_n &= g_n\left(x_n\right) = F_{X_n|X_1=x_1,\ldots,X_{n-1}=x_{n-1}}\left(x_n\right). \end{aligned} \tag{3.43}$$

Note that we are using here an application of the chain rule (3.34).

Show that $(Y_1, \ldots, Y_n)$ are independent and that $Y_i \in U(0, 1)$, $i = 1, 2, \ldots, n$.

[2] The author thanks Dr. Thomas Dersjö from Scania, Södertälje for pointing out this.

In structural safety and solid mechanics this transformation is an instance of the **isoprobabilistic transformations** . In econometrics and risk management[3] this transformation is known as **PIT** = **probability integral transform**. PIT is applied for evaluating density forecasts[4] and assessing a model's validity. Thus the PIT is used for transforming joint probabilities for stochastic processes in discrete time. Here the arbitrariness of the ordering in $X_1, \ldots, X_n$, that is regarded as a difficulty of the Rosenblatt transformation, is automatically absent.

13. Let (X, Y) be a bivariate random variable, where both X and Y are binary, i.e., their values are 0 or 1. The p.m.f of (X, Y) is

$$p_{X,Y}(x,y) = \tau^x (1-\tau)^{1-x} \left(\theta^y (1-\theta)^{(1-y)}\right)^x \left(\lambda^y (1-\lambda)^{(1-y)}\right)^{1-x}, \quad x \in \{0,1\}, y \in \{0,1\}. \tag{3.44}$$

Here $0 \leq \tau \leq 1$, $0 \leq \theta \leq 1$, and $0 \leq \lambda \leq 1$.

(a) Check that $p_{X,Y}(x,y)$ is a p.m.f..

(b) Find the marginal p.m.f. $p_X(x)$.

(c) Find the p.m.f. $p_{Y|X=x}(y)$ for all $x \in \{0,1\}$, $y \in \{0,1\}$.

(d) What is the meaning of θ? What is the meaning of λ?

(e) Find $E[Y \mid X = x]$ for $x = 0$ and $x = 1$.

14. (From [49]) Let (X, Y) be a bivariate r.v. such that

$$Y \mid X = x \in \mathrm{Fs}(x), \quad f_X(x) = 3x^2, \quad 0 \leq x \leq 1.$$

Compute $E[Y]$, $\mathrm{Var}[Y]$, $\mathrm{Cov}(X,Y)$ and the p.m.f. of Y. *Answers:* $E[Y] = \frac{3}{2}$,$\mathrm{Var}[Y] = \frac{9}{4}$, $\mathrm{Cov}(X,Y) = -\frac{1}{8}$, and $p_Y(k) = \frac{18}{(k+3)(k+2)(k+1)k}$, $k \geq 1$.

3.8.4 Miscellaneous

1. (From [20]) Let A and B be sets in $\mathcal{F}$ and let χ_A and χ_B be the respective indicator functions, see equation (3.12). Assume that $0 < \mathbf{P}(B) < 1$. Show that

$$E[\chi_A \mid \chi_B](\omega) = \begin{cases} \mathbf{P}(A \mid B) & \text{if } \omega \in B \\ \mathbf{P}(A \mid B^c) & \text{if } \omega \notin B. \end{cases} \tag{3.45}$$

2. (From [20]) Let $B \in \mathcal{G}$, $\mathbf{P}(B) > 0$ and let X be such that $E[|X|] < \infty$. Show that

$$E[E[X \mid \mathcal{G}] \mid B] = E[X \mid B].$$

3. (From the Exam in sf2940, 23^{rd} of October 2007) $X \in \mathrm{Po}(\lambda)$. Show that $E\left[e^{tX} \mid X > 0\right] = \frac{e^{-\lambda}}{1-e^{-\lambda}}\left(e^{\lambda e^t} - 1\right)$.

4. **Mean Square Error** Let $H(x)$ be a Borel function and X random variable such that $E\left[(H(X))^2\right] < \infty$. Then show that

$$E\left[(Y - H(X))^2\right] = E[\mathrm{Var}(Y \mid X)] + E\left((E[Y \mid X] - H(X))^2\right). \tag{3.46}$$

[3] Hampus Engsner, when writing his M.Sc-thesis, pointed out PIT for the author.

[4] Diebold F.X., Gunther T.A & Tay A.S.: Evaluating density forecasts, 1997, National Bureau of Economic Research Cambridge, Mass., USA

Aid: We start with the identity

$$E\left[(Y - H(X))^2\right] = E\left[(Y - E[Y \mid X] + E[Y \mid X] - H(X))^2\right].$$

When we square this and compute the expectation we get

$$E\left[(Y - H(X))^2\right] = A + 2B + C, \tag{3.47}$$

where

$$A = E\left[(Y - E[Y \mid X])^2\right],$$

$$B = E\left[(Y - E[Y \mid X]) \cdot (E[Y \mid X] - H(X))\right],$$

and

$$C = E\left[(E[Y \mid X] - H(X))^2\right].$$

Now use double expectation for the three terms in the right hand side of (3.47). For A we get

$$E\left[E\left[(Y - E[Y \mid X])^2\right] \mid X\right],$$

for B

$$E\left[E\left[(Y - E[Y \mid X] \cdot E[Y \mid X] - H(X))\right] \mid X\right],$$

and for C,

$$E\left[E\left[(E[Y \mid X] - H(X))^2 \mid X\right]\right] = E\left[(E[Y \mid X] - H(X))^2\right],$$

where the known condition dropped out. Now show that the term B is $= 0$, and then draw the desired conclusion.

When one takes $H(X) = E[Y]$, a constant function of X, (3.46) yields the **law of total variance** in (3.10)

$$\text{Var}[Y] = E\left[(Y - E[Y])^2\right] = E[\text{Var}(Y \mid X)] + \text{Var}[E[Y \mid X]]. \tag{3.48}$$

5. (From [12]) Let X and Y be independent random variables and assume that $E\left[(XY)^2\right] < \infty$. Show that

$$\text{Var}[XY] = (E[X])^2 \text{Var}(Y) + (E[Y])^2 \text{Var}[X] + \text{Var}[Y]\,\text{Var}[X].$$

Aid: Set $Z = XY$, and then use the law of total variance, equation (3.48) above, via

$$\text{Var}[Z] = E[\text{Var}[Z \mid X]] + \text{Var}(E[Z \mid X]),$$

and continue using the properties of variance and conditional expectation.

6. The *linear estimator* $\widehat{Y}_L$, of Y by means of X, optimal in the mean square sense is given (as will be shown in section 7.5) by

$$\widehat{Y}_L = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_Y),$$

where $\mu_Y = E[Y]$, $\mu_X = E[X]$, $\sigma_Y^2 = \text{Var}[Y]$, $\sigma_X^2 = \text{Var}[X]$, $\rho = \frac{\text{Cov}(Y,X)}{\sigma_Y \cdot \sigma_X}$.

(a) Show that

$$E\left[\left(Y - \widehat{Y}_L\right)X\right] = 0. \tag{3.49}$$

This says that the optimal linear estimation error is *orthogonal* to X.

(b) Show that $Y - \widehat{Y}_L$ is uncorrelated with X.

(c) Show that if $\widehat{Y} = E\left[Y \mid X\right]$, then

$$E\left[\left(Y - \widehat{Y}\right) h(X)\right] = 0 \tag{3.50}$$

for any Borel function h such that $E\left[(h(X))^2\right] < \infty$.

7. (From [12])

(a) Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (I.I.D.) random variables and let

$$S = X_1 + X_2 + \ldots + X_n.$$

Show that

$$E\left[X_1 \mid S\right] = \frac{S}{n}. \tag{3.51}$$

(b) Let $X \in N(0, k)$, $W \in N(0, m)$ and be independent, where k and m are positive integers. Show that

$$E\left[X \mid X + W\right] = \frac{k}{k+m}(X + W).$$

Aid: The result in (a) can turn out to be helpful.

8. Let $X \in \mathrm{Pa}(k, \alpha)$ as in example 2.2.20. Show for $b > a > k$ that

$$P\left(X > b \mid X > a\right) = \left(\frac{a}{b}\right)^{\alpha}.$$

The term **scale-free** is used of any distribution (discrete or continuous or mixed) that looks essentially the same when looked at any scale, or such that

$$P\left(X > b \mid X > a\right)$$

depends only on the ratio a/b and not on the individual scales a and b. Zipf's law is also scale-free in this sense.

Recently the scale-free property has been observed for the degree distribution of many networks, where it is associated with the so-called small world phenomenon[5]. Examples are the World Wide Web, and human web of sexual contacts and many networks of interaction in molecular biology.

9. Let $N \in \mathrm{Po}\left(\frac{v^2}{2\sigma^2}\right)$. Let

$$X \mid N = n \in \chi^2(2n + 2).$$

Set $R = \sigma\sqrt{X}$. Compute directly the density of R and show that you obtain (2.114), i.e., $R \in \mathrm{Rice}\left(v, \sigma\right)$. *Aid:* You will eventually need a series expansion of a modified Bessel function of the first kind with order 0, as a real function see, e.g., [92, section 12.4][6] or [3, p. 288].

10. Assume that $X \mid P = p \in \mathrm{Ge}(p)$ $(= \mathrm{NBin}(1, p))$ and $P \in \beta\left(\alpha, \rho\right)$. Show that $X \in \mathrm{War}(\rho, \alpha)$, as defined in example 2.3.14. We apply here the Bayesian integral of (3.5). This fact should explain why the **Waring** distribution is known under the name **Negative-Binomial Beta** distribution.

[5] A small world network is a graph in which the distribution of connectivity is not confined to any scale and where every node can be reached from each other by a small number of steps.

[6] Or, see p.9 of **Formelsamling i Fysik**, Institutionen för teoretisk fysik, KTH, 2006 `http://courses.theophys.kth.se/SI1161/formelsamling.pdf`.

11. let $X \in N(0,1)$ and $Y \in N(0,1)$ and X and Y be independent. Take a real number λ. Set
$$Z = \begin{cases} Y, & \text{if } \lambda Y \geq X \\ -Y, & \text{if } \lambda Y < X. \end{cases}$$
Show that $Z \in \mathrm{SN}\,(\lambda)$. Hence we have here a generative model of $\mathrm{SN}\,(\lambda)$.

12. $X \in N(0,\sigma_x^2)$ and $f(x)$ is the p.d.f. of $N(0,\sigma_c^2)$. $U \in U(0,f(0))$ and is independent of X.
Show that $X \mid U \leq f(X) \in N\left(0,s^2\right)\big)$, where $\frac{1}{s^2} = \frac{1}{\sigma_x^2} + \frac{1}{\sigma_c^2}$.

13. $X \in \mathrm{SymBe}$. Let $X = x$ and $\rho \in [-1,1]$. Then set
$$Y = \begin{cases} x & \text{with probability } 1/2 + \rho/2 \\ -x & \text{with probability } 1/2 - \rho/2. \end{cases}$$
Show that $\mathrm{Cov}(X,Y) = \rho$.

3.8.5 Martingales

The exercises below are straightforward applications of the rules of computation in theorem 3.7.1 on a sequence of random variables with an assorted sequence of sigma fields, to be called martingales, and defined next.

Definition 3.8.1 Let $\mathcal{F}$ be a sigma field of subsets of Ω. Let for each integer $n > 0$ $\mathcal{F}_n$ be a sigma field $\subset \mathcal{F}$ and such that
$$\mathcal{F}_n \subset \mathcal{F}_{n+1}. \tag{3.52}$$
Then we call the family of sigma fields $(\mathcal{F}_n)_{n \geq 1}$ a **filtration**.

■

A important example of a filtration is given by
$$\mathcal{F}_n = \sigma\left(X_1, \ldots, X_n\right),$$
i.e., the sigma field generated by $X_1, \ldots, X_n$. This means intuitively that if $A \in \mathcal{F}$, then we are able to decide whether $A \in \mathcal{F}_n$ or $A \notin \mathcal{F}_n$ by observing $X_1, \ldots, X_n$.

Definition 3.8.2 Let $\mathbf{X} = (X_n)_{n=1}^{\infty}$ be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbf{P})$. Then we call $\mathbf{X}$ a **martingale with respect to the filtration** $(\mathcal{F}_n)_{n \geq 1}$, if

1. $E\left[\mid X_n \mid\right] < \infty$ for all n.
2. X_n is measurable with respect to $\mathcal{F}_n$ for each n.
3. For $n \geq 1$ the **martingale property** holds:
$$E\left[X_{n+1} \mid \mathcal{F}_n\right] = X_n. \tag{3.53}$$

■

The word martingale can designate several different things, besides the definition above. Martingale is, see figure 3.1[7], a piece of equipment that keeps a horse from raising its head too high, or, keeps the head in a constant position, a special collar for dogs and other animals and a betting system.

It is likely that the preceding nomenclature of probability theory is influenced by the betting system (which may have received its name from the martingale for horses ...).

[7] `http://commons.wikimedia.org/wiki/User:Malene`

Figure 3.1: Shannon Mejnert riding on Sandy in Baltic Cup Show on 28th of May 2006 at Kallehavegaard Rideklub, Randers in Denmark. The horse, Sandy, is wearing a **martingale**, which, quoting the experts, consists of: ..' *a strap attached to the girth and passes between the horse's front legs before dividing into two pieces. At the end of each of these straps is a small metal ring through which the reins pass.'*

1. Let $(\mathcal{F}_n)_{n\geq 1}$ be a filtration and $E\left[|X|\right] < \infty$. Set
$$X_n = E\left[X \mid \mathcal{F}_n\right].$$
Show that $(X_n)_{n\geq 1}$ is a martingale with respect to the filtration $(\mathcal{F}_n)_{n\geq 1}$.

2. Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent, nonnegative random variables with $E\left[X_n\right] = 1$ for every n. Let
$$M_0 = 1, \mathcal{F}_0 = (\Omega, \emptyset),$$
$$M_n = X_1 \cdot X_2 \cdot \ldots \cdot X_n,$$
and
$$\mathcal{F}_n = \sigma\left(X_1, \ldots, X_n\right).$$
Show that $(M_n)_{n\geq 0}$ is a martingale with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$.

3. Let $\mathbf{X}$ be a martingale with respect to the filtration $(\mathcal{F}_n)_{n\geq 1}$. Show that then for every n
$$E\left[X_{n+1}\right] = E\left[X_n\right] = \ldots = E\left[X_1\right].$$
(Recall that a martingale in the sense of figure 3.1 keeps the horse's head in a constant position.)

4. Let $\mathbf{X}$ be a martingale with respect to the filtration $(\mathcal{F}_n)_{n\geq 1}$. Show that then for every $n \geq m \geq 1$
$$E\left[X_n \mid \mathcal{F}_m\right] = X_m.$$

5. $\{X_n\}_{n=1}^{\infty}$ are independent and identically distributed with $E[X_n] = \mu$ and $\mathrm{Var}[X_n] = \sigma^2$. Define

$$W_0 = 0, W_n = \sum_{i=1}^{n} X_i,$$

$$\mathcal{F}_n = \sigma(X_1, \ldots, X_n),$$

and

$$S_n = (W_n - n\mu)^2 - n\sigma^2.$$

Show that $\{S_n\}_{n=0}^{\infty}$ is a martingale w.r.t. $\{\mathcal{F}_n\}_{n=0}^{\infty}$.

6. Let $\{X_n\}_{n=0}^{\infty}$ be a sequence of independent random variables. In many questions of statistical inference, signal detection e.t.c. there are two different probability distributions for $\{X_n\}_{n=0}^{\infty}$. Let now f and g be two distinct probability densities on the real line. The **likelihood ratio** L_n is defined as

$$L_n \overset{\text{def}}{=} \frac{f(X_0) \cdot f(X_1) \cdot \ldots \cdot f(X_n)}{g(X_0) \cdot g(X_1) \cdot \ldots \cdot g(X_n)}, \tag{3.54}$$

where we assume that $g(x) > 0$ for all x.

(a) Show that L_n is a martingale with respect to $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, if we think that g is the *p.d.f. of the true probability distribution* for $\{X_n\}_{n=0}^{\infty}$. That g is the the p.d.f. of the true probability distribution is here simply to be interpreted as the instruction to compute the required expectations using g. For example, for any Borel function H of X

$$E[H(X)] = \int_{-\infty}^{\infty} H(x)g(x)dx.$$

(b) Let

$$l_n = -\ln L_n.$$

The function $l(n)$ is known as the (- 1 ×) **loglikelihood** ratio. Show that (w.r.t. the p.d.f. of the true distribution g)

$$E[l_{n+1} \mid \mathcal{F}_n] \geq l_n.$$

Aid: Consider Jensen's inequality for conditional expectation in theorem 3.7.4.

7. **Stochastic Integrals** We say that $(X_n)_{n\geq 0}$ is **predictable**, if X_n is $\mathcal{F}_{n-1}$ measurable, where $(\mathcal{F}_n)_{n\geq 0}$ is a filtration. Let us define the **increment process** $\triangle X$ as

$$(\triangle X)_n \overset{\text{def}}{=} X_n - X_{n-1},$$

with the convention $X_{-1} = 0$.

(a) Show that a sequence of random variables $(X_n)_{n\geq 0}$ is a martingale if and only if

$$E[(\triangle X)_n \mid \mathcal{F}_{n-1}] = 0, \tag{3.55}$$

for $n = 0, 1, \ldots$.

(b) For any two random sequences $\mathbf{X} = (X_n)_{n\geq 0}$ and $\mathbf{M} = (M_n)_{n\geq 1}$ the **discrete stochastic integral** is a sequence defined by

$$(\mathbf{X} \star \mathbf{M})_n \overset{\text{def}}{=} \sum_{k=0}^{n} X_k (\triangle M)_k \tag{3.56}$$

Assume that $\mathbf{X}$ is predictable, $\mathbf{M}$ is a martingale and that $E[| X_k (\triangle M)_k |] < \infty$. Show that $(\mathbf{X} \star \mathbf{M})$ is a martingale.

Aid: Set $Z_n = \sum_{k=0}^{n} X_k (\triangle M)_k$ and find the expression for $(\triangle Z)_n$ and use (3.55).

8. **Why Martingales?** We have above worked on some examples of martingales with the verification of the martingale property as the main activity. Apart from the potential charm of applying the rules of conditional expectation, why are martingales worthy of this degree of attention? The answer is that there are several general results (the stopping theorem, maximal inequalities, convergence theorems e.t.c.) that hold for martingales. Thus, it follows by martingale convergence, e.g., that in (3.54) the likelihood ratio $L_n \to 0$ almost surely, as $n \to \infty$.

   What is the 'practical' benefit of knowing the the convergence $L_n \to 0$? *Aid:* Think of how you would use L_n to decide between $H_0 : X_i \in g$, $H_1 : X_i \in f$.

   More about martingales and their applications in statistics can be studied, e.g., in [102, ch. 9.2.]. Applications of martingales in computer science are presented in [79].

Chapter 4

Characteristic Functions

4.1 On Transforms of Functions

Several of the readers are presumably informed about the multifarious advances in science and engineering obtained by Fourier, Laplace, Mellin transforms and other transforms. Clumped together the aforementioned techniques constitute a branch of mathematics broadly referred to as transform theory. It would be very surprising, were transforms of some kind not to turn out to be important in probability theory, too.

Many of the transforms are integrals of an exponential function multiplied by a function $f(x)$ to be transformed. The key to success is that the exponential function converts sums into products. We set $i = \sqrt{-1}$ (so that $i^2 = -1$). In electrical engineering one writes $j = \sqrt{-1}$, but we do not follow this practice here.

1. The **Fourier transform** $\widehat{f}(t)$ of $f(x)$ is defined as

$$\widehat{f}(t) = \int_{-\infty}^{\infty} e^{-itx} f(x)dx. \tag{4.1}$$

This requires that f is integrable, or, that $\int_{-\infty}^{\infty} \mid f(x) \mid dx < +\infty$, [100, p.166].

The operation of Fourier transform in (4.1) can be understood as

$$f \overset{\mathcal{F}}{\mapsto} \widehat{f},$$

which means that a function of x is transformed to a (transform) function of t (=the transform variable).

Remark 4.1.1 The literature in mathematical physics and mathematical analysis uses often the definition

$$\widehat{f}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} f(x)dx.$$

There is also

$$\widehat{f}(t) = \int_{-\infty}^{\infty} e^{-2\pi itx} f(x)dx$$

widely used, with j in place of i, in electrical engineering. This state of affairs is without doubt a bit confusing. Of course, any variant of the definition can be converted to another by multiplying by the appropriate power of 2π, or by replacing t with $2\pi t$. When encountered with any document or activity involving the Fourier transform one should immediately identify, which particular definition is being used. We are, moreover, going to add to confusion by modifying (4.1) to define the characteristic function of a random variable.

■

2. The **Laplace transform** $\widehat{f}_{\mathcal{L}}(t)$ of $f(x)$ is

$$\widehat{f}_{\mathcal{L}}(t) = \int_0^{\infty} e^{-tx} f(x)dx.$$

This is a simplified formal expression, we are neglecting considerations of existence and the region of convergence, c.f., [100, p. 39].

3. The **Mellin transform** $\widehat{f}_{\mathcal{M}}(t)$ of $f(x)$ is

$$\widehat{f}_{\mathcal{M}}(t) = \int_0^{\infty} x^{t-1} f(x)dx.$$

Here t is a complex variable.

An important desideratum is that we should be able to uniquely recover f from $\widehat{f}$, or, that there should be an inverse transform. There is, under some conditions, see [100, p.171], the **Fourier inversion formula** given by

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itx} \widehat{f}(t)dt. \tag{4.2}$$

This is the operation

$$\widehat{f} \overset{\mathcal{F}^{-1}}{\mapsto} f. \tag{4.3}$$

Therefore we can talk in about unique Fourier transform pairs

$$\left(f, \widehat{f}\right),$$

which have in the past been collected in printed volumes of tables of Fourier transforms.

Since the distribution function

$$F_X(x) = \mathbf{P}\left(\{X \leq x\}\right).$$

completely determines the probabilistic behaviour and properties of a random variable X, we are obviously lead to work with transforms of $F_X(x)$, or more precisely, we deal with the transform of its p.d.f. $f_X(x)$, when it exists, or with the transforms of the probability mass function $p_X(x_k)$.

The Fourier transform exercises its impact by the fact that, e.g., differentiation and integration of f correspond to simple algebraic operations on $\widehat{f}$, see [100, Appendix C 4.]. Hence we can in many cases easily solve, e.g., differential equations in f by algebraic equations in the transform $\widehat{f}$ and then invert back to obtain the desired solution f. We shall meet with several applications of this interplay between the transformed function and its original function in probability theory, as soon as a suitable transform has been agreed upon.

For another illustration of the same point, the Mellin transform is important in probability theory for the fact that if X and Y are two independent non negative random variables, then the Mellin transform of the density of the product XY is equal to the product of the Mellin transforms of the probability densities of X and of Y. Or, if the Mellin transform of a probability density $f_X(x)$ of a r.v. $X \geq 0$ is $\widehat{f}_{\mathcal{M}_X}(t) = \int_0^{\infty} x^{t-1} f_X(x)dx$, then

$$\widehat{f}_{\mathcal{M}_{XY}}(t) = \widehat{f}_{\mathcal{M}_X}(t)\widehat{f}_{\mathcal{M}_Y}(t).$$

4.2 Characteristic Functions: Definition and Examples

4.2.1 Definition and Necessary Properties of Characteristic Functions

We begin with the formal definition.

Definition 4.2.1 (Characteristic Function) The characteristic function $\varphi_X(t)$ of the random variable X is for $t \in \mathbf{R}$ given by

$$\varphi_X(t) = E\left[e^{itX}\right] = \begin{cases} \sum\limits_{k=-\infty}^{\infty} e^{itx_k} p_X(x_k) & \text{discrete r.v.} \\ \int\limits_{-\infty}^{\infty} e^{itx} f_X(x)\, dx & \text{continuous r.v..} \end{cases} \tag{4.4}$$

■

This is the complex conjugate of the Fourier transform, needless to say. Let us recall that $e^{itx} = \cos(tx) + i\sin(tx)$. Then we have

$$E\left[e^{itX}\right] = E\left[\cos(tX)\right] + iE\left[\sin(tX)\right].$$

We can regard the right hand side of the last expression as giving meaning to the expectation of the complex random variable e^{itX} in terms of expectations of two real random variables. By definition of the modulus of a complex number $\mid e^{itx} \mid = \sqrt{\cos^2(tx) + \sin^2(tx)} = \sqrt{1}$. Therefore

$$E\left[\mid e^{itX} \mid\right] = 1, E\left[\mid e^{itX} \mid^2\right] = 1.$$

Hence the function e^{itx} is integrable (w.r.t. to dF_X), and $\varphi_X(t)$ **exists for all** t. In other words, **every distribution function/random variable has a characteristic function**.

We are thus dealing with an operation that transforms, e.g., a probability density f_X (or probability mass function) to a complex function $\varphi_X(t)$,

$$f_X \overset{\text{Ch}}{\mapsto} \varphi_X.$$

The following theorem deals with the inverse of a characteristic function.

Theorem 4.2.1 If the random variable X has the characteristic function $\varphi_X(t)$, then for any interval $(a, b]$

$$P(a < X < b) + \frac{P(X = a) + P(X = b)}{2} = \lim_{T \to +\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t) dt.$$

Proof: The interested reader can prove this by a modification of the proof of the Fourier inversion theorem found in [100, p.172-173]. ■

Here we have in other words established that there is the operation

$$\varphi_X \overset{\text{Ch}^{-1}}{\mapsto} f_X.$$

The following theorem is nothing but a simple consequence of the preceding explicit construction of the inverse.

Theorem 4.2.2 (Uniqueness) If two random variables X_1 and X_2 have the same characteristic functions, i.e.,

$$\varphi_{X_1}(t) = \varphi_{X_2}(t) \quad \text{for all } t \in \mathbf{R},$$

then they have the same distribution functions

$$F_{X_1}(x) = F_{X_2}(x) \quad \text{for all } x \in \mathbf{R},$$

which we write as

$$X_1 \overset{\text{d}}{=} X_2.$$

■

There are several additional properties that follow immediately from the definition of the characteristic function.

Theorem 4.2.3 (a) $\varphi_X(t)$ exists for any random variable.

(b) $\varphi_X(0) = 1$.

(c) $\mid \varphi_X(t) \mid \leq 1$.

(d) $\varphi_X(t)$ is uniformly continuous.

(e) The characteristic function of $a + bX$, where a and b are real numbers, is

$$\varphi_{a+bX}(t) = e^{iat}\varphi_X(bt).$$

(f) The characteristic function of $-X$ is the complex conjugate $\overline{\varphi}_X(t)$.

(g) The characteristic function is real valued if and only if $X \overset{\mathrm{d}}{=} -X$ (the distribution of X is symmetric about zero).

(h) For any n, any complex numbers z_l, $l = 1, 2, \ldots, n$, and any real t_l, $l = 1, 2, \ldots, n$ we have

$$\sum_{l=1}^{n}\sum_{k=1}^{n} z_l \overline{z}_k \varphi_X\left(t_l - t_k\right) \geq 0. \tag{4.5}$$

■

Proof:

(a) This was proved above.

(b) $e^{itX} \mid_{t=0} = e^0 = 1$. We have $\varphi_X(0) = E\left[e^0\right] = 1$.

(c) This is a part of the proof of (a).

(d) Let us pause to think what we are supposed to prove. A function $\varphi_X(t)$ is by definition *uniformly continuous* in $\mathbf{R}$ [69, p. 68], if it holds that for all $\epsilon > 0$, there exists a $\delta > 0$ such that $|\varphi_X(t+h) - \varphi_X(t)| \leq \epsilon$ for all $|h| \leq \delta$ and all $t \in \mathbf{R}$. The point is that δ is independent of t, i.e., that δ depends only on ϵ. In order to prove this let us assume, without restriction of generality that $h > 0$. Then we have

$$|\varphi_X(t+h) - \varphi_X(t)| = \mid E\left[e^{itX}\left(e^{ihX} - 1\right)\right] \mid \leq E\left[\mid e^{itX}\left(e^{ihX} - 1\right) \mid\right]$$

$$\leq E\left[\underbrace{\mid e^{itX} \mid}_{=1} \mid \left(e^{ihX} - 1\right) \mid\right] = E\left[\mid \left(e^{ihX} - 1\right) \mid\right].$$

From the expression in the right hand side the claim about uniform continuity is obvious, if $E\left[\mid \left(e^{ihX} - 1\right) \mid\right] \to 0$, as $h \to 0$, since we can then make $E\left[\mid \left(e^{ihX} - 1\right) \mid\right]$ arbitrarily small by choosing h sufficiently small independently of t.

It is clear that $e^{ihX} - 1 \to 0$ (almost surely), as $h \to 0$. Since $\mid \left(e^{ihX} - 1\right) \mid \leq 2$, we can apply dominated convergence theorem 1.8.7 to establish $E\left[\mid \left(e^{ihX} - 1\right) \mid\right] \to 0$. Hence we have proved the assertion in part (d).

(e) The characteristic function of $a+bX$ is by definition

$$\varphi_{a+bX}(t) = E\left[e^{it(a+bX)}\right] = E\left[e^{ita}e^{itbX}\right]$$

$$= e^{ita}E\left[e^{itbX}\right] = e^{iat}\varphi_X(bt).$$

(f) The characteristic function of $-X$ is by (e) with $a=0$ and $b=-1$ equal to

$$\varphi_{-X}(t) = \varphi_X(-t) = E\left[e^{-itX}\right]$$

$$= E\left[\cos(-tX)\right] + iE\left[\sin(-tX)\right] = E\left[\cos(tX)\right] - iE\left[\sin(tX)\right],$$

where we used $\cos(-x) = \cos(x)$ and $\sin(-x) = -\sin(x)$,

$$= \overline{E\left[\cos(tX)\right] + iE\left[\sin(tX)\right]},$$

where $\overline{z}$ stands for the conjugate of the complex number z, and then

$$= \overline{E\left[e^{itX}\right]} = \overline{\varphi}_X(t).$$

(g) Let us first suppose that the characteristic function of X is real valued, which implies that $\overline{\varphi}_X(t) = \varphi_X(t)$. But we have found in the proof of (f) that $\overline{\varphi}_X(t)$ is the characteristic function of $-X$. By uniqueness of the characteristic functions, theorem 4.2.2 above, this means that $X \overset{\text{d}}{=} -X$, as was to be shown.

Let us next suppose that $X \overset{\text{d}}{=} -X$. Then $\varphi_X(t) = \varphi_{-X}(t)$ and by (f) $\varphi_{-X}(t) = \overline{\varphi}_X(t)$, and therefore $\varphi_X(t) = \overline{\varphi}_X(t)$, and the characteristic function of X is real valued.

(h) Take any n and complex numbers z_l and real t_l, $l = 1, 2, \ldots, n$. Then we write using the properties of complex numbers and the definition of φ_X

$$\sum_{l=1}^{n}\sum_{k=1}^{n} z_l\overline{z}_k\varphi_X\left(t_l - t_k\right) = \sum_{l=1}^{n}\sum_{k=1}^{n} z_l\overline{z}_k E\left[e^{i(t_l-t_k)X}\right]$$

$$= \sum_{l=1}^{n}\sum_{k=1}^{n} E\left[z_l e^{it_lX}\overline{z}_k e^{-it_kX}\right] = E\left[\sum_{l=1}^{n}\sum_{k=1}^{n} z_l e^{it_lX}\overline{z}_k e^{-it_kX}\right]$$

$$= E\left[\sum_{l=1}^{n}\sum_{k=1}^{n} z_l e^{it_lX}\overline{z_k e^{it_kX}}\right] = E\left[\sum_{l=1}^{n} z_l e^{it_lX}\overline{\sum_{k=1}^{n} z_k e^{it_kX}}\right]$$

and as $\mid w \mid^2 = w\cdot\overline{w} \geq 0$ for any complex number w,

$$= E\left[\mid \sum_{l=1}^{n} z_l e^{it_lX} \mid^2\right] \geq 0,$$

which proves (4.5).

■

The properties (a)-(h) in the preceding theorem are **necessary** conditions, i.e., they will be fulfilled, if a function is a characteristic function of a random variable. The condition (h), i.e., (4.5) says that a characteristic function is **non negative definite**.

There are several sets of **necessary and sufficient** conditions for a complex valued function to be a characteristic function of some random variable. One of these is known as *Bochner's theorem.* This theorem states that an arbitrary complex valued function φ is the characteristic function of some random variable if and only if (i) -(iii) hold, where (i) φ is non-negative definite, (ii) φ is continuous at the origin, (iii) $\varphi(0) = 1$. Unfortunately the condition (i),i.e., (4.5) is in practice rather difficult to verify.

4.2.2 Examples of Characteristic Functions

Example 4.2.4 (Standard Normal Distribution) $X \in N(0,1)$. The p.d.f. of X is, as stated,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < +\infty. \tag{4.6}$$

Then by (4.4)

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

and if we are allowed to move differentiation w.r.t. t inside the integral sign, we get

$$\varphi_X^{(1)}(t) = \frac{d}{dt}\varphi_X(t) = \int_{-\infty}^{\infty} \frac{d}{dt} e^{itx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

$$= \int_{-\infty}^{\infty} ixe^{itx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \int_{-\infty}^{\infty} -ie^{itx} \frac{1}{\sqrt{2\pi}} \left(-xe^{-x^2/2}\right) dx =$$

and by integration by parts we obtain

$$= -ie^{itx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mid_{-\infty}^{+\infty} - \int_{-\infty}^{\infty} \left(-ti^2 e^{itx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right) dx$$

$$= 0 - t\varphi_X(t).$$

In other words we have encountered the differential equation $\varphi_X^{(1)}(t) + t\varphi_X(t) = 0$. This equation has the integrating factor $e^{t^2/2}$, or, in other words we have the equation

$$\frac{d}{dt}\left(e^{t^2/2}\varphi_X(t)\right) = 0.$$

We solve this with $\varphi_X(t) = Ce^{-t^2/2}$. Since $\varphi_X(0) = 1$ by (b) in theorem 4.2.3 above, we get $C = 1$. Thus we have obtained the result

$$X \in N(0,1) \Leftrightarrow \varphi_X(t) = e^{-t^2/2}. \tag{4.7}$$

We observe that $e^{-t^2/2}$ is a real valued function. Hence theorem 4.2.3 (g) shows that if $X \in N(0,1)$, then $-X \in N(0,1)$, which is also readily checked without transforms. Indeed,

$$\mathbf{P}(-X \leq x) = \mathbf{P}(X \geq -x) = 1 - \mathbf{P}(X \leq -x)$$

$$= 1 - \Phi(-x) = 1 - (1 - \Phi(x)) = \Phi(x) = \mathbf{P}(X \leq x),$$

where we used a well known property of $\Phi(x)$, which in its turn rests upon the fact that $\phi(-x) = \phi(x)$. Actually we have by this provided the solution to an exercise in section 2.6.2.

■

Example 4.2.5 (Normal Distribution $X \in N(\mu, \sigma^2)$) Let $Z \in N(0,1)$ and set $X = \sigma Z + \mu$, where $\sigma > 0$ and μ is an arbitrary real number. Then we find that

$$F_X(x) = P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

where $\Phi(x)$ is the distribution function of $Z \in N(0,1)$ and $\frac{d}{dx}\Phi(x) = \phi(x)$. Thus we obtain by by (4.6) that

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}.$$

Hence $X \in N\left(\mu, \sigma^2\right)$. But by (e) in theorem 4.2.3 we have

$$\varphi_X(t) = \varphi_{\sigma Z+\mu}(t) = e^{i\mu t}\varphi_Z(\sigma t) = e^{i\mu t}e^{-\frac{\sigma^2 t^2}{2}},$$

where we used (4.7). Without any doubt we have shown that

$$X \in N(\mu, \sigma^2) \Leftrightarrow \varphi_X(t) = e^{i\mu t - \frac{\sigma^2 t^2}{2}}. \tag{4.8}$$

■

Example 4.2.6 (Poisson Distribution) Let $X \in \text{Po}(\lambda)$, $\lambda > 0$. Due to definition (4.4)

$$\varphi_X(t) = \sum_{k=0}^{\infty} e^{itk}e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\left(e^{it}\lambda\right)^k}{k!} = e^{-\lambda}e^{e^{it}\lambda}$$

$$= e^{\lambda\left(e^{it}-1\right)},$$

where we invoked the standard series expansion of e^z for any complex z. In other words, we have found the following:

$$X \in \text{Po}(\lambda) \Leftrightarrow \varphi_X(t) = e^{\lambda\left(e^{it}-1\right)}. \tag{4.9}$$

■

Some of the next few examples are concerned with the continuous case of the definition by evaluating the integral in (4.4). The reader with a taste for mathematical rigor may become constersed for the fact that we will be proceeding as if everything was real valued. This is a pragmatic simplification of the presentation, and the results in the cases below will equal those obtained, when using a more rigorous approach. The computation of Fourier transforms and inverse Fourier transforms can then, of course, require contour integration and residue calculus, which we do not enter upon in the main body of the text. An exception is the section on Mellin transforms.

Example 4.2.7 (Exponential Distribution) Let $X \in \text{Exp}(\lambda)$, $\lambda > 0$. By definition (4.4)

$$\varphi_X(t) = E\left[e^{itX}\right] = \int_0^{\infty} e^{itx}\frac{1}{\lambda}e^{-x/\lambda}\,dx$$

$$= \frac{1}{\lambda}\int_0^{\infty} e^{-x((1/\lambda)-it)}\,dx = \frac{1}{\lambda}\left[\frac{-1}{((1/\lambda)-it)}e^{-x((1/\lambda)-it)}\right]_0^{\infty} = \frac{1}{\lambda}\frac{1}{(it-(1/\lambda))} = \frac{1}{(1-i\lambda t)}.$$

Thus we have

$$X \in \text{Exp}(\lambda) \Leftrightarrow \varphi_X(t) = \frac{1}{1-i\lambda t}. \tag{4.10}$$

■

Example 4.2.8 (Laplace Distribution) $X \in L(1)$ says that X has the p.d.f.

$$f_X(x) = \frac{1}{2}e^{-|x|}, \quad -\infty < x < +\infty. \tag{4.11}$$

The definition in (4.4) gives

$$\varphi_X(t) = \frac{1}{2}\int_{-\infty}^{\infty} e^{itx}e^{-|x|}\,dx$$

We compute the integral by applying the definition of $|x|$ to get

$$\int_{-\infty}^{\infty} e^{itx}e^{-|x|}\,dx = \int_{-\infty}^{0} e^{itx}e^{x}\,dx + \int_{0}^{\infty} e^{itx}e^{-x}\,dx. \tag{4.12}$$

We change the variable $x = -u$, in the first integral in the right hand side of (4.12), which yields

$$\int_{-\infty}^{0} e^{itx}e^{x}\,dx = \int_{\infty}^{0} e^{-itu}e^{-u}\,(-1)du$$

$$= \int_{0}^{\infty} e^{-itu}e^{-u}\,du = \overline{\int_{0}^{\infty} e^{itu}e^{-u}\,du},$$

which is seen to be the complex conjugate of the second integral in the right hand side of (4.12). This second integral is in its turn recognized from the directly preceding example as the characteristic function of Exp(1). Thus we get by (4.10)

$$\int_{0}^{\infty} e^{itx}e^{-x}\,dx = \frac{1}{1-it}.$$

Hence

$$\frac{1}{2}\int_{-\infty}^{\infty} e^{itx}e^{-|x|}\,dx = \frac{1}{2}\left(\overline{\frac{1}{1-it}} + \frac{1}{1-it}\right)$$

$$= \frac{1}{2}\left(\frac{1}{1+it} + \frac{1}{1-it}\right) = \frac{1}{2}\left(\frac{1+it+1-it}{1+t^2}\right) = \frac{1}{1+t^2}.$$

In summary,

$$X \in L(1) \Leftrightarrow \varphi_X(t) = \frac{1}{1+t^2}. \tag{4.13}$$

The theorem 4.2.3 (g) shows that if $X \in L(1)$, then $X \overset{d}{=} -X$.

■

Example 4.2.9 ($X \in \mathrm{Exp}(1)$, $Y \in \mathrm{Exp}(1)$, X and Y independent, Distribution of $X - Y$) Let $X \in \mathrm{Exp}(1)$, $Y \in \mathrm{Exp}(1)$. In addition, X and Y are assumed independent. We want to find the distribution of $X - Y$. The rules of computation with characteristic functions above entail

$$\varphi_{X-Y}(t) = \varphi_X(t)\cdot\varphi_{-Y}(t) = \varphi_X(t)\cdot\varphi_Y(-t),$$

and by (4.10)

$$= \frac{1}{1-it}\cdot\frac{1}{1-i(-t)} = \frac{1}{1-it}\cdot\frac{1}{1+it}$$

$$= \frac{1}{1+t^2}.$$

Here a reference to (4.13) gives that if $X \in \mathrm{Exp}(1)$, $Y \in \mathrm{Exp}(1)$, X and Y independent, then

$$X - Y \in L(1). \tag{4.14}$$

We have $X - Y \overset{d}{=} Y - X$, too.

■

Example 4.2.10 (Gamma Distribution) Let $X \in \Gamma(p, a)$, $p > 0$, $a > 0$. The p.d.f. is

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(p)} \frac{x^{p-1}}{a^p} e^{-x/a} & 0 \le x \\ \\ 0 & x < 0. \end{cases} \tag{4.15}$$

By definition (4.4)

$$\varphi_X(t) = \int_0^\infty e^{itx} \frac{1}{\Gamma(p)} \frac{x^{p-1}}{a^p} e^{-x/a}\, dx$$

$$= \int_0^\infty \frac{1}{\Gamma(p)} \frac{x^{p-1}}{a^p} e^{-x((1/a)-it)}\, dx.$$

We change the variable $u = x\left((1/a) - it\right) \leftrightarrow x = u/\left((1/a) - it\right)$ and get

$$= \frac{1}{a^p} \frac{1}{\Gamma(p)} \int_0^\infty \frac{u^{p-1}}{((1/a)-it)^{p-1}} e^{-u} \frac{du}{((1/a)-it)}$$

$$= \frac{1}{a^p} \frac{1}{\Gamma(p)} \frac{1}{((1/a)-it)^p} \int_0^\infty u^{p-1} e^{-u}\, du.$$

By definition of the Gamma function $\Gamma(p) = \int_0^\infty u^{p-1} e^{-u}\, du$, and the desired characteristic function is

$$= \frac{1}{a^p} \frac{1}{((1/a)-it)^p} = \frac{1}{(1-iat)^p}.$$

Thus we have found that

$$X \in \Gamma(p, a) \Leftrightarrow \varphi_X(t) = \frac{1}{(1-iat)^p}. \tag{4.16}$$

■

Example 4.2.11 (Standard Cauchy) $X \in C(0, 1)$ says that X is a continuous r.v., and has the p.d.f.

$$f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2}, \quad -\infty < x < \infty. \tag{4.17}$$

We are going to find the characteristic function of $X \in C(0, 1)$ by the **duality argument** or the **symmetry property** of the Fourier transforms, see [100, p. 252]. Since all transforms involved are real, we have no difficulty for the fact that the characteristic function is the complex conjugate of the Fourier transform.

Remark 4.2.1 The **symmetry** or **duality** property of the Fourier transform in (4.1) is as follows.

$$\text{If } f(x) \overset{\mathcal{F}}{\to} \widehat{f}(t), \text{ then } \widehat{f}(x) \overset{\mathcal{F}}{\to} 2\pi f(-t).$$

■

By (4.13) we know that

$$X \in L(1) \Leftrightarrow \varphi_X(t) = \frac{1}{1+t^2},$$

Let us hence apply the symmetry property first with $\varphi_X(x) = \frac{1}{1+x^2}$. Then the symmetry property tells that

$$\varphi_X(x) = \frac{1}{1+x^2} \overset{\text{Ch}}{\to} 2\pi \cdot f_X(-t) = 2\pi \cdot \frac{1}{2}e^{-|-t|} = \pi e^{-|t|}.$$

But it is an obvious property of scaling of the Fourier transform (by (4.1)) that if $f(x) \overset{\mathcal{F}}{\to} \widehat{f}(t)$, then $af(x) \overset{\mathcal{F}}{\to} a\widehat{f}(t)$ for any real constant a. By the scaling $a = 1/\pi$ we get

$$\frac{1}{\pi}\varphi_X(x) = \frac{1}{\pi}\frac{1}{1+x^2} \overset{\text{Ch}}{\to} \frac{1}{\pi}\pi e^{-|t|} = e^{-|t|}.$$

$$X \in C(0,1) \Leftrightarrow \varphi_X(t) = e^{-|t|}. \tag{4.18}$$

Once more we find that $X \overset{d}{=} -X$.

■

Example 4.2.12 (Point Mass Distribution) For the purposes of several statements in the sequel we introduce a probability mass function with a notation reminiscent of the Dirac pulse.

$$\delta_c(x) = \begin{cases} 1 & x = c \\ 0 & x \neq c. \end{cases} \tag{4.19}$$

Then δ_c is a distribution such that all mass is located at c. In the terminology of appendix 2.5 δ_c defines a purely discrete measure with one atom at c. Then, if $X \in \delta_c$,

$$\varphi_X(t) = e^{itc}. \tag{4.20}$$

■

Example 4.2.13 (Bernoulli Distribution) Let $X \in \text{Be}(p)$. Here $p = P(X = 1)$. Then we apply again the discrete case of the definition (4.4) and get

$$\varphi_X(t) = E\left[e^{itX}\right] = e^{it0}(1-p) + e^{it}p = (1-p) + e^{it}p.$$

$$X \in \text{Be}(p) \Leftrightarrow \varphi_X(t) = (1-p) + e^{it}p. \tag{4.21}$$

■

Example 4.2.14 (Symmetric Bernoulli Distribution) The characteristic function of $X \in$ SymBe with p.m.f. in (2.50) is computed as

$$\varphi_X(t) = E\left[e^{itX}\right] = e^{-it}\frac{1}{2} + e^{it}\frac{1}{2} = \cos(t).$$

$$X \in \text{SymBe} \Leftrightarrow \varphi_X(t) = \cos(t). \tag{4.22}$$

■

Example 4.2.15 (Binomial Distribution) Let $X \in \text{Bin}\,(n,p)$. The discrete case of the definition (4.4) yields

$$\varphi_X(t) = \sum_{k=0}^{n} e^{itk} P(X=k) = \sum_{k=0}^{n} e^{itk} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) \left(e^{it}p\right)^k (1-p)^{n-k}$$

$$= \left(e^{it}p + (1-p)\right)^n,$$

where we used the binomial theorem. We have thus found

$$X \in \text{Bin}\,(n,p) \Leftrightarrow \varphi_X(t) = \left((1-p) + e^{it}p\right)^n. \tag{4.23}$$

■

4.3 Characteristic Functions and Moments of Random Variables

One can compute moments by differentiation of the characteristic function.

Theorem 4.3.1 If the random variable X has the expectation, $E\left[\mid X \mid\right] < \infty$, then

$$\frac{d}{dt}\varphi_X(t) \mid_{t=0} = \frac{d}{dt}\varphi_X(0) = iE\left[X\right]. \tag{4.24}$$

If $E\left[\mid X \mid^k\right] < \infty$, then

$$\frac{d^k}{dt^k}\varphi_X(0) = i^k E\left[X^k\right]. \tag{4.25}$$

Proof: Formally, $\frac{d}{dt}\varphi_X(t) = E\left[\frac{d}{dt}e^{itX}\right] = E\left[iXe^{itX}\right]$. Hence $\frac{d}{dt}\varphi_X(0) = iE\left[X\right]$. The legitimacy of changing the order of diffentiation and expectation is taken for granted. ■

We can do some simple examples.

Example 4.3.2 (The Cauchy Distribution) In (4.18) $X \in C\,(0,1)$ was shown to have the characteristic function $\varphi_X(t) = e^{-|t|}$. Let us note that $|t|$ does not have a derivative at $t=0$.

■

Example 4.3.3 (Mean and Variance of the Poisson Distribution) We have in (4.9)

$$\varphi_X(t) = e^{\lambda\left(e^{it}-1\right)}.$$

Then

$$\frac{d}{dt}\varphi_X(t) = e^{\lambda\left(e^{it}-1\right)} \cdot i\lambda e^{it}$$

and by (4.24)

$$E\left[X\right] = \frac{1}{i}\frac{d}{dt}\varphi_X(0) = \lambda,$$

as is familiar from any first course in probability and/or statistics.

$$\frac{d^2}{dt^2}\varphi_X(t) = e^{\lambda\left(e^{it}-1\right)} \cdot i^2\lambda^2 e^{i2t} + e^{\lambda\left(e^{it}-1\right)} i^2 \lambda e^{it},$$

and from (4.25)

$$E\left[X^2\right] = \frac{1}{i^2}\frac{d^2}{dt^2}\varphi_X(0) = \lambda^2 + \lambda.$$

Thus

$$\text{Var}\left[X\right] = E\left[X^2\right] - \left(E\left[X\right]\right)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda,$$

which again agrees with the expression derived in any first course in probability and/or statistics.

■

4.4 Characteristic Functions of Sums of Independent Random Variables

Let X_1, X_2, $\ldots, X_n$ be n independent random variables. We consider their sum

$$S_n = X_1 + X_2 + \ldots + X_n = \sum_{k=1}^{n} X_k.$$

Theorem 4.4.1 X_1, X_2, $\ldots, X_n$ are independent random variables with respective characteristic functions $\varphi_{X_k}(t)$, $k = 1, 2, \ldots, n$. Then the characteristic function $\varphi_{S_n}(t)$ of their sum $S_n = \sum_{k=1}^{n} X_k$ is given by

$$\varphi_{S_n}(t) = \varphi_{X_1}(t) \cdot \varphi_{X_2}(t) \cdot \ldots \cdot \varphi_{X_n}(t). \tag{4.26}$$

Proof: $\varphi_{S_n}(t) = E\left[e^{itS_n}\right] = E\left[e^{it(X_1+X_2+\ldots+X_n)}\right] = E\left[e^{itX_1}e^{itX_2} \cdot \ldots \cdot e^{itX_n}\right]$. Then we can recall theorem 1.6.1 above, and suitably applied this gives by independence that

$$= E\left[e^{itX_1}\right]E\left[e^{itX_2}\right] \cdot \ldots \cdot E\left[e^{itX_n}\right]$$

$$= \varphi_{X_1}(t) \cdot \varphi_{X_2}(t) \cdot \ldots \cdot \varphi_{X_n}(t).$$

■

Corollary 4.4.2 X_1, X_2, $\ldots, X_n$ are independent and identically distributed random variables with the characteristic function $\varphi_X(t)$, $X \overset{d}{=} X_k$. Then the characteristic function $\varphi_{S_n}(t)$ of their sum $S_n = \sum_{i=1}^{k} X_i$ is given by

$$\varphi_{S_n}(t) = \left(\varphi_X(t)\right)^n. \tag{4.27}$$

■

If X and Y are independent random variables with probability densities f_X and f_Y, respectively, then their sum $Z = X + Y$ has, as is checked in (2.110), the p.d.f. given by the **convolutions**

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx = \int_{-\infty}^{\infty} f_Y(y)f_X(z-y)dy.$$

If we write the convolution symbolically as, say,

$$f_Z = f_X \oplus f_Y,$$

then we have in the theorem 4.4.1 established the rule of transformation

$$f_X \oplus f_Y \overset{\text{Ch}}{\mapsto} \varphi_X \cdot \varphi_Y.$$

This is plainly nothing but a well known and important property (convolution theorem) of Fourier transforms, [100, p. 177].

As applications of the preceding we can prove a couple of essential theorems.

Theorem 4.4.3 $X_1, X_2, \ldots, X_n$ are independent and $X_k \in N\left(\mu_k, \sigma_k^2\right)$ for $k = 1, 2, \ldots, n$. Then for any real constants $a_1, \ldots, a_n$

$$S_n = \sum_{k=1}^{n} a_k X_k \in N\left(\sum_{k=1}^{n} a_k \mu_k, \sum_{k=1}^{n} a_k^2 \sigma_k^2\right). \tag{4.28}$$

Proof: By (4.26) we obtain

$$\varphi_{S_n}(t) = \varphi_{a_1 X_1}(t) \varphi_{a_2 X_2}(t) \cdot \ldots \cdot \varphi_{a_n X_n}(t)$$

and by (e) in theorem 4.2.3 we get

$$= \varphi_{X_1}(a_1 t) \cdot \varphi_{X_2}(a_2 t) \cdot \ldots \cdot \varphi_{X_n}(a_n t).$$

By assumption and (4.8) we have $\varphi_X(a_k t) = e^{i\mu_k a_k t - \frac{\sigma_k^2 a_k^2 t^2}{2}}$. This yields

$$\varphi_{X_1}(a_1 t) \cdot \varphi_{X_2}(a_2 t) \cdot \ldots \cdot \varphi_{X_n}(a_n t) = e^{i\mu_1 a_1 t - \frac{\sigma_1^2 a_1^2 t^2}{2}} e^{i\mu_2 a_2 t - \frac{\sigma_2^2 a_2^2 t^2}{2}} \cdot \ldots \cdot e^{i\mu_n a_n t - \frac{\sigma_n^2 a_n^2 t^2}{2}}$$

and some elementary rearrangements using the properties of the exponential function we find

$$= e^{i \sum_{k=1}^{n} \mu_k a_k t - \frac{\sum_{k=1}^{n} a_k^2 \sigma_k^2 t^2}{2}}$$

or

$$\varphi_{S_n}(t) = e^{i \sum_{k=1}^{n} \mu_k a_k t - \frac{\sum_{k=1}^{n} a_k^2 \sigma_k^2 t^2}{2}}.$$

A comparison with (4.8) identifies $\varphi_{S_n}(t)$ as the characteristic function of $N\left(\sum_{k=1}^{n} a_k \mu_k, \sum_{k=1}^{n} a_k^2 \sigma_k^2\right)$. By uniqueness of the characteristic function we have shown the assertion as claimed. ■

Example 4.4.4 Let $X_1, \ldots, X_n$ are I.I.D. and $\in N(\mu, \sigma^2)$. Set $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Thus $\overline{X} \in N\left(\mu, \frac{\sigma^2}{n}\right)$.

■

Next we deal with sums of independent Poisson random variables.

Theorem 4.4.5 $X_1, X_2, \ldots, X_n$ are independent and $X_k \in \text{Po}(\lambda_k)$ for $k = 1, 2, \ldots, n$. Then

$$S_n = \sum_{k=1}^{n} X_k \in \text{Po}\left(\sum_{k=1}^{n} \lambda_k\right). \tag{4.29}$$

Proof: By (4.26) we obtain

$$\varphi_{S_n}(t) = \varphi_{X_1}(t) \cdot \varphi_{X_2}(t) \cdot \ldots \cdot \varphi_{X_n}(t)$$

and when we invoke (4.9)

$$= e^{\lambda_1 \left(e^{it} - 1\right)} e^{\lambda_2 \left(e^{it} - 1\right)} \cdot \ldots \cdot e^{\lambda_n \left(e^{it} - 1\right)}$$

$$= e^{(\lambda_1 + \ldots + \lambda_n)\left(e^{it} - 1\right)}.$$

Thus

$$\varphi_{S_n}(t) = e^{(\lambda_1 + \ldots + \lambda_n)\left(e^{it} - 1\right)}.$$

But another look at (4.9) shows that the right hand side of the last equality is the characteristic function of $\text{Po}\left(\sum_{k=1}^{n} \lambda_k\right)$. Thus by uniqueness of characteristic functions the claim follows as asserted. ■

Example 4.4.6 (Binomial Distribution as a Sum of Independent Be(p) **Variables)** By a comparison of (4.21) with (4.23) it follows by uniqueness of characteristic functions that if $X \in \text{Bin}\,(n,p)$, then

$$X \overset{d}{=} U_1 + U_2 + \ldots + U_n,$$

where U_k are independent and identically distributed (I.I.D.) $U_k \in \text{Be}(p)$, $k = 1, 2, \ldots, n$.

■

Example 4.4.7 (Sum of Two Independent Binomial Random Variables with the same p**)** $X_1 \in \text{Bin}\,(n_1,p)$, $X_2 \in \text{Bin}\,(n_2,p)$, X_1 and X_2 are independent. Then

$$X_1 + X_2 \in \text{Bin}\,(n_1 + n_2, p)\,. \tag{4.30}$$

To check this, by (4.23) and (4.26) it holds

$$\begin{aligned}\varphi_{X_1+X_2}(t) &= \left((1-p) + e^{it}p\right)^{n_1} \cdot \left((1-p) + e^{it}p\right)^{n_2} \\ &= \left((1-p) + e^{it}p\right)^{n_1+n_2},\end{aligned}$$

which proves the assertion.

■

Example 4.4.8 (Poisson binomial Distribution) $X \in \text{Pobin}\,(p_1, p_2, \ldots, p_n)$, $0 \leq p_i \leq 1$, $i = 1, 2, \ldots, n$, of Example 2.3.7 is naturally defined as the sum of independent $U_i \in \text{Be}\,(p_i)$, $i = 1, 2, \ldots, n$ that are independent, or

$$X = U_1 + \ldots + U_n.$$

From this the mean and variance given in Example 2.3.7 are immediate. In addition, the characteristic function is

$$\varphi_X(t) = \prod_{j=1}^{n} \left(1 - p_j + p_j e^{it}\right). \tag{4.31}$$

■

Example 4.4.9 (Gamma Distribution a Sum of Independent Exp(λ) **Variables)** Let $X \in \Gamma\,(n, \lambda)$, where n is a positive integer. Then the finding in (4.16) shows in view of (4.10) that X is in distribution equal to a sum of n independent Exp(λ)-distributed variables. In view of (2.2.10) we can also state that a sum of n independent Exp(λ)-distributed variables has an Erlang distribution.

■

Example 4.4.10 (Sum of Two Independent Gamma Distributed Random Variables) Let $X_1 \in \Gamma\,(n_1, \lambda)$ and $X_2 \in \Gamma\,(n_2, \lambda)$. Then in view of (4.16) and (4.26) we get that

$$X_1 + X_2 \in \Gamma\,(n_1 + n_2, \lambda)\,.$$

■

4.5 Expansions of Characteristic Functions

4.5.1 Expansions and Error Bounds

We recall the complex exponential for a purely imaginary argument, or with a real t and x,

$$e^{itx} = 1 + \sum_{k=1}^{\infty} \frac{(itx)^k}{k!} = 1 + itx + \frac{(itx)^2}{2} + \frac{(itx)^3}{3!} + \frac{(itx)^4}{4!} + \ldots .$$

Lemma 4.5.1 For real x we have

$$\mid e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!} \mid \leq \min\left(\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right) \tag{4.32}$$

Proof: The proof is by complete induction. For $n = 0$ we claim that

$$\mid e^{ix} - 1 \mid \leq \min\left(|x|, 2\right),$$

which is easily seen by drawing a picture of the complex unit circle and a chord of it to depict $e^{ix} - 1$.

We make the induction assumption that (4.32) holds for n. We wish to prove the assertion for $n+1$. By complex conjugation we find that it suffices to consider $x > 0$. We proceed by expressing the function to be bounded as a definite integral.

$$e^{ix} - \sum_{k=0}^{n+1} \frac{(ix)^k}{k!} = e^{ix} - 1 - \sum_{k=1}^{n+1} \frac{(ix)^k}{k!}$$

$$= e^{ix} - 1 - \sum_{k=0}^{n} \frac{(ix)^{k+1}}{(k+1)!} = \int_0^x \left[e^{it} - \sum_{k=0}^{n} \frac{(it)^k}{k!} \right] d(it).$$

Thus

$$\mid e^{ix} - \sum_{k=0}^{n+1} \frac{(ix)^k}{k!} \mid \leq \int_0^x \mid e^{it} - \sum_{k=0}^{n} \frac{(it)^k}{k!} \mid d(it).$$

At this point of the argument we use the induction hypothesis, i.e., (4.32) holds for n. This yields

$$\int_0^x \mid e^{it} - \sum_{k=0}^{n} \frac{(it)^k}{k!} \mid dt \leq \int_0^x \min\left(\frac{|t|^{n+1}}{(n+1)!}, \frac{2|t|^n}{n!} \right) dt$$

$$\leq \min\left(\frac{|x|^{n+2}}{(n+2)!}, \frac{2|x|^{n+1}}{(n+1)!} \right).$$

(Why does the last inequality hold ?) In summary, we have shown that

$$\mid e^{ix} - \sum_{k=0}^{n+1} \frac{(ix)^k}{k!} \mid \leq \min\left(\frac{|x|^{n+2}}{(n+2)!}, \frac{2|x|^{n+1}}{(n+1)!} \right),$$

which evidently tells that the assertion (4.32) holds for $n+1$. The proof by induction is complete. ■

The bound (4.32) leads immediately to the next bound for expansion of characteristic function.

Lemma 4.5.2 For a random variable X such that $E\left[\mid X \mid^n\right] < \infty$ we have

$$\mid \varphi_X(t) - \sum_{k=0}^{n} \frac{(it)^k E\left[X^k\right]}{k!} \mid \leq E\left[\min\left(\frac{|tX|^{n+1}}{(n+1)!}, \frac{2|tX|^n}{n!} \right) \right] \tag{4.33}$$

Proof: By the definition of $\varphi_X(t)$ we have

$$\mid \varphi_X(t) - \sum_{k=0}^{n} \frac{(it)^k E\left[X^k\right]}{k!} \mid = \mid E\left[e^{itX}\right] - \sum_{k=0}^{n} \frac{(it)^k E\left[X^k\right]}{k!} \mid$$

$$= \mid E\left[e^{itX} - \sum_{k=0}^{n} \frac{(it)^k X^k}{k!}\right] \mid \leq E\left[\mid e^{itX} - \sum_{k=0}^{n} \frac{(it)^k X^k}{k!} \mid\right]$$

Now we apply the error bound in (4.32) on the expression inside the expectation, and the upper bound in (4.33) follows. ■

For ease of effort in the sequel we isolate an important special case of the preceding. With $n = 2$ we have in (4.33) the error bound

$$E\left[\min\left(\frac{|tX|^3}{3!}, \frac{2|tX|^2}{2!}\right)\right] \leq |t|^2 E\left[\frac{\min\left(|t||X|^3, 2|X|^2\right)}{3!}\right].$$

Let $o(t)$ denote any function such that $\lim_{t\to 0} \frac{o(t)}{t} \to 0$ (**Landau's o -notation**). $o(t)$ is also called 'small ordo'. In our case we construct a small ordo by observing that

$$o\left(t^2\right) = |t|^2 E\left[\frac{\min\left(|t||X|^3, 2|X|^2\right)}{3!}\right]$$

fulfills $\lim_{t\to 0} \frac{o(t^2)}{t^2} \to 0$. Thus we can write

$$\varphi_X(t) = 1 + itE[X] - \frac{t^2}{2}E\left[X^2\right] + o\left(t^2\right). \tag{4.34}$$

In view of the preceding there is also the following series expansion.

Theorem 4.5.3 Suppose that the random variable X has the nth moment $E[| X |^n] < \infty$ for some n. Then

$$\varphi_X(t) = 1 + \sum_{k=1}^{n} E\left[X^k\right] \frac{(it)^k}{k!} + o(|t|^n). \tag{4.35}$$

■

4.5.2 A Scaled Sum of Standardised Random Variables (Central Limit Theorem)

Let us now consider the following problem. X_1, X_2, $\ldots, X_n$ $\ldots$ is an infinite sequence of independent and identically distributed random variables with $E[X_k] = \mu$ and $\text{Var}[X_k] = \sigma^2$ for $k = 1, 2, \ldots,$. We *standardise* the X_ks by subtracting the common mean and then dividing the difference by the common standard deviation or

$$Y_k = \frac{X_k - \mu}{\sigma}.$$

Thereby Y_k's are independent and identically distributed. In addition, we assume the standardization $E[Y_k] = 0$ and $\text{Var}[Y_k] = 1$. Let us furthermore add the first n of the Y_k's and scale the sum by the factor $\frac{1}{\sqrt{n}}$ so that

$$W_n \overset{\text{def}}{=} \frac{1}{\sqrt{n}} \sum_{k=1}^{n} Y_k = \sum_{k=1}^{n} \frac{Y_k}{\sqrt{n}}$$

We shall now compute the characteristic function of W_n and then see what happens to this function, as $n \to \infty$. It turns out that the scaling must be taken exactly as $\frac{1}{\sqrt{n}}$ for anything useful to emerge.

By (4.27) with $Y \overset{d}{=} Y_k$ it follows that

$$\varphi_{W_n}(t) = \varphi_{\sum_{k=1}^{n} \frac{Y_k}{\sqrt{n}}}(t) = \left(\varphi_{\frac{Y}{\sqrt{n}}}(t)\right)^n.$$

By property (e) in theorem 4.2.3 we have that $\varphi_{\frac{Y}{\sqrt{n}}}(t) = \varphi_Y\left(\frac{t}{\sqrt{n}}\right)$. Thus

$$\varphi_{W_n}(t) = \left(\varphi_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n.$$

When we expand $\varphi_Y\left(\frac{t}{\sqrt{n}}\right)$ as in (4.34) we obtain, as $E[Y_k] = 0$ and $\text{Var}[Y_k] = 1$,

$$\varphi_Y\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right).$$

Thereby we get

$$\varphi_{W_n}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n.$$

It is shown in the Appendix 4.6, see (4.42), that now

$$\lim_{n\to\infty} \varphi_{W_n}(t) = e^{-t^2/2}. \tag{4.36}$$

In view of (4.7) we observe that the characteristic function of the scaled sum W_n of random variables converges by the above for all t to the characteristic function of $N(0,1)$. We have now in essence proved a version of the **Central Limit Theorem**, but the full setting of convergence of sequences of random variables will be treated in chapter 6.

4.6 An Appendix: A Limit

4.6.1 A Sequence of Numbers with the Limit e^x

The following statement appears under various guises in several of the proofs and exercises.

Proposition 4.6.1

$$c_n \to c \Rightarrow \left(1 + \frac{c_n}{n}\right)^n \to e^c, \quad \text{as } n \to \infty. \tag{4.37}$$

Proof: Let us consider

$$n \ln\left(1 + \frac{c_n}{n}\right) = c_n \cdot \frac{\ln\left(1 + \frac{c_n}{n}\right)}{\frac{c_n}{n}}. \tag{4.38}$$

Here we recall a standard limit in calculus (or, the derivative of $\ln x$ at $x = 1$)

$$\lim_{h\to 0} \frac{\ln(1+h)}{h} = 1.$$

Since $c_n \to c$ by assumption, we have that $\frac{c_n}{n} \to 0$, as $n \to \infty$. Thus

$$\lim_{n\to\infty} \frac{\ln\left(1 + \frac{c_n}{n}\right)}{\frac{c_n}{n}} = 1,$$

we get in (4.38) that

$$\lim_{n\to\infty} n \ln\left(1 + \frac{c_n}{n}\right) = c.$$

Let $x_n = n \ln\left(1 + \frac{c_n}{n}\right)$. Since e^x is a continuous function we have

$$\left(1 + \frac{c_n}{n}\right)^n = e^{x_n} \to e^c.$$

■

The proof above is strictly speaking valid for sequences of real numbers. We shall next present two additional arguments.

4.6.2 Some Auxiliary Inequalities

Lemma 4.6.2 For any complex numbers w_i, z_i, if $|w_i| \leq 1$ and $|z_i| \leq 1$, then

$$|\prod_{i=1}^{n} z_i - \prod_{i=1}^{n} w_i| \leq \sum_{i=1}^{n} |z_i - w_i|. \tag{4.39}$$

Proof We have the identity

$$| \prod_{i=1}^{n} z_i - \prod_{i=1}^{n} w_i \mid = \mid (z_n - w_n) \prod_{i=1}^{n-1} z_i + w_n \left(\prod_{i=1}^{n-1} z_i - \prod_{i=1}^{n-1} w_i \right) |$$

and the right hand side of this inequality is upper bounded by

$$\leq \mid z_n - w_n \mid + \mid \prod_{i=1}^{n-1} z_i - \prod_{i=1}^{n-1} w_i \mid,$$

since $|w_n| \leq 1$ and $|z_i| \leq 1$. Then we use an induction hypothesis. ■

Lemma 4.6.3 For any complex numbers u and v,

$$|u^n - v^n| \leq |u - v| n \max(|u|, |v|)^{n-1} \tag{4.40}$$

Proof We have the identity

$$u^n - v^n = (u - v)u^{n-1} + v(u^{n-1} - v^{n-1})$$

and then

$$|u^n - v^n| \leq |u - v||u^{n-1}| + |v||u^{n-1} - v^{n-1}|. \tag{4.41}$$

Our induction hypothesis is that

$$|u^{n-1} - v^{n-1}| \leq |u - v|(n - 1) \max(|u|, |v|)^{n-2}.$$

When we apply this in right hand side of (4.41) we get

$$|u^n - v^n| \leq |u - v||u^{n-1}| + |v||u - v|(n - 1) \max(|u|, |v|)^{n-2}.$$

We note that

$$|u^{n-1}| \leq \max(|u|, |v|)^{n-1}$$

and that

$$|v| \cdot \max(|u|, |v|)^{n-2} \leq \max(|u|, |v|) \cdot \max(|u|, |v|)^{n-2} = \max(|u|, |v|)^{n-1}.$$

Thus we have obtained

$$|u^n - v^n| \leq |u - v| \max(|u|, |v|)^{n-1} + |u - v|(n - 1) \max(|u|, |v|)^{n-1},$$

which proves (4.40) as asserted. ■

4.6.3 Applications

The situation corresponding to (4.37) is often encountered as

$$\lim_{n\to\infty}\left(1-\frac{t^2}{2n}+o\left(\frac{t^2}{n}\right)\right)^n = e^{-t^2/2}. \tag{4.42}$$

1. Let us set
$$c_n = -\left(\frac{t^2}{2} - n\cdot o\left(\frac{t^2}{n}\right)\right).$$
Then $n\cdot o\left(\frac{t}{n}\right) = \frac{o\left(\frac{t}{n}\right)}{\frac{1}{n}} \to 0$, as $n\to\infty$. Thus $c_n \to -\frac{t^2}{2}$, as $n\to\infty$, and (4.42) follows by (4.37), since
$$\frac{c_n}{n} = -\left(\frac{t^2}{2n} - o\left(\frac{t^2}{n}\right)\right).$$

2. Let us now check (4.42) using the inequalities in the preceding section. With regard to lemma 4.6.2 we take for all i
$$z_i = \left(1-\frac{t^2}{2n}+o\left(\frac{t^2}{n}\right)\right)$$
and
$$w_i = \left(1-\frac{t^2}{2n}\right).$$
Then $|w_i|\leq 1$ and $|z_i|\leq 1$ and
$$|\prod_{i=1}^n z_i - \prod_{i=1}^n w_i| = |\left(1-\frac{t^2}{2n}+o\left(\frac{t^2}{n}\right)\right)^n - \left(1-\frac{t^2}{2n}\right)^n|,$$
and in view of the lemma 4.6.2 above we get
$$|\left(1-\frac{t^2}{2n}+o\left(\frac{t^2}{n}\right)\right)^n - \left(1-\frac{t^2}{2n}\right)^n| \leq n \mid o\left(\frac{t^2}{n}\right)\mid =$$
$$=\mid \frac{o\left(\frac{t^2}{n}\right)}{\frac{1}{n}} \mid .$$
When $n\to\infty$, $\frac{o\left(\frac{t^2}{n}\right)}{\frac{1}{n}} \to 0$, by definition of Landau's o. Since
$$\left(1-\frac{t^2}{2n}\right)^n \to e^{-t^2/2}, \quad \text{as } n\to\infty,$$
it now follows that as $n\to\infty$
$$\left(1-\frac{t^2}{2n}+o\left(\frac{t^2}{n}\right)\right)^n \to e^{-t^2/2},$$
as was to be proved.

3. If $\mid u \mid< 1$ and $\mid v \mid< 1$, then we get in (4.40)
$$|u^n - v^n| \leq |u-v|n$$
and thus with
$$u = \left(1-\frac{t^2}{2n}+o\left(\frac{t^2}{n}\right)\right), v = \left(1-\frac{t^2}{n}\right),$$
we obtain again as $n\to\infty$, that
$$\left(1-\frac{t^2}{2n}+o\left(\frac{t^2}{n}\right)\right)^n \to e^{-t^2/2}.$$

4.7 Exercises

4.7.1 Additional Examples of Characteristic Functions

1. Let $a < b$. Show that
$$X \in U(a,b) \Leftrightarrow \varphi_X(t) = \frac{e^{itb} - e^{ita}}{it(b-a)}. \tag{4.43}$$

2. Let $X \in \text{Tri}(-1,1)$, which means that the p.d.f. of X is
$$f_X(x) = \begin{cases} 1 - |x| & |x| < 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{4.44}$$
Show that
$$X \in \text{Tri}(-1,1) \Leftrightarrow \varphi_X(t) = \left(\frac{\sin \frac{t}{2}}{\frac{t}{2}} \right)^2. \tag{4.45}$$

3. Let $X_1 \in U\left(-\frac{1}{2}, \frac{1}{2}\right)$ and $X_2 \in U\left(-\frac{1}{2}, \frac{1}{2}\right)$. Assume that X_1 and X_2 are independent. Find the distribution of $X_1 + X_2$.

4. $X \in N(0,1)$. Show that the characteristic function of X^2 is
$$\varphi_{X^2}(t) = \frac{1}{\sqrt{1-2it}}. \tag{4.46}$$

5. Assume that $X_1, \ldots, X_n$ are independent and $N(0,1)$ distributed. Show that
$$\sum_{i=1}^{n} X_i^2 \in \chi^2(n). \tag{4.47}$$
Aid: You can do this with little effort by inspection of (4.16).

6. **Stable Distributions**

   (a) $X_1, \ldots, X_n$ are independent and $C(0,1)$ -distributed. Set $S_n = X_1 + \ldots + X_n$. Show that
   $$\frac{1}{n} S_n \in C(0,1).$$
   In other words, $\frac{1}{n} S_n \stackrel{d}{=} X$.

   (b) $X_1, \ldots, X_n$ are independent and $N(0,1)$-distributed. Set $S_n = X_1 + \ldots + X_n$. Show that
   $$\frac{1}{\sqrt{n}} S_n \in N(0,1).$$
   In other words, in this case $\frac{1}{\sqrt{n}} S_n \stackrel{d}{=} X$.

   (c) X is a r.v. and $X_1, \ldots, X_n$ are independent r.v.'s and $X_k \stackrel{d}{=} X$ for all k. Set $S_n = X_1 + \ldots + X_n$. If there exist sequences of real numbers $a_n > 0$ and b_n such that for all $n \geq 1$
   $$S_n \stackrel{d}{=} a_n X + b_n,$$
   then the distribution of X is said to be **stable**. Show that if
   $$\varphi_X(t) = e^{-c|t|^{\alpha}}, \quad 0 < \alpha \leq 2, c > 0, \tag{4.48}$$
   then the distribution of X is stable. (It can be verified that $\varphi_X(t)$ is in fact a characteristic function.) Interpret (a) and (b) in terms of (4.48).

7. Let $X_1, \ldots, X_n, \ldots$ be independent and $C(0,1)$ -distributed. Set $S_n = X_1 + \ldots + X_n$. Show that

$$\frac{1}{n}\sum_{k=1}^{n}\frac{S_k}{k} \in C(0,1).$$

Note that the r.v.'s $\frac{S_k}{k}$ are not independent.

8. (From [35]) Here we study the product of two independent standard Gaussian variables. More on products of independent random variables is given in section 4.7.4.

   (a) $X_1 \in N(0,1)$ and $X_2 \in N(0,1)$ are independent. Show that the characteristic function of their product $Y = X_1 \cdot X_2$ is

   $$\varphi_Y(t) = \frac{1}{\sqrt{1+t^2}}. \tag{4.49}$$

   (b) $Z_1 \in \Gamma(a,b)$ and $Z_2 \in \Gamma(a,b)$ are independent. We set

   $$U = Z_1 - Z_2,$$

   and suppose we know that $U \overset{d}{=} Y$, where Y has the distribution in part (a) of this exercise. What are the values of a and b ? *Answer:* $a = 1/2$, $b = 1$.

9. (From [35]) The r.v. X has the characteristic function $\varphi(t)$. Show that $|\varphi(t)|^2$ is a characteristic function. *Aid:* Take X_1 and X_2 as independent and $X_1 \overset{d}{=} X$ as well as $X_2 \overset{d}{=} X$. Check the characteristic function of $Y = X_1 - X_2$.

10. $X \in \mathrm{IG}(\mu, \lambda)$ with p.d.f. given in (2.37). Find its characteristic function as

$$\varphi_X(t) = e^{\left(\frac{\lambda}{\mu}\right)\left[1-\sqrt{1-\frac{2\mu^2 it}{\lambda}}\right]}.$$

11. $X \in \mathrm{K}(L, \mu, \nu)$ as in example 2.2.22. What could $\varphi_X(t)$ be ? *Aid:* None known.

12. $X \in \mathrm{Ske}(\mu_1, \mu_2)$ as in example 2.3.15.

    (a) Show that

    $$\varphi_X(t) = e^{-(\mu_1+\mu_2)+\mu_1 e^{it}+\mu_2 e^{-it}}.$$

    (b) Find $E[X]$ and $\mathrm{Var}[X]$ using $\varphi_X(t)$.

    (c) Show that the sum and the difference of two independent Skellam-distributed variables are Skellam-distributed.

4.7.2 Selected Exam Questions from the Past Decades

1. (5B1540 02-08-21, slightly simplified) The random variables $X_1, X_2, \ldots, X_n, \ldots,$ are I.I.D. and have the probability mass function

$$p_X(-1) = \frac{1}{4}, p_X(0) = \frac{1}{2}, p_X(1) = \frac{1}{4}.$$

We define the random variable (random time) N by

$$N = \min\{n \mid X_n = 0\},$$

i.e., N is the smallest (or first) n such that $X_n = 0$.

(a) Show that $N \in \mathrm{Fs}\left(\frac{1}{2}\right)$.

(b) Show that the characteristic function of $S_N = \sum_{k=1}^{N} X_k$ is $\varphi_{S_N}(t) = 1/(2-\cos t)$.

(c) Find $\mathrm{Var}\,(S_N)$ (Answer: $\mathrm{Var}\,(S_N) = 1$).

2. (5B1540 04-05-26) The random variable Y_n is uniformly distributed on the numbers $\{j/2^n | j = 0, 1, 2, \ldots, 2^n - 1\}$. The r.v. $X_{n+1} \in \mathrm{Be}\left(\frac{1}{2}\right)$ is independent of Y_n.

   (a) Show that
$$Y_n + \frac{X_{n+1}}{2^{n+1}} \overset{d}{=} Y_{n+1}.$$

   (b) Show that
$$\prod_{k=1}^{n} \frac{1+e^{it/2^k}}{2} = \sum_{l=0}^{2^n-1} \frac{e^{itl/2^n}}{2^n}.$$

3. (5B1540 00-08-29) $X \in \mathrm{Exp}(1)$, $Y \in \mathrm{Exp}(1)$ are independent random variables. Show by means of characteristic functions that
$$X + \frac{Y}{2} \overset{d}{=} \max(X, Y).$$

4. (an intermediate step of an exam question in FA 181 1981-02-06) Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed. Furthermore, $a_1, a_2, \ldots, a_n$ are arbitrary real numbers. Set
$$Y_1 = a_1X_1 + a_2X_2 + \ldots + a_nX_n$$
and
$$Y_2 = a_nX_1 + a_{n-1}X_2 + \ldots + a_1X_n.$$
Show that
$$Y_1 \overset{d}{=} Y_2.$$

4.7.3 Various Applications of the Characteristic Function

1. In section 10.4.1 and elsewhere we shall require the following result.
$X \in N(0, \sigma^2)$. Show that
$$E\left[X^n\right] = \begin{cases} 0 & n \text{ is odd} \\ \frac{(2k)!}{2^k k!}\sigma^{2k} & n = 2k,\ k = 0, 1, 2, \ldots. \end{cases} \tag{4.50}$$
Aid: ([66, pp. 23-24]): We have
$$\varphi_X(t) = e^{-\frac{1}{2}t^2\sigma^2}.$$
Let
$$\varphi_X^{(n)}(t) = \frac{d^n}{dt^n}\varphi_X(t),$$
where $\varphi_X^{(0)}(t) = \varphi_X(t)$, $\varphi_X^{(1)}(t) = -t\sigma^2\varphi_X(t)$. Show by induction that for $n \geq 2$,
$$\varphi_X^{(n)}(t) = -(n-1)\sigma^2\varphi_X^{(n-2)}(t) - t\sigma^2\varphi_X^{(n-1)}(t).$$
Then we get
$$\varphi_X^{(n)}(0) = -(n-1)\varphi_X^{(n-2)}(0), \quad n \geq 2, \tag{4.51}$$
which is regarded as a difference equation with the initial value $\varphi_X^{(1)}(0) = 0$. Solve (4.51).

2. The **Rice Method** is a technique of computing moments of nonlinear transformations of random variables by means of characteristic functions [104, pp. 378-]. Let $H(x)$ be a (Borel) function such that its Fourier transform $\widehat{H}(t)$ exists. X is a random variable such that $E[H(X)]$ exists. Then we recall the formula for inverse Fourier transform in (4.2) as
$$H(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{itx}\widehat{H}(t)dt.$$
Then it follows, if the interchange of integral and expectation is taken for granted, that
$$E[H(X)] = \frac{1}{2\pi}E\left[\int_{-\infty}^{\infty} e^{itX}\widehat{H}(t)dt\right] = \frac{1}{2\pi}\int_{-\infty}^{\infty} E\left[e^{itX}\right]\widehat{H}(t)dt,$$
and by definition of the characteristic function
$$E[H(X)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi_X(t)\widehat{H}(t)dt. \tag{4.52}$$
This is the tool of the Rice method. It may turn out that the integration in the right hand side can be performed straightforwardly (often by means of contour integration and residue calculus).
Assume that $X \in N(0,\sigma^2)$. Show that
$$E[\cos(X)] = e^{\frac{\sigma^2}{2}}.$$
Aid 1.: An engineering formula for the Fourier transform of $\cos(x)$ is, [101, p.413],
$$H(x) = \cos(x) \overset{\mathcal{F}}{\mapsto} \widehat{H}(t) = \frac{1}{2}\left(\delta(t-1)+\delta(t+1)\right),$$
where $\delta(t)$ is the Dirac's delta 'function'.
Aid 2.: If you do not feel comfortable with Dirac's delta, write $\cos(x)$ by Euler's formula, in this attempt you do not really need (4.52).

4.7.4 Mellin Transform in Probability

The transform is named after Hjalmar Mellin[1]. The Mellin transform of probability densities is being applied in communications engineering, econometrics, biology, classification, analytic combinatorics and other fields. The point in this context is that products of random variables are part of the problem at hand, and that the conclusion about the distribution of these products can be derived by Mellin transforms.

Example 4.7.1 A financial portfolio is valued in a domestic currency (e.g., SEK). The prices of shares and other instruments are uncertain and are modeled as random variables. In addition the exchange rates are uncertain, hence the value of the portfolio in, say, JPY may be modelled by a product of two random variables.

■

Example 4.7.2 In statistical methodology an important role is played by the following result. Suppose $X \in N(0,1)$, $Y \in \chi^2(f)$, X and Y are independent. Then we know (presumably without a proof (?)) by any first course in statistics that
$$\frac{X}{\sqrt{Y/f}} \in t(f) \tag{4.53}$$

[1] Robert Hjalmar Mellin (1854 - 1933) studied at the University of Helsinki, where his teacher was Gösta Mittag-Leffler, who left Helsinki for having been appointed to professor of mathematics at the University of Stockholm. Mellin did post-doctoral studies in Berlin under Karl Weierstrass and in Stockholm. From 1908 till retirement Mellin served as professor of mathematics at the Polytechnic Institute in Helsinki, which subsequently became Helsinki University of Technology, currently merged to a constituent of the Aalto University.

Figure 4.1: Mellin Auditorium, Aalto University Main Building, Otaniemi.

i.e., the ratio follows the Student's t -distribution. We hint thereby that this can shown by a Mellin transformation.

■

For a random variable $X \geq 0$ with the probability density $f_X(x)$ we define the Mellin transform as

$$\widehat{f}_{\mathcal{M}_X}(z) = \int_0^\infty x^{z-1} f_X(x) dx. \tag{4.54}$$

Considered as a function of the complex variable z, $\widehat{f}_{\mathcal{M}_X}(z)$ is a function of the exponential type and is analytic in a strip parallel to the imaginary axis. The inverse transformation is

$$f_X(x) = \frac{1}{2\pi i} \int_L x^{-z} \widehat{f}_{\mathcal{M}_X}(z) dz, \tag{4.55}$$

for all x, where $f_X(x)$ is continuous. The contour of integration is usually $L = \{c - i\infty, c + i\infty\}$ and lies in the strip of analycity of $\widehat{f}_{\mathcal{M}_X}(z)$.

General Properties of the Mellin Transform

Several of the exercises below require proficiency in complex analysis to the extent provided in [93].

1. Let $X \geq 0$ be a random variable. Then show that

   (a) For any real constant $a > 0$,

$$\widehat{f}_{\mathcal{M}_{(aX)}}(z) = a^{z-1} \widehat{f}_{\mathcal{M}_X}(z). \tag{4.56}$$

(b) For any constant α,

$$\widehat{f}_{\mathcal{M}_{X^\alpha}}(z) = \widehat{f}_{\mathcal{M}_X}(\alpha z - \alpha + 1) \tag{4.57}$$

In particular, the Mellin transform of $\frac{1}{X}$ is

$$\widehat{f}_{\mathcal{M}_{X^{-1}}}(z) = \widehat{f}_{\mathcal{M}_X}(-z + 2) \tag{4.58}$$

2. Let $X \geq 0$ and $Y \geq 0$ be independent continuous random variables. Show that

$$\widehat{f}_{\mathcal{M}_{XY}}(z) = \widehat{f}_{\mathcal{M}_X}(z)\widehat{f}_{\mathcal{M}_Y}(z). \tag{4.59}$$

3. Let $X \geq 0$ and $Y \geq 0$ be independent continuous random variables. Show that

$$\widehat{f}_{\mathcal{M}_{\frac{X}{Y}}}(z) = \widehat{f}_{\mathcal{M}_X}(z)\widehat{f}_{\mathcal{M}_Y}(-z + 2). \tag{4.60}$$

4. Let $f_X(x)$ and $f_Y(y)$ be two probability densities on $(0, \infty)$. Let

$$h(x) = \int_0^\infty \frac{1}{y} f_X\left(\frac{x}{y}\right) f_Y(y) dy = \int_0^\infty \frac{1}{y} f_X(y) f_Y\left(\frac{x}{y}\right) dy. \tag{4.61}$$

Compute the Mellin transform of $h(x)$. *Aid:* Recall (2.108) in the preceding.

The function $h(x)$ is called the **Mellin convolution** of $f_X(x)$ and $f_Y(y)$.

5. $X \in U(0, 1)$. Show that

$$\widehat{f}_{\mathcal{M}_X}(z) = \frac{1}{z}, \tag{4.62}$$

where the strip of analycity is the half-plane $\text{Re}(z) > 0$.

6. $X \in \Gamma(p, 1)$. Show that

$$\widehat{f}_{\mathcal{M}_X}(z) = \frac{\Gamma(z + p - 1)}{\Gamma(p)}, \tag{4.63}$$

where the strip of analycity is the half-plane $\text{Re}(z) > 0$.

7. The Mellin transform of a probability density is

$$\widehat{f}_{\mathcal{M}_X}(z) = \Gamma(z), \tag{4.64}$$

where the strip of analycity is the half-plane $\text{Re}(z) > 0$. Find $f_X(x)$. A piece of residue calculus is required for the inversion in (4.55).

8. The random variables X_k, $k = 1, 2, \ldots, n$ are independent and have the density

$$f_X(x) = \begin{cases} (a+1)x^a & \text{if } 0 \leq x \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

(a) Show that

$$\widehat{f}_{\mathcal{M}_{\prod_{k=1}^n X_k}}(z) = \left(\frac{a+1}{z+a}\right)^n. \tag{4.65}$$

(b) Show that

$$f_{\prod_{k=1}^n X_k}(x) = \begin{cases} \frac{(a+1)^n}{(n-1)!} x^a \left(\ln \frac{1}{x}\right)^{n-1} & \text{if } 0 \leq x \leq 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{4.66}$$

9. In example 2.2.22 it was claimed that if

$$X = X_1 \cdot X_2,$$

where $X_1 \in \Gamma(1/L, L)$, and $X_2 \in \Gamma(\mu/\nu, \nu)$ are independent, then X has the p.d.f. in (2.38). Verify this by means of the appropriate Mellin transforms. *Aid:* None available.

The Mellin Transform of the Product of n Independent $N(0,1)$ Variables

The requirement $X \geq 0$ would seem to be a serious impediment to usefulness the Mellin transform in probability calculus. However, let $X^{+} = \max\{X, 0\}$ denote the positive part of X and $X^{-} = \max\{-X, 0\}$ denote its negative part. Thus $X = X^{+} - X^{-}$, and

$$XY = X^{+}Y^{+} - X^{+}Y^{-} - X^{-}Y^{+} + X^{-}Y^{-},$$

and then the Mellin transform of X can be defined for XY. This or other similar tricks enable us to extend the transform to the general case[2]. Then we can show, e.g., that the product of n independent $N(0,1)$ variables is (the student is not required to do this)

$$f_{\prod_{k=1}^{n} X_k}(x) = \frac{1}{(2\pi)^{n/2}} \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \left(\frac{x^2}{2^n}\right)^{-z} \Gamma^{n}(z)\, dz, \tag{4.67}$$

where the contour of integration is a line parallel to the imaginary axis and to the right of origin. The integral may be evaluated by residue calculus to give

$$f_{\prod_{k=1}^{n} X_k}(x) = \sum_{j=0}^{\infty} \frac{1}{(2\pi)^{n/2}} R(z,n,j),$$

where $R(z,n,j)$ denotes the residue of $\left(\frac{x^2}{2^n}\right)^{-z} \Gamma^{n}(z)$ at the nth order pole at $z = -j$. People knowledgeable in special functions recognize by (4.67) also that $f_{\prod_{k=1}^{n} X_k}(x)$ is an instance of what is called **Meijer's G-function** (or H-function) [3, pp.419−425], which is a generalization of the hypergeometric function. The residues can be evaluated by numerical algorithms, and therefore the probability density and the corresponding distribution function are available computationally, and by virtue of compilation efforts in the past, in tables of of function values.

10. Let $X_1, \ldots, X_n$ be independent $N(0, \sigma^2)$ variables. Show that

$$f_{\prod_{k=1}^{n} X_k}(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \left(\frac{x^2}{(2\sigma)^n}\right)^{-z} \Gamma^{n}(z)\, dz. \tag{4.68}$$

11. Establish the result in (4.49) by means of (4.68).

The Mellin Transform is a Fourier Transform

Make the change of variable $x = e^{u}$ and $z = c - it$ in (4.54). Then we get the transform

$$\widehat{f}_{\mathcal{M}X}(c - it) = \int_{0}^{\infty} e^{u(c-it)} f_X(e^{u})\, dx, \tag{4.69}$$

and the inverse in (4.55) as

$$f_X(e^{u}) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{iut} e^{-uc} \widehat{f}_{\mathcal{M}X}(c - it) dt. \tag{4.70}$$

This shows in view of (4.1) and (4.2) that we have in fact the pair of a function and its Fourier transform as in (4.3),

$$\left(f_X(e^{u})\, e^{uc}, \widehat{f}_{\mathcal{M}X}(c - it)\right).$$

[2] M.D. Springer & W.E. Thompson: The Distribution of Products of Independent Random Variables. *SIAM Journal on Applied Mathematics*, Vol. 14, No.3, 1966, pp. 511−526.

Chapter 5

Generating Functions in Probability

5.1 Introduction

The topic in this chapter will be the probability generating functions and moment generating functions in probability theory. Generating functions are encountered in many areas of mathematics, physics, finance and engineering. For example, in [3] one finds the generating functions for Hermite, Laguerre, Legendre polynomials and other systems of polynomials. The calculus of generating functions for problems of discrete mathematics (e.g., combinatorics) is expounded in [41]. In control engineering and signal processing generating functions are known plainly as **z-transforms**, see [93, 100]. The generic concept is as follows.

Consider a sequence of real numbers $(a_k)_{k=0}^{\infty}$, e.g., a_k could be the value of the kth Hermite polynomial H_k at x. The (ordinary) generating function of $(a_k)_{k=0}^{\infty}$ is defined as

$$G(t) = \sum_{k=0}^{\infty} a_k t^k$$

for those values of t, where the sum converges. For a given series there exists a *radius of convergence* $R > 0$ such that the sum converges absolutely for $\mid t \mid < R$ and diverges for $\mid t \mid > R$. $G(t)$ can be differentiated or integrated term by term any number of times, when $\mid t \mid < R$, [69, section 5.4]. We recall also **Abel's Theorem**: if $R \geq 1$ then $\lim_{t \uparrow 1} G(t) = \sum_{k=0}^{\infty} a_k$. In the sequel limits with $t \uparrow 1$ will often be written as $t \to 1$.

In many cases the $G(t)$ can be evaluated in a closed form. For example, the generating function of (probabilist's) Hermite polynomials $H_k(x)$ in (2.97) is

$$G(t) = e^{xt - \frac{1}{2}t^2} = \sum_{k=0}^{\infty} H_k(x) t^k.$$

The individual numbers, $H_k(x)$, in the sequence can be recovered (generated) from the explicit $G(t)$ by differentiation.

5.2 Probability Generating Functions

In this section we consider only discrete random variables X that have the non negative integers (or a subset thereof) as values. We have the probability mass function

$$p_X(k) = \mathbf{P}(X = k), \quad k = 0, 1, 2, \ldots$$

The first example that comes to mind is $X \in \text{Po}(\lambda)$, see example 2.3.8. In the case of a finite set of values we take $p_X(k) = 0$ for those non negative integers that cannot occur, e.g., when X takes only a finite number of values.

Definition 5.2.1 (Probability generating function) X is a non negative integer valued random variable. The probability generating function (p.g.f.) $g_X(t)$ of X is defined by

$$g_X(t) \overset{\text{def}}{=} E\left[t^X\right] = \sum_{k=0}^{\infty} t^k p_X(k). \tag{5.1}$$

■

We could be more precise and talk about the p.g.f. of the probability mass function $\{p_X(k)\}_{k=0}^{\infty}$, but it is customary and acceptable to use phrases like 'p.g.f. of a random variable' or 'p.g.f. of a distribution'.

Example 5.2.1 (P.g.f. for Poisson distributed random variables) $X \in \text{Po}(\lambda)$, $\lambda > 0$. The p.g.f. is

$$g_X(t) = \sum_{k=0}^{\infty} t^k e^{-\lambda} \frac{k^{\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(t\lambda)^k}{k!} = e^{-\lambda} \cdot e^{t\lambda},$$

where we used the series expansion of $e^{t\lambda}$, which converges for all t. In summary,

$$X \in \text{Po}(\lambda) \Rightarrow g_X(t) = e^{\lambda(t-1)}. \tag{5.2}$$

We write also

$$g_{\text{Po}}(t) = e^{\lambda(t-1)}, \quad t \in \mathbf{R}.$$

■

Note that $g_X(1) = \sum_{k=0}^{\infty} p_X(k) = 1$, so the series converges absolutely at least for $\mid t \mid \leq 1$. In addition, $g_X(0) = p_X(0)$. By termwise differentiation we get

$$g_X^{(1)}(t) = \frac{d}{dt} g_X(t) = \sum_{k=0}^{\infty} k t^{k-1} p_X(k) = \sum_{k=1}^{\infty} k t^{k-1} p_X(k).$$

Then it follows that

$$g_X^{(1)}(0) = p_X(1).$$

If we differentiate successively and evaluate the kth derivative $g_X^{(k)}(t)$ at $t = 0$, we get

$$p_X(k) = \mathbf{P}(X = k) = \frac{g_X^{(k)}(0)}{k!}, \quad k = 0, 1, 2, \ldots. \tag{5.3}$$

In this sense we can recover (generate) the probability mass function $p_X(n)$ from $g_X(t)$.

Example 5.2.2 $X \in \text{Po}(\lambda)$, then by (5.2) $g_X^{(1)}(t) = \lambda e^{\lambda(t-1)}$ and $g_X^{(1)}(0) = e^{-\lambda}\lambda$, as should be.

■

Theorem 5.2.3 (Uniqueness) If X and Y are two non negative integer valued random variables such that

$$g_X(t) = g_Y(t) \quad \text{for all } t\ ,$$

then

$$p_X(k) = p_Y(k) \quad k = 0, 1, 2, \ldots$$

We write this as

$$X \overset{d}{=} Y.$$

('X and Y are equal in distribution').

Proof: Since $g_X(t) = g_Y(t)$ holds for a region of convergence, we can take that the equality holds for some region around origin. Then we have by (5.3) for all k

$$p_X(k) = \frac{g_X^{(k)}(0)}{k!}, p_Y(k) = \frac{g_Y^{(k)}(0)}{k!}.$$

But the assumption implies that $g_X^{(k)}(0) = g_Y^{(k)}(0)$, and hence $p_X(k) = p_Y(k)$ for all k. ■
The uniqueness theorem means in the example above that (5.2) can be strengthened to

$$X \in \text{Po}(\lambda) \Leftrightarrow g_X(t) = e^{\lambda(t-1)}. \tag{5.4}$$

We can think of the generating functions of functions of X. The 'p.g.f. of $Y = H(X)$' would then be

$$g_Y(t) = g_{H(X)} = E\left[t^{H(X)}\right] = \sum_{k=0}^{\infty} t^{H(k)} p_X(k).$$

Example 5.2.4 Let $Y = a + bX$, where X is a non negative integer valued random variable and a and b are non negative integers. Then

$$g_Y(t) = E\left[t^{a+bX}\right] = t^a E\left[t^{bX}\right] = t^a E\left[\left(t^b\right)^X\right] = t^a g_X\left(t^b\right), \tag{5.5}$$

if t^b is in the domain of convergence of g_X.

■

Let us next compute additional examples of p.g.f.'s.

Example 5.2.5 (P.g.f. for Bernoulli random variables) $X \in \text{Be}(p)$, $0 < p < 1$. Here $\mathbf{P}(X = 1) = p$, $\mathbf{P}(X = 0) = 1 - p = q$ (and . The p.g.f. is

$$g_X(t) = t^0(1-p) + tp = q + pt.$$

Hence we have

$$X \in \text{Be}(p) \Leftrightarrow g_X(t) = q + pt. \tag{5.6}$$

We write also

$$g_{\text{Be}}(t) = q + pt.$$

We note that $g_X(0) = q = \mathbf{P}(X = 0)$, $g_X^{(1)}(t) = p$ and thus $g_X^{(1)}(0) = p = \mathbf{P}(X = 1)$ and $g_X^{(k)}(0) = 0$ for $k = 2, 3 \ldots$, as should be.

■

Example 5.2.6 (P.g.f. for Binomial random variables) $X \in \text{Bin}(p)$, $0 < p < 1$, $q = 1 - p$. The p.g.f. is

$$\begin{aligned} g_X(t) &= \sum_{k=0}^{n} t^k \left(\begin{array}{c} n \\ k \end{array}\right) p^k (1-p)^{n-k} = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array}\right) (tp)^k (1-p)^{n-k} \\ &= ((1-p) + tp)^n = (q + tp)^n, \end{aligned}$$

where we used the binomial theorem.

$$X \in \mathrm{Bin}(p) \Leftrightarrow g_X(t) = (q+tp)^n . \tag{5.7}$$

We write also

$$g_{\mathrm{Bin}}(t) = (q+tp)^n .$$

When both (5.6) and (5.7) are taken into account, we find

$$g_{\mathrm{Bin}}(t) = (g_{\mathrm{Be}}(t))^n . \tag{5.8}$$

■

Example 5.2.7 (P.g.f. for Geometric random variables) $X \in \mathrm{Ge}(p)$, $0 < p < 1$, $q = 1-p$. $p_X(k) = q^k p$, $k = 0,1,2,\ldots$. The p.g.f. is

$$g_X(t) = \sum_{k=0}^{\infty} t^k p(1-p)^k = p\sum_{k=0}^{n}(tq)^k,$$

which we sum as a geometric series to get

$$= \frac{p}{1-qt},$$

if $\mid t \mid< \frac{1}{q}$, where the radius of convergence is obtained from the radius of convergence of geometric series.

$$X \in \mathrm{Ge}(p) \Leftrightarrow g_X(t) = g_{\mathrm{Ge}}(t) = \frac{p}{1-qt}, \quad \mid t \mid< \frac{1}{q}. \tag{5.9}$$

■

Example 5.2.8 (P.g.f. for First Success random variables) $X \in \mathrm{Fs}(p)$, $0 < p < 1$, $q = 1-p$. $p_X(k) = q^{k-1}p$, $k = 1,2,\ldots$. The p.g.f. is

$$g_X(t) = \sum_{k=1}^{\infty} t^k pq^{k-1} = p\sum_{k=1}^{\infty} t^k q^{k-1} = \frac{p}{q}\sum_{k=1}^{n}(tq)^k = \frac{p}{q}\left(\sum_{k=0}^{n}(tq)^k - 1\right)$$

and we sum the geometric series, if $\mid t \mid< \frac{1}{q}$, to get

$$= \frac{p}{q}\left(\frac{1}{1-qt} - 1\right) = \frac{p}{q}\left(\frac{qt}{1-qt}\right) = \frac{pt}{1-qt}.$$

$$X \in \mathrm{Fs}(p) \Leftrightarrow g_X(t) = g_{\mathrm{Fs}}(t) = \frac{pt}{1-qt}, \quad \mid t \mid< \frac{1}{q}. \tag{5.10}$$

■

Example 5.2.9 (P.g.f. for $X+1$, $X \in \mathrm{Ge}(p)$) Let $X \in \mathrm{Ge}(p)$, $0 < p < 1$, $q = 1-p$. We set $Y = X+1$. Since X has values $k = 0,1,2,\ldots,$, the values of Y are $k = 1,2,\ldots,$. To compute the p.g.f. of Y we can use (5.5) with $a = 1$ and $b = 1$ and apply (5.9)

$$g_Y(t) = t \cdot g_X\,(t) = t \cdot \frac{p}{1-qt} = \frac{pt}{1-qt}.$$

Here a look at (5.10) and the uniqueness of p.g.f. entail

$$X+1 \in \mathrm{Fs}(p).$$

This makes perfect sense by our definitions. If $X \in \mathrm{Ge}(p)$, then X is the number of independent attempts in a binary trial *until* one gets the first success NOT INCLUDING the successful attempt. The first success distribution $\mathrm{Fs}(p)$ is the distribution of the number of independent attempts in a binary trial *until* one gets the first success INCLUDING the successful attempt. Clearly these very conceptions imply that $X+1 \in \mathrm{Fs}(p)$, if $X \in \mathrm{Ge}(p)$. Hence we have re-established this fact by a mechanical calculation. Or, we have checked that p.g.f. corresponds to the right thing.

■

5.3 Moments and Probability Generating Functions

We find first a formula for $g_X^{(r)}(1)$. We call

$$E\left[X(X-1)\cdot\ldots\cdot(X-(r-1))\right]$$

the rth (descending) factorial moment of X.

Theorem 5.3.1 (Factorial Moments by p.g.f.) X is a non negative integer valued random variable, and $E\left[X^r\right] < \infty$ for some $r > 0$. Then

$$g_X^{(r)}(1) = E\left[X(X-1)\cdot\ldots\cdot(X-(r-1))\right]. \tag{5.11}$$

Proof: By successive differentiations

$$g_X^{(r)}(t) = \sum_{k=r}^{\infty} k(k-1)\cdot\ldots\cdot(k-(r-1))t^{k-r}p_X(k)$$

Then we observe that

$$\sum_{k=r}^{\infty} k(k-1)\cdot\ldots\cdot(k-(r-1))p_X(k) = E\left[X(X-1)\cdot\ldots\cdot(X-(r-1))\right].$$

As a clarification, by the law of the unconscious statistician (2.4)

$$E\left[X(X-1)\cdot\ldots\cdot(X-(r-1))\right] = \sum_{k=0}^{\infty} k(k-1)\cdot\ldots\cdot(k-(r-1))p_X(k),$$

but the terms corresponding to $k = 0, 1, \ldots, r-1$ contribute obviously by a zero to the sum in the right hand side, and hence the claim in the theorem is true. ■

A number of special cases of the preceding result are of interest as well as of importance. We assume that the moments required below exist.

- $$g_X^{(1)}(1) = E\left[X\right]. \tag{5.12}$$

- $$g_X^{(2)}(1) = E\left[X(X-1)\right] = E\left[X^2\right] - E\left[X\right].$$

  As we have

  $$\mathrm{Var}[X] = E\left[X^2\right] - \left(E\left[X\right]\right)^2,$$

it follows that

$$\text{Var}[X] = g_X^{(2)}(1) + E[X] - (E[X])^2$$

or

$$\text{Var}[X] = g_X^{(2)}(1) + g_X^{(1)}(1) - \left(g_X^{(1)}(1)\right)^2. \tag{5.13}$$

Example 5.3.2 $X \in \text{Po}(\lambda)$, and from (5.4)

$$g_X^{(1)}(1) = \lambda e^{\lambda(t-1)} \mid_{t=1} = \lambda,$$

$$g_X^{(2)}(1) = \lambda^2 e^{\lambda(t-1)} \mid_{t=1} = \lambda^2,$$

and

$$\text{Var}[X] = g_X^{(2)}(1) + g_X^{(1)}(1) - \left(g_X^{(1)}(1)\right)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

■

5.4 Probability Generating Functions for Sums of Independent Random Variables

Let again $X_1, X_2, \ldots, X_n$ be n independent non negative integer valued random variables and consider their sum

$$S_n = X_1 + X_2 + \ldots + X_n = \sum_{k=1}^{n} X_k.$$

Then clearly S_n has, by basic algebra, the non negative integers as values. The results about the p.g.f. of the sum follow exactly as the analogous results for characteristic functions of the sum .

Theorem 5.4.1 If $X_1, X_2, \ldots, X_n$ are independent non negative integer valued random variables with respective p.g.f.s $g_{X_k}(t)$, $k = 1, 2, \ldots, n$. Then the p.g.f. $g_{S_n}(t)$ of their sum $S_n = \sum_{k=1}^{n} X_k$ is given by

$$g_{S_n}(t) = g_{X_1}(t) \cdot g_{X_2}(t) \cdot \ldots \cdot g_{X_n}(t). \tag{5.14}$$

Proof: $g_{S_n}(t) = E\left[t^{S_n}\right] = E\left[t^{(X_1+X_2+\ldots+X_n)}\right] = E\left[t^{X_1} t^{X_2} \cdot \ldots \cdot t^{X_n}\right]$. Then theorem 1.6.1 and independence (of Borel functions of independent random variables) entail together that

$$= E\left[t^{X_1}\right] E\left[t^{X_2}\right] \cdot \ldots \cdot E\left[t^{X_n}\right]$$

$$= g_{X_1}(t) \cdot g_{X_2}(t) \cdot \ldots \cdot g_{X_n}(t).$$

■

Example 5.4.2 (Sums of Independent Poisson Random Variables) $X_1, X_2, \ldots, X_n$ are independent and $X_k \in \text{Po}(\lambda_k)$, $\lambda_k > 0$ for $k = 1, 2, \ldots, n$. Then (5.4) and (5.14) entail

$$g_{S_n}(t) = e^{\lambda_1(t-1)} \cdot e^{\lambda_2(t-1)} \cdot \ldots \cdot e^{\lambda_n(t-1)} = e^{(\lambda_1+\lambda_2+\ldots+\lambda_n)(t-1)}.$$

Thus $S_n \in \text{Po}(\lambda_1 + \lambda_2 + \ldots + \lambda_n)$, as was already found by means of characteristic functions.

■

Corollary 5.4.3 $X_1,\ X_2,\ \ldots, X_n$, are independent and identically distributed non negative integer valued random variables with the p.g.f. $g_X(t)$, $X \overset{d}{=} X_k$. Then the p.g.f. $g_{S_n}(t)$ of their sum $S_n = \sum_{i=1}^{k} X_i$ is given by

$$g_{S_n}(t) = \left(g_X(t)\right)^n. \tag{5.15}$$

■

The assertions in (5.15) and (5.8) give another proof of the fact in Example 4.4.6.

5.5 Sums of a Random Number of Independent Random Variables

We consider N,$X_1,\ X_2,\ \ldots, X_n, \ldots$, which are independent random variables with non negative integers as values. $X_1,\ X_2,\ \ldots, X_n, \ldots$, are furthermore identically distributed with the p.g.f. $g_X(t)$. The p.g.f. of N is $g_N\left((t)\right)$. We want to study the sum of a random number of X_k's, or,

$$S_N = \begin{cases} 0, & \text{if } N = 0 \\ X_1 + X_2 + \ldots + X_N, & \text{if } N \geq 1. \end{cases} \tag{5.16}$$

In operational terms of a computer simulation, we generate first an outcome $N = n$, then the independent outcomes of X_1,$X_2 \ldots X_n$ and thereafter add the latter outcomes.

Theorem 5.5.1 (Composition Formula) The p.g.f. $g_{S_N}(t)$ of S_N defined in (5.16) is

$$g_{S_N}(t) = g_N\left(g_X(t)\right). \tag{5.17}$$

Proof: By definition of p.g.f. and double expectation in (3.9)

$$g_{S_N}(t) = E\left[t^{S_N}\right] = E\left[E\left[t^{S_N} \mid N\right]\right].$$

Since $E\left[t^{S_N} \mid N\right]$ is measurable with respect to the sigma field generated by N, we can write by the Doob-Dynkin theorem 1.5.5 that $H(N) = E\left[t^{S_N} \mid N\right]$. Then

$$E\left[E\left[t^{S_N} \mid N\right]\right] = E\left[H(N)\right],$$

and by the law of the unconscious statistician (2.4)

$$E\left[H(N)\right] = \sum_{n=0}^{\infty} H(n)\mathbf{P}\left(N = n\right) = \sum_{n=0}^{\infty} E\left[t^{S_N} \mid N = n\right]\mathbf{P}\left(N = n\right)$$

and as $p_N\left(n\right) = \mathbf{P}\left(N = n\right)$, this equals

$$= \sum_{n=0}^{\infty} E\left[t^{X_1+X_2+\ldots+X_n} \mid N = n\right] p_N\left(n\right) = \sum_{n=0}^{\infty} E\left[t^{X_1+X_2+\ldots+X_n}\right] p_N\left(n\right),$$

where we took advantage of the assumed independence between the r.v.'s in the sum and the variable N (an independent condition drops out). But then (5.15) yields

$$= \sum_{n=0}^{\infty} \left(g_X(t)\right)^n p_N\left(n\right).$$

In view of the definition of the p.g.f. of N the last expression is seen to equal

$$= g_N\left(g_X(t)\right).$$

■

We refer to $g_{S_N}(t) = g_N\left(g_X(t)\right)$ as the **composition formula**. An inspection of the preceding proof shows that the following more general composition formula is also true.

Theorem 5.5.2 (Composition Formula with Characteristic Function) $X_1,\ X_2,\ \ldots, X_n, \ldots$ are independent and identically distributed random variables with the characteristic function $\varphi_X(t)$. N is independent of the X_ks and has the non negative integers as values with the p.g.f. $g_N(t)$. The characteristic function $\varphi_{S_N}(t)$ of S_N defined in (5.16) is

$$\varphi_{S_N}(t) = g_N(\varphi_X(t)). \tag{5.18}$$

■

Example 5.5.3 Let $N \in \text{Po}(\lambda)$, $X_k \in \text{Be}(p)$ for $k = 1, 2, \ldots$. From (5.6) $g_X(t) = q + pt$ and from (5.4) $g_N(t) = e^{\lambda(t-1)}$. Then (5.17) becomes

$$g_{S_N}(t) = g_N(g_X(t)) = e^{\lambda(q+pt-1)} = e^{\lambda(1-p+pt-1)} = e^{\lambda p(t-1)},$$

i.e.,

$$g_{S_N}(t) = e^{\lambda p(t-1)}.$$

By uniqueness of p.g.f.s we have thus obtained that $S_N \in \text{Po}(\lambda p)$. The result is intuitive: we can think of first generating N ones (1) and then deciding for each of these ones, whether to keep it or not by drawing independently from a Bernoulli random variable. Then we add the ones that remain. This can be called 'thinning' of the initial Poisson r.v.. Therefore thinning of Po(λ) is probabilistically nothing else but drawing an integer from Poisson r.v. with the intensity λ modulated by p, Po(λp).

■

The result in theorem 5.5.1 has many nice consequences, when combined with the moment formulas in section 5.3. Let us assume that all required moments exist.

- By (5.12) $g_X^{(1)}(1) = E[X]$ and thus

  $$E[S_N] = g_{S_N}^{(1)}(1) = g_N^{(1)}(g_X(1)) \cdot g_X^{(1)}(1)$$

  and since $g_X(1) = 1$,

  $$= g_N^{(1)}(1) \cdot g_X^{(1)}(1) = E[N] \cdot E[X].$$

  In summary

  $$E[S_N] = E[N] \cdot E[X]. \tag{5.19}$$

- We shall next show

  Lemma 5.5.4

  $$\text{Var}[S_N] = \text{Var}[N](E[X])^2 + E[N]\,\text{Var}[X]. \tag{5.20}$$

  Proof: We start handling this by (5.13) and get

  $$\text{Var}[S_N] = g_{S_N}^{(2)}(1) + g_{S_N}^{(1)}(1) - \left(g_{S_N}^{(1)}(1)\right)^2. \tag{5.21}$$

  We need the second derivative of $g_N(g_X(t))$, or

  $$g_{S_N}^{(2)}(t) = g_N^{(2)}(g_X(t)) \cdot \left(g_X^{(1)}(t)\right)^2 + g_N^{(1)}(g_X(t))\,g_X^{(2)}(t),$$

  and find its value at $t = 1$ as

  $$g_{S_N}^{(2)}(1) = g_N^{(2)}(1) \cdot \left(g_X^{(1)}(1)\right)^2 + g_N^{(1)}(1)\,g_X^{(2)}(1). \tag{5.22}$$

Now we have by the rule for factorial moments (5.11) both

$$g_N^{(2)}(1)) = E[N(N-1)] = E\left[N^2\right] - E[N],$$

and

$$g_X^{(2)}(1)) = E\left[X^2\right] - E[X].$$

By inserting these formulas in (5.22) we get

$$g_{S_N}^{(2)}(1) = \left(E\left[N^2\right] - E[N]\right)(E[X])^2 + E[N]\left(E\left[X^2\right] - E[X]\right).$$

In addition by (5.19)

$$g_{S_N}^{(1)}(1) = E[N] \cdot E[X].$$

When we insert these in (5.21), i.e., in

$$\mathrm{Var}[S_N] = g_{S_N}^{(2)}(1) + g_{S_N}^{(1)}(1) - \left(g_{S_N}^{(1)}(1)\right)^2$$

we get

$$= \left(E\left[N^2\right] - E[N]\right)(E[X])^2 + E[N]\left(E\left[X^2\right] - E[X]\right) + E[N] \cdot E[X] - (E[N] \cdot E[X])^2.$$

We simplify this step-by-step (one simplification per line), e.g., with eventual applications of Steiner$'$s formula (2.6),

$$= E\left[N^2\right](E[X])^2 - E[N](E[X])^2 + E[N]\left(E\left[X^2\right] - E[X]\right) + E[N] \cdot E[X] - (E[N] \cdot E[X])^2$$

$$= \left(E\left[N^2\right] - (E[N])^2\right)(E[X])^2 - E[N](E[X])^2 + E[N]\left(E\left[X^2\right] - E[X]\right) + E[N] \cdot E[X]$$

$$= \mathrm{Var}[N](E[X])^2 - E[N](E[X])^2 + E[N]\left(E\left[X^2\right] - E[X]\right) + E[N] \cdot E[X]$$

$$= \mathrm{Var}[N](E[X])^2 - E[N](E[X])^2 + E[N]E\left[X^2\right] - E[N]E[X] + E[N] \cdot E[X]$$

$$= \mathrm{Var}[N](E[X])^2 - E[N](E[X])^2 + E[N]E\left[X^2\right]$$

$$= \mathrm{Var}[N](E[X])^2 + E[N]\left(E\left[X^2\right] - (E[X])^2\right)$$

$$= \mathrm{Var}[N](E[X])^2 + E[N]\,\mathrm{Var}[X],$$

which is (5.20), as was to be shown. ■

5.6 The Probability of an Even Number of Successes

One of the powerful applications of ordinary generating functions is to solve various recurrences or difference equations, see [41, 45]. As one demonstration of these capabilities, we compute the probability of an even number of successes in the first k of an infinite Bernoulli sequence.

We consider an infinite sequence $\{X_n\}_{n\geq 1}$ of independent $\mathrm{Be}(p)$ -distributed random variables, which means that p is the probability of success and $q = 1 - p$ is the probability of failure for every X_n. We refer to the X_n as (Bernoulli) trials. For any $k \geq 1$

$$E_k = \{\text{an even number of successes in the first } k \text{ trials}\}.$$

Since the infinite sequence lacks memory due to independence, we can always drop a finite number of trials in the beginning and yet, in this new infinite sequence, the probability $\mathbf{P}(E_k)$ is for any k unaffected.

If the first trial is a failure, in order for the outcome to be in E_k, there must be an even number of successes in the next $k-1$ trials (lack of memory), or, in other words the outcome of the next $k-1$ trials is in E_{k-1}. If the first trial is a success, then there must be an odd number of successes in the next $k-1$ trials, or, in other words the outcome of the next $k-1$ trials is in the complement E_{k-1}^c. Thus we can write

$$E_k = (E_{k-1} \cap \{\text{failure in the first trial }\}) \cup \left(E_{k-1}^c \cap \{\text{success in the first trial }\}\right).$$

This expresses E_k as a union of two disjoint events, and therefore

$$= \mathbf{P}(E_{k-1} \cap \{\text{failure in the first trial }\}) + \mathbf{P}\left(E_{k-1}^c \cap \{\text{success in the first trial }\}\right).$$

But as the trials are independent, we get

$$\mathbf{P}(E_k) = \mathbf{P}(\{\text{failure in the first trial }\})\,\mathbf{P}(E_{k-1}) + \mathbf{P}(\{\text{success in the first trial }\})\,\mathbf{P}\left(E_{k-1}^c\right). \tag{5.23}$$

We let p_k be defined by

$$p_k \overset{\text{def}}{=} \mathbf{P}(E_k).$$

Then we can write (5.23) as the **difference equation** or **recursion**

$$p_k = qp_{k-1} + p(1 - p_{k-1}). \tag{5.24}$$

This is actually valid only for $k \geq 2$. Namely, if $k = 1$, an even number of successes can come about in only one way, namely by making zero successes, and thus we take $p_1 = q$. If the equation in (5.24) is to hold for $k = 1$, i.e.,

$$q = p_1 = qp_0 + p(1 - p_0),$$

we must take $p_0 = 1$. The initial conditions for the equation in (5.24) must therefore be taken as

$$p_1 = q, p_0 = 1. \tag{5.25}$$

Here is a first method of solution of (5.24). We write (5.24) as

$$p_k - (q-p)p_{k-1} = p. \tag{5.26}$$

Hence we are dealing with a non homogeneous linear difference equation of first order with constant coefficients. One can solve (5.26) with the analytic techniques of difference equations [45, pp. 13−14]. We consider first the homogeneous equation

$$p_k - (q-p)p_{k-1} = 0.$$

The standard 'Ansatz' for its solution is $p_k = c_1 z^k$, where c_1 is a constant to be determined. This gives clearly the general solution of the homogeneous difference equation as $p_k^H = c_1(q-p)^k$. We need next to find a particular solution of the nonhomogenous equation

$$p_k - (q-p)p_{k-1} = p.$$

In this situation one seeks for a constant as a particular solution. One sees that $p_k^S = c_2\frac{1}{2}$ is a particular solution, where c_2 is a constant to be determined. Then we have by linearity the complete solution of (5.24) as the sum of the two solutions

$$p_k = p_k^H + p_k^S = c_1(q-p)^k + c_2\frac{1}{2}.$$

The constants c_1 and c_2 are next determined by the two initial conditions $p_0 = 1$ and $p_1 = q$. This gives the system of equations $c_1 + \frac{c_2}{2} = 1$, $c_1(1-2p) + \frac{c_2}{2} = (1-p)$. Its solution is $c_1 = \frac{1}{2}$ and $c_2 = 1$. Hence we have obtained the complete solution as

$$p_k = \frac{1}{2}(q-p)^k + \frac{1}{2} = \frac{1}{2}\left(1 + (q-p)^k\right).$$

This is the expression we should rediscover by using the generating function.

Let us introduce the (ordinary) generating function, see [55, pp. 86-87] or [35],

$$G(t) = \sum_{k=0}^{\infty} p_k t^k.$$

When we first multiply both sides of (5.24) by t^k and then sum over $k = 1, 2, \ldots$

$$\begin{aligned}\sum_{k=1}^{\infty} p_k t^k &= qt\sum_{k=1}^{\infty} p_{k-1}t^{k-1} + pt\sum_{k=1}^{\infty} t^{k-1} - pt\sum_{k=1}^{\infty} t^{k-1}p_{k-1} \\ &= qt\sum_{k=0}^{\infty} p_k t^k + pt\sum_{k=0}^{\infty} t^k - pt\sum_{k=0}^{\infty} t^k p_k. \qquad (5.27)\end{aligned}$$

By (5.25) we observe that

$$\sum_{k=1}^{\infty} p_k t^k = G(t) - p_0 = G(t) - 1.$$

Then we have in (5.27) that

$$G(t) - 1 = qtG(t) + \frac{pt}{1-t} - ptG(t),$$

where we have symbolically used $\sum_{k=0}^{\infty} t^k = \frac{1}{1-t}$. We solve algebraically w.r.t. $G(t)$ to get

$$G(t) = \frac{1}{1-qt+pt} + \frac{pt}{(1-t)(1-qt+pt)}.$$

An expansion by partial fractions yields

$$\begin{aligned}G(t) &= \frac{1}{1-qt+pt} + \frac{p}{1-q+p}\frac{1}{1-t} + \frac{p}{1-q+p}\frac{1}{1-qt+pt} \\ &= \frac{1}{2}\frac{1}{1-t} + \frac{1}{2}\frac{1}{1-qt+pt},\end{aligned}$$

where we used $1 - q + p = 2p$. Thereby

$$2G(t) = \frac{1}{1-t} + \frac{1}{1-qt+pt}.$$

If we recast the two terms in the right hand side as sums of respective geometric series we obtain

$$2\sum_{k=0}^{\infty} p_k t^k = \sum_{k=0}^{\infty} t^k + \sum_{k=0}^{\infty}(q-p)^k t^k = \sum_{k=0}^{\infty}(1+(q-p)^k)t^k. \qquad (5.28)$$

When we identify the coefficients of t^k in the power series in the both sides of (5.28) we get

$$p_k = \frac{1}{2}\left(1 + (q-p)^k\right) \quad k \geq 0, \qquad (5.29)$$

which agrees with the expression found by the first method.

5.7 Moment Generating Functions

5.7.1 Definition and First Properties

In this section we consider general random variables X in the sense that X need not have non negative integers as values.

Definition 5.7.1 (Moment generating function) The moment generating function (m.g.f.) $g_X(t)$ for a random variable X is defined by

$$\psi_X(t) \overset{\text{def}}{=} E\left[e^{tX}\right] = \begin{cases} \sum\limits_{k=-\infty}^{\infty} e^{tx_k} p_X(x_k) & \text{discrete r.v.,} \\ \int\limits_{-\infty}^{\infty} e^{tx} f_X(x)\, dx & \text{continuous r.v.,} \end{cases} \tag{5.30}$$

if there is a positive real number h such that $E\left[e^{tX}\right]$ exists for $\mid t \mid < h$.

■

The requirement of existence of $E\left[e^{tX}\right]$ is not satisfied for any $h > 0$, for example, by a random variable $X \in C(0,1)$. Thus $X \in C(0,1)$ has no m.g.f. and, as has been pointed out in example 2.2.16, has no moments either for that matter. Having said that, let us note that the analysis of optical fiber communication in [33] is completely based on m.g.f.s. The pertinent uniqueness theorem is there, but we omit the proof.

Theorem 5.7.1 (Uniqueness) If X and Y are two random variables such that

$$\psi_X(t) = \psi_Y(t) \quad \text{for all } |t| < h \ ,$$

then

$$X \overset{d}{=} Y$$

('X and Y are equal in distribution').

■

The proof of the following theorem should be obvious in view of the proofs of the analogous theorems for characteristic and probability generating functions in the preceding .

Theorem 5.7.2 If $X_1,\ X_2,\ \ldots, X_n$ are independent random variables with respective m.g.f.s $\psi_{X_k}(t)$, $k = 1, 2, \ldots, n$, which all exist for $|t| < h$, for some $h > 0$. Then the m.g.f. $\psi_{S_n}(t)$ of the sum $S_n = \sum_{k=1}^{n} X_k$ is given by

$$\psi_{S_n}(t) = \psi_{X_1}(t) \cdot \psi_{X_2}(t) \cdot \ldots \cdot \psi_{X_n}(t). \tag{5.31}$$

■

There is again the immediate corollary.

Corollary 5.7.3 If $X_1,\ X_2,\ \ldots, X_n$ are independent and identically distributed random variables with the m.g.f. $\psi_X(t)$, which exists for $|t| < h$, $h > 0$. Then the m.g.f. $\psi_{S_n}(t)$ of the sum $S_n = \sum_{k=1}^{n} X_k$ is given by

$$\psi_{S_n}(t) = \left(\psi_X(t)\right)^n . \tag{5.32}$$

■

Example 5.7.4 (M.g.f. for Random Variables Taking Values in Non Negative Integers) If X is a r.v. taking values in non negative integers the m.g.f. is by definition (5.30) in the discrete case, assuming existence of $\psi_X(t)$,

$$\psi_X(t) = \sum_{k=0}^{\infty} e^{tk} p_X(k) = \sum_{k=0}^{\infty} \left(e^t\right)^k p_X(k) = g_X\left(e^t\right),$$

where $g_X\left(e^t\right)$ is the p.g.f. of X with e^t in the domain of convergence of the p.g.f.. In view of this several examples of m.g.f.s are immediate. We get by (5.4)

$$X \in \text{Po}(\lambda) \Leftrightarrow \psi_X(t) = e^{\lambda(e^t-1)},$$

from (5.7)

$$X \in \text{Bin}(p) \Leftrightarrow \psi_X(t) = \left(q + e^t p\right)^n,$$

and from (5.10)

$$X \in \text{Fs}(p) \Leftrightarrow \psi_X(t) = \frac{pe^t}{1-qe^t}, \quad t < -\ln(1-p).$$

■

Example 5.7.5 (M.g.f. for $Y = aX + b$) If X is a r.v. with the m.g.f. $\psi_X(t)$, which exists for $|at| < h$, and $Y = aX + b$, where a and b are real numbers, then

$$\psi_Y(t) = e^{tb} \cdot \psi_X(at). \tag{5.33}$$

■

Example 5.7.6 (M.g.f. for $X \in N(0,1)$) If X is $\in N(0,1)$, we have by the definition (5.30)

$$\psi_X(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

and complete the square to get

$$= e^{\frac{t^2}{2}} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-t)^2/2}\, dx}_{=1} = e^{\frac{t^2}{2}}.$$

Here we used the fact that the integrand in the underbraced integral is the probability density of $N(t,1)$. This m.g.f. exists for all t.

$$X \in N(0,1) \Leftrightarrow \psi_X(t) = e^{\frac{t^2}{2}}. \tag{5.34}$$

■

Example 5.7.7 (M.g.f. for $X \in N(\mu, \sigma^2)$) If $X \in N(\mu, \sigma^2)$, we have shown in example 4.2.5 above that if $Z \in N(0,1)$ and $X = \sigma Z + \mu$, then $X \in N(\mu, \sigma^2)$. Then as in (5.33)

$$\psi_X(t) = e^{t\mu} \cdot \psi_Z(\sigma t),$$

and this gives by (5.34)

$$\psi_X(t) = e^{t\mu} e^{\frac{\sigma^2 t^2}{2}} = e^{t\mu + \frac{\sigma^2 t^2}{2}}.$$

$$X \in N(\mu, \sigma^2) \Leftrightarrow \psi_X(t) = e^{t\mu + \frac{\sigma^2 t^2}{2}}. \tag{5.35}$$

■

Example 5.7.8 (M.g.f. for a sum of independent normal random variables) Let $X_1 \in N(\mu_1, \sigma_1^2)$ and $X_2 \in N(\mu_2, \sigma_2^2)$ be independent. Then by (5.31)

$$\psi_{X_1+X_2}(t) = \psi_{X_1}(t) \cdot \psi_{X_2}(t) =$$

and by (5.35)

$$= e^{t\mu_1+\frac{\sigma_1^2 t^2}{2}} \cdot e^{t\mu_2+\frac{\sigma_2^2 t^2}{2}} = e^{t(\mu_1+\mu_2)+\frac{(\sigma_1^2+\sigma_2^2)t^2}{2}}.$$

Hence we have again established that

$$X_1 + X_2 \in N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2). \tag{5.36}$$

■

Example 5.7.9 (M.g.f. for an Exponential Random Variable) Let $X \in \text{Exp}(a)$, $a > 0$. The p.d.f. is

$$f_X(x) = \begin{cases} \frac{1}{a} e^{-x/a} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

The definition in (5.30) entails

$$\psi_X(t) = \int_0^\infty e^{tx} \frac{1}{a} e^{-x/a}\, dx = \frac{1}{a} \int_0^\infty e^{-x\left(\frac{1}{a}-t\right)}\, dx$$

$$= \frac{1}{a} \left[\frac{-1}{\left(\frac{1}{a}-t\right)} e^{-x\left(\frac{1}{a}-t\right)} \right]_0^{+\infty},$$

and if $\frac{1}{a} - t > 0$, i.e., if $\frac{1}{a} > t$, we have

$$= \frac{1}{a} \frac{1}{\left(\frac{1}{a}-t\right)} = \frac{1}{(1-at)}.$$

Thereby we have found

$$X \in \text{Exp}(a) \Leftrightarrow \psi_X(t) = \frac{1}{(1-at)}, \quad \frac{1}{a} > t. \tag{5.37}$$

■

Example 5.7.10 (M.g.f. for a Gamma (Erlang) Random Variable) $X \in \Gamma(n, a)$, where n is a positive integer. In other words, we consider an Erlang distribution. Then example 4.4.9 and (5.37) yield

$$X \in \Gamma(n, \lambda) \Leftrightarrow \psi_X(t) = \left(\frac{1}{1-at} \right)^n, \quad \frac{1}{a} > t. \tag{5.38}$$

■

The proof of the statement in theorem 5.5.1 can be modified in an obvious manner to establish the following composition rule.

Theorem 5.7.11 (Composition Rule with m.g.f) $X_1, X_2, \ldots, X_n, \ldots$ are independent and identically distributed random variables with the m.g.f. $\psi_X(t)$ for $|t| < h$. N is independent of the X_ks and has the non negative integers as values and with the p.g.f. $g_N(t)$. The m.g.f. $\psi_{S_N}(t)$ of S_N defined in (5.16) is

$$\psi_{S_N}(t) = g_N\left(\psi_X(t)\right), \tag{5.39}$$

wheer we assume that $\psi_X(t)$ is in the domain of convergence of $g_N(t)$.

■

5.7.2 M.g.f. is really an Exponential Moment Generating Function, E.m.g.f !

The introduction to this chapter stated that ordinary generating functions of sequences of real numbers $(a_k)_{k=0}^{\infty}$ are functions (power series) of the form

$$G(t) = \sum_{k=0}^{\infty} a_k t^k.$$

Yet, we have welcomed the m.g.f. as defined in (5.30), which is not at face value compatible the idea of ordinary generating functions. For the sake of pedagogic coherence it should be appropriate to settle the issue[1].

Let us suppose that we have a random variable X such that all moments $E\left[X^k\right]$ $k = 1, 2, \ldots$, exist. Then a generating function for the sequence $\left(E\left[X^k\right]\right)_{k=0}^{\infty}$ in the ordinary sense would be

$$\sum_{k=1}^{\infty} E\left[X^k\right] t^k.$$

This does not produce the moment generating function as defined in (5.30). Symbolically we have that

$$E\left[\frac{1}{1-tX}\right] = \sum_{k=0}^{\infty} E\left[X^k\right] t^k,$$

and this is the ordinary generating function of $\left(E\left[X^k\right]\right)_{k=0}^{\infty}$. On the other hand, if we set

$$\sum_{k=0}^{\infty} \frac{E\left[X^k\right]}{k!} t^k,$$

we can apply the series expansion of e^{tx} to obtain by termwise expectation

$$E\left[e^{tX}\right] = \sum_{k=0}^{\infty} \frac{E\left[X^k\right]}{k!} t^k, \tag{5.40}$$

which equals the moment generating function $\psi_X(t)$, as treated above. However, in mathematics, see, e.g., [41, p. 350], the power series

$$\mathrm{EG}(t) = \sum_{k=0}^{\infty} \frac{a^k}{k!} t^k$$

is called the **exponential generating function** of the sequence of real numbers $(a_k)_{k=0}^{\infty}$. In order to adhere to the standard mathematical terminology we should hence call any $\psi_X(t) = E\left[e^{tX}\right]$ the exponential moment generating function (e.m.g.f.).

But the practice of talking about moment generating functions has become well established and is thereto time-honoured. There is in other words neither a pragmatic reason to campaign for a change of terminology to e.m.g.f.'s, nor a realistic hope of any success in that endeavour.

The **take-home message** is the following theorem.

[1] The point is made by J.P. Hoyt in *The American Statistician*, vol. 26, June 1972, pp. 45−46.

Theorem 5.7.12 Let X be a random variable with m.g.f. $\psi_X(t)$ that exists for $|t| < h$ for some $h > 0$. Then

(i) For all $k > 0$, $E\left[|X|^k\right] < \infty$, i.e, moments of all orders exist.

(ii)
$$E\left[X^k\right] = \psi_X^{(k)}(0). \tag{5.41}$$

Proof: We omit the proof of (i). To prove (ii) we observe that

$$\psi_X(t) = \sum_{k=0}^{\infty} \frac{E\left[X^k\right]}{k!} t^k, \quad |t| < h, \tag{5.42}$$

by successive differentiation one finds that the coefficient of $\frac{t^k}{k!}$ is equal to $\psi_X^{(k)}(0)$. ■

5.8 Exercises

5.8.1 Probability Generating Functions

1. **Pascal Distribution** Let $X \in \text{Pascal}(n, p)$, $n = 1, 2, 3, \ldots$, $0 < p < 1$ and $q = 1 - p$, see Example 2.3.10. Its probability mass function is then

$$p_X(k) = \mathbf{P}(X = k) = \left(\begin{array}{c} k-1 \\ n-1 \end{array}\right) p^n q^{k-n}, \quad k = n, n+1, n+2, \ldots. \tag{5.43}$$

Show that the p.g.f. of X is

$$g_X(t) = \left(\frac{pt}{1-qt}\right)^n, \quad |t| < q^{-1}.$$

Note: Consider also the examples 5.2.8 and 5.2.9.

2. **Negative Binomial Distribution** Let X_i, $i = 1, 2, \ldots, n$, be independent and have the distribution $X_i \in \text{Ge}(p)$. Define

$$Y = X_1 + X_2 + \ldots + X_n.$$

(a) Find the p.g.f. of Y.

(b) Show that if $Z \in \text{Pascal}(n, p)$, then $Y \stackrel{d}{=} Z - n$.

(c) Show that the probability mass function of Y is

$$p_Y(k) = \left(\begin{array}{c} n+k-1 \\ k \end{array}\right) p^n q^k, \quad k = 0, 1, 2, \ldots$$

Hence Y has the Negative Binomial distribution, $Y \in \text{NBin}(n, p)$. *Aid:* The part (c) does not require a generating function. Use the finding in (b) and (5.43).

3. X_1,$X_2, \ldots, X_n$ are independent Poisson distributed random variables with $E\left[X_k\right] = \frac{1}{k}$. Show that the p.g.f. of $Y_n = \sum_{k=1}^{n} kX_k$ is

$$g_{Y_n}(t) = e^{\sum_{k=1}^{n} \frac{t^k - 1}{k}}.$$

4. N assumes values in the nonnegative integers.

(a) Show that $\frac{g_N(t)-1}{t-1} = \sum_{k=0}^{\infty} \mathbf{P}(N > k) t^k, \quad$ for $|t| < 1$.

(b) Show that $E[N] = \sum_{k=0}^{\infty} \mathbf{P}(N > k)$.

5. ([35]) The r.v.'s X_1,$X_2, \ldots, X_n$ are independent and identically disributed. Their common distribution function is $F_X(x)$. We consider the random variable N, which has the positive integers as values and has the p.g.f. $g_N(t)$. N is independent of X_1,$X_2, \ldots, X_n$. Set

$$Y = \max\{X_1, X_2, \ldots, X_N\}.$$

Show that

$$F_Y(y) = g_N(F_X(y)).$$

Aid: The law of total probability (3.35) may turn out to be useful.

6. (From [49]) The r.v. X has the p.g.f. $g_X(t) = \ln\left(\frac{1}{1-qt}\right)$. Determine $E[X]$, $\text{Var}[X]$, and the p.m.f. of X. *Answers:* $E[X] = \text{Var}[X] = e - 1$, $p_X(k) = \frac{(1-e^{-1})^k}{k}$, $k \geq 1$.

7. Let us introduce

$$\tilde{g}_{X_1-X_2}(t) \stackrel{\text{def}}{=} E\left[t^{X_1-X_2}\right],$$

where X_1 and X_2 are two independent r.v.'s with non negative integers as values. Hence $\tilde{g}$ is an extension of the notion of p.g.f. to a random variable with integers as values. Let $X_1 \in \text{Po}(\mu_1)$ and $X_2 \in \text{Po}(\mu_2)$, X_1 and X_2 are independent.

(a) Show that

$$\tilde{g}_{X_1-X_2}(t) = e^{-(\mu_1+\mu_2)+\mu_1 t+\mu_2/t}. \tag{5.44}$$

(b) Find the p.m.f of $X_1 - X_2$ by means of the extended p.g.f. in (5.44).
Aid: The generating function of modified Bessel functions $I_k(x)$ of the first kind is

$$e^{x\left(t+\frac{1}{t}\right)} = \sum_{k=-\infty}^{\infty} I_k(x)t^k, t \neq 0,$$

see [3, pp. 292–293].
Answer: $X_1 - X_2 \in \text{Ske}(\mu_1, \mu_2)$, see (2.62) [2] .

5.8.2 Moment Generating Functions

1. X is a random variable with values in the non negative integers. We know that

$$E[X] = 1.$$

Let B be the event $B = \{X > 0\}$. We consider X truncated to the positive integers, $X|B$, i.e., X conditioned on B (recall section 3.3). We have in addition that

$$X|B \in \text{Fs}(p).$$

Find the m.g.f. of X as

$$\psi_X(t) = 1 - p + \frac{p^2 e^t}{1-(1-p)e^t}.$$

[2] This is how John Gordon Skellam (1914-1979), a statistician and ecologist, derived the p.m.f. (2.62).

2. $X \in \text{Ra}\,(a)$, c.f., example 2.2.17. Compute the m.g.f. of X. *Answer:*
$$\psi_X(t) = 1 + \frac{a}{2} t\, e^{(\frac{a}{2})^2 t^2/2} \sqrt{\frac{\pi}{2}} \left(\text{erf} \left(\frac{\frac{a}{2}t}{\sqrt{2}} \right) + 1 \right),$$
where erf is the error function
$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt, \quad -\infty < x < \infty$$

3. **M.g.f.** of the **Gumbel distribution** Let distribution function of the r.v. X be
$$F_X(x) = e^{-e^{-x}}, \quad -\infty < x < \infty.$$
Or, $X \in$ Gumbel, as defined in example 2.2.19.
   (a) Find the m.g.f. of X. What is the region of existence ? *Answer:* $\psi_X(t) = \Gamma(1-t)$, $|t| < 1$.
   Aid: Find the p.d.f. of X and use the appropriate part of the definition in (5.30). In the resulting integral make the change of variable $u = e^x$ and be sure to find the right limits of integration.
   (b) Show that $E\,[X] = \gamma =$ Euler's constant.
   Aid: Karl Weierstrass[3] re-expressed (You need not check this) the Gamma function in (2.7) with
$$\frac{1}{\Gamma(x)} = x e^{\gamma x} \prod_{r=1}^{\infty} \left(1 + \frac{x}{r}\right) e^{-\frac{x}{r}},$$
   where γ is Euler's constant. Show now that
$$\frac{\frac{d}{dx}\Gamma(x)}{\Gamma(x)} = -\frac{1}{x} - \gamma + \sum_{r=1}^{\infty} \left(\frac{1}{r} - \frac{1}{r+x} \right).$$
   The function $\frac{\frac{d}{dx}\Gamma(x)}{\Gamma(x)}$ is known as the **Digamma** function.
   (c) Show that $\text{Var}\,[X] = \frac{\pi^2}{6}$.

4. Find the m.g.f. of the logistic distribution with p.d.f. in (2.39). *Answer*: $B(1-t, 1+t)$, $-1 < t < 1$, where B is the Beta function given in (2.31).

5. **Difference of two independent Gumbel variables** $V \in$ Gumbel and $W \in$ Gumbel are independent. In other words their common distribution function is found in (2.35). Show that
$$U = V - W \in \text{logistic}(0, 1),$$
where the distribution logistic$(0, 1)$ is given in Example 2.2.23.
Hint: The two directly preceding exercises should be useful.

6. (a) Find the m.g.f. of $X \in U(0, 2)$. *Answer:* $\psi_X(t) = \frac{1}{2t}\left(e^{2t} - 1\right)$.
   (b) $E\,[X^n] = \frac{2^n}{n+1}$. Determine the distribution of X. *Answer:* It is an easy guess that case (a) has something to do with this.

7. $E\,[X^n] = \frac{1}{n+1}$. Find the distribution of X.

8. $E\,[X^n] = c$ for $n = 1, 2, \ldots,$. Find the distribution of X. *Answer:* $X \in \text{Be}(c)$, if $0 \leq c \leq 1$. There is no solution for $c \notin [0, 1]$.

[3] (1815−1897) spent a long career as teacher of mathematics at a gymnasium (high school) in Germany. He became Professor of mathematics at Technical University of Berlin. Weierstrass has a fame in posterity as the 'father of modern analysis'.

5.8.3 Sums of a Random Number of Independent Random Variables

1. Let X be a random variable assuming values in $0, 1, 2, \ldots$. Assume that

$$\begin{aligned} p_X(0) &= 0, \\ p_X(k) &= \int_0^1 u \cdot (1-u)^{k-1} du, \quad k = 1, 2, \ldots. \end{aligned} \tag{5.45}$$

   (a) Show that
$$p_X(k) = \frac{1}{k(k+1)}, \quad k = 1, 2, \ldots.$$
   Aid: Apply a suitable Beta function (2.31).

   (b) Show that the p.g.f. $g_X(t)$ is
$$g_X(t) = 1 + \frac{(1-t)\ln(1-t)}{t}.$$

   (c) Let X_1,X_2, $\ldots$, $X_n, \ldots$ be I.I.D. with the probability mass function in (a). $N \in \text{Po}\,(m)$ and N is independent of X_k:s. Set
$$S_N = X_1 + X_2 + \ldots + X_N, \quad S_0 = 0.$$
   Show that the p.g.f. of S_N is
$$g_{S_N}(t) = (1-t)^{m\frac{(1-t)}{t}}.$$
   Check that you get $g_{S_N}(0) = 1$. What is the distribution of S_N ? *Hint:* Try the world wide web with **Lea-Coulson Model for Luria -Delbrück Distribution** or **Lea-Coulson Probability Generating Function for Luria -Delbrück Distribution** as search words.

2. (5B1540 02-08-21, reconsidered) The random variables $X_1, X_2, \ldots, X_n, \ldots,$ are I.I.D. and have the probability mass function
$$p_X(-1) = \frac{1}{4}, p_X(0) = \frac{1}{2}, p_X(1) = \frac{1}{4}.$$
   Let $N \in \text{Fs}\left(\frac{1}{2}\right)$ and be independent of the X_k's. Find the characteristic function of $S_N = \sum_{k=1}^{N} X_k$. (Aid: Use (5.18).)

   In an exercise to chapter 4 we defined for the same r.v.'s X_n the random variable $N^{'}$ by
$$N^{'} = \min\{n \mid X_n = 0\},$$
   so that $N^{'}$ is the smallest (or first) n such that $X_n = 0$. It was found that the characteristic function of is $\varphi_{S_{N^{'}}}(t) = 1/(2 - \cos t)$. What is the reason for the difference in the results about N and $N^{'}$?

3. (FA 181 1982-02-05) Let $X_1, X_2, \ldots, X_n, \ldots$ be independent and identically distributed with $X_k \in N(0, 1)$, $k = 1, 2, \ldots, n$. N is a random variable with values in the positive integers $\{1, 2, \ldots\}$. N is independent of the variables $X_1, X_2, \ldots, X_n, \ldots$. We set
$$S_N = X_1 + X_2 + \ldots + X_N.$$
   We assume that
$$\mathbf{P}\,(N = k) < 1, k = 1, 2, \ldots.$$
   Show now that S_N cannot be a normal random variable, no matter what distribution N has, as long as this distribution satisfies our assumptions. *Aid:* The result in (5.18) should turn out to be useful.

4. Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of independent and identically distributed r.v.'s $\in$ Po (2). N is independent of the X_n, $N \in$ Po(1). Set $S_N = X_1 + X_2 + \ldots + X_N$, $S_0 = 0$. Find using the appropriate p.g.f.'s that

$$\mathbf{P}\left(S_N = 0\right) = e^{e^{-2}-1}. \tag{5.46}$$

Compare with (3.39) in the preceding.

Answer: $1 - 1.4 \cdot 0.9^2$.

6. (Due to Harald Lang) $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of independent and identically distributed r.v.'s that assume values in the non negative integers. We have $\mathbf{P}\left(X_n = 0\right) = p$. N assumes also values in the non negative integers, is independent of the X_n and has the p.g.f. $g_N(t)$. Set $S_N = X_1 + X_2 + \ldots + X_N$, $S_0 = 0$. Express $\mathbf{P}\left(S_N = 0\right)$ in terms of $g_N(t)$ and p. *Answer:* $g_N(p)$.

7. (Due to Harald Lang) Let p be the probability that when tossing a thumbtack (North American English), or a drawing pin (British English) [4] it falls on its pin (not on its head). A person tosses a thumbtack a number of times, until the toss results in falling on the pin for the first time. Let X be the serial number of the toss, when the falling on the pin occurred for the first time. After that the person tosses the thumbtack an additional X times. Let Y be the number of times the thumbtack falls on its pin in the latter sequence of tosses. Find the p.m.f. of Y,

   (a) by a reasoning that evokes conditional probability,

   (b) by finding the p.g.f. of Y.

   Answer: $p_Y(k) = \frac{(1-p)^{k-1}}{(2-p)^{k+1}}$, $k \geq 1$.

8. **Compound Poisson Distribution** Let X_k, $k = 1, 2, \ldots$, be independent and have the distribution $X_k \in \text{Po}(\mu)$. Let $N \in \text{Po}(\lambda)$. N is independent of the X_k's. Set

$$S_N = X_1 + X_2 + \ldots + X_N.$$

   (a) Show that the p.g.f. of S_N is

$$g_{S_N}(t) = e^{\lambda\left(e^{\mu(t-1)}-1\right)}. \tag{5.47}$$

   (b) Show that

$$E\left[S_N\right] = \lambda\mu, \text{Var}\left[S_N\right] = \lambda\mu(1+\mu).$$

   (c) In fact a good definition of the compound Poisson distribution is that it is the probability distribution on the non negative integers with the p.g.f. in (5.47). In example 2.3.9 the compound Poisson distribution was defined in terms of the p.m.f. in (2.54). Explain why the two definitions actually deal with one and the same thing, i.e., $S_N \in \text{ComPo}(\lambda, \mu)$ in the sense of example 2.3.9.

9. Let $X_1, X_2, \ldots, X_n, \ldots$ be independent and identically distributed with $X_k \in \text{Exp}(1/a)$, $k = 1, 2, \ldots,$. $N \in \text{Fs}(p)$. N is independent of the variables $X_1, X_2, \ldots, X_n, \ldots$. We set

$$S_N = X_1 + X_2 + \ldots + X_N.$$

   Show that $S_N \in \text{Exp}\left(\frac{1}{pa}\right)$.

[4] a short nail or pin with usually a circular head, used to fasten items such as documents to a wall or board for display. In Swedish: **häftstift**.

10. (From [49]) Let $0 < p < 1$. $q = 1 - p$. $X_1, X_2, \ldots, X_n, \ldots$ are independent and identically distributed with $X_k \in \mathrm{Ge}(q)$, $k = 1, 2, \ldots,$. $N \in \mathrm{Ge}(p)$. N is independent of the variables $X_1, X_2, \ldots, X_n, \ldots$. We set

$$S_N = X_1 + X_2 + \ldots + X_N.$$

(a) Show that the p.m.f. of S_N is $p_{S_N}(k) = \frac{(1-p)^2}{(2-p)^{k+1}}$, $k \geq 1$, $p_{S_N}(0) = \frac{1}{2-p}$.

(b) Show that $S_N \mid S_N > 0 \in \mathrm{Fs}(a)$, and show that $a = \frac{1-p}{2-p}$.

11. (From [49]) Let $X_1, X_2, \ldots, X_n, \ldots$ be independent and identically distributed with $X_k \in L(a)$, $k = 1, 2, \ldots,$. $N_p \in \mathrm{Fs}(p)$. N_p is independent of the variables $X_1, X_2, \ldots, X_n, \ldots$. We set

$$S_{N_p} = X_1 + X_2 + \ldots + X_{N_p}.$$

Show that $\sqrt{p} S_{N_p} \in L(a)$.

12. (From [49]) Let $X_1, X_2, \ldots, X_n, \ldots$ be independent and identically distributed with $X_k \in \mathrm{Po}(2)$, $k = 1, 2, \ldots,$. $N \in \mathrm{Po}(1)$. N is independent of the variables $X_1, X_2, \ldots, X_n, \ldots$. We set $S_0 = 0$, and

$$S_N = X_1 + X_2 + \ldots + X_N.$$

Show that

$$E\left[S_N\right] = 2, \mathrm{Var}\left[S_N\right] = 6.$$

13. (From [49]) Let $X_1, X_2, \ldots, X_n, \ldots$ be independent and identically distributed. N is independent of the variables and has the non negative integers as values.

$$S_N = X_1 + X_2 + \ldots + X_N.$$

Show that

$$\mathrm{Cov}\left(X_1 + X_2 + \ldots + X_N, N\right) = E\left[X\right] \cdot \mathrm{Var}\left[N\right].$$

14. (From [49]) Let $X_1, X_2, \ldots, X_n, \ldots$ be independent. $E\left[X_k\right] = m^k$, where $m \neq 1$, $k = 1, 2, \ldots,$. N is independent of the variables and $\in \mathrm{Po}(\lambda)$. We set

$$S_N = X_1 + X_2 + \ldots + X_N.$$

Show that

$$E\left[X_1 + X_2 + \ldots + X_N\right] = \frac{m}{m-1}\left(e^{\lambda(m-1)} - 1\right).$$

5.8.4 Various Additional Generating Functions in Probability

1. (From [41]) The **Dirichlet probability generating function**
Dirichlet probability generating function $D_X(t)$ of a random variable X with values in the positive integers is defined as

$$D_X(t) = \sum_{k=1}^{\infty} \frac{p_X(k)}{k^t}.$$

Find $E\left[X\right]$, $\mathrm{Var}\left[X\right]$ and $E\left[\ln X\right]$ expressed in terms of $D_X(t)$ and its derivatives.

2. The **exponential generating function for factorial moments**
The kth (descending) factorial moment is denoted by $E\left[X^{[k]}\right]$ and defined by
$$E\left[X^{[k]}\right] = E\left[X(X-1)(X-2)\cdot\ldots\cdot(X-(k-1))\right].$$
Let X be a random variable that assumes values in the nonnegative integers. Show that the exponential generating function for factorial moments is
$$\mathrm{EG}_X(t) = E\left[(1+s)^X\right].$$

3. The **ordinary moment generating function** was in the preceding argued to be
$$E\left[\frac{1}{1-tX}\right] = \sum_{k=0}^{\infty} E\left[X^k\right] t^k.$$
What is one main disadvantage of this function as a tool in probability ?

4. Check that the solution in (5.29) satisfies (5.24) as well as (5.25).

5.8.5 The Chernoff Inequality

1. Let X be a random variable with the m.g.f. $\psi_X(t)$. Show that for any constant c
$$\mathbf{P}(X \geq c) \leq \min_{t\geq 0} e^{-tc}\psi_X(t). \tag{5.48}$$
Aid: Try to find a suitable way of using the Markov inequality (1.38). The inequality in (5.48) is known as the **Chernoff Inequality** or the **Chernoff Bound**.

2. Let $X_1, \ldots X_n$ be independent and identically Be(p)-distributed. Show that
$$\mathbf{P}\left(\frac{1}{n}\sum_{k=1}^{n} X_k \geq c\right) \leq \min_{t\geq 0}\left[pe^{(1-c)t} + (1-p)e^{-ct}\right]^n. \tag{5.49}$$

3. We continue with (5.49). Show that
$$\mathbf{P}\left(\frac{1}{n}\sum_{k=1}^{n} X_k \geq p+\epsilon\right) \leq \left(\left(\frac{p}{p+\varepsilon}\right)^{p+\varepsilon}\left(\frac{1-p}{1-p-\varepsilon}\right)^{1-p-\varepsilon}\right)^n. \tag{5.50}$$
Aid: Minimize the upper bound in (5.49) as a function of t by differential calculus.

We define for $0 \leq x \leq 1$ and $0 \leq y \leq 1$ the number
$$D(x|y) \overset{def}{=} x\ln\frac{x}{y} + (1-x)\ln\frac{1-x}{1-y},$$
which is non negative, as can be checked. Then we can recast the bound in (5.50) as
$$\mathbf{P}\left(\frac{1}{n}\sum_{k=1}^{n} X_k \geq p+\epsilon\right) \leq e^{-nD(p|p+\epsilon)}. \tag{5.51}$$
The number $D(p|p+\epsilon)$ is called the **Kullback distance** between the probability distributions Be(p) and Be$(p+\epsilon)$, see [23].

4. (From [33, p. 324]) $N \in \mathrm{Po}(\lambda)$. Show that
$$\mathbf{P}(N \geq a) \leq \left(\frac{\lambda}{a}\right)^a e^{a-\lambda},$$
where $a \geq \lambda$.

Chapter 6

Convergence of Sequences of Random Variables

6.1 Introduction

This chapter introduces and deals with the various modes in which a sequence of random variables defined in a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$ can be said to converge. We start by three examples (for as many different senses of convergence, there will be a fourth mode, almost sure convergence, later in this chapter).

Results about convergence of sequences are important for the same reasons as limits are important everywhere in mathematics. In probability theory we can find simple approximations to complicated or analytically unaccessible probability distributions. In section 6.5 we clarify the formulas of propagation of error by convergence of sequences. In section 7.4.2 we will give meaning to a sum of a countably infinite number of random variables that looks like $\sum_{i=0}^{\infty} a_i X_i$. In section 10.5 we will define by a limit for a Wiener process an integral that looks like $\int_a^b f(t) dW(t)$.

Example 6.1.1 (Convergence to Gumbel Distribution) Let $X_1, X_2, \ldots, X_n, \ldots$ be an I.I.D. sequence of Exp(1) -distributed r.v.'s. Let us consider the random variable

$$X_{\max} = \max\{X_1, X_2, \ldots, X_n\}.$$

It is clear that $X_{\max}$ is a well defined r.v., since it holds for any $x \in \mathbf{R}$ that $\{X_{\max} \leq x\} = \cap_{i=1}^{n} \{X_i \leq x\} \in \mathcal{F}$. We wish to understand or approximate the probabilistic behaviour of $X_{\max}$ for large values of n, which we study by letting $n \to \infty$. Let $x > 0$. By independence

$$\mathbf{P}\left(X_{\max} \leq x\right) = \mathbf{P}\left(\cap_{i=1}^{n} \{X_i \leq x\}\right) = \prod_{i=1}^{n} \mathbf{P}\left(\{X_i \leq x\}\right)$$

$$= \left(F_X(x)\right)^n = \left(1 - e^{-x}\right)^n,$$

since all $X_k \in$ Exp(1). Then

$$\mathbf{P}\left(X_{\max} \leq x\right) = \left(1 - e^{-x}\right)^n \to 0,$$

as $n \to \infty$. This is an intuitive result, but it does not contribute much to any the purpose of useful approximation we might have had in mind. We need a more refined apporoach. The trick turns out to be to shift $X_{\max}$ by a suitable amount depending on n, or precisely by $-\ln n$,

$$Y_n = X_{\max} - \ln n, \quad n = 1, 2, \ldots \tag{6.1}$$

Then for any $x \in \mathbf{R}$

$$F_{Y_n}(x) = \mathbf{P}(Y_n \leq x) = \mathbf{P}(X_{\max} \leq x + \ln n)$$

and by the computation above this equals

$$= \left(1 - e^{-(x+\ln n)}\right)^n = \left(1 - \frac{e^{-x}}{n}\right)^n.$$

Now we get, as $n \to \infty$,

$$F_{Y_n}(x) = \left(1 - \frac{e^{-x}}{n}\right)^n \to e^{-e^{-x}}.$$

Let us write

$$F_Y(x) = e^{-e^{-x}}, \quad -\infty < x < \infty. \tag{6.2}$$

This is the distribution function of a Gumbel distributed random variable Y, c.f. example 2.2.19. Hence it should be permissible to say that there is the convergence of Y_n to Y in the sense that $F_{Y_n}(x) \to F_Y(x)$.

■

Example 6.1.2 (Weak Law of Large Numbers) $X_1, X_2, \ldots$ are independent, identically distributed (I.I.D.) random variables with finite expectation μ and with variance σ^2. We set $S_n = X_1 + X_2 + \ldots + X_n, \quad n \geq 1$. We want to understand the properties of the arithmetic mean $\frac{1}{n}S_n$ for large values of n, which we again study by letting $n \to \infty$.

We need to recall that by I.I.D. $E\left[\frac{1}{n}S_n\right] = \mu$ and $\text{Var}\left[\frac{1}{n}S_n\right] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$. Then Chebyshev's inequality in (1.27) yields for any $\epsilon > 0$ that

$$\mathbf{P}\left(\mid \frac{S_n}{n} - \mu \mid > \epsilon\right) \leq \frac{1}{\epsilon^2}\text{Var}\left(\frac{S_n}{n}\right) = \frac{1}{\epsilon^2}\frac{\sigma^2}{n}.$$

Thus we have for any $\epsilon > 0$

$$\mathbf{P}\left(\mid \frac{S_n}{n} - \mu \mid > \epsilon\right) \to 0,$$

as $n \to \infty$. Again it should be correct to say that there is the convergence of $\frac{S_n}{n}$ to μ in the sense that the probability of an arbitrary small deviation of $\frac{S_n}{n}$ from μ goes to zero with increasing n.

> For example, we know by the (weak) law of large numbers that $\frac{1}{n}\sum_{k=1}^{n} X_k \overset{\text{P}}{\to} p$, as $n \to \infty$, if $X_1, \ldots X_n$ are independent and identically Be(p)-distributed. Therefore (5.51) tells that the probability of $\frac{1}{n}\sum_{k=1}^{n} X_k$ being larger than p goes to zero exponentially in n, and that the rate of convergence is determined by the Kullback distance $D(p|p+\epsilon)$.

■

Example 6.1.3 (Convergence of Second Order Moments) $X_1, X_2, \ldots$ is a sequence three point random variables such that

$$\mathbf{P}(X_n = -1) = \frac{1}{2n}, \mathbf{P}(X_n = 0) = 1 - \frac{1}{n}, \mathbf{P}(X_n = +1) = \frac{1}{2n}.$$

It is immediate that $E[X] = 0$ and that $E\left[X_n^2\right] = (-1)^2 \cdot \frac{1}{2n} + 0^2 \cdot \left(1 - \frac{1}{n}\right) + (+1)^2 \cdot \frac{1}{2n} = \frac{1}{n}$. Hence

$$E\left[X_n^2\right] \to 0,$$

as $n \to \infty$. Again we can regard this convergence of the second moments as a notion of probabilistic convergence of the sequence $X_1, X_2, \ldots$ to 0. To be quite accurate, we are saying that X_n converges to zero in the sense that

$$E\left[(X_n - 0)^2\right] \to 0,$$

as $n \to \infty$.

■

6.2 Definitions of Modes of Convergence, Uniqueness of the Limit

Now we launch the general formal definitions of the three modes of convergence suggested by the three examples in the section above in this chapter.

Definition 6.2.1 (Convergence in Distribution) A sequence of random variables $(X_n)_{n=0}^{+\infty}$ **converges in distribution** to the random variable X, if and only if it holds for the sequence of respective distribution functions that

$$F_{X_n}(x) \to F_X(x) \quad \text{as } n \to \infty$$

for all x that are points of continuity of $F_X(x)$.

■

We write convergence in distribution compactly as

$$X_n \xrightarrow{d} X, \qquad \text{as } n \to \infty.$$

Remark 6.2.1 We try next to justify the presence of points of continuity in the definition above. Let X_n be a random variable which induces, see (2.80), on the real line the total mass at the point $\frac{1}{n}$,

$$\mathbf{P}\left(X_n = \frac{1}{n}\right) = 1.$$

Then for any real x

$$F_{X_n}(x) = \mathbf{P}(X_n \leq x) = \begin{cases} 1 & x \geq \frac{1}{n} \\ 0 & x < \frac{1}{n}. \end{cases}$$

Then we consider the distribution function

$$F_X(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0. \end{cases}$$

Thus we see that, as $n \to \infty$

$$x \neq 0 : F_{X_n}(x) \to F_X(x),$$

but

$$x = 0 : F_{X_n}(0) = 0 \quad \text{does not converge to} \quad F_X(0) = 1.$$

But it is reasonable that the convergence $F_{X_n}(x) \to F_X(x)$ for $x \neq 0$ should matter, and therefore we require convergence of the sequence of distribution functions only for the points of continuity of the limit distribution function.

■

The notation $X_n \overset{d}{\to} X$ will be systematically distorted in the sequel, as we are going to write, e.g.,

$$X_n \overset{d}{\to} N(0,1), \quad X_n \overset{d}{\to} \text{Po}\,(\lambda)\,,$$

and so on, if $X \in N(0,1)$, $X \in \text{Po}\,(\lambda)$ and so on. In terms of the assumed licence to distort we have in the example 6.1.1 found that

$$X_{\max} - \ln n \overset{d}{\to} \text{Gumbel.}$$

A second mode of convergence is formulated by the next definition.

Definition 6.2.2 (Convergence in Probability) A sequence of random variables $(X_n)_{n=0}^{+\infty}$ **converges in probability** to the random variable X, if and only if it holds for all $\epsilon > 0$ that

$$\mathbf{P}\left(\mid X_n - X \mid > \epsilon\right) \to 0,$$

as $n \to \infty$.

■

We write this compactly as

$$X_n \overset{P}{\to} X, \quad \text{as } n \to \infty.$$

The limiting random variable may be a degenerate on, i.e., a constant. This is the case in example 6.1.2, where we found that

$$\frac{S_n}{n} \overset{P}{\to} \mu, \quad \text{as } n \to \infty.$$

Definition 6.2.3 (Convergence in r-mean) A sequence of random variables $(X_n)_{n=0}^{+\infty}$ **converges in r-mean** to the random variable X, if and only if it holds that

$$E\left[\mid X_n - X \mid^r\right] \to 0,$$

as $n \to \infty$. .

■

We have the compact expression

$$X_n \overset{r}{\to} X.$$

If $r > s$, then $X_n \overset{r}{\to} X$ implies $X_n \overset{s}{\to} X$. In the sequel we shall be exclusively concerned with the case $r = 2$. Here we talk about **convergence in mean square**

$$E\left|X_n - X\right|^2 \to 0, \quad \text{as } n \to \infty.$$

The chapter 7 below will be devoted to this convergence and its applications. Obviously, $X_n \overset{2}{\to} X$ is the case encountered in Example 6.1.3.

The limiting random variable in all of these modes of convergence is unique in distribution. This will now be proved in the case of convergence in probability.

Theorem 6.2.1 if $X_n \overset{P}{\to} X$, as $n \to \infty$ and $X_n \overset{P}{\to} Y$, as $n \to \infty$, then

$$X \overset{d}{=} Y.$$

Proof: We apply in this the inequality in (1.33). For given $\epsilon > 0$ we take $C = \{|X - Y| \leq \epsilon\}$, $A = \{|X_n - Y| \leq \epsilon/2\}$ and $B = \{|X_n - Y| \leq \epsilon/2\}$. We check first that the condition

$$A \cap B \subseteq C,$$

required for (1.33) is valid here. We note by the triangle inequality that

$$|X - Y| = |(X - X_n) + (X_n - Y)| \leq |X - X_n| + |X_n - Y)|$$

But $A \cap B$ is the event that both $A = \{|X_n - Y| \leq \epsilon/2\}$ and $B = \{|X_n - Y| \leq \epsilon/2\}$ hold. Hence if the event $A \cap B$ holds,

$$|X - X_n| + |X_n - Y)| \leq \epsilon/2 + \epsilon/2 = \epsilon,$$

i.e., if the event $A \cap B$ holds, then

$$|X - Y| \leq \epsilon.$$

Thus we have checked that $A \cap B \subseteq C$.

The assumptions $X_n \overset{P}{\to} X$, as $n \to \infty$ and that $X_n \overset{P}{\to} Y$, as $n \to \infty$ mean that

$$\mathbf{P}(A^c) = \mathbf{P}(\{|X_n - Y| > \epsilon/2\}) \to 0$$

and

$$\mathbf{P}(B^c) = \mathbf{P}(\{|X_n - Y| > \epsilon/2\}) \to 0$$

as $n \to \infty$. Hence the inequality (1.33) implies that

$$\mathbf{P}(C^c) = \mathbf{P}(\{|X - Y| > \epsilon\}) = 0$$

for any $\epsilon > 0$. Hence we have shown that

$$\mathbf{P}(X \neq Y) = 0,$$

as was desired. ■

6.3 Relations between Convergences

The modes of convergence formulated above are related to each other by an easily memorized catalogue of implications. The big picture is the following:

$$X_n \overset{r}{\to} X \Rightarrow X_n \overset{P}{\to} X$$

as $n \to \infty$.

$$X_n \overset{P}{\to} X \Rightarrow X_n \overset{d}{\to} X$$

as $n \to \infty$. If c is a constant,

$$X_n \overset{P}{\to} c \Leftrightarrow X_n \overset{d}{\to} c$$

as $n \to \infty$. The last implication can also be written as, c.f. (4.19),

$$X_n \overset{P}{\to} c \Leftrightarrow X_n \overset{d}{\to} \delta_c.$$

We shall now prove each of these statements.

Theorem 6.3.1

$$X_n \overset{r}{\to} X \Rightarrow X_n \overset{P}{\to} X \tag{6.3}$$

as $n \to \infty$.

Proof: We use Markov's inequality, (1.38). We check readily that this implies for a non negative random variable U $(U \geq 0)$ and $a > 0$, that for $r \geq 1$

$$\mathbf{P}\left(U \geq a\right) \leq \frac{1}{a^r} E\left[U^r\right].$$

Then we apply this to $U = \mid X_n - X \mid$ and get

$$\mathbf{P}\left(\mid X_n - X \mid \geq \epsilon\right) \leq \frac{1}{\epsilon^r} E\left[\mid X_n - X \mid^r\right].$$

Thus the desired conclusion follows. ■

Theorem 6.3.2

$$X_n \overset{P}{\to} X \Rightarrow X_n \overset{d}{\to} X \tag{6.4}$$

as $n \to \infty$.

Proof: Let us set $F_{X_n}(x) = \mathbf{P}\left(X_n \leq x\right)$. By some basic set operations we get

$$\mathbf{P}\left(X_n \leq x\right) = \mathbf{P}\left(\{X_n \leq x\} \cap \{\mid X_n - X \mid \leq \epsilon\}\right) + \mathbf{P}\left(\{X_n \leq x\} \cap \{\mid X_n - X \mid > \epsilon\}\right)$$

(a case of the obvious application of finite additivity: $\mathbf{P}\left(A\right) = \mathbf{P}\left(A \cap B\right) + \mathbf{P}\left(A \cap B^c\right)$). We observe that

$$\mid X_n - X \mid \leq \epsilon \Leftrightarrow -\epsilon \leq X_n - X \leq \epsilon$$

$$\Leftrightarrow -X_n - \epsilon \leq -X \leq -X_n + \epsilon$$

$$\Leftrightarrow X_n - \epsilon \leq X \leq X_n + \epsilon$$

Hence we can conclude that $X_n \leq x \Rightarrow X \leq x + \epsilon$, for the event $\{X_n \leq x\} \cap \{\mid X_n - X \mid \leq \epsilon\}$, which implies that $\{X_n \leq x\} \cap \{\mid X_n - X \mid \leq \epsilon\} \subseteq \{X \leq x + \epsilon\} \cap \{\mid X_n - X \mid \leq \epsilon\}$ and then

$$\mathbf{P}\left(\{X_n \leq x\} \cap \{\mid X_n - X \mid \leq \epsilon\}\right) \leq \mathbf{P}\left(\{X \leq x + \epsilon\} \cap \{\mid X_n - X \mid \leq \epsilon\}\right).$$

Thus we have obtained

$$\mathbf{P}\left(X_n \leq x\right) \leq \mathbf{P}\left(\{X \leq x + \epsilon\} \cap \{\mid X_n - X \mid \leq \epsilon\}\right) + \mathbf{P}\left(\{X_n \leq x\} \cap \{\mid X_n - X \mid > \epsilon\}\right),$$

$$\leq \mathbf{P}\left(\{X \leq x + \epsilon\}\right) + \mathbf{P}\left(\{\mid X_n - X \mid > \epsilon\}\right),$$

where we implemented twice the generic rule $\mathbf{P}\left(A \cap B\right) \leq \mathbf{P}\left(A\right)$. Hence we have obtained

$$F_{X_n}(x) \leq F_X(x + \epsilon) + \mathbf{P}\left(\{\mid X_n - X \mid > \epsilon\}\right). \tag{6.5}$$

If we change $X_n \mapsto X$, $x \mapsto x - \epsilon$, $X \mapsto X_n$, we can as above prove that

$$F_X(x - \epsilon) \leq F_{X_n}(x) + \mathbf{P}\left(\{\mid X_n - X \mid > \epsilon\}\right). \tag{6.6}$$

As $n \to \infty$, the two inequalities (6.5) and (6.6) and the assumption $X_n \overset{P}{\to} X$ entail (c.f., appendix 1.9) that

$$F_X(x - \epsilon) \leq \liminf_{n \to \infty} F_{X_n}(x) \leq \limsup_{n \to \infty} F_{X_n}(x) \leq F_X(x + \epsilon) \tag{6.7}$$

If we now let $\epsilon \downarrow 0$, we get by existence of left limits and right continuity (theorem 1.5.6) required of distribution functions

$$F_X(x-) \leq \liminf_{n\to\infty} F_{X_n}(x) \leq \limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x). \tag{6.8}$$

But the definition 6.2.1 requires us to consider any point x of continuity of $F_X(x)$. For such a point

$$F_X(x-) = F_X(x)$$

and we have obtained in (6.8)

$$F_X(x) \leq \liminf_{n\to\infty} F_{X_n}(x) \leq \limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x), \tag{6.9}$$

and therefore

$$\liminf_{n\to\infty} F_{X_n}(x) = \limsup_{n\to\infty} F_{X_n}(x) = F_X(x). \tag{6.10}$$

Thus the desired limit exists and

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x). \tag{6.11}$$

This is the assertion that was to be proved. ■

For the proof of the next theorem exploits the point mass distribution δ_c in (4.19).

Theorem 6.3.3 Let c be a constant. Then

$$X_n \overset{P}{\to} c \Leftrightarrow X_n \overset{d}{\to} c \tag{6.12}$$

as $n \to \infty$.

Proof:

- $\Rightarrow$: $X_n \overset{P}{\to} c \Rightarrow X_n \overset{d}{\to} c$, as $n \to \infty$, is true by (6.4).
- $\Leftarrow$: We assume in other words that $X_n \overset{d}{\to} \delta_c$, as $n \to \infty$. Let us take $\epsilon > 0$ and consider in view of definition 6.2.2

$$\begin{aligned} \mathbf{P}\left(\mid X_n - c \mid > \epsilon\right) &= 1 - \mathbf{P}\left(-\epsilon \leq X_n - c \leq \epsilon\right) \\ \overbrace{=}^{\text{by the rule (2.91)}} &\; 1 - \left(F_{X_n}(c+\epsilon) - F_{X_n}(c-\epsilon) + \mathbf{P}\left(X_n = c - \epsilon\right)\right) \\ &= 1 - F_{X_n}(c+\epsilon) + F_{X_n}(c-\epsilon) - \mathbf{P}\left(X_n = c - \epsilon\right) \\ &\leq 1 - F_{X_n}(c+\epsilon) + F_{X_n}(c-\epsilon), \end{aligned}$$

since $\mathbf{P}\left(X_n = c - \epsilon\right) \geq 0$. Now, by assumption

$$F_{X_n}(c+\epsilon) \to F_X(c+\epsilon) = 1,$$

since

$$F_X(x) = \begin{cases} 1 & x \geq c \\ 0 & x < c \end{cases}$$

and $c + \epsilon$ is a point of continuity of $F_X(x)$. By assumption

$$F_{X_n}(c-\epsilon) \to F_X(c-\epsilon) = 0,$$

where $c - \epsilon$ is a point of continuity of $F_X(x)$. Thus

$$1 - F_{X_n}(c+\epsilon) + F_{X_n}(c-\epsilon) \to 1 - 1 + 0 = 0,$$

as $n \to \infty$, and we have proved the assertion as claimed. ■

6.4 Some Rules of Computation

The following statements contain useful rules of computation, but the pertinent proofs, except the last one, are left to the interested reader.

Theorem 6.4.1 $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ are two sequences such that $X_n \overset{P}{\rightarrow} X$ and $Y_n \overset{P}{\rightarrow} Y$, as $n \rightarrow \infty$. Then

$$X_n + Y_n \overset{P}{\rightarrow} X + Y.$$

■

Theorem 6.4.2 $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ are two sequences such that $X_n \overset{r}{\rightarrow} X$ and $Y_n \overset{r}{\rightarrow} Y$, as $n \rightarrow \infty$ for some $r > 0$. Then

$$X_n + Y_n \overset{r}{\rightarrow} X + Y.$$

■

The following theorem has been accredited to two past researchers in probability[1].

Theorem 6.4.3 (Cramér -Slutzky Theorem) $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ are two sequences such that $X_n \overset{d}{\rightarrow} X$ and $Y_n \overset{P}{\rightarrow} a$, as $n \rightarrow \infty$, where a is a constant. Then, as $n \rightarrow \infty$,

(i)

$$X_n + Y_n \overset{d}{\rightarrow} X + a.$$

(ii)

$$X_n - Y_n \overset{d}{\rightarrow} X - a.$$

(iii)

$$X_n \cdot Y_n \overset{d}{\rightarrow} X \cdot a.$$

(iv)

$$\frac{X_n}{Y_n} \overset{d}{\rightarrow} \frac{X}{a} \quad \text{for } a \neq 0.$$

■

The proof of the next assertion is an instructive exercise in probability calculus and the definition of continuity of a function.

Theorem 6.4.4 $(X_n)_{n\geq 1}$ is a sequence such that $X_n \overset{P}{\rightarrow} a$, as $n \rightarrow \infty$, where a is a constant. Suppose that $h(x)$ is a function that is continuous at a. Then

$$h\left(X_n\right) \overset{P}{\rightarrow} h(a), \tag{6.13}$$

as $n \rightarrow \infty$.

Proof: Take an arbitrary $\epsilon > 0$. We are to show that

$$\mathbf{P}\left(\mid h\left(X_n\right) - h(a) \mid > \epsilon\right) \rightarrow 0,$$

[1] Harald Cramér (1893-1985) was a mathematician and actuary. He was professor of mathematical statistics at the University of Stockholm. Evgeny Evgenievich Slutzky (1880-1948) was a Russian mathematical statistician, economist and political economist.

as $n \to \infty$. We shall, as several times above, find an upper bound that converges to zero, if $X_n \overset{P}{\to} a$ is assumed. We write on this occasion the expression in the complete form

$$\mathbf{P}\left(\mid h\left(X_n\right) - h(a) \mid > \epsilon\right) = \mathbf{P}\left(\{\omega \in \Omega \mid \mid h\left(X_n(\omega)\right) - h(a) \mid > \epsilon\}\right).$$

Since $h(x)$ is continuous at a we have that for all $\epsilon > 0$ there exists a $\delta = \delta(\epsilon) > 0$ such that

$$\mid x - a \mid \leq \delta \Rightarrow \mid h(x) - h(a) \mid \leq \epsilon.$$

If we take the logical negation of this implication we get that

$$\mid h(x) - h(a) \mid > \epsilon \Rightarrow \mid x - a \mid > \delta.$$

For the corresponding events this gives the inclusion

$$\{\omega \in \Omega \mid \mid h\left(X_n(\omega)\right) - h(a) \mid > \epsilon\} \subseteq \{\omega \in \Omega \mid \mid X_n(\omega) - a \mid > \delta\}.$$

Thus we get

$$\mathbf{P}\left(\{\omega \in \Omega \mid \mid h\left(X_n(\omega)\right) - h(a) \mid > \epsilon\}\right) \leq \mathbf{P}\left(\{\omega \in \Omega \mid \mid X_n(\omega) - a \mid > \delta\}\right).$$

But by assumption we have that

$$\mathbf{P}\left(\{\omega \in \Omega \mid \mid X_n(\omega) - a \mid > \delta\}\right) \to 0,$$

as $n \to \infty$, which proves the claim as asserted. ■

The next example shows how these results are applied in statistics.

Example 6.4.5 Let $(X_n)_{n \geq 1}$ be a sequence of independent r.v.,s $X_n \in \text{Be}(p)$, $0 < p < 1$. Let $S_n = X_1 + X_2 + \ldots + X_n$. We wish to study the distribution of

$$Q = \frac{\frac{1}{n} S_n - p}{\sqrt{\frac{\frac{1}{n} S_n \left(1 - \frac{1}{n} S_n\right)}{n}}},$$

as $n \to \infty$. The r.v. Q is used for building confidence intervals for p, where p is carries a statement about an unknown population proportion. In order to handle this expression we note the following. We can write

$$\frac{1}{n} S_n - p = \frac{1}{n} \sum_{k=1}^{n} \left(X_k - p\right).$$

Thus

$$Q = \frac{\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \left(X_k - p\right)}{\sqrt{\frac{1}{n} S_n \left(1 - \frac{1}{n} S_n\right)}} = \frac{\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{(X_k - p)}{\sqrt{p(1-p)}}}{\sqrt{\frac{\frac{1}{n} S_n \left(1 - \frac{1}{n} S_n\right)}{p(1-p)}}}.$$

The rationale for the introducing this identity will become soon clear. We define for $x \in [0, 1]$ the continuous function

$$h(x) = \sqrt{\frac{x\left(1 - x\right)}{p\left(1 - p\right)}}.$$

Then

$$Q = \frac{\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{(X_k - p)}{\sqrt{p(1-p)}}}{h\left(\frac{1}{n} S_n\right)}. \tag{6.14}$$

By properties of Bernoulli variables

$$E\left[X_i\right] = p, \text{Var}\left[X_i\right] = p\left(1 - p\right).$$

Hence we observe in the numerator of Q that $\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{(X_k-p)}{\sqrt{p(1-p)}}$ is a scaled sum of exactly same form as the scaled sum in section 4.5.2 above (replace $\mu \mapsto p$, $\sigma \mapsto \sqrt{p\,(1-p)}$ in $\frac{X_k-\mu}{\sigma}$). The provisional argument in loc.cit. entails that

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{(X_k-p)}{\sqrt{p\,(1-p)}} \xrightarrow{d} N(0,1),$$

as $n \to \infty$. In the denominator of (6.14) we observe that the weak law of large numbers, example 6.1.2 above, implies

$$\frac{1}{n}S_n \xrightarrow{P} p,$$

as $n \to \infty$. Then we get by (ii) in Cramér -Slutzky theorem that $\left(1-\frac{1}{n}S_n\right) \xrightarrow{P} (1-p)$. Thus (6.13) in the theorem above implies, as $0 < p < 1$, that

$$h\left(\frac{1}{n}S_n\right) \xrightarrow{P} h(p) = 1.$$

But then case (iv) in the Cramér -Slutzky Theorem entails that

$$Q = \frac{\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{(X_k-p)}{\sqrt{p(1-p)}}}{h\left(\frac{1}{n}S_n\right)} \xrightarrow{d} N(0,1),$$

as $n \to \infty$, which resolves the question posed. In the section 6.6.3 we ascertain that the central limit theorem suggested in section 4.5.2 by means of characteristic functions is valid.

6.5 Asymptotic Moments and Propagation of Error

As an application of the preceding rules for computing limits we shall next consider in terms of asymptotic moments what is known as propagation of error. Propagation of error is formally stated concerned with expressing the mean and variance of a (suitably smooth) function $Y = g(X)$ of a random variable X in terms of $\mu = E[X]$ and $\sigma^2 = \text{Var}[X]$. Two well known formulas [15, pp.273−274] or [51, section 9.9] in propagation of error are

$$E[g(X)] \approx g(\mu), \text{Var}[g(X)] \approx \sigma^2 \frac{d}{dx}g(x)|_{x=\mu}. \tag{6.15}$$

It can be a difficult task to judge or justify in a general way when these formulas should be applicable. Calculations of the propagation of error based on the formulas above have for many decades provided practically very accurate approximations, e.g., in instrument technology [52, 83][2]. We shall clarify the approximations above by the most common justification of approximations in probability calculus, i.e., by means of convergence in distribution of a sequence of random variables.

Remark 6.5.1 If we approximate the expectations as in (6.15) with $g(x) = 1/x$ we get

$$E\left[\frac{1}{X}\right] \approx \frac{1}{E[X]}, \tag{6.16}$$

as a practically minded rule for computing $E\left[\frac{1}{X}\right]$. But if $X \in C(0,1)$, then $\frac{1}{X} \in C(0,1)$. For $C(0,1)$, the expectation does not exist, as shown in example 2.2.16. Hence, an approximation like (6.16) makes no sense in this situation.

[2] see also H.H. Ku: Notes on the Use of Propagation of Error Formulas. *Journal of Research of the National Bureau of Standards - C. Engineering and Instrumentation.* Vol 70C, no 4, October - December, 1966

■

Suppose that $\{X_n\}_{n\geq 1}$ is a sequence of random variables such that

$$\sqrt{n}\left(X_n-\mu\right)\xrightarrow{d} N\left(0,\sigma^2\right), \tag{6.17}$$

as $n\to\infty$. We say that μ and σ^2/n are the **asymptotic mean** and **asymptotic variance**, respectively, of the sequence $\{X_n\}_{n\geq 1}$. The obvious example is $X_n=\frac{1}{n}\sum_{i=1}^{n}Z_i$ of Z_i, I.I.D. variables with $\mu=E\left[Z_i\right]$ and $\sigma^2=\text{Var}[Z_i]$.

Note that we do not suppose that $\mu_n=E\left[X_n\right]$, $\sigma_n^2=\text{Var}[X_n]$ and that $\mu_n\to\mu$ and $\sigma_n^2\to\sigma^2$. In fact $E\left[X_n\right]$, and $\text{Var}[X_n]$ are not even required to exist.

Theorem 6.5.1 (Propagation of Error) Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables such that (6.17) holds. Let $g(x)$ be a differentiable function with the first derivative $g^{'}(x)$ which is continuous and that $g^{'}(\mu)\neq 0$. Then it holds that

$$\sqrt{n}\left(g(X_n)-g(\mu)\right)\xrightarrow{d} N\left(0,\sigma^2\left(g^{'}(\mu)\right)^2\right), \tag{6.18}$$

as $n\to\infty$.

Proof: By the mean value theorem of calculus [69, p.100] there exists for every x and μ a number ξ between x and μ such that

$$g(x)-g(\mu)=g^{'}(\xi)(x-\mu). \tag{6.19}$$

Thus there is a well defined function of ω, Z_n, such that $\mid Z_n-\mu\mid\leq\mid X_n-\mu\mid$ and by (6.19)

$$g\left(X_n\right)-g(\mu)=g^{'}\left(Z_n\right)\left(X_n-\mu\right). \tag{6.20}$$

In fact, Z_n is a random variable, a property we must require, but this will be proved after the theorem.
By (i) and (iii) of the Cramér -Slutzky Theorem 6.4.3

$$\left(X_n-\mu\right)=\frac{1}{\sqrt{n}}\left[\sqrt{n}\left(X_n-\mu\right)\right]\xrightarrow{d} 0\cdot N\left(0,\sigma^2\right),$$

as $n\to\infty$. Hence $X_n-\mu\xrightarrow{P} 0$ in view of theorem 6.3.3 and (6.12). Since $\mid Z_n-\mu\mid\leq\mid X_n-\mu\mid$, we get that $\mid Z_n-\mu\mid\xrightarrow{P} 0$ (it is here we need the fact that Z_n is a sequence of random variables), as $n\to\infty$. But then (6.13) in theorem 6.4.4 implies, as $n\to\infty$, that

$$g^{'}\left(Z_n\right)\xrightarrow{P} g^{'}(\mu), \tag{6.21}$$

by the assumed continuity of the derivative $g^{'}(x)$. Now we have in (6.20)

$$\begin{aligned}\sqrt{n}\left(g(X_n)-g(\mu)\right)&=\sqrt{n}g^{'}\left(Z_n\right)\left(X_n-\mu\right)\\&=g^{'}\left(Z_n\right)\sqrt{n}\left(X_n-\mu\right).\end{aligned}$$

By (6.17) and (iii) of the Cramér -Slutzky Theorem 6.4.3 and by (6.21)

$$\sqrt{n}\left(g(X_n)-g(\mu)\right)\xrightarrow{d} g^{'}(\mu)X,$$

where $X\in N(0,\sigma^2)$. Hence we have established (6.18) as claimed. ■

It is, of course, still a matter of judgement to decide whether the approximation in the theorem above can be used in any given situation.

The following indented section of text verifies that Z_n defined in (6.20) is a random variable and can be skipped at the first reading.

We have earlier maintained that unmeasurable sets lack importance in practice, as they are very hard to construct. The claim that Z_n in (6.20) is a random variable seems innocuous for such a setting of mind. However, we can ask, if there is a proof the claim. The mean value theorem of calculus gives Z_n as a well defined function of ω by (6.20). According to the definition 1.5.1 we need in addition to show that Z_n is a measurable map from $\mathcal{F}$ to the Borel σ algebra. To that end we fix an arbitrary n and drop it temporarily from the notation. Let us define for each $\omega \in \Omega$,

$$H(X(\omega)) \stackrel{\text{def}}{=} \begin{cases} g^{'}(\mu) & \text{if } X(\omega) = \mu \\ \frac{g(X(\omega))-g(\mu)}{X(\omega)-\mu} & \text{if } X(\omega) \neq \mu. \end{cases}$$

$H(X(\omega))$ is a random variable, as $g(X(\omega))$ is a random variable and a ratio of two random variables is a random variable. We set for each $\omega \in \Omega$

$$G(z) \stackrel{\text{def}}{=} H(X(\omega)) - g^{'}(z).$$

We should actually be writing $G_\omega(z)$, as there is a different function $G(z)$ for each ω, but we abstain from this for simplicity. Then $G(z)$ is a random variable and is continuous as a function of z and (6.20) corresponds to finding for fixed ω a root (at least one exists by the mean value theorem of calculus) $Z(\omega)$ to the equation

$$G(z) = 0. \tag{6.22}$$

(i) We assume first that $G(X(\omega))G(\mu) < 0$. Then we can apply the **method of bisection** to construct a root to (6.22). We assume first that $X(\omega) < \mu$. Then we set for $k = 1, 2, \ldots,$

$$\begin{aligned} a_0(\omega) &= X(\omega), & b_0 &= \mu \\ a_k(\omega) &= a_{k-1}(\omega) & b_k(\omega) &= m_{k-1}(\omega), \text{ if } G(a_{k-1}(\omega))G(m_{k-1}(\omega)) < 0 \\ a_k(\omega) &= m_{k-1}(\omega) & b_k(\omega) &= b_{k-1}(\omega), \text{ if } G(a_{k-1}(\omega))G(m_{k-1}(\omega)) > 0. \end{aligned} \tag{6.23}$$

Here

$$m_k(\omega) = \frac{a_{k-1}(\omega) + b_{k-1}(\omega)}{2},$$

which explains the name bisection (draw a picture) given to this iterative method of solving equations. By the construction it holds that

$$G(a_k(\omega))G(b_k(\omega)) < 0 \tag{6.24}$$

Since X is a random variable, m_1 is a random variable, and since $G(z)$ is a random variable, too, both a_1 and b_1 are by (6.23) random variables. Therefore, by the steps of construction of the bisection, each m_k is a random variable. It holds that $a_{k-1}(\omega) \leq a_k(\omega)$ and $b_k(\omega) \leq b_{k-1}(\omega)$, $a_k(\omega) < b_k(\omega)$ and $a_k(\omega) - b_k(\omega) = \frac{1}{2^k}(\mu - X(\omega))$. Thus we have the limit, which we denote by $Z(\omega)$,

$$Z(\omega) = \lim_{k\to\infty} m_k(\omega) = \lim_{k\to\infty} a_k(\omega) = \lim_{k\to\infty} b_k(\omega).$$

Since $Z(\omega)$ is a pointwise limit of random variables, it is a random variable (this can be verified by writing the statement of convergence by means of unions and intersections of events).

Then it follows by continuity of $G(z)$ and (6.24) that

$$G^2(Z(\omega)) = \lim_{k\to\infty} G(a_k(\omega))G(b_k(\omega)) \leq 0$$

or that $G^2(Z(\omega)) = 0$, i.e., $G(Z(\omega)) = 0$ so that $Z(\omega)$ is a root of (6.22) between $X(\omega)$ and μ, and Z is random variable, as was claimed.

If we assume that $X(\omega) > 0$ we get the same result by trivial modifications of the proof above.

(ii) The case $G(X(\omega))G(\mu) \geq 0$ contains two special cases. First, there is a unique root to $G(z) = 0$, which we can find by a hill-climbing technique of as limit of measurable approximations. Or, we can move over to a subdomain with a root, where the bisection technique of case (i) again applies.

The method of bisection is a simple (and computationally ineffective) algorithm of root solving, but in fact it can be evoked analogously in a constructive proof the theorem of intermediate values [69, pp.71-73] of differential calculus.

6.6 Convergence by Transforms

6.6.1 Theorems on Convergence by Characteristic Functions

Let us start again by some examples of what we shall be studying in this section.

Example 6.6.1 As in example 6.1.3 we have the sequence $(X_n)_{n=1}^{+\infty}$ of three point random variables

$$\mathbf{P}(X_n = -1) = \frac{1}{2n}, \mathbf{P}(X_n = 0) = 1 - \frac{1}{n}, \mathbf{P}(X_n = +1) = \frac{1}{2n}.$$

We have found, loc.cit., that

$$X_n^2 \overset{2}{\to} 0,$$

as $n \to \infty$. We know by (6.3) and (6.12) that

$$X_n^2 \overset{2}{\to} 0 \Rightarrow X_n \overset{P}{\to} 0 \Leftrightarrow X_n \overset{d}{\to} 0.$$

If we compute the characteristic function of X_n we get

$$\varphi_{X_n}(t) = \frac{1}{2n}e^{-it} + \left(1 - \frac{1}{n}\right)e^{-i0} + \frac{1}{2n}e^{it}$$

$$= \left(1 - \frac{1}{n}\right) + \frac{1}{2n}\left(e^{it} + e^{-it}\right)$$

and by Euler's formula for $\cos t$

$$= \left(1 - \frac{1}{n}\right) + \frac{1}{2n}(2\cos t).$$

As $n \to \infty$, we see that

$$\varphi_{X_n}(t) \to 1 = e^{i0}.$$

We have in the preceding introduced the distribution δ_c, c.f., (4.19) above. The characteristic function of δ_c is by (4.20)

$$\varphi_{\delta_c}(t) = 1 \cdot e^{itc}.$$

Hence we have obtained that

$$\varphi_{X_n}(t) \to \varphi_{\delta_0}(t),$$

as $n \to \infty$. Clearly this corresponds to the fact that $X_n \overset{d}{\to} 0$.

■

Example 6.6.2 $(X_n)_{n=1}^{+\infty}$ is a sequence of random variables such that $X_n \in \text{Bin}\left(n, \frac{\lambda}{n}\right)$ for $n = 1, 2, \ldots,, \lambda > 0$. We have by (4.23) that

$$\varphi_{X_n}(t) = \left(\left(1 - \frac{\lambda}{n}\right) + e^{it}\frac{\lambda}{n}\right)^n,$$

which we rewrite as

$$= \left(1 + \frac{\lambda}{n}\left(e^{it} - 1\right)\right)^n,$$

and then by a standard limit as $n \to \infty$,

$$\to e^{\lambda\left(e^{it}-1\right)} = \varphi_{\text{Po}}(\lambda),$$

where we recognized the result (4.9). In words, we should be allowed to draw the conclusion that

$$X_n \xrightarrow{d} \text{Po}(\lambda).$$

This result tells rigorously that we can approximate $X \in \text{Bin}\left(n, \frac{\lambda}{n}\right)$ for small p and large n by $\text{Po}(np)$.

■

In fact these two examples present two respective examples of the workings of the following fundamental theorem.

Theorem 6.6.3 (Continuity Theorem for Characteristic Functions) (a) If $X_n \xrightarrow{d} X$, and X is a random variable with the characteristic function $\varphi_X(t)$, then

$$\varphi_{X_n}(t) \to \varphi_X(t), \quad \text{for all } t,$$

as $n \to \infty$.

(b) If $\{\varphi_{X_n}(t)\}_{n=1}^{\infty}$ is a sequence of characteristic functions of random variables X_n, and

$$\varphi_{X_n}(t) \to \varphi(t), \quad \text{for all } t,$$

and $\varphi(t)$ is continuous at $t = 0$, then $\varphi(t)$ is the characteristic function of some random variable X ($\varphi(t) = \varphi_X(t)$) and

$$X_n \xrightarrow{d} X.$$

■

The proof is omitted. We saw an instance of case(a) in example 6.6.1. In addition, we applied correctly the case (b) in example 6.6.2, since $e^{\lambda\left(e^{it}-1\right)}$ is continuous at $t = 0$.

With regard to the 'converse statement' in (b) it should be kept in mind that one can construct sequences of characteristic functions that converge to a function that is not a characteristic function.

By means of characteristic functions we can easily prove (proof omitted) the uniqueness theorem for convergence in distribution.

Theorem 6.6.4 (Uniqueness of convergence in distribution) If $X_n \xrightarrow{d} X$, and $X_n \xrightarrow{d} Y$, then

$$X \overset{d}{=} Y.$$

■

6.6.2 Convergence and Generating Functions

We note the following facts.

Theorem 6.6.5 If $\{X_n\}_{n\geq 1}$ is a sequence of random variables with values in the non negative integers and p.g.f.'s $g_{X_n}(t)$. If

$$g_{X_n}(t) \to g_X(t),$$

as $n \to \infty$, then $X_n \overset{d}{\to} X$, as $n \to \infty$.

■

Theorem 6.6.6 $\{X_n\}_{n\geq 1}$ is a sequence of random variables such that the m.g.f.'s $\psi_{X_n}(t)$ exist for $|t| < h$ for some $h > 0$. Suppose that X is a random variable such that its m.g.f. $\psi_X(t)$ exists for $|t| < h_1 \leq h$ for some $h_1 > 0$ and that

$$\psi_{X_n}(t) \to \psi_X(t),$$

as $n \to \infty$, then $X_n \overset{d}{\to} X$, as $n \to \infty$.

■

6.6.3 Central Limit Theorem

We can now return to the sum scaled by $\frac{1}{\sqrt{n}}$ in the section 4.5.2 above and formulate and prove the finding there as a theorem.

Theorem 6.6.7 (Central Limit Theorem) X_1, X_2, $\ldots, X_n$ $\ldots$ is an infinite sequence of independent and identically distributed random variables with $E[X_k] = \mu$ and $\text{Var}[X_k] = \sigma^2$ for $k = 1, 2, \ldots,$. Define

$$W_n \overset{\text{def}}{=} \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{X_k - \mu}{\sigma}.$$

Then

$$W_n \overset{d}{\to} N(0,1), \quad \text{as } n \to \infty. \tag{6.25}$$

Proof: In section 4.5.2 we have shown that for all t

$$\lim_{n\to\infty} \varphi_{W_n}(t) = e^{-t^2/2}. \tag{6.26}$$

Since the function $e^{-t^2/2}$ is the characteristic function of $N(0,1)$ and is continuous at $t = 0$, it follows in view of case (b) of theorem 6.6.3 and by uniqueness of characteristic functions that $W_n \overset{d}{\to} N(0,1)$, as $n \to \infty$. ■

The **Berry-Esseen**[3] **theorem** (1941, 1942, respectively) gives us the speed of convergence in the central limit theorem :

$$|F_{W_n}(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}},$$

where $\rho = E\left[|X|^3\right]$. Since the 1940's there has been an intensive activity for finding the best value of the constant C. By the year 2011 the best estimate is known to be $C < 0.4784$.

There are several more complex versions, extensions and generalizations of the central limit theorem, e.g., to martingales.

[3] Carl Gustav Esseen, (1918-2001), appointed in 1949 to professor of applied mathematics at KTH, the Royal Institute of Technology. In 1962 his professorship was transferred to mathematical statistics and in 1967, he obtained the first chair in mathematical statistics at Uppsala University.

6.7 Almost Sure Convergence

6.7.1 Definition

The final mode of convergence of sequences of random variables to be introduced is almost sure convergence.

Definition 6.7.1 (Almost Sure Convergence) A sequence of random variables $(X_n)_{n=1}^{+\infty}$ **converges almost surely** or **with probability one** to the random variable X, (X exists and has values in $\mathbf{R}$) if and only if it holds that

$$\mathbf{P}\left(\{\omega \in \Omega | X_n(\omega) \to X(\omega) \text{ as } n \to \infty \ \}\right) = 1.$$

■

We express this more compactly as

$$X_n \overset{a.s.}{\to} X.$$

Let us set

$$C = \{\omega \in \Omega | X_n(\omega) \to X(\omega) \text{ as } n \to \infty \ \}.$$

This means, in the language of real analysis [36], that the sequence of measurable functions X_n converges 'pointwise' to the limiting measurable function X on a set of points (=elementary events, ω), which has probability one. We are thus stating that $\mathbf{P}(C) = 1$ if and only if $X_n \overset{a.s.}{\to} X$. We shall next try to write the set C more transparently.

Convergence of a sequence of numbers $(x_n)_{n\geq 1}$ to a real number x means by definition that for all $\epsilon > 0$ there exists an $n(\epsilon)$ such that for all $n > n(\epsilon)$ it holds that $|x_n - x| < \epsilon$. By this understanding we can write C in countable terms, i.e. we replace the arbitrary ϵ's with $1/k$'s, as

$$C = \cap_{k=1}^{\infty} \cup_{m=1}^{\infty} \cap_{n\geq m} \left\{\omega \in \Omega | \mid X_n(\omega) - X(\omega) \mid \leq \frac{1}{k}\right\}. \tag{6.27}$$

By properties of σ-fields it holds that $C \in \mathcal{F}$, and thus $\mathbf{P}(C)$ is well defined.

6.7.2 Almost Sure Convergence Implies Convergence in Probability

Next we augment (6.3) and (6.4) by one more implication.

Theorem 6.7.1

$$X_n \overset{a.s.}{\to} X \Rightarrow X_n \overset{P}{\to} X \tag{6.28}$$

as $n \to \infty$.

Proof: Let us look at the complement C^c or, by De Morgan's rules, from (6.27)

$$C^c = \cup_{k=1}^{\infty} \cap_{m=1}^{\infty} \cup_{n\geq m} \left\{\omega \in \Omega | \mid X_n(\omega) - X(\omega) \mid > \frac{1}{k}\right\}. \tag{6.29}$$

Let us set (revert from arbitrary $1/k$ to arbitrary $\epsilon > 0$)

$$A_n(\epsilon) = \{\omega \in \Omega | \mid X_n(\omega) - X(\omega) \mid > \epsilon\}$$

and

$$B_m(\epsilon) = \cup_{n\geq m} A_n(\epsilon). \tag{6.30}$$

Then we set

$$A(\epsilon) = \cap_{m=1}^{\infty} B_m(\epsilon) = \cap_{m=1}^{\infty} \cup_{n\geq m} A_n(\epsilon).$$

Then clearly

$$C^c = \cup_{k=1}^{\infty} A\left(\frac{1}{k}\right).$$

In view of (1.14) and the discussion around it

$$A(\epsilon) = \{\omega \in \Omega | A_n(\epsilon) \text{ infinitely often } \}.$$

Of course, $X_n(\omega) \to X(\omega)$ if and only if $\omega \notin A(\epsilon)$. Hence, if $\mathbf{P}(C) = 1$, then $\mathbf{P}(A(\epsilon)) = 0$ for all $\epsilon > 0$.

We have as in section 1.7 that $B_m(\epsilon)$ is a decreasing sequence of events with limit $A(\epsilon)$. Therefore, by continuity of probability measures from above, i.e., theorem 1.4.9, it follows that if $\mathbf{P}(A(\epsilon)) = 0$, then $\mathbf{P}(B_m(\epsilon)) \to 0$. But by construction in (6.30), $A_n(\epsilon) \subseteq B_m(\epsilon)$. Hence

$$\mathbf{P}(B_m(\epsilon)) \geq \mathbf{P}(A_n(\epsilon)) = \mathbf{P}(\{\omega \in \Omega | \mid X_n(\omega) - X(\omega) \mid > \epsilon\}).$$

Hence $\mathbf{P}(B_m(\epsilon)) \to 0$ implies

$$\mathbf{P}(\{\omega \in \Omega | \mid X_n(\omega) - X(\omega) \mid > \epsilon\}) \to 0,$$

as $n \to \infty$, which we have as $m \to \infty$. ■

6.7.3 A Summary of the General Implications between Convergence Concepts and One Special Implication

We have thus shown that as $n \to \infty$,

$$X_n \overset{a.s.}{\to} X \Rightarrow X_n \overset{P}{\to} X$$

$$X_n \overset{r}{\to} X \Rightarrow X_n \overset{P}{\to} X$$

$$X_n \overset{P}{\to} X \Rightarrow X_n \overset{d}{\to} X$$

If c is a constant,

$$X_n \overset{P}{\to} c \Leftrightarrow X_n \overset{d}{\to} \delta_c.$$

There are no further implications that hold in general. It is shown in the exercises in section 6.8.5 that almost sure convergence does not imply convergence in mean square, and vice versa. It can be shown by examples that $X_n \overset{P}{\to} X$ does not imply $X_n \overset{a.s.}{\to} X$.

Additional implications between convergence concepts can be established under special assumptions. The following result shows in this regard that if we have a sequence of r.v.'s that are almost surely bounded by a constant, and the sequence converges in probability to the r.v. X, then the sequence converges in mean square to X, too.

Theorem 6.7.2 $(X_n)_{n\geq 1}$ is a sequence of r.v.'s such that i) and ii) below are satisfied:

i) There is a positive real number L such that $\mathbf{P}(\mid X_n \mid \leq L) = 1$ for every n.

ii) $X_n \overset{P}{\to} X$, as $n \to +\infty$.

Then

$$X_n \overset{2}{\to} X \tag{6.31}$$

as $n \to \infty$.

■

The steps of the required proof are the exercise of subsection 6.8.4 below.

6.7.4 The Strong Law of Large Numbers

The statement

We let $X_1, X_2, \ldots$ be I.I.D. with $E[X_i] = m$ and $\text{Var}[X_i] = \sigma^2 < \infty$. We define $S_n = X_1 + X_2 + \cdots + X_n$. We are interested in showing the **strong form of the law of large numbers (SLLN)**, i.e., a law of large numbers such that $S_n/n \to m$ as $n \to \infty$ with probability one or almost surely. This means that we want to prove that

$$P\left(\lim_{n\to\infty} \frac{S_n}{n} = m\right) = 1,$$

i.e., that there exists a set C with $P(C) = 1$, where

$$C = \left\{\omega | \lim_{n\to\infty} |\frac{S_n(\omega)}{n} - m| = 0\right\}.$$

We need in other words to prove that for every $\omega \in C$ and for every $\varepsilon > 0$ there is $N(\omega, \varepsilon)$ so that if $n \geq N(\omega, \varepsilon)$ holds that $|S_n/n - m| \leq \varepsilon$.

It suffices to prove that $|\frac{S_n}{n} - m| > \varepsilon$ can occur only a finite number of times, i.e.,

$$\lim_{N\to\infty} P\left(|\frac{S_n}{n} - m| > \varepsilon \quad \text{some } n \geq N\right) = 0.$$

Note the distinction with regard to the law of large numbers in the weak form, which says that that for all $\varepsilon > 0$

$$P\left(|\frac{S_n}{n} - m| > \varepsilon\right) \to 0 \text{ as } n \to \infty.$$

In words: for the law of large numbers in the strong form $|S_n/n - m|$ must be small for all sufficiently large n for all $\omega \in C$, where $P(C) = 1$.

> In tossing a coin we can code heads and tails with 1 and 0, respectively, and we can identify an ω with a number in the interval $[0, 1]$ drawn at random, where binary expansion gives the sequence of zeros and ones. The law of large numbers says in this case that we will obtain with probability 1 a number such that the proportion of 1:s in sequence converges towards $1/2$. There can be "exceptional" -ω - for example the sequence $000\ldots$ is possible, but such exceptional sequences have the probability 0.

After these deliberations of pedagogic nature let us get on with the proof[4].

The Proof of SLLN

Without restriction of generality we can assume that $E(X_i) = m = 0$, since we in any case can consider $X_i - m$. We have $\text{Var}[S_n] = n\sigma^2$. By Chebyshev's inequality (1.27) it holds that

$$P(|S_n| > n\varepsilon)) \leq \frac{\text{Var}[S_n]}{(n\varepsilon)^2} = \frac{n\sigma^2}{(n\varepsilon)^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Unfortunately the harmonic series $\sum_1^\infty 1/n$ is divergent so we cannot use Borel-Cantelli lemma 1.7.1 directly. But it holds that $\sum_1^\infty 1/n^2 < \infty$ and this means that we can use the lemma for n^2, $n = 1, 2, \ldots$. We have

$$P(|S_{n^2}| > n^2\varepsilon) \leq \frac{\sigma^2}{n^2\varepsilon^2}.$$

[4] Gunnar Englund is thanked for pointing out this argument.

In other words it holds by Borel-Cantelli lemma 1.7.1, that $P(|\frac{S_{n^2}}{n^2}| > \varepsilon \text{ i.o.}) = 0$ which proves that $S_{n^2}/n^2 \to 0$ almost surely. We have in other words managed to establish that for the subsequence n^2, $n = 1, 2, \ldots$ there is convergence with probability 1. It remains to find out what will happen between these n^2. We define therefore

$$D_n = \max_{n^2 \leq k < (n+1)^2} |S_k - S_{n^2}|,$$

i.e., the largest of the deviation from S_{n^2} that can occur between n^2 and $(n+1)^2$. We get

$$D_n^2 = \max_{n^2 \leq k < (n+1)^2} (S_k - S_{n^2})^2 \leq \sum_{k=n^2}^{(n+1)^2-1} (S_k - S_{n^2})^2,$$

where we used the rather crude inequality $\max(|x|, |y|) \leq (|x| + |y|)$. This entails

$$E\left[D_n^2\right] \leq \sum_{k=n^2}^{(n+1)^2-1} E\left[(S_k - S_{n^2})^2)\right].$$

But $E\left[(S_k - S_{n^2})^2\right] = (k - n^2)\sigma^2 \leq 2n\sigma^2$ as $n^2 \leq k < (n+1)^2$ and there are $2n$ terms in the sum and this entails

$$E\left[D_n^2\right] \leq (2n)(2n)\sigma^2 = 4n^2\sigma^2.$$

With Chebyshev's inequality (1.27) this gives

$$P\left(D_n > n^2\varepsilon\right) \leq \frac{4n^2\sigma^2}{(n^2\varepsilon)^2} = \frac{4\sigma^2}{n^2\varepsilon^2}.$$

In other words, $D_n/n^2 \to 0$ holds almost surely. Finally this yields for k between n^2 and $(n+1)^2$ that

$$|\frac{S_k}{k}| \leq \frac{|S_{n^2}| + D_n}{k} \leq \frac{|S_{n^2}| + D_n}{n^2} \to 0.$$

This means that we have succeeded in proving that $S_n/n \to 0$ with probability 1. We have done this under the condition that $\text{Var}(X_i) = \sigma^2 < \infty$, but with a painstaking effort we can in fact prove that this condition is not necessary. ■

6.8 Exercises

6.8.1 Convergence in Distribution

1. (5B1540 2003-08-27) The random variables $X_1, X_2, \ldots$ be I.I.D. with the p.d.f. $f_X(x) = \frac{1-\cos x}{\pi x^2}$.

   (a) Check that $f_X(x) = \frac{1-\cos x}{\pi x^2}$ is a probability density. *Aid:* First, note $1 - \cos x = 2\left(\sin \frac{x}{2}\right)^2$. Then recall (4.45), and the inverse transform (4.2), i.e.,

   $$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itx} \widehat{f}(t) dt.$$

   (b) Show that $\frac{1}{n}(X_1 + X_2 + \ldots + X_n) \xrightarrow{d} C(0, 1)$, as $n \to \infty$. *Aid:* Use (4.2) to find $\varphi_X(t)$.

2. Assume that $X \in \text{Ge}(p)$. Show that

   $$pX \xrightarrow{d} \text{Exp}(1),$$

   as $p \downarrow 0$.

3. (5B1540 2004-08-25) Y_n is uniformly distributed over the set $\{j/2^n; j = 0, 1, 2, \ldots, 2^n - 1\}$, $n = 1, 2, \ldots$. Show by means of a sequence of characteristic functions that

$$Y_n \overset{d}{\to} U(0, 1),$$

as $n \to \infty$.

4. (From [35]) This exercise studies convergence in distribution in relation to convergence of the corresponding sequence of expectations.
$\{X_n\}_{n \geq 1}$ is a sequence of r.v.'s such that for a real number r

$$\mathbf{P}(X_n = x) = \begin{cases} 1 - \frac{1}{n} & x = 0 \\ \frac{1}{n} & x = n^r. \end{cases}$$

(a) Show that

$$X_n \overset{d}{\to} \delta_0,$$

as $n \to \infty$.

(b) Investigate $\lim_{n \to \infty} E[X_n]$ for $r < 1$, $r = 0$ and $r > 1$. Is there convergence to the expectation of the limiting distribution δ_0?

5. $X \in \text{Po}(\lambda)$. Show that

$$\frac{X - \lambda}{\sqrt{\lambda}} \overset{d}{\to} N(0, 1),$$

as $\lambda \to \infty$.

6. (From [35]) $\{X_n\}_{n \geq 1}$ is a sequence of independent r.v.'s such that

$$\mathbf{P}(X_n = x) = \begin{cases} \frac{1}{2} & x = -\frac{1}{2^n} \\ \frac{1}{2} & x = \frac{1}{2^n}. \end{cases}$$

Set $S_n = X_1 + X_2 + \ldots + X_n$. Show that

$$S_n \overset{d}{\to} U(-1, 1),$$

as $n \to \infty$.

7. (From [35]) $\{X_n\}_{n \geq 1}$ is a sequence of independent r.v.'s such that

$$\mathbf{P}(X_n = x) = \begin{cases} \frac{1}{2} & x = -1 \\ \frac{1}{2} & x = 1. \end{cases}$$

Let $N \in \text{Po}(\lambda)$. N is independent of $\{X_n\}_{n \geq 1}$. Set $Y = X_1 + X_2 + \ldots + X_N$. Show that

$$\frac{Y}{\sqrt{\lambda}} \overset{d}{\to} N(0, 1),$$

as $\lambda \to \infty$.

8. (From [35]) $\{X_l^{(n)}\}_{l \geq 1}$ is for each n a sequence of independent r.v.'s such that

$$\mathbf{P}\left(X_l^{(n)} = x\right) = \begin{cases} 1 - \frac{1}{n} & x = 0 \\ \frac{1}{n} & x = 1. \end{cases}$$

Let N assume values in the non negative integers. N is independent of $\{X_l^{(n)}\}_{l\geq 1}$ for each n. Set

$$S_N^{(n)} = X_1^{(n)} + X_2^{(n)} + \ldots + X_{N+n}^{(n)}.$$

Show that

$$S_N^{(n)} \xrightarrow{d} \text{Po}(1),$$

as $n \to \infty$.

9. (From [35]) $\{X_n\}_{n\geq 1}$ is a sequence of independent r.v.'s, $X_n \in \text{Po}(\lambda)$ for each n. N is independent of $\{X_n\}_{n\geq 1}$, and $N \in \text{Ge}(p)$. Set

$$S_N = X_1 + X_2 + \ldots + X_N, \quad S_0 = 0.$$

Let now $\lambda \to 0$, while at the same time $p \to 0$, so that $\frac{p}{\lambda} \to \alpha$, where α is a pre-selected positive number. Show that

$$S_N \xrightarrow{d} \text{Fs}\left(\frac{\alpha}{\alpha+1}\right).$$

10. (From [49]) $\{X_n\}_{n\geq 1}$ is a sequence of independent r.v.'s, $X_n \in C(0,1)$. Show that

$$Y_n \overset{\text{def}}{=} \frac{1}{n}\max\left(X_1, \ldots, X_n\right) \xrightarrow{d} F_Y(y) = e^{-\frac{1}{\pi y}}, y > 0,$$

as $n \to \infty$. *Aid:* $\arctan(x) + \arctan(1/x) = \frac{\pi}{2}$ and $\arctan y = y - \frac{y^3}{3!} + \frac{y^5}{5!} - \frac{y^7}{7!} \ldots$.

11. (From [49]) $\{X_n\}_{n\geq 1}$ is a sequence of independent r.v.'s, $X_n \in \text{Pa}(1,2)$.

   (a) Show that

   $$Y_n \overset{\text{def}}{=} \min\left(X_1, \ldots, X_n\right) \xrightarrow{P} 1,$$

   as $n \to \infty$.

   (b)

   $$n(Y_n - 1) \xrightarrow{d} \text{Exp}\left(\frac{1}{2}\right),$$

   as as $n \to \infty$.

12. (From [49]) $X_n \in \text{Ge}(\lambda/(n+\lambda))$, where $\lambda > 0$. Show that

$$\frac{X_n}{n} \xrightarrow{d} \text{Exp}\left(\frac{1}{\lambda}\right),$$

as $n \to \infty$.

13. (From [49]) $X_n \in \text{Bin}(n^2, m/n)$, where $m > 0$. Show that

$$\frac{X_n - m \cdot n}{\sqrt{mn}} \xrightarrow{d} N\left(0,1\right),$$

as $n \to \infty$.

14. (From [49]) $\{X_n\}_{n\geq 1}$ is a sequence of independent and identically distributed r.v.'s, with the characteristic function

$$\varphi(t) = \begin{cases} 1 - \sqrt{|t|(2-|t|)} & |t| \leq 1 \\ 0 & |t| \geq 1. \end{cases}$$

Show that

$$\frac{1}{n^2}\sum_{k=1}^{n} X_k \xrightarrow{d} X,$$

as $n \to \infty$, where $\varphi_X(t) = e^{-\sqrt{2|t|}}$ and compare with (4.48).

15. (From [49]) $\{X_n\}_{n\geq 1}$ is a sequence of independent r.v.'s, $X_n \in \text{La}(a)$ for each n. N is independent of $\{X_n\}_{n\geq 1}$, and $N \in \text{Po}(m)$. Set

$$S_N = X_1 + X_2 + \ldots + X_N, \quad S_0 = 0.$$

Let now $m \to +\infty$, when at the same time $a \to 0$, so that $ma^2 \to 1$. Show that then

$$S_N \xrightarrow{d} N(0,2).$$

16. (From [49]) $\{X_n\}_{n\geq 1}$ is a sequence of independent r.v.'s, $X_n \in \text{Po}(\mu)$ for each n. N is independent of $\{X_n\}_{n\geq 1}$, and $N \in \text{Po}(\lambda)$. Set

$$S_N = X_1 + X_2 + \ldots + X_N, \quad S_0 = 0.$$

Let now $\lambda \to \infty$, while at the same time $\mu \to 0$, so that $\mu\lambda \to \upsilon > 0$. Show that

$$S_N \xrightarrow{d} \text{Po}(\upsilon).$$

17. (From [49]) $\{X_n\}_{n\geq 1}$ is a sequence of independent r.v.'s, $X_n \in \text{Po}(\mu)$ for each n. N is independent of $\{X_n\}_{n\geq 1}$, and $N \in \text{Ge}(p)$. Set

$$S_N = X_1 + X_2 + \ldots + X_N, \quad S_0 = 0.$$

Let now $\mu \to 0$, while at the same time $p \to 0$, so that $\frac{p}{\mu} \to \alpha > 0$. Show that then

$$S_N \xrightarrow{d} \text{Ge}\left(\frac{\alpha}{\alpha+1}\right).$$

18. (From [83])

Let $\{T_n\}_{n\geq 1}$ be a sequence of random variables such that

$$\sqrt{n}\,(T_n - \theta) \xrightarrow{d} N\left(0, \sigma^2(\theta)\right)$$

as $n \to \infty$. Let $g(x)$ be a differentiable function with the first derivative $g^{'}(x)$ which is continuous and $g^{'}(\theta) \neq 0$. Show that

$$\frac{\sqrt{n}\,(g(T_n) - g(\theta))}{g^{'}(T_n)\sigma(T_n)} \xrightarrow{d} N(0,1), \tag{6.32}$$

as $n \to \infty$. We can think of a sequence of statistical estimators T_n of the parameter θ that has the asymptotic mean zero and the asymptotic variance $\sigma^2(\theta)/n$. The difficulty is that the asymptotic variance depends on the parameter to be estimated. The result above provides of way of overcoming this so that we can, e.g., find approximate confidence intervals for $g(\theta)$, which are independent of θ.

19. [**Propagation of Error**]

(a) $X \in N(\mu, \sigma^2)$. Find the exact value of $\text{Var}\left[e^X\right]$.

(b) $X \in N(\mu, \sigma^2)$. Find $\text{Var}\left[e^X\right]$ using (6.15) and compare with (a).

6.8.2 Central Limit Theorem

1. (From [35]) $X \in \Gamma(a,b)$. Show that
$$\frac{X - E[X]}{\sqrt{\mathrm{Var}(X)}} \xrightarrow{d} N(0,1),$$
as $a \to \infty$. Use both of the following two methods:

   (a) The central limit theorem.

   (b) The continuity theorem 6.6.3.

2. (From [35]) Use the result in the preceding example to show that the $\chi^2(n)$ distribution with a large number of degrees of freedom is approximately $N(0,1)$.

3. (From [35]) $\{X_n\}_{n\geq 1}$ is a sequence of independent r.v.'s, $X_n \in U(0,1)$ for every n. Take
$$Y_n = e^{\sqrt{n}} \left(X_1 \cdot X_2 \cdot \ldots \cdot X_n\right)^{1/\sqrt{n}}.$$
Show that
$$Y_n \xrightarrow{d} \text{Log-Normal},$$
as $n \to \infty$. The Log-Normal distribution is found in (2.93). Here the parameters of the Log-Normal distribution are $\mu = 0$ and $\sigma^2 = 1$.

4. (From [35]) Let $\{X_n\}_{n\geq 1}$ be a sequence of independent r.v.'s, $X_n \in U(0,e)$ for every n. Show that
$$\left(X_1 \cdot X_2 \cdot \ldots \cdot X_n\right)^{1/\sqrt{n}} \xrightarrow{d} \text{Log-Normal},$$
as $n \to \infty$. The Log-Normal distribution is in (2.93). Here the parameters of the Log-Normal distribution are $\mu = 0$ and $\sigma^2 = 1$.

5. $\{X_n\}_{n\geq 1}$ is a sequence of independent and identically distributed SymBer - r.v.'s, i.e., they have with the common p.m.f.
$$p_X(k) = \begin{cases} \frac{1}{2} & k = -1 \\ \frac{1}{2} & k = 1. \end{cases}$$
Set
$$S_n = \sum_{k=1}^{n} \frac{X_k}{\sqrt{k}}.$$
Show that the following statements of convergence hold, as $n \to \infty$:

   (a) $\frac{\mathrm{Var}[S_n]}{\ln n} \to 1$. For this statement it is an advantage to know that $\sum_{k=1}^{n} \frac{1}{k} - \ln n \to \gamma$, where γ is Euler's constant $= 0.577\ldots$.

   (b)
$$\frac{S_n - E[S_n]}{\ln n} \xrightarrow{d} N(0,1).$$

6. (From [49]) $\{X_n\}_{n\geq 1}$ is an I.I.D. sequence of r.v.'s and $E[X] = \mu < \infty$ for for each n. N_n is independent of $\{X_n\}_{n\geq 1}$, and $N_n \in \mathrm{Ge}(p_n)$.

   Let now $p_n \to 0$, as $n \to \infty$. Show that
$$p_n\left(X_1 + X_2 + \ldots + X_{N_n}\right) \xrightarrow{d} \mathrm{Exp}(\mu),$$
as $n \to \infty$.

7. (From [49]) $\{X_n\}_{n\geq 1}$ is a sequence of positive I.I.D. r.v.'s with $E[X_n] = 1$ and $\text{Var}[X_n] = \sigma^2$. For $n \geq 1$

$$S_n \overset{\text{def}}{=} X_1 + X_2 + \ldots + X_n.$$

Show that

$$\sqrt{S_n} - \sqrt{n} \xrightarrow{d} N\left(0, \frac{\sigma^2}{4}\right),$$

as $n \to \infty$.

8. $\{X_i\}_{i\geq 1}$ are I.I.D. $N(\mu, \sigma^2)$. Find the asymptotic distribution of $\{e^{T_n}\}_{n\geq 1}$, where $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

6.8.3 Convergence in Probability

1. (From [35]) Let $X_k \in \text{Be}(p_k)$, $k = 1, 2, \ldots, n$. The variables X_k are independent.

$$S_n = \sum_{k=1}^{n} X_k.$$

Show that

$$\frac{1}{n}\left(S_n - \sum_{k=1}^{n} p_k\right) \xrightarrow{P} 0,$$

as $n \to \infty$. *Aid:* Use Chebychev's inequality (1.27).

2. (From [35]) X_k, $k = 1, 2, \ldots$, is a sequence of independent random variables such that

$$\mathbf{P}\left(X_k = 2^k\right) = \frac{1}{2}, \mathbf{P}\left(X_k = -2^k\right) = \frac{1}{2}.$$

Investigate, whether the weak law of large numbers holds for this sequence. *Aid:* Check first that $\sum_{k=1}^{n} X_k < 0$ with probability 1/2 and $\sum_{k=1}^{n} X_k > 0$ with probability 1/2. Then you can deduce that the weak law of large numbers does not hold.

3. (From [35]) Let $X_1, X_2, \ldots, X_{2n+1}$ be independent and identically distributed r.v.'s. They have a distribution function $F_X(x)$ such that the equation $F_X(m) = \frac{1}{2}$ has a unique solution. Set

$$M_n = \text{median}\left(X_1, X_2, \ldots X_{2n+1}\right).$$

The median of an odd number of numerical values is the middle one of the numbers. Median is thus algorithmically found by sorting the numerical values from the lowest value to the highest value and picking the middle one, i.e., the one separating the higher half of the list from the lower half. Show that

$$M_n \xrightarrow{P} m,$$

as $n \to \infty$.

4. (sf2940 2012-02-11) $X_1, X_2, \ldots, X_n, \ldots$ are independent and identically distributed r.v.'s. $X_n \overset{d}{=} X$, and $E[X] = \mu$, $\text{Var}[X] = \sigma^2 > 0$. Set

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k,$$

Show that

$$\frac{1}{n}\sum_{k=1}^{n}\left(X_k - \overline{X}_n\right)^2 \xrightarrow{P} \sigma^2, \tag{6.33}$$

as $n \to \infty$. *Aid:* In order to do this you may prefer considering the following

(a) Check that

$$\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2=\sum_{i=1}^{n}\left(X_i-\mu\right)^2-n\left(\overline{X}_n-\mu\right)^2.$$

(b) Part (a) yields

$$S_n^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2=$$

$$=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\mu\right)^2-\frac{n}{n-1}\left(\overline{X}_n-\mu\right)^2.$$

Apply the weak law of large numbers and a suitable property of convergence in probability to prove the assertion.

5. $X_1, X_2, \ldots, X_n, \ldots$ are independent and identically distributed r.v.'s., $X_n \overset{d}{=} X$ for each n, and $E[X]=\mu$, $\text{Var}[X]=\sigma^2>0$. Set

$$\overline{X}_n=\frac{1}{n}\sum_{k=1}^{n}X_k,$$

and

$$S_n^2=\frac{1}{n-1}\sum_{k=1}^{n}\left(X_k-\overline{X}_n\right)^2.$$

Show that

$$\frac{\sqrt{n}\left(\overline{X}_n-\mu\right)}{S_n}\xrightarrow{d}N(0,1).$$

as $n\to\infty$. *Aid:* The result in (6.33) is definitely useable here.

6. Show that the $t(n)$ distribution converges to $N(0,1)$, as $n\to\infty$. *Aid:* Consider exercise 5. in this section.

7. $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of r.v.'s such that as $n\to\infty$,

$$X_n\xrightarrow{P}X,\quad X_n\xrightarrow{P}Y. \tag{6.34}$$

Show that $\mathbf{P}(X=Y)=1$. *Aid:* Convince yourself of the inclusion of events

$$\{|X+Y|>2\epsilon\}\subset\{|X|>\epsilon\}\cup\{|Y|>\epsilon\}.$$

8. Let $X_1, X_2, \ldots,$ be independent r.v's and $\in U(-1,1)$. Let

$$Y_n\overset{\text{def}}{=}\frac{\sum_{k=1}^{n}X_k}{\sum_{k=1}^{n}X_k^2+\sum_{k=1}^{n}X_k^3}.$$

Show that

$$Y_n\xrightarrow{P}0,$$

as $n\to\infty$.

9. (From [49]) $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of independent r.v.'s. $X_k\in\text{Exp}(k!)$. Let $S_n=X_1+\ldots+X_n$. Show that

$$\frac{S_n}{n}\xrightarrow{d}\text{Exp}(1),$$

as $n\to+\infty$.

6.8.4 Proof of Theorem 6.7.2

This exercise consists of the steps (a) -((d) for establishing Theorem 6.7.2, c.f., [50].

(a) Show that even the limiting r.v. X is bounded almost surely by L, or,

$$\mathbf{P}\left(\mid X \mid \leq L\right) = 1.$$

Aid: Show that for any $\epsilon > 0$

$$\mathbf{P}\left(\mid X \mid \geq L + \epsilon\right) \leq \mathbf{P}(\mid X - X_n \mid \geq \epsilon).$$

and draw the desired conclusion.

(b) Justify by the preceding that $\mathbf{P}(|X - X_n|^2 \leq 4L^2) = 1$.

(c) Let I be the indicator function

$$I_{|X-X_n|\geq\epsilon} = \begin{cases} 1, \text{if } |X - X_n| \geq \epsilon \\ 0, \text{if } |X - X_n| < \epsilon\ . \end{cases}$$

Show that the inequality

$$|X - X_n|^2 \leq 4L^2 I_{|X-X_n|\geq\epsilon} + \epsilon^2$$

holds almost surely.

(d) Determine now the limit of

$$E\left[\mid X - X_n \mid^2\right],$$

as $n \to +\infty$.

6.8.5 Almost Sure Convergence, The Interrelationship Between Almost Sure Convergence and Mean Square Convergence, Criteria for Almost Sure Convergence

The exercises here rely on the Borel-Cantelli lemmas in the section 1.7.

1. Let, as in the proof of theorem 6.28,

$$A_n\left(\varepsilon\right) \overset{\text{def}}{=} \{\mid X_n - X \mid > \varepsilon\}$$

and

$$B_m\left(\varepsilon\right) = \cup_{n\geq m} A_n\left(\varepsilon\right).$$

Then show that

(a) $X_n \overset{\text{a.s.}}{\to} X$, as $n \to \infty \Leftrightarrow \mathbf{P}\left(B_m\left(\varepsilon\right)\right) \to 0$, as $m \to \infty$. Aid: Part of this is imbedded in the proof of theorem 6.28.

(b) $X_n \overset{\text{a.s.}}{\to} X$, as $n \to \infty$, if $\sum_{n\geq 1} \mathbf{P}\left(A_n\left(\varepsilon\right)\right) < \infty$ for all $\varepsilon > 0$.

2. (From [48, p.279]) Define

$$X_n = \begin{cases} n^3 & \text{with probability } n^{-2} \\ 0 & \text{with probability } 1 - n^{-2} \end{cases}$$

Then show that $X_n \overset{\text{a.s.}}{\to} 0$, but that the sequence X_n does not converge in L_2.
Aid: You need a result in the preceding exercise 1. in this section.

3. (From [48, p.279]) Define a sequence of independent r.v.'s

$$X_n = \begin{cases} 1 & \text{with probability } n^{-1} \\ 0 & \text{with probability } 1 - n^{-1} \end{cases}$$

Then show that $X_n \overset{2}{\to} 0$, but that the sequence X_n does not converge almost surely.
Aid: You need a result in the preceding exercise 1. in this section.

4. (From [35]) $X \in U(0,1)$. Write an outcome $X = x$ in its binary expansion

$$x = 0.a_1 a_2 a_3 a_4 a_5 \ldots$$

where $a_k = 0$ or $a_k = 1$. Show that

$$\frac{1}{n} \sum_{k=1}^{n} a_k \overset{\text{a.s}}{\to} \frac{1}{2},$$

as $n \to \infty$.

5. $\{X_n\}_{n \geq 1}$ is a sequence of random variables such that there is a sequence of (non negative) numbers $\{\epsilon_n\}_{n \geq 1}$ such that $\sum_{n=1}^{\infty} \epsilon_n < \infty$ and

$$\sum_{n=1}^{\infty} \mathbf{P}\left(\mid X_{n+1} - X_n \mid > \epsilon_n\right) < +\infty. \tag{6.35}$$

Show that there is a random variable X such that $X_n \overset{\text{a.s.}}{\to} X$, as $n \to \infty$.

Chapter 7

Convergence in Mean Square and a Hilbert Space

7.1 Convergence in Mean Square; Basic Points of View

7.1.1 Definition

We restate the definition of convergence in mean square.

Definition 7.1.1 A random sequence $\{X_n\}_{n=1}^{\infty}$ with $E\left[X_n^2\right] < \infty$ is said to **converge in mean square** to a random variable X, if

$$E\left[|X_n - X|^2\right] \to 0 \tag{7.1}$$

as $n \to \infty$.

■

We write also

$$X_n \overset{2}{\to} X.$$

This definition is silent about convergence of individual *sample paths* $(X_n(\omega))_{n=1}^{\infty}$ (a fixed $\omega \in \Omega$). By a sample path we mean that we take a **fixed** $\omega \in \Omega$ and obtain the sequence of outcomes $(X_n(\omega))_{n=1}^{\infty}$. Hence, by the above we can not in general claim that $X_n(\omega) \to X(\omega)$ for an arbitrarily chosen ω or almost surely, as shown in the preceding.

7.1.2 The Hilbert Space $L_2(\Omega, \mathcal{F}, \mathbf{P})$

Convergence in mean square, as defined above, deals with random variables X such that $E\left[X^2\right] < \infty$. Then we say that $X \in L_2(\Omega, \mathcal{F}, \mathbf{P})$. For $X \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ and $Y \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ we can set

$$\langle X, Y \rangle \overset{\text{def}}{=} E[XY]. \tag{7.2}$$

We can easily verify that

(i) $\langle X, Y \rangle = \langle Y, X \rangle$,

(ii) $\langle X, X \rangle \geq 0$, $\langle X, X \rangle = 0 \Leftrightarrow X = 0$ almost surely.

(iii) $\langle aX + bY, Z \rangle = a\langle X, Z \rangle + b\langle Y, Z \rangle$, where $Z \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ and a and b are real constants.

In view of (i)-(iii) we can regard random variables $X \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ as elements in a real linear vector space with the *scalar product* $\langle X, Y \rangle$. Hence $L_2(\Omega, \mathcal{F}, \mathbf{P})$ equipped with the scalar product $\langle X, Y \rangle$ is a *pre-Hilbert space*, see e.g., in [96, Appendix H p. 252][1] or [89, ch. 17.7] and [92, pp. 299−301]. Thus we define the norm (or length)

$$\parallel X \parallel \overset{\text{def}}{=} \sqrt{\langle X, X \rangle}. \tag{7.3}$$

and the *distance* or *metric*

$$\delta(X, Y) \overset{\text{def}}{=} \parallel X - Y \parallel = \sqrt{E(X - Y)^2}. \tag{7.4}$$

Then we can write

$$X_n \overset{2}{\to} X \Leftrightarrow \delta(X_n, X) \to 0.$$

In fact one can prove the *completeness* of our pre-Hilbert space, [63, p. 22]. Completeness means that if

$$\delta(X_n, X_m) \to 0, \quad \text{as } \min(m, n) \to \infty \ ,$$

then there exists $X \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ such that $X_n \overset{2}{\to} X$. In other words, $L_2(\Omega, \mathcal{F}, \mathbf{P})$ equipped with the scalar product $\langle X, Y \rangle$ is a **Hilbert space**. Hence several properties in this chapter are nothing but special cases of general properties of Hilbert spaces.

Hilbert spaces are important, as, amongst other things, they possess natural notions of length, *orthogonality* and *orthogonal projection*, see [36, chapter 6.] for a full account. Active knowledge about Hilbert spaces in general will NOT be required in the examination of this course.

7.2 Cauchy-Schwartz and Triangle Inequalities

The norm of any Hilbert space, like here (7.3) in $L_2(\Omega, \mathcal{F}, \mathbf{P})$, satisfies two famous and useful inequalities.

Lemma 7.2.1 *$X \in L_2(\Omega, \mathcal{F}, \mathbf{P})$, $Y \in L_2(\Omega, \mathcal{F}, \mathbf{P})$.*

$$|E[XY]| \leq E[|XY|] \leq \sqrt{E[|X|^2]} \cdot \sqrt{E[|Y|^2]}. \tag{7.5}$$

$$\sqrt{E[|X \pm Y|^2]} \leq \sqrt{E[|X|^2]} + \sqrt{E[|Y|^2]}. \tag{7.6}$$

■

The inequality (7.5) is known as the *Cauchy-Schwartz* inequality, and is but a special case of Hölder's inequality in (1.25) for $p = q = 2$. The inequality (7.6) is known as the *triangle* inequality.

7.3 Properties of Mean Square Convergence

Theorem 7.3.1 The two random sequences $\{X_n\}_{n=1}^{\infty}$ and $\{Y_n\}_{n=1}^{\infty}$ are defined in the same probability space and $X_n \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ for all n and $Y_n \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ for all n. Let

$$X_n \overset{2}{\to} X, Y_n \overset{2}{\to} Y.$$

Then it holds that

(a)

$$E[X] = \lim_{n \to \infty} E[X_n]$$

[1] The reference is primarily to this book written in Swedish, as it is the textbook for SI1140 Mathematical Methods in Physics `http://www.kth.se/student/kurser/kurs/SI1140?l=en_UK`, which is a mandatory course for the programme CTFYS at KTH.

(b)

$$E\left[|X|^2\right] = \lim_{n\to\infty} E\left[|X_n|^2\right]$$

(c)

$$E\left[XY\right] = \lim_{n\to\infty} E\left[X_n \cdot Y_n\right]$$

(d) If $Z \in L_2\left(\Omega, \mathcal{F}, \mathbf{P}\right)$, then

$$E\left[X \cdot Z\right] = \lim_{n\to\infty} E\left[X_n Z\right].$$

Proof We prove (c), when (a) and (b) have been proved. First, we see that $|E\left[X_n Y_n\right]| < \infty$ and $|E\left[XY\right]| < \infty$ by virtue of the Cauchy - Schwartz inequality and the other assumptions. In order to prove (c) we consider

$$|E\left[X_n Y_n\right] - E\left[XY\right]| \leq E|\left[(X_n - X)Y_n + X\left(Y_n - Y\right)|\right],$$

since $|E\left[Z\right]| \leq E\left[|Z|\right]$. Now we can use the ordinary triangle inequality for real numbers and obtain:

$$E|\left[(X_n - X)Y_n + X\left(Y_n - Y\right)|\right] \leq E\left[|(X_n - X)Y_n|\right] + E\left[|X\left(Y_n - Y\right)|\right].$$

But Cauchy-Schwartz entails now

$$E\left[|(X_n - X)Y_n|\right] \leq \sqrt{E\left[|X_n - X|^2\right]}\sqrt{E\left[|Y_n|^2\right]}$$

and

$$E\left[|(Y_n - Y)X|\right] \leq \sqrt{E\left[|Y_n - Y|^2\right]}\sqrt{E\left[|X|^2\right]}.$$

But by assumption $\sqrt{E\left[|X_n - X|^2\right]} \to 0$, $E\left[|Y_n|^2\right] \to E\left[|Y|^2\right]$ (part (b)), and $\sqrt{E\left[|Y_n - Y|^2\right]} \to 0$, and thus the assertion (c) is proved. ■

We shall often need **Cauchy's criterion for mean square convergence** , which is the next theorem.

Theorem 7.3.2 Consider the random sequence $\{X_n\}_{n=1}^{\infty}$ with $X_n \in L_2\left(\Omega, \mathcal{F}, \mathbf{P}\right)$ for every n. Then

$$E\left[|X_n - X_m|^2\right] \to 0 \tag{7.7}$$

as $\min(m,n) \to \infty$ if and only if there exists a random variable X such that

$$X_n \overset{2}{\to} X.$$

■

The assertion here is nothing else but that the pre-Hilbert space defined section 7.1.2 above is complete. A useful form of Cauchy's criterion is known as *Loève's criterion*:

Theorem 7.3.3

$$E\left[|X_n - X_m|^2\right] \to 0 \Longleftrightarrow E\left[X_n X_m\right] \to C. \tag{7.8}$$

as $\min(m,n) \to \infty$, where the constant C is finite and independent of the way $m, n \to \infty$.

Proof Proof of $\Longleftarrow$: We assume that $E\left[X_n X_m\right] \to C$. Thus

$$E\left[|X_n - X_m|^2\right] = E\left[X_n \cdot X_n + X_m \cdot X_m - 2X_n \cdot X_m\right]$$

$$\to C + C - 2C = 0.$$

Proof of $\Longrightarrow$: We assume that $E\left[|X_n - X_m|^2\right] \to 0$. Then for any m and n

$$E\left[X_n X_m\right] = E\left[(X_n - X)\, X_m\right] + E\left[X X_m\right].$$

Here,

$$E\left[(X_n - X)\, X_m\right] \to E\left[0X\right] = 0,$$

by theorem 7.3.1 (c), since $X_n \overset{2}{\to} X$ according to Cauchy's criterion. Also,

$$E\left[X X_m\right] \to E\left[X^2\right] = C$$

by theorem 7.3.1 (d). Hence

$$E\left[X_n X_m\right] \to 0 + C = C.$$

■

7.4 Applications

7.4.1 Mean Ergodic Theorem

Although the definition of converge in mean square encompasses convergence to a random variable, in many applications we shall encounter convergence to a constant.

Theorem 7.4.1 The random sequence $\{X_n\}_{n=1}^{\infty}$ is uncorrelated and with $E\left[X_n\right] = \mu < \infty$ for every n and $\mathrm{Var}\left[X_n\right] = \sigma^2 < \infty$ for every n. Then

$$\frac{1}{n}\sum_{j=1}^{n} X_n \overset{2}{\to} \mu,$$

as $n \to \infty$.

Proof Let us set $S_n = \frac{1}{n}\sum_{j=1}^{n} X_n$. We have $E\left[S_n\right] = \mu$ and $\mathrm{Var}\left[S_n\right] = \frac{1}{n}\sigma^2$, since the variables are uncorrelated. For the claimed mean square convergence we need to consider

$$E\left[|S_n - \mu|^2\right] = E\left[(S_n - E\left[S_n\right])^2\right] = \mathrm{Var}\left[S_n\right] = \frac{1}{n}\sigma^2$$

so that

$$E\left[|S_n - \mu|^2\right] = \frac{1}{n}\sigma^2 \to 0$$

as $n \to \infty$, as was claimed. ■

7.4.2 Mean Square Convergence of Sums

Consider a sequence $\{X_n\}_{n=0}^{\infty}$ of independent random variables in $L_2\left(\Omega, \mathcal{F}, \mathbf{P}\right)$ with $E\left[X_i\right] = \mu$ and $\mathrm{Var}\left[X_i\right] = \sigma^2$. We wish to find conditions such that we may regard an infinite linear combination of random variables as a mean square convergent sum, i.e.,

$$\sum_{i=0}^{n} a_i X_i \overset{2}{\to} \sum_{i=0}^{\infty} a_i X_i,$$

as $n \to \infty$. The symbol $\sum_{i=0}^{\infty} a_i X_i$ is a notation for a random variable in $L_2(\Omega, \mathcal{F}, \mathbf{P})$ defined by the converging sequence. The Cauchy criterion in theorem 7.3.2 gives for $Y_n = \sum_{i=0}^{n} a_i X_i$ and $n < m$ that

$$E\left[|Y_n - Y_m|^2\right] = E\left[|\sum_{i=n+1}^{m} a_i X_i|^2\right] = \sigma^2 \sum_{i=n+1}^{m} a_i^2 + \mu^2 \left(\sum_{i=n+1}^{m} a_i\right)^2, \tag{7.9}$$

since by Steiner's formula $EZ^2 = \text{Var}(Z) + (E[Z])^2$ for any random variable that has variance. We need to recall a topic from mathematical analysis.

Remark 7.4.1 The **Cauchy sequence criterion for convergence of sums** states that a sum of real numbers a_i

$$\sum_{i=0}^{\infty} a_i$$

converges if and only if the sequence of partial sums is a **Cauchy sequence**. By a partial sum we mean a finite sum like

$$S_n = \sum_{i=0}^{n} a_i.$$

That the sequence of partial sums is a Cauchy sequence says that for every $\varepsilon > 0$, there is a positive integer N such that for all $m \geq n \geq N$ we have

$$|S_m - S_n| = \left|\sum_{i=n+1}^{m} a_i\right| < \varepsilon,$$

which is equivalent to

$$\lim_{\substack{n\to\infty \\ k\to\infty}} \sum_{i=n}^{n+k} a_i = 0. \tag{7.10}$$

This can be proved as in [69, p.137−138]. The advantage of checking convergence of $\sum_{i=0}^{\infty} a_i$ by partial sums is that one does not need to guess the value of the limit in advance.

■

By the Cauchy sequence criterion for convergence of sums we see in the right hand side of (7.9) by virtue of (7.10) that $E\left[|Y_n - Y_m|^2\right]$ converges by the Cauchy sequence convergence of sums to zero if and only if

- in case $\mu \neq 0$

  $$\sum_{i=0}^{\infty} |a_i| < \infty,$$

  (which implies $\sum_{i=0}^{\infty} a_i^2 < \infty$)

- in case $\mu = 0$

  $$\sum_{i=0}^{\infty} a_i^2 < \infty.$$

7.4.3 Mean Square Convergence of Normal Random Variables

Let us suppose that we have

$$X_n \in N\left(\mu_n, \sigma_n^2\right) \tag{7.11}$$

and, as $n \to \infty$,

$$X_n \overset{2}{\to} X. \tag{7.12}$$

Thus (7.12) implies in view of Theorem 7.3.1 (a) and (b) that there are numbers μ and σ^2 such that

$$\mu_n \to \mu = E[X], \quad \sigma_n^2 \to \sigma^2 = \text{Var}[X].$$

Then the characteristic functions for X_n are

$$\varphi_{X_n}(t) = e^{i\mu_n t - \frac{1}{2}\sigma_n^2 t^2}.$$

Therefore we have for all real t that

$$\varphi_{X_n}(t) \to e^{i\mu t - \frac{1}{2}\sigma^2 t^2},$$

and thus $X \in N\left(\mu, \sigma^2\right)$ by the continuity theorem 6.6.3 for characteristic functions

Theorem 7.4.2 If $X_n \in N\left(\mu_n, \sigma_n^2\right)$ and $X_n \overset{2}{\to} X$, as $n \to \infty$, then X is a normal random variable.

■

As an application, we can continue with the sums in section 7.4.2. If X_i are independent $N\left(\mu, \sigma^2\right)$, and $\sum_{i=0}^{\infty} |a_i| < \infty$, then

$$\sum_{i=0}^{\infty} a_i X_i \in N\left(\mu \sum_{i=0}^{\infty} a_i, \sigma^2 \sum_{i=0}^{\infty} a_i^2\right). \tag{7.13}$$

7.5 Subspaces, Orthogonality and Projections in $L_2(\Omega, \mathcal{F}, \mathbf{P})$

A *subspace* M of $L_2(\Omega, \mathcal{F}, \mathbf{P})$ is a subset such that

- If $X \in M$ and $Y \in M$, then $aX + bY \in M$, for all real constants a and b.

If M_α is a subspace for α in an arbitrary index set I, then $\cap_{\alpha \in I} M_\alpha$ is a subspace.

If a subspace M is such that if $X_n \in M$ and if $X_n \overset{2}{\to} X$, then $X \in M$, we say that M is **closed**.

Example 7.5.1 Let $M_0 = \{X \in L_2(\Omega, \mathcal{F}, \mathbf{P}) \mid E[X] = 0\}$. This is clearly a subspace. By Theorem 7.3.1 (a) M_0 is also a closed subspace. It is also a Hilbert space in its own right.

■

Example 7.5.2 Let $\{\mathbf{0}\} = \{X \in L_2(\Omega, \mathcal{F}, \mathbf{P}) \mid X = 0 \quad \text{a.s.}\}$. Then $\{\mathbf{0}\}$ is a subspace, and a subspace of any other subspace.

■

Let $\mathbf{X} = (X_1, X_2, \ldots)$ be a sequence of random variables in $L_2(\Omega, \mathcal{F}, \mathbf{P})$. We define the subspace *spanned by* $X_1, X_2, \ldots, X_n$, which is the subspace $\mathcal{L}_n^X$ consisting of all linear combinations $\sum_{i=1}^n a_i X_i$ of the random variables, and their limits in the mean square, or

$$\mathcal{L}_n^X = \overline{\text{sp}}\{X_1, X_2, \ldots, X_n\}. \tag{7.14}$$

Since we here keep the number of random variables fixed and finite, the limits in the mean square are limits of

$$Y_m = \sum_{i=1}^n a_i(m) X_i, \quad \text{as } m \to \infty.$$

Next we define orthogonality [96, p. 253];

Definition 7.5.1 Two random variables $X \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ and $Y \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ are said to be **orthogonal**, if

$$\langle X, Y \rangle = 0. \tag{7.15}$$

■

If $X \in M_0$ and $Y \in M_0$, the subspace in example 7.5.1, then orthogonality means that

$$\langle X, Y \rangle = E[XY] = 0,$$

and we are more used to saying that X and Y are uncorrelated.

Definition 7.5.2 Let $X \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ and M be a subspace of $L_2(\Omega, \mathcal{F}, \mathbf{P})$. If it holds for all Y in M that $\langle X, Y \rangle = 0$, we say that X is **orthogonal to the subspace** M, and write this as

$$X \perp M. \tag{7.16}$$

We define the subspace $M^{\perp}$

$$M^{\perp} \overset{\text{def}}{=} \{X \in L_2(\Omega, \mathcal{F}, \mathbf{P}) \mid X \perp M\}. \tag{7.17}$$

■

One might also want to check that $M^{\perp}$ is actually a subspace, as is claimed above.

The following theorem is fundamental for many applications, and holds, of course, for any Hilbert space, not just for $L_2(\Omega, \mathcal{F}, \mathbf{P})$, where we desire to take advantage of it.

Theorem 7.5.3 Let M be a closed subspace of $L_2(\Omega, \mathcal{F}, \mathbf{P})$ Then any $X \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ has a unique decomposition

$$X = \text{Proj}_M(X) + Z \tag{7.18}$$

where $\text{Proj}_M(X) \in M$ and $Z \in M^{\perp}$. In addition it holds that

$$\| X - \text{Proj}_M X \| = \min_{V \in M} \| X - V \| \tag{7.19}$$

Proof is omitted, and can be found in many texts and monographs, see, e.g., [26, pp. 35−36] or [36, p.204−206]. The theorem and proof in [96, p. 262] deals with a special case of the result above. ■

For our immediate purposes the interpretation of theorem 7.5.3 is of a higher priority than expediting its proof. We can think of $\text{Proj}_M(X)$ as an *orthogonal projection* of X to M or as an estimate of X by means of M. Then Z is the estimation error. $\text{Proj}_M(X)$ is optimal in the sense that it minimizes the mean squared error $\| X - V \|^2 = E\left[(X - V)^2\right]$.

This interpretation becomes more obvious if we take $M = \mathcal{L}_n^X$ as in (7.14). Then $\text{Proj}_M(X) \in M$ must be of the form

$$\text{Proj}_M(X) = \sum_{i=1}^{n} a_i X_i. \tag{7.20}$$

which is an **optimal linear mean square error estimate** of X by means of X_1, $X_2 \ldots X_n$. The coefficients a_i can be found as a solution to a system of linear equations, see the exercises below.

Example 7.5.4 We reconsider M_0 in example 7.5.1 above. Then the random variable $\mathbf{1}$, i.e., $\mathbf{1}(\omega) = 1$ for almost all ω, is orthogonal to M_0, since

$$E[X \cdot \mathbf{1}] = E[X] = 0,$$

for any $X \in M_0$. The orthogonal subspace $M_0^{\perp}$ is in fact spanned by $\mathbf{1}$,

$$M_0^{\perp} = \{Z \mid Z = c \cdot \mathbf{1}, c \in \mathbf{R}\}.$$

Every X in $L_2(\Omega, \mathcal{F}, \mathbf{P})$ can then be uniquely decomposed as

$$X = \text{Proj}_M(X) + Z, \quad \text{Proj}_M(X) = X - E[X], Z = E[X] \cdot \mathbf{1}.$$

■

Example 7.5.5 X and Y are random variables in $L_2(\Omega, \mathcal{F}, \mathbf{P})$. Let us consider the subspace $M = \mathcal{L}_1^Y \subset M_0$ (example 7.5.1 above) spanned by $Y - \mu_Y$, where $\mu_Y = E[Y]$. Thus $\text{Proj}_M(X - \mu_X)$, $\mu_X = E[X]$, is a random variable that is of the form

$$\text{Proj}_M(X - \mu_X) = a(Y - \mu_Y)$$

for some real number a. Let

$$Z = (X - \mu_X) - a(Y - \mu_Y).$$

Then we know by theorem 7.5.3 that for the optimal error Z

$$Z \perp \mathcal{L}_1^Y,$$

which is the same as saying that we must find a that satisfies

$$\langle Z, a(Y - \mu_Y)\rangle = 0.$$

When we write out this in full terms we get

$$E[[(X - \mu_X) - a(Y - \mu_Y)) \cdot a(Y - \mu_Y)] = 0 \tag{7.21}$$

$$\Leftrightarrow$$

$$aE[(X - \mu_X) \cdot (Y - \mu_Y)] - a^2 E\left[(Y - \mu_Y)^2\right] = 0$$

$$\Leftrightarrow$$

$$a\text{Cov}(X, Y) = a^2\text{Var}(Y),$$

which gives

$$a = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)}. \tag{7.22}$$

This makes good sense, since if X and Y are independent, then $\text{Cov}(X, Y) = 0$, and $\text{Proj}_M(X - \mu_X) = \mathbf{0}$ (= the random variable $\mathbf{0}(\omega) = 0$ for all $\omega \in \Omega$). Clearly, if X and Y are independent, there is no information about X in Y (and vice versa), and there is no effective estimate that would depend on Y. Let us write

$$a = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} \cdot \frac{\sqrt{\text{Var}(X)}}{\sqrt{\text{Var}(Y)}}$$

$$= \rho_{X,Y} \cdot \frac{\sqrt{\text{Var}(X)}}{\sqrt{\text{Var}(Y)}},$$

where $\rho_{X,Y}$ is the coefficient of correlation between X and Y. Then we have

$$X - \mu_X = \rho_{X,Y} \cdot \frac{\sqrt{\text{Var}(X)}}{\sqrt{\text{Var}(Y)}} \cdot (Y - \mu_Y) + Z$$

$$\Leftrightarrow$$

$$X = \mu_X + \rho_{X,Y} \cdot \frac{\sqrt{\text{Var}(X)}}{\sqrt{\text{Var}(Y)}} \cdot (Y - \mu_Y) + Z.$$

Therefore, the *the best* **linear** *mean square estimator* of X by means of Y is

$$\widehat{X} = \mu_X + \rho_{X,Y} \cdot \frac{\sqrt{\text{Var}(X)}}{\sqrt{\text{Var}(Y)}} \cdot (Y - \mu_Y). \tag{7.23}$$

■

7.6 Exercises

7.6.1 Mean Square Convergence

1. Assume $X_n \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ for all n and $Y_n \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ for all n and

$$X_n \overset{2}{\to} X, Y_n \overset{2}{\to} Y,$$

as $n \to \infty$. Let a and b be real constants. Show that

$$aX_n + bY_n \overset{2}{\to} aX + bY,$$

as $n \to \infty$. You should use the definition of mean square convergence and suitable properties of $\parallel X \parallel$ as defined in (7.3).

2. Consider

$$X_n = \sum_{k=1}^{n} \frac{1}{k} W_k, \quad n \geq 1.$$

where W_k are independent and $N(0, \sigma^2)$ -distributed.

(a) Determine the distribution of X_n.

(b) Show that there is the convergence

$$X_n \overset{2}{\to} X \quad \text{as } n \to \infty,$$

and that $X \in N\left(0, \frac{\sigma^2 \pi^2}{6}\right)$.

3. The sequence $\{X_n\}_{n=1}^{\infty}$ of random variables is such that $E[X_i] = \mu$ for all i, $\text{Cov}(X_i, X_j) = 0$, if $i \neq j$ and such that $\text{Var}(X_i) \leq c$ and for all i. Observe that the variances are thus uniformly bounded but not necessarily equal to each other for all i. This changes the setting from that in theorem 7.4.1 above. Show that

$$\frac{1}{n} \sum_{j=1}^{n} X_j \overset{2}{\to} \mu,$$

as $n \to \infty$.

7.6.2 Optimal Estimation as Projection on Closed Linear Subspaces in $L_2(\Omega, \mathcal{F}, \mathbf{P})$

1. Let $X \in M_0$, see example 7.5.1. Assume also that $Y_1, \ldots, Y_N$ are in M_0. The closed subspace in M_0 spanned by $Y_1, \ldots, Y_N$ is

$$M = \mathcal{L}_N^Y.$$

We want to find the optimal projection of X to $\mathcal{L}_N^Y$, which means to find $\text{Proj}_M(X) = \sum_{k=1}^{N} a_k Y_k$ such that $E\left[(X - V)^2\right]$ is minimized for $V \in \mathcal{L}_N^Y$. We set

$$\begin{aligned} \gamma_{mk} &= \langle Y_m, Y_k \rangle, \qquad m = 1, \ldots, N; k = 1, \ldots, N \\ \gamma_{om} &= \langle Y_m, X \rangle, \qquad m = 1, \ldots, N. \end{aligned} \tag{7.24}$$

(a) Show first that if $a_1, \ldots, a_N$ are solutions to the linear system of equations

$$\sum_{k=1}^{N} a_k \gamma_{mk} = \gamma_{om}; m = 1, \ldots, N, \tag{7.25}$$

then

$$X - \text{Proj}_M(X) \perp \mathcal{L}_N^Y, \tag{7.26}$$

c.f., (7.21) above.

(b) Show that

$$E\left[(X - (a_1 Y_1 + \ldots + a_N Y_N))^2\right] \tag{7.27}$$

is minimized, if the coefficients $a_1, \ldots, a_N$ satisfy the system of equations (7.25).

Aid: Let $b_1, \ldots, b_N$ be an arbitrary set of real numbers. set

$$\text{Proj}_M(X) = \sum_{k=1}^{N} a_k Y_k.$$

for $a_1, \ldots, a_N$ that satisfy the system of equations (7.25). Then we can write the estimation error ε using an arbitrary linear estimator $b_1 Y_1 + \ldots + b_N Y_N$ in $\mathcal{L}_N^Y$ as

$$\varepsilon = X - (b_1 Y_1 + \ldots + b_N Y_N) = (X - \text{Proj}_M(X)) + \sum_{k=1}^{N} (a_k - b_k) Y_k.$$

Expand now $E\left[\varepsilon^2\right]$ and recall (7.26).

2. Let $\mathcal{P} = \{A_1, A_2, \ldots, A_k\}$ be partition of Ω, i.e., $A_i \in \mathcal{F}$, $i = 1, 2, \ldots, k$, $A_i \cap A_j = \emptyset$, $j \neq i$ and $\cup_{i=1}^{k} A_i = \Omega$. Let χ_{A_i} $i = 1, \ldots, k$ be the indicator functions of the cells A_i $i = 1, \ldots, k$, respectively. Note that every $\chi_{A_i} \in L_2(\Omega, \mathcal{F}, \mathbf{P})$. We take the subspace spanned by all linear combinations of the indicator functions

$$\mathcal{L}_n^{\mathcal{P}} = \overline{\text{sp}}\left\{\chi_{A_1}, \chi_{A_2}, \ldots, \chi_{A_n}\right\}. \tag{7.28}$$

In other words, $\mathcal{L}_n^{\mathcal{P}}$ is spanned by random variables of the form

$$\sum_{i=1}^{k} c_i \chi_{A_i}(\omega),$$

where c_is are real constants.

Let $X \in M_0$, (example 7.5.1). Find the optimal projection $\text{Proj}_{\mathcal{L}_n^{\mathcal{P}}}(X)$ of X to $\mathcal{L}_n^{\mathcal{P}}$. A good hint is that the answer should coincide with the expression for the conditional expectation $E[X \mid \mathcal{P}]$ in section 3.4 above.

3. (From [50]) Let X have a Rayleigh distribution with parameter $2\sigma^2 > 0$, $X \in \text{Ra}\left(2\sigma^2\right)$, or

$$f_X(x) = \begin{cases} \frac{x}{\sigma^2} e^{-x^2/2\sigma^2} & x \geq 0 \\ 0 & \text{elsewhere.} \end{cases}$$

Let $Z \in U(0,1)$ (=the uniform distribution on $(0,1)$). X and Z are independent. We multiply these to get

$$Y = Z \cdot X.$$

(a) Consider $M = \mathcal{L}_1^Y$, the subspace spanned by $Y - E[Y]$. Find that the best linear estimator in mean square sense, $\mathrm{Proj}_M(X - E[X])$, is

$$\widehat{X} = \sigma\sqrt{\frac{\pi}{2}} + \frac{\left(1 - \frac{\pi}{4}\right)}{\left(\frac{2}{3} - \frac{\pi}{4}\right)}\left(Y - \frac{\sigma}{2}\sqrt{\frac{\pi}{2}}\right).$$

Aid: This is an application of the results in example 7.5.5, see (7.23). Preferably use the expression for a from (7.22).

(b) Show that

$$E[X \mid Y] = \frac{\sigma}{\sqrt{2\pi}} \cdot \frac{e^{-\frac{-Y^2}{2\sigma^2}}}{Q\left(\frac{Y}{\sigma}\right)},$$

where the Q**-function** $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{t^2}{2}}dt$ is the complementary distribution function for the standard normal distribution, i.e., $\Phi(x) = 1 - Q(x)$.
Aid: Find the joint p.d.f. $f_{X,Y}(x, y)$ and the marginal p.d.f. $f_Y(y)$ and compute $E[X \mid Y = y]$ by its definition.

Remark 7.6.1 This exercise shows that the **best estimator in the mean square sense**, $E[X \mid Y]$, see section 3.7.3 in chapter 2., and the **best *linear* estimator in the mean square sense**, $\mathrm{Proj}_M(X - E[X])$, by no means have to be identical.

Chapter 8

Gaussian Vectors

8.1 Multivariate Gaussian Distribution

8.1.1 Why Gaussianity ?

The Gaussian distribution is **central in probability theory**, since it is the final and stable or equilibrium distribution to which other distributions gravitate under a wide variety of smooth operations, e.g., convolutions and stochastic transformations, and which, once attained, is maintained through an even greater variety of transformations.

In the sequel, see the chapters 10 and 11, we shall discuss the probability theory in relation to molecular motion, [10, 17], and physical noise in a physical system. The pertinent events in the system could be the individual impacts of small molecules, or the electric force from many electrons moving in a conductor. The total force applied by these small molecules or electrons is the sum of the random forces applied by an individual particle. Since the total force is a sum of many random variables and the microscopic fluctuations are fast as compared to the motion of the system, we can think of evoking the **central limit theorem** to model the noise with a Gaussian distribution.
Let us collect from the preceding chapters the following facts;

- X is a normal a.k.a. Gaussian random variable, if

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2},$$

  where μ is real and $\sigma > 0$.

- Notation: $X \in N(\mu, \sigma^2)$.

- Properties: $X \in N(\mu, \sigma^2) \Rightarrow E[X] = \mu$, $\text{Var}[X] = \sigma^2$.

- $X \in N(\mu, \sigma^2)$, then the moment generating function is

$$\psi_X(t) = E\left[e^{tX}\right] = e^{t\mu + \frac{1}{2}t^2\sigma^2}, \tag{8.1}$$

  and the characteristic function is

$$\varphi_X(t) = E\left[e^{itX}\right] = e^{it\mu - \frac{1}{2}t^2\sigma^2}. \tag{8.2}$$

- $X \in N(\mu, \sigma^2) \Rightarrow Y = aX + b \in N(a\mu + b, a^2\sigma^2)$.

- $X \in N(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \in N(0,1)$.

We shall next see that all of these properties are special cases of the corresponding properties of a multivariate normal/Gaussian random variable as defined below, which bears witness to the statement that the normal distribution is central in probability theory.

8.1.2 Notation for Vectors, Mean Vector, Covariance Matrix & Characteristic Functions

An $n \times 1$ random vector or a multivariate random variable is denoted by

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = (X_1, X_2, \ldots, X_n)^{'},$$

where $'$ is the vector transpose. A vector in $\mathbf{R}^n$ is designated by

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = (x_1, x_2, \ldots, x_n)^{'}.$$

For the clarity of expression we note that

$$\mathbf{x}^{'}\mathbf{x} = \sum_{i=1}^{n} x_i^2$$

is a **scalar product** (i.e. a number) and

$$\mathbf{x}\mathbf{x}^{'} = (x_i x_j)_{i=1,j=1}^{n,n}$$

is $n \times n$-matrix. The same statements hold for the random variable $\mathbf{X}^{'}\mathbf{X}$ and the random matrix, or matrix of random variables, $\mathbf{X}\mathbf{X}^{'}$.

We denote by $F_{\mathbf{X}}(\mathbf{x})$ the joint distribution function of $\mathbf{X}$, which means that

$$F_{\mathbf{X}}(\mathbf{x}) = \mathbf{P}(\mathbf{X} \leq \mathbf{x}) = \mathbf{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n).$$

The following definitions are natural. We have the **mean vector**

$$\mu_{\mathbf{X}} = E[\mathbf{X}] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{pmatrix},$$

which is a $n \times 1$ column vector of means (=expected values) of the components of $\mathbf{X}$.
The **covariance matrix** is a square $n \times n$ -matrix

$$C_{\mathbf{X}} := E\left[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^{'}\right],$$

where the entry $c_{i,j}$ at the position (i,j) of $C_{\mathbf{X}}$ is

$$c_{i,j} \overset{\text{def}}{=} E[(X_i - \mu_i)(X_j - \mu_j)],$$

that is the covariance of X_i and X_j. Every covariance matrix, now designated by $\mathbf{C}$, is by construction symmetric

$$\mathbf{C} = \mathbf{C}^{'} \tag{8.3}$$

and nonnegative definite, i.e, for all $\mathbf{x} \in \mathbf{R^n}$

$$\mathbf{x}^{'}\mathbf{C}\mathbf{x} \geq 0. \tag{8.4}$$

It is shown on courses in linear algebra that nonnegative definiteness implies $\det \mathbf{C} \geq 0$. In terms of the entries $c_{i,j}$ of a covariance matrix $\mathbf{C} = (c_{i,j})_{i=1,j=1}^{n,n,}$ the preceding implies the following necessary properties.

1. $c_{i,j} = c_{j,i}$ (symmetry).
2. $c_{i,i} = \text{Var}\,(X_i) = \sigma_i^2 \geq 0$ (the elements in the main diagonal are the variances, and thus all elements in the main diagonal are nonnegative).
3. $c_{i,j}^2 \leq c_{i,i} \cdot c_{j,j}$ (Cauchy-Schwartz' inequality, c.f., (7.5)). Note that this yields another proof of the fact that the absolute value of a coefficient of correlation is ≤ 1.

Example 8.1.1 The covariance matrix of a bivariate random variable $\mathbf{X} = (X_1, X_2)^{'}$ is often written in the following form

$$C = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \tag{8.5}$$

where $\sigma_1^2 = \text{Var}\,(X_1)$, $\sigma_2^2 = \text{Var}\,(X_2)$ and $\rho = \text{Cov}(X,Y)/(\sigma_1\sigma_2)$ is the coefficient of correlation of X_1 and X_2. C is invertible ($\Rightarrow$ positive definite) if and only if $\rho^2 \neq 1$.

■

Linear transformations of random vectors are Borel functions $\mathbf{R}^n \mapsto \mathbf{R}^m$ of random vectors. The rules for finding the mean vector and the covariance matrix of a transformed vector are simple.

Proposition 8.1.2 $\mathbf{X}$ is a random vector with mean vector $\mu_{\mathbf{X}}$ and covariance matrix $C_{\mathbf{X}}$. B is a $m \times n$ matrix. If $\mathbf{Y} = B\mathbf{X} + \mathbf{b}$, then

$$E\mathbf{Y} = B\mu_{\mathbf{X}} + \mathbf{b} \tag{8.6}$$

$$C_{\mathbf{Y}} = BC_{\mathbf{X}}B^{'}. \tag{8.7}$$

Proof For simplicity of writing, take $\mathbf{b} = \mu = \mathbf{0}$. Then

$$C_{\mathbf{Y}} = E\mathbf{Y}\mathbf{Y}^{'} = EB\mathbf{X}\,(B\mathbf{X})^{'} =$$

$$= EB\mathbf{X}\mathbf{X}^{'}B^{'} = BE\left[\mathbf{X}\mathbf{X}^{'}\right]B^{'} = BC_{\mathbf{X}}B^{'}.$$

■

We have

Definition 8.1.1

$$\phi_{\mathbf{X}}\,(\mathbf{s}) \overset{\text{def}}{=} E\left[e^{i\mathbf{s}^{'}\mathbf{X}}\right] = \int_{\mathbf{R}^n} e^{i\mathbf{s}^{'}\mathbf{x}} dF_{\mathbf{X}}\,(\mathbf{x}) \tag{8.8}$$

is the **characteristic function** of the random vector $\mathbf{X}$.

■

In (8.8) $\mathbf{s}^{'}\mathbf{x}$ is a scalar product in $\mathbf{R}^n$,

$$\mathbf{s}^{'}\mathbf{x} = \sum_{i=1}^{n} s_i x_i.$$

As $F_{\mathbf{X}}$ is a joint distribution function on $\mathbf{R}^n$ and $\int_{\mathbf{R}^n}$ is a notation for a multiple integral over $\mathbf{R}^n$, we know that

$$\int_{\mathbf{R}^n} dF_{\mathbf{X}}\left(\mathbf{x}\right) = 1,$$

which means that $\phi_{\mathbf{X}}\left(\mathbf{0}\right) = 1$, where $\mathbf{0}$ is a $n \times 1$ -vector of zeros.

Theorem 8.1.3 *[***Kac's theorem**] $\mathbf{X} = \left(X_1, X_2, \cdots, X_n\right)^{'}$. The components $X_1, X_2, \cdots, X_n$ are independent if and only if

$$\phi_{\mathbf{X}}\left(\mathbf{s}\right) = E\left[e^{i\mathbf{s}^{'}\mathbf{X}}\right] = \prod_{i=1}^{n} \phi_{X_i}(s_i),$$

where $\phi_{X_i}(s_i)$ is the characteristic function for X_i.

Proof Assume that $\mathbf{X} = \left(X_1, X_2, \cdots, X_n\right)^{'}$ is a vector with independent X_i, $i = 1, \ldots, n$, that have, for convenience of writing, the joint p.d.f. $f_{\mathbf{X}}\left(\mathbf{x}\right)$. We have in (8.8)

$$\begin{aligned} \phi_{\mathbf{X}}\left(\mathbf{s}\right) &= \int_{\mathbf{R}^n} e^{i\mathbf{s}^{'}\mathbf{x}} f_{\mathbf{X}}\left(\mathbf{x}\right) d\mathbf{x} \\ &= \int_{\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{i(s_1 x_1 + \ldots + s_n x_n)} \prod_{i=1}^{n} f_{X_i}\left(x_i\right) dx_1 \cdots dx_n \\ &= \int_{\infty}^{\infty} e^{is_1 x_1} f_{X_1}\left(x_1\right) dx_1 \cdots \int_{-\infty}^{\infty} e^{is_n x_n} f_{X_n}\left(x_n\right) dx_n = \phi_{X_1}(s_1) \cdots \phi_{X_n}(s_n), \end{aligned} \tag{8.9}$$

where $\phi_{X_i}(s_i)$ is the characteristic function for X_i.
The more complicated proof of the assertion in the other direction is found in [61, pp.155−156]. ■

8.1.3 Multivariate Normal/Gaussian Distribution

Definition 8.1.2 $\mathbf{X}$ has a multivariate normal or Gaussian distribution with mean vector μ and covariance matrix $\mathbf{C}$, written as $\mathbf{X} \in N\left(\mu, \mathbf{C}\right)$, if and only if the characteristic function is given as

$$\phi_{\mathbf{X}}\left(\mathbf{s}\right) = e^{i\mathbf{s}^{'}\mu - \frac{1}{2}\mathbf{s}^{'}\mathbf{C}\mathbf{s}}. \tag{8.10}$$

■

The next statement is a manifestation of the **Cramér-Wold theorem**[1] or the **Cramér-Wold device**, [67, p. 87], which states that a probability measure on $\left(\mathbf{R}^n, \mathcal{B}\left(\mathbf{R}^n\right)\right)$ is uniquely determined by the totality of its one-dimensional projections. Seen from this angle a multivariate normal distribution is characterized by the totality of its one dimensional linear projections.

[1] Hermann Wold, 1908 - 1992, was a doctoral student of Harald Cramér, then Professor of statistics at Uppsala University and later at Gothenburg University `http://en.wikipedia.org/wiki/Herman_Wold`

Theorem 8.1.4 $\mathbf{X}$ has a multivariate normal distribution $N(\mu, \mathbf{C})$ if and only of

$$\mathbf{a}^{'}\mathbf{X} = \sum_{i=1}^{n} a_i X_i \tag{8.11}$$

has a normal distribution for **all** vectors $\mathbf{a}^{'} = (a_1, a_2, \ldots, a_n)$.

Proof Assume that $\mathbf{a}^{'}\mathbf{X}$ has a multivariate normal distribution for all $\mathbf{a}$ and that μ and $\mathbf{C}$ are the mean vector and covariance matrix of $\mathbf{X}$, respectively. Here (8.6) and (8.7) with $B = \mathbf{a}^{'}$ give

$$Ea^{'}\mathbf{X} = \mathbf{a}^{'}\mu, \text{Var}\left[\mathbf{a}^{'}\mathbf{X}\right] = \mathbf{a}^{'}\mathbf{C}\mathbf{a}.$$

Hence, if we set $Y = \mathbf{a}^{'}\mathbf{X}$, then by assumption $Y \in N\left(\mathbf{a}^{'}\mu, \mathbf{a}^{'}\mathbf{C}\mathbf{a}\right)$ and the characteristic function of Y is by (8.2)

$$\varphi_Y(t) = e^{it\mathbf{a}^{'}\mu - \frac{1}{2}t^2\mathbf{a}^{'}\mathbf{C}\mathbf{a}}.$$

The characteristic function of $\mathbf{X}$ is by definition

$$\varphi_{\mathbf{X}}(\mathbf{s}) = Ee^{i\mathbf{s}^{'}\mathbf{X}}.$$

Thus

$$\varphi_{\mathbf{X}}(\mathbf{a}) = Ee^{i\mathbf{a}^{'}\mathbf{X}} = \varphi_Y(1) = e^{i\mathbf{a}^{'}\mu - \frac{1}{2}\mathbf{a}^{'}\mathbf{C}\mathbf{a}}.$$

Thereby we have established that the characteristic function of $\mathbf{X}$ is

$$\varphi_{\mathbf{X}}(\mathbf{s}) = e^{i\mathbf{s}^{'}\mu - \frac{1}{2}\mathbf{s}^{'}\mathbf{C}\mathbf{s}}.$$

In view of definition 8.1.2 this shows that $\mathbf{X} \in N(\mu, \mathbf{C})$. The proof of the statement in the other direction is obvious. ■

Example 8.1.5 In this example we study a bivariate random variable $(X, Y)^{'}$ such that both X and Y have normal marginal distribution but there is a linear combination (in fact, $X + Y$), which does not have a normal distribution. Therefore $(X, Y)^{'}$ is not a bivariate normal random variable. This is an exercise stated in [80].
Let $X \in N\left(0, \sigma^2\right)$. Let $U \in \text{Be}\left(\frac{1}{2}\right)$ and be independent of X. Define

$$Y = \begin{cases} X & \text{if } U = 0 \\ -X & \text{if } U = 1. \end{cases}$$

Let us find the distribution of Y. We compute the characteristic function by double expectation

$$\begin{aligned}\varphi_Y(t) &= E\left[e^{itY}\right] = E\left[E\left[e^{itY} \mid U\right]\right] \\ &= E\left[e^{itY} \mid U = 0\right] \cdot \frac{1}{2} + E\left[e^{itY} \mid U = 1\right] \cdot \frac{1}{2} \\ &= E\left[e^{itX} \mid U = 0\right] \cdot \frac{1}{2} + E\left[e^{-itX} \mid U = 1\right] \cdot \frac{1}{2}\end{aligned}$$

and since X and U are independent, the independent condition drops out, and $X \in N\left(0, \sigma^2\right)$,

$$= E\left[e^{itX}\right] \cdot \frac{1}{2} + E\left[e^{-itX}\right] \cdot \frac{1}{2} = \frac{1}{2} \cdot e^{-\frac{t^2\sigma^2}{2}} + \frac{1}{2} \cdot e^{-\frac{t^2\sigma^2}{2}} = e^{-\frac{t^2\sigma^2}{2}},$$

which by uniqueness of characteristic functions says that $Y \in N\left(0, \sigma^2\right)$. Hence both marginal distributions of the bivariate random variable (X, Y) are normal distributions. Yet, the sum

$$X + Y = \begin{cases} 2X & \text{if } U = 0 \\ 0 & \text{if } U = 1 \end{cases}$$

is *not* a normal random variable. Hence (X, Y) is according to theorem 8.1.4 not a bivariate Gaussian random variable. Clearly we have

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & (-1)^U \end{pmatrix} \begin{pmatrix} X \\ X \end{pmatrix}. \tag{8.12}$$

Hence we multiply $(X, X)^{'}$ once by a random matrix to get $(X, Y)^{'}$ and therefore should not expect $(X, Y)^{'}$ to have a joint Gaussian distribution. We take next a look at the details. If $U = 1$, then

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} X \\ X \end{pmatrix} = A_1 \begin{pmatrix} X \\ X \end{pmatrix}$$

and if $U = 0$,

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ X \end{pmatrix} = A_0 \begin{pmatrix} X \\ X \end{pmatrix}.$$

The covariance matrix of $(X, X)^{'}$ is clearly

$$\mathbf{C}_X = \sigma^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

We set

$$\mathbf{C}_1 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad \mathbf{C}_0 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

One can verify, c.f. (8.7), that $\sigma^2 \mathbf{C}_1 = A_1 \mathbf{C}_X A_1^{'}$ and $\sigma^2 \mathbf{C}_0 = A_0 \mathbf{C}_X A_0^{'}$. Hence $\sigma^2 \mathbf{C}_1$ is the covariance matrix of (X, Y), if $U = 1$, and $\sigma^2 \mathbf{C}_0$ is the covariance matrix of (X, Y), if $U = 0$.

It is clear by the above that the joint distribution $F_{X,Y}$ should actually be a *mixture* of two distributions $F^{(1)}_{X,Y}$ and $F^{(0)}_{X,Y}$ with mixture coefficients $\left(\frac{1}{2}, \frac{1}{2}\right)$,

$$F_{X,Y}(x, y) = \frac{1}{2} \cdot F^{(1)}_{X,Y}(x, y) + \frac{1}{2} \cdot F^{(0)}_{X,Y}(x, y).$$

We understand this as follows. We draw first a value u from $\text{Be}\left(\frac{1}{2}\right)$, which points out one of the distributions, $F^{(u)}_{X,Y}$, and then draw a sample of (X, Y) from $F^{(u)}_{X,Y}$. We can explore these facts further.

> Let us determine the joint distribution of $(X, Y)^{'}$ by means of the joint characteristic function, see eq.(8.8). We get
>
> $$\begin{aligned} \varphi_{X,Y}(t, s) = E\left[e^{i(tX+sY)}\right] &= E\left[e^{i(tX+sY)} \mid U = 0\right] \cdot \frac{1}{2} + E\left[e^{i(tX+sY)} \mid U = 1\right] \cdot \frac{1}{2} \\ &= E\left[e^{i(t+s)X)}\right] \cdot \frac{1}{2} + E\left[e^{i(t-s)X}\right] \cdot \frac{1}{2} \\ &= \frac{1}{2} e^{-\frac{(t+s)^2\sigma^2}{2}} + \frac{1}{2} e^{-\frac{(t-s)^2\sigma^2}{2}}. \end{aligned}$$
>
> From the above
>
> $$(t-s)^2 = (t, s)\mathbf{C}_1 \begin{pmatrix} t \\ s \end{pmatrix} \quad (t+s)^2 = (t, s)\mathbf{C}_0 \begin{pmatrix} t \\ s \end{pmatrix}.$$

We see that $\mathbf{C}_1$ and $\mathbf{C}_2$ are non-negative definite matrices. (It holds also that $\det \mathbf{C}_1 = \det \mathbf{C}_0 = 0$.) Therefore

$$\frac{1}{2}e^{-\frac{(t+s)^2\sigma^2}{2}} + \frac{1}{2}e^{-\frac{(t-s)^2\sigma^2}{2}} = \frac{1}{2}e^{-\frac{\sigma^2\mathbf{s}^{'}\mathbf{C}_0\mathbf{s}}{2}} + \frac{1}{2}e^{-\frac{\sigma^2\mathbf{s}^{'}\mathbf{C}_1\mathbf{s}}{2}},$$

where $\mathbf{s} = (t,s)^{'}$. This shows by uniqueness of characteristic functions that the joint distribution of (X,Y) is a mixture of $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2\mathbf{C}_0\right)$ and $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2\mathbf{C}_1\right)$ with the mixture coefficients $\left(\frac{1}{2},\frac{1}{2}\right)$.

■

Additional properties are:

1. **Theorem 8.1.6** If $\mathbf{Y} = B\mathbf{X} + \mathbf{b}$, and $\mathbf{X} \in N(\mu, \mathbf{C})$, then

$$\mathbf{Y} \in N\left(B\mu + \mathbf{b}, B\mathbf{C}B^{'}\right).$$

Proof We check the characteristic function of $\mathbf{Y}$; some linear algebra gives

$$\varphi_{\mathbf{Y}}(\mathbf{s}) = E\left[e^{i\mathbf{s}^{'}\mathbf{Y}}\right] = E\left[e^{i\mathbf{s}^{'}(\mathbf{b}+B\mathbf{X})}\right] =$$

$$= e^{i\mathbf{s}^{'}\mathbf{b}}E\left[e^{i\mathbf{s}^{'}B\mathbf{X}}\right] = e^{i\mathbf{s}^{'}\mathbf{b}}E\left[e^{i\left(B^{'}\mathbf{s}\right)^{'}\mathbf{X}}\right]$$

or

$$\varphi_{\mathbf{Y}}(\mathbf{s}) = e^{i\mathbf{s}^{'}\mathbf{b}}E\left[e^{i\left(B^{'}\mathbf{s}\right)^{'}\mathbf{X}}\right]. \tag{8.13}$$

Here

$$E\left[e^{i\left(B^{'}\mathbf{s}\right)^{'}\mathbf{X}}\right] = \varphi_{\mathbf{X}}\left(B^{'}\mathbf{s}\right).$$

Furthermore

$$\varphi_{\mathbf{X}}\left(B^{'}\mathbf{s}\right) = e^{i\left(B^{'}\mathbf{s}\right)^{'}\mu - \frac{1}{2}\left(B^{'}\mathbf{s}\right)^{'}\mathbf{C}\left(B^{'}\mathbf{s}\right)}.$$

Since

$$\left(B^{'}\mathbf{s}\right)^{'}\mu = \mathbf{s}^{'}B\mu, \quad \left(B^{'}\mathbf{s}\right)^{'}\mathbf{C}\left(B^{'}\mathbf{s}\right) = \mathbf{s}^{'}B\mathbf{C}B^{'}\mathbf{s},$$

we get

$$e^{i\left(B^{'}\mathbf{s}\right)^{'}\mu - \frac{1}{2}\left(B^{'}\mathbf{s}\right)^{'}\mathbf{C}\left(B^{'}\mathbf{s}\right)} = e^{i\mathbf{s}^{'}B\mu - \frac{1}{2}\mathbf{s}^{'}B\mathbf{C}B^{'}\mathbf{s}}.$$

Therefore

$$\varphi_{\mathbf{X}}\left(B^{'}\mathbf{s}\right) = e^{i\mathbf{s}^{'}B\mu - \frac{1}{2}\mathbf{s}^{'}B\mathbf{C}B^{'}\mathbf{s}} \tag{8.14}$$

and by (8.14) and (8.13) above we get

$$\varphi_{\mathbf{Y}}(\mathbf{s}) = e^{i\mathbf{s}^{'}\mathbf{b}}\varphi_{\mathbf{X}}\left(B^{'}\mathbf{s}\right) = e^{i\mathbf{s}^{'}\mathbf{b}}e^{i\mathbf{s}^{'}B\mu - \frac{1}{2}\mathbf{s}^{'}B\mathbf{C}B^{'}\mathbf{s}}$$

$$= e^{i\mathbf{s}^{'}(\mathbf{b}+B\mu) - \frac{1}{2}\mathbf{s}^{'}B\mathbf{C}B^{'}\mathbf{s}},$$

which by uniqueness of characteristic functions proves the claim as asserted. ■

2. **Theorem 8.1.7** A Gaussian multivariate random variable has independent components if and only if the covariance matrix is diagonal.

   Proof Let $\mathbf{\Lambda}$ be a diagonal covariance matrix with λ_is on the main diagonal, i.e.,

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ 0 & 0 & \lambda_3 & \ldots & 0 \\ 0 & \ddots & \vdots & \ldots & 0 \\ 0 & 0 & 0 & \ldots & \lambda_n \end{pmatrix}.$$

   Then

$$\varphi_{\mathbf{X}}(\mathbf{t}) = e^{i\mathbf{t}^{'}\mu - \frac{1}{2}\mathbf{t}^{'}\mathbf{\Lambda}\mathbf{t}} =$$
$$= e^{i\sum_{i=1}^{n}\mu_i t_i - \frac{1}{2}\sum_{i=1}^{n}\lambda_i t_i^2}$$
$$= e^{i\mu_1 t_1 - \frac{1}{2}\lambda_1 t_1^2} e^{i\mu_2 t_2 - \frac{1}{2}\lambda_2 t_2^2} \cdots e^{i\mu_n t_n - \frac{1}{2}\lambda_n t_n^2}$$

   is the product of the characteristic functions of $X_i \in N(\mu_i, \lambda_i)$, which are by theorem 8.1.3 seen to be independent. ■

3. **Theorem 8.1.8** If $\mathbf{C}$ is positive definite ($\Rightarrow \det \mathbf{C} > 0$), then it can be shown that there is a simultaneous p.d.f. of the form

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \mathbf{C}}} e^{-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{X}})^{'}\mathbf{C}^{-1}(\mathbf{x}-\mu_{\mathbf{X}})}. \tag{8.15}$$

   Proof It can be checked by a lengthy but straightforward computation that

$$e^{i\mathbf{s}^{'}\mu - \frac{1}{2}\mathbf{s}^{'}\mathbf{C}\mathbf{s}} = \int_{\mathbf{R}^n} e^{i\mathbf{s}^{'}\mathbf{x}} \frac{1}{(2\pi)^{n/2}\sqrt{\det(\mathbf{C})}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^{'}\mathbf{C}^{-1}(\mathbf{x}-\mu)} d\mathbf{X}.$$

   ■

4. **Theorem 8.1.9** $(X_1, X_2)^{'}$ is a bivariate Gaussian random variable. The conditional distribution for X_2 given $X_1 = x_1$ is

$$N\left(\mu_2 + \rho \cdot \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1-\rho^2)\right), \tag{8.16}$$

   where $\mu_2 = E(X_2)$, $\mu_1 = E(X_2)$, $\sigma_2 = \sqrt{\mathrm{Var}(X_2)}$, $\sigma_1 = \sqrt{\mathrm{Var}(X_1)}$ and $\rho = \mathrm{Cov}(X_1, X_2)/(\sigma_1 \cdot \sigma_2)$.

   Proof is done by an explicit evaluation of (8.15) followed by an explicit evaluation of the pertinent conditional p.d.f. and is deferred to Appendix 8.4. ■

   Hence for bivariate Gaussian variables the **best estimator in the mean square sense**, $E[X_2 \mid X_1]$, and the **best *linear* estimator in the mean square sense** are one and the same random variable, c.f., example 7.5.5 and remark 7.6.1.

Definition 8.1.3 $\mathbf{Z} \in N(\mathbf{0}, I)$ is a standard Gaussian vector, where I is $n \times n$ identity matrix.

■

Let $\mathbf{X} \in N(\mu_{\mathbf{X}}, \mathbf{C})$. Then, if $\mathbf{C}$ is positive definite, we can factorize $\mathbf{C}$ as

$$\mathbf{C} = AA^{'},$$

for $n \times n$ matrix A, where A is lower triangular, see [80, Appendix 1]. Actually we can always decompose

$$\mathbf{C} = LDL',$$

where L is a unique $n \times n$ lower triangular, D is diagonal with positive elements on the main diagonal, and we write $A = L\sqrt{D}$. Then A^{-1} is lower triangular. Then

$$\mathbf{Z} = A^{-1}\left(\mathbf{X} - \mu_{\mathbf{X}}\right)$$

is a standard Gaussian vector. In some applications, like, e.g., in time series analysis and signal processing, one refers to A^{-1} as a **whitening** matrix. It can be shown that A^{-1} is lower triangular, thus we have obtained $\mathbf{Z}$ by a **causal** operation, in the sense that Z_i is a function of $X_1, \ldots, X_i$. $\mathbf{Z}$ is known as the **innovations** of $\mathbf{X}$. Conversely, one goes from the innovations to $\mathbf{X}$ through another causal operation by $\mathbf{X} = A\mathbf{Z} + \mathbf{b}$, and then

$$\mathbf{X} = N\left(\mathbf{b}, AA^{'}\right).$$

Example 8.1.10 (Factorization of a 2×2 Covariance Matrix) Let

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \in N\left(\mu, \mathbf{C}\right).$$

Let Z_1 och Z_2 be independent $N(0,1)$. We consider the lower triangular matrix

$$\mathbf{B} = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{pmatrix}, \tag{8.17}$$

which clearly has an inverse, as soon as $\rho \neq \pm 1$. Moreover, one verifies that $\mathbf{C} = \mathbf{B} \cdot \mathbf{B}^{'}$, when we write $\mathbf{C}$ as in (8.5). Then we get

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \mu + \mathbf{B}\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}, \tag{8.18}$$

where, of course,

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \in N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

■

8.2 Partitioned Covariance Matrices

Assume that $\mathbf{X}$, $n \times 1$, is **partitioned** as

$$\mathbf{X} = \left(\mathbf{X}_1, \mathbf{X}_2\right)^{'},$$

where $\mathbf{X}_1$ is $p \times 1$ and $\mathbf{X}_2$ is $q \times 1$, $n = q + p$. Let the covariance matrix $\mathbf{C}$ be **partitioned** in the sense that

$$\mathbf{C} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{8.19}$$

where Σ_{11} is $p \times p$, Σ_{22} is $q \times q$ e.t.c.. The mean is partitioned correspondingly as

$$\mu := \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}. \tag{8.20}$$

Let $\mathbf{X} \in N_n(\mu, \mathbf{C})$, where N_n refers to a normal distribution in n variables, $\mathbf{C}$ and μ are partitioned as in (8.19)-(8.20). Then the marginal distribution of $\mathbf{X}_2$ is

$$\mathbf{X}_2 \in N_q(\mu_2, \Sigma_{22}),$$

if Σ_{22} is invertible. Let $\mathbf{X} \in N_n(\mu, \mathbf{C})$, where $\mathbf{C}$ and μ are partitioned as in (8.19)-(8.20). Assume that the inverse Σ_{22}^{-1} exists. Then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$ is normal, or,

$$\mathbf{X}_1 \mid \mathbf{X}_2 = \mathbf{x}_2 \in N_p\left(\underline{\mu}_{1|2}, \Sigma_{1|2}\right), \tag{8.21}$$

where

$$\underline{\mu}_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2) \tag{8.22}$$

and

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

By virtue of (8.21) and (8.22) the **best estimator in the mean square sense** and the **best *linear* estimator in the mean square sense** are one and the same random variable .

8.3 Appendix: Symmetric Matrices & Orthogonal Diagonalization & Gaussian Vectors

We quote some results from [5, chapter 7.2] or, from any textbook in linear algebra. An $n \times n$ matrix $\mathbf{A}$ is **orthogonally diagonalizable**, if there is an orthogonal matrix $\mathbf{P}$ (i.e., $\mathbf{P}'\mathbf{P} = \mathbf{P}\mathbf{P}' = \mathbf{I}$) such that

$$\mathbf{P}'\mathbf{A}\mathbf{P} = \mathbf{\Lambda},$$

where $\mathbf{\Lambda}$ is a diagonal matrix. Then we have

Theorem 8.3.1 If $\mathbf{A}$ is an $n \times n$ matrix, then the following are equivalent:

(i) $\mathbf{A}$ is orthogonally diagonalizable.

(ii) $\mathbf{A}$ has an orthonormal set of eigenvectors.

(iii) $\mathbf{A}$ is symmetric.

■

Since covariance matrices are symmetric, we have by the theorem above that **all covariance matrices are orthogonally diagonalizable**.

Theorem 8.3.2 If $\mathbf{A}$ is a symmetric matrix, then

(i) Eigenvalues of $\mathbf{A}$ are all real numbers.

(ii) Eigenvectors from different eigenspaces are orthogonal.

■

That is, **all eigenvalues of a covariance matrix are real**. Hence we have for any covariance matrix the **spectral decomposition**

$$\mathbf{C} = \sum_{i=1}^{n} \lambda_i e_i e_i', \tag{8.23}$$

where $\mathbf{C}e_i = \lambda_i e_i$. Since $\mathbf{C}$ is nonnegative definite, and its eigenvectors are orthonormal,

$$0 \leq e_i^{'}\mathbf{C}e_i = \lambda_i e_i^{'} e_i = \lambda_i,$$

and thus **the eigenvalues of a covariance matrix are nonnegative**.
Let now $\mathbf{P}$ be an orthogonal matrix such that

$$\mathbf{P}^{'}\mathbf{C_X}\mathbf{P} = \mathbf{\Lambda},$$

and $\mathbf{X} \in N(\mathbf{0}, \mathbf{C_X})$, i.e., $\mathbf{C_X}$ is a covariance matrix and $\mathbf{\Lambda}$ is diagonal (with the eigenvalues of $\mathbf{C_X}$ on the main diagonal). Then if $\mathbf{Y} = \mathbf{P}^{'}\mathbf{X}$, we have by theorem 8.1.6 that

$$\mathbf{Y} \in N(\mathbf{0}, \mathbf{\Lambda}).$$

In other words, $\mathbf{Y}$ is a Gaussian vector and has by theorem 8.1.7 independent components. This method of producing independent Gaussians has several important applications. One of these is the **principal component analysis**, c.f. [59, p. 74]. In addition, the operation is invertible, as

$$\mathbf{X} = \mathbf{P}\mathbf{Y}$$

recreates $\mathbf{X} \in N(\mathbf{0}, \mathbf{C_X})$ from $\mathbf{Y}$.

8.4 Appendix: Proof of (8.16)

Let $\mathbf{X} = (X_1, X_2)^{'} \in N(\mu_{\mathbf{X}}, C)$, $\mu_{\mathbf{X}} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and C in (8.5) with $\rho^2 \neq 1$. The inverse of C in (8.5) is

$$C^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}.$$

Then we get by straightforward evaluation in (8.15)

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \frac{1}{2\pi\sqrt{\det C}} e^{-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{X}})^{'}C^{-1}(\mathbf{x}-\mu_{\mathbf{X}})} \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}Q(x_1,x_2)}, \end{aligned} \tag{8.24}$$

where

$$Q(x_1, x_2) =$$
$$\frac{1}{(1-\rho^2)} \cdot \left[\left(\frac{x_1-\mu_1}{\sigma_1} \right)^2 - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2} \right)^2 \right].$$

Now we claim that

$$f_{X_2|X_1=x_1}(x_2) = \frac{1}{\tilde{\sigma}_2\sqrt{2\pi}} e^{-\frac{1}{2\tilde{\sigma}_2^2}(x_2-\tilde{\mu}_2(x_1))^2},$$

a p.d.f. of a Gaussian random variable $X_2|X_1 = x_1$ with the (conditional) expectation $\tilde{\mu}_2(x_1)$ and the (conditional) variance $\tilde{\sigma}_2$

$$\tilde{\mu}_2(x_1) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1), \tilde{\sigma}_2 = \sigma_2\sqrt{1-\rho^2}.$$

To prove these assertions about $f_{X_2|X_1=x_1}(x_2)$ we set

$$f_{X_1}(x_1) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2}, \tag{8.25}$$

and compute the ratio $\frac{f_{X_1,X_2}(x_1,x_2)}{f_X(x_1)}$. We get from the above by (8.24) and (8.25) that

$$\frac{f_{X_1,X_2}(x_1,x_2)}{f_X(x_1)} = \frac{\sigma_1\sqrt{2\pi}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-\frac{1}{2}Q(x_1,x_2)+\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2},$$

which we organize, for clarity, by introducing the auxiliary function $H(x_1,x_2)$ by

$$-\frac{1}{2}H(x_1,x_2) \overset{\text{def}}{=} -\frac{1}{2}Q(x_1,x_2) + \frac{1}{2\sigma_1^2}(x_1-\mu_1)^2.$$

Here we have

$$H(x_1,x_2) =$$

$$\frac{1}{(1-\rho^2)} \cdot \left[\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right] - \left(\frac{x_1-\mu_1}{\sigma_1}\right)^2$$

$$= \frac{\rho^2}{(1-\rho^2)}\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2(1-\rho^2)} + \frac{(x_2-\mu_2)^2}{\sigma_2^2(1-\rho^2)}.$$

Evidently we have now shown

$$H(x_1,x_2) = \frac{\left(x_2-\mu_2-\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)\right)^2}{\sigma_2^2(1-\rho^2)}.$$

Hence we have found that

$$\frac{f_{X_1,X_2}(x_1,x_2)}{f_X(x_1)} = \frac{1}{\sqrt{1-\rho^2}\sigma_2\sqrt{2\pi}}e^{-\frac{1}{2}\frac{\left(x_2-\mu_2-\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)\right)^2}{\sigma_2^2(1-\rho^2)}}.$$

This establishes the properties of bivariate normal random variables claimed in (8.16) above. ■

As an additional exercise on the use of (8.16) (and conditional expectation) we make the following check of correctness of our formulas.

Theorem 8.4.1 $\mathbf{X} = (X_1,X_2)^{'} \in N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, C\right) \Rightarrow \rho = \rho_{X_1,X_2}$.

Proof We compute by double expectation

$$E\left[(X_1-\mu_1)(X_2-\mu_2)\right] = E(E(\left[(X_1-\mu_1)(X_2-\mu_2)\right]|X_1)$$

and by taking out what is known,

$$= E((X_1-\mu_1)E\left[X_2-\mu_2\right]|X_1)) = E(X_1-\mu_1)\left[E(X_2|X_1)-\mu_2\right]$$

and by (8.16)

$$= E((X_1-\mu_1)\left[\mu_2+\rho\frac{\sigma_2}{\sigma_1}(X_1-\mu_1)-\mu_2\right]$$

$$= \rho\frac{\sigma_2}{\sigma_1}E(X_1-\mu_1)((X_1-\mu_1))$$

$$= \rho\frac{\sigma_2}{\sigma_1}E(X_1-\mu_1)^2 = \rho\frac{\sigma_2}{\sigma_1}\sigma_1^2 = \rho\sigma_2\sigma_1.$$

In other words, we have established that

$$\rho = \frac{E\left[(X_1-\mu_1)(X_2-\mu_2)\right]}{\sigma_2\sigma_1},$$

which says that ρ is the coefficient of correlation of $(X_1,X_2)^{'}$. ■

8.5 Exercises

8.5.1 Bivariate Gaussian Variables

1. (From [42]) Let $(X_1, X_2)^{'} \in N(\mu, \mathbf{C})$, where

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and

$$\mathbf{C} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

(a) Set $Y = X_1 - X_2$. Show that $Y \in N(0, 2-2\rho)$.

(b) Show that for any $\varepsilon > 0$

$$\mathbf{P}(|Y| \leq \varepsilon) \to 1,$$

if $\rho \uparrow 1$.

2. (From [42]) Let $(X_1, X_2)^{'} \in N(\mu, \mathbf{C})$, where

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and

$$\mathbf{C} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

(a) We want to find the distribution of the random variable $X_1 \mid X_2 \leq a$. Show that

$$\mathbf{P}(X_1 \leq x \mid X_2 \leq a) = \frac{1}{\Phi(a)} \int_{-\infty}^{x} \phi(u) \Phi\left(\frac{a - \rho u}{\sqrt{1-\rho^2}}\right) du, \tag{8.26}$$

where $\Phi(x)$ is the distribution function of $N(0,1)$ and $\phi(x)$ the p.d.f. of $N(0,1)$, i.e., $\frac{d}{dx}\Phi(x) = \phi(x)$. We sketch two different solutions.

Aid 1. We need to find

$$\mathbf{P}(X_1 \leq x \mid X_2 \leq a) = \frac{\mathbf{P}(\{X_1 \leq x\} \cap \{X_2 \leq a\})}{\mathbf{P}(X_2 \leq a)}.$$

Then

$$\mathbf{P}(\{X_1 \leq x\} \cap \{X_2 \leq a\}) = \int_{-\infty}^{x} \int_{-\infty}^{a} f_{X_1,X_2}(u,v) du dv =$$

$$= \int_{-\infty}^{x} f_{X_2}(v) \int_{-\infty}^{a} f_{X_1 \mid X_2 = v}(u) du dv.$$

Now find $f_{X_2}(v)$ and $f_{X_1 \mid X_2=v}(u)$ and make a change of variable in $\int_{-\infty}^{a} f_{X_1 \mid X_2 = v}(u) du$.

Aid 2. Use (8.18), which shows how to write $(X_1, X_2)^{'}$, as a linear transformation of $(Z_1, Z_2)^{'}$ with $N(\mathbf{0}, I)$, or as

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \mathbf{B} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}.$$

Then you can, since $\mathbf{B}$ is invertible, write the event

$$\{X_1 \leq x\} \cap \{X_2 \leq a\}$$

as an event using (the innovations) Z_1 and Z_2 and then compute the desired probability using the joint distribution of Z_1 and Z_2.

(b) Show using (8.26) that

$$\lim_{\rho \uparrow 1} \mathbf{P}\left(X_1 \leq x \mid X_2 \leq a\right) = \begin{cases} \frac{\Phi(x)}{\Phi(a)} & \text{if } x \leq a \\ 1 & \text{if } x > a. \end{cases}$$

3. (From [42]) Determine the constant c so that the function

$$c \cdot e^{-(x^2 - xy + y^2)}$$

becomes the p.d.f. of a bivariate normal distribution, and determine its parameters, that is, its mean vector and covariance matrix.
Answer: $c = \frac{\sqrt{3}}{2\pi}$, $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}\right)$.

4. (From [42]) $(X_1, X_2)' \in N(\mathbf{0}, \mathbf{C})$, where $\mathbf{0} = (0,0)'$.

(a) Show that

$$\text{Cov}\left(X_1^2, X_2^2\right) = 2\left(\text{Cov}\left(X_1, X_2\right)\right)^2 \tag{8.27}$$

(b) Find the mean vector and the covariance matrix of $\left(X_1^2, X_2^2\right)'$.

5. [**Estimation theory**] Let

$$\begin{pmatrix} X \\ Y \end{pmatrix} \in N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right).$$

As in section 3.7.3 we have the estimator

$$\widehat{Y} = E\left[Y \mid \mathcal{F}_X\right] = E\left[Y \mid X\right]$$

and the estimation error

$$\widetilde{Y} = Y - \widehat{Y}.$$

(a) Find $E\left(\widetilde{Y}\right)$, and show that

$$\text{Var}\left[\widetilde{Y}\right] = \sigma_Y^2(1 - \rho^2).$$

Aid: Recall theorem 3.7.3 and (8.16).

(b) What is the distribution of $\widetilde{Y}$?

6. **Rosenblatt Transformation for Bivariate Gaussian Variables** Let

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \in N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right).$$

Find the Rosenblatt transform (3.43) from (X_1, X_2) to (Z_1, Z_2), i.e.

$$z_1 = F_{X_1}(x_1) \tag{8.28}$$

$$z_2 = F_{X_2 \mid X_1 = x_1}(x_2). \tag{8.29}$$

Note that $Z_1 \in U(0,1)$ and $Z_2 \in U(0,1)$, and must be independent.

7. [**Estimation theory and Tower Property**] Let

$$\begin{pmatrix} X \\ Z \end{pmatrix} \in N\left(\begin{pmatrix} \mu_X \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right).$$

In addition we have for an interval $[a,b]$ and some $c \in [a,b]$

$$Y = \chi_A(X) = \begin{cases} 1 & a \leq X \leq b \\ 0 & X \notin [a,b]. \end{cases}$$

Suppose now that we want to estimate Z by means of $Y = 1$ and take

$$E\left[Z|Y=1\right]$$

as our chosen estimator.

(a) Show that

$$E\left[Z|Y=1\right] = \mu_Z + \rho \cdot \frac{\sigma_2}{\sigma_1}(H_1 - \mu_X),$$

where

$$H_1 = \frac{1}{\Phi\left(\frac{b-\mu_x}{\sigma_1}\right) - \Phi\left(\frac{a-\mu_x}{\sigma_1}\right)} \int_a^b x \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_X)^2}{2\sigma_1^2}} dx.$$

Aid: Start by recalling section 3.32 and the formula (3.7.4).

(b) Find

$$\mathrm{Var}\left[Z - E\left[Z|Y=1\right]\right].$$

8. In the mathematical theory of communication, see [23], (communication in the sense of transmission of messages via systems designed by electrical and computer engineers, not in the sense of social competence and human relations or human-computer interaction (HCI)) one introduces the **mutual information** $I(X,Y)$ between two continuous random variables X and Y by

$$I(X,Y) \overset{\text{def}}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} dxdy, \tag{8.30}$$

where $f_{X,Y}(x,y)$ is the joint p.d.f. of (X,Y), $f_X(x)$ and $f_Y(y)$ are the marginal p.d.f.s of X and Y, respectively. $I(X,Y)$ is in fact a measure of dependence between random variables, and is theoretically speaking superior to correlation, as we measure with $I(X,Y)$ more than the mere degree of linear dependence between X and Y.

Assume now that $(X,Y) \in N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix} \right)$. Check that

$$I(X,Y) = -\frac{1}{2}\log\left(1-\rho^2\right). \tag{8.31}$$

Aid: The following steps solution are in a sense instructive, as they rely on the explicit conditional distribution of $Y \mid X = x$, and provide an interesting decomposition of $I(X,Y)$ as an intermediate step. Someone may prefer other ways. Use

$$\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)} = \frac{f_{Y|X=x}(y)}{f_Y(y)},$$

and then

$$I(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y) \log f_{Y|X=x}(y) dxdy$$

$$-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\log f_Y(y)dxdy.$$

Then one inserts in the first term on the right hand side

$$f_{X,Y}(x,y) = f_{Y|X=x}(y)\cdot f_X(x).$$

Observe that the conditional distribution of $Y \mid X = x$ is here

$$N\left(\rho x, \sigma^2(1-\rho^2)\right),$$

and take into account the marginal distributions of X and Y.

Interpret the result in (8.31) by considering $\rho = 0$, $\rho = \pm 1$. Note also that $I(X,Y) \geq 0$.

9. (From [101]) The matrix

$$\mathbf{Q} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \tag{8.32}$$

is known as the *rotation matrix*[2]. Let

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \in N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}\right)$$

and let

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \mathbf{Q}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

and $\sigma_2^2 \geq \sigma_1^2$.

(i) Find $\mathrm{Cov}(Y_1, Y_2)$ and show that Y_1 and Y_2 are independent for all θ if and only if $\sigma_2^2 = \sigma_1^2$.

(ii) Supppose $\sigma_2^2 > \sigma_1^2$. For which values of θ are Y_1 and Y_2 are independent ?

10. (From [101]) Let

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \in N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix}\right).$$

Set

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \mathbf{Q}\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix},$$

where $\mathbf{Q}$ is the rotation matrix (8.32) with $\theta = \frac{\pi}{4}$. Show that

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \in N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Hence we see that by rotating two independent Gaussian variables with variances $1+\rho$ and $1-\rho$, $\rho \neq 0$, with 45 degrees, we get a bivariate Gaussian vector, where covariance of the two variables is equal to ρ.

11. (X,Y) is a bivariate Gaussian r.v. with $\mathrm{Var}[X] = \mathrm{Var}[Y]$. Show that $X+Y$ and $X-Y$ are independent r.v.'s.

12. Let

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \in N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right).$$

Show that $\mathrm{Var}[X_1X_2] = \sigma_1^2\sigma_2^2\left(1+\rho^2\right)$.

[2] $\mathbf{y} = \mathbf{Q}\mathbf{x}$ is a rotation of $\mathbf{x}$ by the angle θ, as explained in any text on linear algebra, see, e.g., [5, p.187].

13. $X \in N(0,1)$, $Y \in N(0,1)$, and are independent.

    (a) Show that $E[X \mid X > Y] = \frac{1}{\sqrt{\pi}}$.

    (b) Show that $E[X + Y \mid X > Y] = 0$.

14. $X \in N(0,\sigma^2)$, $Y \in N(0,\sigma^2)$, and are independent. Show that

$$\frac{X - Y}{X + Y} \in C(0,1).$$

Aid: Recall the exercise 2.6.3.4..

15. X_1, X_2, X_3 are independent and $\in N(1,1)$. We set

$$U = X_1 + X_2 + X_3, V = X_1 + 2X_2 + 3X_3.$$

Determine $V \mid U = 3$. *Answer:* $N(6,2)$.

16. $\mathbf{X} = (X_1, X_2, X_3)^{'}$ has the mean vector $\mu = (0,0,0)^{'}$ and the covariance matrix

$$\mathbf{C} = \begin{pmatrix} 3 & -2 & 1 \\ -2 & 2 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

Find the distribution of $X_1 + X_3$ given that

   (a) $X_2 = 0$.

   (b) $X_2 = 2$.

Answers: (a) $N(0,4)$, (b) $N(-2,4)$.

17. $\mathbf{X} = (X_1, X_2)^{'}$ has the mean vector $\mu = (0,0)^{'}$ and the covariance matrix

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

Find the distribution of the random variable

$$\frac{X_1^2 - 2\rho X_1 X_2 + X_2^2}{1 - \rho^2}$$

by computing its m.g.f.. *Answer:* $\chi^2(2)$.

18. **Return to the Mean**

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \in N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right).$$

Check that

$$\frac{E[X_2 \mid X_1] - E[X_2]}{\sigma_2} = \rho \frac{X_1 - E[X_1]}{\sigma_1}. \tag{8.33}$$

This equality provides a strict mathematical expression for an important statistical phenomenon, namely return to the mean, or **regression**, as discovered by Francis Galton[3].

> Assume that $|\rho| < 1$. Then (8.33) tells us that the standardized distance between $E[X_2 \mid X_1]$ and its mean $E[X_2]$ is smaller than than the standardized distance between X_1 and its mean $E[X_1]$. Here we think of X_1 and X_2 as a first and second measurement, respectively, of some property, like the height of a parent and the height of an adult child of that parent.

[3] Sir Francis Galton, 1822 −1911, contributed to statistics, sociology, psychology, anthropology, geography, meteorology, genetics and psychometry, was active as tropical explorer and inventor, and one of the first proponents of eugenics.

8.5.2 Covariance Matrices & The Four Product Rule

1. $\mathbf{C}$ is a positive definite covariance matrix. Show that $\mathbf{C}^{-1}$ is a covariance matrix.

2. $\mathbf{C}_1$ and $\mathbf{C}_2$ are two $n \times n$ covariance matrices. Show that

   (a) $\mathbf{C}_1 + \mathbf{C}_2$ is a covariance matrix.

   (b) $\mathbf{C}_1 \cdot \mathbf{C}_2$ is a covariance matrix.

Aid: The symmetry of $\mathbf{C}_2 \cdot \mathbf{C}_1$ is immediate. The difficulty is to show that $\mathbf{C}_1 \cdot \mathbf{C}_2$ is nonnegative definite. We need a piece of linear algebra here, c.f. appendix 8.3. Any symmetric and nonnegative definite matrix can written using the **spectral decomposition**, see (8.23),

$$\mathbf{C} = \sum_{i=1}^{n} \lambda_i e_i e_i^{'},$$

where e_i is a real (i.e., has no complex numbers as elements) $n \times 1$ eigenvector, i.e., $\mathbf{C}e_i = \lambda_i e_i$ and $\lambda_i \geq 0$. The set $\{e_i\}_{i=1}^{n}$ is a complete orthonormal basis in $\mathbf{R}^n$, which amongst other things implies that every $\mathbf{x} \in \mathbf{R}^n$ can be written as

$$\mathbf{x} = \sum_{i=1}^{n} (\mathbf{x}^{'} e_i) e_i,$$

where the number $\mathbf{x}^{'} e_i$ is the coordinate of $\mathbf{x}$ w.r.t. the basis vector e_i. In addition, orthonormality is recalled as the property

$$e_j^{'} e_i = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases} \tag{8.34}$$

We make initially the *simplifying assumption* that $\mathbf{C}_1$ and $\mathbf{C}_2$ have the same eigenvectors, so that $\mathbf{C}_1 e_i = \lambda_i e_i$, $\mathbf{C}_2 e_i = \mu_i e_i$. Then we can diagonalize the quadratic form $\mathbf{x}^{'} \mathbf{C}_2 \mathbf{C}_1 \mathbf{x}$ as follows.

$$\begin{aligned} \mathbf{C}_1 \mathbf{x} &= \sum_{i=1}^{n} (\mathbf{x}^{'} e_i) \mathbf{C}_1 e_i = \sum_{i=1}^{n} \lambda_i (\mathbf{x}^{'} e_i) e_i \\ &= \sum_{i=1}^{n} \lambda_i (\mathbf{x}^{'} e_i) e_i. \end{aligned} \tag{8.35}$$

Also, since $\mathbf{C}_2$ is symmetric

$$\mathbf{x}^{'} \mathbf{C}_2 = (\mathbf{C}_2 \mathbf{x})^{'} = \left(\sum_{j=1}^{n} (\mathbf{x}^{'} e_j) \mathbf{C}_2 e_j \right)^{'}$$

or

$$\mathbf{x}^{'} \mathbf{C}_2 = \sum_{j=1}^{n} \mu_j (\mathbf{x}^{'} e_j) e_j^{'}. \tag{8.36}$$

Then for any $\mathbf{x} \in \mathbf{R}^n$ we get from (8.35) and (8.36) that

$$\mathbf{x}^{'} \mathbf{C}_2 \mathbf{C}_1 \mathbf{x} = \sum_{j=1}^{n} \sum_{i=1}^{n} \mu_j \lambda_i (\mathbf{x}^{'} e_j)(\mathbf{x}^{'} e_i) e_j^{'} e_i$$

and because of (8.34)

$$= \sum_{i=1}^{n} \mu_j \lambda_i \left(\mathbf{x}^{'} e_i\right)^2 .$$

But since $\mu_j \geq 0$ and $\lambda_i \geq 0$, we see that

$$\sum_{i=1}^{n} \mu_j \lambda_i \left(\mathbf{x}^{'} e_i\right)^2 \geq 0,$$

or, for any $\mathbf{x} \in \mathbf{R}^n$,

$$\mathbf{x}^{'} \mathbf{C}_2 \mathbf{C}_1 \mathbf{x} \geq 0.$$

One may use the preceding approach to handle the general case, see, e.g., [8, p.8]. The remaining work is left for the interested reader.

(c) $\mathbf{C}$ is a covariance matrix. Show that $e^{\mathbf{C}}$ is a covariance matrix.
Aid: Use a limiting procedure based on that for any square matrix A

$$e^A \overset{\text{def}}{=} \sum_{k=0}^{\infty} \frac{1}{k!} A^n.$$

(see, e.g., [8, p.9]). Do not forget to prove symmetry.

2. **Four product rule** Let $(X_1, X_2, X_3, X_4)^{'} \in N(\mathbf{0}, \mathbf{C})$. Show that

$$E[X_1 X_2 X_3 X_4] =$$

$$E[X_1 X_2] \cdot E[X_3 X_4] + E[X_1 X_3] \cdot E[X_2 X_4] + E[X_1 X_4] \cdot E[X_2 X_3] \tag{8.37}$$

The result is a special case of **Isserli's theorem**, which is in particle physics known as Wick's theorem [4].
Aid : Take the characteristic function of $(X_1, X_2, X_3, X_4)^{'}$. Then use

$$E[X_1 X_2 X_3 X_4] = \frac{\partial^4}{\partial s_1 \partial s_2 \partial s_3 \partial s_4} \phi_{(X_1, X_2, X_3, X_4)}(\underline{s}) \mid_{\underline{s} = \mathbf{0}}.$$

As an additional aid one may say that this requires a lot of manual labour. Note also that we have

$$\frac{\partial^k}{\partial s_i^k} \phi_{\underline{\mathbf{X}}}(\underline{s}) \mid_{\underline{s}=\mathbf{0}} = i^k E\left[X_i^k\right], \quad i = 1, 2, \ldots, n. \tag{8.38}$$

8.5.3 Bussgang's Theorem & Price's Theorem

In this section we assume that

$$\begin{pmatrix} X \\ Y \end{pmatrix} \in N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right).$$

1. **Bussgang's Theorem**

Let $g(y)$ be a Borel function such that $E[|g(Y)|] < \infty$. We are interested in finding

$$\text{Cov}(X, g(Y)).$$

Establish now Bussgang's theorem or Bussgang's formula, which says that

$$\text{Cov}(X, g(Y)) = \frac{\text{Cov}(Y, g(Y))}{\sigma_2^2} \cdot \text{Cov}(X, Y). \tag{8.39}$$

[4] http://en.wikipedia.org/wiki/Isserlis'_theorem

Aid: Use double expectation, so that You write

$$\text{Cov}\,(X, g(Y)) = E\left[(X - \mu_X)\,(g(Y) - E\left[g(Y)\right])\right]$$

$$= E\left[E\left[(X - \mu_X)\,(g(Y) - E\left[g(Y)\right]) \mid Y\right]\right].$$

2. **Bussgang's Theorem and Stein's Lemma** Assume next that $g(y)$ is such that

$$\lim_{y\to\infty} g(y) = g_{\infty} < \infty, \lim_{y\to-\infty} g(y) = g_{-\infty} < \infty,$$

and that $g(y)$ is (almost everywhere) differentiable with the first derivative $g^{'}(y)$ such that $E\left[g^{'}(Y)\right] < \infty$. Show that

$$E\left[g^{'}(Y)\right] = \frac{\text{Cov}\,(Y, g(Y))}{\sigma_2^2}. \tag{8.40}$$

Aid: Use an integration by parts in the integral expression for $\text{Cov}\,(Y, g(Y))$.

Remark 8.5.1 If we write from (8.39) and (8.40) we get

$$\text{Cov}\,(X, g(Y)) = E\left[g^{'}(Y)\right] \cdot \text{Cov}\,(X, Y). \tag{8.41}$$

In statistics (8.41) known as **Stein's lemma**, whereas the (electrical) engineering literature refers to (8.39) and/or (8.40) as Bussgang's theorem[5], see, e.g., [85, p. 340]. In the same way one can also prove that if $X \in N(\mu, \sigma^2)$,

$$E\big(g(X)(X - \mu)\big) = \sigma^2 E\big(g'(X)\big),$$

which is known as Stein's lemma, too. Stein's lemma has a 'Poissonian' counterpart in **Chen's lemma** (2.123). A repeated application of Stein's lemma on the function $g(x) = x^{2n-1}$ yields the moment identity (4.50), too.

The formula (8.41) has been applied as a test of Gaussianity in time series and signal analysis.

■

3. **Bussgang's Theorem and Clipping**

Let next $\mu_X = \mu_Y = 0$ and let $g(y)$ be

$$g(y) = \begin{cases} L & L \leq y \\ y & |y| \leq L \\ -L & y \leq -L. \end{cases}$$

This represents 'clipping' of y at the levels ±L. Show that

$$\text{Cov}\,(X, g(Y)) = 2\text{erf}\left(\frac{L}{\sigma_2}\right) \cdot \text{Cov}\,(X, Y), \tag{8.42}$$

where $\text{erf}\,(x)$ is defined in (2.19). What happens, if $L \to \infty$?

[5] J. Bussgang: Cross-correlation function of amplitude-distorted Gaussian signals. Res.Lab. Elec. MIT, Tech. Rep. 216, March 1952.

4. **Bussgang's Theorem and Hard Limiting**

   Let next $\mu_X = \mu_Y = 0$, let $q > 0$ and let $g(y)$ be the sign function with scaling of levels, i.e.,

$$g(y) = \begin{cases} \frac{q}{2} & 0 < y \\ 0 & y = 0 \\ -\frac{q}{2} & y < 0. \end{cases} \tag{8.43}$$

   This is called hard limiting. Show that

$$\mathrm{Cov}\left(X, g(Y)\right) = \frac{q}{2}\sqrt{\frac{2}{\pi\sigma_2^2}} \cdot \mathrm{Cov}\left(X, Y\right). \tag{8.44}$$

   Can Bussgang's theorem-Stein's lemma from (8.41) be used here, and if yes, how ? The formula in (8.44) is known in circuit theory as the 'input/ouput moment equation for relay correlator'.

5. (From [85, p. 340]) **Price's Theorem**

$$\begin{pmatrix} X \\ Y \end{pmatrix} \in N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

   (a) Show that if $f_{X,Y}(x,y)$ is the joint bivariate p.d.f., then

$$\frac{\partial^n}{\partial^n \rho} f_{X,Y}(x,y) = \frac{\partial^{2n}}{\partial^n x \partial^n y} f_{X,Y}(x,y).$$

   (b) Show that if $Q(x,y)$ is a sufficiently differentiable function integrable with its derivatives w.r.t. x, y, then

$$\frac{\partial^n}{\partial^n \rho} E\left[Q(X,Y)\right] = E\left[\frac{\partial^{2n}}{\partial^n x \partial^n y} Q(X,Y)\right]. \tag{8.45}$$

   This is known as **Price's Theorem**.

   (c) Let $Q(x,y) = xg(y)$, where $g(y)$ is differentiable. Deduce (8.41) by means of (8.45).

Remark 8.5.2 In the applications of Bussgang's and Price's theorems the situation is mostly that $X \leftrightarrow X(t)$ and $Y \leftrightarrow X(t+h)$, where $X(t)$ and $X(t+h)$ are random variables in a Gaussian weakly stationary stochastic process, which is the topic of the next chapter.

■

Chapter 9

Stochastic Processes: Weakly Stationary and Gaussian

9.1 Stochastic Processes

9.1.1 Definition and Terminology, Consistency Theorem

In intuitive terms, a stochastic process is a probabilistic model for evolvement in time of some system that is regarded as being subject to randomly varying influences. We can think that a stochastic process is an ensemble of waveforms (sample functions or sample paths), a waveform chosen at random. A stochastic process is mathematically speaking a family of (infinitely many) random random variables defined on the same probability space.

Definition 9.1.1 A **stochastic process** is a family of random variables $X(t)$,

$$\mathbf{X} = \{X(t) \mid t \in T\},$$

where T is the index set of the process. All random variables $X(t)$ are defined on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

■

In these lecture notes the set T is $\mathbf{R}$ or a subset of $\mathbf{R}$, e.g., $T = [0, \infty)$ or $T = (-\infty, \infty)$ or $T = [a, b]$, $a < b$, and is not countable. We shall thus talk about **stochastic processes in continuous time.**[1]

There are three ways to view a stochastic process;

- For each fixed $t \in T$, $X(t)$ is a random variable $\Omega \mapsto R$.
- $\mathbf{X}$ is a measurable function from $T \times \Omega$ with value $X(t, \omega)$ at (t, ω).
- For each fixed $\omega \in \Omega$, $T \ni t \mapsto X(t, \omega)$ is a function of t called the **sample path** (corresponding to ω).

The mathematical theory deals with these questions as follows. Let now $t_1, \ldots, t_n$ be n points in T and $X(t_1), \ldots, X(t_n)$ be the n corresponding random variables in $\mathbf{X}$. Then for an arbitrary set of real numbers $x_1, x_2, \ldots, x_n$ we have the joint distribution

$$F_{t_1,t_2,\ldots,t_n}(x_1, x_2, \ldots, x_n) = \mathbf{P}(X(t_1) \leq x_1, X(t_2) \leq x_2, \ldots, X(t_n) \leq x_n).$$

[1] A discrete time stochastic process with $T \subseteq \{0, \pm 1, \pm 2 \ldots\}$ is often called a **time series**.

We denote a joint distribution function by

$$F_{t_1,\ldots,t_n}.$$

Suppose that we have a family $\mathbf{F}$ of joint distribution functions or finite dimensional distributions $F_{t_1,\ldots,t_n}$ given for all n and all $t_1,\ldots,t_n \in T$

$$\mathbf{F} = \{F_{t_1,\ldots,t_n}\}_{(t_1,\ldots,t_n)\in T^n, n\in Z_+}.$$

The question is, when we can claim that there exists a stochastic process with $\mathbf{F}$ as its family of finite dimensional distributions.

Theorem 9.1.1 (Kolmogorov Consistency Theorem) Suppose that $\mathbf{F}$ is given and $F_{t_1,\ldots,t_n} \in \mathbf{F}$, and $F_{t_1,\ldots,t_{i-1},t_{i+1},\ldots,t_n} \in \mathbf{F}$. If it holds that

$$F_{t_1,\ldots,t_{i-1},t_{i+1},\ldots,t_n}(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) = \lim_{x_i\uparrow\infty} F_{t_1,\ldots,t_n}(x_1,\ldots,x_n), \tag{9.1}$$

then there exists a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a stochastic process of random variables $X(t)$, $t \in T$, on $(\Omega, \mathcal{F}, \mathbf{P})$ such that $\mathbf{F}$ is its family of finite dimensional distributions.

Proof is omitted here. A concise and readable proof is found in [68, Chapter 1.1]. ■

The condition (9.1) says in plain words that if one takes the joint distribution function for n variables from $\mathbf{F}$, it has to coincide with the marginal distribution for these n variables obtained by marginalization of a joint distribution function from $\mathbf{F}$ for a set of $n+1$ (or, any higher number of) variables that contains these n variables.

Example 9.1.2 Let $\phi \in U(0, 2\pi)$ and

$$X(t) = A\sin(wt+\phi), \quad -\infty < t < \infty,$$

where the amplitude A and the frequency w are fixed. This is a stochastic process, a sine wave with a random phase. We can specify the joint distributions. Take $\mathbf{X} = (X(t_1), X(t_2), \ldots, X(t_n))'$, the characteristic function is

$$\phi_{\mathbf{X}}(\mathbf{s}) = E\left[e^{i\mathbf{s}'\mathbf{X}}\right] = E\left[e^{iR\sin(\phi+\theta)}\right],$$

where

$$R = A\sqrt{\left(\sum_{k=1}^{n} s_k\cos(wt_k)\right)^2 + \left(\sum_{k=1}^{n} s_k\sin(wt_k)\right)^2}$$

$$= A\sqrt{\sum_{k=1}^{n}\sum_{j=1}^{n} s_k s_j\cos(w(t_k-t_j)}$$

and

$$\theta = \tan^{-1}\left(\frac{\sum_{k=1}^{n} s_k\sin(wt_k)}{\sum_{k=1}^{n} s_k\cos(wt_k)}\right).$$

The required details are left for the diligent reader. Hence, see [8, p. 38 -39],

$$\phi_{\mathbf{X}}(\mathbf{s}) = \frac{1}{2\pi}\int_0^{2\pi} e^{iR\sin(\phi+\theta)}d\phi$$

$$= \frac{1}{2\pi}\int_0^{2\pi} e^{iR\sin(\phi)}d\phi = J_0(R),$$

where J_0 is the **Bessel function of first kind of order zero**, [3, pp. 248−249, eq. (6.30)] or in [96, sats 8.1 eq. (12), p. 327] or [92, p. 270]. Needless to say, the joint distribution is not a multivariate Gaussian distribution.

The figures 9.1 and 9.2 illustrate the ways to view a stochastic process stated in the above. We have the probability space $(\Omega, \mathcal{F}, \mathbf{P})$, where $\Omega = [0, 2\pi]$, $\mathcal{F}$ = restriction of the Borel sigma field $\mathcal{B}(\mathbf{R})$ to $[0, 2\pi]$, and $\mathbf{P}$ is the uniform distribution on Borel sets in $[0, 2\pi]$. Thus $\phi \leftrightarrow \omega$. For one ϕ drawn from $U(0, 2\pi)$, we have in figure 9.1 one sample path (or, a random waveform) of $X(t) = \sin(0.5t + \phi)$ ($w = 0.5$ and $A = 1$). In figure 9.2 the graphs are plots of an ensemble of five sample paths of the process corresponding to five samples from $U(0, 2\pi)$. If we focus on the random variable $X(20)$, we see in figure 9.2 five outcomes of $X(20)$. For the third point of view, we see, e.g., $X(20, \omega_5) = 0.9866$, the green path at $t = 20$. In the figure 9.3 we see the histogram for 1000 outcomes of $X(20)$.

■

Remark 9.1.1 The histogram in figure 9.3 can be predicted analytically. We have for $w = 0.5$, $A = 1$, $t = 20$,

$$X(20) = \sin(20 + \phi), \quad \phi \in U(0, 2\pi), \tag{9.2}$$

i.e., $X(20) = H(\phi)$. Since we can by periodicity move to any interval of length 2π, we can consider $X(20) = \sin(\phi)$. It is shown in the example 2.4.2 that the p.d.f. of $X(20)$ is

$$f_{X(20)}(x) = \begin{cases} \frac{1}{\pi\sqrt{1-x^2}}, & -1 < x < 1 \\ 0, \text{ elsewhere.} & \end{cases}$$

Alternatively, [8, p. 18]), the characteristic function of any random variable like $X(20)$ in the random sine wave is, since we can by periodicity move to any interval of length 2π,

$$\varphi_{X(20)}(t) = E\left[e^{it\sin(\phi)}\right] = \frac{1}{2\pi}\int_0^{2\pi} e^{it\sin(\phi)}d\phi = J_0(t)$$

and by inverse Fouriertransformation, (4.2), we get

$$f_{X(20)}(x) = \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} e^{-itx}J_0(t)dt.$$

and this gives $f_{X(20)}$ by a change of variable.

Example 9.1.3 We generalize the example 9.1.2 above. Let $\phi \in U(0, 2\pi)$ and $A \in \mathrm{Ra}(2\sigma^2)$, which means that A has the p.d.f.

$$f_A(x) = \begin{cases} \frac{x}{\sigma^2}e^{-x^2/2\sigma^2} & x \geq 0 \\ 0 & \text{elsewhere.} \end{cases}$$

Let ϕ and A be independent. We have a stochastic process, which is a sine wave with a random amplitude and a random phase. Then we invoke the sine addition formulas

$$X(t) = A\sin(wt + \phi) = A\sin(\phi)\cos(wt) + A\cos(\phi)\sin(wt).$$

It follows that $A\sin(\phi)$ and $A\cos(\phi)$ are independent random variables

$$X_1 = A\sin(\phi) \in N(0, \sigma^2), \quad X_2 = A\cos(\phi) \in N(0, \sigma^2).$$

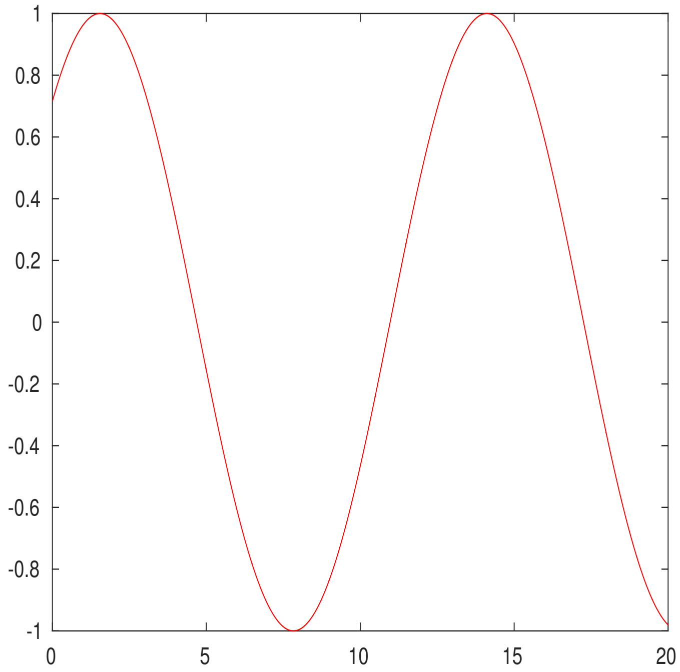

Figure 9.1: One sample path of $X(t) = \sin(0.5t + \phi)$ for $t \in [0, 20]$, $\phi \in U(0, 2\pi)$.

To verify this, we make the change of variables

$$x_1 = A\sin(\phi), \quad x_2 = A\cos(\phi),$$

solve

$$A = \sqrt{x_1^2 + x_2^2}, \quad \tan(\phi) = \frac{x_1}{x_2},$$

compute the Jacobian J, and evaluate $f_A(\sqrt{x_1^2 + x_2^2})\frac{1}{2\pi}|J|$.
The characteristic function for $\mathbf{X} = (X(t_1), X(t_2), \ldots, X(t_n))^{'}$ is

$$\phi_{\mathbf{X}}(\mathbf{s}) = E\left[e^{i\left(A\sin(\phi)\sum_{k=1}^{n} s_k\cos(wt_k)\right)}\right] \cdot E\left[e^{i\left(A\cos(\phi)\sum_{k=1}^{n} s_k\sin(wt_k)\right)}\right]$$

$$= E\left[e^{-i\frac{\sigma^2}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} s_j s_k\cos(w(t_k - t_j))}\right].$$

A second glance at the formula obtained reveals that this should be the characteristic function of a multivariate normal distribution, where the covariance matrix depends on the time points $\{t_k\}_{k=1}^{n}$ only through their mutual differences $t_k - t_j$. As will be understood more fully below, this means that the random sine wave $\{X(t) \mid -\infty < t < \infty\}$ in this example is a **weakly stationary Gaussian stochastic process** with the *autocorrelation* function $\text{Cov}_{\mathbf{X}}(t, s) = \cos(wh)$ for $h = t - s$. We shall now define in general terms the autocorrelation functions and related quantities for stochastic processes.

■

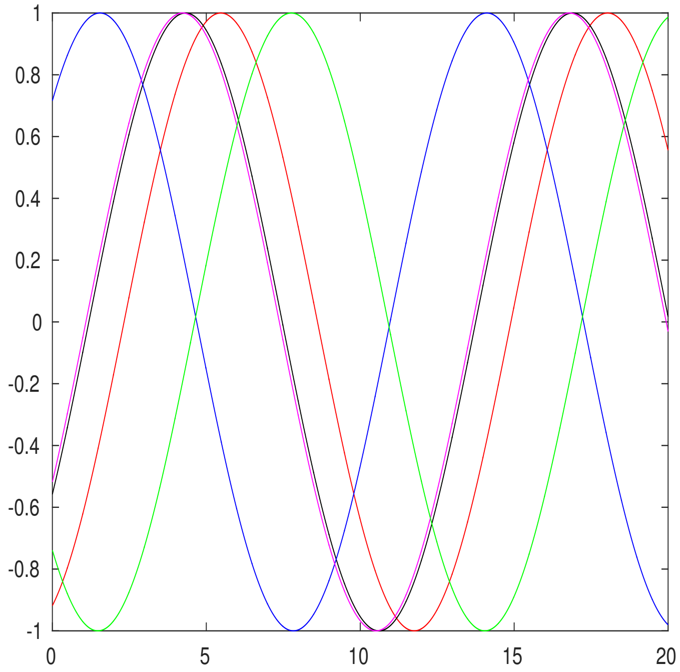

Figure 9.2: Five sample paths of $X(t) = \sin(0.5t + \phi)$ for $t \in [0, 20]$, for five outcomes of $\phi \in U(0, 2\pi)$. Of the random variable $X(20)$ we see five outcomes, $(\omega_i \equiv \phi_i)$, $X(20, \omega_1) = 0.5554$, $X(20, \omega_2) = 0.0167$, $X(20, \omega_3) = -0.9805$, $X(20, \omega_4) = -0.0309$, $X(20, \omega_5) = 0.9866$.

9.1.2 Mean Function, Autocorrelation Function

We shall in the sequel be preoccupied with the moment functions, as soon as these these exist, of a stochastic process $\{X(t) \mid t \in T\}$. Let us assume[2] that $X(t) \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ for all $t \in T$.

> Here, and in the sequel the computational rules (2.47) and (2.48) find frequent and obvious applications without explicit reference.

The *mean function* of the stochastic process $\mathbf{X} = \{X(t) \mid t \in T\}$, $\mu_X(t)$, is

$$\mu_{\mathbf{X}}(t) = E[X(t)], \quad t \in T. \tag{9.3}$$

The *variance function* is

$$\text{Var}_{\mathbf{X}}(t) = E\left[X^2(t)\right] - \mu_{\mathbf{X}}^2(t), \quad t \in T,$$

and the *autocorrelation function* $R_{\mathbf{X}}(t, s)$, is

$$R_{\mathbf{X}}(t, s) = E\left[X(t) \cdot X(s)\right], \quad t, s \in T. \tag{9.4}$$

The *autocovariance* function (*a.c.f.*) is

$$\text{Cov}_{\mathbf{X}}(t, s) = R_{\mathbf{X}}(t, s) - \mu_{\mathbf{X}}(t) \cdot \mu_{\mathbf{X}}(s). \tag{9.5}$$

[2] In [25, 46] such stochastic processes are called *curves* in $L_2(\Omega, \mathcal{F}, \mathbf{P})$.

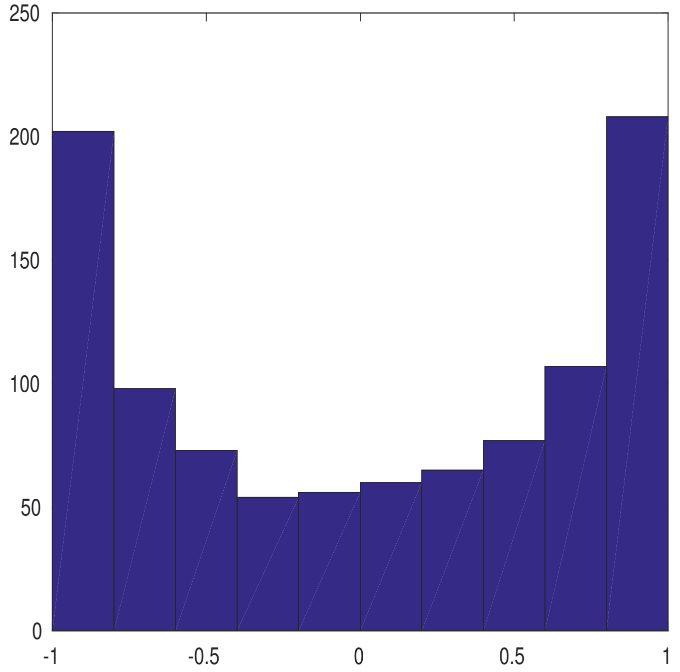

Figure 9.3: The histogram for 1000 outcomes of $X(20)$, $X(t) = \sin(0.5t + \phi)$, $\phi \in U(0, 2\pi)$.

These moment functions depend only on the bivariate joint distributions $F_{t,s}$. We talk also about **second order distributions** and about **second order properties** of a stochastic process.

The terminology advocated above is standard in the engineering literature, e.g., [38, 50, 56, 71, 80, 85, 97, 101], but for a statistician the autocorrelation function would rather have to be $\text{Cov}_{\mathbf{X}}(t,s)/\sqrt{\text{Var}_{\mathbf{X}}(t) \cdot \text{Var}_{\mathbf{X}}(s)}$.

Example 9.1.4 We revisit example 9.1.3 above. The process is

$$X(t) = X_1 \cos(wt) + X_2 \sin(wt),$$

where $X_1 \in N(0, \sigma^2)$ and $X_2 \in N(0, \sigma^2)$ are independent. Then the mean function is

$$\mu_X(t) = \cos(wt) E[X_1] + \sin(wt) E[X_2] = 0,$$

and the autocorrelation function is

$$R_{\mathbf{X}}(t,s) = E[X(t) \cdot X(s)] =$$

$$E[(X_1 \cos(wt) + X_2 \sin(wt)) \cdot (X_1 \cos(ws) + X_2 \sin(ws))] =$$

and since $E[X_1 \cdot X_2] = E[X_1] \cdot E[X_2] = 0$,

$$= \sigma^2 (\cos(wt)\cos(ws) + \sin(wt)\sin(ws))$$

$$= \sigma^2 \cos(w(t-s)),$$

as was already suggested via the characteristic function in example 9.1.3.

■

The autocorrelation function has several distinct properties that are necessary for a function to be an autocorrelation function. For example, if $R_{\mathbf{X}}(t,s)$ is an autocorrelation function, then the following Cauchy-Schwarz inequality holds.

$$\mid R_{\mathbf{X}}(t,s) \mid \leq \sqrt{R_{\mathbf{X}}(t,t)}\sqrt{R_{\mathbf{X}}(s,s)}, \qquad \text{for all } t,s \in T. \tag{9.6}$$

A characterization of autocorrelation functions is given in the next theorem.

Theorem 9.1.5 $R_{\mathbf{X}}(t,s)$ is the autocorrelation function of a process $\mathbf{X} = \{X(t) \mid t \in T\}$, if and only if it has the following properties.

1. **Symmetry**

$$R_{\mathbf{X}}(t,s) = R_{\mathbf{X}}(s,t), \qquad \text{for all } t,s \in T. \tag{9.7}$$

2. **Nonnegative definiteness**

$$\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j R_{\mathbf{X}}(t_i,t_j) \geq 0 \tag{9.8}$$

for all $x_1, x_2, \ldots, x_n$, all $t_1, t_2, \ldots, t_n$ and all n.

Proof: We show the necessity of the property in (9.8).

$$\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j R_{\mathbf{X}}(t_j,t_i) = E\left[\left(\sum_{i=1}^{n} x_i X(t_i)\right)^2\right] \geq 0.$$

■

Clearly (9.8) means that every $n \times n$ - matrix $(R_{\mathbf{X}}(t_i,t_j))_{i=1,j=1}^{n,n}$ is nonnegative definite as in (8.4).

The important question raised by theorem 9.1.5 above is, how to check that a given symmetric function $R(t,s)$ of $(t,s) \in T \times T$ is nonnegative definite.

Example 9.1.6 Consider the function of (t,s) given by

$$\frac{1}{2}\left(\frac{H}{\alpha}\right)^{2H}\left(e^{\alpha(t-s)} + e^{-\alpha(t-s)} - e^{\alpha(t-s)}\left(1 - e^{-\frac{\alpha(t-s)}{H}}\right)^{2H}\right)$$

for $\alpha > 0$ and $0 < H < 1$. Is this an autocorrelation function?[3]

■

One way to decide the question in example above and elsewhere is to find a random process that has $R(t,s)$ as its autocorrelation function. This can, on occasion, require a lot of ingenuity and effort and is prone to errors. We shall give several examples of autocorrelation functions and corresponding underlying processes. It should be kept in mind right from the start that there can be many different stochastic processes with the same autocovariance function.

There is a class of processes with random variables $X(t) \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ called **weakly stationary processes**, that has been extensively evoked in the textbook and engineering literature and practice, c.f., [1, 38, 50, 56, 71, 80, 85, 89, 97, 101, 103]. Weakly stationary processes can be constructed by means of linear analog filtering of (white) noise, as is found in the exercises of section 9.7.5. The weakly stationary processes are defined as having a constant mean and an autocorrelation function which is a function of the difference between t and s, c.f., example 9.1.4. The weakly stationary processes will be defined and treated in section 9.3.

We begin with a few examples of families of autocorrelation functions.

[3] The answer may be found in `http://arxiv.org/abs/0710.5024`

Example 9.1.7 (Bilinear forms of autocorrelation functions) Take **any** real function $f(t)$, $t \in T$. Then

$$R(t,s) = f(t) \cdot f(s) \tag{9.9}$$

is an autocorrelation function of a stochastic process. In fact, take $X \in N(0,1)$, and set $X(t) = f(t)X$. Then $R(t,s)$ is the autocorrelation of the process $\{X(t) \mid t \in T\}$. The mean function is the constant $= 0$. Thus

$$R(t,s) = \sum_{i=0}^{n} f_i(t) \cdot f_i(s)$$

and even

$$R(t,s) = \sum_{i=0}^{\infty} f_i(t) \cdot f_i(s) \tag{9.10}$$

are autocorrelation functions.

■

The next example is a construction of a stochastic process that leads to the bilinear $R(t,s)$ as given in (9.10), see [7, pp. 6−10] or [103, pp. 82−88].

Example 9.1.8 Let $X_i \in N(0,1)$ be I.I.D. for $i = 0, 1, \ldots$. Take for $i = 0, 1, \ldots$ the real numbers $\lambda_i \geq 0$ such that $\sum_{i=0}^{\infty} \lambda_i < \infty$. Let $e_i(t)$ for $i = 0, 1, \ldots$ be a sequence of functions of $t \in [0, T]$ such that

$$\int_0^T e_i(t)e_j(t)dt = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{9.11}$$

and that $(e_i)_{i=0}^{\infty}$ is an orthonormal basis in $L_2([0,T])$, [96, pp. 279−286]. We set

$$X_N(t) \stackrel{\text{def}}{=} \sum_{i=0}^{N} \sqrt{\lambda}_i X_i e_i(t).$$

Then one can show using the methods in section 7.4.2 that for every $t \in [0, T]$

$$X_N(t) \stackrel{2}{\to} X(t) = \sum_{i=0}^{\infty} \sqrt{\lambda}_i X_i e_i(t), \tag{9.12}$$

as $N \to \infty$. Clearly, by theorem 7.4.2 $X(t)$ is a Gaussian random variable. The limit is in addition a stochastic process such that

$$E\left[X(t)X(s)\right] = \sum_{i=0}^{\infty} \sqrt{\lambda}_i e_i(t) \sqrt{\lambda}_i e_i(s),$$

where we used theorem 7.3.1. But this is (9.10) with $f_i(t) = \sqrt{\lambda}_i e_i(t)$. This example will be continued in example 9.2.3 in the sequel and will eventually yield a construction of the Wiener process, see section 10.3 below.

■

Example 9.1.9 A further bilinear form of autocorrelation functions By some further extensions of horizons, [46, chapter 2.3], we can show that integrals of the form

$$R(t,s) = \int_a^b f(t,\lambda) \cdot f(s,\lambda) d\lambda$$

are autocorrelation functions.

■

Example 9.1.10 *[***Separable autocorrelation functions***]* We have here a family of autocorrelation functions that turn out to correspond to certain important processes.

1. Let $T = [0, \infty)$ and $\sigma > 0$. Then
$$R(t,s) = \sigma^2 \min(t,s) \tag{9.13}$$
is an autocorrelation function of a stochastic process. How can one claim this ? The answer is deferred till later, when it will be shown that this is the autocorrelation function of a *Wiener process.*

2. $T = [0,1]$ and
$$R(t,s) = \begin{cases} s(1-t) & s \leq t \\ (1-s)t & s \geq t. \end{cases} \tag{9.14}$$
This is the autocorrelation function of a process known as the *Brownian bridge* or the *tied down Wiener process.*

3. Let $T = (-\infty, \infty)$ and $a > 0$
$$R(t,s) = e^{-a|t-s|} = \begin{cases} e^{as}e^{-at} & s \leq t \\ e^{at}e^{-as} & s \geq t. \end{cases} \tag{9.15}$$
This is the autocorrelation function of a weakly stationary process, as it is a function of $|t-s|$. One process having this autocorrelation function is a stationary *Ornstein-Uhlenbeck process* in the chapter 11, another is the random telegraph signal in chapter 12.5.

4. Let $T = (-\infty, \infty)$.
$$R(t,s) = \begin{cases} u(s)v(t) & s \leq t \\ u(t)v(s) & s \geq t \end{cases} \tag{9.16}$$
Here $u(s) > 0$ for all $s \in T$. We can write (9.16) more compactly as
$$R(t,s) = u\left(\min(t,s)\right) v\left(\max(t,s)\right).$$
By this we see that all the preceding examples (9.13) - (9.15) are special cases of (9.16) for appropriate choices of $u(\cdot)$ and $v(\cdot)$. Processes with this kind of autocorrelation functions are the so called **Gauss-Markov** processes.

5. Let $T = (-\infty, \infty)$ and
$$R(t,s) = \sum_{i=1}^{k} u_i\left(\min(t,s)\right) v_i\left(\max(t,s)\right) \tag{9.17}$$
These autocovariance functions constitute the class of **separable autocorrelation** functions. We shall say more of the processes corresponding to separable autocorrelation functions (i.e. Gauss-Markov processes) in the in section 9.5.

Example 9.1.11 (Periodic Autocorrelation) [97, pp. 272−273] Let $\{B_i\}_{i \geq 1}$ be a sequence of independent random variables with $B_i \in \text{Be}(1/2)$ for all i. Define
$$\Theta_i = \begin{cases} \frac{\pi}{2} & \text{if } B_i = 1 \\ -\frac{\pi}{2} & \text{if } B_i = 0. \end{cases}$$

Take for $T > 0$ and for any integer k

$$\Theta(t) = \Theta_k \quad \text{for } kT \leq t < (k+1)T.$$

Note that in this example T is the length of a time interval, actually the time for transmission of one bit, not the overall index set, as elsewhere in this text. Set

$$X(t) = \cos\left(2\pi f_c t + \Theta(t)\right), \quad -\infty < t < \infty.$$

This process is known as **Phase-Shift Keying** (PSK) , which is a basic method of modulation in transmission of binary data.

We determine the mean function and the autocorrelation function of PSK. It is helpful to introduce two auxiliary functions

$$s_I(t) = \begin{cases} \cos\left(2\pi f_c t\right) & \text{if } 0 \leq t < T \\ 0 & \text{else} \end{cases}$$

and

$$s_Q(t) = \begin{cases} \sin\left(2\pi f_c t\right) & \text{if } 0 \leq t < T \\ 0 & \text{else.} \end{cases}$$

Then we get by the cosine addition formula that

$$\cos\left(2\pi f_c t + \Theta(t)\right) = \cos\left(\Theta(t)\right)\cos\left(2\pi f_c t\right) - \sin\left(\Theta(t)\right)\sin\left(2\pi f_c t\right)$$

$$= \sum_{k=-\infty}^{\infty} \left[\cos\left(\Theta_k\right) s_I(t-kT) - \sin\left(\Theta_k\right) s_Q(t-kT)\right].$$

This looks like an infinite sum, but actually there is no need to prove any convergence.

The mean function follows easily, since $\cos\left(\Theta_k\right) = 0$ and $\sin\left(\Theta_k\right) = \pm 1$ with equal probabilities. Hence

$$E\left[X(t)\right] = 0 \quad \text{for all } t.$$

The autocorrelation function is thus

$$R_{\mathbf{X}}(t,s) = \sum_{k,l} E\left[\sin\left(\Theta_k\right)\sin\left(\Theta_l\right)\right] s_Q(t-kT) s_Q(s-lT).$$

Here we have, if $k \neq l$,

$$E\left[\sin\left(\Theta_k\right)\sin\left(\Theta_l\right)\right] = 1 \cdot 1 P\left(\Theta_k = \frac{\pi}{2}\right) P\left(\Theta_k = \frac{\pi}{2}\right) + 1 \cdot (-1) P\left(\Theta_k = \frac{\pi}{2}\right) P\left(\Theta_k = -\frac{\pi}{2}\right)$$

$$+(-1) \cdot 1 P\left(\Theta_k = \frac{-\pi}{2}\right) P\left(\Theta_k = \frac{\pi}{2}\right) + (-1)\cdot(-1) P\left(\Theta_k = -\frac{\pi}{2}\right) P\left(\Theta_k = -\frac{\pi}{2}\right)$$

$$= \frac{1}{4} - \frac{1}{4} - \frac{1}{4} + \frac{1}{4} = 0.$$

If $k = l$, then $E\left[\sin^2\left(\Theta_k\right)\right] = \frac{1}{2} + \frac{1}{2} = 1$.

Therefore we have

$$R_{\mathbf{X}}(t,s) = \sum_{k=-\infty}^{\infty} s_Q(t-kT) s_Q(s-lT).$$

Since the **support**[4] of $s_Q(t)$ is $[0, T[$, there is no overlap, i.e., for any fixed pair (t,s) only one of the product terms in the sum can be nonzero. Also, if t and s are not in the same period, then this term is not zero.

[4] by the support of a function $f(t)$ we mean the set of points t, where $f(t) \neq 0$.

If we put

$$(t) = t/T - \lfloor t/T \rfloor, \qquad \lfloor t/T \rfloor = \text{integer part of } t/T,$$

then we can write

$$R_{\mathbf{X}}(t,s) = \begin{cases} s_Q((t))s_Q((s)) & \text{for } \lfloor t/T \rfloor = \lfloor s/T \rfloor \\ 0 & \text{else} \end{cases}$$

Thus the autocorrelation function $R_{\mathbf{X}}(t,s)$ of PSK is a periodic function in the sense that $R_{\mathbf{X}}(t,s) = R_{\mathbf{X}}(t+T, s+T)$ (i.e., periodic with the same period in both variables). The textbook [38, chapter 12] and the monograph [60] contain specialized treatments of the theory and applications of stochastic processes with periodic autocorrelation functions.

■

One way of constructing stochastic processes that have a given autocorrelation function is by mean square integrals of stochastic processes in continuous time, as defined next.

9.2 Mean Square Calculus: The Mean Square Integral

There is a *mean square calculus* of stochastic processes, see [1, 46, 56, 66, 71, 89, 97], that is nicely applicable to weakly stationary processes.

9.2.1 Definition and Existence of the Mean Square Integral

Definition 9.2.1 Let $\{X(t) | t \in T\}$ be a stochastic process in continuous time with $X(t) \in L_2(\Omega, \mathcal{F}, \mathbf{P})$ for every $t \in T$. The *mean square integral* $\int_a^b X(t)dt$ of $\{X(t) | t \in T\}$ over $[a,b] \subseteq T$ is defined as the mean square limit (when it exists)

$$\sum_{i=1}^{n} X(t_i)(t_i - t_{i-1}) \overset{2}{\to} \int_a^b X(t)dt, \tag{9.18}$$

where $a = t_0 < t_1 < \ldots < t_{n-1} < t_n = b$ and $\max_i |t_i - t_{i-1}| \to 0$, as $n \to \infty$.

■

The sample paths of a process $\{X(t) | t \in T\}$ need not be integrable in Riemann's sense[5]. Since a mean square integral does not involve sample paths of $\{X(t) | t \in T\}$, we are elaborating an easier notion of integration.

Theorem 9.2.1 The mean square integral $\int_a^b X(t)dt$ of $\{X(t) | t \in T\}$ exists over $[a,b] \subseteq T$ if and only if the double integral

$$\int_a^b \int_a^b E\left[X(t)X(u)\right] dtdu$$

exists as an integral in Riemann's sense. We have also

$$E\left[\int_a^b X(t)dt\right] = \int_a^b \mu_{\mathbf{X}}(t)dt \tag{9.19}$$

and

$$\text{Var}\left[\int_a^b X(t)dt\right] = \int_a^b \int_a^b \text{Cov}_{\mathbf{X}}(t,u)dtdu. \tag{9.20}$$

[5] The Riemann integral is the integral handed down by the first courses in calculus, see, e.g., [69, chapter 6].

Proof Let $Y_n = \sum_{i=1}^{n} X(t_i)(t_i - t_{i-1})$. Evoking **Loéve's criterion** in theorem 7.3.3 we study

$$E\left[Y_n Y_m\right] = \sum_{i=1}^{n}\sum_{j=1}^{m} E\left[X(t_i)X(u_j)\right](t_i - t_{i-1})(u_j - u_{j-1}),$$

where the right hand side is an ordinary Riemann's sum. Hence

$$\sum_{i=1}^{n}\sum_{j=1}^{m} E\left[X(t_i)X(u_j)\right](t_i - t_{i-1})(u_j - u_{j-1}) \to \int_a^b \int_a^b E\left[X(t)X(u)\right]dtdu = C,$$

in case the double integral exists, when $a = t_0 < t_1 < \ldots < t_{n-1} < t_n = b$ and $\max_i |t_i - t_{i-1}| \to 0$ as $n \to \infty$. So the assertion follows, as claimed.

The expectation $E\left[\int_a^b X(t)dt\right]$ is obtained as

$$E\left[\int_a^b X(t)dt\right] = E\left[\lim_{\triangle}\sum_{i=1}^{n} X(t_i)(t_i - t_{i-1})\right],$$

where the auxiliary notion $\lim_{\triangle}$ refers to mean square convergence as $a = t_0 < t_1 < \ldots < t_{n-1} < t_n = b$ and $\max_i |t_i - t_{i-1}| \to 0$ as $n \to \infty$, and by theorem 7.3.1 (a)

$$= \lim_{n\to\infty} E\left[Y_n\right] = \lim_{n\to\infty}\sum_{i=1}^{n} E\left[X(t_i)\right](t_i - t_{i-1})$$

and then

$$= \lim_{n\to\infty}\sum_{i=1}^{n} \mu_{\mathbf{X}}(t_i)(t_i - t_{i-1}) = \int_a^b \mu_{\mathbf{X}}(t)dt.$$

The variance is computed by

$$\text{Var}\left[\int_a^b X(t)dt\right] = E\left[\left(\int_a^b X(t)dt\right)^2\right] - \left(\int_a^b \mu_{\mathbf{X}}(t)dt\right)^2.$$

Here

$$E\left[\left(\int_a^b X(t)dt\right)^2\right] = E\left[\left(\int_a^b X(t)dt \cdot \int_a^b X(u)du\right)\right]$$

and from theorem 7.3.1 (c)

$$E\left[\int_a^b X(t)dt \cdot \int_a^b X(u)du\right] = E\left[\lim_{\triangle} Y_n \cdot Y_m\right]$$

$$= \lim_{\min(m,n)\to\infty} E\left[Y_n \cdot Y_m\right],$$

where $E\left[Y_n \cdot Y_m\right] = \sum_{i=1}^{n}\sum_{j=1}^{m} E\left[X(t_i)X(u_j)\right](t_i - t_{i-1})(u_j - u_{j-1})$. Thus

$$\text{Var}\left[\int_a^b X(t)dt\right] \to \int_a^b \int_a^b E\left[X(t)X(u)\right]dtdu - \left(\int_a^b \mu_X(t)dt\right)^2$$

$$= \int_a^b \int_a^b \left(E\left[X(t)X(u)\right] - \mu_{\mathbf{X}}(t)\mu_{\mathbf{X}}(u)\right)dtdu,$$

which is the assertion as claimed. ■

One can manipulate mean square stochastic integrals much in the same way as ordinary integrals.

Theorem 9.2.2 (a)

$$\int_a^b (\alpha X(t) + \beta Y(t))\, dt = \alpha \int_a^b X(t)dt + \beta \int_a^b Y(t)dt$$

(b) $a < b < c$

$$\int_a^b X(t)dt + \int_b^c X(t)dt = \int_a^c X(t)dt$$

■

Hence we may define new stochastic process $\mathbf{Y} = \{Y(t) \mid t \in T\}$ by a stochastic integral. For each $t \in T$ we set

$$Y(t) = \int_a^t X(s)ds.$$

The mean function of the process $\mathbf{Y}$ is

$$\mu_Y(t) = \int_a^t \mu_{\mathbf{X}}(s)ds, \quad t \in T,$$

and its autocovariance is

$$\mathrm{Cov}_Y(t,s) = \int_0^t \int_0^s \mathrm{Cov}_{\mathbf{X}}(u,v)dudv, \quad (t,s) \in T \times T.$$

Example 9.2.3 We continue with example 9.1.8, where we constructed the random variables

$$X(t) = \sum_{i=0}^{\infty} \sqrt{\lambda_i} X_i e_i(t), \quad t \in [0,T] \tag{9.21}$$

and found their autocorrelation $R(t,s)$ function as a bilinear form. The expression (9.21) is known as the **Karhunen-Loéve expansion** of $X(t)$. When we consider the mean square integral

$$\int_0^T X(t)e_i(t)dt,$$

we obtain by the results on this category of integrals above and by the results on convergence in mean square underlying (9.21) that

$$\int_0^T X(t)e_j(t)dt = \sum_{i=0}^{\infty} \sqrt{\lambda_i} X_i \int_0^T e_j(t)e_i(t)dt = \sqrt{\lambda_j} X_j, \tag{9.22}$$

where we used (9.11). Then

$$\int_0^T R(t,s)e_j(s)ds = \int_0^T E\left[X(t)X(s)\right] e_j(s)ds = E\left[X(t)\int_0^T X(s)e_j(s)ds\right]$$

and from (9.22)

$$= E\left[X(t)\sqrt{\lambda_j} X_j\right]$$

and from (9.21)

$$= \sum_{i=0}^{\infty} \sqrt{\lambda_i}\sqrt{\lambda_j} E\left[X_i X_j\right] e_i(t) = \lambda_j e_j(t),$$

since $X_i \in N(0,1)$ and I.I.D.. In summary, we have ascertained that

$$\int_0^T R(t,s)e_j(s)ds = \lambda_j e_j(t). \tag{9.23}$$

This is an **integral equation**, which is to be solved w.r.t. e_i and λ_j. It holds in fact that we can regard e_i's as **eigenfunctions** and λ_is as the corresponding **eigenvalues** of the autocorrelation function $R(t,s)$. If $R(t,s)$ is continuous in $[0,T] \times [0,T]$, we can always first solve (9.23) w.r.t. λ_i and e_i and then construct $X(t) = \sum_{i=0}^{\infty} \sqrt{\lambda_i} X_i e_i(t)$. For the rigorous mathematical details we refer to [46, pp. 62−69]. The insights in this example will be made use of in section 10.3.

■

9.3 Weakly Stationary Processes

Definition 9.3.1 A process $\mathbf{X} = \{X(t) \mid t \in T =]-\infty, \infty[\}$ is called **(weakly) stationary** if

1. The mean function $\mu_{\mathbf{X}}(t)$ is a constant function of t, $\mu_{\mathbf{X}}(t) = \mu$.
2. The autocorrelation function $R_{\mathbf{X}}(t,s)$ is a function of $(t-s)$, so that

$$R_{\mathbf{X}}(t,s) = R_{\mathbf{X}}(h) = R_{\mathbf{X}}(-h), \quad h = (t-s).$$

■

It follows that for a weakly stationary process even the variance functions is a constant, say $\sigma^2_{\mathbf{X}}$, as a function of t, since

$$\text{Var}_{\mathbf{X}}(t) = E\left[X^2(t)\right] - \mu^2_{\mathbf{X}}(t) = R_{\mathbf{X}}(0) - \mu^2 \overset{\text{def}}{=} \sigma^2_{\mathbf{X}}.$$

In addition, the autocovariance is

$$\text{Cov}_X(h) = R_{\mathbf{X}}(h) - \mu^2,$$

and then

$$\text{Cov}_X(0) = \sigma^2_{\mathbf{X}}.$$

By (9.6) we get here

$$\mid R_{\mathbf{X}}(h) \mid \leq R_{\mathbf{X}}(0). \tag{9.24}$$

This is another necessary condition for a function $R_{\mathbf{X}}(h)$ to be an autocorrelation function of a weakly stationary process.

We have already encountered an example of a weakly stationary process in example 9.1.4.

9.3.1 Bochner's Theorem, Spectral Density and Examples of Autocorrelation Functions for Weakly Stationary Processes

The following theorem gives an effective criterion for deciding, when a function $R(h)$ is nonnegative definite. A simpler version of it is sometimes referred to as the **Einstein-Wiener-Khinchin theorem** .

Theorem 9.3.1 *[***Bochner's Theorem***]* A function $R(h)$ is nonnegative definite if and only if it can be represented in the form

$$R(h) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ihf} dS(f), \tag{9.25}$$

where $S(f)$ is real, nondecreasing and bounded.

Proof of $\Leftarrow$, or we assume that we have a function $R(h)$ that is given by (9.25). Then we show that $R(h)$ is nonnegative definite. Assume that

$$\frac{d}{df}S(f) = s(f)$$

exists, thus $s(f) \geq 0$. Then take any $x_1, \ldots x_n$ and $t_1, \ldots t_n$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} x_i R(t_i - t_j) x_j = \frac{1}{2\pi}\int_{-\infty}^{\infty}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i e^{i(t_i - t_j)f} x_j s(f) df$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\sum_{i=1}^{n} x_i e^{it_i f}\sum_{j=1}^{n} x_j e^{-it_j f} s(f) df = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left|\sum_{i=1}^{n} x_i e^{it_i f}\right|^2 s(f) df \geq 0.$$

Here $|z|^2 = z \cdot \overline{z}$ is the squared modulus of a complex number so that $\overline{z}$ is the complex conjugate of z.

One elegant and pedagogical proof of the converse statement, namely that if $R(h)$ is nonnegative definite function, then we can express it as in (9.25), is due to H. Cramér and can be found in [25, pp. 126−128]. ■

The function $S(f)$ is called the **spectral distribution function**. If $S(f)$ has a derivative,

$$\frac{d}{df}S(f) = s(f),$$

then $s(f)$ is called the **spectral density**. Clearly $s(f) \geq 0$, as $S(f)$ is nondecreasing. Since $R(h) = R(-h)$, we get also that $s(f) = s(-f)$ is to be included in the set of necessary conditions for $R(h)$ to be an autocorrelation function.

Another term used for $s(f)$ is **power spectral density**, as

$$E\left[(X(t))^2\right] = R_{\mathbf{X}}(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} s_{\mathbf{X}}(f) df.$$

The electrical engineering[6] statement is that $s_{\mathbf{X}}(f)$ is the density of power at the frequency f.

Operationally, if one can find a Fourier transform $s(f)$ of a function $R(h)$ with the properties

- $s(f) \geq 0$, and $s(-f) = s(f)$,

then $R(h)$ is the autocorrelation of a weakly stationary process.

Some examples of pairs of autocorrelation functions and (power) spectral densities are given in the table below quoted from [42]. When reading certain details of this table one should remember that $\frac{\sin(0)}{0} = 1$ in view of $\lim_{h \to 0} \frac{\sin(h)}{h} = 1$.

Autocorrelation functions $R(h)$	Spectral densities $s(f)$
$e^{-a\|h\|}$	$\frac{2a}{a^2+f^2}$
$e^{-a\|h\|}\cos(bh)$	$\frac{a}{a^2+(f-b)^2} + \frac{a}{a^2+(f+b)^2}$
$\frac{a}{\pi}\frac{\sin(ah)}{ah}$	$\begin{cases} 1 & \|f\| \leq a \\ 0 & \|f\| > a \end{cases}$
$\begin{cases} 1 - a\|h\| & \|h\| < 1/a \\ 0 & \|h\| \geq 1/a \end{cases}$	$\frac{1}{a}\left(\frac{\sin\left(\frac{f}{2a}\right)}{\frac{f}{2a}}\right)^2$
$\frac{e^{-a\|h\|}}{4(a^2+b^2)}\left(\frac{\cos(bh)}{b} + \frac{\sin(bh)}{b}\right)$	$\frac{1}{(a^2+(f-b)^2)(a^2+(f+b)^2)}$

[6] If $X(t)$ is a voltage or current developed across a 1-ohm resistor, then $(X(t))^2$ is the instantaneous power absorbed by the resistor.

9.3.2 Mean Square Continuity of Weakly Stationary Processes

In the category of second order properties we note the following.

Definition 9.3.2 Let $\mathbf{X} = \{X(t) | t \in T\}$ be a stochastic process in continuous time. Then the process is *mean square continuous* if, when $t + \tau \in T$,

$$E\left[(X(t+\tau) - X(t))^2\right] \to 0$$

as $\tau \to 0$.

■

In order to see what the definition implies let us expand

$$\begin{aligned} E\left[(X(t+\tau) - X(t))^2\right] &= E\left[X(t+\tau)X(t+\tau)\right] - E\left[X(t+\tau)X(t)\right] \\ &\quad -E\left[X(t)X(t+\tau)\right] + E\left[X(t)X(t)\right] \\ &= \mathrm{Cov}_{\mathbf{X}}\left(t+\tau, t+\tau\right) - \mathrm{Cov}_{\mathbf{X}}\left(t, t+\tau\right) - \mathrm{Cov}_{\mathbf{X}}\left(t, t+\tau\right) + \mathrm{Cov}_{\mathbf{X}}\left(t, t\right) + \\ &\quad + \left(\mu_{\mathbf{X}}\left(t+\tau\right) - \mu_{\mathbf{X}}(t)\right)^2. \end{aligned}$$

We get a neat result from this, if we assume that $\mathbf{X}$ is weakly stationary, as then from the above

$$E\left[(X(t+\tau) - X(t))^2\right] = \mathrm{Cov}_{\mathbf{X}}(0) - 2\mathrm{Cov}_{\mathbf{X}}(\tau) + \mathrm{Cov}_{\mathbf{X}}(0).$$

Theorem 9.3.2 A weakly stationary process is mean square continuous, if and only if $\mathrm{Cov}_{\mathbf{X}}(\tau)$ is continuous in the origin.

■

Continuity in mean square does not without further requirements imply continuity of sample paths.

9.4 Gaussian Processes

9.4.1 Definition and Existence

The chapter on Gaussian processes in [67, ch. 13] is an up-to-date and powerful presentation of the topic as a part of modern probability theory.

Definition 9.4.1 A stochastic process $\mathbf{X} = \{X(t) \mid t \in T\}$ is called **Gaussian**, if every stochastic n-vector $(X(t_1), X(t_2), \cdots, X(t_n))'$ is a multivariate normal vector for all n and all $t_1, t_2, \cdots, t_n$.

■

In more detail, this definition says that all stochastic $n \times 1$-vectors

$$(X(t_1), X(t_2), \cdots, X(t_n))'$$

have a multivariate normal distribution

$$F_{t_1,t_2,\ldots,t_n} \leftrightarrow N\left(\underline{\mu}(t_1, t_2, \cdots, t_n), \mathbf{C}(t_1, t_2, \cdots, t_n)\right),$$

where $\underline{\mu}(t_1, t_2, \cdots, t_n)$ is an $n \times 1$ expectation vector, with elements given by

$$\mu_{\mathbf{X}}(t_i) = E(X(t_i)) \quad i = 1, \ldots, n$$

and the $n \times n$ covariance matrix $\mathbf{C}(t_1, t_2, \cdots, t_n) = (c_{ij})_{i=1,j=1}^{n,n}$ has the entries

$$c_{ij} = \mathrm{Cov}_{\mathbf{X}}(t_i, t_j) = R_{\mathbf{X}}(t_i, t_j) - \mu_X(t_i)\mu_X(t_j), i = 1, \ldots, n; j = 1, \ldots, n,$$

i.e., the entries in the covariance matrix are the appropriate values of the autocovariance function.

We show next a theorem of existence for Gaussian processes. This sounds perhaps like a difficult thing to do, but by the Kolmogorov consistency theorem, or (9.1), all we need to show in this case is effectively that

- **All marginal distributions of a multivariate normal distribution are normal distributions**.

Remark 9.4.1 Evidently, if $\mu_{\mathbf{Z}}(t) = E(Z(t)) = 0$ for all $t \in T$, then

$$X(t) = f(t) + Z(t)$$

has mean function $\mu_{\mathbf{X}}(t) = E(X(t)) = f(t)$. Hence we may without loss of generality prove the existence of Gaussian processes by assuming zero as mean function.

Theorem 9.4.1 Let $R(t, s)$ be a symmetric and nonnegative definite function. Then there exists a Gaussian stochastic process $\mathbf{X}$ with $R(t, s)$ as its autocorrelation function and the constant zero as mean function.

Proof Since the mean function is zero, and since $R(t, s)$ is a symmetric and nonnegative definite function, we find that

$$\mathbf{C}(t_1, t_2, \cdots, t_{n+1}) = (R(t_i, t_j))_{i=1,j=1}^{n+1,n+1}$$

is the covariance matrix of a random vector $(X(t_1), X(t_2), \cdots, X(t_{n+1}))'$. We set for ease of writing

$$\mathbf{C}_{\mathbf{t_{n+1}}} = (R(t_i, t_j))_{i=1,j=1}^{n+1,n+1}.$$

We know by [49, p. 123] that we can take the vector $(X(t_1), X(t_2), \cdots, X(t_{n+1}))'$ as multivariate normal. We set for simplicity of writing

$$\mathbf{X}_{\mathbf{t_{n+1}}} = (X(t_1), X(t_2), \cdots, X(t_{n+1}))',$$

and let

$$F_{\mathbf{t_{n+1}}} \leftrightarrow N\left(\mathbf{0}, \mathbf{C}_{\mathbf{t_{n+1}}}\right)$$

denote its distribution function. Then the characteristic function for $\mathbf{X}_{\mathbf{t_{n+1}}}$ is with $\mathbf{s}_{n+1} = (s_1, \ldots, s_{n+1})$ given by

$$\phi_{\mathbf{X}_{\mathbf{t_{n+1}}}}(\mathbf{s}_{n+1}) := E\left[e^{i\mathbf{s}'_{n+1}\mathbf{X}_{\mathbf{t_{n+1}}}}\right] = \int_{R^{n+1}} e^{i\mathbf{s}'_{n+1}\mathbf{x}} dF_{\mathbf{t_{n+1}}}(\mathbf{x}). \tag{9.26}$$

Let us now take

$$\mathbf{s}'_{(i)} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots s_{n+1}).$$

The proof has two steps.

1. We show that $\phi_{\mathbf{X}_{\mathbf{t_{n+1}}}}((s_1, \ldots, s_{i-1}, 0, s_{i+1}, \ldots s_{n+1}))$ gives us the characteristic function of

$$\mathbf{X}_{\mathbf{t_{(i)}}} = (X(t_1), X(t_2), \ldots, X(t_{i-1}), X(t_{i+1}), \ldots, X(t_{n+1}))'.$$

   We denote by $F_{\mathbf{t_{(i)}}}$ its distribution function.

2. We show that $\phi_{\mathbf{X}_{\mathbf{t_{n+1}}}}((s_1, \ldots, s_{i-1}, 0, s_{i+1}, \ldots s_{n+1}))$ is the characteristic function of a normal distribution for the $n-1$ variables in $\mathbf{X}_{\mathbf{t_{(i)}}}$.

The details of the steps outlined are as follows.

1. Set

$$\mathbf{x}_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_{n+1}).$$

   We get that

$$\phi_{\mathbf{X}_{\mathbf{t_{n+1}}}}\left((s_1, \ldots, s_{i-1}, 0, s_{i+1}, \ldots s_{n+1})\right)$$

$$= \int_{\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{i(s_1 x_1 + \cdots + s_{i-1}x_{i-1} + s_{i+1}x_{i+1} + \cdots + s_{n+1}x_{n+1})} \int_{x_i=-\infty}^{x_i=\infty} dF_{\mathbf{t_{n+1}}}(\mathbf{x})$$

$$= \int_{\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{i(s_1 x_1 + \cdots + s_{i-1}x_{i-1} + s_{i+1}x_{i+1} + \cdots + s_{n+1}x_{n+1})} dF_{\mathbf{t_{(i)}}}\left(\mathbf{x}_{(i)}\right) \tag{9.27}$$

   is the characteristic function of $\mathbf{X}_{\mathbf{t_{(i)}}}$.

2. On the other hand, we have the quadratic form

$$(s_1, \ldots, s_{i-1}, 0, s_{i+1}, \ldots s_{n+1})\, \mathbf{C}_{\mathbf{t_{n+1}}}\, (s_1, \ldots, s_{i-1}, 0, s_{i+1}, \ldots s_{n+1})^{'}$$

$$= \sum_{l=1, l\neq i}^{n+1} \sum_{k=1, k\neq i}^{n+1} s_l s_k R(t_l, t_k)$$

$$= \mathbf{s}_{(i)}^{'} \mathbf{C}_{\mathbf{t_{(i)}}} \mathbf{s}_{(i)}.$$

   In this $\mathbf{C}_{\mathbf{t_{(i)}}} = (R(t_l, t_k))_{l=1, l\neq i, k=1, k\neq i}^{n+1, n+1}$ is recognized as the covariance matrix of $\mathbf{X}_{\mathbf{t_{(i)}}}$. By (8.10) we have here, since $\mathbf{X}_{\mathbf{t_{n+1}}}$ is a Gaussian random variable,

$$\phi_{\mathbf{X}_{\mathbf{t_{n+1}}}}\left((s_1, \ldots, s_{i-1}, 0, s_{i+1}, \ldots s_{n+1})\right)$$

$$= e^{-\frac{1}{2}\mathbf{s}_{(i)}^{'} \mathbf{C}_{\mathbf{t_{(i)}}} \mathbf{s}_{(i)}}. \tag{9.28}$$

   But in view of (9.27) this is the characteristic function of $\mathbf{X}_{\mathbf{t_{(i)}}}$, and by (8.10) the expression in (9.28) defines a multivariate normal distribution, with the covariance matrix of $\mathbf{X}_{\mathbf{t_{(i)}}}$ inserted in the quadratic form.

This establishes the consistency condition in (9.1). ■

The message from the above in a nutshell is that

- **There exists a Gaussian process for every symmetric nonnegative definite function** $R(t, s)$.
- **A Gaussian process is uniquely determined by its mean function and its autocorrelation function**.

9.4.2 Weakly Stationary Gaussian Processes

When the property

$$(X(t_1 + h), X(t_2 + h), \ldots, X(t_n + h)) \stackrel{d}{=} (X(t_1), X(t_2), \ldots, X(t_n))$$

holds for all n, all $h \in \mathbf{R}$ and all $t_1, t_2, \ldots, t_n$ points in T for a stochastic process (not necessarily only Gaussian), we call the process **strictly stationary**. In general, weak stationarity does not imply strict stationarity. But if the required moment functions exist, strict stationarity obviously implies weak stationarity. Since the required moment functions exist and uniquely determine the finite dimensional distributions for a Gaussian process, it turns out that a Gaussian process is weakly stationary if and only if it is strictly stationary, as will be demonstrated next.

Theorem 9.4.2 A Gaussian process $\mathbf{X} = \{X_t \mid t \in T =]-\infty, \infty[\}$ is weakly stationary if and only if the property

$$(X(t_1+h), X(t_2+h), \ldots, X(t_n+h)) \overset{d}{=} (X(t_1), X(t_2), \ldots, X(t_n)) \tag{9.29}$$

holds for all n, all h and all $t_1, t_2, \ldots, t_n$ points in T.

Proof $\Rightarrow$: The process is weakly stationary, $(\mu_{\mathbf{X}}(t_1), \cdots, \mu_{\mathbf{X}}(t_n)^{'}$ is a vector with all entries equal to the same constant value, say μ. The entries in $\mathbf{C}(t_1, t_2, \cdots, t_n)$ are of the form

$$R_{\mathbf{X}}\left(|t_i - t_j|\right) - \mu^2.$$

For the same reasons the entries of the mean vector for $(X(t_1+h), \ldots, X(t_n+h))$ are $= \mu$ for all h. Hence the covariance matrix for $(X(t_1+h), \ldots, X(t_n+h))$ has the entries

$$R_{\mathbf{X}}\left(|(t_i+h) - (t_j+h)|\right) - \mu^2 = R_{\mathbf{X}}\left(|t_i - t_j|\right) - \mu^2.$$

That is, $(X(t_1+h), X(t_2+h), \ldots, X(t_n+h))$ and $(X(t_1), X(t_2), \ldots, X(t_n))$ have the same mean vector and same covariance matrix. Since these are vectors with multivariate normal distribution, they have the same distribution.

$\Leftarrow$: If the process is Gaussian, and (9.29) holds, then the desired conclusion follows as above. ■

The computational apparatus mobilized by Gaussian weakly stationary processes is illustrated by the next two examples.

Example 9.4.3 The Gaussian weakly stationary process $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ has expectation function $= 0$ and a.c.f.

$$R_{\mathbf{X}}(h) = \sigma^2 e^{-\lambda|h|}, \quad \lambda > 0.$$

What is the distribution of $(X(t), X(t-1))^{'}$? Since $\mathbf{X}$ is Gaussian and weakly stationary, $(X(t), X(t-1))^{'}$ has a bivariate normal distribution, we need to find the mean vector and the covariance matrix.

The mean vector is found by $E[X(t)] = E[X(t-1)] = 0$. Furthermore we can read the covariance matrix from the autocorrelation function $R_X(h)$. Thereby $E[X(t)X(t-1)] = E[X(t-1)X(t)] = R_X(1) = \sigma^2 e^{-\lambda}$, as $t-(t-1) = 1$, and $E\left[X^2(t)\right] = E\left[X^2(t-1)\right] = R_X(0) = \sigma^2 e^{-\lambda \cdot 0} = \sigma^2$. This says also that $X(t) \in N\left(0, \sigma^2\right)$ and $X(t) \overset{d}{=} X(t-1)$. Thus, the coefficient of correlation is

$$\rho_{X(t),X(t-1)} = \frac{R_X(1)}{R_X(0)} = e^{-\lambda}.$$

Therefore we have established

$$(X(t), X(t-1))^{'} \in N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & e^{-\lambda} \\ e^{-\lambda} & 1 \end{pmatrix}\right).$$

In view of (8.16), we get

$$X(t) \mid X(t-1) = x \in N\left(e^{-\lambda}x, \sigma^2(1-e^{-2\lambda})\right).$$

Then for any real numbers $a < b$

$$\begin{aligned} &\mathbf{P}\left(a < X(t) \le b \mid X(t-1) = x\right) \\ &= \mathbf{P}\left(\frac{a - e^{-\lambda}x}{\sigma\sqrt{1-e^{-2\lambda}}} < \frac{X(t) - e^{-\lambda}x}{\sigma\sqrt{1-e^{-2\lambda}}} \le \frac{b - e^{-\lambda}x}{\sigma\sqrt{1-e^{-2\lambda}}} \mid X(t-1) = x\right) \\ &= \Phi\left(\frac{b - e^{-\lambda}x}{\sigma\sqrt{1-e^{-2\lambda}}}\right) - \Phi\left(\frac{a - e^{-\lambda}x}{\sigma\sqrt{1-e^{-2\lambda}}}\right), \end{aligned}$$

since $\frac{X(t)-e^{-\lambda}x}{\sigma\sqrt{1-e^{-2\lambda}}} \mid X(t-1) = x \in N(0,1)$.

■

Example 9.4.4 The Gaussian weakly stationary process $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ has expectation function $= 0$ and a.c.f.

$$R_{\mathbf{X}}(h) = \frac{1}{1+h^2}.$$

We want to find the probability

$$P\left(3X(1) > 1 - X(2)\right).$$

The standard trick is to write this as

$$\mathbf{P}\left(3X(1) + X(2) > 1\right).$$

Let us first set

$$Y \overset{\text{def}}{=} 3X(1) + X(2).$$

We can use the matrix formulas in the preceding and write Y as

$$Y = B\begin{pmatrix} X(1) \\ X(2) \end{pmatrix} = (3 \quad 1)\begin{pmatrix} X(1) \\ X(2) \end{pmatrix}$$

Since the mean vector of $(X(1), X(2))^{'}$ is the zero vector, we get by (8.6)

$$E[Y] = B\mu_{\mathbf{X}} = B\begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0.$$

Next, the formula in (8.7) entails

$$C_Y = BC_{\mathbf{X}}B^{'}, \tag{9.30}$$

or, since Y is a scalar random variable, its variance is $\text{Var}(Y) = C_Y = BC_{\mathbf{X}}B^{'}$, (9.30) yields

$$BC_{\mathbf{X}}B^{'} = (3 \quad 1)\begin{pmatrix} 1 & \frac{1}{1+1^2} \\ \frac{1}{1+1^2} & 1 \end{pmatrix}\begin{pmatrix} 3 \\ 1 \end{pmatrix}.$$

When we perform the requested matrix multiplications, we obtain

$$\text{Var}(Y) = 13.$$

Hence $Y \in N(0, 13)$. Then the probability sought for is

$$P\left(3X(1) > 1 - X(2)\right) = \mathbf{P}\left(Y > 1\right) = \mathbf{P}\left(\frac{Y}{\sqrt{13}} > \frac{1}{\sqrt{13}}\right) = 1 - \Phi\left(\frac{1}{\sqrt{13}}\right),$$

because $\frac{Y}{\sqrt{13}} \in N(0, 1)$.

■

9.4.3 Distribution of Mean Square Integrals of Gaussian Processes

Let $\{X(t) | t \in T\}$ be a Gaussian stochastic process. It follows that the Riemann sum $\sum_{i=1}^{n} X(t_i)(t_i - t_{i-1})$ is a Gaussian random variable. If the mean square integral exists, then it has to have a normal distribution in view of Theorem 7.4.2 above.

Theorem 9.4.5 If the mean square integral $\int_a^b X(t)dt$ exists for a Gaussian process $\{X(t)|t \in T\}$ for $[a,b] \subseteq T$, then

$$\int_a^b X(t)dt \in N\left(\int_a^b \mu_{\mathbf{X}}(t)dt, \int_a^b \int_a^b \mathrm{Cov}_{\mathbf{X}}(t,u)dtdu\right). \tag{9.31}$$

■

To state an obvious fact, let us note that (9.31) implies that

$$\frac{\int_a^b X(t)dt - \int_a^b \mu_{\mathbf{X}}(t)dt}{\sqrt{\int_a^b \int_a^b \mathrm{Cov}_{\mathbf{X}}(t,u)dtdu}} \in N(0,1).$$

9.5 The Gauss-Markov Processes and Separable Autocorrelation Functions

9.5.1 The Markov Property Defined and Characterized

Let us assume that $\mathbf{X} = \{X(t) \mid t \in T\}$ is a Gaussian stochastic process with zero as mean function and with autocorrelation $R_{\mathbf{X}}(t,s)$. We are not restricting ourselves to weakly stationary processes.

We define now the **Markov property** for $\mathbf{X}$ as follows. For any $t_1 < \ldots < t_{n-1} < t_n$ in T and any $x_1, \ldots, x_{n-1}, x_n$

$$\begin{aligned}\mathbf{P}\left(X(t_n) \leq x_n \mid X(t_1) = x_1, \ldots, X(t_{n-1}) = x_{n-1}\right)\\ = \mathbf{P}\left(X(t_n) \leq x_n \mid X(t_{n-1}) = x_{n-1}\right).\end{aligned} \tag{9.32}$$

The Markov property is saying in other words that the conditional distribution of $X(t_n)$ depends only on $X(t_{n-1})$, not on any chosen history $X(t_1) \ldots, X(t_{n-2})$ prior to t_{n-1}. The Markov property provides an environment for numerous effective algorithms of prediction and Kalman filtering, [90]. We say that a Gaussian process that satisfies (9.32) is a **Gauss-Markov process** .

Another way of writing (9.32) is in view of (3.23) that

$$\mathbf{P}\left(X(t_n) \leq x_n \mid \sigma\left(X(t_1), \ldots, X(t_{n-1})\right)\right) = \mathbf{P}\left(X(t_n) \leq x_n \mid X(t_{n-1})\right),$$

where $\sigma\left(X(t_1), \ldots, X(t_{n-1})\right)$ is the sigma field generated by $X(t_1), \ldots, X(t_{n-1})$.

The process $\mathbf{X}$ has a family of **transition densities** $f_{X(t)|X(t_0)=x_0}(x)$, which are conditional densities so that for $t_0 < t$ and for any Borel set A

$$\mathbf{P}\left(X(t) \in A \mid X(t_0) = x_0\right) = \int_A f_{X(t)|X(t_0)=x_0}(x)dx. \tag{9.33}$$

We shall now characterize the Gauss-Markov processes by a simple and natural property [48, p. 382].

Theorem 9.5.1 The Gaussian process $\mathbf{X} = \{X(t) \mid t \in T\}$ is a Markov process if and only if

$$E\left[X(t_n) \mid X(t_1) = x_1, \ldots, X(t_{n-1}) = x_{n-1}\right] = E\left[X(t_n) \mid X(t_{n-1}) = x_{n-1}\right]. \tag{9.34}$$

Proof If (9.32) holds, then (9.34) obtains per definition of conditional expectation.
Let us assume conversely that (9.34) holds for a Gaussian process $\mathbf{X}$. By properties of Gaussian vectors we know that both

$$\mathbf{P}\left(X(t_n) \leq x_n \mid X(t_1) = x_1, \ldots, X(t_{n-1}) = x_{n-1}\right)$$

and

$$\mathbf{P}\left(X(t_n) \leq x_n \mid X(t_{n-1}) = x_{n-1}\right)$$

are Gaussian distribution functions, and are thus determined by their respective means and variances. We shall now show that these are equal to each other. The statement about means is trivial, as this is nothing but (9.34), which is the assumption.

We introduce some auxiliary notation.

$$\widetilde{Y} \overset{\text{def}}{=} X(t_n) - E\left[X(t_n) \mid X(t_1) = x_1, \ldots, X(t_{n-1}) = x_{n-1}\right]$$

$$= X(t_n) - E\left[X(t_n) \mid X(t_{n-1}) = x_{n-1}\right].$$

But then $\widetilde{Y}$ is the estimation error, when estimating the random variable $X(t_n)$ by $E\left[X(t_n) \mid X(t_{n-1}) = x_{n-1}\right]$, or, which is the same thing here, estimating by $E\left[X(t_n) \mid X(t_1) = x_1, \ldots, X(t_{n-1}) = x_{n-1}\right]$, as expounded in section 3.7.3. We know that for Gaussian random variables we can find $E\left[X(t_n) \mid X(t_{n-1}) = x_{n-1}\right]$ by the projection of $X(t_n)$ to the closed subspace spanned by $X(t_{n-1})$ (and $X(t_1), \ldots, X(t_{n-2})$), as explained after theorem 8.1.9. Then we get by the orthogonality property of projections, see theorem 7.5.3, that for $1 \leq k \leq n-1$

$$E\left[\widetilde{Y} X(t_k)\right] = 0.$$

But $\widetilde{Y}$ and $X(t_k)$ are Gaussian random variables, and $E\left[\widetilde{Y}\right] = 0$ by (3.29). Therefore $\widetilde{Y}$ and $X(t_k)$ are independent for $1 \leq k \leq n-1$ and $\widetilde{Y}$ is independent of (the sigma field spanned by) $X(t_1), \ldots, X(t_{n-1})$ by properties of Gaussian vectors.

Set next $G_k = \{X(t_k) = x_k\}$ and

$$G = G_1 \cap G_2 \cap \ldots \cap G_{n-1}.$$

Then, by the independence just proved

$$E\left[\widetilde{Y}^2 \mid G\right] = E\left[\widetilde{Y}^2 \mid G_{n-1}\right],$$

and this is

$$\text{Var}\left[\widetilde{Y} \mid G\right] = \text{Var}\left[\widetilde{Y} \mid G_{n-1}\right].$$

This says that the distributions in the right and left hand sides of (9.32) have the same variance, and we have consequently proved our assertion as claimed. ■

9.5.2 The Chapman-Kolmogorov (or Smoluchowski) Equation for Transition Densities

Let us take another look at (9.33) with $t_0 < t$, $A =]-\infty, y]$. For any s such that $t_0 < s < t$, marginalization or the law of total probability (3.35) gives

$$\mathbf{P}\left(X(t) \leq y \mid X(t_0) = x\right) = \int_{-\infty}^{\infty} \mathbf{P}\left(X(s) = u, X(t) \leq y, \mid X(t_0) = x\right) du$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{y} f_{X(s),X(t)|X(t_0)=x}\left(u, v\right) dv du = \int_{-\infty}^{y} \int_{-\infty}^{\infty} f_{X(s),X(t)|X(t_0)=x}\left(u, v\right) du dv,$$

and by definition of conditional p.d.f.

$$= \int_{-\infty}^{y} \int_{-\infty}^{\infty} \frac{f_{X(t_0),X(s),X(t)}\left(x, u, v\right)}{f_{X(t_0)}(x)} du dv.$$

Now we invoke twice the definition of conditional p.d.f. (chain rule)

$$f_{X(t_0),X(s),X(t)}\left(x, u, v\right) = f_{X(t)|X(s)=u,X(t_0)=x}\left(v\right) \cdot f_{X(s)|X(t_0)=x}\left(u\right) \cdot f_{X(t_0)}\left(x\right).$$

By the Markov property (9.32) this equals

$$f_{X(t_0),X(s),X(t)}\left(x,u,v\right) = f_{X(t)|X(s)=u}\left(v\right) \cdot f_{X(s)|X(t_0)=x}\left(u\right) \cdot f_{X(t_0)}\left(x\right).$$

If we insert this in the integral above we get

$$\mathbf{P}\left(X(t) \leq y \mid X(t_0) = x\right) = \int_{-\infty}^{y} \int_{-\infty}^{\infty} f_{X(t)|X(s)=u}\left(v\right) \cdot f_{X(s)|X(t_0)=x}\left(u\right) dudv.$$

When we differentiate in both sides of this equality w.r.t. y, we get the following equation for the transition densities

$$f_{X(t)|X(t_0)=x}(y) = \int_{-\infty}^{\infty} f_{X(t)|X(s)=u}\left(y\right) \cdot f_{X(s)|X(t_0)=x}\left(u\right) du. \tag{9.35}$$

It is hoped that a student familiar with finite Markov chains recognizes in (9.35) a certain similarity with the **Chapman-Kolmogorov equation**[7], now valid for probability densities. In statistical physics this equation is called the **Smoluchowski equation**, see [58, p.200]. Regardless of the favoured name, the equation (9.35) can be regarded as a **consistency condition**.

9.5.3 Gauss-Markov Processes and Separable Autocorrelation Functions

Since $\mathbf{X}$ is a Gaussian process with mean zero, we know that (see, e.g.,(8.16))

$$E\left[X(t) \mid X(s) = u\right] = \rho_{X(s),X(t)} \frac{\sigma_{X(t)}}{\sigma_{X(s)}} u =$$

$$= \frac{R_{\mathbf{X}}(t,s)}{\sqrt{R_{\mathbf{X}}(t,t)}\sqrt{R_{\mathbf{X}}(s,s)}} \frac{\sqrt{R_{\mathbf{X}}(t,t)}}{\sqrt{R_{\mathbf{X}}(s,s)}} u$$

i.e,

$$E\left[X(t) \mid X(s) = u\right] = \frac{R_{\mathbf{X}}(t,s)}{R_{\mathbf{X}}(s,s)} u. \tag{9.36}$$

Thus by (9.36) we get for $t_0 < t$

$$\frac{R_{\mathbf{X}}(t,t_0)}{R_{\mathbf{X}}(t_0,t_0)} x_0 = E\left[X(t) \mid X(t_0) = x_0\right] = \int_{-\infty}^{\infty} x \underbrace{f_{X(t)|X(t_0)=x_0}(x)}_{=\int_{-\infty}^{\infty} f_{X(s)|X(t_0)=x_0}(u) \cdot f_{X(t)|X(s)=u}(x) du} dx,$$

and, as indicated, from (9.35)

$$= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X(s)|X(t_0)=x_0}(u) \cdot f_{X(t)|X(s)=u}(x) dudx$$

$$= \int_{-\infty}^{\infty} f_{X(s)|X(t_0)=x_0}(u) \cdot \underbrace{\int_{-\infty}^{\infty} x f_{X(t)|X(s)=u}(x) dx}_{=E[X(t)|X(s)=u]} du$$

$$= \int_{-\infty}^{\infty} f_{X(s)|X(t_0)=x_0}(u) E\left[X(t) \mid X(s) = u\right] du$$

and by two applications of (9.36) in the above

$$= \frac{R_{\mathbf{X}}(t,s)}{R_{\mathbf{X}}(s,s)} \int_{-\infty}^{\infty} u f_{X(s)|X(t_0)=x_0}(u) du$$

[7] This Chapman-Kolmogorov equation for densities was probably first discovered by Louis Bachelier in his theory of speculation, [27].

$$= \frac{R_{\mathbf{X}}(t,s)}{R_{\mathbf{X}}(s,s)} \frac{R_{\mathbf{X}}(s,t_0)}{R_{\mathbf{X}}(t_0,t_0)} x_0,$$

or, equivalently

$$R_{\mathbf{X}}(t,t_0) = \frac{R_{\mathbf{X}}(t,s) R_{\mathbf{X}}(s,t_0)}{R_{\mathbf{X}}(s,s)}. \tag{9.37}$$

Therefore we have found a necessary condition for an autocorrelation function to be the autocorrelation function of a Gaussian Markov process.

Example 9.5.2 Consider a Gaussian process with the autocorrelation function is $R(t,s) = \min(t,s)$. It is shown in an exercise of this chapter that $\min(t,s)$ is an autocorrelation function. Then, if $t_0 < s < t$ we check (9.37) by

$$\frac{R(t,s)R(s,t_0)}{R(s,s)} = \frac{\min(t,s)\min(s,t_0)}{\min(s,s)} = \frac{s \cdot t_0}{s} = t_0,$$

which equals $R(t,t_0) = \min(t,t_0) = t_0$. We shall show in the next chapter that $\min(t,s)$ corresponds, e.g., to the Wiener process, and that the indicated process is a Gaussian Markov process.

■

The equation (9.37) is an example of a **functional equation**, i.e., an equation that in our case specifies the function $R_{\mathbf{X}}(t,s)$ in implicit form by relating the value of $R_{\mathbf{X}}(t,s)$ at a pair of points with its values at other pairs of points. It can be shown [103, p. 72] that if $R_{\mathbf{X}}(t,t) > 0$, then there are functions $v(t)$ and $u(t)$ such that

$$R_{\mathbf{X}}(t,s) = v\left(\max(t,s)\right) u\left(\min(t,s)\right). \tag{9.38}$$

We demonstrate next that (9.38) is sufficient for a Gaussian random process $\mathbf{X}$ to be a Markov process. Let us set

$$\tau(t) = \frac{u(t)}{v(t)}.$$

Because it must hold by Cauchy-Schwartz inequality (7.5) that $R_{\mathbf{X}}(t,s) \leq \sqrt{R_{\mathbf{X}}(t,t) \cdot R_{\mathbf{X}}(s,s)}$, we ge that $\tau(s) \leq \tau(t)$, if $s < t$. Hence, as soon as $\mathbf{X}$ is a Gaussian random process with zero mean and autocovariance function given by (9.38), we can represent it as a **Lamperti transform**

$$X(t) = v(t) W\left(\tau(t)\right), \tag{9.39}$$

where $W(t)$ is a variable in a Gaussian Markov process with autocorrelation function $R_{\mathbf{W}}(t,s) = \min(t,s)$, as in example 9.5.2. To see this we compute for $s < t$

$$E\left[X(t)X(s)\right] = E\left[v(t)W\left(\tau(t)\right) v(s) W\left(\tau(s)\right)\right] =$$

$$= v(t)v(s)E\left[W\left(\tau(t)\right) W\left(\tau(s)\right)\right] =$$

$$= v(t)v(s)\min(\tau(t),\tau(s)) = v(t)v(s)\frac{u(s)}{v(s)} = v(t)u(s) = v\left(\max(t,s)\right) u\left(\min(t,s)\right).$$

Thus, since $\mathbf{W}$ is a Gaussian Markov process (as will be proved in the next chapter) and since $v(t)$ is a deterministic (= non-random) function, the process with variables $X(t) = v(t)W\left(\tau(t)\right)$ is a Gaussian Markov process.

In the preceding, see example 9.1.10, autocorrelation functions of the form $R_{\mathbf{X}}(t,s) = v\left(\max(t,s)\right) u\left(\min(t,s)\right)$ were called **separable**. We have now shown that Gaussian Markov processes lead to separable autocorrelation functions.

Example 9.5.3 Assume next that the Gaussian Markov process is also weakly stationary and mean square continuous. Then $R_{\mathbf{X}}$ is in fact continuous and (9.37) becomes

$$R_{\mathbf{X}}(t-t_0)=\frac{R_{\mathbf{X}}(t-s)R_{\mathbf{X}}(s-t_0)}{R_{\mathbf{X}}(0)}. \tag{9.40}$$

We standardize without loss of generality by $R_{\mathbf{X}}(0)=1$. But then we have

$$R_{\mathbf{X}}(t-t_0)=R_{\mathbf{X}}(t-s)R_{\mathbf{X}}(s-t_0),$$

which is with $x=t-s$, $y=s-t_0$ of the form

$$G(x+y)=G(x)\cdot G(y).$$

This is one of Cauchy's classical functional equations (to be solved w.r.t. $G(\cdot)$). The requirements of autocorrelation functions for weakly stationary processes impose the additional condition $\mid G(x)\mid\leq G(0)$. The continuous autocovariance function that satisfies the functional equation under the extra condition is

$$R_{\mathbf{X}}(h)=e^{-a|h|}. \tag{9.41}$$

In chapter 11 below we shall construct a Gaussian Markov process (the '**Ornstein-Uhlenbeck process**') that, up to a scaling factor, possesses this autocorrelation function.

■

9.6 What Can Not Be Computed by the Methods Introduced Above ?

In the preceding, and later in the exercises, one finds certain straightforward means of computation of probabilities of events that depend on a finite number (most often two) of stochastic variables in a Gaussian process. This is hardly the only kind of situation, where one in practice needs to compute a probability using a random process. We take a cursory look at two reasonable problems, where we are concerned with events that do not depend on a finite or even countable number of random variables. Thus the methods discussed above and in the exercises below do not suffice and development of further mathematical tools is desired.

Let $\{X(t)\mid -\infty\leq t\leq\infty\}$ be a Gaussian stationary stochastic process. There are often reasons to be interested in the **sojourn time**

$$L_b \overset{\text{def}}{=} \text{Length}\left(\{t\mid X(t)\geq b\}\right), \tag{9.42}$$

that is, the time spent at or above a high level b. Or, we might want to find the **extreme value distribution**

$$\mathbf{P}\left(\max_{0\leq s\leq t}X(s)\leq b\right).$$

To hint at what can be achieved, it can be shown, see, e.g., [2, 76], that if

$$R_{\mathbf{X}}(h)\sim 1-\frac{1}{2}\theta t^2, \qquad \text{as } t\to 0,$$

then the sojourn times L_b in (9.42) have approximately the distribution

$$L_b \overset{d}{\approx} \frac{2V}{\theta b}, \tag{9.43}$$

where

$$V \overset{d}{=} \sqrt{\theta} Y,$$

where $Y \in \text{Ra}(2)$ (= Rayleigh distribution with parameter 2). One can also show that

$$\mathbf{P}\left(\max_{0 \leq s \leq t} X(s) \leq b \right) \approx e^{-\lambda_b t},$$

where there is some explicit expression for λ_b, [2]. But to pursue this topic any further is beyond the scope and possibilities of this text.

9.7 Exercises

The bulk of the exercises below consists of specimen of golden oldies from courses related to sf2940 run at KTH once upon time.

9.7.1 Autocovariances and Autocorrelations

1. (From [8, p. 58]) Let $R(t,s) = \min(t,s)$ for $t, s \in T = [0,\infty)$. Show that $R(t,s)$ is a covariance function.

   Aid: The difficulty is to show that the function is nonnegative definite. Use induction. If $t_1 < t_2$, then the matrix

   $$(\min(t_i,t_j))_{i=1,j=1}^{2,2} = \begin{pmatrix} t_1 & t_1 \\ t_1 & t_2 \end{pmatrix}$$

   is symmetric and has the determinant $t_2 t_1 - t_1^2 > 0$.
   Assume that the assertion is true for all $n \times n$ matrices $(\min(t_i,t_j))_{i=1,j=1}^{n,n}$. Then we prove it for t_i for $i = 1, 2, \ldots, n+1$. Arrange or renumber t_i's in increasing order

   $$\min t_i = t_1 \leq t_i \leq t_{i+1}.$$

   Hence

   $$\sum_{i=1}^{n+1} \sum_{j=1}^{n+1} x_i \min(t_i,t_j) x_j =$$

   $$= \sum_{i=2}^{n+1} \sum_{j=2}^{n+1} x_i \min(t_i,t_j) x_j + t_1 x_1^2 + t_1 x_1 \sum_{j=2}^{n+1} x_j.$$

   For $i, j \geq 2$

   $$\min(t_i,t_j) - t_1 = \min(t_i - t_1, t_j - t_1),$$

   and thus

   $$\sum_{i=2}^{n+1} \sum_{j=2}^{n+1} x_i \min(t_i,t_j) x_j - \sum_{i=2}^{n+1} \sum_{j=2}^{n+1} x_i t_1 x_j$$

   $$= \sum_{i=2}^{n+1} \sum_{j=2}^{n+1} x_i \min(t_i - t_1, t_j - t_1) x_j \geq 0$$

   by the induction hypothesis. Now draw the desired conclusion.

2. **Toeplitz Matrices, Toeplitz Forms, Centrosymmetric Matrices**

A **Toeplitz matrix** is defined by the property, a.k.a the Toeplitz property, that the entries on each descending diagonal from left to right are the same ('constant on all diagonals') [47]. Or, if $A = (A_{i,j})_{i=1,j=1}^{n,n}$ is a Toeplitz matrix, then

$$A = \begin{pmatrix} a_0 & a_{-1} & a_{-2} & \ldots & \ldots & a_{-(n-1)} \\ a_1 & a_0 & a_{-1} & \ddots & \ldots & \vdots \\ a_2 & a_1 & a_0 & a_{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & a_{-2} \\ \vdots & \ddots & \ddots & a_1 & a_0 & a_{-1} \\ a_{n-1} & \ldots & \ldots & a_2 & a_1 & a_0 \end{pmatrix}.$$

The Toeplitz property means that

$$A_{i,j} = A_{i+1,j+1} = a_{i-j}.$$

(a) Let $R(h)$ be the autocovariance function of a weakly stationary process and $(R(t_i - t_j))_{i=1,j=1}^{n,n}$ be the covariance matrix (assume zero means) corresponding to equidistant times, i.e., $t_i - t_{i-1} = h > 0$. Convince yourself of the fact that $(R(t_i - t_j))_{i=1,j=1}^{n,n}$ is a Toeplitz matrix. E.g., take one of the autocovariance functions for a weakly stationary process in the text above, and write down the the corresponding covariance matrix for $n = 4$.

(b) An $n \times n$ matrix $A = (a_{ij})_{i=1,j=1}^{n,n}$ is called **centrosymmetric** [8], when its entries a_{ij} satisfy

$$a_{ij} = a_{n+1-i,n+1-j}, \qquad \text{for } 1 \leq i, j \leq n. \tag{9.44}$$

An equivalent way of saying this is that $A = RAR$, where R is the permutation matrix with ones on the cross diagonal (from bottom left to top right) and zero elsewhere, or

$$R = \begin{pmatrix} 0 & 0 & \ldots & 0 & 0 & 1 \\ 0 & 0 & \ldots & 0 & 1 & 0 \\ 0 & 0 & \ldots & 1 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 1 & \ldots & 0 & 0 & 0 \\ 1 & 0 & \ldots 0 & 0 & 0 & 0 \end{pmatrix}.$$

Show that a centrosymmetric matrix is symmetric.

(c) Show that the Toeplitz matrix $(R(t_i - t_j))_{i=1,j=1}^{n,n}$ in (a) above is centrosymmetric. To get a picture of this, take one of the autocovariance functions for a weakly stationary process in the text above, and write down the the corresponding covariance matrix for $n = 4$ and check what (9.44) means.

Therefore we may generalize the class of weakly stationary Gaussian processes by defining a class of Gaussian processes with centrosymmetric covariance matrices.

9.7.2 Examples of Stochastic Processes

1. (From [42]) Let

$$X(t) = \begin{cases} 2 & \text{for all } t \in]-\infty, \infty[\text{ with probability } \frac{1}{2} \\ 1 & \text{for all } t \in]-\infty, \infty[\text{ with probability } \frac{1}{2}. \end{cases}$$

[8] J.R. Weaver: Centrosymmetric (cross-symmetric) matrices, their basic properties, eigenvalues, and eigenvectors. *American Mathematical Monthly*, pp. 711−717, 1985.

Show that $\{X(t) \mid -\infty < t < \infty\}$ is strictly stationary.

2. **The Lognormal Process** (From [8]) $\{X(t)| - \infty < t < \infty\}$ is a weakly stationary Gaussian stochastic process. The process $\mathbf{Y} = \{Y(t)| - \infty < t < \infty\}$ is defined by

$$Y(t) = e^{X(t)}, \quad -\infty < t < \infty.$$

Find the mean function and the autocovariance function of the lognormal process $\mathbf{Y}$.
Aid: Recall the moment generating function of a Gaussian random variable.

3. **The Suzuki Process** Let $\mathbf{X}_i = \{X_i(t)| - \infty < t < \infty\}$, be three ($i = 1, 2, 3$) independent weakly stationary Gaussian processes with mean function zero. $\mathbf{X}_2$ and $\mathbf{X}_3$ have the same autocovariance functions. Let

$$Y(t) = e^{X_1(t)} \cdot \sqrt{X_2^2(t) + X_3^2(t)}.$$

The stochastic process thus defined is known as the Suzuki process[9] and is fundamental in wireless communication (fading distribution for mobile radio) and widely used in dozens of other fields of engineering and science.

Aid: A mnemonic for the Suzuki process is that it is a product of a lognormal process and a Rayleigh process.

(a) Compute $E[Y(t)]$.

(b) Find the p.d.f. of $Y(t)$ (i.e, the Suzuki p.d.f.). Your answer will contain a mixture of densities, of the type eq. (3.7) in [49, p. 41].

4. $\mathbf{X} = \{X(t) \mid -\infty < t < \infty\}$ is a strictly stationary process. Let $g(x)$ be a Borel function. Define a new process $\mathbf{Y} = \{Y(t) \mid -\infty < t < \infty\}$ via

$$Y(t) = g\left(X(t)\right), \quad -\infty < t < \infty.$$

Show that $\mathbf{Y}$ is a strictly stationary process, too.

9.7.3 Autocorrelation Functions

1. [**Periodic Autocorrelation** [60]] Let $\{X(t) \mid -\infty < t < \infty\}$ is a weakly stationary process, with $E[X(t)] = 0$ for all t and autocorrelation $R_X(h)$.

(a) Let $f(t)$ be a function, which is periodic with period T, i.e, $f(t) = f(t + T)$ for all t. Set

$$Y(t) = f(t) \cdot X(t).$$

Show that the process $\mathbf{Y} = \{Y(t) \mid -\infty < t < \infty\}$ is periodically correlated in the sense that

$$R_{\mathbf{Y}}(t, s) = R_{\mathbf{Y}}(t + T, s + T).$$

One can say that the process $\mathbf{Y}$ is produced by amplitude modulation.

(b) Let $f(t)$ be a function, which is periodic with period T, i.e., $f(t) = f(t + T)$ for all t. Set

$$Y(t) = X(t + f(t)).$$

Show that the time modulated process $\{Y(t) \mid -\infty < t < \infty\}$ is periodically correlated. Show also that the variance function is a constant function of time.

[9] H. Suzuki: A statistical model for urban radio propagation, *IEEE Transactions on Communications*, 25, pp. 673–680, 1977.

2. **Band-limited Noise** Show that

$$R(h) = \frac{1}{W_2 - W_1}\left(\frac{\sin(W_2 h)}{h} - \frac{\sin(W_1 h)}{h}\right)$$

is an autocorrelation function.

9.7.4 Weakly Stationary Processes

1. [**Prediction**] $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ is a weakly stationary Gaussian stochastic process with the mean function $\mu_{\mathbf{X}}(t) = m$ and the autocovariance function $\mathrm{Cov}_X(h)$. We want to predict $X(t+\tau)$, $\tau > 0$, by means of a predictor of the form $a \cdot X(t)$. Find a so that

$$E\left[(X(t+\tau) - a \cdot X(t))^2\right]$$

is minimized. Note that the situation is the same as in example 7.5.5 above. Thus check that the optimal parameter a is given by (7.22), or

$$a = \frac{\mathrm{Cov}_{\mathbf{X}}(\tau)}{\mathrm{Cov}_{\mathbf{X}}(0)}.$$

What is the optimal value of $E\left[(X(t+\tau) - a \cdot X(t))^2\right]$?

2. [**An Ergodic Property in Mean Square**] Ergodicity in general means that certain time averages are asymptotically equal to certain statistical averages.

Let $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ be weakly stationary with the mean function $\mu_{\mathbf{X}}(t) = m$. The process $\mathbf{X}$ is mean square continuous. We are interested in the mean square convergence of

$$\frac{1}{t}\int_0^t X(u)du$$

as $t \to \infty$.

(a) (From [50, p.206],[89]) Show that

$$\mathrm{Var}\left(\frac{1}{t}\int_0^t X(u)du\right) = \frac{2}{t}\int_0^t \left(\frac{t-\tau}{t}\right)\mathrm{Cov}_{\mathbf{X}}(\tau)d\tau.$$

(b) Show that if the autocovariance function $\mathrm{Cov}_{\mathbf{X}}(h)$ is such that

$$\mathrm{Cov}_{\mathbf{X}}(h) \to 0, \quad \text{as } h \to \infty,$$

then we have

$$\lim_{t\to\infty} E\left[\left(\frac{1}{t}\int_0^t X(u)du - m\right)^2\right] = 0.$$

Hint: The result in (a) should be useful.

3. (From [57]) Let $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ be weakly stationary with the mean function $\mu_{\mathbf{X}}(t) = m$ and autocovariance function $\mathrm{Cov}_{\mathbf{X}}(h)$ such that

$$\mathrm{Cov}_{\mathbf{X}}(h) \to 0, \quad \text{as } h \to \infty.$$

Define a new process $\mathbf{Y} = \{Y(t) | 0 \leq t < \infty\}$ by

$$Y(t) = X(t) - X(0).$$

Find the autocovariance function $\text{Cov}_{\mathbf{Y}}(t,s)$ and show that

$$\lim_{t\to\infty} \text{Cov}_{\mathbf{Y}}(t,t+h) = \text{Cov}_{\mathbf{X}}(h) + \text{Cov}_{\mathbf{X}}(0).$$

In this sense $\mathbf{Y}$ becomes asymptotically weakly stationary.

4. (From [57] Let $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ be a Gaussian random process such that

$$(X(t), X(s))' \in N\left(\begin{pmatrix} \alpha+\beta t \\ \alpha+\beta s \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & e^{-\lambda|t-s|} \\ e^{-\lambda|t-s|} & 1 \end{pmatrix}\right).$$

This is obviously **not** a weakly stationary process, as there is a **linear trend** in the mean function $\mu_{\mathbf{X}}(t)$. Let us define a new process $\mathbf{Y}$ by **differencing**, by which we mean the operation

$$Y(t) = X(t) - X(t-1).$$

Show that the process $\mathbf{Y}$ is strictly stationary.

Aid: One of the intermediate results you should obtain here is that

$$\text{Cov}_{\mathbf{Y}}(t,s) = \sigma^2 \left[2 \cdot e^{-\lambda|t-s|} - e^{-\lambda|(t-s)+1|} - e^{-\lambda|(t-s)-1|}\right].$$

Comment: Differencing removes here a linear trend and produces a stationary process. This recipe, called **de-trending**, is often used in time series analysis.

5. (From [42]) Let $s(f)$ be a real valued function variable that satisfies

$$s(f) \geq 0, \quad s(-f) = s(f), \tag{9.45}$$

and

$$\int_{-\infty}^{\infty} s(f)df = K < \infty. \tag{9.46}$$

Let X_1, X_2, Y_1 and Y_2 be independent stochastic variables such that

$$E[X_1] = E[X_2] = 0, \quad E\left[X_1^2\right] = E\left[X_2^2\right] = \frac{K}{2\pi},$$

and Y_1 and Y_2 have the p.d.f. $f_Y(y) = \frac{s(y)}{K}$.

Show that the process $\{X(t) | -\infty < t < \infty\}$ defined by

$$X(t) = X_1 \cos(Y_1 t) + X_2 \cos(Y_2 t)$$

is weakly stationary and has the spectral density $s(f)$.

We have in this manner shown that if a function $s(f)$ satisfies (9.45) and (9.46), then there exists at least one stochastic process that has $s(f)$ as spectral density.

6. (From [42]) $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ is a weakly stationary process and $Z \in U(0, 2\pi)$ is independent of $\{\mathbf{X}\}$. Set

$$Y(t) = \sqrt{2}X(t)\cos(f_o t + Z) \quad -\infty < t < \infty.$$

Show that $\mathbf{Y} = \{Y(t) | -\infty < t < \infty\}$ has mean function $\mu_{\mathbf{Y}} = 0$, and

$$R_{\mathbf{Y}}(h) = \left(R_{\mathbf{X}}(h) + \mu_{\mathbf{X}}^2\right)\cos(f_o h).$$

Comment This is a mathematical model for amplitude modulation, c.f., [71, kap.7].

9.7.5 Gaussian Stationary Processes

1. (From [42]) $\{X(t) | -\infty < t < \infty\}$ is a weakly stationary Gaussian stochastic process with

$$\mu_{\mathbf{X}} = 0, \quad R_{\mathbf{X}}(h) = \frac{1}{1+h^2}.$$

   (a) The probability

$$P\left(3X(1) > 1 - X(2)\right) = Q\left(\frac{1}{\sqrt{13}}\right),$$

   where $Q(x)$ is the error function in (2.21), was found in the example 9.4.4. Re-verify this without applying any matrix formula like (8.7). You should, of course, use $Y = 3X(1) + X(2)$.

   (b) Check that

$$P\left(\mid X(2) - \left(\frac{X(1)+X(3)}{2}\right) \mid > 1\right) = 2 \cdot Q\left(\frac{1}{\sqrt{0.6}}\right).$$

   where $Q(x)$ is the error function in (2.21).

   (c) Check that

$$P\left(X(2) - X(1) > 1 \mid X(3) - X(0) = 1\right) = Q\left(\frac{1-\frac{1}{3}}{\sqrt{0.8}}\right).$$

2. (From [42]) $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ is a weakly stationary Gaussian stochastic process with

$$\mu_{\mathbf{X}} = m,$$

which is an unknown statistical parameter, and

$$\mathrm{Cov}_{\mathbf{X}}(h) = \begin{cases} 1 - \frac{|h|}{2} & \text{if } |h| < 2 \\ 0 & \text{if } |h| \geq 2. \end{cases}$$

We try to estimate the mean m by time discrete samples of $\mathbf{X}$ via

$$Y_n = \frac{1}{n}\sum_{k=1}^{n} X(k).$$

Show that

   (a)

$$E\left[Y_n\right] = m$$

   (b)

$$\mathrm{Var}\left[Y_n\right] = \frac{2n-1}{n^2}.$$

   (c) for $\varepsilon > 0$

$$P\left(\mid Y_n - m \mid \leq \varepsilon\right) = \Phi\left(\frac{\varepsilon}{\frac{2n-1}{n^2}}\right) - \Phi\left(-\frac{\varepsilon}{\frac{2n-1}{n^2}}\right).$$

   Is it true that $Y_n \xrightarrow{\mathrm{P}} m$, as $n \to \infty$?

3. [**Bandlimited Gaussian White Noise**] A weakly stationary stationary Gaussian process $\mathbf{Z} = \{Z(t) \mid -\infty < t < \infty\}$ with mean zero that has the power spectral density

$$s_{\mathbf{Z}}(f) = \begin{cases} \frac{N_o}{2} & -W \leq f \leq W, \\ 0 & \text{elsewhere}, \end{cases}$$

is called bandlimited white noise. W is referred to as the bandwidth (in radians).

(a) Show that the a.c.f. of $\mathbf{Z}$ is

$$R_{\mathbf{Z}}(h) = N_o W \cdot \frac{\sin(Wh)}{\pi W h}.$$

(b) Sample the process $\mathbf{Z}$ at time points $\frac{\pi k}{W}$ for $k = 0, \pm 1, \pm 2, \ldots$, so that

$$Z_k = Z\left(\frac{\pi k}{W}\right).$$

Find the autocorrelations

$$r_{k,l} = E\left[Z_k \cdot Z_l\right].$$

(c) Show that for any t

$$\sum_{k=-n}^{k=n} Z_k \cdot \frac{\sin(W(t-\frac{\pi k}{W}))}{\pi W(t-\frac{\pi k}{W})} \overset{2}{\to} Z(t),$$

as $n \to \infty$. This is a stochastic version, [85, pp. 332– 336], of the celebrated **sampling theorem**[10], [100, pp. 187]. It predicts that we can reconstruct completely the band-limited process $\mathbf{Z}$ from its time samples $\{Z_k\}_{k=-\infty}^{\infty}$, also known as Nyquist samples.

Aid: (C.f. [103, p. 106]). The following result ('Shannon's sampling theorem') on covariance interpolation is true (and holds in fact for all bandlimited functions)

$$R_{\mathbf{Z}}(h) = \sum_{k=-\infty}^{\infty} R_{\mathbf{Z}}\left(\frac{\pi k}{W}\right) \cdot \frac{\sin(W(t-\frac{\pi k}{W}))}{\pi W(t-\frac{\pi k}{W})},$$

and can be shown by an application of some Fourier transforms.

As for proving the stochastic version, the complex periodic function e^{ift} is first expanded as a Fourier series

$$e^{ift} = \sum_{k=-\infty}^{\infty} e^{ikft} \frac{\sin(W(t-\frac{\pi k}{W}))}{\pi W(t-\frac{\pi k}{W})}, \tag{9.47}$$

which is uniformly convergent in the interval $\mid f \mid \leq W$.

Then we study

$$E\left[\left(Z(t) - \sum_{k=-n}^{k=n} Z_k \cdot \frac{\sin(W(t-\frac{\pi k}{W}))}{\pi W(t-\frac{\pi k}{W})}\right)^2\right]$$

and obtain (check this)

$$= R_{\mathbf{Z}}(0) - 2\sum_{k=-n}^{k=n} R_{\mathbf{Z}}\left(t - \frac{\pi k}{W}\right) \cdot \frac{\sin(W(t-\frac{\pi k}{W}))}{\pi W(t-\frac{\pi k}{W})}$$

$$+ \sum_{k=-n}^{k=n}\sum_{j=-n}^{j=n} R_{\mathbf{Z}}\left((j-k)\frac{\pi}{W}\right) \cdot \frac{\sin(W(t-\frac{\pi k}{W}))}{\pi W(t-\frac{\pi k}{W})} \frac{\sin(W(t-\frac{\pi j}{W}))}{\pi W(t-\frac{\pi j}{W})}.$$

We represent this by power spectral densities and get (verify)

$$= \int_{-W}^{W} \mid e^{ift} - \sum_{k=-n}^{n} e^{ikft} \frac{\sin(W(t-\frac{\pi k}{W}))}{\pi W(t-\frac{\pi k}{W})} \mid^2 s_{\mathbf{Z}}(f) df.$$

Then the conclusion follows by (9.47).

[10] http://en.wikipedia.org/wiki/Nyquist-Shannon_sampling_theorem

9.7.6 Mean Square Integrals of Processes

1. (From [42]) $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ is a weakly stationary stochastic process with
$$\mu_{\mathbf{X}} = 2, \quad \text{Cov}_{\mathbf{X}}(h) = e^{-|h|}.$$
Set
$$Y = \int_0^1 X(t)dt.$$
Check that
$$E[Y] = 2, \text{Var}(Y) = \frac{2}{e}.$$

2. [**Linear Time Invariant Filters**] Let $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ be a stochastic process with zero as mean function and with the autocorrelation function $R_{\mathbf{X}}(t,s)$. Let
$$Y(t) = \int_{-\infty}^{\infty} G(t-s)X(s)ds = \int_{-\infty}^{\infty} G(s)X(t-s)ds,$$
assuming existence. One can in fact show that the two mean square integrals above are equal (almost surely).

   (a) Check that
$$R_{\mathbf{Y}}(t,s) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(t-u)G(s-v)R_X(u,v)dudv. \tag{9.48}$$

   (b) Show that if $\mathbf{X}$ is (weakly) stationary with zero as mean function and
$$Y(t) = \int_{-\infty}^{\infty} G(t-s)X(s)ds,$$
then $\mathbf{Y} = \{Y(t) | -\infty < t < \infty\}$ is (weakly) stationary.

   (c) Assume that $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ is a Gaussian weakly stationary stochastic process and with the autocorrelation function $R_{\mathbf{X}}(h)$ and
$$Y(t) = \int_{-\infty}^{\infty} G(t-s)X(s)ds.$$
Show that $\{Y(t) | -\infty < t < \infty\}$ is a Gaussian process and find the distribution of $Y(t)$ for any t.

   Remark 9.7.1 The findings in this exercise provide a key, viz. the mathematical representations of analog filters, for understanding of the pre-eminence of weakly stationary processes in [50, 56, 71, 80, 85, 97, 101]. One thinks of $G(t)$ as the **impulse response** of a **linear time-invariant filter** with the process as $\{X(t) | -\infty < t < \infty\}$ input and the process $\{Y(t) | -\infty < t < \infty\}$ as output. An instance of applications is described in *IEEE standard specification format guide and test procedure for single- axis interferometric fiber optic gyros,* IEEE Std 952-1997(R2008), c.f. Annexes B & C, 1998.

   ■

3. [**The Superformula**](From [71, 101] and many other texts) Let $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ be a weakly stationary stochastic process with zero as mean function and with the autocorrelation function $R_{\mathbf{X}}(h)$. Let
$$Y(t) = \int_{-\infty}^{\infty} G(t-s)X(s)ds = \int_{-\infty}^{\infty} G(s)X(t-s)ds,$$

assuming existence. Suppose that the spectral density of $\mathbf{X}$ is $s_{\mathbf{X}}(f)$. Show that the spectral density of $\mathbf{Y} = \{Y(t) | -\infty < t < \infty\}$ (recall the preceding exercise showing that $\mathbf{Y}$ is weakly stationary) is

$$s_{\mathbf{Y}}(f) = | g(f) |^2 \, s_{\mathbf{X}}(f), \quad -\infty < f < \infty, \tag{9.49}$$

where $g(f)$ is the **transfer function**

$$g(f) = \int_{-\infty}^{\infty} e^{-ifh} G(h) dh.$$

Note the connection of (9.49) to (9.48). In certain quarters at KTH the formula in (9.49) used to be referred to in Swedish as the **superformel**.

4. (From [42]) $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ is a weakly stationary process with mean function $= \mu_{\mathbf{X}}$ and with the autocorrelation function $R_{\mathbf{X}}(h) = \sigma^2 e^{-|h|}$. Let

$$G(t) = \begin{cases} e^{-2t} & 0 \leq t \\ 0 & t < 0. \end{cases}$$

Set

$$Y(t) = \int_{-\infty}^{\infty} G(t-s) X(s) ds.$$

Show that if $\mathbf{Y} = \{Y(t) | -\infty < t < \infty\}$, then

$$\mu_{\mathbf{Y}} = \frac{\mu_X}{2},$$

$$R_{\mathbf{Y}}(h) = \frac{\sigma^2}{2}\left(2e^{-|h|} - e^{-|h|}\right),$$

and that the spectral density is

$$s_{\mathbf{Y}}(f) = \frac{2\sigma^2}{(f^2+4)(f^2+1)}.$$

5. (From [71]) $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ is a weakly stationary process with zero as mean function and with the autocorrelation function $R_{\mathbf{X}}(h) = e^{-c|h|}$. Let

$$G(t) = \begin{cases} \frac{1}{T} & 0 \leq t \leq T \\ 0 & \text{otherwise.} \end{cases}$$

Set

$$Y(t) = \int_{-\infty}^{\infty} G(t-s) X(s) ds.$$

Show that

$$R_{\mathbf{Y}}(h) = \begin{cases} \frac{2}{c^2T^2}\left[(C(T-|h|)) - e^{-c|h|} + e^{-cT}\cosh(ch)\right] & |h| < T \\ \frac{2}{c^2T^2} e^{-c|h|}\left(\cosh(ch) - 1\right) & |h| \geq T. \end{cases}$$

6. (From [42]) $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ is a Gaussian stationary process with $\mu_{\mathbf{X}} = 1$ as mean function and with the autocorrelation function $R_{\mathbf{X}}(h) = e^{-h^2/2}$. Show that

$$\int_{-\infty}^{\infty} X(t) e^{-t^2/2} dt \in N\left(\sqrt{2\pi}, \frac{2\pi}{\sqrt{3}}\right).$$

7. (From [105]) Let $\mathbf{W} = \{W(t) \mid t \geq 0\}$ be a Gaussian process with mean function zero and the autocorrelation (i.e., autocovariance)

$$R_{\mathbf{W}}(t,s) = \min(t,s).$$

(a) Show that the mean square integral

$$Y(t) = \int_0^t W(s)ds$$

exists for all $t \geq 0$.

(b) Show that the autocovariance function of the process $\mathbf{Y} = \{Y(t) \mid t \geq 0\}$ is

$$R_{\mathbf{Y}}(t,s) = \begin{cases} \frac{s^2(3t-s)}{6} & t \geq s \\ \frac{t^2(3s-t)}{6} & t < s. \end{cases}$$

9.7.7 Mean Square Continuity

1. Show that if the autocovariance function of a weakly stationary process is continuous at the origin, then it is continuous everywhere.
   Aid: Apply in a suitable manner the Cauchy-Schwarz inequality, eq. (7.5), in the preceding.

2. A stochastic process $\{X(t) \mid t \in T\}$ is said to be **continuous in probability**, if it for every $\varepsilon > 0$ and all $t \in T$ holds that

   $$\mathbf{P}\left(\mid X(t+h) - X(t) \mid > \varepsilon\right) \to 0,$$

   as $h \to 0$. Let $\{X(t) \mid t \in T\}$ be a weakly stationary process. Suppose that the autocovariance function is continuous at the origin. Show that then the process is continuous in probability.
   Aid: Recall Markov's inequality (1.38).

9.7.8 Memoryless Nonlinear Transformations of Gaussian Processes

Introduction

By a **memoryless** nonlinear transformation of a stochastic process $\{X(t)|t \in T\}$, we mean a stochastic process $\{Y(t)|t \in T\}$ defined by

$$Y(t) = Q\left(X(t)\right), \quad t \in T,$$

where Q is a Borel function.

In order to give a specific example, in lightwave technology[11] a lot of attention is paid to **clipping**, c.f., section 8.5.3, where

$$Q\left(x\right) = \begin{cases} x & \text{if } \mid x \mid < x_o \\ x_o \cdot \text{sign(x)} & \text{if } \mid x \mid > x_o. \end{cases}$$

For this to be of interest, it is argued, c.f. [33, p.212−217], that a stationary Gaussian process can represent a broadband analog signal containing many channels of audio and video information (e.g., cabletelevision signals over optical fiber).

Let us incidentally note that in the context of clipping (e.g., of laser) it is obviously important for the engineer to know the distribution of the **sojourn time** in (9.42) or

$$L_{x_o} \overset{\text{def}}{=} \text{Length}\left(\{t \mid X(t) \geq x_0\}\right). \tag{9.50}$$

The approximation in (9.43) is well known to be the practical man's tool for this analysis.

[11] see, e.g., A.J. Rainal: Laser Intensity Modulation as a Clipped Gaussian Process. *IEEE Transactions on Communications*, Vol. 43, 1995, pp. 490−494.

By a nonlinear transformation of a stochastic process $\mathbf{X} = \{X(t) | t \in T\}$ **with memory** we mean, for one example, a stochastic process $\mathbf{Y} = \{Y(t) | t \in T\}$ defined by

$$Y(t) = \int_0^t Q\left(X(s)\right) ds, \quad [0, t] \subset T,$$

where Q is an integrable Borel function. In this case the value of $Y(t)$ depends on the process $\mathbf{X}$ between 0 and t, i.e., has at time t a memory of the 'past' of the process. One can also say that $Y(t)$ is a nonlinear **functional** of the process $\mathbf{X}$ over $[0, t]$.

Exercises, Hermite Expansions

1. $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ is a weakly stationary Gaussian stochastic process with $\mu_{\mathbf{X}} = 0$, variance $\sigma^2_{\mathbf{X}}$ and autocorrelation $R_{\mathbf{X}}(h)$. Let for every t

$$Y(t) = X^2(t).$$

   Let the cofficient of correlation between $X(s)$ and $X(t)$, $h = t - s$, be

$$\rho_{\mathbf{X}}(h) = \frac{R_{\mathbf{X}}(h)}{R_{\mathbf{X}}(0)}$$

   Show that the autocorrelation function of $\mathbf{Y} = \{Y(t) | -\infty < t < \infty\}$ is

$$R_{\mathbf{Y}}(h) = \sigma^4_{\mathbf{X}} \cdot \left[1 + 2\left(\rho_{\mathbf{X}}(h)\right)^2\right] \tag{9.51}$$

   Hint: The four product rule in (8.37) of section 8.5 can turn out to be useful.

2. $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ is a weakly stationary Gaussian stochastic process with $\mu_{\mathbf{X}} = 0$, variance $\sigma^2_{\mathbf{X}}$ and autocorrelation $R_{\mathbf{X}}(h)$ such that $R_{\mathbf{X}}(0) = 1$. We observe a binarization of the process, c.f., (8.43),

$$Y(t) = \begin{cases} 1 & X(t) \geq 0, \\ -1 & X(t) < 0. \end{cases} \tag{9.52}$$

   Show that

$$R_{\mathbf{X}}(h) = \sin\left(\frac{\pi}{2} R_{\mathbf{Y}}(h)\right).$$

   Aid: (From [89]). You may use the fact that (c.f., (8.24))

$$\int_0^\infty \int_0^\infty \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(x^2 - 2\rho xy + y^2\right)} dxdy = \frac{1}{4} + \frac{\arcsin(\rho)}{2\pi}.$$

3. **Hermite Expansions of Nonlinearities** We shall next present a general technique for dealing with a large class of memoryless nonlinear transformations of Gaussian processes. This involves the first properties of *Hermite polynomials*, as discussed in section 2.6.2. The next theorem forms the basis of analysis of non-linearities in section 9.7.8.

   Theorem 9.7.1 Suppose $h(x)$ is a function such that

$$\int_{-\infty}^{\infty} h^2(x) e^{-x^2/2} dx < \infty. \tag{9.53}$$

   Then

$$h(x) = \sum_{n=0}^{\infty} \frac{c_n}{n!} H_n(x), \tag{9.54}$$

where

$$c_n = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(x) e^{-x^2/2} H_n(x) dx, \quad n = 0, 1, 2, \ldots, \tag{9.55}$$

and the series converges with respect to the Hilbert space norm

$$\parallel f \parallel = \sqrt{\int_{-\infty}^{\infty} f^2(x) e^{-x^2/2} dx}.$$

■

We have the following result.

Theorem 9.7.2 Let $\mathbf{X} = \{X(t) | -\infty < t < \infty\}$ be a weakly stationary process with the mean value function $= 0$. Let $Q(x)$ satisfy (9.53) and define

$$Y(t) = Q(X(t)), \quad -\infty < t < \infty. \tag{9.56}$$

Then $\mathbf{Y} = \{Y(t) | -\infty < t < \infty\}$ has the autocorrelation function

$$R_{\mathbf{Y}}(h) = \sum_{n=0}^{\infty} \frac{C_n^2}{n!} \left(\frac{R_{\mathbf{X}}(h)}{R_{\mathbf{X}}(0)} \right)^n, \tag{9.57}$$

where

$$C_n = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} Q\left(x\sigma_{\mathbf{X}}\right) e^{-x^2/2} H_n(x) dx.$$

Proof is outlined. These assertions are derived by **Mehler's Formula** [24, p.133], which says the following. Let

$$\left(X_1, X_2\right)^{'} \in N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

Then the joint p.d.f. of $\left(X_1, X_2\right)^{'}$ is by (8.24)

$$f_{X_1,X_2}\left(x_1, x_2\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2}\frac{1}{(1-\rho^2)} \cdot \left[x_1^2 - 2\rho \cdot x_1 x_2 + x_2^2\right]}$$

and can be written as

$$f_{X_1,X_2}\left(x_1, x_2\right) = \frac{e^{-x_1^2/2}}{\sqrt{2\pi}} \frac{e^{-x_2^2/2}}{\sqrt{2\pi}} \sum_{n=0}^{\infty} \frac{\rho^n}{n!} H_n\left(x_1\right) H_n\left(x_2\right). \tag{9.58}$$

Hence, from (9.56), with $h = t - s$,

$$R_{\mathbf{Y}}(h) = E\left[Y(t)Y(s)\right] = E\left[Q(X(t))Q(X(s))\right] =$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Q\left(x_1\right) Q\left(x_2\right) f_{X_t,X_s}\left(x_1, x_2\right) dx_1 dx_2. \tag{9.59}$$

Now we use in (9.59) the expansions (9.54), (9.58), (2.100) and (2.101) to obtain (9.57) in the special case $\sigma_X = 1$.

■

(a) Verify by symbol manipulation (i.e., do not worry about the exchange of order between integration and the infinite sums) that (9.59) leads to (9.57), as indicated in the last lines of the proof outlined above.

(b) Let now

$$Q(x) = ax^3.$$

(i) Show that (9.53) is satisfied.

(ii) If $\mathbf{X}$ and $\mathbf{Y}$ are as above with $Q(x) = ax^3$ in (9.56), then show that

$$R_{\mathbf{Y}}(h) = 3a^2\sigma_{\mathbf{X}}^6 \left[3\frac{R_{\mathbf{X}}(h)}{R_{\mathbf{X}}(0)} + 2\left(\frac{R_{\mathbf{X}}(h)}{R_{\mathbf{X}}(0)}\right)^3 \right].$$

(c) Let now

$$Q(x) = x^2.$$

Verify that You get the result in (9.51) using (9.57).

9.7.9 Separable Autocovariances

Let us consider $T = [0, 1]$ and

$$R(t, s) = \begin{cases} s(1-t) & s \le t \\ (1-s)t & s \ge t. \end{cases} \tag{9.60}$$

1. Check, whether (9.37) holds for $R(t, s)$ in (9.60).

2. Let $W(t)$ be a random variable in a Gaussian process (i.e., Wiener process, see next chapter) with the autocorrelation function $R_{\mathbf{W}}(t, s) = \min(t, s)$. Find a function $\tau(t) = u(t)/v(t)$ so that the process defined by

$$B(t) = v(t)W\left(\tau(t)\right), \tag{9.61}$$

has the autocorrelation function $R(t, s)$ in (9.60).

3. Find functions $s(t)$ and $h(t)$ such that for $0 \le t < \infty$

$$W(t) = h(t)B\left(s(t)\right), \tag{9.62}$$

where $W(t)$ and $B(t)$ are as in (9.61).

Chapter 10

The Wiener Process

10.1 Introduction

10.1.1 Background: The Brownian Movement, A. Einstein

The British botanist Robert Brown examined[1] in the year 1827 pollen grains and the spores of mosses suspended in water under a microscope, see figure **??**, and observed minute particles within vacuoles in the pollen grains executing a continuous jittery motion. Although he was not the first to make this observation, the phenomenon or the movement became known as the **Brownian movement**.

A note on terminology is at place here. Here we shall refer to the physical phenomenon as the Brownian movement, and to the mathematical model, as derived below, as Brownian motion/Wiener process thus minding of the accusations about 'mind projection fallacies'.

In one of his three great papers published in 1905 Albert Einstein carried a probabilistic analysis of molecular motion and its effect on particles suspended in a liquid. Einstein admits to begin with [31, p.1] that he does not know much of Brown's movement. His purpose was not, as pointed out by L. Cohen[2], to explain the Brownian movement but to prove that atoms existed. In 1905, many scientists did not believe in atomic theory. Einstein's approach was to derive a formula from the atomic theory, and to expect that someone performs the experiments that verify the formula.

In the 1905 paper, see [31], Einstein derives the governing equation for the p.d.f. $f(x,t)$ of the particles, which are influenced by the invisible atoms. The equation of evolution of $f(x,t)$ is found as

$$\frac{\partial}{\partial t} f(x,t) = D \frac{\partial^2}{\partial x^2} f(x,t).$$

Two brief and readable and non-overlapping recapitulations of Einstein's argument for this are [58, pp. 231−234] and [78, chapter 4.4]. Then Einstein goes on to show that the square root of the expected squared displacement of the particle is **proportional to the square root of time** as

$$\sigma_X = \sqrt{E\left[X(t)^2\right]} = \sqrt{2Dt}. \tag{10.1}$$

It is generally overlooked that Einstein's 'coarse time' approach to thermodynamics implies that his finding in (10.1) is valid only for very large t. Then Einstein derives the formula for the diffusion coefficient D as

$$D = \frac{RT}{N} \frac{1}{6\pi k P}, \tag{10.2}$$

[1] A clarification of the intents of Brown's work and a demonstration that Brown's microscope was powerful enough for observing movements so small is found in `http://www.brianjford.com/wbbrowna.htm`

[2] The History of Noise. *IEEE Signal Processing Magazine*, vol. 1053, 2005.

where R is the gas constant, T is the temperature, k is the coefficient of viscosity (Einstein's notation) and P is the radius of the particle. The constant N had in 1905 no name, but was later named **Avogadro's number**[3], see [58, p. 236]. Next Einstein explains how to estimate N from statistical measurements. We have

$$\sigma_X = \sqrt{t}\sqrt{\frac{RT}{N}\frac{1}{3\pi kP}}$$

so that

$$N = \frac{1}{\sigma_X^2}\frac{RT}{3\pi kP},$$

where σ_X^2 is measured 'per minute'.

Besides the formulas and ideas stated, Einstein invoked the Maxwell & Boltzmann statistics, see, e.g., [17, p. 39, p. 211], [58, chapter 6], and saw that the heavy particle is just a big atom pushed around by smaller atoms, and according to energy equipartition, c.f., [17, chapter 19], the statistical properties of the big particle are the same as of the real invisible atoms. More precisely, the mean kinetic energy of the pollen is the same as the mean kinetic energy of the atoms. Therefore we can use the heavy particle as a probe of the ones we cannot see. If we measure the statistical properties of the heavy particle, we know the statistical properties of the small particles. Hence the atoms exist by the erratic movements of the heavy particle[4].

J.B. Perrin[5] was an experimentalist, who used (amongst other experimental techniques) direct measurements of the mean square displacement and Einstein's formula to determine Avogadro's number, and was awarded Nobel Prize in Physics in 1926 in large part, it is said, due to this. Actually, Perrin proceeded to determine Boltzmann's constant and the electronic charge by his measurement of Avogadro's number, [58, p. 239].

10.1.2 Diffusion, Theory of Speculation & the Wiener Process

Another physical description of the background to the mathematical model to be introduced and analysed in this chapter is diffusion of microscopic particles. There are two aspects in a diffusion: very rough particle trajectories, c.f., figure 10.2, at the microscopic level giving rise to a very smooth behaviour of the density of an entire cloud of particles.

The Wiener process[6] $\mathbf{W} = \{W(t) \mid t \geq 0\}$ to be defined below is a mathematical device designed as a model of the motion of individual particles in a diffusion. The paths of the Wiener process exhibit an erratic behaviour, while the density $f_{W(t)}$ of the random variable $W(t)$ is for $t > 0$ given by

$$f_{W(t)}(x) = \frac{1}{\sqrt{2\pi t}} e^{-\dfrac{x^2}{2t}}.$$

We set $p(t,x) = \frac{1}{\sqrt{2\pi t}} e^{-\dfrac{x^2}{2t}}$. Then $p(t,x)$ is the solution of the partial differential equation known as the diffusion (or the heat) equation [96, pp.130−134]

$$\frac{\partial}{\partial t}p(t,x) = \frac{1}{2}\frac{\partial^2}{\partial x^2}p(t,x), \tag{10.3}$$

[3] The Avogadro constant expresses the number of elementary entities per mole of substance, c.f. [17, p.3].

[4] More on this and the history of stochastic processes is found in L. Cohen: The History of Noise. *IEEE Signal Processing Magazine*, vol. 1053, 2005.

[5] Jean Baptiste Perrin 1870-1942, Perrin's Nobel lecture with a discussion of Einstein's work and Brownian movement is found on
`http://nobelprize.org/nobel_prizes/physics/laureates/1926/perrin-lecture.html`

[6] is named after Norbert Wiener, 1894−1964, who constructed it as a stochastic process in mathematical terms, as given here, and proved that the process has continuous sample paths that are nowhere differentiable.
`http://www-groups.dcs.st-and.ac.uk/~history/Biographies/Wiener_Norbert.html`

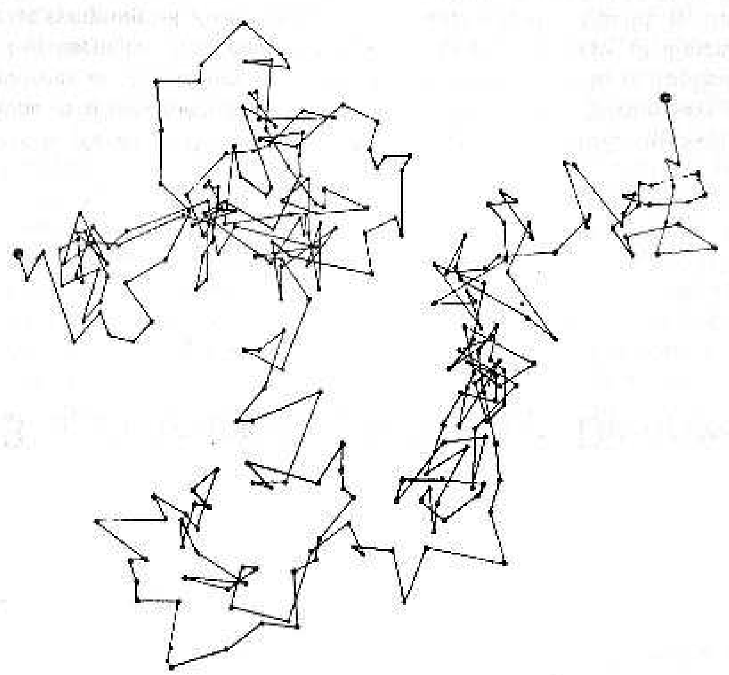

Figure 10.1: A Path of a Brownian Movement Particle

and can be interpreted as the density (in fact p.d.f.) at time t of a cloud issuing from a single point source at time 0.

We shall study the one-dimensional Wiener process starting from the mathematical definition in 10.2.1 below and derive further properties from it. The Wiener process can be thought of as modelling the projection of the position of the Brownian particle onto one of the axes of a coordinate system. A sample path of the one-dimensional Wiener process is given in figure 10.3. In the literature, especially that emanating from British universities, see, e.g., [26], this stochastic process is also known as the **Brownian motion**.

Apart from describing the motion of diffusing particles, the Wiener process is widely applied in mathematical models involving various noisy systems, for example asset pricing at financial markets, c.f. [13, chapter 4].

Actually, Louis Bachelier (1870−1946)[7] is nowadays acknowledged as the first person to define the stochastic process called the Wiener process. This was included in his doctoral thesis with the title *Théorie de la spéculation*, 1900[8] reprinted, translated and commented in [27]. This thesis, which treated Wiener process to evaluate stock options, is historically the first contribution to use advanced mathematics in the study of finance. Hence, Bachelier is appreciated as a pioneer in the study of both financial mathematics and of stochastic processes.

[7] `http://www-groups.dcs.st-and.ac.uk/~history/Biographies/Bachelier.html`

[8] R. Mazo, an expert in statistical mechanics, chooses to write in [78, p. 4]:

> The subject of the thesis (by Bachelier) was a stochastic theory of speculation on the stock market, hardly a topic likely to excite interest among physical scientists (or among mathematicians either).

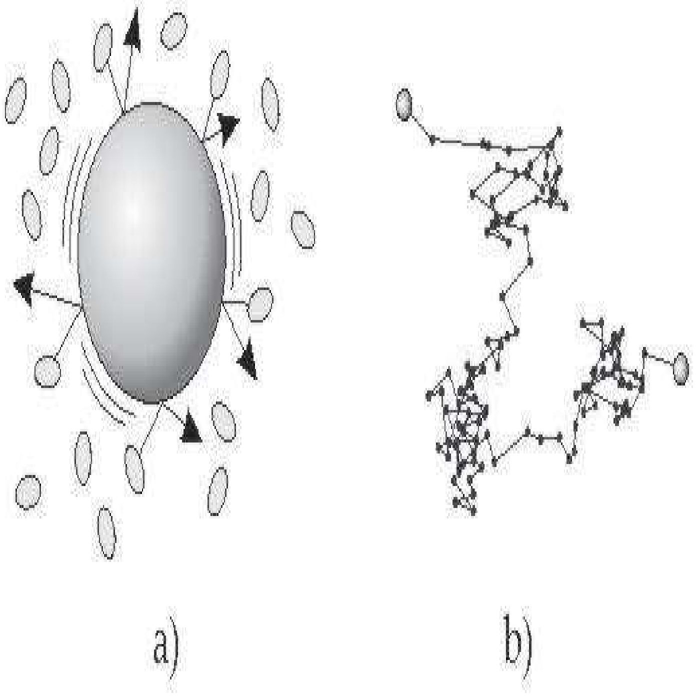

Figure 10.2: A Brownian Movement Particle

10.2 The Wiener Process: Definition and First Properties

We need an auxiliary notation:

$$p(t,y,x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}}, \quad t > 0, -\infty < x < \infty, -\infty < y < \infty. \tag{10.4}$$

Clearly $p(t,x,y)$ is the p.d.f. of a random variable with the distribution $N(x,t)$. This $p(t,x,y)$ is in fact the **transition p.d.f.** of a Wiener process , as will be explained below.

Remark 10.2.1 If we with $\sigma > 0$ set

$$p(t,y,x;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-\frac{(y-x)^2}{2\sigma^2 t}}, \quad t > 0, -\infty < x < \infty, -\infty < y < \infty, \tag{10.5}$$

we shall get a process that is also called the Wiener process. In fact, scaling of time, i.e., the definition in (10.4), which has $\sigma = 1$, is known as the **standard Wiener process**, but we shall not add the qualifier to our statements.

■

Definition 10.2.1 The **Wiener process** a.k.a. **Brownian motion** is a stochastic process $\mathbf{W} = \{W(t) \mid t \geq 0\}$ such that

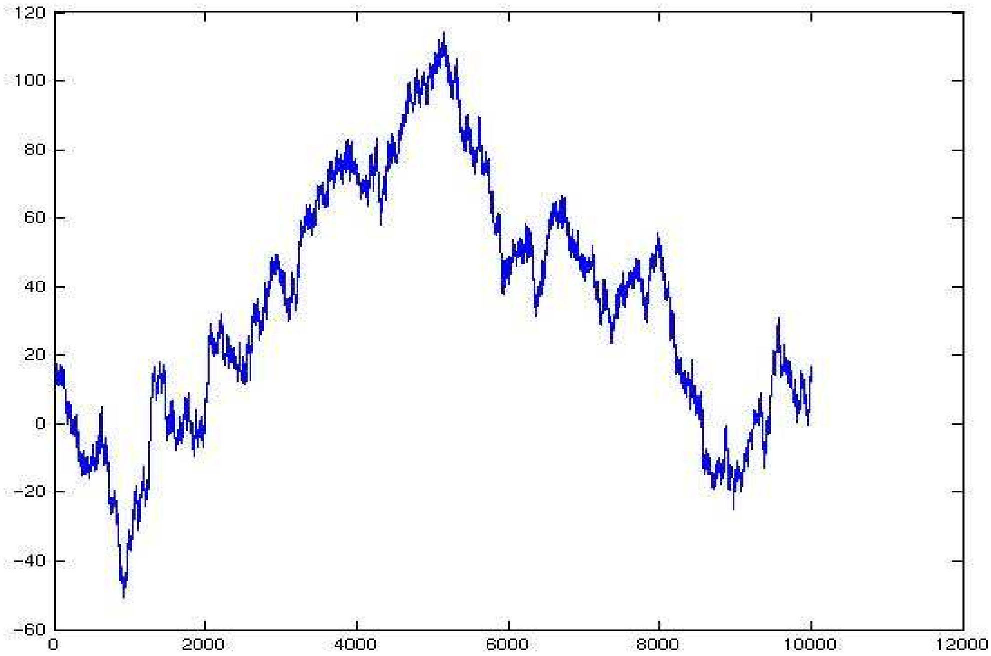

Figure 10.3: A Sample Path of A Wiener Process

i) $W(0) = 0$ almost surely.

ii) for any n and any finite suite of times $0 < t_1 < t_2 < \ldots < t_n$ and any $x_1, x_2, \ldots, x_n$ the joint p.d.f. of $W(t_1), W(t_2), \ldots, W(t_n)$ is

$$f_{W(t_1),W(t_2),\ldots,W(t_n)}(x_1, x_2, \ldots, x_n)$$
$$= p(t_1, x_1, 0)p(t_2 - t_1, x_2, x_1) \cdots p(t_n - t_{n-1}, x_n, x_{n-1}). \tag{10.6}$$

■

Let us next record a few of the immediate consequences of this definition.

1. We should perhaps first verify that (10.6) is in fact a joint p.d.f.. It is clear that $f_{W(t_1),W(t_2),\ldots,W(t_n)} \geq 0$. Next from (10.6)

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(t_1, x_1, 0) \prod_{i=1}^{n-1} p(t_{i+1} - t_i, x_{i+1}, x_i) dx_1 \cdots dx_n$$
$$= \int_{-\infty}^{\infty} p(t_1, x_1, 0) \cdots \int_{-\infty}^{\infty} p(t_n - t_{n-1}, x_n, x_{n-1}) dx_n \cdots dx_1.$$

We integrate first with respect to x_n and get

$$\int_{-\infty}^{\infty} p(t_n - t_{n-1}, x_n, x_{n-1}) dx_n = 1,$$

since we have seen that $p(t_n - t_{n-1}, x_n, x_{n-1})$ is the p.d.f. of $N(x_{n-1}, t_n - t_{n-1})$. An important thing is that the integral is not a function of x_{n-1}. Hence we can next integrate the factor containing x_{n-1} w.r.t

x_{n-1}, whereby the second of the two factors containing x_{n-2} will disappear, and continue successively in this way, and get that

$$\int_{-\infty}^{\infty} p(t_1, x_1, 0) \cdots \int_{-\infty}^{\infty} p(t_n - t_{n-1}, x_n, x_{n-1}) dx_n \cdots dx_1 = 1.$$

The preceding computation indicates also how to prove that the Wiener process exists by the Kolmogorov Consistency Theorem 9.1.1.

2. Take $n = 1$ and $t_1 = t$, then (10.6) and (10.4) give

$$W(t) \in N(0, t), \quad t > 0. \tag{10.7}$$

3. $n = 2$, $t_1 = s < t_2 = t$. Then the joint p.d.f. of $(W(s), W(t))$ is by (10.6)

$$f_{W(s),W(t)}(x, y) = p(s, x, 0) \cdot p(t - s, y, x)$$

$$= \frac{1}{\sqrt{2\pi s}} e^{-\frac{x^2}{2s}} \frac{1}{\sqrt{2\pi(t-s)}} e^{-\frac{(y-x)^2}{2(t-s)}}. \tag{10.8}$$

In view of (10.7) $p(s, x, 0)$ is the marginal p.d.f. $f_{W(s)}(x)$ of $W(s)$. Hence it holds for all x, y that

$$\frac{f_{W(s),W(t)}(x, y)}{f_{W(s)}(x)} = \frac{1}{\sqrt{2\pi(t-s)}} e^{-\frac{(y-x)^2}{2(t-s)}},$$

which tells us that

$$f_{W(t)|W(s)=x}(y) = \frac{1}{\sqrt{2\pi(t-s)}} e^{-\frac{(y-x)^2}{2(t-s)}}, \quad t > s, \tag{10.9}$$

or, equivalently,

$$W(t) \mid W(s) = x \in N(x, t - s), \quad t > s. \tag{10.10}$$

Inherent in the preceding is evidently that for $t > s$

$$W(t) = Z + W(s),$$

where $Z \in N(0, t - s)$ and Z is independent of $W(s)$. We shall, however, in the sequel obtain this finding as a by-product of a general result.

4. Hence we have the interpretation of $p(t - s, x, y)$ as a **transition p.d.f.**, since for any Borel set A and $t > s$

$$\mathbf{P}(W(t) \in A \mid W(s) = x) = \int_A p(t - s, y, x) dy = \int_A \frac{1}{\sqrt{2\pi(t-s)}} e^{-\frac{(y-x)^2}{2(t-s)}} dy.$$

This gives the probability of transition of the Wiener process from x at time s to the set A at time t.

The preceding should bring into mind the properties of a Gaussian process.

Theorem 10.2.1 The Wiener process $\mathbf{W}$ is a Gaussian process.

Proof: We make a change of variables in (10.6). We recall (2.71): if $\mathbf{X}$ has the p.d.f. $f_{\mathbf{X}}(\mathbf{x})$, $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$, and $\mathbf{A}$ is invertible, then $\mathbf{Y}$ has the p.d.f.

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|\det \mathbf{A}|} f_{\mathbf{X}}\left(\mathbf{A}^{-1}(\mathbf{y}-\mathbf{b})\right). \tag{10.11}$$

We are going to define a one-to-one linear transformation (with Jacobian = 1) between the n variables of the Wiener process and its increments. We take any n and any finite suite of times $0 < t_1 < t_2 < \ldots < t_n$. We recall first

$$Z_0 = W(t_0) = W(0) = 0$$

so that

$$Z_1 = W(t_1) - W(t_0) = W(t_1)$$

and then

$$Z_i \overset{\text{def}}{=} W(t_i) - W(t_{i-1}), \quad i = 2, \ldots n.$$

The increments $\{Z_i\}_{i=1}^n$ are a linear transformation of $(W(t_i))_{i=1}^n$, or in matrix form

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ \vdots \\ Z_{n-1} \\ Z_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 1 & \vdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & -1 & 1 & 0 \\ 0 & 0 & \ldots & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} W(t_1) \\ W(t_2) \\ \vdots \\ \vdots \\ W(t_{n-1}) \\ W(t_n) \end{pmatrix}. \tag{10.12}$$

We write this as

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_{n-1} \\ Z_n \end{pmatrix} = \mathbf{A} \begin{pmatrix} W(t_1) \\ W(t_2) \\ \vdots \\ W(t_{n-1}) \\ W(t_n) \end{pmatrix}.$$

The matrix $\mathbf{A}$ is lower triangular, and therefore its determinant is the product of the entries on the main diagonal, see [92, p. 93]. Thus in the above $\det \mathbf{A} = 1$, and the inverse $\mathbf{A}^{-1}$ exists and $\det \mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} = 1$. Hence the Jacobian determinant J is $= 1$.

It looks now, in view of (10.11), as if we are compelled to invert $\mathbf{A}^{-1}$ and insert in $f_{W(t_1),\ldots,W(t_n)}$. However, due to the special structure of $f_{W(t_1),\ldots,W(t_n)}$ in (10.6), we have a kind of stepwise procedure for this. By (10.6)

$$f_{W(t_1),W(t_2),\ldots,W(t_n)}(x_1, x_2, \ldots, x_n) = p(t_1, x_1, 0) \prod_{i=2}^{n} p(t_i - t_{i-1}, x_i, x_{i-1}).$$

Here, by (10.9),

$$p(t_i - t_{i-1}, x_i, x_{i-1}) = f_{W(t_i)|W(t_{i-1})=x_{i-1}}(x_i) = \frac{1}{\sqrt{2\pi(t_i - t_{i-1})}} e^{-\frac{(x_i - x_{i-1})^2}{2(t_i - t_{i-1})}}.$$

Hence, if we know that $W(t_{i-1}) = x_{i-1}$, then $Z_i = W(t_i) - x_{i-1}$ or $W(t_i) = Z_i + x_{i-1}$ and since we are evaluating the p.d.f. at the point $Z_i = z_i$ and $W(t_i) = x_i$, we get

$$p(t_i - t_{i-1}, x_i, x_{i-1}) = \frac{1}{\sqrt{2\pi(t_i - t_{i-1})}} e^{-\frac{z_i^2}{2(t_i - t_{i-1})}} = f_{Z_i}(z_i).$$

Thus

$$Z_i \in N(0, t_i - t_{i-1})$$

and

$$f_{Z_1,Z_2,\ldots,Z_n}(z_1, z_2, \ldots, z_n) = \prod_{i=1}^{n} f_{Z_i}(z_i).$$

This shows that the increments are independent, and that

$$f_{Z_1,Z_2,\ldots,Z_n}(z_1, z_2, \ldots, z_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \mathbf{\Lambda}}} e^{-\mathbf{z}^{'}\mathbf{\Lambda}^{-1}\mathbf{z}/2}$$

where $\mathbf{\Lambda}$ is the diagonal matrix

$$\mathbf{\Lambda} = \begin{pmatrix} t_1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & t_2 - t_1 & \ldots & 0 & 0 & 0 \\ 0 & 0 & t_3 - t_2 & \vdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & t_{n-1} - t_{n-2} & 0 \\ 0 & 0 & \ldots & 0 & 0 & t_n - t_{n-1} \end{pmatrix}. \tag{10.13}$$

This matrix $\mathbf{\Lambda}$ is clearly symmetric. In addition, for any $\mathbf{x} \in \mathbf{R}^n$ we have

$$\mathbf{x}^{'}\mathbf{\Lambda}\mathbf{x} = \sum_{i=1}^{n} x_i^2 \cdot (t_i - t_{i-1}) > 0, \quad (t_0 = 0).$$

Hence the matrix $\mathbf{\Lambda}$ is a covariance matrix. In other words, $Z_1, Z_2, \ldots, Z_n$ has a joint Gaussian distribution $N\left(\underline{0}, \mathbf{\Lambda}\right)$ and since

$$\begin{pmatrix} W(t_1) \\ W(t_2) \\ \vdots \\ W(t_{n-1}) \\ W(t_n) \end{pmatrix} = \mathbf{A}^{-1} \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_{n-1} \\ Z_n \end{pmatrix},$$

then $(W(t_1), W(t_2), \ldots, W(t_{n-1}), W(t_n))$ has a joint Gaussian distribution

$$N\left(\underline{0}, \mathbf{A}^{-1}\mathbf{\Lambda}(\mathbf{A}^{-1})^{'}\right). \tag{10.14}$$

Since n and $t_1, \ldots, t_n$ were arbitrary, we have now shown that the Wiener process is a Gaussian process. ■

Remark 10.2.2 The proof above is perhaps overly arduous, as the main idea is simple. The increments $\{Z_i\}_{i=1}^n$ and $W(t_1), \ldots, W(t_n)$, correspond to each other

$$\{Z_i\}_{i=1}^n \leftrightarrow \{W(t_i)\}_{i=1}^n$$

by an invertible linear transformation, since the inverse is given by

$$W(t_i) = \sum_{k=1}^{i} Z_k, \tag{10.15}$$

which yields uniquely the Wiener process variables from the increments, remembering that $W_0 = W(0) = 0$. Thus, when we know that $W(t_{i-1}) = x_{i-1}$, then $Z_i = W(t_i) - x_{i-1}$ and in the conditional p.d.f. $p(t_i - t_{i-1}, x_i, x_{i-1})$ the change of variable is simple.

■

A Gaussian process is uniquely determined by its mean function and its autocovariance function. We can readily find the mean function $\mu_{\mathbf{W}}(t)$ and the autocorrelation function $R_{\mathbf{W}}(t,s)$. This will give us the matrices $\mathbf{A}^{-1}\mathbf{\Lambda}(\mathbf{A}^{-1})^{'}$ in (10.14), too, but without any matrix operations. The mean function is from (10.7) and i) in the definition

$$\mu_{\mathbf{W}}(t) = E\left[W(t)\right] = 0 \quad t \geq 0. \tag{10.16}$$

Lemma 10.2.2 The $R_{\mathbf{W}}(t,s)$ of the Wiener process is

$$R_{\mathbf{W}}(t,s) = \min(t,s) \tag{10.17}$$

for any $t, s \geq 0$.

Proof Let us assume that $t > s$. Then by double expectation

$$E\left[W(t)W(s)\right] = E\left[E\left[W(t)W(s) \mid W(s)\right]\right] =$$

and by taking out what is known

$$= E\left[W(s)E\left[W(t) \mid W(s)\right]\right].$$

We invoke here (10.10), i.e., $E\left[W(t) \mid W(s)\right] = W(s)$, and obtain

$$= E\left[W^2(s)\right] = s,$$

where we used (10.7).
Exactly in the same manner we can show that if $s > t$

$$E\left[W(t)W(s)\right] = t.$$

Thus we have established (10.17), as claimed. ■

Remark 10.2.3 The definition (10.5) gives instead

$$R_{\mathbf{W}}(t,s) = \sigma^2 \min(t,s). \tag{10.18}$$

■

The equation (10.17) implies that the covariance matrix $\mathbf{C_W}$ of $(W(t_1),\ldots,W(t_n))^{'}$, $0 < t_1 < t_2 < \ldots < t_n$, is

$$\mathbf{C_W} = \begin{pmatrix} t_1 & t_1 & \ldots & t_1 & t_1 & t_1 \\ t_1 & t_2 & \ldots & t_2 & t_2 & t_2 \\ t_1 & t_2 & \ldots & t_3 & t_3 & t_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ t_1 & t_2 & \ldots & t_{n-2} & t_{n-1} & t_{n-1} \\ t_1 & t_2 & \ldots & t_{n-2} & t_{n-1} & t_n \end{pmatrix}. \tag{10.19}$$

One could check that $\mathbf{C_W} = \mathbf{A}^{-1}\mathbf{\Lambda}(\mathbf{A}^{-1})^{'}$, as it should by (10.14). We have encountered the matrix $\mathbf{C_W}$ in an exercise on autocovariance in section 9.7.1 of chapter 9 and shown without recourse to the Wiener process that $\mathbf{C_W}$ is indeed a covariance matrix.

Lemma 10.2.3

$$E\left[(W(t)-W(s))^2\right] = |t-s| \tag{10.20}$$

for any $t, s \geq 0$.

Proof

$$\begin{aligned} E\left[(W(t)-W(s))^2\right] &= E\left[W^2(t) - 2W(t)W(s) + W^2(s)\right] \\ &= E\left[W^2(t)\right] - 2E\left[W(t)W(s)\right] + E\left[W^2(s)\right] \\ &= t - 2\min(t,s) + s \end{aligned}$$

by (10.17) and (10.7). Then we have

$$= \begin{cases} t-2s+s = t-s & \text{if } t > s \\ t-2t+s = s-t & \text{if } s > t. \end{cases}$$

By definition of absolute value,

$$\mid t-s \mid = \begin{cases} t-s & t > s \\ -(t-s) = s-t & s > t. \end{cases} \tag{10.21}$$

Thus

$$E\left[(W(t)-W(s))^2\right] = |t-s|.$$

■

Lemma 10.2.4 For a Wiener process and for $t \geq s$

$$W(t) - W(s) \in N(0, t-s). \tag{10.22}$$

Proof Because the Wiener process is a Gaussian process, $W(t) - W(s)$ is a Gaussian random variable. The rest of the proof follows by (10.16) and (10.20). ■

The result in the following lemma is already found in the proof of theorem 10.2.1, but we state and prove it anew for ease of reference and benefit of learning.

Lemma 10.2.5 For a Wiener process and for $0 \leq u \leq v \leq s \leq t$

$$W(t) - W(s) \text{ is independent of } W(v) - W(u) \tag{10.23}$$

Proof We can write

$$\begin{pmatrix} W(v) - W(u) \\ W(t) - W(s) \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} W(v) \\ W(u) \\ W(t) \\ W(s) \end{pmatrix}.$$

Therefore $W(t) - W(s)$ and $W(v) - W(u)$ are, by theorem 8.1.6, jointly Gaussian random variables with zero means. It is enough to show that they are uncorrelated.

$$E\left[(W(t) - W(s))(W(v) - W(u))\right] =$$

$$\begin{aligned} &= E\left[W(t)W(v)\right] - E\left[W(t)W(u)\right] - E\left[W(s)W(v)\right] + E\left[W(s)W(u)\right] \\ &= \min(t,v) - \min(t,u) - \min(s,v) + \min(s,u) \\ &= v - u - v + u = 0. \end{aligned}$$

■

In fact we have by this last lemma shown that

Theorem 10.2.6 For a Wiener process and any finite suite of times $0 < t_1 < t_2 < \ldots < t_n$ the **increments**

$$W(t_1) - W(t_0), W(t_2) - W(t_1), \ldots, W(t_n) - W(t_{n-1})$$

are independent and Gaussian.

■

It follows also by the above that the **increments of the Wiener process are strictly stationary**, since for all n and h

$$W(t_1) - W(t_0), W(t_2) - W(t_1), \ldots, W(t_n) - W(t_{n-1})$$
$$\overset{d}{=} W(t_1 + h) - W(t_0 + h), W(t_2 + h) - W(t_1 + h), \ldots, W(t_n + h) - W(t_{n-1} + h),$$

by (10.20).

10.3 A Construction of the Wiener Process

Let us recall example 9.2.3, which was based on example 9.1.8. There we obtained the integral equation (9.23) or

$$\int_0^T R(t,s)e_i(s)ds = \lambda_i e_i(t). \tag{10.24}$$

Let us solve this with $R(t,s) = \min(t,s)$ in $[0,T]$, we follow [103, p. 87]. We insert to get

$$\int_0^T \min(t,s)e_i(s)ds = \lambda_i e_i(t), \tag{10.25}$$

or

$$\int_0^t s e_i(s)ds + t\int_t^T e_i(s)ds = \lambda_i e_i(t). \tag{10.26}$$

This is a case, where we can solve an integral equation by reducing it to an ordinary differential equation. We differentiate thus once w.r.t. t in (10.26) and get

$$\int_t^T e_i(s)ds = \lambda_i e_i^{'}(t). \tag{10.27}$$

We differentiate once more, which yields

$$-e_i(t) = \lambda_i e_i^{''}(t). \tag{10.28}$$

We have also the obvious boundary conditions $e_i(0) = 0$ from (10.26) and $e_i^{'}(T) = 0$ from (10.27). Equation (10.28) with $e_i(0) = 0$ gives

$$e_i(t) = A\sin\left(\frac{1}{\sqrt{\lambda_i}}t\right).$$

When we check $e_i^{'}(T) = 0$ we obtain

$$\cos\left(\frac{1}{\sqrt{\lambda_i}}T\right) = 0.$$

In other words

$$\lambda_i = \frac{T^2}{\pi^2\left(i+\frac{1}{2}\right)^2}, \quad i = 0, 1, \ldots$$

Then the normalized eigenfunctions are

$$e_i(t) = \sqrt{\frac{2}{T}}\sin\left(\left(i+\frac{1}{2}\right)\pi\frac{t}{T}\right),$$

and we have by the computations in example 9.2.3

$$\min(t,s) = \frac{2}{T}\sum_{i=0}^{\infty}\frac{T^2}{\pi^2\left(i+\frac{1}{2}\right)^2}\sin\left(\left(i+\frac{1}{2}\right)\pi\frac{t}{T}\right)\sin\left(\left(i+\frac{1}{2}\right)\pi\frac{s}{T}\right),$$

which is an interesting expression for $\min(t,s)$ in $[0,T]\times[0,T]$ in its own right. In addition, by example 9.2.3 we can **construct the Wiener process** as

$$W(t) = \sum_{i=0}^{\infty}\frac{T}{\pi\left(i+\frac{1}{2}\right)}\cdot X_i\cdot\sqrt{\frac{2}{T}}\sin\left(\left(i+\frac{1}{2}\right)\pi\frac{t}{T}\right) \quad t\in[0,T],$$

where X_i are I.I.D. and $N(0,1)$. We are omitting further details that would enable us to prove almost sure convergence of the series [7, pp. 7−9].

10.4 The Sample Paths of the Wiener Process

The Wiener process shares with the Poisson process the status of being the most important process in probability theory. Its sample paths display an astonishing range of behaviour, see, e.g., [19]. Here we are concentrating on the mean square properties, which are straightforward by comparison. We shall, however, next indicate some of the basic sample path properties of the Wiener process, as one needs to be sufficiently well-informed about these in order to decide intelligently in what ways the Wiener process can, and in what ways it cannot, be expected to reflect realistically the properties of some physical processes.

In view of (10.20) we see that

$$E\left[(W(t+h)-W(t))^2\right] = |h| \tag{10.29}$$

and hence the Wiener process is continuous in quadratic mean in the sense of the definition 9.3.2. As is known, convergence in quadratic mean does not imply convergence almost surely. Hence the result in the following section requires a full proof, which is of a higher degree of sophistication than (10.29). As we shall see below, the actual proof does exploit (10.29), too.

10.4.1 The Sample Paths of the Wiener Process are Almost Surely Continuous

We need an additional elementary fact.

$$Z\in N(0,\sigma^2)\Rightarrow E\left[Z^4\right] = 3\sigma^4. \tag{10.30}$$

This can be found by the fourth derivative of either the moment generating or the characteristic function and evaluation of this fourth derivative at zero[9].

We shall now start the proof of the statement in the title of this section following [103, p.57−58] and [7, chap.1]. The next proof can be omitted at first reading.

The Markov inequality (1.38) gives for every $\varepsilon>0$ and $h>0$

$$\mathbf{P}\left(\mid W(t+h)-W(t)\mid\geq\varepsilon\right)\leq\frac{E\left[\mid W(t+h)-W(t)\mid^4\right]}{\varepsilon^4}.$$

[9] By (4.50) the general rule is given as follows. If $Z\in N(0,\sigma^2)$, then

$$E\left[Z^n\right] = \begin{cases} 0 & n \text{ is odd} \\ \frac{(2k)!}{2^k k!}\sigma^{2k} & n=2k,\ k=0,1,2,\ldots. \end{cases}$$

The reason for selecting above the exponent $= 4$ becomes clear eventually. By (10.29) and (10.30) we get

$$E\left[(W(t+h)-W(t))^4\right] = 3h^2.$$

Therefore

$$\mathbf{P}\left(\mid W(t+h)-W(t) \mid \geq h^{\gamma}\right) \leq 3h^{2-4\gamma}.$$

Let now $0 \leq \gamma < 1/4$ and set $\delta = 1-4\gamma > 0$. We get

$$\mathbf{P}\left(\mid W(t+h)-W(t) \mid \geq h^{\gamma}\right) \leq 3h^{1+\delta}. \tag{10.31}$$

These are preparations for an application of the **Borel-Cantelli lemma**. With that strategy in mind we consider for nonnegative integers v the random variables

$$Z_v \overset{\text{def}}{=} \sup_{0\leq k\leq 2^v-1} \mid W((k+1)/2^{\gamma}) - W(k/2^{\gamma}) \mid .$$

Then

$$\mathbf{P}\left(Z_v \geq \left(\frac{1}{2^v}\right)^{\gamma}\right) \leq \mathbf{P}\left(\cup_{k=0}^{2^v-1} \mid W((k+1)/2^{\gamma}) - W(k/2^{\gamma}) \mid \geq \left(\frac{1}{2^v}\right)^{\gamma}\right),$$

since if $Z_v \geq \left(\frac{1}{2^v}\right)^{\gamma}$, then there is at least one increment such that $\mid W((k+1)/2^{\gamma}) - W(k/2^{\gamma}) \mid \geq \left(\frac{1}{2^v}\right)^{\gamma}$. Then by **subadditivity**, or $A \subset \cup_i A_i$ then $\mathbf{P}(A) \leq \sum_i \mathbf{P}(A_i)$, see chapter 1,

$$\leq \sum_{k=0}^{2^v-1} \mathbf{P}\left(\mid W((k+1)/2^{\gamma}) - W(k/2^{\gamma}) \mid \geq \left(\frac{1}{2^v}\right)^{\gamma}\right)$$

$$\leq 3\cdot 2^v \left(\frac{1}{2^v}\right)^{1+\delta} = 3\cdot 2^{-\delta v},$$

where we used (10.31). Since $\sum_{v=0}^{\infty} 2^{-\delta v} < \infty$ we have

$$\sum_{v=0}^{\infty} \mathbf{P}\left(Z_v \geq \frac{1}{2^{v\gamma}}\right) < \infty.$$

By the **Borel-Cantelli lemma**, lemma 1.7.1 above, the event

$$Z_v \geq \frac{1}{2^{v\gamma}}$$

occurs with probability one only a finite number of times. In other words, it holds that there is almost surely an $N(\omega)$ such that for all $v \geq N(\omega)$,

$$Z_v \leq \frac{1}{2^{v\gamma}}$$

and therefore

$$\lim_{n\to\infty} \sum_{v=n+1}^{\infty} Z_v = 0, \quad \text{a.s..}$$

This entails that

$$\sup_{t,s\in T; |t-s|<2^{-n}} \mid W(t)-W(s) \mid \overset{a.s.}{\to} 0,$$

as $n \to \infty$, where T is any finite interval $\subset [0,\infty)$. This assertion is intuitively plausible, but requires a detailed analysis omitted here, see [103, p. 86] for details.

■

By the preceding we have in bits and pieces more or less established the following theorem, which is frequently evoked as the very definition of the Wiener process, see [13, chapter 2].

Theorem 10.4.1 A stochastic process $\{W(t) \mid t \geq 0\}$ is a Wiener process if and only if the following four conditions are true:

1) $W(0) = 0$.

2) The sample paths $t \mapsto W(t)$ are almost surely continuous.

3) $\{W(t) \mid t \geq 0\}$ has stationary and independent increments.

4) $W(t) - W(s) \in N(0, t-s)$ for $t > s$.

■

10.4.2 The Sample Paths of the Wiener Process are Almost Surely Nowhere Differentiable; Quadratic Variation of the Sample Paths

We are not going to prove the following theorem.

Theorem 10.4.2 The Wiener process $\{W(t) \mid t \geq 0\}$ is almost surely non-differentiable at any $t \geq 0$.

■

We shall next present two results, namely lemma 10.4.3 and theorem 10.4.4, that contribute to the understanding of the statement about differentiation of the Wiener process. Let for $i = 0, 1, 2, \ldots, n$

$$t_i^{(n)} = \frac{iT}{n}.$$

Clearly $0 = t_0^{(n)} < t_1^{(n)} < \ldots < t_n^{(n)} = T$ is a partition of $[0, T]$ into n equal parts. We denote by

$$\triangle_i^n W \overset{\text{def}}{=} W\left(t_{i+1}^{(n)}\right) - W\left(t_i^{(n)}\right) \tag{10.32}$$

the corresponding increment of the Wiener process. For future reference we say that the **random quadratic variation** of the Wiener process is the random variable

$$\sum_{i=0}^{n-1} \left(\triangle_i^n W\right)^2.$$

Lemma 10.4.3 The random quadratic variation converges in mean square to T, or

$$\sum_{i=0}^{n-1} \left(\triangle_i^n W\right)^2 \xrightarrow{2} T, \tag{10.33}$$

as $n \to \infty$.

Proof By the definition in chapter 7.1 we need to show that

$$E\left[\left(\sum_{i=0}^{n-1} \left(\triangle_i^n W\right)^2 - T\right)^2\right] \to 0$$

as $n \to \infty$. First we do a simple manipulation of sums

$$\sum_{i=0}^{n-1} (\triangle_i^n W)^2 - T = \sum_{i=0}^{n-1} \left((\triangle_i^n W)^2 - \frac{T}{n} \right).$$

Thus

$$\begin{aligned} E\left[\left(\sum_{i=0}^{n-1} (\triangle_i^n W)^2 - T \right)^2 \right] &= E\left[\left(\sum_{i=0}^{n-1} (\triangle_i^n W)^2 - \frac{T}{n} \right)^2 \right] \\ &= \sum_{i=0}^{n-1} E\left[\left((\triangle_i^n W)^2 - \frac{T}{n} \right)^2 \right] \\ +2\sum_{i<j} E &\left[\left((\triangle_i^n W)^2 - \frac{T}{n} \right) \left((\triangle_j^n W)^2 - \frac{T}{n} \right) \right]. \end{aligned} \tag{10.34}$$

By Theorem 10.2.6 the increments of the Wiener process are independent, when considered over non-overlapping intervals. Thus

$$E\left[\left((\triangle_i^n W)^2 - \frac{T}{n} \right) \left((\triangle_j^n W)^2 - \frac{T}{n} \right) \right] = E\left[\left((\triangle_i^n W)^2 - \frac{T}{n} \right) \right] E\left[\left((\triangle_j^n W)^2 - \frac{T}{n} \right) \right].$$

By (10.20) we get

$$E\left[(\triangle_i^n W)^2 \right] = E\left[(\triangle_j^n W)^2 \right] = \frac{T}{n},$$

and the cross products in (10.34)vanish.
Thus we have obtained

$$E\left[\left(\sum_{i=0}^{n-1} (\triangle_i^n W)^2 - T \right)^2 \right] = \sum_{i=0}^{n-1} E\left[\left((\triangle_i^n W)^2 - \frac{T}{n} \right)^2 \right].$$

We square the term in the sum in the right hand side

$$E\left[\left((\triangle_i^n W)^2 - \frac{T}{n} \right)^2 \right] = E\left[(\triangle_i^n W)^4 \right] - 2\frac{T}{n} \cdot E\left[(\triangle_i^n W)^2 \right] + \frac{T^2}{n^2}.$$

In view of (10.20)

$$E\left[(\triangle_i^n W)^2 \right] = \frac{T}{n},$$

and thus (10.30) entails

$$E\left[(\triangle_i^n W)^4 \right] = \frac{3T^2}{n^2}.$$

Thus

$$\begin{aligned} E\left[\left(\sum_{i=0}^{n-1} (\triangle_i^n W)^2 - T \right)^2 \right] &= \sum_{i=0}^{n-1} \left(\frac{3T^2}{n^2} - \frac{2T^2}{n^2} + \frac{T^2}{n^2} \right) \\ &= \sum_{i=0}^{n-1} \frac{2T^2}{n^2} = \frac{2T^2}{n}. \end{aligned}$$

Hence the assertion follows as claimed, when $n \to \infty$. ■

In the theory of stochastic calculus, see e.g., [70, p.62], one introduces the notation

$$[W, W]([0, T]) \stackrel{\text{def}}{=} \lim \sum_{i=0}^{n-1} (\triangle_i^n W)^2$$

or in [29, p.86],

$$< W >_T \overset{\text{def}}{=} \lim \sum_{i=0}^{n-1} (\triangle_i^n W)^2$$

and refers to $[W, W]([0, T])$ as **quadratic variation**, too, but for our purposes we need not load the presentation with these brackets.
We need to recall a definition from mathematical analysis [36, p.54].

Definition 10.4.1 The **total variation** of a function f from $[0, T]$ to $\mathbf{R}$, is defined by

$$\limsup_{\triangle \to 0} \sum_{i=0}^{n-1} \mid f(t_{i+1}) - f(t_i) \mid,$$

where $0 = t_0 < t_1 < \ldots < t_n = T$ is a partition of $[0, T]$ and

$$\triangle = \max_{i=0,\ldots,n} \mid t_{i+1} - t_i \mid .$$

■

The following theorem 10.4.4 implies that the length of a sample path of the Wiener process in any finite interval is infinite. Hence we understand that a simulated sample path like the one depicted in figure 10.3 cannot be but a computer approximation.

At this point we should pay attention to **Brownian Scaling**. If $\{W(t)|t \geq 0\}$ is the Wiener process, we define for $c > 0$ a new process by

$$V(t) \overset{\text{def}}{=} \frac{1}{c} W(c^2 t).$$

An exercise below shows that $\{V(t) \mid t \geq 0\}$ is a Wiener process. In words, if one magnifies the process $\{W(t)|t \geq 0\}$, i.e., chooses a small c, while at the same time looking at the process in a small neighborhood of origin, then one sees again a process, which is statistically identical with the original Wiener process. In another of the exercises we study **Time Reversal**

$$V(t) \overset{\text{def}}{=} tW\left(\frac{1}{t}\right),$$

in which we, for small values of t, we look at the Wiener process at infinity, and scale it back to small amplitudes, and again we are looking at the Wiener process.
These phenomenona are known as **self-similarity** and explain intuitively that the length of a sample path of the Wiener process in any finite interval must be infinite.

Theorem 10.4.4 The total variation of the sample paths of the Wiener process on any interval $[0, T]$ is infinite.

Proof As in lemma 10.4.3 we consider the sequence of partitions $\left(t_0^{(n)}, t_1^{(n)}, \ldots, t_n^{(n)}\right)$ of $[0, T]$ into n equal parts. Then with the notation of (10.32) we get

$$\sum_{i=0}^{n-1} \mid \triangle_i^n W \mid^2 \leq \max_{i=0,1,\ldots,n} \mid \triangle_i^n W \mid \sum_{i=0}^{n-1} \mid \triangle_i^n W \mid . \tag{10.35}$$

Since the sample paths of the Wiener process are almost surely continuous on $[0, T]$, we must have

$$\lim_{n\to\infty} \max_{i=0,1,\ldots,n} \mid \triangle_i^n W \mid = 0, \tag{10.36}$$

almost surely, as the partitions of $[0, T]$ become successively refined. as n increases.

On the other hand, by lemma 10.4.3

$$\sum_{i=0}^{n-1} (\triangle_i^n W)^2 \overset{2}{\to} T > 0,$$

as $n \to \infty$, which implies (this is a general fact about the relationship between almost sure and mean square convergence) that **there is a subsequence** n_k such that

$$\sum_{i=0}^{n_k-1} (\triangle_i^n W)^2 \overset{a.s.}{\to} T, \tag{10.37}$$

as $k \to \infty$.
Next, from (10.35)

$$\frac{\sum_{i=0}^{n-1} \mid \triangle_i^n W \mid^2}{\max_{i=0,1,\ldots,n} \mid \triangle_i^n W} \mid \leq \sum_{i=0}^{n-1} \mid \triangle_i^n W \mid .$$

But then (10.36) and (10.37) entail

$$\lim_{k\to\infty} \mid \sum_{i=0}^{n-1} \mid \triangle_i^n W \mid \to \infty,$$

as the subsequences of partitions of $[0, T]$ become more and more refined as k increases. ■

A Motivational Argument Concerning Quadratic Variation

We make a summary and an interpretation of the preceding. Take the partition of thetime axis used in lemma 10.4.3 and set

$$S_n = \sum_{i=0}^{n-1} (\triangle_i^n W)^2 .$$

The important fact that emerged above is that the variance of S_n is negligible compared to its expectation, or

$$E[S_n] = T,$$

while the proof of lemma 10.4.3 shows that

$$\mathrm{Var}[S_n] = \frac{2T^2}{n}.$$

Thus, the expectation of S_n is constant, whereas the variance of S_n converges to zero, as n grows to infinity. Hence S_n must converge to a **non-random quantity**. We write this as

$$\int_0^t [dW]^2 = t$$

or

$$[dW]^2 = dt. \tag{10.38}$$

The formula (10.38) is a starting point for the intuitive handling of the differentials behind **It$\widehat{o}$'s formula** in stochastic calculus, see [13, pp. 50−55], [62, pp. 32−36], [68, chapter 5] and [29, 70].

10.4.3 The Wiener Process is a Markov Process

Next we show that the Wiener process has the **Markov property**.

Theorem 10.4.5 For any $t_1 < \ldots < t_{n-1} < t_n$ and any $x_1, \ldots, x_{n-1}, x_n$

$$\mathbf{P}\left(W(t_n) \leq x_n \mid W(t_1) = x_1, \ldots, W(t_{n-1}) = x_{n-1}\right)$$

$$= \mathbf{P}\left(W(t_n) \leq x_n \mid W(t_{n-1}) = x_{n-1}\right). \tag{10.39}$$

Proof

$$\begin{aligned} &\mathbf{P}\left(W(t_n) \leq x_n \mid W(t_1) = x_1, \ldots, W(t_{n-1}) = x_{n-1}\right) \\ &= \int_{-\infty}^{x_n} f_{W(t_n)|W(t_1)=x_1,\ldots,W(t_{n-1})=x_{n-1}}(v)\, dv \\ &= \int_{-\infty}^{x_n} \frac{f_{W(t_1),\ldots,W(t_{n-1}),W(t_n)}(x_1, \ldots, x_{n-1}, v)}{f_{W(t_1),\ldots,W(t_{n-1})}(x_1, \ldots, x_{n-1})} dv \end{aligned}$$

and we use (10.6) to get

$$\int_{-\infty}^{x_n} \frac{p(t_1, x_1, 0)p(t_2 - t_1, x_2, x_1) \cdots p(t_n - t_{n-1}, v, x_{n-1})}{p(t_1, x_1, 0)p(t_2 - t_1, x_2, x_1) \cdots p(t_{n-1} - t_{n-2}, x_{n-1}, x_{n-2})} dv$$

$$= \int_{-\infty}^{x_n} p(t_n - t_{n-1}, v, x_{n-1}) dv = \mathbf{P}\left(W(t_n) \leq x_n \mid W(t_{n-1}) = x_{n-1}\right).$$

■

The Wiener process is a Gaussian Markov process and its autocorrelation function is $R_{\mathbf{W}}(t, s) = \min(t, s)$. Then, if $t_0 < s < t$ we check (9.37) by

$$\frac{R_{\mathbf{W}}(t, s) R_{\mathbf{W}}(s, t_0)}{R_{\mathbf{W}}(s, s)} = \frac{\min(t, s) \min(s, t_0)}{\min(s, s)} = \frac{s \cdot t_0}{s} = t_0,$$

which equals $R_{\mathbf{W}}(t, t_0) = \min(t, t_0) = t_0$, as it should.

10.5 The Wiener Integral

This section draws mainly upon [26, chapter 3.4].

10.5.1 Definition

Definition 10.5.1 Let $\{W(t) | t \geq 0\}$ be a Wiener process and $f(t)$ be a function such that $\int_a^b f^2(t) dt < \infty$, where $0 \leq a < b \leq +\infty$. The mean square integral with respect to the Wiener process or the **Wiener integral** $\int_a^b f(t) dW(t)$ is defined as the mean square limit

$$\sum_{i=1}^{n} f(t_{i-1}) \left(W(t_i) - W(t_{i-1})\right) \overset{2}{\to} \int_a^b f(t) dW(t), \tag{10.40}$$

where $a = t_0 < t_1 < \ldots < t_{n-1} < t_n = b$ and $\max_i |t_i - t_{i-1}| \to 0$ as $n \to \infty$.

■

In general, we know that the sample paths of the Wiener process have unbounded total variation, but have by lemma 10.4.3 finite quadratic variation. Hence we must define $\int_a^b f(t) dW(t)$ using mean square convergence, which means that we **are looking at all sample paths simultaneously**.

The reader should note the similarities and differences between the left hand side of (10.40) and the the discrete stochastic integral in (3.56) above.

In physics the Wiener integral is a name for a different mathematical concept, namely that of a **path integral**. By this we refer to an integral of a functional of the Wiener process with respect to the Wiener measure, which is a probability measure on the set of continuous functions over $[0, T]$, see [78, chapter 6].

Remark 10.5.1 As pointed out in [105, p. 88], Wiener himself introduced the integral later named after him by the formula of integration by parts

$$\int_a^b f(t)dW(t) = [f(t)W(t)]_a^b - \int_a^b W(t)df(t), \tag{10.41}$$

where the function $f(t)$ is assumed have *bounded total variation* in the sense of definition 10.4.1. As the sample functions of a Wiener process are continuous, the right hand side is well-defined, inasmuch the integral is a Stieltjes integral [69, chapter 6.8].

■

Example 10.5.1 We consider

$$X_n = \sum_{i=1}^{n} e^{-\lambda \frac{i-1}{n}} \left(W\left(\frac{i}{n}\right) - W\left(\frac{(i-1)}{n}\right) \right), \quad n \geq 1.$$

That is, we have $t_i = \frac{i}{n}$ in definition, see eq. (10.40), and $0 = t_0 < t_1 < \ldots < t_{n-1} < t_n = 1$. We expect this to converge to

$$X_n \overset{2}{\to} \int_0^1 e^{-\lambda u} dW(u),$$

as $n \to \infty$. This implies convergence in distribution. We shall find the limiting distribution.
We set for convenience of writing for all $1 \leq i \leq n$

$$Y_i \overset{\text{def}}{=} W\left(\frac{i}{n}\right) - W\left(\frac{(i-1)}{n}\right)$$

and then

$$Y_i \in N\left(0, \frac{1}{n}\right)$$

for all $1 \leq i \leq n$. Thus, as a linear combination of normal random variables,

$$X_n = \sum_{i=1}^{n} e^{-\lambda \frac{i-1}{n}} Y_i$$

is a normal random variable. Its expectation and variance are as follows.

$$E[X_n] = \sum_{i=1}^{n} e^{-\lambda \frac{i-1}{n}} E[Y_i] = 0,$$

and since the increments (i.e., here Y_i) of a Wiener process are independent for non overlapping intervals

$$\text{Var}(X_n) = \sum_{i=1}^{n} e^{-2\lambda \frac{i-1}{n}} \text{Var}(Y_i) = \frac{1}{n} \sum_{i=1}^{n} e^{-2\lambda \frac{i-1}{n}}.$$

Therefore the characteristic function of X_n is

$$\varphi_{X_n}(t) = e^{-\frac{t^2}{2}\frac{1}{n}\sum_{i=1}^{n} e^{-2\lambda \frac{i-1}{n}}}.$$

We can check the convergence in distribution by means of this without invoking Riemann sums. In fact we have

$$\frac{1}{n}\sum_{i=1}^{n} e^{-2\lambda\frac{i-1}{n}} = \frac{1}{n}\sum_{i=0}^{n-1} e^{-2\lambda\frac{i}{n}} = \frac{1}{n}\frac{1-e^{-2\lambda}}{1-e^{-2\lambda\frac{1}{n}}}.$$

We write this as

$$\frac{1}{n}\frac{1-e^{-2\lambda}}{1-e^{-2\lambda\frac{1}{n}}} = \frac{1-e^{-2\lambda}}{\frac{1-e^{-2\lambda\frac{1}{n}}}{\frac{1}{n}}}.$$

Then we set $f(t) = e^{-2\lambda t}$, and recognize the difference ratio

$$\frac{1-e^{-2\lambda\frac{1}{n}}}{\frac{1}{n}} = -\frac{\left(f\left(\frac{1}{n}\right) - f(0)\right)}{\frac{1}{n}} \to -f^{'}(0) = 2\lambda,$$

as $n \to \infty$. Hence

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} e^{-2\lambda\frac{k-1}{n}} = \frac{1-e^{-2\lambda}}{2\lambda}.$$

We note that $\int_0^1 e^{-2\lambda u}du = \frac{1-e^{-2\lambda}}{2\lambda}$. Thus we have shown that

$$X_n \xrightarrow{d} \int_0^1 e^{-\lambda u}dW(u) \in N\left(0, \int_0^1 e^{-2\lambda u}du\right),$$

as $n \to \infty$.

■

10.5.2 Properties

Since (10.40) defines the Wiener integral in terms of convergence in mean square, we can easily adapt the techniques in section 9.2 to this case and derive some of the basic properties of the Wiener integral defined in (10.40).

1. The following property is readily verified:

$$\int_a^b (f(t) + g(t))\, dW(t) = \int_a^b f(t)dW(t) + \int_a^b g(t)dW(t),$$

if $\int_a^b f^2(t)dt < \infty$ and $\int_a^b g^2(t)dt < \infty$.

2.

$$E\left[\int_a^b f(t)dW(t)\right] = 0. \tag{10.42}$$

This follows in the same way as the proof of the analogous statement in theorem 9.2.1, since

$$E\left[\sum_{i=1}^{n} f(t_{i-1})\left(W(t_i) - W(t_{i-1})\right)\right] = \sum_{i=1}^{n} f(t_{i-1})E\left[(W(t_i) - W(t_{i-1}))\right] = 0,$$

by (10.22) of lemma 10.2.4 above.

3.
$$\mathrm{Var}\left[\int_a^b f(t)dW(t)\right] = \int_a^b f^2(t)dt \tag{10.43}$$

This follows again as in the proof of the analogous statement in theorem 9.2.1, since by theorem 10.2.6 the increments of the Wiener process over non-overlapping intervals are independent,

$$\mathrm{Var}\left[\sum_{i=1}^{n} f(t_{i-1})\left(W(t_i) - W(t_{i-1})\right)\right] = \sum_{i=1}^{n} f^2(t_{i-1})\mathrm{Var}\left[\left(W(t_i) - W(t_{i-1})\right)\right],$$

$$= \sum_{i=1}^{n} f^2(t_{i-1})(t_i - t_{i-1}),$$

by (10.22) of lemma 10.2.4 above. Then

$$\sum_{i=1}^{n} f^2(t_{i-1})(t_i - t_{i-1}) \to \int_a^b f^2(t)dt,$$

as $a = t_0 < t_1 < \ldots < t_{n-1} < t_n = b$ and $\max_i |t_i - t_{i-1}| \to 0$ as $n \to \infty$.

4. Evidently $\sum_{i=1}^{n} f(t_{i-1})\left(W(t_i) - W(t_{i-1})\right)$ is a Gaussian random variable. By properties of convergence in mean square of sequences of Gaussian random variables, see theorem 7.4.2 in section 7.4.3, and by (10.42) and (10.43) we obtain

$$\int_a^b f(t)dW(t) \in N\left(0, \int_a^b f^2(t)dt\right). \tag{10.44}$$

5. If $\int_a^b f^2(t)dt < \infty$ and $\int_a^b g^2(t)dt < \infty$,

$$E\left[\int_a^b f(t)dW(t)\int_a^b g(t)dW(t)\right] = \int_a^b f(t)g(t)dt. \tag{10.45}$$

Here we see a case of the heuristics in (10.38) in operation, too. To prove this, we fix $a = t_0 < t_1 < \ldots < t_{n-1} < t_n = b$ and start with the approximating sums, or,

$$E\left[\sum_{i=1}^{n} f(t_{i-1})\left(W(t_i) - W(t_{i-1})\right) \cdot \sum_{j=1}^{n} g(t_{j-1})\left(W(t_j) - W(t_{j-1})\right)\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} f(t_{i-1})g(t_{j-1})E\left[\left(W(t_i) - W(t_{i-1})\right) \cdot \left(W(t_j) - W(t_{j-1})\right)\right]$$

and as by theorem 10.2.6 the increments of the Wiener process over non-overlapping intervals are independent,

$$= \sum_{i=1}^{n} f(t_{i-1})g(t_{i-1})(t_i - t_{i-1}) \to \int_a^b f(t) \cdot g(t)dt,$$

as $a = t_0 < t_1 < \ldots < t_{n-1} < t_n = b$ and $\max_i |t_i - t_{i-1}| \to 0$ as $n \to \infty$.

6. By the preceding we can define a new process with variables $Y(t)$ by

$$Y(t) = \int_0^t h(s)dW(s).$$

Then (10.45) can be manipulated to deliver

$$E\left[Y(t)\cdot Y(s)\right]=\int_{0}^{\min(t,s)}h^{2}(u)du. \tag{10.46}$$

To establish this claim, let $\mathbf{I}_{[0,t]}(u)=1$, if $0\leq u\leq t$ and $\mathbf{I}_{[0,t]}(u)=0$ otherwise, and

$$f(u)=\mathbf{I}_{[0,t]}(u)\cdot h(u),\quad g(u)=\mathbf{I}_{[0,s]}(u)\cdot h(u), \tag{10.47}$$

take $a=0,b=+\infty$, and then

$$Y(t)=\int_{0}^{\infty}f(u)dW(u),\quad Y(s)=\int_{0}^{\infty}g(u)dW(u).$$

By insertion we see that (10.46) is a special case of (10.45).

Example 10.5.2 The Wiener integral satisfies

$$W(t)\overset{d}{=}\int_{0}^{t}dW(s). \tag{10.48}$$

This is natural, but cannot be argued by differentiation. To establish (10.48) we write using (10.47)

$$Y(t)\overset{\text{def}}{=}\int_{0}^{t}dW(s)=\int_{0}^{\infty}\mathbf{I}_{[0,t]}(s)dW(s).$$

Clearly $\{Y(t)\mid t\geq 0\}$ is a Gaussian process. Then (10.46) entails

$$E\left[Y(t)\cdot Y(s)\right]=\int_{0}^{\min(t,s)}du=\min(t,s), \tag{10.49}$$

which shows that $\{\int_0^t dW(s)\mid t\geq 0\}$ is a Wiener process, and (10.48) is verified (as an equality in distribution).

■

Example 10.5.3 By (10.48) and the first property of the Wiener integral we can write for any $\tau>0$

$$W(t+\tau)-W(t)\overset{d}{=}\int_{t}^{t+\tau}dW(s). \tag{10.50}$$

We note in this regard, e.g., that by (10.43)

$$\text{Var}\left[W(t+\tau)-W(t)\right]=\text{Var}\left[\int_{t}^{t+\tau}dW(s)\right]=\int_{t}^{t+\tau}ds=\tau,$$

as it should, c.f., (10.20). The integral in (10.50) is sometimes called a **gliding window smoother**, see [97].

■

10.5.3 The Wiener Integral is a Scrambled Wiener Process

Suppose now that $\int_0^\infty f^2(t)dt<\infty$ and

$$Y(t)=\int_{a}^{t}f(u)dW(u).$$

We assume for the sake of simplicity that $f(u) > 0$ for all u. We let

$$\tau(t) = \inf\left\{ s \mid \int_0^s f^2(u)du = t \right\},$$

or, $\tau(t)$ is the time, when $\int_0^s f^2(u)du$ as an increasing function of s first reaches the level $t > 0$. Evidently $t \mapsto \tau(t)$ is one-to-one, or, invertible, and the inverse is

$$\tau^{-1}(s) = \int_0^s f^2(u)du.$$

Let us look at $Y(\tau(t))$. Then from (10.43)

$$E\left[Y^2(\tau(t))\right] = \int_a^{\tau(t)} f^2(u)du = \tau^{-1}(\tau(t)) = t.$$

Hence, if we define

$$V(t) = Y(\tau(t)),$$

then $\{V(t) \mid t \geq 0\}$ is a Wiener process. Furthermore,

$$Y(t) = V\left(\tau^{-1}(t)\right) \overset{d}{=} W\left(\tau^{-1}(t)\right), \tag{10.51}$$

which shows that a Wiener integral is a Wiener process on a distorted or scrambled time scale.

10.5.4 White Noise

In the engineering literature, see, e.g., [8, 32, 56, 71, 80, 85, 97] as well as in physics [58, p.255], [62, pp. 66−69], one encounters the 'random process' with variables $\overset{o}{W}$ such that

$$E\left[\overset{o}{W}(t)\overset{o}{W}(s)\right] = \delta(t-s), \tag{10.52}$$

where $\delta(t-s)$ is the **Dirac delta**, see [96, p. 354]. As stated in [96, loc.cit], $\delta(t-s)$ is not a function in the ordinary sense, but has to be regarded as a distribution, not in the sense of probability theory, but in the sense of the theory of generalized functions (which is a class of functionals on the set of infinitely differentiable functions with support in a bounded set).

Let us, as a piece of formal treatment, c.f., (10.48), set

$$W(t) = \int_0^t \overset{o}{W}(u)du. \tag{10.53}$$

Then we get by a formal manipulation with the rules for integrals above that

$$\begin{aligned} E\left[W(t)W(s)\right] &= \int_0^t \int_0^s E\left[\overset{o}{W}(u)\overset{o}{W}(v)\right] dudv \\ &= \int_0^t \int_0^s \delta(u-v)dudv. \end{aligned}$$

The 'delta function' $\delta(u-v)$ is zero if $u \neq v$ and acts (inside an integral) according to

$$\int_{-\infty}^{\infty} f(v)\delta(u-v)dv = f(u).$$

Then with $f(v) = \mathbf{I}_{[0,s]}(v)$ we get

$$\int_0^s \delta(u-v)dv = \int_{-\infty}^{\infty} \mathbf{I}_{[0,s]}(v)\delta(u-v)dv = \begin{cases} 0 & \text{if } u > s \\ 1 & \text{if } u < s. \end{cases}$$

Thus

$$\int_0^t \int_0^s \delta(u-v)dudv = \int_0^{\min(t,s)} dv = \min(t,s).$$

Hence, if we think of the process with variables $\overset{o}{W}(u)$ as being Gaussian, then the process introduced by the variables $W(t)$ in (10.53) is like a Wiener process ! Of course, by (10.53) one should get then

$$\frac{d}{dt}W(t) = \overset{o}{W}(t),$$

which is not possible, as the Wiener process has almost surely non-differentiable sample paths. Hence, the white noise makes little, or perhaps, should make no sense. One can, nevertheless, introduce linear time invariant filters, c.f. the exercises in section 9.7.6, with white noise as input, or

$$Y(t) = \int_{-\infty}^{\infty} G(t-s)\overset{o}{W}(s)ds,$$

and compute the autocovariances and spectral densities of the output process in a very convenient way. Thus, despite of the fact that the white noise does not exist as a stochastic process in our sense, it can be formally manipulated to yield useful results, at least as long as one does not try to do any non-linear operations. A consequence of (10.52) is that the spectral density of the white noise is a constant for all frequencies,

$$s_{\overset{o}{W}}(f) = 1, \quad -\infty < f < \infty. \tag{10.54}$$

(To 'check' this, insert $s_{\overset{o}{W}}(f) = 1$ in the right hand side of (9.25).) In engineering, see, e.g., [77, 105], the white noise is thought of as an approximation of a weakly stationary process that has a power spectral density which is constant over very wide band of frequencies and then equals to, or decreases rapidly to, zero. An instance of this argument will be demonstrated later in section 11.4 on thermal noise.

10.6 Martingales and the Wiener Process

This section on martingales in continuous time is, for one more time, just a scratch on the surface of an extensive theory, [29], [70, chapter 7], [67, chapter 7]. We extend here the concepts in section 3.8.5. We shall first need to define the notion of a **filtration**.

Definition 10.6.1 Let $\mathcal{F}$ be a sigma field of subsets of Ω. Let T be an index set, so that for every $t \in T$, $\mathcal{F}_t$ is a sigma field $\subset \mathcal{F}$ and that

$$\mathcal{F}_s \subset \mathcal{F}_t \quad s < t. \tag{10.55}$$

Then we call the family of sigma fields $(\mathcal{F}_t)_{t\in T}$ a **filtration**.

■

This leads immediately to the next definition, that of a martingale.

Definition 10.6.2 Let $\mathbf{X} = \{X(t) \mid t \in T\}$ is a stochastic process on $(\Omega, \mathcal{F}, \mathbf{P})$. Then we call $\mathbf{X}$ a **martingale with respect to the filtration** $(\mathcal{F}_t)_{t\in T}$, if

1. $E[\mid X(t) \mid] < \infty$ for all $t \in T$.

2. $X(t)$ is measurable with respect to $\mathcal{F}_t$ for each $t \in T$.

3. For $s \leq t$ the **martingale property** holds:

$$E[X(t) \mid \mathcal{F}_s] = X(s). \tag{10.56}$$

■

Let now $\mathbf{W} = \{W(t) \mid 0 \leq t\}$ be a Wiener process. We define the filtration

$$\mathcal{F}_t^{\mathbf{W}} \stackrel{\text{def}}{=} \text{ the sigma field generated by } W(s) \text{ for } 0 \leq s \leq t.$$

We write this as

$$\mathcal{F}_t^{\mathbf{W}} = \sigma\left(W(s); 0 \leq s \leq t\right).$$

We should read this according to the relevant definition 1.5.3 in chapter 1. We take any number of indices $t_1, \ldots, t_n$, all $t_i \leq s$. The sigma field $\mathcal{F}_{t_1,\ldots,t_n,s}^{\mathbf{W}}$ generated by the random variables $W(t_i)$ $i = 1, \ldots, n$, is defined to be the smallest σ field containing all events of the form $\{\omega : W(t_i)(\omega) \in A\} \in \mathcal{F}$, $A \in \mathcal{B}$, where $\mathcal{B}$ is the Borel σ field over $\mathbf{R}$.

By independent increments, theorem 10.2.6, and lemma 10.2.4, eq. (10.22), we get that

$$E\left[W(t) - W(s) \mid \mathcal{F}_{t_1,\ldots,t_n,s}^{\mathbf{W}}\right] = E\left[W(t) - W(s)\right] = 0. \tag{10.57}$$

Clearly sigma fields like $\mathcal{F}_{t_1,\ldots,t_n,s}^{\mathbf{W}}$ generate $\mathcal{F}_s^{\mathbf{W}}$, so that $\mathcal{F}_{t_1,\ldots,t_n,s}^{\mathbf{W}} \subset \mathcal{F}_s^{\mathbf{W}}$ entails by double expectation

$$E\left[W(t) - W(s) \mid \mathcal{F}_s^{\mathbf{W}}\right] = E\left[E\left[W(t) - W(s) \mid \mathcal{F}_s^{\mathbf{W}}\right] \mid \mathcal{F}_{t_1,\ldots,t_n,s}^{\mathbf{W}}\right]$$

and by the tower property and (10.57)

$$= E\left[W(t) - W(s) \mid \mathcal{F}_{t_1,\ldots,t_n,s}^{\mathbf{W}}\right] = 0,$$

$$\Leftrightarrow E\left[W(t) \mid \mathcal{F}_s^{\mathbf{W}}\right] = E\left[W(s) \mid \mathcal{F}_s^{\mathbf{W}}\right],$$

but since $W(s)$ is by construction $\mathcal{F}_s^{\mathbf{W}}$ -measurable, the rule of taking out what is known gives the martingale property

$$E\left[W(t) \mid \mathcal{F}_s^{\mathbf{W}}\right] = W(s). \tag{10.58}$$

Since $E\left[\mid W(t) \mid\right] < \infty$, we have the following theorem.

Theorem 10.6.1 $\mathbf{W} = \{W(t) \mid t \geq 0\}$ is a Wiener process and the sigma field is $\mathcal{F}_t^{\mathbf{W}} = \sigma\left(W(s); 0 \leq s \leq t\right)$, then $\mathbf{W}$ is a martingale with respect to the filtration $\left(\mathcal{F}_t^{\mathbf{W}}\right)_{t \geq 0}$.

■

This has to be regarded as a very significant finding, because there is a host of inequalities and convergence theorems e.t.c., that hold for martingales in general, and thus for the Wiener process. In addition, the martingale property is of crucial importance for stochastic calculus.

While we are at it, we may note the following re-statement of the Markov property (10.39) in theorem 10.4.5.

Theorem 10.6.2 $\mathbf{W}$ is a Wiener process and the filtration is $\left(\mathcal{F}_t^{\mathbf{W}}\right)_{t \geq 0}$, where $\mathcal{F}_t^{\mathbf{W}} = \sigma\left(W(s); 0 \leq s \leq t\right)$. Then, if $s < t$ and $y \in R$, it holds almost surely that

$$\mathbf{P}\left(W(t) \leq y \mid \mathcal{F}_s^{\mathbf{W}}\right) = \mathbf{P}\left(W(t) \leq y \mid W(s)\right). \tag{10.59}$$

■

10.7 Exercises

10.7.1 Random Walks

Random walk is a mathematical statement about a trajectory of an object that takes successive random steps. Random walk is one of the most important and most studied topics in probability theory. The exercises on random walks in this section are adapted from [10, 48] and [78, chapter 9.1]. We start with the first properities of the (unrestricted) random walk, and then continue to find the connection to the Wiener process, whereby we can interpret a random walk as the path traced by a molecule as it travels in a liquid or a gas and collides with other particles [10, 78].

1. Let $\{X_i\}_{i=1}^{\infty}$ be I.I.D. random variables with two values so that $X_i = +1$ with probability p and $X_i = -1$ with probability $q = 1 - p$. We let

$$S_n = X_1 + X_2 + \ldots + X_n, S_0 = 0 \tag{10.60}$$

The sequence of random variables $\{S_n\}_{i=0}^{\infty}$ is called a **random walk**. We can visualize the random walk as a particle jumping on a lattice of sites $j = 0, \pm 1, \pm 2, \ldots$ starting at time zero in the site 0. At any n the random walk currently at j jumps to the right to the site $j + 1$ with probability p or to the left to the site $j - 1$ with probability q. A random walk is thus constructed also a time- and space-homogeneous finite Markov chain, see [95, lecture 6] for a treatment from this point of view.

(a) Show that

$$\mathbf{P}\left(S_n = j\right) = \left(\begin{array}{c} n \\ \frac{n+j}{2} \end{array}\right) p^{\frac{n+j}{2}} q^{\frac{n-j}{2}}. \tag{10.61}$$

Aid: We hint at a combinatorial argument. Consider the random variable $R(n)$ defined by

$$R(n) \overset{\text{def}}{=} \text{the number of steps to the right in } n \text{ steps.}$$

Then it clearly (c.f., figure **??**) holds that

$$S_n = R(n) - (n - R(n)),$$

and hence, if $S_n = j$, then $R(n) = \frac{n+j}{2}$. Next, find the number of paths of the random walk such that $R(n) = \frac{n+j}{2}$ and $S_n = j$. Find the probability of each of these paths and sum up them.

(b) **Reflection principle** Let

$$S_n = X_1 + X_2 + \ldots + X_n + a.$$

Thus $S_0 = a$ and we suppose that $S_n = b$. Consider

$$N_n(a, b) = \text{the number of possible paths from } a \text{ to } b.$$

Let

$$N_n^0(a, b) = \text{the number of paths from } a \text{ to } b\text{, which pass } 0.$$

Show that if $a, b > 0$, then

$$N_n^0(a, b) = N_n(-a, b).$$

(c) Show that

$$N_n(a, b) = \left(\begin{array}{c} n \\ \frac{n+b-a}{2} \end{array}\right).$$

2. (From [43, 78]) Determine the characteristic function $\varphi_{S_n}(t)$ of the random walk S_n in (10.60), and use $\varphi_{S_n}(t)$ to find the probability expression (10.61).

3. Here we impose a parameter ($= \delta$) on the random walk in the preceding exercises. Thus, for all $i \geq 1$, $X_i = \delta > 0$ with probability $\frac{1}{2}$, $X_i = -\delta$ with probability $\frac{1}{2}$, and $\{X_i\}_{i=1}^{\infty}$ are independent.

   We let
   $$S_n = X_1 + X_2 + \ldots + X_n, S_0 = 0.$$

   This sequence of random variables $\{S_n\}_{i=0}^{\infty}$ is called a **symmetric random walk**. We can visualize $\{S_n\}_{i=0}^{\infty}$ as a particle jumping on sites $j\delta$ on a lattice with $\{0, \pm\delta, \pm 2\delta, \ldots\}$ with δ as **lattice spacing** or a **step size**. The random walk starts at time zero in the site 0.

   (a) Show that
   $$(S_n, \sigma(X_1, X_2, \ldots, X_n))_{n\geq 0}$$
   is a martingale.

   (b) Show that for $m \geq 0, n \geq 0$,
   $$E[S_n S_m] = \delta^2 \min(n, m). \tag{10.62}$$

   (c) Show that
   $$\frac{S_n}{\sqrt{n}\delta} \xrightarrow{d} N(0,1),$$
   as $n \to \infty$, i.e., that approximately for large n,
   $$S_n \in N(0, n\delta^2). \tag{10.63}$$

   Remark 10.7.1 In view of (10.15) we can write the random variables $W(t_1), \ldots, W(t_n)$ of a Wiener process in the form
   $$W(t_i) = \sum_{k=1}^{i} Z_k,$$
   where $Z_k = W(t_k) - W(t_{k-1})$s are I.I.D. for $t_k - t_{k-1}$ =constant. Since we have (10.62) and (10.63), the symmetric random walk has a certain similarity with the Wiener process sampled at equidistant times, recall even (10.18).

   ■

4. The random walk is as in the preceding exercise except that for any integer $n \geq 0$ and for all $i \geq 1$, $X_i^{(n)} = \delta_n > 0$ with probability $\frac{1}{2}$, $X_i^{(n)} = -\delta_n$ with probability $\frac{1}{2}$, and $\{X_i^{(n)}\}_{i=1}^{\infty}$ are independent. Here δ_n is a sequence of positive numbers such that
   $$\delta_n \to 0,$$
   as $n \to \infty$. In words, the spacing of the lattice becomes denser (taken as a subset of the real line) or, in other words, the step size becomes smaller.

   We impose a second sequence of parameters, $\tau_n > 0$ for the purpose of imbedding the random walk in continuous time. For a fixed n we can think of a particle undergoing a random walk moving to right or to left on the lattice $\{0, \pm\delta_n, \pm 2\delta_n, \ldots\}$ at every τ_n seconds. We assume that
   $$\tau_n \to 0,$$

as $n \to \infty$. We have, for some time t in seconds,

$$k = \lfloor \frac{t}{\tau_n} \rfloor.$$

Hence, it takes an inceasing number of steps to make a walk of length t in time. Let us now assume that for some $\delta > 0$ and $\tau > 0$

$$\lfloor \frac{t}{\tau_n} \rfloor \delta_n^2 \to t\frac{\delta^2}{2\tau}, \tag{10.64}$$

as $n \to \infty$. E.g., the choice $\delta_n = \frac{\delta}{\sqrt{n}}$, $\tau_n = \frac{2\tau}{n}$ satisfies all of the requirements above.

For a fixed n we have the random walk

$$S_k^{(n)} = X_1^{(n)} + X_2^{(n)} + \ldots + X_k^{(n)}, S_0^{(n)} = 0.$$

In view of (10.63) we get for fixed n and large k that

$$S_k^{(n)} \in N(0, k\delta_n^2). \tag{10.65}$$

Define for $t \geq 0$

$$W^{(n)}(t) \overset{\text{def}}{=} \sum_{i=1}^{\lfloor \frac{t}{\tau_n} \rfloor} X_i^{(n)}, \tag{10.66}$$

so that $W^{(n)}(k\tau_n) = S_k^{(n)}$. The process $\{W^{(n)}(t) \mid t \geq 0\}$ is a random walk in continuous time, or, a stepwise constant interpolation of $\{S_k^{(n)}\}_{k=0}^{\infty}$.

(a) Show that for any $t \geq 0$

$$W^{(n)}(t) \overset{d}{\to} N\left(0, t\frac{\delta^2}{2\tau}\right), \tag{10.67}$$

as $n \to \infty$.

(b) Show that for any $t > 0$ and $s \geq 0$

$$\lim_{n\to\infty} E\left[W^{(n)}(t)W^{(n)}(s)\right] = \frac{\delta^2}{2\tau}\min(t,s). \tag{10.68}$$

In the above we have produced evidence for the statement that a sequence $\mathbf{W}_n$ of symmetric random walks interpolated in continuous time and with decreasing stepsize and with increasing number of steps (per unit of time) will (c.f.,(10.64)), for increasing n approximate a (non-standard, see (10.18)) Wiener process, as would seem natural in view of the remark 10.7.1.

A pedagogical point of the exercise is that one can model and/or understand diffusion without thermodynamics. We can regard $D \overset{d}{=} \frac{\delta^2}{2\tau}\left[\frac{\text{cm}^2}{\text{sec}}\right]$ as a diffusion coefficient, and $\frac{\delta}{\tau}$ as instantaneous velocity, c.f., [10]. We have found for large n again the result (10.1) on root-mean squared displacement of a diffusing particle,

$$\sqrt{E\left[\left(W^{(n)}(t)\right)^2\right]} = \sqrt{2Dt}.$$

We shall rediscover this finding using the Langevin dynamics in chapter 11.

■

The random walks above are known as **unrestricted random walks**. Random walks with **absorbing and/or reflecting boundaries**, are concisely analyzed in [15]. The mathematical results on random walks have been applied in computer science, physics, molecular biology, ecology, economics, psychology and a number of other fields. For example,the price of a fluctuating stock and the financial status of a gambler have been studied by random walks.

10.7.2 Wiener Process

1. **The Bivariate Distribution for** $(W(s), W(t))$ Let $0 < s < t$. We know that for the Wiener process

$$(W(s), W(t))' \in N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} s & s \\ s & t \end{pmatrix} \right).$$

Set $C_{\mathbf{W}} = \begin{pmatrix} s & s \\ s & t \end{pmatrix}$. Then there is the bivariate joint normal p.d.f. of the form

$$f_{W(s),W(t)}(x,y) = \frac{1}{2\pi\sqrt{\det C_{\mathbf{W}}}} e^{-\frac{1}{2}(x,y)C_{\mathbf{W}}^{-1}\begin{pmatrix} x \\ y \end{pmatrix}}. \tag{10.69}$$

On the other hand, by our definition of the Wiener process, we have (10.8) or

$$f_{W(s),W(t)}(x,y) = \frac{1}{\sqrt{2\pi s}} e^{-\frac{x^2}{2s}} \frac{1}{\sqrt{2\pi(t-s)}} e^{-\frac{(y-x)^2}{2(t-s)}}. \tag{10.70}$$

Verify by explicit expansions that (10.69) and (10.70) are one and the same p.d.f..

2. **Shift** Let $\tau > 0$. $\{W(t) | t \geq 0\}$ is a Wiener process. Define

$$V(t) \stackrel{\text{def}}{=} W(t+\tau) - W(\tau). \tag{10.71}$$

Show that $\mathbf{V} = \{V(t) | t \geq 0\}$ is a Wiener process.

3. **Brownian Scaling** Let $c > 0$ and $\{W(t) | t \geq 0\}$ be a Wiener process. Define

$$V(t) \stackrel{\text{def}}{=} \frac{1}{c} W(c^2 t). \tag{10.72}$$

Show that $\mathbf{V} = \{V(t) | t \geq 0\}$ is a Wiener process.

4. **Time Inversion** $\{W(t) | t \geq 0\}$ is a Wiener process. Define

$$V(t) \stackrel{\text{def}}{=} tW\left(\frac{1}{t}\right). \tag{10.73}$$

Show that $\mathbf{V} = \{V(t) | t \geq 0\}$ is a Wiener process.

5. **Time Reversal** $\{W(t) | t \geq 0\}$ is a Wiener process. Define

$$V(t) \stackrel{\text{def}}{=} W(1) - W(1-t), \quad 0 \leq t \leq 1. \tag{10.74}$$

Show that $V(t) \stackrel{d}{=} W(t)$ for $0 \leq t \leq 1$.

6. **Brownian Bridge** We give a first a general definition from [19, p.64]. Let $x, y \in \mathbf{R}$ and $l > 0$. A Gaussian process $\mathbf{X} = \{X(t) \mid 0 \leq t \leq l\}$ with continuous sample paths and $X(0) = x$ such that

$$\mu_{\mathbf{X}}(t) = x + (y-x)\frac{t}{l}, \quad \text{Cov}\,(X(t), X(s)) = \min(s,t) - \frac{st}{l}$$

is called a **Brownian Bridge from** x **to** y **of length** l or a tied-down Wiener process. Note that $\mu_{\mathbf{X}}(l) = y$ and $\text{Cov}\,(X(s), X(t)) = 0$ if $s = l$ or $t = l$, and hence $X(l) = y$.

(a) (From [19, p.64]) Let $\{W(t)|t \geq 0\}$ be a Wiener process. Show that if $\mathbf{X}^{x,l}$ is a Brownian Bridge from x to x of length l, then

$$X^{x,l}(t) \overset{d}{=} x + W(t) - \frac{t}{l} W(l), \quad 0 \leq t \leq l. \tag{10.75}$$

(b) (From [19, p.64]) Let $\{W(t)|t \geq 0\}$ be a Wiener process. Show that if $\mathbf{X}^{x,l}$ is a Brownian Bridge from x to x of length l, then

$$X^{x,l}(t) \overset{d}{=} x + \frac{l-t}{l} W\left(\frac{lt}{l-t}\right), \quad 0 \leq t \leq l. \tag{10.76}$$

(c) (From [19, p.64]) Show that if $\mathbf{X}$ is a Brownian Bridge from x to y of length l, then

$$X(t) \overset{d}{=} X^{x,l}(t) + (y-x)\frac{t}{l}, \quad 0 \leq t \leq l, \tag{10.77}$$

where $X^{x,l}(t)$ is a random variable of $\mathbf{X}^{x,l}$, as in (a) and (b) above.

(d) Define the process

$$B(t) \overset{\text{def}}{=} W(t) - tW(1), \quad 0 \leq t \leq 1. \tag{10.78}$$

This is a process is tied down by the condition $B(0) = B(1) = 0$, Brownian bridge from 0 to 0 of length 1. Compare with (9.61) in the preceding.

(i) Show that the autocorrelation function of $\{B(t)|0 \leq t \leq 1\}$ is

$$R_B(t,s) = \begin{cases} s(1-t) & s \leq t \\ (1-s)t & s \geq t. \end{cases} \tag{10.79}$$

(ii) Show that the increments of the Brownian bridge are not independent.

(iii) Let $\mathbf{B} = \{B(t) \mid 0 \leq t \leq 1\}$ be a Brownian bridge. Show that the Wiener process in $[0,T]$, in the sense of (10.18), can be decomposed as

$$W(t) = B\left(\frac{t}{\sqrt{T}}\right) + \frac{t}{\sqrt{T}} \cdot Z,$$

where $Z \in N(0,1)$ and is independent of $\mathbf{B}$.

(iv) Show that for $0 \leq t < \infty$

$$W(t) = (1+t)B\left(\frac{t}{1+t}\right).$$

Compare with (9.62) in the preceding.

7. **The Reflection Principle and The Maximum of a Wiener Process in a Finite Interval** Let us look at the collection of all sample paths $t \mapsto W(t)$ of a Wiener process $\{W(t)|t \geq 0\}$ such that $W(T) > a$, where $a > 0$ and $T > 0$, here T is a time point. Since $W(0) = 0$, there exists a time τ, a random variable depending on the particular sample path, such that $W(\tau) = a$ for the first time.

For $t > \tau$ we **reflect** $W(t)$ around the line $x = a$ to obtain

$$\widetilde{W}(t) \overset{\text{def}}{=} \begin{cases} W(t) & \text{if } t < \tau \\ a - (W(t) - a) & \text{if } t > \tau. \end{cases} \tag{10.80}$$

Note that if $t > \tau$, then $\widetilde{W}(t) - a = -(W(t) - a)$, reflection. Thus $\widetilde{W}(T) < a$, since $W(T) > a$ (draw a picture).

In fact we are saying that

$$\widetilde{W}(t) = \begin{cases} W(t) & \text{if } t < \tau \\ 2W(\tau) - W(t) & \text{if } t > \tau. \end{cases} \tag{10.81}$$

and that $\widetilde{W}(t)$ is a Wiener process, but we have to admit that the proof is beyond the resources of these lectures[10]

(a) Show that for $t > \tau$,

$$\mathbf{P}\left(W(t) \leq x + a \mid W(\tau) = a\right) = \mathbf{P}\left(W(t) \geq a - x \mid W(\tau) = a\right), \tag{10.82}$$

where it may be useful to employ $\widetilde{W}(t)$.

The equation (10.82) says that the probability law of the process for $t > \tau$, given $W(\tau) = a$, is symmetrical with respect to $x > a$ and $x < a$. In addition, the process is independent of the past of the process prior to τ.

(b) Set now

$$M_{0 \leq u \leq T} \stackrel{\text{def}}{=} \max_{0 \leq u \leq T} W(u), \widetilde{M}_{0 \leq u \leq T} \stackrel{\text{def}}{=} \max_{0 \leq u \leq T} \widetilde{W}(u).$$

$t \mapsto W(t)$ and $t \mapsto \widetilde{W}(t)$ with the same probability of occurrence and such that

$$M_{0 \leq u \leq T} \geq a, \widetilde{M}_{0 \leq u \leq T} \geq a.$$

Conversely, by the nature of this correspondence, every sample function $t \mapsto W(t)$ for which $M_{0 \leq u \leq T} \geq a$ results from either of the two sample functions $t \mapsto W(t)$ and $t \mapsto \widetilde{W}(t)$ with equal probability, one of which is such that $W(T) > a$ unless $W(T) = a$, but $\mathbf{P}(W(T) = a) = 0$. Show now that

$$\mathbf{P}\left(M_{0 \leq u \leq T} \geq a\right) = \frac{2}{T\sqrt{2\pi}} \int_a^\infty e^{-x^2/2T} dx. \tag{10.83}$$

8. **The Ornstein-Uhlenbeck Process** $\{W(t) | t \geq 0\}$ is a Wiener process. Let $a > 0$. Define for all $t \geq 0$

$$X(t) \stackrel{\text{def}}{=} e^{-at} W\left(e^{2at}\right). \tag{10.84}$$

Show that $\mathbf{X} = \{X(t) | t \geq 0\}$ is a Gaussian process with mean function

$$\mu_{\mathbf{X}}(t) = 0$$

and autocorrelation function

$$R_{\mathbf{X}}(t, s) = e^{-a|t-s|}. \tag{10.85}$$

A comparison of this with (11.12) below shows (with scaling $\sigma^2 = 2a$) that the process $\mathbf{X} = \{X(t) | t \geq 0\}$ is an Ornstein-Uhlenbeck process. How is (10.84) related to the representation (9.39) ? Are the sample paths of $\mathbf{X} = \{X(t) | t \geq 0\}$ almost surely continuous ? The trick in (10.84) is known as the **Lamperti transform**.

9. **The Geometric Brownian Motion** $\{W(t) | t \geq 0\}$ is a Wiener process. Let a, $\sigma > 0$ and x_0 be real constants. Then

$$X(t) \stackrel{\text{def}}{=} x_0 e^{\left(\alpha - \frac{1}{2}\sigma^2\right)t + \sigma W(t)}.$$

Show that

[10] The statements to follow are true, but the complete analysis requires strictly speaking the so called **strong Markov property** [70, p. 73]. For handling the strong Markov property one has definite advantage of both the techniques and the 'jargon of modern probability', but an intuitive physics text like [62, p.57−58] needs, unlike us, to pay no lip service to this difficulty in dealing with the reflection principle.

(a)

$$E\left[X(t)\right] = x_0 e^{\alpha t}.$$

(b)

$$\mathrm{Var}(X(t) = x_0^2 e^{2\alpha t}\left(e^{\sigma^2 t} - 1\right).$$

10. $\{W_i(t) | t \geq 0\}$, $i = 1, 2$, are two independent Wiener processes. Define a new stochastic process $\{V(t)| - \infty < t < +\infty\}$ by

$$V(t) = \begin{cases} W_1(t) & t \geq 0 \\ W_2(-t) & t < 0. \end{cases}$$

   (a) Find $P\left(V(1/2) - V(-1/2) > 1\right)$.

   (b) Show that $\{V(t+1) - V(t)| - \infty < t < +\infty\}$ is a stationary Gaussian process.

11. **Differentiation** $\mathbf{W}$ is a Wiener process. Show that

$$\frac{\mid h \mid}{\mid |W(t+h) - W(t) \mid} \xrightarrow{p} 0,$$

as $h \to 0$.

12. **Partial diffential equation for a functional of the Wiener process** Let $h(x)$ be a bounded and continuous function defined in the whole real line. $\{W(t)|t \geq 0\}$ is a Wiener process. Set

$$u(t,x) = E\left[h(x + W(t))\right].$$

Show that $u(t,x)$ satisfies

$$\frac{\partial}{\partial t}u(t,x) = \frac{1}{2}\frac{\partial^2}{\partial x^2}u(t,x), \quad u(0,x) = h(x). \tag{10.86}$$

Aid: Recall (10.3).

13. **Fractional Brownian Motion** $\mathbf{W}_H = \{W_H(t) \mid 0 \leq t < \infty\}$ is a Gaussian stochastic process. Its expectation is $= 0$ and its autocorrelation function equals

$$R_{\mathbf{W}_H}(t,s) = E\left[W_H(t)W_H(s)\right] = \frac{1}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right), \tag{10.87}$$

where $0 < H < 1$ (H is known as the Hurst parameter).

   (a) Show that $W_H(t) \overset{d}{=} \frac{1}{a^H} W_H(at)$, where $a > 0$.

   (b) Which process do we obtain for $H = \frac{1}{2}$?

   (c) Define the random variable

$$Y = W_H(t+h) - W_H(t)$$

   where $h > 0$. What is the distribution of Y ?

   (d) Show that

$$Y = W_H(t+ah) - W_H(t) \overset{d}{=} a^H W_H(h),$$

   which means that $\mathbf{W}_H$ is the same in all time scales. This implies also that its **sample paths are fractals**[11].

[11] For this and other statements given here, see
B.B. Mandelbrot & J.W. van Ness: Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Review*, vol. 10, 1968, pp. 422−437.

14. **Fractional Gaussian Process** Let $1/2 < H < 1$

$$W_{H,\delta}(t) \stackrel{\text{def}}{=} \frac{1}{\delta}\left(W_H(t+\delta) - W_H(t)\right),$$

where $\mathbf{W}_H$ is the fractional Brownian motion with the autocorrelation function

$$R_{\mathbf{W}_H}(t,s) = E\left[W_H(t)W_H(s)\right] = \frac{v_H}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right), \tag{10.88}$$

where $\frac{1}{2} < H < 1$ and

$$v_H = -\frac{\Gamma(2-2H)\cos(\pi H)}{\pi H(2H-1)}.$$

Show that $\mathbf{W}_{H,\delta}$ is a stationary Gaussian process with zero mean and with autocorrelation function

$$R_{\mathbf{W}_{H,\delta}}(h) = \frac{v_H \delta^{2H-2}}{2}\left(\left(\frac{|\,h\,|}{\delta}+1\right)^{2H} - 2\left(\frac{|\,h\,|}{\delta}\right)^{2H} + \left|\frac{|\,h\,|}{\delta}-1\right|^{2H}\right).$$

It can be shown that the power spectral density of $s_{\mathbf{W}_{H,\delta}}(f)$ is approximately

$$s_{\mathbf{W}_{H,\delta}}(f) \approx f^{1-2H}.$$

This implies that the increments of $\mathbf{W}_H$ are a good model for so called $1/f$ **-type noise** encountered, e.g., in electric circuits [11]. The $1/f$ -type noise models physical processes with long range dependencies.

10.7.3 The Wiener Integral

1. Let $\tau > 0$ and

$$Y(t) \stackrel{d}{=} \frac{1}{\tau}\int_t^{t+\tau} dW(s). \tag{10.89}$$

Note that (10.48) implies that

$$Y(t) \stackrel{d}{=} \frac{1}{\tau}\left(W(t+\tau) - W(t)\right).$$

(a) Show that the autocorrelation function is

$$R_{\mathbf{Y}}(h) = \frac{1}{\tau}\max\left(0, 1-\frac{|h|}{\tau}\right), \quad h = t-s. \tag{10.90}$$

(b) Use the Fourier transform of $R_{\mathbf{Y}}(h)$ (use the table of transform pairs in section 9.3.1 above) to argue that $Y(t)$ approaches white noise as $\tau \to 0$.

2. Let

$$Z(t) \stackrel{d}{=} \int_0^t s^2 dW(s).$$

for $t \in [0,T]$.

(a) Find the scrambled representation (10.51) for $Z(t)$.

(b) (From [62, p.52]) Set for a constant a

$$X(t) \stackrel{d}{=} aZ(t) + t, \quad 0 \le t \le T.$$

Find the p.d.f. of $X(t)$.

3. **Fractional Brownian Motion** Let $1/2 < H < 1$ and

$$W_H(t) \overset{\text{def}}{=} \frac{1}{\Gamma(H+1/2)} \{I_1 + I_2\},$$

where I_1 and I_2 are the Wiener integrals

$$I_1 = \int_{-\infty}^{0} \left(| t - u |^{H-1/2} - |u|^{H-1/2} \right) dW(u)$$

and

$$I_2 = \int_{0}^{t} \left(| t - u |^{H-1/2} \right) dW(u),$$

where the Wiener process has been extended to the negative values as in exercise 10. above. When $t < 0$ the notation $\int_0^t$ should be read as $-\int_t^0$. Show that this is a zero mean Gaussian process with autocorrelation function given by (10.88).

4. **Separable Autocorrelation Functions** Use the representation (10.51) to show that

$$Y(t) = \int_0^t f(u)dW(u), \quad t \geq 0,$$

has a separable autocorrelation function. Has $\mathbf{Y} = \{Y(t) \mid t \geq 0\}$ the Markov property ?

5. **Martingales and Wiener integrals** Let $\mathbf{W}$ be a Wiener process, and let

$$\mathcal{F}_t^{\mathbf{W}} = \sigma\left(W(s); 0 \leq s \leq t\right).$$

Set for $t \geq$

$$Y(t) = \int_0^t f(s)dW(s),$$

where $\int_0^\infty f^2(t)dt < \infty$. Show that the process

$$\{Y(t) \mid 0 \leq t\}$$

is a martingale w.r.t. $\left(\mathcal{F}_t^{\mathbf{W}}\right)_{t\geq 0}$.

6. **Discrete Stochastic Integrals w.r.t. the Wiener Process** Let $\mathbf{W}$ be a Wiener process, and let $0 = t_0 < t_1 < \ldots t_{i-1} < t_i < \ldots < t_n$ and

$$\mathcal{F}_{t_i}^{\mathbf{W}} = \sigma\left(W(t_0), W(t_1), \ldots W(t_i)\right).$$

Let $\int_{-\infty}^{\infty} | f(x) |^2 \, dx < \infty$ and $X(t_i) = f(W(t_{i-1}))$ and $\mathbf{X} = \{X(t_i) \mid i = 0, \ldots, n\}$. Consider as in (3.56)

$$(\mathbf{X} \star \mathbf{W})_n \overset{\text{def}}{=} \sum_{i=1}^{n} X_{t_i} \left(\triangle W\right)_{t_i}. \tag{10.91}$$

Show that this is a well defined discrete stochastic integral and that it is a martingale w.r.t. $\left(\mathcal{F}_{t_i}^{\mathbf{W}}\right)_{i=0}^{n}$.

10.7.4 A Next Step: Stochastic Calculus

1. Let $\mathbf{W} = \{W(t)|t \geq 0\}$ be a Wiener process and

$$X(t) = x_0 e^{W(t)}, t \geq 0.$$

(a) Check first using the Taylor expansion of e^x that for $h > 0$

$$X(t+h)-X(t) = X(t)\left((W(t+h)-W(t)) + \frac{1}{2!}(W(t+h)-W(t))^2 + \frac{1}{3!}(W(t+h)-W(t))^3 + \ldots\right).$$

(b) Show that

$$E\left[X(t+h) - X(t) - X(t)(W(t+h) - W(t))\right] = O(h),$$

where $O(h)$ is a function of h such that $\frac{O(h)}{h} \to M(=$ a finite limit). You will need (4.50). Thus, if one tried to express $X(t)$ formally by the seemingly obvious differential equation

$$dX(t) = X(t)dW(t),$$

the error would in the average be of order dt, which makes no sense, as truncation errors add linearly.

(c) Show that

$$E\left[(X(t+h) - X(t) - X(t)(W(t+h) - W(t)))^2\right] = O(h^2).$$

(d) Show that

$$E\left[X(t+h) - X(t) - X(t)(W(t+h) - W(t)) - \frac{1}{2!}X(t)(W(t+h) - W(t))^2\right] = o(h),$$

where $o(h)$ is a function of h such that $\frac{o(h)}{h} \to 0$ as $h \to 0$. Show that

$$E\left[\left(X(t+h) - X(t) - X(t)(W(t+h) - W(t)) - \frac{1}{2!}X(t)(W(t+h) - W(t))^2\right)^2\right] = o(h^2).$$

By the preceding, we can write in the mean square sense

$$X(t+h) - X(t) = X(t)(W(t+h) - W(t)) + \frac{X(t)}{2}(W(t+h) - W(t))^2 + Z, \tag{10.92}$$

where the remainder Z satisfies $E[Z] = o(h)$ and $E\left[Z^2\right] = o(h^2)$.

With (10.38) in mind we choose to express (10.92) ($h \to dt$) as

$$dX(t) = X(t)dW(t) + \frac{X(t)}{2}dt, \quad t > 0 \quad X(0) = x_0 \tag{10.93}$$

or

$$dX(t) = X(t)\left(dW(t) + \frac{1}{2}dt\right), \quad t > 0 \quad X(0) = x_0.$$

This is referred to as the *stochastic differential equation* satisfied by $X(t) = x_0 e^{W(t)}$. Conversely, the stochastic differential equation (10.93) is solved by $X(t) = x_0 e^{W(t)}$. Here we have derived an instance of *Itô's rule* in **stochastic calculus**.

In the basic calculus the function $x(t) = x_0 e^{a(t)}$, where $a(t)$ is a differentiable function with $a(0) = 0$, satisfies the first order homogeneous differential equation with an initial value,

$$\frac{d}{dt}x(t) = x(t)\frac{d}{dt}a(t), \quad t > 0, \quad x(0) = x_0.$$

Stochastic calculus has, by what we have found above, to differ from the ingrained rules of basic calculus by force of the properties of the Wiener process. Moreover, we are actually to understand (10.93) in terms of $X(t) = x_0 e^{W(t)}$ satisfying

$$X(t) - x_0 = \int_0^t X(s)dW(s) + \frac{1}{2}\int_0^t X(s)ds, \tag{10.94}$$

in which $\int_0^t X(s)dW(s)$ must to be properly defined, the expression in (10.91) is a first step for this. For further steps see, e.g., [13, 29, 70, 94].

Chapter 11

The Langevin Equations and the Ornstein-Uhlenbeck Process

11.1 On Langevin Equations

The French physicist Paul Langevin (1872−1946)[1] published in 1908, see [73, Appendix A], a description of the Brownian movement different from Einstein's. Langevin's approach is based on the Newtonian equations of motion. In fact one talks in physics more generally about **Langevin dynamics** as a technique for mathematical modelling of the dynamics of molecular systems. The Langevin approach applies simplified models accounting for omitted degrees of freedom by the use of stochastic differential equations. Thus this piece of work exercises a notable influence also on the theory of stochastic processes, and we shall now discuss it. We follow [39, 40], see also [17, pp. 390−392], [58, pp.262−264], [62, pp. 71−74] and [78, chapter 4.6], [94, chapter 1].

Both Einstein and Langevin obtained by their respective mathematical methods the same physical statement, namely (10.1), that the root-mean-squared displacement of a Brownian particle increases with the square root of time for large times. We derive (10.1) by the Langevin theory.

Langevin introduced a stochastic force that pushes the Brownian particle in the velocity space, while Einstein worked in the configuration space. In the terminology of this course Langevin described the Brownian particle's velocity as an Ornstein-Uhlenbeck process (to be defined below) and its position as the time integral of its velocity, whereas Einstein described it as a Wiener process. The former is a more general theory and reduces to the latter via a special limit procedure (see section 11.4 below).

Thus, $X(t)$ is the position of the large suspended particle at time $t > 0$ and is given by

$$X(t) = X(0) + \int_0^t U(s)ds, \tag{11.1}$$

where $U(s)$ is the velocity of the particle. The Newtonian second law of motion gives

$$\frac{d}{dt}U(t) = -aU(t) + \sigma F(t), \tag{11.2}$$

where $a > 0$ is a coefficient that reflects the drag force that opposes the particle's motion through the solution and $F(t)$ is a random force representing the random collisions of the particle and the surrounding molecules.

[1] http://en.wikipedia.org/wiki/Paul_Langevin
There are streets, high schools ('lycée'), town squares and residential areas in France named after Paul Langevin. He is buried in the Parisian Panthéon.

We can also write these equations, in very formal fashion, as

$$\frac{d^2}{dt^2}X(t) = -a\frac{d}{dt}X(t) + \sigma F(t).$$

The expression (11.2) is called the **Langevin equation** (for the velocity of the Brownian motion). In physical terms the parameters are

$$a = \frac{\gamma}{m}, \sigma = \frac{\sqrt{g}}{m},$$

where γ is the friction coefficient, and is by *Stokes law* given as $\gamma = 6\pi kP$, $P =$ radius of the diffusing particle, $k =$ viscosity of the fluid, m is the mass of the particle. g is measure of the strength of the force $F(t)$. Additionally, $\tau_B = \frac{m}{\gamma}$ is known as the relaxation time of the particle. We obtain by (10.2) the Einstein relation between the diffusion coefficient D and the friction coefficient γ as

$$D = \frac{RT}{N}\frac{1}{6\pi kP} = \frac{k_B T}{\gamma},$$

where $k_B = R/N$ is the **Boltzmann constant**.
Following Langevin one can argue, since the time scale for random collisions is fast, that for Brownian movement

$$F(t) \approx \overset{o}{W}(t),$$

or, the random force is white noise as described in section 10.5.4 above. In view of (10.53) we write the Langevin theory of Brownian movement as

$$X(t) = X(0) + \int_0^t U(s)ds, \tag{11.3}$$

and

$$dU(t) = -aU(t)dt + \sigma dW(t). \tag{11.4}$$

By virtue of lessons in the theory of ordinary differential equations we surmise that (11.4) be solved with

$$U(t) = e^{-at}U_0 + \sigma \int_0^t e^{-a(t-u)}dW(u). \tag{11.5}$$

The stochastic interpretation of this requires the Wiener integral from section 10.5 above.
From now on the treatment of the Langevin theory of Brownian movement will differ from the physics texts like [17, 62, 73, 78]. The quoted references do not construct an expression like (11.5). By various computations the physics texts do, however, obtain the desired results for $E\left[(X(t) - X(0))^2\right]$ that will be found using (11.5).
We shall next study the random process in (11.5), known as the **Ornstein-Uhlenbeck process** without paying attention to physics and eventually show, in the set of exercises for this chapter, that $U(t)$ in (11.5) does satisfy (11.4) in the sense that it solves

$$U(t) - U_0 = -a\int_0^t U(s)ds + \sigma\int_0^t dW(s).$$

11.2 The Ornstein-Uhlenbeck Process

We use the Wiener integral as defined in section 10.5.1 above to define a Gaussian stochastic process with variables $\widetilde{U}(t)$ with $a > 0$, $\sigma > 0$ and $t > 0$ by

$$\widetilde{U}(t) = \sigma \int_0^t e^{-a(t-u)}dW(u)$$

$$= \sigma e^{-at} \int_0^t e^{au} dW(u), \tag{11.6}$$

where $\int_0^t e^{au} dW(u)$ is evidently well defined both in the sense of definition 10.5.1 and by (10.41). Then $E\left[\widetilde{U}(t)\right] = 0$ for all t, and from (10.46)

$$E\left[\widetilde{U}(t) \cdot \widetilde{U}(s)\right] = \sigma^2 e^{-a(t+s)} \int_0^{\min(t,s)} e^{2au} du. \tag{11.7}$$

Suppose $t > s$, then we have in (11.7)

$$= \sigma^2 e^{-a(t+s)} \int_0^s e^{2au} du = \frac{\sigma^2}{2a} e^{-a(t+s)} \left[e^{2as} - 1\right]$$

$$= \frac{\sigma^2}{2a} \left(e^{-a(t-s)} - e^{-a(t+s)}\right). \tag{11.8}$$

If $t < s$

$$E\left[\widetilde{U}(t) \cdot \widetilde{U}(s)\right] = \frac{\sigma^2}{2a} \left(e^{-a(s-t)} - e^{-a(t+s)}\right). \tag{11.9}$$

Then we observe that

$$\frac{\sigma^2}{2a} e^{-a|t-s|} = \begin{cases} \frac{\sigma^2}{2a} e^{-a(t-s)} & \text{if } t > s \\ \frac{\sigma^2}{2a} e^{-a(s-t)} & \text{if } s > t, \end{cases}$$

and thus

$$E\left[\widetilde{U}(t) \cdot \widetilde{U}(s)\right] = \frac{\sigma^2}{2a} \left(e^{-a|t-s|} - e^{-a(t+s)}\right). \tag{11.10}$$

Let us next take $U_0 \in N(0, \frac{\sigma^2}{2a})$, independent of the Wiener process, and define

$$U(t) = e^{-at} U_0 + \widetilde{U}(t). \tag{11.11}$$

Then again $E\left[U(t)\right] = 0$ and

$$E\left[U(t) \cdot U(s)\right] = \frac{\sigma^2}{2a} e^{-a(t+s)} + E\left[\widetilde{U}(t) \cdot \widetilde{U}(s)\right],$$

and by (11.10)

$$E\left[U(t) \cdot U(s)\right] = \frac{\sigma^2}{2a} e^{-a|t-s|}. \tag{11.12}$$

As a summary, with $U_0 \in N(0, \frac{\sigma^2}{2a})$,

$$U(t) = e^{-at} U_0 + \sigma \int_0^t e^{-a(t-u)} dW(u) \tag{11.13}$$

defines a **Gaussian weakly stationary process**, known as the **Ornstein-Uhlenbeck process**. This implies in view of the derivation of (9.41), or (11.12), from the functional equation (9.40) that the Ornstein-Uhlenbeck process given in (11.13) is the only weakly stationary and mean square continuous Gaussian Markov process. Let us note that from (11.12) $U(t) \in N\left(0, \frac{\sigma^2}{2a}\right)$ and thus $U(t) \stackrel{d}{=} U_0$ for all $t \geq 0$. In statistical physics this is called **equilibrium** (of the dynamical system with the environment).

The following result can be directly discerned from (8.16) by means of (11.12), but we give an alternative narrative as a training exercise on computing with Wiener integrals.

Lemma 11.2.1 For $h > 0$

$$U(t+h) \mid U(t) = u \in N\left(e^{-ah} u, \frac{\sigma^2}{2a}\left(1 - e^{-2ah}\right)\right). \tag{11.14}$$

Proof: We shall first find $E[U(t+h) \mid U(t)]$ with $h>0$. In order to do this we show the intermediate result that for any $h>0$,

$$U(t+h) = e^{-ah}U(t) + \sigma e^{-ah}\int_t^{t+h} e^{-a(t-u)}dW(u). \tag{11.15}$$

We write

$$\begin{aligned}
U(t+h) &= e^{-a(t+h)}U_0 + \sigma\int_0^{t+h} e^{-a((t+h)-u)}dW(u)\\
&= e^{-a(t+h)}U_0 + \sigma e^{-ah}\int_0^{t+h} e^{-a(t-u)}dW(u)\\
&= e^{-a(t+h)}U_0 + \sigma e^{-ah}\int_0^{t} e^{-a(t-u)}dW(u) + \sigma e^{-ah}\int_t^{t+h} e^{-a(t-u)}dW(u)\\
&= e^{-ah}e^{-at}U_0 + \sigma e^{-ah}\int_0^{t} e^{-a(t-u)}dW(u) + \sigma e^{-ah}\int_t^{t+h} e^{-a(t-u)}dW(u)\\
&= e^{-ah}\left[e^{-at}U_0 + \sigma\int_0^{t} e^{-a(t-u)}dW(u)\right] + \sigma e^{-ah}\int_t^{t+h} e^{-a(t-u)}dW(u)
\end{aligned}$$

i.e., when we use (11.13) we get

$$U(t+h) = e^{-ah}U(t) + \sigma e^{-ah}\int_t^{t+h} e^{-a(t-u)}dW(u),$$

which is (11.15), as desired.
We observe now by (11.15) that

$$\begin{aligned}
E[U(t+h) \mid U(t)] &= E\left[e^{-ah}U(t) + \sigma e^{-ah}\int_t^{t+h} e^{-a(t-u)}dW(u) \mid U(t)\right]\\
&= E\left[e^{-ah}U(t) \mid U(t)\right] + E\left[\sigma e^{-ah}\int_t^{t+h} e^{-a(t-u)}dW(u) \mid U(t)\right],
\end{aligned}$$

and since we can take out what is known and since the increments of the Wiener process are independent of the sigma field $\sigma(U_0, W(s) \mid s \leq t)$ generated by the Wiener process up to time t and by the initial value,

$$= e^{-ah}U(t) + E\left[\sigma e^{-ah}\int_t^{t+h} e^{-a(t-u)}dW(u)\right] = e^{-ah}U(t),$$

by a property of the Wiener integral, see (10.42), i.e., we have

$$E[U(t+h) \mid U(t)] = e^{-ah}U(t). \tag{11.16}$$

Thus we have from (11.15) that

$$U(t+h) = E[U(t+h) \mid U(t)] + \sigma e^{-ah}\int_t^{t+h} e^{-a(t-u)}dW(u). \tag{11.17}$$

Or, $\sigma e^{-ah}\int_t^{t+h} e^{-a(t-u)}dW(u)$ is the error, when we estimate $U(t+h)$ by $E[U(t+h) \mid U(t)]$, and is independent of $E[U(t+h) \mid U(t)]$, as already established in section 3.7.3. We have that

$$E\left[\sigma e^{-ah}\int_t^{t+h} e^{-a(t-u)}dW(u)\right] = 0$$

and

$$\mathrm{Var}\left[\sigma e^{-ah}\int_t^{t+h} e^{-a(t-u)}dW(u)\right] = \sigma^2 e^{-2ah}\int_t^{t+h} e^{-2a(t-u)}du$$

$$= \frac{\sigma^2 e^{-2ah}e^{-2at}}{2a}\left[e^{2au}\right]_t^{t+h} = \frac{\sigma^2 e^{-2ah}e^{-2at}}{2a}\left[e^{2a(t+h)} - e^{2at}\right]$$

$$= \frac{\sigma^2}{2a}\left(1 - e^{-2ah}\right).$$

Thus we have

$$\frac{e^{-ah}}{\sqrt{2a}}\int_t^{t+h} e^{-a(t-u)}dW(u) \in N\left(0, \frac{\sigma^2}{2a}\left(1 - e^{-2ah}\right)\right).$$

The assertion in the lemma follows now by (11.16) and (11.17). ■

Note that (11.14) defines a Gaussian **transition p.d.f.**

$$f_{U(t+h)|U(t)=u}(v) = \frac{1}{\sqrt{2\pi\frac{\sigma^2}{2a}(1-e^{-2ah})}} e^{-\frac{1}{2}\frac{\left(v-e^{-ah}u\right)^2}{\frac{\sigma^2}{2a}(1-e^{-2ah})}}.$$

11.3 Mean-Squared Displacement: The Langevin Theory

Let us consider the general case of (11.11)

$$U(t) = e^{-at}U^* + \widetilde{U}(t), \tag{11.18}$$

where $U^* \in N(0, \sigma^*)$, independent of the Wiener process. Then from (11.3)

$$X(t) - X(0) = U^*\int_0^t e^{-as}ds + \int_0^t \widetilde{U}(s)ds$$

$$= U^*\frac{1}{a}\left[1 - e^{-at}\right] + \sigma\int_0^t e^{-as}\int_0^s e^{au}dW(u)ds.$$

As this is not a set of lecture notes in physics, we need to change the order of integration by proving (or referring to the proof of) the following Fubini-type lemma [26, p.109] or [89, p.43]:

Lemma 11.3.1 If $g(t,s)$ is a continuous function of $(t,s) \in R \times R$, then

$$\int_{t_1}^{t_2}\int_{s_1}^{s_2} g(t,s)dW(s)dt \overset{d}{=} \int_{s_1}^{s_2}\int_{t_1}^{t_2} g(t,s)dtdW(s), \tag{11.19}$$

for all finite intervals $[t_1, t_2]$ and $[s_1, s_2]$.

■

This lemma entails

$$= U^*\frac{1}{a}\left[1 - e^{-at}\right] + \sigma\int_0^t\int_u^t e^{-a(s-u)}dsdW(u)$$

$$= U^*\frac{1}{a}\left[1 - e^{-at}\right] + \frac{\sigma}{a}\int_0^t\left(1 - e^{-a(t-u)}\right)dW(u). \tag{11.20}$$

Next, the mean squared displacement is from (11.20), as U^* is independent of the Wiener process,

$$E\left[(X(t) - X(0))^2\right] = E\left[(U^*)^2\right]\frac{1}{a^2}\left[1 - e^{-at}\right]^2 + E\left[\left(\frac{\sigma}{a}\int_0^t\left(1 - e^{-a(t-u)}\right)dW(u)\right)^2\right].$$

By the properties of the Wiener integral in section 10.5.2 we get

$$E\left[\left(\frac{\sigma}{a}\int_0^t \left(1-e^{-a(t-u)}\right) dW(u)\right)^2\right] = \frac{\sigma^2}{a^2}\int_0^t \left(1-e^{-a(t-u)}\right)^2 du.$$

Some elementary algebra gives that

$$\frac{\sigma^2}{a^2}\int_0^t \left(1-e^{-a(t-u)}\right)^2 du = \frac{\sigma^2}{a^2}\left[t-\frac{2}{a}\left(1-e^{-at}\right)+\frac{1}{2a}\left(1-e^{-2at}\right)\right].$$

Then some additional elementary algebra yields

$$E\left[(X(t)-X(0))^2\right] = \left[\frac{1}{a^2}E\left[(U^*)^2\right]-\frac{\sigma^2}{2a^3}\right]\left[1-e^{-at}\right]^2+\frac{\sigma^2}{a^2}\left[t-\frac{1}{a}\left(1-e^{-at}\right)\right]. \tag{11.21}$$

Here we note that if $U^* \stackrel{d}{=} U_0 \in N\left(0,\frac{\sigma^2}{2a}\right)$, or we are in the equilibrium, then the first term in the right hand side is equal to zero. Thus, in the equilibrium,

$$E\left[(X(t)-X(0))^2\right] = \frac{\sigma^2}{a^2}\left[t-\frac{1}{a}\left(1-e^{-at}\right)\right]. \tag{11.22}$$

Then we expand

$$\frac{1}{a}\left(1-e^{-at}\right) = t-\frac{at^2}{2}+O(t^3)$$

to get for small values of t

$$\begin{aligned} E\left[(X(t)-X(0))^2\right] &= \frac{\sigma^2}{a^2}\frac{at^2}{2} \\ &= \frac{\sigma^2}{2a}t^2 = \frac{g}{2\gamma m}t^2. \end{aligned} \tag{11.23}$$

Hence, for small values of t, we get $X(t) = X(0)+U_0 t$, which is like a free particle. For very large values of t

$$E\left[(X(t)-X(0))^2\right] = \frac{\sigma^2}{a^2}t \tag{11.24}$$

Next we invoke some of the pertinent physics. Let us here recall that

$$a = \frac{\gamma}{m}, \sigma = \frac{\sqrt{g}}{m}, \gamma = 6\pi kP, k_B = R/N.$$

The mean kinetic energy of the particle is $\frac{m}{2}E\left[U^2(t)\right]$. By **equipartition of energy** (c.f., [17, chapter 19.1], [58, p. 76]) it must hold in the thermal equilibrium that

$$\frac{m}{2}E\left[U^2(t)\right] = \frac{k_BT}{2} \Leftrightarrow E\left[U^2(t)\right] = \frac{k_BT}{m}. \tag{11.25}$$

Thus, from $E\left[U^2(t)\right] = \frac{\sigma^2}{2a}$ we get

$$g = 2\gamma k_BT,$$

which is an instance of a **fluctuation - dissipation formula/theorem**, see [17, chapter 33.3] [40], [73, p. 59], connecting the fluctuation parameter g to the dissipation parameter a. When we insert this fluctuation - dissipation formula in (11.23) we get for small t

$$E\left[(X(t)-X(0))^2\right] = k_BTt^2.$$

When we insert the fluctuation - dissipation formula in the constant in (11.24), we get

$$\frac{\sigma^2}{a^2} = \frac{g}{\gamma^2} = \frac{2k_BT}{\gamma}$$

or for very large values of t

$$E\left[(X(t)-X(0))^2\right] = \frac{2k_BT}{\gamma}t = \frac{2k_BT}{6\pi kP}t = 2Dt,$$

where D is Einstein's diffusion coefficient in (10.2), i.e.

$$D = \frac{RT}{N6\pi kP},$$

now derived by Langevin's description of the Brownian movement. Moreover, we have shown the crucial statement

$$\sqrt{E\left[(X(t)-X(0))^2\right]} \propto \sqrt{t}.$$

In view of (11.25), so that $U(t) \in N\left(0, \frac{k_BT}{m}\right)$, we have thus in equilibrium that the p.d.f. of $U(t)$ is (recall (1.9))

$$f_U(x) = \sqrt{\frac{m}{2\pi k_BT}}e^{-\frac{m}{2k_BT}x^2},$$

which shows that the distribution of the velocities of the particles in the Langevin model of the Brownian movement is the **Maxwell-Boltzman velocity distribution** [17, pp. 48−49]. A selection of Boltzmann's very readable) studies in these topics is [18].

Remark 11.3.1 Langevin is said to have claimed that his approach to Brownian movement is 'infinitely simpler' than Einstein's. Of course, for us the simplicity is due to an investment in Wiener integrals, Wiener processes, convergence in mean square, multivariate Gaussianity and ultimately, sigma-additive probability measures and the prerequisite knowledge in [16]. None of the mathematical machinery hereby summoned up was available to Langevin himself, who anyhow figured out a way to deal correctly with the pertinent analysis.

■

11.4 The Langevin Equations for Thermal or Nyquist- Johnson Noise

Any circuit element that is above absolute zero will produce thermal noise. It is caused by electrons within a conductor's lattice and is an **electrical analogy of the Brownian movement**. Thermal noise was modelled by Harry T. Nyquist[2] in 1928 and experimentally measured by John E. Johnson (née Johan Erik Johansson from Göteborg (=Gothenburg)), also in 1928, and is as a consequence widely known as Nyquist-Johnson noise, see [17, chapter 33.2].

Nyquist demonstrated that thermal noise has a mean-square voltage value of $4k_BTR(BW)$. In this expression k_B is Boltzmann's constant, T is temperature in degrees Kelvin, R is resistance in ohms, and BW is bandwidth. We shall now find this formula and other results using the Langevin method following [40][3].

The problem is that of describing thermally generated electrical noise in a rigid wire loop of inductance L, resistance R and absolute temperature T. The interactions between the conducting electrons and the thermally vibrating atomic lattice of the wire give rise to temporally varying electromotive force, to be called the *thermal emf*. The **Langevin hypothesis** is that that the thermal emf can be split into a sum of two separate forces

$$-RI(t) + V(t),$$

[2] For the life and work of Harry T. Nyquist (born in Nilsby in Värmland), see the lecture by K.J. Åström in [106].

[3] A survey of thermal noise and its physics with another pedagogical method of deriving the above formula is D. Abbott, B.R. Davis, N.J. Phillips and K. Eshragian: Simple Derivation of the Thermal Noise Formula Using Window-Limited Fourier Transforms and Other Conundrums. *IEEE Transactions on Education*, 39, pp. 1−13, 1996.

where $I(t)$ is the instantaneous electrical current in the loop and $V(t)$ is the random force. $RI(t)$ is called the dissipative voltage, and $V(t)$ is called the *Johnson emf.* By Faraday's law there will be an induced emf., $-L\frac{d}{dt}I(t)$, in the loop, and since the integral of the electric potential around the loop must vanish, we get the circuit equation

$$-RI(t) + V(t) - L\frac{d}{dt}I(t) = 0,$$

or

$$\frac{d}{dt}I(t) = -aI(t) + \frac{1}{L}V(t), \tag{11.26}$$

where

$$a = \frac{R}{L}.$$

We argue (c.f, [44, pp. 392−394]) that

$$V(t) = L\sqrt{c}\overset{o}{W}(t).$$

What this means is that a simple resistor can produce white noise in any amplifier circuit. Thus we have arrived at the Ornstein-Uhlenbeck process, c.f., (11.4), or

$$dI(t) = -aI(t)dt + \sqrt{c}dW(t),$$

with the solution

$$I(t) = e^{-at}I_0 + \sqrt{c}\int_0^t e^{-a(t-u)}dW(u).$$

Mathematically seen the equation is the same as obtained for velocity in Langevin's theory of Brownian movement, but a diffusion coefficient is not of interest here. If the solution is in equilibrium, we get $E\left[I^2(t)\right] = \frac{c}{2a}$. By the equipartition theorem (for the velocity $I(t)$ in equilibrium, no quantum effects) it must hold that

$$E\left[I^2(t)\right] = \frac{k_BT}{L}, \tag{11.27}$$

and thus

$$c = \frac{2k_BTR}{L^2}.$$

Then we get that

$$E\left[V^2(t)\right] = L^2 \cdot c = 2k_BTR,$$

which yields the (celebrated) **Nyquist formula**, namely that the spectral density of the thermal emf satisfies for any $f_0 > 0$, $\triangle f$ is the bandwidth, (recall (10.54)),

$$\int_{f_0}^{f_0+\triangle f} s_{\mathbf{V}}(f)df = 4k_BTR\triangle f.$$

Here $s_{\mathbf{V}}(f)$ is the spectral density of theorem 9.3.1 restricted to the positive axis and multiplied by 2 (recall that $s_{\mathbf{V}}(f) = s_{\mathbf{V}}(-f)$), so the presence of the factor 4 seems to have no deeper physical meaning. The standard derivation of the Nyquist formula without the Langevin hypothesis is given, e.g., in [71, p. 131] or [11]. Today, as pointed out in [106], the Nyquist formula is used daily by developers of micro and nano systems and for space communication.

In view of (11.12) and the model above we get

$$E\left[I(t) \cdot I(s)\right] = \frac{\sigma^2}{2a}e^{-a|t-s|} = \frac{k_BT}{L}e^{-\frac{R}{L}|t-s|}. \tag{11.28}$$

By the table of pairs of autocorrelation functions and spectral densities in section 9.3.1 we have that

$$R_{\mathbf{I}}(h) = \frac{k_BT}{L}e^{-\frac{R}{L}|h|} \overset{\mathcal{F}}{\leftrightarrow} s_{\mathbf{I}}(f) = \frac{k_BT}{L}\frac{2\frac{R}{L}}{\left(\frac{R}{L}\right)^2 + f^2}. \tag{11.29}$$

This form of spectral density is sometimes called the *Lorentzian distribution* [58, p.257]. Then

$$s_{\mathbf{I}}(f) = \frac{2k_BT}{R}\frac{1}{1+\frac{f^2}{\left(\frac{R}{L}\right)^2}}.$$

If $L \approx 0$, then $s_{\mathbf{I}}(f)$ is almost constant up to very high frequencies and the autocorrelation function $R_{\mathbf{U}}(h)$ is almost zero except at $h = 0$. For all L

$$\int_{-\infty}^{\infty} R_{\mathbf{I}}(h)dh = \frac{2k_BT}{R},$$

and we can regard $R_{\mathbf{I}}(h)$ as behaving for small L almost like the scaled Dirac's delta $\frac{2k_BT}{R}\delta(h)$.

Or, for small L we get $RI(t) \approx V(t)$, and the practical distinction between the dissipative voltage and the Johnson emf $V(t)$ disappears. Thus, for $L \approx 0$,

$$Q(t) = \int_0^t I(s)ds \approx \sqrt{\frac{2k_BT}{R}}W(t),$$

where $Q(t)$ is something like the net charge around the loop and the factor $\sqrt{\frac{2k_BT}{R}}$ reminds us of Einstein's diffusion formula (10.2).

11.5 Exercises

1. Consider the Ornstein-Uhlenbeck process in (11.13). Find the coefficient of correlation $\rho_{U(t+h),U(t)}$ for $h > 0$. Use this to check (11.14) w.r.t. (8.16).

2. **A Linear Stochastic Differential Equation** Let us reconsider the representation in (11.13)

$$U(t) = e^{-at}U_0 + \sigma\int_0^t e^{-a(t-u)}dW(u), \tag{11.30}$$

where we now take $U_0 \in N\left(m, \sigma_o^2\right)$, independent of the Wiener process. Then the process $\{U(t) \mid t \geq 0\}$ is no longer stationary, but has continuous sample paths as before.

(a) Show that the process $\{U(t) \mid t \geq 0\}$ satisfies the equation

$$U(t) - U_0 = -a\int_0^t U(s)ds + \sigma\int_0^t dW(s). \tag{11.31}$$

A heuristic way of writing this is

$$dU(t) = -aU(t)dt + \sigma dW(t), \tag{11.32}$$

which is a linear stochastic differential equation. Because the sample paths of a Wiener process are nowhere differentiable, the expression in (11.32) is merely a (very good) notation for (11.31), c.f., (10.93) and (10.94) above. Since this is a linear differential equation, we do not need the full machinery of stochastic calculus [13, 29, 70].

Aid: Start with $-a\int_0^t U(s)ds$ so that You write from (11.30)

$$-a\int_0^t U(s)ds = -a\int_0^t e^{-as}U_0ds + -a\sigma\int_0^t\int_0^s e^{-a(s-u)}dW(u)ds.$$

Then obtain by using lemma 11.3.1 above that

$$-a\int_0^t U(s)ds = U(t) - U_0 - \sigma\int_0^t dW(u),$$

which is (11.31).

(b) Let the mean function be $\mu_{\mathbf{U}}(t) = E[U(t)]$. Show that

$$\frac{d}{dt}\mu_{\mathbf{U}}(t) = -a\mu_{\mathbf{U}}(t), \quad t > 0, \quad \mu_{\mathbf{U}}(0) = m.$$

(c) Let the variance function be $\text{Var}_{\mathbf{U}}(t) = \text{Var}[U(t)]$. Show that

$$\frac{d}{dt}\text{Var}_{\mathbf{U}}(t) = -2a\text{Var}_{\mathbf{U}}(t) + \sigma^2 \quad t > 0, \quad \text{Var}_{\mathbf{U}}(0) = \sigma_o^2.$$

(d) Let the covariance function be $\text{Cov}_{\mathbf{U}}(t,s)$. Show that

$$\frac{\partial}{\partial s}\text{Cov}_{\mathbf{U}}(t,s) = -a\text{Cov}_{\mathbf{U}}(t,s), \quad s > t.$$

(e) Show that the limiting distribution in $U(t) \overset{d}{\to} U^*$, as $t \to \infty$ is

$$U^* \in N\left(0, \frac{\sigma^2}{2a}\right). \tag{11.33}$$

Thus (11.33) shows that the velocities of the particles in the Langevin model of the Brownian movement eventually attain a Maxwell-Boltzmann distribution from any initial normal distribution. What happens with the limiting distribution, if we only assume in (11.30) $E[U_0] = m$ and $\text{Var}[U_0] = \sigma_0^2$ but no Gaussianity?

3. **A Linear Stochastic Differential Equation with Time Variable Coefficients** (From [62, p. 53]) Solve the stochastic differential equation

$$dX(t) = -at^2X(t)dt + \sigma dW(t), \tag{11.34}$$

and find $E[X(t)]$ and $\text{Var}[X(t)]$.
Aid: Apply the same method used in solving (11.32): You solve first

$$dX(t) = -at^2X(t)dt$$

and then invoke this solution formula to suggest a representation using a Wiener integral (c.f., (11.31)). Then proceed as in the exercise above.

4. Let $X(t)$ be the position of the Brownian movement particle according to the Langevin theory.

(a) Show that the expected displacement, conditionally on U^* and $X(0)$, has the mean function

$$\mu_{X(t)|U^*,X(0)}(t) = E[X(t) \mid U^*, X(0)] = X(0) + U^*\frac{1}{a}\left[1 - e^{-at}\right]. \tag{11.35}$$

(b) Show that

$$\text{Var}[X(t) - X(0)] = \frac{2k_BT}{ma^2}\left[at - \frac{3}{2} + 2e^{-at} - \frac{1}{2}e^{-2at}\right]. \tag{11.36}$$

This is more often written with the relaxation factor $\tau_B = \frac{1}{a}$.

5. Give the expression for the transition p.d.f. of $X(t)$, the position of the Brownian movement particle according to the Langevin theory.

6. **The Brownian Projectile** [4] This exercise is adapted from [73, pp. 69−73]. Suppose that a Brownian particle with $x - y$ -coordinates $X(t), Y(t)$ (= horizontal displacement, vertical height) is initialized as a projectile with

$$X(0) = Y(0) = Z(0) = 0,$$

and initial velocities

$$U_X(0) = U_{X,0} > 0, U_Y(0) = U_{Y,0}, U_Z(0) = 0.$$

Here $U_{Y,0}$ is the initial vertical velocity. The equations of motion are

$$X(t) = \int_0^t U_X(s)ds, \quad Y(t) = \int_0^t U_Y(s)ds \tag{11.37}$$

with

$$\begin{aligned} dU_X(t) &= -aU_X(t)dt + \beta dW_X(t) \\ dU_Y(t) &= -a\left(U_Y(t) + \frac{g}{a}\right)dt + \beta dW_Y(t). \end{aligned} \tag{11.38}$$

where g is the acceleration due to gravity, and $\mathbf{W}_X$ and $\mathbf{W}_Y$ are *two independent (standard) Wiener processes* associated with fluctuations in the respective directions. As expounded in one of the earlier exercises in this section, the stochastic differential equations in (11.38) are not taken literally, but are to be understood by means of the Wiener integral.

(a) We set for economy of expression

$$\sigma^2(t) \overset{\text{def}}{=} \frac{\beta^2}{a^3}\left(ta - 2\left(1 - e^{-at}\right) + \frac{1}{2}\left(1 - e^{-a2t}\right)\right).$$

Show that

$$X(t) \in N\left(\frac{U_{X,0}}{a}\left(1 - e^{-at}\right), \sigma^2(t)\right)$$

and

$$Y(t) \in N\left(\frac{U_{Y,0}}{a}\left(1 - e^{-at}\right) - \frac{g}{a^2}\left(at + e^{-at} - 1\right), \sigma^2(t)\right).$$

(b) Suppose $at << 1$, or the time is close to the initial value. Show that you get (leading terms)

$$E\left[X(t)\right] = U_{X,0}t,$$

and

$$E\left[Y(t)\right] = U_{Y,0}t - \frac{gt^2}{2},$$

and

$$\text{Var}\left[Y(t)\right] = \text{Var}\left[X(t)\right] = \frac{\beta^2 t^3}{3}.$$

Observe that $E\left[Y(t)\right]$ is the expected height of the Brownian projectile at time t. Thus, close to the start, the Brownian projectile preserves the effect of the initial conditions and reproduces the deterministic projectile motion familiar from introduction to physics courses.

[4] In Swedish this might be 'brownsk kastparabel', the Swedish word 'projektilbana' corresponds to a different physical setting.

(c) Suppose $at >> 1$, or that the time is late. Show that you get (leading terms)

$$E[X(t)] = 0,$$

and

$$E[Y(t)] = -\frac{gt^2}{2},$$

and

$$\mathrm{Var}[Y(t)] = \mathrm{Var}[X(t)] = \frac{\beta^2 t}{a^2}.$$

This recapitulates the statistical behaviour of two standard Wiener processes with superimposed constant **drift** downwards in the height coordinate, or for large t

$$dX(t) = \frac{\beta}{a} dW_X(t), \quad dY(t) = -gtdt + \frac{\beta}{a} dW_Y(t).$$

7. **Stochastic Damped Harmonic Oscillator** The harmonic oscillator is a multipurpose workhorse of theoretical physics. We shall now give a description of the damped harmonic oscillator using the Langevin dynamics. This discussion is adapted from [73, pp. 75−80], but is originally due to Subrahmanyan Chandrasekhar in 1943 [5].

A massive object of mass $= m$ is attached to a spring and submerged in a viscous fluid. If it is set into motion, the object will oscillate back and forth with an amplitude that decays in time. The collisions that cause the oscillations to decay will also cause the object to oscillate randomly. γ is the friction coefficient, not the Euler gamma. The Chandrasekhar equations for the random motion are

$$X(t) = \int_0^t U(s)ds, \tag{11.39}$$

where

$$\frac{d}{dt} U(t) = -\frac{\gamma}{m} U(t) - \frac{\gamma}{m} X(t) + \frac{\sqrt{2k_B T}}{m} dW(t). \tag{11.40}$$

(a) Let $\mu_U(t) = E[U(t)]$ and $\mu_X(t) = E[X(t)]$. Show that

$$\frac{d}{dt} \mu_U(t) = -\frac{\gamma}{m} \mu_U(t) - \frac{\gamma}{m} \mu_X(t),$$

and

$$\frac{d}{dt} \mu_X(t) = \mu_U(t).$$

Solve these equations !

(b) Show that the autocorrelation function of $\mathbf{X}$ is

$$R_{\mathbf{X}}(h) = \frac{k_B T}{m\omega_o^2} e^{-\frac{\gamma}{2m}h} \left[\cos(\omega_1 h) + \frac{\gamma}{2m\omega_1^2} \sin(\omega_1 h) \right],$$

where $\omega_o = \frac{\gamma}{m}$ and $\omega_1 = \sqrt{\omega_o^2 - \frac{\gamma^2}{4m^2}}$. We have written this expression so that we emphasize the case $2\omega_o >> \frac{\gamma}{m}$ (lightly damped oscillator), so that ω_1 is defined.

Aid: This is tricky ! Some help can be found in [17, p. 440. example 33.5]. Find first the suitable Fourier transforms.

What sort of formula is obtained by means of $R_{\mathbf{X}}(0)$?

[5] There are several well known Indian born scientists with the family name Chandrasekhar. Subrahmanyan C. was an applied mathematician, who worked with astrophysics, and became a laureate of the Nobel Prize in Physics in 1983 http://nobelprize.org/nobel_prizes/physics/laureates/1983/chandrasekhar.html

Chapter 12

The Poisson Process

12.1 Introduction

The Poisson process is an important example of a **point process**. In probability theory one talks about a point process, when any sample path of the process consists of a set of separate points. For example, the sample paths are in continuous time, and assume values in integers. The Poisson process will be connected to the Poisson distribution in the same way as the Wiener process is connected to the normal distribution: namely as the distribution of the independent increments. We shall start with the definition and basic properties of the Poisson process and then proceed with (engineering) models, where the Poisson process has been incorporated and found useful. Poisson processes are applied in an impressing variety of topics. The applications range from queuing, telecommunications, and computer networks to insurance and astronomy.

12.2 Definition and First Properties

12.2.1 The Counter Process and the Poisson Process

Let $N(t) =$ number of occurrences of some event in in $(0, t]$, i.e., $N(t)$ has the nonnegative integers as values. A physical phenomen of this kind is often called a **counter process**. The following definition gives a probabilistic model for the counter process $\{N(t) \mid t \geq 0\}$, which is obviously a point process.

Definition 12.2.1 $\mathbf{N} = \{N(t) \mid t \geq 0\}$ is a Poisson process with parameter $\lambda > 0$, if

(1) $N(0) = 0$.

(2) The increments $N(t_k) - N(t_{k-1})$ are independent stochastic variables for non-overlapping intervals, i.e., $1 \leq k \leq n$, $0 \leq t_0 \leq t_1 \leq t_2 \leq \ldots \leq t_{n-1} \leq t_n$ and all n.

(3) $N(t) - N(s) \in \text{Po}(\lambda(t-s)), \quad 0 \leq s < t.$

■

There are alternative equivalent definitions, but this text will be restricted to the one stated above. Following our line of approach to stochastic processes we shall first find the mean function, the autocorrelation function and the autocovariance function of $\mathbf{N}$, i.e., some of the the second order properties.

12.2.2 Second Order Properties of the Poisson Process

The **mean function** of $\mathbf{N}$ is by definition (9.3)

$$\mu_{\mathbf{N}}(t) = E\left[N(t)\right] = \lambda t, \tag{12.1}$$

since $N(t) = N(t) - N(0) \in \text{Po}(\lambda t)$ by (3). The **autocorrelation function** is by definition (9.4)

$$R_{\mathbf{N}}(t,s) = E\left[N(t)N(s)\right].$$

We assume first that $t > s$. Then

$$\begin{aligned} R_{\mathbf{N}}(t,s) &= E\left[\left(N(t) - N(s) + N(s)\right)N(s)\right] = \\ &= E\left[\left(N(t) - N(s)\right)N(s)\right] + E\left[N^2(s)\right] \\ &= E\left[\left(N(t) - N(s)\right)\right]E\left[N(s)\right] + E\left[N^2(s)\right], \end{aligned}$$

where we applied the premise (2), i.e., $N(t) - N(s)$ is independent of $N(s) = N(s) - N(0)$, and now

$$= \underbrace{E\left[\left(N(t) - N(s)\right)\right]}_{=\lambda(t-s)}\underbrace{E\left[N(s)\right]}_{=\lambda s} + \underbrace{E\left[N^2(s)\right]}_{=\lambda s + (\lambda s)^2}$$

in view of (3), and since $E\left[N^2(s)\right] = \text{Var}\left[N(s)\right] + E^2\left[N(s)\right]$, and by the properties of the Poisson distribution. Then we get

$$= \lambda(t-s)\cdot\lambda s + \lambda s + (\lambda s)^2 = \lambda^2 t\cdot s + \lambda s.$$

If we repeat this for $s > t$, then we find

$$R_{\mathbf{N}}(t,s) = \lambda^2 t\cdot s + \lambda t.$$

The preceding expressions can be summarized as

$$R_{\mathbf{N}}(t,s) = \lambda^2 t\cdot s + \lambda\min(t,s). \tag{12.2}$$

In view of (9.5), (12.1) and (12.2) we get the autocovariance function of $\mathbf{N}$ as

$$\text{Cov}_{\mathbf{N}}(t,s) = R_{\mathbf{N}}(t,s) - \mu_{\mathbf{N}}(t)\mu_{\mathbf{N}}(s) = \lambda\min(t,s).$$

Hence the autocovariance function of $\mathbf{N}$ is equal the autocovariance function in (10.18). Hence it must be clear that the autocovariance function does not tell much about a stochastic process.

12.2.3 Occurrence/Arrival Times and Occupation/Interarrival Times of a Poisson Process

Let us define T_k = the time of occurrence/arrival of the kth event. $T_0 = 0$. The we have that

$$\tau_k = T_k - T_{k-1},$$

is the kth occupation/interarrival time. In words, τ_k is the random time the process occupies or visits the value $N(t) = k-1$, or the random time between the kth and the $k-1$th arrival. In view of these definitions we can (after a moment's reflection) write

$$N(t) = \max\{k | t \geq T_k\}. \tag{12.3}$$

In the same way we observe that

$$\{N(t) = 0\} = \{T_1 > t\}. \tag{12.4}$$

The contents of the following theorem are important for understanding the Poisson processes.

Theorem 12.2.1 1. $\tau_1, \tau_2 \ldots, \tau_k \ldots$ are independent and identically distributed, $\tau_i \in \text{Exp}\left(\frac{1}{\lambda}\right)$.

2. $T_k \in \Gamma\left(k, \frac{1}{\lambda}\right)$, $k = 1, 2, \ldots$.

■

The full proof will not be given. We shall check the assertion for τ_1 and then for (τ_1, τ_2). We start with deriving the distribution of $\tau_1 = T_1$. Clearly $\tau_1 \geq 0$. Then (12.4) yields $\mathbf{P}(T_1 > t) = \mathbf{P}(N(t) = 0) = e^{-\lambda t}$, since $N(t) \in \text{Po}(\lambda)$. Then for $t > 0$

$$1 - F_{T_1}(t) = \mathbf{P}(T_1 > t) = e^{-\lambda t},$$

i.e.,

$$F_{T_1}(t) = \begin{cases} 1 - e^{-\lambda t} & t > 0 \\ 0 & t \leq 0. \end{cases}$$

Hence $\tau_1 = T_1 \in \text{Exp}(1/\lambda)$.

Next we consider (τ_1, τ_2) and find first $F_{T_1,T_2}(s,t)$. We assume $t > s$. It will turn out to be useful to consider the following expression.

$$\mathbf{P}(T_1 \leq s, T_2 > t) = \mathbf{P}(N(s) \geq 1, N(t) < 2) = \mathbf{P}(N(s) = 1, N(t) = 1)$$

$$= \mathbf{P}(N(s) = 1, N(t) - N(s) = 0) = \mathbf{P}(N(s) = 1)\, \mathbf{P}(N(t) - N(s) = 0)$$

$$= e^{-\lambda s} \lambda s e^{-\lambda(t-s)} = s\lambda e^{-\lambda t}. \tag{12.5}$$

Then we have by the law of total probability (3.35) that

$$\mathbf{P}(T_1 \leq s, T_2 > t) + \mathbf{P}(T_1 \leq s, T_2 \leq t) = \mathbf{P}(T_1 \leq s).$$

Thus, by (12.5),

$$F_{T_1,T_2}(s,t) = \mathbf{P}(T_1 \leq s, T_2 \leq t) = \mathbf{P}(T_1 \leq s) - \mathbf{P}(T_1 \leq s, T_2 > t)$$

$$= \mathbf{P}(T_1 \leq s) - s\lambda e^{-\lambda t}.$$

Therefore

$$\frac{\partial}{\partial s} F_{T_1,T_2}(s,t) = f_{T_1}(s) - \lambda e^{-\lambda t}.$$

By a second partial differentiation we have established the joint p.d.f. of (T_1, T_2) as

$$f_{T_1,T_2}(s,t) = \frac{\partial^2}{\partial t \partial s} F_{T_1,T_2}(s,t) = \lambda^2 e^{-\lambda t}. \tag{12.6}$$

Now, we consider the change of variables $(T_1, T_2) \mapsto (\tau_1, \tau_2) = (u, v)$ by

$$u = \tau_1 = T_1 = g_1(T_1, T_2),$$

$$v = \tau_2 = T_2 - T_1 = g_2(T_1, T_2)$$

and its inverse $(\tau_1, \tau_2) \mapsto (T_1, T_2)$,

$$T_1 = u = h_1(u, v)$$

$$T_2 = T_1 + \tau_2 = u + v = h_2(u, v).$$

Then by the formula for change of variable in (2.69) applied to (12.6)

$$f_{\tau_1,\tau_2}(u,v) = f_{T_1,T_2}(u, v+u) \underbrace{|J|}_{=1}$$

$$= \lambda^2 e^{-\lambda(v+u)}$$

$$= \underbrace{\lambda e^{-\lambda u}}_{\text{p.d.f. of Exp}(1/\lambda)} \cdot \underbrace{\lambda e^{-\lambda v}}_{\text{p.d.f. of Exp}(1/\lambda)}.$$

Thus $f_{\tau_1,\tau_2}(u,v) = f_{T_1}(u) \cdot f_{\tau_2}(v)$ for all pairs (u,v) and this ascertains that τ_1 and τ_2 are independent. Thus τ_1 and τ_2 are independent. By example 4.4.9, the distribution of the sum of two I.I.D. r.v.'s $\in$ Exp$(1/\lambda)$ is $T_2 \in \Gamma\left(2, \frac{1}{\lambda}\right)$. We should now continue by considering in an analogous manner T_1, T_2, T_3 to derive $f_{\tau_1,\tau_2,\tau_3}(u,v,w)$ and so on, but we halt at this point. The Gamma distributions in this theorem are all **Erlang**, c.f., example 2.2.10.

12.2.4 Two Auxiliary Probability Formulas

In this section we recapitulate two generally valid formulas for Poisson distribution.

Lemma 12.2.2 For $t \geq 0$ we have

$$P\left(N(t) = \text{ even}\right) = \frac{1}{2}\left(1 + e^{-2\lambda t}\right). \tag{12.7}$$

$$P\left(N(t) = \text{ odd }\right) = \frac{1}{2}\left(1 - e^{-2\lambda t}\right). \tag{12.8}$$

Proof: We regard 0 as an even number, and get

$$P\left(N(t) = \text{ even}\right) = P\left(N(t) = 0 \text{ or } N(t) = 2 \text{ or } N(t) = 4 \ldots\right)$$

$$= \sum_{k=0}^{\infty} P\left(N(t) = 2k\right) = \sum_{k=0}^{\infty} e^{-\lambda t} \frac{(t\lambda)^{2k}}{2k!} = e^{-\lambda t} \sum_{k=0}^{\infty} \frac{(t\lambda)^{2k}}{2k!}.$$

We know the two series expansions:

$$e^{\lambda t} = \sum_{k=0}^{\infty} \frac{(t\lambda)^k}{k!}, e^{-\lambda t} = \sum_{k=0}^{\infty} (-1)^k \frac{(t\lambda)^k}{k!}.$$

The trick here is to add these two series, which yields

$$e^{\lambda t} + e^{-\lambda t} = 2 \sum_{k=0}^{\infty} \frac{(t\lambda)^{2k}}{2k!},$$

or

$$\sum_{k=0}^{\infty} \frac{(t\lambda)^{2k}}{2k!} = \frac{1}{2}\left(e^{\lambda t} + e^{-\lambda t}\right).$$

The desired probability becomes

$$P\left(N(t) = \text{ even}\right) = e^{-\lambda t} \sum_{k=0}^{\infty} \frac{(t\lambda)^{2k}}{2k!}$$

$$= e^{-\lambda t} \cdot \frac{1}{2}\left(e^{\lambda t} + e^{-\lambda t}\right) = \frac{1}{2}\left(1 + e^{-2\lambda t}\right),$$

which proves (12.7). Since $N(0) = 0$, the formulas above hold even for $t = 0$. Next we observe

$$1 = P\left(N(t) = \text{ even}\right) + P\left(N(t) = \text{ odd }\right),$$

and from (12.7) for $t \geq 0$

$$P\left(N(t) = \text{ odd }\right) = 1 - \frac{1}{2}\left(1 + e^{-2\lambda t}\right) = \frac{1}{2}\left(1 - e^{-2\lambda t}\right).$$

■

By the above

$$P\left(N(t) = \text{ even}\right) = e^{-\lambda t} \cdot \frac{1}{2}\left(e^{\lambda t} + e^{-\lambda t}\right) = e^{-\lambda t} \cdot \cosh(\lambda t),$$

and

$$P\left(N(t) = \text{ odd}\right) = e^{-\lambda t} \cdot \frac{1}{2}\left(e^{\lambda t} - e^{-\lambda t}\right) = e^{-\lambda t} \cdot \sinh(\lambda t).$$

12.3 Restarting the Poisson Process

Theorem 12.3.1 $\mathbf{N} = \{N(t) \mid t \geq 0\}$ is a Poisson process with parameter $\lambda > 0$. Then the process $\mathbf{N}_s = \{N(t+s) - N(s) \mid t \geq 0\}$ is a Poisson process.

Proof: The process $\mathbf{N}_s$ has clearly nonnegative integers as values and has nondecreasing sample paths, so we are talking about a counter process. We set

$$N_s(t) \overset{\text{def}}{=} N(t+s) - N(s),$$

and get $N_s(0) = 0$, so the premise (1) in the definition is there. Next we show (3). For $t > u > 0$ we get

$$N_s(t) - N_s(u) = N(t+s) - N(u+s) \in \text{Po}(\lambda(t-u)),$$

which proves the claim vis-á-vis (3).

It remains to show that the increments are independent. Let $0 \leq t_1 \leq t_2 \leq \ldots \leq t_{n-1} \leq t_n$. Then the increments of $\mathbf{N}_s$ are $N_s(t_i) - N_s(t_{i-1}) = N(t_i+s) - N(t_i+s)$. But this corresponds to watching the increments of $\mathbf{N}$ over the intervals determined by the points $s \leq t_1 + s \leq t_2 + s \leq \ldots \leq t_{n-1} + s \leq t_n + s$. But the latter increments are independent, as $\mathbf{N}$ is a Poisson process. ■

What does restarting mean? Observing $N(t+s) - N(s)$ corresponds to moving the origin to $(s, N(s))$ in the coordinate plane and running the process from there. The new process is again a Poisson process. The next theorem shows that we can restart the process as new Poisson process from occurrence/arrival times, too.

Theorem 12.3.2 $\mathbf{N} = \{N(t) \mid t \geq 0\}$ is a Poisson process with parameter $\lambda > 0$, and T_k is the kth occurrence/arrival time. Then the process $\mathbf{N}_k = \{N(t+T_k) - N(T_k) \mid t \geq 0\}$ is a Poisson process.

Proof: The process $\mathbf{N}_{T_k}$ has clearly nonnegative integers as values and has nondecreasing sample paths, and therefore we are dealing with a counter process. We set for $t \geq 0$

$$N_k(t) \overset{\text{def}}{=} N(t+T_k) - N(T_k).$$

We get immediately $N_{T_k}(0) = 0$, so condition (1) in the definition is fulfilled.

We prove next the condition (3), i.e., that the increments of $\mathbf{N}_k$ have the Poisson distribution. If $t > s$, then

$$N_k(t) - N_k(s) \in \text{Po}(\lambda(t-s)). \tag{12.9}$$

For any $l = 0, 1, \ldots,$

$$\mathbf{P}\left(N_k(t) - N_k(s) = l\right) = \mathbf{P}\left(N(t+T_k) - N(s+T_k) = l\right)$$

$$= \int_0^\infty \mathbf{P}\left(N(t+T_k) - N(s+T_k) = l, T_k = u\right) du = \int_0^\infty \mathbf{P}\left(N(t+T_k) - N(s+T_k) = l \mid T_k = u\right) f_{T_k}(u) du$$

$$= \int_0^\infty \mathbf{P}\left(N(t+u) - N(s+u) = l \mid T_k = u\right) f_{T_k}(u) du$$

But we have that the events $\{N(t+u)-N(s+u)=l\}$ and $\{T_k=u\}$ are independent for any u. This follows, since $T_k=u$ is an event measurable w.r.t. the sigma field generated by $N(v)$, $0<v\leq s$ and the $N(t+u)-N(s+u)=l$ is an event measurable w.r.t. the sigma field generated by $N(v)$, $u+s<v\leq t+u$, and the Poisson process has independent increments. Hence

$$= \int_0^\infty \mathbf{P}\left(N(t+u)-N(s+u)=l\right) f_{T_k}(u)du$$

$$= \int_0^\infty \mathbf{P}\left(\underbrace{N(t+u)-N(s+u)}_{\in \mathrm{Po}(\lambda(t-s))}=l\right) f_{T_k}(u)du$$

$$= \int_0^\infty e^{-\lambda(t-s)}\frac{(\lambda(t-s))^l}{l!} f_{T_k}(u)du = e^{-\lambda(t-s)}\frac{(\lambda(t-s))^l}{l!}\underbrace{\int_0^\infty f_{T_k}(u)du}_{=1} = e^{-\lambda(t-s)}\frac{(\lambda(t-s))^l}{l!},$$

as $f_{T_k}(u)$ is a p.d.f.! Hence we have shown (12.9). Finally we should show that the increments $N_k(t_i)-N_k(t_{i-1})$ $=N(t_i+T_k)-N(t_{i-1}+T_k)$ are independent over nonoverlapping intervals. But if we use an analogous consideration as in the corresponding step of the proof 12.3.1 , we are watching the increments over the nonoverlapping intervals with the endpoints $T_k\leq t_1+T_k\leq t_2+T_k\leq\ldots\leq t_{n-1}+T_k\leq t_n+T_k$. In the proof above we have found that

$$N(t+T_k)-N(s+T_k)\mid T_k=u\in \mathrm{Po}(\lambda(t-s)). \tag{12.10}$$

Hence, the probabilistic properties of the increments of $\mathbf{N}_{T_k}$ are independent of $\{T_k=u\}$ for any $u\geq 0$, and (2) follows for $\mathbf{N}_{T_k}$, of course, by property (2) of the restarted Poisson process $\mathbf{N}$. ■

12.4 Filtered Shot Noise

We refer here to [33, pp. 317−319] or [71, pp. 133−136]. Shot noise[1] is a type of electronic noise, which originates from the random movements of the carriers of the electric charge. The term also used to photon counting in optical devices. A good mathematical model for this has been found to be valid both theoretically and experimentally, if we regard the number of electrons emitted from a cathode as a Poisson process.

Let T_k, $k=1,2,\ldots$, be the times of occurrence/arrival of the kth event, respectively, in a Poisson process with intensity $\lambda>0$. Let $h(t)$ be function such that $h(t)=0$ for $t<0$. Then the stochastic process defined by the random variables

$$Z(t)=\sum_{k=1}^{\infty} h\left(t-T_k\right),\quad t\geq 0, \tag{12.11}$$

is a model called a **filtered shot noise**. We shall once more find the mean function of the process thus defined. In addition we shall derive the m.g.f. of $Z(t)$.

> Borrowing from control engineering and signal processing we should/could refer to $h(t)$ with $h(t)=0$ for $t<0$ in (12.11) as a **causal impulse response of a linear filter**. With
>
> $$U(t)=\begin{cases}1 & t\geq 0,\\ 0 & t<0\end{cases}$$
>
> an example is
>
> $$h(t)=e^{-t}U(t).$$

[1] Shot noise is **hagelbrus** in Swedish.

Since $h(t) = 0$ for $t < 0$, the sum in (12.11) contains for any t only a finite number of terms, since there is only a finite number of arrivals in a Poisson process in a finite interval. The word 'causal' means thus simply that $Z(t)$ does not for a given t depend on the arrivals of events in the future beyond t, i.e., on $T_j > t$, recall (12.3).

We start with the mean function. Since the sum in (12.11) consists for any t only of a finite number of terms, there is no mathematical difficulty in computing as follows.

$$E\left[Z(t)\right] = \sum_{k=1}^{\infty} E\left[h\left(t - T_k\right)\right]. \tag{12.12}$$

The individual term in the sum is by the law of the unconscious statistician (2.4)

$$E\left[h\left(t - T_k\right)\right] = \int_0^{\infty} h\left(t - x\right) f_{T_k}(x) dx,$$

where we know by theorem 12.2.1 that $T_k \in \Gamma\left(k, \frac{1}{\lambda}\right)$ (Erlang distribution, example 2.2.10). Thus

$$\int_0^{\infty} h\left(t - x\right) f_{T_k}(x) dx = \int_0^{\infty} h\left(t - x\right) \frac{\lambda^k x^{k-1}}{(k-1)!} e^{-\lambda x} dx,$$

$$= \lambda^k \int_0^{\infty} h\left(t - x\right) \frac{x^{k-1}}{(k-1)!} e^{-\lambda x} dx.$$

When we insert this in (12.12) we get

$$E\left[Z(t)\right] = \sum_{k=1}^{\infty} \lambda^k \int_0^{\infty} h\left(t - x\right) \frac{x^{k-1}}{(k-1)!} e^{-\lambda x} dx$$

$$= \lambda \int_0^{\infty} h\left(t - x\right) \sum_{k=1}^{\infty} \frac{\lambda^{k-1} x^{k-1}}{(k-1)!} e^{-\lambda x} dx = \lambda \int_0^{\infty} h\left(t - x\right) \underbrace{\sum_{k=0}^{\infty} \frac{\lambda^k x^k}{k!}}_{=e^{\lambda x}} e^{-\lambda x} dx$$

$$= \lambda \int_0^{\infty} h\left(t - x\right) dx.$$

Thus we have obtained **Campbell's Formula**, i.e.,

$$E\left[Z(t)\right] = \lambda \int_0^{\infty} h\left(t - x\right) dx = \lambda \int_0^{t} h\left(t - x\right) dx. \tag{12.13}$$

In the final step above we exploited the causality of the impulse response. Next we shall derive the m.g.f. of filtered shot noise. We attack the problem immediately by double expectation to get

$$\psi_{Z(t)}(s) = E\left[e^{sZ(t)}\right] = E\left[E\left[e^{sZ(t)} \mid N(t)\right]\right] =$$

and continue the assault by the law of the unconscious statistician (2.4)

$$= \sum_{l=0}^{\infty} E\left[e^{sZ(t)} \mid N(t) = l\right] \mathbf{P}\left(N(t) = l\right)$$

$$= \sum_{l=0}^{\infty} E\left[e^{s \sum_{k=1}^{l} h(t-T_k)} \mid N(t) = l\right] \mathbf{P}\left(N(t) = l\right),$$

since in (12.3), $N(t) = \max\{k | t \geq T_k\}$,

$$= \sum_{l=0}^{\infty} E\left[E\left[e^{s\sum_{k=1}^{l} h(t-T_k)} \mid N(t) = l\right]\right] \mathbf{P}\left(N(t) = l\right). \tag{12.14}$$

Here we have in view of the result in (12.30) in one of the exercises that

$$E\left[e^{s\sum_{k=1}^{l} h(t-T_k)} \mid N(t) = l\right] = \prod_{k=1}^{l} \int_0^t e^{sh(t-x)} \frac{1}{t} dx = \left(\int_0^t e^{sh(t-x)} \frac{1}{t} dx\right)^l.$$

When we insert this in (12.14) we get

$$\psi_{Z(t)}(s) = \sum_{l=0}^{\infty} \left(\int_0^t e^{sh(t-x)} \frac{1}{t} dx\right)^l e^{-\lambda t} \frac{(t\lambda)^l}{l!}$$

$$= \sum_{l=0}^{\infty} e^{-\lambda t} \frac{\left(\lambda \int_0^t e^{sh(t-x)} dx\right)^l}{l!} = e^{-\lambda t} e^{\lambda \int_0^t e^{sh(t-x)} dx} = e^{\lambda \int_0^t \left(e^{sh(t-x)} - 1\right) dx}.$$

In summary, the **m.g.f. of the filtered shot noise** $Z(t)$ is

$$\psi_{Z(t)}(s) = e^{\lambda \int_0^t \left(e^{sh(t-x)} - 1\right) dx}. \tag{12.15}$$

12.5 Random Telegraph Signal

12.5.1 Background: Popcorn Noise

A **random telegraph signal** (RTS) is a name used for a physical process, whose values at time t is either one of only two possible values (say, UP and DOWN). Many processes including chemical reactions, cell membrane ion channels, and electronic noise generate such signals. By this we referred to real life phenomena that we try to model using stochastic processes, that is, in the final analysis using sigma-additive probability measures.

The material in this chapter deals with a probability model of RTS derived from the Poisson process. This probabilistic model of RTS can be found as an exercise/example in several of the pertinent course texts [56, pp. 392-394], [74, pp. 348−349], [85, pp. 211−212] or [89, pp. 354−356]. These references are occupied by a drill in probability calculus.

In semiconductors RTS is also modelling what is known as **popcorn noise**. It consists of sudden step-like transitions between two voltage or current levels at random and unpredictable times. Each switch in offset voltage or current may last from several milliseconds to seconds, and when connected to an audio speaker, it sounds, as is claimed, like popcorn popping.

Let $Y \in \mathrm{Be}(1/2)$ and $\mathbf{N} = \{N(t) | t \geq 0\}$ be a Poisson process with intensity $\lambda > 0$. Y is independent of $\mathbf{N}$.

Definition 12.5.1 Random Telegraph Signal

$$X(t) = (-1)^{Y + N(t)}, \quad t \geq 0. \tag{12.16}$$

■

Then $\mathbf{X} = \{X(t) \mid t \geq 0\}$ is a process in continuous time flipping between UP $(+1)$ and DOWN (-1) with a random initial value and with the Poisson process generating the flips in time . The figure 12.1 shows in the upper part a sample path of UPs and DOWNs and in the lower part the corresponding sample path of Poisson process. In the figure we have $Y = 0$, since $X(0) = 1$.

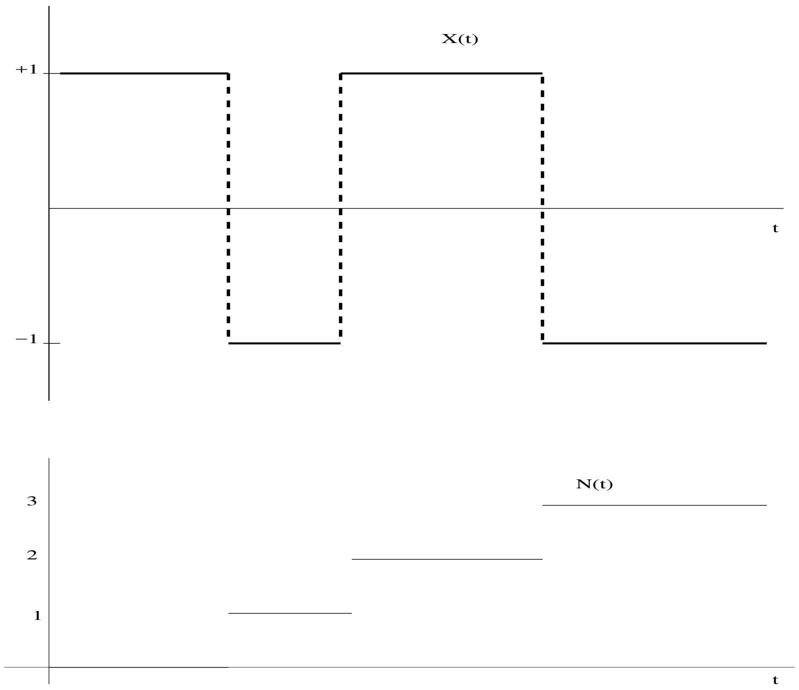

Figure 12.1: A sample path of the random telegraph signal and the corresponding sample path of a Poisson process

Because **N** has the nonnegative integers as values, the definition 12.5.1 implies that:

$$X(t) = \begin{cases} (-1)^Y \cdot 1 & N(t) \text{ is even} \\ (-1)^Y \cdot (-1) & N(t) \text{ is odd.} \end{cases} \tag{12.17}$$

Hence it seems that lemma 12.2.2 must be instrumental here.

12.5.2 The Marginal Distribution and the Mean Function of RTS Modelled by Poisson Flips

Lemma 12.5.1

$$P\left(X(t) = +1\right) = P\left(X(t) = -1\right) = \frac{1}{2}, \quad t \geq 0. \tag{12.18}$$

Proof The law of total probability (3.35) gives

$$P\left(X(t) = +1\right) = P\left(X(t) = +1 \mid Y = 0\right) P\left(Y = 0\right) + P\left(X(t) = +1 \mid Y = 1\right) P\left(Y = 1\right). \tag{12.19}$$

We have that

$$P\left(X(t) = +1 \mid Y = 0\right) = P\left((-1)^{0+N(t)} = +1 \mid Y = 0\right) = P\left((-1)^{N(t)} = +1 \mid Y = 0\right)$$

$$= P\left((-1)^{N(t)} = +1\right) = P\left(N(t) = \text{ even}\right) = \frac{1}{2}\left(1 + e^{-2\lambda t}\right),$$

since Y is independent of **N** and by (12.7). In the same way we get

$$P\left(X(t) = +1 \mid Y = 1\right) = P\left((-1) \cdot (-1)^{N(t)} = +1\right)$$

$$= P\left(N(t) = \text{ odd }\right) = \frac{1}{2}\left(1 - e^{-2\lambda t}\right)$$

from (12.8). Since $Y \in \mathrm{Be}(1/2)$ by construction

$$P\left(Y = 0\right) = P\left(Y = 1\right) = \frac{1}{2}.$$

Insertion of the preceding results in (12.19) gives

$$P\left(X(t) = +1\right) = \frac{1}{2}\left[\frac{1}{2}\left(1 + e^{-2\lambda t}\right) + \frac{1}{2}\left(1 - e^{-2\lambda t}\right)\right] = \frac{1}{2},$$

as was claimed. ■

The mean function is by definition

$$\begin{aligned}\mu_{\mathbf{X}}(t) = E\left[X(t)\right] &= (+1) \cdot P\left(X(t) = +1\right) + (-1) \cdot P\left(X(t) = -1\right) \\ &= \frac{1}{2} - \frac{1}{2} = 0\end{aligned}$$

by (12.18).

Lemma 12.5.2

$$\mu_{\mathbf{X}}(t) = 0, \quad t \geq 0. \tag{12.20}$$

■

12.5.3 The RTS as a Weakly Stationary Process

The autocorrelation function of RTS

We shall compute

$$R_{\mathbf{X}}(t, s) = E\left[X(t)X(s)\right].$$

With double expectation

$$\begin{aligned}E\left[X(t)X(s)\right] = E\left[E\left[X(t)X(s) \mid Y\right]\right] = \\ = E\left[X(t)X(s) \mid Y = 0\right] P\left(Y = 0\right) + E\left[X(t)X(s) \mid Y = 1\right] P\left(Y = 1\right).\end{aligned} \tag{12.21}$$

We work out case by case the conditional expectations in the right hand side. By (12.16)

$$\begin{aligned}E\left[X(t)X(s) \mid Y = 0\right] &= E\left[(-1)^{0+N(t)}(-1)^{0+N(s)} \mid Y = 0\right] \\ &= E\left[(-1)^{N(t)+N(s)}\right],\end{aligned}$$

since $\mathbf{N}$ is independent of Y. Hereafter we must distinguish between two different cases, i) $s > t$ och ii) $s < t$.

i) Let $s > t$, and write

$$N(t) + N(s) = N(s) - N(t) + 2N(t).$$

Then $N(s) - N(t)$ is independent of $2N(t)$ and $2N(t)$ is an even integer, so that

$$(-1)^{2N(t)} = 1.$$

This implies

$$E\left[(-1)^{N(t)+N(s)}\right] = E\left[(-1)^{N(s)-N(t)+2N(t)}\right]$$

$$= E\left[(-1)^{N(s)-N(t)}\right] \cdot E\left[(-1)^{2N(t)}\right] = E\left[(-1)^{N(s)-N(t)}\right].$$

The law of the unconscious statistician (2.4) entails

$$E\left[(-1)^{N(s)-N(t)}\right] = (+1) \cdot P\left(N(s) - N(t) = \text{ even}\right)$$

$$+(-1) \cdot P\left(N(s) - N(t) = \text{ odd }\right).$$

But $N(s) - N(t) \in \text{Po}\left(\lambda(s-t)\right)$, and the very same reasoning that produced (12.7) and (12.8) entails also

$$P\left(N(s) - N(t) = \text{ even}\right) = \frac{1}{2}\left(1 + e^{-2\lambda(s-t)}\right),$$

as well as

$$P\left(N(s) - N(t) = \text{ odd }\right) = \frac{1}{2}\left(1 - e^{-2\lambda(s-t)}\right).$$

Therefeore

$$(+1) \cdot P\left(N(s) - N(t) = \text{ even}\right) + (-1) \cdot P\left(N(s) - N(t) = \text{ odd }\right)$$

$$= \frac{1}{2}\left(1 + e^{-2\lambda(s-t)}\right) - \frac{1}{2}\left(1 - e^{-2\lambda(s-t)}\right) = e^{-2\lambda(s-t)}.$$

For the second term in the right hand side of (12.21) it is ascertained in the same manner

$$E\left[X(t)X(s) \mid Y = 1\right] = (-1)^2 E\left[(-1)^{N(t)+N(s)}\right]$$

$$= e^{-2\lambda(s-t)}.$$

Therefore we have for $s > t$

$$E\left[X(t)X(s)\right] = e^{-2\lambda(s-t)} P\left(Y = 0\right) + e^{-2\lambda(s-t)} P\left(Y = 1\right) =$$

$$= e^{-2\lambda(s-t)}, \quad s - t > 0. \tag{12.22}$$

ii) If $s < t$, $t - s > 0$, we write

$$N(t) + N(s) = N(t) - N(s) + 2N(s).$$

Again, $N(t) - N(s)$ is independent of $N(s)$,

$$P\left(N(t) - N(s) = \text{ even}\right) = \frac{1}{2}\left(1 + e^{-2\lambda(t-s)}\right)$$

$$P\left(N(t) - N(s) \text{ odd }\right) = \frac{1}{2}\left(1 - e^{-2\lambda(t-s)}\right).$$

Thus, as above we get

$$E\left[X(t)X(s)\right] = e^{-2\lambda(t-s)}, \quad s < t. \tag{12.23}$$

The results in (12.22) and (12.23) are expressed by [2] a single formula.

Lemma 12.5.3

$$E\left[X(t)X(s)\right] = e^{-2\lambda|t-s|}. \tag{12.24}$$

■

[2]

$$\mid t - s \mid = \begin{cases} t - s & t > s \\ -(t-s) & s \geq t \end{cases} \Leftrightarrow - \mid t - s \mid = \begin{cases} -(t-s) & t > s \text{ case ii) in eq. (12.23)} \\ t - s = -(s-t) & s \geq t \text{ case i) in eq. (12.22)} \end{cases}$$

Proposition 12.5.4 The RTS is Weakly Stationary

Proof The mean function is a constant (=0), as established in (12.20). The autocorrelation function is

$$R_{\mathbf{X}}(h) = e^{-2\lambda|h|} = R_{\mathbf{X}}(-h), \quad h = t - s,$$

as given in (12.24). ■

In fact we are going to show that the RTS modelled by Poissonian flips is strictly stationary, which implies proposition 12.5.4. We shall, however, first establish mean square continuity and find the power spectral density.

We know, see theorem 9.3.2, that a weakly stationary process is **mean square continuous**, if its autocovariance function is continuous in origin. Autocovariance function is $e^{-2\lambda|h|}$ and is continuous in the origin, and the conclusion follows. In other words

$$E\left[(X(t+h) - X(t))^2\right] \to 0, \quad \text{as } h \to 0.$$

Here we see very clearly that continuity in mean square does not tell about continuity of sample paths. Every sample path of the weakly stationary RTS is discontinuous, or, more precisely, every sample path has a denumerable number of discontinuities of the first kind[3]. The discontinuities are the level changes at random times.

The astute reader recognizes in (12.24) the same expression as in (11.13), the autocorrelation function of an Ornstein-Uhlenbeck process. This shows once more that identical second order properties can correspond to processes with quite different sample path properties.

The Power Spectral Density

By the table of pairs of autocorrelation functions and spectral densities in section 9.3.1 we get the Lorentzian distribution

$$R_{\mathbf{X}}(h) = e^{-2\lambda|h|} \overset{\mathcal{F}}{\leftrightarrow} s_{\mathbf{X}}(f) = \frac{4\lambda}{4\lambda^2 + f^2}. \tag{12.25}$$

The figure 12.2 depicts the spectral density $s_{\mathbf{X}}(f)$ for $\lambda = 1$ and $\lambda = 2$. This demonstrates that the random telegraph signal moves to higher frequencies, i.e., the spectrum is less concentrated at frequences f around zero, for higher values of the intensity λ, as seems natural.

12.5.4 The RTS is Strictly Stationary

The plan is to show the equality ,

$$P\left(X\left(t_n\right) = x_n, X\left(t_{n-1}\right) = x_{n-1}, \ldots, X\left(t_1\right) = x_1\right) =$$

$$P\left(X\left(t_n + h\right) = x_n, X\left(t_{n-1} + h\right) = x_{n-1}, \ldots, X\left(t_1 + h\right) = x_1\right)$$

for all n, $h > 0$ and all $0 \leq t_1 \leq t_2 \ldots \leq t_n$. This is nothing but a consequence of the fact that the Poisson process has independent increments over non-overlapping intervals, and that the increments have a distribution that depends only on the mutual differences of the times, and that lemma 12.5.1 above holds.

We observe by (12.16)

$$X(t+h) = (-1)^{Y + N(t+h)}, \quad h > 0.$$

[3] We say that a function $f(t)$ for $t \in [0, T]$, has only **discontinuities of the first kind**, if the function is 1) bounded and 2) for every $t \in [0, T]$, the limit from left $\lim_{s \uparrow t} f(s) = f(t-)$ and the limit from the right $\lim_{s \downarrow t} f(s) = f(t+)$ exist [91, p. 94].

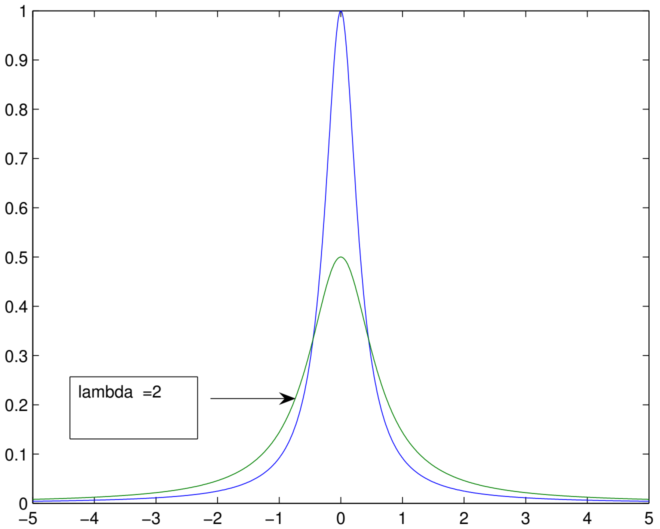

Figure 12.2: Spectral densities for RTS with $\lambda = 1$, $\lambda = 2$.

Then

$$X(t+h) = (-1)^{Y+N(t+h)-N(t)+N(t)} = (-1)^{N(t+h)-N(t)}\,(-1)^{Y+N(t)}$$
$$= (-1)^{N(t+h)-N(t)}\,X(t),$$

where we used (12.16) once more. Thus we have

$$X(t+h) = (-1)^{N(t+h)-N(t)}\,X(t), \quad h > 0. \tag{12.26}$$

This expression for $X(t+h)$ implies the following. If the UP-DOWN status of $X(t)$ is known and given, the status of $X(t+h)$ is determined by $N(t+h) - N(t)$. But $N(t+h) - N(t)$ is independent of $X(t)$, because the increments of the Poisson process are independent. Hence we have shown that

$$P\left(X(t+h) = a \mid X(t) = b\right) = P\left(N(t+h) - N(t) \text{ odd/even }\right) \tag{12.27}$$

with respective combinations of $a = \pm 1, b = \pm 1$ and of even/odd. But as the increments of the Poisson process are independent,

$$P\left(X(t+h) \mid X\left(t_n\right), \ldots, X\left(t_1\right)\right) = P\left(X(t+h) \mid X\left(t_n\right)\right)$$

for $t_1 \leq \ldots \leq t_n < t+h$, which is a **Markov property**. Then the **chain rule** in (3.34) implies that

$$P\left(X\left(t_n\right) = x_n, X\left(t_{n-1}\right) = x_{n-1}, \ldots, X\left(t_1\right) = x_1\right)$$

$$= P\left(X\left(t_n\right) = x_n \mid X\left(t_{n-1}\right) = x_{n-1}, \ldots, X\left(t_1\right) = x_1\right)$$

$$\cdot P\left(X\left(t_{n-1}\right) = x_{n-1} \mid X\left(t_{n-2}\right) = x_{n-2}, \ldots, X\left(t_1\right) = x_1\right)$$

$$\cdot \ldots \cdot P\left(X\left(t_2\right) = x_2 \mid X\left(t_1\right) = x_1\right) P\left(X\left(t_1\right) = x_1\right).$$

$$= P\left(X\left(t_n\right) = x_n \mid X\left(t_{n-1}\right) = x_{n-1}\right) \cdot \ldots \cdot P\left(X\left(t_2\right) = x_2 \mid X\left(t_1\right) = x_1\right) \cdot P\left(X\left(t_1\right) = x_1\right).$$

Every factor in the last product is one of the combinations of the form in (12.27)

$$P\left(X(t_i) = a \mid X(t_{i-1}) = b\right) = P\left(N\left(t_i\right) - N\left(t_{i-1}\right) \text{ odd/even}\right).$$

Let now time be shifted by

$$t_i \mapsto t_i + h$$

for every t_i. For these times it holds again

$$P\left(X\left(t_n + h\right) = x_n, X\left(t_{n-1} + h\right) = x_{n-1}, \ldots, X\left(t_1 + h\right) = x_1\right) =$$

$$P\left(X\left(t_n + h\right) = x_n \mid X\left(t_{n-1} + h\right) = x_{n-1}\right) \cdot \ldots \cdot P\left(X\left(t_1 + h\right) = x_1\right).$$

Every factor in this last product is one of the combinations of the same form as prior to the time shift, and are of the form in (12.27)

$$P\left(X(t_i + h) = a \mid X(t_{i-1} + h) = b\right) = P\left(N\left(t_i + h\right) - N\left(t_{i-1} + h\right) \text{ odd/even}\right).$$

But we recall that

$$N\left(t_i + h\right) - N\left(t_{i-1} + h\right) \in \text{Po}\left(\lambda(t_i - t_{i-1})\right)$$

and

$$N\left(t_i\right) - N\left(t_{i-1}\right) \in \text{Po}\left(\lambda(t_i - t_{i-1})\right)$$

for every i. In addition, lemma 12.5.1 yields that

$$P\left(X\left(t_1 + h\right) = x_1\right) = P\left(X\left(t_1\right) = x_1\right).$$

These observations imply unequivocally that

$$P\left(X\left(t_n\right) = x_n, X\left(t_{n-1}\right) = x_{n-1}, \ldots, X\left(t_1\right) = x_1\right) =$$

$$P\left(X\left(t_n + h\right) = x_n, X\left(t_{n-1} + h\right) = x_{n-1}, \ldots, X\left(t_1 + h\right) = x_1\right)$$

for all n, $h > 0$ and all $0 \leq t_1 \leq t_2 \ldots \leq t_n$. By this we have shown that the Poisson (and Markov) model of RTS is strictly stationary. ■

12.6 Exercises

12.6.1 Basic Poisson Process Probability

1. $\mathbf{N} = \{N(t) \mid t \geq 0\}$ is a Poisson process with parameter $\lambda > 0$. Find

$$\mathbf{P}\left(N(3) = 2 \mid N(1) = 0, N(5) = 4\right).$$

Answer: $\frac{3}{8}$.

2. $\mathbf{N} = \{N(t) \mid t \geq 0\}$ is a Poisson process with parameter $\lambda > 0$. Show that

$$\frac{N(t)}{t} \overset{P}{\to} \lambda,$$

as $t \to \infty$.

3. What is the probability that one of two independent Poisson processess reaches the level 2, before the other reaches the level 1. *Answer:* $\frac{1}{2}$.

4. We say that $\mathbf{N} = \{N(t) \mid t \geq 0\}$ is a Poisson process with random intensity Λ, if $\mathbf{N} \mid \Lambda = \lambda$ is a Poissonprocess with intensity $\Lambda = \lambda$. Or, for every $t \geq 0$, $N(t) \mid \Lambda = \lambda \in \text{Po}(\lambda t)$. We assume that $\Lambda \in \text{Exp}(1/\alpha)$.

   (a) Show that
   $$P\left(N(t) = k\right) = \frac{\alpha}{t+\alpha}\left(\frac{t}{t+\alpha}\right)^k, k = 0, 1, 2, \ldots \tag{12.28}$$

   (b) Find the m.g.f. $\psi_{N(t)}(s)$ of $N(t)$.

5. $\mathbf{N}_1 = \{N_1(t) \mid t \geq 0\}$ is a Poisson process with intensity λ_1, and $\mathbf{N}_2 = \{N_2(t) \mid t \geq 0\}$ is another Poisson process with intensity λ_2. $\mathbf{N}_1$ and $\mathbf{N}_2$ are independent of each other.

   Consider the probability that the first event occurs for $\mathbf{N}_1$, i.e., that $\mathbf{N}_1$ jumps from zero to one before $\mathbf{N}_2$ jumps for the first time. Show that
   $$P(\text{ first jump for } \mathbf{N}_1 \text{ }) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

6. (From [35]) Let $\mathbf{N} = \{N(t) \mid t \geq 0\}$ be a Poisson process with intensity $\lambda = 2$. We form a new stochastic process by
   $$X(t) = \left\lfloor \frac{N(t)}{2} \right\rfloor, \quad t \geq 0,$$
   where $\lfloor x \rfloor$ is the floor function, the integer part of the real number x, i.e., if k is an integer,
   $$\lfloor x \rfloor = k, \qquad \text{for } k \leq x < k+1,$$
   i.e., the largest integer smaller than or equal to x.

   (a) Find the probability
   $$\mathbf{P}\left(X(1) = 1, X(2) = 1\right).$$

   (b) Find the conditional probability
   $$\mathbf{P}\left(X(3) = 3 \mid X(1) = 1, X(2) = 1\right).$$

7. Show that the Poisson process is continuous in probability, i.e.,
   $$N(s) \overset{P}{\to} N(t),$$
   as $s \to t$.

8. $\mathbf{N} = \{N(t) \mid t \geq 0\}$ is a Poisson process with parameter $\lambda > 0$. $T =$ the time of occurrence of the first event. Determine for all $t \in (0,1)$ the probability $\mathbf{P}\left(T \leq t \mid N(1) = 1\right)$. Or, what is the distribution of $T \mid N(t) = 1$? *Answer:* $U(0,1)$.

9. $\mathbf{N} = \{N(t) \mid t \geq 0\}$ is a Poisson process with parameter $\lambda > 0$. We take the conditioning event $\{N(t) = k\}$. Recall (12.3), i.e, $N(t) = \max\{k | t \geq T_k\}$. $T_j =$, $j = 1, \ldots, k$ are the times of occurrence of the jth event, respectively.

Show that for $1 \leq j \leq k$, $0 \leq x \leq t$,

$$\mathbf{P}\left(T_j \leq x \mid N(t) = k\right) = \left(\begin{array}{c} k \\ j \end{array}\right) \left(\frac{x}{t}\right)^j \left(1 - \frac{x}{t}\right)^{k-j} \tag{12.29}$$

In the right hand side we recognize the p.m.f. of Bin $\left(k, \frac{x}{t}\right)$. This is a significant finding, let us see why. If $T_j \leq x$, then $T_1 \leq x$, $T_2 \leq x$, $\ldots$, $T_{j-1} \leq x$, too. We think of drawing independently k points from $U[0,t]$. Then the probability of any of them landing to the left of x $(\leq t)$ is $\frac{x}{t}$. Hence the probability of j of the points landing to the left of x is equal to the binomial probability in the right hand side of 12.29, recall the generative model of the Binomial distribution. Thus we have found

$$T_1, \ldots, T_k \text{ are independent and } T_j \in U(0,t),\ j = 1, \ldots, k, \text{ conditioned on } \{N(t) = k\}. \tag{12.30}$$

10. **The Distribution Function of the Erlang Distribution** Let $X \in \text{Erlang}\,(n, 1/\lambda)$. Show that

$$F_X(t) = \mathbf{P}\left(X \leq t\right) = 1 - \sum_{j=0}^{n-1} e^{-\lambda t} \frac{(\lambda t)^j}{j!}.$$

Aid: Let $\mathbf{N} = \{N(t) \mid t \geq 0\}$ be a Poisson process with parameter $\lambda > 0$. T_n is its nth arrival time of $\mathbf{N}$, then convince yourself first of the fact that, $\{T_n \leq t\} = \{N(t) \geq n\}$.

11. $\mathbf{N} = \{N(t) \mid t \geq 0\}$ is a Poisson process with intensity λ. T_k and T_{k+1} are the kth and $k+1$th occurrence/arrival times.

(a) Show that for $t > s \geq 0$

$$\mathbf{P}\left(T_k \leq s, T_{k+1} > t\right) = \frac{(s\lambda)^k}{k!} e^{-\lambda t}.$$

(b) Show that for $t > s \geq 0$

$$f_{T_k, T_{k+1}}(s,t) = \lambda^{k+1} \frac{s^{k-1}}{(k-1)!} e^{-\lambda t}.$$

(c) Let τ_{k+1} be the $k+1$th occupation/interarrival time. Show that

$$f_{T_k, \tau_{k+1}}(u,v) = \lambda^{k+1} \frac{u^{k-1}}{(k-1)!} e^{-\lambda(v+u)}$$

What is the conclusion?

12.6.2 Various Processes Obtained from the Poisson Process

1. Let $\mathbf{N} = \{N(t) \mid t \geq 0\}$ be a Poisson process with intensity $\lambda > 0$. We define N_k for $k = 0, 1, 2, \ldots$ by sampling at the non negative integers and by subtraction of the mean function at the sampling points, i.e.,

$$N_k = N(k) - \lambda \cdot k.$$

Let $\mathcal{F}_k^{\mathbf{N}} = \sigma\{N_0, N_1, \ldots, N_k\}$. Show that $\{N_k, \mathcal{F}_k^{\mathbf{N}}\}_{k=0}^{\infty}$ is a martingale.

2. Let X_i, $i = 1, 2, \ldots$, be $X_i \in \text{Fs}\,(p)$ and I.I.D.. Let $\mathbf{N} = \{N(t) \mid t \geq 0\}$ be a Poisson process with intensity $\lambda > 0$. The process $\mathbf{N}$ is independent of X_i, $i = 1, 2, \ldots,$.

We define a new stochastic process $X = \{X(t) \mid t \geq 0\}$ with

$$X(t) = \sum_{i=1}^{N(t)} X_i, \quad X(0) = 0, X(t) = 0, \text{ if } N(t) = 0.$$

We say that $\mathbf{X} = \{X(t) \mid t \geq 0\}$ is a **Pólya-Aeppli process** or a **Compound Poisson process** with parameters p and λ.

(a) Show that the m.g.f. of $X(t)$ is

$$\psi_{X(t)}(s) = e^{\lambda t\left(\frac{e^s p}{1-e^s(1-p)}-1\right)}, \quad s \leq -\ln(1-p).$$

(b) Find using $\psi_{X(t)}(s)$ that

$$E[X(t)] = \frac{\lambda}{p} \cdot t, \quad \mathrm{Var}[X(t)] = \frac{\lambda \cdot (2-p)}{p^2} \cdot t.$$

(c) It is being said that a Pólya-Aeppli process is a generalisation of the Poisson process (or that the Poisson process is a special case of the Pólya-Aeppli process). Explain what this means. *Aid*: Consider a suitable value of p.

(d) The process

$$Z(t) = ct - X(t), \quad t \geq 0, c > 0,$$

where $\{X(t) \mid t \geq 0\}$ is a Pólya-Aeppli process, is often used as a model of an insurance business and is thereby called the **risk process of Pólya and Aeppli**. How should one interpret c, N and X_i:s with respect to the needs of an insurance company ?

3. (From [99] and sf2940 2012-10-17) $\mathbf{N} = \{N(t) \mid t \geq 0\}$ is a Poisson process with intensity $\lambda > 0$. We define the new process $\mathbf{Y} = \{Y(t) \mid 0 \leq t \leq 1\}$ by

$$Y(t) \overset{\text{def}}{=} N(t) - tN(1), \quad 0 \leq t \leq 1.$$

(a) Are the sample paths of $\mathbf{Y}$ nondecreasing? Justify your answer. *Answer:* No.

(b) Find $E[Y(t)]$. *Answer:* 0.

(c) Find $\mathrm{Var}[Y(t)]$. *Answer:* $\lambda t(1-t)$.

(d) Find the autocovariance of $\mathbf{Y}$. *Answer:*

$$\mathrm{Cov}_{\mathbf{Y}}(t,s) = \begin{cases} \lambda s(1-t) & s < t, \\ \lambda t(1-s) & t \leq s. \end{cases}$$

(e) Compare the autocovariance function in (d) with the autocovariance function in (10.79). What is Your explanation?

4. (From [99]) $\mathbf{N}_1 = \{N_1(t) \mid t \geq 0\}$ is a Poisson process with intensity λ_1, and $\mathbf{N}_2 = \{N_2(t) \mid t \geq 0\}$ is another Poisson process with intensity λ_2. $\mathbf{N}_1$ and $\mathbf{N}_2$ are independent of each other. Let T_1 and T_2 be the times of occurrence/arrival of the first two events in $\mathbf{N}_1$. Let

$$Y = N_2(T_2) - N_2(T_1)$$

be the number of events in $\mathbf{N}_2$ during $[T_1, T_2]$. Show that

$$\mathbf{P}(Y = k) = \frac{\lambda_1}{\lambda_1 + \lambda_2}\left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^k, \quad k = 0, 1, 2 \ldots$$

Aid: Note that $T_2 = \tau_2 + T_1$. Then

$$Y = N_2(\tau_2 + T_1) - N_2(T_1).$$

where τ_2 is independent of $T_1 = \tau_1$ by theorem 12.2.1. By extending the argument in the proof of the restarting theorems 12.3.1 and 12.3.2 we have that

$$N_2(t+T_1) - N_2(T_1) \mid \tau_2 = t \in \text{Po}(\lambda_2 t).$$

Now use the argument leading to (12.28).

5. **M.g.f. of the Filtered Shot Noise and Campbell's Formula** Use the m.g.f. in (12.15) derive (12.13).
6. Find the mean and variance of $Z(t) = \sum_{k=1}^{\infty} h\left(t - T_k\right), \quad t \geq 0$, when
$$h(t) = e^{-t}U(t),$$
and
$$U(t) = \begin{cases} 1 & t \geq 0, \\ 0 & t < 0. \end{cases}$$

12.6.3 RTS

1. **A Modified RTS** If the start value Y is removed, the modified RTS is
$$X_o(t) \overset{\text{def}}{=} (-1)^{N(t)}, \quad t \geq 0. \tag{12.31}$$
Show that
$$\mu_{\mathbf{X}_0}(t) = E\left[X_o(t)\right] = e^{-2\lambda t}. \tag{12.32}$$
2. Give a short proof of (12.20) without using (12.7), (12.8) and of (12.18).
3. Give a short proof of (12.24) without using (12.7), (12.8) and of (12.18).
4. Show that the m.g.f. of $X(t)$ in the RTS is
$$\psi_{X(t)}(s) = \cosh(t) \cdot \left[e^{-(1-\lambda)t}\cosh(\lambda t) + e^{-(1+\lambda)t}\sinh(\lambda t)\right].$$
Find $E\left(X(t)\right)$ and $\text{Var}\left(X(t)\right)$ by this m.g.f. and compare with the determination without a generating function.
5. **RTS as a Markov Chain in Continuous Time** In midst of the proof of strict stationarity we observed that for any $h > 0$ and any $t_1 < \ldots < t_n$
$$P\left(X(t+h) \mid X\left(t_n\right), \ldots, X\left(t_1\right)\right) = P\left(X(t+h) \mid X\left(t_n\right)\right). \tag{12.33}$$
As already asserted, this says that the process $\mathbf{X} = \{X(t) \mid t \geq 0\}$ is a **Markov chain in continuous time**. As for all Markov chains in continuous time, we can define the **transition probabilities**
$$P_{ij}(t) \overset{\text{def}}{=} P\left(X(t) = j | X(0) = i\right), \quad i,j \in \{+1,-1\}.$$
This is the conditional probability that the RTS will be in state j at time t given that the RTS was in state i at time $t = 0$. These functions of t are arranged in the matrix valued function
$$t \mapsto \mathbf{P}_{\text{RTS}}(t) = \{P_{ij}(t)\}_{i\in\{+1,-1\},j\in\{+1,-1\}}.$$
   (a) Show that
$$\mathbf{P}_{\text{RTS}}(t) = \begin{pmatrix} \frac{1}{2}\left(1+e^{-2\lambda t}\right) & \frac{1}{2}\left(1-e^{-2\lambda t}\right) \\ \frac{1}{2}\left(1-e^{-2\lambda t}\right) & \frac{1}{2}\left(1+e^{-2\lambda t}\right) \end{pmatrix}.$$
   (b) Check the **Chapman - Kolmogorov equations**
$$\mathbf{P}_{\text{RTS}}(t+s) = \mathbf{P}_{\text{RTS}}(t)\mathbf{P}_{\text{RTS}}(s). \tag{12.34}$$

Chapter 13

The Kalman-Bucy Filter

13.1 Background

The recursive algorithm known as the Kalman filter was invented by Rudolf E. Kalman[1]. His original work was on random processes in discrete time, the extension to continuous time is known as the Kalman-Bucy Filter. The Kalman-Bucy filter produces an optimal (in the sense of mean squared error) estimate of the sample path (or, trajectory) of a stochastic process, which is observed in additive noise. The estimate is given by a stochastic differential equation. We consider only the linear model of Kalman-Bucy filtering.

The Kalman filter was prominently applied to the problem of trajectory estimation for the Apollo space program of the NASA (in the 1960s), and implemented in the Apollo space navigation computer. It was also used in the guidance and navigation systems of the NASA Space Shuttle and the attitude control and navigation systems of the International Space Station.

Robotics is a field of engineering, where the Kalman filter (in discrete time) plays an important role [98]. Kalman filter is in the phase-locked loop found everywhere in communications equipment. New applications of the Kalman Filter (and of its extensions like particle filters) continue to be discovered, including global positioning systems (GPS), hydrological modelling, atmospheric observations.

It has been understood only relatively recently that the Danish mathematician and statistician Thorvald N. Thiele[2] discovered the principle (and a special case) of the Kalman filter in his book published in Copenhagen in 1889: *Forelæsningar over Almindelig Iagttagelseslære: Sandsynlighedsregning og mindste Kvadraters Methode.* A translation of the book and an exposition of Thiele's work is found in [72].

13.2 The One-dimensional Kalman Filter in Continuous Time

Let us consider an Ornstein-Uhlenbeck process

$$dU(t) = aU(t)dt + \sigma dW(t), \quad t \geq 0. \tag{13.1}$$

and a process $\mathbf{Y} = \{Y(s) \mid 0 \leq t\}$ of noisy observations of it

$$dY(t) = cU(t)dt + gdV(t). \tag{13.2}$$

[1] b. 1930 in Budapest, but studied and graduated in electrical engineering in USA, Professor Emeritus in Mathematics at ETH, the Swiss Federal Institute of Technology in Zürich.

[2] 1838−1910, a short biography is in
`http://www-groups.dcs.st-ac.uk./~history/Biographies/Thiele.html`

where $\mathbf{V} = \{V(s) \mid 0 \leq t\}$ is another Wiener process, which is independent of the Wiener process $\mathbf{W} = \{W(s) \mid 0 \leq t\}$. Let us set

$$\widehat{U}(t) \overset{\text{def}}{=} E\left[U(t) \mid \mathcal{F}_t^Y\right] \tag{13.3}$$

where $\mathcal{F}_t^Y$ is the sigma field generated by $\{Y(s) \mid 0 \leq s \leq t\}$.

The Kalman-Bucy Filter for an Ornstein-Uhlenbeck process observed in additive Wiener noise is then found as follows.

$$d\widehat{U}(t) = \left(a - \frac{c^2 S(t)}{g^2}\right)\widehat{U}(t)dt + \frac{cS(t)}{g^2}dY(t), \quad \widehat{U}(0) = E\left[U(0)\right]. \tag{13.4}$$

Here

$$S(t) = E\left[\left(U(t) - \widehat{U}(t)\right)^2\right], \tag{13.5}$$

and $S(t)$ satisfies the deterministic first order nonlinear differential equation known as *Riccati equation*

$$\frac{d}{dt}S(t) = 2aS(t) - \frac{c^2}{g^2}S(t)^2 + \sigma^2, \quad S(0) = E\left[\left(U(0) - E\left[U(0)\right]\right)^2\right]. \tag{13.6}$$

The Riccati equation can be solved as

$$S(t) = \frac{\alpha_1 - K\alpha_2 e^{\frac{(\alpha_1-\alpha_2)c^2 t}{g^2}}}{1 - Ke^{\frac{(\alpha_1-\alpha_2)c^2 t}{g^2}}}, \tag{13.7}$$

where

$$\alpha_1 = c^{-2}\left(ag^2 - g\sqrt{a^2g^2 + c^2\sigma^2}\right)$$
$$\alpha_2 = c^{-2}\left(ag^2 + g\sqrt{a^2g^2 + c^2\sigma^2}\right)$$

and

$$K = \frac{S(0) - \alpha_1}{S(0) - \alpha_2}.$$

To derive these expressions we need the results about $\widehat{U}(t)$ as a projection on the linear span of $\{Y(s) \mid 0 \leq s \leq t\}$ (these hold by Gaussianity), c.f. section 7.5, its representation as a Wiener integral and in the mean square. (Forthcoming, $\approx$ 10 pages))

Bibliography

[1] P. Albin: *Stokastiska processer.* (Stochastic Processes; in Swedish), Studentlitteratur, Lund 2003.

[2] D. Aldous: *Probability Approximations via the Poisson Clumping Heuristic.* Springer-Verlag, New York 1989.

[3] L.C. Andrews: *Special Functions of Mathematics for Engineers.* SPIE Optical Engineering Press, Bellingham; Washington, and Oxford University Press. Oxford, Tokyo, Melbourne, 1998.

[4] A. H. S. Ang & W. H. Tang. *Probability Concepts in Engineering: Emphasis on Applications to Civil and Environmental Engineering 2nd Edition.* John Wiley & Sons, New York, 2007.

[5] H. Anton & C. Rorres: *Elementary Linear Algebra with Supplemental Applications.* John Wiley & Sons (Asia) Pte Ltd, 2011.

[6] C. Ash: *The Probability Tutoring Book. An Intuitive Course for Engineers and Scientists (and everyone else!).* IEEE Press, Piscataway, New Jersey, 1993.

[7] A.V. Balakrishnan: *Stochastic Differential Systems I.* Springer Verlag, Berlin 1973.

[8] A.V. Balakrishnan: *Introduction to Random Processes in Engineering.* John Wiley & Sons, Inc., New York, 1995.

[9] A. Barbour, L. Holst & S. Jansson: *Poisson Approximation.* Clarendon Press, Oxford, 1992.

[10] H.C. Berg: *Random Walks in Biology. Expanded Edition.* Princeton University Press, Princeton, New Jersey, 1993.

[11] A. Bernow, T. Bohlin, C. Davidson, R. Magnusson, G. Markesjö & S-O. Öhrvik : *Kurs i elektroniskt brus.* (A Course in Electronic Noise; in Swedish) Svenska teknologföreningen, Stockholm, 1961.

[12] D.P. Bertsekas & J.N. Tsitsiklis: *Introduction to Probability.* Athena Scientific, Belmont, Massachusetts, 2002.

[13] T. Björk: *Arbitrage Theory in Continuous Time. Third Edition.* Oxford University Press, Oxford, 2009.

[14] G. Blom: *Sannolikhetsteori för FEMV.* (Lecture notes in Probability for Engineering Physics, Electrical, Mechanical and Civil Engineering; in Swedish) Lund, 1968.

[15] G. Blom, L. Holst & D. Sandell: *Problems and Snapshots from the World of Probability.* Springer Verlag, Berlin, New York, Heidelberg, 1994.

[16] G. Blom, J. Enger, G. Englund, J. Grandell & L. Holst: *Sannolikhetsteori och statistikteori med tillämpningar.* (Probability Theory and Statistical Theory with Applications; in Swedish) Studentlitteratur, Lund 2005.

[17] S.J. Blundell & K.M. Blundell: *Concepts in Thermal Physics (Second Edition).* Oxford University Press, Oxford, New York, Auckland, Cape Town, Dar es Salaam, Hong Kong, Karachi, Kuala Lumpur, Madrid, Melbourne, Mexico City, Nairobi, New Delhi, Shanghai, Taipei, Toronto, 2010.

[18] L. Boltzmann: *Entropie und Wahrscheinlichkeit.* Oswalds Klassiker der Exakten Wissenschaften Band 286. Verlag Harri Deutsch, Frankfurt am Main, 2008.

[19] A.N. Borodin & P. Salminen: *Handbook of Brownian Motion. Second Edition*, Birkhäuser, Basel, Boston & Berlin, 2002.

[20] Z. Brzeźniak & T. Zastawniak: *Basic Stochastic Processes.* Springer London Ltd, 2005.

[21] G. Chaitin: *Meta Math. The Quest for Omega.* Vintage Books, A Division of Random House Inc., New York, 2005.

[22] C.V.L. Charlier: *Vorlesungen über die Grundzüge der Mathematischen Statistik.* Verlag Scientia, Lund, 1920.

[23] T.M. Cover & J.A. Thomas: *Elements of Information Theory.* J. Wiley & Sons, Inc., New York, 1991.

[24] H. Cramér: *Mathematical Methods of Statistics.* Princeton Series of Landmarks in Mathematics and Physics. Nineteenth Printing & First Paperback Edition. Princeton University Press, Princeton, 1999.

[25] H. Cramér & M.R. Leadbetter: *Stationary and Related Stochastic Processes: Sample Function Properties and Their Applications*, Dover Publications Inc., Mineola, New York 2004 (a republication of the work originally published in 1967 by John Wiley & Sons Inc., New York).

[26] M.H.A. Davis: *Linear Estimation and Stochastic Control.* Chapman and Hall, London, 1977.

[27] M. Davis & A. Etheridge: *Louis Bachelier$'$s Theory of $peculation. The Origins of Modern Finance.* Princeton University Press, Princeton and Oxford, 2006.

[28] J.L. Devore: *Probability and Statistics for the Engineering and Sciences. Fourth Edition.* Duxbury Press, Pacific Grove, Albany, Belmont, Bonn, Boston, Cincinnati, Detroit, Johannesburg, London, Madrid, Melbourne, Mexico City, New York, Paris, Singapore, Tokyo, Toronto, Washington, 1995.

[29] B. Djehiche: *Stochastic Calculus. An Introduction with Applications.* Lecture Notes, KTH, 2000.

[30] A.Y. Dorogovtsev, D.S. Silvestrov, A.V. Skorokhod & M.I. Yadrenko: *Probability Theory: Collection of Problems*, Translations of Mathematical Monographs, vol. 163, American Mathematical Society, Providence, 1997.

[31] A. Einstein: *Investigations on the Theory of the Brownian Movement.* Dover Publications Inc., Mineola, New York 1956 (a translation of the work in German originally published in 1926).

[32] I. Elishakoff: *Probabilistic Theory of Structures. Second Edition.* (Dover Civil and Mechanical Engineering), Dover Inc., Mineola, N.Y., 1999.

[33] G. Einarsson: *Principles of Lightwave Communications.* John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 1996.

[34] J.D. Enderle, D.C. Enderle & D.J. Krause: *Advanced Probability Theory for Biomedical Engineers.* Morgan & Claypool Publishers, 2006.

[35] G. Englund & M. Möller: *Problemsamling i sannolikhetsteori.* (A Collection of Problems in Probability; in Swedish), Matematisk statistik, KTH, 1982.

[36] A. Friedman: *Foundations of Modern Analysis.* Dover Publications Inc., New York 1982.

[37] A.G.Frodesen, O. Skjeggestad & H. TØfte: *Probability and statistics in particle physics.* Universitetsforlaget, Bergen, 1979.

[38] W. Gardner: *Introduction to Random Processes. With Applications to Signals and Systems. Second Edition.* McGraw-Hill Publishing Company, New York, St.Louis, San Francisco, Auckland, Bogotá, Caracas, Lisbon, London, Madrid, Mexico City, Milan, Montreal, New Delhi, Oklahoma City, Paris, San Juan, Singapore, Sydney, Tokyo, Toronto, 1990.

[39] D.T. Gillespie: Fluctuation and dissipation in Brownian motion. *American Journal of Physics.* vol. 61, pp. 1077−1083, 1993.

[40] D.T. Gillespie: The mathematics of Brownian motion and Johnson noise. *American Journal of Physics.* vol. 64, pp. 225−240, 1996.

[41] R.L. Graham, D.E. Knuth & O. Patashnik: *Concrete Mathematics. A Foundation for Computer Science.* Addison-Wesley Publishing Company, Reading Massachusetts, Menlo Park California, New York, Don Mills, Ontario, Wokingham England, Amsterdam, Bonn, Sydney, Singapore, Tokyo, Madrid, San Juan, 1989.

[42] J. Grandell, B. Justusson & G. Karlsson: *Matematisk statistik för E-linjen. Exempelsamling 3* (Mathematical Statistics for Electrical Engineers, A Collection of Examples; in Swedish), KTH/avdelning för matematisk statistik, Stockholm 1984.

[43] R.M. Gray & L.D. Davisson: *Random Processes. A Mathematical Introduction to Engineers.* Prentice-Hall, Inc., Englewood Cliffs, 1986.

[44] R.M. Gray & L.D. Davisson: *An Introduction to Statistical Signal Processing.* Cambridge University Press, Cambridge, 2004.

[45] D.H. Green & D.E. Knuth & O. Patashnik: *Mathematics for the Analysis of Algorithms. Third Edition* Birhäuser, Boston, Basel, Berlin, 1990.

[46] U. Grenander: *Abstract Inference.* John Wiley & Sons, New York, Chichester, Brisbance, Toronto, 1981.

[47] U. Grenander & G. Szegö: *Toeplitz Forms and Their Applications.* American Mathematical Society, Chelsea, 2nd (textually unaltered) edition, 2001.

[48] G.R. Grimmett & D.R. Stirzaker: *Probability and Random Processes. Second Edition.* Oxford Science Publications, Oxford, New York, Toronto, Delhi, Bombay, Calcutta, Madras, Karachi, Kuala Lumpur, Singapore, Hong Kong, Tokyo, Nairobi, Dar es Salaam, Cape Town, Melbourne, Auckland, Madrid. Reprinted Edition 1994.

[49] A. Gut: *An Intermediate Course in Probability. 2nd Edition.* Springer Verlag, Berlin 2009.

[50] B. Hajek: *Random Processes for Engineers.* Cambridge University Press, Cambridge, 2015.

[51] A. Hald: *Statistical theory with engineering applications*, John Wiley & Sons, New York, 1952.

[52] B. Hallert: *Elementär felteori för mätningar.* (Elementary Error Theory for Measurement; in Swedish) P.A. Norstedt & Söners Förlag, Stockhom, 1967.

[53] R.W. Hamming: *The art of probability for scientists and engineers.* Addison-Wesley Publishing Company, Reading Massachusetts, Menlo Park California, New York, Don Mills, Ontario, Wokingham England, Amsterdam, Bonn, Sydney, Singapore, Tokyo, Madrid, San Juan, Paris, Seoul, Milan, Mexico City, Taipei, 1991.

[54] J. Havil: *Gamma. exploring euler's constant.* Princeton University Press, Princeton, Oxford, 2003.

[55] L.L. Helms; *Introduction to Probability Theory with Contemporary Applications.* W.H. Freeman and Company, New York, 1997.

[56] C.W. Helstrom: *Probability and Stochastic Processes for Engineers. Second Edition.* Prentice-Hall, Upper Saddle River, 1991.

[57] U. Hjorth: *Stokastiska Processer. Korrelations- och spektralteori.* (Stochastic Processes. Correlation and Spectral Theory; in Swedish) Studentlitteratur, Lund, 1987.

[58] K. Huang: *Introduction to Statistical Physics.* CRC Press, Boca Raton, London, New York, Washington D.C., 2001.

[59] H. Hult, F. Lindskog, O. Hammarlid & C.J. Rehn: *Risk and Portfolio Analysis: Principles and Methods.* Springer, New York, Heidelberg, Dordrecht, London, 2012.

[60] H.L. Hurd & A. Miamee: *Periodically Correlated Random Sequences. Spectral Theory and Practice.* John Wiley & Sons, Inc., Hoboken, New Jersey, 2007.

[61] K. Itô: *Introduction to probability theory.* Cambridge University Press, Cambridge, New York, New Rochelle, Melbourne, Sydney, 1986.

[62] K. Jacobs: *Stochastic Processes for Physicists: Understanding Noisy Systems.* Cambridge University Press, Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi, Dubai, Tokyo, 2010.

[63] J. Jacod & P. Protter: *Probability Essentials.* Springer Verlag, Berlin 2004.

[64] F. James: *Statistical Methods in Experimental Physics. 2nd Edition.* World Scientific, New Jersey, London, Singapore, Beijing, Shanghai, Hong Kong, Taipei, Chennai, 2006.

[65] E. Jaynes: *Probability Theory. The Logic of Science.* Cambridge University Press, Cambridge, 2003.

[66] A.H. Jazwinski: *Stochastic Processes and Filtering Theory.* Dover Publications Inc., Mineola, New York, 1998.

[67] O. Kallenberg: *Foundations of Modern Probability. Second Edition.* Springer Verlag, Berlin, 2001.

[68] G. Kallianpur & P. Sundar: *Stochastic Analysis and Diffusion Processes.* Oxford Graduate Texts in Mathematics, Oxford University Press, Oxford 2014.

[69] A. Khuri: *Advanced Calculus with Applications to Statistics.* John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore, 1987.

[70] F.M Klebaner: *An Introduction to Stochastic Calculus with Applications.* Imperial College Press, Singapore, 1998.

[71] L. Kristiansson & L.H. Zetterberg: *Signalteori I −II* (Signal Theory; in Swedish). Studentlitteratur, Lund 1970.

[72] S.L. Lauritzen: *Thiele: pioneer in statistics.* Oxford University Press, 2002.

[73] D.S. Lemons: *An Introduction to Stochastic Processes in Physics (containing* 'On the Theory of Brownian Motion' by Paul Langevin translated by Anthony Gythiel*).* The Johns Hopkins University Press. Baltimore, London, 2002.

[74] A. Leon-Garcia: *Probability and Random Processes for Electrical Engineers.* Addison-Wesley Publishing Company. Reading, 1989.

[75] J.W. Lindeberg: *Todennäköisyyslasku ja sen käytäntö tilastotieteessä. Alkeellinen esitys.* (Probability calculus and its practice in statistics. An elementary presentation. in Finnish) Otava, Helsinki, 1927.

[76] G. Lindgren: *Stationary Stochastic Processes: Theory and Applications.* Chapman & Hall CRC Texts in Statistical Science Series, Boca Raton, 2013.

[77] H.O. Madsen, S. Krenk & N.C. Link: *Methods of Structural Safety.* Dover Publications Inc., Mineola, New York, 2006.

[78] R.M. Mazo: *Brownian Motion. Fluctuations, Dynamics, and Applications.* International Series of Monographs on Physics No. 112, Oxford Science Publications, Oxford University Press, Oxford 2002, (First Paperback Edition 2009).

[79] M. Mitzenmacher, & E. Upfal: *Probability and computing: Randomized algorithms and probabilistic analysis.* Cambridge University Press, Cambridge, New York, Port Melbourne, Madrid, Cape Town, 2005.

[80] R.E. Mortensen: *Random Signals and Systems.* John Wiley & Sons, Inc., New York, 1987.

[81] J. Neveu: *Bases Mathématiques du Calcul des Probabilites*, Masson et Cie., Paris 1964.

[82] M. Neymark: *Analysens Grunder Del 2* (Foundations of Mathematical Analysis; in Swedish). Studentlitteratur, Lund, 1970.

[83] E. Ohlsson: *Felfortplantningsformlerna* (Propagation of Error; in Swedish). Matematisk statistik, Stockholms universitet, January 2005.

[84] C. Palm: *Intensity Variations in Telephone Traffic.* A Translation of *Intensitätsschwankungen im Fernsprechverkehr, Ericsson Technics, No. 44, 1943.* Translated by C. Jacobaeus. North-Holland, Amsterdam, New York, Oxford, Tokyo, 1988.

[85] A. Papoulis: *Probability, Random Variables, and Stochastic Processes. Second Edition.* McGraw-Hill Book Company. New York, St.Louis, San Francisco, Auckland, Bogotá, Caracas, Lisbon, London, Madrid, Mexico City, Milan, Montreal, New Delhi, San Juan, Singapore, Sydney, Tokyo, Toronto, 1984.

[86] J. Rissanen: *Lectures on statistical modeling theory.* Helsinki Institute of Information Technology, Helsinki, 2004.
`http://www.lce.hut.fi/teaching/S-114.300/lectures.pdf`

[87] B. Rosén: *Massfördelningar och integration. 2:a upplagan* (Mass distributions and integration; in Swedish) Unpublished Lecture Notes, Department of Mathematics, KTH, 1978.

[88] M. Rudemo & L. Råde: *Sannolikhetslära och statistik med tekniska tillämpningar. Del 1* (Probability and Statistics with Engineering Applications Part 1; in Swedish) Biblioteksförlaget, Stockholm, 1970.

[89] M. Rudemo & L. Råde: *Sannolikhetslära och statistik med tekniska tillämpningar. Del 2* (Probability and Statistics with Engineering Applications Part 2; in Swedish) Biblioteksförlaget, Stockholm, 1967.

[90] M. Rudemo: *Prediction and Filtering for Markov Processes. Lectures at the Department of Mathematics, Royal Institute of Technology*, TRITA-MAT-6 (Feb.), Stockholm, 1974.

[91] W. Rudin: *Principles of Mathematical Analysis. Third Edition.* McGraw-Hill Inc., New York, St.Louis, San Francisco, Auckland, Bogotá, Caracas, Lisbon, London, Madrid, Mexico City, Milan, Montreal, New Delhi, San Juan, Singapore, Sydney, Tokyo, Toronto, 1973.

[92] L. Råde & B. Westergren: *Mathematics Handbook for Science and Engineering. Second Edition.* Studentlitteratur, Lund & Beijing 2004.

[93] E.B. Saff & A.D. Snider: *Fundamentals of Complex Analysis for Mathematics, Science, and Engineering. Third Edition.* Pearson Educational International. Upper Saddle River, New Jersey 2003.

[94] Z. Schuss: *Theory and Applications of Stochastic Processes An Analytical Approach.* Springer Verlag, New York, Dordrecht, Heidelberg, London, 2010.

[95] Y.G. Sinai: *Probability Theory. An Introductory Course. Springer Textbook.* Springer Verlag, Berlin, Heidelberg, 1992.

[96] G. Sparr & A. Sparr: *Kontinuerliga system.* (Continuous Systems; in Swedish) Studentlitteratur, Lund 2010.

[97] H. Stark & J.W. Woods: *Probability, Random Processes and Estimation Theory for Engineers.* Prentice - Hall, 1986.

[98] S. Thrun, W. Burgard & D. Fox: *Probabilistic Robotics.* MIT Press, Cambridge, London, 2005.

[99] Y. Viniotis: *Probability and Random Processes for Electrical Engineers.* WCB McGraw- Hill, Boston, Burr Ridge, IL, Dubuque, WI, New York, San Francisco, St.Louis, Bangkok, Bogotá, Caracas, Lisbon, London, Madrid, Mexico City, Milan, New Delhi, Seoul, Singapore, Sydney, Taipei, Toronto, 1998.

[100] A. Vretblad: *Fourier Analysis and Its Applications.* Springer Verlag, New York, 2003.

[101] R.D. Yates & D.J. Goodman: *Probability and Stochastic Processes. A Friendly Introduction for Electrical and Computer Engineers. Second Edition.* John Wiley & Sons, Inc., New York, 2005.

[102] D. Williams: *Weighing the Odds. A Course in Probability and Statistics.* Cambridge University Press, Cambridge, 2004.

[103] E. Wong & B. Hajek: *Stochastic Processes in Engineering Systems.* McGraw-Hill Book Company, New York, 1985.

[104] A. Zayezdny, D. Tabak & D. Wulich: *Engineering Applications of Stochastic Processes. Theory, Problems and Solutions.* Research Studies Press Ltd., Taunton, Somerset, 1989.

[105] K.J. Åström: *Introduction to Stochastic Control Theory.* Academic Press, New York, San Francisco, London, 1970.

[106] K.J. Åström: *Harry Nyquist (1889 – 1976): A Tribute to the Memory of an Outstanding Scientist.* Royal Swedish Academy of Engineering Sciences (IVA), January 2003.

Index