SELECTED TOPICS IN PROBABILITY

Lectures notes by

Marco C. Campi
Dept. of Electrical Eng.
University of Brescia
Italy

marco.campi@ing.unibs.it

Abstract: This text contains a treatment of some specific topics in probability theory. The presentation is basically self-contained.
In preparing these notes, I have put a significant amount of effort into combining requirements of conciseness and mathematical rigor with those of readability and clarity of presentation. I shall be grateful to anyone that will provide feedback on how to improve these notes.

3rd March 2008

Contents

Chapter 1

MEASURE SPACES AND INTEGRATION

1.1 Measurable spaces and measurable functions

Measurable spaces

The notion of measurable space makes use of the concept of σ-algebra. A σ-algebra is a collection of subsets of a given set, with certain set-theoretic properties, and is introduced first.

DEFINITION 1.1 (σ-algebra) *Given a set X, a collection $\mathscr{X}$ of subsets of X is called a σ-algebra if*

(a) $X \in \mathscr{X}$;
(b) if $A_k \in \mathscr{X}$, $k = 1,2,\ldots$, then $\cup_{k=1}^{\infty} A_k \in \mathscr{X}$;
(c) if $A \in \mathscr{X}$, then $A^c \in \mathscr{X}$ (A^c is the complement of set A). □

Condition (b) states that a σ-algebra is closed under countably infinite union, that is union of an infinite number of subsets A_k, where k runs over the integers. Since $\emptyset = X^c$ and $X \in \mathscr{X}$ by (a), condition (c) implies that $\emptyset \in \mathscr{X}$. Taking $A_{n+1} = A_{n+2} = \cdots = \emptyset$ in (b), we then see that $\cup_{k=1}^{n} A_k \in \mathscr{X}$, that is $\mathscr{X}$ is also closed under finite union. Moreover, a σ-algebra is also closed under intersection, as it follows from relation $\cap_k A_k = (\cup_k A_k^c)^c$.

Given any collection $\mathscr{A}$ of subsets of a set X, consider all σ-algebras containing $\mathscr{A}$. Their intersection (that is the collection of the sets that belong to all σ-algebras) is easily seen to be a σ-algebra too. It is called the minimal σ-algebra containing $\mathscr{A}$ and is denoted by $\sigma(\mathscr{A})$.

We can now introduce the notion of measurable space.

DEFINITION 1.2 (measurable space) *A couple $(X, \mathscr{X})$ where X is any set and $\mathscr{X}$ is a σ-algebra of subsets of X is called a measurable space.* □

It is sometimes important to consider measurable spaces that are the product of other measurable spaces. This is formalized in the following definition.

DEFINITION 1.3 (product measurable space) *Given the measurable spaces $(X_k, \mathscr{X}_k), k = 1, 2, \ldots, n$, their product measurable space is $(X_1 \times X_2 \times \cdots \times X_n, \mathscr{X}_1 \otimes \mathscr{X}_2 \otimes \cdots \otimes \mathscr{X}_n)$, where $X_1 \times X_2 \times \cdots \times X_n$ is the direct product of the X_k's, i.e. the set of ordered n-tuples $(x_1, x_2, \cdots, x_n)$ with $x_k \in X_k$, $k = 1, 2, \ldots, n$, and $\mathscr{X}_1 \otimes \mathscr{X}_2 \otimes \cdots \otimes \mathscr{X}_n$ is the direct product of the $\mathscr{X}_k$'s, i.e. the smallest σ-algebra in $X_1 \times X_2 \times \cdots \times X_n$ that contains all sets of the form $A_1 \times A_2 \times \cdots \times A_n = \{(x_1, x_2, \cdots, x_n)$ such that $x_k \in A_k, k = 1, 2, \ldots, n\}$.* □

Measurable functions

DEFINITION 1.4 (measurable function – see Figure 1.1) *Given two measurable spaces $(X, \mathscr{X})$ and $(X', \mathscr{X}')$, a function $g : X \to X'$ is measurable if, for all $A' \in \mathscr{X}'$, the inverse image of A' through g, that is $g^{-1}(A') := \{x \in X : g(x) \in A'\}$, belongs to $\mathscr{X}$.* □

Sometimes, we emphasize one or both σ-algebras by writing $\mathscr{X}$-measurable or $\mathscr{X}/\mathscr{X}'$-measurable.

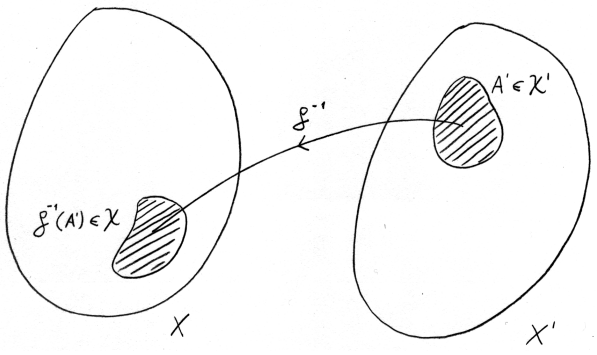

Figure 1.1: Measurable function.

The importance of measurability becomes apparent when speaking of measures and measure spaces, as we shall do in the next section.

The next two theorems study the measurability of functions constructed from other measurable functions.

THEOREM 1.5 (composition of measurable functions) *Given three measurable spaces $(X,\mathscr{X})$, $(X^{'},\mathscr{X}^{'})$, and $(X^{''},\mathscr{X}^{''})$, and two measurable functions $g:X\to X^{'}$ and $h:X^{'}\to X^{''}$, the composition of g and h, i.e. the function $h\cdot g:X\to X^{''}$ defined through relation $h\cdot g(x):=h(g(x))$, is a $\mathscr{X}/\mathscr{X}^{''}$-measurable function.*

PROOF. For any $A^{''}\in\mathscr{X}^{''}$, we have

$$(h\cdot g)^{-1}(A^{''})=g^{-1}(h^{-1}(A^{''})).$$

Since h is $\mathscr{X}^{'}/\mathscr{X}^{''}$-measurable, $h^{-1}(A^{''})\in\mathscr{X}^{'}$; in turn, the $\mathscr{X}/\mathscr{X}^{'}$-measurability of g implies that $g^{-1}\left(h^{-1}\left(A^{''}\right)\right)\in\mathscr{X}$, so showing the $\mathscr{X}/\mathscr{X}^{''}$-measurability of $h\cdot g$. □

THEOREM 1.6 (product and marginal measurable function)
Consider the measurable space $(X,\mathscr{X})$ and two other measurable spaces $(X_1,\mathscr{X}_1)$, and $(X_2,\mathscr{X}_2)$ along with their product $(X_1\times X_2,\mathscr{X}_1\otimes\mathscr{X}_2)$.
i) *Given two measurable functions $g_1:X\to X_1$ and $g_2:X\to X_2$, the function $g:X\to X_1\times X_2$ defined according to the relation $g(x)=(g_1(x),g_2(x)),x\in X$, is a $\mathscr{X}/\mathscr{X}_1\otimes\mathscr{X}_2$-measurable function from X to $X_1\times X_2$;*
ii) *conversely, if $g:X\to X_1\times X_2$ is $\mathscr{X}/\mathscr{X}_1\otimes\mathscr{X}_2$-measurable, then $g_1:X\to X_1$ such that $g_1(x)$ is the first component of $g(x)$ and the similarly defined $g_2:X\to X_2$ are measurable functions from X to X_1 and from X to X_2, respectively.*
The theorem extends in an obvious way to the product of more measurable spaces.

PROOF.

i) We need to show that $g^{-1}(A)\in\mathscr{X},\forall A\in\mathscr{X}_1\otimes\mathscr{X}_2$.

Consider the collection $\mathscr{D}$ of all sets $A\subset X_1\times X_2$ such that $g^{-1}(A)\in\mathscr{X}$. We prove the following two facts:
(a) $\mathscr{D}$ contains all sets of the form $A=A_1\times A_2$, $A_1\in\mathscr{X}_1$ and $A_2\in\mathscr{X}_2$;
(b) $\mathscr{D}$ is a σ-algebra.
Facts (a) and (b) imply the theorem thesis. Indeed, since $\mathscr{X}_1\otimes\mathscr{X}_2$ is the smallest σ-algebra that contains the sets of the form $A_1\times A_2$, $A_1\in\mathscr{X}_1$ and $A_2\in\mathscr{X}_2$, it follows from (a) and (b) that $\mathscr{X}_1\otimes\mathscr{X}_2\subseteq\mathscr{D}$, so that any set in $\mathscr{X}_1\otimes\mathscr{X}_2$ has an inverse image through g^{-1} which is in $\mathscr{X}$ and the thesis is proven.
(a) and (b) can be proven as follows:

(a) $g^{-1}(A_1\times A_2)=g^{-1}((A_1\times X_2)\cap(X_1\times A_2))=g^{-1}(A_1\times X_2)\cap g^{-1}(X_1\times A_2)=g_1^{-1}(A_1)\cap g_2^{-1}(A_2)\in\mathscr{X}$;
(b) if $A=\cup_{k=1}^{\infty}A_k$ with $A_k\in\mathscr{D}$, $k=1,2,\ldots$, then $g^{-1}(A)=g^{-1}(\cup_{k=1}^{\infty}A_k)=\cup_{k=1}^{\infty}g^{-1}(A_k)\in\mathscr{X}$, so that $\mathscr{D}$ is closed under union. The fact that $\mathscr{D}$ is closed under complementation and contains the entire $X_1\times X_2$ is proven in a similar way.

ii) For any $A_1 \in \mathscr{X}_1$, we have $g_1^{-1}(A_1) = g^{-1}(A_1 \times X_2) \in \mathscr{X}$, so that g_1 is $\mathscr{X}/\mathscr{X}_1$-measurable. Similarly, g_2 is $\mathscr{X}/\mathscr{X}_2$-measurable. □

1.2 Measures and measure spaces

A measure is a function that associates to any set belonging to a σ-algebra a non-negative number, the measure of the set. It has to satisfy a certain set-theoretic property called σ-additivity.

DEFINITION 1.7 (measure) *Let $\mathscr{X}$ be a σ-algebra. A function $m : \mathscr{X} \to [0,\infty]$ is called a measure if, for any countable collection of pairwise disjoint sets $A_k \in \mathscr{X}$, $k = 1,2,\ldots$, the following property (σ-additivity) holds*

$$m(\cup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} m(A_k). \tag{1.1}$$

□

Definition 1.7 forces $m(\emptyset) = 0$. To see this, take $A_1 = A$ and $A_2 = A_3 = \cdots = \emptyset$ in (6.13). Moreover, by taking $A_{k+1} = A_{k+2} = \ldots = \emptyset$ in (6.13) we see that the measure of the union of a finite number of sets equals the sum of their measures.

DEFINITION 1.8 (measure space) *A triple $(X, \mathscr{X}, m)$, where X is any set, $\mathscr{X}$ is a σ-algebra of subsets of X, and m is a measure on $\mathscr{X}$ is called a measure space.* □

Image measures

Consider two measurable spaces $(X, \mathscr{X})$ and $(X', \mathscr{X}')$ and a measurable function $g : X \to X'$. If the first space is endowed with a measure m, i.e. $(X, \mathscr{X}, m)$ is a measure space, then function g permits to define a measure m' on $\mathscr{X}'$ according to the following definition:

$$m'(A') := m(g^{-1}(A')), \quad A' \in \mathscr{X}'.$$

(Here, one should note the importance of the fact that g is measurable. If not, $g^{-1}(A')$ need not be a set of $\mathscr{X}$ so that $m(g^{-1}(A'))$ could be undefined.) Measure m' is named the image measure of m through g.

Methods for introducing measures on measurable spaces

When defining a measure on a σ-algebra, it is sometimes convenient to first define the measure on a simpler system of sets and then to extend it to the σ-algebra.

A typical case is when the simpler system is an algebra $\mathscr{A}$ (an algebra is a system of sets with the same properties as for a σ-algebra - see Definition 1.1 - where, however, property (b) is only required to hold for a finite number of sets) and the σ-algebra is $\sigma(\mathscr{A})$, the smallest σ-algebra containing $\mathscr{A}$. This situation is studied in the fundamental Theorem 1.9 below.

Before stating the theorem, we need some terminology. Given an algebra $\mathscr{A}$, a function $m_0 : \mathscr{A} \to [0,\infty]$ satisfying (6.13) for all (possibly countably infinite) collection of pairwise disjoint sets $A_k \in \mathscr{A},\ k = 1,2,\ldots$, is called a premeasure (namely, a premeasure has identical properties as a measure but it is defined over an algebra instead of a σ-algebra.) The reason for calling it a premeasure is that it extends naturally to a measure, as Theorem 1.9 states. A premeasure on $\mathscr{A}$ is called σ-finite if there exits a sequence of sets $A_k \in \mathscr{A},\ k = 1,2,\ldots$, such that $\cup_{k=1}^{\infty} A_k = X$ (i.e. the entire set) and $m_0(A_k) < \infty, \forall k$.

THEOREM 1.9 (Caratheodory's) *Consider a σ-finite premeasure m_0 on an algebra $\mathscr{A}$. Then, there is one and only one measure m that extends m_0 to $\sigma(\mathscr{A})$, that is a measure on $\sigma(\mathscr{A})$ such that*

$$m(A) = m_0(A), \quad for\ A \in \mathscr{A}.$$

□

A proof can be found e.g. in the texts [3], [5], [1].

As an application of Caratheodory's theorem, we next introduce the notion of product measure space.

Product measure spaces

Let us consider the product $(X_1 \times X_2 \times \cdots \times X_n, \mathscr{X}_1 \otimes \mathscr{X}_2 \otimes \cdots \otimes \mathscr{X}_n)$ of n measurable spaces $(X_k, \mathscr{X}_k), k = 1,2,\ldots,n$. We assume that each space $(X_k, \mathscr{X}_k)$ is endowed with a σ-finite measure m_k and we want to introduce a product measure $m_1 \times m_2 \times \cdots \times m_n$ on the product space.

For notational convenience, assume $n = 2$, the extension to $n > 2$ is straightforward. Consider the system of subsets of $X_1 \times X_2$ of the form $A_1 \times A_2$, with $A_1 \in \mathscr{X}_1$ and $A_2 \in \mathscr{X}_2$. Such a system is not an algebra since it is not closed under union. However, it is easily seen that the system of subsets consisting of finite unions of disjoint sets of the form $A_1 \times A_2$ is indeed an algebra (though, not a σ-algebra.) For each set

$A = \cup_{k=1}^{p}(A_1^k \times A_2^k)$ of this algebra we define $m_0(A) = \sum_{k=1}^{p} m_1(A_1^k)m_2(A_2^k)$. It is a simple (but cumbersome) exercise to show that m_0 is a σ-finite premeasure. Thus, by the Caratheodory's extension Theorem 1.9, m_0 can be extended in a unique way to a measure on the σ-algebra $\mathscr{X}_1 \otimes \mathscr{X}_2$. This measure is by definition the product measure and is denoted by $m_1 \times m_2$.

DEFINITION 1.10 (product measure space) *Given n measure spaces* $(X_k, \mathscr{X}_k, m_k), k = 1, 2, \ldots, n$, *the measure space* $(X_1 \times X_2 \times \cdots \times X_n, \mathscr{X}_1 \otimes \mathscr{X}_2 \otimes \cdots \otimes \mathscr{X}_n, m_1 \times m_2 \times \cdots \times m_k)$ *is called their product measure space.* □

The measure spaces $(\mathbb{R}^{\mathbf{n}}, \mathscr{B}(\mathbb{R}^{\mathbf{n}}), \lambda^{\mathbf{n}})$

We start by considering the measure space $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \lambda)$, i.e. we take $n = 1$, a space that plays a prominent role in measure theory.

Here, $\mathbb{R}$ is the set of real numbers and $\mathscr{B}(\mathbb{R})$ is the σ-algebra generated by all open intervals (a, b). $\mathscr{B}(\mathbb{R})$ is named the Borel σ-algebra on the real line. Measure λ is the Lebesgue measure on the real line and is defined as follows. Consider the intervals of the form $(a, b] := \{x \in \mathbb{R} : a < x \leq b\}$, or $(-\infty, b] := \{x \in \mathbb{R} : x \leq b\}$, or $(a, \infty) := \{x \in \mathbb{R} : a < x\}$. Next, consider the system $\mathscr{A}$ of subsets A of $\mathbb{R}$ that are finite unions of disjoint intervals A_k, where each A_k has one of the indicated forms: $A = \cup_{k=1}^{p} A_k$. $\mathscr{A}$ is an algebra. Define $\lambda_0(A) := \sum_{k=1}^{p}(b_k - a_k)$, where a_k, b_k are the extremes of interval A_k. It can be seen that λ_0 is a σ-finite premeasure, so that, by the Caratheodory's Theorem 1.9, it can be extended in a unique way to a measure on $\sigma(\mathscr{A})$. Finally, it is an easy fact to prove that $\sigma(\mathscr{A}) = \mathscr{B}(\mathbb{R})$, so that the above construction has generated a measure on $\mathscr{B}(\mathbb{R})$, to which the name of Lebesgue measure λ is given.

For $n \geq 2$, $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \lambda^n)$ is simply defined as the n-fold product measure space of $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \lambda)$.

Sometimes, it is of interest to consider the restriction of $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \lambda)$ to a set in $\mathscr{B}(\mathbb{R})$. An example is $([0,1], \mathscr{B}[0,1], \lambda)$. Here, by definition, $\mathscr{B}[0,1]$ is the σ-algebra of all sets of the form $A \cap [0,1]$, $A \in \mathscr{B}(\mathbb{R})$ and λ is the restriction of the Lebesgue measure to these sets.

1.3 Integration

Given a measure space $(X, \mathscr{X}, m)$ and a $\mathscr{X}/\mathscr{B}(\mathbb{R})$-measurable function $g : X \to \mathbb{R}$, we want to define the integral of g with respect to the measure m, for which we use the symbol

$$\int_X g(x)dm(x).$$

This is done in 3 steps. For the first step we need the following definition.

DEFINITION 1.11 (simple measurable function) *Given a measurable space $(X, \mathscr{X})$, a measurable function $g : X \to \mathbb{R}$ is said to be simple if it has the form*

$$g = \sum_{k=1}^{N} \alpha_k \cdot 1(A_k), \quad A_k \in \mathscr{X}, \ \alpha_k \in \mathbb{R}, \quad k = 1, 2, \ldots, n, \tag{1.2}$$

where N is finite and $1(A_k)$ is the indicator function of set A_k. □

STEP 1: Integral of non-negative simple measurable functions.

Consider function g of the form in (1.2) and assume $\alpha_k \geq 0$, $k = 1, 2, \ldots, n$. Then, we define $\int_X g(x)dm(x) = \sum_{k=1}^{N} \alpha_k m(A_k)$ (if $\alpha_k = 0$ and $m(A_k) = \infty$, we let $\alpha_k m(\alpha_k) = 0 \cdot \infty = 0$).

STEP 2: Integral of non-negative measurable functions.

Let now $g : X \to \mathbb{R}$ be measurable and $g \geq 0$. Consider a sequence of simple measurable functions g_n such that $g_n(x) \uparrow g(x), \forall x$ (that is, for all x, $g_n(x)$ tends to $g(x)$ and is increasing with n; in words, it converges from below.) One such sequence is the following:

$$g_n := \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \cdot 1(A_{n,k}) + n \cdot 1(A_n),$$

where $A_{n,k} := \{x : (k-1)/2^n \leq g(x) < k/2^n\}$ and $A_n := \{x : g(x) \geq n\}$. Clearly, $\int_X g_n(x)dm(x)$ increases with n, so that it converges (either to a finite value or to $+\infty$.) We let $\int_X g(x)dm(x) := \lim_{n\to\infty} \int_X g_n(x)dm(x)$. One can prove that this definition is consistent, that is the limit is independent of the choice of the approximating sequence g_n.

STEP 3: Integral of measurable functions.

Let $g^+ := \max\{g, 0\}$ and $g^- := -\min\{g, 0\}$ and note that $g = g^+ - g^-$. By definition, the integral of g is given by the formula

$$\int_X g(x)dm(x) = \int_X g^+(x)dm(x) - \int_X g^-(x)dm(x),$$

provided that not both integrals in the right-hand-side are $+\infty$ (in which case we say that the integral is not defined.)

Notations

– When this generates no confusion, we drop the domain of integration and/or the arguments in the integral. So, for example, we write $\int g dm$ for $\int_X g(x)dm(x)$.

– When the integral is performed over $\mathbb{R}$ with respect to the Lebesgue measure λ, it is customary to write $\int_{\mathbb{R}} g(x)dx$ for $\int_{\mathbb{R}} g(x)d\lambda(x)$.

– Take a set $A \in \mathscr{X}$. Then, $g \cdot 1(A)$ (where $1(A)$ is the indicator function of set A) is measurable, provided that g is. We write $\int_A g dm$ for $\int_X g \cdot 1(A)dm$. When A is the interval $[a,b]$ and integration is with respect to the Lebesgue measure, we also write $\int_a^b g(x)dx$.

Functions with value in $\overline{\mathbb{R}} = [-\infty,\infty]$

It is often of interest to integrate functions taking value on the extended real line $[-\infty,\infty]$. let $\mathscr{B}(\overline{\mathbb{R}})$ be the σ-algebra generated by all intervals of the type (a,b), $[-\infty,b)$, and $(a,\infty]$ and consider a $\mathscr{X}/\mathscr{B}(\overline{\mathbb{R}})$-measurable function g. The definition of integral extends to this case with no modifications with the agreement that $0 \cdot \infty = 0$.

Properties of the integral

The following properties of the integral are easy to prove (all integrals appearing in the formulas are assumed to exist by hypothesis).

1. $\int 1(A)dm = m(A)$;
2. if $g_1 \leq g_2$, then $\int g_1 dm \leq \int g_2 dm$;
3. $\int (\alpha g_1 + \beta g_2)dm = \alpha \int g_1 dm + \beta \int g_2 dm$, provided that the right-hand-side does not result in the indeterminate form $\infty - \infty$;
4. if $g = 0$ m-almost surely (i.e. $m(\{g \neq 0\}) = 0$), then $\int g dm = 0$;
5. if $\int g \cdot 1(A)dm = 0, \forall A \in \mathscr{X}$, then $g = 0$ m-almost surely.

Change of space of integration

The following theorem permits to change space of integration in integrals.

THEOREM 1.12 (change of space of integration in integrals)
Consider the three measurable spaces $(X,\mathscr{X})$, $(X',\mathscr{X}')$, and $(\mathbb{R},\mathscr{B}(\mathbb{R}))$, and two measurable functions $g : X \to X'$ and $h : X' \to \mathbb{R}$. Further, assume that $(X,\mathscr{X})$ is endowed with a measure m. Then,

$$\int_X h(g(x))dm(x) = \int_{X'} h(x')dm'(x'), \tag{1.3}$$

where m' is the image measure of m through g, in the sense that if one integral exists, the other one also exists and the two are equal.

PROOF. For a simple non-negative function $h = \sum_{k=1}^{N} \alpha_k \cdot 1(A_k)$, we have

$$\begin{aligned} \int_X h(g(x))dm(x) &= \sum_{k=1}^{N} \alpha_k m(g^{-1}(A_k)) \\ &= \sum_{k=1}^{N} \alpha_k m'(A_k) \\ &= \int_{X'} h(x')dm'(x'). \end{aligned}$$

For a generic non-negative h, take a sequence of non-negative simple measurable functions h_n converging to h from below. Then, $h_n \cdot g$ is a sequence of non-negative simple measurable functions defined on X converging from below to $h \cdot g$ and, from what we have prove above,

$$\int_X h_n(g(x))dm(x) = \int_{X'} h_n(x')dm'(x'), \quad \forall n.$$

Since the limit of the left-hand-side as $n \to \infty$ is by definition $\int_X h(g(x))dm(x)$ and that of the right-hand-side is $\int_{X'} h_n(x')dm'(x')$, (1.3) is proven for non-negative h's.

Finally, for a generic h, the result follows from the usual decomposition $h = h^+ - h^-$. □

Integration with respect to a product measure

Suppose that $(X, \mathscr{X}, m) = (X_1 \times X_2, \mathscr{X}_1 \otimes \mathscr{X}_2, m_1 \times m_2)$. The following theorem permits to reduce the integral over $X_1 \times X_2$ to an iterated integral (the proof can be found e.g. in [2], [5], [7], [1].)

THEOREM 1.13 (Fubini's) *Let $(X_1, \mathscr{X}_1, m_1)$ and $(X_2, \mathscr{X}_2, m_2)$ be measure spaces with σ-finite measures m_1 and m_2 (a measure on $(X, \mathscr{X})$ is σ-finite if X is the countable union of sets $X_k \in \mathscr{X}$ with $m(X_k) < \infty$.) Further, let $g : X_1 \times X_2 \to \mathbb{R}$ be a $\mathscr{X}_1 \otimes \mathscr{X}_2$-measurable function and assume that*

$$\int_{X_1 \times X_2} |g| d(m_1 \times m_2) < \infty.$$

Then,

i) $g(x_1, x_2)$ is a $\mathscr{X}_1$-measurable function of x_1 for each fixed x_2 and a $\mathscr{X}_2$-measurable function of x_2 for each fixed x_1;
ii) the integral $\int_{X_1} g(x_1, x_2)dm_1$ is defined m_2-almost surely (i.e. the set where it is not

defined is in $\mathscr{X}_2$ and has m_2 measure zero.) Moreover, if we define $\int_{X_1} g(x_1,x_2)dm_1$ to be zero - or, equivalently, any other real number - where the integral is undefined, then $\int_{X_1} g(x_1,x_2)dm_1$ is a $\mathscr{X}_2$-measurable function. A similar statement holds for $\int_{X_2} g(x_1,x_2)dm_2$;
iii)

$$\begin{aligned}\int_{X_1\times X_2} g(x_1,x_2)d(m_1\times m_2) &= \int_{X_1}\left[\int_{X_2} g(x_1,x_2)dm_2\right]dm_1\\ &= \int_{X_2}\left[\int_{X_1} g(x_1,x_2)dm_1\right]dm_2,\end{aligned}$$

in the sense that the integral in the left-hand-side and the external integrals in the right-hand-side exist and equality holds.
The integral on the left is often referred to as the "double integral", while those on the right are called the "iterated integrals". □

If we assume that g is nonnegative, the same result as in Fubini's theorem holds without requiring that the integral of $|g|$ be bounded:

THEOREM 1.14 (Tonelli's) *Let $(X_1,\mathscr{X}_1,m_1)$ and $(X_2,\mathscr{X}_2,m_2)$ be measure spaces with σ-finite measures m_1 and m_2 (a measure on $(X,\mathscr{X})$ is σ-finite if X is the countable union of sets $X_k\in\mathscr{X}$ with $m(X_k)<\infty$.) Further, let $g: X_1\times X_2\to\mathbb{R}$ be a $\mathscr{X}_1\otimes\mathscr{X}_2$-measurable function with $g\geq 0$. Then,*

i) $g(x_1,x_2)$ is a $\mathscr{X}_1$-measurable function of x_1 for each fixed x_2 and a $\mathscr{X}_2$-measurable function of x_2 for each fixed x_1;
ii) $\int_{X_1} g(x_1,x_2)dm_1$ is a $\mathscr{X}_2$-measurable function and $\int_{X_2} g(x_1,x_2)dm_2$ is a $\mathscr{X}_1$-measurable function;
iii)

$$\begin{aligned}\int_{X_1\times X_2} g(x_1,x_2)d(m_1\times m_2) &= \int_{X_1}\left[\int_{X_2} g(x_1,x_2)dm_2\right]dm_1\\ &= \int_{X_2}\left[\int_{X_1} g(x_1,x_2)dm_1\right]dm_2.\end{aligned}$$

□

Chapter 2

RANDOM VARIABLES

2.1 Random variables

In this Chapter, we make continuous reference to notions like measure, measurable space, and integration that are discussed in Chapter 1.

DEFINITION 2.1 (probability and probability space) *Given a measurable space* $(\Omega, \mathscr{F})$*, a probability P is a measure on the* σ*-algebra* $\mathscr{F}$ *such that* $P(\Omega) = 1$*. The measure space* $(\Omega, \mathscr{F}, P)$ *is called a probability space.* □

DEFINITION 2.2 (random variable) *Given a probability space* $(\Omega, \mathscr{F}, P)$*, a* $\mathscr{F}/\mathscr{B}(\mathbb{R}^n)$*-measurable function* $v : \Omega \to \mathbb{R}^n$ *is called a n-dimensional random variable. In the case* $n = 1$*, we simply speak of a random variable.* □

From Theorem 1.6, it is clear that *n* random variables form a *n*-dimensional random variable and viceversa.

Interpretation and use of random variables

Random variables are used to model phenomena where a variable can take different real values in dependence of causes that one does not want to explicitly describe in the model. Corresponding to different points in Ω, *v* assumes different real values and one can ask the question: what is the probability that $v \in [a,b]$? This probability is computable from the model thanks to the assumption that *v* is $\mathscr{F}/\mathscr{B}(\mathbb{R})$-measurable: due to this assumption, the set of $\omega \in \Omega$ such that $v(\omega) \in [a,b]$ is an element of $\mathscr{F}$ and so probability $P(\omega : v(\omega) \in [a,b])$ is provided by the model. Thus, *v* is directly linked to the phenomenon under consideration and the value assumed by *v* describes

the particular occurrence of the phenomenon, while Ω and P are used as instruments to describe the likelihood with which different occurrences of the phenomenon take place.

More on $\mathscr{B}(\mathbb{R}^{\mathbf{n}})$ and the measurability of random variables

By definition, $\mathscr{B}(\mathbb{R})$ contains all open intervals (a,b). Since any open set on the real line is the countable union of open intervals, it is clear that $\mathscr{B}(\mathbb{R})$ contains all open sets and it is in fact the σ-algebra generated by open sets (in measure theory, the term "Borel σ-algebra" is assigned to the σ-algebra generated by open sets in a given topological space. Thus, $\mathscr{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$.) Similarly, $\mathscr{B}(\mathbb{R}^n)$ is the σ-algebra generated by the open sets in $\mathbb{R}^n$.

Suppose we want to prove that a given function $v : \Omega \to \mathbb{R}^n$ is a random variable, i.e. it is $\mathscr{F}/\mathscr{B}(\mathbb{R}^n)$-measurable. In principle, we have to show that the inverse image of any set $A \in \mathscr{B}(\mathbb{R}^n)$ is in $\mathscr{F}$. However, an easier test can be formulated: verify that the inverse image of just any open set is in $\mathscr{F}$. To see that this is enough, note that the system $\mathscr{D}$ of all sets A in $\mathbb{R}^n$ such that $v^{-1}(A) \in \mathscr{F}$ is a σ-algebra (indeed, if $A = \cup_{k=1}^{\infty} A_k$ with $A_k \in \mathscr{D}$, then $v^{-1}(A) = v^{-1}(\cup_{k=1}^{\infty} A_k) = \cup_{k=1}^{\infty} v^{-1}(A_k) \in \mathscr{F}$, so that $\mathscr{D}$ is closed under union. The fact that $\mathscr{D}$ is closed under complementation and contains the entire $\mathbb{R}^n$ can be proven in a similar way.) Now, since $\mathscr{B}(\mathbb{R}^n)$ is the smallest σ-algebra that contains the open sets, it is clear that $\mathscr{B}(\mathbb{R}^n) \subseteq \mathscr{D}$, showing that the inverse image of any set in $\mathscr{B}(\mathbb{R}^n)$ is in $\mathscr{F}$, that is the measurability of v.

Following the same reasoning, it is possible to conclude that v is measurable provided that the inverse image of any set in a system generating $\mathscr{B}(\mathbb{R}^n)$ is in $\mathscr{F}$. For example, for $n = 1$ this leads to the following test:

TEST OF MEASURABILITY 2.3 *$v : \Omega \to \mathbb{R}$ is measurable provided that $v^{-1}(a,b) \in \mathscr{F}$ for any $a, b \in \mathbb{R}$.* □

Suppose now that $f : \mathbb{R}^2 \to \mathbb{R}$ is continuous and v_1, v_2 are two random variables. Then, $f(v_1, v_2)$ is a random variable. So, e.g. $v_1 + v_2$, $v_1 \cdot v_2$, $sin(v_1 \cdot v_2)$ are random variables. To see this, recall the definition of a continuous function: f is continuous if the inverse image of any open set is an open set. Thus, if f is continuous, it is $\mathscr{B}(\mathbb{R}^2)/\mathscr{B}(\mathbb{R})$-measurable. Appealing to Theorem 1.5 on the measurability of composition functions, we then conclude that $f(v_1, v_2)$ is measurable. This fact extends in a natural way to functions $f : \mathbb{R}^n \to \mathbb{R}$.

Probability distribution function and probability density function

To make notations easier, we consider first the case of 1-dimensional random variables. The extension to the n-dimensional case is discussed later in this section.

DEFINITION 2.4 (probability distribution function) *The probability distribution function (or, more simply, the probability distribution or the distribution) of a 1-dimensional random variable v is the function $F : \mathbb{R} \to [0,1]$ defined as $F(x) = P(v^{-1}(-\infty, x])$.* □

The following properties of F are a direct consequence of the properties of P (the reader may want to try to detail a proof):

(a) $F(x)$ is nondecreasing;
(b) $F(x)$ is continuous on the right;
(c) $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

If we let P' be the image probability of P on $\mathscr{B}(\mathbb{R})$ through v, it is clear that $F(x) = P'(-\infty, x]$, so that the probability distribution can be calculated from the image probability. It is an important fact that the converse is also true: the image probability P' can be completely reconstructed from F. To see this, note that the system of subsets consisting of finite unions of disjoint sets of the form $(a,b]$, or $(-\infty,b]$, or (a,∞) is an algebra (let us call it $\mathscr{A}$) and an element $A := \cup_{k=1}^{n} A_k$ of this algebra has a probability that can be computed from F by the formula $P'(A) = \sum_{k=1}^{n} [F(b_k) - F(a_k)]$ (here, a_k, b_k are the extremes of interval A_k and $F(\infty)$ is short for $\lim_{x\to\infty} F(x) = 1$ and similarly for $F(-\infty)$.) Then, by virtue of the Caratheodory's Theorem 1.9, this probability can be extended in a unique way to $\sigma(\mathscr{A}) = \mathscr{B}(\mathbb{R})$, so reconstructing P'.

Next, we define the probability density function.

DEFINITION 2.5 (probability density function) *Suppose there exists a measurable function $p : \mathbb{R} \to \mathbb{R}$ such that $F(x) = \int_{-\infty}^{x} p(t)dt$, where F is the probability distribution function of a random variable v. Then, p is called the probability density function (or, more simply, the probability density or the density) of the random variable v.* □

Given a random variable $v : \Omega \to \mathbb{R}$ with distribution F and image probability P' and a measurable function $g : \mathbb{R} \to \mathbb{R}$, sometimes the notation $\int_{\mathbb{R}} g(x)dF(x)$ is used in place of $\int_{\mathbb{R}} g(x)dP'(x)$. It is easy to see that, if v admits density p, then $\int_{\mathbb{R}} g(x)dF(x)$ is also equal to $\int_{\mathbb{R}} g(x)p(x)dx$ in the sense that if one integral exists, also the other one exists and the two are equal (prove this by first considering simple non-negative g's, then non-negative g's and finally arbitrary measurable g's.)

From the definition of probability density, it follows that, if p exists, then F is the integral of p and is therefore an absolute continuous function. But then F is λ-almost surely differentiable and p is λ-almost surely the derivative of F (this is the "fundamental theorem of calculus", see e.g. Theorem 7.20 in [6]). Moreover, p is unique up

to a set of zero Lebesgue measure, that is, if p_1 and p_2 are two densities associated with the same distribution F, then $\lambda(x : p_1(x) \neq p_2(x)) = 0$. We refer to different densities as "versions" of the density and a phrase like "consider the density of v" means: "consider a version of the density of v".

It is also worth mentioning that the density needs not exist. An example is given by a random variable taking always value 0. In this case

$$F(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0, \end{cases}$$

and p clearly does not exist.

The above is an example of discrete distribution. In general, we can identify three classes of distributions (discrete, absolutely continuous, and singular) and it turns out that any distribution is the convex combination of elements of these classes. This classification is made explicit next.

1. DISCRETE DISTRIBUTIONS

A distribution F is discrete if it is piecewise constant, that is it is constant except for certain points where it is discontinuous. The density p is not defined for discrete distributions.

By the following simple argument, it is possible to see that the number of points of discontinuity is countable, that is they can be enumerated as $x_1, x_2, \ldots$. There can be at most one point of discontinuity with jump bigger than $1/2$ (with two jumps bigger than $1/2$, F would reach a value bigger than 1); similarly, there can be at most three points of discontinuity with jump in $(1/4, 1/2]$ and seven points with jump in $(1/8, 1/4]$ and so on. Summing up, the points of discontinuity can be at most countable.

2. ABSOLUTELY CONTINUOUS DISTRIBUTIONS

A distribution F is absolutely continuous if it has probability density.

Importantly, discrete and absolutely continuous distributions, or a combination of them, do not cover the set of all possibilities, that is not all F can be written as $F = \alpha F_d + (1-\alpha)F_{ac}$, for some $\alpha \in [0,1]$, with F_d discrete and F_{ac} absolutely continuous. A universal decomposition is obtained by the introduction of a a third type of distributions, called singular:

$$F = \alpha F_d + \beta F_{ac} + (1-\alpha-\beta)F_s, \tag{2.1}$$

where F_s is a singular distribution.

3. SINGULAR DISTRIBUTIONS

A distribution F is singular if it is continuous (and therefore any single point x has probability zero) but there exist a set of zero Lebesgue measure whose probability is 1.

Though singular distributions have a somehow peculiar behavior, examples of singular distributions are not difficult to construct, see e.g. [7]. Decomposition (2.1) is proven in many textbooks, among which [7]. □

More on probability distribution functions

Consider a probability measure P on $\mathbb{R}$. Following the discussion after Definition 2.4 (in the discussion after Definition 2.4 we had the symbol P' in place of P), we see that function $F(x)$, $x \in \mathbb{R}$, defined as $F(x) = P(-\infty, x]$ satisfies properties (a), (b), (c), which we report here for the reader's convenience:

(a) $F(x)$ is nondecreasing;
(b) $F(x)$ is continuous on the right;
(c) $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

Moreover, P can be reconstructed from F.

Let us now ask a converse question: is it true that to an F satisfying (a), (b), (c), there always corresponds a (unique) probability P such that $P(-\infty, x] = F(x)$? The answer is indeed positive.

THEOREM 2.6 *If $F(x)$, $x \in \mathbb{R}$, satisfies (a), (b), (c), then there exists a unique probability measure P on $\mathbb{R}$ such that $P(-\infty, x] = F(x)$, $\forall x \in \mathbb{R}$.* □

Thus, there is a one-to-one correspondence between functions $F(x)$ satisfying (a), (b), (c) and probability measures on $\mathbb{R}$.
Moreover, given a function $F(x)$ satisfying (a), (b), (c) and the corresponding probability measure P, the identity function $v(x) = x$ is a random variable between $(\mathbb{R}, \mathscr{B}(\mathbb{R}), P)$ and $\mathbb{R}$ with distribution $F(x)$. We thus see that any function $F(x)$ satisfying (a), (b), (c) is a distribution function and conditions (a), (b), (c) characterize the set of all distribution functions.

PROOF. Consider the system of subsets of $\mathbb{R}$ consisting of finite unions of disjoint sets of the form $(a, b]$, or $(-\infty, b]$, or (a, ∞) and note that it is an algebra (let us call it $\mathscr{A}$.) To ease the notation, we write $(a, \infty]$ for (a, ∞), so that all intervals are written as $(a, b]$ where a can possibly be $-\infty$ and b can possibly be $+\infty$. For any set in $\mathscr{A}$, define

$$P_0\left(\cup_{k=p}^{n}(a_k, b_k]\right) = \sum_{k=1}^{p}[F(b_k) - F(a_k)] \tag{2.2}$$

($F(\infty)$ is short for $\lim_{x\to\infty} F(x) = 1$ and similarly for $F(-\infty)$.) If we prove that P_0 is countably additive, then by Caratheodory's Theorem 1.9 it can be extended in a unique way to a measure P defined over $\sigma(\mathscr{A}) = \mathscr{B}(\mathbb{R})$. Moreover, for such a P the following holds:

- $P(-\infty, x] = F(x) - 0 = F(x)$;
- $P(-\infty, \infty) = F(\infty) - F(-\infty) = 1$, showing that P is a probability;
- no other probability $P_1 \neq P$ exists which satisfies $P_1(-\infty, x] = F(x)$, $x \in \mathbb{R}$. Indeed, the restriction of P_1 to $\mathscr{A}$ would satisfy (2.2) (with P_1 replacing P_0); since P also satisfies (2.2) (with P replacing P_0), by the uniqueness of the Caratheodory's extension we would have $P_1 = P$.

Thus, what is left to prove is that P_0 is countably additive on $\mathscr{A}$.

Let

$$A = \cup_{k=1}^{p}(a_k, b_k], \quad A_j = \cup_{k=1}^{p_j}(a_k^j, b_k^j], j = 1, 2, \ldots,$$

where the A_j's are disjoint and $\cup_{j=1}^{\infty} A_j = A$. Proving the countable additivity of P_0 amounts to show that

$$P_0(A) = \sum_{j=1}^{\infty} P_0(A_j). \tag{2.3}$$

Rewriting (2.3) as follows

$$0 = P_0(A) - \sum_{j=1}^{\infty} P_0(A_j) = P_0(A) - \lim_{m\to\infty} \sum_{j=1}^{m} P_0(A_j) = \lim_{m\to\infty} P_0(A - \cup_{j=1}^{m} A_j)$$

and observing that $A - \cup_{j=1}^{m} A_j =: B_m \downarrow \emptyset$ ("$\downarrow$" means that $B_m \supseteq B_{m+1}$, $m = 1, 2, \ldots$, and $\cap_{m=1}^{\infty} B_m = \emptyset$, the empty set), we see that (2.3) can be rewritten as

$$\lim_{m\to\infty} P_0(B_m) = 0, \quad \text{for any sequence } B_m \in \mathscr{A}, m = 1, 2, \ldots, \text{ such that } B_m \downarrow \emptyset. \tag{2.4}$$

The proof is now completed by showing the validity of (2.4).

Let us suppose first that $B_m \in [-M, M]$, $m = 1, 2, \ldots$, for some $M < \infty$. Since $F(x)$ is continuous on the right, the left extremes of the intervals forming B_m can be slightly moved to the right without a significant change of the F value. Therefore, we can find sets $C_m \in \mathscr{A}$ such that $closure(C_m) \subseteq B_m$ and $P_0(B_m - C_m) \leq \frac{1}{2^m}\varepsilon$, where $\varepsilon > 0$ is a preassigned small number. One fundamental property of the sets $closure(C_m)$, $m = 1, 2, \ldots$, is that the intersection of a finite number of them is already empty:

$$\cap_{m=1}^{r} closure(C_m) = \emptyset, \quad r < \infty. \tag{2.5}$$

The reason is that sets $[-M,M] - closure(C_m)$, $m = 1,2,\ldots$, form an open covering of $[-M,M]$. But, $[-M,M]$ is compact (by Heine-Borel theorem) so that a finite subcovering exists:

$$\cup_{m=1}^{r}([-M,M] - closure(C_m)) = [-M,M],$$

and this implies (2.5).

Thus,

$$\begin{aligned} P_0(B_r) &= P_0(B_r - \cap_{m=1}^{r} C_m) \quad (using\ (2.5)) \\ &\leq P_0(\cup_{m=1}^{r}(B_m - C_m)) \\ &\leq \sum_{m=1}^{r} P_0(B_m - C_m) \\ &\leq \sum_{m=1}^{r} \frac{1}{2^m}\varepsilon \\ &= \varepsilon, \end{aligned}$$

from which (2.4) follows.

Suppose now that sets B_m are not confined to an interval $[-M,M]$. Then, conclusion (2.4) can still be drawn by taking an interval $[-M,M]$ such that $F(-M)$ and $1-F(M)$ are smaller than ε and then following the same line of reasoning as before (details are left to the reader.) □

Multidimensional random variables

The notions of probability distribution and density can be extended with just some notational complications to multidimensional random variables. Here, it suffices to say that the probability distribution function is defined as

$$\begin{aligned} F(x_1,x_2,\ldots,x_n) &= P(v_1^{-1}(-\infty,x_1] \cap v_2^{-1}(-\infty,x_2] \cap \ldots \cap v_n^{-1}(-\infty,x_n]) \\ &= P'((-\infty,x_1] \times (-\infty,x_2] \times \cdots \times (-\infty,x_n]), \end{aligned}$$

where v_k is the k-th component of the n-dimensional random variable v and P' is the image probability. Again, it is possible to see that F uniquely defines P'.

In the multidimensional case, the probability distribution has the following properties that extend those valid for the 1-dimensional case:

(a) Given two points $x^{(1)} = (x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)})$ and $x^{(2)} = (x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)})$ in $\mathbb{R}^n$ with $x_1^{(2)} \geq x_1^{(1)}, x_2^{(2)} \geq x_2^{(1)}, \ldots, x_n^{(2)} \geq x_n^{(1)}$, and a function $f : \mathbb{R}^n \to \mathbb{R}$, introduce the notation $\Delta_k^{x_k^{(1)}, x_k^{(2)}} f := f(x_1, \ldots, x_{k-1}, x_k^{(2)}, x_{k+1}, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, x_k^{(1)}, x_{k+1}, \ldots, x_n)$. Then, it is a matter of a cumbersome computation to show that $\Delta_1^{x_1^{(1)}, x_1^{(2)}} (\Delta_2^{x_2^{(1)}, x_2^{(2)}} (\cdots (\Delta_n^{x_n^{(1)}, x_n^{(2)}} F))) = P'((x_1^{(1)}, x_1^{(2)}] \times (x_2^{(1)}, x_2^{(2)}] \times \cdots \times (x_n^{(1)}, x_n^{(2)}])$. Since the left-hand-side is clearly nonnegative, we then have

$$\Delta_1^{x_1^{(1)}, x_1^{(2)}} \cdots \Delta_n^{x_n^{(1)}, x_n^{(2)}} F \geq 0. \tag{2.6}$$

To help visualize the situation, for $n = 2$ we have $\Delta_1^{x_1^{(1)}, x_1^{(2)}} (\Delta_2^{x_2^{(1)}, x_2^{(2)}} F) = F(x_1^{(2)}, x_1^{(2)}) - F(x_1^{(1)}, x_2^{(2)}) - F(x_1^{(2)}, x_2^{(1)}) + F(x_1^{(1)}, x_2^{(1)})$ and it represents the measure P' of the rectangle in Figure 2.1.

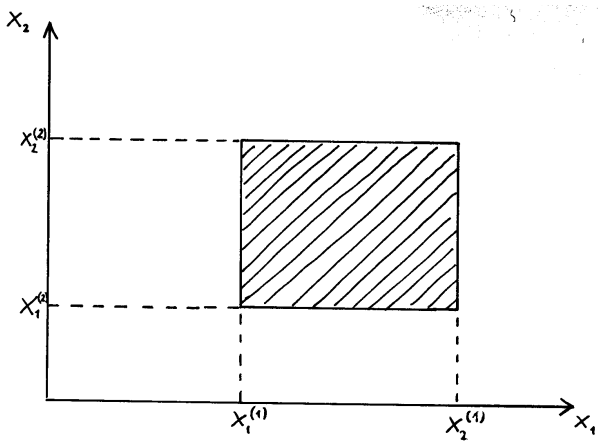

Figure 2.1: $\Delta_1^{x_1^{(1)}, x_1^{(2)}} (\Delta_2^{x_2^{(1)}, x_2^{(2)}} F)$ is the measure P' of the rectangle.

For $n = 1$, (2.6) is equivalent to say that $F(x)$ is nondecreasing;
(b) $F(x_1, x_2, \ldots, x_n)$ is continuous on the right in the sense that if $x_k \downarrow \bar{x_k}$, $k = 1, 2, \ldots, n$, then $F(x_1, x_2 \ldots, x_n) \to F(\bar{x_1}, \bar{x_2} \ldots, \bar{x_n})$;
(c) if $x_k \to \bar{x_k}$, $k = 1, 2, \ldots, n$, and at least one of the $\bar{x_k}$'s is $-\infty$, then $F(x_1, x_2, \ldots, x_n) \to 0$. Moreover, $\lim_{x_1 \to \infty, \ldots, x_n \to \infty} F(x_1, x_2, \ldots, x_n) = 1$. □

A n-dimensional probability density function is a nonnegative measurable function p such that $F(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} [\int_{-\infty}^{x_2} \cdots [\int_{-\infty}^{x_n} p(t_n, t_2, \ldots, t_n) dt_n] \cdots dt_2] dt_1$.

Expectation, variance and moments

Given a random variable v, the integral $\int_\Omega v dP$ (assuming it is defined) is called the expectation of v and is also written $E[v]$. Other significant integral characteristics of

v are its moment or order r : $E[v^r]$ (that is the expectation of the random variable obtained by the composition of v with the $\mathbb{R} \to \mathbb{R}$ function of elevation to the r-th power), and its variance: $E[(v - E[v])^2]$. To indicate the variance of v, the symbol $var(v)$ is also used.

When v is a matrix of random variables with entries v_{kj}, by the symbol $E[v]$ we mean the matrix with entries $E[v_{kj}]$. If v is a n-dimensional random variable, $E[v]$ is the vector listing the expectation of its components. Similarly, $var(v)$ is a matrix with entries $E[(v_j - E[v_j])(v_k - E[v_k])^T]$, where v_j, v_k are the components of v.

Note that the expectation and the other integral quantities can be computed in different ways. For example, letting $F^{(v)}$ be the probability distribution of v and $F^{(v^2)}$ that of v^2, we have

$$E[v^2] = \int_\Omega v^2 dP = \int_{\mathbb{R}} x^2 dF^{(v)}(x) = \int_{\mathbb{R}} x dF^{(v^2)}(x),$$

where the last two equalities are justified in the light of Theorem 1.12.

2.2 Independence and incorrelation

We start by considering two 1-dimensional random variables v_1 and v_2.

DEFINITION 2.7 (independence) *We say that v_1 and v_2 are independent if*

$$P(v_1^{-1}(A_1) \cap v_2^{-1}(A_2)) = P(v_1^{-1}(A_1)) \cdot P(v_2^{-1}(A_2)), \quad \forall A_1, A_2 \in \mathscr{B}(\mathbb{R}). \tag{2.7}$$

□

The interpretation is that, if v_1 and v_2 are independent, then the probability that they simultaneously taken on value in given ranges A_1 and A_2 equals the probabilities that the first one takes value in A_1 times the probability that the second one takes on value in A_2.

If v_1 and v_2 are independent, so are $f(v_1)$ and $g(v_2)$, with f and g arbitrary measurable functions (show this.)

Let $P_1^{'}$ and $P_2^{'}$ be the image probabilities on $\mathbb{R}$ induced by v_1 and v_2, respectively. Also, consider the 2-dimensional random variable defined through relation $v = (v_1, v_2)$ and let $P^{'}$ be the corresponding image probability on $\mathbb{R}^2$. It turns out that $P^{'} = P_1^{'} \times P_2^{'}$. To prove this, it suffices to show that $P^{'}$ and $P_1^{'} \times P_2^{'}$ agrees over the algebra of finite unions of disjoint sets of the form $A_1 \times A_2$ with $A_1, A_2 \in \mathscr{B}(\mathbb{R})$. In fact, by the Caratheodory's

theorem 1.9 we then have that the extension to $\mathscr{B}(\mathbb{R}^2)$ is unique and, therefore, coincident. Take $A = \cup_{k=1}^{p}(A_1^k \times A_2^k)$. We have: $P^{'}(A) = P^{'}(\cup_{k=1}^{p}(A_1^k \times A_2^k)) = \sum_{k=1}^{p} P^{'}(A_1^k \times A_2^k) =$ [due to the independence of v_1 and v_2] $= \sum_{k=1}^{p} P_1^{'}(A_1^k) \cdot P_2^{'}(A_2^k) = (P_1^{'} \times P_2^{'})(A)$, so that P' and $P_1^{'} \times P_2^{'}$ indeed agree over the considered algebra.

Letting F_1 and F_2 be the probability distributions of v_1 and v_2 and F that of v, as a direct consequence of (2.7) we have that $F(x_1,x_2) = F_1(x_1) \cdot F_2(x_2)$. Moreover, if F_1 and F_2 admit density, say p_1 and p_2, we then have that v has density too and it is given by $p(x_1,x_2) = p_1(x_1) \cdot p_2(x_2)$, as it is shown by direct inspection:

$$\begin{aligned}
\int_{-\infty}^{x_1}\left[\int_{-\infty}^{x_2} p_1(t_1)p_2(t_2)dt_2\right]dt_1 &= \int_{-\infty}^{x_1} p_1(t_1)\left[\int_{-\infty}^{x_2} p_2(t_2)dt_2\right]dt_1 \\
&= \int_{-\infty}^{x_1} p_1(t_1)F_2(x_2)dt_1 \\
&= F_1(x_1)F_2(x_2) \\
&= F(x_1,x_2).
\end{aligned}$$

The converse also holds true: if $F(x_1,x_2) = F_1(x_1) \cdot F_2(x_2)$, or $p(x_1,x_2) = p_1(x_1) \cdot p_2(x_2)$, then v_1 and v_2 are independent (providing details is a useful exercise.)

DEFINITION 2.8 (incorrelation) *We say that v_1 and v_2 are uncorrelated if $E[v_1v_2]$, $E[v_1]$ and $E[v_2]$ exist finite and*

$$E[v_1v_2] = E[v_1] \cdot E[v_2]. \tag{2.8}$$

□

Incorrelation is an integral notion. Not surprisingly, independence is a stronger notion than incorrelation and the former implies the latter, while the opposite is in general false. To state the implication between independence and incorrelation precisely, suppose that v_1 and v_2 are independent and that $E[v_1v_2]$, $E[v_1]$ and $E[v_2]$ exist finite; then, v_1 and v_2 are uncorrelated, as the following calculation shows:

$$\begin{aligned}
E[v_1 v_2] &= \int_\Omega v_1(\omega) v_2(\omega) dP(\omega) \\
&= \int_{\mathbb{R}^2} xy \, dP'(x,y) \quad (\textit{using Theorem } 1.12) \\
&= \int_{\mathbb{R}^2} xy \, d(P_1' \times P_2')(x,y) \\
&= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} xy \, dP_1'(x) \right] dP_2'(y) \quad (\textit{using Theorem } 1.13) \\
&= \left[\int_{\mathbb{R}} x \, dP_1'(x) \right] \left[\int_{\mathbb{R}} y \, dP_2'(y) \right] \\
&= E[v_1] E[v_2].
\end{aligned}$$

An example of two random variables that are uncorrelated but not independent is shown in Figure 2.2.

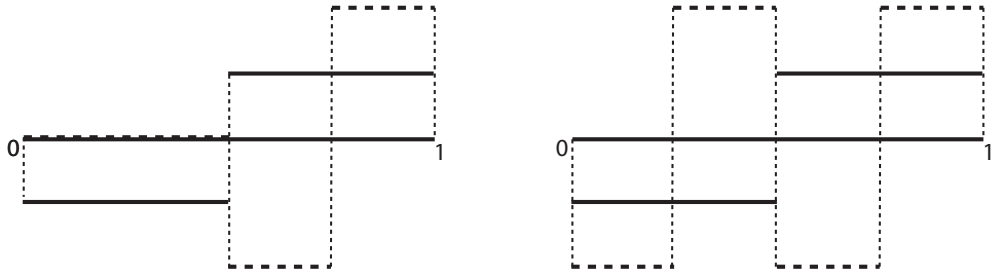

Figure 2.2: Probability space: $(\Omega, \mathscr{F}, P) = ([0,1], \mathscr{B}[0,1], \lambda)$. Solid line = v_1; dashed line = v_2. Left: v_1 and v_2 are uncorrelated, but not independent; right: v_1 and v_2 are independent, and therefore also uncorrelated.

The notions of independence and incorrelation carry over to the multidimensional case in a straightforward way. Given $v_1 : \Omega \to \mathbb{R}^{n_1}$ and $v_2 : \Omega \to \mathbb{R}^{n_2}$, we say that v_1 is independent of v_2 if $P(v_1^{-1}(A_1) \cap v_2^{-1}(A_2)) = P(v_1^{-1}(A_1)) \cdot P(v_2^{-1}(A_2)), \forall A_1 \in \mathscr{B}(\mathbb{R}^{n_1}), A_2 \in \mathscr{B}(\mathbb{R}^{n_2})$. Note that this definition only establish a cross-property of v_1 and v_2; different components of e.g. v_1 can well be dependent one on the others. By identifying v_1 and v_2 with the vectors of their components, we say that v_1 and v_2 are uncorrelated if $E[v_1 v_2^T] = E[v_1] E[v_2^T]$.

2.3 Characteristic functions

The method of characteristic functions is one of the main tools in probability theory. Though a characteristic function carries exactly the same information content as the

corresponding probability distribution, in many contexts it is more handy to use than the distribution itself. Here, we merely define characteristic functions and derive some basic properties of use in the book. The interested reader is referred to textbooks on probability for a broader treatment.

A complex-valued random variable v is, by definition, given by $v = v_{\mathbb{R}} + iv_{\mathbb{I}}$, where $v_{\mathbb{R}}$ and $v_{\mathbb{I}}$ are real-valued random variables. We also let $E[v] = E[v_{\mathbb{R}}] + iE[v_{\mathbb{I}}]$.

DEFINITION 2.9 (characteristic function) *The characteristic function of a random variable v is defined as $\varphi(t) := E[e^{itv}]$, $t \in \mathbb{R}$.* □

Fort a given t, $E[e^{itv}]$ is a complex number; as t varies over $\mathbb{R}$, $\varphi(t) = E[e^{itv}]$ is a complex-valued function. It is clear that $\varphi(t)$ can also be expressed as $\varphi(t) = \int_{\mathbb{R}} e^{itx} dF(x)$, where F is the distribution of v. Thus, φ is determined by F. It is a crucial fact that the converse is also true: F can be completely reconstructed from φ, as the next theorem states.

THEOREM 2.10 *Let F and G be probability distributions on $\mathbb{R}$ with the same characteristic function, viz.*

$$\int_{\mathbb{R}} e^{itx} dF(x) = \int_{\mathbb{R}} e^{itx} dG(x), \quad \forall t \in R. \tag{2.9}$$

Then, $F(x) = G(x)$, $\forall x \in \mathbb{R}$.

PROOF. Given arbitrary $\alpha, \beta \in \mathbb{R}$ and $\varepsilon > 0$, consider the function f^{ε} in Figure 2.3.

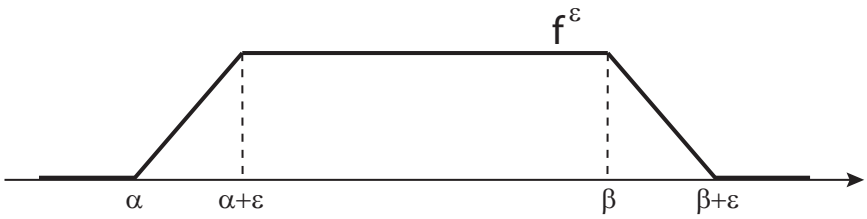

Figure 2.3:

We first prove that

$$\int_{\mathbb{R}} f^{\varepsilon} dF = \int_{\mathbb{R}} f^{\varepsilon} dG, \tag{2.10}$$

and then show that the theorem thesis follows from this equality.

Consider a sequence of real numbers $\rho_n \downarrow 0$ and pick a n large enough so that $[\alpha, \beta + \varepsilon] \subseteq [-n, n]$. In $[-n, n]$, f^ε can be uniformly approximated by a finite trigonometric sum (Weierstrass's theorem). Precisely, there exists a function

$$f_n^\varepsilon(x) := \sum_{k=-N(n)}^{N(n)} a_k e^{i\pi \frac{k}{n} x},$$

where a_k are complex coefficients, such that

$$\sup_{-n \le x \le n} |f^\varepsilon(x) - f_n^\varepsilon(x)| \le \rho_n. \tag{2.11}$$

Observe also that function $f_n^\varepsilon(x)$ is periodic so that

$$\sup_x |f_n^\varepsilon(x)| = \sup_{-n \le x \le n} |f_n^\varepsilon(x)| \le 1 + \rho_n, \tag{2.12}$$

and that, by (2.9),

$$\int_{\mathbb{R}} f_n^\varepsilon dF = \int_{\mathbb{R}} f_n^\varepsilon dG. \tag{2.13}$$

Thus,

$$\begin{aligned}
&\left| \int_{\mathbb{R}} f^\varepsilon dF - \int_{\mathbb{R}} f^\varepsilon dG \right| \\
&= \left| \int_{[-n,n]} f^\varepsilon dF - \int_{[-n,n]} f^\varepsilon dG \right| \\
&\le \left| \int_{[-n,n]} f_n^\varepsilon dF - \int_{[-n,n]} f_n^\varepsilon dG \right| + 2\rho_n \quad (\textit{using } (2.11)) \\
&\le \left| \int_{\mathbb{R}} f_n^\varepsilon dF - \int_{\mathbb{R}} f_n^\varepsilon dG \right| + (1+\rho_n) \int_{[-n,n]^c} dF + (1+\rho_n) \int_{[-n,n]^c} dG + 2\rho_n \\
&\qquad\qquad (\textit{using } (2.12); [-n,n]^c = \textit{complement of } [-n,n]) \\
&\le (1+\rho_n) \int_{[-n,n]^c} dF + (1+\rho_n) \int_{[-n,n]^c} dG + 2\rho_n \quad (\textit{using } (2.13)).
\end{aligned}$$

The right-hand-side tends to zero as $n \to \infty$. Since the left-hand-side does not depend on n, it must then be equal to zero and (2.10) is proven.

We turn now to prove that (2.10) implies that $F = G$. As $\varepsilon \to 0$, we have $\int_{\mathbb{R}} f^\varepsilon dF \to F(\beta) - F(\alpha)$ and $\int_{\mathbb{R}} f^\varepsilon dG \to G(\beta) - G(\alpha)$. Hence, from (2.10), $F(\beta) - F(\alpha) =$

$G(\beta) - G(\alpha)$. Letting $\alpha \to -\infty$, we conclude that $F(\beta) = G(\beta)$, $\forall \beta \in \mathbb{R}$. □

Now, the question we want to address is : why are characteristic functions so useful? One reason is that characteristic functions allow one to deal more easily with independence, since the characteristic function of the sum of two independent random variables is simply given by the product of the characteristic functions of the two random variables:

- *If v_1 and v_2 are independent, then $\varphi^{(v_1+v_2)}(t) = \varphi^{(v_1)}(t) \cdot \varphi^{(v_2)}(t)$.*

To see this, write: $\varphi^{(v_1+v_2)}(t) = E[e^{it(v_1+v_2)}] = E[e^{itv_1}e^{itv_2}] = E[e^{itv_1}]E[e^{itv_2}] = \varphi^{(v_1)}(t)\varphi^{(v_2)}(t)$. So, when dealing with independent variables, we can move from distributions (where independence translates into the awkward condition that the distribution of $v_1 + v_2$ is the convolution of the distributions of v_1 and v_2) to characteristic functions and use the handy product rule. In doing so, no information is lost, as the distribution can be reconstructed from the characteristic function, as stated in Theorem 2.10.

We know that $\varphi_1(t) = \varphi_2(t)$ implies $F_1(x) = F_2(x)$. Now, we ask: suppose that $\varphi_n(t) \to \varphi(t)$; is it true that $F_n(x) \to F(x)$? A precise answer is given by the following theorem, which plays an important role in proving limit results in probability theory.

THEOREM 2.11 *Let F_n be a sequence of probability distributions on $\mathbb{R}$ and let φ_n be the corresponding sequence of characteristic functions.*

(a) If $F_n \to F$ weakly, where F is a probability distribution with characteristic function φ (see Section 3.4 for the notion of weak convergence), then $\varphi_n(t) \to \varphi(t)$, $\forall t \in \mathbb{R}$;

(b) if $\varphi_n(t) \to \varphi(t)$, $\forall t \in \mathbb{R}$, and $\varphi(t)$ is continuous at $t = 0$, then $\varphi(t)$ is a characteristic function (i.e. $\varphi(t) = \int_{\mathbb{R}} e^{itx} dF(x)$ for some probability distribution F) and $F_n \to F$ weakly.

PROOF.
(a) Write $\varphi_n(t) = \int_{\mathbb{R}} (\cos(tx) + i\sin(tx)) dF_n(x)$. The weak convergence of $F_n \to F$ means that $\int_{\mathbb{R}} f(x) dF_n(x) \to \int_{\mathbb{R}} f(x) dF(x)$, for any continuous and bounded $f(x)$. The thesis then follows by taking in turn $f(x) = \cos(tx)$ and $f(x) = \sin(tx)$.

(b) The proof proceeds as follows. Thanks to the continuity of $\varphi(t)$ at $t = 0$, we prove that F_n is tight (see Theorem 3.23 for the definition of tightness.) Due to tightness, by Theorem 3.23, F_n admits a subsequence weakly convergent to some F and this F has

$\varphi(t)$ as characteristic function. Finally, by the convergence $\varphi_n(t) \to \varphi(t)$ we establish that the whole sequence F_n converges to F.

To prove the tightness of F_n, pick any real M and let $\beta := \inf_{|\alpha|\geq 1}\left(1-\frac{\sin\alpha}{\alpha}\right) \geq \frac{1}{7}$. We have

$$\begin{aligned}
\beta\int_{|x|\geq M} dF_n(x) &= \inf_{|\alpha|\geq 1}\left(1-\frac{\sin\alpha}{\alpha}\right)\int_{|x|\geq M} dF_n(x)\\
&\leq \int_{|x|\geq M}\left(1-\frac{\sin(x/M)}{x/M}\right) dF_n(x)\\
&\leq \int_{\mathbb{R}}\left(1-\frac{\sin(x/M)}{x/M}\right) dF_n(x)\\
&= \int_{\mathbb{R}}\left[M\int_0^{1/M}(1-\cos(tx))dt\right] dF_n(x)\\
&= M\int_0^{1/M}\left[\int_{\mathbb{R}}(1-\cos(tx))dF_n(x)\right] dt\\
&\quad (\textit{using Fubini's Theorem } 1.13)\\
&= M\int_0^{1/M}(1-Re(\varphi_n(t)))\, dt\\
&\overset{n\to\infty}{\longrightarrow} M\int_0^{1/M}(1-Re(\varphi(t)))\, dt
\end{aligned}$$

(*using a slight variant of the dominated convergence Theorem* 3.8)

The right-hand-side represents the mean value of $1-Re(\varphi(t))$ in a right neighborhood of the origin. Since $\varphi(0) = E[e^{i0x}] = 1$ and $\varphi(t)$ is continuous at $t = 0$, we have

$$M\int_0^{1/M}(1-Re(\varphi(t)))dt \to 0, \quad as\ M\to\infty. \tag{2.14}$$

We show that (2.14) implies the tightness of F_n. Indeed, given an arbitrarily small $\varepsilon > 0$, take $M(\varepsilon)$ such that $M(\varepsilon)\int_0^{1/M(\varepsilon)}(1-Re(\varphi(t)))dt \leq \frac{\varepsilon}{2}$. Then, by (2.14), $\beta\int_{|x|\geq M(\varepsilon)} dF_n(x) \leq \varepsilon$ for any n large enough, say $n \geq n(\varepsilon)$. Since $\beta\int_{|x|\geq M} dF_n(x)$ is decreasing with M, we then have $\sup_{n\geq n(\varepsilon)}\beta\int_{|x|\geq M} dF_n(x) \leq \varepsilon$, as $M\to\infty$, and, by the arbitrariness of ε, we obtain that $\sup_{n\geq n(\varepsilon)}\beta\int_{|x|\geq M} dF_n(x) \to 0$, as $M\to\infty$. On the other hand, over the finite set of integers n with $n < n(\varepsilon)$, we have $\max_{n<n(\varepsilon)}\beta\int_{|x|\geq M} dF_n(x) \to 0$, as $M\to\infty$. Putting together these two facts yields: $\sup_n \beta\int_{|x|\geq M} dF_n(x) \to 0$, as $M\to\infty$, i.e. the tightness of F_n.

Having proven the tightness of F_n, appeal now to Helly's Theorem 3.23 to conclude that a subsequence F_{n_k} of F_n exists converging weakly to some limit distribution F.

Since $F_{n_k} \to F$ weakly, from part (a) of this theorem, we also have that $\varphi_{n_k}(t)$ tends for every t to the characteristic function of F. But, by assumption, $\varphi_{n_k}(t) \to \varphi(t)$, $\forall t \in \mathbb{R}$, so that $\varphi(t)$ must be the characteristic function of F.

We conclude the proof by showing that the whole sequence $F_n \to F$ weakly. Suppose not. Then, there exist a subsequence $F_{n_k^{(1)}}$ and a continuous and bounded $f : \mathbb{R} \to \mathbb{R}$ such that, for some $\varepsilon > 0$,

$$\left| \int_{\mathbb{R}} f dF_{n_k^{(1)}} - \int_{\mathbb{R}} f dF \right| \geq \varepsilon, \quad k = 1, 2, \ldots. \tag{2.15}$$

But $F_{n_k^{(1)}}$ is tight (being a subsequence of F_n), so that again by Helly's theorem there is a subsequence of indeces $\{n_k^{(2)}\} \subset \{n_k^{(1)}\}$ such that $F_{n_k^{(2)}}$ converges weakly to some distribution Q. Certainly, $Q \neq F$, since, otherwise, (2.15) would be violated. Now, $\varphi_{n_k}(t) \to \varphi^{(F)}(t)$, the characteristic function of F, and $\varphi_{n_k^{(2)}}(t) \to \varphi^{(Q)}(t)$, where $\varphi^{(F)}(t) \neq \varphi^{(Q)}(t)$ since $F \neq Q$. These two convergence are contradictory since, by hypothesis, the whole $\varphi_n(t)$ sequence converges to the same limiting function $\varphi(t)$. Thus, $F_n \to F$ weakly and this completes the proof. □

EXAMPLE 2.12 *It is possible that $\varphi_n(t) \to \varphi(t)$, $\forall t \in \mathbb{R}$, but $\varphi(t)$ is not continuous at $t = 0$, $\varphi(t)$ is not a characteristic function and F_n is not weakly convergent to any F. The reader can gain insigt in this fact by considering $F_n =$ uniform distribution on $[-n, n]$, whose characteristic function is*

$$\varphi_n(t) = \begin{cases} 1, & t = 0 \\ \frac{1}{nt}\sin(tn), & t \neq 0. \end{cases}$$

Here, $\varphi_n(t) \to \varphi(t)$ with

$$\varphi(t) = \begin{cases} 1, & t = 0 \\ 0, & t \neq 0, \end{cases}$$

but $\varphi(t)$ is not a characteristic function (as it can be easily shown, a characteristic function has to be continuous) and F_n does not converge weakly to any F.

Thus, $\varphi_n(t) \to \varphi(t)$ is not sufficient to conclude that F_n converges weakly to some F. However, part (b) of Theorem 2.11 tells us that whenever convergence fails, $\varphi(t)$ has to be discontinuous in 0. □

2.4 Gaussian random variables

A n-dimensional random variable is "Gaussian" (or "normal") if it has probability density

$$p(x) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} e^{-\frac{1}{2}(x-m)^T V^{-1}(x-m)},$$

where $x = [x_1\ x_2\ \cdots\ x_n]^T \in \mathbb{R}^n$, $m \in \mathbb{R}^n$, $V \in \mathbb{R}^{n\times n}$ is symmetric and positive definite (we write $V \succ 0$), and $|\cdot|$ indicates determinant.

It can be computed that

$$\begin{aligned} m &= E[v]; \\ V &= var(v). \end{aligned}$$

Thus, the density of a Gaussian random variable is fully described by its mean and its variance.

Gaussian random variables have notable properties, as listed below.

(i) If the n-dimensional random variable v is Gaussian, then, given a matrix $A \in \mathbb{R}^{m\times n}$ such that $AA^T \succ 0$, Av is Gaussian too.

This can be proven by a direct computation, which we here omit. Condition $AA^T \succ 0$ prevents Av from concentrating in a subspace of $\mathbb{R}^m$, in which case the density of Av does not exist. Also, we have: $E[Av] = AE[v] = Am$ and $var(Av) = E[(Av - Am)(Av - Am)^T] = AE[(v-m)(v-m)^T]A^T = AVA^T$.

(ii) Incorrelation implies independence.

Suppose that the variance matrix V has the form

$$V = \begin{bmatrix} V_{11} & 0 \\ 0 & V_{22} \end{bmatrix},$$

where V_{11} and V_{22} are matrices of size $n_1 \times n_1$ and $n_2 \times n_2$, respectively, that is v is formed by two uncorrelated components $v_1 \in \mathbb{R}^{n_1}$ and $v_2 \in \mathbb{R}^{n_2}$. Then, by splitting $x - m$ into two components of suitable dimensions, we have

$$\begin{aligned} p(x) &= \frac{1}{(2\pi)^{n/2}|V|^{1/2}} e^{-\frac{1}{2}(x-m)^T V^{-1}(x-m)} \\ &= \frac{1}{(2\pi)^{n_1/2}|V_{11}|^{1/2}(2\pi)^{n_2/2}|V_{22}|^{1/2}} e^{-\frac{1}{2}[(x_1-m_1)^T V_{11}^{-1}(x_1-m_1)+(x_2-m_2)^T V_{22}^{-1}(x_2-m_2)]} \\ &= p_1(x_1)p_2(x_2), \end{aligned}$$

where $p_1(x_1)$ and $p_2(x_2)$ are the densities of v_1 and v_2, so proving that v_1 and v_2 are independent.

The above two properties are certainly one reason of the success of Gaussian variables: because of these properties, many statistical problems find an easier solution within the Gaussian framework. A second reason of success is that Gaussian variables provide a universal paradigm for the description of natural phenomena involving many stochastic sources, a fact that has a theoretical foundation in the central limit Theorem (see Section 3.3.)

Thus far, we have considered Gaussian variables having a positive definite variance V. It is sometimes convenient to have at our disposal a more general definition that allows for positive semidefinite variance as well. In this case, however, a density does not exist and we have to move to probability distributions or – as we prefer to do – to characteristic functions.

Let us start by observing that the characteristic function of a Gaussian variable with mean m and variance $V \succ 0$ is given by (we omit the lengthy and conceptually uninteresting derivation):

$$\varphi(t_1,t_2,\ldots,t_n) := E[e^{it^T v}] = e^{it^T m - \frac{1}{2}t^T V t}$$

(in fact we have referred here to the characteristic function of a multidimensional random variable, a definition which naturally extends that valid for the 1-dimensional case. Similarly to Theorem 2.10, multidimensional distributions are in a 1-to-1 correspondence with multidimensional characteristic functions.) Suppose now that V is only positive semidefinite and consider again expression

$$e^{it^T m - \frac{1}{2}t^T V t}. \tag{2.16}$$

(2.16) still identifies a characteristic function (i.e. it equals $\int_{\mathbb{R}} e^{it^T x} dF(x)$ for some n-dimensional probability distribution F.) To show this, consider

$$e^{it^T m - \frac{1}{2}t^T (V + \frac{1}{n}I) t}, \tag{2.17}$$

I being the identity matrix. Since $V + \frac{1}{n}I \succ 0$, (2.17) is the characteristic function $\varphi_n(t)$ of a Gaussian distribution $G(m, V + \frac{1}{n}I)$. When we let $n \to \infty$, (2.17) $\to$ (2.16), and the limit $\varphi(t) =$ (2.16) is continuous at $t = 0$. Then, by appealing to Theorem 2.11 (actually, to an extension of this theorem to multidimensional distributions), we conclude that (2.17) is indeed the characteristic function of some distribution. This justifies the following definition.

DEFINITION 2.13 (Gaussian random variable)
A n-dimensional random variable is "Gaussian" (or "normal") if its characteristic function is

$$\varphi(t_1,t_2,\ldots,t_n) := E[e^{it^T v}] = e^{it^T m - \frac{1}{2}t^T V t},$$

where $t = [t_1\ t_2\ \cdots\ t_n]^T, m \in \mathbb{R}^n$, $0 \preceq V \in \mathbb{R}^{n\times n}$. □

2.5 Computing the density induced by a function

Sometimes, it is necessary to compute the density of a random variable obtained by applying a function f to another random variable whose density is known. The following theorem provides an answer to this problem.

THEOREM 2.14 (density induced by a function) *Consider a n-dimensional random variable v with density p. Given a function* $f : \mathbb{R}^n \to \mathbb{R}^n$ *such that: i)* f *is* 1*-to-*1*; and ii)* $g = f^{-1}$ *is everywhere differentiable, then* $v' = f(v)$ *is a n-dimensional random variable and it has a density given by the relation*

$$p'(y) = p(g(y)) \cdot \left|J_g\right|,$$

where J_g *is the Jacobian of g, namely*

$$J_g = \det \begin{bmatrix} \frac{\partial g_1}{\partial y_1} & \cdots & \frac{\partial g_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial y_1} & \cdots & \frac{\partial g_n}{\partial y_n} \end{bmatrix}$$

(subscripts denote components) and $|\cdot|$ *is absolute value.*

PROOF. The proof uses results on absolute continuous functions which we take here for granted. Moreover, we only consider the 1-dimensional case since the multi-dimensional case is conceptually similar but notationally more complicated.

For $n = 1$, the thesis writes

$$p'(y) = p(g(y)) \left|\frac{dg}{dy}\right|. \tag{2.18}$$

We first establish that v' is a random variable. Start by observing that f, being the inverse of a differentiable and therefore continuous function g, is continuous, so that, by definition of continuity, the inverse image through f of an open set is open, and therefore in $\mathscr{B}(\mathbb{R})$. Now, it is not difficult to see that the collection of all sets whose

inverse image through f is in $\mathscr{B}(\mathbb{R})$ is a σ-algebra. Since Borel sets are the smallest σ-algebra containing the open sets, we conclude that the inverse image of any Borel set is in $\mathscr{B}(\mathbb{R})$, i.e. f is $\mathscr{B}(\mathbb{R})/\mathscr{B}(\mathbb{R})$-measurable. Hence, $v' = f(v)$ is a random variable in view of Theorem 1.5.

We turn now to prove the validity of (2.18).

Since g is 1-to-1, it is either increasing or decreasing. Suppose g is increasing (the decreasing case goes through similarly.) Fix an interval $[-M,M]$, where M is an integer. Then,

$$g(x) - g(-M) = \int_{-M}^{x} \frac{dg}{dy} dy \qquad for \ -M \le x \le M \tag{2.19}$$

(this follows from Lemma 7.25 and Theorem 7.18 in [6] that prove that a g increasing and everywhere differentiable over $[-M,M]$ is absolutely continuous over the same interval and from the fact that, for an absolutely continuous function g, $\frac{dg}{dy}$ is measurable and (2.19) holds - Theorem 7.20 in [6].)

Now, put $\mu(B) = \lambda(g(B))$, $B \in \mathscr{B}[-M,M]$ (λ is Lebesgue measure.) Since g is 1-to-1, the σ-additivity of λ implies the σ-additivity of μ, so that μ is a measure on $\mathscr{B}[-M,M]$. $g(x) - g(-M)$ is its distribution and, by virtue of (2.19), $\frac{dg}{dy}$ is its density (the notions of distribution and density used here are the same as in Definitions 2.4 and 2.5 expect for the scaling factor $\mu[-M,M]$.) Thus,

$$\lambda(g(B)) = \mu(B) = \int_{[-M,M]} 1(B) d\mu = \int_{[-M,M]} 1(B) \frac{dg}{dy} dy = \int_B \frac{dg}{dy} dy, \quad \forall B \in \mathscr{B}[-M,M], \tag{2.20}$$

where $1(\cdot)$ is the indicator function and the third "=" is justified in view of the comment that follows Definition 2.5.

Turn now to consider the density p. Assume first that $p = 1(A)/\lambda(A)$, where A is a Borel set with $\lambda(A) > 0$. Then

$$\begin{aligned}
\int_{-M}^{x} p(g(y)) \frac{dg}{dy} dy &= \frac{1}{\lambda(A)} \int_{[-M,x] \cap f(A)} \frac{dg}{dy} dy \\
&= \frac{1}{\lambda(A)} \lambda\left(g([-M,x] \cap A)\right) \quad (using \ (2.20)) \\
&= P\{g(-M) \le v \le g(x)\} \\
&= P\left\{-M \le v' \le x\right\}.
\end{aligned}$$

Let $M \to \infty$ to conclude that

$$\int_{-\infty}^{x} p(g(y))\frac{dg}{dy}dy = P\left\{v' \leq x\right\}, \tag{2.21}$$

from which we see that $p(g(y))\frac{dg}{dy}$ is the density of v', that is (2.18) is established.

In the case of a generic density p, to arrive to equation (2.21) one has to first extend the derivation in (2.21) to simple functions, and then pass to the limit by the monotone convergence Theorem 3.7. □

Theorem 2.14 can also be applied to functions $f : \mathbb{R}^n \to \mathbb{R}^m$, with $m < n$, provided that $\mathbb{R}^m$ can be augmented with dummy variables such that the augmented transformation becomes invertible. This is illustrated by an example.

EXAMPLE 2.15 *Given a bi-dimensional v with density p, suppose we want to compute the density of η defined as the sum of the two components of v: $\eta = v_1 + v_2$. Since the transformation $v \to \eta$ is from $\mathbb{R}^2$ to $\mathbb{R}^1$, Theorem 2.14 cannot be directly applied. However, introducing the dummy variable $\xi = v_2$ and letting $v' = [\eta \;\; \xi]^T$, we have*

$$v' = Av,$$

where $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, *and this is an invertible transformation. Now,* $A^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$, *so that* $\left|J_g\right| = \left|\det \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}\right| = 1$, *and Theorem 2.14 gives:*

$$p'(y_1, y_2) = p(y_1 - y_2, y_2).$$

The density of η can then be recovered by integration:

$$p_\eta(y_1) = \int_{\mathbb{R}} p(y_1 - y_2, y_2)dy_2.$$

The reader is invited to complete this example by computing p_η when v is uniformly distributed in $[0,1]^2$. □

Chapter 3

STOCHASTIC CONVERGENCE

3.1 Probabilistic notions of convergence

We introduce a number of probabilistic notions of convergence of a sequence of random variables v_n to a limit random variable v and relate each one to the others.

DEFINITION 3.1 (stochastic convergence) *Given a sequence of random variables v_n and an additional random variable v defined on a probability space $(\Omega, \mathscr{F}, P)$, we say that $v_n \to v$*

(a) uniformly, if $\sup_{\omega\in\Omega} |v_n(\omega) - v(\omega)| \to 0$*;*
(b) surely, if $v_n(\omega) - v(\omega) \to 0$, $\forall \omega \in \Omega$*;*
(c) almost surely, if $P\{\omega \text{ such that } v_n(\omega) - v(\omega) \to 0\} = 1$ *(when we want to emphasize probability P we write P-almost surely.) Another expression equivalent to "almost surely" is "with probability 1";*
(d) in $\mathbb{L}^2$*, if* $E[(v_n - v)^2] \to 0$*;*
(e) in $\mathbb{L}^1$*, if* $E[|v_n - v|] \to 0$*;*
(f) in probability, if $\forall \varepsilon > 0$, $P\{\omega \text{ such that } |v_n(\omega) - v(\omega)| \geq \varepsilon\} \to 0$*;*
(g) weakly, if for any continuous and bounded function $f : \mathbb{R} \to \mathbb{R}$*, we have* $E[f(v_n)] \to E[f(v)]$*. "$v_n \to v$ weakly" is also expressed as "$v_n \to v$ in distribution". When v_n converges in distribution to a variable with distribution F, we also write $v_n \sim AsF$.* □

Definitions (a)-(f) are concerned with the behavior of $v_n - v$ and require that this difference goes to zero as $n \to \infty$ in different ways as specified by the different definitions. Thus, for instance, v_n tends to v almost surely if $v_n - v$ tends to zero almost surely. In contrast, the fact that $v_n \to v$ weakly in no way implies that $v_n - v \to 0$.

To understand this, suppose e.g. that $v_n = \xi$, $n = 1,2,\ldots$, where ξ is a fixed random variable different from v but sharing with v the same distribution. Then clearly $E[f(v_n)] = E[f(\xi)] = E[f(v)]$, $\forall n$, so that $v_n \to v$ weakly, but $v_n - v = \xi - v$ does not converge to zero.

Weak convergence is in fact a property of the random variables distributions. Indeed, $E[f(v_n)] \to E[f(v)]$ can be rewritten as $\int_{\mathbb{R}} f dF_n \to \int_{\mathbb{R}} f dF$ (where F_n is the distribution of v_n and F that of v) and we see that weak convergence means that F_n approaches F. Weak convergence is discussed in detail in Section 3.4.

The different notions of convergence are related to each other by the following theorem.

THEOREM 3.2 *The implications shown in Figure 3.1 hold true.*

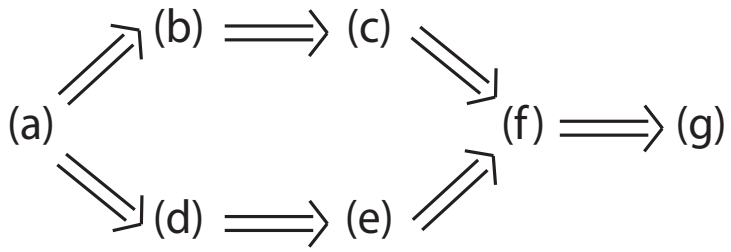

Figure 3.1: Implications among convergence properties

PROOF. Implications $(a) \Rightarrow (b) \Rightarrow (c)$ and $(a) \Rightarrow (d)$ are obvious.

Let $\xi_n := v_n - v$.

$(d) \Rightarrow (e)$] By Schwarz inequality 4.8 applied to $\mathbb{L}^2$ (see Example 4.5): $E[|\xi_n|] = E[1 \cdot |\xi_n|] \leq (E[1^2])^{1/2}(E[\xi_n^2])^{1/2} = (E[\xi_n^2])^{1/2}$, showing that $E[\xi_n^2] \to 0$ implies $E[|\xi_n|] \to 0$.
$(c) \Rightarrow (f)$] Let $A_j^\varepsilon := \{\omega \textit{ such that } |\xi_n| < \varepsilon, \forall n \geq j\}$. A_j^ε is increasing with j and $\cup_j A_j^\varepsilon =: A^\varepsilon$ is the set where the tail of $|\xi_n|$ is below ε. Since $\xi_n \to 0$ almost surely, $P(A^\varepsilon) = 1$, from which $P(A_j^\varepsilon) \to 1$ as $j \to \infty$. Now, since $\{\omega \textit{ such that } |\xi_j| \geq \varepsilon\} \subseteq \Omega - A_j^\varepsilon$, we obtain that $P\{\omega \textit{ such that } |\xi_j| \geq \varepsilon\} \to 0$.
$(e) \Rightarrow (f)$] From (e), $\varepsilon P\{\omega \textit{ such that } |\xi_n| \geq \varepsilon\} \leq E[|\xi_n|] \to 0$ and (f) follows.
$(f) \Rightarrow (g)$] Given $\varepsilon_1, \varepsilon_2 > 0$, fix M and ε such that $P\{\omega \textit{ such that } |v| \geq M\} \leq \varepsilon_1$ and $|f(x) - f(y)| \leq \varepsilon_2$ for $|y| < M$ and $|x - y| < \varepsilon$ (such ε exists since a continuous function is uniformly continuous on a bounded set.) Then,

$$\begin{aligned} &|E[f(v_n)] - E[f(v)]| \\ &\quad \leq \; E[|f(v_n) - f(v)|] \\ &\quad \leq \; (2 \max_x |f(x)|)\,[\varepsilon_1 + P\{\omega \textit{ such that } |v_n - v| \geq \varepsilon\}] + \varepsilon_2. \end{aligned}$$

Since ε_1 and ε_2 are arbitrarily small, f is bounded, and $P\{\omega \text{ such that } |v_n - v| \geq \varepsilon\} \to 0$ by assumption, (g) follows. □

No other implications than the ones stated in Theorem 3.2 hold true. In particular, almost sure convergence does not imply and is not implied by $\mathbb{L}^2$-convergence, as the following example shows.

EXAMPLE 3.3 *Consider the following sequence of random variables defined on the probability space* $([0,1], \mathscr{B}[0,1], \lambda)$*:*

$$v_n = \begin{cases} \sqrt{n}, & \text{on } [0,1/n] \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

Letting $v := 0$*, clearly* $v_n \to v$ *almost surely, but* $E[(v_n - v)^2] = \frac{1}{n} n = 1 \not\to 0$*, so that almost sure convergence does not imply* $\mathbb{L}^2$*-convergence.*

Conversely, consider the sequence $v_1^1, v_2^1, v_2^2, v_3^1, v_3^2, v_3^3, \ldots$ *with*

$$v_n^k = \begin{cases} 1, & \text{on } [\frac{k-1}{n}, \frac{k}{n}] \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

This sequence is $\mathbb{L}^2$*-convergent to zero, but, for every* $\omega \in [0,1]$*, the sequence keeps oscillating between* 0 *and* 1 *so that it does not converge for any* ω*.* □

The reason why we had $\mathbb{L}^2$-convergence but not almost sure convergence in the latter example was that the intervals where $v_n^k = 1$ in (3.2) had two properties: i) their size shrinks (so that $\mathbb{L}^2$ convergence to zero holds); and ii) each point in $[0,1]$ falls infinitely many times in the intervals (and, thus, almost sure convergence fails.) The latter property is possible since the sum of the interval sizes $1, \frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \ldots$ where $v_n^k = 1$ diverges. The following theorem showing a converse result that when this sum is finite almost sure convergence does take place is important to assess almost sure convergence in many contexts.

THEOREM 3.4 *Let* v_n*,* $n = 1, 2, \ldots$*, and* v *be random variables. Suppose that for any* $\varepsilon > 0$*,*

$$\sum_{n=1}^{\infty} P\{\omega \text{ such that } |v_n - v| \geq \varepsilon\} < \infty, \tag{3.3}$$

then $v_n \to v$ *almost surely.*

PROOF. Letting

$$A_j^k := \{\omega \text{ such that } |v_n - v| \geq \frac{1}{k}, \quad \forall n \geq j\},$$

we have that $\{\omega \text{ such that } v_n - v \not\to 0\} = \cup_{k=1}^{\infty} \cap_{j=1}^{\infty} A_j^k$. Thus,

$$P\{\omega \text{ such that } v_n - v \not\to 0\} \leq \sum_{k=1}^{\infty} \lim_{j\to\infty} \sum_{n=j}^{\infty} P\{\omega \text{ such that } |v_n - v| \geq \frac{1}{k}\}.$$

Since (3.3) holds, each single term $\lim_{j\to\infty} \sum_{n=j}^{\infty} P\{\omega \text{ such that } |v_n - v| \geq \frac{1}{k}\}$ is zero and the right-hand-side of the previous inequality is null, so proving that $v_n \to v$ almost surely. □

Measurability of the limit of random variables

The next result relates the measurability properties of a sequence of random variables to those of its limit.

THEOREM 3.5 *Let v_n be a sequence of random variables on $(\Omega, \mathscr{F}, P)$ (so that each v_n is $\mathscr{F}$-measurable) and let v be an additional variable that is not required to be $\mathscr{F}$-measurable by assumption.*

(i) if $v_n(\omega) \to v(\omega)$, $\forall \omega \in \Omega$, then v is $\mathscr{F}$-measurable;

(ii) if the set $\{v_n \not\to v\}$ of the ω where v_n does not converge to v is measurable and $P\{v_n \not\to v\} = 0$, then v need not be $\mathscr{F}$-measurable. However, $\bar{v}$ so defined

$$\bar{v} = \begin{cases} v, & \text{where } v_n \to v \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

is $\mathscr{F}$-measurable.

The idea underlying (i) is that v has to behave well (i.e. it has to be $\mathscr{F}$-measurable) since it is the limit of $\mathscr{F}$-measurable variables. In point (ii), $P\{v_n \not\to v\} = 0$ leaves open the possibility that v exhibits a weird behavior on the set where $v_n \not\to v$, so losing the measurability property of v. If however v is flattened to 0 on $\{v_n \not\to v\}$, the so-obtained $\bar{v}$ is $\mathscr{F}$-measurable.

PROOF.
(i) For any $a, b \in \mathbb{R}$, we have

$$\{\omega \text{ such that } v \in (a,b)\} = \cup_{p=1}^{\infty} \cap_{n\geq p} \{\omega \text{ such that } v_n \in (a,b)\}.$$

Since $\{\omega \textit{ such that } v_n \in (a,b)\} \in \mathscr{F}$ and a σ-algebra is closed under countable intersection and union, we have that $\{\omega \textit{ such that } v \in (a,b)\} \in \mathscr{F}$, from which the measurability of v follows by applying the test of measurability 2.3.

(ii) Define

$$\bar{v}_n = \begin{cases} v_n, & \textit{where } v_n \to v \\ 0, & \textit{otherwise.} \end{cases}$$

$\bar{v}_n$ is $\mathscr{F}$-measurable and convergent for any $\omega \in \Omega$ and thus, by an application of point (i), its limit $\bar{v}$ as given by (3.4) is $\mathscr{F}$-measurable. □

If $\mathscr{F}$ is a complete σ-algebra (a σ-algebra $\mathscr{F}$ is complete when all subsets of sets in $\mathscr{F}$ with zero probability are also in $\mathscr{F}$), then v itself is measurable and introducing $\bar{v}$ is not necessary. Note also that a complete σ-algebra can always be obtained from a non-complete one by augmenting it with all subsets of zero measure sets. Thus, considering complete σ-algebras represents a valid alternative to constructing $\bar{v}$.

The next theorem which analyzes the measurability of the limit with respect to a coarser σ-algebra $\mathscr{G} \subseteq \mathscr{F}$ is somehow more involved.

THEOREM 3.6 *Let v_n be a sequence of random variables on $(\Omega, \mathscr{F}, P)$ that are $\mathscr{G}$-measurable for some σ-algebra $\mathscr{G} \subseteq \mathscr{F}$ and let v be an additional random variable. If $v_n \to v$ in probability, then v need not be $\mathscr{G}$-measurable, but there exists $\mathscr{G}$-measurable $\bar{v}$ such that $P\{v \neq \bar{v}\} = 0$.*

PROOF. Fix a sequence of real numbers $\varepsilon_k \downarrow 0$ (ε_k is decreasing and tends to zero) such that and extract from v_n a subsequence v_{n_k} such that

$$\sum_{k=1}^{\infty} P\{\omega \textit{ such that } |v_{n_k} - v| \geq \varepsilon_k\} < \infty, \tag{3.5}$$

(such a sequence exists since $v_n \to v$ in probability.) Equation (3.5) implies that $v_{n_k} \to v$ almost surely, as it can be proven by applying Theorem 3.4. Indeed, given $\varepsilon > 0$, let $\bar{k}$ be such that $\varepsilon_{\bar{k}} \leq \varepsilon$ and write

$$\begin{aligned}
&\sum_{k=1}^{\infty} P\{\omega \textit{ such that } |v_{n_k} - v| \geq \varepsilon\} \\
&= \sum_{k=1}^{\bar{k}-1} P\{\omega \textit{ such that } |v_{n_k} - v| \geq \varepsilon\} + \sum_{k=\bar{k}}^{\infty} P\{\omega \textit{ such that } |v_{n_k} - v| \geq \varepsilon\} \\
&\leq \sum_{k=1}^{\bar{k}-1} P\{\omega \textit{ such that } |v_{n_k} - v| \geq \varepsilon\} + \sum_{k=\bar{k}}^{\infty} P\{\omega \textit{ such that } |v_{n_k} - v| \geq \varepsilon_k\} \\
&< \infty,
\end{aligned}$$

so that the assumption (3.3) of Theorem 3.4 is satisfied.

We next show that the set of ω's where v_{n_k} does not converge to a finite limit is $\mathscr{G}$-measurable. The set where $v_{n_k} \to \infty$ can be written as $\cap_{M=1}^{\infty} \cup_{p=1}^{\infty} \cap_{k \geq p} \{\omega \textit{ such that } v_{n_k} \geq M\}$, representing the ω's where, for any fixed M, the tail of v_{n_k} from a certain p onward is above M. Since $\{\omega \textit{ such that } v_{n_k} \geq M\} \in \mathscr{G}$ and a σ-algebra is closed under countable intersection and union, we have that $\{\omega \textit{ such that } v_{n_k} \to \infty\} \in \mathscr{G}$. Similarly, $\{\omega \textit{ such that } v_{n_k} \to -\infty\} \in \mathscr{G}$. Consider then a ω where $v_{n_k}(\omega)$ oscillates. Then, there exist two rational numbers α and β with $\alpha < \beta$ such that $v_{n_k}(\omega)$ is below α infinitely many times and above β infinitely many times. The set of ω's where $v_{n_k}(\omega) < \alpha$ infinitely many times writes $\cap_{p=1}^{\infty} \cup_{k \geq p} \{\omega \textit{ such that } v_{n_k} < \alpha\}$ and is $\mathscr{G}$-measurable. Likewise, is $\mathscr{G}$-measurable the set where $v_{n_k}(\omega) > \beta$ infinitely many times, so that we obtain the $\mathscr{G}$-measurability of the set where $v_{n_k}(\omega)$ oscillates between α and β infinitely many times, which is the intersection of the two previous sets. Finally, the set where v_{n_k} oscillates is obtained as union over all rationals α and β and is therefore $\mathscr{G}$-measurable.

Now, let

$$\bar{v}_{n_k} = \begin{cases} v_{n_k}, & \textit{where } v_{n_k} \textit{ converges to a finite value} \\ 0, & \textit{otherwise.} \end{cases}$$

$\bar{v}_{n_k}$ is $\mathscr{G}$-measurable (that is it is a random variable on $(\Omega, \mathscr{G}, P)$) and it converges for any $\omega \in \Omega$. By an application of part (i) of Theorem 3.5, its limit, say $\bar{v}$, is therefore $\mathscr{G}$-measurable. Moreover, for all ω's where $v_{n_k}(\omega) \to v(\omega)$, $v_{n_k}(\omega)$ converges to the finite value $v(\omega)$ and so $\bar{v}_{n_k}(\omega) = v_{n_k}(\omega)$. Thus, $\bar{v}_{n_k}(\omega) \to v(\omega)$ and $v(\omega) = \bar{v}(\omega)$. Since $v_{n_k}(\omega) \to v(\omega)$ holds with probability 1, we conclude that $v(\omega) = \bar{v}(\omega)$ with probability 1 and the proof is complete. □

3.2 Limit under the sign of expectation

Suppose that $v_n \to v$ almost surely. Under what conditions is it true that $E[v_n] \to E[v]$? The following theorems provide an answer.

THEOREM 3.7 (monotone convergence) *Let $v_n, n = 1,2,\ldots$, and v be random variables such that $v_n \uparrow v$ almost surely (i.e. v_n is increasing and tends to v almost surely), and assume that $v_n \geq z, n = 1,2,\ldots$, for some random variable z with $E[z] > -\infty$. Then,*

$$E[v_n] \uparrow E[v].$$

□

THEOREM 3.8 (dominated convergence) *Let $v_n, n = 1,2,\ldots$, and v be random variables such that $v_n \to v$ almost surely, and assume that $|v_n| \leq z, n = 1,2,\ldots$, for some random variable z with $E[z] < \infty$. Then,*

$$E[v_n] \to E[v].$$

□

A proof of these theorems can be found in any textbook on probability.

In the statements of the theorems, two types of conditions are present: v_n is required to approach v; and v_n is bounded by z. The latter condition serves the purpose to limit the importance of the mismatch between v_n and v on events of small probability. An example clarifies this matter.

EXAMPLE 3.9 (Example 3.3 continued) *Consider again the v_n's in (3.1). Clearly, $v_n^2 \to 0$ almost surely, but $E[v_n^2] = 1 \not\to E[0] = 0$. Here, no dominating z exists with $E[z] < \infty$, so that the conditions of Theorem 3.8 are violated.* □

3.3 Convergence results for independent random variables

We commence by proving probabilistic inequalities. Besides being in use later in this section when proving convergence results, these inequalities are of interest in their own right.

MARKOV'S INEQUALITY 3.10

For any nonnegative random variable v and real number $\varepsilon > 0$,

$$P\{v \geq \varepsilon\} \leq \frac{E[v]}{\varepsilon}. \tag{3.6}$$

PROOF. The proof is elementary: $E[v] = \int_\Omega v dP \geq \int_{\{v \geq \varepsilon\}} v dP \geq \varepsilon P\{v \geq \varepsilon\}$. □

An application of Markov's inequality gives

CHEBYSHEV'S INEQUALITY 3.11

For any $\varepsilon > 0$,

$$P\{|v| \geq \varepsilon\} \leq \frac{E[v^2]}{\varepsilon^2}.$$

PROOF.

$$\begin{aligned} P\{|v| \geq \varepsilon\} &= P(v^2 \geq \varepsilon^2) \\ &\leq \frac{E[v^2]}{\varepsilon^2}. \quad (\textit{using } (3.6)) \end{aligned}$$

□

In Markov's inequality, the idea is to lowerbound $E[v]$ by squeezing the tail of v to the value ε. Thus, the bound is tight only when the tail rapidly vanishes after ε. A similar observation applies to Chebyshev's inequality. Better bounds can be found by redressing the random variable distribution through some transformation before Markov's inequality is applied. One such example is given by the following inequality of Chernoff. In this inequality, s is a free parameter that can be used to retune the distribution and an example of use of s is found in the proof of Hoeffding's inequality (Theorem 3.14.)

CHERNOFF'S INEQUALITY 3.12

For any $s > 0$ and $\varepsilon > 0$,

$$P\{v \geq \varepsilon\} \leq \frac{E[e^{sv}]}{e^{s\varepsilon}}. \tag{3.7}$$

PROOF.

$$\begin{aligned} P\{v \geq \varepsilon\} &= P\{e^{sv} \geq e^{s\varepsilon}\} \\ &\leq \frac{E[e^{sv}]}{e^{s\varepsilon}}. \quad (\textit{using } (3.6)) \end{aligned}$$

□

Concentration inequalities

Consider a sequence of independent random variables $v_k, k = 1, 2, \ldots$. Concentration inequalities study how a function $f(v_1, v_2, \ldots, v_n)$ concentrates around its expected value $E[f(v_1, v_2, \ldots, v_n)]$.

Here, we are mainly concerned with the deviation of sums of random variables from their means, that is our interest is on function $f(v_1, v_2, \ldots, v_n) = \frac{1}{n}\sum_{k=1}^{n} v_k$ and we study the behavior of

$$S_n := \frac{1}{n}\sum_{k=1}^{n} v_k - E\left[\frac{1}{n}\sum_{k=1}^{n} v_k\right]. \tag{3.8}$$

A first bound is obtained by means of Chebyshev's inequality:

$$P\{|S_n| \geq \varepsilon\} \leq \frac{E[S_n^2]}{\varepsilon^2} = \frac{\frac{1}{n^2}\sum_{k=1}^{n} var(v_k)}{\varepsilon^2}. \tag{3.9}$$

EXAMPLE 3.13 *For an independent and identically distributed sequence of Bernoulli random variables (i.e. $P\{v_k = 1\} = 1 - P\{v_k = 0\} = p$), from (3.9) we have*

$$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n} v_k - p\right| \geq \varepsilon\right\} \leq \frac{p(1-p)}{n\varepsilon^2}. \tag{3.10}$$

□

Do we expect that bound (3.9) is tight? Remember that Chebishev's inequality is tight when the distribution tail vanishes rapidly after ε. On the other hand, applying the central limit Theorem 3.19 leads to the conclusion that, under mild assumptions, the distribution of S_n tends weakly to a Gaussian, a long-tailed distribution. For example, in the case of the Bernulli sequence of Example 3.13, letting $\Phi(x) = \int_{-\infty}^{x}(2\pi)^{-1/2}e^{-r^2/2}dr$, the central limit theorem states that

$$P\left\{\sqrt{\frac{n}{p(1-p)}}\left(\frac{1}{n}\sum_{k=1}^{n}v_k-p\right)\geq x\right\}\to 1-\Phi(x)\leq\frac{e^{-x^2/2}}{x\sqrt{2\pi}},$$

from which we would expect something like

$$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}v_k-p\right|\geq\varepsilon\right\}\sim e^{-\frac{n\varepsilon^2}{2p(1-p)}}, \tag{3.11}$$

that is the probability decays exponentially with n. The gap between inversely linear (see (3.10)) and exponential convergence in n can be filled in by resorting to Hoeffding's inequality.

THEOREM 3.14 (Hoeffding's inequality) *Let $v_k, k=1,2,\ldots$, be independent bounded random variables taking value in $[\alpha_k,\beta_k]$ and let S_n be defined as in (3.8). Then, for any $\varepsilon>0$,*

$$P\{S_n\geq\varepsilon\}\leq e^{-\frac{2n^2\varepsilon^2}{\sum_{k=1}^{n}(\beta_k-\alpha_k)^2}}; \tag{3.12}$$

and

$$P\{S_n\leq-\varepsilon\}\leq e^{-\frac{2n^2\varepsilon^2}{\sum_{k=1}^{n}(\beta_k-\alpha_k)^2}}. \tag{3.13}$$

PROOF. For ease of notation, we assume $[\alpha_k,\beta_k]=[0,1]$, in which case we prove that

$$P\{S_n\geq\varepsilon\}\leq e^{-2n\varepsilon^2}; \tag{3.14}$$

and

$$P\{S_n\leq-\varepsilon\}\leq e^{-2n\varepsilon^2}. \tag{3.15}$$

The extension is easy.

We start by observing that for any random variable v with $E[v]=0$ and $\alpha\leq v\leq 1+\alpha$, and for any $h>0$, we have

$$E[e^{hv}]\leq e^{\frac{h^2}{8}}. \tag{3.16}$$

In fact, by convexity of the exponential function, $e^{hv}\leq(v-\alpha)e^{(1+\alpha)h}+(1+\alpha-v)e^{\alpha h}$, so that

$$\begin{aligned} E[e^{hv}] &\leq E\left[(v-\alpha)e^{(1+\alpha)h}+(1+\alpha-v)e^{\alpha h}\right] \\ &= E\left[-\alpha e^{(1+\alpha)h}+(1+\alpha)e^{\alpha h}\right] \quad (since\ E[v]=0) \\ &= -\alpha e^{(1+\alpha)h}+(1+\alpha)e^{\alpha h} \\ &= e^{\Gamma(h)}, \end{aligned}$$

where $\Gamma(h)=\alpha h+\ln(1+\alpha-\alpha e^{h})$. The derivative of $\Gamma(h)$ is $\Gamma^{'}(h)=\alpha-\alpha/[(1+\alpha)e^{-h}-\alpha]$, so that $\Gamma^{'}(0)=0$. Moreover,

$$\begin{aligned} \Gamma^{''}(h) &= \frac{-\alpha(1+\alpha)e^{-h}}{[(1+\alpha)e^{-h}-\alpha]^2} \\ &= \frac{ab}{[a+b]^2} \quad (with\ a=(1+\alpha)e^{-h}, b=-\alpha) \\ &\leq \frac{1}{4}, \quad \forall h. \end{aligned}$$

Thus, by Taylor series expansion, for some $\xi\in[0,h]$:

$$\Gamma(h)=\Gamma(0)+\Gamma^{'}(0)h+\frac{1}{2}\Gamma^{''}(\xi)h^2\leq\frac{h^2}{8},$$

which, used in (3.17), yields (3.16).

Thanks to (3.16), (3.14) is now easily obtained from Chernoff's inequality:

$$\begin{aligned} P\{S_n\geq\varepsilon\} &\leq \frac{E[e^{sS_n}]}{e^{s\varepsilon}} \quad (using\ Chernoff's\ inequality\ 3.7) \\ &= \frac{E[e^{s\frac{1}{n}\sum_{k=1}^{n}(v_k-E[v_k])}]}{e^{s\varepsilon}} \\ &= \frac{\prod_{k=1}^{n}E[e^{s\frac{1}{n}(v_k-E[v_k])}]}{e^{s\varepsilon}} \quad (by\ the\ independence\ of\ the\ v_k's) \\ &\leq \frac{\prod_{k=1}^{n}e^{\frac{s^2}{8n^2}}}{e^{s\varepsilon}} \quad (using\ (3.16)\ with\ h=s/n) \\ &\leq e^{-2n\varepsilon^2}. \quad (by\ choosing\ s=4\varepsilon n) \end{aligned}$$

Equation (3.15) is obtained similarly. □

EXAMPLE 3.15 (Example 3.13 continued) *Using Hoeffding's inequality yields*

$$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n} v_k - p\right| \geq \varepsilon\right\} \leq 2e^{-2n\varepsilon^2}$$

(compare with (3.11).) □

Hoeffding's inequality deals specifically with empirical means, showing that the empirical mean rapidly concentrates around the true mean value. The reason why this is so is that in the empirical mean each single variable has a moderate influence in determining the empirical mean value and, moreover, different variables do not cooperate because they are independent.

It is a fact that this Hoeffding's inequality can be extended to more general functions provided that each variable has a marginal importance in determining the total value of the function, as given in the following theorem which we state without providing a proof (for a proof see e.g. [4].)

THEOREM 3.16 (the bounded difference inequality) *Let $v_k, k = 1,2,\ldots,$ be independent random variables taking value in a set A and assume that $\forall x_1,\ldots,x_n \in A$, $x'_k \in A$, and $\forall k \in [1,n]$, the measurable function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the condition:*

$$|f(x_1,\ldots,x_{k-1},x_k,x_{k+1},\ldots,x_n) - f(x_1,\ldots,x_{k-1},x'_k,x_{k+1},\ldots,x_n)| \leq \gamma_k. \tag{3.17}$$

Then, for any $\varepsilon > 0$,

$$P\{f(v_1,v_2,\ldots,v_n) - E[f(v_1,v_2,\ldots,v_n)] \geq \varepsilon\} \leq e^{-\frac{2\varepsilon^2}{\sum_{k=1}^{n}\gamma_k^2}}; \tag{3.18}$$

and

$$P\{f(v_1,v_2,\ldots,v_n) - E[f(v_1,v_2,\ldots,v_n)] \leq -\varepsilon\} \leq e^{-\frac{2\varepsilon^2}{\sum_{k=1}^{n}\gamma_k^2}}; \tag{3.19}$$

□

Note that (3.18) and (3.19) reduce to (3.12) and (3.13) when we consider empirical means.

Laws of large numbers

The laws of large numbers study the convergence of $\frac{1}{n}\sum_{k=1}^{n} v_k$ to $E\left[\frac{1}{n}\sum_{k=1}^{n} v_k\right]$. This is probably the most studied problem in probability theory and the literature offers an abundant supply of results under varying assumptions and according to different notions of convergence. Here, we only present a standard result and prove it by means of concentration inequalities.

THEOREM 3.17 (law of large numbers) *Let $v_k, k = 1, 2, \ldots$, be independent random variables with uniformly bounded variance: $var(v_k) \leq C, \forall k$. Then,*

$$\frac{1}{n}\sum_{k=1}^{n} v_k \to E\left[\frac{1}{n}\sum_{k=1}^{n} v_k\right] \quad \textit{almost surely.}$$

Before proving the theorem, we would like to note that the result is immediate from Hoeffding's inequality if the variables are uniformly bounded, In fact if $v_k \in [\alpha, \beta]$, $\forall k$,

$$P\{|S_n| \geq \varepsilon\} \leq 2e^{-\frac{2n\varepsilon^2}{(\beta-\alpha)^2}},$$

where $S_n := \frac{1}{n}\sum_{k=1}^{n} v_k - E\left[\frac{1}{n}\sum_{k=1}^{n} v_k\right]$, so that $S_n \to 0$ almost surely follows from Theorem 3.4. On the other hand, if we only assume the boundedness of the variance of the v_k's (as is the case in the theorem), Hoeffding's inequality does not apply. On the other hand, resorting to Chebyshev's inequality 3.9 yields

$$P\{|S_n| \geq \varepsilon\} \leq \frac{E[S_n^2]}{\varepsilon^2} = \frac{C}{n\varepsilon^2},$$

which is not enough to prove that $S_n \to 0$ almost surely by way of Theorem 3.4 since $\sum_{n=1}^{\infty} \frac{C}{n\varepsilon^2} = \infty$. The proof of the theorem given below suggests a way to get around this difficulty.

Instead of directly proving the theorem, we prefer to state the following lemma, from which the theorem immediately follows, because the lemma is useful in other contexts as well.

LEMMA 3.18 *Consider the doubly indexed set of random variables S_r^p such that $S_r^p = 0$ for $r > p$ and assume that $E[(S_r^p)^2] \leq C(p+1-r)$ for some constant C and that, for $m < n$, $|S_1^n| \leq |S_1^m| + |S_{m+1}^n|$. Then,*

$$\frac{1}{n}S_1^n \to 0 \quad \textit{almost surely.}$$

Note that Theorem 3.17 immediately follows from the lemma by the position $S_r^p := \sum_{k=r}^p (v_k - E[v_k])$.

PROOF OF THE LEMMA. Given an integer n, let N be the integer such that $N^2 \leq n < (N+1)^2$ and write:

$$\left|\frac{1}{n}S_1^n\right| \leq \frac{1}{N^2}\left|S_1^{N^2}\right| + \frac{1}{N^2}\left|S_{N^2+1}^n\right|. \tag{3.20}$$

The lemma is proven by showing that both terms in the last expression go to zero almost surely.

As for the first term, by the Chebishev's inequality we have

$$\begin{aligned}
\sum_{N=1}^{\infty} P\left\{\frac{1}{N^2}\left|S_1^{N^2}\right| \geq \varepsilon\right\} &\leq \sum_{N=1}^{\infty} \frac{var\left(\frac{1}{N^2}\left|S_1^{N^2}\right|\right)}{\varepsilon^2} \\
&\leq \sum_{N=1}^{\infty} \frac{CN^2}{N^4\varepsilon^2} \\
&< \infty,
\end{aligned}$$

so that almost sure convergence to zero follows from Theorem 3.4. For the second term in (3.20) we instead have

$$\begin{aligned}
\sum_{n=1}^{\infty} P\left\{\frac{1}{N^2}\left|S_{N^2+1}^n\right| \geq \varepsilon\right\} &\leq \sum_{n=1}^{\infty} \frac{var\left(\frac{1}{N^2}\left|S_{N^2+1}^n\right|\right)}{\varepsilon^2} \\
&\leq \sum_{n=1}^{\infty} \frac{C(n-N^2)}{N^4\varepsilon^2} \\
&\leq \sum_{N=1}^{\infty} \sum_{n=N^2}^{(N+1)^2-1} \frac{C(n-N^2)}{N^4\varepsilon^2} \\
&\leq \sum_{N=1}^{\infty} ((N+1)^2 - N^2)\frac{C((N+1)^2 - 1 - N^2)}{N^4\varepsilon^2} \\
&\leq \sum_{N=1}^{\infty} (2N+1)\frac{C2N}{N^4\varepsilon^2} \\
&< \infty
\end{aligned}$$

and, again, Theorem 3.4 can be resorted to to prove almost sure convergence to zero. □

Central limit theorems

As for the laws of large numbers, the literature on central limit theorems is truly vast. We only provide a basic treatment of the topic.

Theorem 3.17 shows that, for independent random variables with boundend variance, the following convergence takes place

$$\frac{1}{n}\sum_{k=1}^{n} v_k \to E\left[\frac{1}{n}\sum_{k=1}^{n} v_k\right] \quad \textit{almost surely},$$

or, equivalently,

$$\frac{1}{n}\sum_{k=1}^{n} (v_k - E[v_k]) \to 0 \quad \textit{almost surely.}$$

A question that arises naturally is as to how fast convergence to zero takes place. This question is answered by the following central limit theorem.

NOTE: In condition (3.21) of the theorem, the integral has to be intended as follows. Consider the total variation $H(x)$ of function $f := F_k - \Phi_k$, namely $H(x) := \sup \sum_{i=1}^{N} |f(t_i) - f(t_{i-1})|$ where supremum is take over all N and over all choices of t_i such that $t_0 < t_1 < \cdots < t_n = x$. $H(x)$ is nondecreasing and tends to a finite constant α as $x \to \infty$. If $\alpha = 0$ (which only happens if $F_k = \Phi_k$), then the integral is taken with respect to the zero measure and its value is zero. Otherwise, $H(x)/\alpha$ is a probability distribution. Integration is with respect to this measure and, to compensate for the division by α, the integrand is multiplied by this same α value.

THEOREM 3.19 (central limit theorem) *Let $v_k, k = 1, 2, \ldots$, be independent random variables with probability distribution F_k, zero mean and variance σ_k^2, and let $V_n^2 := \sum_{k=1}^{n} \sigma_k^2$. Moreover, let $\Phi_k(x) = \int_{-\infty}^{x} (2\pi)^{-1/2} \sigma_k^{-1} e^{-r^2/2\sigma_k^2} dr$ be the Gaussian distribution with zero mean and same variance as v_k. If*

$$\forall \varepsilon > 0, \quad \frac{1}{V_n^2}\sum_{k=1}^{n} \int_{|x|>\varepsilon V_n} x^2 d|F_k - \Phi_k| \to 0, \quad \textit{as } n \to \infty \tag{3.21}$$

(see the NOTE before the theorem on how to interpret this integral), then,

$$\frac{1}{V_n}\sum_{k=1}^{n} v_k \sim AsG(0,1).$$

Based on the results in Chapter 2.4, we know that the sum of jointly Gaussian random variables is Gaussian too. In words, Gaussianity is a "closed world": once we are in it, we cannot get out by applying linear operations. Theorem 3.19 tells us more: this world is also "attractive" and the sum of independent variables tend to it under general conditions.

Before proving the theorem we make an observation on the theorem assumptions.

Observation

Assumption (3.21) in the theorem is implied by each of the following handier sufficient conditions:

1. $\forall \varepsilon > 0, \quad \frac{1}{V_n^2} \sum_{k=1}^n \int_{|x|>\varepsilon V_n} x^2 dF_k(x) \to 0, \quad as\ n \to \infty$ **(Lindeberg's condition)**;

2. $\frac{1}{V_n^{2+\delta}} \sum_{k=1}^n E[|v_k|^{2+\delta}] \to 0$, for some $\delta > 0$ **(Lyapunov's condition)**.

We show that Lyapunov's condition implies Lindeberg's condition which, in turn, implies (3.21).

Lyapunov's condition $\Rightarrow$ Lindeberg's condition) We have:

$$\begin{aligned} E[|v_k|^{2+\delta}] &\geq \int_{|x|>\varepsilon V_n} |x|^{2+\delta} dF_k(x) \\ &\geq \varepsilon^\delta V_n^\delta \int_{|x|>\varepsilon V_k} x^2 dF_k(x), \end{aligned}$$

which gives

$$\begin{aligned} \frac{1}{V_n^2} \sum_{k=1}^n \int_{|x|>\varepsilon V_n} x^2 dF_k(x) &\leq \frac{1}{\varepsilon^\delta V_n^{2+\delta}} \sum_{k=1}^n E[|v_k|^{2+\delta}] \\ &\to 0, \end{aligned}$$

that is Lindeberg's condition.

Lindeberg's condition $\Rightarrow$ (3.21)) Note first that Lindeberg's condition implies

$$\frac{\max_{1\leq k\leq n} \sigma_k^2}{V_n^2} \to 0. \tag{3.22}$$

Indeed,

$$\begin{aligned} \frac{\max_{1\leq k\leq n} \sigma_k^2}{V_n^2} &\leq \frac{\sum_{k=1}^n \int_{|x|>\varepsilon V_n} x^2 dF_k(x) + \varepsilon^2 V_n^2}{V_n^2} \\ &\to \varepsilon^2, \end{aligned}$$

where Lindeberg's condition has been applied in computing the limit. Since ε is arbitrary, (3.22) follows. Now, letting $\Phi(x)$ be the Gaussian distribution with zero mean and unitary variance, we have:

$$\begin{aligned}
\frac{1}{V_n^2}\sum_{k=1}^{n}\int_{|x|>\varepsilon V_n} x^2 d\Phi_k(x) &= \frac{1}{V_n^2}\sum_{k=1}^{n}\int_{|z|>\frac{\varepsilon V_n}{\sigma_k}} \sigma_k^2 z^2 d\Phi(z) \quad (\textit{where } z = x/\sigma_k) \\
&\leq \frac{1}{V_n^2}\sum_{k=1}^{n}\int_{|z|>\frac{\varepsilon V_n}{max_{1\leq k\leq n}\sigma_k}} \sigma_k^2 z^2 d\Phi(z) \\
&= \int_{|z|>\frac{\varepsilon V_n}{max_{1\leq k\leq n}\sigma_k}} z^2 d\Phi(z) \cdot \frac{1}{V_n^2}\sum_{k=1}^{n}\sigma_k^2 \\
&= \int_{|z|>\frac{\varepsilon V_n}{max_{1\leq k\leq n}\sigma_k}} z^2 d\Phi(z) \\
&\to 0, \qquad (3.23)
\end{aligned}$$

where the limit to zero in the last step follows from the fact that (3.22) implies divergence of the integration boundary: $\frac{\varepsilon V_n}{max_{1\leq k\leq n}\sigma_k} \to \infty$. Finally, observing that $\frac{1}{V_n^2}\sum_{k=1}^{n}\int_{|x|>\varepsilon V_n} x^2 d|F_k(x) - \Phi_k(x)| \leq \frac{1}{V_n^2}\sum_{k=1}^{n}\int_{|x|>\varepsilon V_n} x^2 dF_k(x) + \frac{1}{V_n^2}\sum_{k=1}^{n}\int_{|x|>\varepsilon V_n} x^2 d\Phi_k(x)$, (3.21) follows from Lindeberg's condition and (3.23), so concluding the proof of the implication.

To help an intuitive understanding of the conditions, we note that Lindeberg's condition implies (3.22) and this means that, for large n, the largest variable variance becomes negligible as compared to the total variance. This fact can be interpreted by saying that each variable is infinitesimal if compared to the sum of the others. One the other hand, this infinitesimal behavior is not necessary for the central limit theorem to hold and condition (3.21) is for example satisfied by

$$v_1 \sim G(0,1),\ v_2 = 0,\ v_3 = 0,\ \ldots.$$

PROOF OF THEOREM 3.19. For a given n, let $f_k(t)$ and $\phi_k(t)$ be the characteristic functions of v_k/V_n and z_k/V_n (where the z_k's are independent and $G(0,\sigma_k^2)$ distributed), and let $f_n(t)$ and $\phi(t)$ be the characteristic functions of $\sum_{k=1}^{n} v_k/V_n$ and $\sum_{k=1}^{n} z_k/V_n$ (see Section 2.3 for the notion of characteristic function; note that we have not used the index n in $\phi(t)$ since $\sum_{k=1}^{n} z_k/V_n$ has $G(0,1)$ distribution for any n.) Note also that, due to independence, $f_n(t) = \Pi_{k=1}^{n} f_k(t)$ and $\phi(t) = \Pi_{k=1}^{n}\phi_k(t)$.

In the following derivations, the notation $\int f d(F_k - \Phi_k)$ is short for $\int f dF_k - \int f d\Phi_k$. Moreover, we use the following equalities:

$$\int_{\mathbb{R}} dF_k = \int_{\mathbb{R}} d\Phi_k = 1; \quad \int_{\mathbb{R}} x dF_k = \int_{\mathbb{R}} x d\Phi_k = 0; \quad \int_{\mathbb{R}} x^2 dF_k = \int_{\mathbb{R}} x^2 d\Phi_k = \sigma_k^2. \quad (3.24)$$

and the following bounds:

$$\left| e^{\frac{itx}{V_n}} - 1 - it\frac{x}{V_n} + \frac{1}{2}t^2\frac{x^2}{V_n^2} \right| \leq t^2\frac{x^2}{V_n^2}; \tag{3.25}$$

$$\left| e^{\frac{itx}{V_n}} - 1 - it\frac{x}{V_n} + \frac{1}{2}t^2\frac{x^2}{V_n^2} \right| \leq \frac{1}{6}|t|^3\frac{|x|^3}{V_n^3} \tag{3.26}$$

(the first bound follows from the Taylor expansion $e^{\frac{itx}{V_n}} = 1 + it\frac{x}{V_n} - \frac{1}{2}t^2\frac{\xi(x)^2}{V_n^2}$, with $|\xi(x)| \leq x$, and the second one from the Taylor expansion $e^{\frac{itx}{V_n}} = 1 + it\frac{x}{V_n} - \frac{1}{2}t^2\frac{x^2}{V_n^2} - \frac{1}{6}it^3\frac{\xi(x)^3}{V_n^3}$, again with $|\xi(x)| \leq x$.) Finally, we also make use of the elementary inequality

$$|\Pi_{k=1}^n \alpha_k - \Pi_{k=1}^n \beta_k| \leq \sum_{k=1}^n |\alpha_k - \beta_k|, \tag{3.27}$$

valid for α_k's and β_k's with $|\alpha_k|, |\beta_k| \leq 1$. In order to prove (3.27), start with $n = 2$ and write

$$\begin{aligned} |\alpha_1\alpha_2 - \beta_1\beta_2| &= |\alpha_1\alpha_2 - \beta_1\alpha_2 + \beta_1\alpha_2 - \beta_1\beta_2| \\ &\leq |(\alpha_1 - \beta_1)\alpha_2| + |\beta_1(\alpha_2 - \beta_2)| \\ &\leq |\alpha_1 - \beta_1| + |\alpha_2 - \beta_2|. \end{aligned}$$

The general case is obtained by repeated application of this same inequality.

We now have:

$$
\begin{aligned}
&|f_n(t)-\phi(t)| \\
&\quad = |\Pi_{k=1}^n f_k(t) - \Pi_{k=1}^n \phi_k(t)| \\
&\quad \le \sum_{k=1}^n |f_k(t)-\phi_k(t)| \quad (\textit{using } (3.27) \textit{ since } |f_k(t)|, |\phi_k(t)| \le 1) \\
&\quad = \sum_{k=1}^n \left| \int_{\mathbb{R}} e^{itx/V_n} d(F_k - \Phi_k) \right| \\
&\quad = \sum_{k=1}^n \left| \int_{\mathbb{R}} \left(e^{\frac{itx}{V_n}} - 1 - it\frac{x}{V_n} + \frac{1}{2}t^2\frac{x^2}{V_n^2} \right) d(F_k - \Phi_k) \right| \quad (\textit{using } (3.24)) \\
&\quad \le \sum_{k=1}^n \int_{|x|\le \varepsilon V_n} \left| e^{\frac{itx}{V_n}} - 1 - it\frac{x}{V_n} + \frac{1}{2}t^2\frac{x^2}{V_n^2} \right| d|F_k - \Phi_k| \\
&\quad + \sum_{k=1}^n \int_{|x|> \varepsilon V_n} \left| e^{\frac{itx}{V_n}} - 1 - it\frac{x}{V_n} + \frac{1}{2}t^2\frac{x^2}{V_n^2} \right| d|F_k - \Phi_k| \\
&\quad \le \sum_{k=1}^n \int_{|x|\le \varepsilon V_n} \frac{1}{6}|t|^3 \frac{|x|^3}{V_n^3} d|F_k - \Phi_k| + \sum_{k=1}^n \int_{|x|>\varepsilon V_n} t^2 \frac{x^2}{V_n^2} d|F_k - \Phi_k| \\
&\qquad\qquad (\textit{using } (3.25) \textit{ and } (3.26)) \\
&\quad \le \frac{1}{6}|t|^3 \varepsilon \frac{1}{V_n^2} \sum_{k=1}^n \int_{|x|\le \varepsilon V_n} x^2 d|F_k - \Phi_k)| + t^2 \frac{1}{V_n^2} \sum_{k=1}^n \int_{|x|>\varepsilon V_n} x^2 d|F_k - \Phi_k| \\
&\quad \le \frac{1}{6}|t|^3 \varepsilon \frac{1}{V_n^2} \sum_{k=1}^n 2\sigma_k^2 + t^2 \frac{1}{V_n^2} \sum_{k=1}^n \int_{|x|>\varepsilon V_n} x^2 d|F_k - \Phi_k| \\
&\quad = \frac{1}{3}|t|^3 \varepsilon + t^2 \frac{1}{V_n^2} \sum_{k=1}^n \int_{|x|>\varepsilon V_n} x^2 d|F_k - \Phi_k| \\
&\quad \overset{n\to\infty}{\longrightarrow} \frac{1}{3}|t|^3 \varepsilon \quad (\textit{using } (3.21)),
\end{aligned}
$$

which, by the arbitrariness of ε, implies that

$$|f_n(t) - \phi(t)| \to 0.$$

The result of the theorem now follows from Theorem 2.11. □

3.4 Weak convergence on $\mathbb{R}$

Weak convergence is a very rich and important topic in measure theory. We discuss only facts related to probability measures on $\mathbb{R}$ and refer the reader to standard text-

books for more comprehensive treatments.

From Definition 3.1 we know that weak convergence of v_n to v means that $E[f(v_n)] \to E[f(v)]$ (or, equivalently, $\int_{\mathbb{R}} f dP_n \to \int_{\mathbb{R}} f dP$, where P_n and P are the image probabilities of v_n and v) for any continuous and bounded function $f : \mathbb{R} \to \mathbb{R}$. Thus, weak convergence is in fact a property of the image probabilities of the random variables. Making a step towards generality, it should not come as a surprise that weak convergence can also be directly defined for probability measures that are not necessarily the image probabilities of given random variables.

DEFINITION 3.20 (weak convergence of probability measures on $\mathbb{R}$)
A probability measure sequence P_n on $\mathbb{R}$ converges weakly to a probability measure P ($P_n \xrightarrow{w} P$) if, for any continuous and bounded function $f : \mathbb{R} \to \mathbb{R}$, it holds that

$$\int_{\mathbb{R}} f dP_n \to \int_{\mathbb{R}} f dP. \tag{3.28}$$

□

When $P_n \xrightarrow{w} P$, we also write $F_n \xrightarrow{w} F$, where F_n and F are the distribution finctions associated to P_n and P.

Thus, weak convergence of v_n to v can be rephrased by saying that the image probability of v_n converges weakly to the image probability of v.

Convergence (3.28) is softened by the smoothing properties of integral and, as a consequence, weak convergence can take place among probabilities of very different nature, as illustrated in the next example.

EXAMPLE 3.21 *Let P_n be the discrete probability with equal mass concentrated in $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}$ (its distribution is in Figure 3.2.) It follows from Theorem 3.22 below that $P_n \xrightarrow{w} \lambda[0,1]$, the uniform distribution in $[0,1]$. For any n, P_n is a discrete distribution and we see that it converges to a distribution of totally different nature: P is an absolutely continuous distribution.*

□

The next theorem relates weak convergence on $\mathbb{R}$ to the behavior of distribution functions.

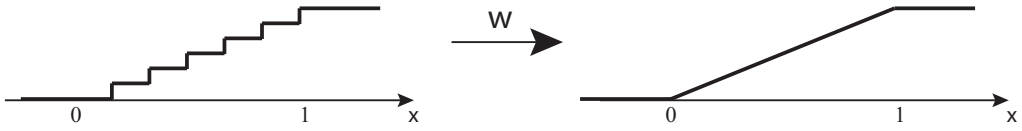

Figure 3.2: An example of weak convergence

THEOREM 3.22 *Let P_n and P be probability measures on $\mathbb{R}$. Then $P_n \xrightarrow{w} P$ if and only if the distribution $F_n(x)$ of P_n converges to the distribution $F(x)$ of P in every x where $F(x)$ is continuous.*

In general, $P_n \xrightarrow{w} P$ does not imply that $F_n(x) \not\to F(x)$ for those x where F is not continuous. One example is shown in Figure 3.3 where P_n is the concentrated mass in $\frac{1}{n}$ and P is the concentrated mass in 0 (show that $P_n \xrightarrow{w} P$.) Clearly, $F_n(0) = 0 \not\to F(0) = 1$.

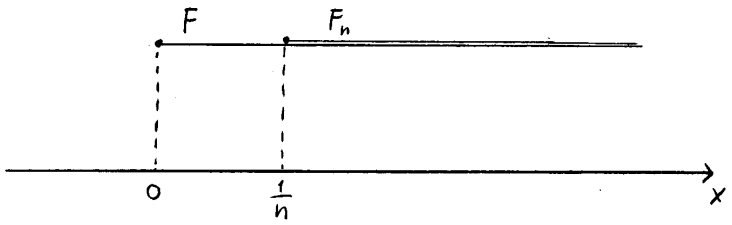

Figure 3.3:

PROOF.

$\mathbf{P_n \xrightarrow{w} P \Rightarrow F_n(x) \xrightarrow{w} F(x)$, for any x : $\mathbf{F(x)}$ continuous)

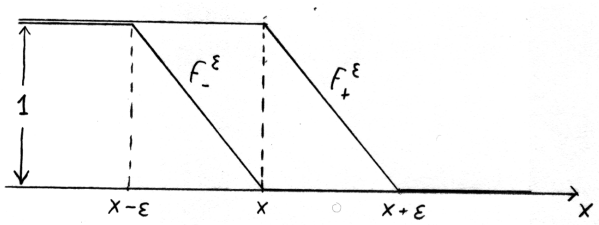

Figure 3.4: ELIMINATE SEMICOLUMN

With x a given point where F is continuous and f_+^ε and f_-^ε as represented in Figure 3.4, we have:

$$\limsup_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} \int_{\mathbb{R}} f_+^{\varepsilon} dP_n = \int_{\mathbb{R}} f_+^{\varepsilon} dP \le F(x+\varepsilon) \xrightarrow{\varepsilon\downarrow 0} F(x);$$

$$\liminf_{n\to\infty} F_n(x) \ge \liminf_{n\to\infty} \int_{\mathbb{R}} f_-^{\varepsilon} dP_n = \int_{\mathbb{R}} f_-^{\varepsilon} dP \ge F(x-\varepsilon) \xrightarrow{\varepsilon\downarrow 0} F(x),$$

so that $F_n(x) \to F(x)$.

$\mathbf{F_n(x) \xrightarrow{w} F(x)}$ at every x where $\mathbf{F(x)}$ is continuous $\Rightarrow$ $\mathbf{P_n \xrightarrow{w} P}$)

Let $A \in \mathscr{B}(\mathbb{R})$ be a set in $\mathbb{R}$ with $P(\partial A) = 0$ (∂A is the boundary of A: $\partial A = (closure\ A) \cap (closure\ A^c)$.) We want to prove that

$$P_n(A) \to P(A).$$

Let $A_0 = interior\ A$ ("*interior A*" is the set of points x in A such that $x \in (a,b)$ for some $(a,b) \subseteq A$.) Since A_0 is open, it can be represented as the union of disjoint open intervals: $A_0 = \cup_{k=1}^{\infty}(a_k,b_k)$. Choose an $\varepsilon > 0$ and, for each interval (a_k,b_k) select a subinterval $(a_k',b_k']$ such that a_k' and b_k' are points where $F(x)$ is continuous and $P(a_k,b_k) \le P(a_k',b_k'] + 2^{-k}\cdot\varepsilon$ (since $F(x)$ has at most finitely many discontinuities, such a_k' and b_k' certainly exist.) Also, fix q such that $P(A_0 - \cup_{k=1}^{q}(a_k,b_k)) \le \varepsilon$. Now,

$$\begin{aligned}
P(A_0) &= P(\cup_{k=1}^{\infty}(a_k,b_k)) = \sum_{k=1}^{\infty} P(a_k,b_k) \le \sum_{k=1}^{q} P(a_k,b_k) + \varepsilon \\
&\le \sum_{k=1}^{q} (P(a_k',b_k'] + 2^{-k}\cdot\varepsilon) + \varepsilon = \sum_{k=1}^{q} (F(b_k') - F(a_k') + 2^{-k}\cdot\varepsilon) + \varepsilon \\
&\le \sum_{k=1}^{q} (F(b_k') - F(a_k')) + 2\varepsilon = \sum_{k=1}^{q} \lim_{n\to\infty} (F_n(b_k') - F_n(a_k')) + 2\varepsilon \\
&= \lim_{n\to\infty} \sum_{k=1}^{q} P_n(a_k',b_k'] + 2\varepsilon \le \liminf_{n\to\infty} P_n(A_0) + 2\varepsilon,
\end{aligned}$$

which, due to the arbitrariness of ε, gives

$$P(A_0) \le \liminf_{n\to\infty} P_n(A_0). \tag{3.29}$$

The same derivation can be applied to A_0^c (the interior of the complement of A) leading to

$$P(A_0^c) \le \liminf_{n\to\infty} P_n(A_0^c). \tag{3.30}$$

Moreover, it holds that

$$P(\partial A) = 0 \leq \liminf_{n\to\infty} P_n(\partial A), \tag{3.31}$$

where $P(\partial A) = 0$ is by assumption and $0 \leq \liminf_{n\to\infty} P_n(\partial A)$ is just because a probability cannot be negative. From (3.29), (3.30), and (3.31) it follows that

$$P(A_0) = \lim_{n\to\infty} P_n(A_0), \quad P(A_0^c) = \lim_{n\to\infty} P_n(A_0^c), \quad 0 = \lim_{n\to\infty} P_n(\partial A). \tag{3.32}$$

To prove this, suppose that one of these equations is false. Say the first one. Using (3.29) we have

$$P(A_0) \leq \liminf_{n\to\infty} P_n(A_0) \leq \limsup_{n\to\infty} P_n(A_0),$$

and equality cannot hold throughout since $P(A_0) = \liminf_{n\to\infty} P_n(A_0) = \limsup_{n\to\infty} P_n(A_0)$ implies that $\lim_{n\to\infty} P_n(A_0)$ exists and it is equal to $P(A_0)$. Thus, it must be that $P(A_0) \leq \limsup_{n\to\infty} P_n(A_0)$. Now, this latter fact gives us the possibility of extracting a subsequence n_k such that $P(A_0) < \lim_{k\to\infty} P_{n_k}(A_0)$. But this, used together (3.30) and (3.31), leads to an absurd inequality:

$$\begin{aligned}
1 &= P(A_0 \cup A_0^c \cup \partial A) = P(A_0) + P(A_0^c) + P(\partial A) \\
&< \lim_{k\to\infty} P_{n_k}(A_0) + \liminf_{n\to\infty} P_n(A_0^c) + \liminf_{n\to\infty} P_n(\partial A) \\
&\leq \lim_{k\to\infty} P_{n_k}(A_0) + \liminf_{k\to\infty} P_{n_k}(A_0^c) + \liminf_{k\to\infty} P_{n_k}(\partial A) \\
&\leq \liminf_{k\to\infty} (P_{n_k}(A_0) + P_{n_k}(A_0^c) + P_{n_k}(\partial A)) = \liminf_{k\to\infty} P_{n_k}(A_0 \cup A_0^c \cup \partial A) \\
&= \liminf_{k\to\infty} 1 = 1.
\end{aligned}$$

Thus, our initial assumption that the first equation in (3.32) was false cannot be correct; proceeding similarly for the other equations in (3.32), (3.32) remains proven.

Result (3.29) now immediately follows from (3.32):

$$\begin{aligned}
\lim_{n\to\infty} P_n(A) &= \lim_{n\to\infty} P_n(A_0 \cup \partial A) = \lim_{n\to\infty} P_n(A_0) + \lim_{n\to\infty} P_n(\partial A) \\
&= P(A_0) + 0 = P(A_0) + P(\partial A) = P(A_0 \cup \partial A) = P(A).
\end{aligned}$$

We now move to prove that $P_n \xrightarrow{w} P$.

Pick any continuous and bounded $f : \mathbb{R} \to \mathbb{R}$ and a number M such that $|f(x)| < M$, $\forall x$. Let $T := \{t \in [-M,M] \textit{ such that } P\{f^{-1}(t)\} \neq 0\}$. T is at most countable.

Consider a partition of $[-M,M]$

$$-M = t_1 < t_2 < \cdots < t_q = M,$$

with $t_k \notin T$, $k = 1,2,\ldots,q$, and let $A_k = f^{-1}[t_k, t_{k+1})$ (i.e. A_k is the set where $f \in [t_k, t_{k+1})$.) Since f is continuous, f^{-1} of the open interval (t_k, t_{k+1}) is open, so that $\partial A_k \subseteq f^{-1}\{t_k\}$ and, being $t_k \in T$, $P(\partial A_k) = 0$, $k = 1,\ldots,q$. So,

$$\begin{aligned}
&\left| \int_{\mathbb{R}} f dP_n - \int_{\mathbb{R}} f dP \right| \\
&\quad \leq \quad \left| \int_{\mathbb{R}} f dP_n - \sum_{k=1}^{q-1} t_k P_n(A_k) \right| + \left| \sum_{k=1}^{q-1} t_k P_n(A_k) - \sum_{k=1}^{q-1} t_k P(A_k) \right| + \left| \sum_{k=1}^{q-1} t_k P(A_k) - \int_{\mathbb{R}} f dP \right| \\
&\quad \leq \quad 2 \max_{1 \leq k \leq q-1} (t_k - t_{k-1}) + \left| \sum_{k=1}^{q-1} t_k (P_n(A_k) - P(A_k)) \right| \\
&\quad \stackrel{n \to \infty}{\longrightarrow} \quad 2 \max_{1 \leq k \leq q-1} (t_k - t_{k-1}),
\end{aligned}$$

where we have used (3.29) when taking the limit. Since the t_k's can be selected to be one close to the next at will, $\max_{1 \leq k \leq q-1}(t_k - t_{k-1})$ can be made arbitrarily small, leading to the conclusion that $\int_{\mathbb{R}} f dP_n \to \int_{\mathbb{R}} f dP$ as $n \to \infty$, and this ends the proof. □

Suppose we are given a sequence of probability distributions F_n on $\mathbb{R}$. It is true that we can certainly extract a subsequence F_{n_k} converging weakly to some probability F? The answer is no, as it can be readily verified by taking F_n to be the concentrated mass in n.

The reason why here F_{n_k} fails to exist is that the mass escapes to infinity. It is an important fact that if such an "escape to infinity" behavior does not take place, then the F_n's are packed in such a way that a converging F_{n_k} certainly exists, as stated in the next theorem (the theorem is a particular case of Prokhorov's theorem, which is valid in generic metric spaces.)

THEOREM 3.23 (Helly) *A sequence of probability distributions F_n on $\mathbb{R}$ is tight if*

$$\sup_n \int_{|x| \geq M} dF_n(x) \to 0, \quad \textit{as } M \to \infty.$$

Given a tight sequence of probability distributions F_n, there always exists a subsequence F_{n_k} of F_n which converges weakly to a limit probability distribution F.

PROOF. Let $X = \{x_j, j = 1,2,\ldots\}$ be a countable dense subset of $\mathbb{R}$.

Since $F_n(x_1) \in [0,1]$, we can find a sequence of integers $n^{(1)} := \{n_1^{(1)}, n_2^{(1)}, \ldots\}$ such that $F_{n_k^{(1)}}(x_1)$ is convergent. Let us denote by $\bar{F}(x_1)$ the limiting value. We now restrict attention to sequence $F_{n_k^{(1)}}$ and extract a subsequence $n^{(2)} := \{n_1^{(2)}, n_2^{(2)}, \ldots\}$ of $n^{(1)}$ such that $F_{n_k^{(2)}}(x_2)$ converges to a limiting value, say $\bar{F}(x_2)$. Clearly, being $F_{n_k^{(2)}}$ a subsequence of $F_{n_k^{(1)}}$, $F_{n_k^{(2)}}(x_1)$ still converges to $\bar{F}(x_1)$. Proceeding along this scheme, we keep constructing nested subsequences of the original sequence F_n with the property that they converge on an increasing number of points x_j.

Now, consider the "diagonal" sequence of integers $n_k := n_k^{(k)}$. Then, for each x_j we have

$$F_{n_k}(x_j) \to \bar{F}(x_j).$$

F can now be defined based on $\bar{F}$. For all $x \in \mathbb{R}$ let

$$F(x) = \inf_{x_j > x} \bar{F}(x_j).$$

It is not difficult (though it requires some verification) to show that the so constructed F is a probability distribution (in particular, $\lim_{x\to\infty} F(x) = 1$ follows from the tightness condition which guarantees that $F_n(x) > 1 - \varepsilon$, $\forall n$, for x large enough) and that $F_{n_k}(x) \to F(x)$ for all x where F is continuous. Then $F_{n_k} \xrightarrow{w} F$ by virtue of Theorem 3.22 and this completes the proof. □

Chapter 4

THE PROJECTION THEOREM

4.1 Hilbert spaces

Since the natural setting of the projection theorem is a Hilbert space, Hilbert spaces are introduced first. A Hilbert space is an inner product space (with a completeness property), and in turn a inner product space is a vector space (with an inner product.) We here introduce these notions in a bottom-up fashion, from vector spaces to Hilbert spaces.

DEFINITION 4.1 (vector space) *A complex vector space (or a vector space over the complex field $\mathbb{C}$) is a set V, in which two operations are defined. The first one, called addition, applies two any pair of vectors $x, y \in V$ and returns a vector $x+y \in V$, while the second one, called scalar multiplication, applies to a pair α, x, with $\alpha \in \mathbb{C}$ and $x \in V$, and returns a vector $\alpha \cdot x \in V$. The following rules apply to the two operations:*

(a) $x+y = y+x$;
(b) $x+(y+z) = (x+y)+z$;
(c) there exists a vector 0 *(the zero vector) such that* $x+0 = x$, $\forall x \in V$;
(d) to each $x \in V$, *there corresponds a vector* $-x$ *(called the opposite vector) such that* $x+(-x) = 0$;
(e) $1 \cdot x = x$, $\forall x \in V$;
(f) $\alpha \cdot (\beta \cdot x) = (\alpha\beta) \cdot x$;
(g) $\alpha \cdot (x+y) = \alpha \cdot x + \alpha \cdot y$;
(h) $(\alpha+\beta) \cdot x = \alpha \cdot x + \beta \cdot x$. □

Remark 4.2 *The "purist" may find the above definition slightly imprecise: a set V remains just a set V even if we attach two operations to it, so that it appears*

unjustified to call it a "vector space" instead of a "set". For the sake of precision, we note that the vector space is the triple $(V,+,\cdot)$, and the sentence "V is a vector space" is short for "$(V,+,\cdot)$ is a vector space", where the operations are either clear from the context or their specification is unimportant. □

We derive some immediate consequences of Definition 4.1.

(I) Because of (b), we can write $x+y+z$ with no ambiguity;

(II) there is no other vector y besides the zero vector such that $x+y=x$, $\forall x \in V$. Indeed, take $x=0$ in relation $x+y=x$ and write:

$$\begin{aligned} 0 &= 0+y \\ &= y+0 \\ &= y, \end{aligned}$$

showing that y has to be the zero vector;

(III) $0\cdot x = 0, \forall x \in V$ (note that 0 in the left hand side is the number zero, while 0 in the right hand side is the vector zero.) In fact:

$$\begin{aligned} y+0\cdot x &= y+0\cdot x+1\cdot x+(-x) \\ &= y+(0+1)\cdot x+(-x) \\ &= y+1\cdot x+(-x) \\ &= y, \quad \forall y \in V, \end{aligned}$$

so, by the uniqueness of the zero vector shown in (II), $0\cdot x$ has to be 0;

(IV) the opposite of a vector x is unique. Suppose there are two: x_1 and x_2. Then,

$$\begin{aligned} x_1 &= x_1+x+(-x) \\ &= (x_1+x)+(-x) \\ &= 0+(-x) \\ &= (x_2+x)+(-x) \\ &= x_2+x+(-x) \\ &= x_2, \end{aligned}$$

so that $x_1 = x_2$;

(V) $-x = -1 \cdot x, \forall x$. In fact:

$$\begin{aligned} x + (-1) \cdot x &= (1-1) \cdot x \\ &= 0 \cdot x \\ &= 0 \quad (\textit{using } (III)). \end{aligned}$$

Since the opposite vector is unique (see (IV)), the conclusion follows.

For short, $x + (-y)$ is also written $x - y$.

DEFINITION 4.3 (inner product space) *A inner product space is a vector space V where, to each ordered pair of vectors x and y, there is associated a complex number (x,y) called the inner product (or the scalar product) of x and y, with the following properties:*

(i) $(x,y) = \overline{(y,x)}$ *(¯ denotes complex conjugation);*
(l) $(x+y,z) = (x,z) + (y,z)$*;*
(m) $(\alpha \cdot x, y) = \alpha(x,y)$*;*
(n) $(x,x) \geq 0$*, and* $(x,x) = 0$ *implies* $x = 0$. □

A vector space over the real field $\mathbb{R}$ is defined identically to a complex vector space, except that α and β are real numbers. In this case, the scalar product (x,y) is a real number too. Throughout, we use notations for the complex case, particularization to the real case is straightforward.

EXAMPLE 4.4 ($\mathbb{R}^m$) *For any fixed m, the set $\mathbb{R}^m$ of real valued p-dimensional vectors with addition and scalar multiplication defined in the usual componentwise manner is a real vector space. It becomes a inner product space by the definition:*

$$(x,y) = \sum_{k=1}^{m} x_k y_k,$$

where x_k [y_k] are the components of vector x [y]. □

EXAMPLE 4.5 ($\mathbb{L}^2$) *Consider the set of square integrable random variables (i.e. random variables ξ with $E[\xi^2] < \infty$) defined over a probability space $(\Omega, \mathscr{F}, P)$. This set is indicated with $\mathbb{L}^2$. With the usual operations of addition and scalar multiplication, $\mathbb{L}^2$ is a real vector space.*

We can try to endow $\mathbb{L}^2$ with an inner product by defining

$$(\xi,\eta) = E[\xi\eta].$$

Along this line, however, a difficulty is encountered. Such a difficulty is merely technical, and we prefer to make it explicit to remove any possibility of confusion.

Conditions (i), (l), (m) and the first part of (n) ($(x,x) \geq 0$) in Definition 4.3 are easily verified. The problem shows up for the second part of (n): $(x,x) = 0$ implies $x = 0$. In fact, condition $(\xi,\xi) = E[\xi^2] = 0$ does not imply that $\xi = 0$, $\forall\omega \in \Omega$, the 0 vector in $\mathbb{L}^2$; it only implies that $\xi = 0$ almost surely. Thus, the condition in (n) of Definition 4.3 that $(x,x) = 0$ implies $x = 0$ is violated in this case. Nevertheless, we insist with the definition $(\xi,\eta) = E[\xi\eta]$ with the understanding that $\mathbb{L}^2$ is not a inner product space in the standard sense of Definition 4.3. It is instead a generalized inner product space where (n) is substituted by:

(n') $(x,x) \geq 0$, and $(x,x) = 0$ implies $x \in Z$, where Z is a set containing the 0 vector.

Z has to be interpreted as the set of vectors that are indistinguishable from 0 in the adopted inner product; in $\mathbb{L}^2$, it is the set of random variables $\xi = 0$ almost surely. Adopting this generalized point of view only introduces minor modifications in the theory of inner product spaces as it will be explicitely indicated when $\mathbb{L}^2$ will be reconsidered at later stages in this chapter.

It is worth mentioning that a different route can also be adopted to fix this difficulty in an alternative way: instead of viewing $\mathbb{L}^2$ as a space of random variables, it can be seen as a space of equivalence classes of random variables, where each class contains all variables that differ only on a zero probability set. This corresponds to a coarser-grained viewpoint where one aggregates all variables that are almost surely coincident. This approach, however, introduces some extra complications with measurability issues and we prefer to adopt the first approach where condition (n') substitutes (n). Still, it should be clear that this choice is purely utilitarian, and has no conceptual motivation. □

EXAMPLE 4.6 ($\mathbf{C}[\mathbf{0},\mathbf{1}]$) *The set of continuous real functions defined on $[0,1]$ with the usual addition and scalar multiplication operations is a real inner product space by the definition*

$$(f,g) = \int_0^1 f(r)g(r)dr.$$

□

Geometry of inner product spaces

Quantity

$$\|x\| := \sqrt{(x,x)}$$

is called the norm of vector x.

PARALLELOGRAM LAW 4.7

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2 \qquad \text{(see Figure 4.1)}$$

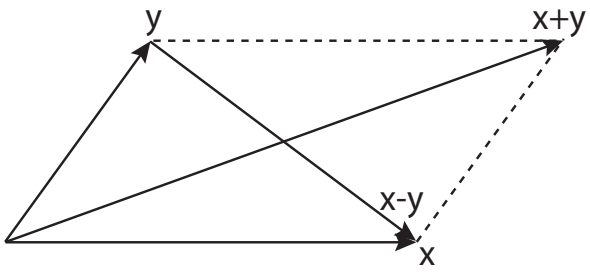

Figure 4.1: The parallelogram law

PROOF. Note first that

$$\begin{aligned} (x,-y) &= (x,-1\cdot y) \quad (\textit{using } (V)) \\ &= -1\overline{(y,x)} \quad (\textit{using } (i) \textit{ and } (m)) \\ &= -(x,y). \end{aligned} \tag{4.1}$$

and, similarly,

$$(-x,-y) = (x,y). \tag{4.2}$$

By the properties of the inner product and (4.1) and (4.2), we then have

$$\begin{aligned} &\|x+y\|^2 + \|x-y\|^2 \\ &= (x+y,x+y) + (x-y,x-y) \\ &= \|x\|^2 + \|y\|^2 + (x,y) + (y,x) + \|x\|^2 + \|y\|^2 - (x,y) - (y,x) \\ &= 2\|x\|^2 + 2\|y\|^2. \end{aligned}$$

□

SCHWARZ INEQUALITY 4.8

$$|(x,y)| \leq \|x\|\|y\| \qquad \text{(see Figure 4.2)} \tag{4.3}$$

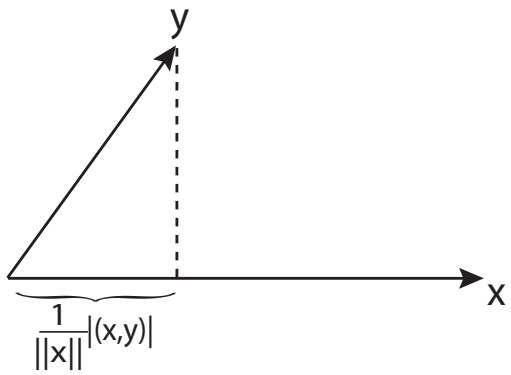

Figure 4.2: The Schwarz inequality

PROOF. If $\|x\| = 0$, then $x = 0$ (use (n)), so that $(x,y) = (0,y) = (0 \cdot x, y) = 0(x,y) = 0$, and (4.3) is true. Suppose $\|x\| \neq 0$, then

$$\begin{aligned} 0 &\leq \left(y - \frac{(y,x)}{\|x\|^2}x, y - \frac{(y,x)}{\|x\|^2}x\right) \\ &= \|y\|^2 + \frac{|(y,x)|^2}{\|x\|^2} - \frac{(y,x)}{\|x\|^2}(x,y) - \frac{\overline{(y,x)}}{\|x\|^2}(y,x) \\ &= \|y\|^2 - \frac{|(x,y)|^2}{\|x\|^2}, \end{aligned}$$

from which (4.3) follows. □

THE TRIANGLE INEQUALITY 4.9

$$\|x+y\| \leq \|x\| + \|y\| \qquad \text{(see Figure 4.3)} \tag{4.4}$$

PROOF.

$$\begin{aligned} \|x+y\|^2 &= (x+y, x+y) \\ &= \|x\|^2 + \|y\|^2 + (x,y) + (y,x) \\ &\leq \|x\|^2 + \|y\|^2 + 2|(x,y)| \\ &\leq \|x\|^2 + \|y\|^2 + 2\|x\|\|y\| \quad (\textit{using } (4.3)) \\ &= (\|x\| + \|y\|)^2, \end{aligned}$$

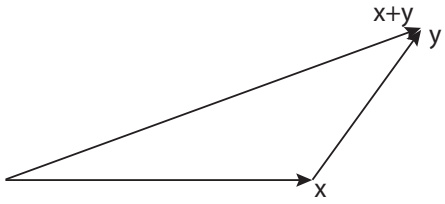

Figure 4.3: The triangle inequality

which implies (4.4). □

DEFINITION 4.10 (orthogonality) *We say that two vectors x and y are orthogonal (and write $x \perp y$) if* $(x,y) = 0$. □

PITAGORA'S THEOREM 4.11

$$\text{If } x \perp y, \text{ then } \|x+y\|^2 = \|x\|^2 + \|y\|^2 \qquad \text{(see Figure 4.4)}$$

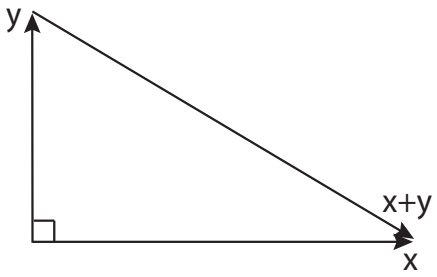

Figure 4.4: The Pitagora's theorem

PROOF.

$$\begin{aligned} \|x+y\|^2 &= (x+y, x+y) \\ &= \|x\|^2 + \|y\|^2 + (x,y) + (y,x) \\ &= \|x\|^2 + \|y\|^2. \end{aligned}$$

□

Topology in inner product spaces

DEFINITION 4.12 (distance) *Given any two vectors x and y, the quantity*

$$\rho(x,y) := \|x-y\|$$

is called the distance between x and y. □

It is easy to verify that $\rho(\cdot,\cdot)$ in the above definition is indeed a distance, that is it satisfies the following usual properties of a distance:

$\rho(x,y) \geq 0$ and $\rho(x,y) = 0$ implies $x = y$;
$\rho(x,y) = \rho(y,x)$;
$\rho(x,y) \leq \rho(x,z) + \rho(z,y)$.

We say that a sequence x_n, $n = 1,2,\cdots$, converges to x (and write $x_n \to x$) if $\rho(x_n,x) \to 0$. It is easy to see that the limit is unique (if $x_n \to x$ and $x_n \to y$, then $x = y$.)

THEOREM 4.13 *If $x_n \to x$, then $(x_n,y) \to (x,y)$, $\forall y$ (this is expressed in words by saying that the inner product is a continuous operator.)*

PROOF. By assumption $\|x_n - x\| = \rho(x_n,x) \to 0$. Then, by the Schwarz inequality 4.8,

$$\begin{aligned} |(x_n,y)-(x,y)| &= |(x_n-x,y)| \\ &\leq \|x_n-x\|\|y\| \\ &\to 0. \end{aligned}$$

□

Hilbert spaces

A sequence x_n is said to be a Cauchy (or "fundamental") sequence if, $\forall \varepsilon > 0, \exists N(\varepsilon)$ such that $\forall p,q \geq N(\varepsilon)$ it holds that $\rho(x_p,x_q) \leq \varepsilon$. The following definition is fundamental.

DEFINITION 4.14 (completeness - Hilbert space) *A inner product space V is complete if every Cauchy sequence converges to a limit point in V. A complete inner product space is also called a Hilbert space.* □

In words, a Cauchy sequence is a sequence where any two vectors in its tail are arbitrarily close to each other. If any such sequence converges, we then say that the space is complete.

The following examples illustrate the concept of completeness.

EXAMPLE 4.15 (Example 4.4 continued - $\mathbb{R}^m$ is complete) *Suppose $\{x_n \in \mathbb{R}^m\}$ is a Cauchy sequence, that is $\forall \varepsilon > 0, \exists N(\varepsilon)$ such that $\forall p,q \geq N(\varepsilon)$ it holds that $\rho(x_p,x_q) = \|x_p - x_q\| = \left(\left(\sum_{k=1}^m (x_{pk} - x_{qk})^2\right)^{1/2} \leq \varepsilon$ (the indexes k identify components.) This relation implies $|x_{pk} - x_{qk}| \leq \varepsilon$, $k = 1,2,\ldots,m$, so that $x_{nk} \to x_k$ for some $x_k \in \mathbb{R}$ (since the real line is complete, see any text on real analysis), from which $x_n \to x$, where x is the vector in $\mathbb{R}^m$ whose components are x_k.* □

EXAMPLE 4.16 (Example 4.5 continued - $\mathbb{L}^2$ is complete) *In standard inner product spaces where condition (n) in Definition 4.3 holds, if the limit of a sequence exists, then it is unique, and Definition 4.14 requires that such a limit point actually exists for any Cauchy sequence. In the $\mathbb{L}^2$ space, property (n) has been substituted by property (n') in Example 4.5. Consequently, if $\xi_n \to \xi$ in $\mathbb{L}^2$, ξ_n also tends to any other variable $\xi + \eta$, with $\eta = 0$ almost surely, and the uniqueness of the limit is lost. In this context, by completeness it is meant that any Cauchy sequence has (at least) one limit point. If so, all other variables almost surely equal to this limit point will be limit points as well.*

Let us then prove that the limit point of a Cauchy sequence always exists. Suppose ξ_n is a Cauchy sequence in $\mathbb{L}^2$, so that $\forall \varepsilon > 0, \exists N(\varepsilon)$ such that $\forall p,q \geq N(\varepsilon)$ we have $E[(\xi_p - \xi_q)^2]^{1/2} \leq \varepsilon$. Then, it is possible to find a sequence of indexes n_k such that

$$E[(\xi_{n_{k+1}} - \xi_{n_k})^2]^{1/2} \leq \frac{1}{2^k}. \tag{4.5}$$

To this end, let $n_1 = N(\frac{1}{2})$, so that, for any choice of n_2, $E[(\xi_{n_2} - \xi_{n_1})^2]^{1/2} \leq \frac{1}{2}$ holds. Then, take $n_2 = N(\frac{1}{4})$ and so on with $n_3, n_4, \ldots$. For sequence ξ_{n_k} we then have

$$\begin{aligned}
\sum_{k=1}^{\infty} E[|\xi_{n_{k+1}} - \xi_{n_k}|] &\leq \sum_{k=1}^{\infty} E[(\xi_{n_{k+1}} - \xi_{n_k})^2]^{1/2} \quad (\textit{using Schwarz inequality } 4.8) \\
&\qquad\qquad \textit{with } x = 1 \textit{ and } y = |\xi_{n_{k+1}} - \xi_{n_k}|\ ) \\
&\leq \sum_{k=1}^{\infty} \frac{1}{2^k} \quad (\textit{using } (4.5)) \\
&= 1. \qquad (4.6)
\end{aligned}$$

Based on (4.6), we first prove that ξ_{n_k} is almost surely convergent to some random variable ξ and then that ξ is the $\mathbb{L}^2$-limit of the initial sequence ξ_n, which shows that a Cauchy sequence in $\mathbb{L}^2$ has limit, so concluding the proof.

Consider the sequence

$$\zeta_k := |\xi_{n_1}| + |\xi_{n_2} - \xi_{n_1}| + \cdots + |\xi_{n_k} - \xi_{n_{k-1}}|.$$

This sequence is increasing and, therefore, convergent (either to a finite value or to infinity.) The set A where $\zeta_k \to \infty$ can be expressed as $A = \cap_{p=1}^{\infty} \cup_{k=1}^{\infty} \{\zeta_k \geq p\}$. Since $\{\zeta_k \geq p\}$ is measurable and a σ-algebra is closed under countable union and intersection, we have that A is measurable too. Moreover, (4.6) implies that $P(A) = 0$ (why?). Consider now

$$\xi_{n_k} = \xi_{n_1} + (\xi_{n_2} - \xi_{n_1}) + \cdots + \left(\xi_{n_k} - \xi_{n_{k-1}}\right).$$

Certainly, ξ_{n_k} converges everywhere on A^c (the complement of A) since a sequence is always convergent whenever the corresponding absolute sequence is convergent to a finite value. Let ξ be the limit. On A, define $\xi = 0$. Then, ξ is a random variable (i.e. it is measurable) as it follows from point (ii) of Theorem 3.5.

Finally, we show that $\xi_n \to \xi$ in $\mathbb{L}^2$. Fix $\varepsilon > 0$ and an integer $n \geq N(\varepsilon)$ and let

$$\eta_j := in f_{k \geq j} \left(\xi_{n_k} - \xi_n\right)^2.$$

Clearly, $\eta_j \leq (\xi_{n_j} - \xi_n)^2$, so that

$$E[\eta_j] \leq \varepsilon^2, \quad \textit{for any } j \textit{ large enough}. \tag{4.7}$$

On the other hand, $\eta_j \uparrow (\xi - \xi_n)^2$ almost surely as $j \to \infty$ and $\eta_j \geq 0$, which, by the monotone convergence Theorem 3.7, implies

$$E[\eta_j] \uparrow E\left[(\xi - \xi_n)^2\right]. \tag{4.8}$$

Putting together (4.7) and (4.8) gives $E\left[(\xi - \xi_n)^2\right] \leq \varepsilon^2$, which, by the arbitrariness of ε, entails that $\xi_n \to \xi$ in $\mathbb{L}^2$. □

EXAMPLE 4.17 (Example 4.6 continued - $\mathbf{C}[\mathbf{0},\mathbf{1}]$ is not complete) *The inner product space $C[0,1]$ of Example 4.6 is not complete. In fact the sequence*

$$f_n(x) = \begin{cases} 1, & x \in \left[0, \frac{1}{2}\right] \\ -n\left(x - \frac{1}{2}\right) + 1, & x \in \left(\frac{1}{2}, \frac{1}{2} + \frac{1}{n+1}\right] \\ 0, & otherwise, \end{cases} \tag{4.9}$$

is Cauchy (verify this), but does not converge to a continuous function. □

Subspaces

DEFINITION 4.18 (subspace) *A subspace S of a vector space V is a subset S of V which is itself a vector space, relative to the operations defined in V.* □

It is easy to verify that a subset S of a vector space V is a subspace if and only if addition and scalar multiplication of vectors in S are still in S: $x+y \in S$ if $x,y \in S$ and $\alpha \cdot x \in S$ if $x \in S$.

DEFINITION 4.19 (closed subspace) *A subspace S of a Hilbert space is closed if any convergent sequence $x_n \in S$ has limit in S.* □

EXAMPLE 4.20 *In $\mathbb{L}^2$, the limit point of a convergent sequence is not unique (see Example 4.16.) We say that S is closed if any convergent sequence has at least one limit point in S.*

The vector space $\mathbb{L}^2[0,1]$ of measurable functions defined on $[0,1]$ such that $\int_0^1 f(r)^2 dr < \infty$ is complete (in fact, this is the $\mathbb{L}^2$ space defined over $(\Omega, \mathscr{F}, P) = ([0,1], \mathscr{B}[0,1], \lambda)$ and completeness has been proven in Example 4.16.) The space $C[0,1]$ of Example 4.17 is a subspace of $\mathbb{L}^2[0,1]$, but it is not closed (to verify this, recall that (4.9) does not converge to any function in $C[0,1]$.) □

In a Hilbert space H, let $x^\perp$ denote the set of all vectors y orthogonal to x. It is easy to prove that $x^\perp$ is a closed subspace. In fact, $(y_1+y_2,x) = (y_1,x)+(y_2,x) = 0$ and $(\alpha \cdot y, x) = \alpha(y,x) = 0$, $\forall y_1, y_2, y \in x^\perp$, showing that $x^\perp$ is a subspace. Its closedness is proven by observing that if $y_n \to y$ and $y_n \in x^\perp$, then $0 = \lim_{n\to\infty}(y_n,x) = (y,x)$ (where the last equality follows from Theorem 4.13), that is $y \in x^\perp$ too.

If A is a subset of H, by the symbol $A^\perp$ we indicate the set of all vectors orthogonal to every $x \in A$. Since $A^\perp = \cap_{x\in A} x^\perp$, it is immediate to verify that $A^\perp$ is a closed subspace.

4.2 The projection theorem

THEOREM 4.21 (projection theorem) *Let S be a closed subspace of a Hilbert space H. Every vector $x \in H$ has a unique decomposition*

$$x = s + z, \tag{4.10}$$

where $s \in S$ and $z \in S^{\perp}$ (s is called the projection of x onto S.) Moreover, s is the unique vector in S nearest to x: $\|x-s\| = \min_{s' \in S} \|x-s'\|$.

PROOF. To prove the uniqueness of the decomposition, suppose that there is a second such a decomposition: $x = s_1 + z_1, s_1 \in S, z_1 \in S^{\perp}$. Then,

$$\begin{aligned} 0 &= \|x-x\|^2 \\ &= \|s-s_1+z-z_1\|^2 \\ &= \|s-s_1\|^2 + \|z-z_1\|^2 \quad (\textit{using Pitagora}'\textit{s Theorem } 4.11); \\ &\Rightarrow \|s-s_1\|^2 = 0 \ \textit{ and } \ \|z-z_1\|^2 = 0 \\ &\Rightarrow s = s_1 \ \textit{ and } \ z = z_1, \end{aligned}$$

so that the two decompositions coincide.

To prove the existence, note first that if there exists a vector $s \in S$ at nearest distance from x, then such a s gives the sought decomposition with $z := x-s$. To show this, we only have to prove that $x-s \in S^{\perp}$. Suppose not. Then, $\exists y \in S : (x-s,y) \neq 0$ and

$$\begin{aligned} \left\| x-s-\frac{(x-s,y)}{\|y\|^2}y \right\|^2 &= \left(x-s-\frac{(x-s,y)}{\|y\|^2}y, x-s-\frac{(x-s,y)}{\|y\|^2}y \right) \\ &= \|x-s\|^2 - \frac{|(x-s,y)|^2}{\|y\|^2}, \end{aligned}$$

showing that $s + \frac{(x-s,y)}{\|y\|^2}y \in S$ would be closer to x than s, so contradicting the fact that s is the vector at nearest distance.

Thus, to prove existence of the decomposition, all we need to show is that a vector $s \in S$ at nearest distance from x does exist.

Let $\delta := \inf_{s' \in S} \|x-s'\|$. Consider a vector sequence $s_n \in S$ such that $\|x-s_n\| \to \delta$. We show that s_n is a Cauchy sequence and that it converges to the sought vector s.

By the parallelogram law 4.7,

$$\begin{aligned}\|s_p - s_q\|^2 &= \|(x-s_q)-(x-s_p)\|^2 \\ &= 2\|x-s_q\|^2 + 2\|x-s_p\|^2 - 4\left\|x - \frac{s_q+s_p}{2}\right\|^2.\end{aligned}$$

Since $\frac{s_q+s_p}{2}$ belongs to S, the last term $4\left\|x-\frac{s_q+s_p}{2}\right\|^2$ is no smaller that $4\delta^2$. On the other hand, the first two terms tend to $2\delta^2$ by construction. Thus, $\|s_q - s_p\|^2$ is arbitrarily small for any p and q large enough, that is s_n is indeed a Cauchy sequence. Letting s be its limit point (which is in S by the assumption that S is closed), a straightforward application of the triangular inequality shows that $\|x-s\| = \delta$, that is s is at nearest distance: $\|x-s\| \leq \|x-s_n\| + \|s_n - s\| \to \delta + 0 = \delta$, but $\|x-s\|$ does not depend on n so that $\|x-s\| \leq \delta$; on the other hand, $\|x-s\|$ cannot be smaller than δ since δ is a lower bound to $\|x-s'\|$, $\forall s' \in S$.

Summing up, we have shown that decomposition $x = s+z$ exists and is unique. Moreover, by construction, $\|x-s\| = \delta = \min_{s'\in S}\|x-s'\|$.

The theorem statement contains a very last point: the vector $s \in S$ nearest to x is unique. This final point is readily established from what we have already proven: suppose for the purpose of contradiction that a $s_1 \neq s$ exists in S at nearest distance; then, $z_1 := x - s_1$ would belong to $S^\perp$, so providing a second decomposition $x = s_1 + z_1$. But this is in contradiction with the already proven uniqueness of the decomposition. □

Interpretation of the projection theorem

The projection theorem tells us two things.

1. If s is the projection of x onto S (i.e. $z := x - s$ is orthogonal to all vectors in S), then s minimizes the distance of the subspace S from x (see Figure 4.5.)

Thus, if we are given the projection, the problem of finding the vector in S closest to x is automatically solved.

2. In addition, the theorem tells us that such a projection actually exists (and is unique) with generality. The idea in the proof is to construct a sequence $s_n \in S$ whose distance from x tends to the minimal distance $\delta := \inf_{s'\in S}\|x-s'\|$. Such a sequence tends to accumulate (i.e. it is a Cauchy sequence), so that, by the fundamental closedness assumption of S, it converges to a vector s, and this s is the projection (see Figure 4.6.)

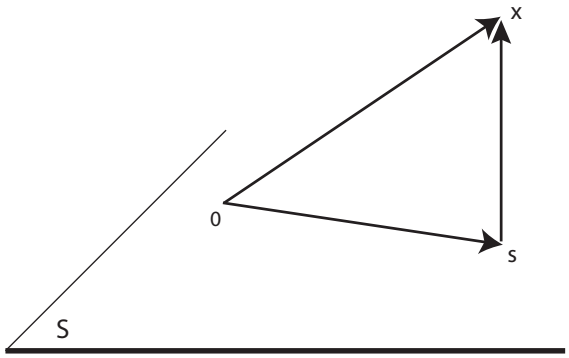

Figure 4.5: The decomposition of x

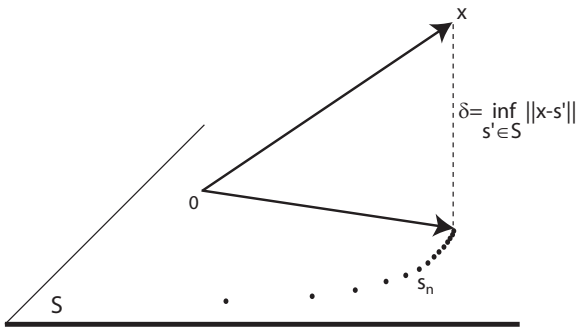

Figure 4.6: Construction of s

4.3 Applications of the projection theorem

We discuss two applications of the projection theorem. The first one refers to $\mathbb{R}^m$ and is presented mostly for didactic reasons. The second, concerning $\mathbb{L}^2$, is of great importance in estimation theory.

Application 1: $\mathbb{R}^m$

As we know, $\mathbb{R}^m$ is a Hilbert space (see Example 4.15.) Given $x \in \mathbb{R}^m$ and r ($r \leq m$) linearly independent vectors $y_1, y_2, \ldots, y_r$ in $\mathbb{R}^m$, consider the problem of finding the vector s closest to x and belonging to $S := span\{y_1, y_2, \ldots, y_r\}$, the subspace linearly spanned by $y_1, y_2, \ldots, y_r$ (this is certainly a closed subspace, since any finite dimensional subspace is closed.)

In matrix notations, vectors $s \in S$ can be expressed as

$$s = Y\alpha,$$

where Y is the matrix $[y_1\ y_2 \cdots y_r]$ stacking the y_k vectors and $\alpha \in \mathbb{R}^m$ is the vector of unknowns.

In view of the projection theorem, s is the projection of x onto S. Thus, we have to impose that $x-s$ is orthogonal to all vectors in S or, equivalently, to vectors $y_1, y_2, \ldots, y_r$:

$$(y_k, x-s) = 0, \quad k = 1,2,\ldots,q.$$

Using the definition of inner product in $\mathbb{R}^m$ and relation $s = Y\alpha$, we then have

$$y_k^T x = y_k^T Y\alpha, \quad k = 1,2,\ldots,q,$$

which can be written in a more compact form as

$$Y^T x = Y^T Y\alpha. \tag{4.11}$$

Solving this equation yields α.

Clearly, the same result can be achieved by direct minimization of $\|x-s\|$. Along this route, we write:

$$\|x-s\|^2 = x^T x + \alpha^T Y^T Y\alpha - 2x^T Y\alpha,$$

whose minimization gives again (4.11).

Application 2: $\mathbb{L}^2$

In Example 4.16, we have seen that the space $\mathbb{L}^2$ of square integrable random variables is complete. Strictly speaking, however, it is not a Hilbert space since condition (n) of Definition 4.3 has been substituted by condition (n') of Example 4.5, so that $\mathbb{L}^2$ is not a standard inner product space. In particular, this implies that in $\mathbb{L}^2$ the limit point of a convergent sequence is not unique.

In this context, we say that a subspace S is closed if any convergent sequence in S has at least one limit point is S.

By inspecting the proof of the projection Theorem 4.21, we see that the results of this theorem are still valid with one single amendment: the decomposition is no longer unique. In fact, given the decomposition $x = s+z$, any $s_1 \in S$ such that $s_1 = s$ almost surely gives another valid decomposition $x = s_1 + (z+s-s_1)$ (note that $z+s-s_1 \in S^{\perp}$.) Moreover, no other decompositions are possible besides these. For short, we express this fact by saying that the decomposition is almost surely unique. Similarly, the vector in S nearest to x is not unique and the set of points minimizing the distance is: any $s_1 \in S$ such that $s_1 = s$ almost surely.

A notable example of this decompositional construction is found when $S = \mathbb{L}^2(\mathscr{G})$, the subset of $\mathbb{L}^2$ formed by all $\mathscr{G}$-measurable random variables (here, $\mathscr{G}$ is any sub σ-algebra of $\mathscr{F}$.) Since the sum of $\mathscr{G}$-measurable random variables and their product

by a scalar α is still $\mathcal{G}$-measurable, $\mathbb{L}^2(\mathcal{G})$ is a subspace. It is in fact a closed subspace. Indeed, any Cauchy sequence v_n in $\mathbb{L}^2(\mathcal{G})$ is certainly convergent to a point v in $\mathbb{L}^2$, since $\mathbb{L}^2$ is complete. But then, by virtue of Theorem 3.6, we can determine a $\mathcal{G}$-measurable $\bar{v}$ such that $v_n \to \bar{v}$ in $\mathbb{L}^2$.

Thus, $\mathbb{L}^2(\mathcal{G})$ is a closed subspace and, by the projection theorem, any $v \in \mathbb{L}^2$ has an almost surely unique projection onto $\mathbb{L}^2(\mathcal{G})$. This projection minimizes the distance (i.e. the second order moment) from v among all variables that are $\mathcal{G}$-measurable.

For convenience, the results valid for $\mathbb{L}^2$ are summarized in the following theorem.

THEOREM 4.22 *Let S be a closed subspace of $\mathbb{L}^2$ (for example, $S = \mathbb{L}^2(\mathcal{G})$, the subset of $\mathbb{L}^2$ formed by all $\mathcal{G}$-measurable random variables.) Then, given any $v \in \mathbb{L}^2$, the projection of v onto S exists, is unique up to almost sure equivalence, that is, given a projection, all other projections are characterized as the set of all random variables in S that are almost surely equal to the given projection. Any projection in the equivalence class minimizes the distance between S and v (i.e. $E[(v - projection\ of\ v)^2] = min_{s \in S} E[(v-s)^2]$.) Moreover, the set of all vectors that minimize the distance coincides with the projection equivalence class, that is no other vector besides those in the projection equivalence class minimizes the distance.* □

Chapter 5

CONDITIONAL EXPECTATION AND CONDITIONAL DENSITY

5.1 Conditional expectation

Elementary conditional expectation

Let v be a random variable and let $\mathscr{D} = \{D_1, D_2, \ldots, D_N\}$ be a finite decomposition of the sample space Ω (that is $\Omega = \cup_{k=1}^{N} D_k$ and $D_k \cap D_j = \emptyset$ for $k \neq j$) such that $P(D_k) > 0$, $k = 1, 2, \ldots, N$. The conditional expectation of v with respect to the σ-algebra $\sigma(\mathscr{D})$ generated by $\mathscr{D}$ (that is the class of all possible unions of D_k sets plus the empty set) is defined as

$$E[v \mid \sigma(\mathscr{D})] := \sum_{k=1}^{N} \frac{E[v \cdot 1(D_k)]}{P(D_k)} \cdot 1(D_k),$$

where $1(D_k)$ denotes the indicator function of set D_k, namely $1(D_k) = 1$ for $\omega \in D_k$ and $1(D_k) = 0$ for $\omega \notin D_k$. In Figure 5.1 the conditional expectation of a random variable v defined over the sample space $\Omega = [0,1]$ is shown. The idea is that the value of $E[v \mid \sigma(\mathscr{D})]$ over each D_k is the mean value of v over the same set.

From the above definition, it is clear that

i) $E[v \mid \sigma(\mathscr{D})]$ is $\sigma(\mathscr{D})$-measurable;
ii) for any $A \in \sigma(\mathscr{D})$, $\int_A E[v \mid \sigma(\mathscr{D})]dP = \int_A v dP$.

The interpretation of i) and ii) is as follows. Because of i), we see that $E[v \mid \sigma(\mathscr{D})]$ is a simpler random variable than v (i.e. it is measurable with respect to a coarser σ-algebra than v is.) On the other hand, fact ii) says that, from the coarser-grained point of view of $\sigma(\mathscr{D})$, $E[v \mid \sigma(\mathscr{D})]$ is indistinguishable from v.

Definition of conditional expectation

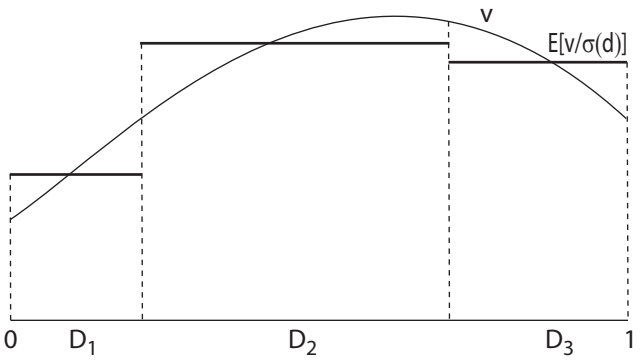

Figure 5.1: The conditional expectation in an elementary case.

In probability theory, it is often necessary to take conditional expectation with respect to non-elementary σ-algebras containing zero probability events. We here address such a generalization.

DEFINITION 5.1 (conditional expectation) *Let v be a random variable defined on a probability space $(\Omega, \mathscr{F}, P)$ such that $E[v]$ exists. Given a σ-algebra $\mathscr{G} \subseteq \mathscr{F}$, the conditional expectation of v given $\mathscr{G}$ is a random variable $E[v \mid \mathscr{G}]$ such that*

i) $E[v \mid \mathscr{G}]$ is $\mathscr{G}$-measurable;
ii) for any $A \in \mathscr{G}$, $\int_A E[v \mid \mathscr{G}]dP = \int_A v dP$. □

Below, we shall see that the conditional expectation exists in full generality. Before that, we prove that it is unique up to almost sure equivalence (i.e. two conditional expectations can only differ from each other on a zero probability set.)

Suppose that there are two conditional expectations $E[v \mid \mathscr{G}]_1$ and $E[v \mid \mathscr{G}]_2$ that satisfy i) and ii). Let $A_+ := \{\omega : E[v \mid \mathscr{G}]_1 > E[v \mid \mathscr{G}]_2\}$. Then, $A_+ \in \mathscr{G}$ and $\int_{A_+}(E[v \mid \mathscr{G}]_1 - E[v \mid \mathscr{G}]_2)dP = \int_{A_+} E[v \mid \mathscr{G}]_1 dP - \int_{A_+} E[v \mid \mathscr{G}]_2 dP = \int_{A_+} v dP - \int_{A_+} v dP = 0$, which implies that $P(A_+) = 0$ since the integrand $(E[v \mid \mathscr{G}]_1 - E[v \mid \mathscr{G}]_2)$ in the first integral is strictly positive over A_+. Similarly, $P(A_-) = P\{\omega : E[v \mid \mathscr{G}]_1 < E[v \mid \mathscr{G}]_2\} = 0$, so that $E[v \mid \mathscr{G}]_1 = E[v \mid \mathscr{G}]_2$ almost surely.

Each random variable satisfying i) and ii) in Definition 5.1 is a "version" of the conditional expectation. To many purposes, specifying the version is immaterial and it is customary to say: "let us consider the conditional expectation $E[v \mid \mathscr{G}]$ of v given $\mathscr{G}$" as a shorthand for "let us consider a version of the conditional expectation $E[v \mid \mathscr{G}]$ of v given $\mathscr{G}$, which we voluntarily do not specify because the specification of the version is immaterial in the considered context".

The fact that the conditional expectation actually exists is now proven in 3 steps.

STEP 1: Conditional expectation of $\mathbf{v} \in \mathbb{L}^2$.

Consider the space $\mathbb{L}^2$ of square integrable random variables, i.e. random variables v with $E[v^2] < \infty$ ($\mathbb{L}^2$ is studied in Chapter 4, Examples 4.5 and 4.16.) Also, consider $\mathbb{L}^2(\mathscr{G})$, the subspace of $\mathbb{L}^2$ formed by all $\mathscr{G}$-measurable random variables. Theorem 4.22 in Chapter 4 proves that the projection of a $v \in \mathbb{L}^2$ onto $\mathbb{L}^2(\mathscr{G})$ exists and is unique (up to equivalence.) We show that such a projection provides a version of the conditional expectation:

$$E[v \mid \mathscr{G}] = \text{ projection of } v \text{ onto } \mathbb{L}^2(\mathscr{G}),$$

where the right-hand-side indicates any projection in the equivalence class. To this end, we need to prove that properties i) and ii) in the Definition 5.1 are fulfilled by the projection.

$\mathscr{G}$-measurability of the projection is a direct consequence of the fact that the projection belongs to $\mathbb{L}^2(\mathscr{G})$. As for ii), by the definition of projection we have the property that

$$v - (\text{ projection of } v \text{ onto } \mathbb{L}^2(\mathscr{G})) \perp g, \quad \forall g \in \mathbb{L}^2(\mathscr{G}).$$

In particular, by taking $g = 1(A), A \in \mathscr{G}$, we get: $\int_A (\text{projection of } v \text{ onto } \mathbb{L}^2(\mathscr{G}))dP = \int_\Omega (\text{projection of } v \text{ onto } \mathbb{L}^2(\mathscr{G})) \cdot 1(A)dP = \int_\Omega (v + (\text{projection of } v \text{ onto } \mathbb{L}^2(\mathscr{G}) - v)) \cdot 1(A)dP = \int_\Omega v \cdot 1(A)dP = \int_A v dP$, that is property ii).

For easy reference, we state the obtained result as a theorem.

THEOREM 5.2 (conditional expectation of $\mathbf{v} \in \mathbb{L}^2$) *If $v \in \mathbb{L}^2$, then $E[v \mid \mathscr{G}]$ is the projection of v onto the subspace $\mathbb{L}^2(\mathscr{G})$ of all $\mathscr{G}$-measurable square integrable random variables. Precisely, any projection in the equivalence class is a version of $E[v \mid \mathscr{G}]$ and all the versions are obtained by varying the projection in the equivalence class.* □

STEP 2: Conditional expectation of nonnegative random variables.

Given $v \geq 0$, define the sequence of bounded random variables $v_n := \min\{v, n\}$, where $n = 1, 2, \ldots$. Clearly $v_n \in \mathbb{L}^2$, so that $E[v_n \mid \mathscr{G}]$ is defined as in Step 1. For any n, take a version of $E[v_n \mid \mathscr{G}]$. The fact that v_n is nondecreasing implies that $E[v_n \mid \mathscr{G}]$ is almost surely nondecreasing too. Indeed, if A is the set where $E[v_{n+1} \mid \mathscr{G}] < E[v_n \mid \mathscr{G}]$, we have: $0 \geq \int_A (E[v_{n+1} \mid \mathscr{G}] - E[v_n \mid \mathscr{G}])dP = \int_A E[v_{n+1} \mid \mathscr{G}]dP - \int_A E[v_n \mid \mathscr{G}]dP = \int_A v_{n+1}dP - \int_A v_n dP = \int_A (v_{n+1} - v_n)dP \geq 0$, from which we find that equality holds throughout so that $P(A) = 0$. Since $E[v_n \mid \mathscr{G}]$ is nondecreasing almost surely, it converges almost surely (to a finite value or to ∞.) Where $E[v_n \mid \mathscr{G}]$ does not converge, redefine the limit to be zero. We claim that the limit is a version of the conditional expectation of v given $\mathscr{G}$, and we verify this in the following.

Remark 5.3 *We need to remark the fact that defining* $E[v \mid \mathscr{G}] = \lim_{n\to\infty} E[v_n \mid \mathscr{G}]$ *leaves open the possibility that* $E[v \mid \mathscr{G}] = \infty$ *on a nonzero probability set even when* $v < \infty$, $\forall \omega \in \Omega$. *To see this, consider the following example: over the probability space* $([0,1], \mathscr{B}[0,1], \lambda)$, *with* $\lambda =$ *Lebesgue measure, consider the random variable*

$$\begin{cases} 0, & x = 0 \\ \frac{1}{x}, & otherwise, \end{cases}$$

and let $\mathscr{G}$ *be the trivial* σ*-algebra that only contains the empty set* $\emptyset$ *and the whole set* $[0,1]$. *Being* $\mathscr{G}$ *trivial,* $E[v_n \mid \mathscr{G}]$ *is constant over* $[0,1]$ *and equal to* $E[v_n]$. *But* $E[v_n] \to \infty$ *as* $n \to \infty$, *showing that* $E[v \mid \mathscr{G}] = \infty$, $\forall \omega \in [0,1]$.

The fact that $E[v \mid \mathscr{G}]$ *can possibly be* ∞ *may seem to pose a difficulty since our definition of random variable (Definition 2.2) and all subsequent developments assume* $v \in \mathbb{R}$, *where* $\mathbb{R}$ *does not include* $\pm\infty$. *This difficulty is however easily circumvented, provided that one instead works with random variables taking on value in* $[-\infty, \infty]$, *the extended set of real. We recall that the arithmetic of* $\mathbb{R}$ *is extended to* $[-\infty, \infty]$ *with the definitions:* $a + \infty = \infty + a = \infty$ *if* $a > -\infty$, *and* $a - \infty = -\infty + a = -\infty$ *if* $a < \infty$; $\infty - \infty$ *is not defined.* $a \cdot \pm\infty = \pm\infty \cdot a = \pm\infty$ *if* $a > 0$, $a \cdot \pm\infty = \pm\infty \cdot a = \mp\infty$ *if* $a < 0$, *and* $a \cdot \pm\infty = \pm\infty \cdot a = 0$ *if* $a = 0$. *One easily verifies that the commutative, associative and distributive laws hold in* $[-\infty, \infty]$. *The definition of random variables and all the subsequent developments can naturally be extended to random variables taking on value in* $[-\infty, \infty]$.

Let us go to verify that properties i) and ii) hold. The measurability property i) follows from Theorem 3.5-(ii) applied to sequence $E[v_n \mid \mathscr{G}]$ seen as random variables on $(\Omega, \mathscr{G}, P)$. As for Property ii), note first that $\int_A E[v_n \mid \mathscr{G}]dP = \int_A v_n dP$, $\forall A \in \mathscr{G}$ (this is Property ii) for random variables in $\mathbb{L}^2$.) Then,

$$\begin{aligned} \int_A E[v \mid \mathscr{G}]dP &= \lim_{n\to\infty} \int_A E[v_n \mid \mathscr{G}]dP \\ &\qquad (by\ the\ monotone\ convergence\ Theorem\ 3.7) \\ &= \lim_{n\to\infty} \int_A v_n dP \\ &= \int_A \lim_{n\to\infty} v_n dP \\ &\qquad (again\ by\ the\ monotone\ convergence\ Theorem\ 3.7) \\ &= \int_A v dP. \end{aligned}$$

STEP 3: Conditional expectation of random variables such that $E[v]$ **exists.**

Let $v^+ := \max\{v, 0\}$ and $v^- := -\min\{v, 0\}$. Clearly, $v = v^+ - v^-$. We show that a version of the conditional expectation is given by the formula

$$E[v \mid \mathscr{G}] = E[v^+ \mid \mathscr{G}] - E[v^- \mid \mathscr{G}], \tag{5.1}$$

where in the right-hand-side we take any version of the conditional expectation of v^+ and of v^- and the left-hand-side is redefined to be zero where both $E[v^+ \mid \mathscr{G}]$ and $E[v^- \mid \mathscr{G}]$ are ∞.

We first show that it is not possible that $E[v^+ \mid \mathscr{G}]$ and $E[v^- \mid \mathscr{G}]$ take both value ∞ on a nonzero probability set , so that $E[v \mid \mathscr{G}]$ takes on expression (5.1) almost surely: were $E[v^+ \mid \mathscr{G}] = \infty$ on a nonzero probability set, we would then have $E[v^+] = \infty$. Similarly, $E[v^- \mid \mathscr{G}] = \infty$ on a nonzero probability set implies $E[v^-] = \infty$. But this would mean that $E[v]$ is not defined (recall that $E[v]$ is by definition $E[v^+] - E[v^-]$ provided not both these expectations are ∞), against our initial assumption. Once we have established that (5.1) holds almost surely, showing that $E[v \mid \mathscr{G}]$ satisfies properties i) and ii) is straightforward (the reader is invited to work out the details.) □

The following example shows the importance of the assumption that $E[v]$ exists when taking conditional expectation.

EXAMPLE 5.4 *Over* $([0,1], \mathscr{B}[0,1], \lambda)$*, consider the random variable*

$$v = \begin{cases} 0, & x = 0 \text{ and } 1 \\ \frac{1}{x}, & 0 < x \leq 0.5 \\ \frac{1}{x-1}, & 0.5 < x < 1, \end{cases}$$

and let $\mathscr{G} = \{\emptyset, [0,1]\}$*. Here,* $E[v^+ \mid \mathscr{G}] = \infty$ *in* $[0,1]$ *and, similarly,* $E[v^- \mid \mathscr{G}] = \infty$ *in* $[0,1]$*, so that* $E[v \mid \mathscr{G}]$ *is not defined. The difficulty arises from the fact that* $E[v]$ *does not exist in this case.* □

Properties of the conditional expectation

The following properties are listed without proof. They are all valid almost surely and it is understood that $E[v]$, $E[v_1]$, and $E[v_2]$ are assumed to exist. The reader is referred, among other textbooks, to [7], Chapter 2, for a proof.

1. If $v_1 \leq v_2$, then $E[v_1 \mid \mathscr{G}] \leq E[v_2 \mid \mathscr{G}]$;
2. If α and β are constants such that $\alpha E[v_1] + \beta E[v_2]$ is defined, then $E[\alpha v_1 + \beta v_2 \mid \mathscr{G}] = \alpha E[v_1 \mid \mathscr{G}] + \beta E[v_2 \mid \mathscr{G}]$;
3. If $\mathscr{G} = \{\emptyset, \Omega\}$, then $E[v \mid \mathscr{G}] = E[v]$;
4. $E[E[v \mid \mathscr{G}]] = E[v]$;
5. If $\mathscr{G}_1 \subseteq \mathscr{G}_2$, then $E[E[v \mid \mathscr{G}_2] \mid \mathscr{G}_1] = E[v \mid \mathscr{G}_1]$;

6. Let v_1 be a $\mathscr{G}$-measurable random variable and assume that $E[v_1 v_2]$ exists. Then, $E[v_1 v_2 \mid \mathscr{G}] = v_1 E[v_2 \mid \mathscr{G}]$.

Conditional expectation of $\mathbf{v_2}$ given $\mathbf{v_1}$

Consider a measurable function $f : \mathbb{R} \to \mathbb{R}$ and a random variable $\eta : \Omega \to \mathbb{R}$ and denote by $\sigma(\eta)$ the σ-algebra generated by η (i.e. $\sigma(\eta)$ is the smallest σ-algebra in Ω with respect to which η is measurable.) Clearly, the random variable $f(\eta) : \Omega \to \mathbb{R}$ is $\sigma(\eta)$-measurable (see Theorem 1.5.) It is a remarkable fact that the converse also holds true: if a random variable ξ is $\sigma(\eta)$-measurable, then there exists a measurable function f such that $\xi = f(\eta)$ for all $\omega \in \Omega$ (see e.g. [7], Theorem 3, Chapter 2, Section 4, for a proof.) For easy reference, we state this fact in the following theorem.

THEOREM 5.5 *Given a random variable η, the set of random variables $\{f(\eta)$, with f measurable function from $\mathbb{R}$ to $\mathbb{R}\}$ coincides with the set of $\sigma(\eta)$-measurable random variables.* □

DEFINITION 5.6 (conditional expectation of $\mathbf{v_2}$ given $\mathbf{v_1}$) *Given two random variables v_1 and v_2 such that $E[v_2]$ exists, consider any version of $E[v_2 \mid \sigma(v_1)]$. Since this conditional expectation is $\sigma(v_1)$-measurable, in view of Theorem 5.5 we can write $E[v_2 \mid \sigma(v_1)] = f(v_1)$, for some measurable function f, where equality holds $\forall \omega \in \Omega$. This function $f : \mathbb{R} \to \mathbb{R}$ is called a conditional expectation of v_2 given v_1.* □

Function f in Definition 5.6 is not unique. This can be easily understood by observing that $f(v_1)$ involves only computing $f(x)$ for values of x that are taken on by $v_1(\omega)$ for some $\omega \in \Omega$. Thus, changing the value of f elsewhere does not affect the value of $f(v_1)$. Moreover, if one considers a different version of $E[v_2 \mid \sigma(v_1)]$, by applying Definition 5.6 one finds that a function f such that $f(v_1) = E[v_2 \mid \sigma(v_1)]$ for this new version of the conditional expectation is still a conditional expectation of v_2 given v_1. This adds an extra degree of freedom in the selection of f. It is not difficult to see that, given a conditional expectation f of v_2 given v_1, the set of all conditional expectations is the collection of all measurable functions f_1 that differ from f on a set having zero P'_{v_1} measure, where P'_{v_1} is the image probability induced on $\mathbb{R}$ by v_1. Any such function is called a version of the conditional expectation of v_2 given v_1.

Sometimes, we use the symbol $E[v_2 \mid v_1 = x]$ for $f(x)$. Intuitively, $E[v_2 \mid v_1 = x]$ represents the mean value assumed by v_2 once we know v_1 has taken on the value $v_1 = x$.

EXAMPLE 5.7 *Let $\Omega = \{(0,0),(0,1),(1,0),(1,1)\}$, $\mathscr{F}$ = all subsets of Ω, and let P be specified by $P(0,0) = P(1,1) = \frac{1}{6}$ and $P(0,1) = P(1,0) = \frac{2}{6}$.*

Consider the random variables v_1 and v_2 that assign to the sample outcome (i, j) the value i and j, respectively (see Figure 5.2):

$$v_1(i,j) = i; \quad v_2(i,j) = j.$$

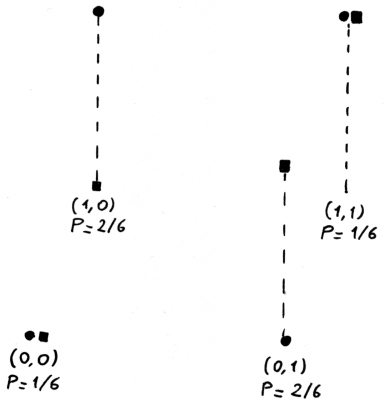

Figure 5.2: Ω, v_1 (•), and v_2 (■) for Example 5.7

We have:

$$E[v_2 \mid \sigma(v_1)] = \begin{cases} \frac{4}{6}, & \text{on } (0,0) \text{ and } (0,1) \\ \frac{2}{6}, & \text{on } (1,0) \text{ and } (1,1), \end{cases}$$

and no other version exists. Moreover, any measurable function $f : \mathbb{R} \to \mathbb{R}$ with $f(0) = \frac{4}{6}$ and $f(1) = \frac{2}{6}$ is a conditional expectation of v_2 given v_1, as it is readily seen by noting that $f(v_1) = E[v_2 \mid \sigma(v_1)]$. □

5.2 Conditional density

We define the conditional density of a random variable v_2 given a second random variable v_1. Here, we assume that the two random variables v_1 and v_2 admit joint probability density $p_{v_1,v_2}(x,y)$.

DEFINITION 5.8 (conditional density) *Given a version of the joint probability density $p_{v_1,v_2}(x,y)$ and a version of the probability density $p_{v_1}(x)$, the conditional density of v_2 given v_1 is defined as*

$$p_{v_2|v_1}(y \mid x) := \begin{cases} \frac{p_{v_1,v_2}(x,y)}{p_{v_1}(x)}, & if\ p_{v_1}(x) \neq 0 \\ 0, & otherwise. \end{cases}$$

□

The conditional density $p_{v_2|v_1}(y \mid x)$ describes how v_2 distributes under the condition that $v_1 = x$.

By varying the version of p_{v_1,v_2} and p_{v_1}, different versions of $p_{v_2|v_1}$ are obtained.

From the definition, it is clear that $p_{v_2|v_1}$ can be constructed from p_{v_1,v_2} and p_{v_1}. On the other hand, if we are given $p_{v_2|v_1}$, then p_{v_1,v_2} and p_{v_1} cannot be uniquely determined. To see this, note that, if we multiply $p_{v_1}(x)$ by a function $f(x) > 0$ such that $\int p_{v_1}(x) f(x) dx = 1$ and repeat a similar operation for $p_{v_1,v_2}(x,y)$ so obtaining $p_{v_1,v_2}(x,y) f(x)$, the so-achieved two new densities $p_{v_1,v_2}(x,y) f(x)$ and $p_{v_1}(x) f(x)$ are different from the original ones but share with these the same conditional density. As we can see, the indetermination lies in the fact that the division by $p_{v_1}(x)$ in the conditional density definition is a normalization operation that hides the relative probability of different x values.

$p_{v_2|v_1}(y \mid x)$ tells us how v_2 distributes when $v_1 = x$. Intuitively, this knowledge is richer than that provided by $E[v_2 \mid v_1 = x]$, the mean of v_2 for $v_1 = x$ and the following theorem provides a way to compute $E[v_2 \mid v_1 = x]$ from $p_{v_2|v_1}(y \mid x)$.

THEOREM 5.9 *Given two random variables v_1 and v_2 such that $E[v_2]$ exists, a version of $E[v_2 \mid v_1 = x]$ is given by*

$$E[v_2 \mid v_1 = x] = \begin{cases} \int_{\mathbb{R}} y p_{v_2|v_1}(y \mid x) dy, & \text{if the } \int \text{ exists} \\ 0, & \text{otherwise.} \end{cases}$$

□

Remark 5.10 *The integral $\int_{\mathbb{R}} y p_{v_2|v_1}(y \mid x) dy$ in (5.2) is not guaranteed to be defined for all x values. To understand this, just note that, in correspondence of a given x, $p_{v_1,v_2}(x,y)$ is substantially arbitrary (the only constraints are due to measurability properties) since a line in $\mathbb{R}^2$ with fixed coordinate x has zero Lebesgue measure. This arbitrariness can be spent so that the integral in (5.2) is undefined for the selected x.* □

PROOF. First, let $A := \{x : p_{v_1}(x) = 0\}$ and note that, by an application of Theorem 1.12, we have

$$\int_{A\times\mathbb{R}} p_{v_1,v_2}(x,y)d(x,y) = \int_\Omega 1(v_1 \in A)dP = \int_A p_{v_1}(x)dx = 0,$$

entailing that $p_{v_1,v_2}(x,y) = 0$ λ^2-*almost surely* on $A\times\mathbb{R}$ (i.e. $p_{v_1,v_2}(x,y)$ may be different from zero at most in a zero Lebesgue measure set in $A\times\mathbb{R}$.) Consequently,

$$p_{v_1,v_2}(x,y) = p_{v_2|v_1}(y \mid x)p_{v_1}(x) \quad \lambda^2-\textit{almost surely}, \tag{5.2}$$

since the two sides are by definition equal outside $A\times\mathbb{R}$ while in $A\times\mathbb{R}$ the right-hand-side and the left hand side are both zero almost surely.

Take now a Borel set B in $\mathbb{R}$. We want to show that

$$\int_{\mathbb{R}^2} 1(x\in B)y\cdot p_{v_2|v_1}(y \mid x)p_{v_1}(x)d(x,y) = \int_{\mathbb{R}}\left[\int_{\mathbb{R}} 1(x\in B)y\cdot p_{v_2|v_1}(y \mid x)p_{v_1}(x)dy\right]dx, \tag{5.3}$$

where it is understood that the inner integral is set to zero at those x's where it is undefined.

To see this, start by noting that the existence of $E[v_2] = E[v_2^+] - E[v_2^-]$ implies that at least one between $E[v_2^+]$ and $E[v_2^-]$ is finite. Suppose e.g. that $E[v_2^+] < \infty$, then:

$$\begin{aligned}
\infty &> E[1(v_1\in B)\cdot v_2^+] \\
&= \int_\Omega 1(v_1\in B)\cdot v_2\cdot 1(v_2\geq 0)dP \\
&= \int_{\mathbb{R}^2} 1(x\in B)y\cdot 1(y\geq 0)p_{v_1,v_2}(x,y)d(x,y) \\
&= \int_{\mathbb{R}\times\{y\geq 0\}} 1(x\in B)y\cdot p_{v_2|v_1}(y \mid x)p_{v_1}(x)d(x,y) \quad (\textit{using } (5.2)) \\
&= \int_{\mathbb{R}}\left[\int_{\{y\geq 0\}} 1(x\in B)y\cdot p_{v_2|v_1}(y \mid x)p_{v_1}(x)dy\right]dx, \quad (\textit{using Tonelli's Theorem } 1.14)
\end{aligned}$$

showing that the inner integral $\int_{\{y\geq 0\}}$ is less than infinity λ_x-*almost surely*. (5.3) can now be proven by first rewriting the integral $\int_{\mathbb{R}^2}$ on the left-hand-side as $\int_{\mathbb{R}\times\{y\geq 0\}} + \int_{\mathbb{R}\times\{y<0\}}$; then, by applying Tonelli's theorem to each of these two integrals as in the last step of the previous derivation; and, finally, by noting that the sum of the two integrals can be rewritten as the right-hand-side of (5.3) since the inner integral does not exist (and therefore is redefined to be zero) where both $\int_{\{y\geq 0\}} = \infty$ and $\int_{\{y<0\}} = -\infty$, which only happens on a zero λ_x measure set.

With the technical results (5.2) and (5.3) in our hands, we can now proceed to write $\int_\Omega 1(v_1 \in B)v_2 dP$ in two different ways and, from a comparison of the results, the theorem conclusion will finally be drawn.

First, we have:

$$\begin{aligned}
&\int_\Omega 1(v_1 \in B)v_2 dP \\
&= \int_{\mathbb{R}^2} 1(x \in B)y \cdot p_{v_1,v_2}(x,y)d(x,y) \\
&= \int_{\mathbb{R}^2} 1(x \in B)y \cdot p_{v_2|v_1}(y \mid x)p_{v_1}(x)d(x,y) \quad (\textit{using } (5.2)) \\
&= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} 1(x \in B)y \cdot p_{v_2|v_1}(y \mid x)p_{v_1}(x)dy \right] dx \quad (\textit{using } (5.3)) \\
&= \int_B \left[\int_{\mathbb{R}} y p_{v_2|v_1}(y \mid x)dy \right] p_{v_1}(x)dx. \qquad (5.4)
\end{aligned}$$

But, we also have:

$$\begin{aligned}
&\int_\Omega 1(v_1 \in B)v_2 dP \\
&= \int_{v_1^{-1}(B)} v_2 dP \\
&= \int_{v_1^{-1}(B)} E[v_2 \mid \sigma(v_1)]dP \\
&= \int_{v_1^{-1}(B)} f(v_1)dP \quad (\textit{where } f \textit{ is a version of the conditional expectation of } v_2 \textit{ given } v_1) \\
&= \int_\Omega 1(v_1 \in B) f(v_1)dP \\
&= \int_{\mathbb{R}} 1(x \in B) f(x)p_{v_1}(x)dx \quad (\textit{using Theorem } 1.12) \\
&= \int_B f(x)p_{v_1}(x)dx. \qquad (5.5)
\end{aligned}$$

Comparing (5.4) and (5.5), since B is arbitrary we conclude that

$$\int_{\mathbb{R}} y p_{v_2|v_1}(y \mid x)dy = f(x) \quad P'_{v_1} - \textit{almost surely},$$

that is $\int_{\mathbb{R}} y p_{v_2|v_1}(y \mid x)dy$ is a version of $E[v_2 \mid v_1 = x]$. □

In this Chapter, we have only considered the conditional expectation in the case of $\mathbb{R}$-valued random variable. The extension to a multi-dimensional (i.e. $\mathbb{R}^n$-valued) random variable is straightforward and is defined as the $\mathbb{R}^n$-valued random variable whose components are the conditional expectations of the components of the random variable. Also, all other concepts extend in a natural way to the multi-dimensional case.

Chapter 6

WIDE-SENSE STATIONARY PROCESSES

6.1 Definitions and examples

Let us consider a sequence of complex-valued random variables v_t defined over some probability space $(\Omega, \mathscr{F}, P)$, where t runs over positive and negative integers: $t = \ldots, -2, -1, 0, 1, 2, \ldots$. The fact that v_t is "complex-valued" simply means that $v_t = v_{\mathbb{R},t} + iv_{\mathbb{I},t}$, where $v_{\mathbb{R},t}$ and $v_{\mathbb{I},t}$ are real-valued random variables. The motivation for considering complex - as opposed to real - v_t's is that certain derivations become notationally easier in a complex setting. v_t is called a discrete-time complex-valued stochastic process. When $v_{\mathbb{I},t} = 0$, the complex-valued process reduces to a real-valued process.

Throughout this chapter, we assume that variables v_t are square integrable, i.e. $E[|v_t|^2] = E[v_{\mathbb{R},t}^2 + v_{\mathbb{I},t}^2] < \infty$. The set of square integrable random variables is indicated with $\mathbb{L}^2$. Thus, $v_t \in \mathbb{L}^2$, for any t.

DEFINITION 6.1 (wide-sense stationary process) *Process v_t is said to be wide-sense stationary if its mean is constant:*

$$E[v_t] = E[v_0], \quad \forall t, \tag{6.1}$$

and the covariance of v_t and $v_{t+\ell}$ only depends on the time lag ℓ:

$$E\left[(v_{t+\ell} - E[v_{t+\ell}])\,\overline{(v_t - E[v_t])}\right] = E\left[(v_\ell - E[v_\ell])\,\overline{(v_0 - E[v_0])}\right], \quad \forall t, \ell. \tag{6.2}$$

($\bar{}$ denotes complex conjugation.) □

Thus, a wide-sense stationary stochastic process has first and second order statistics that are invariant under time shift.

Function

$$\gamma_\ell := E[(v_\ell - E[v_\ell])\overline{(v_0 - E[v_0])}] \tag{6.3}$$

is called the *auto-covariance function* of process v_t.

From definition (6.3), it is not difficult to see that γ_ℓ is symmetric, i.e. $\gamma_\ell = \overline{\gamma}_{-\ell}$, and *positive semidefinite*, that is, for any given integer n and for all complex numbers $a_1, \ldots, a_n$, it holds that

$$\sum_{k,j=1}^{n} a_k \gamma_{k-j} \overline{a}_j \geq 0. \tag{6.4}$$

Without any loss of generality, from now on we shall assume that $E[v_t] = 0$ (if this is not the case, it is sufficient to subtract the mean from the original stochastic process in order to conform to this assumption.) Then, $\gamma_\ell = E[v_\ell \overline{v}_0]$ can be interpreted as the scalar product between the random variables v_ℓ and v_0 in the $\mathbb{L}^2$ space (the reader is referred to Chapter 4 for the notion of scalar product and, particularly, to Example 4.5 for the scalar product in $\mathbb{L}^2$ - in fact, in Example 4.5 real-valued random variables are considered, but the extension to complex-valued variables presents no difficulties.)

The concept of stationary process is now illustrated through examples.

EXAMPLE 6.2 *Given a real random variable z with $E[z^2] < \infty$, let*

$$v_t = z, \quad \forall t.$$

It is immediately seen that v_t is a wide-sense stationary process. Its realizations are constant functions. □

EXAMPLE 6.3 (white process) *A real stochastic process v_t such that*

$$E[v_t] = 0, \quad \forall t,$$

and

$$E[v_{t+\ell} v_t] = \begin{cases} \sigma^2, & if\ \ell = 0 \\ 0, & otherwise, \end{cases}$$

is clearly wide-sense stationary.

Such a process is called a white process and its characteristic is that each random variable is uncorrelated with all others. A white process can be equivalently seen as a sequence of orthogonal real functions with constant norm in the $\mathbb{L}^2$ space. □

EXAMPLE 6.4 (one-harmonic process) *Consider the stochastic process v_t defined through the relation*

$$v_t = ze^{-i\omega t} + \bar{z}e^{i\omega t}, \tag{6.5}$$

where ω is a fixed frequency belonging to the interval $[-\pi, \pi]$ and $z = z_{\mathbb{R}} + iz_{\mathbb{I}}$ is a complex random variable such that: $E[z_{\mathbb{R}}] = E[z_{\mathbb{I}}] = 0$, $E[z_{\mathbb{R}}^2] = E[z_{\mathbb{I}}^2] = \sigma^2/4$, $E[z_{\mathbb{R}} z_{\mathbb{I}}] = 0$.

In (6.5), process v_t has been defined through complex quantities for the sake of notational compactness. However, an easy computation shows that process v_t is in fact real:

$$\begin{aligned} v_t &= (z_{\mathbb{R}} + iz_{\mathbb{I}})(\cos(\omega t) - i\sin(\omega t)) + (z_{\mathbb{R}} - iz_{\mathbb{I}})(\cos(\omega t) + i\sin(\omega t)) \\ &= 2z_{\mathbb{R}}\cos(\omega t) + 2z_{\mathbb{I}}\sin(\omega t) \\ &= 2\sqrt{z_{\mathbb{R}}^2 + z_{\mathbb{I}}^2}\sin(\omega t + atan(z_{\mathbb{R}}/z_{\mathbb{I}})) \\ &= A\sin(\omega t + \phi), \end{aligned} \tag{6.6}$$

where, in the last equality, we have defined $A = 2\sqrt{z_{\mathbb{R}}^2 + z_{\mathbb{I}}^2}$ and $\phi = atan(z_{\mathbb{R}}/z_{\mathbb{I}})$ (here, the appropriate determination for atan has to be taken depending on the sign of $z_{\mathbb{R}}$ and $z_{\mathbb{I}}$.) Expression (6.6) reveals the nature of process v_t: all its realizations are sinusoids with fixed frequency ω and random amplitude A and phase ϕ.

The stationarity of process v_t can be verified by a direct computation of its mean and auto-covariance function:

$$\begin{aligned} E[v_t] &= E\left[ze^{-i\omega t} + \bar{z}e^{i\omega t}\right] \\ &= E[z]e^{-i\omega t} + E[\bar{z}]e^{i\omega t} \\ &= 0, \quad \forall t; \end{aligned}$$

$$\begin{aligned} E[v_{t+\ell}\overline{v_t}] &= E\left[\left(ze^{-i\omega(t+\ell)} + \bar{z}e^{i\omega(t+\ell)}\right)\left(\bar{z}e^{i\omega t} + ze^{-i\omega t}\right)\right] \\ &= E[|z|^2]e^{-i\omega\ell} + E[z^2]e^{-i\omega(2t+\ell)} + E[\bar{z}^2]e^{i\omega(2t+\ell)} + E[|z|^2]e^{i\omega\ell}. \end{aligned}$$

In the latter expression, the expectations are given by

$$\begin{aligned} E[|z|^2] &= E[z_{\mathbb{R}}^2 + z_{\mathbb{I}}^2] = \sigma^2/2 \\ E[z^2] &= E[z_{\mathbb{R}}^2 - z_{\mathbb{I}}^2 + 2iz_{\mathbb{R}}z_{\mathbb{I}}] = E[z_{\mathbb{R}}^2] - E[z_{\mathbb{I}}^2] = 0 \\ E[\overline{z}^2] &= E[z_{\mathbb{R}}^2 - z_{\mathbb{I}}^2 - 2iz_{\mathbb{R}}z_{\mathbb{I}}] = E[z_{\mathbb{R}}^2] - E[z_{\mathbb{I}}^2] = 0, \end{aligned}$$

which, substituted in the expression for $E[v_{t+\ell}\overline{v_t}]$, give

$$E[v_{t+\ell}\overline{v_t}] = \sigma^2 \cos(\omega\ell), \quad \forall t, \ell.$$

Since this last expression only depends on ℓ, the stationarity of process v_t follows. □

EXAMPLE 6.5 (multi-harmonic process) *The example above can be straightforwardly generalized to the case when the stochastic process is formed by more harmonic components, with different frequencies.*

Let us define

$$v_t = \sum_{k=1}^{N} \left(z_k e^{-i\omega_k t} + \bar{z}_k e^{i\omega_k t} \right),$$

where the z_k's satisfy conditions similar to those for z in Example 6.4 and, in addition, $E[z_{\mathbb{R},k} z_{\mathbb{R},j}] = E[z_{\mathbb{R},k} z_{\mathbb{I},j}] = E[z_{\mathbb{I},k} z_{\mathbb{I},j}] = 0, \forall k \neq j$.

Computations entirely similar to those for the case of a single harmonic component show that process v_t is stationary and that its realizations are formed by the sum of sinusoids with frequencies ω_k, $k = 1, \ldots, N$. Each sinusoid has random amplitude and phase and the variance of the k-th sinusoidal component is σ_k^2. Moreover, each sinusoidal component is uncorrelated with all others. □

In the above example, the stationary stochastic process is the sum of uncorrelated sinusoidal stochastic components. A truly remarkable fact is that this holds true in full generality: any wide sense stationary stochastic process admits a decomposition in uncorrelated harmonical components. This will be proven at a later stage as Theorem 6.10.

6.2 Elementary spectral theory of stationary processes

The elementary spectral theory is not universally applicable, but it is easy to establish and yet it can be used in many circumstances. A general spectral theory is differed to the next Section 6.3.

Assume that $\gamma_\ell \in l^1$ (i.e. $\sum_{\ell=-\infty}^{\infty} |\gamma_\ell| < \infty$.) This assumption requires that the correlation vanishes for large time lags and is not satisfied e.g. in the processes of Examples 6.2, 6.4, and 6.5. Function

$$f(\omega) = \frac{1}{\sqrt{2\pi}} \sum_{\ell=-\infty}^{\infty} \gamma_\ell e^{-i\omega\ell} \tag{6.7}$$

(the discrete Fourier transform of γ_ℓ) is then pointwise convergent for any ω and is called the *spectral density* of the process.

Clearly, $f(\omega)$ is periodic of period 2π, so that it is enough to regard it as a function defined over $(-\pi, \pi]$. Moreover, from the properties of γ_ℓ it can be seen that $f(\omega)$ is real and nonnegative.

Given γ_ℓ, one can compute the spectral density $f(\omega)$ via (6.7). Viceversa, given $f(\omega)$, γ_ℓ can be reconstructed by relation

$$\gamma_\ell = \frac{1}{\sqrt{2\pi}} \int_{(-\pi,\pi]} e^{i\omega\ell} f(\omega) d\omega,$$

(verify this) so that γ_ℓ and $f(\omega)$ carry exactly the same information content.

The interpretation of $f(\omega)$ is that it describes the harmonic content of the stochastic process. A full justification of this interpretation requires a more-in-depth analysis along the lines provided in the next section.

6.3 Spectral theory of stationary processes

Spectral measure

The spectral measure is a way to prescribe the correlation pattern of a stationary process alternative to γ_ℓ. Though the spectral measure conveys exactly the same information as γ_ℓ, it has an extra intuitive appeal as it directly describes the harmonic content of the stationary process. When $\gamma_\ell \in l^1$ as in Section 6.2, the spectral measure has a density and such a density is given by (6.7).

We start with the following fundamental theorem.

THEOREM 6.6 (Herglotz) *Let γ_ℓ be a positive semidefinite function (i.e. γ_ℓ satisfies (6.4).) Then, there exists a finite measure m on $((-\pi,\pi],\mathscr{B}(-\pi,\pi])$ such that, for any $\ell = \ldots,-2,-1,0,1,2,\ldots$:*

$$\gamma_\ell = \int_{(-\pi,\pi]} e^{i\omega\ell} dm(\omega). \tag{6.8}$$

□

When γ_ℓ is the auto-covariance function of a wide-sense stationary process v_t, measure m in Theorem 6.6 is called the *spectral measure* of v_t. Its distribution $F(\omega) = \int_{-\pi}^{\omega} dm(\omega)$ is called the *spectral distribution*.

PROOF. This proof uses the notion of weak convergence. All results used are provided in Chapter 3.

Define

$$f_n(\omega) := \frac{1}{\sqrt{2\pi}n} \sum_{k,j=1}^{n} e^{-i\omega k}\gamma_{k-j}e^{i\omega j} = \frac{1}{\sqrt{2\pi}} \sum_{j=-n+1}^{n-1} \left(1 - \frac{|j|}{n}\right) \gamma_j e^{-i\omega j} \tag{6.9}$$

(the second equality follows from a simple computation.)

Since γ_ℓ is positive semidefinite, $f_n(\omega) \geq 0$. Next, let

$$F_n(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\omega} f_n(x)dx,$$

for $\omega \in (-\pi,\pi]$, while $F_n(\omega) = 0$ for $\omega \leq -\pi$, and $F_n(\omega) = \int_{-\pi}^{\pi} f_n(x)dx$ for $\omega > \pi$. $F_n(\omega)$ is nondecreasing and continuous; moreover, observing that

$$F_n(\pi) = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} f_n(x)dx = \gamma_0, \quad \forall n,$$

we conclude that F_n/γ_0 is a sequence of probability distributions on $\mathbb{R}$. It is in fact a tight sequence according to the definition of tightness given in Helly's Theorem 3.23 (take $M = \pi$ in that definition.) Thus, due to Theorem 3.23, we conclude that there exists a subsequence F_{n_k}/γ_0 which converges weakly to a limit probability distribution F/γ_0. This distribution is supported on the closed interval $[-\pi,\pi]$. To such a distribution, a probability measure is associated (see Theorem 2.6 and the comment that follows the statement of this theorem.) Now, recalling the Definition 3.20 of weak convergence and noting that $e^{i\omega\ell}$ is a function with continuous and bounded real and complex parts, we obtain

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}}\int_{[-\pi,\pi]} e^{i\omega\ell} d\frac{F(\omega)}{\gamma_0} &= \frac{1}{\sqrt{2\pi}}\lim_{k\to\infty}\int_{[-\pi,\pi]} e^{i\omega\ell} d\frac{F_{n_k}(\omega)}{\gamma_0} \\
&= \frac{1}{\sqrt{2\pi}}\lim_{k\to\infty}\frac{1}{\gamma_0}\int_{[-\pi,\pi]} e^{i\omega\ell} f_{n_k}(\omega) d\omega \\
&= \frac{1}{\sqrt{2\pi}}\lim_{k\to\infty}\frac{1}{\gamma_0}\int_{[-\pi,\pi]} \frac{1}{2\pi}\sum_{j=-n_k+1}^{n_k-1}\left(1-\frac{|j|}{n_k}\right)\gamma_j e^{i\omega(-j+\ell)} d\omega \\
&= \frac{\gamma_\ell}{\gamma_0}. \qquad (6.10)
\end{aligned}$$

Finally, rescale the measure associated to F/γ_0 by a factor γ_0. The so-obtained measure is supported on $[-\pi,\pi]$, but we can reduce it to a measure supported on $(-\pi,\pi]$ by transferring the mass in $-\pi$ to π, and this operation does not change the integral of function $e^{i\omega\ell}$. The latter is the measure m of the theorem statement, where (6.8) easily follows from (6.10). □

A few comments on the spectral measure are now in order.

1. The spectral measure m is defined by use of the auto-covariance function γ_ℓ only. On the other hand, Herglotz's theorem gives an inversion formula to reconstruct γ_ℓ from m. This entails that m and γ_ℓ carry the same content of information: they both completely define the correlation pattern of the stationary process;
2. the spectral measure is unique. This claim requires a little proof.

   Suppose there are two spectral measures with distributions F_1 and F_2 (we define $F_1(\omega) = F_2(\omega) = 0$ for $\omega \le -\pi$ and $F_1(\omega) = F_2(\omega) = \gamma_0$ for $\omega > \pi$.) Then,

$$\int_{(-\pi,\pi]} e^{i\omega\ell} dF_1(\omega) = \gamma_\ell = \int_{(-\pi,\pi]} e^{i\omega\ell} dF_2(\omega). \qquad (6.11)$$

   Given an arbitrary $\overline{\omega} \in (-\pi,\pi]$, consider the sequence of functions

$$g_n(\omega) = \begin{cases} 1, & in\ [-\pi,\overline{\omega}] \\ 1-n(\omega-\overline{\omega}), & in\ (\overline{\omega},\overline{\omega}+\frac{1}{n}] \\ 0, & in\ (\overline{\omega}+\frac{1}{n},\pi]. \end{cases}$$

   Since every bounded continuous function can be uniformly approximated on $(-\pi,\pi]$ by trigonometric polynomials (see e.g. [6]), from (6.11) we have

$$\int_{(-\pi,\pi]} g_n(\omega) dF_1(\omega) = \int_{(-\pi,\pi]} g_n(\omega) dF_2(\omega).$$

Sending $n \to \infty$, this last equation gives $F_1(\overline{\omega}) = F_2(\overline{\omega})$, which, owing to the arbitrariness of $\overline{\omega}$, yields the desired result.

Spectral density

Suppose there exists a measurable function f defined on $(-\pi, \pi]$ such that F, the spectral distribution, is given by $F(\omega) = \int_{-\pi}^{\omega} f(x)dx$. Then, f is called the *spectral density* of the process. In this case, F is λ-almost surely differentiable and f is λ-almost surely the derivative of F (this is the "fundamental theorem of calculus", see e.g. Theorem 7.20 in [6].)

When $\gamma_\ell \in l^1$, f is given by (6.7) (and this justifies our calling "spectral density" the function in (6.7)), a fact that is proven in the next theorem.

THEOREM 6.7 *If $\gamma_\ell \in l^1$, then*

$$f(\omega) = \frac{1}{\sqrt{2\pi}} \sum_{\ell=-\infty}^{\infty} \gamma_\ell e^{-i\omega\ell} \tag{6.12}$$

is the spectral density of the process.

PROOF. Recall that

$$F_n(\omega) = \int_{-\pi}^{\omega} \frac{1}{\sqrt{2\pi}} \sum_{j=-n+1}^{n-1} \left(1 - \frac{|j|}{n}\right) \gamma_j e^{-ixj} dx.$$

Since $\gamma_\ell \in l^1$, the integrand can be uniformly (with respect to n) bounded as follows:

$$\left| \frac{1}{\sqrt{2\pi}} \sum_{j=-n+1}^{n-1} \left(1 - \frac{|j|}{n}\right) \gamma_j e^{-ixj} \right| \leq \frac{1}{\sqrt{2\pi}} \sum_{j=-\infty}^{\infty} |\gamma_j| < \infty.$$

Thus, by the dominated convergence Theorem 3.8 (in fact, here we are integrating with respect to a finite measure instead of a probability measure as in Theorem 3.8, but this difference can be leveled by a rescaling factor), we obtain

$$\begin{aligned} \lim_{n\to\infty} \int_{-\pi}^{\omega} \frac{1}{\sqrt{2\pi}} \sum_{j=-n+1}^{n-1} \left(1 - \frac{|j|}{n}\right) \gamma_j e^{-ixj} dx &= \int_{-\pi}^{\omega} \lim_{n\to\infty} \frac{1}{\sqrt{2\pi}} \sum_{j=-n+1}^{n-1} \left(1 - \frac{|j|}{n}\right) \gamma_j e^{-ixj} dx \\ &= \int_{-\pi}^{\omega} \frac{1}{\sqrt{2\pi}} \sum_{j=-\infty}^{\infty} \gamma_j e^{-ixj} dx, \end{aligned}$$

showing that the sequence $F_n(\omega)$ converges for any ω to $F(\omega) = \int_{-\pi}^{\omega} \frac{1}{\sqrt{2\pi}} \sum_{j=-\infty}^{\infty} \gamma_j e^{-ixj} dx$. Thus, $F(\omega)$ is the weak limit of $F_n(\omega)$ and so it is the spectral distribution and $\frac{1}{\sqrt{2\pi}} \sum_{j=-\infty}^{\infty} \gamma_j e^{-i\omega j}$ is its density. This concludes the proof. □

EXAMPLE 6.8 (Example 6.3 continued) *For the white process of Example 6.3 we have*

$$F_n(\omega) = \int_{-\pi}^{\omega} \sum_{j=-n+1}^{n-1} \left(1 - \frac{|j|}{n}\right) \gamma_j e^{-ixj} dx = \int_{-\pi}^{\omega} \sigma^2 dx = \sigma^2(\omega + \pi), \quad for\ \omega \in (-\pi, \pi].$$

Taking derivative with resepct to ω, we find the spectral density to be $f(\omega) = \sigma^2$, λ-almost surely. The same result can be obtained by relation $f(\omega) = \sum_{\ell=-\infty}^{\infty} \gamma_\ell e^{-i\omega\ell} = \sigma^2$. □

EXAMPLE 6.9 (Example 6.2 continued) *For the process of Example 6.2, assume $E[z^2] = 1$. Then, the auto-covariance function is $\gamma_\ell = 1$, $\forall \ell$, and is not in l^1. In this case, the Fourier transform $\frac{1}{\sqrt{2\pi}} \sum_{j=-\infty}^{\infty} \gamma_j e^{-ixj} \sum_{\ell=-\infty}^{\infty} \gamma_\ell e^{-i\omega\ell}$ does not converge (take e.g. $\omega = 0$) and the limit of $F_n(\omega) = \int_{-\pi}^{\omega} \frac{1}{\sqrt{2\pi}} \sum_{j=-\infty}^{\infty} \gamma_j e^{-ixj} \sum_{j=-n+1}^{n-1} \left(1 - \frac{|j|}{n}\right) \gamma_j e^{-ixj} dx$ cannot be computed by sending the limit under the sign of integral. Still, convergence holds in a weak sense as indicated in the proof of Herglotz Theorem 6.6.*

Figure 6.1 displays $F_n(\omega)$ for some values of n. It can be noted that $F_n(\omega)$ seems to converge to the step function with a step in 0 of hight $\sqrt{2\pi}$. This is in fact true and $F_n(\omega)$ converges to this step function for any $\omega \neq 0$, while $F_n(0) = \sqrt{2\pi}/2$, $\forall n$ (verifying this claim requires some lengthy computations and the reader can go through the calculations along the following line: first show that $F_n(\omega) = \frac{1}{\sqrt{2\pi}} \sum_{j=-n+1}^{-1} \left(1 - \frac{|j|}{n}\right) \frac{i}{j} e^{-i\omega j} + \frac{1}{\sqrt{2\pi}} \sum_{j=1}^{n-1} \left(1 - \frac{|j|}{n}\right) \frac{i}{j} e^{-i\omega j} + \frac{\omega}{\sqrt{2\pi}} + \frac{\sqrt{2\pi}}{2}$; then, term $\frac{\omega}{\sqrt{2\pi}}$ can be seen as the restriction to $(-\pi, \pi]$ of a periodic saw-tooth function with period 2π and this function can be expressed by means of its Fourier expansion; finally, after substituting the Fourier expansion for $\frac{\omega}{\sqrt{2\pi}}$ in the expression of $F_n(\omega)$, one can recognize that the so-obtained expansion for $F_n(\omega)$ tends to the expansion for the step function.)

What is F, the spectral distribution? It is the step function with a step in 0 of hight equal to $\sqrt{2\pi}$. This distribution is not absolutely continuous and the spectral density does not exist in this case. The spectral measure has concentrated mass in 0.

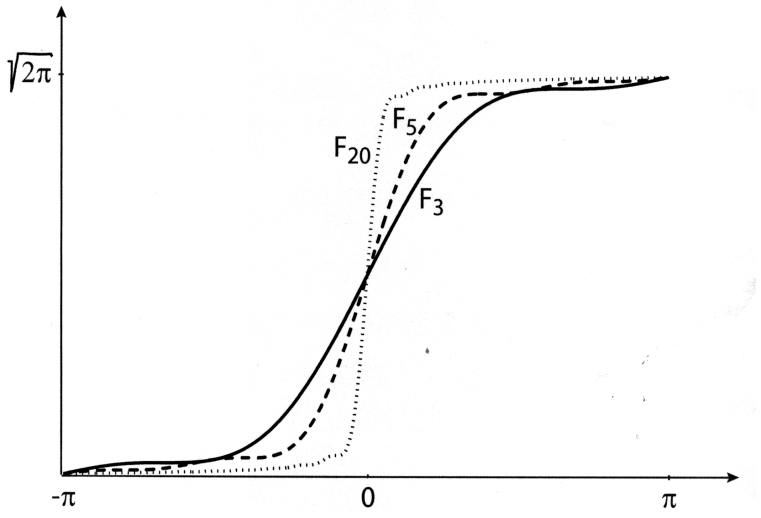

Figure 6.1:

We conclude with a technical remark relative to a point that may have attracted the reader's attention. In order for F to be a distribution, $F(0)$ must be equal to $\sqrt{2\pi}$, for, otherwise, F would not be continuous on the right in $\omega = 0$. If we go back to the proof of Helly's Theorem 3.23 (Helly's theorem is used in the proof of Herglotz Theorem 6.6), we see that $F(x)$ is defined for any x as $\inf_{x_j > x} \overline{F}(x_j)$. This definition indeed gives $F(0) = \sqrt{2\pi}$. □

A remark on the usefulness of the spectral distribution

When $\gamma_\ell \in l^1$, the tail of γ_ℓ is vanishing fast enough and $\frac{1}{\sqrt{2\pi}} \sum_{\ell=-n}^{n} \gamma_\ell e^{-i\omega\ell}$ gently converges as $n \to \infty$, allowing us to work with its limit $\frac{1}{\sqrt{2\pi}} \sum_{\ell=-\infty}^{\infty} \gamma_\ell e^{-i\omega\ell}$. This situation, however, does not cover all possible cases and the γ_ℓ can as well have a powerful tail so that taming the limit behavior of $\frac{1}{\sqrt{2\pi}} \sum_{\ell=-n}^{n} \gamma_\ell e^{-i\omega\ell}$ becomes difficult. The idea behind the spectral distribution is to control the roughness of $\frac{1}{\sqrt{2\pi}} \sum_{\ell=-n}^{n} \gamma_\ell e^{-i\omega\ell}$ by the smoothing properties of integration. The integrated function converges to the spectral distribution in a weak sense only, but this convergence is strong enough to secure the fundamental inversion rule (6.8).

Spectral representation of stationary processes

We now introduce an alternative representation of a wide-sense stationary process where the process is viewed as a stochastic integral with respect to an orthogonal stochastic measure. This representation shows that any stationary process can be interpreted as the sum of elementary harmonic components, so extending the results in Example 6.5. Moreover, as a side result, we obtain a new interpretation of the spectral measure m as a measure of the harmonic content of the stationary process.

Let us consider the space $\mathbb{L}^2(m)$ of the complex-valued, measurable and square integrable functions defined on $((-\pi,\pi],\mathscr{B}(-\pi,\pi],m)$, where m is the spectral measure associated to process v_t. This space is entirely similar to the space $\mathbb{L}^2$ studied in Examples 4.5 and 4.16 in Chapter 4, to which we refer the reader for definitions and explanation (in Chapter 4, real-valued functions are considered. Extending the results therein to the present complex-valued setting with a measure m presents no difficulties and the reader is invited to work out the details.) Here, we merely recall that $\mathbb{L}^2(m)$ is a vector space which can be endowed with a generalized inner product by the definition

$$(f,g) = \int_{(-\pi,\pi]} f(\omega)\overline{g}(\omega)dm(\omega).$$

This inner product is "generalized" since $(f,f)=0$ does not imply that $f=0$, it simply yields $f=0$ m-almost surely. Moreover, $\mathbb{L}^2(m)$ is complete.

Let $\mathbb{L}^2{}_0(m)\subseteq\mathbb{L}^2(m)$ be the linear subspace spanned by $e^{i\omega k}$, $k=\ldots,-2,-1,0,1,2,\ldots$ (i.e. each element in $\mathbb{L}^2{}_0(m)$ is simply a linear combination of functions $e^{i\omega k}$ of the form $\sum_{k\in K}\alpha_k e^{i\omega k}$, where K is any finite set of integers and α_k are complex numbers.) The closure of $\mathbb{L}^2{}_0(m)$ coincides with $\mathbb{L}^2(m)$ itself (see e.g. [6].)

Also, let $\mathbb{L}^2{}_0(v)$ be the linear subspace in the space $\mathbb{L}^2$ of the square integrable random variables spanned by v_k, and denote by $\mathbb{L}^2(v)$ its closure. In general, $\mathbb{L}^2(v)\neq\mathbb{L}^2$.

We want to establish a one-to-one correspondence T between $\mathbb{L}^2(m)$ and $\mathbb{L}^2(v)$. To be precise, T is one-to-one up to equivalence, i.e. we identify functions of $\mathbb{L}^2(m)$ which are m-almost surely equal and random variables of $\mathbb{L}^2(v)$ which are P-almost surely equal.

To this end, define first the following correspondence between elements of $\mathbb{L}^2{}_0(m)$ and elements of $\mathbb{L}^2{}_0(v)$:

$$\sum_{k\in K}\alpha_k e^{i\omega k} \leftrightarrow \sum_{k\in K}\alpha_k v_k.$$

This correspondence is an isometry, that is it preserves the inner product. In fact:

$$\begin{aligned}
\left(\sum_{k\in K}\alpha_k e^{i\omega k}, \sum_{j\in J}\beta_j e^{i\omega j}\right) &= \sum_{k\in K}\sum_{j\in J}\int_{(-\pi,\pi]}\alpha_k e^{i\omega k}\overline{\beta}_j e^{-i\omega j}dm(\omega) \\
&= \sum_{k\in K}\sum_{j\in J}\alpha_k\overline{\beta}_j\int_{(-\pi,\pi]}e^{i\omega(k-j)}dm(\omega) \\
&= \sum_{k\in K}\sum_{j\in J}\alpha_k\overline{\beta}_j\gamma_{k-j} \quad (\textit{using Theorem } 6.6) \\
&= \sum_{k\in K}\sum_{j\in J}\alpha_k\overline{\beta}_j E[v_k\overline{v}_j] \\
&= \left(\sum_{k\in K}\alpha_k v_k, \sum_{j\in J}\beta_j v_j\right).
\end{aligned}$$

Now, let $f \in \mathbb{L}^2(m)$. Since $\mathbb{L}^2(m)$ is the closure of $\mathbb{L}^2{}_0(m)$, there is a sequence $f_n \in \mathbb{L}^2{}_0(m)$ that converges to f. Let z_n be the sequence of random variables in $\mathbb{L}^2{}_0(v)$ corresponding to f_n. Since the correspondence is an isometry, it is immediate to verify that z_n is a Cauchy sequence, so that, in view of the completeness of $\mathbb{L}^2$, it converges to some limit point in $\mathbb{L}^2(v)$. There is an evident converse to this construction: if $z \in \mathbb{L}^2(v)$, then one can find a sequence in $\mathbb{L}^2{}_0(v)$ converging to z and the corresponding sequence in $\mathbb{L}^2{}_0(m)$ converges to a function $f \in \mathbb{L}^2(m)$.

By definition, we let $T : f \leftrightarrow z$. It is easily verified that this correspondence between $\mathbb{L}^2(m)$ and $\mathbb{L}^2(v)$ is one-to-one, linear and isometric. The construction is illustrated in Figure 6.2.

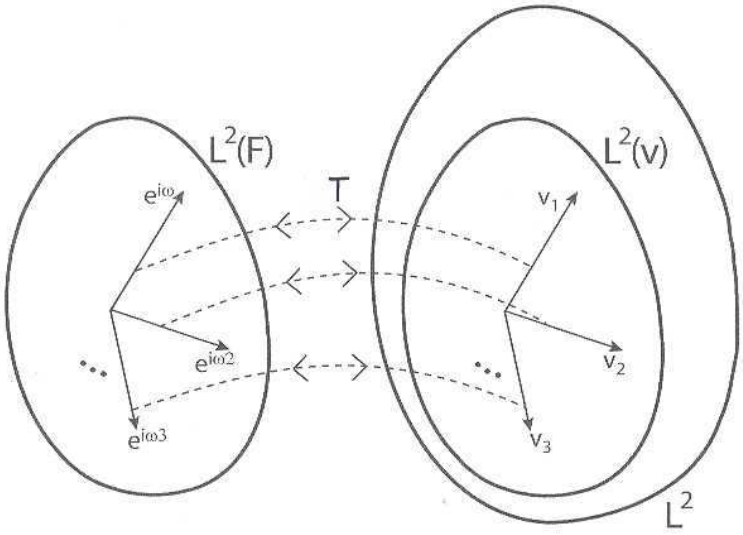

Figure 6.2: The T correspondence.

We next introduce the notion of stochastic integral. For any Borel set $B \in \mathscr{B}(-\pi,\pi]$, denote by $T(B)$ the random variable corresponding to $1(B)$ (i.e. the indicator function of set B equal to 1 on B and 0 elsewhere) in the isometry T. Note that $E[T(B_1)\overline{T}(B_2)] = 0$ whenever $B_1 \cap B_2 = \emptyset$. This is immediately verified by noting that $E[T(B_1)\overline{T}(B_2)] = \int_{(-\pi,\pi]} 1(B_1)1(B_2)dm(\omega) = \int_{(-\pi,\pi]} 0\, dm(\omega) = 0$. Function $T : \mathscr{B}(-\pi,\pi] \to \mathbb{L}^2(v)$ is called an orthogonal stochastic measure.

We first define the stochastic integral for elementary functions. Given a finite set of N disjoint Borel sets B_k, $k = 1,\ldots,N$, and N complex numbers α_k, $k = 1,\ldots,N$, the integral of the elementary function $f = \sum_{k=1}^{N} \alpha_k 1(B_k)$ is simply defined as $\sum_{k=1}^{N} \alpha_k \cdot T(B_k)$. Here, one should note that the stochastic integral of f is nothing but the random variable that corresponds to f in the isometry T.

Next, the definition is extended to any function $f \in \mathbb{L}^2(m)$ as follows. Take a sequence f_n of elementary functions that converges to f in $\mathbb{L}^2(m)$. Since the integrals of the f_n's are the random variables corresponding to f_n in the isometry, such integrals form a Chauchy sequence and therefore converge to a limit point in $\mathbb{L}^2(v)$. This limit is by definition the integral of f and it is denoted by $\int_{(-\pi,\pi]} f(\omega)dT(\omega)$. Again, the integral of f is nothing but the random variable corresponding to f in the isometry T.

The notion of stochastic integral can now be applied to the functions $e^{i\omega t}$. By construction, we obtain $\int_{(-\pi,\pi]} e^{i\omega t} dT(\omega) = v_t$.

We have proved the following result.

THEOREM 6.10 *There exists an orthogonal stochastic measure T on $((-\pi,\pi],\mathscr{B}(-\pi,\pi])$ such that*

$$v_t = \int_{(-\pi,\pi]} e^{i\omega t} dT(\omega).$$

Moreover, $E[|Z(B)|^2] = m(B), \forall B \in \mathscr{B}(-\pi,\pi]$. □

A new interpretation of stationary processes

The above definition of stochastic integral delivers a new interpretation of a stationary process that points directly to its inborn structure and provides an insightful standpoint in many applications.

Consider equation $v_t = \int_{(-\pi,\pi]} e^{i\omega t} dT(\omega)$. Let us partition the interval $(-\pi,\pi]$ in a large, though finite, number of subintervals of equal length: $(-\pi,\pi] = (-\pi,-\pi + 2\pi\frac{1}{N}] \cup (-\pi + 2\pi\frac{1}{N}, -\pi + 2\pi\frac{2}{N}] \cup \cdots \cup (\pi - 2\pi\frac{1}{N}, 2\pi] = \cup_{k=1}^{N} B_k$. Then, function $e^{i\omega t}$ can be approximated by $\sum_{k=1}^{N} e^{i\omega_k t} 1(B_k)$, $\omega_k \in B_k$. Correspondingly, v_t can be approximated by the stochastic integral of this latter function, leading to

$$v_t \approx \sum_{k=1}^{N} e^{i\omega_k t} T(B_k).$$

This expression delivers an interpretation of process v_t which is very useful for an intuitive understanding of its structure:

> A stationary process v_t is given by the linear combination of uncorrelated random variables $T(B_k)$. Each variable has a variance $m(B_k)$ (remember that T is an isometry) and it is modulated in time by the harmonic function $e^{i\omega_k t}$, which oscillates at the frequency ω_k.

6.4 Multivariable stationary processes

The theory of wide-sense stationary processes extends in a natural way to the multivariable case, and this is dealt with here in brief summary. We consider 2-component processes solely, as this case captures all relevant apsects.

Let $v_t : \Omega \to \mathbb{C}^2$ have components $v_t^{(1)}$ and $v_t^{(2)}$. v_t is wide-sense stationary if

$$E[v_t] = E[v_0], \quad \forall t,$$

and

$$E\left[(v_{t+\ell} - E[v_{t+\ell}])\,\overline{(v_t - E[v_t])^T}\right] = E\left[(v_\ell - E[v_\ell])\,\overline{(v_0 - E[v_0])^T}\right], \quad \forall t, \ell,$$

where a bi-dimensional stochastic variable is identified with the vector of its components where it needs be.

Notations and terminology are the same as in the mono-variate case, so e.g. we call $\gamma_\ell := E\left[(v_\ell - E[v_\ell])\,\overline{(v_0 - E[v_0])^T}\right]$ the process auto-covariance function. γ_ℓ is a 2×2 matrix, where the diagonal elements are the auto-covariance functions of $v_t^{(1)}$ and $v_t^{(2)}$ and the extra-diagonal elements measure the cross-covariance between $v_t^{(1)}$ and $v_t^{(2)}$. When we want to emphasize this fact, we also call $\gamma_\ell^{(1,2)}$ (the $(1,2)$ element of γ_ℓ) the cross-covariance function between $v_t^{(1)}$ and $v_t^{(2)}$.

If $v_t : \Omega \to \mathbb{C}^2$ is wide-sense stationary, so is $\alpha v_t^{(1)} + \beta v_t^{(2)}$ for any choice of complex numbers α and β.

The spectral theory is extended more easily to the bi-dimensional case by first introducing the spectral representation and then the spectral measure (so reversing the order adopted in the one-dimensional case), so we follow this route.

The spectral representation is simply a componentwise concept: from the one-dimensional theory, $v_t^{(1)}$ has associated an orthogonal stochastic measure $T^{(1)}$ such that $v_t^{(1)} = \int_{(-\pi,\pi]} e^{i\omega t} dT^{(1)}(\omega)$; similarly $v_t^{(2)} = \int_{(-\pi,\pi]} e^{i\omega t} dT^{(2)}(\omega)$. It is worth remarking that $T^{(1)}$ and $T^{(2)}$ carry all information on the bi-dimensional process since $v_t^{(1)}$ and $v_t^{(2)}$ can be fully reconstructed from $T^{(1)}$ and $T^{(2)}$. So, one can e.g. reconstruct the cross-covariance function from $T^{(1)}$ and $T^{(2)}$.

In contrast to the orthogonal stochastic measure, the spectral measure is a matrix concept: it is a 2×2 matrix of the form

$$m = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}.$$

It needs to be so because one cannot reconstruct the cross-covariance from the auto-covariance $\gamma_\ell^{(1,1)}$ and $\gamma_\ell^{(2,2)}$ only. The reader is invited to reflect on this point by considering the following two situations:

(i) $v_t^{(1)}$ and $v_t^{(2)}$ are zero-mean, unitary-variance, white and mutually uncorrelated: $E[v_t^{(1)} \overline{v}_\tau^{(2)}] = 0, \forall t, \tau$;

(ii) $v_t^{(1)}$ and $v_t^{(2)}$ are zero-mean, unitary variance, white and $v_t^{(1)} = v_t^{(2)}, \forall t$.

Here, both in (i) and (ii) $\gamma_\ell^{(1,1)} = \gamma_\ell^{(1,1)} = 1$ for $\ell = 0$ and $\gamma_\ell^{(1,1)} = \gamma_\ell^{(1,1)} = 0$ for $\ell \neq 0$. However, $\gamma_\ell^{(1,2)}$ are different in the two cases.

Before proceeding any further in the definition of m, we need to extend our notion of measure to complex-valued signed-measures, as m_{12} and m_{21} are measures of this type.

DEFINITION 6.11 (complex-valued signed-measure) *Let $\mathscr{X}$ be a σ-algebra. A function $m : \mathscr{X} \to [-\infty, \infty]$ such that no two sets B_1 and B_2 exist with $m(B_1) = \infty$ and $m(B_2) = -\infty$ is called a signed-measure if, for any countable collection of pairwise disjoint sets $B_k \in \mathscr{X}$, $k = 1, 2, \ldots$, the following property (σ-additivity) holds*

$$m(\cup_{k=1}^{\infty} B_k) = \sum_{k=1}^{\infty} m(B_k). \tag{6.13}$$

A complex-valued signed-measure m is given by $m_1 + im_2$, where m_1 and m_2 are signed-measures. □

The assumption that no two sets B_1 and B_2 exist with $m(B_1) = \infty$ and $m(B_2) = -\infty$ rules out the possibility of indeterminate forms $\infty - \infty$. By a comparison with Definition 1.7,

we see that a signed-measure is simply a measure where the positivity requirement has been relaxed.

We are now ready to define the spectral measure. For $B \in \mathscr{B}(-\pi,\pi]$, let:

$$\begin{aligned} m_{11}(B) &= E[|T^{(1)}(B)|^2], & m_{22}(B) &= E[|T^{(2)}(B)|^2], \\ m_{12}(B) &= E[T^{(1)}(B)\overline{T}^{(2)}(B)], & m_{21}(B) &= E[\overline{T}^{(1)}(B)T^{(2)}(B)]. \end{aligned}$$

m_{11} and m_{22} are the usual spectral measures for processes $v_t^{(1)}$ and $v_t^{(2)}$. From the definition, $m_{12} = \overline{m}_{21}$ and these are complex-valued signed-measures. To see this, we need to verify the σ-additivity property, i.e. $m_{12}\left(\cup_{k=1}^{\infty} B_k\right) = \sum_{k=1}^{\infty} m_{12}(B_k)$, and this requires a bit of extra investigation: we shall prove that

$$E[T^{(1)}(B_1)\overline{T}^{(2)}(B_2)] = 0, \quad \textit{whenever } B_1 \cap B_2 = \emptyset, \tag{6.14}$$

from which the σ-additivity follows:

$$\begin{aligned} m_{12}(\cup_{k=1}^{\infty} B_k) &= E[T^{(1)}(\cup_{k=1}^{\infty} B_k)\,\overline{T}^{(2)}(\cup_{k=1}^{\infty} B_k)] \\ &= E\left[\left(\sum_{k=1}^{\infty} T^{(1)}(B_k)\right)\left(\sum_{k=1}^{\infty} \overline{T}^{(2)}(B_k)\right)\right] \\ &= \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} E\left[T^{(1)}(B_k)\overline{T}^{(2)}(B_j)\right] \\ &= \sum_{k=1}^{\infty} E\left[T^{(1)}(B_k)\overline{T}^{(2)}(B_k)\right] \quad (\textit{using } (6.14)) \\ &= \sum_{k=1}^{\infty} m_{12}(B_k). \end{aligned}$$

To prove (6.14), start by considering the stationary process $v_t = \alpha v_t^{(1)} + \beta v_t^{(2)}$ and note that the orthogonal stochastic measure T associated to v_t is $\alpha T^{(1)} + \beta T^{(2)}$. Thus,

$$\begin{aligned} 0 &= E[T(B_1)\overline{T}(B_2)] \quad (\textit{since } B_1 \cap B_2 = \emptyset) \\ &= E[(\alpha T^{(1)}(B_1) + \beta T^{(2)}(B_1))(\overline{\alpha}\overline{T}^{(1)}(B_2) + \overline{\beta}\overline{T}^{(2)}(B_2))] \qquad (6.15) \\ &= \alpha\overline{\beta}E[T^{(1)}(B_1)\overline{T}^{(2)}(B_2)] + \beta\overline{\alpha}E[T^{(2)}(B_1)\overline{T}^{(1)}(B_2)]. \qquad (6.16) \\ &\quad (\textit{since } E[T^{(1)}(B_1)\overline{T}^{(1)}(B_2)] = E[T^{(2)}(B_1)\overline{T}^{(2)}(B_2)] = 0) \qquad (6.17) \end{aligned}$$

Taking $\alpha = \beta = 1$ first, and $\alpha = 1$, $\beta = i$ then, yields

$$\begin{aligned} 0 &= E[T^{(1)}(B_1)\overline{T}^{(2)}(B_2)] + E[T^{(2)}(B_1)\overline{T}^{(1)}(B_2)]) \\ 0 &= -iE[T^{(1)}(B_1)\overline{T}^{(2)}(B_2)] + iE[T^{(2)}(B_1)\overline{T}^{(1)}(B_2)], \end{aligned}$$

from which $E[T^{(1)}(B_1)\overline{T}^{(2)}(B_2)] = 0$ follows, so proving (6.14).

We want to finally recall that Herglotz Theorem 6.6 extends naturally to the multivariable case:

$$\gamma_\ell^{(i,j)} = \int_{(-\pi,\pi]} e^{i\omega\ell} dm_{ij}(\omega), \quad \ell = \ldots,-2,-1,0,1,2,\ldots, \quad i,j = 1,2. \tag{6.18}$$

For $i = j = 1$ and $i = j = -1$, this is the standard Herglotz theorem applied componentwise. As for $i \neq j$, take a partition $B_k = (-\pi + 2\pi\frac{k-1}{N}, -\pi + 2\pi\frac{k}{N}]$, $k = 1,\ldots,N$, of $(-\pi,\pi]$ and let $\sum_{k=1}^N e^{i\omega_k\ell} T^{(1)}(B_k)$ and $\sum_{k=1}^N e^{i\omega_k 0} T^{(2)}(B_k) = \sum_{k=1}^N T^{(2)}(B_k)$ be ε-approximations (in the $\mathbb{L}^2$-norm) of $v_\ell^{(1)}$ and $v_0^{(2)}$. We have

$$\begin{aligned} \gamma_\ell^{(1,2)} &= E[v_\ell^{(1)}\overline{v}_0^{(2)}] \\ &\approx E\left[\left(\sum_{k=1}^N e^{i\omega_k\ell} T^{(1)}(B_k)\right)\left(\sum_{k=1}^N \overline{T}^{(2)}(B_k)\right)\right] \\ &= \sum_{k=1}^N \sum_{j=1}^N e^{i\omega_k\ell} E[T^{(1)}(B_k)\overline{T}^{(2)}(B_j))] \\ &= \sum_{k=1}^N e^{i\omega_k\ell} E[T^{(1)}(B_k)\overline{T}^{(2)}(B_k))] \quad (\textit{using } (6.14)) \\ &= \sum_{k=1}^N e^{i\omega_k\ell} m_{12}(B_k) \\ &\approx \int_{(-\pi,\pi]} e^{i\omega\ell} dm_{12}(\omega), \end{aligned}$$

where the "$\approx$" become "$=$" in the limit when $\varepsilon \to 0$, so proving (6.18).

Bibliography

[1] P. Billingsley. *Probability and measure*. John Wiley and Sons, 1995.

[2] K.L. Chung. *A course in probability theory*. Academic Press, 1974.

[3] P.R. Halmos. *Measure theory*. Van Nostrand, New York, 1950.

[4] L.Devroye, L.Gyorfi, and G.Lugosi. *A probabilistic theory of pattern recognition*. Springer-Verlag, New York, 1996.

[5] M. Loeve. *Probability theory*. Springer-Verlag, New York, 1978.

[6] W. Rudin. *Real and complex analysis*. McGraw-Hill, New York, 1966.

[7] A. N. Shiryayev. *Probability*. Springer Verlag, New York, 1984.