Probability Theory II

Spring 2014

Peter Orbanz

Contents

CHAPTER 1

Martingales, continued

Martingales are first and foremost a tool to establish the existence and properties of random limits. The basic limit theorems of probability (the law of large numbers and Glivenko-Cantelli) establish that averages of i.i.d. variables converge to their expected value. (The central limit theorem also falls into this class—it additionally establishes a rate of convergence.) The sequence of averages is a random sequence, but it *completely derandomizes in the limit.* For more complicated processes—typically, if the variables are stochastically dependent—the limit might be random. In general, random limits are very hard to handle mathematically, since we have to precisely quantify the effect of dependencies and cope with the randomness aggregating as we get further into the sequence.

It turns out that it is possible to control dependencies and randomness if a process has the martingale property (or is at least a sub- or supermartingale). Hence, one perspective on martingales is the following:

Martingales are tools for working with random limits.

They are not the only such tools, but there are very few others.

1.1. Martingales indexed by partially ordered sets

Martingales have already been discussed in Probability Theory I, where they were indexed by values in $\mathbb{N}$ (so-called "discrete-time martingales"). We will briefly discuss martingales in "continuous time"—that is, martingales indexed by $\mathbb{R}_+$. However, martingales can much more generally be defined for index sets that need not be totally ordered, and we will proof two useful results for such general martingales.

Partially ordered index sets. Let $\mathbb{T}$ be a set. Recall that a binary relation $\preceq$ on $\mathbb{T}$ is called a **partial order** if it is

(1) reflexive, that is $s \preceq s$ for every $s \in \mathbb{T}$.
(2) antisymmetric, that is if $s \preceq t$ and $t \preceq s$, then $s = t$.
(3) transitive, that is if $s \preceq t$ and $t \preceq u$, then $s \preceq u$.

In general, a partially ordered set may contain elements that are not comparable, i.e. some s, t for which neither $s \preceq t$ nor $t \preceq s$ (hence "partial"). If all pairs of elements are comparable, the partial order is called a **total order**.

In various contexts—including martingales and the construction of stochastic processes—we need partially ordered index sets. We have to be careful, though: Using arbitrary partially ordered sets can lead to all kinds of pathologies. Roughly speaking, the problem is that a partially ordered set can decompose into subsets between which elements cannot be compared at all, as if we were indexing arbitrarily by picking indices from completely unrelated index sets. For instance, a partially ordered set could contain two sequences $s_1 \preceq s_2 \preceq s_3 \preceq \dots$ and $t_1 \preceq t_2 \preceq t_3 \preceq \dots$ of elements which both grow larger and larger in terms of the partial order, but whose elements are completely incomparable between the sequences. To avoid such pathologies, we impose an extra condition:

$$\text{If } s, t \in \mathbb{T}, \text{ there exists } u \in \mathbb{T} \text{ such that } s \preceq u \text{ and } t \preceq u\ . \tag{1.1}$$

A partially ordered set $(\mathbb{T}, \preceq)$ which satisfies (1.1) is called a **directed set**.

1.1 Example. Some examples of directed sets:

(a) The set of all subsets of an arbitrary set, ordered by inclusion.
(b) The set of all finite subsets of an infinite set, ordered by inclusion.
(c) The set of all positive definite $n \times n$ matrices over $\mathbb{R}$, in the Löwner partial order.
(d) Obviously, any totally ordered set (such as $\mathbb{N}$ or $\mathbb{R}$ in the standard order) is directed.

◁

Just as we can index a family of variables by $\mathbb{N}$ and obtain a sequence, we can more generally index it by a directed set; the generalization of a sequence so obtained is called a *net*. To make this notion precise, let $\mathcal{X}$ be a set. Recall that, formally, an **(infinite) sequence** in $\mathcal{X}$ is a mapping $\mathbb{N} \to \mathcal{X}$, that is, each index s is mapped to the sequence element x_i. We usually denote such a sequence as $(x_i)_{i \in \mathbb{N}}$, or more concisely as (x_i).

1.2 Definition. Let $(\mathbb{T}, \preceq)$ be a directed set. A **net** in a set $\mathcal{X}$ is a function $x : \mathbb{T} \to \mathcal{X}$, and we write $x_t := x(t)$ and denote the net as $(x_t)_{t \in \mathbb{T}}$. ◁

Clearly, the net is a sequence if $(\mathbb{T}, \preceq)$ is the totally ordered set $(\mathbb{N}, \leq)$. Just like sequences, nets may converge to a limit. The definition is analogous to that of a Cauchy sequence:

1.3 Definition. A net $(x_t)_{t \in I}$ is said to **converge** to a point x if, for every open neighborhood U of x, there exists an index $t_0 \in \mathbb{T}$ such that

$$x_t \in U \qquad \text{whenever } t_0 \preceq t\ . \tag{1.2}$$

◁

Nets play an important role in real and functional analysis: Too establish certain properties in spaces more general than $\mathbb{R}^d$, we have to demand e.g. that every net satisfying certain properties converges (not just every sequence). Sequences have much stronger properties than nets; for example, in any topological space, the set consisting of all elements of a convergent sequence and its limit is a compact set. The same needs not be true for for nets.

Filtrations and martingales. The discrete-time martingales discussed in Probability I are a special type of *random sequences*. Analogously, the more general martingales we discuss in the following are a form of *random nets*.

Let $(\mathbb{T}, \preceq)$ be a directed set. By a **filtration**, we mean a family $\mathcal{F} = (\mathcal{F}_t)_{t \in \mathbb{T}}$ of σ-algebras $\mathcal{F}_i$, indexed by the elements of $\mathbb{T}$, which satisfy

$$s \preceq t \quad \Longrightarrow \quad \mathcal{F}_s \subset \mathcal{F}_t\ . \tag{1.3}$$

The case $\mathbb{T} = \mathbb{N}$, which you have already encountered in Probability Theory I, is often called the discrete-time case; similarly, $\mathbb{T} = \mathbb{R}_+$ is called the continuous-time case. The filtration property states that each σ-algebra $\mathcal{F}_s$ contains all preceding ones. For a filtration, there is also a uniquely determined, smallest σ-algebra which contains all σ-algebras in $\mathcal{F}$, namely

$$\mathcal{F}_\infty := \sigma\Bigl(\bigcup_{s\in\mathbb{T}}\mathcal{F}_s\Bigr) \ . \tag{1.4}$$

Clearly, every random sequence or net $(X_s)_{s\in\mathbb{T}}$ is adapted to the filtration

$$\mathcal{F}_t := \sigma\Bigl(\bigcup_{s\preceq t}\sigma(X_s)\Bigr) \ , \tag{1.5}$$

which is called the **canonical filtration** of (X_s).

Now, let $(\mathbb{T},\preceq)$ be a partially ordered set and $\mathcal{F}=(\mathcal{F}_s)_{s\in\mathbb{T}}$ a filtration. As in the discrete-time case, we call a family $(X_s)_{s\in\mathbb{T}}$ of random variables **adapted** to $\mathcal{F}$ if X_s is $\mathcal{F}_s$-measurable for every s. Then $(X_s,\mathcal{F}_s)_{s\in\mathbb{T}}$ is called a **martingale** if it satisfies

$$X_s =_{\text{a.s.}} \mathbb{E}[X_t|\mathcal{F}_s] \qquad \text{whenever } s\preceq t \ , \tag{1.6}$$

or equivalently,

$$\forall A\in\mathcal{F}_s : \qquad \int_A X_s d\mathbb{P} = \int_A X_t d\mathbb{P} \qquad \text{whenever } s\preceq t \ . \tag{1.7}$$

If it the conditions only with equality weakened to $\leq$ (that is, $X_s\leq\mathbb{E}(X_t|\mathcal{F}_s)$ or $\int X_s d\mathbb{P}\leq\int X_t d\mathbb{P}$), it is called a **submartingale**; for $\geq$, it is called a **supermartingale**.

Our objective in the following. The objective of martingale convergence results is ideally to establish, for a given martingale $(X_s,\mathcal{F}_s)$, the existence of a limit random variable X_∞ which satisfies

$$X_s =_{\text{a.s.}} \mathbb{E}[X_\infty|\mathcal{F}_s] \qquad \text{for all } s\in\mathbb{T} \ . \tag{1.8}$$

Note that a result of the form (1.8) is more than just a convergence result: (1.8) is a representation theorem.

1.2. Notions of convergence for martingales

You will recall from the martingale convergence results discussed in Probability I that two types of convergence are of interest: Convergence almost surely, and in $\mathbf{L}_1$. We should briefly review why so.

In terms of convergence of a random variable, the strongest result we can hope for is almost sure convergence, i.e.

$$X_s(\omega)\to X_\infty(\omega) \qquad \mathbb{P}\text{-a.s.} \tag{1.9}$$

However, if (X_s) is a martingale, we also know that

$$X_s =_{\text{a.s.}} \mathbb{E}[X_t|\mathcal{F}_s] \qquad \text{for all } s \text{ with } s\preceq t \ . \tag{1.10}$$

This holds no matter how large (in the partial order) we make t. We hence might be forgiven to hope that (1.8) holds, too—but that is not generally true. What do we need to establish (1.8)? Suppose (X_s) converges to X_∞ in $\mathbf{L}_1$, i.e.

$$\lim\int_\Omega |X_s(\omega)-X_\infty(\omega)|\mathbb{P}(d\omega) = \lim\mathbb{E}\bigl[|X_s-X_\infty|\bigr]\to 0 \ . \tag{1.11}$$

(Note: If $\mathbb{T}$ is a directed set, then $(\mathbb{E}[|X_s-X_\infty|])_{s\in\mathbb{T}}$ is a net in $\mathbb{R}$, so $\mathbf{L}_1$ convergence means this net converges to the point 0 in the sense of Definition 1.3.) It is easy to verify the following fact:

1.4 Fact. If (X_s) converges in $\mathbf{L}_1$ to X_∞, then $(X_s \mathbb{I}_A)_i$ converges in $\mathbf{L}_1$ to $X_\infty \mathbb{I}_A$ for every measurable set A. $\triangleleft$

Hence, for every index $s \in \mathbb{T}$, $\mathbf{L}_1$ convergence implies

$$\lim \int_A X_s d\mathbb{P} = \lim \int_\Omega X_s \mathbb{I}_A d\mathbb{P} = \int_\Omega X_\infty \mathbb{I}_A d\mathbb{P} = \int_A X_\infty d\mathbb{P} \tag{1.12}$$

for every $A \in \mathcal{F}_s$. By the martingale property (1.6), the sequence or net of integrals is additionally constant, that is

$$\int_A X_s d\mathbb{P} = \int_A X_t d\mathbb{P} \quad \text{for all pairs } s \leq t \text{ and hence } \int_A X_s d\mathbb{P} = \int_A X_\infty d\mathbb{P} \, , \tag{1.13}$$

which is precisely (1.8).

Recall from your discussion of notions of convergence for random variables in [Probability I, Chapter 17] how the different types of convergence relate to each other:

$$\begin{tikzcd}
& \boxed{\text{weakly}} & & \\
\boxed{\text{almost surely}} \arrow[r, shift left] & \boxed{\text{in probability}} \arrow[u] \arrow[l, shift left, "\text{subsequence}"] & \boxed{\mathbf{L}_1} \arrow[l] & \boxed{\mathbf{L}_p} \arrow[l]
\end{tikzcd} \tag{1.14}$$

Neither does almost sure convergence imply $\mathbf{L}_1$ convergence, nor vice versa. This is why "strong" martingale convergence results are typically of the form, "If (...), the martingale converges to a limit almost surely and in $\mathbf{L}_1$".

1.3. Uniform integrability

Recall from e.g. Theorem 27.1 in Probability I that the martingale property and a very mild supremum condition suffice to establish almost sure convergence: If $(X_n, \mathcal{F}_n)$ is a submartingale satisfying $\sup_n \mathbb{E}[\max\{0, X_n\}] < \infty$, then $X_\infty = \lim X_n$ exists almost surely. Moreover, the limit is finite almost surely, and integrable. The result does not yield $\mathbf{L}_1$ convergence. The property we need for martingales to converge in $\mathbf{L}_1$, and hence to establish (1.8), is uniform integrability. As a reminder:

1.5 Definition. Let $\mathbb{T}$ be an index set and $\{f_s | s \in \mathbb{T}\}$ a family of real-valued functions. The family is called **uniformly integrable** with respect to a measure μ if, for every $\varepsilon > 0$, there exists a positive function $g \geq 0$ such that

$$\int_{\{|f_s| \geq g\}} |f_s| d\mu \leq \varepsilon \qquad \text{for all } s \in \mathbb{T} \, . \tag{1.15}$$

$\triangleleft$

Clearly, any finite set of random variables is uniformly integrable. The definition is nontrivial only if the index set is infinite. Here is a primitive example: Suppose the functions f_s in (1.15) are the constant functions on $[0, 1]$ with values $1, 2, \ldots$. Each function by itself is integrable, but the set is obviously not uniformly integrable. If the functions are in particular random variables $X_s : \Omega \to \mathbb{R}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, Definition 1.5 reads: For every ε, there is a positive random variable Y such that

$$\mathbb{E}\big[|X_s|\cdot\mathbb{I}\{|X_s|\geq Y\}\big]\leq\varepsilon\qquad\text{for all } s\in\mathbb{T}\;. \tag{1.16}$$

The definition applies to martingales in the obvious way: $(X_s,\mathcal{F}_s)_{s\in\mathbb{T}}$ is called a **uniformly integrable martingale** if it is a martingale and the family $(X_s)_{s\in\mathbb{T}}$ of functions is uniformly integrable.

Verifying (1.15) for a given family of functions can be pretty cumbersome, but can be simplified using various useful criteria. We recall two of them from [Probability I, Theorem 27.2]:

1.6 Lemma. *A family $(X_s)_{s\in\mathbb{T}}$ of real-valued random variables with finite expectations is uniformly integrable if there is a random variable Y with $\mathbb{E}[Y]<\infty$ such that either of the following conditions holds:*

(1) *Each X_s satisfies*

$$\int_{\{|X_s|\geq\alpha\}}|X_s|d\mu\leq\int_{\{|X_s|\geq\alpha\}}Y\,d\mu\qquad\textit{for all } \alpha>0\;. \tag{1.17}$$

(2) *Each X_s satisfies $|X_s|\leq Y$.*

◁

We now come to our first result for martingales on general index sets, which shows two important facts:

(1) We can easily obtain uniformly integrable martingales by positing a limit variable X and "filtering" it.
(2) Uniform integrability is *necessary* to for (1.8) to hold.

1.7 Theorem. *Let X be an integrable, real-valued random variable on a probability space $(\Omega,\mathcal{A},\mathbb{P})$. Let $(\mathbb{T},\preceq)$ be a partially ordered set, and $\mathcal{F}=(\mathcal{F}_s)_{s\in\mathbb{T}}$ a filtration in $\mathcal{A}$. Then*

$$\big(\mathbb{E}[X|\mathcal{F}_s],\mathcal{F}_s\big)_{s\in\mathbb{T}} \tag{1.18}$$

is a uniformly integrable martingale. ◁

Proof. $(X_s,\mathcal{F}_s)$ is a martingale by construction: For any pair $s\preceq t$ of indices,

$$\mathbb{E}[X_t|\mathcal{F}_s]=_{\text{a.s.}}\mathbb{E}[\mathbb{E}[X|\mathcal{F}_t]|\mathcal{F}_s]\overset{\overset{\mathcal{F}_s\subseteq\mathcal{F}_t}{\downarrow}}{=}_{\text{a.s.}}\mathbb{E}[X|\mathcal{F}_s]=_{\text{a.s.}}X_s\;. \tag{1.19}$$

To show that it is also uniformly integrable, we use Jensen's inequality [e.g. Probability I, Theorem 23.9]—recall: $\phi(\mathbb{E}[X|\mathcal{C}])\leq\mathbb{E}[\phi(X)|\mathcal{C}]$ for any convex function ϕ—which implies

$$|X_s|=_{\text{a.s.}}\big|\mathbb{E}[X|\mathcal{F}_s]\big|\leq_{\text{a.s.}}\mathbb{E}\big[|X|\big|\mathcal{F}_s\big]\;. \tag{1.20}$$

Hence, for every $A\in\mathcal{F}_s$, we have

$$\int_A|X_s|dP\leq\int_A|X|dP\;, \tag{1.21}$$

which holds in particular for $A:=\{|X_s|\geq\alpha\}$. By Lemma 1.6, the martingale is uniformly integrable. □

1.4. Convergence of martingales with directed index sets

You have already seen martingale convergence results for the discrete time case $\mathbb{T} = \mathbb{N}$ in the previous class. The next result generalizes the martingale convergence theorem [Probability I, Theorem 27.3] to the case of arbitrary directed index sets.

1.8 Martingale convergence theorem. *Let $(X_s, \mathcal{F}_s)_{s\in\mathbb{T}}$ be a martingale with directed index set $\mathbb{T}$. If the martingale is uniformly integrable, there exists a random variable X_∞ such that $(X_s, \mathcal{F}_s)_{s\in\mathbb{T}\cup\{\infty\}}$ is a martingale; that is,*

$$X_s =_{\text{a.s.}} \mathbb{E}[X_\infty|\mathcal{F}_s] \qquad \textit{for all } s \in \mathbb{T} \; . \tag{1.22}$$

The variable X_∞ is integrable, and uniquely determined up to almost sure equivalence. ◁

Proof. The key idea is to use the directed structure of the index set to reduce to the martingale convergence theorem for the discrete-time case, Theorem 27.3 in [Probability I].

Step 1: The net satisfies the Cauchy criterion. We have to show that the random net $(X_s)_{s\in\mathbb{T}}$ converges in $\mathbf{L}_1$; in other words, that

$$\forall\varepsilon > 0 \; \exists s_0 \in \mathbb{T} : \quad \mathbb{E}[|X_t - X_u|] \leq \varepsilon \quad \text{ for all } t, u \text{ with } s_0 \preceq t \text{ and } s_0 \preceq u \; . \tag{1.23}$$

This follows by contradiction: Suppose (1.23) was not true. Then we could find an $\varepsilon > 0$ and a sequence $s_1 \preceq s_2 \preceq \ldots$ of indices such that $\mathbb{E}[|X_{s_{n+1}} - X_{s_n}|] \geq \varepsilon$ for all n. Since the $(X_s, \mathcal{F}_s)_{s\in\mathbb{T}}$ is a uniformly integrable martingale, so is the sub-family $(X_{s_n}, \mathcal{F}_{s_n})_{n\in\mathbb{N}}$—but we have just shown that is does not converge, which contradicts Theorem 27.3. Thus, (1.23) holds.

Step 2: Constructing the limit. Armed with (1.23), we can explicitly construct the limit, which we do using a specifically chosen subsequence: Choose ε in (1.23) consecutively as $1/1, 1/2, \ldots$. For each such $\varepsilon = 1/n$, choose an index s_n satisfying (1.23). Since $\mathbb{T}$ is directed, we can choose these indices increasingly in the partial order, $s_{1/1} \preceq s_{1/2} \preceq \ldots$. Again by [Probability I, Theorem 27.3], this makes $(X_{s_{1/n}})_n$ a convergent martingale, and there is a limit variable X.

Step 3: The entire net converges to the limit X. For the sequence constructed above, if $n \leq m$, then $s_{1/n} \preceq s_{1/m}$. Substituting into (1.23) shows that

$$\mathbb{E}[|X_{s_{1/m}} - X_{s_{1/n}}|] \leq \frac{1}{n} \qquad \text{and hence} \qquad \mathbb{E}[|X - X_s|] \leq \frac{1}{n} \qquad \text{for all } s \text{ with } s_n \preceq s \; .$$

Hence, the entire net converges to X.

Step 4: X is determined almost everywhere. We have to show that the event

$$N := \{\omega \,|\, (X_{t_m}(\omega)) \text{ does not converge}\} \tag{1.24}$$

is a null set in $\mathcal{F}_\infty$. Recall from the convergence diagram (1.14) that $\mathbf{L}_1$ is pretty high up in the food chain: Convergence of a sequence of random variables in $\mathbf{L}_1$ implies convergence in probability, which in turn implies existence of an almost surely convergent subsequence. Since $(X_{s_{1/n}})$ by construction converges to X in $\mathbf{L}_1$, this means there is a subsequence $(t_1, t_2, \ldots)$ of $(s_{1/n})_n$ such that $(X_{t_m})_m$ converges almost surely to X.

Since X_s is a martingale, we also have $X_{s_{1/n}} =_{\text{a.s.}} \mathbb{E}[X_{t_m}|\mathcal{F}_{s_{1/n}}]$ for any index $s_{1/n}$ in the entire sequence which satisfies $s_{1/n} \preceq t_m$, and so the *entire* sequence

$(X_{1/n})$ converges almost surely. Hence, N is a null set. Since all random variables X_s and X are $\mathcal{F}_\infty$-measurable, $N \in \mathcal{F}_\infty$.

Step 5: X does as we want. We have to convince ourselves that X indeed satisfies (1.22), and hence that

$$\int_A X_s d\mathbb{P} = \int X_A d\mathbb{P} \qquad \text{for all } A \in \mathcal{F}_s \; . \tag{1.25}$$

Since the entire net $(X_s)_{s\in\mathbb{T}}$ converges to X in $\mathbf{L}_1$, we can use Fact 1.4: $\mathbb{I}_A X_s$ also converges to $\mathbb{I}_A X$ in $\mathbf{L}_1$ for any $A \in \mathcal{F}_s$. Hence,

$$\int_A X_s d\mathbb{P} = \int_\Omega \mathbb{I}_A X_s d\mathbb{P} = \int_\Omega \mathbb{I}_A X d\mathbb{P} = \int_A X d\mathbb{P} \; . \tag{1.26}$$

Step 6: X is unique up to a.s. equivalence. Finally, suppose X' is another $\mathcal{F}_\infty$-measurable random variable satisfying (1.22). We have to show $X =_{\text{a.s.}} X'$. Since both variables are $\mathcal{F}_\infty$-measurable, we have to show that

$$\int_A X d\mathbb{P} = \int_A X' d\mathbb{P} \tag{1.27}$$

holds for all $A \in \mathcal{F}_\infty$. We will proof this using a standard proof technique, which I do not think you have encountered before. Since it is important and very useful, let me briefly summarize it in general terms before we continue:

1.9 Remark [Proof technique]. What we have to show in this step is that some property—in this case, (1.27)—is satisfied on all sets in a given σ-algebra $\mathcal{C}$ (here: $\mathcal{F}_\infty$). To solve problems of this type, we define two set systems:

(1) The set $\mathcal{D}$ of all sets $A \in \mathcal{C}$ which *do* satisfy the property. At this point, we do not know much about this system, but we know that $\mathcal{D} \subset \mathcal{C}$.
(2) The set $\mathcal{E}$ of all $A \in \mathcal{C}$ for which we *already know* the property is satisfied.

Then clearly,

$$\mathcal{E} \subset \mathcal{D} \subset \mathcal{C} \; . \tag{1.28}$$

What we have to show is $\mathcal{D} = \mathcal{C}$.

The proof strategy is applicable if we can show that: (1) $\mathcal{E}$ is a generator of $\mathcal{C}$, i.e. $\sigma(\mathcal{E}) = \mathcal{C}$; (2) $\mathcal{E}$ is closed under finite intersections; and (3) $\mathcal{D}$ is closed under differrences and increasing limits. If (2) and (3) are true, the monotone class theorem [Probability I, Theorem 6.2] tells us that $\sigma(\mathcal{E}) \subset \mathcal{D}$. In summary, (1.28) then becomes

$$\mathcal{C} = \sigma(\mathcal{E}) \subset \mathcal{D} \subset \mathcal{C} \; , \tag{1.29}$$

and we have indeed shown $\mathcal{D} = \mathcal{C}$, i.e. our property holds on all of $\mathcal{C}$. ◁

Now back to the proof at hand: In this case, we define $\mathcal{D}$ as the set of all $A \in \mathcal{F}_\infty$ which satisfy (1.27). We note that (1.27) is satisfied whenever $A \in \mathcal{F}_s$ for some index s, and so we choose $\mathcal{E}$ as

$$\mathcal{E} = \bigcup_{s\in\mathbb{T}} \mathcal{F}_s \; . \tag{1.30}$$

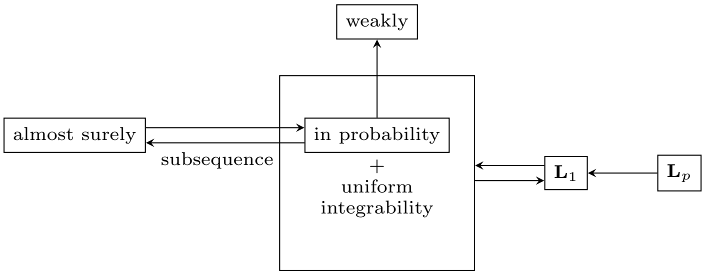
Then (1.28) holds (for $\mathcal{C} = \mathcal{F}_\infty$), and we have left to show $\mathcal{D} = \mathcal{F}_\infty$. Recall that $\sigma(\mathcal{E}) = \mathcal{F}_\infty$ by definition, so one requirement is already satisfied.

The system $\mathcal{D}$ is closed under differences and increasing limits: Suppose $A, B \in \mathcal{D}$. Then (1.27) is satisfied for $A \setminus B$, since we only have to substract the equations for A and B, so $\mathcal{D}$ is closed under differences. Similarly, suppose $A_1 \subset A_2 \subset \dots$ is a sequence of sets which are all in $\mathcal{D}$, and $A := \cup_n A_n$. By definition of the integral, $\int_A X d\mathbb{P} = \lim_n \int_{A_n} X d\mathbb{P}$. Applying the limit on both sides of (1.27) shows $A \in \mathcal{D}$. The set system $\mathcal{E}$ is closed under finite intersections: Suppose $A \in \mathcal{F}_s$ and $B \in \mathcal{F}_t$ for any two $s, t \in \mathbb{T}$. Since $\mathbb{T}$ is directed, there is some $u \in \mathbb{T}$ with $s, t \preceq u$, and hence $A, B \in \mathcal{F}_u$ and $A \cap B \in \mathcal{F}_u \subset \mathcal{E}$. Hence, we have $\sigma(\mathcal{E}) = \mathcal{D}$ by the monotone class theorem, and

$$\mathcal{F}_\infty = \sigma(\mathcal{E}) = \mathcal{D} \subset \mathcal{F}_\infty \ , \tag{1.31}$$

so (1.27) indeed holds for all $A \in \mathcal{F}_\infty$. □

Theorem 1.8 above shows that uniform integrability is not only necessary but also sufficient to establish (1.8). This is an example of a more general fact regarding the convergence of random variables (whether or not they form martingales): The diagram (1.14) shows that convergence in probability does not imply $\mathbf{L}_1$ convergence. To obtain the latter from the former, we need extra conditions. One possible condition is to assume that the random sequence or net in question (1) converges almost surely and (2) is dominated, i.e. bounded in absolute value by some random variable ($|X_s| \leq Y$ for some Y and all s). This fairly strong assumption can be weakened to: The net (1) converges in probability and (2) is uniformly integrable. We can hence augment diagram (1.8) as follows:



(1.32)

For martingales in particular, we can use the martingale property (plus mild regularity conditions) to ensure almost sure convergence to a limit (e.g. Theorem 27.1 in Probability I). Almost sure convergence implies convergence in probability; by also imposing an uniform integrability assumption, we then obtain the desired $\mathbf{L}_1$ convergence.

1.5. Application: The 0-1 law of Kolmogorov

Recall the 0-1 law from Probability I: If $X_1, X_2, \ldots$ is an infinite sequence of independent random variables, and a measurable event A does not depend on the value of the initial sequence $X_1, \ldots, X_n$ for any n, then A occurs with probability either 0 or 1. The prototypical example is the convergence of a series: If the random variables take values in, say, $\mathbb{R}^n$, the event

$$\left\{ \textstyle\sum_i X_i \text{ converges } \right\} \tag{1.33}$$

does not depend on the first n elements of the sequence for any finite n. Hence, the theorem states that the series either converges almost surely, or almost surely does not converge. However, the limit value it converges to *does* depend on every X_i. Thus, the theorem may tell us that the sequence converges, but not what it converges to.

In formal terms, the set of events which do not depend on values of the first n variables is the σ-algebra $\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, \ldots)$. The set of all events which do not depend on $(X_1, \ldots, X_n)$ for *any* n is $\mathcal{T} := \cap_n \mathcal{T}_n$, which is again a σ-algebra, and is called the **tail σ-algebra**, or simply the **tail field**.

1.10 Kolmogorov's 0-1 law. *Let $X_1, X_2, \ldots$ be independent random variables, and let A be an event such that, for every $n \in \mathbb{N}$, A is independent of the outcomes of $X_1, \ldots, X_n$. Then $\mathbb{P}(A)$ is 0 or 1. That is, if $A \in \mathcal{T}$, then $\mathbb{P}(A) \in \{0, 1\}$.* ◁

This result was also proven as Theorem 10.6 in [Probability I]. It can be proven very concisely using martingales. Arguably the key insight underlying the theorem is that every set in $\mathcal{T}$ is $\sigma(X_1, X_2, \ldots)$-measurable. The martingale proof shows this very nicely:

Proof. For any measurable set A, we have $\mathbb{P}(A) = \mathbb{E}[\mathbb{I}_A]$. Suppose $A \in \mathcal{T}$. Since A is independent of $X_1, \ldots, X_n$, we have

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{I}_A] =_{\text{a.s.}} \mathbb{E}[\mathbb{I}_A | X_{1:n}] \qquad \text{for all } n \in \mathbb{N} \,. \tag{1.34}$$

We use martingales because they let us determine the conditional expectation $\mathbb{E}[\mathbb{I}_A | X_{1:\infty}]$ given the entire sequence: The sequence $(\mathbb{E}[\mathbb{I}_A | X_{1:n}], \sigma(X_{1:n}))_n$ is an uniformly integrable martingale by Theorem 1.7, and by Theorem 1.8 converges almost surely to an a.s.-unique limit. Since

$$\mathbb{E}\big[\mathbb{E}[\mathbb{I}_A | X_{1:\infty}] \big| X_{1:n}\big] =_{\text{a.s.}} \mathbb{E}[\mathbb{I}_A | X_{1:n}] \,, \tag{1.35}$$

(1.22) shows that the limit is $\mathbb{E}[\mathbb{I}_A | X_{1:\infty}]$, and hence

$$\mathbb{E}[\mathbb{I}_A | X_{1:\infty}] =_{\text{a.s.}} \lim_n \mathbb{E}[\mathbb{I}_A | X_{1:n}] =_{\text{a.s.}} \lim_n \mathbb{P}(A) = \mathbb{P}(A) \,. \tag{1.36}$$

Since $\mathcal{T} \subset \sigma(X_{1:\infty})$, the function $\mathbb{I}_A$ is $\sigma(X_{1:\infty})$-measurable, and hence

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{I}_A | X_{1:\infty}] = \mathbb{I}_A \in \{0, 1\} \qquad \text{almost surely.} \tag{1.37}$$

□

1.6. Continuous-time martingales

The so-called *continuous-time case* is the special case where the index set is chosen as $\mathbb{T} := \mathbb{R}_+$, so we can think of the martingale (X_t) as a time series, started at time $t = 0$. For any fixed $\omega \in \Omega$, we can the interpret the realization $(X_t(\omega))$ of the martingale as a random function $t \mapsto X_t(\omega)$. Each realization of this function is called a **sample path**. We can then ask whether this function is continuous, or at least piece-wise continuous—this is one of the aspects which distinguish the continuous-time case from discrete time. Rather than continuity, we will use a notion of piece-wise continuity:

1.11 Reminder [rcll functions]. Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a function. Recall that f is continuous at x if, for every sequence (x_n) with $x_n \to x$, we have $\lim_n f(x_n) = f(x)$. We can split this condition into two parts: For every sequence $x_n \to x$, (1) $\lim_n f(x_n)$ exists and (2) equals $f(x)$.

Now suppose that, instead of all sequence with limit x, we consider only those which **converge from above** to x, i.e. sequences with $x_n \to x$ and $x_n \geq x$ for all n; we denote convergence from above as $x_n \searrow x$. If condition (1) is satisfied for all sequences which converge to x from above, i.e. if $\lim_n f(x_n)$ exists for all $x_n \searrow x$, we say that f has a **right-hand limit** at x. If (2) is also satisfied, i.e. if

$\lim_n f(x_n) = f(x)$ for all such sequence, we call f **right-continuous** at x. **Left-hand limits** and **left-continuity** are defined similarly, considering only sequence which converge to x from below.

We say that a function on $\mathbb{R}_+$ is **right-continuous with left-hand limits**, or **rcll** for short, if it is right-continuous at every point in $[0, \infty)$ and has a left-hand limit at every point in $(0, \infty]$. ◁

Intuitively, rcll functions are functions that are piece-wise continuous functions which jump at an at most countable number of points (otherwise, they would not have right- and left-hand limits). If the function jumps at x, the function value $f(x)$ is part of the "right-hand branch" of the function (which is condition (2) in right-continuity).

Filtrations for continuous-time martingales. In this section, we look for conditions which ensure a martingale has rcll sample paths. To formulate such conditions, we have to impose additional requirements on filtrations. One requirement is that filtrations contain all negligible sets.

1.12 Reminder [Negligible sets and completions]. If $(\Omega, \mathcal{A})$ is a measurable space, the σ-algebra $\mathcal{A}$ does not usually contain all subsets of Ω. For a given probability measure, there may hence be a non-measurable set B which is contained in a $\mathbb{P}$-null set $A \in \mathcal{A}$. Sets which are contained in null sets are called **negligible sets**. (In other words, a null set is a negligible set which is also measurable.)

Even if a negligible set is not technically measurable, we might still argue that it is morally measurable, since we know what its measure would be if it happened to be in $\mathcal{A}$: $B \subset A$ and $\mathbb{P}(A) = 0$ implies the measure would have to be zero. With this rationale, we can simply regard all negligible sets as null sets, and add them to the σ-algebra. It is easy to check that the resulting set system is again a σ-algebra. It is called the $\mathbb{P}$**-completion** of $\mathcal{A}$, and denoted $\overline{\mathcal{A}}^{\mathbb{P}}$. Note that we cannot define a completion before specifying a measure on $(\Omega, \mathcal{A})$. ◁

To work with rcll sample paths, we need a similar requirement for filtrations:

1.13 Definition. A filtration $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$ in a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is called **complete** if it contains all $\mathbb{P}$-negligible sets, i.e. if $\mathcal{F}_t = \overline{\mathcal{F}}_t^{\mathbb{P}}$ for all t. ◁

A second requirement is that the filtration itself is "smooth": Suppose for some index $s \in \mathbb{R}_+$, all filtrations $\mathcal{F}_t$ with $t > s$ suddenly contain much more information than $\mathcal{F}_s$—roughly speaking, the amount of information available "jumps up" at s. Such cases are excluded by the following definition:

1.14 Definition. A filtration $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$ is called **right-continuous** if

$$\mathcal{F}_t = \cap_{s>t}\mathcal{F}_s \qquad \text{for all } t \in \mathbb{R}_+ \,. \tag{1.38}$$

◁

Martingales with rcll paths. Theorem 1.16 below shows under which conditions a martingale, or even a submartingale, defined with respect to a complete and right-continuous filtration has rcll sample paths. To proof that result, we will need the following lemma. I will cheat a little and proof the theorem but not the lemma.

1.15 Lemma. *Let $\mathcal{F}$ be a filtration indexed by $\mathbb{R}_+$, and let $(X_t, \mathcal{F}_t)_{t\in\mathbb{R}_+}$ be a submartingale. Then there is a null set N such that the following holds: For all $t \in \mathbb{R}_+$,* *there is a real-valued random variable X_{t+} such that*

$$X_{t+}(\omega) = \lim_{s \in \mathbb{Q}_+, s \searrow t} X_s(\omega) \qquad \text{whenever } \omega \notin N \,. \tag{1.39}$$

Modify X_{t+} on the null set N by defining $X_{t+}(\omega) := 0$ for $\omega \in N$. If $\mathcal{F}$ is complete and right-continuous, then X_{t+} is integrable and

$$X_t \leq X_{t+} \qquad \text{almost surely} \tag{1.40}$$

for each $t \in \mathbb{R}_+$, with equality almost surely if and only if the function $\mu(t) := \mathbb{E}[X_t]$ is right-continuous at t. ◁

Two remarks on Lemma 1.15:

(1) Note that the assertion in (1.39) is stronger than just almost sure convergence for each t: The latter would mean that, for each t, there is a null set N_t outside of which (1.39) holds. Since the index set is uncountable, $\cup_t N_t$ would *not* be guaranteed to be a null set. The lemma shows, however, that there is a single null set N outside of which (1.39) holds for all t.

(2) The lemma holds for a general submartingale. Recall that, if (X_t) is a martingale, then all X_t have identical mean $\mathbb{E}[X_t] = \mu_t = \mu$, so the function $t \mapsto \mathbb{E}[X_t]$ is constant and hence rcll. By the last assertion in the lemma, equality in (1.40) therefore holds automatically if (X_t) is a martingale.

1.16 Theorem [Submartingales with rcll sample paths]. *Let $(X_t, \mathcal{F}_t)_{t \in \mathbb{R}_+}$ be a submartingale, where $\mathcal{F}$ is right-continuous and complete, and the function $\mu(t) := \mathbb{E}[X_t]$ is right-continuous. Then there exists a submartingale $(Y_t, \mathcal{F}_t)_{t \in \mathbb{R}_+}$ satisfying $X_t =_{\text{a.s.}} Y_t$ for all t whose paths $t \mapsto Y_t(\omega)$ are rcll almost surely.* ◁

Note that the result does not quite state that (X_t) is almost surely rcll, but rather that there is a martingale Y which is equivalent to X—in the sense that $X_t =_{\text{a.s.}} Y_t$, i.e. we are not able to distinguish Y from X by probabilistic means—and this equivalent martingale is rcll almost surely. The process Y is called a **version** or **modification** of X (since we modify the measurable function X on a null set to obtain Y; a precise definition will follow in a later chapter). Theorem 1.16 is our first example of a regularity result for a stochastic process, and we will see later on, in the chapter on stochastic processes, that most regularity results are stated in terms of the existence of almost surely regular versions.

Proof. Since (X_t) is a submartingale, Lemma 1.15 guarantees that the random variable X_{t+} defined in (1.39) exists for each t. Define $Y_t := X_{t+}$. Then the paths of Y_t are rcll by construction. Since $\mu(t)$ is right-continuous by hypothesis, Lemma 1.15 shows that $Y_t = X_t$ almost surely (equality holds in (1.40)). The only thing left to show is hence that $(Y_t, \mathcal{F}_t)$ is a submartingale, i.e. that $\int_A Y_s d\mathbb{P} \leq \int_A Y_t d\mathbb{P}$ for all $A \in \mathcal{F}_s$ and all $s < t$.

Let $s < t$. Then there are sequence $s_1 > s_2 > \ldots$ and $t_1 > t_2 > \ldots$ in $\mathbb{Q}_+$ such that $s_n \searrow s$ and $t_n \searrow t$. By Lemma 1.15,

$$X_{s+} =_{\text{a.s.}} \lim_n X_{s_n} \qquad \text{and} \qquad X_{t+} =_{\text{a.s.}} \lim_n X_{t_n} \,. \tag{1.41}$$

Random variables that are almost surely equal integrate identically over measurable sets; hence in particular for all $A \in \mathcal{F}_s$

$$\int_A X_{s+} d\mathbb{P} = \lim_n \int_A X_{s_n} d\mathbb{P} \qquad \text{and} \qquad \int_A X_{t+} d\mathbb{P} = \lim_n \int_A X_{t_n} d\mathbb{P} \,. \tag{1.42}$$

Since $s < t$, we can always choose the sequences such that $s_n < t_n$ for all n, which by the submartingale property implies

$$\int_A X_{s+} d\mathbb{P} = \lim_n \int_A X_{s_n} d\mathbb{P} \leq \lim_n \int_A X_{t_n} d\mathbb{P} = \int_A X_{t+} d\mathbb{P} \ , \tag{1.43}$$

and so $(X_{t+}, \mathcal{F}_t)$ is a submartingale. □

Note well: We cannot deduce directly from the definition $Y_t := X_{t+}$ and almost sure equality in (1.40) that (Y_t) is a submartingale, since (1.40) holds only pointwise—there is in general a separate null set N_t of exceptions for every $t \in \mathbb{R}_+$, and the union of this uncountable number of null sets need not be null. In the proof, we negotiate this problem by choosing a dense countable subset of $\mathbb{R}_+$, in this case $\mathbb{Q}_+$. Since the set is countable, we *can* conclude that there is a single null $N_{\mathbb{Q}_+}$ such that (1.40) holds simultanuously for all $t \in \mathbb{Q}_+$ whenever $\omega \notin N_{\mathbb{Q}_+}$.

1.7. Tail bounds for martingales

A **tail bound** for a (real-valued) random variable X is an inequality of the form

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \ \leq \ f(\varepsilon) \ . \tag{1.44}$$

Almost all distributions—at least on unbounded sample spaces—concentrate most of their probability mass in some region around the mean, even if they are not unimodal. If we move sufficiently far away from the mean, the distribution decays. A tail bound quantifies how rapidly it does so; of interest is not so much how far out we have to move before the probability decays, but rather the shape of f in the region far away from the mean (in the "tails"). In particular, if f is of the form $f(\varepsilon) \leq ce^{-g(\varepsilon)}$ for some positive polynomial g, the distribution is said to exhibit **exponential decay**. If $f(\varepsilon) \sim \varepsilon^{-\alpha}$ for some $\alpha > 0$, it is called **heavy-tailed**.

1.17 Example [Hoeffding's inequality]. One of the most widely used tail bounds is the **Hoeffding bound**: Suppose $X_1, X_2, \ldots$ are independent, real-valued random variables (which need not be identically distributed), and each is bounded in the sense that $X_n \in [a_i, b_i]$ almost surely for some constants $a_i < b_i$. Then the empirical average $S_n = \frac{1}{n}(X_1 + \ldots + X_n)$ has tails bounded as

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq \varepsilon) \leq 2 \exp\Bigl(-\frac{-2n^2\varepsilon^2}{\sum_i (b_i - a_i)^2}\Bigr) \ . \tag{1.45}$$

◁

If the X_n are dependent, we cannot generally hope for a bound of this form to hold. Remarkably, it still does if the X_n form a martingale. In this case, each X_n is required to be *conditionally* bounded given the previous value X_{n-1}, i.e. X_n must lie in the interval $[X_n - c_{n+1}, X_n + c_{n+1}]$.

1.18 Azuma-Hoeffding Inequality. *Let $(X_n, \mathcal{F}_n)_{i \in \mathbb{N}}$ be a martingale. Require that there exists a sequence of non-negative constants $c_n \geq 0$ such that*

$$|X_{n+1} - X_n| \leq c_{n+1} \qquad \textit{almost surely} \tag{1.46}$$

and $|X_1 - \mu| \leq c_1$ a.s. Then for all $\varepsilon > 0$

$$\mathbb{P}(|X_n - \mu| \geq \varepsilon) \leq 2 \exp\Bigl(-\frac{\varepsilon^2}{2\sum_{j=0}^n c_n^2}\Bigr) \ . \tag{1.47}$$

◁

I can hardly stress enough how useful this result can be in applications: On the one hand, the only requirement on the martingales is the boundedness of increments, which makes the bound very widely applicable. On the other hand, when it is applicable, it is often also very sharp, i.e. we will not get a much better bound using more complicated methods. One application that has become very important over the past twenty or so years is the analysis of randomized algorithms:

1.19 Example [Method of bounded differences]. Suppose an iterative randomized algorithm computes some real-valued quantity X; since the algorithm is randomized, X is a random variable. At each iteration, the algorithm computes a candidate quantity X_n (which we usually hope to be a successively better approximation of some "true" value as n increases). If we can show that (1) the intermediate results X_n form a martingale and (2) the change of X_n from one step to the next is bounded, we can apply Theorem 1.18 to bound X. ◁

1.8. Application: The Pólya urn

Recall that an urn is a stochastic process defined by starting with a certain number of white and black balls, and repeatedly drawing a ball uniformly at random. You will be familiar with *sampling with replacement* (an urn in which the ball is replaced after having been drawn) and *sampling without replacement* (the ball is removed).

More generally, an urn is a process where, each time we draw a ball, we may or may not replace it, and may or may not add additional balls to the urn. It can be parametrized as

$$\begin{pmatrix} w & a \\ d & b \end{pmatrix} \qquad \text{where} \qquad \begin{aligned} w &= \text{ \# initial white balls} \\ b &= \text{ \# initial black balls} \end{aligned} \;. \tag{1.48}$$

Each time a ball is drawn, we replace it by a balls of the same color and d balls of the opposite color. Important examples are:

$a = 0$	$d = 0$	Sampling without replacement
$a = 1$	$d = 0$	Sampling with replacement
$a > 1$	$d = 0$	Pólya urn
$a = 0$	$d = 1$	Ehrenfest urn (or Ehrenfest heat transfer model)

In particular, a **Pólya urn** with parameters (w_0, b_0, a) is a stochastic process defined by an urn initially containing w_0 white and b_0 black balls. At each step, draw a ball from the urn at random; then replace the ball, and add an additional a balls of the same color. We define X_n as the fraction of white balls after n steps,

$$X_n = \frac{\text{\# white balls after } n \text{ steps}}{(\text{\# white balls } + \text{ \# black balls) after } n \text{ steps}} \;. \tag{1.49}$$

1.20 Proposition. *The proportions X_n converge almost surely: There exists a random variable X_∞ such that $\lim_{n\to\infty} X_n(\omega) = X_\infty(\omega)$ almost surely.* ◁

Before we prove this existence result, I want to complement it with a result on the form of the limit, which we will not proof (since it does not involve a martingale argument):

1.21 Fact. The limiting distribution of proportions, i.e. the law of X_∞, is the beta distribution on $[0, 1]$ with density

$$p(x) = B(w/(a-1), b/(a-1))x^{w/(a-1)-1}(1-x)^{b/(a-1)-1} \;. \tag{1.50}$$

where B denotes the beta function. $\triangleleft$

Before we proof the existence of the limit, reconsider for a moment what Proposition 1.20 tells us:

- We know from basic calculus that a sequence need not converge—the proportions could fluctuate perpetually. Proposition 1.20 shows that this is not the case here: Even though the sequence is generated at random by the urn, it *always* converges to a limit. Roughly speaking, if we would run the process for an infinite amount of time to obtain the proportions X_∞, and then restart it with those proportions, they would never change again (which of course can only be true since the urn has swollen to contain an infinite number of balls).
- On the other hand, the limit is random—if we start the process from the same initial values twice, we will typically obtain two distinct limiting proportions. (In fact, since the limiting distribution is continuous, they differ with probability 1.)

Proof of Proposition 1.20. We will show that (X_n) is a martingale, and then apply the martingale convergence theorem to verify existence of the limit. Let W_n and B_n respectively denote the number of white and black balls after n draws. The probability of observing a white ball in the $(n+1)$st draw is, conditionally on (W_n, B_n),

$$p_{n+1} = \frac{W_n}{W_n + B_n} \;. \tag{1.51}$$

In each step, the number of balls of the color that was drawn increases by $(a-1)$. Hence,

$$X_{n+1}|W_n, B_n = \begin{cases} \frac{W_n+(a-1)}{W_n+B_n+(a-1)} & \text{with probability } p_n \\ \frac{W_n}{W_n+B_n+(a-1)} & \text{with probability } 1-p_n \end{cases} \;. \tag{1.52}$$

The history of the process, up to step n, is given by the nth σ-algebra in the filtration $\mathcal{F}_n$. The conditional expectation of X_{n+1} given the history of the process is hence

$$\begin{aligned} \mathbb{E}[X_{n+1}|\mathcal{F}_n] =& \frac{W_n+(a-1)}{W_n+B_n+(a-1)}p_{n+1} + \frac{W_n}{W_n+B_n+(a-1)}(1-p_{n+1}) \\ =& \ldots = \frac{W_n}{W_n+B_n} = X_n \;. \end{aligned} \tag{1.53}$$

Since X_n is also clearly integrable, it is hence a martingale. We now use the martingale convergence theorem, in the form of Corollary 27.1 in [Probability I]—recall: If (X_n) is a non-negative supermartingale, then $\lim_{n\to\infty} X_n$ exists almost surely—which completes the proof. □

1.22 Remark [Preferential attachment networks]. The Pólya urn may seem primitive, but it has many important applications. One example are random graphs used as models for certain social networks: A **preferential attachment graph** is generated as follows. Fix an integer $m \geq 1$. Start with a graph consisting of a single vertex. At each step, insert a new vertex, and connect it to m randomly selected vertices in the current graph. These vertices are selected by **degree-biased sampling**, i.e. each vertex is selected with probability proportional to the number of edges currently attached to it. You will notice that (1) the placement of the next edge depends only on the vertex degrees (not on which vertex is connected to which), and (2) the model is basically a Pólya urn (where each vertex represents a color, and the degrees are the number of balls per color). It is hence not surprising that most proofs on asymptotic properties of this model involve martingales. This, in turn, is one of the reasons why this model is as well-studied as it is in the applied probability literature—the applicability of martingales makes it tractable, so we study it because we can. ◁

1.9. Application: The Radon-Nikodym theorem

Let P be probability measure on a measurable space $(\mathcal{X}, \mathcal{A})$, and let μ be a finite measure on the same space (that is, $\mu(\mathcal{X}) < \infty$). Recall that a **density** of μ with respect to P is an integrable function $f : \mathcal{X} \to \mathbb{R}_{\geq 0}$ satisfying

$$\mu(dx) =_{\text{a.e.}} f(x)P(dx) \; . \tag{1.54}$$

When does a density exist for a given pair μ and P?

Equation (1.54) says that f transforms the set function P into μ by reweighting it point-wise. Since it we cannot transform 0 into a positive number by multiplication with any value, this clearly requires that μ vanishes wherever P vanishes, that is,

$$P(A) = 0 \qquad \Rightarrow \qquad \mu(A) = 0 \tag{1.55}$$

for all measurable sets A in $\mathcal{X}$. Recall that μ is called **absolutely continuous** with respect to P if μ and P satisfy (1.55)—in symbols, $\mu \ll P$. The term "absolute continuity" derives from the following:

1.23 Fact. $\nu \ll \mu$ holds if and only if

$$\text{for all } \varepsilon > 0 \text{ exists } \delta > 0 \text{ such that } \mu(A) \leq \delta \Rightarrow \nu(A) \leq \varepsilon \tag{1.56}$$

holds for all measurable sets A. ◁

That absolute continuity is a necessary condition for (1.54) to hold is obvious. Remarkably, it is also the only condition required:

1.24 Radon-Nikodym theorem (for probability measures). *Let P be a probability measure and μ a finite measure on a measurable space $\mathcal{X}$. Then μ has a density with respect to P if and only if $\mu \ll P$. Any two such densities differ only on a P null set.* ◁

Proof of the theorem. The idea of the proof is to subdivide the space $\mathcal{X}$ into a partition of n disjoint sets A_j, and define

$$Y_{(A_1,\ldots,A_n)}(x) := \sum_{j=1}^{n} f(A_j)\mathbb{I}_{A_j}(x) \qquad \text{where } f(A_j) := \begin{cases} \frac{\mu(A_j)}{P(A_j)} & P(A_j) > 0 \\ 0 & P(A_j) = 0 \end{cases} . \tag{1.57}$$

Think of Y as a "discretization" of the density f whose existence we wish to establish. Roughly speaking, we will make the partition finer and finer (by making the sets A_j smaller and increasing n), and obtain f as the limit of Y. Since Y is a measurable function on the space $\mathcal{X}$, which forms a probability space with P, we can regard the collection of Y we obtain for different partitions as a martingale.

More formally, we construct a directed index set $\mathbb{T}$ as follows: A **finite measurable partition** $H = (A_1, \ldots, A_n)$ of $\mathcal{X}$ is a subdivision of $\mathcal{X}$ into a finite number of disjoint measurable sets A_i whose union is $\mathcal{X}$. Let $\mathbb{T}$ be the set of all finite measurable partitions of $\mathcal{X}$. Now we have to define a partial order: We say that a partition $H_2 = (B_1, \dots, B_m)$ is a **refinement** of another partition $H = (A_1, \dots, A_n)$ if every set B_j in H_2 is a subset of some set A_i in H_1; in words, H_2 can be obtaine from H_1 by splitting sets in H_1 further, without changing any of the existing set boundaries in H_1. We then define a partial order on $\mathbb{T}$ as

$$H_1 \preceq H_2 \qquad \Leftrightarrow \qquad H_2 \text{ is a refinement of } H_1 \ . \tag{1.58}$$

Since each index $s \in \mathbb{T}$ is now a measurable partition, we can define $\mathcal{F}_s$ as the σ-algebra generated by the sets in s,

$$\mathcal{F}_s := \sigma(A_1, \dots, A_n) \qquad \text{if } s = (A_1, \dots, A_n) \ . \tag{1.59}$$

1.25 Lemma. $(Y_s, \mathcal{F}_s)_{s \in \mathbb{T}}$ *is a uniformly integrable martingale.* ◁

Proof. It is easy to check the martingale property; we will show uniform integrability. Let $\alpha > 0$ and choose some index $s = (A_1, \dots, A_n)$. (Recall the definition of uniform integrability in (1.15); we choose g as the constant function with value α.) Then

$$\begin{aligned}\int_{\{|Y_s| \geq \alpha\}} |Y_s(x)| P(dx) &\overset{Y_s \geq 0}{=} \int_{\{Y_s \geq \alpha\}} Y_s(x) P(dx) \\ &= \int_{\mathcal{X}} \sum_{j=1}^{n} \frac{\mu(A_j)}{P(A_i)} \mathbb{I}\{x \in A_i \text{ and } Y_s(x) \geq \alpha\} P(dx) \\ &= \mu\{Y_s \geq \alpha\} \ .\end{aligned} \tag{1.60}$$

Since Y_s is a positive random variable, Markov's inequality for Y_s reads

$$P\{Y_s \geq \alpha\} \leq \frac{1}{\alpha} \mathbb{E}[Y_s] = \frac{1}{\alpha} \mu(\mathcal{X}) \ . \tag{1.61}$$

Now we use (1.56): For a given $\varepsilon > 0$, choose some δ which satisfies (1.56), and set $\alpha > \frac{\mu(\mathcal{X})}{\delta}$. Then (1.61) implies $P\{Y_s \geq \alpha\} \leq \delta$, and hence

$$\int_{\{|Y_s| \geq \alpha\}} |Y_s(x)| P(dx) \overset{(1.60)}{=} \mu\{Y_s \geq \alpha\} \overset{(1.56)}{\leq} \varepsilon \ . \tag{1.62}$$

The choice of ε and δ is independent of the index s (since the rightmost term in (1.61) does not depend on s). Hence, $(Y_s, \mathcal{F}_s)$ is uniformly integrable. □

The proof of uniform integrability is the only real leg work in the proof of the Radon-Nikodym theorem. The rest is easy:

Proof of Theorem 1.24. Since $(Y_s, \mathcal{F}_s)$ is a uniformly integrable martingale, Theorem 1.8 shows that an integrable random variable Y_∞ with $\mathbb{E}[Y_\infty | \mathcal{F}_s] =_{\text{a.s.}} Y_s$ exists and is uniquely determined, up to almost sure equivalence. To verify that Y_∞ is a density, we have to show that $\mu(A) = \int Y_\infty(x) P(dx)$, and that Y_∞ is non-negative almost surely. The identity $\mathbb{E}[Y_\infty | \mathcal{F}_s] =_{\text{a.s.}} Y_s$ means

$$\int_A Y_\infty(x) P(dx) = \int_A Y_s(x) P(dx) \qquad \text{for all } A \in \mathcal{F}_s \ . \tag{1.63}$$

For each A, the index set $\mathbb{T}$ contains in particular the partition $s = (A, \bar{A})$ consisting only of A and its complement $\bar{A}$. For this s, the previous equation becomes

$$\begin{aligned}\int_A Y_\infty(x)P(dx) &= \int_A Y_s(x)P(dx) \\ &= \int_A \Big(\frac{\mu(A)}{P(A)}\mathbb{T}_A(x) + \frac{\mu(\bar{A})}{P(\bar{A})}\mathbb{T}_{\bar{A}}(x)\Big)P(dx) = \mu(A)\ .\end{aligned} \tag{1.64}$$

This also implies that $Y_\infty \geq 0$ almost everywhere—otherwise, there would be a non-null set A (i.e. $P(A) > 0$) on which Y_∞ takes only negative values, and by the previous equation, that would yield $\mu(A) < 0$. □

FYI: The general case. The existence of densities is of course not limited to the case where P is a probability measure, or even finite; I have stated the result in the form above in order to prove it using martingales (and because the case where P is not normalized is not particularly relevant in the following). Nonetheless, I should stress that Theorem 1.24 still holds in precisely this form if μ and P are both σ-finite measures:

1.26 Radon-Nikodym theorem. *Let μ and ν be σ-finite measures on a measurable space $(\mathcal{X}, \mathcal{A})$. Then there exists a measurable function $f : \Omega \to [0, \infty)$ with $\mu(A) = \int_A f d\nu$ for all $A \in \mathcal{A}$ if and only if $\mu \ll \nu$.* ◁

Indeed, there is a generalization beyond even the σ-finite case: ν need not be σ-finite, and μ need not even be a measure. I have found this result very useful at times, and since it is not widely known and almost impossible to find in textbooks, I state it here without proof (which you can read up in [5, 232E], if you feel so inclined):

1.27 Generalized Radon-Nikodym theorem. *Let ν be a measure on a measurable space $(\mathcal{X}, \mathcal{A})$, and let $\mu : \mathcal{A} \to \mathbb{R}_{\geq 0}$ be a finitely additive set function. Then there is a measurable function $f : \mathcal{X} \to \mathbb{R}_{\geq 0}$ satisfying $\mu(A) = \int_A f d\nu$ for all $A \in \mathcal{A}$ if and only if:*

(i) *μ is absolutely continuous with respect to ν.*

(ii) *For each $A \in \mathcal{A}$ with $\mu(A) > 0$, there exists a set $B \in \mathcal{A}$ such that $\nu(B) < \infty$ and $\mu(A \cap B) > 0$.*

If so, f is uniquely determined ν-a.e. ◁

CHAPTER 2

Measures on nice spaces

We will now start to talk about probability measures on very general spaces—the law of a stochastic process, for example, is usually a probability measure on an infinite-dimensional space (provided that we can even define in a straightforward way what a dimension is). Two problems we encounter in this case are the following:

(1) How do we define a σ-algebra on the space?

On the line, we can generate the Lebesgue σ-algebra using intervals, or hypercubes in $\mathbb{R}^d$. That does not work in infinite dimensions; roughly speaking, the volume of a hypercube with fixed edge length s is s^d, and if $d \to \infty$, this volume converges to 0, 1 or ∞ (depending on whether s is smaller, equal or larger than 1). The example illustrates that our finite-dimensional conceptions of volume do not really work in infinite dimensions. On more abstract spaces, there is not even a simple notion of dimensions that we could use as an exponent.

(2) Many properties of measures that hold automatically on $\mathbb{R}^d$ do not hold on arbitrary measurable spaces. How do we ensure those properties?

There is a common answer to both problems, which is to define a topological space with suitable properties, and to use the open sets in this space to generate the σ-algebra. The topological spaces which have emerged as the golden mean between generality and tractability for most purposes in analysis and probability are *Polish spaces.*

2.1. Topology review

A function between Euclidean spaces is continuous if $\lim_n f(x_n) = f(x)$ for every convergent sequence $x_n \to x$. This definition is requires a metric, since the definition of the limit involves a metric. Assuming a metric as given is a fairly strong condition, but it turns out that continuity can be formulated in much more general terms: A function between Euclidean spaces is continuous if and only if the preimage of every open set is open. Provided we have a definition of what an open set is, this statement does not involve a metric, so we can substitute it for the definition of continuity. In order to ensure that functions defined as "continuous" still have properties resembling those we are used to form continuous functions in Euclidean space, we need to make sure that the sets we call open have some of the properties that open sets in $\mathbb{R}^d$ have. The set of all open sets is called a *topology.* The properties we need to ensure are the following:

2.1 Definition. A **topology** τ on a set $\mathcal{X}$ is a system τ of subsets of $\mathcal{X}$ which satisfies:

(1) $\varnothing \in \tau$.

(2) τ is closed under countable unions.

(3) τ is closed under finite intersections.

The sets in τ are called **open sets**. The pair $\mathbf{X} := (\mathcal{X}, \tau)$ is called a **topological space**. A function $f : \mathbf{X} \to \mathbf{X}'$ between two topological spaces is **continuous** if

$$f^{-1}\tau(\mathbf{X}') \subset \tau(\mathbf{X}) \;, \tag{2.1}$$

that is if the preimage $f^{-1}A'$ of every open set A' in $\mathbf{X}'$ is an open set in $\mathbf{X}$. $\triangleleft$

The definition of a topology above is very general; we could, for example, just define three or four sets, fill in all unions and intersections, and call these the open sets. To obtain a space with useful properties, we need to make sure it contains a sufficient number of open sets. Usually, a minimal requirement is that any two points can be separated by an open set:

2.2 Definition. A **Hausdorff space** is a topological space in every pair of points is separated by disjoint open neighborhoods: For any $x, x' \in \mathbf{X}$, there exist open sets $A, A' \in \tau$ such that $x \in A$, $x' \in A'$ and $A \cap A' = \varnothing$. $\triangleleft$

There is a concept analogous to the generator of a σ-algebra: If $\mathcal{G}$ is a system of sets, the topology **generated** by $\mathcal{G}$ is the smallest topology which contains all sets in $\mathcal{G}$, and denoted $\tau(\mathcal{G})$. Every set in $\tau(\mathcal{G})$ can be represented as a (possibly uncountable) union of finite intersections of sets in $\mathcal{G}$ (except possibly $\varnothing$ and $\mathcal{X}$). In this sense, topologies are simpler than σ-algebras—a set in a σ-algebra need not have such an explicit representation in terms of sets in a generator. Because of this fact, we distinguish two particular types of generators:

2.3 Definition. If every set in τ can be represented as a union of sets in $\mathcal{G}$, then $\mathcal{G}$ is called a **base** of τ. Any generator of τ which contains $\varnothing$ and $\mathcal{X}$ is called a **subbase** of τ. (Note that this means that every set in τ is a union of finite intersections of set in $\mathcal{G}$.) $\triangleleft$

We have not yet discussed how we can actually define a topology on a given space. There are two topologies everybody has encountered, at least implicitly:

- The standard topology on $\mathbb{R}^d$. This is the topology generated by all open balls in $\mathbb{R}^d$ (the open intervals in case of $\mathbb{R}$). Under this topology, the topological definitions of open sets, closed sets, continuous functions etc. coincide with the Euclidean definitions used in every calculus class.
- Finite or countable sets fit into the picture by endowing them with the **discrete topology**, which is the topology generated by all subsets. That is, every set is open, every set is closed, and every function is continuous.

To define more general topological spaces, there are two important ways to endow a given set $\mathcal{X}$ with a topology:

(1) By defining a notion of convergence of sequences or nets in $\mathcal{X}$. If we know which sequences converge, we know which sets are closed, and hence (by taking complements) which sets are open. The most important way to define convergence is to define a metric d on $\mathcal{X}$. In this case, the resulting topology is called a **metric topology**.

(2) By defining a family $\mathcal{F}$ of (usually real-valued) functions on $\mathbf{X}$, and choosing the smallest topology which makes all $f \in \mathcal{F}$ continuous. Such a topology is called a **weak topology**.

The standard topology on Euclidean space is, of course, the metric topology defined by the Euclidean distance; it is also the weak topology generated by the set of all continuous functions (defined in the ε-δ-sense). Clearly, every topology is the weak topology generated by its continuous functions. The discrete topology is a metric topology defined by the metric

$$d(x,y) := \mathbb{I}\{x \neq y\} \ . \tag{2.2}$$

There are two important types of topologies that are derived from given topologies: The product and the relative topology. Consider two topological spaces $\mathbf{X} = (\mathcal{X}, \tau(\mathbf{X}))$ and $\mathbf{Y} = (\mathcal{Y}, \tau(\mathbf{Y}))$. Is there a natural way to equip the product space $\mathbf{X} \times \mathbf{Y}$ with a topology that preserves the topological properties of $\mathbf{X}$ and $\mathbf{Y}$? By "preserving the topologies", we mean that if we project from the product $\mathbf{X} \times \mathbf{Y}$ back to, say, $\mathbf{X}$, we should recover the original topology $\tau(\mathbf{X})$. To formalize this idea, we formalize the projection as a mapping, and then use the weak topology generated by these maps.

2.4 Definition. Let $\mathbf{X}_t = (\mathcal{X}_t, \tau_t)$ be topological spaces, for all t in a (possibly uncountable) index set $\mathbb{T}$. For each $t \in \mathbb{T}$, define the **projection map**

$$\begin{aligned} \mathrm{pr}_{\mathbf{X}_t} : \prod_{s \in \mathbb{T}} \mathbf{X}_s &\to \mathbf{X}_t \\ (x_s)_{s \in \mathbb{T}} &\mapsto x_t \end{aligned} \tag{2.3}$$

be topological spaces. The topology on the product space $\prod \mathbf{X}_s$ generated by the family $\{\mathrm{pr}_{\mathbf{X}_t}, t \in \mathbb{T}\}$ is called the **product topology**. ◁

If the product consists of finitely many factors, the product topology is generated by all Cartesian products $A \times B$ of open sets A in $\mathbf{X}$ and B in $\mathbf{Y}$. If $\mathbb{T}$ is countably infinite or even uncountable, the generator also consists of products $\times_s A_s$, where A_s is open in $\mathbf{X}_s$, but A_s equals $\mathbf{X}_s$ for all but finitely many s.

The second important case is when we consider a subset $\mathcal{Y}$ of a topological space $\mathbf{X}$. Such a subset "inherits" a topology from $\mathbf{X}$, namely the restriction of all open sets in $\mathbf{X}$ to $\mathcal{Y}$:

$$\tau \cap \mathcal{Y} := \{A \cap \mathcal{Y} | A \in \tau\} \tag{2.4}$$

Once again, this can be elegantly formulated as a weak topology:

2.5 Definition. Let $\mathcal{Y}$ be a subset of a $\mathbf{X} = (\mathcal{X}, \tau)$, and let $\mathcal{I}_{\mathcal{Y}} : \mathcal{Y} \hookrightarrow \mathbf{X}$ denote the **canonical inclusion map** (i.e. the map which is defined as $x \mapsto x$ for $x \in \mathcal{Y}$ and undefined outside $\mathcal{Y}$). The weak topology on $\mathcal{Y}$ generated by $\mathcal{I}_{\mathcal{Y}}$ is called the **relative topology** (or **subspace topology**, or **trace topology**) on $\mathcal{Y}$. ◁

You will find it easy to see that, given a product topology on e.g. $\mathbf{X} \times \mathbf{Y}$, the original topology on $\mathbf{X}$ coincides with its relative topology under the product topology.

2.6 Definition. The σ-algebra $\mathcal{B}(\mathbf{X}) := \sigma(\tau)$ generated by a the open sets is called the **Borel σ-algebra** of the topological space $\mathbf{X}$, and its elements are called **Borel sets**. ◁

When we refer to a measure on a topological space $\mathbf{X}$ (without further qualification), we always mean a measure defined on the Borel σ-algebra of $\mathbf{X}$.

It is worth noting that the Borel σ-algebra is in general *much, much larger* than the topology of $\mathbf{X}$: Recall that, if $\mathcal{G}$ is a generator of a topology, then every open set is a countable union of finite intersections of sets in $\mathcal{G}$. As a generator of the topology, $\mathcal{G}$ is also a generator of the Borel σ-algebra, but there is no similarly explicit representation of arbitrary Borel sets. This is so because σ-algebras are closed under both countable unions and countable intersections, and the two do not commute. Hence, to represent an arbitrary Borel set, we need countable unions of countable intersections of countable unions of..., which leads to a structure within the Borel σ-algebra known as the "Borel hierarchy". In particular, recall that there is no such thing as a countably infinite σ-algebra; any σ-algebra is either finite or uncountable.

2.7 Lemma. *If two measures defined on the Borel σ-algebra of $\mathbf{X}$ coincide on all open sets, or if they coincide on all closed sets, then they are identical.* $\triangleleft$

Proof. Both the open and the closed sets form generators of $\mathcal{B}(\mathbf{X})$ that are closed under finite intersections, and hence completely determine measures on $\mathcal{B}(\mathbf{X})$ (cf. [Probability I, Corollary 6.1]). $\square$

2.2. Metric and metrizable spaces

The minimal condition we need to obtain one of the "nice" spaces in the title of this chapter is metrizability. Metrizability has many important consequences for analysis, and in particular two fundamental consequences for probability: On metrizable spaces, all probability measures have an important property called *regularity*, and weak convergence of distributions is well-defined on such spaces.

2.8 Definition. A function $d : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ is called a **metric** on $\mathcal{X}$ if:

(1) It is positive definite: $d(x, y) = 0$ if and only if $x = y$.
(2) It is symmetric: $d(x, y) = d(y, x)$ for all $x, y \in \mathcal{X}$.
(3) It satisfies the triangle inequality: $d(x, y) + d(y, z) \leq d(x, z)$ for all $x, y, z \in \mathcal{X}$.

$\triangleleft$

Recall that a net $(x_t)_{t\in\mathbb{T}}$ converges to a limit x with respect to a given metric d if the net $d(x_t, x)$ converges to 0 in $\mathbb{R}$. Also recall that a set A is called closed with respect to d if the limit of every d-convergent sequence of points in A is also in A. The topology induced by d is the set system

$$\tau := \{A \subset \mathcal{X} \,|\, \overline{A} \text{ closed with respect to } d\} \ . \tag{2.5}$$

We call a metric **compatible** with a given topology τ if d induces τ.

2.9 Exercise. Let $\mathcal{X}$ be a set and d a metric on $\mathcal{X}$. Show that the set system τ defined in (2.5) is a topology. $\triangleleft$

2.10 Definition. A topological space $\mathbf{X} = (\mathcal{X}, \tau)$ is called **metrizable** if there exists a metric on $\mathcal{X}$ which induces τ. For a specific compatible metric d, the pair $(\mathbf{X}, d)$ is called a **metric space**. $\triangleleft$

The distinction between metrizable and metric spaces is more useful than it may seem at first glance: Simply the fact that a topology can be generated by a metric implies a whole range of nice topological properties of $\mathbf{X}$, but these properties *do not depend on the metric*. We hence refer to $\mathbf{X}$ as metrizable if only such metric-independent properties are concerned. On the other hand, two metrics that metrize the same topology can have rather different properties. For two specific compatible metrics d_1 and d_2, we hence regard $(\mathbf{X}, d_1)$ and $(\mathbf{X}, d_2)$ as two distinct metric spaces.

2.11 Remark. Mathematical statistics provides an important example of where this distinction matters: If two metrics metrize the same topology, then sequences converge in one if and only if they converge in the other; but two convergent sequences may converge at very different rates. This is of no consequence in parametric statistics, where the parameter space can always be regarded as a subset of Euclidean space, and all metrics that metrize the Euclidean topology yield identical rates. In nonparametric statistics, however, parameter spaces are infinite-dimensional, and the properties of metrics can differ substantially. Given a topology on such a parameter space, any two compatible metrics yield identical notions of consistency (since consistency depends only on whether or not an estimator converges), but may yield very different convergence rates. $\triangleleft$

2.12 Lemma. *If* $\mathbf{X}$ *is metrizable, it is a Hausdorff space.* $\triangleleft$

Proof. Homework. □

Before we move and do probability theory again, we should mention a special type of set that we will use repeatedly in the following: A set of the form

$$B_r(x) := \{y \in \mathbf{X} | d(x,y) < r\} \tag{2.6}$$

is called an **open ball** of radius r, or an **open** d**-ball** if we wish to emphasize the metric.

2.13 Lemma. *In a metric space* $(\mathbf{X}, d)$*, a set* A *is open if and only if, for every* $x \in A$ *there is some* $\varepsilon > 0$ *such that* $B_\varepsilon(x) \subset A$*. Hence, every open set is the union of all open balls it contains.* $\triangleleft$

Proof. A closed set F is precisely the set of all limits of sequences in F. If A is open, its complement $\overline{A}$ is closed. Hence, $x \in A$ if and only if x is not the limit of any sequence in $\overline{A}$. By the definition of d-convergence, this means that $\varepsilon := \inf_{y \in \overline{A}} d(x,y) > 0$. Thus, $B_\varepsilon(x) \subset A$. The converse holds by definition, since arbitrary unions of open sets are open. □

2.3. Regularity of measures

We have not so far established any specific link between a measure on the Borel sets of a topological space and the underlying topology. We know of course, by Lemma 2.7, that a measure on the Borel sets is uniquely determined by its values on the open sets—but the same is true for any measure and any generator of the underlying σ-algebra, provided the generator is closed under finite intersections [Probability I, Corollary 6.1]. We have not yet used the open sets for anything we could not do just as well with any other generator.

However, Lemma 2.7 is not constructive: The open sets determine a measure μ on $\mathcal{B}(\mathbf{X})$ *abstractly*, in the sense that any other measure ν with identical values on all open sets must be identical. The next definition makes determination by open sets *constructive*, in the sense that we can derive the value of a probability measure P on any set to arbitrary precision from its values on open sets, or similarly, by its values on closed sets.

2.14 Definition. A measure μ on a topological space $\mathbf{X}$ is called:

(1) **Inner regular** if, for every Borel set A,

$$\mu(A) = \sup\{\mu(F) | F \subset A \text{ and } F \text{ closed}\} \ . \tag{2.7}$$

(2) **Outer regular** if, for every Borel set A,

$$\mu(A) = \inf\{\mu(G) | A \subset G \text{ and } G \text{ open}\} . \tag{2.8}$$

(3) **Regular** if it is inner and outer regular.

◁

Note that we can alternatively formulate these definitions as follows: P is inner regular if, for any $A \in \mathcal{B}(\mathbf{X})$,

$$\forall \varepsilon > 0 \quad \exists \text{ closed } F \subset A : \qquad \mu(A \setminus F) < \varepsilon . \tag{2.9}$$

Similarly, outer regularity requires

$$\forall \varepsilon > 0 \quad \exists G \in \tau, A \subset G : \qquad \mu(G \setminus A) < \varepsilon . \tag{2.10}$$

Combining these two, we see that regularity means: For any $A \in \mathcal{B}(\mathbf{X})$ and any $\varepsilon > 0$, there exist a closed F and an open G such that

$$F \subset A \subset G \qquad \text{and} \qquad \mu(G \setminus F) < \varepsilon . \tag{2.11}$$

If we can establish that a probability measure P is regular, we have indeed established a much closer link between the topology and P than Lemma 2.7. One reason why metrizable spaces play such key role in probability theory is that metrizability suffices to ensure regularity:

2.15 Theorem. *Every probability measure on a metrizable space is regular.* ◁

The proof is based on the fact that the existence of a compatible metric lets us define the distance between points and sets. To do so, we define the **distance function** of a set A in a metric space $(\mathbf{X}, d)$ as

$$d(x, A) := \inf_{y \in A} d(x, y) . \tag{2.12}$$

2.16 Lemma. *For any subset A of a metric space $(\mathbf{X}, d)$, the function $x \mapsto d(x, A)$ is Lipschitz (with Lipschitz constant 1). In particular, it is continuous.* ◁

Proof. Exercise. (Hint: Use the triangle inequality.) □

The proof of Theorem 2.15 now proceeds in a manner similar to the monotone class arguments we have already encountered; however, we do not have to invoke the monotone class theorem, since in this case we can verify the properties of a σ-algebra directly.

Proof of Theorem 2.15. Let $\mathcal{C}$ be the set of all measurable sets A which satisfy (2.11). If F is a closed set, the set

$$F^\delta := \{x \in \mathbf{X} | d(x, F) < \delta\} \tag{2.13}$$

is open. Hence, if the set A in (2.11) is closed, then A holds for $F := A$ and $G := F^\delta$ with sufficiently small δ. Hence, $\mathcal{C}$ contains all closed sets. Since the closed sets generate $\mathcal{B}(\mathbf{X})$, the set $\mathcal{C}$ is a generator of $\mathcal{B}(\mathbf{X})$. Hence, if we can show that $\mathcal{C}$ is even a σ-algebra, then $\mathcal{C} = \mathcal{B}(\mathbf{X})$, and the claim follows.

Clearly, $\mathcal{C}$ is closed under complements, since the complement of a set approximated by F and G is approximated by $\overline{G}$ and $\overline{F}$. Now suppose A_n is a sequence of sets in $\mathcal{C}$ with union A. We construct sets F and G approximating A as follows: Since $A_n \in \mathcal{C}$, we can choose an closed F_n and an open G_n such that

$$P(G_n \setminus F_n) < \frac{\varepsilon}{2^{n+1}} . \tag{2.14}$$

We then choose n_0 such that

$$P\Big(\bigcup_{n\in\mathbb{N}}F_n \setminus \bigcup_{n\leq n_0}F_n\Big) < \frac{\varepsilon}{2}\;, \tag{2.15}$$

and define $F := \cup_{n\leq n_0}F_n$ (which is closed) and $G := \cup_n G_n$ (which is open). Then $P(G \setminus F) < \varepsilon$, so $\mathcal{C}$ is closed under countable unions. Clearly, it is also closed under complements, so it is indeed a σ-algebra. □

2.4. Weak convergence

A second reason why metrizable spaces are important is that they are exactly the spaces on which we can meaningfully define weak convergence of probability measures; this fact is closely related to regularity. In this chapter, we will have to endure many expressions of the form $\int f dP_n \to \int f dP$ etc, so this is a good time to introduce more concise notation: Whenever a function f on $\mathbf{X}$ is integrable with respect to a measure μ, we write

$$\mu f := \int_{\mathbf{X}} f(x)\mu(dx) \tag{2.16}$$

for the integral.[1] Also, for any topological space $\mathbf{X}$, we denote by $\mathbf{C}_b(\mathbf{X})$ the set of real-valued, bounded, continuous functions on $\mathbf{X}$.

2.17 Definition. A sequence (or net) P_n of probability measures on a metrizable space $\mathbf{X}$ is said to **converge weakly** to a probability measure P if

$$P_n f \to Pf \qquad \text{for all } f \in \mathbf{C}_b(\mathbf{X})\;. \tag{2.17}$$

We denote weak convergence as $P_n \xrightarrow{w} P$. ◁

Weak convergence was of course defined already for $\mathbf{X} = \mathbb{R}^d$ in Chapter 18 of [Probability I], in exactly the same manner. Since the set $\mathbf{C}_b(\mathbf{X})$ is well-defined for any topological space $\mathbf{X}$, it seems tempting to simply generalize the definition to arbitrary topological spaces. That does not work, however, since the definition is based on the fact that a probability measure μ is uniquely determined by all its integrals $\int f d\mu$ for all $f \in \mathbf{C}_b(\mathbf{X})$, which is *not* the case for arbitrary $\mathbf{X}$. However, it turns out that the only condition we need to impose is metrizability:

2.18 Theorem. *Let P and Q be probability measures on a metrizable space $\mathbf{X}$. Then*

$$P = Q \qquad \Leftrightarrow \qquad Pf = Qf \qquad \textit{for all } f \in \mathbf{C}_b(\mathbf{X})\;. \tag{2.18}$$

◁

Proof of Theorem 2.18. Let d be a compatible metric on $\mathbf{X}$. For any open set U, we approximate the indicator function $\mathbb{I}_U$ by a sequence of bounded continuous functions as follows: Since U is open, its complement $F := \overline{U}$ is closed, and we again define the sets F^δ as in (2.13) for any $\delta > 0$. We then define the function

$$f_\delta(x) := \min\{1, \delta^{-1}d(x, \overline{U})\}\;. \tag{2.19}$$

[1] The notation suggests that μ is an operator acting on f. Note that this is more than a shorthand: By the properties of integrals, we have $\mu(\alpha f + \beta g) = \alpha\mu f + \beta\mu g$, so a measure can indeed be regarded as a linear functional acting on functions defined on $\mathbf{X}$.

This function takes value 0 on $\overline{U}$, value 1 outside F^δ, and on $F^\delta \setminus \overline{U}$, it increases from 0 to 1 as $\delta^{-1}d(x,\overline{U})$. Since $d(\,.\,,\overline{U})$ is continuous, so is each f_δ, hence $f_\delta \in \mathbf{C}_b(\mathbf{X})$.

Now we use the fact that U is open, which implies that every point x in U has positive distance $\varepsilon := d(x,\overline{U})$ to the complement $\overline{U}$. Hence, for $\delta < \varepsilon$, we have $f_\delta(x) = 1 = \mathbb{I}_U(x)$. Thus, $f_\delta \nearrow \mathbb{I}_U$ point-wise as $\delta \to 0$. By monotone convergence of the integral, Pf_n converges to $P\mathbb{I}_U = P(U)$. The same holds for Q, and since the integrals coincide for all $f \in \mathbf{C}_b(\mathbf{X})$ by hypothesis, the two measures coincide on all open sets. By Lemma 2.7, that means they are identical. □

2.19 Remark. If you are familiar with Billingsley's textbook [3], you will notice that my proof of Theorem 2.15, and also of Theorem 2.20 below, are precise copies of his; they are the most concise and elegant proofs of these results I am aware of. As Billingsley also shows, Theorem 2.18 can be deduced directly from the regularity result in Theorem 2.15. I have chosen a slightly different argument here, to emphasize the fact that (2.18) is not quite a consequence of regularity. Rather, on metrizable spaces, both are a consequence of the same fact, namely the existence of a continuous distance function. ◁

Our main technical result on weak convergence is the collection of criteria summarized by the next theorem. To state the result, and to prove it, we need a few more definitions: The **interior** A° of a set A is the largest open set contained in A, i.e. the union of all open sets contained in A. The **closure** $\mathrm{cl}(A)$ of A is the smallest (with respect to inclusion) closed set containing A. A point x is called a **boundary point** of A if

$$V \cap A \neq \varnothing \qquad \text{and} \qquad V \cap \overline{A} \neq \varnothing \tag{2.20}$$

for all neighborhoods V of x. The set of all boundary points of A is called the **boundary** ∂A of A. The closure, interior, and boundary of any set satisfy the following relations:

$$\mathrm{cl}(A) = A^\circ \cup \partial A \qquad \text{and} \qquad \partial A = \partial\overline{A} = \mathrm{cl}(A) \cap \mathrm{cl}(\overline{A}) \tag{2.21}$$

Also, recall that every bounded sequence has a limit point. The same is true for bounded nets. The smallest limit point is called the **limit inferior** ($\liminf$), the largest one the **limit superior** ($\limsup$). A sequence or net in a Hausdorff space converges if and only if $\liminf$ and $\limsup$ coincide, in which case this point is the limit.

2.20 Theorem [Criteria for weak convergence]. *Let $(P_s)_{s\in\mathbb{T}}$ be a net of probability measures on a metrizable space $\mathbf{X}$. Then the following are equivalent:*

(1) $P_s \xrightarrow{w} P$.
(2) $P_s f \to Pf$ *for every bounded uniformly continuous function* f.
(3) $\limsup_s P_s(F) \leq P(F)$ *for every closed set* F.
(4) $\liminf_s P_s(G) \geq P(G)$ *for every open set* G.
(5) $P_s(A) \to P(A)$ *for every Borel set* A *with* $P(\partial A) = 0$.

◁

Proof. By definition, (1) implies (2), and clearly (3) ⇔ (4). We will proof

$$(2) \Rightarrow (3) \qquad \text{and} \qquad (3) + (4) \Rightarrow (5) \qquad \text{and} \qquad (5) \Rightarrow (1)\,.$$

Step 1: (2)⇒(3). Let F be closed. For any $\delta > 0$, we again define the function f_δ as in (2.19). We have already established that f_δ is bounded and continuous. Since $|f_\delta(x) - f_\delta(y)| \leq \delta^{-1}d(x,y)$, it is even uniformly continuous. We also know

$$\mathbb{I}_F \leq 1 - f_\delta \leq \mathbb{I}_{F^\delta} \qquad \text{for any } \delta > 0 \; . \tag{2.22}$$

Assuming that (2) holds, we hence have

$$\begin{aligned}\limsup P_s\mathbb{I}_F \overset{(2.22)}{\leq} \limsup P_s(1-f_\delta) &\overset{\substack{\text{limit exists}\\ \text{by (2)}}}{=} \lim P_s(1-f_\delta)\\ &\overset{(2)}{=} P(1-f_\delta) \overset{(2.22)}{\leq} P\mathbb{I}_{F^\delta} = P(F^\delta) \; .\end{aligned}$$

Since P, as a probability measure on a metrizable space, is regular, and since F is closed, we have $P(F^\delta) \searrow P(F)$ for $\delta \to 0$, and (3) hence holds.

Step 2: (3)+(4)⇒(5). Let A be a Borel set.. Since its interior A° is open and its closure $\text{cl}(A)$ is closed, we can apply (3) and (4) and obtain

$$\begin{aligned}P(A^\circ) &\overset{(4)}{\leq} \liminf P_s(A^\circ) \overset{A^\circ \subset A}{\leq} \liminf P_s(A)\\ &\leq \limsup P_s(A) \overset{A\subset \text{cl}(A)}{\leq} \limsup P_s(\text{cl}(A)) \overset{(3)}{\leq} P(\text{cl}(A)) \; .\end{aligned} \tag{2.23}$$

Now assume $P(\partial A) = 0$. By (2.21), this implies $P(\text{cl}(A)) = P(A^\circ)$, so all inequalities in (2.23) are equalities, so

$$P(A) = \liminf P_s(A) = \limsup P_s(A) = \lim P_s(A) \; . \tag{2.24}$$

Step 3: (5)⇒(1). In previous proofs, we have started with sets and constructed bounded continuous functions to approximate their measure; now we have to do the converse. To express $f \in \mathbf{C}_b(\mathbf{X})$ by means of a set, choose any constant t, and consider the set $[f > t]$ of all points x for which $f(x) > t$. Since f is by assumption upper-bounded by some constant c, we have

$$Pf = \int_0^\infty P[f > t]dt \overset{\text{bounded}}{=} \int_0^c P[f>t]dt \; . \tag{2.25}$$

Now we have to find a way to make use of (5). Continuity of f implies $\partial[f > t] \subset [f = t]$. The set $[f = t]$ has non-zero measure only if f assumes constant value t on some non-null set A. Since we can subdivide $\mathbf{X}$ into at most countably many disjoint non-null sets, there are at most countably many such constants t. For all other t, $[f = t]$ is a null set, and hence $P(\partial[f > t]) = 0$. By (5),

$$P_s(\partial[f > t]) \to P(\partial[f > t]) \qquad \text{for all but countably many } t \; . \tag{2.26}$$

We substitute this into (2.25). Note that we are integrating over t in (2.25) with respect to Lebesgue measure, so the countable set of exceptional constants t is a null set. Since the sequence of functions $t \mapsto P_s[f > t]$ is dominated (by the constant function with value 1), and converges to the function $t \mapsto P[f > t]$ for Lebesgue-almost all t, we have

$$P_sf \overset{(2.25)}{=} \int_0^c P_n[f > t]dt \longrightarrow \int_0^c P[f > t]dt \overset{(2.25)}{=} Pf \; , \tag{2.27}$$

where the convergence is due to (5) and Lebesgue's dominated convergence theorem [e.g. Probability I, Theorem 9.1(f)]. □

2.5. Polish spaces and Borel spaces

In this section, we ask which additional properties we have to add to metrizability to obtain spaces on which probability theory works without pitfalls. One property we would like to ensure is that the σ-algebra of every mesurable space we work with is countably generated. That is, for example, a key requirement for the existence of well-behaved conditional probabilities, and for various concepts in statistics (such as well-behaved sufficient statistics).

On a topological space $\mathbf{X} = (\mathcal{X}, \tau)$, it seems we should be able to ensure $\mathcal{B}(\mathbf{X})$ is countably generated by requiring that τ has a countable generator. We have to be careful, though:

2.21 Remark [Generated topology vs. generated σ-algebra]. If we use a generator $\mathcal{G}$ of the topology τ to generate a σ-algebra, we do *not generally obtain the Borel σ-algebra.* The problem is that we may need uncountable unions to construct sets in τ from those in $\mathcal{G}$. Hence, $\sigma(\mathcal{G})$ need not contain every open set. ◁

It turns out, however, that if a generator of τ is (1) a base and (2) countable, then there exists a countable generator of $\mathcal{B}(\mathbf{X})$:

2.22 Lemma. *Let $\mathbf{X}$ be a topological space. If $\mathcal{G}$ is a countable base for the topology of $\mathbf{X}$, it is a generator for the Borel σ-algebra.* ◁

Proof. Obvious: Every open set is a union of sets in the base, so in this case a countable union, and hence also contained in $\sigma(\mathcal{G})$. □

Spaces with a a countable base play an important role in topology, and come with their own terminology:

2.23 Definition. A topological space is called **second-countable** if its topology possesses a countable base. ◁

If $\mathbf{X}$ is metrizable, we can always construct a base explicitly by choosing any compatible metric d. Then the set of all open d-balls, as defined in (2.6), is a base of the topology by Lemma 2.13. We do not even have to use all open balls: Those with centers in a dense subset suffice.

2.24 Definition. A subset D of a topological space $\mathbf{X}$ is called **dense** if every non-empty open set contains a point in D. ◁

Every set is clearly dense in itself. By far the most important example of a non-trivial, dense subset is the set $\mathbb{Q}$ of rational numbers, which is dense in $\mathbb{R}$. (Note that $\mathbb{Q}$ is much smaller than $\mathbb{R}$, since it is countable.) In a metric space $(\mathbf{X}, d)$, dense means equivalently that, for every $x \in \mathbf{X}$ and every $\varepsilon > 0$, there is a point $x' \in D$ with $d(x, x') < \varepsilon$.

2.25 Lemma. *Let $(\mathbf{X}, d)$ be a metric space, D a dense subset and define the system of open balls with centers in A and rational radii,*

$$\mathcal{G} := \{B_r(x) \,|\, x \in D \text{ and } r \in \mathbb{Q}_+\} . \tag{2.28}$$

Then $\mathcal{G}$ is a base of the topology of $\mathbf{X}$. ◁

Proof. We know by Lemma 2.13 that every open set G is the union of all open balls it contains. Since D is dense, every point $x \in G$ is also contained in an open ball centered at a point in D. Since $\mathbb{Q}_+$ is dense in $\mathbb{R}_+$, this is even true if we consider only balls centered at points in D with rational radii. Hence, every open set G is the union of all elements of $\mathcal{G}$ which are subsets of G, which is precisely the definition of a base. □

We see immediately that the base constructed in (2.28) becomes countable if the dense subset D is countable. A space in which such a subset exists is called separable.

2.26 Definition. A topological space $\mathbf{X}$ is **separable** if it possesses a dense subset that is countable. ◁

Since separability makes the set $\mathcal{G}$ in (2.28) a countable base, it makes $\mathbf{X}$ second-countable. Since conversely any second-countable space is clearly separable, we hence the following:

2.27 Theorem. *A metrizable space is second-countable if and only if it is separable. In particular, the Borel σ-algebra of a separable metrizable space is countably generated.* ◁

2.28 Example [The ball σ-algebra and nonparametric statistics]. In any metric space, the σ-algebra $\sigma(\mathcal{G})$ generated by the system $\mathcal{G}$ of *all* open balls is called the **ball σ-algebra**. The last theorem implies that $\sigma(\mathcal{G}) = \mathcal{B}(\mathbf{X})$ if the space is separable. In non-separable spaces, the caveat of Remark 2.21 applies; in this case, we generally have $\sigma(\mathcal{G}) \subset \sigma(\tau)$, but equality does not necessarily hold.

The theory of nonparametric statistical models is a topic in which you may regularly encounter spaces $\mathbf{X}$ which are *not* separable. This is, roughly speaking, due to the fact that meaningful convergence rates require metrics defined by supremum-type norms or "uniform" norms. These norms have a habit of making interesting infinite-dimensional spaces non-separable. Consequently, the ball σ-field is smaller than the Borel σ-field. This means that functions *into* $\mathbf{X}$ which are not Borel-measurable may still be ball-measurable (and hence be valid random variables with respect to the ball σ-field). In mathematical statistics and empirical process theory, the ball σ-field turns out to be a possible alternative to the Borel sets, mostly for three reasons:

(1) Many interesting functions are ball- but not Borel-measurable.
(2) The regularity theorem (Theorem 2.15) also holds on the ball σ-field.
(3) The basic results on weak convergence (i.e. Theorem 2.18 and Theorem 2.20) still hold on the ball σ-field for measures whose support is separable—that means we can still salvage weak convergence theory for measures concentrated on a sufficiently small subset of the space.

On the other hand, odd things happen; for example, since there are open sets which are not measurable, continuous functions need not be measurable. See [10] for more on this topic. ◁

2.29 Remark [Continuous functions on separable spaces]. The existence of a countable base aside, separability also has a number of other useful consequences. One is due to the fact that, if $\mathbf{X}$ is a topological space and D a dense subset, then every continuous mapping on $\mathbf{X}$ is uniquely determined by its values on D. Hence, every continuous function on separable space is completely determined by its value at a countable number of points. In this sense, continuous functions on such spaces are objects of "countable complexity", and the space of continuous functions on a separable space can be regarded as a space of countable dimension. ◁

Overall, metrizability and separability are the two most important properties of topological spaces for the purposes of probability and statistics, and many relevant properties can be established assuming only these two. Indeed, especially in older research articles, a separable metric space is often the standard assumption. Modern probability usually adds one more assumption, called *completeness*, to exclude certain pathologies. To understand this property, we need to point out a certain odd way in which Cauchy sequences can behave on general topological spaces. Recall the definition: A sequence (x_n) in a metric space $(\mathbf{X}, d)$ is called a d-**Cauchy sequence** if, for every $\varepsilon > 0$, there exists some n_ε such that

$$d(x_n, x_m) < \varepsilon \qquad \text{for all } n, m > n_\varepsilon \ . \tag{2.29}$$

Compare this to the definition of a sequence converging to a limit x with respect to d:

$$d(x_n, x) \to 0 \qquad \text{for } n \to \infty \ . \tag{2.30}$$

These two are not quite the same: The convergent sequence clusters more and more tightly around the limit point as n grows, whereas the Cauchy sequence becomes more and more "densely" concentrated. Every d-convergent sequence is clearly d-Cauchy. Intuitively, the definitions should be equivalent—and they are of course equivalent in Euclidean space—but in general, a d-Cauchy sequence need not be convergent with respect to d. Here is an example from [A&B], using a rather odd metric:

2.30 Example [Cauchy sequences need not converge]. Define a function d on $\mathbb{N} \times \mathbb{N}$ as $d(n, m) := |\frac{1}{m} - \frac{1}{n}|$. Then d is a metric on $\mathbb{N}$ that metrizes the discrete topology. The sequence $(1, 2, 3, \dots)$ is a d-Cauchy sequence, even though it is clearly not convergent. $\triangleleft$

To avoid nasty surprises, we exclude spaces where Cauchy sequences might not converge. This requirement has a name:

2.31 Definition. A metric space (X, d) is called **complete** if every d-Cauchy sequence converges. A metrizable space is called **complete**, or **completely metrizable** if there exists a compatible metric d such that (X, d) is complete. $\triangleleft$

Note that convergence is a *topological* concept, whereas completeness is a *metric* concept (although complete metrizability has topological implications). The natural habitat of modern probability theory are spaces which have all three properties discussed above:

2.32 Definition. A separable, completely metrizable space is called a **Polish space**. $\triangleleft$

Polish spaces and mappings between Polish spaces have many very useful properties that we can unfortunately (and I mean unfortunately!) not discuss in detail; if you are interested in this topic, I would recommend the account given by Aliprantis and Border [A&B], which is simply excellent.

We will see below (Theorem 2.54) that two distinct Polish topologies on a given set may generate the same Borel sets. If we are only interested in measurability properties, it can therefore be useful to abstract from the specific choice of Polish topology and consider only the σ-algebra it generates:

2.33 Definition. A measurable space $(\mathcal{X}, \mathcal{B})$ is called a **Borel space** (or, by some authors, a **standard Borel space**) if there is a Polish topology on $\mathcal{X}$ whose Borel σ-algebra is $\mathcal{B}$. $\triangleleft$

In this section, we have convinced ourselves that Polish spaces are the spaces on which we would like to work. In summary:

- Metrizability ensures regularity of probability measures and a meaningful definition of weak convergence.
- Separability ensures the Borel σ-algebra is countably generated. We can even explicitly construct a countable generator using open balls.
- Completeness avoids pathologies regarding Cauchy sequences. Why this is useful is a little less obvious from the perspective of probability, and I have to ask you to take the importance of completeness on faith. Basically, it has been recognized in analysis in the last 50 or so years that Polish spaces have nicer properties than just separable metric spaces (some of which are listed in Section 2.9 below). Hence, we work on Polish spaces whenever possible. In statistics and probability, assuming completeness is usually no restrictive. The property of Polish spaces we sometimes have to sacrifice rather tends to be separability, cf. Example 2.28.

To derive the properties of Polish spaces, we have started with useful properties of $\mathbb{R}^d$ and asked for those to be preserved. However: We have not so far verified that any space other than $\mathbb{R}^d$ is actually Polish, so all of this may just be wishful thinking. It turns out that many of the most important spaces arising in probability and related fields *are* Polish, however. Instead of proving our way through every single one of them, I have listed some of the key examples in Table 2.1. I will only state the following result, which shows under which conditions the derived topologies—product and relative topologies—are again Polish. Recall that a countable intersection of open sets is called a $\boldsymbol{G_\delta}$ **set** (and a countable union of closed sets a $\boldsymbol{F_\sigma}$ **set**).

2.34 Theorem [Derived topologies and Polish spaces]. *Every countable product of Polish spaces is Polish in the product topology. A subset of a Polish space* $\mathbf{X}$ *is Polish in the relative topology if and only if it is a* G_δ *set in* $\mathbf{X}$. ◁

2.6. The space of probability measures

For any topological space $\mathbf{X}$, we denote the set of all probability measures on the Borel σ-algebra of $\mathbf{X}$ as $\mathcal{PM}(\mathbf{X})$. Weak convergence of probability measures defines a topology on $\mathcal{PM}(\mathbf{X})$, called the **topology of weak convergence**. This topology turns $\mathcal{PM}(\mathbf{X})$ into a topological space $\mathbf{PM}(\mathbf{X})$, which we call the **space of probability measures** on $\mathbf{X}$.

When we refer to the space of probability measures, we always refer to the topology of weak convergence, as is common practice in the literature. Otherwise, we state so explicitly.

Note that the topology of weak convergence is a weak topology, namely the one generated by the family of mappings $\{\mu \mapsto \int f d\mu \,|\, f \in \mathbf{C}_b(\mathbf{X})\}$. (Why?) If $\mathbf{X}$ is separable, it is also a metric topology, and again separable. More generally, the space $\mathbf{PM}(\mathbf{X})$ inherits key topological properties from $\mathbf{X}$:

2.35 Theorem [Topological properties of PM(X)]. *Let* $\mathbf{X}$ *be a metrizable space. Then the following hold:*

(1) $\mathbf{PM}(\mathbf{X})$ *is separable and metrizable if and only if* $\mathbf{X}$ *is separable.*
(2) $\mathbf{PM}(\mathbf{X})$ *is compact if and only if* $\mathbf{X}$ *is compact.*

Name/Symbol	Set	Topology	Properties
	Finite or countable set	Discrete topology	Polish
	Countable product of Polish spaces	Product topology	Polish
	Borel subset of Polish space	Relative topology	Polish
$\mathbb{R}^d$	Euclidean space	Standard topology	Polish
$\mathbf{C}(\mathbf{X}, \mathbf{Y})$	Continuous functions $\mathbf{X} \to \mathbf{Y}$	Uniform convergence on compact sets	Polish if X compact and Polish, Y Polish [A&B, §3.19]
$\mathbf{D}(\mathbf{X}, \mathbb{R})$	Rcll functions, $\mathbf{X}$ Polish	Skorohod topology	Polish [3, Theorem 12.2]
Banach space	Normed, complete topological vector space	Metric topology of norm	Polish iff separable
$\mathbb{R}^{\mathbb{N}}$	Real-valued sequences	Product topology	Polish
ℓ_p, $1 \leq p < \infty$	Real-valued sequences (x_n) with finite series $\sum x_n^p$	ℓ_p norm	Polish (separable Banach)
$\mathbf{L}_p(\mu)$, $1 \leq p \leq \infty$	Functions with finite integral $\int f^p d\mu$	$\mathbf{L}_p$ norm	Polish (separable Banach) if μ σ-finite measure on countably generated σ-algebra
Separable Hilbert space	Isomporhic to $\mathbb{R}^d$ if dimension finite	Euclidean norm	Polish (separable Banach)
	Isomorphic to ℓ_2 if dimension infinite	ℓ_2 norm	
Cantor space	$\{0, 1\}^\infty$	Product topology	Polish and compact
$\mathbf{PM}(\mathbf{X})$	Space of probability measures on metrizable $\mathbf{X}$	Weak convergence	Separable iff $\mathbf{X}$ separable Compact iff $\mathbf{X}$ compact Polish iff $\mathbf{X}$ Polish Borel iff $\mathbf{X}$ Borel

Table 2.1. Useful Polish spaces.

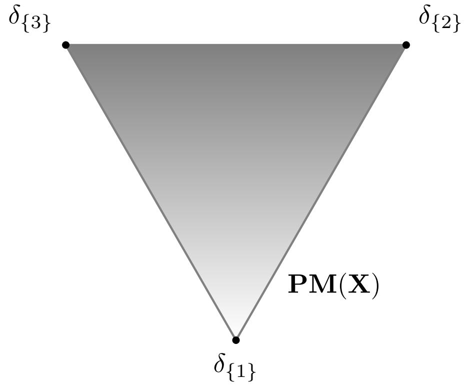


Figure 2.1. The space $\mathbf{PM(X)}$ for the finite discrete space $\mathbf{X} = \{1, 2, 3\}$.

(3) $\mathbf{PM(X)}$ *is Polish if and only if* $\mathbf{X}$ *is Polish.*
(4) $\mathbf{PM(X)}$ *is Borel if and only if* $\mathbf{X}$ *is Borel.*

◁

Proof. See [A&B, §15.3]. □

2.36 Remark [Prokhorov metric]. If $\mathbf{X}$ is a metrizable space and d a compatible metric, the the topology of weak convergence can be metrized by the **Prokhorov metric** (also known as **Lévy-Prokhorov metric**): We denote by A^δ the set $\{x \in \mathbf{X} | d(x, A) \leq \delta\}$ for any measurable set A. (Note that, in contrast to the similar sets F^δ used in the proof of the regularity theorem, we now use $\leq$, i.e. the sets A^δ are closed.) The Prokhorov metric is defined for any two probability measures P and Q on $\mathbf{X}$ as

$$d_{\text{LP}}(P, Q) := \inf\{\delta > 0 \,|\, P(A) \leq Q(A^\delta) + \delta \text{ for all } A \in \mathcal{B}(\mathbf{X})\} \ . \tag{2.31}$$

◁

The space $\mathbf{PM(X)}$ has very pretty geometric properties: If P and Q are two probability measures, then the two-component mixture $(\lambda P + (1 - \lambda)Q)$ (for $\lambda \in [0, 1]$) is obviously again a probability measure. Hence, $\mathbf{PM(X)}$ is convex. The extreme points of a convex set are precisely those points not expressable as convex combinations of other points in the set, which in the case of probability measures on $\mathbf{X}$ are precisely the point masses δ_x. In the simple case $\mathbf{X} := \{1, 2, 3\}$, equipped with the discrete topology, $\mathbf{PM(X)}$ looks like the polytope shown in Figure 2.1.

The center of the polytope corresponds to the uniform distribution. The faces of the convex set (in this case: the three boundary lines of the triangle) are those measures which can be obtained as a convex combinations of a subset of extreme points (in this case, mixtures of two point masses). For each additional point in $\mathbf{X}$, we must add a point mass, and hence a dimension. If $\mathbf{X}$ is finite with n elements, we obtain a polytope in $\mathbb{R}^{n-1}$. In general, when $\mathbf{X}$ is uncountable, we obtain an infinite-dimensional space, but the triangle above is still its closest finite-dimensional analogue—even if $\mathbf{X}$ is infinite, any two point masses are still connected by an edge, since their mixtures are not representable using any other point masses.

If you are interested in the geometric properties of $\mathbf{PM}(\mathbf{X})$, I recommend Chapter 15.2 and 15.4 in [A&B].

2.37 Remark [$\mathcal{PM}(\mathbf{X})$ is not separable in total variation]. The distance on probability measures most of you may be familiar with is the total variation distance

$$d_{\mathrm{TV}}(P,Q) := 2 \sup_{A\in\mathcal{B}(\mathbf{X})} |P(A) - Q(A)| \,. \tag{2.32}$$

This distance has some nice properties; for instance, if S is a sufficient statistic for the set $\{P,Q\}$, then it preserves total variation, i.e. $d(S(P),S(Q)) = d(P,Q)$. It also has a nice probabilistic interpretation:

$$d_{\mathrm{TV}}(P,Q) := 2\inf\{\mathbb{P}(X \neq Y)|\mathcal{L}(X) = P, \mathcal{L}(Y) = Q\} \tag{2.33}$$

The infimum is taken over all possible pairs of random variables X and Y whose laws are respectively P and Q—note that such pairs can have arbitrary dependence structure, since we only constrain the marginals P and Q of their joint distribution, i.e. the infimum is taken over all possible joint distributions with the given marginals.

The set of probability measures on a Polish space is *not* separable in topology defined by total variation. Total variation can nonetheless be useful on certain subset of $\mathbf{PM}(\mathbf{X})$: If μ is a σ-finite measure on $\mathbf{X}$, let $\mathcal{P}(\mu)$ denote the set of all probability measures which are absolutely continuous with respect to μ. Then the restriction of d_{TV} to $\mathcal{P}(\mu)$ is precisely the $\mathbf{L}_1$ distance, and in particular, $d_{\mathrm{TV}}(P,Q) = \int |p(x) - q(x)|\mu(dx)$ if p and q are the μ-densities of P and Q. Moreover, d_{TV} metrizes the relative topology of $\mathbf{PM}(\mathbf{X})$ on $\mathcal{P}(\mu)$. $\triangleleft$

2.7. Compactness and local compactness

Ask ten mathematicians which topological property they would consider the most useful one, and chances are nine or so will name compactness. Compact sets are, roughly speaking, the topological generalization of finite sets: Recall that there are two ways in a which a sequence in, say, $\mathbb{R}$ may not converge: It could oscillate back and forth, or it could "escape to infinite". The latter can only happen because $\mathbb{R}$ is unbounded. In a finite set, in particular, a sequence that does not converge must oscillate. Whether or not it does, it still must visit at least one point infinitely often, and so even a sequence that does not converge has a subsequence which does. Sets with this property—every sequence (net) has a convergent subsequence (subnet)—are called compact. There are several equivalent definitions of compactness; the following one emphasizes the intuition that compact sets are generalization of finite sets.

Let K be a set in a topological space $\mathbf{X}$. A family $\{G_s|s \in \mathbb{T}\}$ of open sets is called an **open cover** of K if $K \subset \bigcup_s G_s$.

2.38 Definition. A subset K of a topological space is **compact** if every open cover has a finite subcover. A topological space is compact if it is a compact set. $\triangleleft$

2.39 Fact. Some helpful properties of compacts sets and spaces:

(1) A topological space $\mathbf{X}$ is compact iff every net in $\mathbf{X}$ has a convergent subnet.
(2) A metrizable space $\mathbf{X}$ is compact iff every sequence in $\mathbf{X}$ has a convergent subsequence.
(3) Compact sets are closed (if $\mathbf{X}$ is Hausdorff).

(4) Continuous images (images, not preimages!) are compact.
(5) If a compact space is metrizable, it is Polish.

◁

Clearly, finite unions and countable intersections of compact sets are compact. Compactness is a fairly strong property, but where it occurs, it is surprisingly robust:

2.40 Theorem [Tychonoff]. *The product of (arbitrarily many) topological spaces is compact in the product topology if and only if each factor is compact.* ◁

The following lemma is a convenient way of verifying compactness in metric spaces: A set A in a metric space is called **totally bounded** if, for every $\varepsilon > 0$, there is a finite set $\{x_1, \dots, x_n\}$ such that $A \subset \bigcup_{i \leq n} B_\varepsilon(x_i)$.

2.41 Lemma. *In a complete metric space, a closed set is compact if and only if it is totally bounded.* ◁

Note that total boundedness is not the same as finite diameter: If a set in a vector space has, say, diameter 1, and we try to cover it with balls of diameter $\frac{1}{2}$, we need more ball to do so the higher the dimension of the space. If the dimension is infinite, we need an infinite number of balls. Indeed, this idea is used to define a notion of dimension in spaces that are not vector spaces (and hence have no axes to count).

The property that sequences in a set cannot "escape to infinity" is true whenever the set is contained in a compact set, even if it is not itself compact. In a Hausdorff space, that is the case if the set is not closed. In this case, every sequence of points in the set still has a convergent subsequence, but the limit of the subsequence may not be in the set. Such sets have a name:

2.42 Definition. A subset of a topological space is called **relatively compact** if its closure is compact. ◁

Compactness of a space $\mathbf{X}$ can be interpreted vaguely as saying that $\mathbf{X}$ is not too large—the real line is not compact because it is too long. In general analysis, there is another way in which a space can become "too large": In a metric vector space, for example, the unit ball is bounded, and it is compact if the dimension of the space is finite. If the dimension is infinite, the ball is still bounded, but it is no longer compact. The qualitative difference between the two cases is that the line is *globally* large, but perfectly well-behaved in a small neighborhood of a given point. In an infinite-dimensional vector space, the space is "too large" even locally around each point. Too distinguish these two cases in a topological way, we use the following definition:

2.43 Definition. A topological space is **locally compact** if every point has a neighborhood that is compact, i.e. if for every point x, there is a compact set which contains x in its interior. ◁

Compactness clearly implies local compactness. The example above seems to suggest that compact spaces should hence behave like finite-dimensional spaces, but we have to be careful: Even spaces which in many regards have infinite-dimensional properties (such as $\mathbf{PM}(\mathbf{X})$ on a compact space $\mathbf{X}$) can be compact, and hence locally compact. In vector spaces, however, there is a one-to-one correspondence between local compactness and dimension.

In order to make this correspondence precise, we have to define the notion of a topological vector space. Basically, for every algebraic structure (a set with operations defined on it) we can define a topological version of that structure, by equipping the set with a topology. However, we have to make sure that the topology and the operation are compatible:

2.44 Definition. A **topological group** is a group with a topology defined on it which makes the group operation continuous. Similarly, a **topological vector space** is a vector space endowed with a topology which makes linear operations (addition and scaling) continuous. $\triangleleft$

2.45 Lemma. *A topological vector space is locally compact if and only if it is finite-dimensional.* $\triangleleft$

One way in which locally compact spaces are important in probability and statistics is that they admit a translation-invariant measure. Such a measure does *not* usually exist, even on well-behaved spaces such as Polish ones.

2.46 Theorem. *Let $\mathbf{X}$ be a locally compact, second-countable Hausdorff space, with an operation $+$ defined on it such that $(\mathbf{X},+)$ is a topological group. Than there is a measure λ on $\mathbf{X}$ which satisfies $\lambda(A+x)=\lambda(A)$ for all Borel sets A and points $x\in\mathbf{X}$. This measure is unique up to scaling by a positive constant.* $\triangleleft$

The measure λ is called **Haar measure** on the group $(\mathbf{X},+)$. Lebesgue measure is Haar measure on the group $(\mathbb{R}^d, d)$, scaled to satisfy $\lambda([0,1]^d)=1$.

2.47 Remark [Translation-invariance and density representations]. In statistics in particular, the existence of Haar measure is relevant since it is arguably a prerequisite for working with densities: Recall that the density $f=d\mu/d\nu$ of some measure μ is always defined with respect to another measure ν. That does not require local compactness, of course, but the interpretation of densities becomes difficult if ν is not translation-invariant—does a large value of f on a set A mean that μ is large on A, or that ν is small? Hence, to keep densities interpretable, ν needs to be "flat", which is precisely what translation-invariance means. $\triangleleft$

2.8. Tightness

Recall the regularity properties of measures defined in Section 2.3, in terms of the behavior of a measure on open or closed sets. If we replace the closed sets by compact once—a stronger hypothesis—we obtain a property called tightness, which is so useful that it has become one of the pivotal concepts in modern probability.

2.48 Definition. A measure μ on a topological space $\mathbf{X}$ is called **tight** if

$$\mu(A)=\sup\{\mu(K)|K\subset A \text{ and } K \text{ compact }\} \qquad \text{for all } A\in\mathcal{B}(\mathbf{X})\ . \tag{2.34}$$

$\triangleleft$

For tightness, there is a direct counterpart to Theorem 2.15. To guarantee regularity, metrizability of $\mathbf{X}$ was sufficient. Tightness is a stronger property, so it is not surprising that we have to strengthen the hypothesis somewhat.

2.49 Theorem. *Every probability measure on a Polish space is tight.* $\triangleleft$

A tight probability measure is called a **Radon measure**. (More generally, an arbitrary measure is a Radon measures if it is tight and also locally finite, i.e.

$\mu(K) < \infty$ for all compact K, which of course is always the case for probability measures.)

Proof of Theorem 2.49. Fix any compatible metric on $\mathbf{X}$. Suppose P is a probability measure on $\mathbf{X}$. Given any $\varepsilon > 0$, we have to find a compact set K such that $P(K) > 1 - \varepsilon$. We construct K as follows: Since $\mathbf{X}$ is separable, it has a countable dense subset $\{x_1, x_2, \ldots\}$. Hence, the d-balls $B_\delta(x_i)$ cover $\mathbf{X}$ for any choice of δ; in particular. For every $k \in \mathbb{N}$, choose $\delta := 1/k$. Since P is a probability measure, there is for each k some $n_k \in \mathbb{N}$ such that

$$P\Big(\bigcup_{i \leq n_k} B_{1/k}(x_i)\Big) > 1 - \frac{\varepsilon}{2^k} \ . \tag{2.35}$$

Now take the intersection over all k: The set

$$A := \bigcap_{k\in\mathbb{N}} \bigcup_{i \leq n_k} B_{1/k}(x_i) \ . \tag{2.36}$$

is totally bounded. Let $K := \mathrm{cl}(A)$ be its closure. Then by Lemma 2.41, K is compact. By (2.35), we have

$$P(K) > 1 - \sum_{k\in\mathbb{N}} \frac{\varepsilon}{2^k} > 1 - \varepsilon \ , \tag{2.37}$$

which is precisely what we had to show. □

Tightness is a particularly powerful property if it holds uniformly for an entire family of measures.

2.50 Definition. A family $\{P_s | s \in \mathbb{T}\}$ of measures on a topological space $\mathbf{X}$ is called **tight** if, for every $\varepsilon > 0$, there is a compact set K such that

$$P_s(K) > 1 - \varepsilon \qquad \text{for all } s \in \mathbb{T} \ . \tag{2.38}$$

◁

Above, we have discussed the relationship between compactness (or relative compactness) and convergence. Since the topology on $\mathbf{PM}(\mathbf{X})$ describes weak convergence, and one of the typical problems arising in proofs of probabilistic results is to establish weak convergence of a given sequence, it would be very useful to know what the relatively compact sets of $\mathbf{PM}(\mathbf{X})$ are. The next result relates relative compactness of sets in $\mathbf{PM}(\mathbf{X})$ to the behavior of the measures in the set on compact sets in $\mathbf{X}$.

2.51 Theorem [Prokhorov]. *Let $\{P_s | s \in \mathbb{T}\}$ be a family of probability measures on a metrizable space $\mathbf{X}$. Then*

$$\{P_s | s \in \mathbb{T}\} \text{ tight} \qquad \Rightarrow \qquad \{P_s | s \in \mathbb{T}\} \text{ relatively compact in } \mathbf{PM}(\mathbf{X}) \ , \tag{2.39}$$

and the two are even equivalent if $\mathbf{X}$ is Polish. ◁

On a metrizable space, we can hence prove weak convergence via tightness. Given a sequence (P_n), we can show that $P_s \xrightarrow{w} P$ by showing:

(1) $\{P_n | n \in \mathbb{N}\}$ is tight.
(2) If a subsequence converges, it converges to P.

Then Theorem 2.51 implies $P_n \xrightarrow{w} P$.

2.9. Some additional useful facts on Polish spaces

A very important result (which I will not prove here) is that any two uncountable Borel spaces are isomorphic, in the following sense: A **Borel isomorphism** of two Borel spaces $(\mathcal{X}_1, \mathcal{B}_1)$ and $(\mathcal{X}_2, \mathcal{B}_2)$ is a bijective mapping $f : \mathcal{X}_1 \to \mathcal{X}_2$ which is measurable and has a measurable inverse.

2.52 Borel isomorphism theorem. *Every Borel space is Borel isomorphic to a Borel subset of* $[0, 1]$*. In particular, for every probability measure* P *on a Borel space* $(\mathcal{X}, \mathcal{B})$*, there exists a Borel isomorphism* $f : [0, 1] \to \mathcal{X}$ *and a probability measure* P_0 *on* $[0, 1]$ *such that* $P = fP_0$*. If* $\mathcal{X}$ *is uncountable,* P_0 *can always be chosen as the uniform distribution.* ◁

To ensure that a bijection between Borel spaces is a Borel isomorphism, it is in fact sufficient to verify that it is measurable:

2.53 Theorem. *If a bijection* $f : \mathcal{X}_1 \to \mathcal{X}_2$ *between two Borel spaces is measurable, its inverse is also measurable.* ◁

Finally, I would like to point out that it can be risky to think of "the" Polish space defining a Borel space, since various different Polish topologies can all define the same Borel sets—this is again a consequence of the fact that the Borel σ-algebra is typically much larger than the topology by which it is generated. We can often change the topology considerably without affecting the Borel sets. This is illustrated in a rather stunning manner by the following result:

2.54 Theorem. *Let* $\mathbf{X} = (\mathcal{X}, \tau)$ *be a Polish space, and* $f : \mathbf{X} \to \mathbf{Y}$ *a Borel function from* $\mathbf{X}$ *into an arbitrary second-countable space* $\mathbf{Y}$*. Then there exists another Polish topology on* $\mathbf{X}$ *such that:*

(1) τ' *generates the same Borel sets as* τ*.*

(2) f *is continuous with respect to* τ'*.*

The same is even true if the single function f *is replaced by a countable family of Borel functions.* ◁

CHAPTER 3

Conditioning

The topic of this chapter are the definition and properties of conditional distributions. Intuitively, it is pretty straightforward what a conditional distribution should be: Suppose X and Y are two random variables. We want to define an object $\mathbb{P}[X \in A|Y = y]$ with the semantics

$$\mathbb{P}[X \in A|Y = y] = \text{ probability of } \{X \in A\} \text{ given that } Y = y\ . \tag{3.1}$$

This quantity depends on two arguments, the measurable set A and the value y of Y, and we can hence abbreviate $\mathbf{p}(A, y) := \mathbb{P}[X \in A|Y = y]$. Denote by P_{Y} the distribution of Y. Assuming (3.1) is true, the function $\mathbf{p}$ should certainly satisfy

$$\int_B \mathbf{p}(A, y) P_{\mathrm{Y}}(dy) = \mathbb{P}\{X \in A, Y \in B\}\ . \tag{3.2}$$

3.1 Definition. Let $\mathbf{X}$ be metrizable and $\mathbf{Y}$ a measurable space. A measurable mapping $\mathbf{Y} \to \mathbf{PM}(\mathbf{X})$ is called a **probability kernel**. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbf{X}$ and $Y : \Omega \to \mathbf{Y}$ two random variables. Then a probability kernel $y \mapsto \mathbf{p}(\bullet, y)$ satisfying (3.2) is called a **conditional probability** of X given Y. ◁

Note that we can regard $\mathbf{p}$ as a function $y \mapsto \mathbf{p}(\bullet, y)$ of a single argument (which takes points to measures), or as a function $(A, y) \mapsto \mathbf{p}(A, y)$ of two arguments (which maps into $[0, 1]$). Both perspectives are useful, and I will use the notations $\mathbf{p}(y)$ and $\mathbf{p}(A, y)$ interchangeably. Now, before I explain the definition in more detail, let me state the most important result of this chapter:

3.2 Theorem [Existence of conditional distributions]. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let $\mathbf{X}$ be Polish, and let $\mathbf{Y}$ be a measurable space. Then for any two random variables $X : \Omega \to \mathbf{X}$ and $Y : \Omega \to \mathbf{Y}$, the following holds:*

(1) **Existence:** *X has a conditional distribution given Y, that is, there exists a probability kernel satisfying (3.2).*

(2) **Uniqueness:** *There exists a null set N such that any two conditional probabilities $\mathbf{p}$ and $\mathbf{p}'$ of X given Y satisfy*

$$\forall \omega \notin N : \qquad \mathbf{p}(A, \omega) = \mathbf{p}'(A, \omega) \qquad \textit{for all } A \in \mathcal{B}(\mathbf{X})\ . \tag{3.3}$$

◁

We will not prove this result yet—it is a special case of a more abstract result which we prove below.

3.1. A closer look at the definition

Let me try to unpack Definition 3.1 a bit. Its rough structure should be pretty accessible—a conditional probability is a mapping which takes points to probability measures. We should try to understand the detailed assumptions, though: (3.1) is not a precise mathematical statement, but it tells us that (1) $\mathbf{p}(\bullet, y)$ should be a measure for every y and (2) that (3.2) should hold. From there, we get to the definition as follows:

(1) For (3.2) to even be well-defined, the function $y \mapsto \mathbf{p}(A, y)$ must be integrable, and hence measurable, for every A.
(2) If we assume metrizability of $\mathbf{X}$ (as the definition does), we can equip the set of probability measures on $\mathbf{X}$ with the topology of weak convergence. That makes $y \mapsto \mathbf{p}(\bullet, y)$ a mapping $\mathbf{Y} \to \mathbf{PM}(\mathbf{X})$. We will see below that we should actually require $\mathbf{X}$ to be Polish to make conditional distributions behave well, so metrizability is no loss of generality.
(3) The set-wise (for every A) measurability of $y \mapsto \mathbf{p}(A, y)$ (which maps y into $[0,1]$) is now equivalent to measurability of the mapping $y \mapsto \mathbf{p}(\bullet, y)$ (which maps y into $\mathbf{PM}(\mathbf{X})$). (This is so because the evaluation functional $e_A : \mu \mapsto \mu(A)$, which takes a probability measure μ to its value on A, is a measurable mapping $\mathbf{PM}(\mathbf{X}) \to [0,1]$ for every Borel set A, and since $\mathbf{p}(A, y) = e_A(\mathbf{p}(\bullet, y))$.)

3.2. Conditioning on σ-algebras

The general notion of a conditional distribution is defined by conditioning on a σ-algebra, rather than on a random variable (just as for conditional expectations, cf. [Probability I, Chapter 23]). We arrive at the definition by showing that the conditional distribution of X given Y does not directly depend on Y, only on the σ-algebra $\sigma(Y)$ generated by Y. This σ-algebra can then be substituted by some other σ-algebra $\mathcal{C}$. The intuitive meaning of the conditional probability $\mathbb{P}[X \in A|\mathcal{C}]$ of X given $\mathcal{C}$ is

$$\begin{aligned}\mathbb{P}[X \in A|\mathcal{C}](\omega) \; &= \text{probability that } X(\omega) \in A \text{ if we know} \\ &\quad\; \text{for every } C \in \mathcal{C} \text{ whether } \omega \in C\;.\end{aligned}$$

For a fixed measurable set A, the function

$$y \mapsto \mathbf{p}(A, y) = \mathbb{P}[X \in A|Y = y] \tag{3.4}$$

is called the *conditional probability* of the event $\{X \in A\}$ given Y. This is not yet our final definition, since a conditional probability can be defined more generally than a conditional distribution: It is a much simpler object, and does not require X to take values in a Polish space. For the purposes of this section, we hold A fixed. We can hence abbreviate further and write

$$f(y) := \mathbb{P}[X \in A|Y = y]\;. \tag{3.5}$$

Equation (3.2) then becomes

$$\mathbb{P}\{X \in A, Y \in B\} = \int_B f(y) P_Y(dy)\;. \tag{3.6}$$

Hence, any measurable function f satisfying (3.6) can be regarded as a conditional probability of $\{X \in A\}$ given Y.

However, the event $\{X \in A\}$ is really the set $X^{-1}A$ in Ω. Aside from defining this event, the random variable does not play any role in (3.6). There is hence no reason to restrict ourselves to sets of the form $X^{-1}A$ for some A; more generally, we can consider any set $A' \in \mathcal{A}$. We hence define:

3.3 Definition. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $(\mathcal{Y}, \mathcal{A}_Y)$ a measurable space and $Y : \Omega \to \mathcal{Y}$ a random variable. Let $A' \in \mathcal{A}$. Then any measurable function $f : \mathbf{Y} \to [0,1]$ satisfying

$$\mathbb{P}(A' \cap Y^{-1}B) = \int_B f(y) P_{\mathrm{Y}}(dy) \tag{3.7}$$

is called a **conditional probability** of A' given Y, denoted $f(y) =: \mathbb{P}[A|Y = y]$. ◁

The law P_{Y} of Y is precisely the image measure $P_{\mathrm{Y}} = Y(\mathbb{P})$. We can therefore re-express f on Ω using the identity[1]

$$\int_B f(y) P_{\mathrm{Y}}(dy) = \int_{Y^{-1}B} f \circ Y(\omega)\mathbb{P}(\omega) \ . \tag{3.8}$$

The function $g : \Omega \to [0,1]$ defined by $g := f \circ Y$ can then be interpreted as

$$g(\omega) = \mathbb{P}[A|Y = Y(\omega)] \ . \tag{3.9}$$

Since (3.8) holds only for those sets in Ω which are of the form $Y^{-1}B$, the function g is $\sigma(Y)$-measurable.

Let me summarize what we can conclude so far: A valid conditional probability of A given Y is any $\sigma(Y)$-measurable function $g : \Omega \to [0,1]$ satisfying

$$\mathbb{P}(A \cap B) = \int_B g(\omega)\mathbb{P}(d\omega) \qquad \text{for all } B \in \sigma(Y) \ . \tag{3.10}$$

As for X above, we note the random variable Y plays no explicit role in this last formulation—it only defines the σ-algebra $\sigma(Y)$. Although we defined g above by means of Y,

whether a given function g satisfies (3.10) depends only on the σ-algebra $\sigma(Y)$.

Since any σ-algebra $\mathcal{C} \subset \mathcal{A}$ is of the form $\mathcal{C} = \sigma(Y)$ for *some* random variable Y, we can choose an arbitrary σ-algebra $\mathcal{C} \subset \mathcal{A}$ and substitute it for $\sigma(Y)$ in (3.10) to obtain

$$\mathbb{P}(A \cap B) = \int_B g(\omega)\mathbb{P}(d\omega) \qquad \text{for all } B \in \mathcal{C} \ . \tag{3.11}$$

We can then regard a $\mathcal{C}$-measurable function g satisfying (3.11) as a conditional probability of A given the σ-algebra $\mathcal{C}$.

Now, a sharp look (3.11) shows that we know a $\mathcal{C}$-measurable function which satisfies the equation, namely the conditional expectation $\mathbb{E}[\mathbb{I}_A|\mathcal{C}]$:

$$\int_B \mathbb{E}[\mathbb{I}_A|\mathcal{C}](\omega)\mathbb{P}(d\omega) \overset{B\in\mathcal{C}}{=} \int_B \mathbb{I}_A \mathbb{P}(d\omega) = \mathbb{P}(A \cap B) \tag{3.12}$$

Since conditional expectations are unique up to a null set, we can always choose $g = \mathbb{E}[\mathbb{I}_A|\mathcal{C}]$.

3.4 Definition. Let $A \in \mathcal{A}$, and let $\mathcal{C}$ be a sub-σ-algebra of $\mathcal{A}$. Then

$$\mathbb{P}[A|\mathcal{C}](\omega) := \mathbb{E}[\mathbb{I}_A|\mathcal{C}](\omega) \ . \tag{3.13}$$

is called a **conditional probability** of A given $\mathcal{C}$. ◁

After all the abstractions above, that is actually quite intuitive, since the probability of a set A can always be expressed as $\mathbb{P}(A) = \mathbb{E}[\mathbb{I}_A]$.

[1] Recall the rules for integration with respect to an image measure: If μ is a measure on $\mathbf{X}$, T a measurable mapping $\mathbf{X} \to \mathbf{X}'$, and g a positive function on $\mathbf{X}'$, then

$$\int_{A'} g(x')(T(\mu))(dx') = \int_{T^{-1}A'} g \circ T(x)\mu(dx) \qquad \text{for all measurable } A' \subset \mathbf{X}' \ . \tag{3.14}$$

3.3. Conditional distributions given σ-algebras

The definition of a conditional distribution can be extended from random variables to σ-algebras in a precisely analogous manner as that of a conditional probability:

3.5 Definition. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $\mathbf{X}$ a Borel space. Let $X : \Omega \to \mathbf{X}$ be a random variable and $\mathcal{C} \subset \mathcal{A}$ a σ-algebra. Then any $\mathcal{C}$-measurable probability kernel $\mathbf{p} : \Omega \to \mathbf{PM}(\mathbf{X})$ satisfying

$$\mathbb{P}(X^{-1}A \cap B) = \int_B \mathbf{p}(A, \omega)\mathbb{P}(d\omega) \qquad \text{for all } A \in \mathcal{B}(\mathbf{X}), B \in \mathcal{C} \tag{3.15}$$

is called a **conditional distribution** of X given $\mathcal{C}$. ◁

The conditional distribution of X given a random variable Y is hence precisely the special case $\mathcal{C} := \sigma(Y)$. The two cases are indeed equivalent, in the sense that $\mathcal{C}$ can always be obtained as $\sigma(Y)$ for a suitably chosen Y, but it can be useful to make $\mathbf{p}$ independent from Y and whatever space Y takes values in, and work directly in Ω.

3.6 Theorem. *Let X be a random variable on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in a Polish space $\mathbf{X}$. Let $\mathcal{C}$ be a sub-σ-algebra of $\mathcal{A}$. Then the following holds:*

(1) **Existence:** *Then X has a conditional distribution given $\mathcal{C}$, that is, there exists a probability kernel satisfying (3.15).*

(2) **Uniqueness:** *The conditional distribution is unique in the following sense: There exists a null set N such that any two conditional probabilities $\mathbf{p}$ and $\mathbf{p}'$ of X given $\mathcal{C}$ satisfy*

$$\forall \omega \notin N : \qquad \mathbf{p}(A, \omega) = \mathbf{p}'(A, \omega) \qquad \textit{for all } A \in \mathcal{B}(\mathbf{X}) \tag{3.16}$$

◁

Our next task is to prove this result. The idea is simple: If $\mathbf{p}$ exists, it certainly satisfies $\mathbf{p}(A, \omega) = \mathbb{E}[\mathbb{I}_A|\mathcal{C}](\omega)$ $\mathbb{P}$-almost surely. We will therefore start with the quantities $\mathbb{E}[\mathbb{I}_A|\mathcal{C}](\omega)$, for all $A \in \mathcal{B}(\mathbf{X})$ and all ω, and try to assemble them into a suitable function $\mathbf{p}$. The problem is that each function $\mathbb{E}[\mathbb{I}_A|\mathcal{C}]$ is only determined up to a null set, so for each $A \in \mathcal{B}(\mathbf{X})$, we pick up a null set $N_A \subset \Omega$ of exceptions. Unless $\mathbf{X}$ is finite, $\mathcal{B}(\mathbf{X})$ is uncountable, and the null sets N_A aggregate into a set which is in general non-null. This is where the Borel space requirement comes in: It guarantees that $\mathcal{B}(\mathbf{X})$ has a countable generator, and we will show that considering only sets A in this generator suffices to define $\mathbf{p}$.

We will break the proof down into a couple of lemmas. Recall first that an **algebra** in a set Ω is a system of subsets which contains $\varnothing$ and Ω, and is closed under set differences and finite unions.

3.7 Lemma. *If a σ-algebra admits a countable generator, this generator is contained in a countable generator which is an algebra. In particular, the Borel σ-algebra $\mathcal{B}(\mathbf{X})$ of a Polish space $\mathbf{X}$ is generated by a countable algebra.* ◁

Proof. Let $\mathcal{F}$ be a countable generator; in the case of a Polish space, choose $\mathcal{F}$ as a countable base of $\mathbf{X}$. Let $\mathcal{F}'$ consist of $\mathcal{F}$ on all complements of sets in $\mathcal{F}$. Then define $\mathcal{G}$ as the system of all sets G of the form $G = \cup_{i=1}^{n} \cap_{j=1}^{n} F_{ij}$, where $n \in \mathbb{N}$ and $F_{ij} \in \mathcal{F}'$. Then $\mathcal{G}$ is a countable algebra containing $\mathcal{F}$. □

The first aspect of Theorem 3.6 we prove is uniqueness:

3.8 Lemma. *Under the conditions of Theorem 3.6, there is a $\mathbb{P}$ null set N such that* (3.16) *holds for any two conditional distributions* $\mathbf{p}$ *and* $\mathbf{p}'$ *of X given $\mathcal{C}$.* ◁

Proof. Let $\mathcal{G}$ be the countable generator constructed in Lemma 3.7. Since $\mathbf{p}$ and $\mathbf{p}'$ are conditional distributions, we have by definition

$$\mathbf{p}(A,\omega) =_{\text{a.s.}} \mathbb{E}[\mathbb{I}_A|\mathcal{C}](\omega) =_{\text{a.s.}} \mathbf{p}'(A,\omega) \tag{3.17}$$

for each $A \in \mathcal{G}$. Note that there is a null set N_A of exceptions for each A, and hence an overall set $N := \cup_{A\in\mathcal{G}} N_A$ of exceptions. Since $\mathcal{G}$ is countable, N is again a null set. We have

$$\mathbf{p}(A,\omega) = \mathbf{p}'(A,\omega) \qquad \text{for all } \omega \notin N \tag{3.18}$$

for all $A \in \mathcal{G}$. By definition, $\mathbf{p}(\bullet,\omega)$ is a measure for each ω. Since $\mathcal{G}$ generates $\mathcal{B}(\mathbf{X})$ and is an algebra (and hence closed under finite intersections), the values this measure completely determine it on all of $\mathcal{B}(\mathbf{X})$ [Probability I, Corollary 6.1]. □

To proof existence in Theorem 3.6, we recall a result which you have proved in Homework 3: Recall that a family $\mathcal{K}$ of a sets is called a **compact class** if it has the finite intersection property, i.e. if every sequence (K_n) of sets in $\mathcal{K}$ satisfies

$$\begin{gathered}\text{every finite subset of sets } K_n \text{ in } (K_n) \text{ has non-empty intersection}\\ \Downarrow \\ (K_n) \text{ has non-empty intersection} \quad .\end{gathered} \tag{3.19}$$

A set function $\mu : \mathcal{G} \to \mathbb{R}_+$ is called **tight** with respect to a compact class $\mathcal{K} \subset \mathcal{G}$ if

$$\mu(A) = \sup\{\mu(K) | K \in \mathcal{K} \text{ and } K \subset A\} \qquad \text{for all } A \in \mathcal{G} \; .$$

3.9 Lemma. *Let $\mathcal{G}$ be an algebra and let $\mu : \mathcal{G} \to [0,\infty)$ be a finitely additive set function on $\mathcal{G}$, with $\mu(G) < \infty$ for all $G \in \mathcal{G}$ and $\mu(\varnothing) = 0$. If μ is tight with respect to some compact class $\mathcal{K} \subset \mathcal{G}$, then it is σ-additive on $\mathcal{G}$.* ◁

The statement is slightly more general than in the homework, since we only require $\mathcal{G}$ to be an algebra, not a σ-algebra. If you cannot find your homework solution anymore, see [A&B, Theorem 10.13] for a proof.

Proof of Theorem 3.6. *Step 1: Defining a candidate object.* Let $\mathcal{G}$ again be the countable generator of $\mathcal{B}(\mathbf{X})$ constructed in Lemma 3.7. Let $P_{\text{X}} := X(\mathbb{P})$ be the law of X. Since P_{X} is a probability measure on a Polish space, it is tight by Theorem 2.49. Let (G_i) be an enumeration of the sets in $\mathcal{G}$. By tightness, there exists for each G_i an increasing sequence of compact sets $K_{i1} \subset K_{i2} \subset \dots$ in $\mathbf{X}$ such that

$$P_{\text{X}}(G_i) = \sup_{j\in\mathbb{N}} P_{\text{X}}(K_{ij}) \; . \tag{3.20}$$

Define $\mathcal{K} = \{K_{ij} | i,j \in \mathbb{N}\}$. Clearly, $\mathcal{K}$ is countable, and by Lemma 3.7, there is hence a countable generator $\mathcal{G}^*$ which contains both $\mathcal{G}$ and $\mathcal{K}$ and forms an algebra. Now define a function on $\mathcal{G}^* \times \Omega$ as

$$\mathbf{q}(G,\omega) := \mathbb{E}[\mathbb{I}_G(X(\omega))|\mathcal{C}] \qquad \text{for all } G \in \mathcal{G}, \omega \in \Omega \; . \tag{3.21}$$

Note this implies $\omega \mapsto \mathbf{q}(G,\omega)$ is measurable for each $G \in \mathcal{G}^*$.

Step 2: $A \mapsto \mathbf{q}(A, \omega)$ is a probability measure. It is straightforward to deduce from the definition (3.21) that $G \mapsto \mathbf{q}(G, \omega)$ satisfies both $\mathbf{q}(\mathbf{X}, \omega) = 1$ and $\mathbf{q}(\varnothing, \omega) = 0$ $\mathbb{P}$-a.s. and is finitely additive almost surely, i.e.

$$\sum_{i=1}^{n} \mathbf{q}(G_i, \omega) = \mathbf{q}(\cup_{i=1}^{n} G_i, \omega) \qquad \mathbb{P}\text{-a.s.} \tag{3.22}$$

for any finite collection $G_1, \ldots, G_n$ of disjoint sets in $\mathcal{G}^*$. For each such collection, we pick up a null set of exceptions, plus one null set each for $\varnothing$ and $\mathbf{X}$. Since $\mathcal{G}^*$ is countable, there are only countably many subsets of finite sets, so overall we obtain a null set $N_{\mathbf{q}}$, with the property that $G \mapsto \mathbf{q}(G, \omega)$ is a finitely additive probability measure on $\mathcal{G}^*$ for all $\omega \notin N_{\mathbf{q}}$. In particular, it satisfies the conditions of Lemma 3.9 for $\omega \notin N_{\mathbf{q}}$.

Now consider the set $\mathcal{K}$: The compact sets in $\mathcal{B}(\mathbf{X})$ form a compact class. Clearly, any subset of a compact class is a compact class, so $\mathcal{K}$ is a compact class contained in $\mathcal{G}^*$. Hence, we will show that $\mathbf{q}(\bullet, \omega)$ is tight on $\mathcal{G}^*$ with respect to $\mathcal{K}$ for all ω outside a null set; countable additivity then follows by Lemma 3.9.

First, let G_i be a set in the original generator $\mathcal{G}$. Then by construction of the sets K_{ij}, we have $\mathbb{I}_{K_{ij}} \to \mathbb{I}_{G_i}$ P_{X}-a.s. Hence, we have

$$\mathbf{q}(G_i, \omega) =_{\text{a.s.}} \mathbb{E}[\mathbb{I}_{G_i}(X(\omega)|\mathcal{C}] =_{\text{a.s.}} \sup_j \mathbb{E}[\mathbb{I}_{K_{ij}}(X(\omega)|\mathcal{C}] =_{\text{a.s.}} \sup_j \mathbf{q}(K_{ij}, \omega) \ . \tag{3.23}$$

Thus, $\mathbf{q}(G, \omega)$ can be approximated to arbitrary precision from within by sets in $\mathcal{K}$ whenever $G \in \mathcal{G}$. Trivially, the same holds if $G \in \mathcal{K}$. Each such approximation of a set G comes with a null set of exceptions. Since both $\mathcal{G}$ and $\mathcal{K}$ are countable, we obtain an overall null set $N_{\mathcal{K},\mathcal{G}}$. Now recall the construction of $\mathcal{G}^*$ from $\mathcal{G} \cup \mathcal{K}$ using Lemma 3.7: If $\mathbf{q}(F_{ij}, \omega)$ can be approximated to arbitrary precision for finitely many sets F_{ij} in $\mathcal{G} \cup \mathcal{K}$, then obviously also $\mathbf{q}(G, \omega)$ for $G = \cup_{i=1}^{n} \cap_{j=1}^{n} F_{ij}$. Since there are only countably many such combinations, we obtain another null set of exceptions. This null set, and the null sets $N_{\mathcal{K},\mathcal{G}}$ and $N_{\mathbf{q}}$ above aggregate into an overall null set N. Hence, $\mathbf{q}(G, \omega)$ is tight on $\mathcal{G}^*$ with respect to $\mathcal{K}$, and by Lemma 3.9, it is countably additive for every $\omega \notin N$. To ensure we obtain a valid probability kernel, we define $\mathbf{q}(A, \omega) := \mathbb{I}_A(X(\omega))$ whenever $\omega \in N$.

Step 3: $\omega \mapsto \mathbf{q}(A, \omega)$ is measurable. What remains to be shown is that $\omega \mapsto \mathbf{q}(A, \omega)$ is measurable for all $A \in \mathcal{B}(\mathbf{X})$. We already know by construction that this is true for $A \in \mathcal{G}^*$. Since $\mathcal{G}^*$ is a generator and an algebra, measurability on all of $\mathcal{B}(\mathbf{X})$ follows using the monotone class theorem. □

3.4. Working with pairs of random variables

In this section, we collect some tools that help us handle pairs (X, Y) of random variables. Recall that is makes a substantial difference whether we know that X and Y satisfy

$$\text{(i)}\ X \overset{\text{d}}{=} Y \qquad \text{or} \qquad \text{(ii)}\ X =_{\text{a.s.}} Y \ .$$

Statement (ii) says that $X(\omega) = Y(\omega)$ for all ω outside a null set. That is, if we think of the variables as mappings $X : \Omega \to \mathbf{X}$ and $Y : \Omega \to \mathbf{X}$, these two mappings are identical, except possibly on a set of measure zero. On the other hand, (i) only states that X and Y have the same distribution, i.e. that the measures $X(\mathbb{P})$ and

$Y(\mathbb{P})$ put the same amount of mass on each measurable set, which is a much weaker statement.

3.10 Example. The difference between (i) and (ii) produces additional fallout when we consider pairs of random variables: If X, X' and Y are random variables,

$$X \overset{\mathrm{d}}{=} X' \qquad \text{does } \textit{not} \text{ imply} \qquad (X,Y) \overset{\mathrm{d}}{=} (X',Y)\ . \tag{3.24}$$

The joint distribution of X and Y is the distribution of the random variable $\omega \to (X(\omega), Y(\omega))$. Even if X and X' are identically distributed, it is perfectly possible that $X(\omega) \neq X'(\omega)$ for all ω, and so $(X'(\omega), Y(\omega))$ can have a different distribution than $(X(\omega), Y(\omega))$. ◁

3.11 Exercise. Elementary, but well worth doing: Give a counterexample. Choose $\Omega = [0,1]$ and $\mathbb{P}$ as Lebesgue measure. Specify random variables X, X' and Y on Ω, say with values in $\mathbb{R}$, such that X and X' are identically distributed, but (X',Y) and (X,Y) are not. ◁

We have defined conditional probability of a random variable X in terms of a conditional expectation. For our first result in this section, we note that the converse is also possible: If X takes values in a Borel space and $\mathbf{p}$ is a conditional distribution of X given a σ-algebra $\mathcal{C}$, we can compute $\mathbb{E}[X|\mathcal{C}]$ as

$$\mathbb{E}[X|\mathcal{C}](\omega) =_{\text{a.s.}} \int_{\mathbf{X}} x\mathbf{p}(dx,\omega)\ . \tag{3.25}$$

The next result generalizes this to functions of two arguments, and can hence be read as a Fubini theorem for conditional distributions.

3.12 Lemma. *If X is a random variable with values in a Borel space $\mathbf{X}$, and a probability kernel $\mathbf{p}$ is a version of the conditional distribution $\mathbb{P}[X \in \bullet|Y=y]$, then*

$$\mathbb{E}[f(X,Y)|Y=y] = \int_{\mathbf{X}} f(x,y)\mathbf{p}(dx,y) \qquad P_{\mathrm{Y}}\text{-}a.s. \tag{3.26}$$

◁

The proof is basically an application of the monotone class theorem, and I will skip it here. If you feel curious:

Proof. See e.g. [K, Theorem 6.4]. □

The next theorem does perhaps not look particularly thrilling at first glance, but it is in my experience one of the most useful tools for problems involving conditionals:

3.13 Theorem. *Let $\mathbf{p} : \mathcal{Y} \to \mathbf{PM}(\mathbf{X})$ be a probability kernel, where $(\mathcal{Y}, \mathcal{A}_{\mathrm{Y}})$ is a measurable space and $\mathbf{X}$ a Borel space. Then there exists a measurable function $f : [0,1] \times \mathcal{Y} \to \mathbf{X}$ such that*

$$\mathbb{P}\{f(U,y) \in A\} = \mathbf{p}(A,y) \qquad \textit{for } U \sim \textit{Uniform}(0,1)\ . \tag{3.27}$$

◁

Proof. We begin by invoking the Borel isomorphism theorem (Theorem 2.52), by which it is sufficient to prove the result for $\mathbf{X} = [0,1]$. To establish the result, we construct f explicitly: Define

$$f(u,y) := \sup\{x \in [0,1] \,|\, \mathbf{p}([0,x],y) < u\}\ . \tag{3.28}$$

We have to show that (i) f is *jointly* measurable in (u, y), and (ii) that it satisfies (3.27). Regarding measurability, we note that:

(1) Since f maps into $\mathbf{X} = [0,1]$, and the Borel sets in $[0,1]$ are generated by the half-open intervals, it suffices to show that the preimages of all sets $(a,b] \subset [0,1]$ are measurable.

(2) The preimage of the interval $[0,x]$ is

$$f^{-1}[0,x] = \{(u,y) | u \leq \mathbf{p}([0,x],y)\} \ .$$

If (u,y) is contained in $f^{-1}[0,x]$, it is also in $f^{-1}[0,x']$ for all $x \leq x'$, which means $f^{-1}[0,x] \subset f^{-1}[0,x']$.

(3) We therefore have

$$f^{-1}(a,b] = f^{-1}([0,b] \setminus [0,a]) = f^{-1}[0,b] \setminus f^{-1}[0,a] \ ,$$

It is hence sufficient to show the set $f^{-1}[0,x]$ is measurable for each $x \in [0,1]$.

(4) The function $y \mapsto \mathbf{p}([0,x],y)$ is measurable by definition of probability kernels.

(5) The supremum of a sequence of measurable functions is measurable. The supremum in (3.28) remains unchanged if we restrict $x \in [0,1]$ to rational values, which indeed yields a sequence of measurable functions $y \mapsto \mathbf{p}([0,x_n],y)$.

(6) The set

$$f^{-1}[0,x] = \{(u,y)|\mathbf{p}([0,x],y) < u\} = \{(u,y)|\mathbf{p}([0,x],y) - u < 0\}$$

is therefore measurable, since, additionally, both differences of measurable functions and the set $\{s \in \mathbb{R} | s < 0\}$ are measurable.

Thus, f is jointly measurable. Now suppose U is a Uniform$(0,1)$ random variable. Then for any y,

$$\mathbb{P}\{f(U,y) \leq x\} = \mathbb{P}\{u \leq \mathbf{p}([0,x],y)\} = \mathbf{p}([0,x],y)$$

for any $x \in [0,1]$. Hence, $x \mapsto \mathbf{p}([0,x],y)$ is a cumulative distribution function for $f(U,y)$, which (since $\mathbf{X} = [0,1]$) implies $f(U,y)$ has law $\mathbf{p}(\bullet, y)$. □

A neat application of this theorem is the following: In modeling problems, especially in applied statistics or in machine learning, you will frequently encounter notation of the form $X|Y$, meant to be read as "the random variable X conditionally on Y". For instance, one could write

$$X|(Y = y) \sim P \tag{3.29}$$

to state that, for a fixed value y, the conditional distribution of X given $Y = y$ is the measure P. We already know by the existence theorem for conditional distributions that we can make this assumption if X is a Borel space (choose $P := \mathbf{p}(\bullet, y)$). However, the existence theorem only tells us that we can choose *some* random variable X^y for every y which is distributed according to $\mathbf{p}(\bullet, y)$. It does *not* tell us that X^y depends measurably on y—which means, for example, that we cannot think of y as being random (we cannot substitute Y for y and write X^Y), and that apparently simple statements like $X|(Y = y) =_{\text{a.s.}} y$ are not well-defined. Using Theorem 3.13, we can give a precise definition of a random variable X^y parametrized by y which does depend measurably on y. We can think of this X^y as a conditional random variable:

3.14 Corollary. *Let X and Y be two random variables with values in Borel spaces $\mathbf{X}$ and $\mathbf{Y}$. Then X and Y can be represented as random variables on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that there exists a measurable function $X' : \Omega \times \mathbf{Y} \to \mathbf{X}$ with*

$$\mathbb{P}\{X'(\omega, y) \in A\} = \mathbf{p}(A, y) \qquad P_{\mathrm{Y}}\text{-a.s.} \; . \tag{3.30}$$

$\triangleleft$

To obtain the form X^y above, set $X^y(\omega) := X'(\omega, y)$, so

$$\mathcal{L}(X^y) = \mathbb{P}[X \in \bullet | Y = y] \; . \tag{3.31}$$

Proof. We can always assume that any countable collection of random variables with values in Borel spaces—in this case X, Y, and the uniform variable U in Theorem 3.13—are defined on a joint probability space. Then choose f as the function in Theorem 3.13, and define $X'(\omega, y) := f(U(\omega), y)$. □

The next result concerns distributional equations; in this case, pairs of random variables that satisfy an equality in distribution.

3.15 Theorem. *Let X and Y be random variables with values in Borel spaces $\mathbf{X}$ and $\mathbf{Y}$. Then for any random variable X' with the same distribution as X, there is a measurable mapping $f : \mathbf{X} \times [0,1] \to \mathbf{Y}$ such that*

$$(X', f(X', U)) \overset{d}{=} (X, Y) \tag{3.32}$$

where $U \sim \text{Uniform}(0,1)$ is independent of X'. $\triangleleft$

In simpler terms, the theorem says: Given two random variables $X \overset{\mathrm{d}}{=} X'$ with values in $\mathbf{X}$ and a random variable Y with values in $\mathbf{Y}$, we can always find another $\mathbf{Y}$-valued random variable Y' satisfying the distributional equation

$$(X', Y') \overset{\mathrm{d}}{=} (X, Y) \; . \tag{3.33}$$

Compare this to Example 3.10 at the beginning of this section.

Proof of Theorem 3.15. Since $\mathbf{Y}$ is Borel, there exists a probability kernel $\mathbf{p}$ with

$$\mathbf{p}(A, x) =_{\text{a.s.}} \mathbb{P}[Y \in A | X = x] \tag{3.34}$$

by Theorem 3.2. By Theorem 3.13 we can hence choose a function $f : [0,1] \times \mathbf{X} \to \mathbf{Y}$ with $\mathcal{L}(f(U, x) = \mathbf{p}(\bullet, x)$ for all x. Let $g(x, y)$ be a real-valued, measurable function. Then

$$\begin{aligned} \mathbb{E}[g(X', f(U, X'))] &= \mathbb{E}_{X'}\left[\mathbb{E}_{U|X'=x'}[g(X', f(X', U))]\right] \\ &\overset{\text{Lemma 3.12}}{=} \mathbb{E}_{X'}\Big[\int_{\mathbf{Y}} g(X', y)\mathbf{p}(dy, X')\Big] \; . \end{aligned} \tag{3.35}$$

This is now simply the expectation of a function of X', and since $X' \overset{\mathrm{d}}{=} X$, we have

$$\begin{aligned} \mathbb{E}_{X'}\Big[\int_{\mathbf{Y}} g(X', y)\mathbf{p}(dy, X')\Big] &= \mathbb{E}_{X}\Big[\int_{\mathbf{Y}} g(X, y)\mathbf{p}(dy, X)\Big] \\ &= \int_{\mathbf{X}}\int_{\mathbf{Y}} g(x, y)\mathbf{p}(dy, x)P_{\mathrm{X}}(dx) = \mathbb{E}[g(X, Y)] \; . \end{aligned} \tag{3.36}$$

Thus, $\mathbb{E}[g(X', f(X', U))] = \mathbb{E}[g(X, Y)]$ for every measurable function, which means $(X', f(X', U))$ and (X, Y) are identically distributed. □

As one consequence of Theorem 3.15, we can replace a pair of random variables that satisfy a relationship in distribution by a pair which satisfy the same relationship almost surely, without changing the marginal distributions of the variables:

3.16 Theorem. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be Borel spaces and* $f : \mathbf{X} \to \mathbf{Y}$ *measurable. Let* X *and* Y *be random variables with values in* $\mathbf{X}$ *and* $\mathbf{Y}$ *such that* $Y \overset{d}{=} f(X)$. *Then there exists a* $\mathbf{X}$*-valued random variable* X' *such that*

$$X' \overset{d}{=} X \qquad \text{and} \qquad Y =_{\text{a.s.}} f(X') \; . \tag{3.37}$$

◁

Note that, although the *marginal* distributions of X and Y do not change when substituting the pair (X', Y), the *joint* distribution changes.

To prove this result we have to compute the probability of events of the form $\{Y = Y'\}$, which means we have to make sure such events are measurable. If Y and Y' are random variables with values in the same space $\mathbf{Y}$, that is the case if the set $\{(y, y) | y \in \mathbf{Y}\}$, also called the **diagonal** of $\mathbf{Y}^2$, is is measurable in $\mathbf{Y}^2$, which we cannot generally assume in a measurable space $\mathbf{Y}$.

This problem is closely related to the problem of joint measurability of functions: Recall that a function $f(x, y)$ on $\mathbf{X} \times \mathbf{Y}$ of two arguments is called jointly measurable if it is measurable in the product σ-algebra on $\mathbf{X} \times \mathbf{Y}$. If we only know it is measurable in each argument (if each of the functions $x \mapsto f(x, y)$ and $y \mapsto f(x, y)$ is measurable), we cannot conclude that the function $(x, y) \mapsto f(x, y)$ is jointly measurable. One widely applicable criterion for joint measurability is the following:

3.17 Fact. Let $f : \mathcal{X} \times \mathbf{Y} \to \mathbf{Z}$ be measurable and its first and continuous in its second argument, where $\mathcal{X}$ is a measurable space, $\mathbf{Y}$ separable metrizable and $\mathbf{Z}$ metrizable. Then f is jointly measurable. ◁

Functions of two arguments that are measurable in one and continuous in the other are also known as **Carathéodory functions**.[2] From the joint measurability of Carathéodory functions, we can conclude:

3.18 Lemma. *If* $\mathbf{Y}$ *is a Borel space, the diagonal in* $\mathbf{Y}^2$ *is measurable.* ◁

Proof. Since $\mathbf{Y}$ is Borel, there exists a Polish topology which generates the Borel sets of $\mathbf{Y}$, and hence a compatible metric $d : \mathbf{Y} \times \mathbf{Y} \to [0, \infty)$. The metric d is continuous (and hence measurable) in each argument, and thus a Carathéodory function. Since the diagonal in $\mathbf{Y}^2$ is precisely the set $d^{-1}\{0\}$, it is measurable. □

Proof of Theorem 3.16. By Theorem 3.15, there exists a random variable X' in $\mathbf{X}$ such that

$$(X', Y) \overset{\mathrm{d}}{=} (X, f(X)) \; , \tag{3.38}$$

which implies that also marginally $X \overset{\mathrm{d}}{=} X'$, which establishes the first claim in (3.37). Applying f on both sides of (3.38) yields $(f(X'), Y) \overset{\mathrm{d}}{=} (f(X), f(X))$. Since the diagonal of $\mathbf{Y} \times \mathbf{Y}$ is measurable, we can compute

$$\mathbb{P}\{f(X') = Y\} = \mathbb{P}\{f(X) = f(X)\} = 1 \; , \tag{3.39}$$

which implies $f(X') = Y$ almost surely. □

[2] If you would like to know more about Carathéodory functions, see [A&B, §4.10].

3.5. Conditional independence

If the statisticians among you feel that the previous section was a little far out there, we now come to a concept of more obvious relevance to statistics, conditional independence: The fundamental modeling assumption in most Bayesian models is that there exists a random variable (the parameter) which renders observations conditionally independent; graphical models are representations of conditional dependence and independence relations; etc.

We again assume that all random variables are defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Recall the definition of independence from [Probability I]:

- Two events $A, B \in \mathcal{A}$ are independent if the probability of them occurring simultaneously factorizes, i.e. if
$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)\ . \tag{3.40}$$
- Two σ-algebras $\mathcal{C}_1, \mathcal{C}_2 \subset \mathcal{A}$ are independent if every possible pair of events $A \in \mathcal{C}_1$ and $B \in \mathcal{C}_2$ is independent.
- Two random variables X and Y on Ω are independent if the σ-algebras $\sigma(X)$ and $\sigma(Y)$ are independent.

Conditional independence is defined in an analogous manner, by substituting $\mathbb{P}$ by a conditional probability:

3.19 Definition. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $\mathcal{G} \subset \mathcal{A}$ a σ-algebra. Then two events $A, B \in \mathcal{A}$ are **conditionally independent** given $\mathcal{G}$ if
$$\mathbb{P}(A \cap B|\mathcal{G}) = \mathbb{P}(A|\mathcal{G})\mathbb{P}(B|\mathcal{G})\ . \tag{3.41}$$
Similarly, σ-algebras $\mathcal{C}_1, \ldots, \mathcal{C}_n \subset \mathcal{A}$ are conditionally independent given $\mathcal{G}$ if
$$\mathbb{P}\Big[\bigcap\nolimits_{k\leq n} A_k \,\Big|\, \mathcal{G}\Big] = \prod_{k\leq n} \mathbb{P}[A_k|\mathcal{G}] \qquad \text{for all } A_k \in \mathcal{C}_k, k = 1, \ldots, n\ . \tag{3.42}$$
◁

If two σ-algebras $\mathcal{C}_1$ and $\mathcal{C}_2$ are conditionally independent given $\mathcal{G}$, we write
$$\mathcal{C}_1 \perp\!\!\!\perp_{\mathcal{G}} \mathcal{C}_2\ . \tag{3.43}$$
Conditional independence of random variables is defined in terms of the σ-algebras they generate, as
$$X \perp\!\!\!\perp_Z Y \qquad \text{iff} \qquad \sigma(X) \perp\!\!\!\perp_{\sigma(Z)} \sigma(Y)\ . \tag{3.44}$$

3.20 Proposition. *Let $\mathcal{C}$, $\mathcal{F}$ and $\mathcal{G}$ be σ-algebras. Then*
$$\mathcal{C} \perp\!\!\!\perp_{\mathcal{G}} \mathcal{F} \qquad \Leftrightarrow \qquad \mathbb{P}[C|\mathcal{F},\mathcal{G}] =_{\text{a.s.}} \mathbb{P}[C|\mathcal{G}] \textit{ for all } C \in \mathcal{C}\ . \tag{3.45}$$
◁

In terms of random variables X, Y and Z, this means
$$X \perp\!\!\!\perp_Z Y \qquad \Leftrightarrow \qquad \mathbb{P}[X \in A|Y, Z] =_{\text{a.s.}} \mathbb{P}[X \in A|Z] \quad P_{\mathrm{Z}}\text{-a.s.} \tag{3.46}$$

For the proof, let me remind you of two fundamental properties of conditional expectations, which can both be found in [Probability I, Chapter 23]: One is the fact that a $\mathcal{C}$-measurable random variable can be pulled out of any conditional expectation given $\mathcal{C}$, to wit
$$\mathbb{E}[XY|\mathcal{C}] =_{\text{a.s.}} X\,\mathbb{E}[Y|\mathcal{C}] \qquad \text{if } X \text{ is } \mathcal{C}\text{-measurable.} \tag{3.47}$$

The second one, also known as the **law of total probability**, says that conditioning first on $\mathcal{C}$ and then on a coarser σ-algebra $\mathcal{D}$ amounts to conditioning only on $\mathcal{D}$:

$$\mathbb{E}\big[\mathbb{E}[X|\mathcal{C}]\big|\mathcal{D}\big] =_{\text{a.s.}} \mathbb{E}[X|\mathcal{D}] \qquad \text{if } \mathcal{C} \subset \mathcal{D}\ . \tag{3.48}$$

Since a conditional probability can be represented as a conditional expectation of an indicator function, we can in particular expand $\mathbb{P}[A|\mathcal{D}]$ as

$$\mathbb{P}[A|\mathcal{D}] =_{\text{a.s.}} \mathbb{E}\big[\mathbb{P}[A|\mathcal{C}]\big|\mathcal{D}\big] \qquad \text{if } \mathcal{C} \subset \mathcal{D}\ , \tag{3.49}$$

whenever that happens to be useful in a computation. Recall also that the defining property of conditional expectation $\mathbb{E}[X|\mathcal{C}]$, namely $\int_C \mathbb{E}[X|\mathcal{C}]d\mathbb{P} = \int_C Xd\mathbb{P}$ for all $C \in \mathcal{C}$, can be written as

$$\mathbb{E}\big[\mathbb{E}[X|\mathcal{C}] \cdot \mathbb{I}_C\big] = \mathbb{E}[X \cdot \mathbb{I}_C] \qquad \text{for all } C \in \mathcal{C}\ . \tag{3.50}$$

(Indeed, this is how the conditional expectation was defined in [Probability I].)

Proof of Proposition 3.20.
Step 1: "⇐". Suppose the right-hand side of (3.45) holds. We have to show that the joint conditional distribution of $C \cap F$ factorizes for all $C \in \mathcal{C}$ and $F \in \mathcal{F}$.

$$\begin{aligned}\mathbb{P}[F \cap C|\mathcal{G}] &\overset{(3.48)}{=} \mathbb{E}\big[\mathbb{P}[F \cap C|\sigma(\mathcal{F} \cup \mathcal{G})]\big|\mathcal{G}\big]\\ &\overset{(3.47)}{=} \mathbb{E}\big[\mathbb{P}[C|\sigma(\mathcal{F} \cup \mathcal{G})] \cdot \mathbb{I}_F\big|\mathcal{G}\big]\\ &\overset{(3.45)}{=} \mathbb{E}\big[\mathbb{P}[C|\mathcal{G}] \cdot \mathbb{I}_F\big|\mathcal{G}\big]\\ &\overset{(3.47)}{=} \mathbb{P}[C|\mathcal{G}] \cdot \mathbb{E}[\mathbb{I}_F|\mathcal{G}] = \mathbb{P}[C|\mathcal{G}] \cdot \mathbb{P}[F|\mathcal{G}]\ .\end{aligned} \tag{3.51}$$

Step 2: "⇒". Now assume $\mathcal{C} \perp\!\!\!\perp_{\mathcal{G}} \mathcal{F}$ holds. Let us clarify what we have to show: On the right-hand side of (3.45), think of $\mathbb{P}[C|\mathcal{G}]$ as a random variable. Then $\mathbb{P}[C|\mathcal{G},\mathcal{F}]$ is its conditional expectation $\mathbb{E}[\mathbb{P}[C|\mathcal{G}]|\mathcal{F}]$. To show that the two are equal almost surely, we have to show that, for any $A \in \sigma(\mathcal{F} \cup \mathcal{G})$,

$$\int_A \mathbb{P}[C|\mathcal{G}]d\mathbb{P} = \int_A \mathbb{P}[C|\mathcal{G},\mathcal{F}]d\mathbb{P} = \mathbb{P}(A \cap C)\ , \tag{3.52}$$

where the second equality is by definition of conditional probabilities. Suppose first that A is in particular of the form $F \cap G$ for some $F \in \mathcal{F}$ and $G \in \mathcal{G}$. Then

$$\begin{aligned}\int_{F \cap G} \mathbb{P}[C|\mathcal{G}]d\mathbb{P} &= \mathbb{E}\big[\mathbb{P}[C|\mathcal{G}] \cdot \mathbb{I}_F\mathbb{I}_G\big]\\ &\overset{(3.50)}{=} \mathbb{E}\big[\mathbb{P}[C|\mathcal{G}]\mathbb{P}[F|\mathcal{G}] \cdot \mathbb{I}_G\big]\\ &\overset{\mathcal{C}\perp\!\!\!\perp_{\mathcal{G}}\mathcal{F}}{=} \mathbb{E}\big[\mathbb{P}[C \cap F|\mathcal{G}] \cdot \mathbb{I}_G\big]\\ &\overset{(3.47)}{=} \mathbb{E}\big[\mathbb{P}[C \cap F \cap G|\mathcal{G}]\big]\\ &\overset{(3.48)}{=} \mathbb{P}(C \cap F \cap G)\ .\end{aligned} \tag{3.53}$$

All that is left to do is to generalize from the case $A = F \cap G$ to any $A \in \mathcal{A}$, which is an application of the monotone class theorem. □

One of the basic rules used in everyday computations involving conditional independence is the following chain rule:

3.21 Proposition [Chain rule for conditional independence]. *If $\mathcal{C}$, $\mathcal{G}$, and $\mathcal{F}_1, \mathcal{F}_2, \ldots$ are σ-algebras, then*

$$\mathcal{C} \perp\!\!\!\perp_{\mathcal{G}} \sigma(\mathcal{F}_1, \mathcal{F}_2, \ldots) \qquad \Leftrightarrow \qquad \mathcal{C} \perp\!\!\!\perp_{\sigma(\mathcal{G},\mathcal{F}_1,\ldots,\mathcal{F}_n)} \mathcal{F}_{n+1} \text{ for all } n \in \mathbb{N} \ . \tag{3.54}$$

In particular,

$$\mathcal{C} \perp\!\!\!\perp_{\mathcal{G}} (\mathcal{F}, \mathcal{F}') \qquad \text{if and only if} \qquad \mathcal{C} \perp\!\!\!\perp_{\mathcal{G}} \mathcal{F} \text{ and } \mathcal{C} \perp\!\!\!\perp_{\mathcal{G},\mathcal{F}} \mathcal{F}' \ . \tag{3.55}$$

◁

PROOF. Homework. □

Aside from Proposition 3.20, another way to verify conditional independence is by means of an independent randomization.

3.22 Theorem [Randomization criterion for conditional independence]. *Let X, Y and Z be random variables with values in Borel spaces $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$. Then $X \perp\!\!\!\perp_Y Z$ iff $X =_{\text{a.s.}} f(Y, U)$ for some measurable function $f : \mathbf{Y} \times [0, 1] \to \mathbf{X}$ and a uniform variable $U \sim$ Uniform$(0, 1)$ which is independent of (Y, Z).* ◁

PROOF. By application of Theorem 3.13. See [K, Proposition 6.13]. □

3.6. Application: Sufficient statistics

This section is optional.

One of the many applications of conditioning and conditional independence in statistics is sufficiency. Recall that a **statistical model** on a sample space $\mathbf{X}$ is any set $\mathcal{P} \subset \mathbf{PM}(\mathbf{X})$ of probability measures on $\mathbf{X}$. A **statistic** is a measurable function $S : \mathbf{X} \to \mathbf{S}$ into some (typically Polish) space $\mathbf{S}$. A statistic S is called **sufficient** for a model $\mathcal{P}$ if all measures in the model have the same conditional distribution given S. As we have seen above, this means for a Borel sample space $\mathbf{X}$: The statistic is sufficient if there exists a probability kernel $\mathbf{p} : \mathbf{S} \to \mathbf{PM}(\mathbf{X})$ such that

$$P[\bullet|S = s] = \mathbf{p}(\bullet, s) \qquad \text{for all } P \in \mathcal{P} \ . \tag{3.56}$$

If we parametrize the model by a parameter θ with values in a parameter space $\mathbf{T}$, i.e. if $\mathcal{P} = \{P_\theta | \theta \in \mathbf{T}\}$, the equation above takes the (perhaps more familiar) form

$$P_\theta[\bullet|S = s] = \mathbf{p}(\bullet, s) \qquad \text{for all } \theta \in \mathbf{T} \ . \tag{3.57}$$

As we have learned in this chapter, we can not just condition on random variables (such as the statistics S), but more generally on σ-algebras. Therefore, a σ-algebra $\mathcal{S} \subset \mathcal{A}$ in the underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is called **sufficient** for the model $\mathcal{P}$ if there is a probability kernel $\Omega \to \mathbf{PM}(\mathbf{X})$

$$P[\bullet|\mathcal{S}](\omega) = \mathbf{p}(\bullet, \omega) \qquad \text{for all } P \in \mathcal{P} \ . \tag{3.58}$$

Some rather surprising things can happen when the spaces and σ-algebras we work with are not benign. One intuitive explanation touted in many textbooks is that σ-algebras represent information. Although this intuition is useful, it can fail dismally if the σ-algebra is not countably generated. In the particular case of sufficiency, consider two σ-algebras $\mathcal{S}$ and $\mathcal{T}$. If $\mathcal{T} \subset \mathcal{S}$, the "information" interpretation tells us that $\mathcal{S}$ in particular contains all information contained in $\mathcal{T}$. If $\mathcal{T}$ is sufficient, we would hence expect $\mathcal{S}$ to be sufficient also. That need not be the case: Classic work by Burkholder constructs an explicit counter-example. However, Burkholder also proved that things do work as expected if the larger σ-algebra $\mathcal{S}$ is countably generated:

3.23 Theorem. *If a σ-algebra $\mathcal{S} \subset \mathcal{A}$ is countably generated and contains a sufficient σ-algebra, then $\mathcal{S}$ is sufficient.* ◁

PROOF. See Burkholder, *Sufficiency in the undominated case*, Ann. Math. Statist., Vol. 32 (1961) pp. 1191-1200, Theorem 5. □

I advocate to think of sufficiency not in terms of the statistic S, but always in terms of the pair $(S, \mathbf{p})$. For example, a fairly common mistake is to assume that, if a statistic S is sufficient for two models $\mathcal{P}$ and $\mathcal{P}'$, then it is also sufficient for $\mathcal{P} \cup \mathcal{P}'$. That is not generally the case, since even if the mapping S in (3.57) is the same for $\mathcal{P}$ and $\mathcal{P}'$, the kernels $\mathbf{p}$ and $\mathbf{p}'$ generally differ. If the two kernels are identical, then S is indeed sufficient for the union. If we specify both S and $\mathbf{p}$, there is a uniquely defined set

$$M(S, \mathbf{p}) := \{P \in \mathbf{PM}(\mathbf{X}) | P[\bullet|S] = \mathbf{p}\} . \tag{3.59}$$

In particular, this is the (uniquely determined) largest model for which the pair $(S, \mathbf{p})$ is sufficient.

3.24 Remark [Sufficiency and symmetry]. The beautiful work of Lauritzen (see e.g. *Extremal families and systems of sufficient statistics*, Springer 1988) shows that $M(S, \mathbf{p})$ is a convex set. Under suitable conditions, it also has much stronger properties: The extreme points of $M(S, \mathbf{p})$ are precisely those measures which are of the form of the form $\mathbf{p}(\bullet, s)$ for some s. Every measure P in $M(S, \mathbf{p})$ has a representation of the form

$$P = \int_{\mathbf{S}} \mathbf{p}(\bullet, s) \nu_{\mathrm{P}}(ds) \tag{3.60}$$

for some probability measure ν_{P} on $\mathbf{S}$. Lauritzen calls the set $\{\mathbf{p}(\bullet, s) | s \in \mathbf{S}\}$ an **extremal family**. For example, the famous representation theorem of de Finetti can be obtained as a special case of (3.60). If S is chosen to take values in a finite-dimensional space, the set of extreme points is an exponential family (and all exponential families can be obtained in this way). The deeper meaning behind this is that sufficiency can be interpreted as a statistical notion of symmetry. ◁

3.7. Conditional densities

Let X and Y be random variables with values in Borel spaces $\mathbf{X}$ and $\mathbf{Y}$. Now choose a σ-finite measure μ on $\mathbf{X}$. Since the conditional probability of X given Y is a probability measure on $\mathbf{X}$ for every $y \in \mathbf{Y}$, we can ask whether it has a density with respect to a suitable measure μ:

3.25 Definition. Let X and Y be random variables, where X takes values in a Borel space $\mathbf{X}$. Let μ be a σ-finite measure on $\mathbf{X}$. Any measurable function p satisfying

$$\mathbb{P}[X \in dx | Y = y] = p(x|y)\mu(dx) \qquad P_{\mathrm{Y}}\text{-a.s.} \tag{3.61}$$

is called a **conditional density** of X given Y. ◁

As a probability measure, each distribution $\mathbb{P}[X \in dx | Y = y]$ is of course absolutely continuous with respect to *some* σ-finite measure, but the question is whether a single μ can be found for all values of y. To ensure that is the case, we formulate a sufficient condition in terms of the joint distribution:

3.26 Theorem. *Let X and Y be random variables with values in Borel spaces $\mathbf{X}$ and $\mathbf{Y}$. Require that there are σ-finite measures μ on $\mathbf{X}$ and ν on $\mathbf{Y}$ such that the joint distribution $P := \mathcal{L}(X, Y)$ satisfies $P \ll \mu \otimes \nu$. Then $\mathbb{P}[X \in dx | Y = y] \ll \mu(dx)$ holds $\mathcal{L}(Y)$-a.s., i.e. the conditional density $p(x|y)$ exists. If we define*

$$p(x, y) := \frac{P(dx \times dy)}{\mu(dx)\nu(dy)} \qquad \textit{and} \qquad f(y) := \int_{\mathbf{X}} p(x, y)\mathbb{P}(X \in dx) \;, \tag{3.62}$$

then $p(x|y)$ is given by

$$p(x|y) = \frac{p(x, y)}{f(y)} \;, \tag{3.63}$$

and f is a density of P_{Y} with respect to ν. ◁

CHAPTER 4

Pushing forward and pulling back

Before moving on to stochastic processes, I will very briefly discuss the concepts of pushforwards and pullbacks of measures. Consider two measures, μ on a space $\mathbf{X}$ and ν on $\mathbf{Y}$. Suppose $\phi : \mathbf{X} \to \mathbf{Y}$ is a mapping, and consider the equation

$$\phi(\mu) = \nu \ . \tag{4.1}$$

Given this equation, we can ask for different types of solutions:

- If μ and ϕ are given, we can ask whether there exists a measure ν satisfying (4.1). That is always the case (provided ϕ is measurable), and ν is then called the *image measure* or the *pushforward* of μ. In probability, it is usually denoted $\phi(\mu)$, or simply $\phi\mu$. In other areas of mathematics, the notation $\phi_{\#}\mu$ is more common.
- If instead ν and ϕ are given, a measure μ satisfying (4.1) is called the *pullback* of ν, denoted $\phi^{\#}\nu$. The pullback need not exist, even if ϕ is measurable.

The results in this chapter show (1) how we integrate with respect to a pushforward and (2) how to guarantee the existence of pullback measures. We will need pullbacks to construct stochastic processes with regular paths; discussing pullbacks in tandem with pushforwards gives me an excuse to state the change of variables theorem, which does not really fit in well anywhere else. First, more formally:

4.1 Definition. Let $(\mathcal{X}, \mathcal{A}_{\mathrm{X}})$ and $(\mathcal{Y}, \mathcal{A}_{\mathrm{Y}})$ be two measurable spaces and $\phi : \mathcal{X} \to \mathcal{Y}$ a measurable mapping.

(1) If μ is a measure on $(\mathcal{X}, \mathcal{A}_{\mathrm{X}})$, the measure $\phi_{\#}\mu$ defined by

$$\phi_{\#}\mu(A) := \mu(\phi^{-1}A) \qquad \text{for } A \in \mathcal{A}_{\mathrm{X}} \tag{4.2}$$

is called the **pushforward** or **image measure** of μ under ϕ.

(2) Let ν be a measure on $(\mathcal{Y}, \mathcal{A}_{\mathrm{Y}})$. If there exists a measure $\phi^{\#}\nu$ satisfying

$$\phi_{\#}(\phi^{\#}\nu) = \nu \ , \tag{4.3}$$

then $\phi^{\#}\nu$ is called the **pullback** of ν under ϕ.

◁

4.1. Integration with respect to image measures

The pushforward $\phi_{\#}\mu$ is well-defined whenever ϕ is measurable. Whenever it exists, we can without further difficulties compute integrals:

4.2 Theorem [Integration with respect to an image measure]. *Let $(\mathcal{X}, \mathcal{A}_{\mathrm{X}})$ and $(\mathcal{Y}, \mathcal{A}_{\mathrm{Y}})$ be measurable spaces, $\phi : \mathcal{X} \to \mathcal{Y}$ measurable, and μ a measure on $\mathcal{X}$. Let g be a measurable, real-valued function on $\mathcal{Y}$. Then g is $\phi_{\#}\mu$-integrable if and* *only if $g \circ \phi$ is μ-integrable, and the integral is given by*

$$\int_B g(y)\phi_{\#}\mu(dy) = \int_{\phi^{-1}B} g \circ \phi(x)\mu(dx) \qquad \textit{for all } B \in \mathcal{A}_{\mathrm{Y}} \; . \tag{4.4}$$

◁

Proof. See e.g. [3, Theorem 16.12]. □

This theorem is useful in proofs and on a conceptual level. When we try to perform specific computations, μ is often defined in terms of a density. The most important case is of course if $\mathbf{X}$ is $\mathbb{R}^d$, and the density is defined with respect to Lebesgue measure λ^d. If ϕ is a mapping from $\mathbb{R}^d$ to itself, we can then ask whether we can directly obtain the image measure $\phi_{\#}\mu$ in terms of its density (again with respect to λ^d). More generally, we can ask: If we transform μ to $\phi_{\#}\mu$, can we express the transformation of the integral $\int f d\mu$ as a transformation of f, rather than of μ as in the previous result? This is indeed possible, provided that $\mathbf{X}$ is Euclidean and ϕ sufficiently smooth. The notion of smoothness we need is the following:

4.3 Definition. A mapping $\phi : V \to W$ between open sets in $\mathbb{R}^d$ is called **diffeomorphism** if it is bijective, continuously differentiable, and if $\mathbf{J}_\phi(x) \neq 0$ for each $x \in V$. ◁

The definition implies in particular that ϕ is a homeomorphism (a continuous bijection with continuous inverse). If ϕ is a diffeomorphism, the requisit transformation of f can be expressed in terms of the derivative of ϕ. Since the domain and range of ϕ are both d-dimensional, the derivative is given by the Jacobian matrix $\mathbf{J}_\phi$:

4.4 Theorem [Change of variables]. *Let $\mathbf{X}$ and $\mathbf{Y}$ be Borel subsets of Euclidean space $\mathbb{R}^d$. Assume there exist open sets $V \subset \mathbf{X}$ and $W \subset \mathbf{Y}$ such that $\lambda(\mathbf{X} \setminus V) = \lambda(\mathbf{Y} \setminus W) = 0$, and let $\phi : V \to W$ be a diffeomorphism. Then for each integrable function f on $\mathbf{Y}$, the function $(f \circ \phi) \cdot |\mathbf{J}_\phi|$ on A is integrable, an*

$$\int_{\mathbf{Y}} f d\lambda^d = \int_{\mathbf{X}} (f \circ \phi) \cdot |\mathbf{J}_\phi| d\lambda^d \; . \tag{4.5}$$

◁

In the one-dimensional case $\mathbb{R}^d = \mathbb{R}$, this reduces to the substitution-of-variables rule of highschool calculus fame. A common application in probability and statistics is the problem of finding the density of a transformed random variable: Suppose Y is a real-valued random variable with known density p. We want to determine the density of p_x of $X = \tau(Y)$, for some invertible and sufficiently smooth $\tau : \mathbb{R} \to \mathbb{R}$. In this case, the function f in Theorem 4.4 is the density p. We hence have to use the theorem "backwards" (since $f = p$ lives on $\mathbf{Y}$, and the transformation in the theorem is of the form $\phi : \mathbf{X} \to \mathbf{Y}$): Choose $\phi := \tau^{-1}$, so $Y = \phi(X)$. Then we can express p as a function of x, since $\mathbf{J}$ is in this case just the derivative ϕ', and Theorem 4.4 gives

$$\int_{\mathbb{R}} p(y)dy = \int_{\mathbb{R}} p(\phi(x))d(f(x)) = \int_{\mathbb{R}} p(\phi(x))\phi'(x)dx \; . \tag{4.6}$$

It helps to remember that what we are really looking for in such a problem is p_y *as a function of* x. The recipe for finding p_x is hence:

(1) Determine τ such that $x = \tau(y)$.

(2) Then p_x as a function of x is

$$p_x(x) = p(\phi(x))\phi'(x) \qquad \text{where } \phi = \tau^{-1} \ . \tag{4.7}$$

4.5 Example. Suppose Y is a positive random variables, p its density (with respect to Lebesgue measure), t a constant, and we want to find the density p_x of $X = \frac{Y}{t}$ (again with respect to Lebesgue measure). We set $\tau(y) := y/t$, and hence $\phi(x) = xt$, so we obtain $p_x(x) = tp(tx)$. ◁

4.6 Exercise. Recall that the gamma distribution with parameters (α, λ) is the distribution on $(0, \infty)$ with Lebesgue density $p(x) = \Gamma(\lambda)^{-1}\alpha^\lambda x^{\lambda-1}e^{-\alpha x}$. Suppose X and Y are independent gamma variables with parameters (α, λ_y) and (α, λ_x). Show that

$$\tfrac{X}{X+Y} \perp\!\!\!\perp \tfrac{Y}{X+Y} \qquad \text{and} \qquad \tfrac{X}{X+Y} \perp\!\!\!\perp Y \ . \tag{4.8}$$

◁

4.2. Pullback measures

The image measure $\phi_\#\mu$ is well-defined whenever ϕ is measurable. Things are more complicated for pullbacks: In principle, we would like to define $\phi^\#\nu$ in terms of ν by setting

$$\phi^\#\nu(\phi^{-1}A) := \nu(A) \ . \tag{4.9}$$

The problem is the following:

- The preimage $\phi^{-1}A$ of a set A is determined only by the points in $A \cap \phi(\mathcal{X})$. Points in A outside $\phi(\mathcal{X})$ do not correspond to any points in the preimage.
- Therefore, two different sets $A \neq B$ can have identical preimages $\phi^{-1}A = \phi^{-1}B$, provided they do not differ inside the subset $\phi(\mathcal{X})$.
- That means that a set in $\mathcal{X}$ can simultanuously be the preimage of two different sets A and B in $\mathcal{Y}$, with different values of $\nu(A)$ and $\nu(B)$, so it is not clear which of those values should be assigned in (4.9).

We note that this is not a problem if A and B have the same measure under ν. One way to ensure that is to require that ν concentrates all its mass on $\phi(\mathcal{X})$—in that case, even if A and B differ outside $\phi(\mathcal{X})$, these distinct sets have measure 0, and so $\nu(A) = \nu(B)$. We could hence require $\nu(\phi(\mathcal{X})) = 1$ if ν is a probability measure, or more generally $\nu(\phi(\mathcal{X})) = \nu(\mathcal{Y})$.

That, again, is problematic, since even if ϕ is measurable, the image of a measurable set such as $\mathcal{X}$ need *not* be measurable, in which case $\nu(\phi(\mathcal{X}))$ is not defined. We can get around this problem by defining the so-called *outer measure* ν^* of ν, which is a set function defined on all subsets in $\mathcal{Y}$:

4.7 Definition. Let $(\mathcal{X}, \mathcal{C}, \mu)$ be a measure space. The set function

$$\mu^* : 2^{\mathcal{X}} \to [0, \infty] \qquad \mu^*(M) := \inf\{\mu(A) | A \in \mathcal{C}, M \subset A\} \tag{4.10}$$

is called the **outer measure** defined by μ. ◁

This definition is of course reminiscent of definitions we have seen before (regularity and tightness). Clearly, $\mu^*(A) = \mu(A)$ whenever $A \in \mathcal{C}$. If a (not necessarily measurable) subset M of $\mathcal{Y}$ satisfies

$$\nu^*(M) = \nu(\mathcal{Y}) \ , \tag{4.11}$$

we say that M has **full outer measure** under ν. If ν is a probability measure, this means of course that $\nu^*(\phi(\mathcal{X})) = 1$. It is not hard to deduce the following properties from the definition of outer measure:

4.8 Lemma. *Let $(\mathcal{Y}, \mathcal{C}, \nu)$ be a measure space.*

(1) *For every subset M of $\mathcal{Y}$ there is a measurable set $A \in \mathcal{C}$ such that $M \subset A$ and $\nu^*(M) = \nu(A)$.*
(2) *A measurable set A is a null set under ν iff $\nu^*(A) = 0$.*
(3) *A set $M \subset \mathcal{Y}$ has full outer measure if and only if $\nu(A) = 0$ for every measurable set $A \subset (\mathcal{Y} \setminus M)$.*

◁

Proof. Homework. □

Using the outer measure, we can formulate a condition which prevents ambiguity in the definition (4.9), even if $\phi(\mathcal{X})$ is not measurable: Instead of demanding that ν assigns all its mass to $\phi(\mathcal{X})$, we require that the outer measure does so:

4.9 Theorem [Existence of pullbacks]. *Let $\mathcal{X}$ be a set, $(\mathcal{Y}, \mathcal{A}_{\mathrm{Y}})$ a measurable space and $\phi : \mathcal{X} \to \mathcal{Y}$ a mapping. If ν is a measure on $\mathcal{Y}$ such that $\phi(\mathcal{X})$ has full outer measure under ν, there exists a measure μ on the σ-algebra $\phi^{-1}\mathcal{A}_{\mathrm{Y}}$ in $\mathcal{X}$ satisfying such that $\phi(\mu) = \nu$.* ◁

Proof. Since $\phi^{-1}\mathcal{A}_{\mathrm{Y}}$ is the preimage of a σ-algebra, it is itself a σ-algebra. We have to show that (4.9) is well-defined: To this end, consider two measurable sets $A, B \in \mathcal{A}_{\mathrm{Y}}$ which have identical preimages $\phi^{-1}A = \phi^{-1}B$. Since the preimages are identical, $A \cap \phi(\mathcal{X})$ and $B \cap \phi(\mathcal{X})$ are identical. Hence, the symmetric difference $A \triangle B$ and $\phi(\mathcal{X})$ are disjoint. Since $\phi(\mathcal{X})$ has full outer measure under ν, this means (by Lemma 4.8) $A \triangle B$ is a null set under ν, and so $\nu(A) = \nu(B)$. Therefore, (4.9) indeed defines a set function $\phi^{\#}\nu : \phi^{-1}\mathcal{A}_{\mathrm{Y}} \to [0, \infty)$. What remains to be shown is that this set function is a measure, which is straightforward to verify. □

CHAPTER 5

Stochastic processes

A stochastic process is a random mapping, from some set U to a set V—a random function, random probability measure, random operator, etc. The distribution $\mathcal{L}(X)$ of the process is a probability measure on the space of mappings. Perhaps the most common case are random functions $\mathbb{N} \to \mathbb{R}$ (in which case the elements of $\mathbb{N}$ are often interpreted as points in time) or $\mathbb{R}_+ \to \mathbb{R}$ (where $\mathbb{R}_+$ is a time axis).

5.1 Definition. A **stochastic process** is a random mapping $U \to V$, where U is a set and V a measurable space. That is, the process is a random variable $X : \Omega \to \{x | x \text{ is mapping } U \to V\}$; we also write $X : U \to V$. We call U the **index set** and V the **state space** of X. It is customary to write

$$X_u := X(u) \ , \tag{5.1}$$

and X is often denoted $(X_u)_{u \in U}$. A realization $X(\omega)$ of the process is called a **path** or a **trajectory** of X. ◁

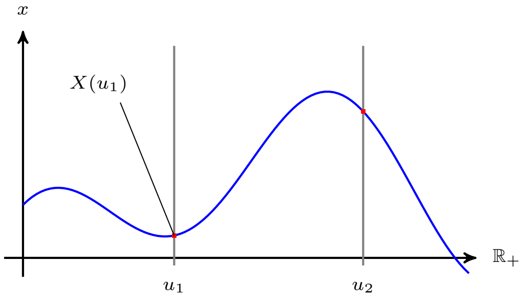
The set of all mappings $U \to V$, for any given sets U and V, is the product set V^U. However, although we have to work with this product set, we will not necessarily work with the product *space*—for the moment, we will leave open which topology or σ-algebra we choose on V^U.

Suppose a stochastic process $X : \mathbb{R}_+ \to \mathbb{R}$ is given. For any point $u \in \mathbb{R}_+$ the function value $X(u)$ is then a random variable with values in $\mathbb{R}$ (see Figure 5.1). More generally, if we pick a finite set $\{u_1, \ldots, u_n\}$, we can ask for the joint distribution of $(X(u_1), \ldots, X(u_n))$. These distributions play an important role in this chapter.

5.2 Definition. Let $X : U \to V$ be a random mapping. For any finite subset $\{u_1, \ldots, u_n\} \subset \mathbb{R}$, define

$$\mu_{u_1,\ldots,u_n} := \mathcal{L}(X(U_1), \ldots, X(U_n)) \ . \tag{5.2}$$

Figure 5.1. A random function $X : \mathbb{R}_+ \to \mathbb{R}$ defines a random scalar $X(u)$ for every $u \in \mathbb{R}$.



The laws $\mu_{u_1,\ldots,u_n}$, for all finite $\{u_1,\ldots,u_n\}$, are called the **finite-dimensional distributions (FDDs)** of the process X. ◁

Here is a brief overview of the fundamental questions addressed in this chapter:

- The set of all mappings $U \to V$ is the product space V^U. A random mapping is hence a random variable with values in V^U, and its law is a measure on V^U. The construction of stochastic processes hence amounts to the definition of probability measures (with nice properties) on such spaces.
- Since V^U is typically infinite-dimensional, the law of X usually does not have a useful density representation (cf. Remark 2.47). We hence have to find another way to represent the law of X. One of the key ideas of stochastic process theory is that $\mathcal{L}(X)$ can be uniquely represented by the family of all FDDs of X. Roughly speaking, we use an infinite number of finite-dimensional distributions to represent a single infinite-dimensional one. Theorems which establish these representations are usually called **extension theorems**; the most important ones are due to Kolmogorov, to Bochner, and to Prokhorov.
- If U is uncountable (which it often is), the space V^U typically does not have nice properties. For example, even if U and V are both Polish, V^U is not if U is uncountable. On the other hand, we are not really interested in all elements of this space: For example, if V^U is the set of all functions $\mathbb{R}_+ \to \mathbb{R}$ as above, the lion's share of these functions jumps at almost every point. We are typically interested only in a subset $\mathbf{X} \subset V^U$ of mappings that are in some sense regular, e.g. continuous or piece-wise continuous (if X is a random function on the line), countably additive (if X is a random set function), etc.
- The basic extension theorem (of Kolmogorov) constructs a measure on V^U. We then have to complement the extension theorem by a result which restricts the constructed measure on V^U to a measure on a subset of interest, e.g. to $\mathbf{C}(\mathbb{R}_+,\mathbb{R}) \subset \mathbb{R}^{\mathbb{R}_+}$.

5.1. Constructing spaces of mappings

We will now discuss a method to discuss more bespoke subsets $\mathcal{X}$ of the product set V^U, using a so-called inverse limit. The extension theorems in the following section will then show how to put a probability distribution on $\mathcal{X}$, to define a stochastic process with paths in $\mathcal{X}$.

5.3 Definition. Let the following components be given:

(1) A directed set $(\mathbb{T},\preceq)$ (cf. Section 1.1).
(2) For each $t \in \mathbb{T}$, a topological space $\mathbf{X}_t$.
(3) For every pair $s,t \in \mathbb{T}$ that satisfies $s \preceq t$, a surjective, continuous mapping $\mathrm{pr}_{ts} : \mathbf{X}_t \to \mathbf{X}_s$.

The **inverse limit** of $\varprojlim(X_t)$ of the sets X_t with respect to the mappings pr_{ts} is the set of all nets $(x_t)_{t\in\mathbb{T}}$ satisfying

$$x_t \in \mathbf{X}_t \qquad \text{and} \qquad \mathrm{pr}_{ts}(x_t) = x_s \quad \text{whenever } s \preceq t\;. \tag{5.3}$$

◁

Note that $\varprojlim(\mathbf{X}_t) \subset \prod_{t\in\mathbb{T}} \mathbf{X}_t$ by definition. Many authors refer to inverse limits as **projective limits**. Once we have constructed an inverse limit set $\varprojlim(\mathbf{X}_t)$, we can define a map $\mathrm{pr}_t : \varprojlim (\mathbf{X}_t) \to \mathbf{X}_t$, simply by defining

$$\mathrm{pr}_t\big((x_s)_{s\in\mathbb{T}}\big) = x_t \ . \tag{5.4}$$

That is: The inverse limit is constructed by regarding each sequence or net satisfying (5.3) as a point, and defining $\varprojlim (\mathbf{X}_t)$ as the set of all these points. The mapping pr_t then takes each point x to the entry x_t of the corresponding net. With the help of these mappings, we can turn $\mathcal{X}$ into a topological space $\mathbf{X}$:

5.4 Definition. The **inverse limit topology** on $\varprojlim (X_t)$ is the smallest topology which makes all mappings pr_t, for $t \in \mathbb{T}$, continuous. ◁

We will next look at a few examples of inverse limit sets. The most straightforward example of an inverse limit is a product space of the form $\mathbf{X}_0^U$. That may seem a bit pointless—of course, we do not need an inverse limit to construct a product set. However, we will see below that the construction of a probability measure on the product does use the inverse limit structure, so it is useful to look at this construction first.

5.5 Example [Product space]. Suppose we want to construct the product space $\mathbb{R}^{\mathbb{R}_+}$. We choose $\mathbb{T}$ as the set of all finite subset of $\mathbb{R}_+$, ordered by inclusion, i.e.

$$(\mathbb{T}, \preceq) := \big(\big\{\{u_1, \dots, u_n\} \,|\, n \in \mathbb{N}, u_1, \dots, u_n \in \mathbb{R}\big\}, \subset\big) \ . \tag{5.5}$$

Let $\mathbf{X}_t = \mathbb{R}^t$ for all $t \in \mathbb{T}$. For $s \preceq t$, we define pr_{ts} as the mapping

$$\mathrm{pr}_{ts} : \mathbb{R}^t \to \mathbb{R}^s \qquad (x_u)_{u\in t} \mapsto (x_u)_{u\in s} \ . \tag{5.6}$$

which, from the finite sequence $(x_u)_{u\in t}$, deletes all elements whose indices are not in $s \subset t$. The mapping (5.6) is called a **product space projection**[1]. Now consider all sequences satisfying (5.3); it will not be hard to convince yourself that each such sequence defines a point in $\mathbb{R}^{\mathbb{R}_+}$, and conversely, that each point in $\mathbb{R}^{\mathbb{R}_+}$ defines one and only one such sequence. Then the inverse limit is hence the product space

$$\varprojlim (\mathbf{X}_t) = \prod_{u\in\mathbb{R}_+} \mathbb{R} = \mathbb{R}^{\mathbb{R}_+} \ , \tag{5.7}$$

By applying the definition (5.4) of the mappings pr_t, we find that each mapping pr_t is the projection to the subspace $\mathbf{X}^t$, i.e. the mapping

$$\mathrm{pr}_t : \mathbb{R}^{\mathbb{R}_+} \to \mathbb{R}^t \qquad (x_u)_{u\in\mathbb{R}_+} \mapsto (x_u)_{u\in t} \ , \tag{5.8}$$

which deletes all elements from the infinite sequence $(x_u)_{u\in\mathbb{R}_+}$ whose indices u are not contained in t. ◁

5.6 Example [Set functions]. Let U be a Borel space. As index set, choose the set of all partitions of U into a finite number of Borel sets, i.e.

$$\mathbb{T} := \{(A_1, \dots, A_n) \,|\, n \in \mathbb{N}, A_i \in \mathcal{B}(U) \text{ and } (A_1, \dots, A_n) \text{ partition of } U\} \ . \tag{5.9}$$

We order $\mathbb{T}$ by defining

$$s \preceq t :\Leftrightarrow t \text{ is refinement of } s \ . \tag{5.10}$$

[1] In the Euclidean spaces used in this example, this is precisely the orthogonal projection onto the subspace $\mathbb{R}^s$, where the direction of projection is parallel to all axes indexed by $u \in t \setminus s$. However, orthogonal projections are only defined on spaces with scalar products. The product space projector (5.6) is well-defined even in spaces without a a scalar product, since axes-parallel projection only requires deletion of entries from a sequence.

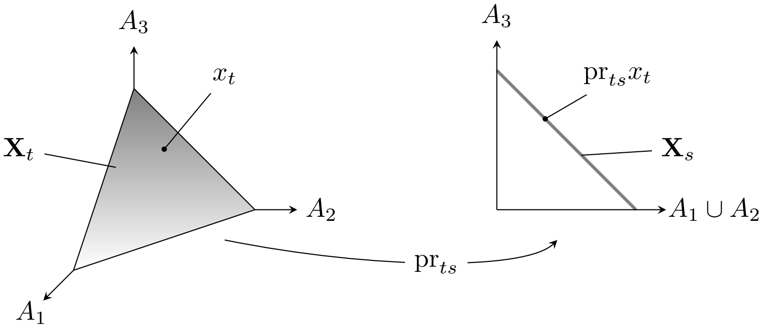


Figure 5.2. The set function construction, for $t = (A_1, A_2, A_3)$ and $s = (A_1 \cup A_2, A_3)$. The set $\mathbf{X}_t$ contains all probability measures on the events A_1, A_2 and A_3, i.e. all vectors $x_t \in \mathbb{R}^3$ with $x_t(A_i) \geq 0$ and $x_t(A_1) + x_t(A_2) + x_t(A_3) = 1$.

The directed set $(\mathbb{T}, \preceq)$ is precisely the index set we already used in Section 1.9. For each $t = (A_1, \dots, A_n) \in \mathbb{T}$, let $\mathbf{X}_t$ be the set of all probability measures $\sigma(t) \to [0, 1]$, defined on the finite σ-algebra $\sigma(t)$ generated by the sets A_i in t. In other words,

$$\mathbf{X}_t := \left\{ x_t \in \mathbb{R}^n \,|\, x_t(A_i) \geq 0 \text{ for } i = 1, \dots, n \text{ and } \textstyle\sum_{i=1}^n x_t(A_i) = 1 \right\} . \tag{5.11}$$

To define the mapping pr_{ts}, suppose t is a refinement of s. Then some of the sets in s are unions of sets in t, e.g. $t = (A_1, \dots, A_n)$ and $s = (A_1 \cup A_2, A_3, \dots, A_n)$. If $x_t \in \mathbf{X}_t$ is a measure on $\sigma(t)$, we define a measure on $\sigma(s)$ as

$$\mathrm{pr}_{ts}(x_t) := (x_t(A_1 \cup A_2), x_t(A_3), \dots, x_t(A_n)) . \tag{5.12}$$

Obviously, this generalizes to arbitrary coarsenings s of t, but notation becomes pretty cumbersome, so I will not write it out.

What is the inverse limit set? An element x of $\varprojlim(\mathbf{X}_t)$ is certainly some form of set function. We can evaluate it on any $A \in \mathcal{B}(U)$ by choosing some $t = (A_1, \dots, A_n)$ which contains A as one of the sets A_i; then $x(A) = (\mathrm{pr}_t x)(A)$. This also means that x is finitely additive: To evaluate $x(A_1) + x(A_2)$, choose a t that contains both A_1 and A_2, and use the finite additivity of the probability measure x_t. However, x lives on the *infinite* set system $\mathcal{B}(U)$, so to be a probability measure, it would have to be countably additive. That need not be the case, since we cannot formulate countable additivity as a property on any $t \in \mathbb{T}$.

The inverse limit set we obtain is hence the set of probability charges on $\mathcal{B}(U)$, i.e. all of *finitely* additive probability measures,

$$\varprojlim(X_t) = \left\{ x : \mathcal{B}(U) \to [0, 1] \,\middle|\, x(\varnothing) = 0, x(U) = 1 \text{ and } x \text{ finitely additive } \right\} . \tag{5.13}$$

Note that this set contains all probability measures (the charges which are also countably additive). Each mapping pr_t takes a charge x to its values on the partition t, i.e. to $(x(A_1), \dots, x(A_n))$. An example of a subset of regular functions we might be interested in would be the subset of those x which are even countably additive. ◁

5.7 Example [Graphs]. This final example illustrates that we do not necessarily have to interpret the elements of the inverse limit set as mappings. Suppose our index set is $(\mathbb{T}, \preceq) := (\mathbb{N}, \leq)$, and for each n, $\mathbf{X}_n$ is the set of all undirected, simple graphs on the vertex set $\{1, \dots, n\}$. For $m \leq n$, we define the map pr_{nm} as the mapping from a graph on n vertices to its induced subgraph on the vertices

$\{1, \ldots, m\}$. Then $\varprojlim(X_n)$ is the set of all undirected, simple graphs on the vertex set $\mathbb{N}$, and pr_n the mapping from an infinite graph to its induced subgraph on the vertices $\{1, \ldots, n\}$. An example of a subset of "regular" elements would be the set of all graphs on $\mathbb{N}$ for that each vertex has finite degree. ◁

5.8 Remark [Warning]. An inverse limit set can be empty, (if the condition (5.3) cannot be satisfied by any net). This can happen even if we assume the mappings pr_{ts} to be surjective—even if pr_{ts} is surjective whenever $s \preceq t$, the mappings pr_t need not be surjective. ◁

If the index set $\mathbb{T}$ is countable, the inverse limit is particularly well-behaved. Any properties that hold for a countable index set also hold in a slightly more general case: A subset $\mathbb{T}'$ of a directed set $\mathbb{T}$ is called **cofinal** if, for every $t \in \mathbb{T}$, there is a $t' \in \mathbb{T}'$ such that $t \preceq t'$. Clearly, the projective limit of a family of spaces indexed by $\mathbb{T}$ is exactly the same as the inverse limit constructed only from those spaces indexed by $\mathbb{T}'$. Hence, whenever $\mathbb{T}$ contains a cofinal subset that is countable, it behaves like a countable set for all purposes of inverse limit constructions.

5.9 Lemma [Inverse limits with countable index sets]. *Let $\varprojlim(\mathbf{X}_t)$ be an inverse limit set. Suppose $\mathbb{T}$ contains a countable cofinal subset.*

(1) *If each of the mappings pr_{ts} is surjective, so are the mappings pr_t. In particular, the inverse limit set is not empty.*

(2) *If additionally each of the spaces $\mathbf{X}_t$ is Polish, the inverse limit set is a Polish space in the inverse limit topology.*

◁

5.2. Extension theorems

If we construct the set of paths using an inverse limit, the index set U of the process and the index set $\mathbb{T}$ of the inverse limit are not necessarily indentically; they may coincide, but more generally, $\mathbb{T}$ is derived from U (such as in the product space example, where each element of $\mathbb{T}$ is a subset of U).

Now consider an inverse limit of topological spaces $\mathbf{X}_t$, with index set $(\mathbb{T}, \preceq)$, and suppose we define a probability measure P_t on each space $\mathbf{X}_t$ (that is, on the Borel sets of $\mathbf{X}_t$). Since each mapping pr_{ts} is continuous, it is measurable. The family $(P_t)_{t\in\mathbb{T}}$ is called **projective** (or an **inverse family**) if

$$\mathrm{pr}_{ts}(P_t) = P_s \qquad \text{whenever } s \preceq t \,. \tag{5.14}$$

Under suitable conditions, a projective family of measures defines a measure P on the inverse limit set—more precisely, on the inverse limit σ-algebra.

Product spaces, as in Example 5.5, are by far the most important case. Recall that each point in a product space can be thought of as a (possibly uncountable) sequence or list $(x_u)_{u\in U}$ of elements, one for each dimension, and the mappings pr_{ts} in this case are the product space projections, i.e. the mappings

$$\mathrm{pr}_{ts} : \prod_{u\in t} \mathbf{V} \to \prod_{u\in s} \mathbf{V} \qquad (x_u)_{u\in t} \mapsto (x_u)_{u\in s} \,. \tag{5.15}$$

which delete some of the elements in the list.

5.10 Kolmogorov's extension theorem. *Let U be an infinite set and $\mathbf{V}$ a Polish space. For each finite subset $t = \{u_1, \ldots, u_n\}$ of U, let P_t be a probability measure* *on the product space* $\mathbf{V}^{\{u_1,\ldots,u_n\}}$. *Require that*

$$\mathrm{pr}_{ts} P_t = P_s \qquad \textit{whenever } s \subset t \tag{5.16}$$

for the product space projections pr_{ts}. *Then there exists a uniquely defined probability measure on* $\mathbf{V}^U$ *satisfying*

$$\mathrm{pr}_t P = P_t \qquad \textit{for all finite } t \subset U\ . \tag{5.17}$$

◁

Formulated in terms of random variables, this means: For each $u \in U$, let X_u be a random variable with values in the Polish space $\mathbf{V}$. Suppose that for each finite $t \subset U$, the joint distribution $P_t = \mathcal{L}(X_{u_1}, \ldots, X_{u_n})$ known, and that for $s \subset t$, P_s is precisely the marginal distribution under P_t of those X_u for which $u \in s$. Then the joint distribution of $\{X_u | u \in U\}$ exists and is uniquely determined.

5.11 Remark [Interpretation of finite-dimensional distributions]. The FDDs of a process are not just an abstract device of mathematical convenience, but have a concrete modeling interpretation: Suppose we observe samples from some physical process in an experiment, say over time. Our mathematical model of this physical process is a stochastic process X, and since it depends on time, we model the index set as continuous. In any given experiment, we can only observe a finite number of samples from this process. The law of the random variables modeling these samples is a FDD (a *single* FDD, indexed by the set of observation times). To say that two processes have the same FDDs is to say that their laws cannot be distinguished from each other by any experiment. Kolmogorov's theorem tells us that these conditions are sufficient to determined the law of the process, even on an uncountable index set. ◁

Our next task is the proof of Kolmogorov's theorem. The inverse limit measure P in the theorem lives on the product σ-algebra $\mathcal{B}(\mathbf{V})^U$. By definition, the product σ-algebra is generated by all sets of the form

$$\mathrm{pr}^{-1}_{\{u\}} A_{\{u\}} \qquad \text{where } u \in U \text{ and } A_{\{u\}} \in \mathcal{B}(\mathbf{X}^{\{u\}}) = \mathcal{B}(\mathbf{V})\ . \tag{5.18}$$

I will call such sets **mono-cylinders**. If $A_{\{u\}} \subset \mathbf{V}^{\{u\}}$ and $A_{\{v\}} \subset \mathbf{V}^{\{v\}}$, for two distinct elements $u \neq v$ of U, are measurable sets, the intersection of the corresponding cylinders is evidently

$$(\mathrm{pr}^{-1}_{\{u\}} A_{\{u\}}) \cap (\mathrm{pr}^{-1}_{\{v\}} A_{\{v\}}) = \mathrm{pr}^{-1}_{\{u,v\}} (A_{\{u\}} \times A_{\{v\}})\ . \tag{5.19}$$

Thus, the intersection of two mono-cylinders with bases in distinct dimensions is non-empty. The same is true if we intersect any finite or even countable number of mono-cylinders with bases in mutually distinct dimensions. For the σ-algebras with this property, the following auxiliary result—which, as far as I can tell, is due to Marczewski—will save us a lot of lengthy diagonalization arguments.

5.12 Lemma [Unions of compact classes]. *Let* $\mathcal{X}$ *be a set and* U *an arbitrary index set. For each* $u \in U$, *let* $\mathcal{C}_u$ *be a* σ*-algebras in* $\mathcal{X}$, *and* $\mathcal{K}_u \subset \mathcal{C}_u$ *a compact class. Require that* $(\mathcal{C}_u)_{u \in U}$ *has the property: For every countable subset* $U_0 \subset U$ *of indices,*

$$A_u \in \mathcal{C}_u \textit{ and } A_u \neq \varnothing \textit{ for all } u \in U_0 \qquad \Rightarrow \qquad \textstyle\bigcap_{u \in U_0} A_u \neq \varnothing\ . \tag{5.20}$$

(1) *Let* $\mathcal{K}$ *be the closure of* $\cup_{u \in U} \mathcal{K}_u$ *under finite unions and intersections. Then* $\mathcal{K}$ *is a compact class.*

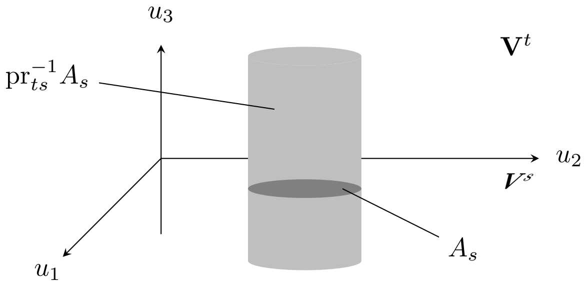


Figure 5.3. A cylinder in $\mathbf{V}^t$ with base A_s in $\mathbf{V}^s$, where $t = \{u_1, u_2, u_3\}$ and $s = \{u_1, u_2\}$: The cylinder is the preimage $\mathrm{pr}_{ts}^{-1}A_s$. The cylinder sets in the proof of Kolmogorov's theorem are the infinite-dimensional analogues, where t is replaced by the entire set U, and preimages are taken under the mappings pr_s rather than pr_{ts}.

(2) *Let μ be a finitely additive probability measure on the smallest algebra containing $\cup_{u\in U}\mathcal{C}_u$. If each restriction $\mu|_{\mathcal{C}_u}$ is tight with respect to $\mathcal{K}_u$, then μ is tight with respect to $\mathcal{K}$.*

◁

PROOF. This is Lemma 6.1 in [8]. □

More generally, a set of the form

$$\mathrm{pr}_t^{-1}A_t \qquad \text{where } t \subset U \text{ is finite and } A_t \in \mathcal{B}(\mathbf{X}^t) \tag{5.21}$$

it is called a **cylinder set with base A_t**, for obvious reasons (see Figure 5.3). Note that the mono-cylinders satisfy condition (5.20), whereas the cylinders do not.

PROOF OF THEOREM 5.10. The inverse limit set is the product space $\mathcal{X} := \mathbf{V}^U$, and the inverse limit σ-algebra is the product σ-algebra $\mathcal{B}(\mathbf{V})^U$. For each finite subset $t \subset U$, the mapping pr_t is a product space projection, and hence surjective. Therefore, $\mathrm{pr}_t\mathcal{X}$ has outer measure 1 under P_t. By Theorem 4.9, the pullback $\mu_t := \mathrm{pr}_t{}^{\#}P_t$ exists and is a probability measure on $\mathrm{pr}_t^{-1}\mathcal{B}(\mathbf{V}^t)$, i.e. on the cylinders with base in $\mathbf{V}^t$.

Let $\mathcal{Z}$ denote the set of all cylinders (for all t). It is not hard to show that $\mathcal{Z}$ is an algebra. Define a set function μ on $\mathcal{Z}$ as

$$\mu(A) := \mu_t(\mathrm{pr}_t^{-1}A) \qquad \text{for any } t \text{ with } A \in \mathrm{pr}_t^{-1}\mathcal{B}(\mathbf{V}^t)\ . \tag{5.22}$$

Since (P_t) is projective, the values $\mu_t(\mathrm{pr}_t^{-1}A)$ coincide for all such t. Since each P_t is a probability measure, it follows immediately that μ is finitely additive (take a finite union of finite index sets t), with $\mu(\varnothing) = 0.$ and $\mu(\mathcal{X}) = 1$. What remains to be shown is that μ extends to a countably additive measure on the product σ-algebra.

Now let α be the smallest algebra containing all mono-cylinders. Since the mono-cylinders are contained in $\mathcal{Z}$, and $\mathcal{Z}$ is an algebra, $\alpha \subset \mathcal{Z}$. Hence, the restriction $\mu|_\alpha$ is a finitely additive probability on α. Let $\mathcal{K}_u$ be the set of compact sets in $\mathbf{V}^{\{u\}}$, and define

$$\mathcal{K} := \text{ closure under finite unions and intersections of } \bigcup_{u\in U} \mathrm{pr}_{\{u\}}^{-1}\mathcal{K}_u\ . \tag{5.23}$$

By Lemma 5.12(i), $\mathcal{K}$ is a compact class. Since $\mathbf{V}$ is Polish, $P_{\{u\}}$ is tight with respect to $\mathcal{K}_u$, and the restriction of μ to the $\{u\}$-mono-cylinders is hence tight with respect to $\mathrm{pr}_{\{u\}}^{-1}\mathcal{K}_u$. By Lemma 5.12(ii), $\mu|_\alpha$ is compact with respect to $\mathcal{K}$. By Lemma 3.9, it is hence countable additive on α.

Since the mono-cylinders generate the product σ-algebra, so does α. Therefore, $\mu|_\alpha$ is a countably additive probability measure on an algebra which is also a generator, and hence has a unique extension to a probability measure on the product σ-algebra by [Probability I, Theorem 6.1]. □

5.13 Remark [The extension theorem as a regularity result]. One possible interpretation of Kolmogorov's theorem is as a regularity result: Since the measures P_t determine the values of P on the cylinder sets (and only on those), the theorem says that P is completely determined by its values on subset (the algebra $\mathcal{Z}$ of cylinder sets) of its domain (the σ-algebra $\mathcal{C}$). This type of property—a mapping is completely determined by its values on a suitable subset—is a hallmark of regular mappings. For example, recall that a continuous function is completely determined by its values on a dense subset of its domain. In Kolmogorov's theorem, the function is a set function and the relevant regularity property is countable additivity. ◁

Kolmogorov's theorem can in particular be applied to construct random sequences $(X_n)_{n\in\mathbb{N}}$, where each X_n takes values in $\mathbf{V}$, by choosing $U := \mathbb{N}$. We obtain:

5.14 Corollary [Random sequences]. *Let $(X_n)_{n\in\mathbb{N}}$ be a random variables, where each X_n takes values in a Borel space $\mathbf{V}$. Assume the joint distribution $\mathcal{L}(X_1,\ldots,X_n)$ is known for each finite n. Then the joint distribution P of the infinite sequence (X_n) exists, is uniquely determined and satisfies the relation*

$$\mathrm{pr}_n P = \mathcal{L}(X_1,\ldots,X_n) \tag{5.24}$$

for every $n \in \mathbb{N}$. ◁

The proof of Kolmogorov's theorem is already quite lengthy, and we have a lot left to discuss, so I will not look at other inverse limit results in detail. In the case of a countable index set (or one with a countable cofinal subset), the general inverse limit theorem does not involve any complicated conditions, so I just state this result here without proof.

5.15 Theorem [Extension theorem for countable index sets]. *Let $(\mathbb{T},\preceq)$ be a directed set containing a countable cofinal sequence. Let $(\mathbf{X}_t)_{t\in\mathbb{T}}$ be a family of topological spaces and pr_{ts}, for $s \preceq t$, continuous surjective maps. For every $t \in \mathbb{T}$, let P_t be a probability measure on $\mathbf{X}_t$. If the family $(P_t)_{t\in\mathbb{T}}$ is projective with respect to the mappings pr_{ts}, there exists a uniquely determined probability measure P on the inverse limit σ-algebra satisfying $\mathrm{pr}_t(P) = P_t$ for all t.* ◁

Note that the set of all finite subsets of a set U, ordered by inclusion, contains a cofinal subset if and only if U is countable. Hence, Kolmogorov's extension theorem becomes a special case of this result if U is countable, but not otherwise.

5.3. Processes with regular paths

In Kolmogorov's theorem, we construct a probability measure P on the set of all mappings $U \to V$. This measure P lives on the product σ-algebra. Since U is usually uncountable, the product σ-algebra is very coarse: It contains precisely those sets which are cylinders with base in a countable-dimensional subspace. If we think of events in this σ-algebra as properties of mappings $x : U \to V$, this means that an event is measurable if it can be expressed in terms of the values of the random mapping X at a *countable* number of points, but not otherwise. For instance:

- $\{X(u) < X(v)\}$, for a fixed pair $u, v \in U$, is measurable.
- $\{X(u)$ is positive at all integers $u\}$ is measurable.
- $\{X$ is continuous$\}$ (or differentiable, or strictly monotone) is not measurable.

The next theorem permits us to restrict the measure P to a measure on a subset $\mathbf{X}$ consisting of "interesting" objects (such as continuous mappings), even if that subset is not measurable.

5.16 Theorem [Processes with regular paths]. *Let $\mathcal{X}$ be an inverse limit $\varprojlim(\mathbf{X}_t)$ of topological spaces $\mathbf{X}_t$ with respect to mappings pr_{ts}, $\mathcal{C}$ the inverse limit σ-algebra, and P the inverse limit of a projective family (P_t) of probability measures. Let $\mathbf{X}$ be a topological space which is contained in $\mathcal{X}$ as a subset. Then there exists a probability measure $\hat{P}$ on $\mathbf{X}$ such that*

$$\mathrm{pr}_t\big|_{\mathbf{X}}(\hat{P}) = P_t \qquad \textit{for all } t \tag{5.25}$$

if

(1) $\mathcal{B}(\mathbf{X}) = \mathcal{C} \cap \mathbf{X} := \{A \cap \mathbf{X} | A \in \mathcal{C}\}$
(2) $\mathbf{X}$ *has outer measure* $P^*(\mathbf{X}) = 1$.

If so, $\hat{P}$ is uniquely determined. ◁

In general, whenever C is a subset of a set A, we can defined a mapping I which maps each point in C to itself and is undefined on points in $A \setminus C$; hence, we can regard I as a mapping $I : C \mapsto A$ that "embeds" C into A. This mapping is called the **canonical inclusion**, and is customarily denoted with a "hooked" arrow,

$$I : C \hookrightarrow A\ . \tag{5.26}$$

Defining a mapping that takes points to itself may not seem particularly thrilling. It allows us, however, to write the restriction of the measure P to the subset $\mathbf{X}$ as a pullback under I.

Proof. Let $I : \mathbf{X} \hookrightarrow \mathcal{X}$ be the canonical inclusion of $\mathbf{X}$ into the inverse limit space. By assumption (2), we have

$$P^*(I(\mathbf{X})) = P^*(\mathbf{X}) = 1, \tag{5.27}$$

so by Theorem 4.9, P has a unique pullback measure $\hat{P} := I^{\#}P$ on $\mathbf{X}$, defined on the σ-algebra $I^{-1}\mathcal{C}$. Since $I^{-1}\mathcal{C} = \mathcal{C} \cap \mathbf{X}$, assumption (1) implies $\mathcal{B}(\mathbf{X}) = I^{-1}\mathcal{C}$. What remains to be shown is that $\hat{P}$ satisfies (5.25). By definition of the inverse limit, each P_t is the image measure of P under pr_t, and by definition of the pullback, P is in turn the image measure of $\hat{P}$ under I. Since the restriction $\mathrm{pr}_t\big|_{\mathbf{X}}$ of pr_t to $\mathbf{X}$ is precisely $\mathrm{pr}_t \circ I$, we have for any set $A \in \mathcal{B}(\mathbf{X}_t)$

$$P_t(A) = P(\mathrm{pr}_t^{-1} A) = \hat{P}(I^{-1} \circ \mathrm{pr}_t^{-1} A)\ , \tag{5.28}$$

and hence (5.25). □

We can now summarize the construction of stochastic processes as a checklist: Our objective is to construct random mappings $U \to V$ that satisfy a given regularity property (continuity, countable additivity, etc). Let $\mathbf{X} \subset V^U$ be the set of all mappings which satisfy this property. To construct a stochastic process X with law $\hat{P}$ whose paths are almost surely elements of $\mathbf{X}$, we:

(1) Define a family of candidate finite-dimensional distributions P_t, for all finite $t \subset U$, as in Theorem 5.10.
(2) Show that the family is projective; then the inverse limit measure P exists by Theorem 5.10.
(3) Show that the Borel σ-algebra of $\mathbf{X}$ is $\mathcal{C} \cap \mathbf{X}$, where $\mathcal{C}$ is the inverse limit σ-algebra.
(4) Show that $\mathbf{X}$ has outer measure $P^*(\mathbf{X}) = 1$ under the inverse limit measure P. The existence of X then follows by Theorem 5.16.

Arguably the most important case is the construction of processes with paths in $\mathbf{C}(\mathbb{R}_+, \mathbf{V})$, the space of continuous functions $\mathbb{R}_+ \to \mathbf{V}$, where $\mathbf{V}$ is a metric space. The standard topology on this space is the **topology of compact convergence**, in which a sequence (x_n) of continuous functions converges if and only if it converges uniformly on every compact set in $\mathbb{R}_+$. To guarantee that a process almost surely takes value in $\mathbf{C}(\mathbb{R}_+, \mathbf{V})$, we need sufficient conditions for the hypothesis in Theorem 5.16 to hold.

Such conditions are provided by a famous result of Kolmogorov and Chentsov for the case where $\mathbf{V} = \mathbb{R}^d$ and the paths of X are locally Lipschitz continuous (i.e. slightly smoother than just continuous). We denote by $\mathbf{C}_\gamma(\mathbb{R}_+, \mathbb{R}^d)$ the topological subspace of $\mathbf{C}(\mathbb{R}_+, \mathbb{R}^d)$ consisting of all functions that are locally Lipschitz of order γ.[2]

5.17 Theorem. *In the setup of Theorem 5.10, let* $\mathbf{V}$ *be a metric space and choose* $U = \mathbb{R}_+$. *Then the following holds:*

(1) *The space* $\mathbf{C}(\mathbb{R}_+, \mathbf{V})$ *is Polish and*

$$\mathcal{B}(\mathbf{C}(\mathbb{R}_+, \mathbf{V})) = (\mathbf{C}(\mathbb{R}_+, \mathbf{V})) \cap \mathcal{B}(\mathbf{V})^{\mathbb{R}_+} \ . \tag{5.32}$$

where $\mathcal{B}(\mathbf{V})^{\mathbb{R}_+}$ *denotes the product* σ*-algebra.*

(2) *If specifically* $\mathbf{V} = \mathbb{R}^d$, *require there exist constants* $\alpha > 0$, $\beta > 0$ *and* $c > 0$ *such that*

$$\mathbb{E}\big[|X_u - X_v|^\alpha\big] \leq c|u - v|^{d+\beta} \qquad \text{for all } u, v \in \mathbb{R}_+ \ , \tag{5.33}$$

[2] Recall that a function $f : \mathbb{R}_+ \to \mathbb{R}$ is Lipschitz continuous of order γ if there is a constant $c > 0$ such that

$$|f(v) - f(w)| \leq c|v - w|^\gamma \qquad \text{for all } v, w \in \mathbb{R}_+ \ . \tag{5.29}$$

We can weaken the Lipschitz condition by making allowing the constant c to vary with location: We require only that, for every point $u \in \mathbb{R}_+$, there exists an open neighborhood $U_\varepsilon(u)$ and a constant $c_u > 0$ such that

$$|f(v) - f(w)| \leq c_u|v - w|^\gamma \qquad \text{for all } v, w \in U_\varepsilon(u) \ . \tag{5.30}$$

Then f is called **locally Lipschitz of order** γ. We write

$$\mathbf{C}_\gamma(\mathbb{R}_+, \mathbb{R}^d) := \{f : \mathbb{R}_+ \to \mathbb{R}^d | f \text{ locally Lipschitz of order } \gamma\} \ . \tag{5.31}$$

Clearly, the local Lipschitz condition becomes strictly stronger with increasing γ, and hence $\mathbf{C}_{\gamma_2}(\mathbb{R}_+, \mathbb{R}^d) \subset \mathbf{C}_{\gamma_1}(\mathbb{R}_+, \mathbb{R}^d)$ whenever $\gamma_1 \leq \gamma_2$. Local Lipschitz continuity (of any order) implies continuity, so these spaces are always contained in $\mathbf{C}(\mathbb{R}_+, \mathbb{R}^d)$. As for general continuous functions, the topology on $\mathbf{C}_\gamma(\mathbb{R}_+, \mathbb{R}^d)$ is that of uniform convergence on compact sets, so $\mathbf{C}_\gamma(\mathbb{R}_+, \mathbb{R}^d)$ is in fact a topological subspace of $\mathbf{C}(\mathbb{R}_+, \mathbb{R}^d)$.

where, for each $u, v \in \mathbb{R}_+$, X_u and X_v are any random variables with joint distribution $\mathcal{L}(X_u, X_s) = P_{\{u,s\}}$. Then $\mathbf{C}_{\beta/\alpha}(\mathbb{R}_+, \mathbb{R})$ has outer measure 1 under the inverse limit $P = \varprojlim (P_t)$.

◁

Since $\mathbf{C}_\gamma(\mathbb{R}_+, \mathbf{V})$ is a topological subspace of $\mathbf{C}(\mathbb{R}_+, \mathbf{V})$, (5.32) remains true if $\mathbf{C}_\gamma(\mathbb{R}_+, \mathbf{V})$ is substituted for $\mathbf{C}(\mathbb{R}_+, \mathbf{V})$, for any $\gamma > 0$.

Proof. The proof is fairly lengthy, and I omit it here. Statements of this result under slightly varying assumptions can be found in almost every introductory textbook, for example in [K, Theorem 3.23], [B, Theorem 39.3], or [7, Theorem 21.6]. □

The statement is of course rather technical, but note that (1) and (2) correspond precisely to the conditions (1) and (2) of Theorem 5.16. By combining the result with Theorem 5.16 and Theorem 5.10, we can phrase it more concretely in terms of processes:

5.18 Corollary [Kolmogorov-Chentsov continuity criterion]. *Let $X = (X_u)_{u \in \mathbb{R}_+}$ be a stochastic process with values in $\mathbb{R}^d$. If there exist constants $\alpha > 0$, $\beta > 0$ and $c > 0$ such that*

$$\mathbb{E}\big[|X_u - X_v|^\alpha\big] \leq c|u - v|^{d+\beta} \qquad \text{for all } u, v \in \mathbb{R}_+ \;, \tag{5.34}$$

there exists a process X' that has the same finite-dimensional distributions as X and whose paths are almost surely locally Lipschitz continuous of order β/α. ◁

Note that the statement of Eq. (5.34) expresses equivalence between the processes X and X' in terms of their FDDs (cf. Remark 5.11). We have refrained from saying that X and X' have the same distribution, even though their FDDs define the same inverse limit measure. The point is that the inverse limit measure lives on the product space, which is not actually the space of interest, and we think of X' as a random variable whose distribution lives on $\mathbf{C}_{\beta/\alpha}(\mathbb{R}_+, \mathbb{R})$.

Although local Lipschitz continuity is a stronger requirement than continuity, it is considerably weaker than Lipschitz continuity; in particular, the paths of Brownian motion are not Lipschitz, but they are locally Lipschitz, and hence within the remit of Theorem 5.17.

5.4. Brownian motion

The most important continuous-time process is, with little doubt, Brownian motion.

5.19 Definition. A stochastic process $X = (X_u)_{u \in \mathbb{R}_+}$ with values in $\mathbb{R}$ is called **Brownian motion** or a **Wiener process** if:

(1) All FDDs are centered Gaussian distributions, and

$$\mathrm{Cov}[X_u, X_v] = \min\{u, v\} \qquad u, v \in U \;. \tag{5.35}$$

(2) With probability 1, each path of X is continuous.

◁

5.20 Theorem. *Brownian motion exists. Its paths are almost surely locally Lipschitz of every order $\gamma < \frac{1}{2}$.* ◁

Recall for the proof that, if Y is a Gaussian variable with law $\mathcal{N}(0, \sigma^2)$, then scaling Y by a positive constant c defines a variable cY with law $\mathcal{N}(0, c^2\sigma^2)$.

Proof. It is not hard to check that the normal FDDs satisfying (5.35) form a projective family, so a probability measure P on the product σ-algebra exists by Theorem 5.10. We apply Corollary 5.18 to show almost sure continuity. Equation (5.35) implies X_s is marginally normal, and in particular $\mathrm{Var}[X_1] = 1$. Hence,

$$X_v - X_u \overset{\mathrm{d}}{=} \sqrt{v-u}X_1 \sim \mathcal{N}(0, v-u) \qquad \text{whenever } u < v\ . \tag{5.36}$$

That implies

$$\mathbb{E}\big[(X_v - X_u)^{2n}\big] = \mathbb{E}\big[(\sqrt{v-u}X_1)^{2n}\big] = c_n|v-u|^n\ , \tag{5.37}$$

where $c_n := \mathbb{E}[X_1^{2n}] < \infty$. Hence, (5.34) holds, regardless of the choice of n. The specific constants are $\alpha = 2n$, $d = 1$ (since the process takes values in one dimension), and hence $\beta = n-1$. Therefore, local Lipschitz continuity holds for every positive $\gamma < \frac{n-1}{2n}$. Since that is true for any n, it holds for every positive $\gamma < \frac{1}{2}$. □

5.21 Exercise [Scaling Brownian motion]. Suppose $X = (X_u)_{u\in\mathbb{R}_+}$ is a Brownian motion. Show that, for any constant $c > 0$, $(c^{-1}X_{c^2u})_{u\in\mathbb{R}_+}$ is again a Brownian motion. ◁

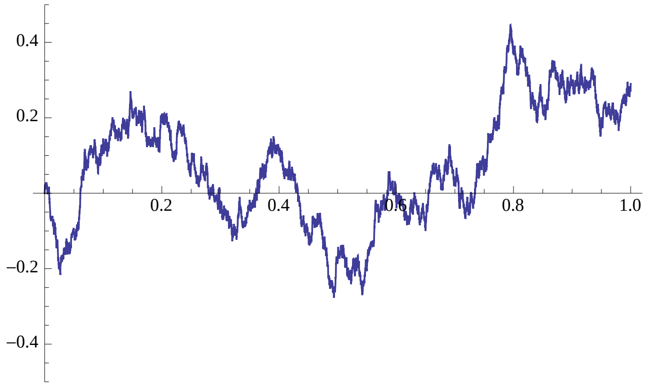
Brownian motion is a process with *independent increments* (see below), which implies it is with probability 1 nowhere differentiable. The study of Brownian motion is an extensive subfield of probability, but we can summarize the fundamental properties:

(1) Each X_u is Gaussian.
(2) With probability 1, each path is continuous at every point $u \in U$.
(3) With probability 1, each path is not differentiable at every point $u \in U$.

5.22 Remark [Modeling with Brownian motion]. Let me add a few imprecise remarks about using Brownian motion as a model. Roughly speaking, it is a model for problems in which we assume that:

(1) At each time, a large number of events takes place on a *microscopic* scale—that is, the effect of each individual event is too small to be observed.
(2) These events aggregate into an effect on a *macroscopic* scale, i.e. which can be observed. This observed effect, over a time interval $[u, u + \Delta u]$ is the change in value of our process X on the interval $[u, u + \Delta u]$.
(3) The events aggregate by summation.

Figure 5.4. A path of Brownian motion on the index set $[0, 1]$.



(4) We assume that the process is not predictable, i.e.

In a finance problem, for example, the microscopic events could be the values of individual shares. Between two fixed times, these values can change, and we hence consider them to be random variables. However, on the macroscopic scale of, say, the entire economy, they are too small to be noticeable. Now suppose we observe the sum of all these variables at each time u (e.g. the value of all shares in a given market). This is our observed process. Its value is much larger than the values of individual shares, so it lives on a macroscopic scale. Since it is a sum of a large number of random variables, it is essentially Gaussian by the central limit theorem (if we assume the individual variables to be independent).

Continuity. Why should the process be continuous? Suppose it is not, i.e. at some time u, the random function exhibits a jump. This would mean that either: (1) Many more of the microscopic events in the sum had positive values than had negative values, or vice versa. (More precisely: The sum of many small events deviated strongly from its expected value one way or another.) For a sufficiently large sum, that does not happen. (2) One of the constituent events had an effect large enough to be visible on a macroscopic scale. Thus, if our modeling assumption is that the constituent events are always microscopic, a process with a continuous path (such as Brownian motion) is an adequate model.[3]

Non-differentiability. From the discussion above, we can conclude that we are looking for a process with Gaussian FDDs and continuous paths, but there processes matching that description which have very smooth, differentiable paths. Suppose our model process has differentiable paths. Differentiability means that the best approximation of the process in a point u by a straight line is locally exact, i.e. the approximation error on a neighborhood of u can be made arbitrarily small by making the the neighborhood sufficiently small (which is just the definition of differentiability). That means that, if a differentiable model is accurate, we can predict future values of the process—and we can make the prediction arbitrarily accurate by making the time interval over which we predict sufficiently short. Thus, if we have reason to assume that the process is not predictable, we need a model whose paths are not differentiable. ◁

5.5. Markov processes

A discrete-time stochastic process (X_n) is called a Markov process if the distribution of each X_n depends on the "past" $(X_1, \dots, X_{n-1})$ of the process only through the value X_{n-1} at the previous time step. Such a discrete-time Markov process is often called a **Markov chain**. More generally, a **Markov chain of order** k is a process where X_n depends only on the past only through the k previous values $X_{n-k}, \dots, X_{n-1}$. Formally, (X_n) is a Markov chain if

$$X_n \perp\!\!\!\perp_{X_{n-1}} X_m \qquad \text{for all } m < n\ . \tag{5.38}$$

Using this formulation, we can immediately generalize the idea to continuous time, or to any totally ordered index set U: The process $(X_u)_{u \in U}$ is Markov if

$$X_u \perp\!\!\!\perp_{X_t} X_s \qquad \text{whenever } s < t < u \in U\ . \tag{5.39}$$

[3] If we want to permit drastic events that generate discontinuities, we could use a process which behaves like a Brownian motion, but jumps at some random times. The standard model of this type is a *Lévy process*.

If we denote the σ-algebra generated by all X_s up to time t as $\sigma_t := \sigma(X_s, s \leq t)$, we can express this more concisely as

$$X_u \perp\!\!\!\perp_{X_t} \sigma_t \qquad \text{whenever } t < u \in U \; . \tag{5.40}$$

Clearly, the family $\{\sigma_u | u \in U\}$ is a filtration, and the family (X_u) is adapted to this filtration. For the general definition of a Markov process, it is customary to generalize a little further, and permit the filtration $\{\sigma_u | u \in U\}$ to be substituted by any filtration $\mathcal{F}$ to which (X_u) is adapted (i.e. which satisfies $\sigma_u \subset \mathcal{F}_u$ for all u). This can be useful because it allows us, for example, to substitute each σ_u by its completion.

5.23 Definition. Let $X = (X_u)_{u \in U}$ be a stochastic process indexed by a totally ordered set U, whose state space is a Borel space $\mathbf{V}$. Let $\mathcal{F} = (\mathcal{F}_u)_{u \in U}$ be a filtration to which X is adapted. Then X is called a **Markov process** (or a **$\mathcal{F}$-Markov process**) if

$$X_u \perp\!\!\!\perp_{X_t} \mathcal{F}_t \qquad \text{whenever } t < u \in U \; . \tag{5.41}$$

◁

The Markov property (5.41) can be stated in terms of conditional distributions as

$$\mathbb{P}[X_u \in \bullet | \mathcal{F}_t] =_{\text{a.s.}} \mathbb{P}[X_u \in \bullet | X_t] \qquad \text{whenever } t < u \; . \tag{5.42}$$

Since $\mathbf{V}$ is Borel, there is a probability kernel representing each conditional distribution by Theorem 3.6: For each pair $s < u$, there exists a probability kernel $\mathbf{p}_{su} : \mathbf{V} \to \mathbf{PM}(\mathbf{V})$ with

$$\mathbb{P}[X_u \in \bullet | X_t = x_t] =_{\text{a.s.}} \mathbf{p}_{tu}(\bullet, x_t) \qquad \text{whenever } t < u \; . \tag{5.43}$$

Suppose X is a Markov process and $\mathbf{p}_{su}$ is the conditional distribution above for two indices $s < u$. If we pick a third index t inbetween, $s < t < u$, we can condition X_t on X_s, and X_u in turn on X_t, which yields kernels $\mathbf{p}_{ts}$ and $\mathbf{p}_{ut}$. If we marginalize out X_t, we must recover the conditional distribution of X_u given X_s, which means the kernels must satisfy

$$\mathbf{p}_{su}(\bullet, x_s) = \int_{\mathbf{V}} \mathbf{p}_{tu}(\bullet, x) \mathbf{p}_{st}(dx, x_s) \tag{5.44}$$

It is customary to denote this marginalization as a product of kernels: For two kernels $\mathbf{p}$ and $\mathbf{q}$, write

$$(\mathbf{p} \cdot \mathbf{q})(A, x) := \int \mathbf{p}(A, y) \mathbf{q}(dy, x) \; . \tag{5.45}$$

Hence, the product of two probability kernels $\mathbf{V} \to \mathbf{PM}(\mathbf{V})$ is again a probability kernel $\mathbf{V} \to \mathbf{PM}(\mathbf{V})$. Using the product notation, (5.44) becomes

$$\mathbf{p}_{su} = \mathbf{p}_{tu} \cdot \mathbf{p}_{st} \qquad \text{whenever } s < t < u \; . \tag{5.46}$$

This identity is called the **Chapman-Kolmogorov equation**.

It is not hard to show that the product operation on kernels is associative. If $\mathcal{P}$ is the family of kernels $\mathbf{p}_{su}$ for a given Markov process, the product of any two elements is again a kernel in $\mathcal{P}$. Thus, $\mathcal{P}$ is a set that is closed under an associative binary operation (the product), and hence a **semigroup** of probability kernels. (Note that this product operation is *not* commutative, i.e. in general $\mathbf{p} \cdot \mathbf{q} \neq \mathbf{q} \cdot \mathbf{p}$.)

So far, we have seen that the conditional probabilities of a Markov process form a semigroup satisfying the Chapman-Kolmogorov equation. This statement can be strengthened considerably: Essentially, each semigroup of kernels satisfying the Chapman-Kolmogorov equation defines a Markov process, and vice versa. We have to distinguish two cases, however: Index sets U that possess a smallest element (such as $\mathbb{R}_+$) and those which do not. We focus on the former case, since it is much more common: Suppose U has a smallest element u_0. In this case, the kernels $\mathbf{p}_{su}$ are well-defined whenever $u > u_0$, but for $u = u_0$, the process has no past. Its behavior at u_0 is hence characterized by the distribution

$$P_0 := \mathcal{L}(X_{u_0}) \ , \tag{5.47}$$

which is called the **initial distribution** of X.

5.24 Theorem [Markov processes and semigroups]. *Let $\mathbf{V}$ be a Borel space and U a totally ordered set with smallest element u_0. Let*

$$\mathcal{P} = \{\mathbf{p}_{su} \colon \mathbf{V} \to \mathbf{PM}(\mathbf{V}) | s, u \in U \text{ and } s < u\} \tag{5.48}$$

be a semigroup of kernels satisfying the Chapman-Kolmogorov condition (5.46), and P_0 a probability measure on $\mathbf{V}$. Then there exists a Markov process X with index set U and state space $\mathbf{V}$ that satisfies (5.43) for the kernels in $\mathcal{P}$ and has initial distribution P_0. Its law is uniquely determined by $\mathcal{P}$ and P_0.

Conversely, a Markov process with index set U and state space $\mathbf{V}$ uniquely uniquely determines a family of probability kernels $\mathbf{p}_{su}$ via (5.43) that satisfies the Chapman-Kolmogorov equations. ◁

Proof. Suppose $\mathcal{P}$ and P_0 are given. Let u_0 be the smallest element of U and $u_1 < \ldots < u_n$ an arbitrary number of points in U. The kernels $\mathbf{p}_{u_i u_{i+1}}$, for $i = 1, \ldots, n-1$, and P_0 then uniquely define a probability measure $P_{\{u_0,\ldots,u_n\}}$ on $\mathbf{V}^{n+1}$; we can additionally integrate out the first dimension to obtain a measure on $P_{\{u_1,\ldots,u_n\}}$ on $\mathbf{V}^n$. A (somewhat lengthy but straightforward) computation shows that, if the kernels satisfy the Chapman-Kolmogorov equation, the family $P_{\{u_1,\ldots,u_n\}}$, for all $u_1, \ldots, u_n \in U$, is projective. By Theorem 5.10, it uniquely defines the law of a stochastic process X with values in $\mathbf{V}^U$. It is again straightforward to show that, since the kernels satisfy the Chapman-Kolmogorov equations, the process X satisfies (5.43), and is hence Markov.

We have already shown the converse statement in the derivation above. □

5.6. Processes with stationary and independent increments

We will now discuss a particularly important class of Markov processes, namely processes with stationary, independent increments. A stochastic process $X = (X_u)_{u \in \mathbb{R}_+}$ is called **stationary** if $(X_u) \overset{\mathrm{d}}{=} (X_{u+t})$ for all $t \in \mathbb{R}_+$. Thus, "shifting the time axis" by an arbitrary finite offset t leaves the law of the process invariant. In particular, the variables X_u all have identical marginal distribution $\mathcal{L}(X_u)$. This means, for instance, that Brownian motion is not stationary.

Rather than in processes that are themself stationary, we are interested in the following in processes whose increments—roughly, the individual steps of the process—are stationary as a random sequence. To define the notion of an increment, we need an addition operation on the state space, and these processes are hence usually defined for Euclidean state spaces $\mathbf{V} = \mathbb{R}^d$.

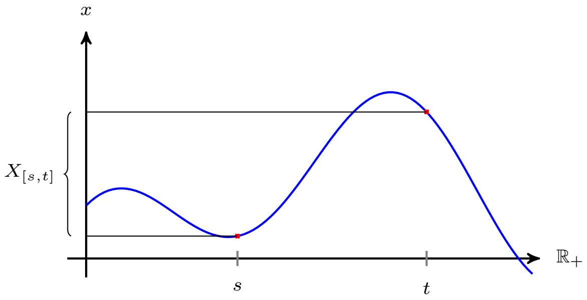


Figure 5.5. Increment $X_{[s,t]}$ of a process X over the interval $[s,t]$.

5.25 Definition. Let $X = (X_u)_{u\in\mathbb{R}_+}$ be a stochastic process with state space $\mathbb{R}^d$. The random variables

$$X_{[s,t]} := X_t - X_s \qquad \text{for } s < t \;, \tag{5.49}$$

are called the **increments** of X. A process has **independent increments** if

$$X_{[s,t]} \perp\!\!\!\perp X_{[u,v]} \qquad \text{whenever } [s,t) \cap [u,v) = \varnothing \;. \tag{5.50}$$

The increments are **stationary** if

$$X_{[s,t]} \overset{\mathrm{d}}{=} X_{[s+u,t+u]} \qquad \text{for all } s < t, u \in \mathbb{R}_+ \;. \tag{5.51}$$

◁

By the definition of increments, any value X_u of a process X is of the form

$$X_u =_{\text{a.s.}} X_{[t,u]} + X_t \qquad \text{for any } 0 \leq t < u \;. \tag{5.52}$$

If the increments are independent, we have for any $s < t < u$

$$X_u =_{\text{a.s.}} X_{[t,u]} + X_t \perp\!\!\!\perp_{X_t} X_t - X_{[s,t]} =_{\text{a.s.}} X_s \;, \tag{5.53}$$

so $X_u \perp\!\!\!\perp_{X_t} X_s$, which is just the definition (5.41) of a Markov process. Thus, processes with independent increments are Markov. If a process has stationary increments, (5.51) implies that for every $t \geq 0$, there is a probability measure

$$\mu_t := \mathcal{L}(X_{[0,t]}) \qquad \text{with} \qquad \mathcal{L}(X_{[s,t]}) = \mu_{t-s} \qquad \text{for all } s < t \;. \tag{5.54}$$

Since $X_{[t,t]} =_{\text{a.s.}} 0$, we have $\mu_0 = \delta_0$.

The next theorem below shows that a process with stationary and independent increments is a Markov process for which all kernels $\mathbf{p}_{su}$ are of the form

$$\mathbf{p}_{su}(A,x) = \mu_{u-s}(A - x) \;, \tag{5.55}$$

where μ_{u-s} is the measure defined in (5.54). If we substitute (5.55) into the Chapman-Kolmogorov equation (5.46), the equation takes the form

$$\mu_{s+t} = \mu_s * \mu_t \qquad \text{for all } s,t \in \mathbb{R}_+ \;, \tag{5.56}$$

where $*$ denotes the convolution[4] operator. A family $(\mu_u)_{u\geq 0}$ of probability measures satisfying (5.56) is called a **convolution semigroup** of measures.

[4] Recall that the the **convolution** of two probability measures P and Q on $\mathbb{R}^d$ is the probability measure

$$(P * Q)(A) := \int_{\mathbb{R}^d} P(A - x)Q(dx) = \int_{\mathbb{R}^d} Q(A - x)P(dx) \;. \tag{5.57}$$

5.26 Theorem [Processes with stationary and independent increments]. *A stochastic process $(X_u)_{u\in\mathbb{R}_+}$ with values in $\mathbb{R}^d$ has stationary and independent increments if and only if it is a Markov process defined by an initial distribution $P_0 = \mathcal{L}(X_0)$ and conditional probabilities*

$$\mathbb{P}[X_u \in A|X_s = x] =_{\text{a.s.}} \mu_{u-s}(A - x)\ , \tag{5.58}$$

for a convolution semigroup of measures $(\mu_u)_{u\geq 0}$. ◁

Proof. We use a simple consequence of Lemma 3.12: For any two random variables Y and Z, we have

$$\mathbb{P}[Z - Y \in A|Y = y] = \mathbb{P}[Z \in A + y|Y = y]\ . \tag{5.59}$$

To show this, simply set $f(y,z) := \mathbb{I}_A(z-y)$ in Lemma 3.12. To prove the theorem, suppose first X is Markov with kernels satisfying (5.58). Then

$$\mathbb{P}[X_t - X_s \in A|X_s = x] \overset{(5.59)}{=} \mathbb{P}[X_t \in A + x|X_s = x] = \mathbf{p}_{st}(A + x, x) \overset{(5.58)}{=} \mu_{t-s}(A)\ ,$$

so the increment $X_{[s,t]} = X_t - X_s$ is independent of X_s (and hence of all previous increments), so X has independent increments. Additionally, the law of $X_{[s,t]}$ depends on $[s,t]$ only through $(t-s)$, and the increments are hence stationary.

Conversely, suppose X has stationary and independent increments. We have already argued above that X is then Markov, and that $\mathcal{L}(X_{[s,t]}) = \mu_{t-s}$ for some measures μ_u. According to Theorem 5.24, there is an initial distribution P_0 and a semigroup of kernels defining the process. What remains to be shown is that these kernels satisfy (5.58). They do:

$$\begin{aligned}\mathbb{P}[X_t \in A|X_s = x] = \mathbb{P}[X_{[s,t]} + X_s \in A|X_s = x] &\overset{(5.59)}{=} \mathbb{P}[X_{[s,t]} \in A - x|X_s = x] \\ \overset{\text{ind. incr.}}{=} \mathbb{P}[X_{[s,t]} \in A - x] &\overset{\text{def. of }\mu_u}{=} \mu_{t-s}(A - x)\ ,\end{aligned} \tag{5.60}$$

which is precisely (5.58). □

Textbook versions of this proof tend to be lengthy, often filling several pages. The argument above owes its remarkable simplicity to identity (5.59); as far as I can tell, this insight is due to Kallenberg (see [K, Proposition 8.5]).

5.27 Example [Brownian convolution semigroup]. Brownian motion has stationary and independent increments, and is hence defined by an initial distribution and a convolution semigroup $(\mu_u)_{u\geq 0}$ of measures: Each μ_u is the Gaussian distribution on $\mathbb{R}$ (or, more generally on $\mathbb{R}^d$) with the centered Gaussian density

$$g_t(x) := \frac{1}{(\sqrt{2\pi t})^d}\exp\Big(-\frac{\|x\|^2}{2t}\Big)\ . \tag{5.61}$$

The family $(\mu_u)_{u\geq 0}$ is called the **Brownian convolution semigroup**. As an exercise, I recommend to convince yourself that this indeed defines Brownian motion as in Definition 5.19. ◁

Recall also that the sum $X + Y$ of two independent random variables X and Y in $\mathbb{R}_d$ has law $\mathcal{L}(X + Y) = \mathcal{L}(X) * \mathcal{L}(Y)$.

5.7. Gaussian processes

A Gaussian process is a stochastic process whose FDDs are all Gaussian. Thus, Brownian motion is a Gaussian process. However, by increasing the covariance between points, we can force the process to have much smoother paths than Brownian motion; in particular, the paths of Gaussian processes can be differentiable.

5.28 Definition. A **Gaussian process** is a stochastic process X with index set $\mathbb{R}_+$ and state space $\mathbb{R}^d$ whose finite-dimensional distributions

$$P_t := \mathcal{L}(X_{u_1}, \ldots, X_{u_n}) \qquad \text{for all finite sets } t := \{u_1, \ldots, u_n\} \in \mathbb{R}_+ \tag{5.62}$$

are Gaussian distributions. ◁

Note that the form of the FDDs implies we are working in the product space setup: Each P_t is the image of $\mathcal{L}(X)$ under the product space projector pr_t; the relevant extension theorem is hence Kolmogorov's theorem Theorem 5.10. Each distribution P_t lives on $(\mathbb{R}^d)^t$. To avoid a notational nightmare, we will assume $d = 1$ henceforth, so P_t is a measure on $\mathbb{R}^t$—just keep in mind that the results carry over immediately to $\mathbb{R}^d$.

Kolmogorov's extension theorem requires the measures P_t to be projective for the process to exist. Since each P_t is Gaussian, it is completely determined by its mean vector $m_t \in \mathbb{R}^t$ and its covariance matrix Σ_t. It is hence obvious to ask whether we can give a condition for projectivity in terms of the means and variances. That is indeed possible: Recall that a function $\mathbf{k}$ on $\mathbb{R}^2_+$ is **positive semidefinite** if, for any set $t = \{u_1, \ldots, u_n\} \in \mathbb{R}_+$, the matrix

$$\big(\mathbf{k}(u_i, u_j)\big)_{i,j \leq n} \tag{5.63}$$

is positive semidefinite.

5.29 Lemma. *For each finite subset $t \subset \mathbb{R}_+$, let P_t be a Gaussian distribution on $\mathbb{R}^t$ with mean m_t and covariance matrix Σ_t. The family (P_t) is projective (satisfies (5.14)) if and only if there exists a function $\mathbf{m} : \mathbb{R} \to \mathbb{R}$ and a positive semidefinite function $\mathbf{k} : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ such that*

$$m_t = (\mathbf{m}(u_1), \ldots, \mathbf{m}(u_n)) \qquad \textit{and} \qquad \Sigma_t := \big(\mathbf{k}(u_i, u_j)\big)_{i,j \leq n} \tag{5.64}$$

◁

Every Gaussian process is therefore uniquely determined by a pair of functions $\mathbf{m}$ and $\mathbf{k}$ and vice versa. If X is a Gaussian process with mean $\mathbf{m}$ and covariance function $\mathbf{k}$, we generically denote its law by $\mathrm{GP}(\mathbf{m}, \mathbf{k})$.

5.30 Exercise. Proof the "if" direction of Lemma 5.29. ◁

We have already encountered the example of Brownian motion, which is continuous but not smooth. There are various applications and modeling problems that use random functions, but require smoothness, for example:

- Spatial statistics: Gaussian processes with index set $\mathbb{R}^2$ (often called Gaussian random fields) are used here to model the distribution of smoothly varying quantity over a region, i.e. temperature as a function of location on a map.
- Bayesian regression: Gaussian process regression represents the solution of a regression problem (i.e. a smooth function) as an unknown quantity, which in a Bayesian approach is modeled as random. A Gaussian process is used as the prior distribution on functions.

5.31 Example [Squared-exponential covariance function]. Probably the most widely used covariance function in such problems is a squared exponential,

$$\mathbf{k}(x,y) := \exp\Big(-\frac{1}{2}\frac{(x-y)^2}{\sigma^2}\Big)\,. \tag{5.65}$$

The paths of the resulting Gaussian process are infinitely often continuously differentiable (almost surely); each derivative is again a Gaussian process [e.g. 1]. ◁

For applications such as spatial statistics and regression, it can be useful to consider Gaussian processes whose paths are almost surely in the space $\mathcal{L}_2(\mathbb{R}_+)$ of functions square-integrable with respect to Lebesgue measure—rather than in the set $\mathbf{C}(\mathbb{R}_+,\mathbb{R})$ we considered in the case of Brownian motion. (There is a deeper connection between Gaussian processes and L_2-spaces, since L_2 spaces are separable Hilbert spaces; the covariance function of a Gaussian process is a so-called Mercer kernel, and such kernels define Hilbert spaces.) We hence need an analogue of the Kolmogorov-Chentsov criterion (Corollary 5.18) for the Hilbert space $\mathcal{L}(\mathbb{R}_+)$, which is given by the following result of Prokhorov, which I just want to mention here without proof:

5.32 Theorem [Gaussian processes with paths in Hilbert space]. *Let (P_t) be a family of Gaussian FDDs defined by a pair $(\mathbf{m},\mathbf{k})$. Then $\mathcal{L}_2(\mathbb{R}_+)$ has outer measure one under $P = \varprojlim(P_t)$ if and only if*

$$\mathbf{m}\in\mathcal{L}_2(\mathbb{R}_+) \qquad \textit{and} \qquad \int_{\mathbb{R}_+}\mathbf{k}(u,u)du<\infty\,. \tag{5.66}$$

◁

Thus, a Gaussian process $\mathrm{GP}(\mathbf{m},\mathbf{k})$ almost surely has paths in $\mathcal{L}_2$ iff its parameter functions satisfy (5.66).

5.33 Remark [Ornstein-Uhlenbeck process]. Brownian motion is a Gaussian process, and at the same time Markov. If a Gaussian process has differentiable path, it can clearly not be Markov—which raises the question whether there is a non-Brownian but Gaussian Markov process. If $(X_u)_{u\in\mathbb{R}_+}$ is Brownian motion, the process defined by

$$Y_u := e^{-u}X_{\frac{1}{2}e^{2u}} \tag{5.67}$$

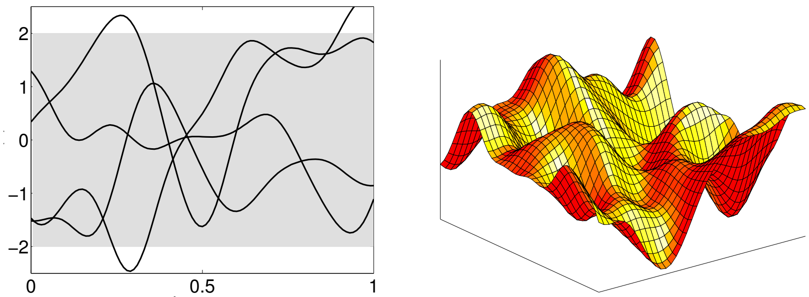


Figure 5.6. Samples from a Gaussian process with squared-exponential covariance function. *Left:* Several sample paths of a Gaussian process with index set $\mathbb{R}_+$. *Right:* A single path of Gaussian process with index set $\mathbb{R}^2$, as used e.g. in spatial statistics. (Illustrations from Rasmussen and Williams, "Gaussian processes for Machine Learning", MIT Press 2006.)

is called an **Ornstein-Uhlenbeck process**. This process is indeed Gaussian and Markov. It is also stationary (unlike Brownian motion, for which only the increments are stationary), and it can be shown that is essentially the only stationary Gaussian Markov process. See e.g. [K, Proposition 13.7]. ◁

5.8. The Poisson process

A point process on an uncountable set $\mathcal{X}$ is random countable subset of $\mathcal{X}$. More formally:

5.34 Definition. Let $(\mathcal{X}, \mathcal{A}_{\mathcal{X}})$ be an uncountable measurable space, such that the diagonal of $\mathcal{X} \times \mathcal{X}$ is measurable (cf. page 48). A **point process** on $\mathcal{X}$ is a random countable subset of $\mathcal{X}$, i.e. a random variable Π with values in the power set $2^{\mathcal{X}}$ satisfying $\#\Pi \leq \#\mathbb{N}$ almost surely. ◁

Note that a point process is defined as a random *set*, not a random *multiset*. That means we cannot distinguish points if they occur more than once: For example,

$$\{x, y, z\} \qquad \text{and} \qquad \{x, y, y, z\} \tag{5.68}$$

are distinct as multisets, but identical as sets. When we specify the law of a point process, we therefore typically require that no point occurs more than once, to avoid inconsistencies. For countably many points, "no point occurs more than once" comes down to a countable number of conditions of the form "$X \neq Y$ almost surely", which is why we require the diagonal to be measurable: $X \neq Y$ holds almost surely iff the event $\{X = Y\}$ is a null set, and this event is precisely the diagonal. Recall that, by Lemma 3.18, the diagonal is automatically meausurable whenever $\mathcal{X}$ is a Borel space.

5.35 Definition. A point process Π on $\mathcal{X}$ is called a **Poisson process** if, for every measurable set $A \in \mathcal{A}_{\mathcal{X}}$,

(1) the number of points $|\Pi \cap A|$ of Π in A is a Poisson random variable[5] and
(2) the variables $\Pi \cap A$ and $\Pi \cap \overline{A}$ are stochastically independent,

where $\overline{A}$ denotes the complement of A. ◁

[5] Recall that the Poisson distribution is the distribution we obtain from the series expansion $e^{\lambda} = \sum_{k=0}^{\infty} \lambda^k/k!$ of the exponential function: If we normalize by multiplication with $e^{-\lambda}$ and multiply in a point mass δ_k at each $k \in \mathbb{N} \cup \{0\}$, we obtain a probability measure

$$P_{\lambda}(\bullet) := \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \delta_k(\bullet) \tag{5.69}$$

on $\mathbb{N} \cup \{0, \infty\}$, called the **Poisson distribution** with parameter λ. The definition implies $P_0 = \delta_0$, and we use the convention $P_{\infty}(\{\infty\}) = 1$. Two very useful properties are:
Additivity: If $N_1 \sim \text{Poisson}(\alpha_1)$ and $N_2 \sim \text{Poisson}(\alpha_2)$, then

$$(N_1 + N_2) \sim \text{Poisson}(\alpha_1 + \alpha_2) \tag{5.70}$$

if and only if N_1 and N_2 are independent.
Thinning: If $N \sim \text{Poisson}(\alpha)$ and $J_1, \ldots, J_N \sim_{\text{iid}} \text{Bernoulli}(p)$, then

$$\textstyle\sum_{i=1}^{N} J_i \sim \text{Poisson}(p\alpha) \;, \tag{5.71}$$

i.e. the number of successes in a Poisson number of coin flips is Poisson.

We do not know yet whether such an object Π exists, but suppose for the moment that it does. For each set $A \in \mathcal{A}_{\mathcal{X}}$, define

$$\mu(A) := \mathbb{E}\big[|\Pi \cap A|\big] \;. \tag{5.72}$$

The definition of the Poisson process then implies μ must be a measure on $(\mathcal{X}, \mathcal{A}_X)$. It is called the **mean measure** of Π.

5.36 Exercise. Deduce from Definition 5.35 that μ is a measure on $\mathcal{X}$. $\triangleleft$

Since the Poisson distribution is completely specified by its mean, the law of the Poisson process Π—if the process exists—is completely determined by μ. We can hence parametrize Π by μ, and we use the notation

$$\Pi^{\mu} := \text{ Poisson process with mean measure } \mu \;. \tag{5.73}$$

You may also encounter references to a Poisson process specified by a "rate": If $\mathcal{X} = \mathbb{R}^d$ and the mean measure is of the form $\mu = c\lambda$, where λ is Lebesgue measure, the constant c is called the **rate** of the process.

5.37 Theorem. *Let $(\mathcal{X}, \mathcal{A}_x)$ be a measurable space such that $\mathcal{X} \times \mathcal{X}$ has measurable diagonal, and let μ be an atomless measure on $\mathcal{X}$ satisfying*

$$\mu = \sum_{n=1}^{\infty} \mu_n \qquad \textit{for some sequence of measures } \mu_n \textit{ with } \mu_n(\mathcal{X}) < \infty \;. \tag{5.74}$$

Then the random set Π generated by Algorithm 5.38 below is a Poisson process on $\mathcal{X}$ with mean measure μ. $\triangleleft$

Recall that **atomless** means $\mu\{x\} = 0$ for every one-point set $\{x\}$ in $\mathcal{X}$. Condition (5.74) says that μ can be infinite, but must be decomposable into a superposition of an at most countably infinite number of finite measures. That is obviously true if μ is σ-finite, but also includes a large class of measures which are not σ-finite; for example, the projection $\text{pr}\lambda^2$ of Lebesgue measure λ^2 on $\mathbb{R}^2$ onto $\mathbb{R}$ is not σ-finite, but it does satisfy (5.74) (why?).

5.38 Algorithm [Sampling a Poisson process].
(a) *If $\mu(\mathcal{X})$ is finite:*
(i) *Generate a random integer $N \sim \text{Poisson}(\mu)$.*
(ii) *Sample N points $X_1, \ldots, X_N$ i.i.d. from the distribution $\mu/\mu(\mathcal{X})$.*
(iii) *Set $\Pi := \{X_1, \ldots, X_N\}$.*
(b) *If $\mu(\mathcal{X})$ is infinite:*
(i) *Decompose μ according to (5.74) into finite measures μ_n.*
(ii) *Generate a random set Π_n independently from each μ_n as in* (a)*.*
(iii) *Set $\Pi := \cup_{n\in\mathbb{N}}\Pi_n$.*

Proof. Throughout, A is any measurable set in $\mathcal{X}$, and $\overline{A}$ its complement. *Case (a): μ finite.* We sample Π according to Algorithm 5.38. For each of the N points in Π, the probability that it ends up in a given measurable set A is $\mu(A)/\mu(\mathcal{X})$. By the thinning property (5.71), with $J_i := \mathbb{I}_A(X_i)$, the number of points in A is a Poisson random variable N_A with

$$\mathbb{E}[N_A] = \frac{\mu(A)}{\mu(\mathcal{X})}\mathbb{E}[N] = \frac{\mu(A)}{\mu(\mathcal{X})}\mu(\mathcal{X}) = \mu(A) \;. \tag{5.75}$$

By the same device, the number of points $N_{\bar{A}}$ in $\overline{A}$ is Poisson$(\mu(\mathcal{X}) - \mu(A))$. Since $N = N_A + N_{\bar{A}}$, the additivity property (5.70) implies $N_A \perp\!\!\!\perp N_{\bar{A}}$, which implies $\Pi \cap A \perp\!\!\!\perp \Pi \cap \overline{A}$. Thus, Π is a Poisson process.

Case (b): μ infinite. Since each μ_n in (5.74) is finite, the random sets Π_n are independent Poisson processes, by the argument above. Each Π_n is Poisson and

$$\Pi \cap A = (\cup_n \Pi_n) \cap A = \cup_n (\Pi_n \cap A) \;, \tag{5.76}$$

which implies $\Pi \cap A$ and $\Pi \cap \overline{A}$ are independent. What remains to be shown is that $|\Pi \cap A|$ is Poisson$(\mu(A))$. Since the sets Π_n each contain an a.s. finite number of points sampled independently from an atomless distribution, they are almost surely disjoint[6], and hence

$$|\Pi \cap A| =_{\text{a.s.}} \textstyle\sum_n |\Pi_n \cap A| \;. \tag{5.77}$$

By Borel-Cantelli, a sum of independent variables almost surely diverges if and only if the sum of its expectations diverges. The variables $|\Pi_n \cap A|$ are independent, each with expectation $\mu_n(A)$, and $\sum \mu_n(A) = \mu(A)$. We hence distinguish two cases:

(1) If $\mu(A) < \infty$, $|\Pi \cap A|$ is almost surely finite, and Poisson$(\mu(A))$ by additivity.
(2) If $\mu(A) = \infty$, then $|\Pi \cap A|$ is infinite almost surely, and hence Poisson(∞)

Thus, $|\Pi \cap A|$ is indeed Poisson$(\mu(A))$ for every measurable set A. In summary, Π is a Poisson process with mean measure μ as claimed. □

Since we have shown how to sample the Poisson process, we have in particular given a constructive proof of its existence:

5.39 Corollary. *Let $(\mathcal{X}, \mathcal{A}_{\mathcal{X}})$ be a measurable space such that $\mathcal{X} \times \mathcal{X}$ has measurable diagonal, and let μ be an atomless measure on $\mathcal{X}$ satisfying* (5.74). *Then the Poisson process Π^μ on $\mathcal{X}$ exists.* ◁

Poisson processes are remarkably robust; almost regardless of what you do to them, they remain Poisson:

5.40 Corollary. *Let μ be a measure and $(\nu_n)_{n \in \mathbb{N}}$ a sequence of measures, all of which are atomless and satisfy* (5.74). *Let $\phi : \mathcal{X} \to \mathcal{X}$ be a measurable mapping. Then the following holds:*

$$\phi(\Pi^\mu) = \Pi^{\phi(\mu)} \qquad \textit{if } \mu \textit{ is } \sigma\textit{-finite} \;. \tag{5.78}$$

$$\Pi^\mu \cap A = \Pi^{\mu(\bullet \cap A)} \qquad \textit{for any set } A \in \mathcal{A}_{\mathcal{X}} \;. \tag{5.79}$$

$$\textstyle\bigcup_n \Pi^{\nu_n} = \Pi^{\sum_n \nu_n} \tag{5.80}$$

◁

Proof. Homework. □

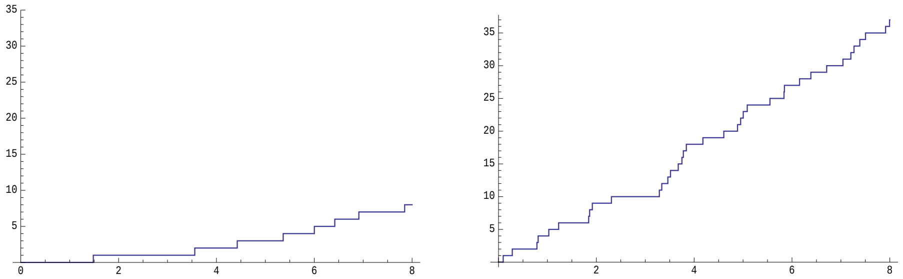
Depending on the context, the term Poisson process may also refer to a time-dependent process, i.e. a random (piece-wise constant) function rather than a random set of points. A process $X = (X_u)$ of this form is defined as follows: For a Poisson process Π on $\mathbb{R}_+ \times \mathbb{R}_+$, define a continuous-time process (X_u) on $\mathbb{R}_+$ as

$$X_u := \textstyle\sum_{(x_1, x_2) \in \Pi | x_1 < u} x_2 \;. \tag{5.81}$$

In other words, if the random set Π contains a point (x_1, x_2), then at time $u = x_1$, the process X increases by x_2 (see Figure 5.7). By the independence property of the Poisson process (Definition 5.35(ii)), the increments of X are independent, so

[6] This is the statement for which measurability of the diagonal is required.

Figure 5.7. Sum of counts of Poisson processes on $\mathbb{R}_+$.
Left: Mean measure λ (i.e. rate 1). *Right*: Mean measure 5λ (i.e. rate 5).



X is Markov. You can easily verify that its increments are also stationary, and that the convolution semigroup $(\mu_u)_{u\geq 0}$ defining this process according to Theorem 5.26 consists simply of the Poisson distributions with parameter $\lambda = u$, that is, $\mu_u = P_u$ (as defined in (5.69)) for all $u \geq 0$.

5.41 Remark. We have not discussed regularity of paths for processes with stationary and independent increments. It can be shown that the only such process whose paths are almost surely continuous is Brownian motion; for more general processes, we have to permit jumps. The adequate choice of for a set of paths turns out to be the space of rcll functions, which we have already encountered in Section 1.6. A process with stationary and independent increments whose paths are almost surely rcll is called a **Lévy process**. The famous Lévy-Khinchine theorem [K, Corollary 15.7] shows that, roughly speaking, a process is Lévy if and only if it is of the form

$$\begin{aligned} X = \ & \text{non-random linear function} + \text{constantly scaled Brownian motion} \\ & \qquad\qquad\qquad + \text{jumps represented by a Poisson process} \ . \end{aligned}$$

In other words, up to a fixed "drift", a Lévy process is always a superposition of a (continuous) Brownian component and a (pure-jump) Poisson component. In this sense, the Brownian motion and the Poisson process are the two fundamental Lévyprocesses from which all other Lévyprocesses can be derived. ◁

Bibliography

Main references

[A&B] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis.* 3rd. Springer, 2006.

[B] H. Bauer. *Probability Theory.* W. de Gruyter, 1996.

[Probability I] J. Jacod and P. Protter. *Probability Essentials.* Springer, 2004.

[K] O. Kallenberg. *Foundations of Modern Probability.* 2nd. Springer, 2001.

Other references

[1] R. J. Adler and J. E. Taylor. *Random Fields and Geometry.* Springer, 2007.

[2] P. Billingsley. *Convergence of Probability Measures.* J. Wiley & Sons, 1999.

[3] P. Billingsley. *Probability and Measure.* J. Wiley & Sons, 1995.

[4] E. Cinlar. *Probability and Stochastics.* Springer, 2011.

[5] D. H. Fremlin. *Measure Theory.* Torres Fremlin, 2003–2006.

[6] J. F. C. Kingman. *Poisson Processes.* Oxford University Press, 1993.

[7] A. Klenke. *Probability Theory.* 2nd. Springer, 2014.

[8] J. Pfanzagl and W. Pierlo. *Compact Systems of Sets.* Springer, 1966.

[9] D. Pollard. *A User's Guide to Measure Theoretic Probability.* Cambridge University Press, 2002.

[10] A. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes.* Springer, 1996.

Index