2 Foundations of Probability (†)

This chapter covers the fundamental concepts of measure-theoretic probability on which the remainder of this book relies. Readers familiar with this topic can safely skip the chapter, but perhaps a brief reading would yield some refreshing perspectives. Measure-theoretic probability is often viewed as a necessary evil, to be used when a demand for rigor combined with continuous spaces breaks the simple approach we know and love from high school. We claim that measure-theoretic probability offers more than annoying technical machinery. In this chapter we attempt to prove this by providing a non-standard introduction. Rather than a long list of definitions, we demonstrate the intuitive power of the notation and tools. For those readers with little prior experience in measure theory this chapter will doubtless be a challenging read. We think the investment is worth the effort, but a great deal of the book can be read without it, provided one is willing to take certain results on faith.

2.1 Probability spaces and random elements

Probability theory is a latecomer to the party of mathematical study. While the ancient Greeks and Romans certainly gambled, there is no evidence they ever formally analyzed the probabilistic nature of the games they played. But probability does have its origins in the study of games of chance and gambling, with early steps taken in the 16th and 17th centuries by famous mathematicians and physicists such as Niccoló Tartaglia, Gerolamo Cardano, Blaise Pascal, Pierre Fermat, Christian Huygens and Jacob Bernoulli. The thrill of gambling comes from the fact that the bet is placed on future outcomes that are uncertain at the time of the gamble. A central question in gambling is the fair value of a game. This can be difficult to answer for all but the simplest games. As an illustrative example, imagine the following moderately complex game: I throw a dice. If the result is four, I throw two more dice, otherwise I throw one dice only. Looking at the newly thrown dice (one or two), I repeat the same, for a total of three rounds (and at most seven dice throws in total). Afterwards, I pay you the sum of the values on the faces of the dice. How much are you willing to pay to play this game with me?

The fact that the number of dice used is random appears to create a messy

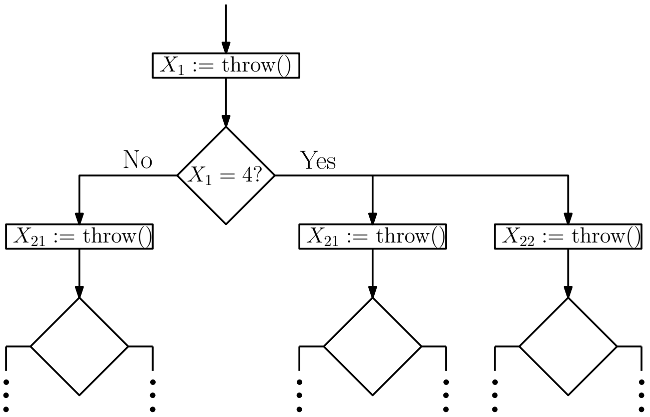
$X_1 := \text{throw}()$
No
$X_1 = 4?$
Yes
$X_{21} := \text{throw}()$
$X_{21} := \text{throw}()$
$X_{22} := \text{throw}()$

Figure 2.1 The initial phase of a gambling game with a random number of dice rolls. Depending on the outcome of a dice roll, one or two dice are rolled for a total of three rounds. The number of dice used will then be random in the range of three to seven.

situation where the outcomes have a complicated dependency structure. This situation is not unusual, with many examples of practical interest exhibiting the same random interdependency between outcomes. The fundamental idea in modern probability is aimed at removing this complication.

Instead of rolling the dice one by one, imagine that sufficiently many dice were rolled before the game has even started. For our game we need to roll seven dice, because this is the maximum number that might be required (see Fig. 2.1). With the dice all rolled, the game can be emulated easily by ordering the dice and revealing the outcomes sequentially. Then the value of the first dice in the chosen ordering is the outcome of the dice in the first round. If we see a four, we look at the next two dice in the ordering, otherwise we look at the single next dice.

⚠ This approach separates the randomness (rolls of the dice) from the mechanism that produces values based on the random outcomes. This idea is one of the cornerstones of modern probability as proposed by Kolmogorov.

By taking this approach we get a simple calculus for the probabilities of all kinds of **events**. Rather than directly calculating the likelihood of each payoff, we first consider the probability of any single outcome of the dice. Since there are seven dice, the set of all possible outcomes is $\Omega = \{1, \dots, 6\}^7$. Because all outcomes are equally probable the probability of any $\omega \in \Omega$ is $(1/6)^7$. The probability of the game payoff taking value v can then be evaluated by calculating the total probability assigned to all those outcomes $\omega \in \Omega$ that would result in the value of v. In principle, this is trivial to do thanks to the separation of everything that is probabilistic from the rest. The set Ω is called the **outcome space** and its elements are the **outcomes**. Fig. 2.2 illustrates this idea: Random outcomes are

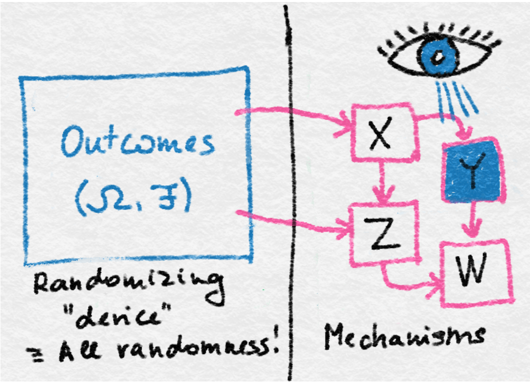

Figure 2.2 A key idea in probability theory is the separation of sources of randomness from game mechanisms. A mechanism creates values from the elementary random outcomes, some of which are visible for observers, while others may remain hidden.

generated on the left, while on the right, various mechanisms are used to arrive at values, some of which of these values may be observed and some not.

There will be much benefit from being a little more formal about how we come up with the value of our artificial game. For this note that the process by which the game gets its value is a function X that maps Ω to the set of natural numbers $\mathbb{N}$ (simply, $X : \Omega \to \mathbb{N}$). While we view the value of the game as random, this map is deterministic. We find it ironic that functions of this type (from the outcome space to subsets of the reals) are called **random variables**. They are neither random nor variables in a programming language sense. The randomness is in the argument that X is acting on, producing randomly changing results. Later we will put a little more structure on random variables, but for now it suffices to think of them as maps from the outcome space to the naturals, or more generally, to the reals.

We will follow the standard convention in probability theory where random variables are denoted by capital letters. Be warned that capital letters are also used for other purposes as demanded by different conventions.

Pick some number $v \in \mathbb{N}$. What is the probability of seeing $X = v$? As described above, this probability is $(1/6)^7$ times the size of the set $X^{-1}(v) = \{\omega \in \Omega : X(\omega) = v\}$. The set $X^{-1}(v)$ is called the **preimage** of v under X. More generally, the probability that X takes its value in some set $A \subseteq \mathbb{N}$ is given by $(1/6)^7$ times the cardinality of $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$, where we have overloaded the definition of X^{-1} to set-valued inputs.

Notice in the previous paragraph we only needed probabilities assigned to subsets of Ω regardless of the question asked. To make this a bit more general, let us introduce a map $\mathbb{P}$ that assigns probabilities to certain subsets of Ω. The intuitive meaning of $\mathbb{P}$ is as follows. Random outcomes are generated in Ω. The probability that an outcome falls into a set $A \subset \Omega$ is $\mathbb{P}(A)$. If A is not in the domain of $\mathbb{P}$, then there is no answer to the question of the probability of the outcome falling in A. But let's postpone the discussion of why $\mathbb{P}$ should be restricted to only certain subsets of Ω later. In the above example with the dice, the set of subsets in the domain of $\mathbb{P}$ are not restricted and, in particular, for any subset $A \subseteq \Omega$, $\mathbb{P}(A) = (1/6)^7|A|$.

With this new notation, the answer to the question of what is the probability of seeing X taking the value of v becomes $\mathbb{P}\left(X^{-1}(v)\right)$. To minimize clutter, the more readable notation for this is $\mathbb{P}(X = v)$. It is important to realize, however, that this familiar form is just a shorthand for $\mathbb{P}\left(X^{-1}(v)\right)$. More generally, we also use

$$\mathbb{P}(\text{predicate}(X, U, V, \dots)) = \mathbb{P}(\{\omega \in \Omega : \text{predicate}(X, U, V, \dots) \text{ is true}\})$$

with any predicate (an expression evaluating to true or false) where $X, U, V, \dots$ are functions with domain Ω.

What properties should $\mathbb{P}$ satisfy? It seems reasonable to expect the probability that *something* happens is equal to one, which is equivalent to saying that $\mathbb{P}$ is defined for Ω and $\mathbb{P}(\Omega) = 1$. Second, probabilities should be nonnegative so $\mathbb{P}(A) \geq 0$ for any $A \subset \Omega$ on which $\mathbb{P}$ is defined. Let $A^c = \Omega \setminus A$ be the **complement** of A. Then we should expect that $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ (negation rule). Finally, if A, B are disjoint so that $A \cap B = \emptyset$ and $\mathbb{P}(A)$, $\mathbb{P}(B)$ and $\mathbb{P}(A \cup B)$ are all defined, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. This is called the **finite additivity property**.

Let $\mathcal{F}$ be the set of subsets of Ω on which $\mathbb{P}$ is defined. It would seem silly if $A \in \mathcal{F}$ and $A^c \notin \mathcal{F}$, since $\mathbb{P}(A^c)$ could simply be defined by $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$. Similarly, if $\mathbb{P}$ is defined on disjoint sets A and B, then it makes sense if $A \cup B \in \mathcal{F}$. By a logical jump, we will also require the additivity property to hold *(i)* regardless of whether the sets are disjoint and *(ii)* even for **countably infinitely many** sets If $\{A_i\}_i$ is a collection of sets and $A_i \in \mathcal{F}$ for all $i \in \mathbb{N}$, then $\cup_i A_i \in \mathcal{F}$ and if these sets are disjoint, $\mathbb{P}(\cup_i A_i) = \sum_i \mathbb{P}(A_i)$. A set of subsets that satisfies all these properties is called a **$\boldsymbol{\sigma}$-algebra**, which is pronounced 'sigma-algebra' and sometimes also called a σ-field.

DEFINITION 2.1 A set $\mathcal{F} \subseteq 2^\Omega$ is a σ-algebra if $\Omega \in \mathcal{F}$ and $A^c \in \mathcal{F}$ for all $A \in \mathcal{F}$ and $\cup_i A_i \in \mathcal{F}$ for all $\{A_i\}_i$ with $A_i \in \mathcal{F}$ for all $i \in \mathbb{N}$. A function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is a **probability measure** if $\mathbb{P}(\Omega) = 1$ and for all $A \in \mathcal{F}$, $\mathbb{P}(A) \geq 0$ and $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ and $\mathbb{P}(\cup_i A_i) = \sum_i \mathbb{P}(A_i)$ for all countable collections of disjoint sets $\{A_i\}_i$ with $A_i \in \mathcal{F}$ for all i. If $\mathcal{F}$ is a σ-algebra and $\mathcal{G} \subset \mathcal{F}$ is also a σ-algebra, then we say $\mathcal{G}$ is a **sub-$\boldsymbol{\sigma}$-algebra** of $\mathcal{F}$.

At this stage, the reader may rightly wonder about why we introduced the notion of sub-σ-algebras. The answer should become clear quite soon. The elements of $\mathcal{F}$ are called **measurable sets**. They are measurable in the sense that $\mathbb{P}$ assigns values to them. The pair $(\Omega, \mathcal{F})$ alone is called a **measurable space**, while the triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**. If the condition that $\mathbb{P}(\Omega) = 1$ is lifted, then $\mathbb{P}$ is called a **measure**. If the condition that $\mathbb{P}(A) \geq 0$ is also lifted, then $\mathbb{P}$ is called a **signed measure**. We note in passing that both for measures and signed measures it would be unusual to use the symbol $\mathbb{P}$, which is mostly reserved for probabilities.

Random variables lead to new probability measures. In particular, $\mathbb{P}_X(A) = \mathbb{P}\left(X^{-1}(A)\right)$ is a probability measure defined for all the subsets A of $\mathbb{N}$ for which $\mathbb{P}\left(X^{-1}(A)\right)$ is defined. The probability measure $\mathbb{P}_X$ is called the **law** of X or the **pushforward** measure of $\mathbb{P}$ under X.

⚠ The significance of the pushforward measure $\mathbb{P}_X$ is that any probabilistic question concerning X can be answered from the knowledge of $\mathbb{P}_X$ alone. Even Ω and the details of the map X are not needed. This is often used as an excuse to not even mention the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

It is worth noting that if we keep X fixed, but change $\mathbb{P}$ (for example, by switching to loaded dice), then the measure induced by X changes. We will often use arguments that do exactly this, especially when proving lower bounds on the limits of how well bandit algorithms can perform.

The astute reader would have noticed that we skipped over some details. Measures are defined as functions from a σ-algebra to $\mathbb{R}$, so if we want to call $\mathbb{P}_X$ a measure, then its domain $\{A \subset \mathbb{N} : X^{-1}(A) \in \mathcal{F}\}$ better be a σ-algebra. This is indeed the case, and in fact holds more generally. You will show in Exercise 2.3 that for functions $X : \Omega \to \mathcal{X}$ with $\mathcal{X}$ arbitrary, the collection $\{A \subset \mathcal{X} : X^{-1}(A) \in \mathcal{F}\}$ is a σ-algebra.

It will be useful to generalize our example a little by allowing X to take on values in sets other than the reals. For example, the range could be vectors or abstract objects like sequences. Let $(\Omega, \mathcal{F})$ be a measurable space, $\mathcal{X}$ be an arbitrary set and $\mathcal{G} \subseteq 2^{\mathcal{X}}$. A function $X : \Omega \to \mathcal{X}$ is called a **$\mathcal{F}/\mathcal{G}$-measurable map** if $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{G}$. Note that $\mathcal{G}$ need not be a σ-algebra. When $\mathcal{F}$ and $\mathcal{G}$ are obvious from the context, X is called a **measurable map**. What are the typical choices for $\mathcal{G}$? When X is real-valued it is usual to let $\mathcal{G} = \{(a, b) : a < b \text{ with } a, b \in \mathbb{R}\}$ be the set of all open intervals. The reader can verify that if X is $\mathcal{F}/\mathcal{G}$-measurable, then it is also $\mathcal{F}/\sigma(\mathcal{G})$-measurable, where $\sigma(\mathcal{G})$ is the smallest σ-algebra that contains $\mathcal{G}$. This smallest σ-algebra can be shown to exist. Furthermore, it contains exactly those sets A that are in every σ-algebra that contains $\mathcal{G}$ (see Exercise 2.5). When $\mathcal{G}$ is the set of open intervals, $\sigma(\mathcal{G})$ is usually denoted by $\mathfrak{B}$ or $\mathfrak{B}(\mathbb{R})$ and is called the **Borel $\boldsymbol{\sigma}$-algebra**. This definition is also extended to $\mathbb{R}^k$ so that the open intervals are replaced by open rectangles of the form $\{x \in \mathbb{R}^k : a_i < x_i < b_i,\ i \in [k]\}$, $(a_i)_i, (b_i)_i \in \mathbb{R}^k$. When $\mathcal{G}$ is the set of such open rectangles then $\sigma(\mathcal{G})$ is denoted by $\mathfrak{B}(\mathbb{R}^k)$.

DEFINITION 2.2 A **random variable** (**random vector**) on measurable space $(\Omega, \mathcal{F})$ is a $\mathcal{F}/\mathfrak{B}(\mathbb{R})$-measurable function $X : \Omega \to \mathbb{R}$ (respectively $\mathcal{F}/\mathfrak{B}(\mathbb{R}^k)$-

measurable function $X : \Omega \to \mathbb{R}^k$). A **random element** between measurable spaces $(\Omega, \mathcal{F})$ and $(\mathcal{X}, \mathcal{G})$ is a $\mathcal{F}/\mathcal{G}$-measurable function $X : \Omega \to \mathcal{X}$.

Thus, random vectors are random elements where the range space is $(\mathbb{R}^k, \mathfrak{B}(\mathbb{R}^k))$ and random vectors are random variables when $k = 1$. The pushforward measure (or law) can be defined for any random element. Furthermore, random variables and vectors work nicely together: If $X_1, \dots, X_k$ are k random variables on the same domain $(\Omega, \mathcal{F})$, then $X(\omega) = (X_1(\omega), \dots, X_k(\omega))$ is an $\mathbb{R}^k$-valued random vector and vice versa (Exercise 2.2). Multiple random variables $X_1, \dots, X_k$ from the same measurable space can thus be viewed as a random vector $X = (X_1, \dots, X_k)$. When the measurable space is equipped with a probability measure $\mathbb{P}$, the pushforward measure (or law) of $\mathbb{P}$ under X is also known as the **joint law** or **joint distribution** of $(X_1, \dots, X_k)$.

Given a map $X : \Omega \to \mathcal{X}$ between measurable spaces $(\Omega, \mathcal{F})$ and $(\mathcal{X}, \mathcal{G})$, we let $\sigma(X) = \{X^{-1}(A) : A \in \mathcal{G}\}$ be the **$\boldsymbol{\sigma}$-algebra generated by $\boldsymbol{X}$**. The map X is $\mathcal{F}/\mathcal{G}$-measurable if and only if $\sigma(X) \subseteq \mathcal{F}$. By checking the definitions one can show that $\sigma(X)$ is a sub-σ-algebra of $\mathcal{F}$ and in fact is the smallest sub-σ-algebra for which X is measurable. If $\mathcal{G} = \sigma(\mathcal{A})$ itself is generated by a set system $\mathcal{A} \subset 2^{\mathcal{X}}$, then to check the $\mathcal{F}/\mathcal{G}$-measurability of X it suffices to check whether $X^{-1}(\mathcal{A}) \doteq \{X^{-1}(A) : A \in \mathcal{A}\}$ is a subset of $\mathcal{F}$. The reason this is sufficient is because $\sigma(X^{-1}(\mathcal{A})) = X^{-1}(\sigma(\mathcal{A}))$ and by definition the latter is $\sigma(X)$. In fact, to check whether a map is measurable, either one uses the composition rule or checks $X^{-1}(\mathcal{A}) \subset \mathcal{F}$ for a 'generator' $\mathcal{A}$ of $\mathcal{G}$.

Random elements can be combined to produce new random elements by composition. One can show that if f is $\mathcal{F}/\mathcal{G}$-measurable and g is $\mathcal{G}/\mathcal{H}$-measurable for σ-algebras $\mathcal{F}, \mathcal{G}$ and $\mathcal{H}$ over appropriate spaces then their composition $g \circ f$ is $\mathcal{F}/\mathcal{H}$-measurable (Exercise 2.1). This is used most often for **Borel functions**, which is a special name for $\mathfrak{B}(\mathbb{R}^m)/\mathfrak{B}(\mathbb{R}^n)$-measurable functions from $\mathbb{R}^m$ to $\mathbb{R}^n$. The reader will find it pleasing that all familiar functions are Borel. First and foremost, all continuous functions are Borel, which includes elementary operations such as addition and multiplication. Continuity is far from essential, however. In fact one is hard-pressed to construct a function that is not Borel. This means the usual operations are 'safe' when working with random variables.

Indicator functions

In the following text we make heavy uses of **indicator functions**. Given an arbitrary set Ω and $A \subseteq \Omega$ the indicator function of A is $\mathbb{I}_A : \Omega \to \{0, 1\}$ given by

$$\mathbb{I}_A(\omega) = \begin{cases} 1\,, & \text{if } \omega \in A\,; \\ 0\,, & \text{otherwise}\,. \end{cases}$$

Since a lot times the set A is in the focus and this set may have a complex structure, by abusing notation with the noble goal of improving readability, we will often write $\mathbb{I}\{\omega \in A\}$ instead of $\mathbb{I}_A(\omega)$. Similarly, we will often write $\mathbb{I}\{\mathrm{p}redicate(X, Y, \dots)\}$ to mean the indicator function of the subset of Ω on which the predicate is true. It is easy to check that an indicator function $\mathbb{I}_A$ is a random variable on $(\Omega, \mathcal{F})$ if and only if $A \in \mathcal{F}$.

Why so complicated?
You may be wondering why we did not define $\mathbb{P}$ on the powerset of Ω, which is equivalent to declaring all sets to be measurable. In many cases this is a perfectly reasonable thing to do, including the example game above where nothing prevents us from defining $\mathcal{F} = 2^\Omega$. There are two justifications not to do this, the first technical and the second conceptual. The technical issue is highlighted by the following surprising theorem, which shows there does not exist a uniform probability distribution on $\Omega = [0,1]$ if $\mathcal{F}$ is chosen to be the powerset of Ω. In other words, if you want to be able to define the uniform measure, then $\mathcal{F}$ cannot be too large. By contrast, the uniform measure can be defined on the Borel σ-algebra, though proving this is not elementary.

THEOREM 2.1 *Let $\Omega = [0,1]$ and $\mathcal{F}$ is the powerset of Ω. Then there does not exist a measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ such that $\mathbb{P}([a,b]) = b - a$ for all $0 \leq a \leq b \leq 1$.*

A second technical reason to prefer the measure-theoretic approach to probabilities is that this approach allows for the unification of distributions on discrete spaces and densities on continuous ones (the uninitiated reader will find the definitions of these later). This unification can be necessary when dealing with random variables that combine elements of both. For example, when dealing with a random variable that is zero with probability $1/2$ and otherwise behaves like a standard Gaussian.

The main conceptual reason not to focus exclusively on the case where $\mathcal{F}$ is the powerset is that σ-algebras are a way of representing information. This is especially useful in the study of bandits where the learner is interacting with an environment and slowly gaining knowledge. One useful way to represent this idea is by the means of a sequence of σ-algebras as we explain in the next section. One might be worried that the Borel σ-algebra does not contain enough measurable sets. Rest assured that this is not a problem and you will not easily find a non-measurable set. For completeness, an example will still be given in the notes, along with a little more discussion on this topic.

From laws to probability spaces and random variables
A big 'conspiracy' in probability theory is that probability spaces are seldom mentioned in theorem statements, despite the fact that a measure cannot be defined without one. Statements are instead given in terms of random elements and constraints on their joint probabilities. For example, suppose that X and Y are random variables such that

$$\mathbb{P}\left(X \in A, Y \in B\right) = \frac{|A \cap [6]|}{6} \cdot \frac{|B \cap [2]|}{2} \qquad \text{for all } A, B \in \mathfrak{B}(\mathbb{R}), \tag{2.1}$$

which represents the joint distribution for the values of a dice ($X \in [6]$) and coin ($Y \in [2]$). The formula describes some constraints on the probabilistic interactions between the outputs of X and Y, but says nothing about their domain. In a way, the domain is an unimportant detail. Nevertheless, one *must* ask whether an appropriate domain exists at all? More generally, one may ask whether an appropriate probability space exists given some constraints on the joint law of a collection $X_1, \dots, X_k$ of random variables. For this to make sense, the constraints should not contradict each other, which means there is a probability measure μ on $\mathfrak{B}(\mathbb{R}^k)$ such that μ satisfies the postulated constraints. But then we can choose $\Omega = \mathbb{R}^k$, $\mathcal{F} = \mathfrak{B}(\mathbb{R}^k)$, $\mathbb{P} = \mu$ and $X_i : \Omega \to \mathbb{R}$ be the ith coordinate map: $X_i(\omega) = \omega_i$. The pushforward of $\mathbb{P}$ under X is μ, which by definition is compatible with the constraints.

A more specific question is whether for a particular set of constraints on the joint law there exists a measure μ compatible with the constraints. Very often the constraints are specified for elements of the cartesian product of finitely many σ-algebras, like in Eq. (2.1). If $(\Omega_1, \mathcal{F}_1), \dots, (\Omega_n, \mathcal{F}_n)$ are measurable spaces, then the cartesian product of $\mathcal{F}_1, \dots \mathcal{F}_n$ is

$$\mathcal{F}_1 \times \cdots \times \mathcal{F}_n = \{A_1 \times \cdots \times A_n \,:\, A_1 \in \mathcal{F}_1, \dots, A_n \in \mathcal{F}_n\} \subseteq 2^{\Omega_1 \times \cdots \times \Omega_n}\,.$$

Elements of this set are known as **measurable rectangles** in $\Omega_1 \times \cdots \times \Omega_n$.

THEOREM 2.2 (Carathéodory's extension theorem) *Let $(\Omega_1, \mathcal{F}_1), \dots, (\Omega_n, \mathcal{F}_n)$ be measurable spaces and $\bar{\mu} : \mathcal{F}_1 \times \cdots \times \mathcal{F}_n \to [0,1]$ be a function such that:*

(a) $\bar{\mu}(\Omega_1 \times \cdots \times \Omega_n) = 1$.
(b) $\bar{\mu}(\cup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} \bar{\mu}(A_k)$ *for all sequences of disjoint sets with* $A_k \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$.

Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ and $\mathcal{F} = \sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_n)$. Then there exists a unique probability measure μ on $(\Omega, \mathcal{F})$ such that μ agrees with $\bar{\mu}$ on $\mathcal{F}_1 \times \cdots \times \mathcal{F}_n$.

The theorem is applied by letting $\Omega_k = \mathbb{R}$ and $\mathcal{F}_k = \mathfrak{B}(\mathbb{R})$. Then the values of a measure on all cartesian products uniquely determines its value everywhere.

Even when $n = 2$ it is not true that $\mathcal{F}_1 \times \mathcal{F}_2 = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$. Take for example, $\mathcal{F}_1 = \mathcal{F}_2 = 2^{\{1,2\}}$. Then, $|\mathcal{F}_1 \times \mathcal{F}_2| = 1 + 3 \times 3 = 10$ (because $\emptyset \times X = \emptyset$), while, since $\mathcal{F}_1 \times \mathcal{F}_2$ includes the singletons of $2^{\{1,2\} \times \{1,2\}}$, $\sigma(\mathcal{F}_1 \times \mathcal{F}_2) = 2^{\{1,2\} \times \{1,2\}}$. Hence, six sets are missing from $\mathcal{F}_1 \times \mathcal{F}_2$. For example, $\{(1,1),(2,2)\} \in \sigma(\mathcal{F}_1 \times \mathcal{F}_2) \setminus \mathcal{F}_1 \times \mathcal{F}_2$.

The σ-algebra $\sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_n)$ is called the **product σ-algebra** of $(\mathcal{F}_k)_{k \in [n]}$ and is also denoted by $\mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n$. The product operation turns out to be associative: $(\mathcal{F}_1 \otimes \mathcal{F}_2) \otimes \mathcal{F}_3 = \mathcal{F}_1 \otimes (\mathcal{F}_2 \otimes \mathcal{F}_3)$, which justifies writing $\mathcal{F}_1 \otimes \mathcal{F}_2 \otimes \mathcal{F}_3$. As it turns out, things work out well again with Borel σ-algebras: For $p, q \in \mathbb{N}^+$, $\mathfrak{B}(R^{p+q}) = \mathfrak{B}(R^p) \otimes \mathfrak{B}(\mathbb{R}^q)$. Needless to say, the same holds when there are more than two terms in the product.

2.2 σ-algebras and knowledge

One of the conceptual advantages of measure-theoretic probability is the relationship between σ-algebras and the intuitive idea of 'knowledge'. Although the relationship is useful and intuitive, it is regrettably not quite perfect. Let $(\Omega, \mathcal{F})$, $(\mathcal{X}, \mathcal{G})$ and $(\mathcal{Y}, \mathcal{H})$ be measurable spaces and $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ be random elements. Having observed the value of X ('knowing X'), one might wonder what this entails about the value of Y. Even more simplistically, under what circumstances can the value of Y be determined exactly having observed X? The situation is illustrated on Fig. 2.3. As it turns out, with some restrictions, the answer can be given in terms of the σ-algebras generated by X and Y. Except for a technical assumption on $(\mathcal{Y}, \mathcal{H})$, the following result shows that Y is a *measurable* function of X if and only if Y is $\sigma(X)/\mathcal{H}$-measurable. The technical assumption mentioned requires $(\mathcal{Y}, \mathcal{H})$ to be a Borel space, which is true of all probability spaces considered in this book, including $(\mathbb{R}^k, \mathfrak{B}(\mathbb{R}^k))$. We leave the exact definition of Borel spaces to the next chapter.

$$\begin{tikzcd}
(\Omega, \mathcal{F}) \arrow[r, "X"] \arrow[dr, "Y"'] & (\mathcal{X}, \mathcal{G}) \arrow[d, "f"] \\
& (\mathcal{Y}, \mathcal{H})
\end{tikzcd}$$

Figure 2.3 The factorization problem asks whether there exists a (measurable) function f that makes the diagram commute.

LEMMA 2.1 (Factorization lemma) *Assume that $(\mathcal{Y}, \mathcal{H})$ is a Borel space. Then Y is $\sigma(X)$-measurable ($\sigma(Y) \subset \sigma(X)$) if and only if there exists a $\mathcal{G}/\mathcal{H}$-measurable map $f : \mathcal{X} \to \mathcal{Y}$ such that $Y = f \circ X$.*

What this means is that Y can be computed from X as $Y = f(x)$ for some measurable function f if and only if Y is $\sigma(X)$-measurable. In this sense $\sigma(X)$ contains all the information that can be extracted from X via *measurable* functions. This is not the same as saying that Y can be deduced from X if and only if Y is $\sigma(X)$-measurable because the set of $\mathcal{X} \to \mathcal{Y}$ maps can be much larger than the set of $\mathcal{G}/\mathcal{H}$-measurable functions. When $\mathcal{G}$ is coarse there are not many $\mathcal{G}/\mathcal{H}$-measurable functions with the extreme case occurring when $\mathcal{G} = \{\mathcal{X}, \emptyset\}$. In cases like this, the intuition that $\sigma(X)$ captures all there is to know about X is not true anymore (Exercise 2.6). The issue is that $\sigma(X)$ does not only depend on X, but also on the σ-algebra of $(\mathcal{X}, \mathcal{G})$ and that if $\mathcal{G}$ is coarse-grained, then $\sigma(X)$ can also be coarse grained and not many functions will be $\sigma(X)$-measurable. If X is a random variable, then by definition $\mathcal{X} = \mathbb{R}$ and $\mathcal{G} = \mathfrak{B}(\mathbb{R})$, which is relatively fine-grained and the requirement that f be measurable is less restrictive. Nevertheless, even in the nicest setting where $\Omega = \mathcal{X} = \mathcal{Y} = \mathbb{R}$ and

$\mathcal{F} = \mathcal{G} = \mathcal{H} = \mathfrak{B}(\mathbb{R})$ it can still occur that $Y = f \circ X$ for some nonmeasurable f. In other words, all the information about Y exists in X but cannot be extracted in a measurable way. These problems only occur when X maps measurable sets in Ω to nonmeasurable sets in $\mathcal{X}$. While such random variables exist, they are never encountered in applications, which provides the final justification for thinking of $\sigma(X)$ as containing all that there is to know about any random variable X that one may ever expect to encounter.

Filtrations

In the study of bandits and other online settings it usually occurs that information is revealed to the learner sequentially. Let $X_1, \dots, X_n$ be a collection of random variables on common measurable space $(\Omega, \mathcal{F})$. We imagine a learner is sequentially observing the values of these random variables. First X_1, then X_2 and so on. The learner needs to make a prediction, or act, based on the available observations. Say, a prediction or an act must produce a real-valued response. Then, having observed $X_{1:t} \doteq (X_1, \dots, X_t)$, the set of maps $f \circ X_{1:t}$ where $f : \mathbb{R}^t \to \mathbb{R}$ is Borel, captures all the possible ways the learner can respond. By Lemma 2.1, this set contains exactly the $\sigma(X_{1:t})/\mathfrak{B}(\mathbb{R})$-measurable maps. Thus, if we need to reason about the set of $\Omega \to \mathbb{R}$ maps available after observing $X_{1:t}$, it suffices to concentrate on the σ-algebra $\mathcal{F}_t = \sigma(X_{1:t})$. Conveniently, $\mathcal{F}_t$ is independent of the space of possible responses, and being a subset of $\mathcal{F}$, it also hides details about the range space of $X_{1:t}$. It is easy to check that $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \mathcal{F}$, which means that more and more functions are becoming $\mathcal{F}_t$-measurable as t increases, which corresponds to increasing knowledge (note that $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and the set of $\mathcal{F}_0$-measurable function is the set of constant functions on Ω). To minimize clutter, we will write $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$. Note that here the ordering of the random variables does not matter (the same σ-algebra is generated by any fixed, non-random ordering).

Bringing these a little further, we will often find it useful to talk about increasing sequences of σ-algebras without constructing them in terms of random variables as above. Given measurable space $(\Omega, \mathcal{F})$ a **filtration** is a sequence $(\mathcal{F}_t)_{t=0}^n$ of sub-σ-algebras of $\mathcal{F}$ where $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$ for all $t < n$. Note that we also allow $n = \infty$ and in this case we define

$$\mathcal{F}_\infty = \sigma\left(\bigcup_{t=0}^{\infty} \mathcal{F}_t\right)$$

to be the smallest σ-algebra containing the union of all $\mathcal{F}_t$. Filtrations can also be defined in continuous time in the obvious way, but we have no need for that here. A sequence of random variables $(X_t)_{t=1}^n$ is **adapted** to filtration $\mathbb{F} = (\mathcal{F}_t)_{t=0}^n$ if X_t is $\mathcal{F}_t$-measurable for each t. We also say in this case that $(X_t)_t$ is $\mathbb{F}$-adapted. Finally, $(X_t)_t$ is $\mathbb{F}$-**predictable** if X_t is $\mathcal{F}_{t-1}$-measurable for each $t \in [n]$. Intuitively (again, with the caveats expressed earlier), we may think of a $\mathbb{F}$-predictable process $X = (X_t)_t$ as one that has the property that X_t can be known (or 'predicted') based on $\mathcal{F}_{t-1}$, while a $\mathbb{F}$-adapted process is one that has the property that X_t can be known based on $\mathcal{F}_t$ only. Since $\mathcal{F}_{t-1} \subseteq \mathcal{F}_t$, a predictable process is also adapted. A **filtered probability space** is the tuple $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\mathbb{F} = (\mathcal{F}_t)_t$ is filtration of $\mathcal{F}$.

2.3 Conditional probabilities

Conditional probabilities are introduced so that we can talk about how probabilities should be updated when one gains some partial knowledge about a random outcome. For the formal definition, let $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the space of outcomes, $\mathcal{F}$ is the collection of events to which the probability measure $\mathbb{P}$ assigns probabilities. Fix some event $B \in \mathcal{F}$ that has a positive probability and consider some other event $A \in \mathcal{F}$. The **conditional probability** $\mathbb{P}(A \mid B)$ of A given B is defined as

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}\,.$$

This answers the question of how knowing that B occurred should change the probability assigned to A. We can think about the outcome $\omega \in \Omega$ as the result of throwing a many-sided dice. The question asked is the probability that the dice landed so that $\omega \in A$ given that it landed with $\omega \in B$. The meaning of the condition $\omega \in B$ is that we focus on dice rolls when $\omega \in B$ is true. All dice rolls when $\omega \in B$ does not hold are discarded. Intuitively, what should matter in the conditional probability of A given B is how large the portion of A is that lies in B and this is indeed what the definition means.

⚠ The importance of conditional probabilities is that they define a calculus of how probabilities are to be updated in the presence of extra information.

To emphasize this relationship to knowledge, the probability $\mathbb{P}(A \mid B)$ is also called the **a posteriori** ('after the fact') probability of A given B. In contrast, its **a priori** probability is $\mathbb{P}(A)$. Note that $\mathbb{P}(A \mid B)$ is defined for every $A \in \mathcal{F}$ as long as $\mathbb{P}(B) > 0$. In fact, $A \mapsto \mathbb{P}(A \mid B)$ is a probability measure over the measure space $(\Omega, \mathcal{F})$ called the a posteriori probability measure given B (see Exercise 2.7). In a way the temporal characteristics attached to the words 'a posteriori' and 'a priori' can be a bit misleading. As discussed before, probabilities are concerned with predictions. They express the degrees of uncertainty one assigns to future events. This is also true for conditional probabilities. The conditional probability of A given B is a prediction of certain properties of the outcome of the random experiment that results in ω given a certain condition. Note that everything is related to a future hypothetical outcome. Once the dice is rolled, ω gets fixed and either $\omega \in A, B$ or not. There is no uncertainty left: predictions are trivial after an experiment is done.

The discussion of conditional probabilities would not be complete without mentioning **Bayes law** or **Bayes rule**, which states that provided $A, B \in \mathcal{F}$ have both positive probabilities,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)} . \tag{2.2}$$

The Bayes rule is useful because it allows one to obtain $\mathbb{P}(A \mid B)$ based on information about the quantities on the right-hand side. Remarkably, this happens to be the case quite often, explaining why this simple formula has quite a status in probability and statistics. Exercise 2.8 asks the reader to verify this law.

2.4 Independence

Independence is another basic concept of probability that relates to knowledge/information. In its simplest form independence is a relation that holds between events on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Two events $A, B \in \mathcal{F}$ are **independent** if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) . \tag{2.3}$$

How is this related to knowledge? Assuming that $\mathbb{P}(B) > 0$, dividing both sides by $\mathbb{P}(B)$ and using the definition of conditional probability we get that the above is equivalent to

$$\mathbb{P}(A \mid B) = \mathbb{P}(A) . \tag{2.4}$$

Of course, we also have that if $\mathbb{P}(A) > 0$, (2.3) is equivalent to $\mathbb{P}(B \mid A) = \mathbb{P}(B)$. Note that both of the latter relations express something intuitive, which is that A and B are independent if the probability assigned to A (or B) remains the same regardless of whether it is known that B (respectively, A) occurred.

⚠ Independence of two events means that observing the outcome of one does not change the likelihood of the other.

We hope our readers will find the definition of independence in terms of a 'lack of influence' to be sensible. The reason not to use Eq. (2.4) as the definition is mostly for the sake of convenience. If we started with (2.4) we would need to separately discuss the case of $\mathbb{P}(B) = 0$, which would be cumbersome. A second reason is that (2.4) suggests an asymmetric relationship, but intuitively we expect independence to be symmetric.

Why care about independence? There are at least two, unrelated reasons: *(i)* Uncertain outcomes are often generated part by part with no interaction between the processes, which naturally leads to an independence structure (think of rolling multiple dice with no interactions between the rolls) and *(ii)* once we discover some independence structure, calculations with probabilities can be immensely simplified. In fact, independence is often used as a way of constructing probability measures of interest (cf. Eq. (2.1), Theorem 2.2 and Exercise 2.9). Independence can also appear serendipitously in the sense that a probability space may hold many more independent events than what its construction may obviously suggest (Exercise 2.10).

⚠ Independence assumptions should not be taken lightly. Whenever independence is brought up, one should carefully judge whether the independence structure is really true. Since this is part of modeling, this reasoning is not mathematical in nature but is concerned with thinking about the physical processes.

A collection of events $\mathcal{G} \subset \mathcal{F}$ is said to be **pairwise independent** if any two distinct elements of $\mathcal{G}$ are independent of each other. The events in $\mathcal{G}$ are said to be **mutually independent** if for any $n > 0$ integer and $A_1, \dots, A_n$ distinct elements of $\mathcal{G}$, $\mathbb{P}(A_1 \cap \dots \cap A_n) = \prod_{i=1}^n \mathbb{P}(A_i)$. This is a stronger restriction than pairwise independence. In the case of mutually independent events the knowledge of joint occurrence of any finitely many events from the collection will not change our prediction of whether some other event happens. But this may not be the case when the events are only pairwise independent (Exercise 2.10). Two collections of events $\mathcal{G}_1, \mathcal{G}_2$ are said to be **independent of each other** if for any $A \in \mathcal{G}_1$ and $B \in \mathcal{G}_2$ it holds that A and B are independent. This definition is often applied to σ-algebras.

When the σ-algebras are induced by random variables, this leads to the definition of **independence between random variables**. Two random variables X and Y are independent if $\sigma(X)$ and $\sigma(Y)$ are independent of each other. As we discussed previously, σ-algebras summarize what knowledge can be gained by learning the value of a random variable. Hence the above definition says that two random variables are independent of each other when learning the value of one of them does not help in any way predicting the value of the other. The notions of pairwise and mutual independence can also be naturally extended to apply to collections of random variables. All these concepts can be and are in fact extended to random elements. The default meaning of independence when multiple events or random variables are involved is mutual independence.

⚠ When we say that $X_1, \dots, X_n$ are independent random variables, we mean that they are mutually independent. Independence is always relative to some probability measure, even when a probability measure is not explicitly mentioned. In such cases the identity of the probability measure should be clear from the context.

2.5 Integration and expectation

A key quantity in probability theory is the **expectation**, or **mean value** of random variables. For the formal definition fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The expectation of a random variable $X : \Omega \to \mathbb{R}$ is often denoted by $\mathbb{E}[X]$. This notation unfortunately obscures the dependence on the measure $\mathbb{P}$. When the underlying measure is not obvious from context we write $\mathbb{E}_{\mathbb{P}}$ to indicate the expectation with respect to $\mathbb{P}$. Mathematically, we define the expected value of X as its Lebesgue integral with respect to $\mathbb{P}$:

$$\mathbb{E}[X] = \int X(\omega) \, \mathrm{d}\mathbb{P}(\omega) \, .$$

The right-hand side is also often abbreviated to $\int X \, \mathrm{d}\mathbb{P}$, suppressing the variable ω that the integration is over. The integral on the right-hand side is constructed to satisfy the following two key properties:

(a) The integral of indicators is the probability of the underlying event. If $X(\omega) = \mathbb{I}\{\omega \in A\}$ is an indicator function for some $A \in \mathcal{F}$, then $\int X \mathrm{d}\mathbb{P} = \mathbb{P}(A)$.

(b) Integrals are linear. For all random variables X_1, X_2 and reals α_1, α_2 such that $\int X_1 \mathrm{d}\mathbb{P}$ and $\int X_2 \mathrm{d}\mathbb{P}$ are defined, $\int (\alpha_1 X_1 + \alpha_2 X_2) \mathrm{d}\mathbb{P}$ is defined and satisfies

$$\int (\alpha_1 X_1 + \alpha_2 X_2) \, \mathrm{d}\mathbb{P} = \alpha_1 \int X_1 \, \mathrm{d}\mathbb{P} + \alpha_2 \int X_2 \, \mathrm{d}\mathbb{P} \, . \tag{2.5}$$

These two properties together tell us that whenever $X(\omega) = \sum_{i=1}^n \alpha_i \mathbb{I}\{\omega \in A_i\}$ for some n, $\alpha_i \in \mathbb{R}$ and $A_i \in \mathcal{F}$, $i = 1, \dots, n$, then

$$\int X \mathrm{d}\mathbb{P} = \sum_i \alpha_i \mathbb{P}(A_i) \, . \tag{2.6}$$

Functions of the form X are called **simple functions**.

In defining the Lebesgue integral of some random variable X, we use (2.6) as the definition of the integral when X is a simple function. The next step is to extend the definition to nonnegative random variables. Let $X : \Omega \to [0, \infty)$ be measurable. The idea is to approximate X using simple functions from below and take the largest value that can be obtained this way.

$$\int_\Omega X \mathrm{d}\mathbb{P} = \sup \left\{ \int_\Omega h \, \mathrm{d}\mathbb{P} : h \text{ is simple and } 0 \le h \le X \right\} \, . \tag{2.7}$$

The meaning of $U \le V$ for random variables U, V is that $U(\omega) \le V(\omega)$ for all $\omega \in \Omega$. The supremum on the right-hand side could be infinite in which case we say the integral of X is not defined. Whenever the integral of X is defined we say that X is **integrable** or, if the identity of the measure $\mathbb{P}$ is unclear, that X is integrable with respect to $\mathbb{P}$.

Integrals for arbitrary random variables are defined by decomposing the random variable into positive and negative parts. Let $X : \Omega \to \mathbb{R}$ be any measurable function. Then define $X^+(\omega) = X(\omega)\mathbb{I}\{X(\omega) > 0\}$ and $X^-(\omega) = -X(\omega)\mathbb{I}\{X(\omega) < 0\}$ so that $X(\omega) = X^+(\omega) - X^-(\omega)$. Now X^+ and X^- are both nonnegative random variables (why?) called the **positive** and **negative** parts of X. Provided that both X^+ and X^- are integrable we define

$$\int_\Omega X \mathrm{d}\mathbb{P} = \int_\Omega X^+ \mathrm{d}\mathbb{P} - \int_\Omega X^- \mathrm{d}\mathbb{P}\,.$$

Note that X is integrable if and only if the nonnegative-valued random variable $|X|$ is integrable. The reader may challenge themselves by proving this in Exercise 2.12.

None of what we have done depends on $\mathbb{P}$ being a probability measure (that is $\mathbb{P}(A) \geq 0$ and $\mathbb{P}(\Omega) = 1$). The definitions all hold more generally for any measure, though for signed measures it is necessary to split Ω into disjoint measurable sets on which the measure is positive/negative, an operation that is possible by the **Hahn decomposition theorem**. We will never need signed measures in this book, however.

A particularly interesting case is when $\Omega = \mathbb{R}$ is the real line, $\mathcal{F}$ is the so-called Lebesgue σ-algebra (defined in the notes below), while the measure is the so-called Lebesgue measure λ, which is the unique measure over the Lebesgue σ-algebra such that $\lambda((a, b)) = b - a$ for any $a \leq b$. In this scenario, if $f : \mathbb{R} \to \mathbb{R}$ is a Borel-measurable function then we can write the Lebesgue integral of f with respect to the Lebesgue measure as

$$\int_\mathbb{R} f \, d\lambda\,.$$

Perhaps unsurprisingly this almost always coincides with the improper Riemann integral of f, which is normally written as $\int_{-\infty}^{\infty} f(x)dx$. Precisely, if $|f|$ is both Lebesgue integrable and Riemann integrable, then the integrals are equal.

There exist functions that are Riemann integrable and not Lebesgue integrable, and also the other way around (although examples of the former are more unusual than the latter).

The Lebesgue measure and its relation to Riemann integration is mentioned because when it comes to actually calculating the value of an expectation or integral, this is often reduced to calculating integrals over the real line with respect to the Lebesgue measure. The calculation is then performed by evaluating the Riemann integral, thereby circumventing the need to rederive the integral of many elementary functions. Integrals (and thus expectations) have a number of important properties. By far the most is their linearity, which was postulated above as the second property in (2.5). To practice using the notation with expectations, we restate the first half of this property. In fact, the statement is slightly more general than what we demanded for integrals above.

PROPOSITION 2.1 *Let $(X_i)_i$ be a (possibly infinite) collection of random variables on the same probability space and assume that $\mathbb{E}[X_i]$ exists for all i and furthermore that $X = \sum_i X_i$ and $\mathbb{E}[X]$ also exist. Then*

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X_i] \,.$$

This exchange of expectations and summation is the source of much magic in probability theory because it holds even if X_i are not independent. This means that (unlike probabilities) we can very often decouple the expectations of dependent random variables, which often proves extremely useful (a collection of random variables is dependent, if they are not independent). We will not prove this statement here, but as usual suggest the reader do so for themselves (Exercise 2.14). The other requirement for linearity is that if $c \in \mathbb{R}$ is a constant, then $\mathbb{E}[cX] = c\,\mathbb{E}[X]$, which is also true and rather easy to prove (Exercise 2.15).

Another important statement is concerned with independent random variables.

PROPOSITION 2.2 *If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.*

Note that in general $\mathbb{E}[XY] \neq \mathbb{E}[X]\,\mathbb{E}[Y]$ (Exercise 2.17). Finally, an important simple result connects expectations of nonnegative random variables to their tail probabilities.

PROPOSITION 2.3 *If $X \geq 0$ is a nonnegative random variable, then*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)\, dx \,.$$

The integrand in Proposition 2.3 is called the **tail probability function** $x \mapsto \mathbb{P}(X > x)$ of X. This is also known as the complementary cumulative distribution function of X. The **cumulative distribution function** (CDF) of X is defined as $x \mapsto \mathbb{P}(X \leq x)$ and is usually denoted by F_X. These functions are defined for all random variables, not just nonnegative ones. One can check that $F_X : \mathbb{R} \to [0, 1]$ is nondecreasing, right-continuous and $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$. The CDF of a random variable captures every aspect of the probability measure $\mathbb{P}_X$ induced by X, while still being just a function on the real line, a property that makes it a little more human-friendly than $\mathbb{P}_X$. One can also generalize CDFs to random vectors: If X is an $\mathbb{R}^k$-valued random vector then its CDF is defined as the $F_X : \mathbb{R}^k \to [0, 1]$ function that satisfies $F_X(x) = \mathbb{P}(X \leq x)$, where, in line with our conventions, $X \leq x$ means that all components of X are less-than-equal to the respective component of x.

2.6 Conditional expectation

Besides the expectation, we will also need **conditional expectation**, which allows us to talk about the expectation of a random variable given the value of another random variable, or more generally, given some σ-algebra. To illustrate with an example, let $(\Omega, \mathcal{F}, \mathbb{P})$ model the outcomes of an unloaded dice: $\Omega = [6]$, $\mathcal{F} = 2^\Omega$ and $\mathbb{P}(A) = |A|/6$. Define two random variables X and Y by $Y(\omega) = \mathbb{I}\{\omega > 3\}$ and $X(\omega) = \omega$. Suppose we are interested in the expectation of X given a specific value of Y. Arguing intuitively, we might notice that $Y = 1$ means that the unobserved X must be either 4, 5 or 6, and that each of these outcomes is equally likely and so the expectation of X given $Y = 1$ should be $(4 + 5 + 6)/3 = 5$. Similarly, the expectation of X given $Y = 0$ should be $(1 + 2 + 3)/3 = 2$. If we want a concise summary, we can just write that 'the expectation of X given Y' is $5Y + 2(1 - Y)$. Notice how this is a random variable itself. The notation for this conditional expectation is $\mathbb{E}[X \mid Y]$. Using this notation, in the above example, we can concisely write $\mathbb{E}[X \mid Y] = 5Y + 2(1 - Y)$. A little more generally, if $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ with $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$ and $|\mathcal{X}|, |\mathcal{Y}| < \infty$. Then $\mathbb{E}[X \mid Y] : \Omega \to \mathbb{R}$ is the random variable given by $\mathbb{E}[X \mid Y](\omega) = \mathbb{E}[X \mid Y = Y(\omega)]$ where

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \mathcal{X}} x\mathbb{P}(X = x \mid Y = y) = \sum_{x \in \mathcal{X}} \frac{x\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} . \tag{2.8}$$

Notice that this is undefined when $\mathbb{P}(Y = y) = 0$ so that $\mathbb{E}[X \mid Y](\omega)$ is undefined on the measure zero set $\{\omega : \mathbb{P}(Y = Y(\omega)) = 0\}$.

The definition in Eq. (2.8) does not generalize to continuous random variables because $\mathbb{P}(Y = y)$ in the denominator might be zero for all y. For example, let Y be a random variable taking values on $[0, 1]$ according to a uniform distribution and $X \in \{0, 1\}$ be Bernoulli with bias Y. This means that the joint measure on X and Y is $\mathbb{P}(X = 1, Y \in [p, q]) = \int_p^q x dx$ for $0 \le p < q \le 1$. Intuitively it seems like $\mathbb{E}[X \mid Y]$ should be equal to Y, but how to define it? Remember that the mean of a Bernoulli random variable is equal to its bias so the definition of conditional probability shows that for $0 \le p < q \le 1$,

$$\begin{aligned}
\mathbb{E}[X = 1 \mid Y \in [p, q]] &= \mathbb{P}(X = 1 \mid Y \in [p, q]) \\
&= \frac{\mathbb{P}(X = 1, Y \in [p, q])}{\mathbb{P}(Y \in [p, q])} \\
&= \frac{q^2 - p^2}{2(q - p)} \\
&= \frac{p + q}{2} .
\end{aligned}$$

The above calculation is not well defined when $p = q$ because $\mathbb{P}(Y \in [p, p]) = 0$. Nevertheless, letting $q = p + \varepsilon$ for $\varepsilon > 0$ and taking the limit as ε tends to zero seems like a reasonable way to argue that $\mathbb{P}(X = 1 \mid Y = p) = p$. Unfortunately this approach does not generalize to abstract spaces because there is no canonical way of taking limits towards a set of measure zero and different choices lead to different answers.

Instead we use Eq. (2.8) as the starting point for a more abstract definition. From Eq. (2.8) we see that $\mathbb{E}[X \mid Y](\omega)$ should only depend on $Y(\omega)$ and so should be measurable with respect to $\sigma(Y)$. The second requirement is called the 'averaging property'. For measurable $A \subseteq \mathcal{Y}$ the above display shows that

$$\begin{aligned}\mathbb{E}[\mathbb{I}_{Y^{-1}(A)}\mathbb{E}[X \mid Y]] &= \sum_{y\in A} \mathbb{P}\left(Y = y\right) \mathbb{E}[X \mid Y = y] \\ &= \sum_{y\in A} \sum_{x\in\mathcal{X}} x\mathbb{P}\left(X = x, Y = y\right) \\ &= \mathbb{E}[\mathbb{I}_{Y^{-1}(A)}X] \,.\end{aligned}$$

This can be viewed as putting a set of linear constraints on $\mathbb{E}[X \mid Y]$ with one constraint for each measurable $A \subseteq \mathcal{Y}$. By treating $\mathbb{E}[X \mid Y]$ as an unknown $\sigma(Y)$-measurable random variable, we can attempt to solve this linear system. As it turns out, this can always be done: The linear constraints and the measurability restriction on $\mathbb{E}\left[X \mid Y\right]$ completely determine $\mathbb{E}[X \mid Y]$ except for a set of measure zero. Notice that both conditions only depend on $\sigma(Y) \subseteq \mathcal{F}$. The abstract definition of conditional expectation takes these properties as the definition and replaces the role of Y with a sub-σ-algebra.

DEFINITION 2.3 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be random variable and $\mathcal{H}$ be a sub-σ-algebra of $\mathcal{F}$. The conditional expectation of X given $\mathcal{H}$ is denoted by $\mathbb{E}[X \mid \mathcal{H}]$ and defined to be any $\mathcal{H}$-measurable random variable on Ω such that for all $H \in \mathcal{H}$,

$$\int_H \mathbb{E}[X \mid \mathcal{H}] \mathrm{d}\mathbb{P} = \int_H X \mathrm{d}\mathbb{P} \,. \tag{2.9}$$

Given a random variable Y, the conditional expectation of X given Y is $\mathbb{E}\left[X \mid Y\right] = \mathbb{E}\left[X \mid \sigma(Y)\right]$.

THEOREM 2.3 *Given any probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*, a sub-*$\sigma$*-algebra* $\mathcal{H}$ *of* $\mathcal{F}$ *and a* $\mathbb{P}$*-integrable random variable* $X : \Omega \to \mathbb{R}$*, there exist a* $\mathcal{H}$*-measurable function* $f : \Omega \to \mathbb{R}$ *that satisfies* (2.9)*. Further, any two* $\mathcal{H}$*-measurable functions* $f_1, f_2 : \Omega \to \mathbb{R}$ *that satisfy* (2.9) *are equal with probability one:* $\mathbb{P}(f_1 = f_2) = 1$.

When random variables X and Y agree with $\mathbb{P}$-probability one we say they agree **$\mathbb{P}$-almost surely** equal, which is often abbreviated to '$X = Y$ $\mathbb{P}$-a.s.' or '$X = Y$ a.s.' when the measure is clear from context. A related notation that will be useful in the future is the concept of **almost surely**.

The reader may find it odd that $\mathbb{E}[X \mid Y]$ is a random variable on Ω rather than the range of Y. Lemma 2.1 and the fact that $\mathbb{E}[X \mid \sigma(Y)]$ is $\sigma(Y)$-measurable shows there exists a measurable function $f : (\mathbb{R}, \mathfrak{B}(\mathbb{R})) \to (\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ such that $\mathbb{E}[X \mid \sigma(Y)](\omega) = (f \circ Y)(\omega)$ (see Fig. 2.4). In this sense $\mathbb{E}[X \mid Y](\omega)$ only depends on $Y(\omega)$ and occasionally we write $\mathbb{E}[X \mid Y](y)$.

At the risk of being overly verbose, what is the meaning of all this? Returning to the dice example above we see that $\mathbb{E}\left[X \mid Y\right] = \mathbb{E}\left[X \mid \sigma(Y)\right]$ and $\sigma(Y) = \{\{1, 2, 3\}, \{4, 5, 6\}, \emptyset, \Omega\}$. The condition that $\mathbb{E}[X \mid \mathcal{H}]$ is $\mathcal{H}$-measurable can only

$$\begin{tikzcd}
& (\Omega, \mathcal{F}) \arrow[dl, "Y"'] \arrow[dr, "{\mathbb{E}[X \mid \sigma(Y)]}"] & \\
(\mathbb{R}, \mathfrak{B}(\mathbb{R})) \arrow[rr, "f"'] & & (\mathbb{R}, \mathfrak{B}(\mathbb{R}))
\end{tikzcd}$$

Figure 2.4 Factorization of conditional expectation.

be satisfied if $\mathbb{E}[X \mid \mathcal{H}](\omega)$ is constant on $\{1,2,3\}$ and $\{4,5,6\}$. Then (2.9) immediately implies that

$$\mathbb{E}\left[X \mid \mathcal{H}\right](\omega) = \begin{cases} 2, & \text{if } \omega \in \{1,2,3\}\,; \\ 5, & \text{if } \omega \in \{4,5,6\}\,. \end{cases}$$

While the definition of conditional expectations given above is non-constructive and $\mathbb{E}[X \mid \mathcal{H}]$ is uniquely defined only up to events of $\mathbb{P}$-measure zero, none of this should be of a significant concern. First, we will rarely need closed form expressions for conditional expectations, but we rather need how they relate to other expectations, conditional or not. This is also the reason why it should not be concerning that they are only determined up to zero probability events: Usually, conditional expectations appear in other expectations or in statements that are concerned with how probable some event is, making the difference between the different 'versions' of conditional expectations disappear.

We close the section by summarizing some additional important properties of conditional expectations. These follow from the definition directly and the reader is invited to prove them in Exercise 2.19, but note the difficulty varies wildly.

THEOREM 2.4 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G}, \mathcal{G}_1, \mathcal{G}_2 \subset \mathcal{F}$ sub-σ-algebras of $\mathcal{F}$, X, Y random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. The following hold true:*

1. *If $X \geq 0$, then $\mathbb{E}\left[X \mid \mathcal{G}\right] \geq 0$ almost surely.*
2. *$\mathbb{E}\left[1 \mid \mathcal{G}\right] = 1$ almost surely.*
3. *$\mathbb{E}\left[X + Y \mid \mathcal{G}\right] = \mathbb{E}\left[X \mid \mathcal{G}\right] + \mathbb{E}\left[Y \mid \mathcal{G}\right]$ almost surely, assuming the expression on the right-hand side is defined.*
4. *$\mathbb{E}\left[XY \mid \mathcal{G}\right] = Y\mathbb{E}\left[X \mid \mathcal{G}\right]$ almost surely if $\mathbb{E}\left[XY\right]$ exists and Y is $\mathcal{G}$-measurable.*
5. *if $\mathcal{G}_1 \subset \mathcal{G}_2$, then $\mathbb{E}\left[X \mid \mathcal{G}_1\right] = \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}_2\right] \mid \mathcal{G}_1\right]$ almost surely.*
6. *if $\mathcal{G}_1$ and $\mathcal{G}_2$ are independent, then $\mathbb{E}\left[X \mid \sigma(\mathcal{G}_1 \cup \mathcal{G}_2)\right] = \mathbb{E}\left[X \mid \mathcal{G}_1\right] = \mathbb{E}\left[X \mid \mathcal{G}_2\right]$ almost surely.*
7. *If $\mathcal{G} = \{\emptyset, \Omega\}$ is the trivial σ-algebra, then $\mathbb{E}\left[X \mid \mathcal{G}\right] = \mathbb{E}\left[X\right]$ almost surely.*

Properties 1 and 2 are self-explanatory. Property 3 generalizes the linearity of expectation. Property 4 shows that a measurable quantity can be pulled outside of a conditional expectation and corresponds to the property that for constants c, $\mathbb{E}\left[cX\right] = c\mathbb{E}\left[X\right]$. Property 5 is called the **tower rule** or the **law of total**

expectations. It says that the fineness of $\mathbb{E}[X \mid \mathcal{G}_2]$ is obliterated when taking the conditional expectation with respect to $\mathcal{G}_1$. Property 6 relates independence and conditional expectations and it says that conditioning on independent quantities does not give further information on expectations. Finally, Property 7 states that conditioning on no information gives the same expectation as not conditioning at all.

⚠ The above list of abstract properties will be used over and over again. We encourage the reader to study the list carefully and convince yourself that all items are intuitive. Playing around with discrete random variables can be invaluable for this. Eventually it will all become second nature.

2.7 Notes

1 The Greek letter σ is often used by mathematicians in association with countable infinities. Hence the term σ-algebra (and σ-field). Note that countable additivity is often called σ-additivity. The requirement that additivity should hold for systems of countably infinitely many sets is made so that probabilities of (interesting) limiting events are guaranteed to exist.

2 **Measure theory** is concerned with measurable spaces and measures on them, and with their properties. An obvious distinction between probability theory and measure theory is that in probability theory one is (mostly) concerned with probability measures. But the distinction does not stop here: In probability theory, the emphasis is on the probability distributions and their relations to each other; the measurable spaces are there in the background, but they are viewed as part of the technical toolkit, rather than as the topic of main interest.

3 In our toy example instead of $\Omega = [6]^7$, we could have chosen $\Omega = [6]^8$ (considering rolling eight dice instead of 7, one dice never used). There are many other possibilities. We can consider coin flips instead of dice rolls (think about how this could be done). To make this easy, we could use weighted coins (for example, a coin that lands on heads with probability 1/6), but we don't actually need weighted coins (this may be a little tricky to see). The main point is that there are many ways to emulate one randomization device by using another. The difference between these is the set Ω. What makes a choice of Ω viable is if we can emulate the game mechanism on the top of Ω so that in the end the probability of seeing any particular value remains the same. But the main point is that the choice of Ω is far from unique. The same is true for the way we calculate the value of the game! For example, the dice could be reordered, if we stay with the first construction. This was noted already, but it cannot be repeated frequently enough: The biggest conspiracy in all