B Mathematical Prerequisites

The beginner... should not be discouraged if... he finds that he does not have the prerequisites for reading the prerequisites.

P. Halmos[1]

In this chapter, we introduce mathematical results that might not be known to all readers, but which are sufficiently standard that they not be put into the actual chapters.

This exposition is almost certainly incomplete, and some readers will inevitably happen upon terms in the book that are unknown to them, yet not explained here. Consequently, we also give some further references.

B.1 Probability

B.1.1 Probability Spaces

Let us start with some basic notions of probability theory. For further detail, we refer to [77, 165, 561]. We do not try to be rigorous; instead, we endeavor to give some intuition and explain how these concepts are related to our present interests.

Domain

Assume we are given a nonempty set $\mathcal{X}$, called the *domain* or *universe*. We refer to the elements x of $\mathcal{X}$ as *patterns*. The patterns are generated by a stochastic source. For instance, they could be handwritten digits, which are subject to fluctuations in their generation best modelled probabilistically. In the terms of probability theory, each pattern x is considered the outcome of a *random experiment*.

We would next like to assign probabilities to the patterns. We naively think of a probability as being the limiting frequency of a pattern; in other words, how often, relative to the number of trials, a certain pattern x comes up in a random experiment, if we repeat this experiment infinitely often?

Event
Probability

It turns out to be convenient to be slightly more general, and to talk about the probability of *sets* of possible outcomes; that is, subsets C of $\mathcal{X}$ called *events*. We denote the *probability* that the outcome of the experiment lies in C by

1. Quoted after [429].

$$\mathrm{P}\{x \in C\}. \tag{B.1}$$

If Υ is a logical formula in terms of x, meaning a mapping from $\mathcal{X}$ to $\{\text{true}, \text{false}\}$, then it is sometimes convenient to talk about the probability of Υ being true. We will use the same symbol P in this case, and define its usage as

$$\mathrm{P}\{\Upsilon(x)\} := \mathrm{P}\{x \in C\} \text{ where } C = \{x \in \mathcal{X} | \Upsilon(x) = \text{true}\}. \tag{B.2}$$

Let us also introduce the shorthand

$$\mathrm{P}(C) := \mathrm{P}\{x \in C\}, \tag{B.3}$$

to be read as "the probability of the event C." If P satisfies some fairly natural conditions, it is called a *probability measure*. It is also referred to as the *(probability) distribution of x*.

Distribution of x

In the case where $\mathcal{X} \subset \mathbb{R}^N$, the patterns are usually referred to as *random variables* ($N = 1$) or *random vectors* ($N > 1$). A generic term we shall sometimes use is *random quantity*.[2]

To emphasize the fact that P is the distribution of x, we sometimes denote it as P_x or $\mathrm{P}(x)$.[3] To give the precise definition of a probability measure, we first need to be a bit more formal about which sets C we are going to allow. Certainly,

$$C = \mathcal{X} \tag{B.4}$$

should be a possibility, corresponding to the event that necessarily occurs ("sure thing"). If C is allowed, then its complement,

$$\overline{C} := \mathcal{X} \setminus C, \tag{B.5}$$

should also be allowed. This corresponds to the event "not C." Finally, if $C_1, C_2, \ldots$ are events, then we would like to be able to talk about the probability of the event "C_1 or C_2 or $\ldots$", hence

$$\bigcup_{i=1}^{\infty} C_i \tag{B.6}$$

should be an allowed event.

σ-Algebra

Definition B.1 (σ-Algebra) *A collection $\mathcal{C}$ of subsets of $\mathcal{X}$ is called a* σ-algebra on $\mathcal{X}$ *if*

(i) $\mathcal{X} \in \mathcal{C}$*; in other words, (B.4) is one of its elements;*

(ii) *it is closed under complementation, meaning if $C \in \mathcal{C}$, then also (B.5); and*

(iii) *it is closed under countable*[4] *unions: if $C_1, C_2, \ldots \in \mathcal{C}$, then also (B.6).*

---

2. For simplicity, we are somewhat sloppy in not distinguishing between a random variable and the values it takes. Likewise, we deviate from standard usage in not having introduced random variables as functions on underlying universes of events.
3. The latter is somewhat sloppy, as it suggests that P takes *elements* of $\mathcal{X}$ as inputs, which it does not: P is defined for *subsets* of $\mathcal{X}$.
4. *Countable* means with a number of elements not larger than that of $\mathbb{N}$. Formally, a set

The elements of a σ-algebra are sometimes referred to as measurable sets.

We are now in a position to formalize our intuitions about the probability measure.

Definition B.2 (Probability Measure) *Let $\mathcal{C}$ be a σ-algebra on the domain $\mathcal{X}$. A function*

$$\mathrm{P} : \mathcal{C} \to [0, 1] \tag{B.7}$$

Probability Measure

is called a probability measure *if it is normalized,*

$$\mathrm{P}(\mathcal{X}) = 1, \tag{B.8}$$

and σ-additive, meaning that for sets $C_1, C_2, \ldots \in \mathcal{C}$ that are mutually disjoint ($C_i \cap C_j = \emptyset$ if $i \neq j$), we have

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} C_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(C_i). \tag{B.9}$$

Measure

As an aside, note that if we drop the normalization condition, we are left with what is called a *measure*.

Probability Space

Taken together, $(\mathcal{X}, \mathcal{C}, \mathrm{P})$ are called a *probability space*. This is the mathematical description of the probabilistic experiment.

B.1.2 IID Samples

Nevertheless, we are not quite there yet, since most of the probabilistic statements in this book do not talk about the outcomes of the experiment described by $(\mathcal{X}, \mathcal{C}, \mathrm{P})$. For instance, when we are trying to learn something about a regularity (that is, about some aspects of P) based on a collection of patterns $x_1, \ldots, x_m \in \mathcal{X}$ (usually called a *sample*), we actually perform the random experiment m times, under identical conditions. This is referred to as *drawing an iid (independent and identically distributed) sample* from P.

Sample

IID Sample

Formally, drawing an iid sample can be described by the probability space $(\mathcal{X}^m, \mathcal{C}^m, \mathrm{P}^m)$. Here, $\mathcal{X}^m$ denotes the m-fold Cartesian product of $\mathcal{X}$ with itself (thus, each element of $\mathcal{X}^m$ is an m-tuple of elements of $\mathcal{X}$), and $\mathcal{C}^m$ denotes the smallest σ-algebra that contains the elements of the m-fold Cartesian product of $\mathcal{C}$ with itself. Likewise, the product measure P^m is determined uniquely by

$$\mathrm{P}^m((C_1, \ldots, C_m)) := \prod_{i=1}^{m} \mathrm{P}(C_i). \tag{B.10}$$

Note that the independence of the "iid" is encoded in (B.10) being a product of measures on $\mathcal{C}$, while the identicality lies in the fact that all the measures on $\mathcal{C}$ are one and the same.

---

is countable if there is a surjective map from $\mathbb{N}$ onto this set; that is, a map with range encompassing the whole set.

By analogy to (B.2), we sometimes talk about the probability of a logical formula involving an m-sample,[5]

$$\mathrm{P}\{Y(x_1, \ldots, x_m)\} := \mathrm{P}^m(\{(x_1, \ldots, x_m) \in \mathcal{X}^m | Y(x_1, \ldots, x_m) = \text{true}\}). \tag{B.11}$$

So far, we have denoted the outcomes of the random experiments as x for simplicity, and have referred to them as patterns. In many cases studied in this book, however, we will not only observe patterns $x \in \mathcal{X}$ but also *targets* $y \in \mathcal{Y}$. For instance, in binary pattern recognition, we have $\mathcal{Y} = \{\pm 1\}$. The underlying regularity is now assumed to generate *examples* (x, y). All of the above applies to this case, with the difference that we now end up with a probability measure on $\mathcal{X} \times \mathcal{Y}$, called the (joint) distribution of (x, y).

B.1.3 Densities and Integrals

We now move on to the concept of a *density*, often confused with the distribution. For simplicity, we restrict ourselves to the case where $\mathcal{X} = \mathbb{R}^N$; in this instance, $\mathcal{C}$ is usually taken to be the *Borel σ-algebra*.[6]

Definition B.3 (Density) *We say that the nonnegative function p is the* density *of the distribution* P *if for all $C \in \mathcal{C}$,*

$$\mathrm{P}(C) = \int_C p(x)dx. \tag{B.12}$$

If such a p exists, it is uniquely determined.[7]

Not all distributions actually *have* a density. To see this, let us consider a distribution that does. If we plug a set of the form $C = \{x\}$ into (B.12), we see that $\mathrm{P}(\{x\}) = 0$; that is, the distribution assigns zero probability to any set of the form $\{x\}$. We infer that only distributions that assign zero probability to individual points can have densities.[8]

It is important to understand the difference between distributions and densities. The distribution takes *sets* of patterns as inputs, and assigns them a probability between 0 and 1. The density takes an individual pattern as its input, and assigns a nonnegative number (possibly larger than 1) to it. Using (B.12), the density can be used to compute the probability of a set C. If the density is a continuous function, and we use a small neighborhood of point x as the set C, then P is approximately

5. Note that there is some sloppiness in the notation: strictly speaking, we should denote this quantity as P^m — usually, however, it can be inferred from the context that we actually mean the m-fold product measure.
6. Readers not familiar with this concept may simply think of it as a collection that contains all "reasonable" subsets of $\mathbb{R}^N$.
7. *Almost everywhere*; in other words, up to a set N with $\mathrm{P}(N) = 0$.
8. In our case, we can show that the distribution P has a density if and only if it is *absolutely continuous* with respect to the Lebesgue measure on $\mathbb{R}^N$, meaning that every set of Lebesgue-measure zero also has P-measure zero.

the size (i.e. , the measure) of the neighborhood times the value of p; in this case, and in this sense, the two quantities are proportional.

Distribution Function

A more fundamental concept, which exists for *every* distribution of a random quantity taking values in $\mathbb{R}^N$, is the *distribution function*,[9]

$$\mathrm{F} : \mathbb{R}^N \to [0,1] \tag{B.13}$$

$$z \mapsto \mathrm{F}(z) = \mathrm{P}\{[x]_1 < [z]_1 \wedge \ldots \wedge [x]_N < [z]_N\}. \tag{B.14}$$

Finally, we need to introduce the notion of an integral with respect to a measure. Consider a function $f : \mathbb{R}^N \to \mathbb{R}$. We denote by

$$\int_C f(x) d\mathrm{P}(x) \tag{B.15}$$

the integral of a function with respect to the distribution (or measure) P, provided that f is *measurable*. For our purposes, the latter means that for every interval $[a,b] \subset \mathbb{R}$, $f^{-1}([a,b])$ (the set of all points in $\mathbb{R}^N$ that get mapped to $[a,b]$) is an element of $\mathcal{C}$. Component-wise extension to vector-valued functions is straightforward.

In the case where P has a density p, (B.15) equals

$$\int_C f(x) p(x) dx, \tag{B.16}$$

which is a standard integral in $\mathbb{R}^N$, weighted by the density function p.

If P does not have a density, we can define the integral by decomposing the range of f into disjoint half-open intervals $[a_i, b_i)$, and computing the measure of each set $f^{-1}([a_i, b_i))$ using P. The contribution of each such set to the integral is determined by multiplying this measure with the function value (on the set), which by construction is in $[a_i, b_i)$. The exact value of the integral is obtained by taking the limit at infinitely small intervals. This construction, which is the basic idea of the Lebesgue integral, does not rely on f being defined on $\mathbb{R}$; it works for general sets $\mathcal{X}$ as long as they are suitably endowed with a measure.

Empirical Measure

Let us consider a special case. If P is the *empirical measure* with respect to $x_1, \ldots, x_m$,[10]

$$\mathrm{P}^m_{\mathrm{emp}}(C) := \frac{|C \cap \{x_1, \ldots, x_m\}|}{m}, \tag{B.17}$$

which represents the fraction of points that lie in C, then the integral takes the form

$$\int_C f(x) d\mathrm{P}^m_{\mathrm{emp}}(x) = \frac{1}{m} \sum_{i=1}^{m} f(x_i). \tag{B.18}$$

As an aside, note that this shows the empirical risk term (1.17) can actually be thought of as an integral, just like the actual risk (1.18).

9. We use $\wedge$ to denote the logical "and" operation, and $[z]_i$ to denote the i^{th}component of z.
10. By $|.|$ we denote the number of elements in a set.

Expectation, Variance, Covariance

If P is a probability distribution (rather than a general measure), then two more special cases of interest are obtained for particular choices of functions f in (B.15). If f is the identity on $\mathbb{R}^N$, we get the *expectation* $\mathbf{E}[x]$. If $f(x) = (x - \mathbf{E}[x])^2$ (on $\mathbb{R}$), we obtain the *variance* of x, denoted by $\mathrm{var}(x)$. In the N-dimensional case, the functions $f_{ij}(x) = (x_i - \mathbf{E}[x_i])(x_j - \mathbf{E}[x_j])$ lead to the *covariance* $\mathrm{cov}(x_i, x_j)$. For a data set $\{x_1, \ldots, x_m\}$, the matrix $(\mathrm{cov}(x_i, x_j))_{ij}$ is called the *covariance matrix*.

B.1.4 Stochastic Processes

A *stochastic process* y on a set $\mathcal{X}$ is a random quantity indexed by $x \in \mathcal{X}$. This means that for every x, we get a random quantity $y(x)$ taking values in $\mathbb{R}$, or more generally, in a *set* $\mathcal{R}$. A stochastic process is characterized by the joint probability distributions of y on arbitrary finite subsets of $\mathcal{X}$; in other words, of $(y(x_1), \ldots, y(x_m))$.[11]

A *Gaussian process* is a stochastic process with the property that for any $\{x_1, \ldots, x_m\} \subset \mathcal{X}$, the random quantities $(y(x_1), \ldots, y(x_m))$ have a joint Gaussian distribution with mean μ and covariance matrix K. The matrix elements K_{ij} are given by a covariance kernel $k(x_i, x_j)$.

When a Gaussian process is used for learning, the *covariance function* $k(x_i, x_j) := \mathrm{cov}(y(x_i), y(x_j))$ essentially plays the same role as the kernel in a SVM. See Section 16.3 and [587, 596] for further information.

B.2 Linear Algebra

B.2.1 Vector Spaces

We move on to basic concepts of linear algebra, which is to say the study of vector spaces. Additional detail can be found in any textbook on linear algebra (e.g., [170]). The feature spaces studied in this book have a rich mathematical structure, which arises from the fact that they allow a number of useful operations to be carried out on their elements: addition, multiplication with scalars, and the product between the elements themselves, called the dot product.

What's so special about these operations? Let us, for a moment, go back to our earlier example (Chapter 1), where we classify sheep. Surely, nobody would come up with the idea of trying to add two sheep, let alone compute their dot product. The set of sheep does not form a vector space; mathematically speaking, it could be argued that it does not have a very rich structure. However, as discussed in Chapter 1 (cf. also Chapter 2), it is possible to *embed* the set of all sheep into a dot product space such that we can think of the dot product as a measure of

11. Note that knowledge of the finite-dimensional distributions (fdds) does not yield complete information on the properties of the sample paths of the stochastic process; two different processes which have the same fdds are known as *versions* of one another.