Chapter 2

Axiomatic Probability Theory

2.1 Axioms of Probability

In the previous chapter, we have introduced sample spaces and event spaces. We would like to be able to express that certain events are more (or less) likely than others. Therefore, we are going to measure the probability of events in a mathematically precise sense.

Definition 2.1 (Finite Measure) *A* finite measure *is a function* $\mu : \mathcal{S} \to \mathbb{R} : S \mapsto \mu(S)$ *that maps elements from a countable set of sets* $\mathcal{S}$ *(formally a* σ*-algebra) to real numbers. Such a measure has the following properties:*

1. $\mu(S) \in \mathbb{R}$ *for* $S \in \mathcal{S}$,
2. $\mu\left(\bigcup_{i=1}^{\infty} S_i\right) = \sum_{i=1}^{\infty} \mu\left(S_i\right)$ *for disjoint sets* $S_1, S_2, \ldots$.

Notice that we are restricting ourselves to finite measures here, i.e. the value of the measure can never be infinite. This restriction makes sense as probabilities are finite as well. Property 2 is known as *countable additivity*.

Let $S = \bigcup_{i=1}^{n} S_i$ for some positive natural number n and disjoint S_i and $S_j = \emptyset$ for $j > n$. By countable additivity, we then get

$$\mu(S) = \mu(\bigcup_{i=1}^{\infty} S_i) = \mu\left(\bigcup_{i=1}^{n} S_i \cup \bigcup_{j=n+1}^{\infty} \emptyset\right) = \sum_{i=1}^{n} \mu(S_i) + \sum_{j=n+1}^{\infty} \mu(\emptyset) \tag{2.1}$$

Since the S_i are disjoint, we must have $\mu(S) = \sum_{i=1}^{n} \mu(S_i)$ and it follows that $\mu(\emptyset) = 0$. We conclude that the empty set has measure 0 for all

measures. Furthermore, we also see from the above derivation that countable additivity implies finite additivity, i.e. $\mu(S) = \sum_{i=1}^{n} \mu(S_i)$ for finite positive n (again, this only holds if the S_i are disjoint).

Examples of measures are not hard to find. In fact, we have already seen a measure, namely the function $|\cdot|$ that counts the elements of a set (check yourself that it really is a measure). Another measure is the Dirac-measure that is related to the characteristic function of a set. While the characteristic function tells you whether any object belongs to a given set, the Dirac-measure tells you whether any set contains a given object. Let us call the object in question a. Then its Dirac measure $\delta_a(S) = 1$ iff $a \in S$ and 0 otherwise (check yourself that the Dirac-measure indeed is a measure).

Apart from these examples, there is one measure, however, that is going to be the star of the rest of this script, namely the **probability measure**.

Definition 2.2 (Probability measure) *A probability measure $\mathbb{P} : \mathcal{A} \to \mathbb{R}, A \mapsto \mathbb{P}(A)$ on an event space $\mathcal{A}$ associated with a sample space Ω has the following properties:*

1. $\mathbb{P}(A) \geq 0$ *for all* $A \in \mathcal{A}$,
2. $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ *for disjoint events* $A_1, A_2, \ldots$,
3. $\mathbb{P}(\Omega) = 1$.

Notice that we only added Property 3 to the general definition of a measure. Hence, a **probability** (the value that the probability measure assigns to an event) will always lie in the real interval $[0, 1]$. The above three axioms for a probability measure are often referred to as *axioms of probability* or *Kolmogorov axioms* after their inventor Andrey Kolmogorov.

We have already discussed uniform probabilities in the previous chapter. We can now formally explain what we meant by that. The uniform probability measure $\mathbb{P}$ has the property that $\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|}$ for all $\omega \in \Omega$. At this point, the distinction between sample and event spaces becomes important. We cannot measure the elements of a sample space, only the elements of an event space! Recall our convention that we will always assume that $\mathcal{A} = \mathcal{P}(\Omega)$ which obviously contains a singleton for each element in Ω. Using this assumption, the uniform probability measure is indeed well-defined. Whenever we talk about *uniform probability*, we either mean the uniform probability measure or, more often, the real value $\frac{1}{|\Omega|}$ to which this measure uniformly evaluates.

In order to create a tight relationship between a sample space, an event space and a probability measure, we introduce the concept of a **probability space**. Probability spaces are also known as **(probabilistic) experiments**.

Definition 2.3 (Probability space) *A probability space is a triple $(\Omega, \mathcal{A}, \mathbb{P})$, consisting of a sample space Ω, an event space $\mathcal{A}$ and a probability measure $\mathbb{P}$.*

If we roll a die, for example, we have the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ and, by convention, the event space $\mathcal{A} = \mathcal{P}(\Omega)$. If we add the uniform probability measure, we have constructed a *probabilistic experiment*. We can use it to answer a couple of questions. For example, we might wonder about the probability of obtaining an even number. By Property 2 of our definition, this probability is given by

$$\mathbb{P}(\{2,4,6)\}) = \mathbb{P}(\{2\} \cup \{4\} \cup \{6\}) \tag{2.2}$$

$$= \mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \tag{2.3}$$

Notice that this calculation is rather cumbersome. After all, we might just have evaluated $\mathbb{P}(\{2, 4, 6\})$ directly. This is because by convention we have $\mathcal{A} = \mathcal{P}(\Omega)$ which certainly contains $\{2, 4, 6\}$. Since the probability measure is defined on $\mathcal{A}$, it must map $\{2, 4, 6\}$ to some real number. However, the above calculation points to an interesting fact. In order to fully specify a probability measure, is suffices to specify the measure on the singleton sets of the event space. By countable additivity, this assignment already specifies the measure on the entire event space, as we can construct any event as a countable union of singletons.

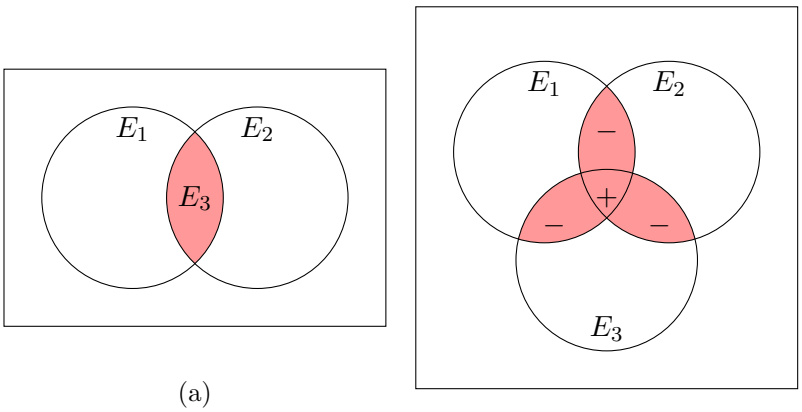
It is important to point out that we just chose the uniform probability measure as the one that seems "natural" for a die roll. However, nobody is forcing us to do so. In fact, Definition 2.3 allows us to impose arbitrary probability measures.

Exercise 2.1 *Let us consider a rigged die. Take $(\Omega, \mathcal{A}, \mathbb{P})$ with Ω and $\mathcal{A} = \mathcal{P}(\Omega)$ as in the uniform die-roll example before, but use the probability measure specified by*
$\mathbb{P} = \{(\{1\}, 0), (\{2\}, \frac{1}{12}), (\{3\}, \frac{1}{6}), (\{4\}, \frac{1}{6}), (\{5\}, \frac{1}{3}), (\{6\}, \frac{1}{4})\}$.

1. *Verify that $\mathbb{P}$ is indeed a probability measure.*
2. *Compute the probability of obtaining a number strictly smaller than 5 in this experiment.*

2.2 Probability of Arbitrary Unions of Events

We have seen how to compute probabilities of events if they can be formed as unions of *disjoint* events. The natural question to ask is what to do if we want to compute the probability of the *union of non-disjoint events*. In order to reason about this problem, we first take a step back and think about



(a) (b)

Figure 2.1: 2.1a: Two overlapping events E_1 and E_2. Their intersection (the coloured region) gets counted twice if we add up their probabilities.
2.1b: Venn diagram with 3 events. First we deduct $E_1 \cap E_2, E_1 \cap E_3, E_2 \cap E_3$ in order to prevent double counting and then we add in $E_1 \cap E_2 \cap E_3$. Deductions and additions are indicated by pluses and minuses.

the outcomes of our probability space. We know that each event with non-zero probability contains at least one outcome (since $\mathbb{P}(\emptyset) = 0$, we can safely ignore the empty event). Let us assume that we take the union of events E_1 and E_2 with $E_1 \cap E_2 = E_3 \neq \emptyset$. This means that the outcomes in E_3 are contained in both E_1 and E_2. This situation is illustrated in Figure 2.1a. If we apply the rule that we have for disjoint events, we get

$$\begin{aligned} \mathbb{P}(E_1 \cup E_2) =& \mathbb{P}(((E_1 \backslash E_3) \cup E_3) \cup ((E_2 \backslash E_3) \cup E_3)) \\ =& \mathbb{P}(E_1 \backslash E_3) + \mathbb{P}(E_3) + \mathbb{P}(E_2 \backslash E_3) + \mathbb{P}(E_3) \end{aligned} \tag{2.4}$$

Exercise 2.2 *Try to justify each of the above equalities using your knowledge about sets, event spaces and the probability measure. In doing so, you will realize that one of the above equalities does not hold. Which one?*

What happens here is that we count E_3, the outcomes that are in both events of interest, twice. Thus, we will need to subtract E_3 one time. This leads us to the following formulation:

$$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2) \tag{2.5}$$

Notice that this is fully general in that it is true even if E_1 and E_2 were disjoint. In that case, their intersection would be empty. We can generalize this principle to the (countable) union of an arbitrary number of events.

This will give us a principled way of calculating the probability of any union of events. This calculation technique is know as the **Inclusion-Exclusion principle**.

Theorem 2.1 (Inclusion-Exclusion principle) *The probability of any (countable) union of events $E_1, \ldots, E_n$ can be computed as*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} (-1)^{i+1} \left(\sum_{j_1 < \ldots < j_i} \mathbb{P}\left(E_{j_1} \cap \ldots \cap E_{j_i}\right) \right) \tag{2.6}$$

We are going to proceed with a combinatorial proof of the Inclusion-Exclusion principle. It is very elegant but invokes the binomial theorem. For completeness sake we will prove the binomial theorem at the end of this chapter. For now, just trust us that it exists and is correct.

Proof We are going to focus on a particular outcome ω that is contained in m events which we call without loss of generality $E_1, \ldots, E_m$ for some $m < n$. Notice that we can safely neglect all events which do not contain ω, since ω is not going to contribute to their probability.

For all the E_i, $1 \leq i \leq m$ in which ω is contained, it is certainly true that ω is also contained in their intersections. The Inclusion-Exclusion-principle adds up or subtracts the probabilities of intersections of a given size. Notice that any intersection of more than m events will not contain ω as we intersect with at least one event that does not contain ω. Thus, we only need to consider intersections of our m ω-containing sets.

When $i = 1$ intersection is trivial, as it just consists of one event. How many ways are there to pick one out of m events? The answer is $\binom{m}{1}$. This is the number of times that ω contributes to the overall probability. At this point we have an overestimate of that probability (compare this to Equation (2.4)). Next we subtract the probabilities of the mutual intersections ($i = 2$). By the same reasoning as before, the contribution of ω is deducted $\binom{m}{2}$ times which gives us an underestimate since $\binom{m}{1} \leq \binom{m}{2}$ for $m \geq 3$. Since we are adding and subtracting in alternation, we will now keep flip-flopping between under- and over estimates. After considering all intersections of up to m sets, we should get the correct result, however.

What we want to prove is that the right-hand side of (2.6) counts ω's contribution to the overall probability exactly once (because this is what happens on the left hand-side of (2.6). That is, we have to prove that

$$1 = \sum_{i=1}^{m} (-1)^{i-1} \binom{m}{i} \tag{2.7}$$

We are right on our way towards exploiting the binomial theorem. Let

us first state it.

$$(p+q)^m = \sum_{i=0}^{m} \binom{m}{i} p^i q^{n-i} \tag{2.8}$$

Setting $p = (-1)$ and $q = 1$, and multiplying both sides with (-1), we obtain

$$-(-1+1)^m = -\sum_{i=0}^{n} \binom{m}{i} (-1)^i$$

which can be rewritten as

$$0 = -1 + \sum_{i=1}^{n} \binom{m}{i} (-1)^{i+1}\,, \tag{2.9}$$

because $\binom{m}{0} = 1$. Equation (2.9) implies (2.7) which we needed to prove. □

At this point we have done our fair share of math and found out how to calculate the probability of a union of events. We should ask ourselves what the probability of a union of events even tells us. Observe that an event occurs whenever we draw an outcome from our sample space that is contained in that event. By taking the union of events $E_1, \ldots, E_n$ we form a new event E that (possibly) contains more outcomes than each of the original events. Thus, the probability of the E will be higher than (or the same as) the probability of each of $E_1, \ldots, E_n$. What we are measuring then, is the probability that *any* of the events $E_1, \ldots, E_n$ occur. Crucially, we do not care anymore which one of them occurs.

What we are missing is a way to express the probability that a given number of events occur *together*. This concept is so important that we have a dedicated name for it, that of **joint probability**.

Definition 2.4 (Joint probability) *The joint probability of a (countable) set of events $\{E_1, \ldots, E_n\}$ is defined as*

$$\mathbb{P}(E_1 \cap \ldots \cap E_n)$$

Wow, that was simple! We don not event need to prove another rule for calculating the joint probability. After all, we already know how to take the intersection of sets. Annoyingly, one problem remains: our definition of event spaces does not guarantee that they contain the intersections of their members. Or does it? Well, let us see whether we can "paraphrase" what an intersection is.

$$E_1 \cap E_2 = \Omega \backslash ((\Omega \backslash E_1) \cup (\Omega \backslash E_2)) \tag{2.10}$$

All the operations on the right hand side are defined for events spaces. We have thus solved our problem since we have shown that we can indeed do intersection in event spaces. To convince yourself that this is correct, you may want to consult Figure 2.1a. Alternatively, you may also just realise that this is an instance of DeMorgan's laws which you should know from set theory. Notice that we do not claim that this is the only valid "paraphrase". Feel free to find others, if you like!

2.3 Probability of Complements of Events

At this point we are capable to do most probabilistic computations that we will encounter in this course. From here on, it is all about making our lives easier. For example, how would you solve the following problem.

Exercise 2.3 *You are observing a panel of 200 light bulbs and you know that at least one of them will light up once you press a button. What is the probability that any except the 87th bulb will light up? Note: this is a conceptual exercise. For the very keen ones, you can obtain the probability for each bulb to be turned on by typing the following into the Python interpreter:*

```
import numpy

probabilities = numpy.random.rand(1,200)
print probabilities/probabilities.sum()
```

The point of the above exercise is that it will be awfully cumbersome to compute the probability of the union of the singletons E_i where $1 \leq i \leq 200$ and $i \neq 87$. On the other hand we can easily look up $\mathbb{P}(E_{87})$. The question is whether we can exploit this simpler calculation to help us answer the original question. Here we will again make use of the properties of event spaces. For any event E in our event space we also have $\Omega \backslash E$ in the same space. Furthermore, E and $\Omega \backslash E$ are disjoint which by our probability axioms means that we can simply add up their probabilities if we want to calculate the probability of their union. But what's the union of E and $\Omega \backslash E$? It's exactly Ω. From axiom 3 we know that $\mathbb{P}(\Omega) = 1$. By simple algebraic manipulations we find that

$$\mathbb{P}(\Omega \backslash E) = 1 - \mathbb{P}(E) \tag{2.11}$$

Thus if we want to find the probability that any but the $87th$ bulb will light up, we simply compute the probability that the $87th$ bulb will light up will light up and subtract that from 1. This is a rather general strategy to simplify calculations whenever the probability of an event is hard to compute. Maybe the probability of the complement of that event will be easier to compute.

Exercise 2.4 *Show that in general*

$$\mathbb{P}(E_1 \backslash (E_1 \cap E_2)) = \mathbb{P}(E_1) - \mathbb{P}(E_1 \cap E_2)$$

2.4 Conditional Probability and Independence

After we have seen how to measure the probability of events, we are going to introduce another tremendously important concept, that of **conditional probability** measures.

Definition 2.5 (Conditional probability measure) *The probability of an event E_i conditioned on another event E_j with $\mathbb{P}(E_j) > 0$ is defined as*

$$\mathbb{P}(E_i|E_j) := \frac{\mathbb{P}(E_i \cap E_j)}{\mathbb{P}(E_j)}$$

Before we get into the math of conditional probabilities, let us try to understand the meaning of this concept. When we are computing the conditional probability of an event E_i, we re-scale with the probability of the conditioning event E_j. If $E_j \neq \Omega$, $\mathbb{P}(E_j)$ might be smaller than 1. Thus, this rescaling assumes *that E_j has already occurred*. In other words, we are excluding all outcomes that are not in E_j from further consideration (even though they may be in E_i). The interpretation of conditional probabilities is that they are the probabilities of events assuming that another event has already occurred.

Another interpretation is that when working with a conditional probability measure, we are in fact working in a new probability space, where $\Omega_{new} = E_2$, i.e. our new sample space is the conditioning event. Notice that this also means that our probability measure will change and become the measure from Definition 2.5.

Here comes the cool part: although we have introduced a new concept, all the properties of probability measures that we know by now will seamlessly carry over to conditional probabilities, if we can prove that the conditional probability measure is a probability measure according to our axioms.

Exercise 2.5 *Use the axioms from Definition 2.2 to prove that $\mathbb{P}(\cdot|E_j)$ is a probability measure.*

We will make use of conditional probabilities quite a lot in this course. We will later see a way in which they help us to decompose joint probability distributions. For now, we are going to focus on the fact that they are also related to the idea of independence of events.

Definition 2.6 (Independence) *Two events E_1, E_2 are said to be independent if*

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1) \times \mathbb{P}(E_2)$$

Independence of two events is denoted as $E_1 \perp E_2$.

This definition relates to conditional probabilities in the following way: assume that $E_1 \perp E_2$. Then we get

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)} = \frac{\mathbb{P}(E_1) \times \mathbb{P}(E_2)}{\mathbb{P}(E_2)} = \mathbb{P}(E_1)\,. \tag{2.12}$$

Hence, independence of two events $E_1 \perp E_2$ is equivalent with $\mathbb{P}(E_1|E_2) = \mathbb{P}(E_1)$.

Exercise 2.6 *Prove that $E_1 \perp E_2$ is also equivalent with $\mathbb{P}(E_2|E_1) = \mathbb{P}(E_2)$.*

Independence will prove to be a useful concept in later chapters. More precisely, we will often just *assume* that two events (or random variables – see the next chapter) are independent. Although such an independence assumption might not always hold in practice, it will allow us to formulate much simpler probabilistic models.

2.5 A Remark on the Interpretation of Probabilities*

This concludes our introduction of axiomatic probability theory. We know that a probability is a real number in $[0, 1]$. For all that we are going to do in this course (and in most follow-up courses) this is fully sufficient. However, some of you may wonder what a "natural" interpretation of probabilities would be. There are two dominating views on that. One postulates that if we were to take A LOT (read: almost infinitely many) samples from a sample space, the probability of an event is its frequency amongst these samples divided by the total number of samples taken. For those of you who know limits, this principle can be formalized as $\mathbb{P}(E) = \lim_{n\to\infty} \frac{\#E}{n}$. This view is known as the *frequentist view*.

The second view postulates that probabilities are an expression for degrees of belief. Basically, if you assign $\mathbb{P}(E)$ to an event E, then $\mathbb{P}(E)$ is the strength of your personal belief that E will occur. This latter view is known as the *Bayesian view*.

Which conception of probability you choose is a philosophical matter and does not really impact the math. That is why we will not care about this issue in this course. However, it is useful to at least be aware of these two views (if only to appear knowledgeable in a conversation you may have with your philosopher friends).

2.6 The Binomial Theorem

The binomial theorem from Equation 2.8 is actually not that hard to prove. We will do so by induction. As a base case we choose $m = 0$. Then the equality is easy to see.

$$(p+q)^0 = 1 = \binom{0}{0} p^0 q^0 \tag{2.13}$$

Next, we assume that the theorem holds for $m = n$. What we want to show is that it also holds for $m = n + 1$. We achieve this by algebraic manipulation.

$$(p+q)^{n+1} = (p+q)^n \times (p+q) \tag{2.14}$$

$$= (p+q)^n p + (p+q)^n q \tag{2.15}$$

$$= p \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} + q \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} \tag{2.16}$$

$$= \sum_{i=0}^{n} \binom{n}{i} p^{i+1} q^{n-i} + \sum_{i=0}^{n} \binom{n}{i} p^i q^{n+1-i} \tag{2.17}$$

$$= \sum_{j=1}^{n+1} \binom{n}{j-1} p^j q^{n+1-j} + \sum_{i=0}^{n} \binom{n}{i} p^i q^{n+1-i} \tag{2.18}$$

$$= \binom{n}{n} p^{n+1} q^{(n+1)-(n+1)} + \sum_{k=1}^{n} \binom{n}{k-1} p^k q^{n+1-k} \tag{2.19}$$

$$+ \binom{n}{0} p^0 q^{n+1} + \sum_{k=1}^{n} \binom{n}{k} p^k q^{n+1-k}$$

$$
\begin{aligned}
&= q^{n+1} + p^{n+1} + \sum_{k=1}^{n} \left(\binom{n}{k} + \binom{n}{k-1} \right) p^i q^{n+1-k} && (2.20)\\
&= q^{n+1} + p^{n+1} + \sum_{k=1}^{n} \left(\frac{n!}{k!(n-k)!} + \frac{n!}{(k-1)!(n-k+1)!} \right) p^i q^{n+1-k} && (2.21)\\
&= q^{n+1} + p^{n+1} + \sum_{k=1}^{n} \left(\frac{n!(n+1-k)}{k!(n+1-k)!} + \frac{n!k}{k!(n-k+1)!} \right) p^k q^{n+1-k} && (2.22)\\
&= q^{n+1} + p^{n+1} + \sum_{k=1}^{n} \left(\frac{n!(n+1)}{k!(n+1-k)!} \right) p^i q^{n+1-k} && (2.23)\\
&= q^{n+1} + p^{n+1} + \sum_{k=1}^{n} \binom{n+1}{k} p^k q^{n+1-k} && (2.24)\\
&= \sum_{i=0}^{n+1} \binom{n}{k} p^k q^{n-k} && (2.25)
\end{aligned}
$$

Let us clarify some parts of the proof. We use the induction hypothesis to expand the terms in Line 2.16. In Line 2.18, we switch the variable i in the first summand to $j = i + 1$. The reason why we do this is because we want to achieve congruence with the exponents of the second summand. In the following line we uniformly name the variables k. Since k has to run over a common range, we chop off the ends of both sums that stick out. In the first sum of line 2.18 that is the summand that corresponds to $j = n+1$ and in the second sum it is the summand that corresponds to $i = 0$. We pull out both of them in line 2.19 and then collapse the sums in line 2.20. The following lines are basically just an exercise in manipulation fractions. The jump from the second-to-last to the last line is allowed because

$$q^{n+1} = \binom{n+1}{0} p^0 q^{n+1-0}$$

and

$$p^{n+1} = \binom{n+1}{n+1} p^{n+1} q^{(n+1)-(n+1)}$$

which are exactly the quantities that we need to add to make our sum reach from 0 to $n + 1$. This completes the proof.

Further Reading

A very quick and dirty introduction to measure theory is provided by Maya Gupta and can be found here. If you are looking for something more extensive that also motivates event spaces and the like you may want to take a look at this script by Ross Leadbatter and Stamatis Cambanis (which has also been published as a book).