1 Probability spaces

We may be already familiar in one way or another with the idea of probability, a value between 0 and 1 we can assign to an outcome that describes its likelihood or chance of happening. The axioms of a probability space are not meant to contradict this intuition, but rather generalize it to make it more powerful.

Definition 1. A probability space consists of a sample space \( \Omega \), a set of outcomes; an event space \( \mathcal{F} \), a collection* of subsets of \( \Omega \); and a probability measure \( \mathbb{P} \), which assigns values in \([0, 1]\) to events in \( \mathcal{F} \). Every probability space satisfies three axioms:

1. Nonnegativity. \( \mathbb{P}(A) \geq 0 \) for every event \( A \in \mathcal{F} \).
2. Unit measure. \( \mathbb{P}(\Omega) = 1 \).
3. Countable additivity. For any finite or countably infinite collection of events \( (A_i)_{i=1}^{\infty} \) pairwise disjoint,
   \[
   \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).
   \]

First of all, why do we choose to assign probabilities to events rather than outcomes? We can continue assigning probabilities to outcomes: \( \mathbb{P}(\omega) := \mathbb{P}(\{\omega\}) \), \( \{\omega\} \) being the event with one outcome. However, we can also describe more general situations. We might know that event \( A \) will happen with probability 1, but not the chance of any particular outcome \( \omega \in A \). This model of the probability space allows us to set \( \mathcal{F} = \{\emptyset, A, A^c, \Omega\} \), \( \mathbb{P}(A) = 1 \), and \( \mathbb{P}(A^c) = 0 \) to describe the situation, without forcing us to define any specific \( \mathbb{P}(\omega) \).

We might also wonder why we need axiom 3 above. The reason becomes clear in the following section: we can find disjoint unions in complements, set differences, partitions, and even general unions. But there is a perhaps more direct reason: it expresses the simple relation that if we combine unrelated sets of outcomes (disjoint events), then we also combine their likelihoods. We require countability because uncountable sums (that are not integrals) are a bit troubling — we might get “1 is an uncountable sum of 0s.”

Finally, note that there are conditions on an event space \( \mathcal{F} \), namely it should be a \( \sigma\)-algebra: a nonempty collection closed under complements, countable unions, and countable intersections. This is for a fairly natural reason: if \( A \) and \( B \) are events, then we would want \{\( A \) but not \( B \)\}, \{\( A \) or \( B \)\}, \{\( A \) and \( B \)\} to also be valid events.
2 Some identities

This is a somewhat comprehensive list of identities that follow from the axioms; you may not need all of them in this course, but they are here for your reference.

a. **Empty set.** \( \mathbb{P}(\emptyset) = 0 \).

b. **Complement.** \( \mathbb{P}(A^c) = 1 - \mathbb{P}(A) \).

c. **Set difference.** \( \mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) \).

d. **Symmetric set difference (exclusive or).** \( \mathbb{P}(A \triangle B) = \mathbb{P}(A) + \mathbb{P}(B) - 2 \cdot \mathbb{P}(A \cap B) \).

e. **Monotonicity.** \( A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B) \), but not the converse.

f. **Law of total probability (partition).** If \( \{B_i\} \) is a countable partition of event \( B \) (possibly \( B = \Omega \)), then
\[
\mathbb{P}(A \cap B) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i) \cdot \mathbb{P}(B_i).
\]

g. **Union bound.** The following inequality holds, with equality if and only if the union is disjoint:
\[
\mathbb{P}\left( \bigcup_{i=1}^{\infty} A_i \right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).
\]

h. **Principle of inclusion-exclusion (union).** For any finite collection of events \( A_i \),
\[
\mathbb{P}\left( \bigcup_{i=1}^{n} A_i \right) = \sum_{k=1}^{n} \binom{n}{k} \left( -1 \right)^{k-1} \sum_{|I|=k} \mathbb{P} \left( \bigcap_{i \in I} A_i \right).
\]

i. **De Morgan’s laws.** \( \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cup B^c) = 1 \) and \( \mathbb{P}(A \cup B) + \mathbb{P}(A^c \cap B^c) = 1 \).

The proofs of these identities are fairly short and left as an exercise (I know, I know) in using the three axioms. Here are a few hints to help you get started:

- We might notice that (a) is a special case of (b) with \( A = \Omega \), and (b) is also a special case of (c) with \( A^c = \Omega \setminus A \). To prove (c), we can use that \((A \setminus B) \cup (A \cap B) = A\) is a disjoint union (why?).
- One definition of the symmetric set difference is \( A \triangle B = (A \setminus B) \cup (B \setminus A)\); is this a disjoint union?
- We’ll continue using axiom 3 for disjoint unions in (e) and (f). Identity (g) is slightly tougher, but it uses a cool idea: we can *disjointize* any countable union as \( A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus (A_1 \cup A_2)) \cup \cdots \)!

Finally, the formula for inclusion-exclusion may seem fairly intimidating, but the underlying idea is not so much: we repeatedly apply \( \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \) to avoid overcounting certain intersections.
\[ P(\bigcup A_i) = \sum_{\text{size 1}} P(A_i) - \sum_{\text{size 2}} P(A_i \cap A_j) + \sum_{\text{size 3}} P(A_i \cap A_j \cap A_k) - \cdots . \]

To find \( P(\bigcup A_i) \), we sum the individual \( P(A_i) \), then subtract out the double-counted \( P(A_i \cap A_j) \). But this leaves out any intersection of 3 events, so we have to add those back! This idea will be more easily seen from drawing out a Venn diagram than any description in words. A formal proof is also possible via induction.

You may notice that we’ve described the probability of complements and unions, but intersections are missing. To address the probabilities of intersections, we’ll need some definitions in the following section.

### 3 Conditional probability and independence

These two related concepts describe probabilities if or when we know an event \( B \) has “happened.” We expect that if an outcome \( \omega \) is not in \( B \), then \( P(\{\omega\}) = 0 \), and the “new probability of \( B \)” should be 1 in some sense, which motivates the following definition.

**Definition 2.** The conditional probability of event \( A \) given event \( B \), if \( P(B) > 0 \), is

\[ P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \]

There is a special case when the probability of \( A \) is unaffected given event \( B \): \( P(A \mid B) = P(A) \). This is logically equivalent to \( P(B \mid A) = P(B) \), and to the following definition:

**Definition 3.** Events \( A \) and \( B \) are independent iff \( P(A \cap B) = P(A) \cdot P(B) \).

We can now, among other things, describe the probability of an intersection:

j. **Independence.** The statement that events \( A \) and \( B \) are independent, denoted \( A \perp \perp B \), is equivalent to \( A \perp \perp B^c \), or \( A^c \perp \perp B \), or \( A^c \perp \perp B^c \).

k. **Bayes’ rule (conditional probability).** If \( P(B) > 0 \), then

\[ P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} = \frac{P(B \mid A) \cdot P(A) + P(B \mid A^c) \cdot P(A^c)}{P(B)} \cdot \]

l. **Chain rule (intersection).** For any finite collection of events \( A_i \),

\[ P\left( \bigcap_{i=1}^n A_i \right) = \prod_{i=1}^n P\left( A_i \mid \bigcap_{k=1}^{i-1} A_k \right). \]

We say that a possibly infinite collection of events \( \{A_i\}_{i \in I} \) is (mutually) independent if the product rule

\[ P\left( \bigcap_{i=1}^n A_i \right) = \prod_{i=1}^n P(A_i) \]
holds for every finite subcollection of events \( \{A_1, \ldots, A_n\} \). Mutual independence is a strictly stronger condition than pairwise independence: for example, consider \( \Omega = \{1, 2, 3, 4\} \), each outcome with probability \( \frac{1}{4} \). Then we can check that \( A = \{1, 2\} \), \( B = \{1, 3\} \), and \( C = \{2, 3\} \) are pairwise, but not mutually, independent.

Here are a few hints for showing the identities above:

- The equivalent definitions and statements of independence, as well as Bayes’ rule, should all follow from the definition of conditional probability.

- After we write out the product a bit more explicitly, induction would be a good approach for the chain rule. Note that \( \mathbb{P}(A_1 | \bigcap_{k=1}^{n-1} A_k) \) just denotes \( \mathbb{P}(A_1) \).

For some more material this note will not cover, the reader may refer to conditional independence and its many identities. Lastly, we leave the following exercise in axioms: show that probability preserves monotone limits.

m. If \( A_n \uparrow A \), then \( \mathbb{P}(A_n) \uparrow \mathbb{P}(A) \). That is,

\[
\text{if } A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots, \text{ then } \lim_{n \to \infty} A_n = \bigcup_{n=1}^{\infty} A_n \text{ and } \lim_{n \to \infty} \mathbb{P}(A_n) = \mathbb{P}\left(\lim_{n \to \infty} A_n\right).
\]

n. Likewise, if \( A_n \downarrow A \), then \( \mathbb{P}(A_n) \downarrow \mathbb{P}(A) \). That is,

\[
\text{if } A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots, \text{ then } \lim_{n \to \infty} A_n = \bigcap_{n=1}^{\infty} A_n \text{ and } \lim_{n \to \infty} \mathbb{P}(A_n) = \mathbb{P}\left(\lim_{n \to \infty} A_n\right).
\]