Note 01. Probability spaces

Alex Fu

Fall 2022

Preface

These course notes are intended to be a reference for most of the content in this course. However, please keep in mind that they may not include all of the material that is considered in scope — homeworks, discussions, labs, and the current semester's lectures. These notes may also contain some interesting but out-of-scope material, which will be indicated using an asterisk (*).

1 Axioms

Definition 1 (Probability space; Kolmogorov's axioms).

A **probability space** $(\Omega, \mathcal{F}, \mathbb{P})$ consists of a *sample space* Ω, a set of outcomes; an *event space* $\mathcal{F}$, a σ-algebra on Ω; 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 countable collection of disjoint events $\{A_i\}_{i=1}^{\infty}$,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Note that *countable* means finite or countably infinite.

Definition 2 (σ-algebra; event).

A σ-**algebra** or σ-*field* $\mathcal{F}$ on a set Ω is a nonempty collection of subsets of Ω that is closed under complements, countable unions, and countable intersections. A subset $A \subseteq \Omega$ is an **event** if and only if $A \in \mathcal{F}$. Every σ-algebra satisfies three axioms:

1. **Nonemptiness**. $\Omega \in \mathcal{F}$.
2. **Closure under complements**. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
3. **Closure under countable unions**. If $A_1, A_2, A_3, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The rest of this section will be devoted to explaining the definitions above. These fairly concise conditions summarize a great deal of insight into our understanding and application of probability, but they certainly can seem quite arbitrary without explanation.

1.1 The sample space Ω

Choice of sample space. The sample space is the basis of probability: there is randomness or uncertainty precisely because there may be multiple possible *outcomes*, *points*, or *samples*. For example, we could set $\Omega = \{\text{heads}, \text{tails}\}$ for a simple coin flip experiment, or we could set $\Omega = \{\text{all possible states of the universe}\}$. Naturally, we would like to choose the simpler sample space for modelling, so Ω is commonly finite, $\mathbb{N}$, $\mathbb{Z}$, or $\mathbb{R}$.

Limitations of fixed sample spaces. The probability space does force us to somehow "know" all of the possible outcomes beforehand when modelling situations. Otherwise, we may need to add a special outcome ω_* for "any unknown outcome," creating slightly unwieldly models. It may also seem difficult to modify sample spaces with dynamic information, like finding new outcomes or rejecting old outcomes. Fortunately, we will soon see a remedy in *conditional probability*.

1.2 Events $A \in \mathcal{F}$

Why is an event a set of outcomes? An event intuitively describes **determinable** or **measurable information**, something whose probability we wish to determine. A more natural definition of an event might be a function $E \colon \Omega \to \{\text{does not happen}, \text{happens}\}$ or $E \colon \Omega \to \{0, 1\}$.

Definition 3 (Indicator function).

The **indicator function** $\mathbb{1}_A \colon \Omega \to \{0, 1\}$ of a subset $A \subseteq \Omega$ is

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

Importantly, there is a bijection between subsets $A \subseteq \Omega$ and functions $\Omega \to \{0, 1\}$.

So, an event is just as well described by a subset $\{\omega : E(\omega) = 1\}$ as a function $E : \Omega \to \{0, 1\}$, and subsets provide a simpler definition. The correspondence $A \leftrightarrow \mathbb{1}_A$ will still be important: it allows us to translate between events and random variables, a key identity being $\mathbb{P}(A) = \mathbb{E}(\mathbb{1}_A)$.

Why do we need events? Or, why assign probabilities to events and not outcomes? Well, there is an easy way to reconcile the probability measure $\mathbb{P} \colon \mathcal{F} \to [0, 1]$ with our original formulation involving the probability function $p \colon \Omega \to [0, 1]$:

Definition 4 (Probability of an outcome).

$\mathbb{P}(\omega) := \mathbb{P}(\{\omega\})$, if the singleton set $\{\omega\}$ is an event in $\mathcal{F}$.

So, events give us more generality. Measuring the probability of events also more closely reflects reality: for instance, a noisy sensor cannot tell us its exact reading, but it can tell us an approximate range of readings. Or, many underlying outcomes result in the event $A = \{\text{device stops working}\}$, but we only have access to a single indicator light. In this case, $\mathcal{F} = \{\varnothing, A, A^c, \Omega\}$ more closely reflects what we can determine — $\mathbb{P}(A)$, but not the probability of any individual outcome.

Definition 5 (Special events).

a. An event happens **almost never** if it has probability 0. Such an event is also called a **null set** or *set of measure zero*.

b. An event happens **never** if it is equal to the empty set $\varnothing$.

c. An event happens **almost surely** (a.s.), *almost certainly*, or *almost always* if it has probability 1.

d. An event happens **surely** if it is equal to the sample space Ω.

Events of probability 0 and 1. A common misconception is that events with probability 0 are *impossible*, and events with probability 1 are *certain* to happen. A common *mis*correction is that events with probability 0 are still possible, only *unobservable*. Both can be true in practice: impossible events, like a coin with two tails sides landing on heads, do have probability 0, but we can be more careful.

An event A is impossible if it happens on no outcomes, i.e. if it is equal to $\varnothing$. However, almost never is not never: $\mathbb{P}(A) = 0$ may not imply that $A = \varnothing$. For example, randomly choose a real number in $[0, 1]$. The probability that you chose the number you did is 0, but this was neither impossible nor unobservable — it just happened, and you just observed it!

So, a better term than "impossible" or "unobservable" is **negligible**. An event that almost never happens is still possible to observe, but its probability is so low that we can treat it as if it never happens. This is really part of a broader measure-theoretic philosophy:

Events with probability measure zero are most often negligible — we only care about an event happening up to *almost surely*.

For a similar example, the integral of $f(x)$ over $[a, b]$ is the same as the integral over (a, b): the endpoints a and b have zero length, so they are negligible to the integral. There is always uncertainty in life — we cannot always be sure of events like $\{X = Y\}$ in *every* possible outcome, so the best we can do is to have $X = Y$ on all but a negligible set of outcomes. "In probability, the most sure we can be is almost sure."

1.3 The event space $\mathcal{F}$

Why is a σ-algebra defined the way it is? If events represent "measurable information," then the event space represents "all measurable information." So nonemptiness, or $\Omega, \varnothing \in \mathcal{F}$, means that we can always measure $\mathbb{P}(\Omega)$ and $\mathbb{P}(\varnothing)$. Closure under complements means that if we can find $\mathbb{P}(A)$, i.e. $A \in \mathcal{F}$, then we can also determine $\mathbb{P}(\text{not } A)$. Closure under unions and intersections means that the events $\{A \text{ or } B\}$ and $\{A \text{ and } B\}$ can be measured if A and B can.

Why countable unions and countable additivity? There are some who define probability spaces to have only *finite* additivity. But, countable additivity allows us to have countable sample spaces like $\mathbb{N}$, where we should be able to find the probability of each outcome:

$$\bigsqcup_{i=1}^{\infty} \{\omega_i\} = \Omega \implies \sum_{i=1}^{\infty} \mathbb{P}(\omega_i) = 1.$$

However, uncountable sample spaces can have subsets with length, area, volume, or even n-dimensional hypervolume, which become more complicated than just adding the probabilities of points. In a previous example, if there was *un*countable additivity, then each outcome in $[0,1]$ has probability 0, but the sample space $[0,1]$ has probability 1:

$$\sum_{\omega \in [0,1]} \mathbb{P}(\omega) = 0 \neq 1 = \mathbb{P}([0,1]).$$

Why not just use the power set 2^Ω? If Ω is countable, then the most common choice of $\mathcal{F}$ is indeed 2^Ω, because the power set *is* a σ-algebra. However, for uncountable sample spaces like $\Omega = \mathbb{R}$, there can exist pathological *nonmeasurable sets*, which cannot have a probability at all — any attempt to define their probability results in contradiction due to axioms of set theory.

Definition 6 (Borel σ-algebra).

The **Borel σ-algebra** on $\mathbb{R}$, denoted Borel$(\mathbb{R})$ or $\mathcal{B}(\mathbb{R})$, is the minimal σ-algebra that contains all of the open and closed intervals of $\mathbb{R}$. In particular, Borel$(\mathbb{R})$ contains all intervals of the form $(-\infty, x]$, and all singleton sets $\{x\}$.

If $\Omega = \mathbb{R}$, then the σ-algebra $\mathcal{F}$ will be taken to be Borel$(\mathbb{R}) \subset 2^{\mathbb{R}}$, unless otherwise stated.

1.4 The probability measure $\mathbb{P}$

Where do we get the probability values from in the first place? After all, the probability space axioms seem to assume that we already have some $\mathbb{P}\colon \mathcal{F} \to [0,1]$. One starting place is, naturally, *symmetry*: every outcome gets the same probability. There is a major problem: *there is no uniform measure over any countable sample space*. (See if you can show this by countable additivity.) We could instead choose common *distributions* to model natural patterns, or use relative frequencies or prior probability: see the following paragraph.

Interpretations of probability. There are at least two major schools: the **frequentist** view that probability is *empirical frequency*, the proportion of occurrences over infinitely many trials, and the

Bayesian view that probability is *degree of belief*, a subjective measure involving prior information and incorporating evidence. This philosophical discussion is far from resolved, but we can sidestep it for now: the axiomatic definition of probability spaces is interpretation-independent!

2 Consequences

You may not need all of the following identities in this course, but they are here for your reference.

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

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)**. $\mathbb{P}(A \triangle B) = \mathbb{P}(A) + \mathbb{P}(B) - 2\,\mathbb{P}(A \cap B)$.

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

f. **Law of total probability (partition)**. If $\{B_i\}_{i=1}^{\infty}$ is a countable partition of the event B, possibly $B = \Omega$, then

$$\mathbb{P}(A \cap B) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i).$$

g. **Union bound (union I)**. The following inequality known as *countable subadditivity* holds in general, 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 II)**. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. In general, for any finite collection of events A_i,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} \left[(-1)^{k-1} \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I| = k}} \mathbb{P}\left(\bigcap_{i \in I} A_i\right) \right].$$

i. **De Morgan's laws (union $\leftrightarrow$ intersection)**. $\mathbb{P}(A \cup B) = 1 - \mathbb{P}(A^c \cap B^c)$; dually, $\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A^c \cup B^c)$. In general, for any countable collection of events A_i,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = 1 - \mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i^c\right) \text{ and } \mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i^c\right).$$

j. **Continuity from below**. If $A_n \uparrow A$, then $\mathbb{P}(A_n) \uparrow \mathbb{P}(A)$. That is,

$$\text{if } A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots \text{ and } \lim_{n\to\infty} A_n := \bigcup_{n=1}^{\infty} A_n = A,$$

$$\text{then } \mathbb{P}(A_1) \leq \mathbb{P}(A_2) \leq \mathbb{P}(A_3) \leq \cdots \text{ and } \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\Big(\lim_{n\to\infty} A_n\Big) = \mathbb{P}(A).$$

k. **Continuity from above**. Likewise, if $A_n \downarrow A$, then $\mathbb{P}(A_n) \downarrow \mathbb{P}(A)$. That is,

$$\text{if } A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots \text{ and } \lim_{n\to\infty} A_n := \bigcap_{n=1}^{\infty} A_n = A,$$

$$\text{then } \mathbb{P}(A_1) \geq \mathbb{P}(A_2) \geq \mathbb{P}(A_3) \geq \cdots \text{ and } \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\Big(\lim_{n\to\infty} A_n\Big) = \mathbb{P}(A).$$

The proofs of the above identities are fairly short and reasonable, so I leave them as exercises to you (I know, I know) in applying the probability axioms. Here are a few hints to get you started:

0. **Venn diagrams** are a great place to start for all of these identities: probability has a really nice interpretation as area, so visualization can be incredibly helpful.

1. Countable additivity implies finite additivity, the weaker condition that the probability of any finite disjoint union is the sum of the individual probabilities: we can just take the countable collection $A_1, A_2, \ldots, A_n, \varnothing, \varnothing, \ldots$.

2. *Disjoint unions* with finite additivity will be your friend. Are any of the following — an event and its complement; A and $B \setminus A$; $A \triangle B := (A \setminus B) \cup (B \setminus A)$; any partition — disjoint? If so, what are their unions?

3. The union bound uses a sweet idea: we can **disjointize** any countable union! Consider the events $B_n = A_n \setminus (\bigcup_{i=1}^{n-1} A_i)$. Are they disjoint, and what is $\bigcup_{i=1}^{n} B_i$? See if you find any interesting parallels to the *Gram–Schmidt procedure* for removing linear dependence.

4. The formula for inclusion-exclusion certainly seems intimidating, but its underlying idea is not so much: we inductively apply the case of $n = 2$ to avoid overcounting certain intersections:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{\text{size 1}} \mathbb{P}(A_i) - \sum_{\text{size 2}} \mathbb{P}(A_i \cap A_j) + \sum_{\text{size 3}} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots .$$

5. The final two identities say that probability has *monotone continuity*: it preserves monotone limits, which are limits of increasing or decreasing sequences. We cannot always exchange the $\lim$ sign and $\mathbb{P}$ like we did, but what about $A_1 \subseteq A_2 \subseteq \cdots$ seems to make it possible?

   Disjointization may come in handy again. This time, the events B_n can be visualized as the shells of Russian nesting dolls, and the sums of their probabilities experience some beautiful telescoping cancellation. Then, can we reduce continity from above to that from below?

There is a great deal more to explore: the Bonferroni inequalities; countable intersections and unions of special events; upper and lower bounds on probabilities of intersections; generators of σ-algebras; product spaces; the language of measure theory; alternative axiomizations. However, these topics are wildly out of scope, and thus left only to the interested reader.

For now, we will take a closer look at the interpretation of "events as information," and in the process find another method of finding the probabilities of intersections, as well as a graphical model for dependencies between events.

$\blacksquare$