Probability Theory on General Spaces

Michael Betancourt

July 2023

Table of contents

Previously in Chapter One we introduced measure and probability theory over sets with only a finite number of elements. We saw in Chapter Two, however, that many of the most mathematical spaces we encounter in practical applications, like the integers and the real line, feature not a finite number of elements but rather countably infinite and even uncountably infinite numbers of elements. Unfortunately extending measure and probability theory to more general spaces like these is not always straightforward.

In this chapter we will investigate the difficulties in defining measure and probability theory on general mathematical spaces, with a focus on concepts instead of technical details. We will first discuss why measures allocated to individual elements does not, in general, provide enough information to define a consistent allocation for all subsets. Then we will consider how certain pathological subsets on some spaces can obstruct consistent allocations over the full power set, and how we can systematically remove these obstructions in practice. Finally we will present the most general form of measure and probability theory that can be applied to any mathematical space and then discuss some common applications.

1 Allocation Over Elements

Recall that in Chapter One we first defined measures and probability distributions as allocations over the individual elements in a finite set. More formally we were able to define a measure as a function that mapped each element to its allocation of the total measure,

$$\begin{aligned} \mu : X &\to [0, \infty] \\ x &\mapsto \mu(x) \quad . \end{aligned}$$

This element-wise allocation then allowed us to define the measure allocated to subsets. In particular the measure allocated to a subset $\mathsf{x} \subset X$ was unambiguously determined by summing up the measures allocated to the included elements,

$$\mu(\mathsf{x}) = \sum_{x \in \mathsf{x}} \mu(x).$$

On finite spaces this construction gives us a *consistent* allocation in the sense that the total measure is always preserved no matter how we might decompose the ambient set into subsets.

Conveniently this construction does extend to spaces with countably infinite numbers of elements, such as the integers. In these spaces every subset contains at most a countably infinite number of elements and sums of measures will always converge to well-defined values. Element-wise measure allocations on finite and countably infinite spaces are also known as **mass functions**, with element-wise probability allocations also known as **probability mass functions**.

Unfortunately the element-wise construction does not extend any further. Once we consider spaces with uncountably infinite numbers of elements, such as the real numbers, we have to confront subsets with uncountably infinite numbers of elements where sums start to misbehave.

Consider for example a subset x where each of the included elements has been allocated exactly zero measure. If x contains only a finite or countably infinite number of elements then the sum of these zero measures *always* yields zero.

When x contains an uncountably infinite number of elements, however, the sum of the individual element measures is not necessarily zero. In fact it can give *any* value between zero and infinity; uncountably infinite spaces have so many elements that we can very much get something from nothing!

Ultimately this means that on general spaces the allocation of measure to individual elements *does not provide enough information* to uniquely determine what measure should be allocated to every combination of those elements. In order to completely define a measure we need to specify what those subset allocations are ourselves.

2 Allocation Over All Subsets

In Chapter One we also considered defining a measure by specifying allocations to each subset in the power set,

$$\begin{aligned} \mu : 2^X &\to [0, \infty] \\ \mathsf{x} &\mapsto \mu(\mathsf{x}) \quad . \end{aligned}$$

Importantly these subsets allocations needed to be consistent with each other to match the behavior of those derived from individual element allocations. For any finite collection of disjoint subsets we should have

$$\mu(\cup_{i=1}^{I} \mathsf{x}) = \sum_{i=1}^{I} \mu(\mathsf{x}_i).$$

For finite spaces this construction is excessive; the subset allocations contain an abundance of redundant information. Because we also can derive subset allocations from element-wise allocations on countably infinite spaces, this construction is unnecessary there as well.

On the other hand at least some subset allocation is strictly necessary for fully defining measures on uncountably infinite spaces, and hence mathematical spaces in general. The only question is whether or not *consistent* subset allocations are even possible on these more sophisticated spaces.

2.1 Consistent Allocations

Before answering this question let's take a second to define exactly what kind of consistency we need. Because finite spaces feature only a finite number of subsets we only ever have to consider the consistency of a finite collection of subsets at a time. More formally if

$$\{\mathsf{x}_1, \ldots, \mathsf{x}_i, \ldots, \mathsf{x}_I\}$$

is any finite collection of disjoint subsets,

$$\mathsf{x}_i \cup \mathsf{x}_{i' \neq i} = \emptyset,$$

then a consistent measure should give

$$\mu(\cup_{i=1}^{I} \mathsf{x}) = \sum_{i=1}^{I} \mu(\mathsf{x}_i).$$

Regardless of how many elements the ambient space contains, consistency of a measure over any finite collection of subsets is known as **finite additivity**.

More general spaces can feature infinitely many subsets, and hence different possible notions of additive consistency. For example on a countably infinite space the subset allocations derived from a mass function are consistent across countably infinite collections of subsets. If

$$\{\mathsf{x}_1, \ldots, \mathsf{x}_i, \ldots\}$$

is any countably infinite collection of disjoint subsets with with

$$\mathsf{x}_i \cup \mathsf{x}_{i' \neq i} = \emptyset$$

then

$$\mu(\cup_i \mathsf{x}) = \sum_i \mu(\mathsf{x}_i).$$

This is known as **countable additivity**.

The question is then whether measures with finite additivity are sufficiently useful for practical application or if we need to consider countably additive measures, let alone measures that might be additive over even larger collections of subsets.

For example a common problem that arises is practice is reconstructing the measure allocated to a general subset from the measures allocated to particularly nice subsets that are easier with which to work. If we could always decompose a generic subset into the disjoint union of a finite number of nice subsets then finite additivity would be sufficient for this task. On the other hand if we could decompose a generic subset into the disjoint union of only a countably infinite number of nice subsets then countable additivity would be sufficient. Potentially some subsets might be decomposable only into an uncountably infinite number of subsets in which case we would need even stronger notions of additivity!

Fortunately for us we don't have to go to that last extreme. In turns out that on most spaces that we'll encounter in practice, and typical notions of "nice" subsets, countable additivity is sufficient for reconstructing the measure allocated to more general subsets.

To demonstrate let's consider the two-dimensional real plane $\mathbb{R}^2$ and a measure that is partially defined through it's allocations to **rectangular** subsets (Figure 1). In general a non-rectangular subset, in this case a disk, can be crudely approximated by a single rectangular subset. The disk can be approximated more precisely as the disjoint union of many different rectangular subsets, but that will never exact reconstruct the disk. Only when we incorporate a countably infinite number of rectangular subsets can be reconstruct the disk without any error.

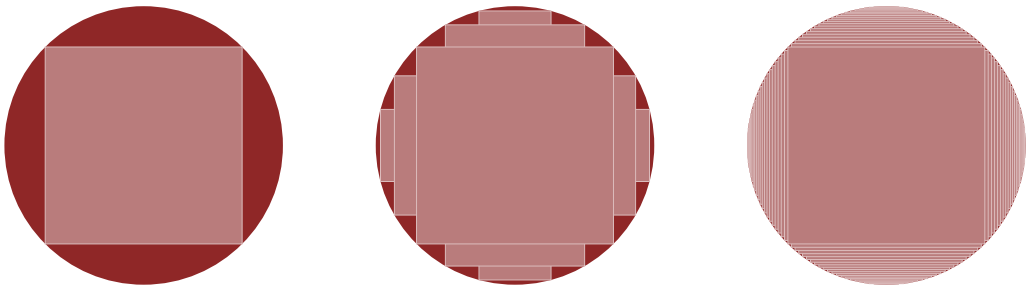

Figure 1: On a two-dimensional real plane $\mathbb{R}^2$ a non-rectangular disc can be approximated, but not exactly reconstructed, by the finite union of different rectangular subsets. In order to exactly reconstruct a non-rectangular subset we need to include a countably infinite number of rectangular subsets. If measures are countably additive then we can use this decomposition to reconstruct the measure allocated to the disk by adding up the measures allocated to the infinitely many rectangular subsets in the reconstruction.

Ultimately countable additive measures give us the mathematical flexibility we need for many practical applications.

2.2 Sub-Additivity and Super-Additivity

Ideally we would be able to define measures that are additive over any countably infinite collection of disjoint subsets on any space. Unfortunately mathematics is not always kind, and many seemingly well-behaved space feature pathological subsets that obstruct countable additivity.

Specifically many uncountably infinite spaces feature disjoint subsets that will always behave

sub-additivity,

$$\begin{aligned} \mathsf{x}_1 \cup \mathsf{x}_2 &= \emptyset \\ \mu(\mathsf{x}_1 \cup \mathsf{x}_2) &< \mu(\mathsf{x}_1) + \mu(\mathsf{x}_2) \end{aligned}$$

no matter how we try to define the allocation! In other words the power set will always be infiltrated by certain subsets that are always less than the sum of their parts and obstruct a consistent definition of measure.

At the same time we can generally show that there exist disjoint subsets that are **super-additive**,

$$\begin{aligned} \mathsf{x}_1 \cup \mathsf{x}_2 &= \emptyset \\ \mu(\mathsf{x}_1 \cup \mathsf{x}_2) &> \mu(\mathsf{x}_1) + \mu(\mathsf{x}_2). \end{aligned}$$

In other words if the measure allocated to these subsets and the measure allocated to their union then we will always appear to end up with measure than what had been initially allocated.

What makes these pathological subsets even more awkward is that we can't actually construct them from explicit conditions. Given typical assumptions about infinity all we can do is prove that these subsets exist. These phantom subsets are known as **non-constructive** objects.

That said because the misbehaving subsets are non-constructive we don't really need to consider them in any practical application of measure theory. If we could consistently filter them out of the full power set then we would be able to define consistent measures over the remaining subsets, and that would be sufficient for any practical application.

3 σ-Algebras

Because the term "σ-algebra" is often thrown around in measure and probability theory without much explanation it is often seen as an impenetrable concept that defies explanation. In reality, however, σ-algebra are simply a way to consistently filter out subsets from the power set.

3.1 Filtering Subsets

We can always filter the power set by removing certain subsets. The difficultly is ensuring that no application of the three set operations would ever lead us back to the excised subsets and reveal a "hole" in the remaining collection of subsets. In other words we need our filtered collection of subsets to be *closed* under the three set operations so that there is no risk of accidentally recreating a subset outside of the collection.

In particular if the subset $\mathsf{x} \subset X$ is in our filtered collection then so too should be the complement x^c. If this is true then anytime we apply the complement operator to a subset in our collection we are guaranteed to always see another subset in our collection.

Similarly for every pair of subsets $\mathsf{x}_1 \subset X$ and $\mathsf{x}_1 \subset X$ in a filtered collection the union $\mathsf{x}_1 \cup \mathsf{x}_2$ and intersection $\mathsf{x}_1 \cap \mathsf{x}_2$ should also be in the collection. In order to ensure closure under repeated applications of the union and intersection operators we need the union and intersection of any countably infinite sequences of subsets to also be in the filtered collection.

A **σ-algebra** is any collection of subsets that is closed under complements, countable unions, and countable intersections. In other words a σ-algebra is just any consistent filtering of the power set. I will use a calligraphic font to refer to σ-algebras so that if X is a space then $\mathcal{X} \subset 2^X$ will denote a σ-algebra defined on that space.

A set equipped with a σ-algebra, $(X, \mathcal{X})$ is known as a **measurable space**. I will refer to X as the **ambient set**, or the **ambient space** if it is also equipped with additional structure. Similarly the elements of a σ-algebra are known as **measurable sets** while any subsets in the power set but not in the σ-algebra are referred to as **non-measurable** subsets.

When non-measurable subsets are misbehaving subsets they reveals the subtle, and often counterintuitive, pathologies inherent to that space. By working with *sigma*-algebras directly we can avoid these awkward pathologies entirely.

3.2 Generating σ-Algebras

Now that we've defined how a consistent sub-collection of subsets behaves we need to consider how to construct these σ-algebras in practice. One particularly useful way to build up σ-algebras is to *generate* them by repeatedly applying the three set operations to an initial collection of subsets.

For example consider an initial collection of two subsets

$$\{\mathsf{x}_1, \mathsf{x}_2\}.$$

Applying the complement operator gives us two subsets that fall outside of the initial collection,

$$\{\mathsf{x}_1^c, \mathsf{x}_2^c\},$$

Similarly applying the union operator gives

$$\{\mathsf{x}_1 \cup \mathsf{x}_2\}$$

while applying the intersection operator gives

$$\{\mathsf{x}_1 \cap \mathsf{x}_2\}.$$

To ensure closure we have to add *all* of these subsets to our initial collection,

$$\{\mathsf{x}_1, \mathsf{x}_2, \mathsf{x}_1^c, \mathsf{x}_2^c, \mathsf{x}_1 \cup \mathsf{x}_2, \mathsf{x}_1 \cap \mathsf{x}_2\}.$$

At this point we iterate, applying the complement operator to every subset and the union and intersection operators to every finite and countably infinite sub-collection of subsets to generate an even larger collection of subsets. When the set operations no longer return new subsets the final collection of subsets defines a σ-algebra.

A convenient feature of this procedure is that if we start with a collection of constructive subsets then we will *always* end up with a σ-algebra that is free of any non-constructive subsets and their pathological behaviors. To ensure that we don't filter out any well-behaved subsets in the process we just have to make sure that our initial collection is sufficiently large.

Conveniently when working on a topological space we already have a natural collection of subsets that we can use to generate a σ-algebra – the defining topology itself! The σ-algebra generated by repeatedly applying all three set operations to the subsets in a topology is known as a **Borel** σ-algebra. In other words a Borel σ-algebra is the unique σ-algebra comprised of all of the open and closed subsets.

Every space that we will consider in this book will be a topological space. Consequently we can always use the corresponding Borel σ-algebra to remove any undesired subsets that would obstruct the definition of a consistent measures and probability distributions. Indeed Borel σ-algebras are so common that they are often take for granted, with any reference to a "measurable space" implicitly assuming a topological space and its corresponding Borel σ-algebras to filter out any inconsistent behavior.

For example finite and countably infinite spaces are almost always equipped with discrete topologies. Because discrete topologies contain all of the atomic sets the σ-algebras derived from them will always be the full power set. There are no pathological behaviors that we have to avoid in these cases.

On the the other hand the Borel σ-algebra derived from the topology that defines the real line filters out all of the non-constructive subsets and their undesired behaviors. This results in a σ-algebra that is strictly smaller than the full power set of the real line.

A Borel σ-algebra is sufficient for removing any counterintuitive behavior from a topological space but in more technical mathematical work there are circumstances where slightly larger σ-algebras may be more convenient. In more applied practice we can safely assume a Borel σ-algebra or any extension that might be needed to avoid any technical problems.

4 Measures and Probability Distributions

With all of that work we are finally ready to define a theory for allocating any conserved, but not necessarily finite, quantity across a general mathematical space.

4.1 Formal Definitions

A **measure** on any measurable space $(X, \mathcal{X})$ is a function from the σ-algebra $\mathcal{X}$ to the extended positive real line,

$$\begin{aligned} \mu : \mathcal{X} &\to [0, \infty] \\ \mathsf{x} &\mapsto \mu(\mathsf{x}) \quad , \end{aligned}$$

that is **countably additive**,

$$\mu(\cup_i \mathsf{x}) = \sum_i \mu(\mathsf{x}_i)$$

for any countably infinite collection of subsets

$$\{\mathsf{x}_1, \ldots, \mathsf{x}_i, \ldots\}$$

that are mutually disjoint,

$$\mathsf{x}_i \cup \mathsf{x}_{i' \neq i} = \emptyset.$$

On finite and countably infinite spaces we can always take $\mathcal{X} = 2^X$ and ensure countable additivity by allocating measure to individual elements and then deriving the measure allocated to subsets by summing over the measure allocated to the included elements. When working with more sophisticated ambient spaces, however, the pair $(X, 2^X)$ may not admit *any* consistent measures. In these cases we have to consider smaller σ-algebras in order for measures to exist.

A set equipped with not only a σ-algebra but also a measure, in other words the triple $(X, \mathcal{X}, \mu)$ is known as a **measure space**. Again I will refer to X as the ambient set or ambient space as appropriate.

If the total measure is finite, $\mu(X) < \infty$, then μ is referred to as a **finite measure**. In this case we can always normalize the measure by $\mu(X)$ to define a proportional allocation.

A **probability distribution** (Figure 2) on any measurable space $(X, \mathcal{X})$ is a function from the σ-algebra $\mathcal{X}$ to the closed unit interval,

$$\begin{aligned} \pi : \mathcal{X} &\to [0, \infty] \\ \mathsf{x} &\mapsto \mu(\mathsf{x}) \quad , \end{aligned}$$

with

$$\pi(X) = 1$$

and

$$\pi(\cup_i \mathsf{x}) = \sum_i \pi(\mathsf{x}_i)$$

for any countably infinite collection of subsets

$$\{\mathsf{x}_1, \ldots, \mathsf{x}_i, \ldots\}$$

that are mutually disjoint,

$$\mathsf{x}_i \cup \mathsf{x}_{i' \neq i} = \emptyset.$$

These properties are also known collectively as the **Kolmogorov axioms**.

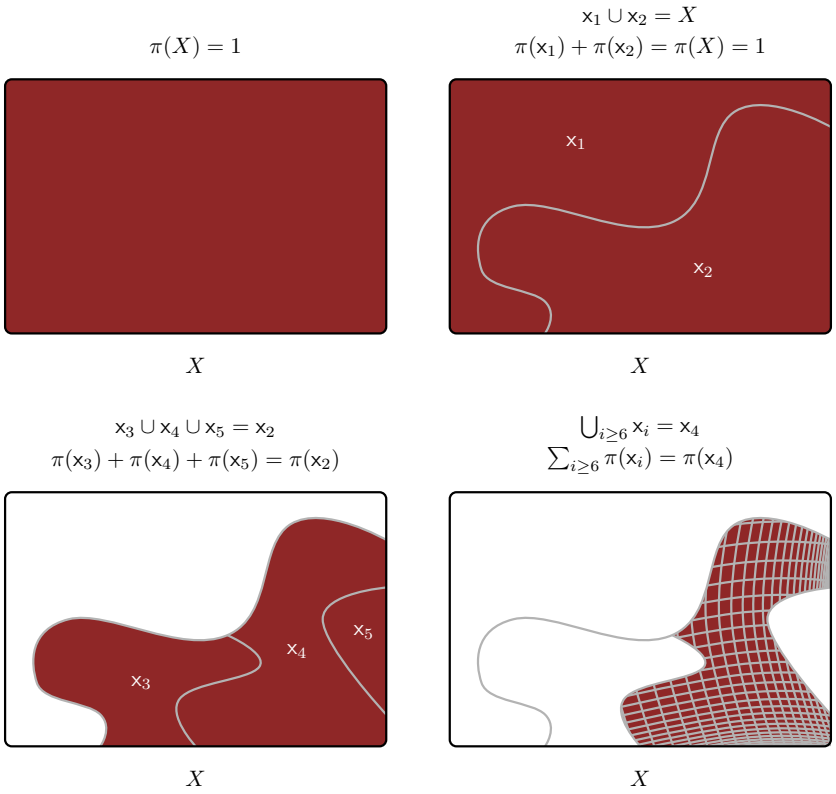

Figure 2: A probability distribution defines a proportional allocation across a measurable space such that no matter how we slice up the ambient set X into measurable subsets the total probability is always preserved.

A set equipped with a σ-algebra and a probability distribution is known as a **probability space**. Sometimes the combination $(X, \mathcal{X}, \pi)$ is also known as a **probability triple**.

Probability spaces are also sometimes denoted by

$$x \sim \pi,$$

where $x \in X$ indicates the ambient set and a σ-algebra is taken for granted. In words this reads "the variable x is distributed according to π" or "the variable x follows the distribution π". That said because probability distributions are generally defined over measurable subsets and not individual elements a more precise description be would be "the variable x takes values in a set X that is equipped with a probability distribution π". The emphasis on variables instead

of spaces in this notation is related to the awkward notion of a "random variable" which we will discuss in more detail in Chapter Ten.

4.2 Derived Properties

Although these definitions might appear to be a bit stark, we can derive all of the usual rules of measure and probability theory from them.

For example consider one measurable subset that is strictly smaller than another,

$$\mathsf{x}_1 \subset \mathsf{x}_2 \in \mathcal{X}.$$

In this case we can always write

$$\mathsf{x}_2 = \mathsf{x}_1 \cup \mathsf{x}_3$$

for the non-empty, measurable subset of elements that are in x_2 but not in x_1. Applying countable additivity then gives

$$\begin{aligned} \pi(\mathsf{x}_2) &= \pi(\mathsf{x}_1 \cup \mathsf{x}_3) \\ &= \pi(\mathsf{x}_1) + \pi(\mathsf{x}_3) \\ &\geq \pi(\mathsf{x}_1) \end{aligned}$$

because $\pi(\mathsf{x}_3) \geq 0$. In other words larger measurable subsets are always allocated more or equal probability than smaller subsets.

Similarly because any subset and its complement are disjoint and combine to reconstruct the full set we always have

$$\begin{aligned} 1 &= \pi(X) \\ &= \pi(\mathsf{x} \cup \mathsf{x}^c) \\ &= \pi(\mathsf{x}) + \pi(\mathsf{x}^c) \end{aligned}$$

or

$$\pi(\mathsf{x}^c) = 1 - \pi(\mathsf{x}).$$

In order to work with two measurable subsets $\mathsf{x}_1, \mathsf{x}_2 \in \mathcal{X}$ that might not be disjoint (Figure 3) we have to consider the elements that are unique to each,

$$\mathsf{x}_{1\,2} = \{x \in X \mid x \in \mathsf{x}_1, x \notin \mathsf{x}_2\}$$

and

$$\mathsf{x}_{2\,1} = \{x \in X \mid x \in \mathsf{x}_2, x \notin \mathsf{x}_1\},$$

and the elements that are shared,

$$\mathsf{x}_1 \cap \mathsf{x}_2 = \{x \in X \mid x \in \mathsf{x}_1, x \in \mathsf{x}_2\}.$$

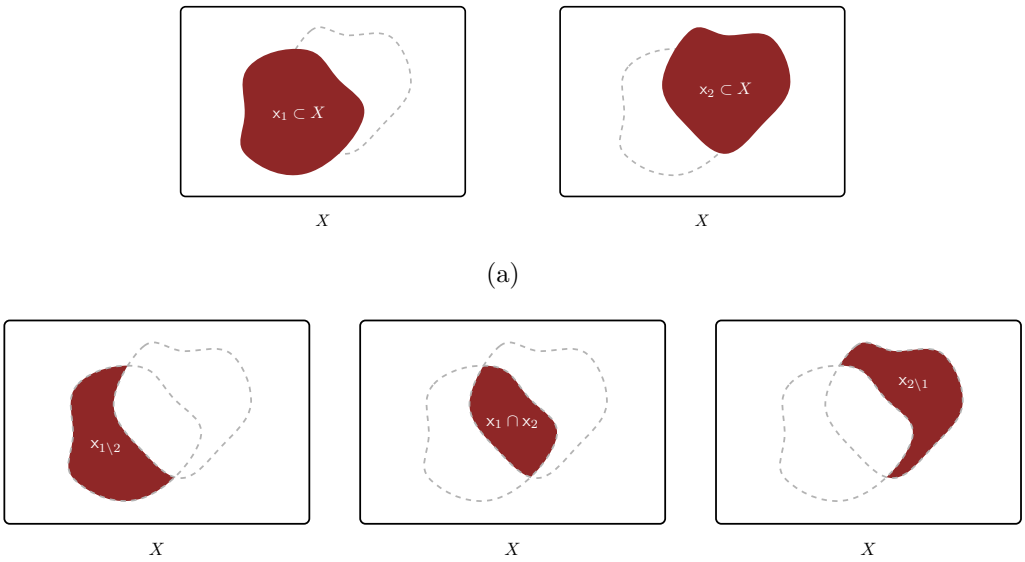

Figure 3: The union of (a) two overlapping subsets x_1 and x_2 can always be decomposed into the union of (b) three disjoint subsets. One disjoint subset $\mathsf{x}_{1\,2}$ encapsulates the elements unique to x_1, another $\mathsf{x}_{2\,1}$ encapsulates the elements unique to x_2, and finally the intersection $\mathsf{x}_1 \cap \mathsf{x}_2$ encapsulates the elements shared by the two input subsets.

This then allows us to decompose x_1, x_2, and their union into disjoint, measurable subsets (Figure 4,)

$$\begin{aligned}
\mathsf{x}_1 &= \mathsf{x}_{1\,2} \cup (\mathsf{x}_1 \cap \mathsf{x}_2) \\
\mathsf{x}_2 &= \mathsf{x}_{2\,1} \cup (\mathsf{x}_1 \cap \mathsf{x}_2) \\
\mathsf{x}_1 \cup \mathsf{x}_2 &= \mathsf{x}_{1\,2} \cup (\mathsf{x}_1 \cap \mathsf{x}_2) \cup \mathsf{x}_{2\,1}.
\end{aligned}$$

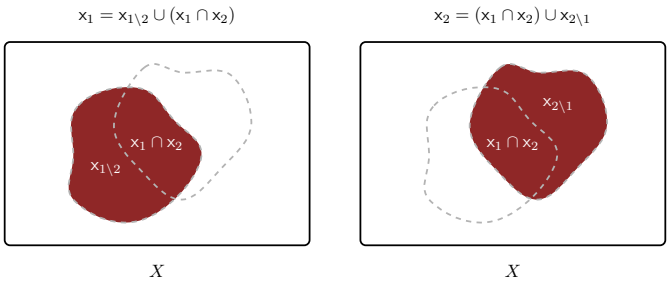

Figure 4: The disjoint subsets introduced in (Figure 3) can also be used to reconstruct the two input subsets individually.

Applying countable additivity to these three decompositions gives a system of equations

$$\begin{aligned}
\pi(\mathsf{x}_1) &= \pi(\mathsf{x}_{1\,2}) + \pi(\mathsf{x}_1 \cap \mathsf{x}_2) \\
\pi(\mathsf{x}_2) &= \pi(\mathsf{x}_{2\,1}) + \pi(\mathsf{x}_1 \cap \mathsf{x}_2) \\
\pi(\mathsf{x}_1 \cup \mathsf{x}_2) &= \pi(\mathsf{x}_{1\,2}) + \pi(\mathsf{x}_1 \cap \mathsf{x}_2) + \pi\mathsf{x}_{2\,1}).
\end{aligned}$$

Adding the first two together gives

$$\pi(\mathsf{x}_1) + \pi(\mathsf{x}_2) = \pi(\mathsf{x}_{1\,2}) + 2\,\pi(\mathsf{x}_1 \cap \mathsf{x}_2) + \pi(\mathsf{x}_1 \cap \mathsf{x}_2)$$

or

$$\pi(\mathsf{x}_{1\,2}) + \pi(\mathsf{x}_1 \cap \mathsf{x}_2) + \pi(\mathsf{x}_1 \cap \mathsf{x}_2) = \pi(\mathsf{x}_1) + \pi(\mathsf{x}_2) - \pi(\mathsf{x}_1 \cap \mathsf{x}_2).$$

Substituting this into the third equation finally gives

$$\pi(\mathsf{x}_1 \cup \mathsf{x}_2) = \pi(\mathsf{x}_1) + \pi(\mathsf{x}_2) - \pi(\mathsf{x}_1 \cap \mathsf{x}_2).$$

4.3 Measures and Probability Distributions In Practice

The formal definition of measures and probability distributions tell us what form the consistent allocation of any quantity on any measurable space has to take, but it does not necessarily provide a way for constructing explicit allocations in practice. Specifically in almost all circumstances it is infeasible, if not outright impossible, to exhaustively specify the measure or

probability allocated to *every* subset in the ambient σ-algebra. Constructing and then storing infinitely large databases linking each measurable subset to their allocations is not particularly practical!

In some cases we can define a useful measures and probability distributions by specifying the allocation to only some of the measurable subsets and then deriving the allocations to the rest with countable additivity. For example in finite and countable spaces we need to specify only the allocations to atomic subsets, and in some uncountable spaces we might need to specify only the allocations to certain nice subsets.

That said in most of these cases those reduced allocations are still impractical to specify one-by-one. Because of that our introduction to general measure and probability theory will have to remain a bit abstract, without any explicit examples, for the time being.

In applied problems measures and probability distributions are almost always defined *algorithmically*, with rules to evaluate the measure or probability allocated to a subset on the fly instead of storing and retrieving the allocation from an exhaustive specification. We will introduce two of the most important algorithm representations of measures and probability distributions, and use them to define many useful allocations, in Chapter Five and Chapter Eight.

5 Interpretations of Measure And Probability

To this point our treatment of measure and probability theory has been purely mathematical. A measures defines the allocations of some abstract conserved quantity across some abstract measurable space; a probability distribution defines a proportional allocations. This mathematical construction cannot be endowed with any particular interpretation until we use it to model something.

In this section we'll review some of the most common applications of measure and probability theory and the particular interpretations those applications create.

5.1 Modeling Physical Distributions

One immediate application of measure theory is to model the behavior of a physical quantity, such as mass or electric charge. For example physical mass can be distributed across a solid object in a variety of different ways, with the exact distribution affecting how that object interacts with the surrounding environment. Similarly the distribution of charge across the surface of a conducting object defines its electrostatic properties.

In some physical systems the distribution can also change with time and influence the dynamics of the system. Time-dependent measures that quantify how the distribution of a physical quantity evolves are a common feature of many mathematical physical theories.

5.2 Modeling Populations

A similar application is modeling the selection of individuals, or the properties of individuals, from a larger population. Each time we *sample* a subset of individuals from the population we will observe a different ensemble of behaviors, such as political or consumer preferences, heights, or ages. The heterogeneity of these characteristics across the population can often be quantified with measures, and their relative occurrences modeled with probability distributions.

For example is 30% of the individuals in a population have a height between 0 feet and 5 feet then a probability distribution modeling the variation in heights would give

$$\pi([0, 5]) = 0.3.$$

5.3 Modeling Frequencies

An application particular to probability theory concerns the frequencies of repeated events.

Consider an abstract event whose outcomes take *unpredictable* values in some space X. Perfectly replicating the circumstances of this event N times defines a sequence of values in Y,

$$\{x_1, \ldots, x_n, \ldots x_N\}.$$

While we cannot predict what values the individual events in this sequence will take, we may be able to characterize how often certain outcomes appear relative to others. In particular we can define the **frequency** of a subset $\mathsf{x} \subset X$ by the number of events that that take values in x,

$$f_N(\mathsf{x}) = \frac{\sum_{n=1}^{N} \mathbb{I}_{\mathsf{x}}(x_n)}{N},$$

where

$$\mathbb{I}_{\mathsf{x}}(x) = \begin{cases} 1, & x \in \mathsf{x} \\ 0, & x \notin \mathsf{x} \end{cases}.$$

Replicating the event a countably infinite number of times defines the **asymptotic** or **long-run** frequency of a subset,

$$\begin{aligned} f(\mathsf{x}) &= \lim_{N \to \infty} f_N(\mathsf{x}) \\ &= \lim_{N \to \infty} \frac{\sum_{n=1}^{N} \mathbb{I}_{\mathsf{x}}(x_n)}{N}. \end{aligned}$$

In other words the more frequent subsets contain more common event outcomes.

If the frequencies are the same for *any* sequence of events then we can model them with probability theory. Specifically we can interpret the allocated probabilities as the proportion of the total event outcomes that fall into each subset of outcome values.

In this case the particular ordering of the event sequences doesn't matter and we can also interpret them as defining a population of possible events. From this perspective the application of probability theory is equivalent to the application in the previous section.

5.4 Modeling Uncertainties

Probability theory can also be used to consistently quantify uncertain information.

Consider a space of possible statements X. Under perfect knowledge we would be able to specify a particular statement $x \in X$ as true with all other statements in X being false. In other words certainty is quantified with binary true/false assignments. When our knowledge is not quite so certain, however, we have to soften those claims.

To quantify uncertain information we have to generalize beyond binary true/false assignments to continuous values that *interpolate* between absolute truth and falsity. The larger the value we assign to a subset of statements the more our uncertain information supports one of those statements being true. Conversely the smaller the value we assign to a subset the more our uncertain information supports all of the included statements being false.

Applying probability theory allows us to enforce consistent uncertainty assignments across all of the possible statements. The individual probability allocations can then be interpreted as quantifying how strongly our information supports that one of the statements within a measurable set is true. In this setting the allocated probabilities are sometimes referred to as "plausibilities", "credibilities", and "beliefs".

For example the property that $\pi(X) = 1$ corresponds to the fact that at least one of the statements in X always has to be true. A probability distribution that concentrates around the statement x encodes confidence that one of the statements near x is true. The singular limit where all of the available probability collapses onto a single statement, $\pi(\{x\}) = 1$, communicates certainty that x is true.

This kind of probabilistic uncertainty quantification can be interpreted in many ways. For example we can use it out model the personal, subjective beliefs that an individual holds about the behavior of a system. In particular we can use it to model our own specific beliefs. At the same time we can use it to model the collective understanding of entire communities. We can also use probability theory to model only certain aspects of individual or community knowledge and not attempt to quantify the entirety of that knowledge at once.

More formally this application defines one way to generalize classical propositional logic to a many-valued logic. Using probability theory to generalize other logical systems can sometimes also be possible, although the technical details quickly become more complicated.

5.5 Everyone Play Nicely

A key point of confusion in probability theory is the confounding of its abstract mathematical structure with the interpretations that arise in particular applications. This confusion is made all the worse by the long history of attempts to *derive* probability theory from these particular applications.

For example many have tried to derive probabilities as asymptotic frequencies of physical events. The key motivations of this approach is that the resulting probabilities would be objective in the sense that everyone who could implement those infinite trials would attain the same probabilities. Even if we ignore the impracticality of perfectly repeating an event an infinite number of times within a finite lifetime it turns out that there are also some subtle mathematical complications with this approach. For a comprehensive discussion see Diaconis and Skyrms (2017).

Similarly many have tried to derive probability theory from uncertainty quantification. For example the Cox postulates (Van Horn, Kevin S. (2003)) define basic intuitions about uncertainty quantification. On simpler spaces these rules are equivalent to probability theory, but that equivalence doesn't persist to more general spaces. Because of that this approach is not able to recover the full generality of probability theory.

A common reaction to these technical difficulties is to resort to a sort of philosophical bait and switch. When one cannot derive probability theory from a particular application one might define probability theory abstractly, as we have done above, but then impose an *arbitrary* restriction that it can only ever be applied to that one application. For example those trying to derive probability theory from frequencies might argue that probability theory can be applied to model only frequencies, in which case all probabilities are frequencies. Others trying to derive probability theory from the Cox axioms might argue that any application of probability theory always models uncertain information.

These interpretational restrictions then force some awkward philosophical contortions when trying to apply probability theory in practice. For example after imposing that all probabilities are frequencies the only way to model uncertainty in the value of some quantity is to treat it as the outcome of some hypothetical, and completely non-existent, event. The introduction of these hypothetical events to real events makes the entire system more difficult to understand.

In this book we will avoid these restrictions, respect the full generality of probability theory, and take advantage of any consistent applications that might be useful in a given problem. Indeed we will often take advantage of multiple applications *at the same time*.

Consider, for example, a binary space $X = 0, 1$ that corresponds to the two sides of a coin. In particular let `0` denote tails and `1` denote heads. Any probability distribution over X can be quantified with the probability $p \in [0, 1]$ allocated to the point `1`, which gives the consistent

probability allocations

$$\begin{aligned}
\pi(\emptyset; p) &= 0 \\
\pi(0; p) &= 1 - p \\
\pi(1; p) &= p \\
\pi(X; p) &= 1.
\end{aligned}$$

There are many ways to flip a coin, but let's say that we flip our coin in a way that results in an unpredictable sequence of heads and tails. The asymptotic frequencies of these outcomes can then be modeled with an application of probability theory. In other words we can use a probability distribution $\pi(\cdot; p)$ to model the physical outcomes of the flips.

At the same time we use probability theory to model any uncertainty in which of the possible frequency models best matches the true behavior of the coin. In particular we can construct a probability distribution over the unit interval to quantify how compatible each probability allocation $p \in [0, 1]$ is with our knowledge of the coin.

If we have a bag of I coins then could also model the variation in the probability parameters

$$\{p_1, \ldots, p_i, \ldots, p_I\}$$

for each coin. In this case we can apply probability theory once again, this time to model the population of coin behaviors.

To be clear the interpretations inherent to particular applications of probability theory are important for ensuring that we implement those applications correctly in practice. Elevating one interpretation to the exclusion of others, however, excludes the corresponding applications and limits the full potential of probability theory. To take full advantage of the practical utility of probability theory we have to respect all of consistent applications!

6 Conclusion

Conceptually measure and probability theory are straightforward. Measure theory quantifies how we can consistently allocate a conserved quantify across a general mathematical space and probability theory considers the special case of proportional allocations. In order to quantify that conceptual simplicity, however, we need to resort to some careful mathematics. In particular we need to incorporate σ-algebras to surgically remove any pathological behavior that can arise, even on seemingly well-behaved spaces such as the real line, and obstruct consistent allocations.

Once we've safely constructed these theories in full generality we can use the apply abstract mathematics to model particular systems. Within these applications the math inherits particular interpretations, but we have to be careful to not take these circumstantial interpretations too seriously lest we abandon the full utility of the abstract mathematics.

The technical exploration of measures and probability distributions goes far beyond the introduction in this chapter. Unfortunately many textbooks that cover this material can be difficult to parse without extensive mathematical experience. My personal favorite is Folland (1999) which, while technically rigorous, provides more exposition and motivation than I have found in other treatments.

7 Acknowledgements

I thank Jeff Helzner for helpful discussion.

A very special thanks to everyone supporting me on Patreon: Adam Fleischhacker, Adriano Yoshino, Alan Chang, Alessandro Varacca, Alexander Bartik, Alexander Noll, Alexander Petrov, Alexander Rosteck, Anders Valind, Andrea Serafino, Andrew Mascioli, Andrew Rouillard, Andrew Vigotsky, Angie_Hyunji Moon, Ara Winter, Austin Rochford, Austin Rochford, Avraham Adler, Ben Matthews, Ben Swallow, Benjamin Glemain, Bradley Kolb, Brandon Liu, Brynjolfur Gauti Jónsson, Cameron Smith, Canaan Breiss, Cat Shark, Charles Naylor, Chase Dwelle, Chris Jones, Chris Zawora, Christopher Mehrvarzi, Colin Carroll, Colin McAuliffe, Damien Mannion, Damon Bayer, dan mackinlay, Dan Muck, Dan W Joyce, Dan Waxman, Dan Weitzenfeld, Daniel Edward Marthaler, Darshan Pandit, Darthmaluus , David Burdelski, David Galley, David Wurtz, Doug Rivers, Dr. Jobo, Dr. Omri Har Shemesh, Ed Cashin, Edgar Merkle, Eric LaMotte, Erik Banek, Ero Carrera, Eugene O'Friel, Felipe González, Fergus Chadwick, Finn Lindgren, Florian Wellmann, Francesco Corona, Geoff Rollins, Granville Matheson, Greg Sutcliffe, Guido Biele, Hamed Bastan-Hagh, Haonan Zhu, Hector Munoz, Henri Wallen, hs, Hugo Botha, Håkan Johansson, Ian Costley, Ian Koller, idontgetoutmuch, Ignacio Vera, Ilaria Prosdocimi, Isaac Vock, J, J Michael Burgess, Jair Andrade, James Hodgson, James McInerney, James Wade, Janek Berger, Jason Martin, Jason Pekos, Jason Wong, Jeff Burnett, Jeff Dotson, Jeff Helzner, Jeffrey Erlich, Jesse Wolfhagen, Jessica Graves, Joe Wagner, John Flournoy, Jonathan H. Morgan, Jonathon Vallejo, Joran Jongerling, Joseph Despres, Josh Weinstock, Joshua Duncan, Joshua Griffith, JU, Justin Bois, Karim Naguib, Karim Osman, Kejia Shi, Kevin Foley, Kristian Gårdhus Wichmann, Kádár András, Lars Barquist, lizzie , LOU ODETTE, Marc Dotson, Marcel Lüthi, Marek Kwiatkowski, Mark Donoghoe, Markus P., Martin Modrák, Matt Moores, Matthew, Matthew Kay, Matthieu LEROY, Maurits van der Meer, Merlin Noel Heidemanns, Michael DeWitt, Michael Dillon, Michael Lerner, Mick Cooney, Márton Vaitkus, N Sanders, Name, Nathaniel Burbank, Nic Fishman, Nicholas Clark, Nicholas Cowie, Nick S, Nicolas Frisby, Octavio Medina, Ole Rogeberg, Oliver Crook, Olivier Ma, Pablo León Villagrá, Patrick Kelley, Patrick Boehnke, Pau Pereira Batlle, Peter Smits, Pieter van den Berg , ptr, Putra Manggala, Ramiro Barrantes Reynolds, Ravin Kumar, Raúl Peralta Lozada, Riccardo Fusaroli, Richard Nerland, Robert Frost, Robert Goldman, Robert kohn, Robert Mitchell V, Robin Taylor, Ross McCullough, Ryan Grossman, Rémi , S Hong, Scott Block, Sean Pinkney, Sean Wilson, Seth Axen, shira, Simon Duane, Simon Lilburn, sssz, Stan_user, Stefan, Stephanie Fitzgerald, Stephen Lienhard, Steve Bertolani, Stew Watts, Stone Chen, Susan Holmes, Svilup, Sören Berg, Tao Ye, Tate Tunstall, Tatsuo Okubo, Teresa

Ortiz, Thomas Lees, Thomas Vladeck, Tiago Cabaço, Tim Radtke, Tobychev , Tom McEwen, Tony Wuersch, Utku Turk, Virginia Fisher, Vitaly Druker, Vladimir Markov, Wil Yegelwel, Will Farr, Will Tudor-Evans, woejozney, yolhaj , yureq , Zach A, Zad Rafi, and Zhengchen Cai.

References

Diaconis, Persi, and Brian Skyrms. 2017. *Ten Great Ideas about Chance*. Princeton University Press.

Folland, G. B. 1999. *Real Analysis: Modern Techniques and Their Applications*. New York: John Wiley; Sons, Inc.

Van Horn, Kevin S. 2003. "Constructing a logic of plausible inference: a guide to Cox's theorem." *International Journal of Approximate Reasoning* 34 (1): 3–24.

License

A repository containing all of the files used to generate this chapter is available on GitHub.