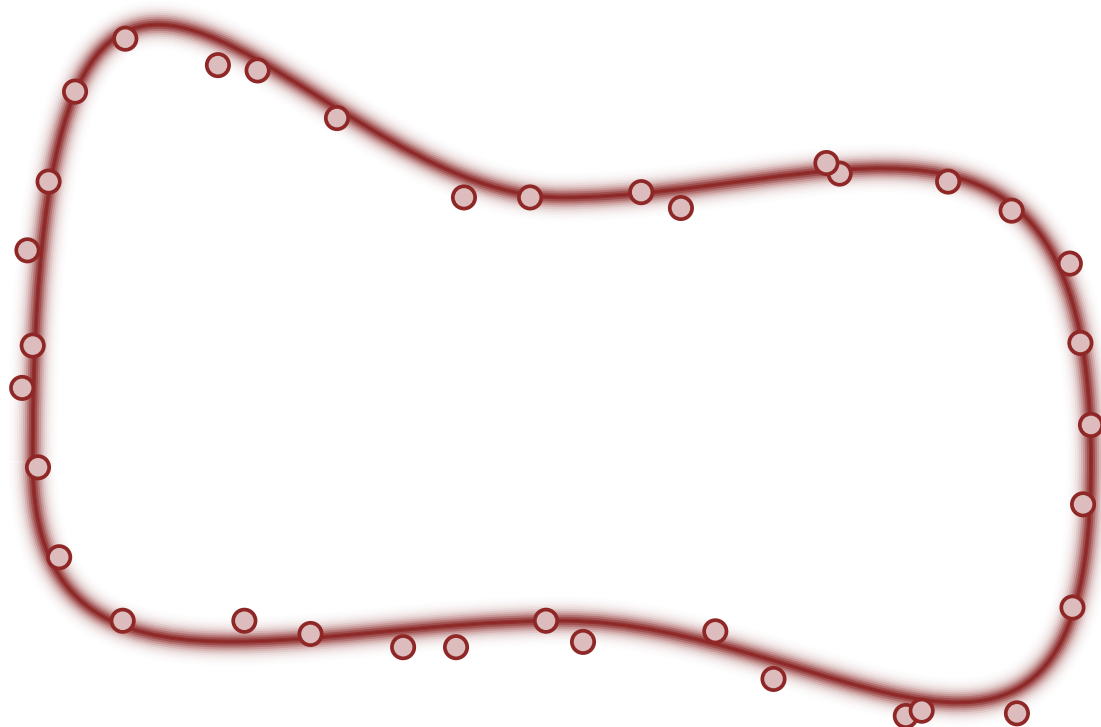# Logic, Probability, and Bayesian Inference:
A Conceptual Introduction to the Foundations
of Applied Inference for Scientists and Engineers

The Stan Development Team

Version 0.0.0

May 11, 2016

https://github.com/betanalpha/stan_intro

Contents

Preface

Naturally curious as a species, we all have a strong intuition for how we can better understand the world around us. We first make observations and then try to gleam insight from those observations. Formalizing this intuition into a robust methodology for inference, however, is a delicate process.

Consider, for example, a ball flying through the air, slowly falling under the force of gravity. After diligently recording the positions of the ball over its trajectory we want to identify a physical model that quantitatively describes its motion. Perfect position measurements would strongly constraint the possible models, allowing us to exactly recover not only the trajectory taken by the ball but also any latent model parameters, such as the acceleration due to gravity, g (Figure 1a).

Of course we cannot make perfect measurements in practice. The observations we can make are limited by an inherent variability that obscures the underlying trajectory; multiple trajectories, and hence multiple physical models, are all consistent with the noisy measurements (Figure 1b). In order to learn from these realistic measurements we need not only models of the ball's motion and models of the variability in the measurements, but also a means of quantifying the uncertainty in which of those models are consistent with the observations. *The only conclusive statement we can make is how uncertain we are.*

This simple example demonstrates a intrinsic principle that underlies science, industry, medicine, and any other field that attempts to learn from observations: uncertainty is inherent to learning and decision making. In particular, if we want to develop any formal methodology for inference and decision making then we first need a formal procedure for quantifying and manipulating uncertainty itself. *Bayesian inference* uses *probability theory* to quantify all forms of uncertainty, including not only the intrinsic variability of measurements but also ignorance in the learning process itself. This unified perspective provides an elegant and powerful approach for first making inferences and then making robust decisions.

0.1   Our Objectives

While reading this review we want the reader to learn how probability theory can be used to quantify uncertainty both in theory and in practice, ultimately motivating Bayesian infer-
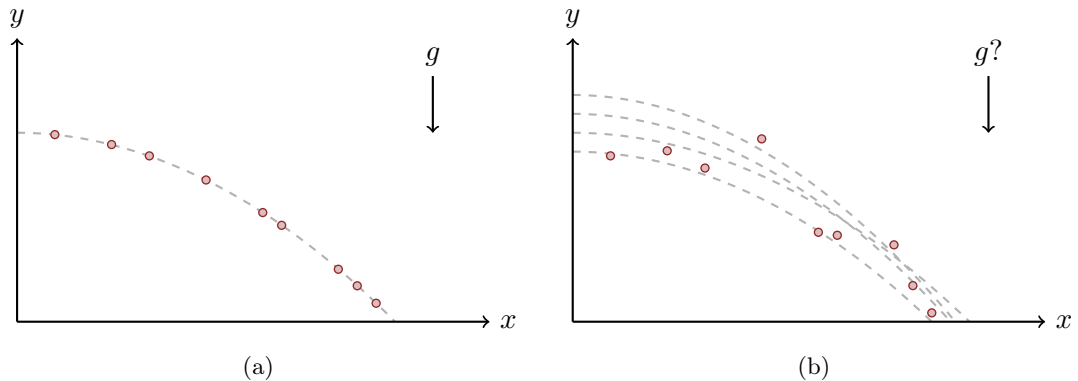
Figure 1: (a) Perfect measurements of a ball falling under the influence of gravity would strongly constrain any physical model of that motion, including latent parameters such as the the strength of gravity, g. (b) In practice, however, measurements are inherently variable, which limits our ability to infer the exact trajectory and hence any model of how the trajectory itself. Here and in all measurements, uncertainty is intrinsic to learning.

ence itself. Unfortunately, this seemingly straightforward objective ends up being quickly complicated by the subtle and often counterintuitive nature of probability theory itself.

Many introductory treatments of Bayesian inference ignore these subtleties entirely, oversimplifying the subject and neglecting many of its finer technical aspects. Unfortunately, these technicalities are not irrelevant – indeed they often have a strong influence on practical applications of the theory. Without at least a conceptual understanding of these technicalities, the readier is then subject to dangerous fallacies and, ultimately, fragile analyses.

On the other hand, the more formal treatments of probability theory that do address these subtleties are typically mired in abstract, mathematical pedantry with little or not discussion of how the theory applies to inference. A few treatments will introduce the application of these ideas to frequentist inference, but rarely is there any discussion of the more general Bayesian perspective.

In this review we attempt an intermediate treatment where we provide a deeper introduction to probability theory and Bayesian inference than most introductory references, but focus entirely on concepts rather than proofs. We hope that this will give Stan users the foundational understanding they need to properly wield Bayesian inference in practice and take full advantage of their measurements.

0.2 Our Strategy

Probability theory, like much of higher mathematics, is challenging because of its abstraction; the theory is defined in complete generality with only abstract objects and manipulations. While much can be proved about the behavior of these abstractions, there is no general way to explicitly specify these objects and implement their manipulations. Without explicit examples it then becomes difficult to develop accurate intuition and understanding.

To apply probability theory in practice we need to first define a *representation* that maps these abstractions into an explicit context where we can compute. For example, we can map into a discrete space where manipulations reduce to counting or a real space where manipulations reduce to integration. The challenge in using these representations is to ensure that the calculations depend only on the properties of the original abstract system and not on any irrelevant details of the representation itself. Understanding exactly what details are relevant or not, however, requires at least some comprehension of the abstract theory.

Many introductions to probability and statistics simply jump into one of these representations immediately, ignoring the abstractions, and the subtitles as to which operations are valid, altogether. This approach inevitably leads to confused frustration. *What do you mean a probability* density *isn't a probability? What do you mean most observations won't be near the mode of the probability density?.* Unfortunately, this confusion manifests not just in frustration but also incorrect statistical analyses that can have serious consequences.

In this introduction we will not ignore the abstractions of probability theory. As with more formal treatments we will instead being by discussion the abstract theory and only then discuss representations and the explicit computations that they allow. Unlike those treatments, however, our presentation will be largely conceptual, meanings lots of definitions of abstract objects and their manipulations but little concern for any technical details that do not end up affecting the application of the theory in practice. As we develop representations we'll discuss more and more specific examples to develop as much intuitive context as can while also emphasizing the limitations of these representations.

We'll first introduce logic as a means of quantifying information with certainty, and then probability theory as a means of quantifying uncertainty about that information. In both cases we'll first consider abstract definitions and then the explicit representations of these concepts needed in practice. Next we'll discuss how to implement probabilistic computations and discuss many popular computational methods. Finally we'll show how all of these ideas come together in Bayesian inference.

This will not be an easy journey. It requires a significant investment from the reader, but that investment will be rewarded with a robust understanding of statistics that will enable some amazing science.

0.3 Mathematical Background and Notation

A thorough review of probability theory and its application requires a nontrivial mathematical background. We have attempted to make this review as self-contained as possible regarding probability theory itself, but we do have to assume that the reader is comfortable with the basics of set theory and differential and integral calculus over the real numbers. We highly encourage anyone whose math might be rusty to brush up before proceeding.

Throughout we will use common set theory notation. If A is a set then any element of the set is written as $a \in A$ while a subset is written as $S \subset A$. Sets are also sometimes denoted by their elements, for example $A = \{a_1, \ldots, a_N\}$. The *set builder notation* is similarly used to denote subsets as $S = \{a \in A \mid \cdot\}$, where $\cdot$ is the condition identifying which elements of A are in the subset $S \subset A$. For example, we can define the positive real numbers as

$$\mathbb{R}^+ = \{x \in \mathbb{R} \mid x \geq 0\}.$$

The *union* of two sets, $A \cup B$ is the combination of all elements in either set while the *intersection* of two sets, $A \cap B$ contains only those elements that appear in both sets. If $S \subset A$ is a subset then its *complement*, S^c is the collection of all elements of A not in S.

Spaces are sets endowed with a structure called a *topology* that allows us to separate "well-behaved subsets" from "pathological subsets". We will assume that all of our sets have such a structure and consequently for all intents and purposes set and space will be used interchangeably. Topologies are also useful for characterizing spaces. For example, the familiar properties of discrete spaces and continuous spaces, such as the real numbers, are defined are defined by their respective topologies.

Throughout we will use the common notation for maps from one set into another, $f : A \to B$ which defines f as a map taking elements of the set A to elements of the set B. In other words, $f(a) \in B$ for any $a \in A$. Sometimes we will be more explicit regarding the action on a given point and write

$$f : A \to B$$
$$a \mapsto f(a).$$

At times we will be less precise. For example, when discussing computation we will liberally use $\approx$ to define when two objects are approximately equal, or when we *assume* that they are approximately equal, without making any effort to formally define what "approximately equal" means. Similarly, we will make no attempt at the full mathematical rigor necessary for a complete understanding of the intricacies of probability theory, and instead focus on developing a high-level, conceptual intuition. In particular, in many places we will appeal to vague notions like "well-behaved", as their technical definitions do not offer much pedagogical benefit.

1 | Logic

Before we can formalize *learning* about the world around us we first need a formal means of *describing* the world. *Logical propositions* provide a mathematically-precise language for quantifying information about a system with certainty. In this section we will review abstract logical propositions, their manipulations, and how we can represent them in practice. We conclude with a brief discussion of implicative statements.

1.1 Logical Propositions

If $\mathcal{S}$ is an abstract system of interest then any description of $\mathcal{S}$ can be defined as a *logical proposition* which asserts that a statement about that system is either true or false. For example, consider describing one of the eight planets in our Solar System: Mercury ($\mathpzc{☿}$), Venus ($\mathpzc{♀}$), Earth ($\mathpzc{♁}$), Mars ($\mathpzc{♂}$), Jupiter ($\mathpzc{♃}$), Saturn ($\mathpzc{♄}$), Uranus ($\mathpzc{♅}$), and Neptune ($\mathpzc{♆}$). Information about the planet is encapsulated in logical propositions asserting both statements that are true, for example

$$P_1 = \text{``the planet has an atmosphere''},$$

and statements that are false, for example

$$P_2 = \text{``the planet does not have rings''}.$$

Logical propositions can be manipulated into other logical propositions with any of three *boolean operations* (Table 1.1). *Negation*, $\neg$, is a unary operator resulting in a proposition asserting the converse of the input proposition:

$$\neg P_1 = \text{``the planet does \textit{not} have an atmosphere''}.$$

Binary operations take propositions about two statements and define a new proposition about the joint statement. The result of a *conjunction*, $\wedge$, is true only when both input propositions are true:

$$P_1 \wedge P_2 = \text{``the planet does not has both an atmosphere and rings''}.$$

1

Negation

P	$\neg P$
False	True
True	False

Conjunction			Disjunction		
P_1	P_2	$P_1 \wedge P_2$	P_1	P_2	$P_1 \vee P_2$
False	False	False	False	False	False
True	False	False	True	False	True
False	True	False	False	True	True
True	True	True	True	True	True

Table 1.1: The three basic Boolean operations transform given logical propositions into new logical propositions, and their action is conveniently summarized in truth tables.

A *disjunction*, $\vee$, however, is true whenever either input proposition is true :

$$P_1 \vee P_2 = \text{"the planet does have an atmosphere } or \text{ rings"}.$$

Naively, we might think that a system would be completely described only once we have defined a proposition for every statement, classifying each as either true or false. In his infamous Incompleteness Theorems, however, Gödel showed that if we try to define a proposition for every possible statement about $\mathcal{S}$ then there will exist some pathological propositions that are not consistent with the Boolean operations. For example, there are always pathological propositions asserting truth whose negations do not assert falseness! These pathological propositions cannot be explicitly constructed in practice, only proven to exist, but in order to ensure a self-consistent description of the system we need some means of removing them.

To avoid these pathological propositions we leverage the fact that propositions about non-pathological statements yield non-pathological propositions when we apply any Boolean operation. In other words, well-behaved propositions are *closed* under these operations, and any collection of well-behaved, consistent propositions forms a logical algebra, $\mathcal{L}(\mathcal{S})$.

We will refer to the pairing of an abstract system with some choice of a logical algebra as a *logical model*, $(\mathcal{S}, \mathcal{L}(\mathcal{S}))$.

Abstract System	Representation	Measurable Maps
Historical Decrees	Ancient Languages	Translations
	(Egyptian, Demotic, Green, ...)	(Rosetta Stone)
Computer Program	Programming Languages	Source-to-Source
	(C++, S, Python, ...)	Compilers
Unordered Categories	Integers	Permutations
Direction	Vector Space	Changes of Basis
Position	Coordinates	Reparameterizations
Quantification	Units	Changes of Units

Table 1.2: Representations provide a concrete language in which we can communicate the descriptions of an abstract system. Because there are many degenerate ways to express these descriptions, these representations are not unique and in practice we have to be careful that the irrelevant details of any particular representation do not affect the descriptions themselves.

1.2 Representations of Logical Models

Logical models provide a very abstract, and hence very genetic, framework for describing arbitrary systems, but that abstraction also makes these descriptions ungainly to specify and manipulate in practice. For our abstract descriptions to be useful we need an explicit language in which we can communicate them, which we will refer to as a *representation* (Table 1.2).

In order to construct a representation we assume that the properties of our abstract system take values in some *sample space*, Θ, such as integers or real numbers. Statements about these properties are represented as subsets of the sample space, and logical propositions are represented as set inclusion. If a statement is represented by $E \subset \Theta$, for example, then a logical proposition asserting that the statement is true is represented by asserting that $\theta \in E$, and one asserting that the statement is false is represented by asserting that $\theta \notin E$.

With propositions represented by set inclusion and set exclusion, the Boolean operations are naturally implemented with set operations. Conjunction becomes inclusion in the subset intersection, disjunction inclusion in the subset union, and negation inclusion in the subset complement negation (Figure 1.1, Table 1.3).

The same inconsistencies that limited us to consider only well-behaved logical propositions, however, also limit our representations to only well-behaved subsets of the sample
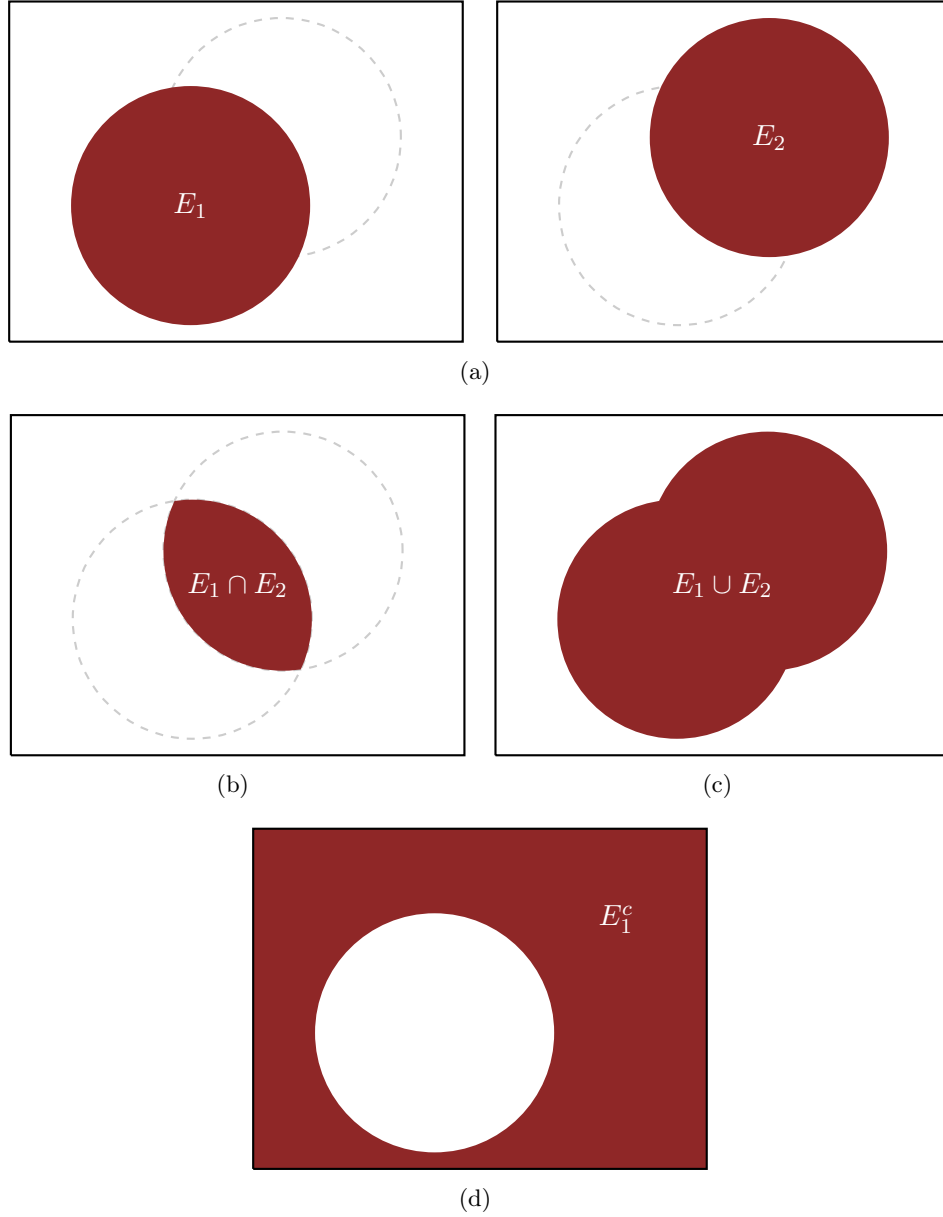
Figure 1.1: In a representation statements are defined as subsets in some sample space and logical propositions are implemented with set inclusion or set exclusion. Boolean operations are then implemented with inclusion in a set complement (negation), set intersection (conjunction), and set union (disjunction). (a) If two propositions are represented by inclusion in the sets E_1 and E_2, then (b) their conjunction is represented as inclusion in the set intersection, $E_1 \cup E_2$, and (c) their disjunction is represented as inclusion in the set union, $E_1 \cap E_2$. (d) Similarly, the negation of E_1 is represented as inclusion in the set complement, E_1^c.

Negation

$\theta \in E$	$\theta \in E^c$
False	True
True	False

Conjunction

$\theta \in E_1$	$\theta \in E_1$	$\theta \in E_1 \cap E_2$
False	False	False
True	False	False
False	True	False
True	True	True

Disjunction

$\theta \in E_1$	$\theta \in E_1$	$\theta \in E_1 \cup E_2$
False	False	False
True	False	True
False	True	True
True	True	True

Table 1.3: In a representation statements are defined as subsets in some sample space and logical propositions are implemented with set inclusion or set exclusion. Boolean operations are then implemented with inclusion in a set complement (negation), set intersection (conjunction), and set union (disjunction), which yield the same truth tables and hence define a valid representation.

space. We denote a collection of subsets of the sample space that is closed under set compliments, set intersections, and set unions as an *event space*, $\mathcal{E}(\Theta)$, with the well-behaved subsets in that space denoted *events*. Event spaces always include the null event, $E = \emptyset$, which can never be true as it is always empty. Similarly, event spaces also always include the trivial event, $E = \Theta$, which can never be false as every element of the sample space is included in this event. The pairing of a sample space and an event space is denoted a *measurable space*, for reasons that will become more clear in the next chapter.

More formally, then, we should define a representation as a map from a logical model into a measurable space. In practice this is defined as a map from the abstract system into the sample space,

$$\mathfrak{r} : \mathcal{S} \to \Theta,$$

that then induces a map from the logical algebra into the event space,

$$\mathfrak{r}_* : \mathcal{L}(\mathcal{S}) \to \mathcal{E}(\Theta) .$$

A representation is called *faithful* if every logical statement maps into an event and vice-versa, and Stone's Theorem guarantees that a faithful representation for any logical model always exists.

For example, descriptions of the planets in our solar system can be represented with the integers, $\Theta = \{1, 2, \ldots, 8\}$, with each logical proposition represented by inclusion in a subset of those integers (Figure 1.2). Similarly, the distance between two objects can be represented by real numbers, (Figure 1.3).

1.3 Equivalent Representations

Because they map abstract systems into explicit contexts, representations are critical to the practical implementation of logical models and their manipulations. Unfortunately representations also bring with them a dangerous subtlety – there is no *unique* representation of a given abstract system.

A map from one sample space into another, $s : \Theta \to \Omega$, is *measurable* if every event in Ω is generated by an event in Θ,

$$s^{-1}(E_\Omega) \in \mathcal{E}(\Theta) .$$

Measurable maps allow us to completely generate the measurable space $(\Omega, \mathcal{E}(\Omega))$ by mapping forward elements of the initial measurable space $(\Theta, \mathcal{E}(\Theta))$; we cannot, however, necessarily achieve the converse as there is no guarantee that every event in $\mathcal{E}(\Theta)$ maps to an event in $\mathcal{E}(\Omega)$. This process is also known as *pushing forward* the measurable space $(\Theta, \mathcal{E}(\Theta))$ into the measurable space $(\Omega, \mathcal{E}(\Omega))$, and in this context $(\Omega, \mathcal{E}(\Omega))$ is denoted the *pushforward* of $(\Theta, \mathcal{E}(\Theta))$ by s.

The planet has a magnetic field

$$\theta \in E_1$$

The planet has moons

$$\theta \in E_2$$

(a)

The planet has a magnetic field and moons

$$\theta \in E_1 \cap E_2$$

(b)

The planet has a magnetic field or moons

$$\theta \in E_1 \cup E_2$$

(c)

The planet does not have a magnetic field

$$\theta \in E_1^c$$

(d)
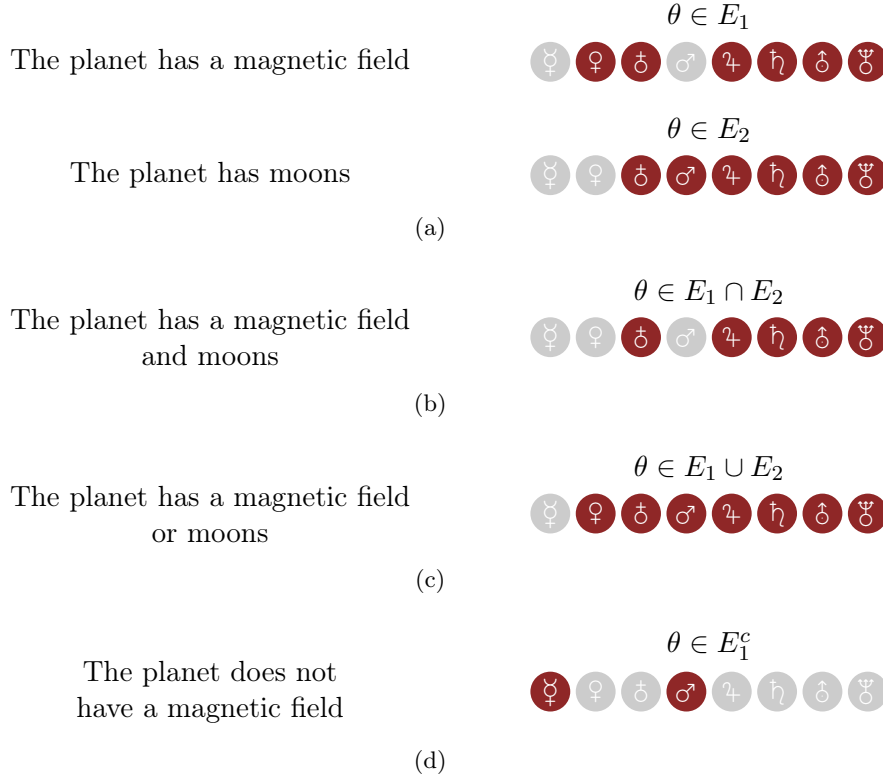
Figure 1.2: (a) Systems that take discrete values, such as the planets in our solar system, can be represented with integers, with any logical proposition about the planets represented by inclusion in a subset of integers. (b) Conjuction of logical propositions is implemented with inclusion in set intersections, (c) disjunction with inclusion in set unions, and (d) negation with inclusion in set complements.

The distance is
less than five inches

$$\theta \in E_1$$

$$\xrightarrow{\quad\quad\quad} x$$

The distance is between
three and seven inches

$$\theta \in E_2$$

$$\xrightarrow{\quad\quad\quad} x$$

(a)

The distance is less than five inches
and between three and seven inches

$$\theta \in E_1 \cap E_2$$

$$\xrightarrow{\quad\quad\quad} x$$

(b)

The distance is less than five inches
or between three and seven inches

$$\theta \in E_1 \cup E_2$$

$$\xrightarrow{\quad\quad\quad} x$$

(c)

The distance is not
less than five inches

$$\theta \in E_1^c$$

$$\xrightarrow{\quad\quad\quad} x$$

(d)
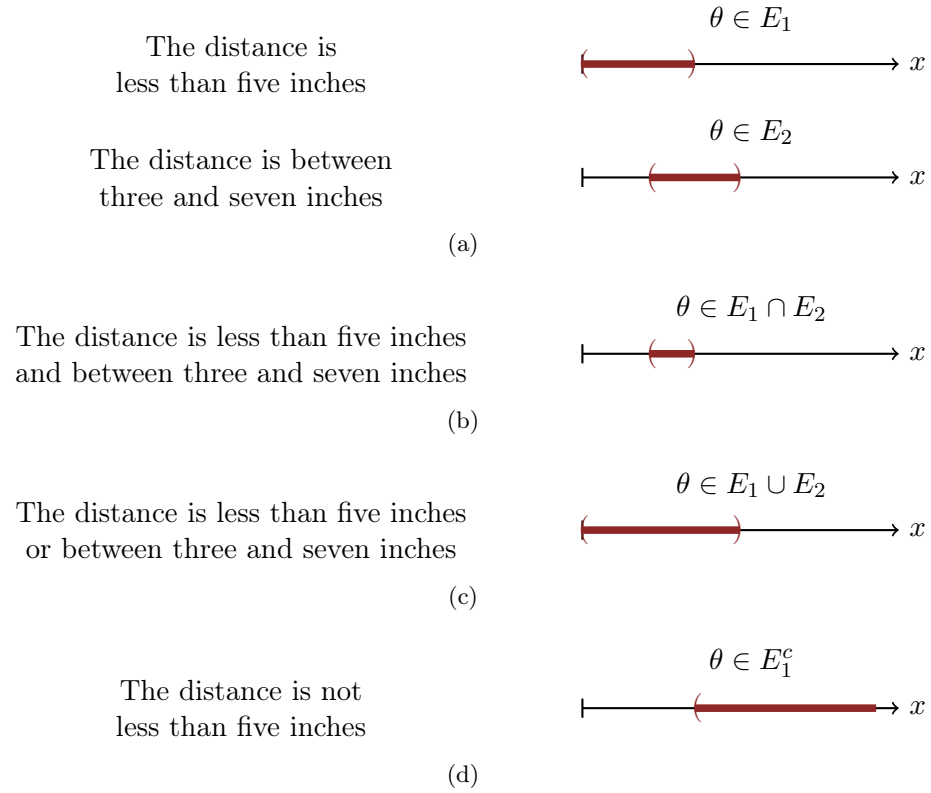
Figure 1.3: (a) Systems that take continuous values, such as distances, can be represented with real numbers, with any logical proposition represented by inclusion in well-behaved subsets of real numbers. (b) Conjuction of logical propositions is implemented with inclusion in set intersections, (c) disjunction with inclusion in set unions, and (d) negation with inclusion in set complements.

Composing a faithful representation with a measurable map,

$$s \circ \mathfrak{e} : \mathcal{S} \to \Theta \to \Omega,$$

yields another valid representation but not necessarily a faithful one. Because there is no guarantee that every event in $\mathcal{E}(\Theta)$ maps to an event in $\mathcal{E}(\Omega)$, this pushforward may be able to represent only a handful of propositions in the logical model. For example, a measurable map might round real numbers down to discrete numbers, losing the ability to represent all but a countable number of logical propositions.

In order to maintain a faithful representation we need to ensure that the events spaces are preserved by both the map and its inverse. Here we will say that a map $s : \Theta \to \Omega$ is *doubly-measurable* if every event in $\mathcal{E}(\Theta)$ maps to an event in $\mathcal{E}(\Omega)$ and vice versa,

$$s(E_\Theta) \in \mathcal{E}(\Omega)$$
$$s^{-1}(E_\Omega) \in \mathcal{E}(\Theta) \, .$$

A doubly-measurable map allows us to recover either measurable space as a pushforward of the other, and the composition of a faithful representation with a doubly-measurable map,

$$s \circ \mathfrak{e} : \mathcal{S} \to \Theta \to \Omega,$$

is then another faithful representation. We warn the reader that there is no conventional terminology for these maps and we have introduced doubly-measurable here as our own terminology.

When a doubly-measurable map exists between two measurable spaces then the two spaces can be used to specify the same logical model, and we say that they are *equivalent representations* as they provide equivalent descriptions of a given system (Table 1.2). The set of measurable spaces equivalent to $(\Theta, \mathcal{E}(\Theta))$ will be denoted $\lfloor \Theta, \mathcal{E}(\Theta) \rfloor$.

For example, if the sample space is discrete then we can always define an equivalent measurable space by simply permuting the labels. It doesn't matter if we order the planets by distance from the sun, by diameter, or by any other metric: the measurable spaces quantify the same information (Figure 1.4). Similarly, we can always apply a transformation that warps real numbers to map between real sample spaces without compromising our ability to represent propositions with inclusion in events. One of the most common ways this manifests in practice is when our descriptions require units. The information we quantify doesn't depend on whether we express distance in ångströms or inches or meters or furlongs: each unit defines a separate but equivalent sample space.

Ultimately, there are many ways to describe of a given logical model and measurable maps translate from one equivalent representation to another. Although certain representations may be more useful than others in practice, we have to be careful to ensure that our final analysis does not depend on the irrelevant details of any particular representation.
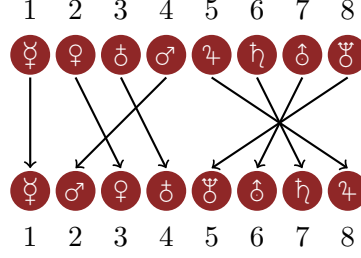
Figure 1.4: Discrete sample spaces are not unique way to express a system as we can always permute the arbitrary labels to yield an equivalent sample space. For example, we could order the planets by distance from the sun or diameter without affecting our ability to describe the planets themselves.

1.4 Implications

There are various Boolean operations that we can derive from the basic operations of negation, conjunction, and disjunction, and one of the most powerful are *implications*, which assert the truth of a statement conditioned on the truth of another statement. For example, we might not be able to assert that "a planet has a magnetic field" universally, but we might be able to assert the implication "if a planet has a rotating core then it has a magnetic field".

Implications are particularly naturally when we want to describe a target system conditioned on the state of some other system. Letting $(\Theta, \mathcal{E}(\Theta))$ be a measurable space capable of representing the target system and $(\Phi, \mathcal{E}(\Phi))$ be a measurable space capable of representing the conditioning system, an implication is represented as a map from points in Φ to events in $\mathcal{E}(\Theta)$,

$$\iota_{\Theta|\Phi} : \Phi \to \mathcal{E}(\Theta)$$
$$\phi \mapsto \iota_{\Theta|\Phi}(\phi) \, .$$

For any event in the conditioning space, $E_\Phi \in \mathcal{E}(\Phi)$, the implication also defines an event in the joint sample space, $\Theta \times \Phi$, as the union of all of the events implied by each point in E_Φ,

$$\iota^*_{\Theta|\Phi} : \mathcal{E}(\Phi) \to \mathcal{E}(\Theta \times \Phi)$$
$$E_\Phi \mapsto \bigcup_{\phi \in E_\Phi} \iota_{\Theta|\Phi}(\phi) \, .$$

Because relationships are often much easier to reason about than the entirety of a complex system, these induced events can drastically facilitate the description of complex systems by decomposing logical propositions into simpler implications. For example, instead

10

of having to enumerate all possible combinations of planetary core behavior and magnetic field, we can utilize the single implication that only a rotating core generates a magnetic field to build any logical proposition about the joint core-magnetic field system from a logical proposition about the core alone.

More formally, any logical proposition representation in a joint sample space $\prod_{n=1}^{N} \Theta_n$ can be built up sequentially, starting with an event in one component and adding implications one after another,

$$E_{\Theta_1}$$
$$\iota^*_{\Theta_2|\Theta_1}(E_{\Theta_1})$$
$$\iota^*_{\Theta_3|\Theta_2,\Theta_1} \circ \iota^*_{\Theta_2|\Theta_1}(E_{\Theta_1})$$
$$\cdots$$
$$\iota^*_{\Theta_N|\Theta_{N-1},\ldots,\Theta_2,\Theta_1} \circ \iota^*_{\Theta_3|\Theta_2,\Theta_1} \circ \iota^*_{\Theta_2|\Theta_1}(E_{\Theta_1}).$$

2 | Probability in Theory

While logic provides a framework for how we can describe systems with certainty, probability theory allow us to assign *uncertainty* to logical propositions. We begin this chapter with an introduction of measures, which allow us to assign arbitrary values to propositions, and then probability measures, which limit those value to probabilities that quantify uncertainty. We end with a discussion of conditional probability distributions over implications.

Unfortunately, the abstraction of probability theory means that we cannot rely on explicit examples, and consequently visualizations, to support the definitions and manipulations introduced in this chapter. That context will have to wait until the next chapter where we discuss how to implement probability theory in practice.

2.1 Measures

The first step towards assigning uncertainty to logical statements is to assign *any* positive value. A *measure* accomplishes this by assigning a positive value to all events in the chosen event space,

$$\mu : \mathcal{E}(\Theta) \to \mathbb{R}^+.$$

The term measure often serves multiple uses depending on context. It can, for example, refer to the object which assigns values to each event or the actual value assigned to a particular event, as in "the measure of the event E is $\mu(E)$".

Because a measure assigns values to only events, we need both a sample space and an event space in order to define measure. In other words, we can define measures only on measurable spaces. This is the ultimate motivation for this term, as measurable spaces are most often introduced not as representations of logical models but only as a precursor to measures themselves.

Provided that the measure of each event is finite,

$$\mu[E] < \infty, \ \forall E \in \mathcal{E}(\Theta) \,,$$

the assignment of measures is naturally compatible with Boolean operations. For example, the measure of conjunctions and disjunction are related as

$$\mu(E_1 \cup E_2) = \mu(E_1) + \mu(E_2) - \mu(E_1 \cap E_2) \,,$$

and negations satisfy,

$$\mu(E) = \mu(\Theta) - \mu(E^c) = 1 - \mu(E^c).$$

By definition all events have finite measure, and all of these results hold, provided that the trivial event has finite measure, $\mu(\Theta) < \infty$.

Measures can also be used to integrate functions over the sample space in a procedure known as *Lebesgue integration*,

$$\mathcal{I}_\mu : \mathcal{F}(\Theta) \to \mathbb{R},$$

where $\mathcal{F}(\Theta)$ is the collection of well-behaved functions $f : \Theta \to \mathbb{R}$. In fact, we can also consider measure assignments themselves as Lebesgue integrals,

$$\mu(E) = \mathcal{I}_\mu[\mathbb{I}_E],$$

where the *indicator function* of the event E, $\mathbb{I}_E$, is defined as

$$\mathbb{I}_E(\theta) = \left\{ \begin{array}{ll} 0, & \theta \notin E \\ 1, & \theta \in E \end{array} \right. .$$

In particular, the Lebesgue integral of every well-behaved function completely specifies a given measure.

In some cases Lebesgue integration can also be used to emulate one measure with other. Consider two measures μ_1 and μ_2 where $\mu_2(E) = 0$ whenever $\mu_1 = 0$. We say that any measure satisfying this property is *absolutely continuous* with respect to μ_1.

When μ_2 is absolutely continuous with respect to μ_1, the Lebesgue integral of any well-behaved function, f, can be recovered as a modified Lebesgue integral with respect to μ_1,

$$\mathcal{I}_{\mu_2}[f] = \mathcal{I}_{\mu_1}[\nu \cdot f], \tag{2.1}$$

where $\nu \in \mathcal{F}(\Theta)$ is a bounded function denoted the *Radon-Nikodym derivative* of μ_2 with respect to μ_1.

The Radon-Nikodym derivative is often written as

$$\nu = \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1}$$

to make the relevant measures more explicit. Moreover, it typical to abuse notation a bit and write

$$\mu_2 = \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1}\mu_1$$

as shorthand for the defining relationship (2.1).

14

2.2 Pushforward Measures

Like events, measures can be defined with respect to many different sample spaces. If $s : \Theta \to \Omega$ is a measurable map and μ_Θ is a probability distribution defined over $(\Theta, \mathcal{E}(\Theta))$, then we can define an equivalent *pushforward* measure over $(\Omega, \mathcal{E}(\Omega))$ by assigning values as

$$s_* \mu_\Theta(E_\Omega) \equiv \mu_\Theta\big(s^{-1}(E_\Omega)\big).$$

As with measurable spaces, the measurable in measurable maps refers to this particular application in pushing measures forward from one space to another.

Furthermore, if the map is doubly-measurable then this whole process can be inverted: a measure μ_Ω defined over $(\Omega, \mathcal{E}(\Omega))$ can be pushed forward onto $(\Theta, \mathcal{E}(\Theta))$ by

$$s_*^{-1} \mu_\Omega(E_\Theta) \equiv \mu_\Omega(s(E_\Theta)).$$

In other words, the same measure can be defined with respect to any of the equivalent measurable spaces and does not depend on the particular details of any one of them.

2.3 Probability Distributions

When we are uncertain about our target system then we cannot guarantee that any particular logical statement is absolutely true or false. In order to quantify uncertainty about our descriptions we attenuate logical propositions by assigning to each statement a *probability* that quantifies how plausible it is be true. Probabilities are bounded between 0, indicating that the statement is absolutely false, and 1, indicating that the statement is absolutely true.

A *probability distribution* quantifies the uncertainty about the entire system by assigning probabilities to each event in a given representation. More formally, a probability distribution is a measure assigning probabilities to each event,

$$\pi : \mathcal{E}(\Theta) \to [0, 1] \,,$$

such that the probability of the null event is zero, $\pi[\emptyset] = 0$, and the probability of the trivial event is one, $\pi[\Theta] = 1$. These latter two conditions follow immediately from the null event always representing a false proposition and the trivial event always representing a true proposition.

When using a probability distribution to quantify uncertainty about a property taking values in the sample space Θ, we often write $\theta \sim \pi$ which is read as, "θ is distributed according to the probability distribution π" but should really read "The plausibility of any logical statement about θ is assigned by the probability distribution π".

The interaction of these probability assignments and the Boolean operations immediately give the familiar rules of probability. For example, the probability of a conjunction

becomes the sum rule,

$$\pi(E_1 \cup E_2) = \pi(E_1) + \pi(E_2) - \pi(E_1 \cap E_2) \,,$$

and the probability of negation becomes the rule of total probability,

$$\pi(E) = \pi(\Theta) - \pi(E^c) = 1 - \pi(E^c) \,.$$

Lebesgue integrals with respect to probability measures are better known as *expectation values*,

$$\mathbb{E}_\pi : \mathcal{F}(\Theta) \to \mathbb{R},$$

Expectations include means, variances, and higher-order moments and are the foundation of any statistical methodology.

The most important consequence of these definitions is that *all of probability theory reduces to computing expectations*. Any other operation that you may have encountered in probability theory can only ever be an intermediate step in computing a final expectation. In particular, many of the more non-intuitive aspects of probability theory can avoided by carefully framing everything as an expectation – don't try to intuit solutions, calculate them!

2.4   Pushforward Probability Distributions

Being measures, probability distributions can also be pushed forward from one measurable space into another using measurable maps.

Because it is invariant when we push it forward between equivalent measurable spaces, a probability distribution is fundamental to the underlying abstract system and not any particular representation of that system. In other words, different but equivalent measurable spaces are just different ways to describe the same system, with inclusion in events quantifying the same, invariant information and probability quantifying the same, invariant uncertainty.

2.5   Conditional Probability Distributions

Conditional probability distributions allow us to quantify uncertainty about implications. While implications assign an event to each value of the conditioning space, Φ, a conditional probability distribution assigns an entire probability distribution to each value in the conditioning space,

$$\pi_{\Theta|\Phi} : \mathcal{E}(\Theta) \times \Phi \to [0, 1]$$
$$(E_\Theta, \phi) \ \mapsto \pi_{\Theta|\Phi}(E_\Theta \mid \phi) \,.$$

In other words, for any value of $\phi \in \Phi$ the conditional probability distribution defines a probability distribution on Θ, and for any event in Θ the conditional probability distribution defines a function from Φ to probabilities.

2.5.1 Joint Distributions and Marginal Distributions

As with implications, conditional probability distributions can be used to construct probability distributions on larger spaces. In particular, by combining a conditional probability distribution with a probability distribution on the conditioning space, Φ, we can construct a probability distribution on the joint sample space, $\Theta \times \Phi$.

This *joint distribution* is defined implicitly by its probability assignments or expectation values. For example, the probability of any joint event, $E_\Theta \times E_\Phi$, is given by first using the conditional probability distribution to assign a probability to E_Θ for each $\phi \in \Phi$,

$$p(\phi) \equiv \pi_{\Theta|\Phi}(E_\Theta \mid \phi),$$

and then taking the expectation of this function with respect to the distribution on Φ,

$$\pi_{\Theta \times \Phi}(E_\Theta \times E_\Phi) = \mathbb{E}_{\pi_\Phi}[p \cdot \mathbb{I}_{E_\Phi}],$$

where the indicator function, $\mathbb{I}_{E_\Phi}$, ensures that we take the expectation only over the event in Φ. Similarly, joint expectations are defined iteratively as

$$\mathbb{E}_{\pi_{\Theta \times \Phi}}[g(\theta, \phi)] = \mathbb{E}_{\pi_\Phi}\left[\mathbb{E}_{\pi_{\Theta|\Phi}}\Big[g(\theta, \phi) \mid \phi\Big]\right].$$

If we consider only the trivial event on the conditioning space $E_\Phi = \Phi$, then this construction also defines a *marginal distribution* on Θ by the probability assignments

$$\pi_\Theta(E_\Theta) \equiv \pi_{\Theta \times \Phi}(E_\Theta \times \Phi)$$
$$= \mathbb{E}_{\pi_\Phi}\left[\pi_{\Theta|\Phi}(E_\Theta \mid \phi)\right],$$

or the expectations

$$\mathbb{E}_{\pi_\Theta}[f(\theta)] \equiv \mathbb{E}_{\pi_{\Theta \times \Phi}}[f(\theta)]$$
$$= \mathbb{E}_{\pi_\Phi}\left[\mathbb{E}_{\pi_{\Theta|\Phi}}\Big[f(\theta) \mid \phi\Big]\right].$$

Alternatively, given the *projection*

$$\varpi : \Theta \times \Omega \to \Theta$$
$$(\theta, \omega) \mapsto \theta,$$

we can also define this marginal distribution as the pushforward of the joint distribution along the projection,

$$\pi_\Theta = \varpi_* \pi_{\Theta \times \Omega}.$$

This marginalization process allows us to collapse a joint probability distribution onto any of the component spaces, explicitly incorporating any correlations between the components in the process.

2.5.2 Bayes' Theorem

While we can decompose any joint probability distribution into a conditional probability distribution and a marginal distribution, there is not unique decomposition. For example, a joint probability distribution on $\Theta \times \Phi$ can be decomposed into $\pi_{\Theta|\Phi}$ and π_Φ or $\pi_{\Phi|\Theta}$ and π_Θ.

The equivalence of these two decompositions immediately implies *Bayes' Theorem* which allows us to define any one of these distributions from the other three using Radon-Nikodym derivatives. If $\pi_{\Theta|\Phi}$ and π_Θ are absolutely continuous with respect to each other then we have

$$\pi_{\Theta|\Phi} = \frac{\mathrm{d}\pi_{\Phi|\Theta}}{\mathrm{d}\pi_\Phi}\pi_\Theta$$

$$\pi_\Theta = \frac{\mathrm{d}\pi_\Phi}{\mathrm{d}\pi_{\Phi|\Theta}}\pi_{\Theta|\Phi}$$

Similarly, if $\pi_{\Phi|\Theta}$ and π_Φ are absolutely continuous with respect to each other then we have the two additional equalities

$$\pi_{\Phi|\Theta} = \frac{\mathrm{d}\pi_{\Theta|\Phi}}{\mathrm{d}\pi_\Theta}\pi_\Phi$$

$$\pi_{\Theta|\Phi} = \frac{\mathrm{d}\pi_\Theta}{\mathrm{d}\pi_{\Theta|\Phi}}\pi.$$

Although an immediate consequence of the axioms of probability theory, Bayes' Theorem has extraordinarily important consequences in the application of probability theory. These equalities allow us to invert any implicative structure – if we have a probabilistic model for how uncertainty of propositions in Φ affects uncertainty of propositions in Θ, then Bayes' Theorem identifies how uncertainty of propositions in Θ affects the uncertainty of propositions in Φ.

2.5.3 Generative Modeling

Just as we can build logical propositions sequentially from simpler implications, we can also build probability distributions over high-dimensional product spaces sequentially from conditional probability distributions. Mirroring the procedure for implications, we start with a probability distribution on one low-dimensional component of the target space and then build up a joint distribution by adding conditional probability distributions for each

new component,

$$\pi_{\Theta_1}$$
$$\pi_{\Theta_2 | \Theta_1}$$
$$\pi_{\Theta_3 | \Theta_2, \Theta_1}$$
$$\ldots$$
$$\pi_{\Theta_N | \Theta_{N-1}, \ldots, \Theta_2, \Theta_1}.$$

As with implications, these conditional probability distributions are often motivated by the natural structure of our target system. In particular, if we think about deterministic processes as degenerate conditional probability distributions that assign all probability to a single event for each conditioning value,

$$\pi_{\Theta | \Phi}[E_\Theta \mid \phi] = \begin{cases} 0, & E_\Theta \neq \hat{E}(\phi) \\ 1, & E_\Theta = \hat{E}(\phi) \end{cases},$$

then these conditional probability distributions can seamlessly incorporate both stochastic and deterministic, causal relationships. This iterative process of building a joint probability distribution from conditional probability distributions is the key building block of *generative modeling.*

3 | Probability in Practice: Probabilistic Computation

The immediate problem with the abstract definitions introduced in the previous chapter is that they do not provide an explicit means of either specifying a probability distribution, at least without considering every single event, or computing expectations in practical applications. When the sample space is structured, however, that structure can be leveraged to provide the explicit implementations we need to apply probability theory in practice. This is particularly evident when the sample space ordered, discrete, or a subset of real numbers.

3.1 Cumulative Distribution Functions

When the sample space is ordered we can completely specify a probability distribution with only the probabilities assigned to *intervals*, $\mathcal{I}(\Theta) \subset \mathcal{E}(\Theta)$. Intervals are events spanning all points less than or equal to some distinguished point, θ,

$$I(\theta) = \left\{ \theta' \in \Theta \mid \theta' \leq \theta \right\}.$$

The function that assigns these probabilities,

$$P : \Theta \to \mathcal{I}(\Theta) \to [0, 1]$$
$$\theta \mapsto I(\theta) \ \mapsto \mathbb{P}[I(\theta)].$$

is called the *cumulative distribution function.*

Cumulative distribution functions naturally pushforward from one sample space into another. Given a measureable map $s : \Theta \to \Omega$, the cumulative distribution function on Ω is defined by

$$P_\Omega(I_\omega) \equiv P_\Theta\left(s^{-1}(I_\omega)\right).$$

3.2 Implementations of Probability Distributions over Discrete Sample Spaces

Every discrete sample space features a natural measure which we can use to not only specify all other measures, and hence probability distributions, but also convert expectations into summations.

3.2.1 The Counting Measure and Mass Functions

The *counting measure* on a discrete space assigns unit measure to the *point events*,

$$\chi(\theta) = 1, \ \forall \theta \in \Theta.$$

In order to compute the measure of any other event we simply count the elements in that event,

$$\chi(E) = \sum_{\theta \in E} 1 = |E|,$$

hence the name. Similarly, Lebesgue integrals reduce to a summation over the sample space weighted by the function values,

$$\mathcal{I}_\chi[f] = \sum_{\theta \in \Theta} f(\theta) \, \chi(\theta) = \sum_{\theta \in \Theta} f(\theta),$$

for any $f \in \mathcal{F}(\Theta)$.

Because the counting measure assigns zero measure only to the null event, all discrete measures are absolutely continuous with respect to it. Consequently, any discrete measure can be defined by its Radon-Nikodym derivative with respect to the counting measure. In this case the Radon-Nikodym derivative is denoted a *mass function*

$$m : \Theta \to \mathbb{R}^+.$$

Measures and Lebesgue integrals then reduce to discrete summations weighted by the mass function,

$$\mu(E) = \sum_{\theta \in E} m(\theta)$$

$$\mathcal{I}_\mu[f] = \sum_{\theta \in \Theta} f(\theta) \, m(\theta).$$

Moreover, the Radon-Nikodym derivative between different discrete measures reduces to the simple ratio of their mass functions,

$$\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1}(\theta) = \frac{m_2(\theta)}{m_1(\theta)}.$$

3.2.2 Transforming Counting Measures and Mass Functions

The counting measure, and hence mass functions, also have the convenient property that they transform naturally with changes to the underlying sample space. Given a measurable map $s : \Theta \to \Omega$, a mass function on Θ pushes forward to a mass function on Ω,

$$m_\Omega(\omega) \equiv m_\Theta\big(s^{-1}(\omega)\big) \,.$$

3.2.3 Probability Mass Functions

Mass functions corresponding to probability distributions reduce to *probability mass functions*, which assigns probability to the point events,

$$p : \Theta \to [0, 1] \,.$$

such that

$$1 = \pi[\Theta] = \sum_{\theta \in \Theta} p(\theta) \,.$$

3.2.4 Conditional Probability Mass Functions

Probability mass functions immediately generalize to specify conditional probability distributions by simply adding a conditioning variable,

$$p_{\Theta|\Phi} : \Theta \times \Phi \to [0, 1] \,.$$

Conditional probabilities and conditional expectations are computed as above,

$$\mathbb{P}_{\Theta|\Phi}[E \mid \phi] = \sum_{\theta \in E} p_{\Theta|\Phi}(\theta \mid \phi) \,,$$

$$\mathbb{E}_{\mathbb{P}_{\Theta|\Phi}}[f \mid \phi] = \sum_{\theta \in \Theta} f(\theta) \, p_{\Theta|\Phi}(\theta \mid \phi) \,,$$

A huge advantage of this representation is that it drastically simplifies the construction of joint and marginal probability distributions. Instead of implicitly defining an abstract joint distribution with probability assignments or expectations, for example, we can construct an explicit joint probability mass function as the product of a conditional mass function and a marginal mass function,

$$p_{\Theta \times \Phi}(\theta, \phi) = p_{\Theta|\Phi}(\theta|\phi) \, p_\Phi(\phi) \,,$$

which readily gives joint probabilities and joint expectations.

Marginalization proceeds similarly – the marginal probability mass function is given by simply summing the joint probability mass function over the nuisance components,

$$p_\Theta(\theta) = \sum_{\phi \in \Phi} p_{\Theta \times \Phi}(\theta, \phi)$$
$$= \sum_{\phi \in \Phi} p_{\Theta | \Phi}(\theta | \phi) \, p_\Phi(\phi) \,.$$

Bayes' Theorem implies that any decomposition of a joint probability mass function is equal,

$$p_{\Phi | \Theta}(\phi | \theta) \, p_\Theta(\theta) = p_{\Theta \times \Phi}(\theta, \phi) = p_{\Theta | \Phi}(\theta | \phi) \, p_\Phi(\phi) \,, \tag{3.1}$$

and from this one equality we can then recover the various implications of Bayes' Theorem. For example,

$$\pi_{\Theta | \Phi} = \frac{\mathrm{d}\pi_{\Phi | \Theta}}{\mathrm{d}\pi_\Phi} \pi_\Theta$$

becomes

$$p_{\Theta | \Phi}(\theta \mid \phi) = \frac{p_{\Phi | \Theta}(\phi \mid \theta)}{p_\Phi(\phi)} p_\Theta(\theta) \,,$$

which is an immediate consequence of dividing both sides of the joint equality (3.1) by $p_\Phi(\phi)$.

3.2.5   Relating Probability Mass Functions and Cumulative Distribution Functions

Because probability mass functions and cumulative distribution functions both specify the same probability distribution, one can always be used to construct the other. Given a probability mass function, for example, we can construct the cumulative distribution function as

$$P(\theta) = \mathbb{P}[I(\theta)] = \sum_{\theta' \in I(\theta)} p(\theta') \,.$$

Similarly, we can construct a probability mass function from a cumulative distribution function as

$$p(\theta) = P(\theta) - P(\theta_-) \,,$$

where θ_- is the largest element of Θ less that θ,

$$\theta_- = \max \left\{ \theta' \in \Theta \mid \theta' < \theta \right\} \,.$$

24

3.3 Implementations of Probability Distributions over Real Numbers

When the sample space are D-dimensional real numbers, or a subset thereof, there are always an uncountably infinite number of point events and we can't assign a finite measure to each without the measure of any dense event becoming infinite. Consequently, the counting measure becomes pathological, assigning either infinite measure or zero measure to every dense event. Fortunately, the real numbers admit their own natural measure on which we can implement probability distributions in practice.

3.3.1 The Lebesgue Measure and Density Functions

Lebesgue measure on R assigns measure based on interval length, then generalize as a product of Lebesgue measures that assigns measure based on volume.
    The Lebesgue measure on the one-dimensional real numbers assigns to each interval,

$$I = [a, b] \subset \mathbb{R},$$

a measure proportional to its length,

$$\lambda(I) = b - a.$$

This generalizes to the general Lebesgue measure, λ, which assigns to each rectangular neighborhood

$$I_D = \prod_{d=1}^{D} I_d = \prod_{d=1}^{D} [a_d, b_d] \subset \mathbb{R}^D,$$

a measure proportional to its volume,

$$\lambda(I_d) = \prod_{d=1}^{D} (b_d - a_d).$$

Although not immediately obvious, this condition implies that the Lebesgue measure of an arbitrary event is also given by the enclosed volume specified by the Riemann integral over that event

$$\lambda(E) = \int_E \mathrm{d}^D \theta.$$

Because of this equivalence, it is common shorthand to write the Lebesgue measure on $\Theta = R^D$ as

$$\lambda = \mathrm{d}^D \theta = \prod_{d=1}^{D} \mathrm{d}\theta_d.$$

Similarly, Lebesgue integrals with respect to the Lebesgue measure (mathematicians have sick sense of humor) also reduce to Riemann integrals,

$$\mathcal{I}_\lambda(f) = \int_\Theta f(\theta)\,\mathrm{d}\theta.$$

An immediate consequence of this result is that all point events have zero Lebesgue measure.

Provided that a measure does not assign non-zero measure to any point events, it will be absolutely continuous with respect to the Lebesgue measure and hence can be completely specified with the corresponding Radon-Nikodym derivative. In this case the Radon-Nikodym derivative is denoted a *density function*,

$$m : \Theta \to \mathbb{R}^+.$$

As above, any measure or Lebesgue integral becomes a standard Riemann integral that can be readily computed with the usual calculus techniques,

$$\mu(E) = \mathcal{I}_\lambda[m \cdot I_E] = \int_E m(\theta)\,\mathrm{d}\theta$$

$$\mathcal{I}_\mu[f] = \mathcal{I}_\lambda[f \cdot m] = \int_\Theta f(\theta)\,m(\theta)\,\mathrm{d}\theta.$$

The Radon-Nikodym derivative between two different real measures behaves exactly as in the discrete case. If μ_2 is absolutely continuous with respect to μ_1 then the density $m_2(\theta)$ can vanish only when the density $m_1(\theta)$ vanishes, and the corresponding Radon-Nikodym derivative is given by

$$\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1}(\theta) = \frac{m_2(\theta)}{m_1(\theta)}.$$

When both densities vanish we have to be careful take an appropriate limit in order to evaluate the Radon-Nikodym derivative.

3.3.2   Transforming Lebesgue Measures and Density Functions

One of the most subtle properties of Lebesgue measures, and hence density functions, is that they do not transform as naturally as their discrete counterparts. The difficulty is that nonlinear transformations warp the real numbers, distorting volumes and, consequently, the Lebesgue measure.

Consider a nonlinear, measurable map $s : \Theta \to \Omega$, and let $\tilde{\lambda}_\Omega$ be the corresponding pushforward of the Lebesgue measure on Θ onto Ω,

$$\tilde{\lambda}_\Omega(E_\Omega) \equiv s_* \lambda_\Theta(E_\Omega) = \mu_\Theta^L\left[s^{-1}(E_\Omega)\right].$$

The pushforward $\tilde{\lambda}_\Omega$ is not the Lebesgue measure on Ω, rather the two are related by the Radon-Nikodym derivative

$$\frac{\mathrm{d}\lambda_\Omega}{\mathrm{d}\tilde{\lambda}_\Omega} = |\mathbf{J}|,$$

where the matrix

$$J_{ij} = \frac{\partial \omega_i}{\partial \theta_j} \equiv \frac{\partial s_i}{\partial \theta_j}$$

is called the *Jacobian* of the transformation. The determinant of the Jacobian matrix quantifies how the nonlinear transformation warps volumes and hence the Lebesgue measure.

This nontrivial relationship between Lebesgue measures and their pushforwards has an important consequence for density functions. Density functions do not transform as we might naively expect,

$$m_\Omega(\omega) \neq m_\Theta\big(s^{-1}(\omega)\big),$$

rather they too are modified by the determinant of the Jacobian matrix,

$$m_\Omega(\omega) = m_\Theta\big(s^{-1}(\omega)\big)\,|\mathbf{J}|^{-1}.$$

We can also motivate this relationship by considering the invariance of Lebesgue integrals. Equivalent sample spaces feature distinct Lebesgue measure and density functions, but they must have the *same Lebesgue integrals* (Figure 3.1). The Lebesgue integrals are invariant only if the Lebesgue measures and densities transform oppositely to each other.

Consequently, density functions depend not just on the abstract system that we're trying to describe but also on the particular representation that we're using. The only application of density functions that are not sensitive to this dependence is integration, and we have to be careful not to take densities too seriously when they are not explicitly under an integral sign.

Univariate Examples of Density Transformations

A common motivation for transforming the sample space, and hence any probability density functions, is when the nominal sample space is contained. Mapping to an unconstrained sample space removes the unconstraint, potentially simplifying calculations.

For example, let $p_\Theta(\theta)$ be a probability distribution function defined on $\Theta = \mathbb{R}^+$. We can unconstrained the positivity constraint by applying a logarithmic transformation,

$$\omega = s(\theta) = \log \theta \in \mathbb{R}.$$

which results in the Jacobian

$$J = \frac{\partial s}{\partial \theta}(\omega) = \frac{1}{\exp(\omega)}$$

and the transformed probability density function on $\Omega = \mathbb{R}$,

$$p_\Omega(\omega) = p_\Theta\big(s^{-1}(\omega)\big)\,|J|^{-1}$$
$$= p_\Theta\big(s^{-1}(\omega)\big)\exp(\omega).$$

27

$$\mu(E_\Theta) = \int_{E_\Theta} m_\Theta(\theta_1, \theta_2) \mathrm{d}\theta_1 \mathrm{d}\theta_2 \qquad \mu(s(E_\Theta)) = \int_{s(E_\Theta)} m_\Omega(\omega_1, \omega_2) \mathrm{d}\omega_1 \mathrm{d}\omega_2$$

(a)   (b)

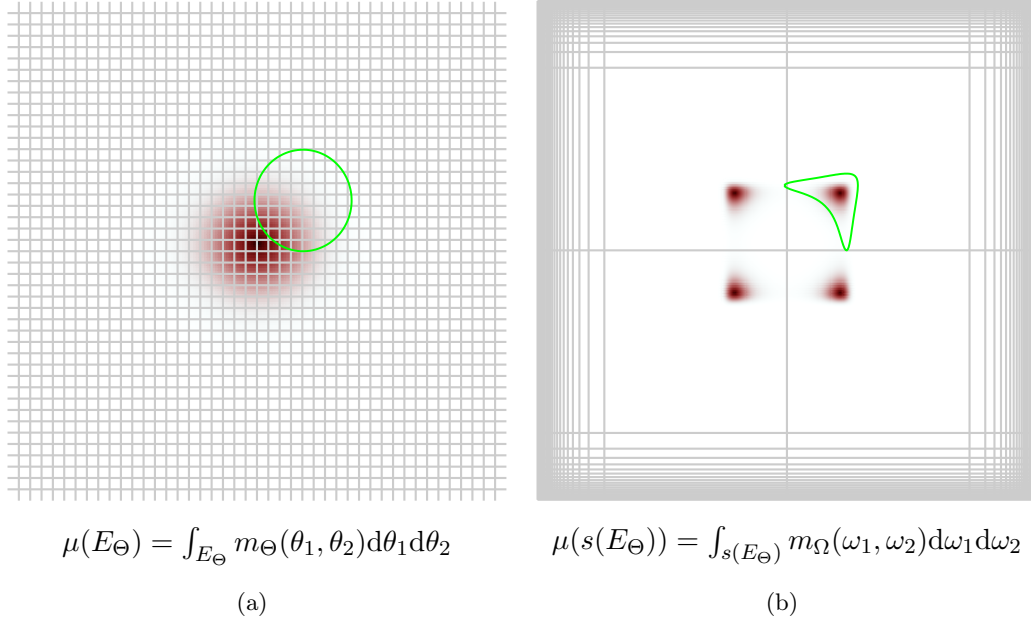Figure 3.1: When sample spaces are real, each sample space has its own density functions (red), events (green), and differential volumes (grey). Here the sample space in (b) is related to the sample space in (a) by the compatible mapping $(\omega_1, \omega_2) = s(\theta_1, \theta_2)$. All of these differences, however, exactly compensate to ensure that integrals always yield the same measures, $\mu(E_\Theta) = \mu(s(E_\Theta))$.

Similarly, we can unconstrain a sample space representing probabilities themselves, $\Theta = [0, 1]$ by applying a logistic transform,

$$\omega = s(\theta) = \frac{1}{1 + \exp(-\theta)} \in \mathbb{R}.$$

This results in the Jacobian

$$J = \frac{\partial s}{\partial \theta}(\omega) = \frac{1}{\omega(1 - \omega)}$$

and the transformed probability density function on $\Omega = \mathbb{R}$,

$$p_\Omega(\omega) = p_\Theta\left(s^{-1}(\omega)\right) |J|^{-1}$$
$$= p_\Theta\left(s^{-1}(\omega)\right) \omega(1 - \omega).$$

A Multivariate Example of Density Transformations

To see how much a probability density function can be warped, consider a two-dimensional sample space with real components (θ_1, θ_2) and the probability density function

$$p_\Theta(\theta_1, \theta_2) \propto \exp\left(-\frac{\theta_1^2 + \theta_2^2}{2}\right).$$

and the measurable map

$$(\omega_1, \omega_2) = r(\theta_1, \theta_2) = (s(\theta_1), s(\theta_2)),$$

where the component maps are given by

$$s(\theta) = \log\left(\frac{\pi + 2\operatorname{atan}(\alpha\,\theta)}{\pi - 2\operatorname{atan}(\alpha\,\theta)}\right)$$

with the inverse

$$s^{-1}(\omega) = \frac{1}{\alpha} \tan\left(\frac{\pi}{2} \frac{e^\omega - 1}{e^\omega + 1}\right)$$

and Jacobian

$$J(\omega) = \frac{\partial s}{\partial \theta}(\omega) = \frac{\alpha}{\pi} \frac{(1 + e^\omega)^2}{e^\omega} \sin^2\left(\frac{\pi}{1 + e^\omega}\right).$$

The Jacobian of the complete map is then given by

$$\mathbf{J} = \begin{bmatrix} J & 0 \\ 0 & J \end{bmatrix},$$

with the determinant $|\mathbf{J}| = J^2$. Hence the transformed probability density function and differential volume are given by

$$p_\Omega(\omega_1, \omega_2) = p_\Theta\left(s^{-1}(\omega_1), s^{-1}(\omega_2)\right) J^{-2}(\omega)$$

and

$$\mathrm{d}\omega_1 \mathrm{d}\omega_2 = J^2 \mathrm{d}\theta_1 \mathrm{d}\theta_2,$$

respectively. The extreme differences between these two representations are evident graphically in Figure 3.1.

3.3.3 Probability Density Functions

Density functions corresponding to probability distributions reduce to *probability density function* which assign a positive value to each point in the sample space,

$$p : \Theta \to \mathbb{R}^+.$$

such that

$$1 = \pi(\Theta) = \int_\Theta p(\theta)\,\mathrm{d}\theta.$$

These values, known as *probability densities*, have no particular meaning of their own and instead exist only to be integrated to give probabilities,

$$\pi(E) = \int_E p(\theta)\,\mathrm{d}\theta,$$

and expectations,

$$\mathbb{E}_\pi[f] = \int_\Theta f(\theta)\,p(\theta)\,\mathrm{d}\theta.$$

This is an important point that is worth repeating – probability densities are meaningless until they have been integrated over some event. To analogize with physics, the event over which we integrate corresponds to a *volume* and the probability given by integrating the density over such a volume corresponds to a *mass*. When we want to be careful to differentiate between probabilities and probability densities we'll use *probability mass* to refer to the former.

3.3.4 Conditional Probability Density Functions

Just as in the discrete case, probability density functions can be immediately extended to implement conditional probability distributions by simply adding a conditioning variable,

$$p_{\Theta|\Phi} : \Theta \times \Phi \to \mathbb{R}^+,$$

with conditional probabilities and conditional expectations reducing to integrals,

$$\pi_{\Theta|\Phi}(E \mid \phi) = \int_E p_{\Theta|\Phi}(\theta \mid \phi)\,\mathrm{d}\theta,$$

$$\mathbb{E}_{\pi_{\Theta|\Phi}}[f \mid \phi] = \int_\Theta f(\theta)\,p_{\Theta|\Phi}(\theta \mid \phi)\,\mathrm{d}\theta,$$

Likewise, probability density functions representing joint and marginal probability distributions are easy to construct for conditional probability density functions. Joint probability density functions are given by a simple multiplication,

$$p_{\Theta \times \Phi}(\theta, \phi) = p_{\Theta|\Phi}(\theta|\phi) \, p_\Phi(\phi) \,,$$

and marginal probability density functions are given by integrating out the nuisance components,

$$
\begin{aligned}
p_\Theta(\theta) &= \int_\Phi p_{\Theta \times \Phi}(\theta, \phi) \, \mathrm{d}\phi \\
&= \int_\Phi p_{\Theta|\Phi}(\theta|\phi) \, p_\Phi(\phi) \, \mathrm{d}\phi.
\end{aligned}
$$

Bayes' Theorem also implies that any decomposition of a joint probability density function is equal,

$$p_{\Phi|\Theta}(\phi|\theta) \, p_\Theta(\theta) = p_{\Theta \times \Phi}(\theta, \phi) = p_{\Theta|\Phi}(\theta|\phi) \, p_\Phi(\phi) \,, \tag{3.2}$$

and from this one equality we can then recover the various implications of Bayes' Theorem. For example,

$$\pi_{\Theta|\Phi} = \frac{\mathrm{d}\pi_{\Phi|\Theta}}{\mathrm{d}\pi_\Phi} \pi_\Theta$$

becomes

$$p_{\Theta|\Phi}(\theta \mid \phi) = \frac{p_{\Phi|\Theta}(\phi \mid \theta)}{p_\Phi(\phi)} p_\Theta(\theta) \,,$$

which is an immediate consequence of dividing both sides of the joint equality (3.2) by $p_\Phi(\phi)$.

3.3.5 Relating Probability Density Functions and Cumulative Distribution Functions

As with their discrete counterparts, probability density functions and cumulative distribution functions can always be derived from each other because they specify the same probability distribution. Cumulative distribution functions, for example, are given by integrating over probability density functions,

$$P(\theta) = \pi[I(\theta)] = \int_{\theta_{\min}}^{\theta} p(\theta') \, \mathrm{d}\theta.$$

Probability density functions, on the other hand, are given by differentiating cumulative distribution functions,

$$p(\theta) = \frac{\partial P(\theta)}{\partial \theta}.$$

31

Note that if we map to an equivalent measurable space with $s : \Theta \to \Omega$, then the derivative acquires a factor of the inverse Jacobian so that the corresponding probability density function transforms as necessary,

$$p(\omega) = \frac{\partial P(\omega)}{\partial \omega} = \frac{\partial P\big(s^{-1}(\omega)\big)}{\partial \theta} \frac{\partial \theta}{\partial \omega} = p\big(s^{-1}(\omega)\big) \left| \mathbf{J}^{-1} \right|.$$

3.3.6 Implementations of Mixed Probability Distributions

Distributions over samples spaces that have both a discrete, Φ, and a real Ψ, component can be implemented by leveraging discrete and continuous representations of conditional distributions.

For example, a distribution over $\Theta = \Phi \times \Psi$ can be specified by conditioning the discrete component with the real component, $\pi_{\Phi|\Psi}$, and providing a marginal distribution over the real component, π_Ψ. Using a conditional probability mass function for the former and a probability density function for the latter, the probability of any event is given by

$$\begin{aligned}
\pi_\Theta(E) &= \pi_\Theta(E_\Phi \times E_\Psi) \\
&= \mathbb{E}_{\pi_\Psi}\Big[\pi_{\Phi|\Psi}(E_\Phi \mid \psi) \cdot \mathbb{I}_{E_\Psi}(\psi)\Big] \\
&= \int_{E_\Psi} \sum_{\phi \in E_\Phi} p_{\Phi|\Psi}(\phi \mid \psi)\, p_\Psi(\psi)\, \mathrm{d}\psi,
\end{aligned}$$

with expectations given similarly by

$$\begin{aligned}
\mathbb{E}_{\pi_\Theta}[f] &= \mathbb{E}_{\pi_\Psi}\Big[\mathbb{E}_{\pi_{\Phi|\Psi}}[f \mid \psi]\Big] \\
&= \int_{E_\Psi} \sum_{\phi \in E_\Phi} f(\phi, \psi)\, p_{\Phi|\Psi}(\phi \mid \psi)\, p_\Psi(\psi)\, \mathrm{d}\psi.
\end{aligned}$$

Equivalently, we could also condition the real component on the discrete component, $\pi_{\Psi|\Phi}$ and provide a marginal distribution over the discrete component, π_Φ. We could then represent the distribution with a conditional probability density function for the former and a probability mass function for the latter. Likewise, the probability of any event is given by

$$\begin{aligned}
\pi_\Theta(E) &= \pi_\Theta(E_\Phi \times E_\Psi) \\
&= \mathbb{E}_{\pi_\Phi}\Big[\pi_{\Psi|\Phi}(E_\Psi \mid \phi) \cdot \mathbb{I}_{E_\Phi}(\phi)\Big] \\
&= \sum_{\phi \in E_\Phi} \int_{E_\Psi} p_{\Psi|\Phi}(\psi \mid \phi)\, p_\Phi(\phi)\, \mathrm{d}\psi,
\end{aligned}$$

with expectations given similarly by

$$\mathbb{E}_{\mathbb{P}_\Theta}[f] = \mathbb{E}_{\mathbb{P}_\Phi}\left[\mathbb{E}_{\mathbb{P}_{\Psi|\Phi}}[f \mid \phi]\right]$$
$$= \sum_{\phi \in E_\Phi} \int_{E_\Psi} f(\phi, \psi)\, p_{\Psi|\Phi}(\psi \mid \phi)\, p_\Phi(\phi)\, \mathrm{d}\psi.$$

Regardless of how we choose to decompose the mixed distribution, combining probability density functions and probability mass functions makes specifying and manipulating these distributions straightforward in practice.

3.4 Stochastic Implementations of Probability Distributions

We can also specify any probability distributions and compute its expectations *stochastically*. A *stochastic process* is any mechanism that generates a sequence of states, or *samples*, from a given sample space, $\{\theta_1, \ldots, \theta_N\} \subset \Theta$. In general, the generation of the n-th state in the sequence, θ_n, can depend on all previous states in the sequence, $\{\theta_1, \ldots, \theta_{n-1}\}$.

A stochastic process implements a given probability distribution if the samples themselves can be used to recover all expectations as the size of the sequence becomes infinitely large. More formally, if

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} f(\theta_n) = \mathbb{E}_\pi[f],$$

for all well-behaved functions, $f : \Theta \to \mathbb{R}$, then the stochastic process implements all computations with respect to the probability distribution π.

This procedure, and hence the resulting samples, are *exact* if every state in the sequence is generated independently of any previous states. In other words, samples are exact when the stochastic process generates samples one at a time, with no dependence on the preceding or following samples. If samples are not exact we refer to them as *correlated*.

Exact stochastic processes are, by construction, perfectly *random* processes: there is no way to predict any element of the sampling sequence given the state of other samples in that sequence. Unfortunately such randomness is impossible to achieve in practice as computers are fundamentally deterministic. Instead we rely on *pseudorandom* processes which generate sequences in a convoluted but ultimately deterministic fashion. When we are ignorant of the precise configuration of a pseudorandom process the resulting samples *appear* to be random for all intents and purposes.

Pseudorandom processes that target specific probability distributions, or *pseudorandom number generators*, can be devised for many simple probability distributions. Unfortunately they are typically impossible to construct for the more complex distributions that are often of practical interest.

4 | Probability in Practice: The Complexity of Computation

The beauty of probability theory is that, once we have selected a *target* probability distribution, the only well-posed computations are expectations. As noted in the previous chapter, the many subtleties and apparent paradoxes of probability theory are readily overcome by ignoring our typically-biased intuition and instead posing questions as expectations and computing.

Once we've settled on an implementation of our probability distribution, these computations reduce to straightforward computations, for example summation in the discrete case or integration in the real case. Unfortunately, the conceptual elegance of these computations does not imply that the calculations themselves are trivial.

Outside of the most simple problems the necessary summations and integrations cannot be calculated analytically and we must be satisfied with only approximations. Moreover, the only way to *guarantee* accurate estimation is to exhaustively survey the sample space, and for large sample spaces the cost of such surveys easily overwhelms our finite computational resources. Consequently, computationally efficient yet accurate estimates require more sophisticated approaches that take exploit the geometry of the target probability distribution itself.

In this section we study the geometry of probability distributions on high-dimensional sample spaces, especially how that geometry frustrates approximate computation, both in theory and with an explicit example. From here on we will consider only real sample spaces; much of the intuition we will develop does carry over to the discrete case, but developing algorithms around that corresponding intuition is still an open problem in statistics.

4.1 Concentration of Measure

The key to constructing robust estimates of any expectation is identifying which neighborhoods of our target sample space contribute to the corresponding integrals; any computation outside of those relevant neighborhoods is wasted. How to identify those relevant neighborhoods for a given target distribution, however, is not immediately obvious. All that we

Figure 4.1: On high-dimensional samples spaces well-behaved probability distributions concentrate in a neighborhood called the *typical set*. In order to estimate the integrals needed to compute probabilities and expectations we have to be able to identify where the typical set lies in the sample space, which is a challenging task.

know is that, because those neighborhoods define expectations, they should be equivalent for equivalent sample spaces.

Naively, we might consider a neighborhood around the mode of our representative probability density function where the density, and presumably also contributions to any integral, is largest. Because this neighborhood is dependent on the particular representation, however, it doesn't have the necessary invariance properties. Our initial intuition must be missing something important.

Indeed, our naive intuition overlooks the *volume* over which we integrate the probability density functions. Although probability density functions concentrate around their modes, the Lebesgue measure over which we integrate does not. As we move away from the mode towards infinity there is more and more volume over which we can integrate, especially in high-dimensions. Consequently the integrand, which incorporates both the density and the volume, concentrates in a singular neighborhood somewhere in between which we denote the *typical set* (Figure 4.1). This ubiquitous phenomenon is known as *concentration of measure*

Transformations between equivalent sample spaces may change where the probability density and volume concentrate, but these changes always cancel exactly to yield an equivalent typical set. Concentration of measure and the typical set are properties of a probability distribution itself and not of any particular representation.

Although this analysis is too vague to help us identify the typical set for a given probability distribution, it does provide crucial understanding of why high-dimensional expectations are so difficult to compute. For example, because probability is distributed across the entire typical set, no *single* point in the sample space yields a good approximation to all

expectations:

> to construct estimators that are accurate for diverse expectations we need to quantify the entire typical set.

Moreover, because points outside of the typical set do not contribute to any integral,

> we need to focus the entirely of our computational resources on evaluations within only the typical set.

Finally, because the typical set is a property of a probability distribution itself,

> all of our computational algorithms should be independent of the details of any particular representation, such as the shape of a probability density function.

4.2   An Explicit Example of Concentration of Measure

Concentration of measure can be difficult to reconcile with our low-dimensional intuition, so let's examine an explicit example. Consider a probability distribution over the D-dimensional real numbers represented by a product of Gaussian probability density functions,

$$p_\Theta(\theta) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\theta_d^2}{2\sigma^2}\right),$$

with the corresponding rectangular Lebesgue measure,

$$\mathrm{d}^n\theta = \prod_{d=1}^{D} \mathrm{d}\theta_d.$$

Because both the probability density function and Lebesgue measure are invariant to rotations, it is easier to analyze this distribution with hyper-spherical coordinates. In hyper-spherical coordinates the original probability density function appears as

$$p_\Theta(r, \phi_1, \ldots, \phi_{D-1}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2}\right),$$

and with respect to the hyper-spherical Lebesgue measure the rectangular Lebesgue measure appears as

$$\mathrm{d}^n\theta = r^{D-1} \prod_{d=1}^{D-1} \sin^{D-d-1}(\phi_d) \cdot \mathrm{d}r \prod_{d=1}^{D} \mathrm{d}\phi_d.$$

From the hyper-spherical perspective, our original probability density function concentrates around the mode and rapidly decays with increasing radius, r. The rectangular Lebesgue
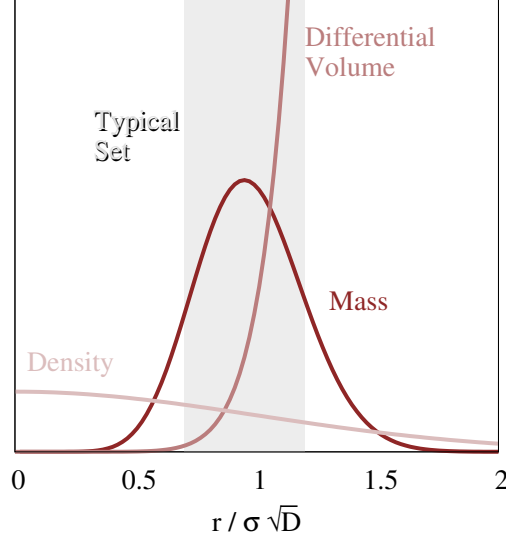
Figure 4.2: In high dimensions real probability distributions generically assign almost all of their probability into a singular neighborhood known as the typical set. This is apparent even from a probability density function representation: although the density concentrates around the corresponding mode, the volume over which we integrate that density is much larger away from the mode. These two opposing trends balance to give the typical set.

measure, however, becomes increasingly larger at higher radius where there is more and more volume over which to integrate (Figure 4.2)

To see where the probability distribution concentrates radially we pushforward the probabiltiy density function,

$$p_\Phi(r, \phi_1, \ldots, \phi_{D-1}) = \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{D}{2}} \exp\left(-\frac{r^2}{2\sigma^2} \right) r^{D-1} \prod_{d=1}^{D-1} \sin^{D-d-1}(\phi_d) \, ,$$

and then marginalize out the hyperspherical angles analytically to give the radial probability density function,

$$p_R(r) = \left(2\sigma^2 \right)^{-\frac{D}{2}} \frac{2}{\Gamma\left(\frac{D}{2}\right)} r^{D-1} \exp\left(-\frac{r^2}{2\sigma^2} \right) ,$$

which is exactly the scaled χ distribution. In particular, for large D all of the probability concentrates in a neighborhood around $r \approx \sigma\sqrt{D}$ with width approximately equal to $\sigma/\sqrt{2}$. In other words, almost all of our target probability can be found in a thin shell at $r = \sigma\sqrt{D}$ and the relative width of that shell decreases as we add more and more dimensions (Figure 4.2).
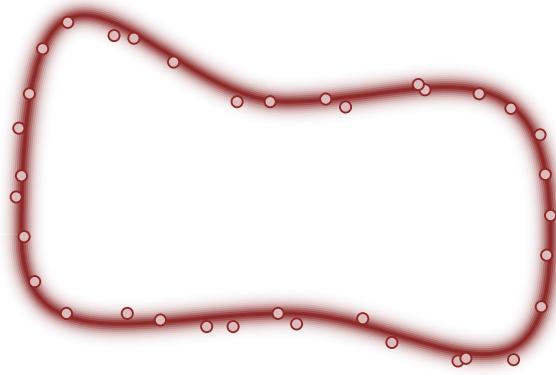
38

Figure 4.3: Because samples recover expectations asymptotically, large sequences of samples must concentrate in the typical set. This provides a means of visualizing concentration of measure, and will prove a powerful way to estimate expectations.

Because samples recover all expectations, they too must concentrate across the typical set (Figure 4.3). Consequently, we can also use samples to simulate concentration of measure. For a given D we generate a sample from our target probability using a univariate Gaussian random number generator available in any computing library,

$$\theta_d \sim \mathcal{N}\left(0, \sigma^2\right),$$

and compute the corresponding radial distance $r = \sqrt{\sum_{d=1}^{D} x_d^2}$. Generating a sequence of samples and then histogramming the radial distance reveals the same χ distribution that we arrived at analytically (Figure 4.4).
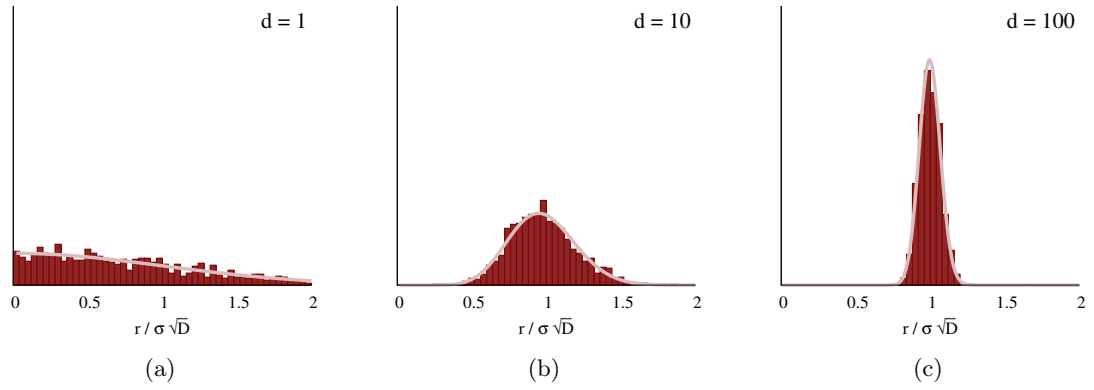
Figure 4.4: Concentration of measure can be visualized with samples from a given distribution, which concentrate across the typical set. For low-dimensions the concentration is weak and the typical set is diffuse, but as the dimensionality of the target distribution grows so too does the concentration of measure.

5 | Probability in Practice: Deterministic Estimators

If we can't compute expectations with respect to our target distribution analytically, one immediate strategy is to replace it with a simpler one. In other words, we can approximate a complex target distribution, π, with a simpler distribution, $\widetilde{\pi}$, whose expectations, or at least some expectations of practical interest, are known analytically,

$$\mathbb{E}_\pi[f] \approx \mathbb{E}_{\widetilde{\pi}}[f].$$

Deterministic estimators use various criteria to identify an optimal approximating distribution so that expectations can be approximated deterministically.

5.1 Modal Estimators

Ideally all expectations with respect to our approximating distribution would be analytic so that we could use it to approximate any expectation with respect to the target distribution. The only probability distribution that fits this criteria is the *Dirac distribution*, $\delta_{\tilde{\theta}}$, that assigns all probability to a single point in the sample space, $\tilde{\theta}$,

$$\delta_{\tilde{\theta}}[E] = \left\{ \begin{array}{ll} 0, & \tilde{\theta} \notin E \\ 1, & \tilde{\theta} \in E \end{array} \right. .$$

Because all probability concentrates at $\tilde{\theta}$, expectations are trivial,

$$\mathbb{E}_\delta[f] = f(\tilde{\theta}).$$

Where, however, should we assign all probability to best approximate the target distribution and its typical set? One of the simplest, and consequently most popular, deterministic estimation strategies is *modal estimation*, where we approximate the target distribution with a Dirac distribution at the mode of the probability density function,

$$\hat{\theta} = \operatorname{argmax} p(\theta).$$

Figure 5.1: (a) In simple cases a prescient choice of representation can yield a modal estimator that well approximates some expectations, such as the mean of the real parameters θ. (b) Poor choices of the representation, however, yield very inaccurate estimates, even in these simple problems.

This approach, however, immediately contradicts the intuition provided by concentration of measure: it utilizes a single point in the sample space that lies outside of the typical set and depends entirely on the choice of representation! Why, then, is modal estimation so ubiquitous?

Modal estimators are seductive because the optimization on which they rely is relatively computationally inexpensive. Moreover, in some very simple cases modal estimators constructed from the probability density function in *some* representations can be reasonably accurate for *some* functions. For example, if the target probability distribution is sufficiently simple that the typical set is approximately symmetric *and* if a representation can be found such that the mode of the probability density function lies in the center of that typical set, then the corresponding modal estimator might yield reasonably accurate estimates for the mean, $\mathbb{E}_\pi[\theta]$ (Figure 5.1a). Other choices of representation, can just as easily yield terribly inaccurate estimators (Figure 5.1b).

Moreover, because they rely on a point estimate, modal estimators are terrible at approximating expectations that depend on the breadth of the typical set, such as the variance. Furthermore, identifying the optimal representation for a particular target function, even if one exists, is extremely challenging in practice. Worse, we have no generic means of even quantifying the error in these estimators for a generic target distribution. Ultimately this strong sensitivity to the choice of a particular representation and the inability to validate the accuracy of the estimators makes modal estimation extremely fragile in practice.

5.2 Laplace Estimators

The fragility of modal estimators can be partially resolved by generalizing them to *Laplace estimators* which approximate the probability density function with a Gaussian density function,

$$p(\theta) \approx \mathcal{N}(\theta \mid \mu, \Sigma),$$

where the mean is given by the modal estimate,

$$\mu = \hat{\theta},$$

and the covariance is given by the Hessian of the probability density function evaluated at the mode,

$$\left(\Sigma^{-1}\right)_{ij} = \left.\frac{\partial^2 p(\theta)}{\partial\theta_i \partial\theta_j}\right|_{\hat{\theta}}.$$

Expectations are then estimated with Gaussian integrals

$$\mathbb{E}_\pi[f] \approx \int_\Theta f(\theta)\,\mathcal{N}(\theta \mid \mu, \Sigma)\,\mathrm{d}^D\theta,$$

which often admit analytic solutions.

The accuracy of a Laplace approximations depends on how well the mode and the Hessian of the probability density function quantify the geometry of the typical set (Figure 5.2). Unfortunately the conditions that are necessary for these estimates to be reasonably accurate hold only for very simple probability distributions, and then only if an appropriate representation can be found.

As with the simpler modal estimators, the dependence on the particular representation manifests as fragility of the corresponding estimators, and we have no generic means of quantifying that error in practice. If we want robust estimation of probabilities and expectations then we need strategies that do not depend on the irrelevant properties of any particular representation.

5.3 Variational Estimators

In order to construct an approximation that is not sensitive to the choice of a particular representation we need to frame the problem as an optimization over a space of approximating distributions directly. Optimizations over spaces of probability distributions fall into a class of algorithms known as *variational methods*.

Variational methods are characterized by two choices: the variational family and a divergence function. The *variational family*, $\mathcal{U}$, is a set of probability distributions such
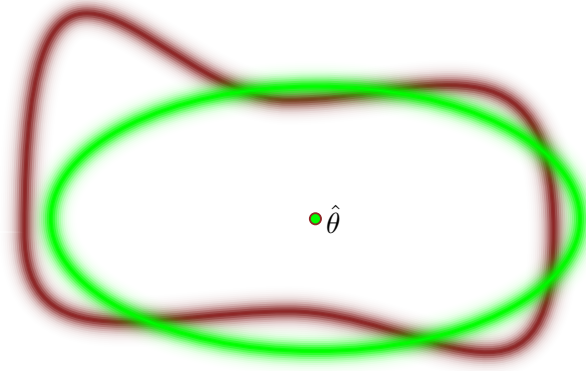
Figure 5.2: For simple probability distributions in well-chosen representations, the local geometry around the mode of a probability density function can quantify the geometry of the entire typical set, yielding accurate Laplace estimators. For more complex probability distributions, however, this local information poorly quantifies the global geometry of the typical set and Laplace estimators suffer from large biases.

that at least some expectations can be computed analytically. In order to identify the best approximation to the target distribution we then define a *divergence function*,

$$D : \mathcal{U} \times \mathcal{U} \to \mathbb{R}^+$$
$$v_1, v_2 \mapsto D(v_1 \| v_2) \,,$$

which is zero if the two arguments are the same and increases as they deviate from each other.

The best approximating distribution is then defined by the variational objective (Figure 5.3).

$$v^* = \underset{v \in \mathcal{U}}{\operatorname{argmin}} \, D(\pi \| v) \,.$$

Although straightforward to define, this variational optimization can be quite challenging in practice. Depending on how the target distribution interacts with the choice of variational distribution and divergence function, the variational objective might feature multiple critical points and we may not be able to find the global optimum (Figure 5.4).

Even if we could find a global optimum, however, there are no guarantee that the chosen variational family is rich enough that the best approximating distribution will yield accurate estimates for all relevant expectations. For example, some divergence functions are biased towards variational solutions that underestimate the breadth of the typical set while others tend to significantly overestimate it (Figure 5.5). Here the means may be approximated adequately well , but the variances would not be.
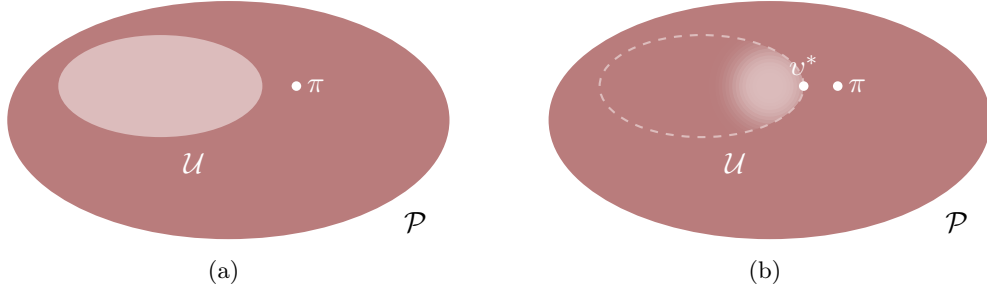
Figure 5.3: (a) A variational family, $\mathcal{U}$, is a set of probability distributions taken from the set of all probability distributions over the same sample space, as the target distribution, $\mathcal{P}$. (b) Adding a divergence function distinguishes which elements of $\mathcal{U}$ are good approximations to the target distribution, allowing us to identify the best approximation, v^*.
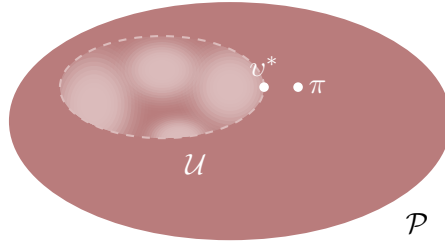
Figure 5.4: Typically different elements of the variational family are able to capture different characteristics of the target distribution and the variational objective manifests multiple optima. Even if a unique global optimum, v^*, exists it will be difficult to find and we may be left with only a suboptimal local optimum.

Variational methods are relatively new to statistics and at the moment there are no generic methods for quantifying the error in variational estimators.
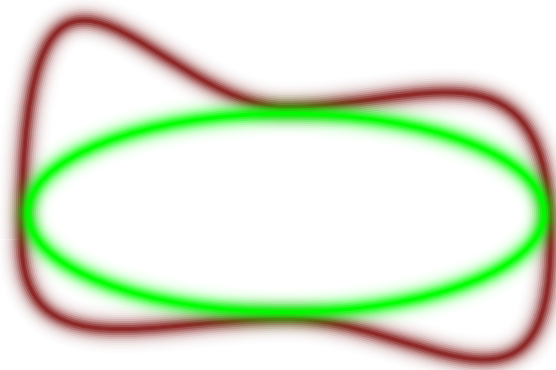
(a)
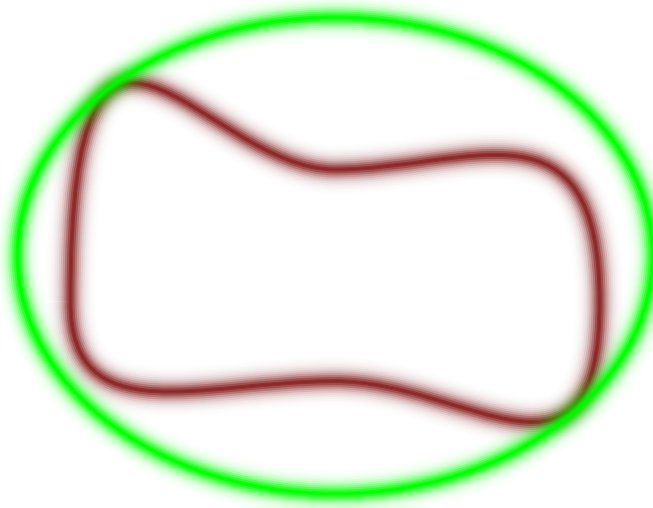
(b)

Figure 5.5: Intuition about the effects of a particular variational divergence function can be developed by considering how the typical set of an approximating distribution (green) interacts with the typical of the target distribution (red). (a) Some divergence functions favor approximating distributions that expand into the interior of the target typical set, resulting in an underestimate of the breadth of the typical set. (b) Others, however, favor approximating distractions that collapse around the exterior of the target typical set, resulting in overestimated expectations.

6 | Probability in Practice: Stochastic Estimators

The accuracy of deterministic estimators will always be limited by the flexibility of the approximating distribution to match the geometry of the exact typical set. The only way to overcome this restriction is to quantify the typical set of the target distribution directly. Unfortunately, this presents problems of its own as in practice we don't know where to find the typical set in the expansive sample space. Because exhaustive search of the sample space is far too expensive we need a more targeted procedure for finding and then exploring the typical set.

By construction an infinite number of samples from the target distribution completely specifies the typical set, and hence the samples themselves provide a natural way to quantify the typical set (Figure 4.3). The utility of samples, however, depends both on how precisely we can quantify the typical set using only a finite number of samples and how well we can generate samples in the first place.

Stochastic estimators use samples, either from the target probability distribution or auxiliary probability distributions, to construct estimators of expectations with respect to the target distribution. Exactly how these samples are generated leads to estimators with substantially different behaviors.

6.1 Monte Carlo Estimators

Monte Carlo estimators use a finite sequence of exact samples from the target distribution to estimate expectations. Given a sequence of exact samples $\{\theta_1, \ldots, \theta_N\}$ we can construct a *Monte Carlo estimator* of the expectation of *any* function $f : \Theta \to \mathbb{R}$ as

$$\hat{f}_N^{\mathrm{MC}} \equiv \frac{1}{N} \sum_{n=1}^{N} f(\theta_n),$$

provided that $\mathbb{E}_\pi\left[f^2\right] < \infty$.

By construction these Monte Carlo estimators recover the exact expectation asymptotically,

$$\lim_{N \to \infty} \hat{f}_N^{\text{MC}} = \mathbb{E}_\pi[f],$$

but they are also quite accurate even when the sequence is finite. Provided that the samples are exact, Monte Carlo estimators follow a Central Limit Theorem – for sufficiently large N the estimators themselves follow a distribution given by a Gaussian density function,

$$\hat{f}_N \sim \mathcal{N}\left(\mathbb{E}_\pi[f], \text{MCSE}\right),$$

where the *Monte Carlo Standard Error* is defined as

$$\text{MCSE} \equiv \sqrt{\frac{\text{Var}[f]}{N}}.$$

Consequently Monte Carlo estimators are unbiased with respect to all of the possible sequences we could have generated, and their precision improves as we generate more and more samples. Moreover, because the Monte Carlo Standard Error doesn't depend on the dimensionality of the target distribution it scales particularly well to high-dimensional problems.

In practice we can use another Monte Carlo estimator to approximate the variance $\text{Var}[f]$, and hence the Monte Carlo Standard Error itself. The error of this approximation is

$$\sqrt{\text{Var}[\text{Var}[f]]/N^2},$$

which is typically negligible compared to the Monte Carlo Standard Error of f. This gives Monte Carlo estimators an important distinction from deterministic approximations in that they naturally come equipped with a procedure for estimating their error.

Of course all of these benefits of Monte Carlo estimators are dependent on our ability to generate exact samples from the target probability distribution. Unfortunately, generating exact samples is infeasible for all but the simplest probability distributions, and we are once again frustrated by our ignorance of the typical set. In order to proceed we need to approximate the exact samples themselves.

6.2   Importance Sampling Estimators

Although we typically can't generate exact samples from the target distribution, often we can generate exact samples from an *auxiliary* probability distribution, γ,

$$\{\vartheta_1, \ldots, \vartheta_N\} \sim \eth \partial \geqslant \geqslant \partial.$$

Importance sampling estimators use these auxiliary samples corrected with *importance weights*, $w(\vartheta)$,

$$\mathbb{E}_\pi[f] \approx \hat{f}_N^{\text{IS}} = \frac{1}{N} \sum_{n=1}^{N} w(\vartheta_n) \, f(\vartheta_n)$$

The importance weights are defined as

$$w(\vartheta_n) = \frac{\mathrm{d}\pi}{\mathrm{d}\gamma}(\vartheta_n)$$

and hence are independent of any representation. A particular representation does facilitate their computation, however, as

$$w(\vartheta_n) = \frac{p(\vartheta_n)}{g(\theta_n)},$$

where p and g are the probability density functions corresponding to the target distribution and auxiliary distribution, respectively.

Given certain regularity conditions, importance sampling estimators also satisfy a Central Limit Theorem

$$\hat{f}_N^{\mathrm{IS}} \sim \mathcal{N}(\mathbb{E}_\pi[f], \mathrm{ISSE}),$$

The *Importance Sampling Standard Error* is given by

$$\mathrm{ISSE} \equiv \sqrt{\frac{\mathrm{Var}[f]}{\mathrm{ESS}}},$$

with the *effective sample size* defined as

$$\mathrm{ESS} = N \frac{\left(\sum_{n=1}^{N} w(\vartheta_n)\right)^2}{\sum_{n=1}^{N} w(\vartheta_n)^2}.$$

Comparing this to the Monte Carlo Central Limit Theorem we can see that the effective sample size quantifies how many exact samples would have yielded the same estimator precision, hence the effective sample size can be interpreted as the effective number of exact samples "contained" in the auxiliary samples.

The challenge with constructing a useful importance sampler is finding an auxiliary distribution that is not too different from the target distribution. Although importance sampling estimators are unbiased, their variance can be so large as to be impractical when the auxiliary distribution deviates too strongly from the target distribution and the weights are large. In fact, when the auxiliary distribution has lighter tails than the target distribution these estimators can easily have infinitely large variance: not only does this make the estimators themselves unreliable, it also makes estimates of the variance and hence any quantification of the estimator error unreliable.

Selecting an auxiliary distribution that yields accurate importance sampling estimators, however, is challenging without knowing the structure of the typical set a priori. For higher-dimensional target distributions, concentration of measure makes this an even more delicate process as any auxiliary distribution that isn't already very close to the target distribution will yield poor importance sampling estimators. Ultimately, importance sampling is most useful as a means to correct a distribution that is already known to be a good approximation to the target distribution.

6.3 Markov Chain Monte Carlo Estimators

Another strategy for approximating the Monte Carlo procedure is to generate samples from the target distribution but relax the requirement that they be exact. Fortunately, correlated samples are readily given by *Markov chains.*

A Markov chain is a stochastic processes generated not by a static probability distribution but by a probability distribution that depends on the last state in the sequence. In other words, each state in the sequence is sampled from a conditional probability distribution known as a *Markov transition kernel, τ,*

$$\tau : \mathcal{E}(\Theta) \times \Theta \to [0, 1]$$
$$(E, \theta) \mapsto \tau(E \mid \theta).$$

When the Markov transition operator preserves the target distribution,

$$\pi(E) = \mathbb{E}_\pi(\tau(E \mid \theta)), \, \forall E \in \mathcal{E}(\Theta),$$

or, with respect to probability density functions,

$$p(\theta) = \int_\Theta t(\theta' \mid \theta) \, \pi(\theta') \, \mathrm{d}^D \theta',$$

then the Markov chain asymptotically recovers expectations with *Markov chain Monte Carlo estimators,*

$$\lim_{N \to \infty} \hat{f}_N^{\mathrm{MCMC}} \equiv \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} f(\theta_n) = \mathbb{E}_\pi[f],$$

and we can interpret the Markov chain itself as a sequence of correlated samples from the target distribution.

More intuitively, the Markov transition kernel quantifies the variability in each step of the Markov chain. When the transition preserves the target distribution, it concentrates closer to the typical set than away from it – it is literally attracted to neighborhoods of high probability. Consequently, the Markov chain will eventually find and then explore the typical set no matter where we start in the sample space, and the Markov chain Monte Carlo estimators will always converge to the true expectations.

The important caveat here is that the Markov chain is guaranteed to find and fully explore the typical set only asymptotically. In practice, however, it is only the finite time performance of Markov chains that matters and, unfortunately, the finite time behavior of Markov chain Monte Carlo is much more subtle than its exact predecessor.

In this section we discuss how Markov chain Monte Carlo behaves under ideal conditions, how it behaves under less-than-ideal conditions, and how to effectively run the algorithm in practice to be robust to the latter.

6.3.1 Markov Chain Monte Carlo Under Ideal Conditions

Under ideal conditions, Markov chains explore the target distribution in three distinct phases. In the first phase the Markov chain converges towards the typical set from its initial position and Markov chain Monte Carlo estimators are highly biased (Figure 6.1a). The second phase begins once the Markov chain finds the typical set and persists through the first sojourn across the typical set. This initial exploration is extremely effective and the accuracy of Markov chain Monte Carlo estimators rapidly improves as the bias from the initial samples is eliminated (Figure 6.1b). The third phase consists of all subsequent exploration where the the Markov chain refines its exploration of the typical set and the precision of the Markov chain Monte Carlo estimators improves, albeit at a slower rate (Figure 6.1c).

Once the Markov chain has entered into this third phase the Markov chain Monte Carlo estimators satisfy a Central Limit Theorem

$$\hat{f}_N^{\mathrm{MCMC}} \sim \mathcal{N}(\mathbb{E}_\pi[f], \mathrm{MCMCSE}),$$

where the *Markov Chain Monte Carlo Standard Error* is given by

$$\mathrm{MCMCSE} \equiv \sqrt{\frac{\mathrm{Var}[f]}{\mathrm{ESS}}}.$$

Here the *effective sample size* is defined as

$$\mathrm{ESS} = \frac{N}{1 + 2\sum_{l=1}^{\infty} \rho_l},$$

where ρ_l is the lag-l autocorrelation of f over the history of the Markov chain. As in the Importance Sampling Central Limit Theorem, the effective sample size quantifies the number of exact samples necessary to give an equivalent estimator precision and hence the effective number of exact samples "contained" in the Markov chain. We can also interpret the effective sample size as the total number of sojourns the Markov chain has made through the typical set.

Because the states of the Markov chain generated during the initial convergence phase mostly bias Markov chain Monte Carlo estimators, we can achieve more precise estimators more quickly by using only those states generated once the Markov chain has begun to explore the typical set. Consequently, it is common practice to throw away those initial converging samples before computing Markov chain Monte Carlo estimators.

### 6.3.2 Markov Chain Monte Carlo Under Less-Than-Ideal Conditions

Under ideal conditions Markov chain Monte Carlo behaves very similarly to Monte Carlo, only with a loss of efficiency due to the correlation in the samples. When the target distribution exhibits more pathological behavior, however, Monte Carlo continues to perform well while Markov chain Monte Carlo begins to fail in spectacular fashion.
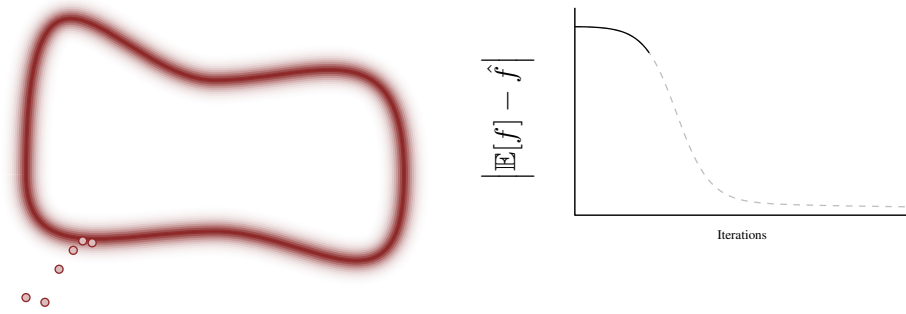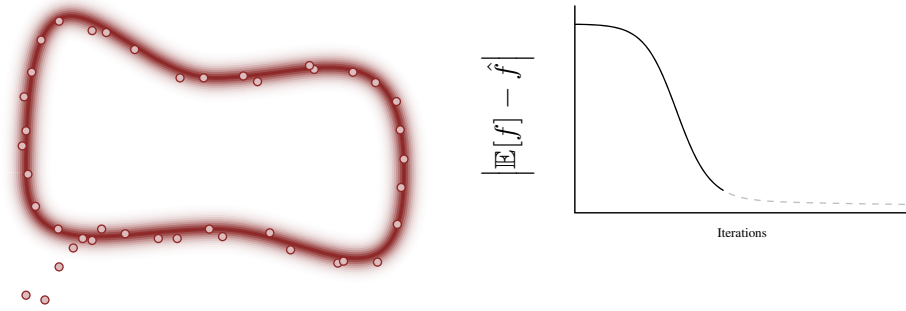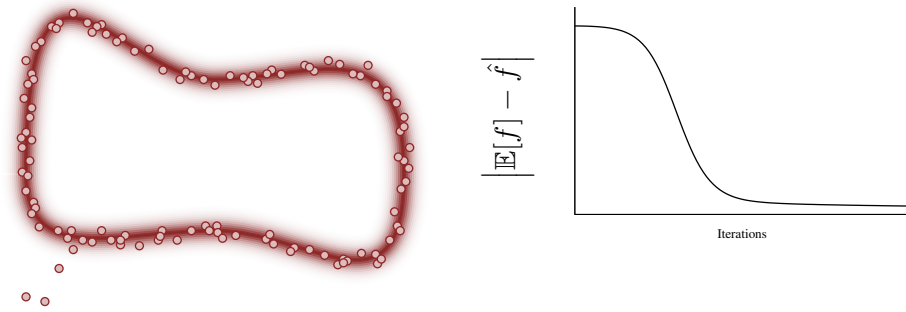
Figure 6.1: Under ideal circumstances, a Markov chain explores the target distribution in three phases. (a) First the Markov chain convergences to the typical set and estimators suffer from initial but ultimately transient biases. (b) Once the Markov chain finds the typical set and makes it first sojourn through it, this initial bias rapidly vanishes and the estimators become much more accurate. (c) As the Markov chain continues it explores more details of the typical set and gradually improves estimator precision.
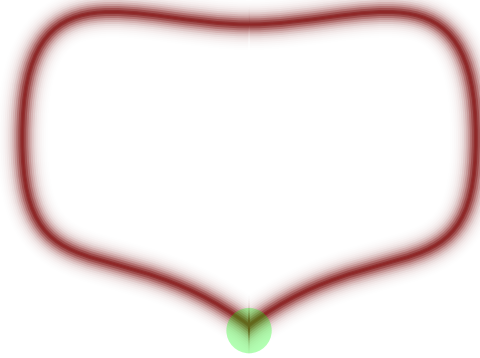
Figure 6.2: Markov chains typically have trouble exploring regions of the typical set with large curvature (green), which induces bias in Markov chain Monte Carlo estimators and spoils useful results such as Central Limit Theorems.

Consider, for example, a target probability distribution where the typical set pinches into a region of high curvature (Figure 6.2). Most Markov transitions do not have the resolution to maneuver into these tight regions and the resulting Markov chains simply ignore them, biasing subsequent Markov chain Monte Carlo estimators. It's as if there are thin but deep cracks hiding a significant amount of probability that the Markov chains pass right over and miss entirely.

Because Markov chains have to recover the exact expectations asymptotically, they have to somehow compensate for not being able to explore these regions. Typically the Markov chain accomplishes this by getting stuck near the boundary of the pathological region: as the chain hovers around the pathological region the estimators are drawn down, almost as if the Markov chain were exploring the pathological region. Eventually the Markov chain escapes to explore the rest of the typical set and the estimator bias begins to increase again (Figure 6.3).

Ultimately this behavior results in estimators that oscillate around the true expectations. Asymptotically the oscillations average out to the true values, but that balance is delicate and any finite time estimator will suffer from substantial biases.

Whether or not features of the target distribution become pathological depends on how the Markov transition kernel interacts with the target distribution. Some transition kernels are more robust than others and some can achieve robust performance with careful tuning of auxiliary kernel parameters.

6.3.3   Markov Chain Monte Carlo in Practice

In order to guarantee that we will not suffer from pathological behavior we have to demonstrate strong *ergodicity* conditions that ensure the Markov chain not only explores the typical set but does so sufficiently fast without being hampered by pathological neighborhoods.
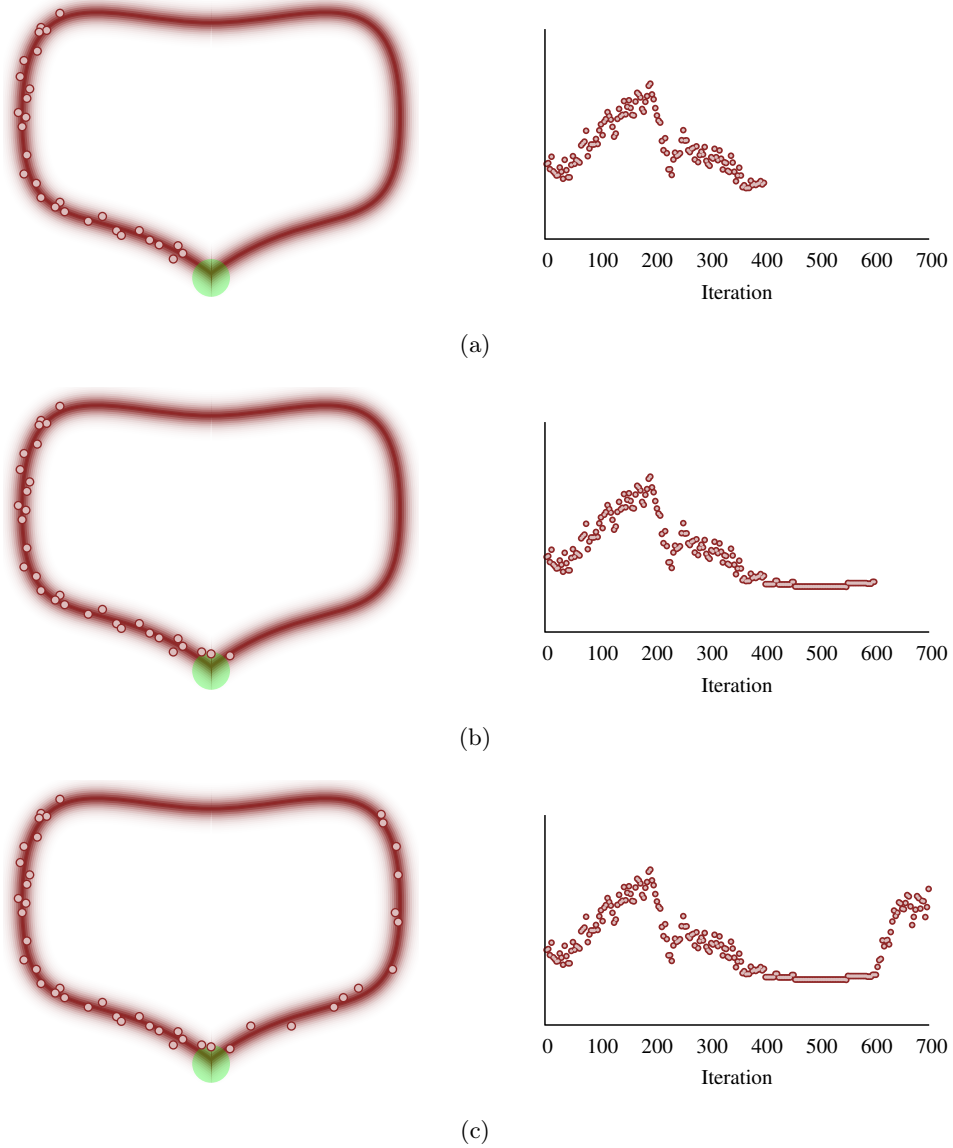
Figure 6.3: In practice, pathological regions of the typical set usually cause Markov chains to get "stuck". (a) Initially the Markov chain explores well-behaved regions of the typical set, avoiding the pathological neighborhood entirely and biasing Markov chain Monte Carlo estimators. (b) If the Markov chain is run long enough then it will get stuck near the boundary of the pathological region, slowly correcting the Markov chain Monte Carlo estimators. (c) Eventually the Markov chain escapes and explores the rest of the typical set. This process repeats, causing the resulting estimators to oscillate around the true expectations in an unstable fashion.

In most cases we need to establish a condition called *geometric ergodicity*.

Although we can identify generic features that often prevent geometric ergodicity, determining whether or not a particular Markov chain will exhibit pathological behavior when targeting a particular distribution is almost always infeasible for nontrivial problems. Moreover, there are no sufficient conditions that we can use to establish geometric ergodicity empirically. Instead we have to rely on an assortment of necessary conditions to provide some evidence that we can trust the resulting Markov chain Monte Carlo estimators.

Consequently we have to take great care when implementing Markov chain Monte Carlo, and to maximize robust performance we proceed in three stages.

Warmup

We begin with *warmup*, where we initialize multiple chains from multiple, ideally diffuse, points in the sample space, and then run long enough for each of them to converge to the typical set. Because we do not include these warmup samples in any Markov chain Monte Carlo estimators, we can also use this period to adaptively tune any parameters in the Markov kernel without biasing our estimates.

Historically this stage has usually been called *burn-in*, but we find that terminology inappropriate for Markov chain Monte Carlo. The problem is that burn-in is a process of stress-testing a system to identify and replace any failing components. In Markov chain Monte Carlo, however, any misbehaving chains identify pathological behavior that is biasing *all* of the chains and should very much not be ignored! Because of this potentially-confusing, false analogy we prefer the term warmup.

Sampling

Once warmup has finished we begin a sampling phase where we run the Markov chain and save all of the resulting samples to construct Markov chain Monte Carlo estimators.

Evaluation

Once both warmup and sampling have completed we can search for any signs of pathological behavior and, if we can't find any, move on to computing any desired estimator.

For what kind of pathological behavior should we be looking? If we don't run warmup long enough for all of the Markov chains to converge then not all of the Markov chains will look the same. Similarly, any pathological regions in the typical set will bias any finite-length Markov chains in different ways. Consequently a necessary condition for robust Markov chain Monte Carlo estimators is that each Markov chain appears identical. In theory we can quantify the homogeneity of our ensemble of Markov chains with the *potential scale reduction factor*, and in practice we can estimate the potential scale reduction factor with the $\hat{R}$ statistic.

In addition to $\hat{R}$, specific Markov transitions may admit their own, unique diagnostics sensitive to various pathologies that can frustrate geometric ergodicity.

If we are confident that our Markov chains are exploring without obstruction then we can finally compute Markov chain Monte Carlo estimators using the samples generated in the sampling phase. If we also estimate the variances and autocorrelations of each function then we can also quantify the error of these estimates using an estimate of the Markov chain Monte Carlo Standard Error.

7 | Bayesian Inference

With the theoretical and practical aspects of probability theory at our disposal, we are finally ready to formally define concepts like measurement, inference, and decision making. Here we consider a Bayesian approach to these ideas, although the same foundation is also critical for constructing a rigorous frequentist approach to inference as well.

7.1 Modeling Measurements

Without measurements we would not have nothing from which to learn, hence we must begin our discussion by defining measurements and then mathematical models of those measurements.

7.1.1 The Data Generating Process

The basic assumption underlying inference is that there is some observable process that we would like to understand, or at least some latent process that has observable consequences.

These observable consequences manifest as logical propositions, but in practice we cannot observe those propositions exactly and are instead limited to only variable measurements of the propositions. In order to formalize these concepts we assume that this variability is sufficiently well-behaved that we can model it with probability theory. More formally, we assume that the process under consideration defines a probability distribution, π_D over some measurement space, D, with measurements defined as events in the corresponding event space.

Although we have assumed the existence of a *data generating process*, π_D, we have intentionally not assumed any philosophical interpretation of it. In particular, we are indifferent to the ultimate source of the variability in the measurements quantified by π_D: it could be some ontological variability inherent to the system or just some epistemological variability due to our ignorance of the underlying system. The only assumption we have made is that the measurements are repeatable and variable, and that this variability is sufficiently well-behaved to be quantified by probability theory.

We can assume the existence of a data generating process, but we don't know anything about it until we start making measurements. An infinite number of measurements would
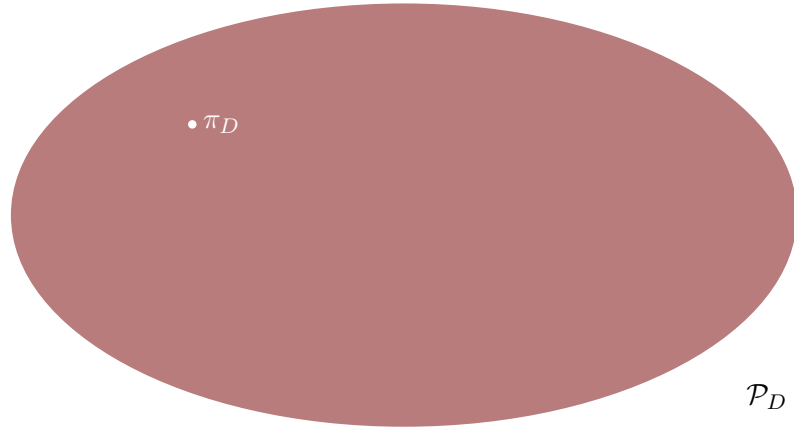
Figure 7.1: Once we have defined a measurement space, D, the latent data generating process, π_D, must be contained in the space of all possible data generating processes over D, $\mathcal{P}_D$.

certainly inform us of the data generating process exactly, but measurements are expensive and in practice we have to learn about the data generating process from only a few measurements, if not just a single measurement. *Inference* is the process of learning about the data generating process using only a finite number measurements.

7.1.2 Big Worlds and Small Worlds

If we want to learn about the data generating process then we have to consider all possible data generating process that we could encounter or, equivalently, all possible probability distributions over the sample space, D. We refer to this massive set, $\mathcal{P}_D$ as the *the big world* (Figure 7.1).

The big world is much too ungainly to be even well-defined in practice, let alone exhaustively explored. Instead we have to limit our consideration to only a subset of probability distributions over the measurement space called a *small world*, $\mathcal{S} \subset \mathcal{P}_D$ (Figure 7.2a). A representation of the small world is called a *parameterization*.

Each point in the small world, $s \in \mathcal{S}$, identifies a unique probability distribution over data. Consequently the small world defines a collection of probability distribution over the measurement space,

$$L : \mathcal{E}(D) \times S \to [0, 1]$$
$$(E_D, s) \mapsto \mathbb{L}(E_D; s).$$

known as the *likelihood*.

Regardless of how it is chosen, the assumption of any specific small world can have drastic limitations on inference. Because any small world is only a shallow approximation of
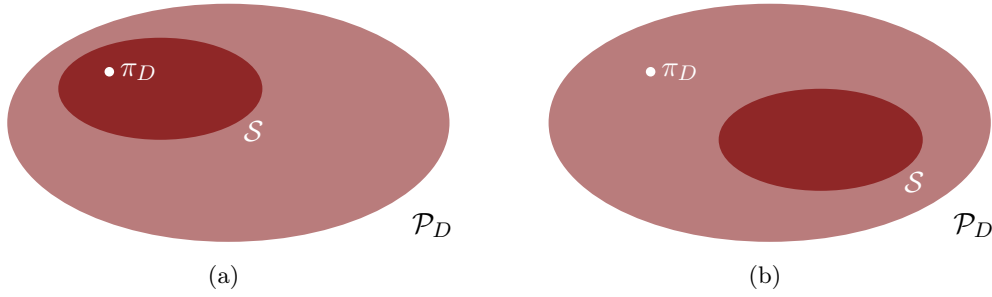
Figure 7.2: Practical inference requires the selection of a distinguished subset of data generating processes called a small world, $\mathcal{S}$, that (a) may or (b) may not contain the latent data generating process, π_D. The Boxian philosophy of "all models are wrong but some are useful" asserts that the former is impossible in practical problems, but even in the latter case the probability distributions in the small world may provide useful approximations of π_D.

reality, for example, it is unlikely to contain the latent data generating process (Figure 7.2b). Consequently even ideal inferences are subject to error, and the utility of any inference will always depend on the viability of our assumptions.

7.2   Uncertainty And Learning In The Small World

Any from of inference uses measurements to learn about the small world, particular to inform which elements of the small world yield good approximations to the data generating process, π_D. In Bayesian inference we make second use of probability theory – in addition to quantifying the variability in the measurements, we also use it to quantify our uncertainty about propositions describing which elements of the small world are good approximations to the latent data generating process.

Assuming a representation of the small world, $(\Theta, \mathcal{E}(\Theta))$, from this perspective the likelihood is not just a collection of probability distributions over the measurement space but a proper conditional probability distribution, $\pi_{D|\Theta}^{\text{like}}$. In other words, the likelihood provides an implicative structure for how uncertainty in the small world affects uncertainty in the measurement space. In order to learn we want to invert this implicative structure and have measurements inform the small world. Fortunately, this is exactly what Bayes' Theorem provides!

In order to apply Bayes' Theorem we need to construct a *prior distribution*, $\pi_\Theta^{\text{prior}}$, over the small world that quantifies our initial uncertainty about propositions describing which elements are most consistent with the latent data generating process. The information embodied by the prior distribution can be motivated from previous measurements, theoretical constraints, or even elicitation of experts.

The prior and likelihood together define a joint distribution over the measurement space and the sample space, which we can marginalize onto the measurement space to produce a *marginal likelihood*,

$$\pi_D^{\mathrm{ml}}(E_D) = \mathbb{E}_{\pi_\Theta^{\mathrm{prior}}}\left[\pi_{D|\Theta}^{\mathrm{like}}(E_D \mid \theta)\right],$$

or, in terms of probability density functions,

$$p_D^{\mathrm{ml}}(d) = \int p_{D|\Theta}^{\mathrm{like}}(d \mid \theta)\, p_\Theta^{\mathrm{prior}}(\theta)\, \mathrm{d}^D\theta.$$

Bayes' Theorem then implements learning in the small world by updating the prior distribution to a *posterior distribution*,

$$\pi_\Theta^{\mathrm{post}} = \frac{\mathrm{d}\pi_{D|\Theta}^{\mathrm{like}}}{\mathrm{d}\pi_D^{\mathrm{ml}}}\pi_\Theta^{\mathrm{prior}},$$

or, in terms of probability density functions,

$$
\begin{aligned}
p_\Theta^{\mathrm{post}}(\theta \mid d) &= \frac{p_{D|\Theta}^{\mathrm{like}}(d \mid \theta)}{p_D^{\mathrm{ml}}(d)}\pi_\Theta^{\mathrm{prior}}(\theta) \\
&= \frac{p_{D|\Theta}^{\mathrm{like}}(d \mid \theta)}{\int p_{D|\Theta}^{\mathrm{like}}(d \mid \theta)\, p_\Theta^{\mathrm{prior}}(\theta)\, \mathrm{d}^D\theta}\, p_\Theta^{\mathrm{prior}}(\theta)\,.
\end{aligned}
$$

Because Bayes' Theorem is just a mathematical consequence of probability theory, its appearance is an inevitability once we apply probabilities to quantify our uncertainty about the small world. Ultimately, all of this abstraction is just a means to formalize the intuition that *what we know after a measurement is what we knew before the measurement plus what we learned from the measurement.*

7.3   Decision Making in the Small World

The posterior distribution not only quantifies what we have learned about the small world, it allows us to use that information to make robust decisions.

Let's say that we have a set of actions or interventions, $\mathcal{A} = \{A_i\}$. A *utility function* quantifies the utility of each action assuming that data are generated from a particular element of the small world,

$$
\begin{aligned}
U : \mathcal{A} \times \Theta &\to \mathbb{R} \\
(A, \theta) &\mapsto U(A, \theta).
\end{aligned}
$$

Given a particular element of the small world, $\theta^* \in \Theta$, the best decision is the action that yields the maximum utility,

$$\hat{A} = \underset{A \in \mathcal{A}}{\operatorname{argmax}} U(A, \theta^*)\,.$$
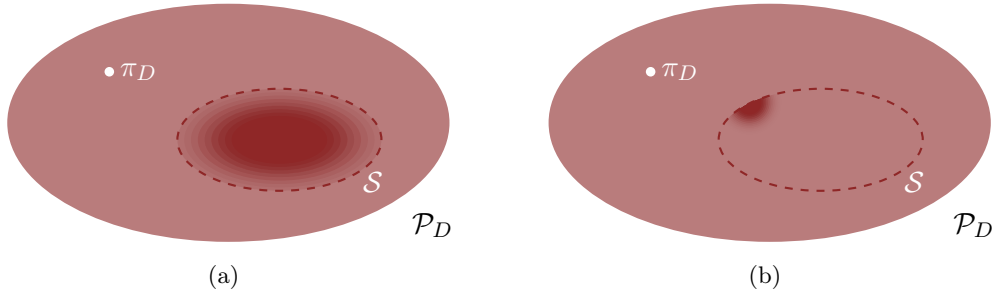
Figure 7.3: Inference in the small world is the process of updating (a) a prior distribution quantifying our initial uncertainty about the small world into (b) a posterior distribution quantifying our uncertainty about the small world after incorporating any information in a measurement. If all of our assumptions are viable then the posterior should concentrate towards the latent data generating process, π_D.

In practice, however, we don't know which element of the small world is true. We could try to identify the element of the small world that best approximates the latent data generating process according to some metric, but that would ignore our uncertainty. In particular, elements of the small world that yield similarly good approximations of the latent data generating process can yield completely different decisions, and decisions based on point estimates in the small world can be very fragile.

We can naturally incorporate our uncertainty into the decision making process, however, by averaging the utility function over the entire small world to give the *expected utility function*,

$$\overline{U} : \mathcal{A} \to \mathbb{R}$$
$$A \mapsto \mathbb{E}_{\pi_\Theta^{\text{post}}}[U(A, \cdot)].$$

We can now make our decision based on the largest expected utility,

$$\hat{A} = \underset{A \in \mathcal{A}}{\text{argmax}}\, \overline{U}(A).$$

Because this incorporates all elements of the small world into the decision making process, it leads to substantially more robust decisions in the presence of uncertainty.

7.4   Checking Model Assumptions with Predictive Performance

As with any form of inference, the efficacy of Bayesian inference depends critically on the underlying assumption of a particular likelihood and prior distribution. For example, our model might not be complex enough to capture the detail of the latent data generating
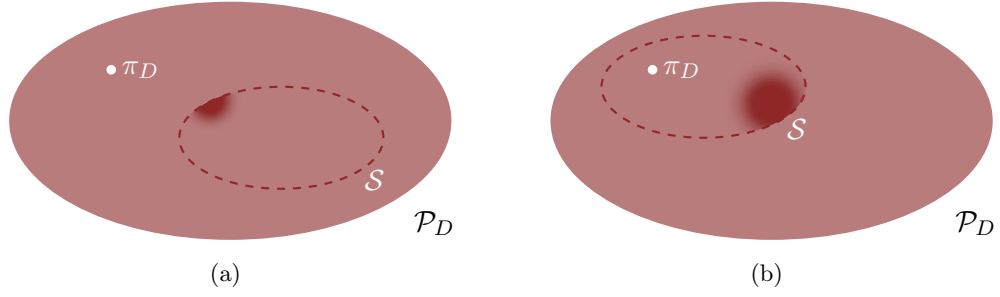
Figure 7.4: Bayesian inference yields optimal learning only within the context of our model assumptions. (a) Misfit occurs when the small world does not contain the latent data generating process and so can never recover it exactly. (b) Even if the small world contains the latent data generating process, it can still overfit by concentrating away from the latent data generating process.

process, leading to *misfit* (Figure 7.4a). On the other hand, our model might also be too complex causing the posterior to be led away from the latent data generating process in *overfitting* (Figure 7.4b).

Because Bayesian inference is a self-consistent framework, we cannot validate those assumptions from a purely Bayesian perspective! In order to check the validity of our modeling assumptions we have to go beyond that context, penetrating the small world and peering into the context of the big world, at least conceptually. Fortunately, this foray into the big world is naturally guided by *predictive performance*.

7.4.1 The Posterior Predictive Distribution

Because we assume that whatever process generates the measurements is stationary, if an element of the small world identifies a distribution that generated our measurements then it should also the *predictive distribution* that that generates new measurements.

A single element of the small world, however, ignores our uncertainty about which elements of the small world are most consistent with the measurements. To fully incorporate our uncertainty we want to consider not a single element of the small world, but rather integrate over its entirety using the posterior distribution.

The *posterior predictive distribution* over new measurements, $\pi^{pp}_{\widetilde{D}|D}$, is constructed in the same way as the marginal likelihood. The posterior distribution and the likelihood together define a condition distribution over the space of new measurements and the small world, and marginalizing over the small world gives a predictive distribution over the space of new measurements,

$$\pi^{pp}_{\widetilde{D}|D}\big(E_{\widetilde{D}}\big) = \mathbb{E}_{\pi^{\text{post}}_{\Theta|D}}\Big[\pi^{\text{like}}_{\widetilde{D}|\Theta}\big(E_{\widetilde{D}} \mid \theta\big)\Big],$$

or, in terms of probability density functions,

$$p^{pp}_{\tilde{D}|D}(\tilde{d} \mid d) = \int p^{\mathrm{like}}_{\tilde{D}|\Theta}(\tilde{d} \mid \theta)\, p^{\mathrm{post}}_{\Theta|D}(\theta \mid d)\, \mathrm{d}^D\theta.$$

In practice we can rarely construct the posterior predictive distribution analytically, but we can readily generate *samples* from it if we are already generating samples from the posterior distribution. Formally, if

$$\{\theta_n\} \sim \pi^{\mathrm{post}}_\Theta$$

is a sequence of samples from the posterior distribution then we can construct a corresponding series of samples from the posterior predictive distribution by sampling from the respective likelihoods,

$$\tilde{d}_n \sim \pi^{\mathrm{like}}_{\tilde{D}|\Theta}.$$

Our model assumptions are valid if and only if the posterior predictive distribution is a sufficiently reasonable approximation to the latent data generating process. We just need a way to quantify how close they are.

7.4.2 Predictive Performance

As we saw in the discussion of variational methods, the natural way to compare distributions is with a divergence, and the natural divergence for comparing probability distributions is the *Kullback-Leibler divergence*,

$$\mathrm{KL}(\pi_1 \mid\mid \pi_2) = \mathbb{E}_{\pi_1}\left[\frac{\mathrm{d}\pi_1}{\mathrm{d}\pi_2}\right],$$

or, in terms of densities,

$$\mathrm{KL}(\pi_1 \mid\mid \pi_2) = \int \frac{p_1(\theta)}{p_2(\theta)} p_1(\theta)\, \mathrm{d}\theta.$$

When the two distributions are the same the divergence vanishes, and as the two distributions deviate the divergence grows.

The Kullback-Leibler divergence between the posterior predictive distribution and the latent data generating process, $\mathrm{KL}(\pi_D \mid\mid p^{pp}_{\tilde{D}|D})$ defines a measure of *predictive performance* which serves as an expected utility function, allowing us to rank all possible sets of assumptions *relative* to each other. Those models with a posterior predictive distribution deviating only slightly from the latent data generating process yield better fits, while those that yield larger divergences yield worse fits (Figure 7.5). What the predictive performance does not do is provide an *absolute* validation of our model assumptions independent of other possible assumptions. An absolute validation would require that we impose a threshold on the divergence below which models are considered valid and above which invalid and, in practice, there is no real motivation for setting such a threshold. This leaves absolute statements as largely ill-posed.
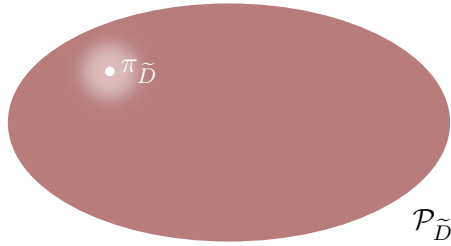
Figure 7.5: The Kullback-Leibler divergence between the latent data generating process and the posterior predictive distribution of a Bayesian model defines a measure of predictive performance which allows us to rank the performance of various model assumptions relative to each other. The smaller the divergence the closer the predictive distribution is to the latent data generating process.

The Kullback-Leibler divergence is an ideal way to compare modeling assumptions in theory, but in practice it is of limited use because we can't compute expectations with respect to the latent data generation process unless we already know it! Consequently a variety of approximations have been developed through the years that estimate the divergence by using a measurement at least twice: once to infer a posterior and again to estimate the divergence. Depending on the exact approach taken, these approximations reduce to many popular model comparison techniques ubiquitous in the literature, including cross validation and the many information criteria.

7.4.3 Posterior Predictive Checks

The predictive performance quantified by the Kullback-Leibler divergence also manifests in visual diagnostics known as *posterior predictive checks*, which formalize the intuitive yet powerful concept of residual analysis.

A posterior predictive check begins by identifying a low-dimensional component of the measurement space, D^*, that is easy to visualize and ideally sensitive to the most suspect model assumptions. We then construct the posterior predictive check by visually comparing the marginalized posterior predictive distribution over D^* to just the measurement itself. If the measurement is far from the predictive distribution then either the measure itself was exceedingly unlikely or our model is exhibiting misfit (Figure 7.6).

We can also construct a *jackknife posterior predictive check* by splitting our measurement into two partitions. Tension between the marginalized predictive distribution constructed from one partition and the second partition indicates either an unlikely measurement or the overfit of the underlying model (Figure 7.7).

As in the general case, these visual predictive checks are not calibrated and so we cannot discriminate between rare measurements and the illegitimacy of our model assumptions. In practice, however, we can appeal to some implicit baseline predictive performance to make
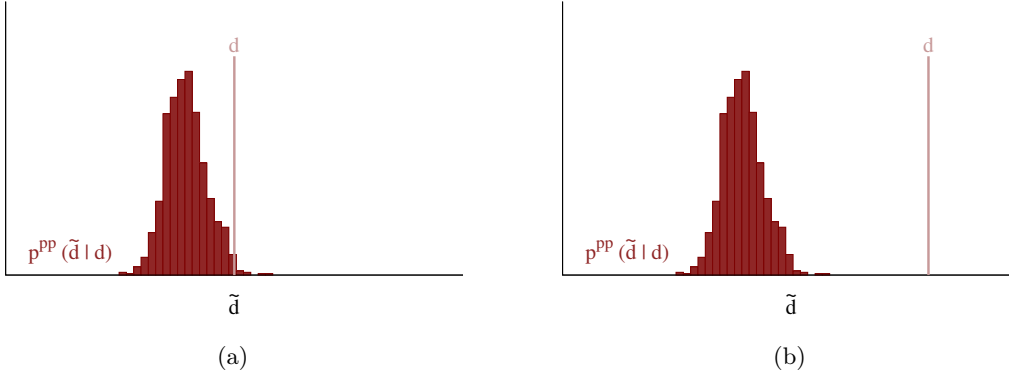
Figure 7.6: (a) When the measurement, d, is consistent with the posterior predictive distribution we have no reason to doubt our model assumptions, but (b) tension between the measurement and the posterior predictive distribution indicates that there may be a problem. Either the measurement was exceedingly rare or our model assumptions are insufficient.
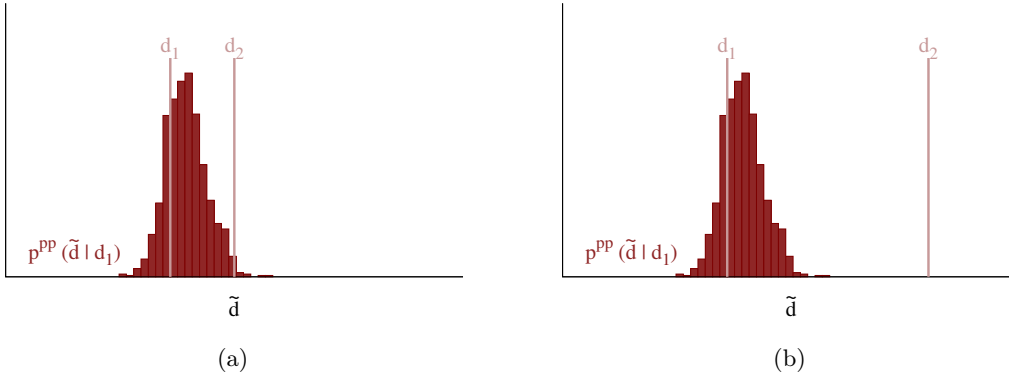

Figure 7.7: To test for potential overfit we need multiple measurements, and given only a single measurement this can emulated by splitting the measurement into two partitions, d_1 and d_2. (a) When both measurements are consistent with the posterior predictive distribution we have no reason to doubt our model assumptions, but (b) tension between the held-out measurement, d_2 and the posterior predictive distribution indicates that the model might be overfitting to d_1. Either the partition generated an exceedingly rare measurement or our model assumptions are insufficient.
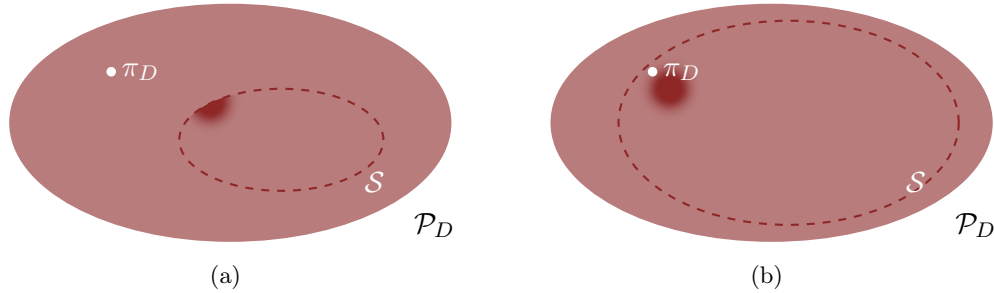
67

Figure 7.8: Posterior predictive checks are powerful ways to make qualitative judgements about the validity of our model assumptions. For example, if (a) they suggest that our model is misfitting the data then (b) we can update our model to better capture the features of the data that are being misfit.

qualitative judgements as to when the tension is strong enough to be suspect. Suspicion then motivates a careful study of the model assumptions that manifest in the chosen component of the measurement, and potentially the updating of our model into one more capable of capturing the intricacies of our data (Figure 7.8).

7.5 Bayesian Inference in Practice

Ultimately, implementing Bayesian inference in practice is surprisingly straightforward. Using our expertise about a system we construct a prior that quantifies our initial assumptions about the system and a likelihood that quantifies our assumptions about the latent data generating process itself. These two distributions then immediately define a posterior distribution quantifying our updated uncertainty about our system, and all inferential summaries and decisions reduce to posterior expectations. Although conceptually simple, neither of these steps is particularly easy.

Building robust models is an immensely challenging task that requires both expertise with a given system and familiarity with the wealth of potential techniques for constructions priors and likelihoods. Many of these techniques are discussed in the Stan manual (and potentially here someday, too), but here we'd like to highlight a particularly powerful approach to building likelihoods: *generative modeling*.

Generative modeling builds up a likelihood from latent parameters to the measurement sequentially, as if we were trying to simulate the data generating process itself. By considering the forward model we can naturally incorporate causal structure, such as a physical model, measurement variation, and systematic effects, such as bias and measurement error. At each stage we model deterministic or stochastic relationships with conditional probability distributions until we have constructed the complete likelihood. The generative nature of these models facilities communication and helps to identify which features may be most

suspect, motivating targeted posterior predictive checks and model updates.

Once we have specified our assumptions and collected our data, there many options for approximating the resulting posterior expectations in practice. We have discussed many of these here, and there are always more being developed. Whichever method we end up using, however, we have to be careful to validate the accuracy of the estimation, lest a biased or highly variable estimator spoil the carefully crafted model. Such validation is easier in some computational techniques than others, and sometimes the fastest algorithms are the most vulnerable to bias!

Although the modeling and computation steps should ideally be completely decoupled, in practice they will always interact. More complex models strain the approximation algorithms in our toolbox, exhausting our computational resources or compromising the accuracy of the estimators and, in either case, undermining the robustness of our analysis. In practice we often have to iterate, using the success or failure of a fit to inform the next generation of model. Provided we do this deliberately, taking care to validate both the model assumptions and the fit themselves, we can use Bayesian inference to make powerful statements about the world around us.