A Technical Introduction to Probability and Bayesian Inference for Stan Users

Stan Development Team

Version 0.0.0
April 10, 2016
https://github.com/betanalpha/stan_intro

Stan Development Team. 2016 *An Technical Introduction to Probability and Bayesian Inference for Stan Users.* Version 0.0.0

Contents

Preface

A ball flies across the room, slowly falling under the force of gravity. Being diligent scientists, we record the positions of the ball as best as we can. Were our measurements perfect we could exactly recover not only the trajectory taken by the ball but also any latent parameters that determined the trajectory, such as the acceleration due to gravity, g (Figure 1a). Of course in practice our measurements are not perfect. They are inherently variable there are consequently there are many trajectories, and values for g, consistent with the noisy measurements (Figure 1b). Any robust analysis must not only infer the latent model of the data generating process but also quantify the uncertainty in that inference.

This simple example demonstrates a principle ubiquitous to science, industry, medicine, and any other field that attempts to learn from measurements: 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 intrinstic 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.

Because probability theory is so subtle and counterintuitive, introductory treatments of Bayesian inference often oversimplify and neglect many of its finer technical aspects. Unfortunately, these technicalities are not irrelevant and often have a strong influence on practical applications of the theory. Without at least a conceptual understanding we are subject to dangerous fallacies and fragile analyses. In this review we attempt a deeper introduction to probability theory and Bayesian inference than usual to provide Stan users with the background necessary to properly wield Bayesian inference and take full advantage of their measurements.

After reviewing some mathematical administration we'll introduce first logic as a procedure for quantifying information and then probability theory as a procedure for quantifying uncertainty about that information. We'll consider both abstract definitions as well as 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.

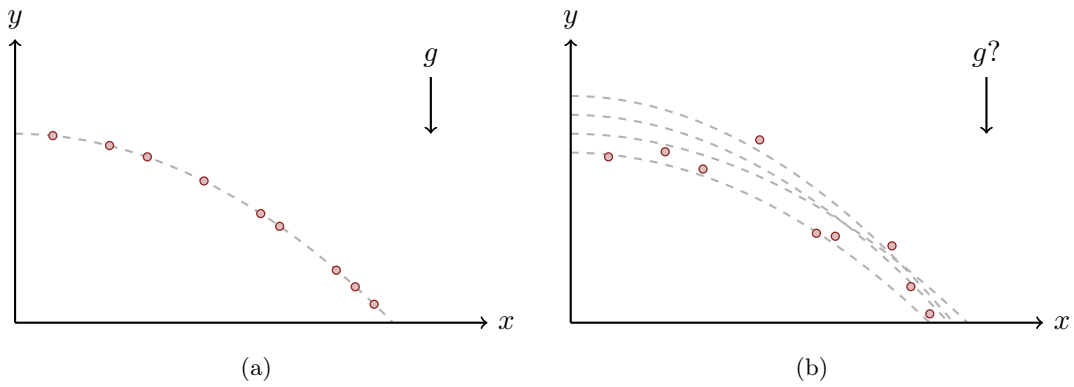


Figure 1: (a) Perfect measurements of a ball falling under the influence of gravity would allow us to exactly recover the trajectory of the ball. (b) Unfortunately, in practice measurements are inherently variable, which limits our ability to infer the exact trajectory and hence any model of how the was generated. Consequently uncertainty is intrinsic to practical learning.

0.1 Mathematical Background and Notation

Regrettably, a thorough review of probability theory requires a nontrivial mathematical background. We have attempted to make this review as self-contained as possible regarding probability theory, 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. The only important consequence of topologies that will be relevant for us is that ultimately it is this topological structure that allows us to characterize spaces as

either discrete or real.

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

$$\begin{aligned} f : A &\to B \\ a &\mapsto f(a) \,. \end{aligned}$$

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

1.1 Logical Statements

Let $\mathcal{S}$ be an abstract system that we would like to describe. Any description of $\mathcal{S}$ can be defined as the assertion of a logical statement about that system. For example, consider describing one of the eight planets in our Solar System: Mercury (☿), Venus (♀), Earth (♁), Mars (♂), Jupiter (♃), Saturn (♄), Uranus (♅), and Neptune (♆). Information about the planet is encapsulated by asserting the truth of logical statements such as "the planet has an atmosphere" or "the planet has rings". We denote the set of well-posed logical statements about $\mathcal{S}$ as $\lambda(\mathcal{S})$.

Logical statements can also be manipulated into other logical statements. A *conjunction* asserts that two statements are both true, "the planet has an atmosphere *and* rings" while a *disjunction* states that either statement is true, "the planet has an atmosphere *or* rings". We can also negate a logical statement by asserting its contrary, "the planet does *not* have rings".

1.2 Events

Logical statements provide a generic framework for describing arbitrary and abstract systems, but that abstraction also makes these statements 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 an *expression* (Table 1.1).

An expression is a map from our abstract system into a concrete space, such as the integers or the real numbers,

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

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

Table 1.1: Expressions provide a concrete representation of the abstract descriptions of a given system. Degenerate ways to express the descriptions of a system, along with the translations between them, are ubiquitous in common experience.

where the concrete space, Θ, is known as the *sample space*. An expression also defines a corresponding map from logical statements into well-behaved subsets of the sample space, of *events*, $E \subset \Theta$,

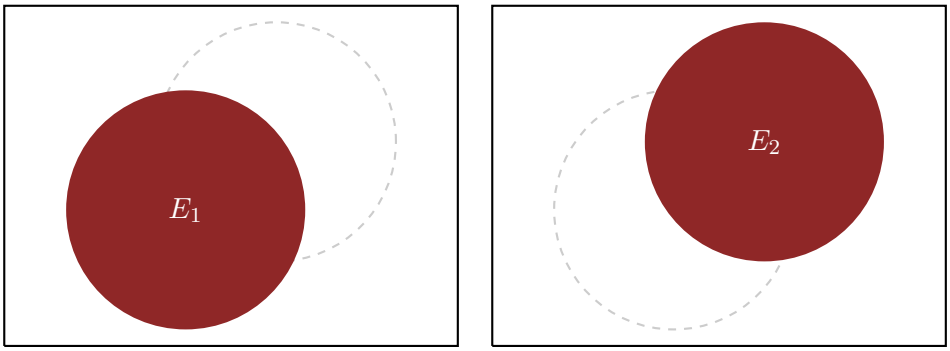
$$\mathfrak{e}^* : \lambda(\mathcal{S}) \to \mathcal{E}(\Theta) \,,$$

where the *event space*, $\mathcal{E}(\Theta)$, collects all well-behaved subsets of the sample space and hence all valid expressions of well-posed logical statements. Event spaces always include the null event, $E = \emptyset$, which represents certain falseness, and the trivial event, $E = \Theta$, which represents certain truth. The manipulation of the logical statements are implemented with in an expression with set operations. Conjunction of statements becomes the intersection of subsets, disjunction the union of subsets, and negation the complement of a subset (Figure 1.1).

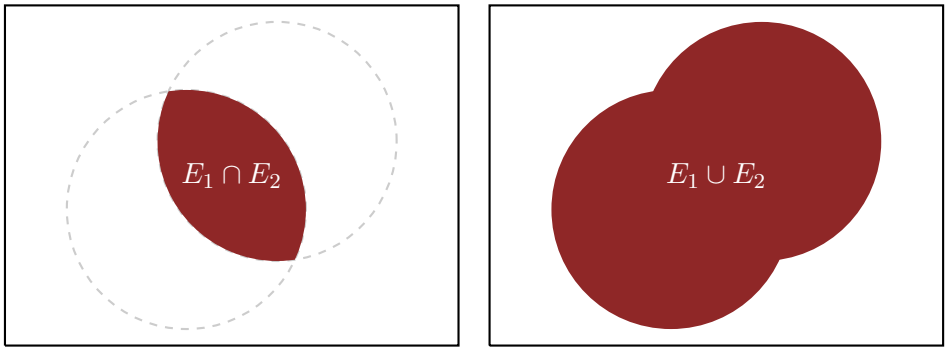
For example, our descriptions of the planets can be expressed with the integers, $\Theta = \{1, 2, \ldots, 8\}$, with each logical statement equivalent to a subset of integers (Figure 1.2). The procedure is equivalent for descriptions that can be represented by real numbers, such as the distance between two objects (Figure 1.3).

One extremely important subtlety with this construction is that expressions are not unique – we can express the same logical statements using events in different sample spaces. To see this consider an invertible map from one sample space into another, $s : \Theta \to \Omega$; if every event in Θ maps to an event in Ω and vice versa,

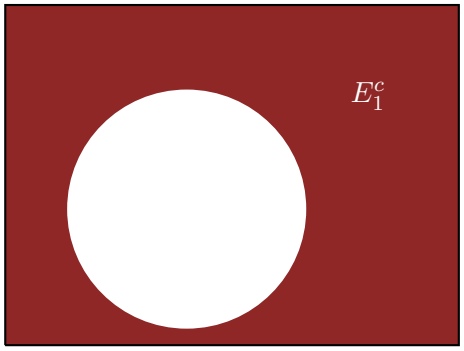
$$s(E_\Theta) \in \mathcal{E}(\Omega)$$
$$s^{-1}(E_\Omega) \in \mathcal{E}(\Theta)$$



(a)



(b) (c)



(d)

Figure 1.1: In an expression logical statements are represented by subsets and the manipulation of these statements is implemented with set operations. (a) If two statements are represented by the sets E_1 and E_2, then (b) their conjunction is the set intersection,and (c) their disjunction is the set union. (d) Simlarly, the negation of E_1 is the set complement.

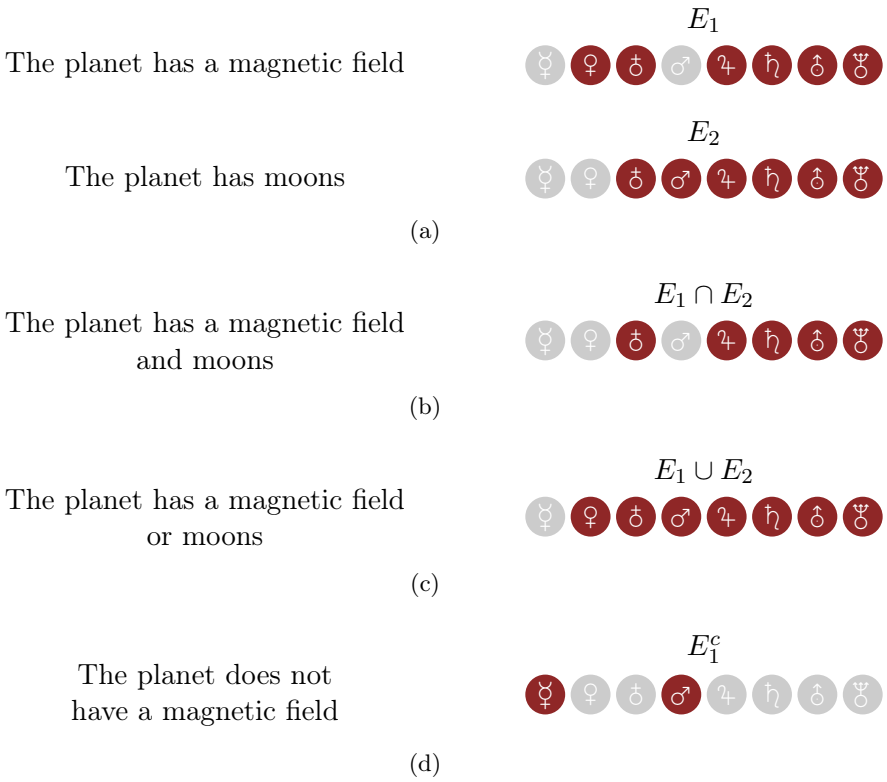


Figure 1.2: (a) Descriptions of a discrete system, such as the collection of planets in our solar system, can be expressed with by the integers, with any of the planets can be represented by a subset of the integers. (b) Conjuction of logical statements is implemented with set intersections, (c) disjunction with set unions, and (d) negation with set negation.

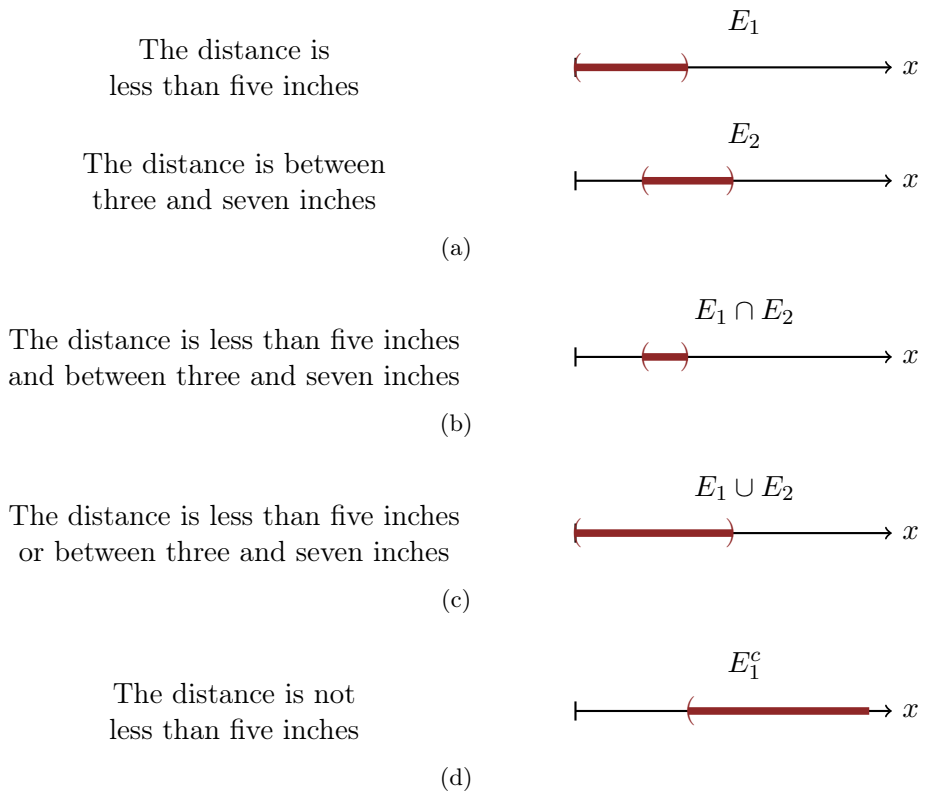


Figure 1.3: Systems whose descriptions can be described by the real numbers are expressed exactly as with discrete systems. (a) Descriptions are expressed as subsets of the real numbers, (b) conjunction with intersections of those subsets, (c) disjunction with unions of those subsets, and (e) negation with complements.

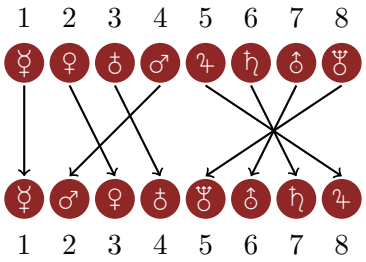


Figure 1.4: Discrete sample spaces are not unique way to express a system as we can always permute the arbitrary labels to yield an equivalent sample space. For example, we could order the planets by distance from the sun or diameter without affecting our ability to describe the planets themselves.

then the map is called *measurable*. The composition of an expression with a measurable map,

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

is another valid expression. Hence when a measurable map exists between two sample spaces they can be used to to specify the same logical statements, and we say that they are *equivalent* as they provide equivalent descriptions of the given system. The set of samples spaces equivalent to Θ is denoted $\lfloor \Theta \rfloor$. **Mention that equivalent sample spaces are typically restricted to common topologies, so we can map discrete to discrete and real to real, but not discrete to real and vice versa? Similar for cardinality of space?**

For example, if the sample space is discrete then we can always define an equivalent sample 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 sample spaces quantify the same information (Figure 1.4). Similarly, we can always apply a transformation that warps the real numbers to map between real samples spaces without compromising our ability to represent statements with 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 express the descriptions of a given system and measurable maps translate from one equivalent expression to another. Although certain expressions may be more useful than others in practice, we have to ensure that our final analysis does not depend on the irrelevant details of any particular expression.

1.3 Implications

Describing simple systems that can expressed with low-dimensional spaces is a relatively straightforward process, but reasoning about more complex systems that can be expressed with only high-dimensional spaces is typically much more subtle. One elegant way to facilitate the description of these complex systems is to decompose each statement into a composition of many simpler but related *implications*.

An implication is a logical statement conditioned on the another logical statement. For example we might not be able to assert that "a planet has a magnetic field" universally, but we can assert that "if a planet has a rotating core then it has a magnetic field". Logical statements about both the planets and their cores can always be decomposed into a statement about one of these systems and an implication from that system into the other. Consequently we can build up logical statements about complex systems by starting with a simple statement and then adding more and more implications.

If Θ is a sample space expressing the target system and Φ is a sample space expressing the conditioning systems, then an implication is map from points in Φ to events in Θ,

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

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

$$\bigcup_{\phi \in E_\Phi} \iota_{\Theta|\Phi}(\phi) \equiv E_{\Theta \times \Phi} \in \mathcal{E}(\Theta \times \Phi) \,.$$

Consequently a logical statement representable in the 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,

$$\begin{aligned} & E_{\Theta_1} \\ & \iota_{\Theta_2|\Theta_1} \\ & \iota_{\Theta_3|\Theta_2,\Theta_1} \\ & \cdots \\ & \iota_{\Theta_N|\Theta_{N-1},\ldots,\Theta_2,\Theta_1} . \end{aligned}$$

2 | Probability in Theory

Logic provides a framework for how we can describe systems with certainty, and the tools of probability theory allow us to assign uncertainty to logical statements and implications. In this section we introduce probability distributions over first logical statements and then implications.

Unfortunately, the abstraction of probability theory means that we cannot rely on helpful visualizations to support the definitions and manipulations in this section. To use a probability distribution in any practical application, including visualization, we need some means of implementing them explicitly, which will have to wait until the next section.

2.1 Probability Distributions

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

In a given expression, probabilities are designated by a *probability distribution* which assigns probability to events,

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

such that the probability of the null event is zero, $\mathbb{P}[\emptyset] = 0$, and the probability of the trivial event is one, $\mathbb{P}[\Theta] = 1$. These latter two conditions are an immediate consequence of our initial assumption that our descriptions can be expressed by the sample space. When using a probability distribution to quantify our uncertainty about a system expressed with the sample space Θ, we often write $\theta \sim \mathbb{P}$ which is read as, "θ is distributed according to the probability distribution $\mathbb{P}$" but should really read "Logical statements about Θ are assigned probabilities according to the probability distribution $\mathbb{P}$".

Probability assignments are naturally compatible with the manipulations of logical statements. For example, the probability of conjunctions and disjunction are related as

$$\mathbb{P}[E_1 \cup E_2] = \mathbb{P}[E_1] + \mathbb{P}[E_2] - \mathbb{P}[E_1 \cap E_2] \,,$$

and negations satisfy,

$$\mathbb{P}[E] = \mathbb{P}[\Theta] - \mathbb{P}[E^c] = 1 - \mathbb{P}[E^c] \,.$$

Probability distributions also allow us to compute *expectation values* of well-behaved functions on the sample space,

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

where $\mathcal{F}(\Theta)$ is the collection of well-behaved functions $f : \Theta \to \mathbb{R}$. Common expectations include means, variances, and higher-order moments. In fact, we can also consider probability assignments themselves as expectations,

$$\mathbb{P}[E] = \mathbb{E}_{\mathbb{P}}[\mathbb{I}_E] \,,$$

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

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

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!

2.2 Densities

Radon-Nikodym derivatives as functions that modify expectations.

Abusive notation of "multiplying" a distribution by a density.

2.3 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 a probability distribution to each value in the conditioning space,

$$\begin{aligned} \mathbb{P}_{\Theta|\Phi} : \mathcal{E}(\Theta) \times \Phi &\to [0, 1] \\ (E_\Theta, \phi) &\mapsto \mathbb{P}_{\Theta|\Phi}[E_\Theta \mid \phi] \,. \end{aligned}$$

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.

As with implications, conditional probability distributions can be used to construct probability distributions on more complex 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, $\mathbb{P}_{\Theta|\Phi}[E_\Theta \mid \phi]$, and then taking the expectation of this assignment over the distribution on Φ,

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

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

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

2.3.1 Marginalization

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

$$\begin{aligned}\mathbb{P}_\Theta[E_\Theta] &\equiv \mathbb{P}_{\Theta\times\Phi}[E_\Theta \times \Phi] \\ &= \mathbb{E}_{\mathbb{P}_\Phi}\left[\mathbb{P}_{\Theta|\Phi}[E_\Theta \mid \phi]\right],\end{aligned}$$

or

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

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

2.3.2 Generative Modeling

Consequently, conditional probability distributions are powerful ways of building probability distributions on high-dimensional spaces. We simply start with a probability distribution on one low-dimensional component and then build up a joint distribution by adding conditional probability distributions for each new component,

$$\begin{aligned}&\mathbb{P}_{\Theta_1} \\ &\mathbb{P}_{\Theta_2|\Theta_1} \\ &\mathbb{P}_{\Theta_3|\Theta_2,\Theta_1} \\ &\cdots \\ &\mathbb{P}_{\Theta_N|\Theta_{N-1},\ldots,\Theta_2,\Theta_1}.\end{aligned}$$

These conditional probability distributions are often motivated by the natural implication 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,

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

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

2.3.3 Bayes' Theorem

Two different decompositions of a joint distribution constrain how the corresponding conditional probability distributions are related. Formally defined in terms of densities.

In practice this allows us to invert implications. Given one conditional distribution and the two marginals, we can construct the opposite conditional distribution via Bayes' Theorem.

Math.

2.4 The Invariance of Probability Distributions

Like events, probability distributions can be defined with respect to many different sample spaces. If $s : \Theta \to \Omega$ is a measurable map and $\mathbb{P}_\Theta$ is a probability distribution defined over events in Θ, then we can define an equivalent probability distribution over events in Ω by assigning probabilities as

$$\mathbb{P}_\Omega[E_\Omega] \equiv \mathbb{P}_\Theta\left[s^{-1}(E_\Omega)\right].$$

Furthermore, this whole process can be inverted: if $\mathbb{P}_\Omega$ is a probability distribution defined over events in Ω then we can define an equivalent probability distribution over events in Θ by assigning probabilities as

$$\mathbb{P}_\Theta[E_\Theta] \equiv \mathbb{P}_\Omega[s(E_\Theta)].$$

Probability distributions are invariant when we move between equivalent sample spaces, and hence are fundamental to the underlying system and not any particular expression of that system. Different but equivalent sample spaces are just different ways to describe the same system, with events quantifying the same, invariant information and probability distributions quantifying the same, invariant uncertainty.

3 | Probability in Practice: Implementing Distributions

The immediate problem with the abstract definitions introduced above is that they do not provide an explicit means of 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 is discrete or a subset of the real numbers.

3.1 Implementations of Probability Distributions over Discrete Sample Spaces

When the sample space is discrete we can completely specify a probability distribution by assigning probability to only a small and manageable set of events. Two particularly convenient sets, point events and interval events, allow us to implement probability distributions with probability mass functions and cumulative distribution functions, respectively.

3.1.1 Probability Mass Functions

Probability mass functions assign probably to point events, those events that are lone elements of the original sample space. Hence a probability mass function is a just function that assigns a probability to each element of the sample space,

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

In this case more general event probabilities are given by simply summing the probability of each element in event,

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

Similarly, expectations are given by summing the probability of each element of the sample

space, weighted by the function value,

$$\mathbb{E}[f] = \sum_{\theta \in \Theta} f(\theta)\, p(\theta)\,,$$

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

Probability mass functions also have the convenient property that they are invariant to the particular choice of sample space. Given a measurable map $s : \Theta \to \Omega$ and a probability mass function on Θ, we can define an equivalent probability mass function on Ω as

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

3.1.2 Conditional Probability Mass Functions

Probability mass functions can be extended to implement conditional probability distributions by simply adding a conditioning variable,

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

with conditional probabilities and conditional expectations computed as above,

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

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, for example, we can construct an explicit joint probability 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,

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

3.1.3 Cumulative Distribution Functions

When the sample space is not only discrete but also ordered then we can also completely specify a probability distribution by assigning probabilities 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,

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

is called the *cumulative distribution function*.

As with the probability mass function, cumulative distributions functions immediately map between sample spaces. For a measureable map $s : \Theta \to \Omega$ we have

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

3.1.4 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

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

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,

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

3.2 Implementations of Probability Distributions over the Real Numbers

When the sample space is the D-dimensional real numbers, or a subset thereof, there is an uncountably infinite number of point events. Not only can we no longer assign a non-zero probability to each point event without having most event probabilities explode,

$$\mathbb{P}[E] \to \infty,$$

we can't even define the sums over the sample space necessary to compute probabilities and expectations!

Instead of assigning to each point a probability we have to assign to each point event a *probability density* which we can *integrate* to give probabilities and expectations.

Assigning probabilities to intervals, however, is still sufficient so we can also define cumulative distribution functions on these spaces.

3.2.1 Probability Density Functions

A *probability density function* assigns a positive value to each point in the sample space

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

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

$$\mathbb{P}[E] = \int_{E} p(\theta) \, \mathrm{d}\theta,$$

and expectations,

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

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.

The most awkward properties of probability density functions is that, unlike probability mass functions, they do not trivially transform between sample spaces. Specifically, for the measurable map $s : \Theta \to \Omega$

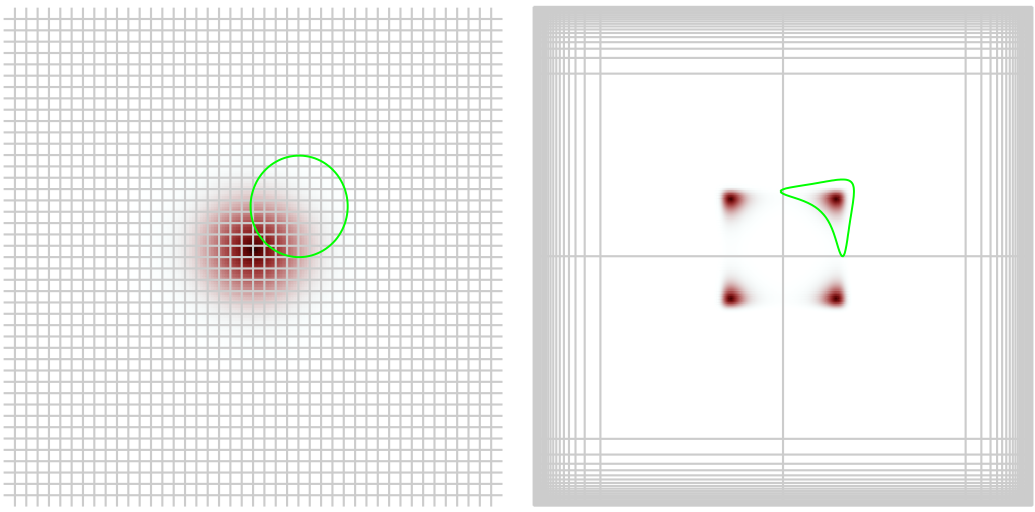
$$p_{\Omega}(\omega) \neq p_{\Theta}\big(s^{-1}(\omega)\big)!$$

The problem with the real numbers is that mapping between sample spaces transforms not only the event space but also how we differentiate and integrate. Under a well-behaved map $s : \Theta \to \Omega$ the corresponding differential volumes are related by

$$\mathrm{d}\omega = |\mathbf{J}| \, \mathrm{d}\theta,$$

where the matrix

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



$\mathbb{P}[E_\Theta] = \int_{E_\Theta} p_\Theta(\theta_1, \theta_2) \mathrm{d}\theta_1 \mathrm{d}\theta_2$

(a)

$\mathbb{P}[s(E_\Theta)] = \int_{s(E_\Theta)} p_\Omega(\omega_1, \omega_2) \mathrm{d}\omega_1 \mathrm{d}\omega_2$

(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 values, here $\mathbb{P}[E_\Theta] = \mathbb{P}[s(E_\Theta)]$.

is called the *Jacobian* of the transformation. Consequently all integrals are invariant to the particular sample space if and only if the probability density functions are related by

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

Each sample space has its own differential volume and probability density function but *the same integrals* and, hence, the same probabilities and expectations (Figure 3.1). This dependence on the sample space is another reason to be careful not to take a probability density function in isolation too seriously.

A helpful mnemonic for the arrangement of the Jacobian in these transformations is to remember that the integrand must be invariant,

$$\begin{aligned} p_\Omega(\omega) \, \mathrm{d}\omega &= p_\Theta(\theta) \, \mathrm{d}\theta \\ p_\Omega(\omega) &= p_\Theta(\theta) \, \frac{\mathrm{d}\theta}{\mathrm{d}\omega} \\ p_\Omega(\omega) &= p_\Theta(\theta) \, |\mathbf{J}|^{-1} . \end{aligned}$$

Some more text here. Jacobians arise naturally when the sample space is *constrained* and we want to map to an unconstrained space to remove the constraint and potentially simplify calculations. For example, let P be a probability distribution on $\mathbb{R}^+$ with density p. We can unconstrained the positivity constraint by applying a logarithmic map

$$map$$

which results in the Jacobian

$$jacobian.$$

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

$$map$$

which results in the Jacobian

$$jacobian.$$

To see how important the Jacobian can be, consider a two-dimensional sample space with real components (θ_1, θ_2) and a probability distribution represented with a Gaussian probability density,

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

We can then introduce a second sample space with the map

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

where the component maps are given by

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

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.

3.2.2 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 computed as integrals,

$$\begin{aligned}
\mathbb{P}_{\Theta|\Phi}[E \mid \phi] &= \int_E p_{\Theta|\Phi}(\theta \mid \phi) \, \mathrm{d}\theta, \\
\mathbb{E}_{\mathbb{P}_{\Theta|\Phi}}[f \mid \phi] &= \int_\Theta f(\theta) \, p_{\Theta|\Phi}(\theta \mid \phi) \, \mathrm{d}\theta,
\end{aligned}$$

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,

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

Cumulative Distribution Functions

Because the real numbers are sufficiently well-ordered, we can also specify probability distributions over these spaces by assigning probability to intervals using a cumulative distribution function,

$$\begin{aligned}
P : \Theta &\to \mathcal{I}(\Theta) \to [0, 1] \\
\theta &\mapsto I(\theta) \;\; \mapsto \mathbb{P}[I(\theta)] \,,
\end{aligned}$$

where each interval, $I(\theta) \in \mathcal{I}(\Theta)$, is defined as before,

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

Unlike probability density functions, and similar to discrete cumulative distribution functions, real cumulative distributions functions immediately map between sample spaces,

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

for a measurable map $s : \Theta \to \Omega$.

3.2.3 Relating Probability Mass Functions and Cumulative Distribution Functions

On the real numbers probability density functions and cumulative distribution functions are also equivalent and can be mapped into each other. Cumulative distribution functions, for example, are given by integrating over probability density functions,

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

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

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

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

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

3.2.4 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, $\mathbb{P}_{\Phi|\Psi}$, and providing a marginal distribution over the real component, $\mathbb{P}_\Psi$. 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

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

with expectations given similarly by

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

Equivalently, we could also condition the real component on the discrete component, $\mathbb{P}_{\Psi|\Phi}$ and provide a marginal distribution over the discrete component, $\mathbb{P}_\Phi$. 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

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

with expectations given similarly by

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

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.

3.3 Stochastic Implementations of Probability Distributions

We can also implement expectations of any probability distribution, including both discrete and real probability distributions, *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

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

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

This procedure, and hence the resulting samples, are *exact* if every state in the sequence is generated independent 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 above, 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, either 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 advantage of 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 know is that, because those neighborhoods define expectations, they should be equivalent

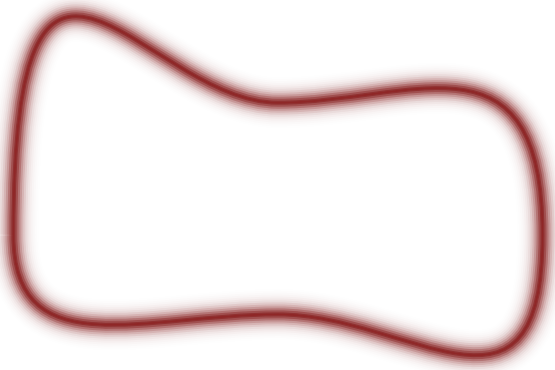


Figure 4.1: On high-dimensional samples spaces all well-behaved probability distributions concentrate in a neighborhood called the *typical set*. In order to estimate the integrals needed to compute probabilities and expectations we have to be able to identify where the typical set lies in the sample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

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. This neighborhood, however, is dependent on the particular sample space and hence doesn't have the necessary invariance properties. Our 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 volume over which we integrate does not. As we move towards infinity there is more and more room 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 the middle denoted 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 sample space we use to express that distribution.

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 many expectations we need to quantify the entire typical set.

Moreover, because points outside of the typical set contribute little to nothing 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 expression, 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,

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

with the corresponding differential volume,

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

In order to identify which neighborhoods of the sample space contribute most to generic expectations, we now transform to a sample space with spherical coordinates. This yields the probability density function

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

with the corresponding differential volume,

$$\mathrm{d}V = r^{D-1}\mathrm{d}r \prod_{d=1}^{D-1} \sin^{D-d-1}(\phi_d)\, \mathrm{d}\phi_d.$$

Because the probability density function does not depend on any of the hyperspherical angles, neither will any probabilities and, consequently, any neighborhood of high probability must be spherically symmetric. To see where these neighborhoods concentrate radially we

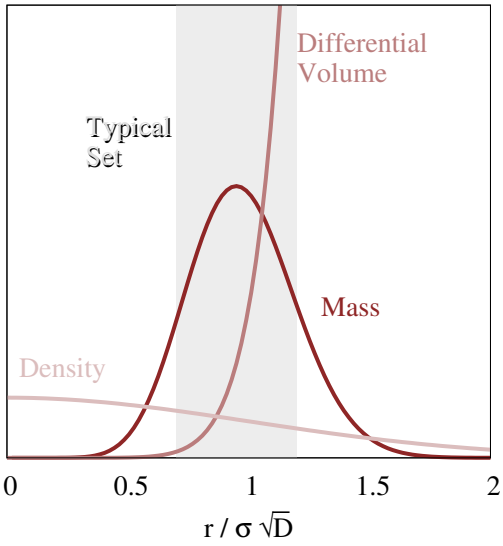


Figure 4.2: In high dimensions real probability distributions generically assign almost all of their probability into a singular neighborhood known as the typical set. This is apparent even from a probability density function representation: although the density concentrates around the corresponding 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.

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

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

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 around $\sigma/\sqrt{2}$. 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 concentrates tighter and tighter as we add more and more dimensions (Figure 4.2).

Because samples recover all expectations, they 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 can generate a sample from our target probability using a univariate Gaussian random number generator available in any computing library,

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

with the corresponding radial distance $r = \sqrt{\sum_{d=1}^{D} x_d^2}$. Generating a sequence of samples

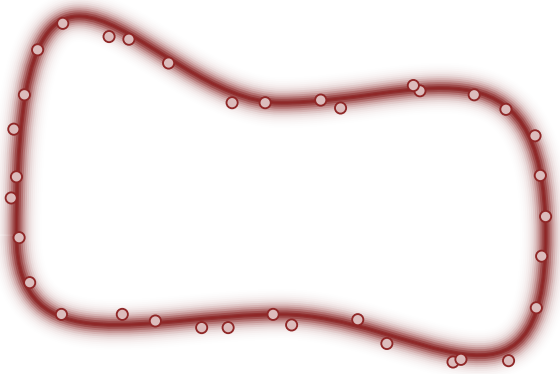


Figure 4.3: Because samples recover expectations asymptotically, large sequences of samples must concentrate in the typical set. This provides a means of visualizing concentration of measure, and will prove a powerful way to estimate expectations.

and then histogramming the radial distance reveals the same χ distribution that we arrived at analytically (Figure 4.4).

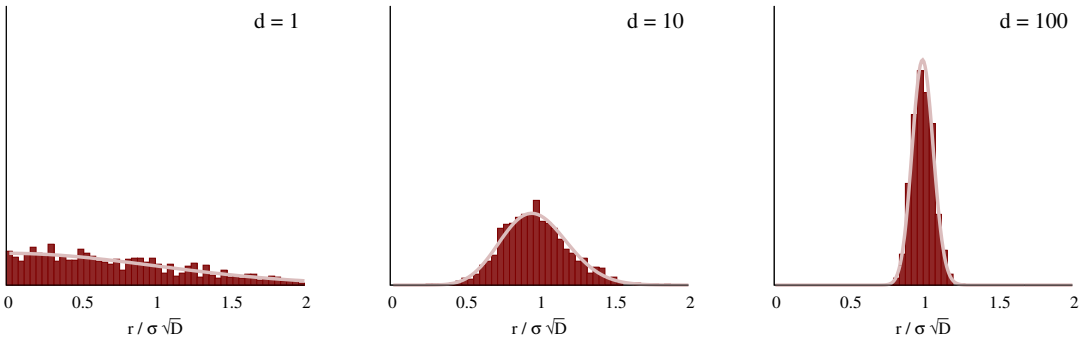


Figure 4.4: Concentration of measure can be visualized with samples from a given distribution, which concentrate across the typical set. For low-dimensions the concentration is weak and the typical set is diffuse, but as the dimensionality of the target distribution grows so too does the concentration 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 just replace it with a simpler one. In other words, we can approximate a complex target distribution, $\mathbb{P}$, with a simpler distribution, $\widetilde{\mathbb{P}}$, whose expectations, or at least some expectations of practical interest, are known analytically,

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

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 our target distribution. The only probability distribution that fits this criteria is the *Dirac distribution*, $\mathbb{D}_{\tilde{\theta}}$, that assigns all probability to a single point in the sample space, $\tilde{\theta}$,

$$\mathbb{D}_{\tilde{\theta}}[E] = \begin{cases} 0, & \tilde{\theta} \notin E \\ 1, & \tilde{\theta} \in E \end{cases} \,.$$

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

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

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,

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

(a) (b)

Figure 5.1: (a) In simple cases a prescient choice of sample space 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 probability density function 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* sample spaces can be reasonably accurate for *some* functions. For example, if the target probability distribution is sufficiently simple that the typical set is convex *and* if a sample space 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, $\mathbb{E}_{\mathbb{P}_\Theta}[\theta]$ (Figure 5.1).

Because they rely on a point estimate, however, modal estimators are terrible at approximating expectations that depend on the breadth of the typical set, such as the variance. Furthermore, identifying the optimal sample space 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 expression 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,

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

Expectations are then estimated with Gaussian integrals

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

which often admit analytic solutions.

The accuracy of 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 sample space can be found.

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

5.3 Variational Estimators

In order to construct an approximation that is not sensitive to the choice of a particular sample space 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*, $\mathcal{Q}$, is a set of probability distributions over the target sample space, Θ, such that at least some expectations can be computed analytically.

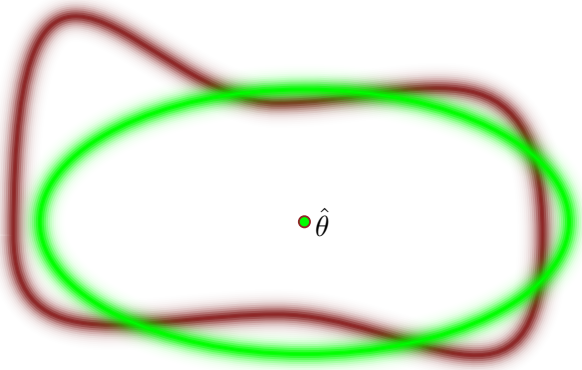


Figure 5.2: For simple probability distributions with well-chosen samples spaces, the local geometry around the mode of a probability density function can quantify the geometry of the entire typical set, yielding accurate Laplace estimators. For more complex probability distributions, however, this local information poorly quantifies the global geometry of the typical set and Laplace estimators suffer from large biases.

In order to identify the best approximation to the target distribution we then define a *divergence function*,

$$\begin{aligned} D : \mathcal{Q} \times \mathcal{Q} &\to \mathbb{R}^{+} \\ \mathbb{P}_1, \mathbb{P}_2 &\mapsto D(\mathbb{P}_1 \,||\, \mathbb{P}_2) \,, \end{aligned}$$

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

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

$$\mathbb{Q}^{*} = \underset{\mathbb{Q} \in \mathcal{Q}}{\text{argmin}} \, D(\mathbb{P} \,||\, \mathbb{Q}) \,.$$

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 in practice (Figure 5.4).

Even if we could find the best approximating distribution, however, there are no guarantees that it will yield accurate estimates for all relevant expectations of our target distribution. 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).

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

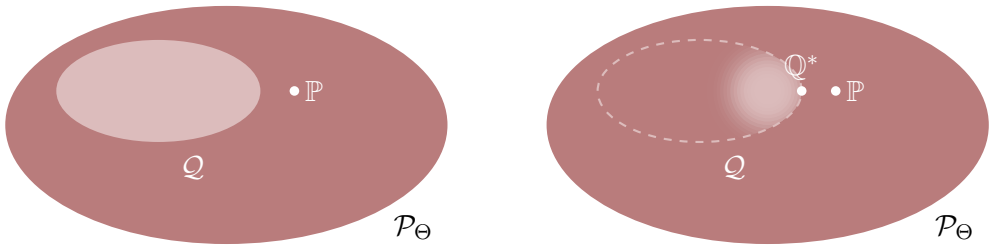


Figure 5.3: (a) A variational family, $\mathcal{Q}$, is a set of probability distributions taken from the set of all probability distributions over the same sample space, as the target distribution, $\mathcal{P}_\Theta$. (b) Adding a divergence function distinguishes which elements of $\mathcal{Q}$ are good approximations to the target distribution, allowing us to identify the best approximation, $\mathbb{Q}^*$.

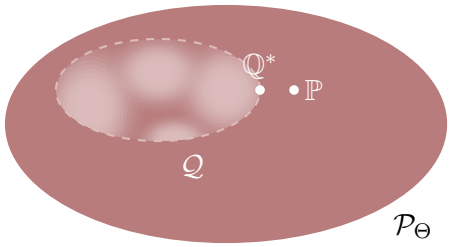
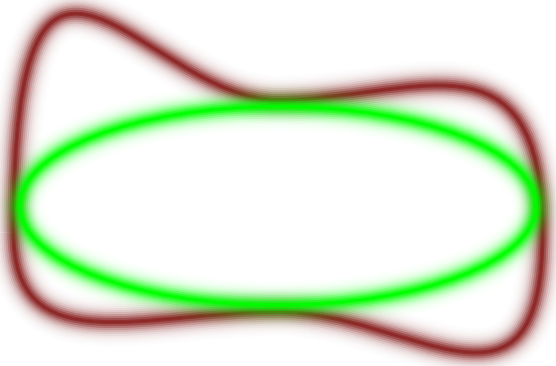
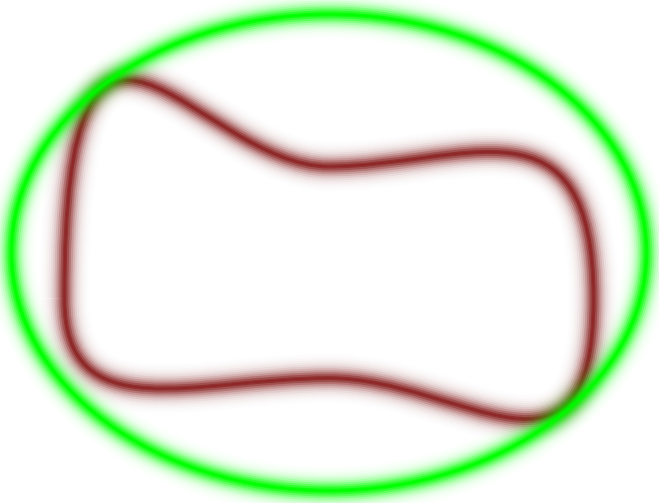


Figure 5.4: Typically different elements of the variational family are able to capture different characteristics of the target distribution and the variational objective manifests multiple optima. Even if a global optimum, $\mathbb{Q}^*$, exists it will be difficult to find and we may be left with only a suboptimal local optimum.



(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 typical set of the target distribution. 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 quantifies the typical set, and hence the samples themselves provide a natural way to identify 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 the expectation 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 **any square integrable function**

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

By construction these Monte Carlo estimators recover the exact expectation asymptot-

ically,

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

but they are also 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,

$$\hat{f}_N \sim \mathcal{N}(\mathbb{E}_{\mathbb{P}}[f], \mathrm{MCSE}),$$

where the *Monte Carlo Standard Error* is defined as

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

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, functions with high variance are more challenging to estimate than those with low variance. **Moreover, because the MCSE doesn't depend on the dimensionality of the sample space it scales particularly well to high-dimensional problems.**

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

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

which is typically negligible compared to the Monte Carlo Standard Error of f. Monte Carlo estimators, then, are distinct from deterministic approximations in that they naturally come equipped with a procedure for at least 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 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, $\mathbb{G}$,

$$\{\vartheta_1, \ldots, \vartheta_N\} \sim \mathbb{G}.$$

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

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

If p and g are the probability density functions corresponding to the target distribution and auxiliary distribution, respectively, then the importance weights are given by

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

Although they are constructed from probability density functions, importance weights, and hence importance sampling estimators, are invariant to the choice of sample space. When we map to an equivalent sample space, the resulting Jacobian is the same in both the numerator and denominator and consequently cancels when evaluating the weights themselves.

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

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

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 useless, it also makes estimates of the variance and hence any quantification of the estimator error useless.

Selecting an auxiliary distribution that yields accurate importance sampling estimators, however, is challenging without knowing the structure of the typical set a priori. **As we move into higher-dimensions, concentration of measure makes this an even more delicate process and 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*, $\mathbb{T}$,

$$\begin{aligned} \mathbb{T} : \mathcal{E}(\Theta) \times \Theta &\to [0, 1] \\ (E, \theta) &\mapsto \mathbb{T}[E \mid \theta] \,. \end{aligned}$$

When the Markov transition operator preserves the target distribution,

$$\mathbb{P}[E] = \mathbb{E}_{\mathbb{P}}[\mathbb{T}[E \mid \theta]]$$

or, with respect to probability density functions,

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

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

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

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, then it concentrates closer to the typical set than away from it – it is literally attracted to the typical set. 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 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 the 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

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

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

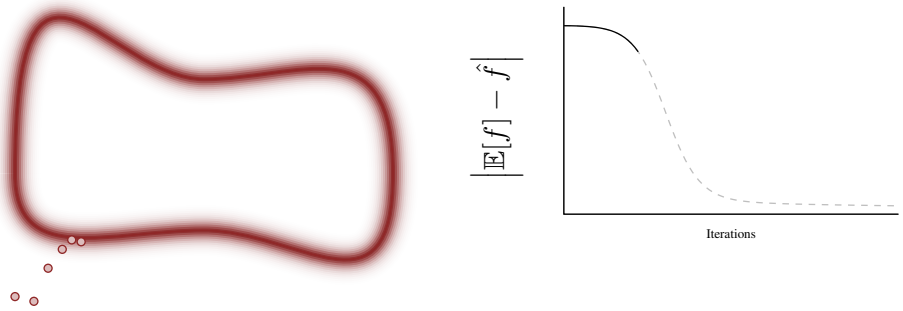
$$\mathrm{ESS} = \frac{N}{1 + 2 \sum_{l=1}^{\infty} \rho_l},$$

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.

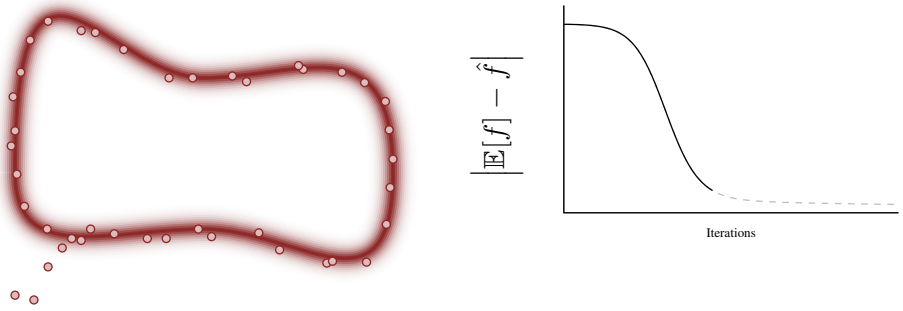
Because the states of the Markov chain during the initial convergence phase mostly bias Markov chain Monte Carlo estimators, we can achieve more precise estimators more quickly by using samples generated only once the Markov chain has begun to explore the typical set. Consequently typical practice is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

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

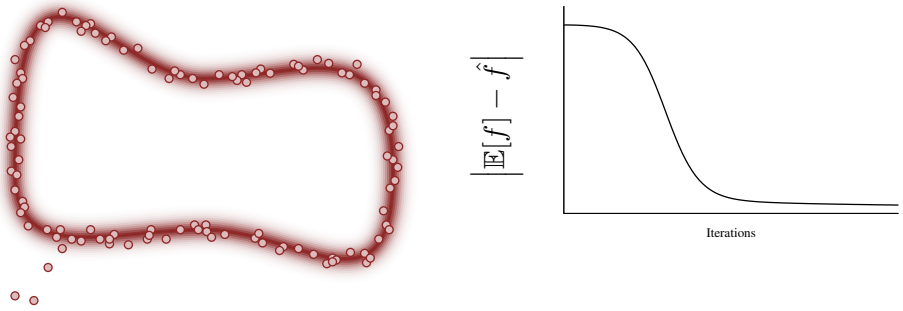
Under ideal conditions Markov chain Monte Carlo behaves very similarly to Monte Carlo, only with a loss of efficiency due to the correlation in the samples. When the target distribution exhibits more pathological behavior, however, Monte Carlo continues to perform well while Markov chain Monte Carlo begins to fail in spectacular fashion.



(a)



(b)



(c)

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 behavior 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.

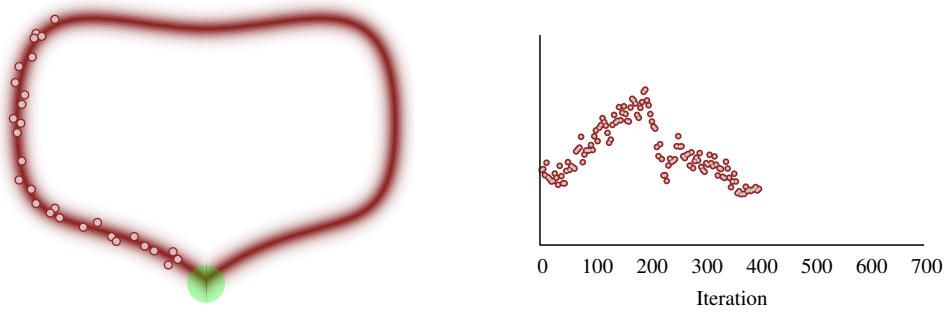
Because Markov chains have to recover the exact expectations asymptotically, they have to somehow compensate for not being able to explore these regions. Typically the Markov chain accomplishes this by getting stuck near the boundary of the pathological region: as it hovers the estimators are drawn down 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.

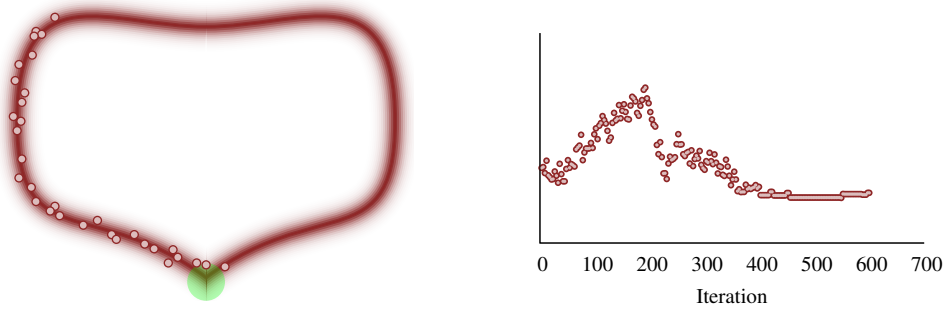
Whether or not features of the target distribution become pathological depends on how the Markov transition kernel interacts with the target distribution. Some transition kernels are more robust than others and some can achieve robust performance with careful tuning of auxiliary kernel parameters.

6.3.3 Markov Chain Monte Carlo in Practice

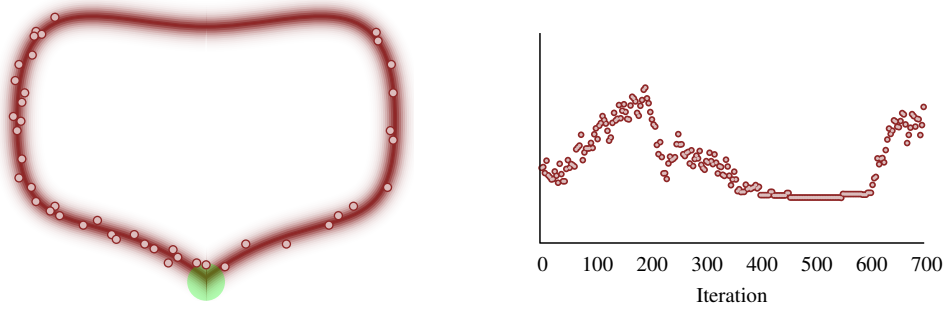
In order to guarantee that we will not suffer from pathological behavior we have to demonstrate strong *ergodicity* conditions that ensure the Markov chain not only explores the typical set but does so sufficiently fast without being hampered by pathological neighborhoods. In most cases we need to establish *geometric ergodicity*.



(a)



(b)



(c)

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.

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 necessary conditions to provide any 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 run long enough for 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 use the term warmup.

Sampling

Once warmup as 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 the 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

We begin by defining measurements before considering how to model that measurements mathematically.

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 statements, but in practice we can observe only variable measurements of those statements. 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, $\mathbb{P}_D$ 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*, $\mathbb{P}_D$, 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 $\mathbb{P}_D$: 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 certainly inform us of the data generating process exactly, but measurements are expensive

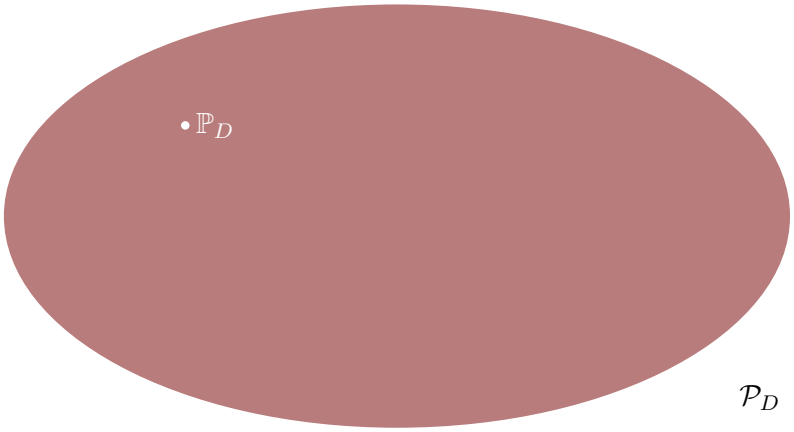


Figure 7.1: Once we have defined a measurement space, D, the latent data generating process, $\mathbb{P}_D$, must be contained in the space of all possible data generating processes over D, $\mathcal{P}_D$.

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 we have to consider all possible data generating process 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).

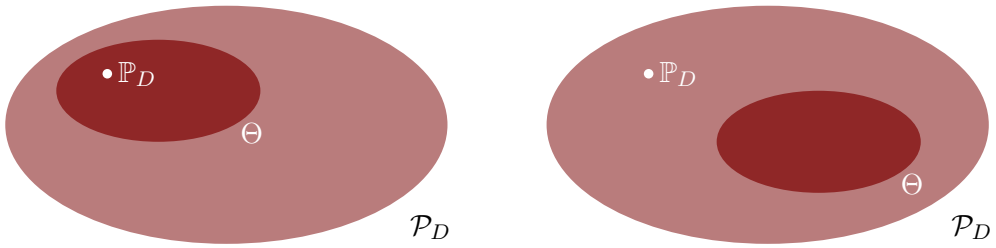
The big world is much too ungainly to be even well-defined in practice, let alone exhaustively explored. Instead we have to limit our consideration to only a subset of probability distributions over the measurement space called a *small world*, $\Theta \subset \mathcal{P}_D$ (Figure 7.2a). An expression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

Each point in the small world, $\theta \in \Theta$, identifies a unique probability distribution over data. Consequently the small world defines a collection of probability distribution over the measurement space,

$$\begin{aligned} \mathbb{L} : \mathcal{E}(D) \times \Theta &\to [0, 1] \\ (E_D, \theta) &\mapsto \mathbb{L}[E_D \mid \theta] \,. \end{aligned}$$

known as the *likelihood*. We can also interpret the likelihood as a conditional probability distribution over implications from our small world model to the measurement space.

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



(a) (b)

Figure 7.2: Practical inference requires the selection of a distinguished subset of data generating processes called a small world, Θ, that (a) may or (b) may not contain the latent data generating process, $\mathbb{P}_D$. 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 $\mathbb{P}_D$.

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

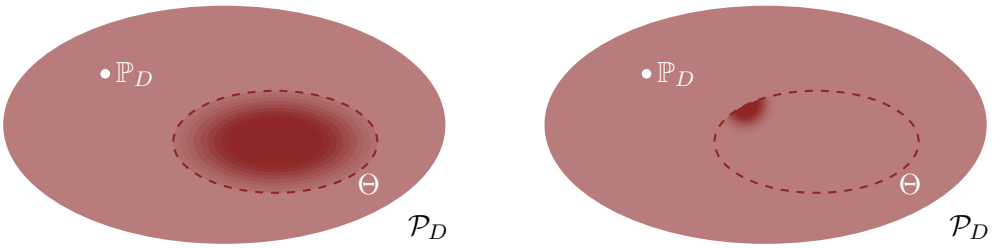
We have already used probability theory to quantify the variability of measurements, but now we can also use probability theory to quantify our uncertainty about which elements of the small world are good approximations to the latent data generating process, both before and after we make a measurement.

The *prior distribution*, $\mathbb{P}_{\Theta}^{\text{prior}}$, is probability distribution over the small world that quantifies our initial uncertainty about 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.

Learning in the small world is the process of updating the prior distribution with any information contained in the measurement to give a *posterior distribution*, $\mathbb{P}_{\Theta}^{\text{post}}$, that quantifies our updated uncertainty about the small world (Figure 7.3). The likelihood implicitly quantifies any information contained in a measurement and then the actual mechanism for this update is immediately given by probability theory. It is most simply written in terms of probability density functions,

$$p^{\text{post}}(\theta \mid d) \propto L(d \mid \theta)\, p^{\text{prior}}(\theta)\,,$$

which can be recognized as the celebrated Bayes' Theorem.



(a) (b)

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, $\mathbb{P}_D$.

Bayes' Theorem, however, is just a mathematical consequence of probability theory and 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

Because it quantifies our uncertainty about the small world, the posterior distribution is a critical component to robust decision making.

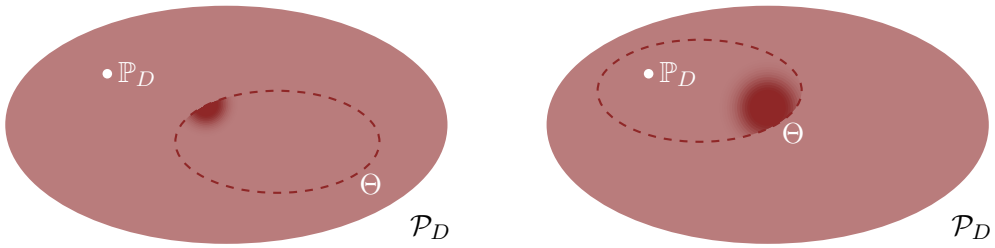
Let's say that we have a set of actions or interventions, $\mathcal{A} = \{A_i\}$. A *utility function* quantifies the utility of each action assuming that a particular element of the small world is true,

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

Given an element of the small world, $\theta^* \in \Theta$, the best decision is the action that maximizes utility,

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

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



(a) (b)

Figure 7.4: Bayesian inference yields optimal learning within only 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 desired distribution.

latent data generating process can yield completely different decisions, making subsequent decision making 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*,

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

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 assumptions 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 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 their 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

Any element of the small world identifies a distribution over the measurement space and, because we assume that whatever process generates the measurements is stationary, we could use that distribution as a *predictive distribution* over new measurements.

A single element of the small world, however, ignores our uncertainty in the small world. 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. This expectation yields the *posterior predictive distribution*, $\mathbb{P}^{pp}_{\tilde{D}|D}$, implemented with the probability density function

$$p^{\text{pp}}\left(\tilde{d} \mid d\right) = \int L\left(\tilde{d} \mid \theta\right) p^{\text{post}}(\theta \mid d) \, \mathrm{d}\theta.$$

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

$$\{\theta_n\} \sim \mathbb{P}^{\text{post}}_{\Theta}$$

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,

$$\tilde{d}_n \sim \mathbb{L}[\cdot \mid \theta_n] \, .$$

Our model assumptions are valid if and only if the posterior predictive distribution is sufficiently close 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 over the measurement space is the *Kullback-Leibler divergence*, $\mathrm{KL}\left(\mathbb{P}_{\tilde{D}} \,||\, \mathbb{P}^{pp}_{\tilde{D}|D}\right)$, which is implemented with an integral over probability density functions,

$$\mathrm{KL}\left(\mathbb{P}_{\tilde{D}} \,||\, \mathbb{P}^{pp}_{\tilde{D}|D}\right) = \int \frac{p^{pp}\left(\tilde{d} \mid d\right)}{p\left(\tilde{d}\right)} p\left(\tilde{d}\right) \mathrm{d}d.$$

When the posterior predictive distribution matches the latent data generating process the divergence vanishes, and as the posterior predictive distribution becomes a worse and worse fit for the latent data generating process the divergence grows.

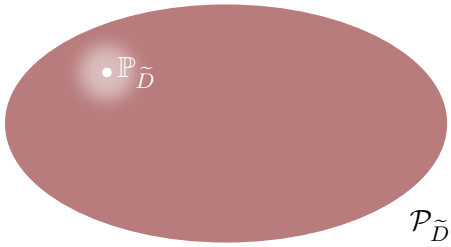


Figure 7.5: The Kullback-Leibler divergence between the latent data generating process and the predictive distribution of a model defines a measure of predictive performance which allows us to rank the performance of model assumptions relative to each other. The smaller the divergence the closer the predictive distribution is to the latent data generating process.

Ultimately the Kullback-Leibler divergence 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 that yield a small divergence with the latent data generating process yield better fits, and 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. In practice, however, there is no real motivation for a setting such a threshold and absolute statements are largely ill-posed.

The Kullback-Leibler divergence is an ideal way to compare modeling assumptions in theory, but in practice it is of limited use because it cannot actually be computed without knowledge of the latent data generating process itself! 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 not the entire latent data generating process but just the measurement itself. If the measurement is far from the

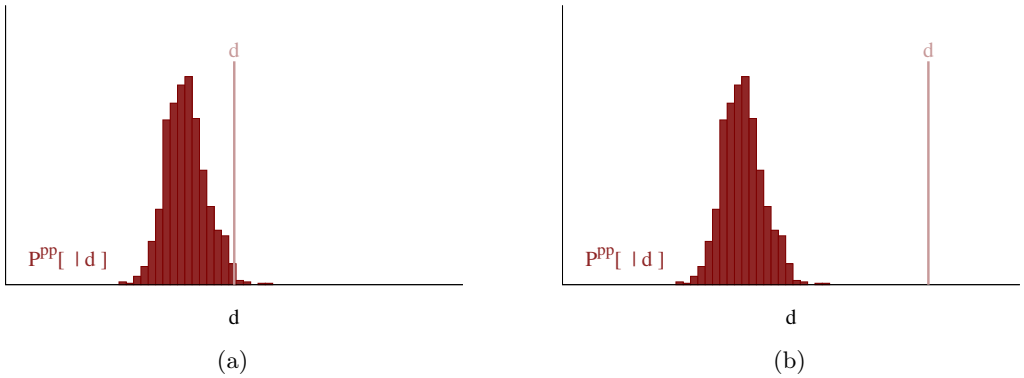


Figure 7.6: (a) When the measurement, d, is consistent with the posterior predictive distribution we have no reason to doubt our model assumptions, but (b) tension between the measurement and the posterior predictive distribution indicates that there may be a problem. Either the measurement was exceedingly rare or our model assumptions are insufficient.

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 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. Given these two inputs we can immediately construct a posterior distribution, and then all inferential summaries and decisions reduce to computing the corresponding expectations. Although conceptually simple, neither of these steps is

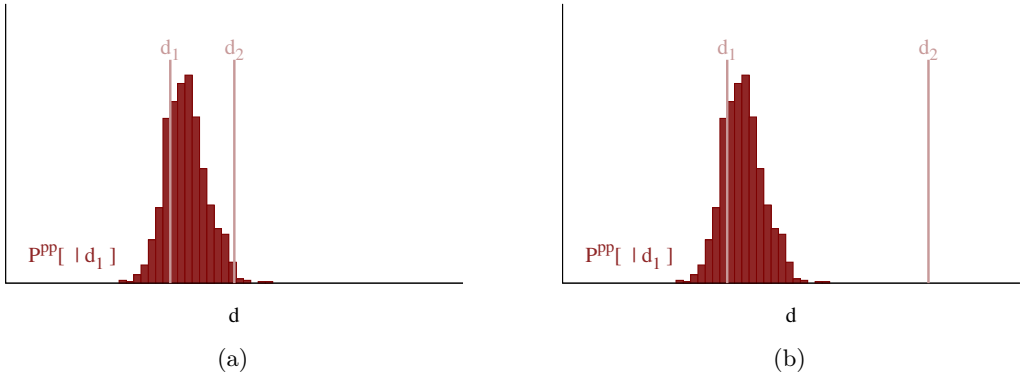


Figure 7.7: To test for potential overfit we need multiple measurements, and given only a single measurement this can emulated by splitting the measurement into two partitions, d_1 and d_2. (a) When both measurements are consistent with the posterior predictive distribution we have no reason to doubt 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.

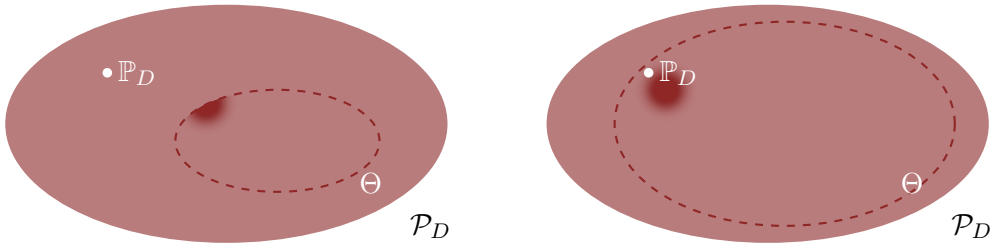


Figure 7.8: Posterior predictive checks are powerful ways to make qualitative judgements about the validity of our model assumptions. For example, if (a) they suggest that our model is misfitting the data then (b) we can update our model to better capture the features of the data that are being misfit.

particularly easy.

Building robust models is an immensely challenging tasks 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. **Constructive example mirroring the motivating example in the introduction.**

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 problems.

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 fits, we can use Bayesian inference to make powerful statements about the world around us.