Uncertainty is inherent to learning and decision making, and it becomes all the more important when we try to formalize these processes. For example, measurement, the means of collecting the data from which we wish to learn, is fundamentally variable, and this variability must propagate to our inferences and any decisions based on them. Consequently any formal technique for learning and decision making must be able to quantify uncertainty in a self-consistent manner. *Bayesian inference* uses probability theory to quantify all forms of uncertainty, including not only the variability in measurements but also ignorance in the learning process itself.

Because probability theory is so subtle and counterintuitive, introductory treatments of Bayesian inference often oversimplify and neglect many of the finer technical aspects. Unfortunately, these technicalities have a strong influence on practical applications of the theory, and without at least a conceptual understanding we are subject to dangerous fallacies. 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 probability theory, keeping the presentation abstract but focusing on concepts instead of mathematical pedantry. We'll then introduce how we implement these abstract ideas in practice, both in terms of concrete representations and computation, before discussing many popular computational methods. Finally we'll show how all of these ideas come together in Bayesian inference.

1 Mathematical Background and Notation

Unfortunately, a thorough review of probability theory requires a nontrivial mathematical background. Although we strive to avoid as much mathematical minutiae as possible, we have to assume that the reader is comfortable with the basics of differential and integral calculus over the real numbers. We highly encourage anyone who's math might be rusty to brush up before proceeding.

We also assume familiarity with 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

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

The *union* of two sets, $A \cup B$ is the combination of all elements in either set while the *intersection* of two sets, $A \cap B$ is only those elements contained in both sets.

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 continuous.

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

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

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.

2 Probability in Theory

Probability theory, and the measure theory on which it is based, is a very detailed and intricate mathematical construction. Here we will make no attempt at the full mathematical rigor necessary for a complete understanding, and instead focus on a high-level, conceptual intuition. This means, for example, that in many places we will appeal to vague notions like "well-behaved", as their technical definitions do not offer much pedagogical benefit.

The fundamental objects in probability theory are probability distributions defined over event spaces. In this section these objects are defined only abstractly, which unfortunately means that we cannot rely on helpful visualizations to support the definitions and manipulations. To use a probability distribution in any practical application, including visualization, we need some means of specifying and utilizing them explicitly, which will have to wait until the next section.

2.1 Logical Statements and Events

Logical statements are the general way to specify information about a system. For example, if $\boldsymbol{\theta}$ is an object that takes values in the space Θ then we can make statements like $\boldsymbol{\theta} = \theta$ or $\theta_1 < \boldsymbol{\theta} < \theta_2$.

One potential use of logical statements is to describe the outcome of a variable measurement, or an event. Indeed this is exactly the application on which modern probability theory was based and, unfortunately, this also means that the terminology in probability theory focuses on this particular application instead of more general logical statements. In order to avoid conflicting with standard terminology, going forward we will refer to logical statements as events, but it is important to keep in mind the more general interpretation.

More formally, consider quantifying information about an object $\boldsymbol{\theta}$ which takes values in the space Θ. Θ is known as the *sample space*, and it allows us to explicitly specify logical statements about $\boldsymbol{\theta}$. Logical statements, or *events*, E, as specified by well-behaved subset of the sample space, $E \subset \Theta$. We also denote the *event space* of well-behaved subsets of the sample space as $\mathcal{E}(\Theta)$.

For example, if $\boldsymbol{\theta}$ took values on the real line, $\Theta = \mathbb{R}$, then events could be points in the sample space, $E = \{\theta = 5\}$, intervals, $E = \{\theta \in \Theta \mid \theta < 5\}$ or $E = \{\theta \in \Theta \mid 1 < \theta < 5\}$, or various combinations of the two, such as

$$E = \{\theta = -5\} \cup \{\theta \in \Theta \mid 1 < \theta < 5\} \cup \{\theta \in \Theta \mid 10 < \theta < 12\}.$$

Event spaces always include the null event, $E = \emptyset$, and the trivial event, $E = \Theta$.

Different ways of specifying logical statements quantify the same information

One subtlety with this construction is that we can specify the same logical statements, and hence events, using different sample spaces. Consider, for example, 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,

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

then the map is called *measurable* and the two sample spaces can be used to specify the same logical statements. When a measurable map exists between two sample spaces we say that they are *equivalent* as they provide equivalent descriptions of the same system.

In other words, there are many ways to describe any given system, and hence different ways to explicitly specify logical statements. These different specifications, however, all quantify the same information.

2.2 Probability Distributions

If we are uncertain about our target system, $\boldsymbol{\theta}$, then we cannot ensure that any particular event is true. Instead all we can do is quantify our uncertainty about $\boldsymbol{\theta}$ by assigning each event a *probability* that quantifies which events are more likely to be true than others. Probabilities themselves are bounded between 0, indicating that an event is absolutely false, and 1, indicating that an event is absolutely true.

In probability theory probabilities are assigned to events by a *probability distribution*,

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

where $[0, 1]$ is the interval

$$[0, 1] = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}.$$

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 $\boldsymbol{\theta}$ has to take values in the sample space.

If logical statements quantify information about the system $\boldsymbol{\theta}$, then probability distributions quantify our uncertainty about those statements and hence our uncertainty about the the system itself. When using a probability distribution to quantify our uncertainty about $\boldsymbol{\theta}$ we often write $\boldsymbol{\theta} \sim \mathbb{P}$ which is usually read as, "$\boldsymbol{\theta}$ is distributed according to the probability distribution $\mathbb{P}$" but should really read "Logical statements about $\boldsymbol{\theta}$ are distributed according to the probability distribution $\mathbb{P}$".

These probability assignments satisfy the usual rules of probability, such as the sum rule,

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

and the exclusion rule,

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

where E^c is the *complement* of E satisfying $E \cup E^c = \Theta$.

All of probability theory reduces to expectations

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!

2.3 Conditional Probability Distributions

Reasoning about probability distributions on low-dimensional spaces is usually straightforward – or at least as straightforward as the mathematics will allow – but reasoning about probability distributions on high-dimensional spaces is much more subtle. One very elegant way to simplify these high-dimensional probability distributions is to treat them as a product of simpler, one-dimensional probability distributions know as *conditional probability distributions*.

A conditional probability distribution is a set of probability distributions indexed by some auxiliary space, Φ. For any value of $\phi \in \Phi$, the conditional probability distribution defines a probability distribution on Θ; for any event in Θ, the conditional probability distribution defines a function from Φ to probabilities:

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

Joint distributions implied by expectations

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 appropriate event in Φ. Similarly, joint expectations are defined iteratively as

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

Marginalization collapses joint distribution onto component spaces

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 taking into account all correlations between the components.

Generative modeling is building joint distribution from conditionals

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 using 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 naturally motivated by the problem at hand and, 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 also incorporate deterministic and even causal relationships. This iterative process of building a joint probability distribution from conditional probability distributions is known as *generative modeling*.

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

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

Just like events, probability distributions are invariant when we move between equivalent sample spaces. 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: Specifying Distributions

As we saw in the previous section, abstract notions of probability distributions are not particularly easy to intuit, and abstract probability distributions are not easily specified or manipulated in practical applications. When the sample space is structured, however, that structure can be leveraged to provide the explicit specifications 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 Representations 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 represent 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 simply 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 represent 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 compute 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,

$$I(\theta) = \big\{\theta' \in \Theta \mid \theta' \leq \theta\big\}\,.$$

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 Representations 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.

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

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

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

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

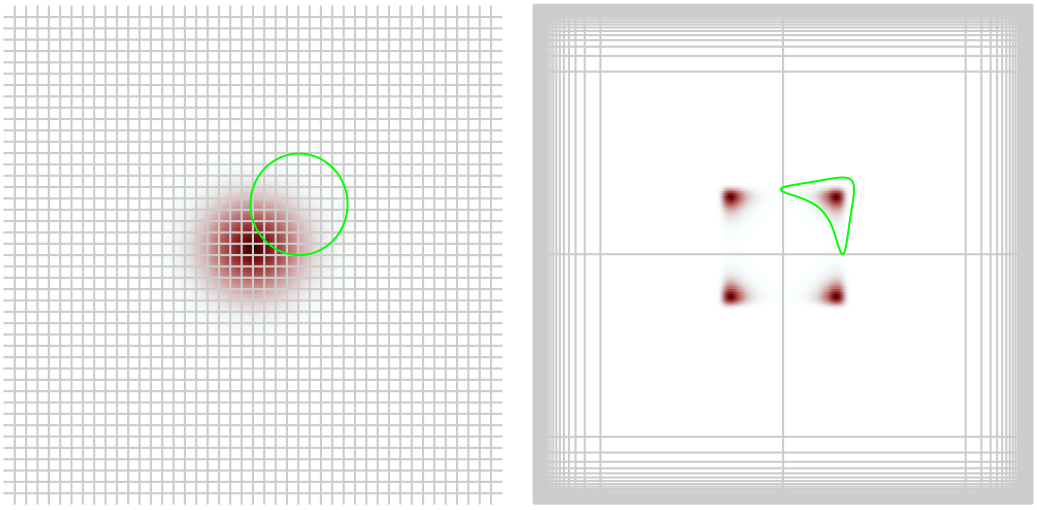
where the matrix

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

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 1). This



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

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

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,

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

For a concrete example, 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

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

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. These two realizations are shown graphically in (Figure 1).

3.2.2 Conditional Probability Density Functions

Just as in the discrete case, probability density functions can be immediately extended to represent 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,

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

3.2.3 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,

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

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,

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

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

3.2.4 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,

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

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,

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

3.3 Representations of Mixed Probability Distributions

Distributions over samples spaces that have both a discrete, Φ, and a real Ψ, component can be specified 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 conditioning 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 distributions straightforward in practice.

3.4 Stochastic Representations of Probability Distributions

We can also represent 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, $\{\theta_1, \ldots, \theta_N\} \subset \Theta$. Such a mechanism is equivalent to 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

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

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

This procedure, and hence the resulting samples, are *exact* if every element in the generated sequence is independent of every other element. 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 are effectively random

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 a representation of our probability distribution, these computations reduce to straightforward manipulations, either summations 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 approximate estimates. 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 overwhelm our finite computational resources. Consequently, computationally efficient yet accurate estimates requires 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 and how that geometry frustrates approximate algorithms, 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 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 inte-

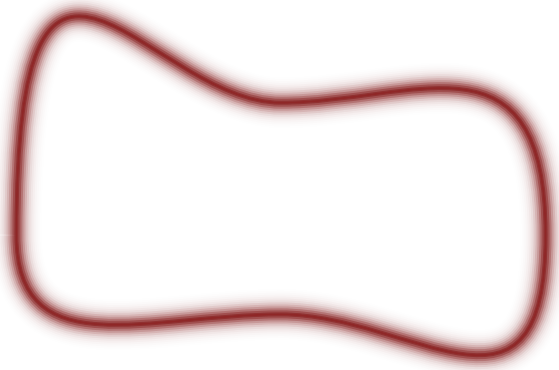


Figure 2: On high-dimensional samples spaces all well-behaved probability distributions concentrate in a neighborhood called the *typical set*. In order to estimate the integrals needed to compute probabilities and expectations we have to be able to identify where the typical set lies in the sample space which is no easy task.

gral, 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 is the *product* of these two contributions, concentrates in a singular neighborhood somewhere in the middle known as the *typical set* (Figure 2) in a 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 specify 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 representation, such as the mode 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 can marginalize out the hyperspherical angles analytically to give the radial probability density function,

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

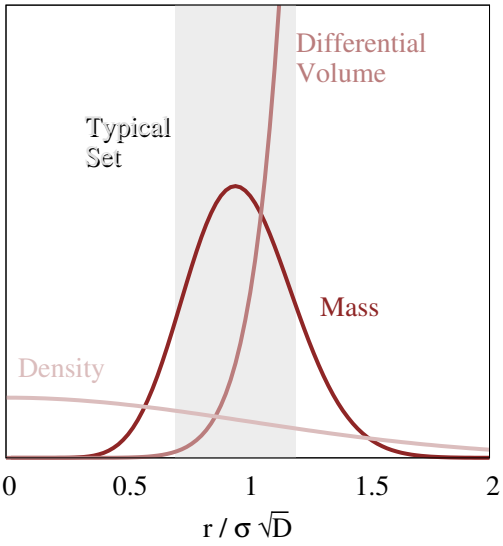


Figure 3: In high dimensions real probability distributions generically assign almost all of their probability into a singular neighborhood known as the typical set. This is apparent even from a probability density function representation: although the density concentrates around the corresponding mode, 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.

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 that neighborhood concentrates tighter and tighter around that shell as we add more and more dimensions (Figure 3).

Because samples recover all expectations asymptotically, they have must concentrate across the typical set (Figure 4). Consequently, we can also use samples to simulate concentrate 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,

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

with the corresponding radial distance $r = \sqrt{\sum_{d=1}^{D} x_d^2}$. Generating a sequence of samples and then histogramming the radial distance reveals the same χ distribution that we arrived at analytically (Figure 5).

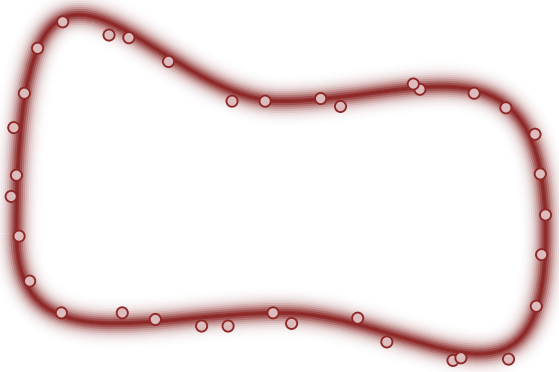


Figure 4: Because samples recover expectations asymptotically, large sequences of samples must concentrate in the typical set. This provides a means of visualizing concentration of measure, and will prove a powerful way to estimate expectations.

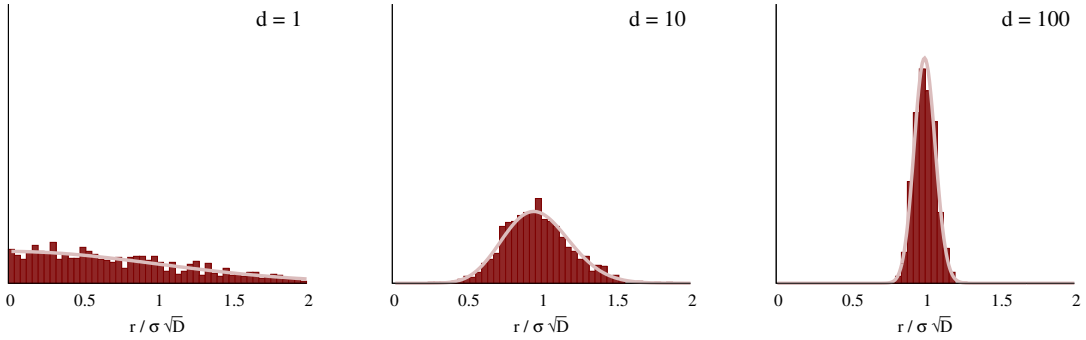


Figure 5: Concentration of measure can be visualized with samples from a given distribution, which concentrate across the typical set. For low-dimensions the concentration is weak the the typical set is diffuse, but as the dimensionality of the target distribution grows so do does the concentration of measure.

5 Probability in Practice: Deterministic Estimators

If we can't quantify the target typical set exactly, one immediately 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 MAP 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 *maximum a posteriori*, or *MAP*, estimation. In MAP estimation we approximate the target distribution with a Dirac distribution at the mode of the probability density function,

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

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 MAP estimation so ubiquitous?

MAP estimators are seductive because the optimization on which they rely is relatively computationally inexpensive. Moreover, in some very simple cases MAP estimators constructed from 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

(a) (b)

Figure 6: (a) In simple cases a prescient choice of sample space can yield a MAP 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.

the mode of the probability density function lies in the center of that typical set, then the corresponding MAP estimator might yield reasonably accurate estimates for the mean, $\mathbb{E}_{\mathbb{P}_\Theta}[\theta]$ (Figure 6).

Because they rely on a point estimate, however, MAP 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, even if one exists, for a particular target function 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 sample space and inability to validate the accuracy of the estimators makes MAP estimation extremely fragile in practice.

5.2 Laplace Estimators

The fragility of MAP 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 MAP estimate,

$$\mu = \theta_{\text{MAP}},$$

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

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

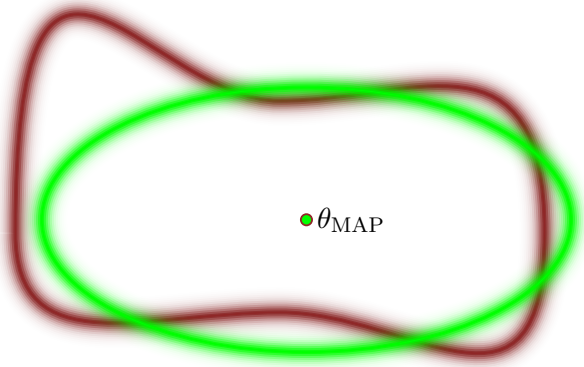


Figure 7: For simple probability distributions with well-chosen samples spaces, the local geometry around the mode of a probability density function can quantify the geometry of the entire typical set, yielding accurate Laplace estimators. For more complex probability distributions, however, this local information poorly quantifies the global geometry of the typical set and Laplace estimators suffer from large biases.

Expectations are then estimated with Gaussian integrals

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

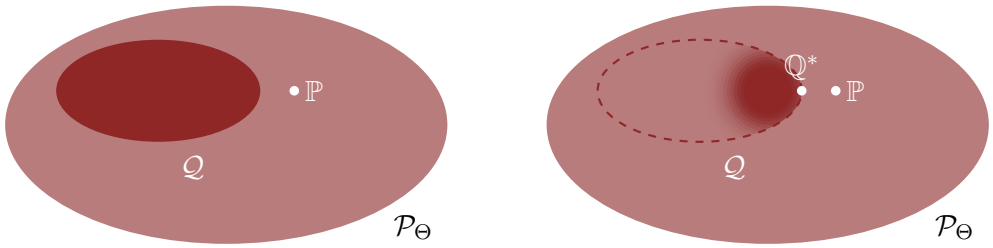
which often admit analytic solutions.

The accuracy of Laplace approximations depends on how well the mode and at the Hessian of the probability density function quantifies the geometry of the typical set (Figure 7). Unfortunately the conditions that are necessary for these estimates to be reasonably accurate hold only for very simple probability distributions, and only then only if an appropriate sample space can be found.

As with the simpler MAP 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 a 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*.



(a) (b)

Figure 8: (a) A variational family, $\mathcal{Q}$, is a set of probability distributions over the same sample space, Θ, as the target distribution, $\mathbb{P}$, taken from the set of all probability distributions over Θ, $\mathcal{P}_\Theta$. (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, $\mathbb{Q}^*$.

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

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

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 8).

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

Although straightforward to define, this variational optimization can 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 9).

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 10).

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

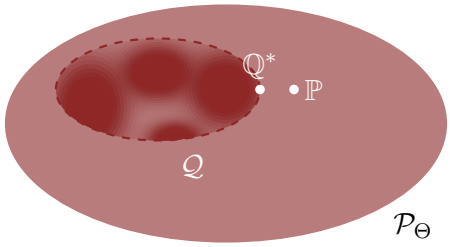


Figure 9: In practice typically different elements of the variational family are able to capture different characteristics of the target distribution and the variational objective manifests multiple optima. Even if a global optimum, $\mathbb{Q}^*$, exists it will be difficult to find and we may be left with only a suboptimal local optimum.

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.

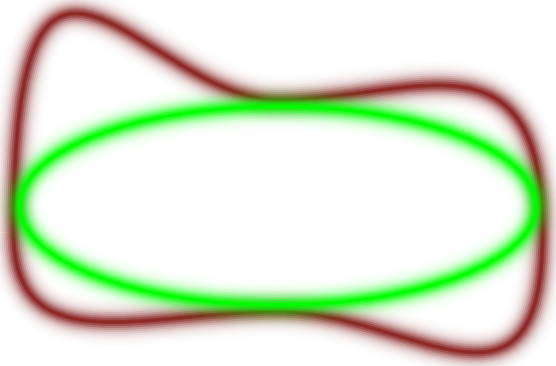
By construction an infinite number of samples from the target distribution quantifies the typical set, and hence the samples themselves provide a naturally way to identify the typical set (Figure 4). The utility of samples, however, depends both on how precisely we can quantify the typical set using 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.

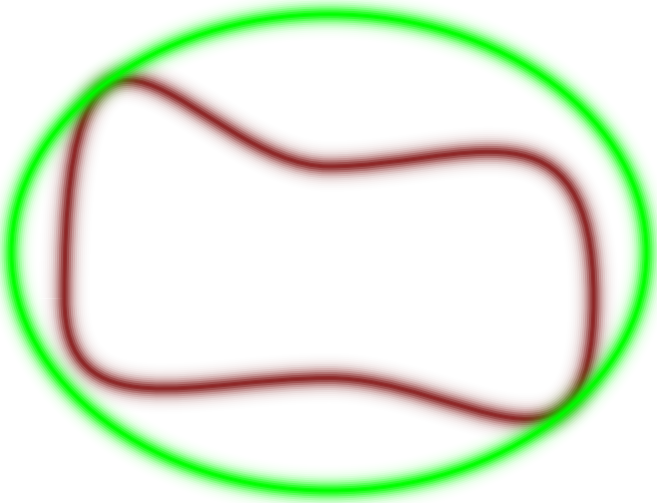
6.1 Monte Carlo Estimators

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

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



(a)



(b)

Figure 10: 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.

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

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

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.

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 $\mathrm{Var}[\,\mathrm{Var}[f]\,]/N$ 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, impotence 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 hence 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 sampler estimators, however, is challenging without knowing the structure of the typical set a priori. 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 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*,

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

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 full explore the typical set only asymptotically. In practice, however, it is the only the finite time performance of Markov chains that matters. 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 11a). 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 (Figure 11b). 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 11c).

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

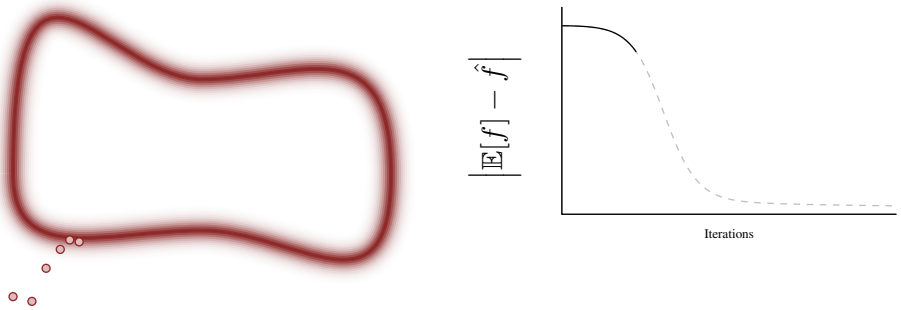
$$\mathrm{ESS} = \frac{N}{1 + 2 \sum_{l=1}^{\infty} \rho_l},$$

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.

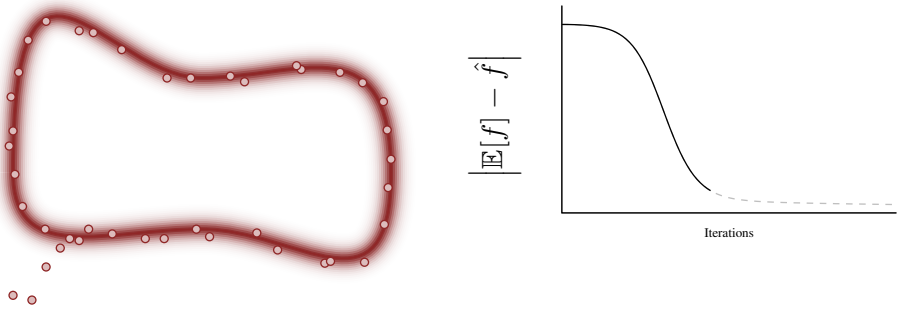
Because the states of the Markov chain during the initial convergence phase mostly bias Markov chain Monte Carlo estimators, we can achieve more precise estimators more quickly by using samples generated only once the Markov chain has begun to explore the typical set. Consequently typical practice is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

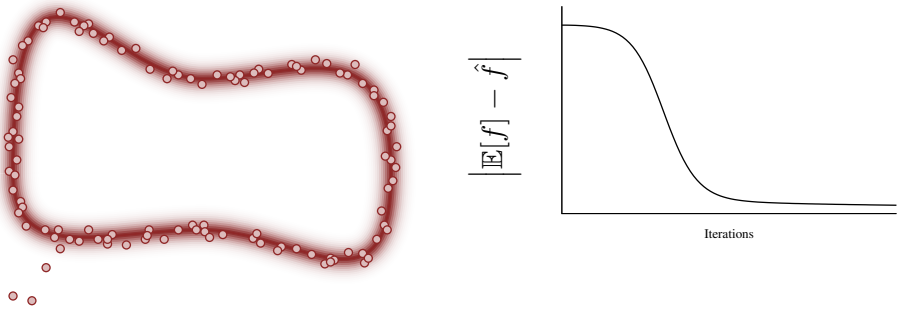
Under ideal conditions Markov chain Monte Carlo behaves very similarly to Monte Carlo with a loss of efficiency due to the correlation in the samples. When the target distribution exhibits more pathological behavior, however, Monte Carlo continues to perform well while Markov chain Monte Carlo begins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



(a)



(b)



(c)

Figure 11: 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 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 mixes, exploring more details of the typical set and gradually improving estimator precision.

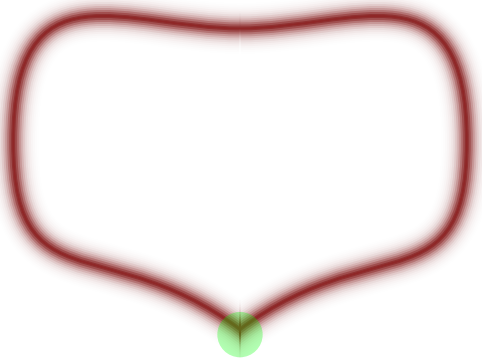


Figure 12: Markov chains typically have trouble exploring regions of the typical set with large curvature (green), which induces bias in Markov chain Monte Carlo estimators and spoils idealized behavior such as Central Limit Theorems.

Consider, for example, a target probability distribution where the typical set pinches into a region of high curvature (Figure 12). 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 chain 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 13).

Ultimately this behavior results in estimators that oscillate around the true expectations. Asymptotically the oscillations average out 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 MCMC 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. In most cases we need to establish *geometric ergodicity*.

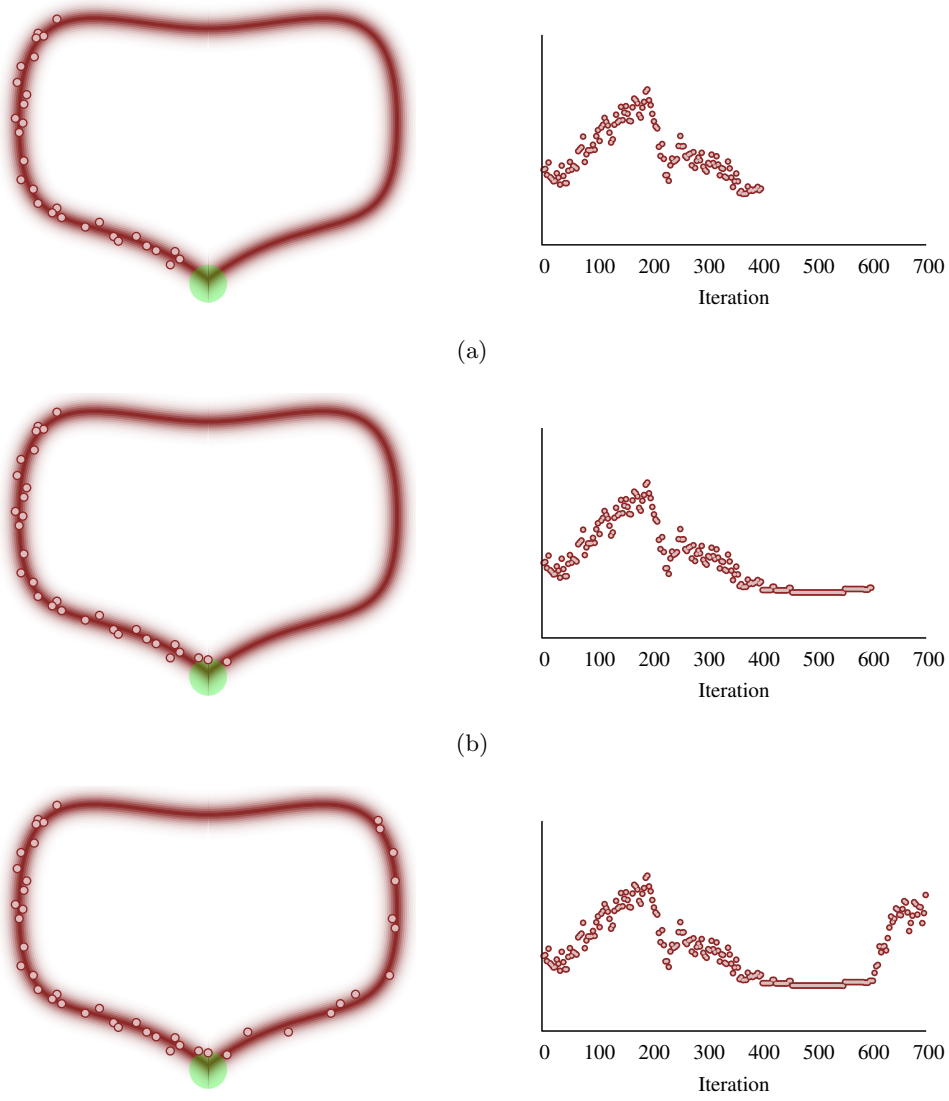


Figure 13: In practice, pathological regions of the typical set usually cause Markov chains to get "stuck". (a) Initial the Markov chain explores well-behaved regions of the typical set, avoiding the pathological neighborhood entirely and biasing Markov chain Monte Carlo estimators. (b) If the Markov 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 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. 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? Well 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 a foundation of probability theory we are now 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 frequentist approach.

We begin by defining measurements and what we want to learn from those measurements, and then consider how we approximate that system in practice with an abstract mathematical model. Once we have defined a model we can define how we learn from that model and then how we make decisions with that model.

7.1 Measurements

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 describe 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 described by a mathematical model.

Although we can assume the existence of a data generating process, 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 and in practice we have to learn about the data generating process from only a few measurements, if not just a single measurements. *Inference* is the process of learning about the data generating process using only a finite number measurements.

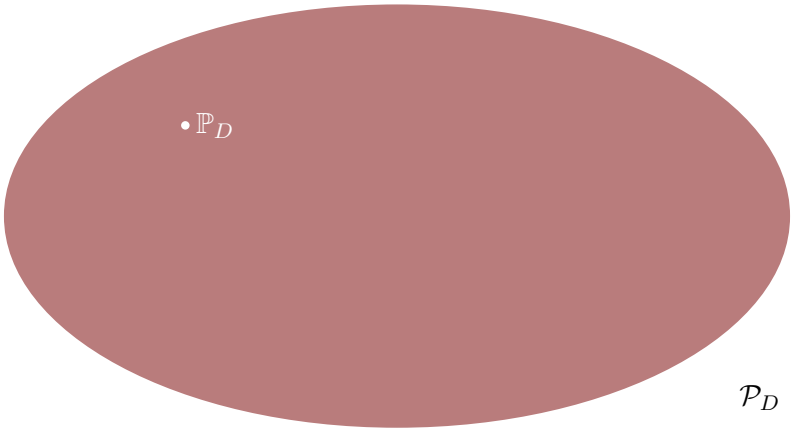


Figure 14: Once we have defined a measurement space, D, the latent data generating process, $\mathbb{P}_D$ can be found in the space of all possible data generating processes over D, $\mathcal{P}_D$.

7.2 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 14).

The big world is much to ungainly to be even be 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 15a).

Each point in the small world, $\theta \in \Theta$, identifies a unique probability distribution over data. Consequently the small world is equivalent to a probability distribution over the measurements space conditioned on the small world,

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

This conditional probability distribution is also known as the *likelihood*.

Regardless of how it is chosen, the assumption of any specific small world can have drastic limitations on inference. Because any small world is likely to be only a shallow approximation of reality, for example, it is probability to contain the latent data generating process (Figure 15b). Consequently even ideal inferences are subject to error, and the utility of any inference always depends on the viability of our assumptions.

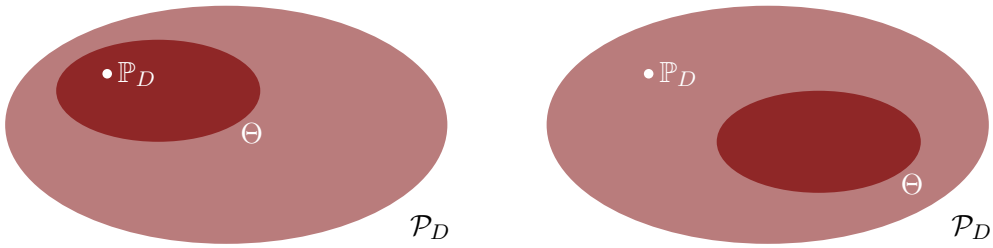


Figure 15: Practical inference requires the selection of a distinguished subset of data generating processes called a small world, Θ, that (a) may or (b) may not contain the latent data generating process, $\mathbb{P}_D$. The Boxian philosophy of "all models are wrong but some are useful" asserts that the former is impossible in practical problems, but even in the latter the probability distributions in the small world may provide useful approximations of $\mathbb{P}_D$.

7.3 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.

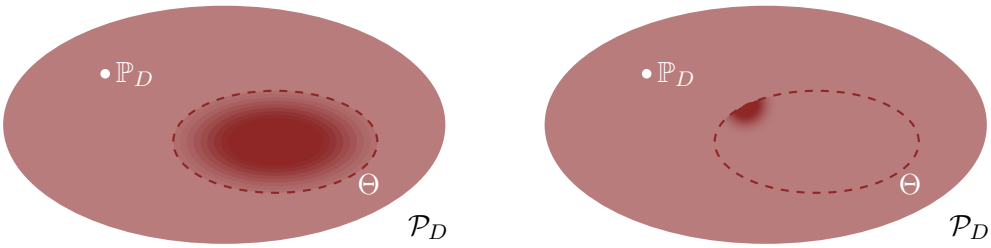
The *prior distribution*, $\mathbb{P}_\Theta^{\text{prior}}$, is a probability distribution over the small world that quantifies our initial uncertainty about which elements are most consistent with the latent data generating process. The information inherent in the prior distribution can come from previous measurements, theoretical constraints, or even elicitation of experts.

Learning in the small world is the process of the updating of the prior distribution with any information contained in the measurement to give a *posterior distribution*, $\mathbb{P}_\Theta^{\text{post}}$, that quantifies our uncertainty about the small world after the measurement (Figure 16). The likelihood implicit defines 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.

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



(a) (b)

Figure 16: 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, $\mathbb{P}_D$.

7.4 TODO: Decision Making in the Small World

Now that we've quantified uncertainty we can make robust decisions.

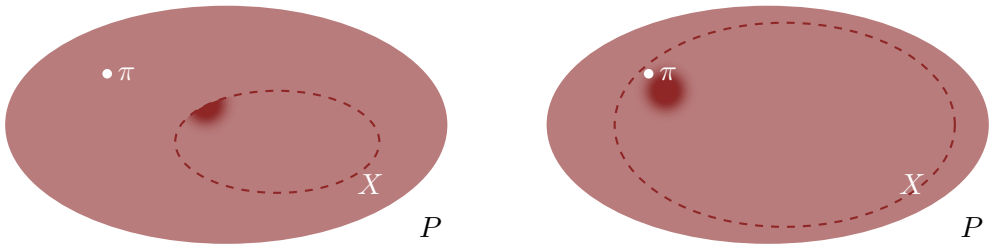
Formalize with a risk/utility function, compute expected risk/utility, then chose the decision that minimizes/maximizes the expected risk/utility.

7.5 TODO: Bayesian Inference in Practice

Model a prior and a likelihood. Posterior is immediately given and all statements about our system, including decisions, are given by expectations. As discussed above we have many options for approximating those expectations in practice.

The biggest challenge in implementing Bayesian inference, then, is the actual modeling of the prior and likelihood. Much can be said about both, but let's take a second to discuss one of the most powerful means of methods of building small worlds: *generative modeling*. Here we build up the small world sequentially by modeling each state of the data generating process. For example, we might build a small world for polling data by modeling the sampling of an individual from a population and then a series of non-responses based on the individual's demographics. Or we might have a strong physical model, which we can wrap in an equality complex measurement model to account for the various systematic effects introduced in the measurement process. Any such model will be an approximation to the true generative process, but this perspective allows us to build better and better approximations by adding more and more detail as necessary. **Natural way to unite user intuition with explicit modeling. More detail in examples, emphasizing modeling of measurement process.**

Generative models as a way to define small worlds with disintegrations. Does not necessarily imply causal structure but the more casual structure the



(a) (b)

Figure 17: Cartoon of model updating.

better!

7.6 TODO: Checking Model Assumptions with Predictive Performance

Checking model assumptions with posterior predictive checks.