A Conceptual Introduction to Markov Chain Monte Carlo Methods

Joshua S. Speagle

Center for Astrophysics | Harvard & Smithsonian, 60 Garden St., Cambridge, MA 02138, USA

jspeagle@cfa.harvard.edu

Abstract

Markov Chain Monte Carlo (MCMC) methods have become a cornerstone of many modern scientific analyses by providing a straightforward approach to numerically estimate uncertainties in the parameters of a model using a sequence of random samples. This article provides a basic introduction to MCMC methods by establishing a strong conceptual understanding of *what* problems MCMC methods are trying to solve, *why* we want to use them, and *how* they work in theory and in practice. To develop these concepts, I outline the foundations of Bayesian inference, discuss how posterior distributions are used in practice, explore basic approaches to estimate posterior-based quantities, and derive their link to Monte Carlo sampling and MCMC. Using a simple toy problem, I then demonstrate how these concepts can be used to understand the benefits and drawbacks of various MCMC approaches. Exercises designed to highlight various concepts are also included throughout the article.

1 Introduction

Scientific analyses generally rest on making inferences about underlying physical models from various sources of observational data. Over the last few decades, the quality and quantity of these data have increased substantially as they become faster and cheaper to collect and store. At the same time, the same technology that has made it possible to collect vast amounts of data has also led to a substantial increase in the computational power and resources available to analyze them.

Together, these changes have made it possible to explore increasingly complex models using methods that can exploit these computational resources. This has led to a dramatic rise in the number of published works that rely on **Monte Carlo** methods, which use a combination of numerical simulation and random number generation to explore these models.

One particularly popular subset of Monte Carlo methods is known as **Markov Chain Monte Carlo (MCMC)**. MCMC methods are appealing because they provide a straightforward, intuitive way to both simulate values from an unknown distribution and use those simulated values to perform subsequent analyses. This allows them to be applicable in a wide variety of domains.

Owing to its widespread use, various overviews of MCMC methods are common both in peer-reviewed and non-peer-reviewed sources. In general, these tend to fall into two groups: articles focused on various statistical underpinnings of MCMC methods and articles focused on implementation and practical usage. Readers interested in reading more details on either topic are encouraged to see Brooks et al. (2011) and Hogg & Foreman-Mackey (2018) along with associated references therein.

This article instead provides an overview of MCMC methods focused instead on building up a strong *conceptual understanding* of the what, why, and how of MCMC based on statistical intuition. In particular, it tries to systematically answer the following questions:

1. *What* problems are MCMC methods trying to solve?
2. *Why* are we interested in using them?
3. *How* do they work in theory and in practice?

When answering these questions, this article generally assumes that the reader is somewhat familiar with the basics of Bayesian inference in theory (e.g., the role of priors) and in practice (e.g., deriving posteriors), basic statistics (e.g., expectation values), and basic numerical methods (e.g., Riemann sums). No advanced statistical knowledge is required. For more details on these topics, please see Gelman et al. (2013) and Blitzstein & Hwang (2014) along with associated references therein.

The outline of the article is as follows. In §2, I provide a brief review of Bayesian inference and posterior distributions. In §3, I discuss what posteriors are used for in practice, focusing on integration and marginalization. In §4, I outline a basic scheme to approximate these posterior integrals using discrete grids. In §5, I illustrate how Monte Carlo methods emerge as a natural extension of grid-based approaches. In §6, I discuss how MCMC methods fit within the broader scope of possible approaches and their benefits

and drawbacks. In §7, I explore the general challenges MCMC methods face. In §8, I examine how these concepts come together in practice using a simple example. I conclude in §9.

2 Bayesian Inference

In many scientific applications, we have access to some **data** $\mathbf{D}$ that we want to use to make inferences about the world around us. Most often, we want to interpret these data in light of an underlying **model** M that can make predictions about the data we expect to see as a function of some **parameters** $\boldsymbol{\Theta}_M$ of that particular model.

We can combine these pieces together to estimate the **probability** $P(\mathbf{D}|\boldsymbol{\Theta}_M, M)$ that we would actually see that data $\mathbf{D}$ we have collected *conditioned on* (i.e. assuming) a specific choice of parameters $\boldsymbol{\Theta}_M$ from our model M. In other words, assuming our model M is right and the parameters $\boldsymbol{\Theta}_M$ describe the data, what is the **likelihood** $P(\mathbf{D}|\boldsymbol{\Theta}_M, M)$ of the parameters $\boldsymbol{\Theta}_M$ based on the observed data $\mathbf{D}$? Assuming different values of $\boldsymbol{\Theta}_M$ will give different likelihoods, telling us which parameter choices appear to best describe the data we observe.

In Bayesian inference, we are interested in inferring the flipped quantity, $P(\boldsymbol{\Theta}_M|\mathbf{D}, M)$. This describes the probability that the underlying *parameters* are actually $\boldsymbol{\Theta}_M$ given our data $\mathbf{D}$ and assuming a particular model M. By using factoring of probability, we can relate this new probability $P(\boldsymbol{\Theta}_M|\mathbf{D}, M)$ to the likelihood $P(\mathbf{D}|\boldsymbol{\Theta}_M, M)$ described above as

$$P(\boldsymbol{\Theta}_M|\mathbf{D}, M)P(\mathbf{D}|M) = P(\boldsymbol{\Theta}_M, \mathbf{D}|M) = P(\mathbf{D}|\boldsymbol{\Theta}_M, M)P(\boldsymbol{\Theta}_M|M) \tag{1}$$

where $P(\boldsymbol{\Theta}_M, \mathbf{D}|M)$ represents the *joint* probability of having an underlying set of parameters $\boldsymbol{\Theta}_M$ that describe the data and observing the particular set of data $\mathbf{D}$ we have already collected.

Rearranging this equality into a more convenient form gives us **Bayes' Theorem**:

$$P(\boldsymbol{\Theta}_M|\mathbf{D}, M) = \frac{P(\mathbf{D}|\boldsymbol{\Theta}_M, M)P(\boldsymbol{\Theta}_M|M)}{P(\mathbf{D}|M)} \tag{2}$$

This equation now describes exactly how our two probabilities relate to each other.

$P(\boldsymbol{\Theta}_M|M)$ is often referred to as the **prior**. This describes the probability of having a particular set of values $\boldsymbol{\Theta}_M$ for our given model M *before conditioning on our data*. Because this is independent of the data, this term is often interpreted as representing our "prior beliefs" about what $\boldsymbol{\Theta}_M$ *should* be based on previous measurements, physical concerns, and other known factors. In practice, this has the effect of essentially "augmenting" the data with other information.

The denominator

$$P(\mathbf{D}|M) = \int P(\mathbf{D}|\boldsymbol{\Theta}_M, M)P(\boldsymbol{\Theta}_M|M)\mathrm{d}\boldsymbol{\Theta}_M \tag{3}$$

is known as the **evidence** or marginal likelihood for our model M **marginalized** (i.e. integrated) over all possible parameter values $\boldsymbol{\Theta}_M$. This broadly tries to quantify how well

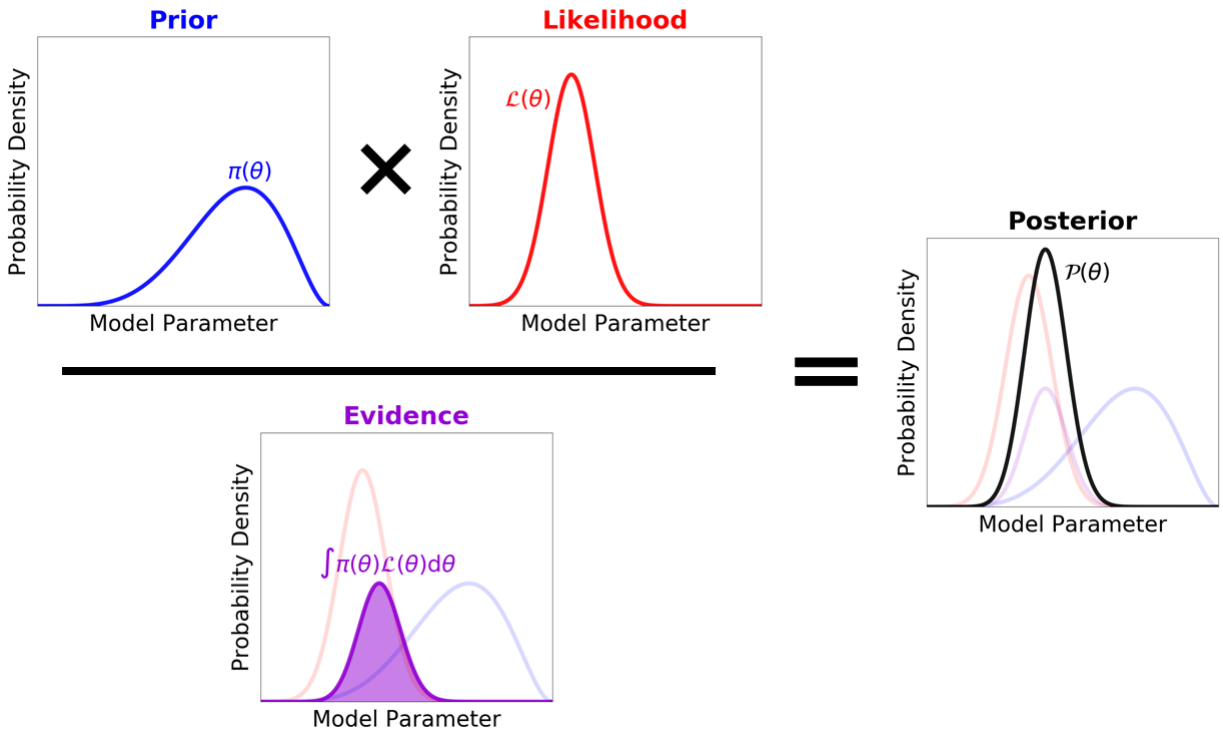

Figure 1: An illustration of Bayes' Theorem. The posterior probability $\mathcal{P}(\boldsymbol{\Theta})$ (black) of our model parameters $\boldsymbol{\Theta}$ is based on a combination of our prior beliefs $\pi(\boldsymbol{\Theta})$ (blue) and the likelihood $\mathcal{L}(\boldsymbol{\Theta})$ (red), normalized by the overall evidence $\mathcal{Z} = \int \pi(\boldsymbol{\Theta})\mathcal{L}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}$ (purple) for our particular model. See §2 for additional details.

our model M explains the data $\mathbf{D}$ after averaging over all possible values $\boldsymbol{\Theta}_M$ of the true underlying parameters. In other words, if the observations predicted by our model look similar to the data $\mathbf{D}$, then M is a good model. Models where this is true more often also tend to be favored over models that give excellent agreement occasionally but disagree most of the time. Since in most instances we take $\mathbf{D}$ as a given, this often ends up being a constant.

Finally, $P(\boldsymbol{\Theta}_M|\mathbf{D}, M)$ represents our **posterior**. This quantifies our belief in $\boldsymbol{\Theta}_M$ after combining our prior intuition $P(\boldsymbol{\Theta}_M|M)$ with current observations $P(\mathbf{D}|\boldsymbol{\Theta}_M, M)$ and normalizing by the overall evidence $P(\mathbf{D}|M)$. The posterior will be some compromise between the prior and the likelihood, with the exact combination depending on the strength and properties of the prior and the quality of the data used to derive the likelihood. A schematic illustration is shown in Figure 1.

Throughout the rest of the paper I will write these four terms (likelihood, prior, evidence, posterior) using shorthand notation such that

$$\mathcal{P}(\boldsymbol{\Theta}) \equiv \frac{\mathcal{L}(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})}{\int \mathcal{L}(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}} \equiv \frac{\mathcal{L}(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})}{\mathcal{Z}} \tag{4}$$

where $\mathcal{P}(\boldsymbol{\Theta}) \equiv P(\boldsymbol{\Theta}_M|\mathbf{D}, M)$ is the posterior, $\mathcal{L}(\boldsymbol{\Theta}) \equiv P(\mathbf{D}|\boldsymbol{\Theta}_M, M)$ is the likelihood,

$\pi(\boldsymbol{\Theta}) \equiv P(\boldsymbol{\Theta}_M|M)$ is the prior, and the constant $\mathcal{Z} \equiv P(\mathbf{D}|M)$ is the evidence. I have suppressed the model M and data $\mathbf{D}$ notation for convenience here since in most cases the data and model are considered fixed, but will re-introduce them as necessary.

Before moving on, I would like to close by emphasizing that *the interpretation of any result is only as good as the models and priors that underlie them*. Trying to explore the implications of any particular model using, for instance, some of the methods described in this article is fundamentally a *secondary concern* behind constructing a reasonable model with well-motivated priors in the first place. I strongly encourage readers to keep this idea in mind throughout the remainder of this work.

Exercise: Noisy Mean

Setup

Consider the case where we have temperature monitoring stations located across a city. Each station i takes a noisy measurement $\hat{T}_i$ of the temperature on any given day with some measurement noise σ_i. We will assume our measurements $\hat{T}_i$ follow a **Normal (i.e. Gaussian) distribution** with mean T and standard deviation σ_i such that

$$\hat{T}_i \sim \mathcal{N}\left[T, \sigma_i\right]$$

This translates into a probability of

$$P(\hat{T}_i|T, \sigma_i) \equiv \mathcal{N}\left[T, \sigma_i\right] = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\frac{(\hat{T}_i - T)^2}{\sigma_i^2}\right]$$

for each observation and

$$P(\{\hat{T}_i\}_{i=1}^n|T, \{\sigma_i\}_{i=1}^n) = \prod_{i=1}^n P(\hat{T}_i|T, \sigma_i)$$

for a collection of n observations.

Let's assume we have five independent noisy measurements of the temperature (in Celsius) from several monitoring stations

$$\hat{T}_1 = 26.3,\ \hat{T}_2 = 30.2,\ \hat{T}_3 = 29.4, \hat{T}_4 = 30.1,\ \hat{T}_5 = 29.8$$

with corresponding uncertainties

$$\sigma_1 = 1.7,\ \sigma_2 = 1.8,\ \sigma_3 = 1.2, \sigma_4 = 0.5,\ \sigma_5 = 1.3$$

Looking at historical data, we find that the typical underlying temperature T during similar days is roughly Normally-distributed with a mean $T_{\rm prior} = 25$ and variation $\sigma_{\rm prior} = 1.5$:

$$T \sim \mathcal{N}\left[T_{\rm prior} = 25, \sigma_{\rm prior} = 1.5\right]$$

Problem

Using these assumptions, compute:

1. the prior $\pi(T)$,
2. the likelihood $\mathcal{L}(T)$, and
3. the posterior $\mathcal{P}(T)$

given our observed data $\{\hat{T}_i\}$ and errors $\{\sigma_i\}$ over a range of temperatures T. How do the three terms differ? Does the prior look like a good assumption? Why or why not?

3 What are Posteriors Good For?

Above, I described how Bayes' Theorem is able to combine our prior beliefs and the observed data into a new posterior estimate $\mathcal{P}(\boldsymbol{\Theta}) \propto \mathcal{L}(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})$. This, however, is only half of the problem. Once we have the posterior, we need to then *use* it to make inferences about the world around us. In general, the ways in which we want to use posteriors fall into a few broad categories:

1. **Making educated guesses**: make a reasonable guess at what the underlying model parameters are.
2. **Quantifying uncertainty**: provide constraints on the range of possible model parameter values.
3. **Generating predictions**: marginalize over uncertainties in the underlying model parameters to predict observables or other variables that depend on the model parameters.
4. **Comparing models**: use the evidences from different models to determine which models are more favorable.

In order to accomplish these goals, we are often more interested in trying to use the posterior to estimate various constraints on the parameters $\boldsymbol{\Theta}$ themselves or other quantities $f(\boldsymbol{\Theta})$ that might be based on them. This often depends on marginalizing over the uncertainties characterized by our posterior (via the likelihood and prior). The evidence $\mathcal{Z}$, for instance, is again just the integral of the likelihood and the prior over all possible parameters:

$$\mathcal{Z} = \int \mathcal{L}(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \equiv \int \tilde{\mathcal{P}}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \tag{5}$$

where $\tilde{\mathcal{P}}(\boldsymbol{\Theta}) \equiv \mathcal{L}(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})$ is the *unnormalized* posterior.

Likewise, if we are investigating the behavior of a subset of "interesting" parameters $\boldsymbol{\Theta}_{\rm int}$ from $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_{\rm int}, \boldsymbol{\Theta}_{\rm nuis}\}$, we want to marginalize over the behavior of the remaining "nuisance" parameters $\boldsymbol{\Theta}_{\rm nuis}$ to see how they can impact $\boldsymbol{\Theta}_{\rm int}$. This process is pretty

straightforward if the entire posterior over $\boldsymbol{\Theta}$ is known:

$$\mathcal{P}(\boldsymbol{\Theta}_{\rm int}) = \int \mathcal{P}(\boldsymbol{\Theta}_{\rm int}, \boldsymbol{\Theta}_{\rm nuis}) \, {\rm d}\boldsymbol{\Theta}_{\rm nuis} = \int \mathcal{P}(\boldsymbol{\Theta}) {\rm d}\boldsymbol{\Theta}_{\rm nuis} \tag{6}$$

Other quantities can generally be derived from the **expectation value** of various parameter-dependent functions $f(\boldsymbol{\Theta})$ with respect to the posterior:

$$\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right] \equiv \frac{\int f(\boldsymbol{\Theta})\mathcal{P}(\boldsymbol{\Theta}){\rm d}\boldsymbol{\Theta}}{\int \mathcal{P}(\boldsymbol{\Theta}){\rm d}\boldsymbol{\Theta}} = \frac{\int f(\boldsymbol{\Theta})\tilde{\mathcal{P}}(\boldsymbol{\Theta}){\rm d}\boldsymbol{\Theta}}{\int \tilde{\mathcal{P}}(\boldsymbol{\Theta}){\rm d}\boldsymbol{\Theta}} = \int f(\boldsymbol{\Theta})\mathcal{P}(\boldsymbol{\Theta}){\rm d}\boldsymbol{\Theta} \tag{7}$$

since $\int \mathcal{P}(\boldsymbol{\Theta}){\rm d}\boldsymbol{\Theta} = 1$ by definition and $\tilde{\mathcal{P}}(\boldsymbol{\Theta}) \propto \mathcal{P}(\boldsymbol{\Theta})$. This represents a weighted average of $f(\boldsymbol{\Theta})$, where at each value $\boldsymbol{\Theta}$ we weight the resulting $f(\boldsymbol{\Theta})$ based on to the chance we believe that value is correct.

Taken together, we see that in almost all cases *we are more interested in computing integrals over the posterior rather than knowing the posterior itself.* To put this another way, the posterior is rarely ever useful on its own; it mainly becomes useful by integrating over it.

This distinction between estimating expectations and other integrals over the posterior versus estimating the posterior in-and-of-itself is a key element of Bayesian inference. This distinction is hugely important when it comes to actually performing inference in practice, since it is often the case that we can get an excellent estimate of $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$ even if we have an extremely poor estimate of $\mathcal{P}(\boldsymbol{\Theta})$ or $\tilde{\mathcal{P}}(\boldsymbol{\Theta})$.

More details are provided below to further illustrate how the particular categories described above translate into particular integrals over the (unnormalized) posterior. An example is shown in Figure 2.

3.1 Making Educated Guesses

One of the core tenets of Bayesian inference is that we don't know the true model M_* or its true underlying parameters $\boldsymbol{\Theta}_*$ that characterize the data we observe: the model M we have is almost always a simplification of what is actually going on. If we assume that our current model M is correct, however, we can try to use our posterior $\mathcal{P}(\boldsymbol{\Theta})$ to propose a **point estimate** $\hat{\boldsymbol{\Theta}}$ that we think is a pretty good guess for the true value $\boldsymbol{\Theta}_*$.

What exactly counts as "good"? This depends on exactly what we care about. In general, we can quantify "goodness" by asking the opposite question: how badly are we penalized if our estimate $\hat{\boldsymbol{\Theta}} \neq \boldsymbol{\Theta}_*$ is wrong? This is often encapsulated through the use of a **loss function** $L(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta}_*)$ that penalizes us when our point estimate $\hat{\boldsymbol{\Theta}}$ differs from $\boldsymbol{\Theta}_*$. An example of a common loss function is $L(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta}_*) = |\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*|^2$ (i.e. squared loss), where an incorrect guess is penalized based on the square of the magnitude of the separation between the guess $\hat{\boldsymbol{\Theta}}$ and the true value $\boldsymbol{\Theta}_*$.

Unfortunately, we don't know what the actual value of $\boldsymbol{\Theta}_*$ is to evaluate the true loss. We can, however, do the next best thing and compute the **expected loss** averaged over all possible values of $\boldsymbol{\Theta}_*$ based on our posterior:

$$L_{\mathcal{P}}(\hat{\boldsymbol{\Theta}}) \equiv \mathbb{E}_{\mathcal{P}}\left[L(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta})\right] = \int L(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta})\mathcal{P}(\boldsymbol{\Theta}){\rm d}\boldsymbol{\Theta} \tag{8}$$

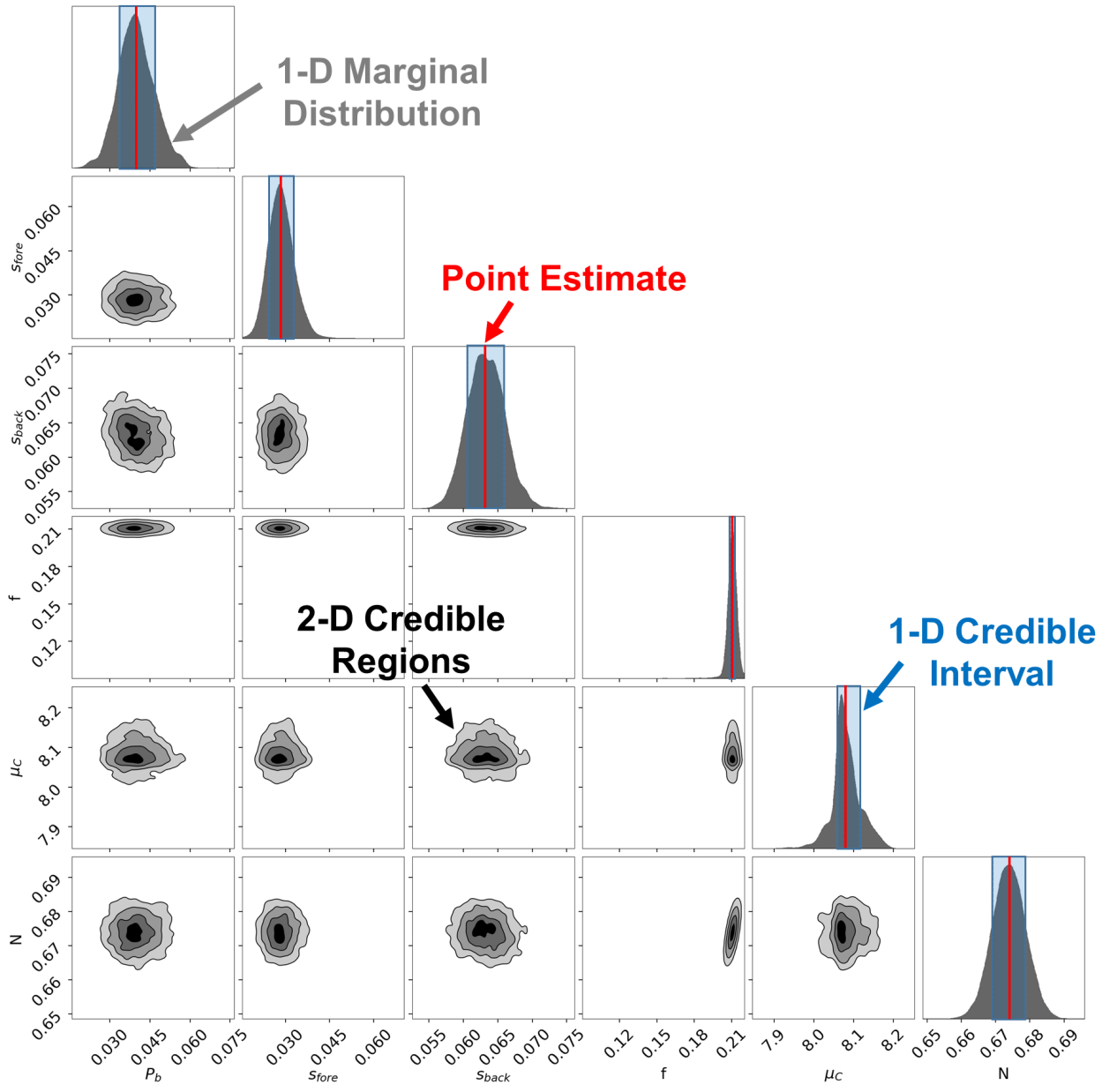

Figure 2: A "corner plot" showing an example of how posteriors are used in practice. Each of the top panels shows the 1-D marginalized posterior distribution for each parameter (grey), along with associated median point estimates (red) and 68% credible intervals (blue). Each central panel shows the 10%, 40%, 65%, and 85% credible regions for each 2-D marginalized posterior distribution. See §3 for additional details.

A reasonable choice for $\hat{\boldsymbol{\Theta}}$ is then the value that minimizes this expected loss in place of the actual (unknown) loss:

$$\hat{\boldsymbol{\Theta}} \equiv \underset{\boldsymbol{\Theta}'}{\operatorname{argmin}} \left[L_{\mathcal{P}}(\boldsymbol{\Theta}') \right] \tag{9}$$

where argmin indicates the value (argument) of $\boldsymbol{\Theta}'$ that minimizes the expected loss $L_{\mathcal{P}}(\boldsymbol{\Theta}')$.

While this strategy can work for any arbitrary loss function, solving for $\hat{\boldsymbol{\Theta}}$ often requires using numerical methods and repeated integration over $\mathcal{P}(\boldsymbol{\Theta})$. However, analytic solutions do exist for particular loss functions. For example, it is straightforward to show (and an insightful exercise for the interested reader) that the optimal point estimate $\hat{\boldsymbol{\Theta}}$ under squared loss is simply the mean.

3.2 Quantifying Uncertainty

In many cases we are not just interested in computing a prediction $\hat{\boldsymbol{\Theta}}$ for $\boldsymbol{\Theta}_*$, but also constraining a region $\mathcal{C}(\boldsymbol{\Theta})$ of possible values within which $\boldsymbol{\Theta}_*$ might lie with some amount of certainty. In other words, can we construct a region $\mathcal{C}_X$ such that we believe there is an $X\%$ chance that it contains $\boldsymbol{\Theta}_*$?

There are many possible definitions for this **credible region**. One common definition is the region above some posterior threshold $\mathcal{P}_X$ where $X\%$ of the posterior is contained, i.e. where

$$\int_{\boldsymbol{\Theta} \in \mathcal{C}_X} \mathcal{P}(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} = \frac{X}{100} \tag{10}$$

given

$$\mathcal{C}_X \equiv \{\boldsymbol{\Theta} : \mathcal{P}(\boldsymbol{\Theta}) \geq \mathcal{P}_X\} \tag{11}$$

In other words, we want to integrate our posterior over all $\boldsymbol{\Theta}$ where the value $\mathcal{P}(\boldsymbol{\Theta}) > \mathcal{P}_X$ is greater than some threshold $\mathcal{P}_X$, where $\mathcal{P}_X$ is set so that this integral encompasses $X\%$ of the full posterior. Common choices for X include 68% and 95% (i.e. "1-sigma" and "2-sigma" credible intervals).

In the special case where our (marginalized) posterior is 1-D, **credible intervals** are often defined using **percentiles** rather than thresholds, where the location x_p of the pth percentile is defined as

$$\int_{-\infty}^{x_p} \mathcal{P}(x) \mathrm{d}x = \frac{p}{100} \tag{12}$$

We can use these to define a credible region $[x_{\rm low}, x_{\rm high}]$ containing $Y\%$ of the data by taking $x_{\rm low} = x_{(1-Y)/2}$ and $x_{\rm high} = x_{(1+Y)/2}$. While this leads to asymmetric thresholds and does not generalize to higher dimensions, it has the benefit of always encompassing the median value x_{50} and having equal tail probabilities (i.e. $(1-Y)/2\%$ of the posterior on each side).

In general, when referring to "credible intervals" throughout the text the percentile definition should be assumed unless explicitly stated otherwise.

3.3 Making Predictions

In addition to trying to estimate the underlying parameters of our model, we often also want to make predictions of other observables or variables that depend on our model parameters. If we think we know the underlying true model parameters $\boldsymbol{\Theta}_*$, then this process is straightforward. Given that we only have access to the posterior distribution $\mathcal{P}(\boldsymbol{\Theta})$ over possible values $\boldsymbol{\Theta}_*$ could take, however, to predict what will happen we will need to marginalize over this uncertainty.

We can quantify this intuition using the **posterior predictive** $P(\tilde{\mathbf{D}}|\mathbf{D})$, which represents the probability of seeing some new data $\tilde{\mathbf{D}}$ based on our existing data $\mathbf{D}$:

$$P(\tilde{\mathbf{D}}|\mathbf{D}) \equiv \int P(\tilde{\mathbf{D}}|\boldsymbol{\Theta})P(\boldsymbol{\Theta}|\mathbf{D})\mathrm{d}\boldsymbol{\Theta} \equiv \int \tilde{\mathcal{L}}(\boldsymbol{\Theta})\mathcal{P}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} = \mathbb{E}_{\mathcal{P}}\left[\tilde{\mathcal{L}}(\boldsymbol{\Theta})\right] \tag{13}$$

In other words, for hypothetical data $\tilde{\mathbf{D}}$, we want to compute the expected value of the likelihood $\tilde{\mathcal{L}}(\boldsymbol{\Theta})$ over all possible values of $\boldsymbol{\Theta}$ based on the current posterior $\mathcal{P}(\boldsymbol{\Theta})$.

3.4 Comparing Models

One final point of interest in many Bayesian analyses is trying to investigate whether the data particularly favors any of the model(s) we are assuming in our analysis. Our choice of priors or the particular way we parameterize the data can lead to substantial differences in the way we might want to interpret our results.

We can compare two models by computing the **Bayes factor**:

$$\mathcal{R}_2^1 \equiv \frac{P(M_1|\mathbf{D})}{P(M_2|\mathbf{D})} = \frac{P(\mathbf{D}|M_1)P(M_1)}{P(\mathbf{D}|M_2)P(M_2)} \equiv \frac{\mathcal{Z}_1}{\mathcal{Z}_2}\frac{\pi_1}{\pi_2} \tag{14}$$

where $\mathcal{Z}_M$ is again the evidence for model M and π_M is our prior belief that M is correct relative to the competing model. Taken together, the Bayes factor $\mathcal{R}$ tells us how much a particular model is favored over another given the observed data, marginalizing over all possible values of the underlying model parameters $\boldsymbol{\Theta}_M$, and our previous relative confidence in the model.

Again, note that computing $\mathcal{Z}_M$ requires computing the integral $\int \tilde{\mathcal{P}}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}$ of the unnormalized posterior $\tilde{\mathcal{P}}(\boldsymbol{\Theta})$ over $\boldsymbol{\Theta}$. Combined with the other examples outlined in this section, it is clear that many common use cases in Bayesian analysis rely on computing integrals over the (possibly unnormalized) posterior.

Exercise: Noisy Mean Revisited

Setup

Let's return to our temperature posterior $\mathcal{P}(T)$ from §2. We want to use this result to derive interesting estimates and constraints on the possible underlying temperature T.

Point Estimates

The **mean** can be defined as the point estimate $\hat{\boldsymbol{\Theta}}$ that minimizes the expected loss $L_{\mathcal{P}}(\hat{\boldsymbol{\Theta}})$ under **squared loss**:

$$L_{\mathrm{mean}}(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta}_*) = |\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*|^2$$

The **median** can be defined as the point estimate that minimizes $L_{\mathcal{P}}(\hat{\boldsymbol{\Theta}})$ under **absolute loss**:

$$L_{\mathrm{med}}(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta}_*) = |\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*|$$

And the **mode** can be defined as the point estimate that minimizes $L_{\mathcal{P}}(\hat{\boldsymbol{\Theta}})$ under **"catastrophic" loss**:

$$L_{\mathrm{mode}}(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta}_*) = -\delta(|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*|)$$

where $\delta(\cdot)$ is the **Dirac delta function** defined such that

$$\int f(x)\delta(x-a)\mathrm{d}x = f(a)$$

Given these expressions for the mean, median, and mode, estimate the corresponding temperature point estimate T_{mean}, T_{med}, and T_{mode} from our corresponding posterior. Feel free to experiment with various analytic and numerical methods to perform these calculations.

We might expect that the historical data we used for our priors might not hold as well today if there have been some long-term changes in the average temperature. For instance, we expect that the average temperature has increased over time, and so we might not want to penalize hotter temperatures $T \geq T_{\mathrm{prior}}$ as much as cooler ones $T < T_{\mathrm{prior}}$. We can encode this information in an asymmetric loss function such as

$$L(\hat{T}|T_*) = \begin{cases} |\hat{T} - T_*|^3 & T < T_{\mathrm{prior}} \\ |\hat{T} - T_*| & T \geq T_{\mathrm{prior}} \end{cases}$$

What is the optimal point estimate T_{asym} that minimizes the expected loss in this case?

Credible Intervals

Next, let's try to quantify the uncertainty. Given the posterior $\mathcal{P}(T)$, compute the 50%, 80%, and 95% credible intervals using posterior thresholds $\mathcal{P}_X$. Next, compute these credible intervals using percentiles. Are there are differences between the credible intervals computed from the two methods? Why or why not?

Posterior Predictive

To propagate our uncertainties into the next observations, compute the posterior predictive $P(\hat{T}_6|\{\hat{T}_1, \ldots, \hat{T}_5\})$ over a range of possible temperature measurements $\hat{T}_6$ for the next observations given the previous five $\{\hat{T}_1, \ldots, \hat{T}_5\}$ assuming an uncertainty of $\sigma_6 = 0$, $\sigma_6 = 0.5$, and $\sigma_6 = 2$.

Model Comparison

Finally, we want to investigate whether our prior appears to be a good assumption. Using numerical methods, compute the evidence $\mathcal{Z}$ for our default prior with mean $T_{\rm prior} = 25$ and standard deviation $\sigma_{\rm prior} = 1.5$. Then compare this to the evidence estimated based on an alternative prior where we assume the temperature has risen by roughly five degrees with mean $T_{\rm prior} = 30$ but with a corresponding larger uncertainty $\sigma_{\rm prior} = 3$. Is one model particularly favored over the other?

4 Approximating Posterior Integrals with Grids

I now want to investigate methods for estimating posterior integrals. While in some cases (e.g., conjugate priors) these can be computed analytically, this is not true in general. To properly estimate quantities such as those outlined in §3 therefore requires the use of numerical methods (highlighted in the previous exercises).

To start, I will first focus on the case where our integral over $\boldsymbol{\Theta}$ is 1-D. In that case, we can approximate it using standard numerical techniques such as a **Riemann sum** over a **discrete grid** of points:

$$\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right] = \int f(\boldsymbol{\Theta})\mathcal{P}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \approx \sum_{i=1}^{n} f(\boldsymbol{\Theta}_i)\mathcal{P}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i \tag{15}$$

where

$$\Delta\boldsymbol{\Theta}_i = \boldsymbol{\Theta}_{j+1} - \boldsymbol{\Theta}_j \tag{16}$$

is simply the spacing between the set of $j = 1, \ldots, n+1$ points on the underlying grid and

$$\boldsymbol{\Theta}_i = \frac{\boldsymbol{\Theta}_{j+1} + \boldsymbol{\Theta}_j}{2} \tag{17}$$

is just defined to be the mid-point between $\boldsymbol{\Theta}_j$ and $\boldsymbol{\Theta}_{j+1}$.[1] As shown in Figure 3, this approach is akin to trying to approximate the integral using a discrete set of n rectangles with heights of $f(\boldsymbol{\Theta}_i)\mathcal{P}(\boldsymbol{\Theta}_i)$ and widths of $\Delta\boldsymbol{\Theta}_i$.

This idea can be generalized to higher dimensions. In that case, instead of breaking up the integral into n 1-D segments, we instead can decompose it into a set of n N-D cuboids. The contribution of each of these pieces is then proportional to the product of the "height" $f(\boldsymbol{\Theta}_i)\mathcal{P}(\boldsymbol{\Theta}_i)$ and the *volume*

$$\Delta\boldsymbol{\Theta}_i = \prod_{j=1}^{d} \Delta\Theta_{i,j} \tag{18}$$

where $\Delta\Theta_{i,j}$ is the width of the ith cuboid in the jth dimension. See Figure 3 for a visual representation of this procedure.

[1] Choosing $\boldsymbol{\Theta}_i$ to be one of the end-points gives consistent behavior (see §4.3) as the number of grid points $n \to \infty$ but generally leads to larger biases for finite n.

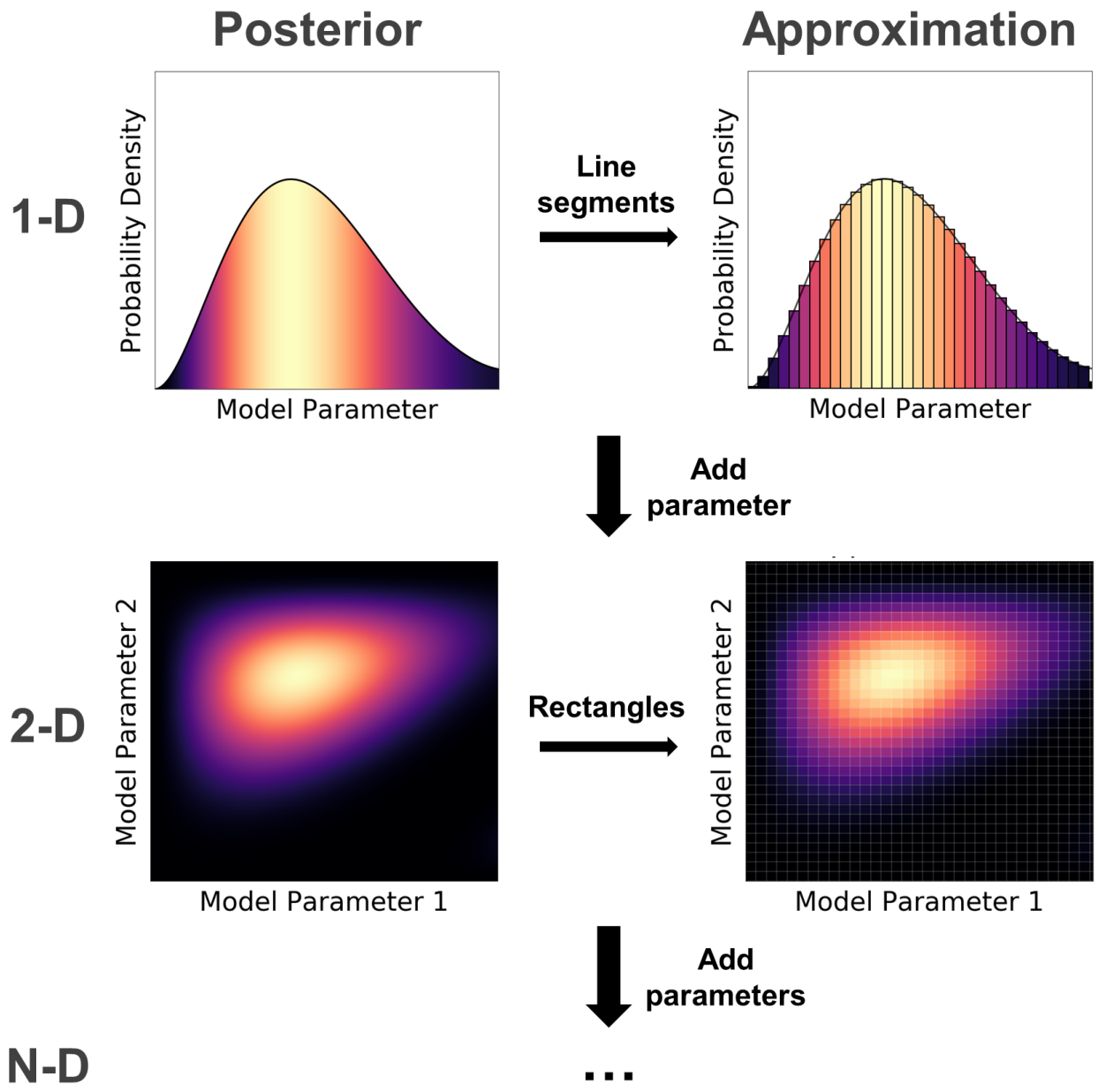

Figure 3: An illustration of how to approximate posterior integrals using a discrete grid of points. We break up the posterior into contiguous regions defined by a position $\boldsymbol{\Theta}_i$ (e.g., an endpoint or midpoint) with corresponding posterior density $\mathcal{P}(\boldsymbol{\Theta}_i)$ and volume $\Delta\boldsymbol{\Theta}_i$ over a grid with $i = 1, \ldots, n$ elements. Our integral can then be approximated by adding up each of these regions proportional to the posterior mass $\mathcal{P}(\boldsymbol{\Theta}_i) \times \Delta\boldsymbol{\Theta}_i$ contained within it. In 1-D (top), these volume elements $\Delta\boldsymbol{\Theta}_i$ correspond to line segments while in 2-D (middle), these correspond to rectangles. This can be generalized to higher dimensions (bottom), where we instead used N-D cuboids. See §4 for additional details.

Substituting $\mathcal{P}(\boldsymbol{\Theta}) = \tilde{\mathcal{P}}(\boldsymbol{\Theta})/\mathcal{Z}$ into the expectation value and replacing any integrals with their grid-based approximations then gives:

$$\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right] = \frac{\int f(\boldsymbol{\Theta})\mathcal{P}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}}{\int \mathcal{P}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}} = \frac{\int f(\boldsymbol{\Theta})\tilde{\mathcal{P}}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}}{\int \tilde{\mathcal{P}}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}} \approx \frac{\sum_{i=1}^{n} f(\boldsymbol{\Theta}_i)\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i}{\sum_{i=1}^{n} \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i} \tag{19}$$

Note the denominator is now an estimate for the evidence:

$$\mathcal{Z} = \int \tilde{\mathcal{P}}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \approx \sum_{i=1}^{n} \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_j \tag{20}$$

This substitution of the unnormalized posterior $\tilde{\mathcal{P}}(\boldsymbol{\Theta})$ for the posterior $\mathcal{P}(\boldsymbol{\Theta})$ is a crucial part of computing expectation values in practice since we can compute $\tilde{\mathcal{P}}(\boldsymbol{\Theta}) = \mathcal{L}(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})$ directly without knowing $\mathcal{Z}$.

4.1 The Curse of Dimensionality

While this approach is straightforward, it has one immediate and severe drawback: the total number of grid points increases *exponentially* as the number of dimensions increases. For example, assuming we have roughly $k \geq 2$ grid points in each dimensions, the total number of points n in our grid scales as

$$n \sim \prod_{j=1}^{d} k = k^d \tag{21}$$

This means that even in the absolute *best* case where $k = 2$, we have 2^d scaling.

This awful scaling is often referred to as the **curse of dimensionality**. This exponential dependence turns out to be a generic feature of high-dimensional distributions (i.e. posteriors of models with larger numbers of parameters) that I will return to later in §7.

4.2 Effective Sample Size

Apart from this exponential scaling of dimensionality, there is a more subtle drawback to using grids. Since we do not know the shape of the distribution ahead of time, the contribution of each portion of the grid (i.e. each N-D cuboid) can be highly uneven depending on the structure of the grid. In other words, the effectiveness of this approach not only depends on the *number* of grid points n but also *where* they are allocated. If we do not specify our grid points well, we can end up with many points located in regions where $\tilde{\mathcal{P}}(\boldsymbol{\Theta})$ and/or $f(\boldsymbol{\Theta})\tilde{\mathcal{P}}(\boldsymbol{\Theta})$ is relatively small. This then implies that their respective sums will be dominated by a small number of points with much larger relative "weights". Ideally, we would want to increase the resolution of the grid in regions where the posterior is large and decrease it elsewhere to mitigate this effect.

Note that our use of the term "weights" in the preceding paragraph is quite deliberate. Looking back at our original approximation, the form of equation (19) is quite similar to one which might be used to compute a **weighted sample mean** of $f(\boldsymbol{\Theta})$. In that case,

where we have n observations $\{f_1, \ldots, f_n\}$ with corresponding weights $\{w_1, \ldots, w_n\}$, the weighted mean is simply:

$$\hat{f}_{\rm mean} \equiv \frac{\sum_{i=1}^{n} w_i f_i}{\sum_{i=1}^{n} w_i} \tag{22}$$

Indeed, if we define

$$f_i \equiv f(\boldsymbol{\Theta}_i), \quad w_i \equiv \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i \tag{23}$$

then the connection between the weighted sample mean in equation (22) and the expectation value from our grid in equation (19) becomes explicit:

$$\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right] \approx \frac{\sum_{i=1}^{n} f(\boldsymbol{\Theta}_i)\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i}{\sum_{i=1}^{n} \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i} \equiv \frac{\sum_{i=1}^{n} w_i f_i}{\sum_{i=1}^{n} w_i} \tag{24}$$

Thinking about our grid as a set of n samples also allows us to consider an associated **effective sample size (ESS)** $n_{\rm eff} \leq n$. The ESS encapsulates the idea that not all of our samples contribute the same amount of information: if we have n samples that are very similar to each other, we expect to have a substantially worse estimate than if we have n samples that are quite different. This is because the information in correlated samples are at least partially redundant with one another, with the amount of redundancy increasing with the strength of the correlation: while two independent samples provide completely unique information about the distribution and no information about each other, two correlated samples instead provide some information about each other at the expense of the underlying distribution.

Returning to grids, this correspondence means that we can in theory come up with an estimate of the expectation value $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$ that is *at least* as good as the one we might currently have using a smaller number $n_{\rm eff} \leq n$ of grid points *if* we were able to allocate them more efficiently. This distinction matters because errors on our estimate of the expectation value generally scale as a function of $n_{\rm eff}$ rather than n. For instance, the error on the mean typically goes as $\propto n_{\rm eff}^{-1/2}$ rather than $\propto n^{-1/2}$.

We can quantify the ideas behind the ESS as discussed above by introducing a formal definition following Kish (1965):

$$n_{\rm eff} \equiv \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2} \tag{25}$$

In line with our intuition, the best case under this definition is one where all the weights are equal ($w_i = w$):

$$n_{\rm eff}^{\rm best} = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2} = \frac{(nw)^2}{\sum_{i=1}^{n} w^2} = \frac{n^2 w^2}{n w^2} = n \tag{26}$$

Likewise, the worst case is one where all the weight is concentrated around a single sample ($w_i = w$ for $i = j$ and $w_i = 0$ otherwise):

$$n_{\rm eff}^{\rm worst} = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2} = \frac{(w)^2}{w^2} = 1 \tag{27}$$

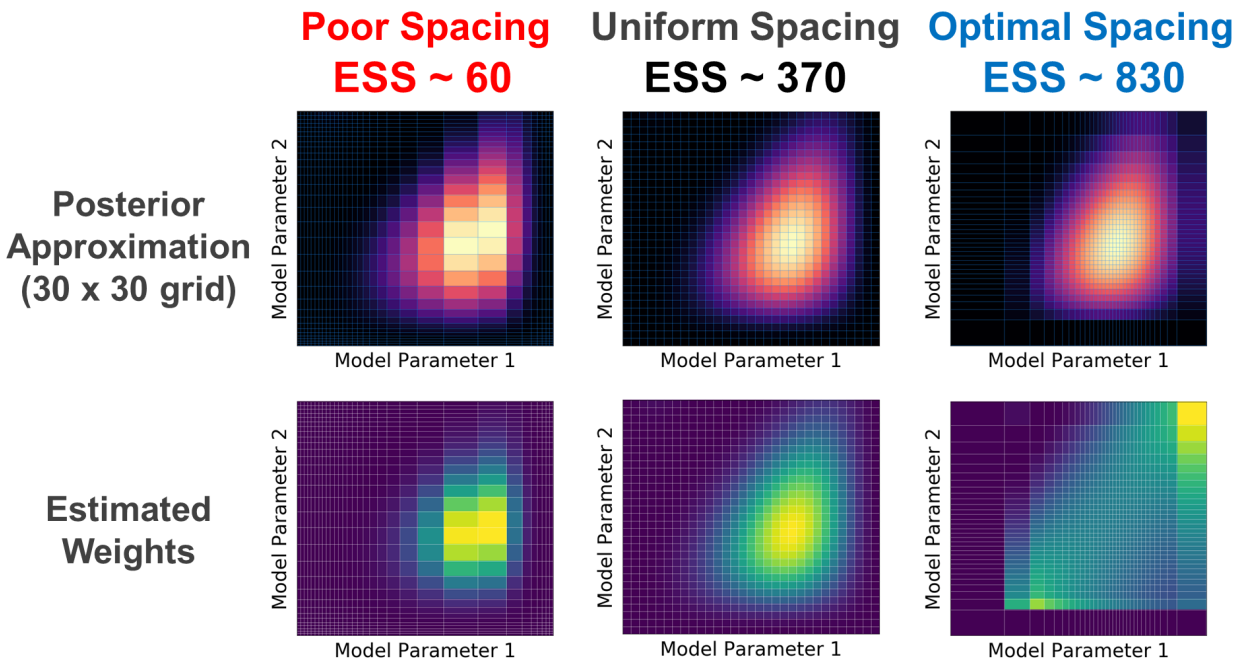

Figure 4: An example of how changing the spacing (volume elements) of the grid can dramatically affect its associated estimate of posterior integrals. On a toy 2-D posterior $\mathcal{P}(\boldsymbol{\Theta})$, simply changing the spacing of the associated 2-D 30×30 grid dramatically affects the effective sample size (ESS) (see §4.2). Differences between poor spacing (left), uniform spacing (middle), and optimal spacing (right) leads to an order of magnitude difference in the ESS, as highlighted by the distribution of weights (bottom) associated with the volume elements of each grid. See §4.2 for additional details.

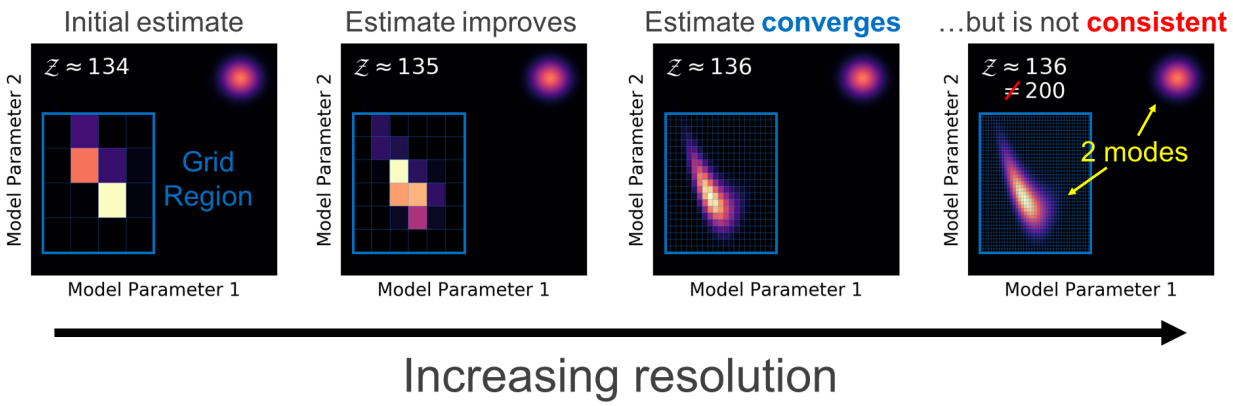

Figure 5: An illustration of how grid-based estimates can be *convergent* (i.e. converge to a single value as the number of grid points increases) but not *consistent* (i.e. the value it converges to is not the correct answer). Our toy 2-D unnormalized posterior $\tilde{\mathcal{P}}(\boldsymbol{\Theta})$ has two modes that are well-separated with a total evidence of $\mathcal{Z} = 200$. If we are not aware of the second mode, we might define a grid region that only encompasses a subset of the entire parameter space (left). While increasing the resolution of the grid within this region allows the estimated $\mathcal{Z}$ to converge to an single answer (left to right), this is not equal to the correct answer of $\mathcal{Z} = 200$ because we have neglected the contribution of the other component (right). See §4.3 for additional details.

This former situation (with $n_{\rm eff}^{\rm best}$) would be the case where each of the elements of our grid all have roughly the same contribution to the integral, while the latter (with $n_{\rm eff}^{\rm worst}$) would be where the entire integral is essentially contained in just one of our n N-D cuboid regions. An illustration of this behavior is shown in Figure 4.

4.3 Convergence and Consistency

Now that I have outlined the relationship between the structure of our grid and the ESS, I want to examine two final issues: **convergence** and **consistency**. Convergence is the idea that, while our estimates using n samples (grid points) might be noisy, it approaches some fiducial value as $n \to \infty$:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} f(\boldsymbol{\Theta}_i)\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i}{\sum_{i=1}^{n} \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i} = C \tag{28}$$

Consistency is subsequently the idea that the value we converge to is the true value we are interested in estimating:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} f(\boldsymbol{\Theta}_i)\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i}{\sum_{i=1}^{n} \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i} = \mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right] \tag{29}$$

It is straightforward to show that *if* the expectation value is well-defined (i.e. it exists) *and* the grid covers the entire domain of $\boldsymbol{\Theta}$ (i.e. spans the smallest and largest possible

values in every dimension) then using a grid is a **consistent** way to estimate the expectation value. This should make intuitive sense: provided our grid is expansive enough in $\boldsymbol{\Theta}$ so that we're not "missing" any region of parameter space, we should be able to estimate $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$ to arbitrary precision by simply increasing the resolution in $\Delta\boldsymbol{\Theta}$.

Unfortunately, we do not know beforehand what range of values of $\boldsymbol{\Theta}$ our grid should span. While parameters can range over $(-\infty, +\infty)$, grids rely on finite-volume elements and so we have to choose some finite sub-space to grid up. So while grids may give estimates that converge to some value over the range spanned by the grid points, there is always a possibility that a significant portion of the posterior lies outside that range. In these cases, grids are not guaranteed to be consistent estimators of $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$. An illustration of this issue is shown in Figure 5. This fundamental problem is not shared by Monte Carlo methods, which I will cover in §5.

Exercise: Grids over a 2-D Gaussian

Setup

Consider an unnormalized posterior well-approximated by a 2-D Gaussian (Normal) distribution centered on (μ_x, μ_y) with standard deviations (σ_x, σ_y):

$$\tilde{\mathcal{P}}(x, y) = \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right\}$$

Assume that we expect to find our posterior has a mean of 0 and a standard deviation of 1. In reality, however, our posterior actually has means $(\mu_x, \mu_y) = (-0.3, 0.8)$ and standard deviations $(\sigma_x^2, \sigma_y^2) = (2, 0.5)$, mimicking the common case where our prior expectations and posterior inferences somewhat disagree.

Grid-based Estimation

We want to use a 2-D grid to estimate various forms of posterior integrals. Starting with an evenly-spaced 5×5 grid from $[-2, 2]$, compute:

1. the evidence $\mathcal{Z}$,
2. the means $\mathbb{E}_{\mathcal{P}}\left[x\right]$ and $\mathbb{E}_{\mathcal{P}}\left[y\right]$,
3. the 68% credible intervals (or closest approximation) $[x_{\rm low}, x_{\rm high}]$ and $[y_{\rm low}, y_{\rm high}]$,
4. and the effective sample size $n_{\rm eff}$.

How accurate are each of these quantities with the values we might expect? What does $n_{\rm eff}/n$ tell us about how efficiently we have allocated our grid points?

Convergence

Repeat the above exercise using an evenly-spaced grid of 20×20 points and 100×100 points. Comment on any differences. How much has the overall accuracy improved? Do the estimates appear convergent?

Consistency

Next, expand the bounds of the grid to be from $[-5, 5]$ and perform the same exercise as above. Do the answers change substantially? If so, what does this tell us about the consistency of our previous estimates? Adjust the density and bounds of the grid until the answers appear both convergent and consistent. Remember that we do not know the exact shape of the posterior ahead of time. What does this imply about general concerns when applying grids in practice?

Effective Sample Size

Finally, explore whether there is a straightforward scheme to adjust the locations of the x and y grid points to maximize the effective sample size based on the definition outlined in §4.2. If so, can you explain why it works? If not, why not? Compared to equivalent evenly-spaced grids, how much can adaptively adjusting the grid spacing improve $n_{\rm eff}$ and the overall accuracy of our estimates?

5 From Grids to Monte Carlo Methods

5.1 Connecting Grid Points and Samples

Earlier, I outlined how we can relate estimating $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$ using a grid of n points to an equivalent estimate using a set of n samples $\{f_1, \ldots, f_n\}$ and a series of associated weights $\{w_1, \ldots, w_n\}$. The main result is that there is an intimate connection between the structure of the posterior and the grid to the relative amplitude of the weights $w_i \equiv \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i$ for each point $f_i \equiv f(\boldsymbol{\Theta}_i)$. Adjusting the resolution of the grid then affects these weights, with a more uniform distribution of weights leading to a larger ESS which can improve our estimate.

The fact that decreasing the spacing (making grid denser) also decreases the weights makes sense: we have more points located in that region, so each point should in general get less relative weight when computing $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$. Likewise, if we have the same spacing but change the relative shape of the posterior, the weight of that point when estimating $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$ should also change accordingly.

I now want to extend this basic relationship further. In theory, adaptively increasing the resolution of our grid allows us more control over the volume elements $\Delta\boldsymbol{\Theta}_i$ used to derive our weights. If we knew the shape of our posterior sufficiently well, for large n we should in theory be able to adjust $\Delta\boldsymbol{\Theta}_i$ such that the weights $w_i = \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i$ are uniform to some amount of desired precision. By inspection, this should happen when

$$\Delta\boldsymbol{\Theta}_i \propto \frac{1}{\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)} \tag{30}$$

for all i.

Taking this reasoning to its conceptual limit, as $n \to \infty$ we can imagine estimating the posterior using a larger and larger number of grid points whose spacing $\Delta\boldsymbol{\Theta}$ changes as

a function of $\boldsymbol{\Theta}$. Using this, we can now define the density of points $\mathcal{Q}(\boldsymbol{\Theta})$ based on the varying resolution $\Delta\boldsymbol{\Theta}(\boldsymbol{\Theta})$ of our infinitely-fine grid as a function of $\boldsymbol{\Theta}$:

$$\mathcal{Q}(\boldsymbol{\Theta}) \propto \frac{1}{\Delta\boldsymbol{\Theta}(\boldsymbol{\Theta})} \tag{31}$$

This result suggests that, in the continuum limit where $n \to \infty$, *the structure of our infinite-resolution grid is equivalent to a new continuous distribution* $\mathcal{Q}(\boldsymbol{\Theta})$. An illustration of this concept is shown in Figure 6. Using $\mathcal{Q}(\boldsymbol{\Theta})$, we can then rewrite our original expectation value as

$$\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right] \equiv \frac{\int f(\boldsymbol{\Theta})\tilde{\mathcal{P}}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}}{\int \tilde{\mathcal{P}}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}} = \frac{\int f(\boldsymbol{\Theta})\frac{\tilde{\mathcal{P}}(\boldsymbol{\Theta})}{\mathcal{Q}(\boldsymbol{\Theta})}\mathcal{Q}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}}{\int \frac{\tilde{\mathcal{P}}(\boldsymbol{\Theta})}{\mathcal{Q}(\boldsymbol{\Theta})}\mathcal{Q}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}} = \frac{\mathbb{E}_{\mathcal{Q}}\left[f(\boldsymbol{\Theta})\tilde{\mathcal{P}}(\boldsymbol{\Theta})/\mathcal{Q}(\boldsymbol{\Theta})\right]}{\mathbb{E}_{\mathcal{Q}}\left[\tilde{\mathcal{P}}(\boldsymbol{\Theta})/\mathcal{Q}(\boldsymbol{\Theta})\right]} \tag{32}$$

For reasons that will soon become clear, I will refer to $\mathcal{Q}(\boldsymbol{\Theta})$ as the **proposal distribution**.

At this point, this may mostly seem like a mathematical trick: all I have done is rewrite our original *single* expectation value with respect to the (unnormalized) posterior $\tilde{\mathcal{P}}(\boldsymbol{\Theta})$ in terms of *two* expectation values with respect to the proposal distribution $\mathcal{Q}(\boldsymbol{\Theta})$. This substitution, however, actually allows us to fully realize the connection between grid points and samples.

Earlier, I showed that the estimate for the expectation value from grid points is exactly analogous to the estimate we would derive assuming the grid points were random samples $\{f_1, \ldots, f_n\}$ with associated weights $\{w_1, \ldots, w_n\}$. Once we have defined our expectation with respect to $\mathcal{Q}(\boldsymbol{\Theta})$, however, this statement can become exact assuming we can explicitly generate samples from $\mathcal{Q}(\boldsymbol{\Theta})$.

Let's quickly review what this means. Initially, we looked at trying to estimate $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$ over a grid with n points. In the limit of infinite resolution, however, our grid becomes equivalent to some distribution $\mathcal{Q}(\boldsymbol{\Theta})$. Using $\mathcal{Q}(\boldsymbol{\Theta})$, we can then rewrite our original expression in terms of two expectations, $\mathbb{E}_{\mathcal{Q}}\left[f(\boldsymbol{\Theta})\tilde{\mathcal{P}}(\boldsymbol{\Theta})/\mathcal{Q}(\boldsymbol{\Theta})\right]$ and $\mathbb{E}_{\mathcal{Q}}\left[\tilde{\mathcal{P}}(\boldsymbol{\Theta})/\mathcal{Q}(\boldsymbol{\Theta})\right]$, over $\mathcal{Q}(\boldsymbol{\Theta})$ instead of $\mathcal{P}(\boldsymbol{\Theta})$. This helps us because we can in theory estimate these final expressions explicitly using a series of n randomly generated samples from $\mathcal{Q}(\boldsymbol{\Theta})$. Due to the randomness inherent in this approach, this is commonly referred to as a **Monte Carlo** approach for estimating $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$ due to historical connections with randomness and gambling.

On the face of it, this should come across as a surprising claim. When we compute an integral of a function $f(\boldsymbol{\Theta})$ on a bounded grid, we know that there is some error in our approximation having to do with the discretization of the grid. This error is entirely *deterministic*: given a number of grid points n and an a particular discretization density $\mathcal{Q}(\boldsymbol{\Theta}) \propto 1/\Delta\boldsymbol{\Theta}(\boldsymbol{\Theta})$, we will get the same result (and error) for $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$ every time.

By contrast, drawing n samples $\{\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_n\}$ from $\mathcal{Q}(\boldsymbol{\Theta})$ is an inherently *random* (i.e. stochastic) process that seems to look nothing like a grid of points. And because these points are inherently random, the actual deviation between our estimate and the true value of $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$ will *also* be random. The "error" from random samples then tells us something about how much we expect our estimate can differ over many possible realizations of our random process given a particular number of samples n generated from

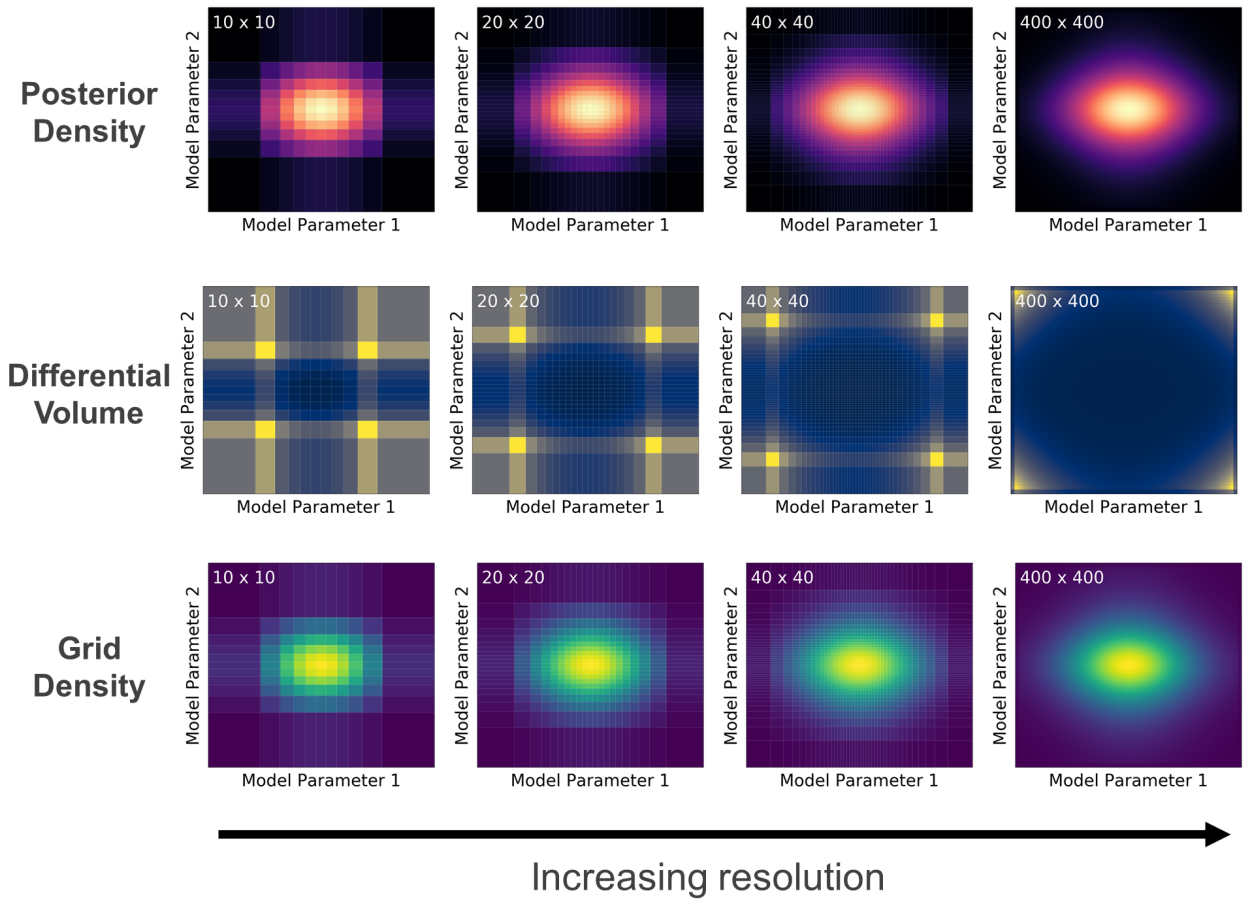

Figure 6: An illustration of the connection between grids and continuous density distributions. As we increase the number of grid points, our estimate of the posterior $\mathcal{P}(\boldsymbol{\Theta})$ improves (top). Since the spacing between the grid points varies to maximize the effective sample size (see Figure 4 and §4.2), the differential volume elements $\Delta\boldsymbol{\Theta}_i$ change depending on our location (middle). As we continue to increase the number of volume elements, the density of grid points at any particular location $\rho(\boldsymbol{\Theta}_i) = [\Delta\boldsymbol{\Theta}_i]^{-1}$ behaves like a continuous function $\mathcal{Q}(\boldsymbol{\Theta})$ whose distribution is similar to $\mathcal{P}(\boldsymbol{\Theta})$ (bottom). This implies we should be able to use $\mathcal{Q}(\boldsymbol{\Theta})$ in some way to estimate $\mathcal{P}(\boldsymbol{\Theta})$. See §5 for additional details.

$\mathcal{Q}(\boldsymbol{\Theta})$. The fact that we can derive roughly equivalent estimates from these these very different approaches as we adjust n and $\mathcal{Q}(\boldsymbol{\Theta})$ lies at the heart of the connection between grid points and samples.

There are three primary benefits from moving from an adaptively-spaced grid to a continuous distribution $\mathcal{Q}(\boldsymbol{\Theta})$. First, a grid will always have some minimum resolution $\Delta\boldsymbol{\Theta}_i$ that makes it difficult to get our weights to be roughly uniform, limiting our maximum ESS in practice. By contrast, we can in theory get $\mathcal{Q}(\boldsymbol{\Theta})$ to more closely match the posterior $\mathcal{P}(\boldsymbol{\Theta})$, giving a larger ESS at fixed n.

Second, because we are now working with *distributions* rather than a finite number of grid points, we are no longer limited to some finite volume when estimating expectations. Since distributions can range over $(-\infty, +\infty)$, we can guarantee $\mathcal{Q}(\boldsymbol{\Theta})$ will provide sufficient **coverage** over all possible $\boldsymbol{\Theta}$ values that our posterior $\mathcal{P}(\boldsymbol{\Theta})$ could be defined over. This means that some of the theoretical issues raised in §4.3 associated with applying grids to posteriors that range over $(-\infty, +\infty)$ no longer apply. Monte Carlo methods therefore can serve as a *consistent* estimator for a wider range of possible posterior expectations than grid-based methods, making them substantially more flexible.

Finally, the minimum number of grid points always scales exponentially with dimensionality (see §4.1), regardless of how many parameters we are interested in marginalizing over. Since Monte Carlo methods do not rely on these, they can take full advantage of marginalizing over parameters when estimating expectations $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$. They are therefore less susceptible to this effect (although see §7.2).

5.2 Importance Sampling

As I have tried to emphasize previously, the core tenet of this article is that *we do not know what $\mathcal{P}(\boldsymbol{\Theta})$ looks like beforehand*. This means we do not know what grid structure will provide an optimal estimate (i.e. maximum ESS) for $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta}\right]$, let alone how this should behave as $\mathcal{Q}(\boldsymbol{\Theta})$ in the continuum limit. This gives us ample motivation to *choose* $\mathcal{Q}(\boldsymbol{\Theta})$ in such a way to make generating samples from it easy and straightforward.

Assuming we have chosen such a $\mathcal{Q}(\boldsymbol{\Theta})$, we can subsequently generate a series of n samples from it. Assuming these samples have weights q_i associated with them and defining

$$f(\boldsymbol{\Theta}_i) \equiv f_i, \quad \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)/\mathcal{Q}(\boldsymbol{\Theta}_i) \equiv \tilde{w}(\boldsymbol{\Theta}_i) \equiv \tilde{w}_i \tag{33}$$

our original expression reduces to

$$\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right] = \frac{\mathbb{E}_{\mathcal{Q}}\left[f(\boldsymbol{\Theta})\tilde{w}(\boldsymbol{\Theta})\right]}{\mathbb{E}_{\mathcal{Q}}\left[\tilde{w}(\boldsymbol{\Theta})\right]} \approx \frac{\sum_{i=1}^{n} f_i \tilde{w}_i q_i}{\sum_{i=1}^{n} \tilde{w}_i q_i} \tag{34}$$

If we further assume that we have chosen $\mathcal{Q}(\boldsymbol{\Theta})$ so that we can simulate samples that are **independently and identically distributed (iid)** (i.e. each sample has the same probability distribution as the others and all the samples are mutually independent), then the corresponding sample weights immediately reduce to $q_i = 1/n$ and our result becomes

$$\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right] \approx \frac{n^{-1}\sum_{i=1}^{n} f_i \tilde{w}_i}{n^{-1}\sum_{i=1}^{n} \tilde{w}_i} \tag{35}$$

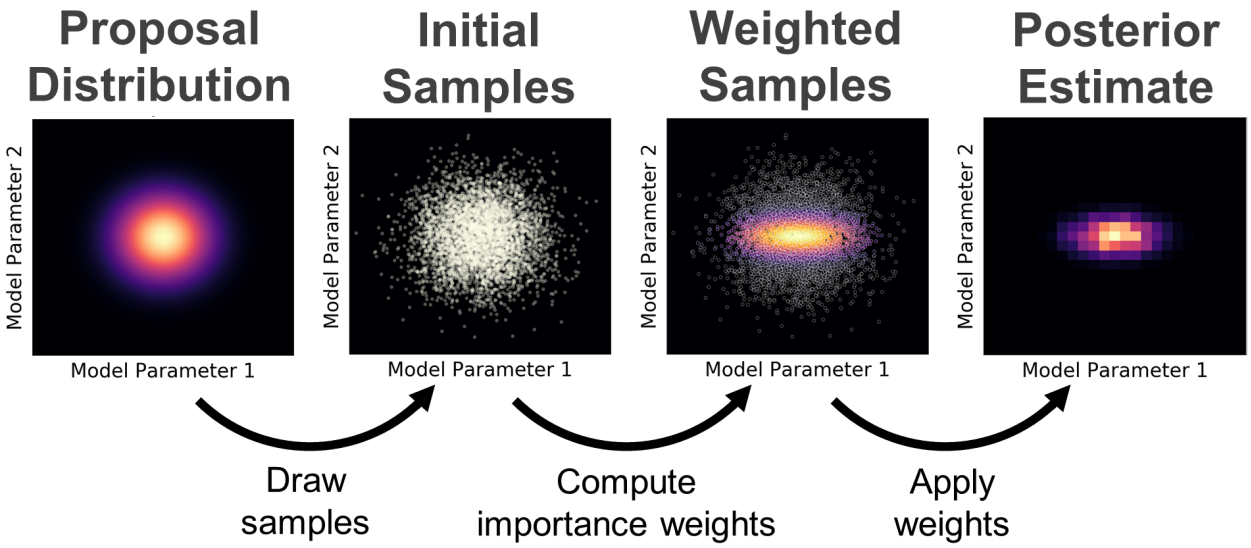

Figure 7: A schematic illustration of Importance Sampling. First, we take a given proposal distribution $\mathcal{Q}(\boldsymbol{\Theta})$ (left) and generate a set of n iid samples from it (middle left). We then weight each sample based on the corresponding "importance" $\tilde{\mathcal{P}}(\boldsymbol{\Theta})/\mathcal{Q}(\boldsymbol{\Theta})$ it has at that location (middle right). We then can use these weighted samples to approximate posterior expectations (right). See §5.2 for additional details.

As with the previous case using grids (§4), the denominator of this expression is again a direct approximation for the evidence

$$\mathcal{Z} = \int \tilde{\mathcal{P}}(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \approx n^{-1} \sum_{i=1}^{n} \tilde{w}_i \tag{36}$$

This gives a straightforward recipe for estimating our original expectation value:

1. Draw n iid samples $\{\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_n\}$ from $\mathcal{Q}(\boldsymbol{\Theta})$.
2. Compute their corresponding weights $\tilde{w}_i = \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)/\mathcal{Q}(\boldsymbol{\Theta}_i)$.
3. Estimate $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$ by computing $\mathbb{E}_{\mathcal{Q}}\left[\tilde{w}(\boldsymbol{\Theta})\right]$ and $\mathbb{E}_{\mathcal{Q}}\left[f(\boldsymbol{\Theta})\tilde{w}(\boldsymbol{\Theta})\right]$ using the weighted sample means.

Since this process just involves "reweighting" the samples based on $\tilde{w}_i$, these weights are often referred to as **importance weights** and the method as **Importance Sampling**. A schematic illustration of Importance Sampling is highlighted in Figure 7.

We can interpret the importance weights as ways to correct for how "far off" our original guess $\mathcal{Q}(\boldsymbol{\Theta})$ is from the truth $\mathcal{P}(\boldsymbol{\Theta})$. If the posterior density is higher at position $\boldsymbol{\Theta}_i$ relative to the proposal density, then we were less likely to generate a sample at that position compared to what we would have seen if we had drawn samples directly from the posterior. As a result, we should increase its corresponding weight to account for this expected deficit of samples at a given position. If the posterior density is lower relative to

the proposal density, then the alternative is true and we want to lower the weight of the corresponding sample to account for the expected excess of samples at a given position.

5.3 Examples of Sampling Strategies

Importance Sampling serves as a useful first step for understanding how the weights $\{\tilde{w}_1, \ldots, \tilde{w}_n\}$ for the corresponding set of n samples are related to different Monte Carlo sampling strategies.

As an example, one common approach is to generate samples uniformly within some cuboid with volume V. The proposal distribution for this will then be

$$\mathcal{Q}^{\rm unif}(\boldsymbol{\Theta}) = \begin{cases} 1/V & \boldsymbol{\Theta} \text{ in cuboid} \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

The corresponding importance weights subsequently will just be proportional to the posterior at a given position:

$$\tilde{w}_i^{\rm unif} = \frac{\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)}{\mathcal{Q}^{\rm unif}(\boldsymbol{\Theta}_i)} = V\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i) \propto \mathcal{P}(\boldsymbol{\Theta}_i) \tag{38}$$

Another possible approach would be if we instead take our proposal to be our prior:

$$\mathcal{Q}^{\rm prior}(\boldsymbol{\Theta}) = \pi(\boldsymbol{\Theta}) \tag{39}$$

This seems like a well-motivated choice: the prior characterizes our knowledge before looking at the data, so it should serve as a useful first guess and encompass the range of all possibilities. Under this assumption, we now find our weights will be equal to the likelihood $\mathcal{L}(\boldsymbol{\Theta})$ at each position:

$$w_i^{\rm prior} = \frac{\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)}{\mathcal{Q}^{\rm prior}(\boldsymbol{\Theta}_i)} = \frac{\mathcal{L}(\boldsymbol{\Theta}_i)\pi(\boldsymbol{\Theta}_i)}{\pi(\boldsymbol{\Theta}_i)} = \mathcal{L}(\boldsymbol{\Theta}_i) \tag{40}$$

Finally, notice that the optimal sampling strategy is to assume that we can take our proposal to be identical to our posterior:

$$\mathcal{Q}^{\rm post}(\boldsymbol{\Theta}) = \mathcal{P}(\boldsymbol{\Theta}) \tag{41}$$

The corresponding weights will then just be constant and equal to the evidence $\mathcal{Z}$:

$$w_i^{\rm post} = \frac{\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)}{\mathcal{Q}^{\rm post}(\boldsymbol{\Theta}_i)} = \frac{\mathcal{Z}\mathcal{P}(\boldsymbol{\Theta}_i)}{\mathcal{P}(\boldsymbol{\Theta}_i)} = \mathcal{Z} \tag{42}$$

As expected, this final result guarantees the maximum possible ESS of $n_{\rm eff} = n$. Getting $\mathcal{Q}(\boldsymbol{\Theta})$ to be as "close" as possible to $\mathcal{P}(\boldsymbol{\Theta})$ therefore becomes a crucial part of analyses when trying to use Importance Sampling to estimate expectation values. It is this result in particular that motivates the use of Markov Chain Monte Carlo (MCMC) methods discussed from §6 onward: if we can somehow generate samples *directly* from $\mathcal{P}(\boldsymbol{\Theta})$ or something close to it, then we can achieve an optimal estimate of our corresponding expectation values.

Exercise: Importance Sampling over a 2-D Gaussian

Setup

Let's return to our exercise from §4, in which our unnormalized posterior is well-approximated by a 2-D Gaussian (Normal) distribution:

$$\tilde{\mathcal{P}}(x, y) = \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right\}$$

where $(\mu_x, \mu_y) = (-0.3, 0.8)$ and $(\sigma_x^2, \sigma_y^2) = (2, 0.5)$.

Importance Sampling

We want to use Importance Sampling to approximate various posterior integrals from this distribution. We will start by choosing our proposal distribution $\mathcal{Q}(x, y)$ to be a 2-D Gaussian with a mean of 0 and standard deviation of 1:

$$\mathcal{Q}(x, y) = \mathcal{N}\left[(\mu_x, \mu_y) = (0, 0), (\sigma_x, \sigma_y) = (1, 1)\right]$$

Using $n = 25$ iid random samples drawn from the proposal distribution, compute an estimate for:

1. the evidence $\mathcal{Z}$,
2. the means $\mathbb{E}_{\mathcal{P}}\left[x\right]$ and $\mathbb{E}_{\mathcal{P}}\left[y\right]$,
3. the 68% credible intervals (or closest approximation) $[x_{\rm low}, x_{\rm high}]$ and $[y_{\rm low}, y_{\rm high}]$,
4. and the effective sample size $n_{\rm eff}$.

How accurate are each of these quantities with the values we might expect? What does $n_{\rm eff}/n$ tell us about how well our proposal $\mathcal{Q}(x, y)$ traces the underlying posterior $\mathcal{P}(x, y)$?

Uncertainty

Repeat the above exercise $m = 100$ times to get an estimate for how much our estimates of each quantity can vary. Is the variation in line with what might be expected given the typical effective sample size? Why or why not?

Convergence

Now repeat the above exercise using $n = 100$, $n = 1000$, and $n = 10000$ points rather than $n = 25$ points and comment on any differences. How much has the overall accuracy improved? Do the estimates appear convergent and consistent as $n_{\rm eff}$ increases? How much do the errors on quantities shrink as a function of n and/or $n_{\rm eff}$? Is this behavior expected? Why or why not?

Consistency

Next, let's expand our proposal distribution to instead have $(\sigma_x, \sigma_y) = (2, 2)$ to get more coverage in the "tails" of the posterior. Perform the same exercise as above with $n = \{100, 1000, 10000\}$ iid random samples. Do the answers change substantially? Why or why not?

While in theory we can choose $\mathcal{Q}(x, y) \approx \mathcal{P}(x, y)$ so that $n_{\rm eff} \approx n$, we do not know the exact shape of the posterior ahead of time. Given that $\tilde{\mathcal{P}}(x, y)$ may differ from our initial expectations, what does this exercise imply about general concerns applying Importance Sampling in practice?

6 Markov Chain Monte Carlo

Now that we see how the weights relate to various Monte Carlo sampling strategies (e.g., generating samples from the prior), I will now outline the idea behind **Markov Chain Monte Carlo (MCMC)**. In brief, MCMC methods try to generate samples in such a way that the importance weights $\{\tilde{w}_1, \ldots, \tilde{w}_n\}$ associated with each sample are constant. Based on the results from §5.3, this means MCMC seeks to generate samples proportional to the posterior $\mathcal{P}(\boldsymbol{\Theta})$ in order to arrive at an *optimal estimate* for our expectation value.

MCMC accomplishes this by creating a **chain** of (correlated) parameter values $\{\boldsymbol{\Theta}_1 \rightarrow \cdots \rightarrow \boldsymbol{\Theta}_n\}$ over n iterations such that the number of iterations $m(\boldsymbol{\Theta}_i)$ spent in any particular region $\delta_{\boldsymbol{\Theta}_i}$ centered on $\boldsymbol{\Theta}_i$ is proportional to the posterior density $\mathcal{P}(\boldsymbol{\Theta}_i)$ contained within that region. In other words, the "density" of samples generated from MCMC

$$\rho(\boldsymbol{\Theta}) \equiv \frac{m(\boldsymbol{\Theta})}{n} \tag{43}$$

at position $\boldsymbol{\Theta}$ integrated over $\delta_{\boldsymbol{\Theta}}$ is approximately

$$\int_{\boldsymbol{\Theta} \in \delta_{\boldsymbol{\Theta}}} \mathcal{P}(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \approx \int_{\boldsymbol{\Theta} \in \delta_{\boldsymbol{\Theta}}} \rho(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \approx n^{-1} \sum_{j=1}^{n} \mathbb{1}\left[\boldsymbol{\Theta}_j \in \delta_{\boldsymbol{\Theta}}\right] \tag{44}$$

where $\mathbb{1}\left[\cdot\right]$ is the **indicator function** which evaluates to 1 if the inside condition is true and 0 otherwise. We can therefore approximate the density by simply adding up the number of samples within $\delta_{\boldsymbol{\Theta}}$ and normalizing by the total number of samples n. A schematic illustration of this concept is shown in Figure 8.

While this will just be approximately true for any finite n, as the number of samples $n \rightarrow \infty$ this procedure generally guarantees that $\rho(\boldsymbol{\Theta}) \rightarrow \mathcal{P}(\boldsymbol{\Theta})$ everywhere.[2] In theory then, once we have a reasonable enough approximation for $\rho(\boldsymbol{\Theta})$, we can also use the samples $\{\boldsymbol{\Theta}_1 \rightarrow \cdots \rightarrow \boldsymbol{\Theta}_n\}$ generated from $\rho(\boldsymbol{\Theta})$ to get an estimate for the evidence using the same substitution trick introduced in §5:

$$\mathcal{Z} = \int \frac{\tilde{\mathcal{P}}(\boldsymbol{\Theta})}{\rho(\boldsymbol{\Theta})} \rho(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \equiv \mathbb{E}_{\rho}\left[\tilde{\mathcal{P}}(\boldsymbol{\Theta})/\rho(\boldsymbol{\Theta})\right] \approx n^{-1} \sum_{i=1}^{n} \frac{\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)}{\rho(\boldsymbol{\Theta}_i)} \tag{45}$$

[2]Discussing the details of exactly when/where this condition holds in theory and in practice is beyond the scope of this paper but can be found in other references such as Asmussen & Glynn (2011) and Brooks et al. (2011).

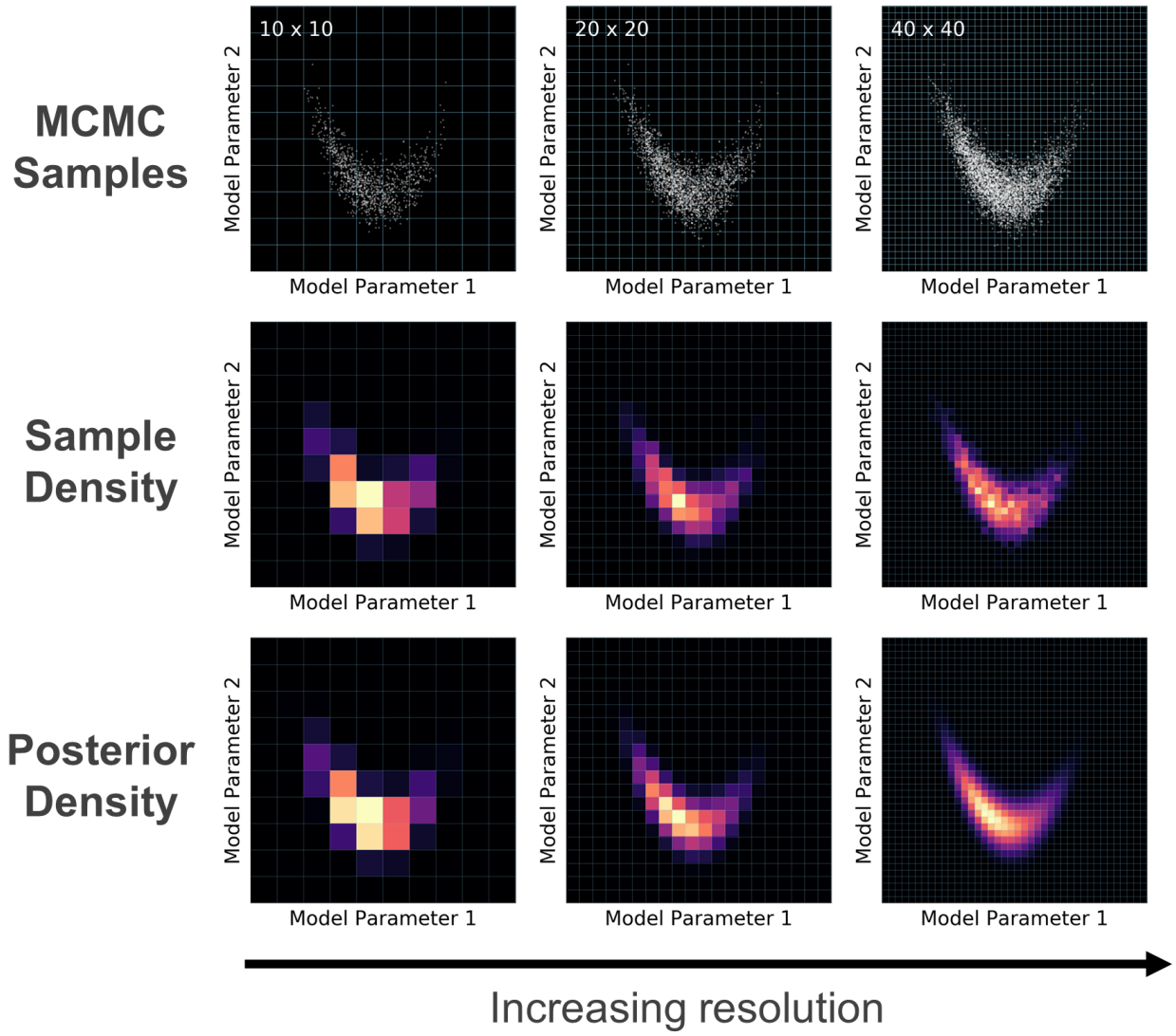

Figure 8: A schematic illustration of Markov Chain Monte Carlo (MCMC). MCMC tries to create a chain of n (correlated) samples $\{\boldsymbol{\Theta}_1 \rightarrow \cdots \rightarrow \boldsymbol{\Theta}_n\}$ (top) such that the number of samples m in some particular volume δ gives a relative density m/n (middle) comparable to the posterior $\mathcal{P}(\boldsymbol{\Theta})$ integrated over the same volume (bottom). See §6 for additional details.

This is just the average of the ratio between $\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)$ and $\rho(\boldsymbol{\Theta}_i)$ over all n samples.

Finally, since our MCMC procedure gives us a series of n samples from the posterior, our expectation value simply reduces to

$$\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right] \approx \frac{n^{-1}\sum_{i=1}^{n} f_i\tilde{w}_i}{n^{-1}\sum_{i=1}^{n}\tilde{w}_i} = \frac{n^{-1}\sum_{i=1}^{n} f_i}{n^{-1}\sum_{i=1}^{n} 1} = n^{-1}\sum_{i=1}^{n} f_i \tag{46}$$

This is just the **sample mean** of the corresponding $\{f_1, \ldots, f_n\}$ values over our set of n samples.

I wish to take a moment here to highlight two features of the above results related to common misconceptions surrounding MCMC methods. First, there is a widespread belief that because MCMC methods generate a chain of samples whose behavior *follows* the posterior, we do not have any ability to use them to estimate normalizing constants such as the evidence $\mathcal{Z}$. As shown above, this is not true at all: not only *can* we do this using $\rho(\boldsymbol{\Theta})$, but the estimate we derive is actually a *consistent* one (although it will converge slowly; see §7.1).

The second misconception is that the primary goal of MCMC is to "approximate" or "explore" the posterior. In other words, to estimate $\rho(\boldsymbol{\Theta})$. However, as shown above, the ability of MCMC methods to estimate $\rho(\boldsymbol{\Theta})$ is really only useful for estimating the evidence $\mathcal{Z}$. In fact, by tracing its heritage from Importance Sampling-based methods, we see its primary purpose is actually *to estimate expectation values* (i.e. integrals *over* the posterior). I have explicitly tried to avoid introducing any mention of "approximating the posterior" up to this point in order to avoid this misconception, but will spend some time discussing this point in more detail in §7.1.

To summarize, the idea behind MCMC is to simulate a series of values $\{\boldsymbol{\Theta}_1 \to \cdots \to \boldsymbol{\Theta}_n\}$ in a way that their density $\rho(\boldsymbol{\Theta})$ after a given amount of time follows the underlying posterior $\mathcal{P}(\boldsymbol{\Theta})$. We can then estimate the posterior within any particular region $\delta_{\boldsymbol{\Theta}}$ by simply counting up how many samples we simulate there and normalizing by the total number of samples n we generated. Because we are also simulating values directly from the posterior, any expectation values also reduce to simple sample averages. This procedure is incredibly intuitive and part of the reason MCMC methods have become so widely adopted.

6.1 Generating Samples with the Metropolis-Hastings Algorithm

There is a vast literature on various approaches to generating samples **(see, e.g., cites)**. Since this article focuses on building up a *conceptual understanding* of MCMC methods, exploring how the majority of these methods behave both in theory and in practice is beyond the scope of this paper.

Instead of an overview, I aim to clarify the basics of how these methods operate. The central idea is that we want a way to generate new samples $\boldsymbol{\Theta}_i \to \boldsymbol{\Theta}_{i+1}$ such that the distribution of the final samples $\rho(\boldsymbol{\Theta})$ as $n \to \infty$ (1) is **stationary** (i.e. it converges to something) and (2) is equal to the $\mathcal{P}(\boldsymbol{\Theta})$. These are essentially analogs to the convergence and consistency constraints discussed in §4.3.

We can satisfy the first condition by invoking **detailed balance**. This is the idea that probability is conserved when moving from one position to another (i.e. the process is reversible). More formally, this just reduces to factoring of probability:

$$P(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)P(\boldsymbol{\Theta}_i) = P(\boldsymbol{\Theta}_{i+1}, \boldsymbol{\Theta}_i) = P(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{i+1})P(\boldsymbol{\Theta}_{i+1}) \tag{47}$$

where $P(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)$ is the probability of moving from $\boldsymbol{\Theta}_i \rightarrow \boldsymbol{\Theta}_{i+1}$ and $P(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{i+1})$ is the probability of the reverse move from $\boldsymbol{\Theta}_{i+1} \rightarrow \boldsymbol{\Theta}_i$. Rearranging then gives the following constraint:

$$\frac{P(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)}{P(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{i+1})} = \frac{P(\boldsymbol{\Theta}_{i+1})}{P(\boldsymbol{\Theta}_i)} = \frac{\mathcal{P}(\boldsymbol{\Theta}_{i+1})}{\mathcal{P}(\boldsymbol{\Theta}_i)} \tag{48}$$

where the final equality comes from the fact that the distribution we are trying to generate samples from is the posterior $\mathcal{P}(\boldsymbol{\Theta})$.

We now need to implement a procedure that enables us to actually move to new positions by computing this probability. We can do this by breaking each move into two steps. First, we want to *propose* a new position $\boldsymbol{\Theta}_i \rightarrow \boldsymbol{\Theta}'_{i+1}$ based on a **proposal distribution** $\mathcal{Q}(\boldsymbol{\Theta}'_{i+1}|\boldsymbol{\Theta}_i)$ similar in nature to the $\mathcal{Q}(\boldsymbol{\Theta})$ used in to Importance Sampling (§5.2). Then we will either decide to **accept** the new position ($\boldsymbol{\Theta}_{i+1} = \boldsymbol{\Theta}'_{i+1}$) or **reject** the new position ($\boldsymbol{\Theta}_{i+1} = \boldsymbol{\Theta}_i$) with some **transition probability** $T(\boldsymbol{\Theta}'_{i+1}|\boldsymbol{\Theta}_i)$. Combining these terms together then gives us the probability of moving to a new position:

$$P(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i) \equiv \mathcal{Q}(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)T(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}i) \tag{49}$$

As with Importance Sampling, we can choose $\mathcal{Q}(\boldsymbol{\Theta}'_{i+1}|\boldsymbol{\Theta}_i)$ so that it is straightforward to propose new samples $\boldsymbol{\Theta}'_{i+1}$ by numerical simulation. We then need to determine the transition probability $T(\boldsymbol{\Theta}'_{i+1}|\boldsymbol{\Theta}_i)$ of whether we should accept or reject $\boldsymbol{\Theta}'_{i+1}$. Substituting into our expression for detailed balance, we find that our form for the transition probability must satisfy the following constraint:

$$\frac{T(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)}{T(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{i+1})} = \frac{\mathcal{P}(\boldsymbol{\Theta}_{i+1})}{\mathcal{P}(\boldsymbol{\Theta}_i)} \frac{\mathcal{Q}(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{i+1})}{\mathcal{Q}(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)} \tag{50}$$

It is straightforward to show that the **Metropolis criterion** Metropolis et al. (1953)

$$T(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i) \equiv \min\left[1, \frac{\mathcal{P}(\boldsymbol{\Theta}_{i+1})}{\mathcal{P}(\boldsymbol{\Theta}_i)} \frac{\mathcal{Q}(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{i+1})}{\mathcal{Q}(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)}\right] \tag{51}$$

satisfies this constraint.

Generating samples following this approach can be done using the **Metropolis-Hastings (MH) Algorithm** (Metropolis et al., 1953; Hastings, 1970):

1. *Propose* a new position $\boldsymbol{\Theta}_i \rightarrow \boldsymbol{\Theta}'_{i+1}$ by generating a sample from the proposal distribution $\mathcal{Q}(\boldsymbol{\Theta}'_{i+1}|\boldsymbol{\Theta}_i)$.

2. *Compute* the transition probability $T(\boldsymbol{\Theta}'_{i+1}|\boldsymbol{\Theta}_i) = \min\left[1, \frac{\mathcal{P}(\boldsymbol{\Theta}'_{i+1})}{\mathcal{P}(\boldsymbol{\Theta}_i)} \frac{\mathcal{Q}(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}'_{i+1})}{\mathcal{Q}(\boldsymbol{\Theta}'_{i+1}|\boldsymbol{\Theta}_i)}\right]$.

3. *Generate* a random number u_{i+1} from $[0, 1]$.

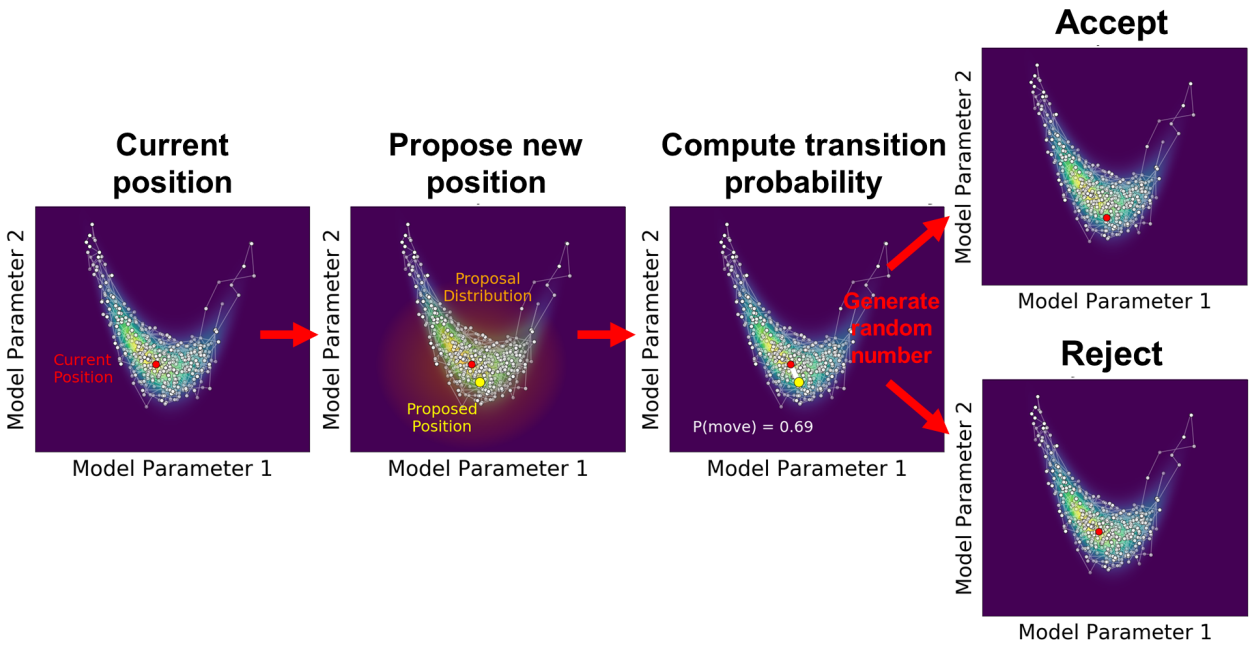

Figure 9: A schematic illustration of the Metropolis-Hastings algorithm. At a given iteration i, we have generated a chain of samples $\{\boldsymbol{\Theta}_1 \rightarrow \cdots \rightarrow \boldsymbol{\Theta}_i\}$ (white) up to the current position $\boldsymbol{\Theta}_i$ (red) whose behavior follows the underlying posterior $\mathcal{P}(\boldsymbol{\Theta})$ (viridis color map). We then propose a new position $\boldsymbol{\Theta}'_{i+1}$ (yellow) from the proposal distribution (orange shaded region). We then compute the transition probability $T(\boldsymbol{\Theta}'_{i+1}|\boldsymbol{\Theta}_i)$ (white) based on the posterior $\mathcal{Q}(\boldsymbol{\Theta})$ and proposal $\mathcal{Q}(\boldsymbol{\Theta}'|\boldsymbol{\Theta})$ densities. We then generate a random number u_{i+1} uniformly from 0 to 1. If $u_{i+1} \leq T(\boldsymbol{\Theta}'_{i+1}|\boldsymbol{\Theta}_i)$, we accept the move and make our next position in the chain $\boldsymbol{\Theta}_{i+1} = \boldsymbol{\Theta}'_{i+1}$. If we reject the move, then $\boldsymbol{\Theta}_{i+1} = \boldsymbol{\Theta}_i$. See §6.1 for additional details.

4. If $u_{i+1} \leq T(\boldsymbol{\Theta}'_{i+1}|\boldsymbol{\Theta}_i)$, *accept* the move and set $\boldsymbol{\Theta}_{i+1} = \boldsymbol{\Theta}'_{i+1}$. If $u_{i+1} > T(\boldsymbol{\Theta}'_{i+1}|\boldsymbol{\Theta}_i)$, *reject* the move and set $\boldsymbol{\Theta}_{i+1} = \boldsymbol{\Theta}_i$.

5. Increment $i = i + 1$ and repeat this process.

See Figure 9 for a schematic illustration of this process.

Because algorithms like the MH algorithm generate a *chain* of states where the next proposed position only depends on the current position rather than any of its past positions (i.e. it "forgets" the past), they are known as **Markov processes**. Combining these two terms with the Monte Carlo nature of simulating new positions is what gives Markov Chain Monte Carlo (MCMC) its namesake.

An issue with generating a chain of samples in practice is the fact that our chain only has finite length and a starting position $\boldsymbol{\Theta}_0$. If our chain were infinitely long, we would expect it to visit every possible position in parameter space, rendering the exact starting position is unimportant. However, since in practice we terminate sampling after only n iterations, starting from a location $\boldsymbol{\Theta}_0$ that has an extremely low probability means an inordinate fraction of our n samples will occupy this low-probability region, possibly

biasing our final results. Since we have limited knowledge beforehand about where $\boldsymbol{\Theta}_0$ is relative to our posterior, in practice we generally want to remove the initial chain of states once we are confident our chain has begun sampling from higher-probability regions. Discussing various approaches for identifying and removing samples from this **burn-in period** is beyond the scope of this article; for additional information, please see Gelman & Rubin (1992), Gelman et al. (2013), and Vehtari et al. (2019) along with references therein.

6.2 Effective Sample Size and Auto-Correlation Time

At this point, MCMC seems like it should be the optimal method for any situation: by simulating samples directly from the (unknown) posterior, we can achieve an optimal estimate for any expectation values we wish to evaluate. In practice, however, this does not hold true. MCMC values rely on specific algorithmic procedures such as the MH algorithm to generate samples, whose limiting behavior *reduces to* a chain of samples $\{\boldsymbol{\Theta}_1 \rightarrow \cdots \rightarrow \boldsymbol{\Theta}_n\}$ whose distribution follows the posterior. Any given sample $\boldsymbol{\Theta}_i$, however, is more likely than not to be **correlated** with both the previous sample in the sequence $\boldsymbol{\Theta}_{i-1}$ and the subsequent sample in the sequence $\boldsymbol{\Theta}_{i+1}$.

This occurs for two reasons. First, new positions $\boldsymbol{\Theta}_i$ drawn from $\mathcal{Q}(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{i-1})$ by construction tend to depend on the current position $\boldsymbol{\Theta}_{i-1}$. This means that the position we propose at iteration $i+1$ from will be correlated with the position at iteration i, which itself will be correlated with the position at iteration $i-1$, etc.

Second, even if we set $\mathcal{Q}(\boldsymbol{\Theta}'|\boldsymbol{\Theta}) = \mathcal{Q}(\boldsymbol{\Theta}')$ so that all of our proposed positions are uncorrelated, our transition probability $T(\boldsymbol{\Theta}'|\boldsymbol{\Theta})$ still ensures that we will eventually reject the new position so that $\boldsymbol{\Theta}_{i+1} = \boldsymbol{\Theta}_i$. Since samples at exactly the same position are maximally correlated, this ensures that samples from our chain will "on average" have non-zero correlations. Note that having low **acceptance fractions** (i.e. the fraction of proposals that are accepted rather than rejected) will lead to a larger fraction of the chain containing these perfectly correlated samples, increasing the overall correlation.

As mentioned in §4.2, correlated samples provide less information about the underlying distribution they are sampled from since their behavior doesn't just depend on the underlying distribution but also the neighboring samples in the sequence. Samples that are more highly correlated then should lead to a reduced ESS.

This intuition can be quantified by introducing the **auto-covariance** $C(t)$ for some integer lag t. Assuming that we have an infinitely long chain $\{\boldsymbol{\Theta}_1 \rightarrow \dots\}$, the auto-covariance $C(t)$ is:

$$C(t) \equiv \mathbb{E}_i \left[(\boldsymbol{\Theta}_i - \bar{\boldsymbol{\Theta}}) \cdot (\boldsymbol{\Theta}_{i+t} - \bar{\boldsymbol{\Theta}}) \right] = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\Theta}_i - \bar{\boldsymbol{\Theta}}) \cdot (\boldsymbol{\Theta}_{i+t} - \bar{\boldsymbol{\Theta}}) \tag{52}$$

where $\cdot$ is the dot product. In other words, we want to know the covariance between $\boldsymbol{\Theta}_i$ at some iteration i and $\boldsymbol{\Theta}_{i+t}$ at some other iteration $i+t$, averaged over all all possible pairs of samples $(\boldsymbol{\Theta}_i, \boldsymbol{\Theta}_{i+t})$ in our infinitely long chain. Note that the amplitude $|C(t)|$ will be maximized at $|C(t=0)|$, where the two samples being compared are identical, and minimized with $|C(t)| = 0$ when $\boldsymbol{\Theta}_i$ and $\boldsymbol{\Theta}_{i+t}$ are completely independent from each other.

Using the auto-covariance, we can define the corresponding **auto-correlation** $A(t)$ as

$$A(t) \equiv \frac{C(t)}{C(0)} \tag{53}$$

This now measures the average degree of correlation between samples separated by an integer lag t. In the case where $t = 0$, both samples are identical and $A(t = 0) = 1$. In the case where the samples are uncorrelated over lag t, $A(t) = 0$.

The overall **auto-correlation time** for our chain is just the auto-correlation $A(t)$ summed over all non-zero lags ($t \neq 0$):

$$\tau \equiv \sum_{t=-\infty}^{\infty} A(t) - 1 = 2 \sum_{t=1}^{\infty} A(t) \tag{54}$$

where the -1 comes from the fact that the auto-correlation with no lag is just $A(t = 0) = 1$ (i.e. each sample perfectly correlates with itself) and the substitution arises from the fact that $A(t) = A(-t)$ by symmetry. If $\tau = 0$, then it takes no time at all for samples to become uncorrelated and the samples can be assumed to be iid. If $\tau > 0$, then it takes on average τ additional iterations for samples to become uncorrelated. An illustration of this process is shown in Figure 10.

Incorporating the auto-correlation time leads directly to a modified definition for the ESS:

$$n'_{\rm eff} \equiv \frac{n_{\rm eff}}{1 + \tau} \tag{55}$$

In practice, we cannot precisely compute τ since we do not have an infinite number of samples and do not know $\mathcal{P}(\boldsymbol{\Theta})$. Therefore we often need to generate an estimate $\hat{\tau}$ of the auto-correlation time using the existing set of n samples we have. While discussing various approaches taken to derive $\hat{\tau}$ is beyond the scope of this work, please see Brooks et al. (2011) for additional details.

The fact that MCMC methods are subject to non-negative auto-correlation times ($\tau \geq 0$) but have optimal importance weights $\tilde{w}_i = 1$ give an ESS of

$$n'_{\rm eff,MCMC} = \frac{n_{\rm eff,MCMC}}{1 + \tau} = \frac{n}{1 + \tau} \leq n \tag{56}$$

This means that *there is no guarantee that MCMC is always the optimal choice to achieve the largest ESS*. In particular, Importance Sampling methods, which can generate fully iid samples with no auto-correlation time ($\tau = 0$) but non-optimal importance weights $\tilde{w}_i$, instead have an ESS of

$$n'_{\rm eff,IS} = \frac{n_{\rm eff,IS}}{1 + \tau} = n_{\rm eff,IS} = \frac{\left(\sum_{i=1}^{n} \tilde{w}_i\right)^2}{\sum_{i=1}^{n} \tilde{w}_i^2} \leq n \tag{57}$$

which can be greater than $n'_{\rm eff,MCMC}$ at fixed n.

Given the results above, it should now be clear that *the central motivating concern of MCMC methods is whether they can generate a chain of samples with an auto-correlation time small enough to outperform Importance Sampling.* Whether or not this is true will depend on the posterior, the approach used to generate the chain of samples (see §6.1 and §8) and the proposal distribution $\mathcal{Q}(\boldsymbol{\Theta})$ used for Importance Sampling (see §5.3).

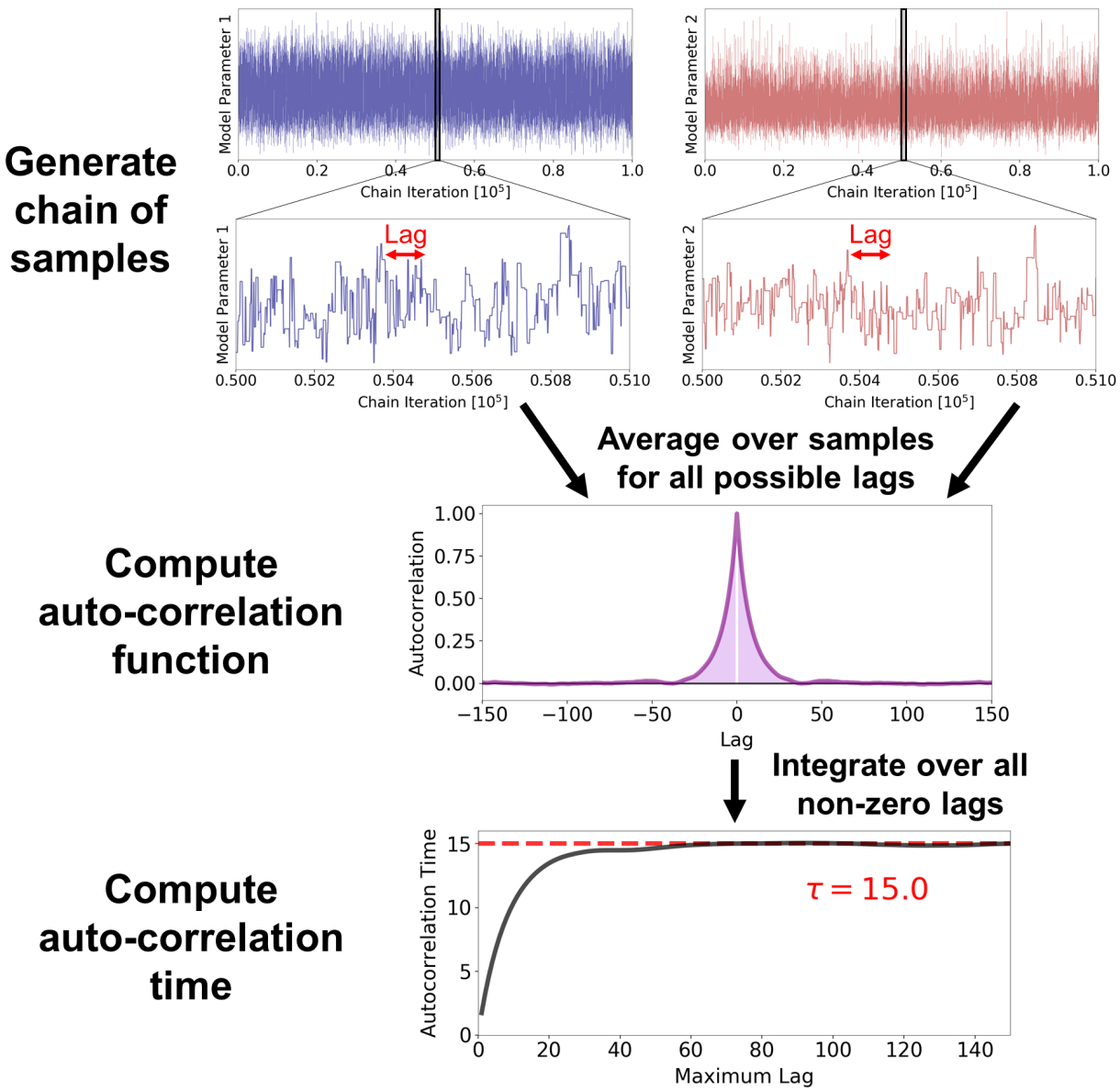

Figure 10: A schematic illustration of the auto-correlation associated with MCMC. MCMC methods generate a chain of samples $\{\boldsymbol{\Theta}_1 \rightarrow \cdots \rightarrow \boldsymbol{\Theta}_n\}$ (top), but these tend to be strongly correlated on small length scales (top middle). We can quantify the degree of correlation by computing the corresponding auto-correlation $A(t)$ over our set of samples and all possible time lags t (bottom middle). This quantity is 1 when $t = 0$ and drops to 0 as $t \rightarrow \pm\infty$. The overall auto-correlation time τ associated with our chain of samples is then just the integrated auto-correlation over $t \neq 0$. See §6.2 for additional details.

Exercise: MCMC over a 2-D Gaussian

Setup

Let's again return to our examples from §4 and §5, in which our unnormalized posterior is well-approximated by a 2-D Gaussian (Normal) distribution:

$$\tilde{\mathcal{P}}(x, y) = \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right\}$$

where $(\mu_x, \mu_y) = (-0.3, 0.8)$ and $(\sigma_x^2, \sigma_y^2) = (2, 0.5)$.

We want to use MCMC to approximate various posterior integrals from this distribution. We will start by choosing our proposal distribution $\mathcal{Q}(x', y'|x, y)$ to be a 2-D Gaussian with a mean of 0 and standard deviation of 1:

$$\mathcal{Q}(x', y'|x, y) = \mathcal{N}\left[(\mu_x, \mu_y) = (x, y), (\sigma_x, \sigma_y) = (1, 1)\right]$$

Parameter Estimation

Using the above proposal, generate $n = 1000$ samples following the MH algorithm starting from the position $(x_0, y_0) = (0, 0)$. Using these samples, compute an estimate of the means $\mathbb{E}_{\mathcal{P}}[x]$ and $\mathbb{E}_{\mathcal{P}}[y]$ as well as the corresponding 68% credible intervals (or closest approximation) $[x_{\rm low}, x_{\rm high}]$ and $[y_{\rm low}, y_{\rm high}]$. How accurate are each of these quantities compared with the values we might expect?

Evidence Estimation

Next, use a set of 10×10 bins from $x = [-5, 5]$ and $y = [-5, 5]$ to construct an estimate $\rho(x, y)$ from the resulting set of samples. Using this estimate for the density, compute an estimate of the evidence $\mathcal{Z}$. How accurate is our approximation? Does it substantially change if we adjust the number and/or size of the bins?

Auto-Correlation Time and Effective Sample Size

Use numerical methods to compute an estimate of the auto-correlation time τ and the corresponding effective sample size $n_{\rm eff}$. How efficient is our sampling ($n_{\rm eff}/n$) compared to the default Importance Sampling approach from the exercise in §5? Does this mirror what we'd expect given the acceptance fraction of our proposals? What do these quantities this tell us about how well our proposal $\mathcal{Q}(x, y)$ matches the structure of the underlying posterior $\mathcal{P}(x, y)$?

Uncertainties

Repeat the above exercises $m = 30$ times to get an estimate for how much our estimates of each quantity can vary. Is the variation in line with what might be expected given the typical effective sample size?

Consistency and Convergence

Now repeat the above exercise using $n = 2500$ and $n = 10000$ samples points and comment on any differences. How much has the overall accuracy improved? Do the estimates appear convergent and consistent as $n_{\rm eff}$ increases? How much do the errors on quantities shrink as a function of n and/or $n_{\rm eff}$? Is this similar or different from the observed dependence from the Importance Sampling exercise in §5?

Sampling Efficiency

Next, adjust the (σ_x, σ_y) of the proposal distribution to try and improve $n_{\rm eff}$ at fixed n. How close is the final ratio σ_x/σ_y of our proposal to that of the underlying posterior? Are there any additional scaling differences between the rough size of our proposal $\mathcal{Q}(x', y'|x, y)$ relative to the underlying posterior $\mathcal{P}(x, y)$? Given that $\tilde{\mathcal{P}}(x, y)$ may differ from the structure assumed when picking $\mathcal{Q}(x', y'|x, y)$, can you think of any possible scheme to try and adjust our proposal using an existing set of samples?

Burn-In

Finally, adjust the starting position to be at $(x_0, y_0) = (10, 10)$ instead of $(0, 0)$ and generate a new chain of samples. Plot the x and y positions of the chain over time. Are there any obvious signs of the burn-in period? How many samples roughly should be assigned to burn-in and subsequently removed from our chain? Are there any possible heuristics that might help to identify the initial burn-in period?

7 Sampling the Posterior with MCMC

The approach by which MCMC methods are able to generate a chain of samples immediately gives a mental image of our chain "exploring" the posterior. While it is true that the density of samples from the chain $\rho(\boldsymbol{\Theta}) \rightarrow \mathcal{P}(\boldsymbol{\Theta})$ as $n \rightarrow \infty$, *the primary purpose of MCMC is estimating expectation values* $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$. Although this might seem like a subtle difference, this distinction is actually crucial for understanding how MCMC algorithms (should) behave in practice. We discuss this in more detail below.

7.1 Approximating the Posterior

Although algorithms such as MH (§6.1) are constructed to ensure the density of the chain of samples $\rho(\boldsymbol{\Theta})$ generated by MCMC converges to the posterior $\mathcal{P}(\boldsymbol{\Theta})$ as $n \rightarrow \infty$, this *does not* necessarily translate into an efficient method to approximate the posterior in practice. In other words, n might need to be extremely large for this constraint to hold. So how many samples do we need to ensure $\rho(\boldsymbol{\Theta})$ is a good approximation to $\mathcal{P}(\boldsymbol{\Theta})$?

To start, we first need to define some metric for what a "good" approximation is. A reasonable one might be that we would like to know the posterior within some region $\delta_{\boldsymbol{\Theta}}$

to within some precision ϵ so that

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \left[\boldsymbol{\Theta}_i \in \delta_{\boldsymbol{\Theta}} \right] - \int_{\delta_{\boldsymbol{\Theta}}} \mathcal{P}(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \right| \equiv |\hat{p}(\delta_{\boldsymbol{\Theta}}) - p(\delta_{\boldsymbol{\Theta}})| < \epsilon \tag{58}$$

where $p(\delta_{\boldsymbol{\Theta}})$ is the total probability contained within $\delta_{\boldsymbol{\Theta}}$ and $\hat{p}(\delta_{\boldsymbol{\Theta}})$ is the fraction of the MCMC chain of samples contained within the same region. While it might seem strange to only estimate this for one region, I will shortly generalize this to encompass the entire[3] posterior.

In the ideal case where our samples are iid and drawn from $\mathcal{P}(\boldsymbol{\Theta})$, our samples each have a probability $p(\delta_{\boldsymbol{\Theta}})$ of being within $\delta_{\boldsymbol{\Theta}}$. The probability that $\hat{p}(\delta_{\boldsymbol{\Theta}}) = m/n$ then follows the **binomial distribution**:

$$P\left(\hat{p}(\delta_{\boldsymbol{\Theta}}) = \frac{m}{n}\right) = \binom{n}{m} \left[p(\delta_{\boldsymbol{\Theta}})\right]^m \left[1 - p(\delta_{\boldsymbol{\Theta}})\right]^{n-m} \tag{59}$$

In other words, our samples end up inside $\delta_{\boldsymbol{\Theta}}$ a total of m times with probability $p(\delta_{\boldsymbol{\Theta}})$ and outside $\delta_{\boldsymbol{\Theta}}$ a total of $n - m$ times with probability $1 - p(\delta_{\boldsymbol{\Theta}})$. The additional binomial coefficient $\binom{n}{m}$ for "n choose m" accounts for all possible unique cases where m samples can end up within $\delta_{\boldsymbol{\Theta}}$ out of our total sample size of n.

This distribution has a mean of $p(\delta_{\boldsymbol{\Theta}})$, so for any finite n we expect $\hat{p}(\delta_{\boldsymbol{\Theta}})$ to be an **unbiased estimator** of $p(\delta_{\boldsymbol{\Theta}})$:

$$\mathbb{E}\left[\hat{p}(\delta_{\boldsymbol{\Theta}}) - p(\delta_{\boldsymbol{\Theta}})\right] = p(\delta_{\boldsymbol{\Theta}}) - p(\delta_{\boldsymbol{\Theta}}) = 0 \tag{60}$$

The variance, however, depends on the sample size:

$$\mathbb{E}\left[|\hat{p}(\delta_{\boldsymbol{\Theta}}) - p(\delta_{\boldsymbol{\Theta}})|^2\right] = \frac{p(\delta_{\boldsymbol{\Theta}})\left[1 - p(\delta_{\boldsymbol{\Theta}})\right]}{n} \tag{61}$$

In practice, we can expect there to be some non-zero auto-correlation time $\tau > 0$. This will increase the number of MCMC samples we will need to generate to be confident that our estimate $\hat{p}(\delta_{\boldsymbol{\Theta}})$ is well-behaved. Inserting a factor of $1 + \tau$ and substituting our expectation value from above into our accuracy constraint then gives a rough constraint for the number of samples n we would require as a function of ϵ:

$$n \gtrsim \frac{p(\delta_{\boldsymbol{\Theta}})\left[1 - p(\delta_{\boldsymbol{\Theta}})\right]}{\epsilon^2/(1+\tau)} \sim \frac{\hat{p}(\delta_{\boldsymbol{\Theta}})\left[1 - \hat{p}(\delta_{\boldsymbol{\Theta}})\right]}{\epsilon^2} \times (1 + \hat{\tau}) \tag{62}$$

The final substitution of $p(\delta_{\boldsymbol{\Theta}})$ and τ with their noisy estimates $\hat{p}(\delta_{\boldsymbol{\Theta}})$ and $\hat{\tau}$ arises from the fact that in practice we don't know $p(\delta_{\boldsymbol{\Theta}})$ or τ (both of which require full knowledge of the posterior). We are therefore forced to rely on estimators derived from our set of n samples.

[3]Technically the procedure outlined in this section only works for finite volumes. The basic intuition, however, holds even when parameters are unbounded although proving those results is beyond the scope of this work.

Let's now examine this result more closely. As expected, the total number of samples is proportional to $1 + \hat{\tau}$: if it takes longer to generate independent samples, then we need more samples to be confident we have characterized the posterior well in a given region. We also see that $n \propto \epsilon^{-2}$, so that if we want to reduce the error by a factor of x we need to increase our sample size by a factor of x^2.

The behavior in the numerator is more interesting. Note that $\hat{p}(\delta_{\boldsymbol{\Theta}})\left[1 - \hat{p}(\delta_{\boldsymbol{\Theta}})\right]$ is maximized for $\hat{p}(\delta_{\boldsymbol{\Theta}}) = 0.5$, and so the largest sample size needed is when we have split our posterior directly in half. In all other cases the sample size needed will be smaller because there will be more samples outside or inside the region of interest whose information we can leverage. The exact value of $\hat{p}(\delta_{\boldsymbol{\Theta}})$ of course depends on both the posterior $\mathcal{P}(\boldsymbol{\Theta})$ and the target region $\delta_{\boldsymbol{\Theta}}$: the sample size needed to approximate the posterior to some ϵ near the peak of the distribution (the small region where $\mathcal{P}(\boldsymbol{\Theta})$ is large) will likely be different than the sample size needed to accurately estimate the tails of the distribution (the large region where $\mathcal{P}(\boldsymbol{\Theta})$ is small).

While the above argument holds if we are looking to estimate the posterior in just *one* region, "converging to the posterior" implies that we want $\rho(\boldsymbol{\Theta})$ to become a good approximation to $\mathcal{P}(\boldsymbol{\Theta})$ *everywhere*. We can enforce this new requirement by splitting our posterior into m different sub-regions $\{\delta_{\boldsymbol{\Theta}_1}, \ldots, \delta_{\boldsymbol{\Theta}_m}\}$ and requiring that each sub-region is well constrained:

$$|\hat{p}(\delta_{\boldsymbol{\Theta}_1}) - p(\delta_{\boldsymbol{\Theta}_1})| < \epsilon_1 \qquad \ldots \qquad |\hat{p}(\delta_{\boldsymbol{\Theta}_m}) - p(\delta_{\boldsymbol{\Theta}_m})| < \epsilon_m \tag{63}$$

Substituting in the expected errors on each of these constraints then gives us an approximate limit on the number of samples n_j that we need to estimate the posterior in each region $\delta_{\boldsymbol{\Theta}_j}$:

$$n_j \gtrsim \frac{\hat{p}(\delta_{\boldsymbol{\Theta}_j})\left[1 - \hat{p}(\delta_{\boldsymbol{\Theta}_j})\right]}{\epsilon_j^2} \times (1 + \hat{\tau}) \tag{64}$$

The total number of samples we need is then simply:

$$n \gtrsim \sum_{j=1}^{m} n_j \tag{65}$$

This approach of dividing up our posterior into sub-regions is conceptually similar to the grid-based approaches described in §4. As such, it is also subject to the same drawbacks: we expect the number of regions m to increase *exponentially* with the number of dimensions d. For instance, if we just wanted to divide our posterior up into m **orthants** we would end up with $m = 2^d$ regions: 2 in 1-D (left-right), 4 in 2-D (upper-left, lower-left, upper-right, lower-right), 8 in 3-D, etc.

This effect implies that we should in general expect the number of samples required to ensure $\rho(\boldsymbol{\Theta})$ is a good approximation to $\mathcal{P}(\boldsymbol{\Theta})$ for some specified accuracy ϵ to scale as

$$n \gtrsim k^d \tag{66}$$

where k is a constant that depends on the accuracy requirements. This puts approximating the full posterior firmly in the "curse of dimensionality" regime (see §4.1).[4]

[4] A direct corollary of this result is that, while the evidence estimates from MCMC *are* consistent, the rate of convergence to the underlying value will proceed exponentially more slowly as d increases.

While many practitioners talk about MCMC being an efficient method to "approximate the posterior", in practice it is rarely used to approximate $\mathcal{P}(\boldsymbol{\Theta})$ directly. As discussed in §3 and shown in Figure 2, almost all quantities that are reported in the literature *do not* rely on approximations to the full d-dimensional posterior, but rather approximations to marginalized distributions that are almost always restricted to no more than $k \lesssim 3$ parameters at a time. The act of marginalizing over the remaining $d - k$ parameters helps to counteract the curse of dimensionality illustrated here. While it is technically fair to say that MCMC can "explore" the marginalized k-D posteriors for certain limited sets of parameters, this type of language can often lead to more misconceptions than insights.

7.2 Posterior Volume

The basic consequences outlined in §7.1 are more general than the specific case where we imagine dividing up the posterior into orthants or other regions. Fundamentally, computing any expectation over the posterior $\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right]$ requires integrating over the *entire domain* of our parameters $\boldsymbol{\Theta}$. We therefore want to understand how the **volume** of this domain behaves (i.e. how many parameter combinations there are). Once we have a grasp on how this behaves, we can then starting trying to quantify how this will impact our estimates.

To start, let's consider the d-dimensional hyper-cube (the d-cube) with side length ℓ in all d dimensions. Its volume scales as

$$V(\ell) = \prod_{i=1}^{d} \ell = \ell^d \tag{67}$$

The differential volume element between ℓ and $\ell + \mathrm{d}\ell$ is

$$\mathrm{d}V(\ell) = (d \times \ell^{d-1}) \times (\mathrm{d}\ell) \propto \ell^{d-1} \tag{68}$$

This exponential scaling with dimensionality means that volume becomes increasingly concentrated in thin shells located in regions located progressively further away from the center of the d-cube. As an example, consider the length-scale

$$\ell_{50} = 2^{-1/d}\ell \tag{69}$$

that divides the d-cube into two equal-sized regions with 50% of the volume contained interior to ℓ_{50} and 50% of the volume exterior to ℓ_{50}. In 1-D, this gives $\ell_{50}/\ell = 0.5$ as we'd expect. In 2-D, this gives $\ell_{50}/\ell \approx 0.7$. In 3-D, $\ell_{50}/\ell \approx 0.8$. In 7-D, $\ell_{50}/\ell \approx 0.9$. By the time we get to 15-D, we have $\ell_{50}/\ell \approx 0.95$, which means that 50% of the volume is located in the last 5% of the length-scale near the boundary of the d-cube. While the constants may change when considering other shapes (e.g., spheres), in general this exponential scaling as a function of d is a generic feature of higher-dimensional volumes. In other words, increasing the number of parameters leads to an exponential increase in the number of available parameter combinations that we have to explore.

In addition to affecting the long-term behavior of MCMC, this exponential increase in volume also directly impacts how MCMC methods operate. To see why this is the

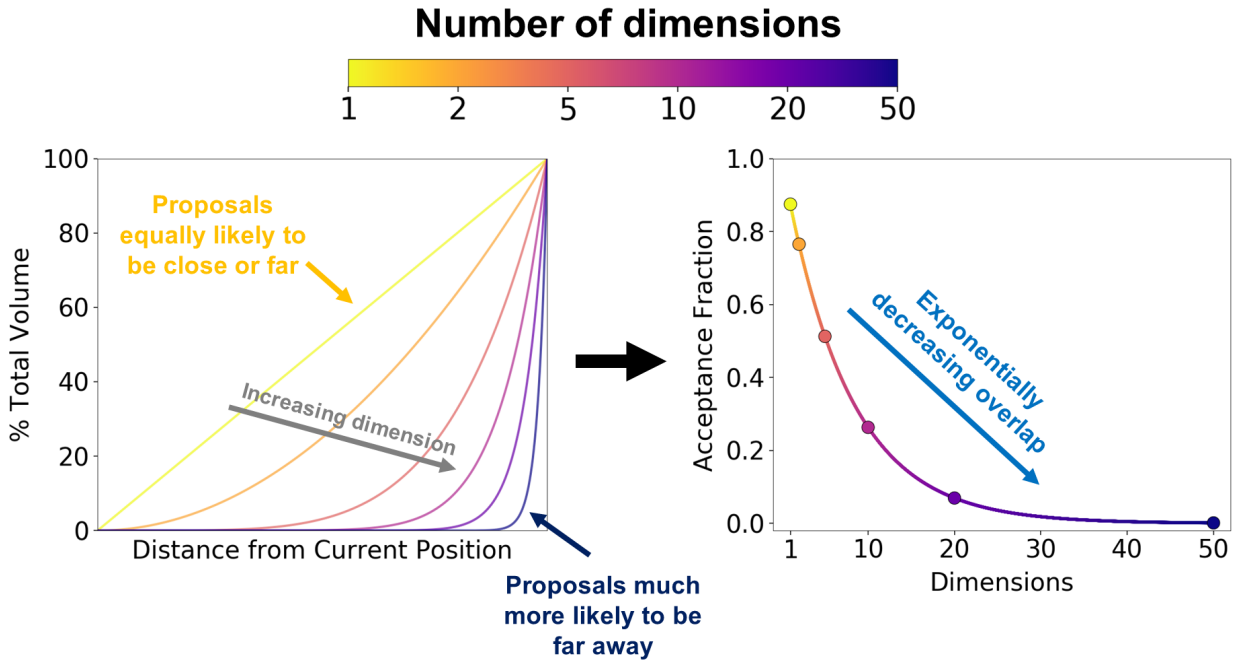

Figure 11: A schematic illustration of how the curse of dimensionality affects MCMC acceptance fractions via posterior volume. At a given position position $\boldsymbol{\Theta}$, the volume increases $\propto r^d$ as a function of distance r away from that position (left). As the dimensionality increases, this implies volume becomes concentrated progressively further out, leading to larger distances between proposed positions $\boldsymbol{\Theta}'$ and the current position $\boldsymbol{\Theta}$. Most of these positions have significantly lower posterior probabilities $\mathcal{P}(\boldsymbol{\Theta}')$ compared to the current value $\mathcal{P}(\boldsymbol{\Theta})$, leading to an exponential decline in the typical acceptance fraction (and a corresponding increase in the auto-correlation time) as the dimensionality increases (right). Adjusting the size and/or shape of the proposal $\mathcal{Q}(\boldsymbol{\Theta}'|\boldsymbol{\Theta})$ can help to counteract this behavior. See §7.2 for additional details.

case, we need look no further than the transition probability used in the MH algorithm discussed in §6.1:

$$T(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i) \equiv \min\left[1, \frac{\mathcal{P}(\boldsymbol{\Theta}_{i+1})}{\mathcal{P}(\boldsymbol{\Theta}_i)}\frac{\mathcal{Q}(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{i+1})}{\mathcal{Q}(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)}\right]$$

The non-trivial portion of this expression cleanly splits into two terms. The first is dependent on the *volume* and is related to how we proposed our next position from $\mathcal{Q}(\boldsymbol{\Theta}'|\boldsymbol{\Theta})$. The second is dependent on the *density* and is related to how the posterior density changes between the two positions.

In practice, our transition probability can be interpreted as a basic corrective approach: after proposing a new position from some nearby volume, we then try to "correct" for differences between our proposal and the underlying posterior by only accepting these moves sometimes based on changes in the underlying density. In high dimensions, this basic "tug of war" between the volume (proposal) and the density (posterior) can break down as the vast majority of an object's volume becomes concentrated near the outer edges.[5] For instance, in the case where our proposal $\mathcal{Q}(\boldsymbol{\Theta}'|\boldsymbol{\Theta})$ is a cube with side-length ℓ centered on $\boldsymbol{\Theta}$, this leads to a median length-scale of $\ell_{50} = 2^{-1/d}\ell$, which increases rapidly from 0.5ℓ to $\approx \ell$ as the dimensionality increases. The same logic also applies to other proposal distributions (see §8). This focus on positions either far away or with very similar separation length-scales as $\ell_{50} \rightarrow \ell$ means that many choices of $\mathcal{Q}(\boldsymbol{\Theta}'|\boldsymbol{\Theta})$ have a tendency to "overshoot", proposing new positions with much smaller posterior densities compared to the current position. These new positions are then almost always rejected, leading to extremely low acceptance fractions and correspondingly long auto-correlation times. An example of this effect is illustrated in Figure 11.

One of the main ways to counteract this behavior is to adjust the size/shape of the proposal $\mathcal{Q}(\boldsymbol{\Theta}'|\boldsymbol{\Theta})$ so that the fraction of proposed positions that are accepted remains sufficiently high. This helps to ensure the posterior density $\mathcal{P}(\boldsymbol{\Theta})$ does not change too drastically when proposing positions new positions, leading to lower overall auto-correlation times. Details of how to implement these schemes in practice are beyond the scope of this article; please see **citation** for additional details.

7.3 Posterior Mass and Typical Sets

Above, I described how the behavior of volume in high dimensions can impact the performance of our MCMC MH sampling algorithm, possibly leading to inefficient proposals and low acceptance fractions. Let's assume that we have resolved this problem and have an efficient way of generating our chain of samples. We now have a secondary question: *where are these samples located?*

From our discussion in §7.1, we know that the highest *density* of samples $\rho(\boldsymbol{\Theta})$ will be located where the posterior density $\mathcal{P}(\boldsymbol{\Theta})$ is also correspondingly high. However, this region $\delta_{\boldsymbol{\Theta}}$ might only correspond to a small portion of the posterior. Indeed, given there is exponentially more volume as the dimensionality increases, it is almost guaranteed

[5] Alternative methods such as Hamiltonian Monte Carlo (Neal, 2012) can get around this problem by smoothly incorporating changes in the density and volume.

that models with many parameters $\boldsymbol{\Theta}$ will have the vast majority of the posterior located outside the region of highest density.

A consequence of this is that the majority of samples in our chain will be located away from the peak density. As a result, *our chain spends the majority of its time generating samples in these regions*. This has a huge impact in the way our chain is expected to behave: while the highest *concentration* of samples will be located in the regions of highest posterior density, the largest *amount* of samples will actually be located in the regions of highest **posterior mass** (i.e. density times volume). Since this implies that a "typical" sample (picked at random) will most likely be located in this region of high posterior mass, this region is also commonly referred to as the **typical set**.

To make this argument a little easier to conceptualize, let's imagine that we have a 3-parameter model $\boldsymbol{\Theta} = (x, y, z)$ and $\mathcal{P}(x, y, z)$ is spherically symmetric. While we could imagine trying to integrate over $\mathcal{P}(x, y, z)$ directly in terms of $\mathrm{d}x\mathrm{d}y\mathrm{d}z$, it is almost always easier to instead integrate over such a distribution in "shells" with differential volume $\mathrm{d}V(r) = 4\pi r^2 \mathrm{d}r$ as a function of radius $r = \sqrt{x^2 + y^2 + z^2}$. This allows us to rewrite the 3-D integral over (x, y, z) as a 1-D integral over r:

$$\int \mathcal{P}(x, y, z)\mathrm{d}x\mathrm{d}y\mathrm{d}z = \int \mathcal{P}(r)4\pi r^2 \mathrm{d}r \equiv \int \mathcal{P}'(r)\mathrm{d}r \tag{70}$$

where $\mathcal{P}'(r) \equiv 4\pi r^2 \mathcal{P}(r)$ is now the 1-D density as a function of r. This "boosts" the contribution as a function of r by the differential volume element of the shell associated with $\mathcal{P}(r)$, and implies that the the posterior should have some sort of shell-like structure (i.e. $\mathcal{P}'(r)$ is maximized for $r > 0$).

Although not all posterior densities can be expected to be spherically-symmetric in this way, in general we can rewrite the d-D integral over $\boldsymbol{\Theta}$ as a 1-D volume integral over V defined by some unknown iso-posterior contours[6]

$$\int \mathcal{P}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} = \int \mathcal{P}(V)\mathrm{d}V \tag{71}$$

As outlined in §7.2, we generically expect the size of each volume element to go as $\mathrm{d}V \sim r^{d-1}\mathrm{d}r$ where r is the distance from the peak of posterior. So the basic intuition we get from the simple spherically-symmetric case still applies and we expect

$$\int \mathcal{P}(V)\mathrm{d}V \sim \int \mathcal{P}(r)r^{d-1}\mathrm{d}r = \int \mathcal{P}'(r)\mathrm{d}r \tag{72}$$

As before, the differential volume element of the shell associated with $\mathcal{P}(r)$ "boosts" its overall contribution as a function of r. This boost also becomes exponentially stronger as d increase. *For even moderately-sized d, we therefore expect the posterior mass to be mostly contained in a thin shell located at a radius r' with some width $\Delta r'$.* See Figure 12 for an illustration of this effect based on the toy problem presented in §8.1.

[6]Indeed, alternative Monte Carlo methods such as Nested Sampling (Skilling, 2004, 2006) or Bridge/Path Sampling (Gelman & Meng, 1998) actually are designed to evaluate this type of volume integral explicitly.

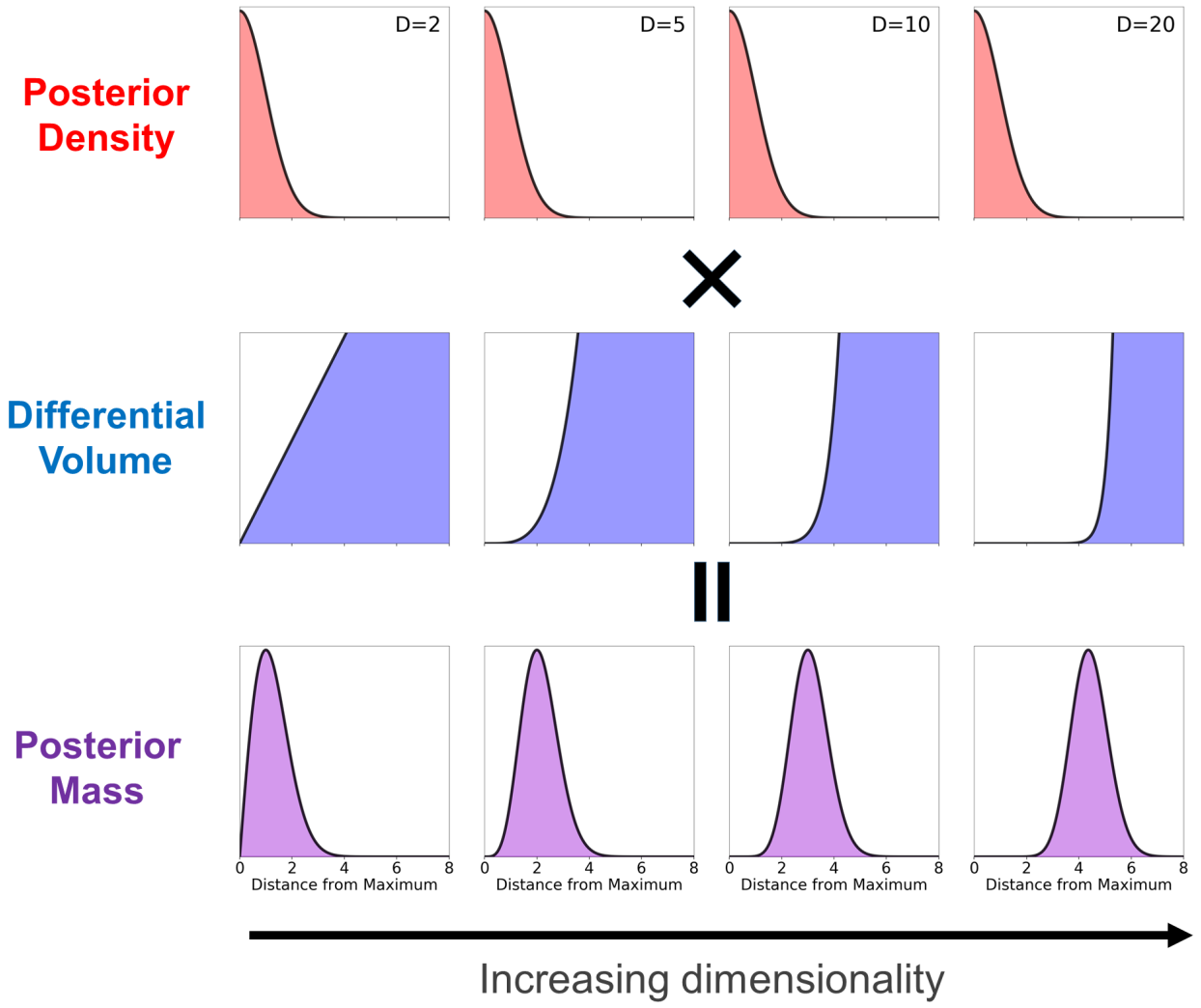

Figure 12: A schematic illustration of how the posterior mass behaves as a function of dimensionality using a d-dimensional Gaussian. The top panel shows the posterior density $\mathcal{P}(r) \propto e^{-r^2/2}$ (red) plotted as a function of distance r from the maximum posterior density at $r = 0$ as the number of dimensions d increases (left to right). As expected, this distribution remains constant. The middle panel shows the differential volume element $\mathrm{d}V(r) \propto r^{d-1}\mathrm{d}r$ (blue) of the corresponding shell at radius r. This illustrates the exponentially increasing volume contributed by shells further away from the maximum. The bottom panel shows corresponding "posterior mass" as a function of radius $\mathcal{P}'(r) \propto r^{d-1}\mathcal{P}(r) \propto r^{d-1}e^{-r^2/2}$ (purple). Due to the increasing amount of volume located further away from the maximum posterior density, we see that the majority of the posterior mass (and therefore of any samples we generate with MCMC) are actually located a shell located far away from the $r = 0$. See §7.3 for additional details.

This result has two immediate implications. First, *the majority of our samples are not located where the posterior density is maximized*. This is the result of an exponentially increasing number of parameter combinations, which allow a small handful of excellent fits to the data to be easily overwhelmed by a substantially larger number of mediocre fits. MCMC methods are therefore generally inefficient at locating and/or characterizing the region of peak posterior density.

Second, as d increases we generally would expect the radius of the shell containing the bulk of the posterior mass to increase, moving further and further away from the peak density due to the exponentially increasing available volume. Since the majority of our samples are located in this region, *our chain will spend the vast majority of time generating samples from this shell*.

This allows us to now outline exactly why it is challenging to propose samples efficiently in high dimensions:

1. To make sure our acceptance fractions remain reasonable, we need to ensure our proposed positions mostly lie within this shell of posterior mass.

2. However, obtaining an independent sample requires being able to (in theory) propose any position within this shell.

3. This means that our auto-correlation time will principally be set by how long it takes to "wander around" the shell, which will be a function of its overall size r', its width $\Delta r'$, and the number of dimensions d.

8 Application to a Simple Toy Problem

I now consider a concrete, detailed example to illustrate how all the concepts discussed in §6 and §7 come together in practice. Throughout this section, I will outline a number of analytic results and utilize several different MCMC sampling strategies to generate chains of samples. I strongly encourage interested readers to implement their own versions of the methods outlined here, which can be used to reproduce the numerical results from this section in their entirety.

8.1 Toy Problem

In this toy problem, we will take our (unnormalized) posterior to be a d-dimensional Gaussian (Normal) distribution with a mean of $\mu = 0$ and a standard deviation of σ in all dimensions:

$$\tilde{\mathcal{P}}(\boldsymbol{\Theta}) = \exp\left[-\frac{1}{2}\frac{|\boldsymbol{\Theta}|^2}{\sigma^2}\right] \tag{73}$$

where $|\boldsymbol{\Theta}|^2 = \sum_{i=1}^{d} \Theta_i^2$ is the squared magnitude of the position vector.

Based on the results from §7.3, we can better understand the properties of this distribution by rewriting the posterior density in terms of the "radius" $r \equiv |\boldsymbol{\Theta}| = \sqrt{\sum_{i=1}^{d} \Theta_i^2}$

away from the center:

$$\tilde{\mathcal{P}}(r) = \exp\left[-\frac{r^2}{2\sigma^2}\right] \tag{74}$$

The corresponding volume contained within a given radius r is then

$$V(r) \propto r^d \tag{75}$$

The corresponding posterior mass is $\tilde{\mathcal{P}}'(r)$ is then defined via

$$\tilde{\mathcal{P}}(V)\mathrm{d}V(r) \propto e^{-r^2/2\sigma^2} r^{d-1}\mathrm{d}r \equiv \tilde{\mathcal{P}}'(r)\mathrm{d}r$$

Note that this is closely related to the **chi-square distribution**.

The **typical radius** $r_{\rm peak}$ where the posterior mass peaks (i.e. is maximized) and a sample is most likely to be located can be derived by setting $\mathrm{d}\tilde{\mathcal{P}}'(r)/\mathrm{d}r = 0$. Solving this gives

$$r_{\rm peak} = \sqrt{d-1}\sigma \tag{76}$$

In other words, while in 1-D a typical sample is most likely to be located at the peak of the distribution with $r_{\rm peak} = 0$, in higher dimensions this changes quite drastically. While $r_{\rm peak} = 1\sigma$ in 2-D, it is 2σ in 5-D, 3σ in 10-D, and 5σ in 26-D. This is a direct consequence of the huge amount of volume at larger radii in high dimensions: although a sample at $r = 5\sigma$ has a posterior density $\mathcal{P}(r)$ orders of magnitude worse than a sample at $r = 0$, the enormous number of parameter combinations (volume) available at $r = 5\sigma$ more than makes up for it.

In general, we expect the posterior mass to comprise a **"Gaussian shell"** centered at some radius

$$r_{\rm mean} \equiv \mathbb{E}_{\mathcal{P}'}\left[r\right] = \int_0^\infty r\mathcal{P}'(r)\mathrm{d}r = \sqrt{2}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\sigma \approx \sqrt{d}\sigma \tag{77}$$

with a standard deviation of

$$\Delta r_{\rm mean} \equiv \sqrt{\mathbb{E}_{\mathcal{P}'}\left[(r - r_{\rm mean})^2\right]} = \sigma\sqrt{d - 2\left(\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\right)^2} \approx \frac{\sigma}{\sqrt{2}} \tag{78}$$

where $\Gamma(d)$ is the Gamma function and the approximations are taken for large d. See Figure 12 for an illustration of this behavior.

8.2 MCMC with Gaussian Proposals

Let us now consider a chain of samples $\{\boldsymbol{\Theta}_1 \rightarrow \cdots \rightarrow \boldsymbol{\Theta}_n\}$. The distance between two samples $\boldsymbol{\Theta}_m$ and $\boldsymbol{\Theta}_{m+t}$ separated by some lag t will be

$$|\boldsymbol{\Theta} - \boldsymbol{\Theta}'| = \sqrt{\sum_{i=1}^{d}(\Theta_{m,i} - \Theta_{m+t,i})^2} \tag{79}$$

Assuming that the lag $t \gg \tau$ is substantially larger than the auto-correlation time τ, we can assume each sample is approximately iid distributed following our Gaussian posterior. This then gives an expected separation of

$$\Delta r_{\rm sep} \equiv \sqrt{\mathbb{E}_{\mathcal{P}}\left[|\boldsymbol{\Theta}_m - \boldsymbol{\Theta}_{m+t}|^2\right]} = \sqrt{\sum_{i=1}^{d} \mathbb{E}_{\mathcal{P}}\left[(\Theta_{m,i} - \Theta_{m+t,i})^2\right]} = \sqrt{2d}\sigma \approx \sqrt{2} r_{\rm mean} \tag{80}$$

We can in theory propose samples in such a way so that the separation $|\boldsymbol{\Theta}_{i+1} - \boldsymbol{\Theta}_i|$ between a proposed position $\boldsymbol{\Theta}_{i+1}$ and the current position $\boldsymbol{\Theta}_i$ follows the ideal separation of $\sqrt{2} r_{\rm mean}$ derived above by using a simple Gaussian proposal distribution:

$$\mathcal{Q}(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i) \propto \exp\left[-\frac{1}{2}\frac{|\boldsymbol{\Theta}_{i+1} - \boldsymbol{\Theta}_i|^2}{2\sigma^2}\right] \tag{81}$$

While this proposal has the same *shape* as the posterior, it is centered on $\boldsymbol{\Theta}_i$ rather than 0. Using our intuition for how volume behaves based on §7.2, we can conclude that the majority of samples proposed from this choice of $\mathcal{Q}(\boldsymbol{\Theta}'|\boldsymbol{\Theta})$ will probably have little overlap with the posterior.

Indeed, numerical simulation suggests the typical fraction of positions that will be accepted given the above proposal roughly scales as

$$\langle f_{\rm acc}(d) \rangle \equiv \exp\left[\mathbb{E}_{\mathcal{P},\mathcal{Q}}\left[\ln T(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)\right]\right] \sim \exp\left[-\frac{d}{4} - \frac{1}{2}\right] \tag{82}$$

which decreases exponentially as the dimensionality increases, similar to Figure 11. Likewise, we find the auto-correlation time roughly scales as

$$\langle \tau(d) \rangle \equiv \exp\left[\mathbb{E}_{\mathcal{P},\mathcal{Q}}\left[\ln \tau\right]\right] \sim \exp\left[\frac{d}{4} + \frac{7}{4}\right] \tag{83}$$

This exponential dependence arises because the overlap between the typical Gaussian proposal $\mathcal{Q}(\boldsymbol{\Theta}'|\boldsymbol{\Theta})$ and the underlying posterior $\mathcal{P}(\boldsymbol{\Theta})$ essentially reduces to the small volume where two thin shells overlap. Since the radii of the shells goes as $\propto \sqrt{d}$ while the widths remain roughly constant, the "fractional size" of the shell (and the corresponding overlap) ends up decreasing exponentially.

To counteract this effect, we need to adjust the σ of our proposal distribution by some factor γ:

$$\mathcal{Q}_{\gamma}(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i) \propto \exp\left[-\frac{1}{2}\frac{|\boldsymbol{\Theta}_{i+1} - \boldsymbol{\Theta}_i|^2}{(\gamma\sigma)^2}\right] \tag{84}$$

where our previous proposal assumes $\gamma = \sqrt{2}$. If we want to ensure our typical acceptance fraction will remain roughly constant as a function of dimension d, γ needs to scale as

$$\langle f_{\rm acc}(\gamma(d)) \rangle \approx C \quad \Rightarrow \quad \gamma(d) \propto \frac{1}{\sqrt{d}} \tag{85}$$

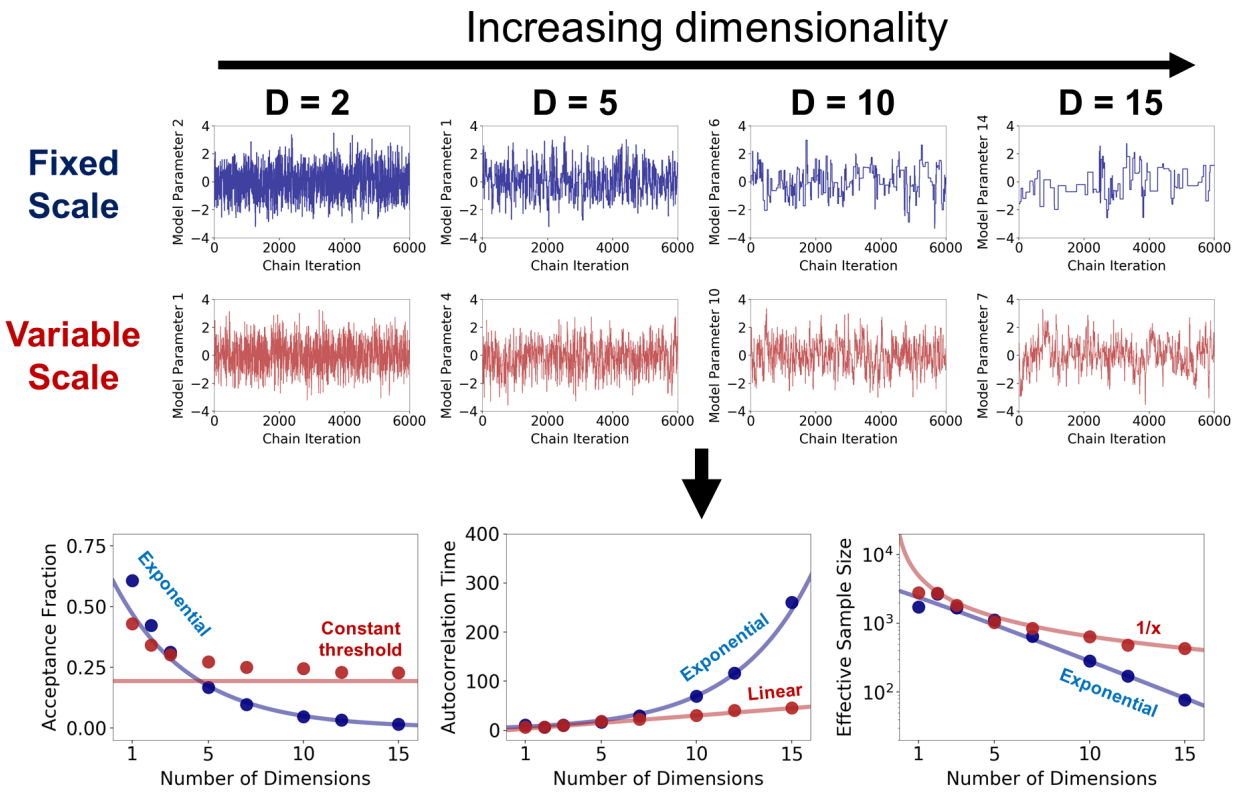

Figure 13: Numerical results showcasing the performance of a simple MH MCMC sampler with Gaussian proposals on our toy problem, a d-dimensional Gaussian with mean $\mu = 0$ and standard deviation $\sigma = 1$ in every dimension. The top series of panels show snapshots of a random parameter from the chain as a function of dimensionality (increasing from left to right) assuming an unchanging proposal with constant scale factor $\gamma = \sqrt{2}$ (blue) and a shrinking proposal with $\gamma = 2.5/\sqrt{d}$ designed to target a constant acceptance fraction of $\sim 25\%$ (red). The bottom panels show the corresponding acceptance fractions (left), auto-correlation times (middle), and effective sample sizes (right) from our chains (colored points) as a function of dimensionality. The approximations from §8.2 are shown as light colored lines. Shrinking the size of the proposal helps to keep samples within the bulk of the posterior mass, substantially reducing the auto-correlation time and increasing the effective sample size. Failing to do so leads to an exponentially decreasing fraction of good proposals and a corresponding exponential increase/decrease in the auto-correlation time/effective sample size. See §8.2 for additional discussion.

which inversely tracks the expected radius $r_{\rm mean}$ of the typical set. We find that taking

$$\gamma = \frac{\delta}{\sqrt{d}} \tag{86}$$

leads to a typical acceptance fraction of

$$\langle f_{\rm acc}(\delta/\sqrt{d}) \rangle \approx \exp\left[-\left(\frac{\delta^2}{4}\right)^2 - \frac{\delta}{2}\right] \tag{87}$$

as d becomes large with a typical auto-correlation time of

$$\langle \tau(\delta/\sqrt{d}) \rangle \approx 3d \tag{88}$$

for reasonable choices of δ. This linear dependence is a substantial improvement over our earlier exponential scaling.

Numerical Tests

To confirm these results, I sample from this d-dimensional Gaussian posterior (assuming $\sigma = 1$ for simplicity) using two MH MCMC algorithms for $n = 20,000$ iterations based on these proposal distributions. The first proposes new points assuming $\gamma = \sqrt{2}$. The second assumes $\gamma = 2.5/\sqrt{d}$ in order to maintain a roughly constant acceptance fraction of 25%. As shown in Figure 13, the chains behave as expected given our theoretical predictions as a function of dimensionality, with the constant proposal quickly becoming stuck while the adaptive proposal continues sampling normally. While the auto-correlation time τ increases in both cases, the increase in the latter case (where it is driven by decreasing size/scale of the proposal distribution) is much more manageable than the former (where it is driven by the exponentially decreasing acceptance fraction).

8.3 MCMC with Ensemble Proposals

One drawback to the Gaussian proposals explored above is that we have to specify the structure of the distribution ahead of time. In this specific case, we assumed that:

1. the width of the posterior in each dimension (parameter) was constant such that $\sigma_1 = \sigma_2 = \cdots = \sigma_n = \sigma$ and
2. the parameters were entirely uncorrelated with each other such that the correlation coefficient $\rho_{ij} = 0$ between any two dimensions i and j.

In general, there is no good reason to assume that either of these are true. This means we have to also estimate the entire set of $d(d+1)/2$ free parameters that determine the overall covariance structure of our unknown posterior distribution. Trying to adjust the covariance structure in order to improve our sampling efficiency and decrease the auto-correlation time (see §5.3 and §6.2) becomes one of the most difficult parts of running MCMC algorithms in practice.

While there are schemes to perform these adjustments during an extended burn-in period (see, e.g., Brooks et al. 2011), there is significant appeal in methods that can "auto-tune" without much additional input from the user. One class of such approaches are known as **ensemble** or **particle** methods. These methods attempt to use many m chains running simultaneously (i.e. in parallel) to improve the performance of any individual chain.

We explore three variations of ensemble methods here that attempt to exploit $m \gtrsim d(d+1)/2$ chains running simultaneously:

1. using the ensemble of particles to condition a Gaussian proposal distribution,
2. using trajectories from multiple particles along with Gaussian "jitter", and
3. using affine-invariant transformations of trajectories from multiple particles.

A schematic illustration of these methods is shown in Figure 14.

As we might expect, an immediate drawback of these methods is they rely on having enough particles to characterize the overall structure of the space (i.e. the curse of dimensionality). While this limits their utility when sampling from high-dimensional spaces, they can be attractive options in moderate-dimensional spaces ($d \lesssim 25$) where a few hundred particles are often sufficient to ensure reasonable performance.

8.3.1 Gaussian Proposal

The first approach is simply a modified Gaussian proposal: at any iteration i for any chain j, we propose a new position $\boldsymbol{\Theta}^j_{i+1}$ based on the current position $\boldsymbol{\Theta}^j_i$ using a Gaussian proposal

$$\mathcal{Q}^j_\gamma(\boldsymbol{\Theta}^j_{i+1}|\boldsymbol{\Theta}^j_i) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\Theta}^j_{i+1} - \boldsymbol{\Theta}^j_i)^{\mathrm{T}}(\gamma^2\mathbf{C}^j_i)^{-1}(\boldsymbol{\Theta}^j_{i+1} - \boldsymbol{\Theta}^j_i)\right] \tag{89}$$

where T is the transpose operator and

$$\mathbf{C}^j_i = \mathrm{Cov}\left[\{\boldsymbol{\Theta}^1_i, \ldots, \boldsymbol{\Theta}^{j-1}_i, \boldsymbol{\Theta}^{j+1}_i, \ldots, \boldsymbol{\Theta}^m_i\}\right] \tag{90}$$

is the empirical covariance matrix estimated from the current positions of the m chains *excluding* chain j. We repeat this process for each of the m chains in turn.

In other words, at each iteration i we want to update all m chains. We do so by updating each chain j in turn based on what the other chains are currently doing. Assuming the current position of each chain is distributed following the underlying posterior $\mathcal{P}(\boldsymbol{\Theta})$, it is straightforward to show that $\mathbf{C}^j_i$ is a reasonable approximation to the unknown covariance structure of our posterior. In addition, because we exclude j when computing $\mathbf{C}^j_i$, this proposal is symmetric going from $\boldsymbol{\Theta}^j_i \rightarrow \boldsymbol{\Theta}^j_{i+1}$ and from $\boldsymbol{\Theta}^j_{i+1} \rightarrow \boldsymbol{\Theta}^j_i$. This means that we satisfy detailed balance and do not have to incorporate any proposal-dependent factors when computing the transition probability.

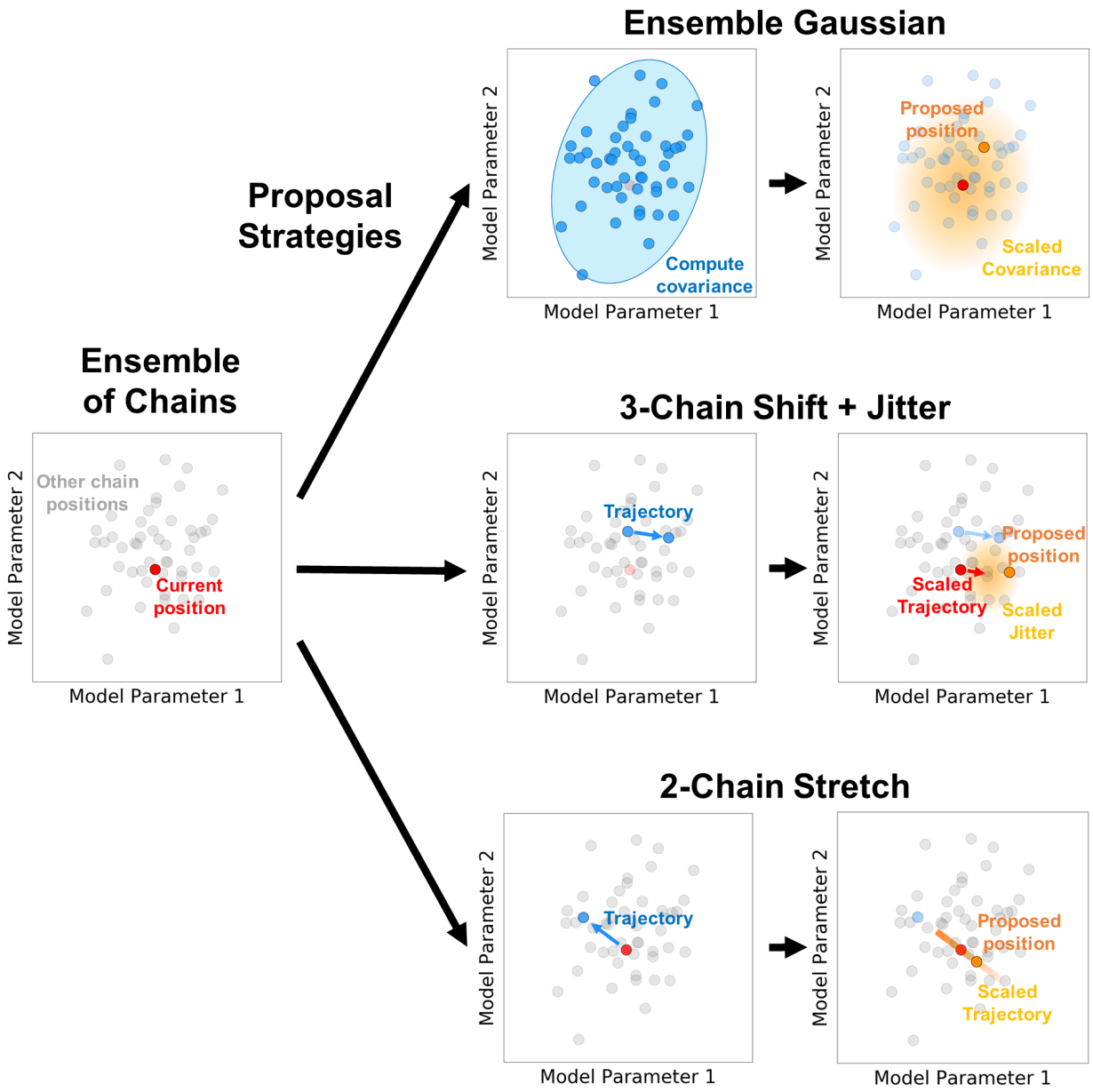

Figure 14: A schematic illustration of the three ensemble MCMC methods described in §8.3. The current state of the chain we are interested in updating (red) and the other chains in the ensemble (gray) are shown on the left. In the top panels (ensemble Gaussian; §8.3.1), we compute the covariance of the other $k \neq j$ chains (middle) and use a scaled version to subsequently propose a new position. In the middle panels (3-chain shift + jitter; §8.3.2), we use two additional chains $k \neq l \neq j$ to compute a trajectory. We then propose a new position based on this scaled trajectory plus a small amount of "jitter". In the bottom panels (2-chain stretch; §8.3.3), we use only one additional chain $k \neq j$ to propose a new trajectory. We then propose a random position along a scaled version of this trajectory with the proposal probability varying as a function of scale. See §8.3 for additional details.

8.3.2 Ensemble Trajectories with a Gaussian Proposal

The approach taken in §8.3.1 solves the problem of trying to tune the covariance of our initial Gaussian proposal. However, it still assumes that a Gaussian proposal is the optimal solution. A more general approach is one that does not rely on assuming a proposal explicitly, but rather only relies on the distribution of the remaining particles.

One such approach used in the literature is **Differential Evolution** MCMC (DE-MCMC; Storn & Price, 1997; Ter Braak, 2006). The main idea behind DE-MCMC is to rely on the *relative positions* of the chains at a given iteration i when making new proposals. We first randomly select two other particles k and l where $\boldsymbol{\Theta}_i^j \neq \boldsymbol{\Theta}_i^k \neq \boldsymbol{\Theta}_i^l$. We then propose a new position based on the vector distance between the other two particles $\boldsymbol{\Theta}_i^k - \boldsymbol{\Theta}_i^l$ with some scaling γ along with some additional "jitter" ϵ:

$$\boldsymbol{\Theta}_{i+1}^j = \boldsymbol{\Theta}_i^j + \gamma \times (\boldsymbol{\Theta}_i^k - \boldsymbol{\Theta}_i^l + \epsilon) \tag{91}$$

In the case where the behavior of chains k and l are approximately independent of each other and assuming the underlying posterior distribution $\mathcal{P}(\boldsymbol{\Theta})$ is Gaussian with some unknown mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$ (and "standard deviation" $\mathbf{C}^{1/2}$), it is straightforward to show that the distribution of $\boldsymbol{\Theta}_i^k - \boldsymbol{\Theta}_i^l$ will then follow

$$\boldsymbol{\Theta}_i^k - \boldsymbol{\Theta}_l \sim \mathcal{N}\left[\mathbf{0}, (2\mathbf{C})^{1/2}\right] \tag{92}$$

Typically, the jitter ϵ is chosen to also be Gaussian distributed with covariance $\mathbf{C}_\epsilon$ such that

$$\epsilon \sim \mathcal{N}\left[\mathbf{0}, \mathbf{C}_\epsilon^{1/2}\right] \tag{93}$$

In general, $\mathbf{C}_\epsilon$ is mostly used to try and avoid issues caused by finite particle sampling: since the number of unique trajectories (ignoring symmetry) is

$$n_{\rm traj} = \binom{m-1}{2} = \frac{(m-1)!}{2!(m-3)!} = \frac{(m-1)(m-2)}{2}$$

if m is sufficiently small the DE-MCMC procedure can only explore a small number of possible trajectories at any given time, leading to extremely inefficient sampling.

Combined, this implies that the proposed position has a distribution of

$$\boldsymbol{\Theta}_{i+1}^j \sim \mathcal{N}\left[\boldsymbol{\Theta}_i^j, \gamma \times (2\mathbf{C} + \mathbf{C}_\epsilon)^{1/2}\right] \tag{94}$$

This shows that the 3-particle DE-MCMC procedure can generate new positions in a manner analogous to the ensemble Gaussian proposal we first discussed.

8.3.3 Affine-Invariant Transformations of Ensemble Trajectories

Another approach used in the literature (e.g., `emcee`; Foreman-Mackey et al., 2013) is the **Affine-Invariant "stretch move"** from Goodman & Weare (2010). This uses only one additional particle $\boldsymbol{\Theta}_i^k$ rather than two:

$$\boldsymbol{\Theta}_{i+1}^j = \boldsymbol{\Theta}_i^k + \gamma \times (\boldsymbol{\Theta}_i^j - \boldsymbol{\Theta}_i^k) \tag{95}$$

In place of the jitter term ϵ from DE-MCMC, the stretch move instead injects some amount of randomness by allowing γ to vary. By sampling γ from some probabilty distribution $g(\gamma)$, we allow the proposals to explore various "stretches" of the direction vector. As shown in Goodman & Weare (2010), if this function is chosen such that

$$g(\gamma^{-1}) = \gamma \times g(\gamma) \tag{96}$$

then this proposal is symmetric. Typically, $g(\gamma)$ is chosen to be

$$g(\gamma|a) = \begin{cases} \gamma^{-1/2} & a^{-1} \leq \gamma \leq a \\ 0 & \text{otherwise} \end{cases} \tag{97}$$

where $a = 2$ is often taken as a typical value. Note that when $\gamma = 1$, this move leaves $\boldsymbol{\Theta}^j_{i+1} = \boldsymbol{\Theta}^j_i$ unchanged.

Compared to DEMCMC, the stretch move appears to have one clear advantage: it doesn't have any reliance on some "jitter" term ϵ that reintroduces scale-dependence into the proposal. That makes the proposal invariant to affine transformations and only sensitive to *a single parameter* a, which governs the range of scales the stretch factor γ is allowed to explore.

This lack of jitter, however, is not substantially advantageous in practice. As noted in §8.3.2, ϵ is really designed to avoid possible degeneracies due to the limited number of available trajectories. In that case we had $(m-1)(m-2)/2 \sim m^2/2$ possible trajectories; here, however, we only have m (since $\boldsymbol{\Theta}^j_i$ is always included). This is a *much* smaller number of possible trajectories at a given m, making this particular proposal more susceptible to that particular effect.

In addition, because this proposal involves adjusting γ and therefore the length of the trajectory itself, we need to consider how changing γ affects the total *volume* of the sphere centered on $\boldsymbol{\Theta}^j_i$ with radius $\boldsymbol{\Theta}^k_i - \boldsymbol{\Theta}^j_i$. As discussed in §7.2, the differential volume increases as r^{d-1}. Therefore, increasing or decreasing γ substantially adjusts the differential volume in our proposal. This involves introducing a steep boost/penalty into our transition probability, which now becomes:

$$T(\boldsymbol{\Theta}^j_{i+1}|\boldsymbol{\Theta}^j_i, \gamma) = \min\left[1, \gamma^{d-1}\frac{\mathcal{P}(\boldsymbol{\Theta}^j_{i+1})}{\mathcal{P}(\boldsymbol{\Theta}^j_i)}\right] \tag{98}$$

This heavily favors proposals with $\gamma > 1$ (outwards) and heavily disfavors proposals with $\gamma < 1$ as d increases to account for the exponentially increasing volume at larger radii.

Finally, while this stretch move actually generates proposals in the right overall *direction*, it is not efficient at generating samples within the bulk of the posterior mass as the dimensionality increases. As discussed in §8.2, given the typical position of $\boldsymbol{\Theta}^j_i$, the typical length-scale of the proposed positions needs to shrink by $\propto 1/\sqrt{d}$ in order to guarantee our new sample remains within the bulk of the posterior mass. However, the form for $g(\gamma|a)$ specified above instead ensures that γ will always be between $1/a$ and a. Even if we attempt to account for this effect by letting $a(d) \to 1$ as $d \to \infty$ in order to target a constant acceptance fraction and ensure more overlap, the asymmetry of our proposal

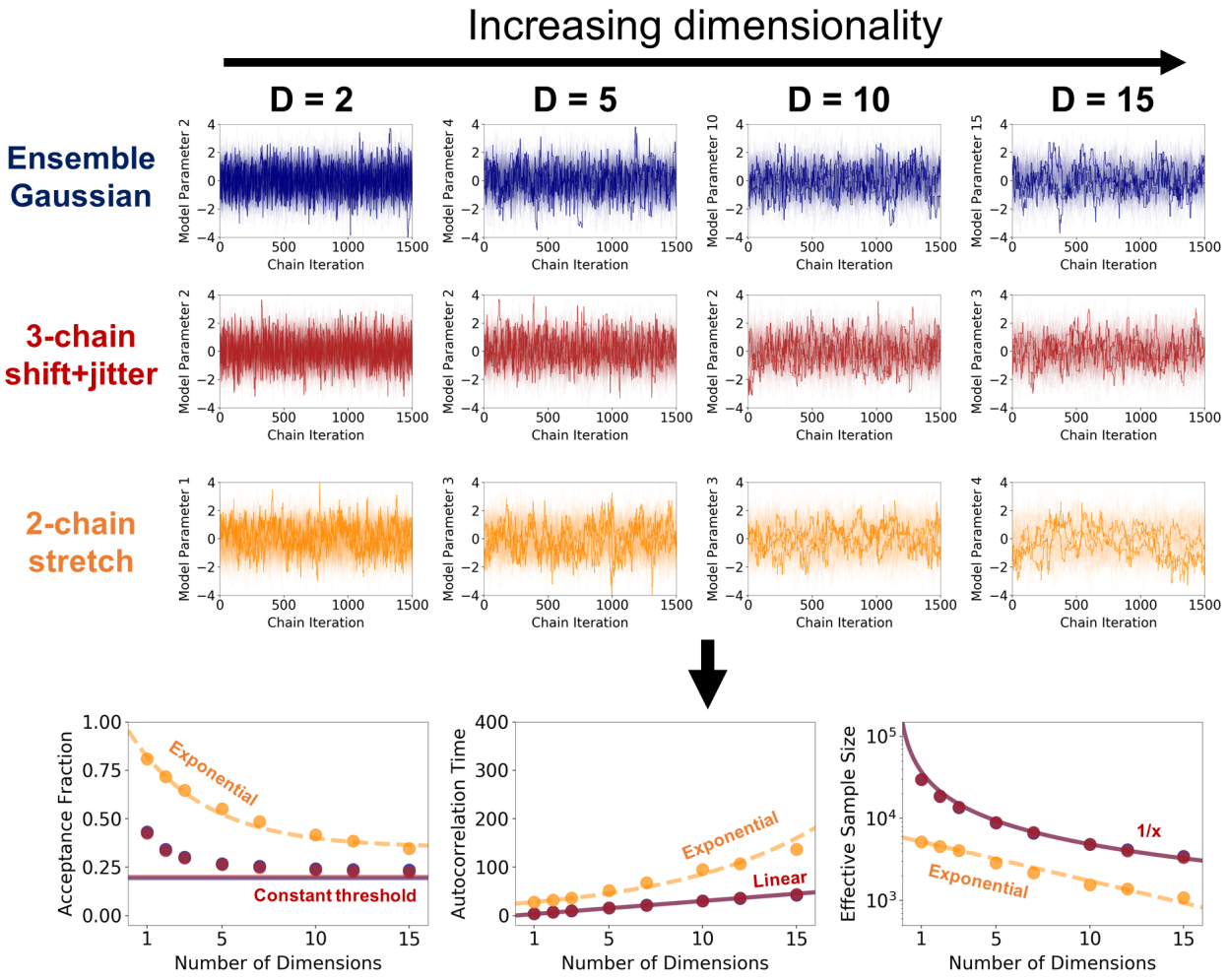

Figure 15: Numerical results showcasing the performance of several ensemble MH MCMC samplers on our toy problem, a d-dimensional Gaussian with mean $\mu = 0$ and standard deviation $\sigma = 1$ in every dimension. The top series of panels show snapshots of a random parameter from the collection of chains (with a few chains highlighted) as a function of dimensionality (increasing from left to right) assuming ensemle Gaussian proposals with $\gamma = 2.5/\sqrt{d}$ (blue), 3-chain "shift and jitter" proposals with $\gamma = 1.7/\sqrt{d}$ (red), and 2-chain "stretch" proposals with γ drawn from the distribution $g(\gamma|a)$ with $a = 2$ as described in §8.3.3 (orange). The bottom panels show the corresponding acceptance fractions (left), auto-correlation times (middle), and effective sample sizes (right) from our chains (colored points) as a function of dimensionality. Approximations based on the §8.2 are shown as light solid colored lines, with dashed lines showing rough fits. The first two methods, which allow the size of the proposal to shrink, are able to propose samples within the bulk of the posterior mass. The last method, which is unable to do so, instead proposes exponentially fewer good positions as the dimensionality increases. See §8.3 for additional details.

and the γ^{d-1} term in the transition probability systematically biases our proposed and accepted positions compared with the ideal distribution. This subsequently leads to lager auto-correlation times, mostly counteracting any expected gains.

Numerical Tests

To confirm these results, I sample from this d-dimensional Gaussian posterior (assuming $\sigma = 1$ for simplicity) using each of these ensemble MH MCMC algorithms with for $n = 1500$ iterations with $m = 100$ chains. In the first case, I propose a new position for chain j at iteration i using a Gaussian distribution with a covariance $\gamma^2 \mathbf{C}_i^j$ computed over the remaining ensemble of $k \neq j$ chains, where the scale factor $\gamma = 2.5/\sqrt{d}$ is chosen to target a constant acceptance fraction of roughly 25%. In the second case, I propose new positions using the DE-MCMC algorithm with a scale factor of $\gamma = 1.7/\sqrt{d}$ and additional Gaussian jitter with covariance $\mathbf{C}_\epsilon = \mathbf{C}_i^j/5$ derived from the remaining chains in the ensemble, again targeting an acceptance fraction of roughly 25%. In the third case, I propose new positions using the affine-invariant stretch move assuming the typical form for $g(\gamma|a)$ with $a = 2$.[7]

As shown in Figure 15, the chains behave as expected given our theoretical predictions as a function of dimensionality. Similar to the adaptive Gaussian case, the first two approaches continue sampling efficiently even as d increases. The affine-invariant stretch move, however, experiences exponentially-decreasing efficiency and struggles to sample the posterior effectively.

8.4 Additional Comments

Before concluding, I wish to emphasize that the toy problem explored in this section should only be interpreted as a *tool* to build intuition surrounding how certain methods are expected to behave in a controlled environment. While the behavior as a function of dimensionality helps to illustrate common issues, in practice the performance of any method will depend on the specific problem, tuning parameters, the time spent on tuning, and many other possible factors. Since it is always possible to find problems for which any particular method will perform well or poorly, I encourage users to try out a variety of approaches to find the ones that work best for their problems.

9 Conclusion

Bayesian statistical methods have become increasingly prevalent in modern scientific analysis as models have become more complex. Exploring the inferences we can draw from these models often requires the use of numerical techniques, the most popular of which is known as **Markov Chain Monte Carlo (MCMC)**.

In this article, I provide a conceptual introduction to MCMC that seeks to highlight the *what*, *why*, and *how* of the overall approach. I first give an overview of Bayesian

[7] Allowing $a(d)$ to vary as a function of dimensionality to target a roughly constant acceptance fraction gives similar results.

inference and discuss *what* types of problems Bayesian inference generally is trying to solve, showing that most quantities we are interested in computing require integrating over the posterior density. I then outline approaches to computing these integrals using grid-based approaches, and illustrate how adaptively changing the resolution of the grid naturally transitions into the use of Monte Carlo methods. I illustrate how different sampling strategies affect the overall efficiency in order to motivate *why* we use MCMC methods. I then discuss various details related to *how* MCMC methods work and examine their expected overall behavior based on simple arguments derived from how volume and posterior density behave as the number of parameters increases. Finally, I highlight the impact this conceptual understanding has in practice by comparing the performance of various MCMC methods on a simple toy problem.

I hope that the material in this article, along with the exercises and applications, serve as a useful resource that helps build up intuition for how MCMC and other Monte Carlo methods work. This intuition should be helpful when making decisions over when to apply MCMC methods to your own problems over possible alternatives, developing novel proposals and sampling strategies, and characterizing what issues you might expect to encounter when doing so.

Acknowledgements

JSS is grateful to Rebecca Bleich for continuing to tolerate his (over-)enthusiasm for sampling during their time together. He would also like to thank a number of people for helping to provide much-needed feedback during earlier stages of this work, including Catherine Zucker, Dom Pesce, Greg Green, Kaisey Mandel, Joel Leja, David Hogg, Theron Carmichael, and Jane Huang. He would also like to thank Ana Bonaca, Charlie Conroy, Ben Cook, Daniel Eisenstein, Doug Finkbeiner, Boryana Hadzhiyska, Will Handley, Locke Patton, and Ioana Zelko for helpful conversations surrounding the material.

JSS also wishes to thank Kaisey Mandel and the Institute of Astronomy at the University of Cambridge, Hans-Walter Rix and the Galaxies and Cosmology Department at the Max Planck Institute for Astronomy, and Renée Hložek, Bryan Gaensler, and the Dunlap Institute for Astronomy and Astrophysics at the University of Toronto for their kindness and hospitality while hosting him over the period where a portion of this work was being completed.

JSS acknowledges financial support from the National Science Foundation Graduate Research Fellowship Program (Grant No. 1650114) and the Harvard Data Science Initiative.

References

Asmussen, S., & Glynn, P. W. 2011, Statistics & Probability Letters, 81, 1482 , doi: `https://doi.org/10.1016/j.spl.2011.05.004`

Blitzstein, J., & Hwang, J. 2014, Introduction to Probability, Chapman & Hall/CRC Texts

in Statistical Science (CRC Press/Taylor & Francis Group). https://books.google.com/books?id=ZwSlMAEACAAJ

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. 2011, Handbook of Markov Chain Monte Carlo (CRC press)

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, Pub. of the Astron. Sco. of the Pac., 125, 306, doi: 10.1086/670067

Gelman, A., Carlin, J., Stern, H., et al. 2013, Bayesian Data Analysis, Third Edition, Chapman & Hall/CRC Texts in Statistical Science (Taylor & Francis). https://books.google.com/books?id=ZXL6AQAAQBAJ

Gelman, A., & Meng, X.-L. 1998, Statist. Sci., 13, 163, doi: 10.1214/ss/1028905934

Gelman, A., & Rubin, D. B. 1992, Statistical Science, 7, 457, doi: 10.1214/ss/1177011136

Goodman, J., & Weare, J. 2010, Communications in Applied Mathematics and Computer Science, 5, 65, doi: 10.2140/camcos.2010.5.65

Hastings, W. 1970, Biometrika, 57, 97, doi: 10.1093/biomet/57.1.97

Hogg, D. W., & Foreman-Mackey, D. 2018, The Astrophys. Journal Supp., 236, 11, doi: 10.3847/1538-4365/aab76e

Kish, L. 1965, Survey sampling, Wiley classics library (J. Wiley). https://books.google.com/books?id=xiZmAAAAIAAJ

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, Journal of Chem. Phys., 21, 1087, doi: 10.1063/1.1699114

Neal, R. M. 2012, arXiv e-prints, arXiv:1206.1901. https://arxiv.org/abs/1206.1901

Skilling, J. 2004, in American Institute of Physics Conference Series, Vol. 735, American Institute of Physics Conference Series, ed. R. Fischer, R. Preuss, & U. V. Toussaint, 395–405

Skilling, J. 2006, Bayesian Anal., 1, 833, doi: 10.1214/06-BA127

Storn, R., & Price, K. 1997, Journal of global optimization, 11, 341

Ter Braak, C. J. 2006, Statistics and Computing, 16, 239

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. 2019, arXiv e-prints, arXiv:1903.08008. https://arxiv.org/abs/1903.08008