Chapter 1

Bayesian Statistics

1.1 Introduction to Bayesian Statistics

In the realm of science, the accessibility of information is inherently limited, making our understanding of nature inherently probabilistic. Scientific analyses typically rely on drawing conclusions about fundamental physical models from diverse sets of observational data. There are two distinct methodologies, grounded in different interpretations of probability. In traditional statistics, the probability of an event is equated with the long-term relative frequency of its occurrence. This perspective is commonly known as the *frequentist* view, wherein probabilities are confined to discussions of random variables—quantities that can meaningfully fluctuate across a sequence of repeated experiments.

In recent years, there has been a shift in the understanding of probability, acknowledging that the mathematical principles of probability extend beyond calculating frequencies of random variables. These principles are now recognized as inherently valid rules of logic for making inferences about any given proposition or hypothesis. This more robust perspective, often referred to as *Probability Theory as Logic* or *Bayesian probability* theory, is gaining prominence in physics and astronomy. The Bayesian approach enables the direct computation of the probability associated with a specific theory or a particular value of a model parameter—issues that the conventional statistical approach can only indirectly address through the use of random variable statistics. The two distinct approaches to statistical inference, along with their underlying definitions of probability, are summarized in Table 1.1. This lecture will predominantly delve into the concepts of Bayesian statistics.

Approach	Frequentist	Bayesian
Definition of Probability	Probability is seen as the long-run relative frequency of an event occurring in repeated, independent experiments. It is based on objective, observable frequencies.	Probability is seen as a measure of belief or certainty about an event. It incorporates both prior knowledge and new evidence to update beliefs.
Parameters	Parameters are fixed, unknown values. Inference is about estimating these fixed values based on observed data.	Parameters are considered random variables with probability distributions. Inference involves updating prior distributions with observed data to obtain posterior distributions.
Subjectivity	It is considered an objective approach, as probabilities are based on observed frequencies, and conclusions are not influenced by subjective beliefs.	Acknowledges subjectivity, as it allows the incorporation of prior beliefs. Bayesian inference is sensitive to the choice of priors.
Hypothesis Testing	Emphasizes hypothesis testing, focusing on rejecting or failing to reject null hypotheses based on the observed data.	While hypothesis testing is possible, Bayesian inference often focuses on estimating parameters and updating beliefs rather than strict hypothesis testing.
Prior Information	Typically does not incorporate prior beliefs or subjective information about parameters.	Incorporates prior information, allowing researchers to include existing knowledge or beliefs about parameters in the analysis.

Table 1.1: Frequentist and Bayesian approaches to probability.

1.1.1 Deductive logic versus plausible reasoning

A schematic representation of deductive logic is show in Figure 1.1(a): given a cause, we can work out its consequences. The sort of reasoning used in pure mathematics is of this type: that is, we can derive many complicated and useful results as the logical consequence of a few well-defined axioms. Most scientists, however, face the reverse of the above situation: Given that certain effects have been observed, what is (are) the underlying cause(s)? This type of question has to do with inductive logic, or plausible reasoning, and is illustrated in Figure 1.1(b); the greater complexity of this diagram is

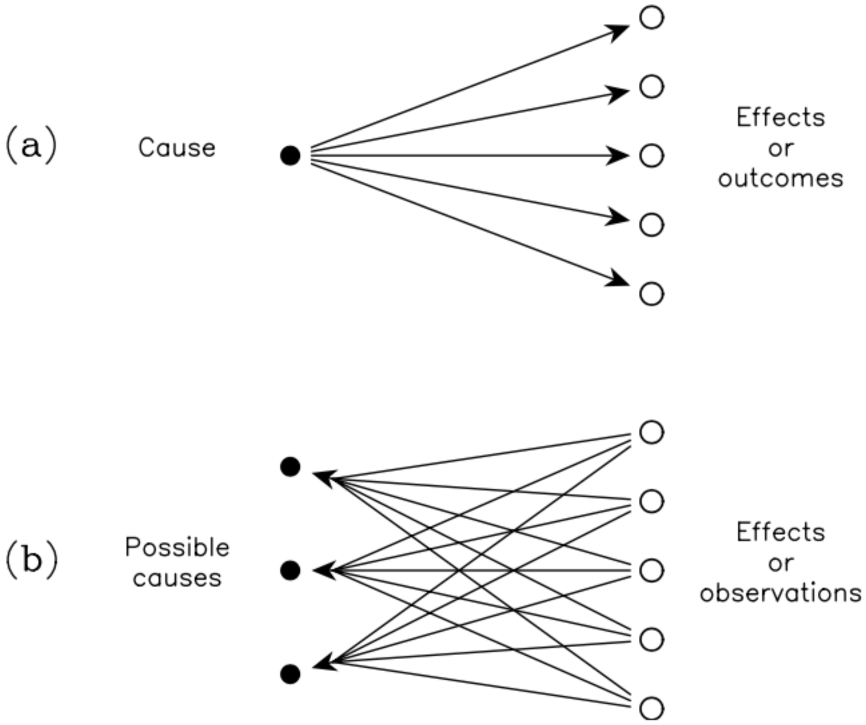


Figure 1.1: A schematic representation of (a) deductive logic, or pure mathematics, and (b) plausible reasoning, inductive logic.

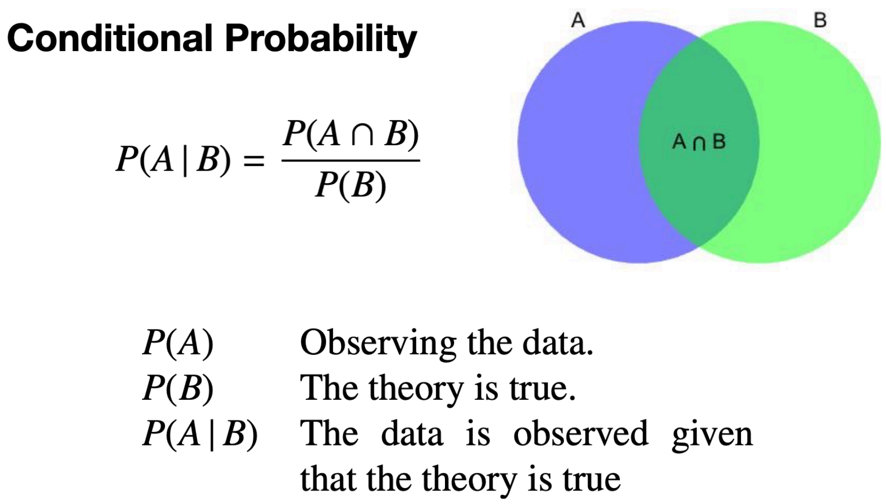
designed to indicate that it is a much harder problem. The most we can hope to do is to make the best inference based on the experimental data and any prior knowledge that we have available, reserving the right the revise our position if new information comes to light.

1.1.2 Bayes's Theorem

The goal of Bayesian probability theory is to provide an extension of logic to handle situations where we have incomplete information so we may arrive at the relative probabilities of competing hypotheses for a given state of information. The main idea in Bayesian probability is that we always based our believes on some prior information which could be assigned a conditional probability as shown in Figure 1.2. The operations for manipulating probabilities that follow from the desiderata are the sum and the product rules.

Sum Rule

$$P(A|B) + P(\overline{A}|B) = 1. \tag{1.1}$$



Conditional Probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$P(A)$	Observing the data.
$P(B)$	The theory is true.
$P(A \mid B)$	The data is observed given that the theory is true

Figure 1.2: A conditional probability of event A assuming that event B occurs.

Product Rule

$$\begin{aligned} P(A \cap B|C) &= P(A|C)P(B|A \cap C) \\ &= P(B|C)P(A|B \cap C), \end{aligned} \tag{1.2}$$

where *conjoint probability* or *conditional probability* $P(A|B)$ is given by

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B|A) \\ &= P(B) \cdot P(A|B), \end{aligned} \tag{1.3}$$

or,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.4}$$

Since in reality we justify our probability based on some prior information which is presumed to be valid. Hence,

$$P(A) \equiv P(A|I), \tag{1.5}$$

where I is proposition regarding our prior information and $P(A|I)$. The Bayes's theorem is based on the product rule and is given by

$$P(A|B,I) \equiv P(A|B \cap I) = \frac{P(A|I) \cdot P(B|A,I)}{P(B|I)}. \tag{1.6}$$

From here on, we shall use notation $P(A,B) \equiv P(A \cap B)$. The Bayes's theorem in Eq. (1.6) has far greater application in science if we replace the abstract notation A and B with something physically more meaningful;

In many scientific applications, we have access to some data **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** $\mathbf{\Theta}_M$ of that particular model. We can combine these pieces together to estimate the probability $P(\mathbf{D}|\mathbf{\Theta}_M, M)$ that we would actually see that data **D** we have collected *conditioned on* (i.e. assuming) a specific choice of parameters $\mathbf{\Theta}_M$ from our model M. In other words, assuming our model M is right and the parameters $\mathbf{\Theta}_M$ describe the data, what is the **likelihood** $P(\mathbf{D}|\mathbf{\Theta}_M, M)$ of the parameters $\mathbf{\Theta}_M$ based on the observed data **D**? Assuming different values of $\mathbf{\Theta}_M$ 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, $P(\mathbf{\Theta}_M|\mathbf{D}, M)$. This describes the probability that the underlying *parameters* are actually $\mathbf{\Theta}_M$ given our data **D** and assuming a particular model M. By using factoring of probability, we can relate this new probability $P(\mathbf{\Theta}_M|\mathbf{D}, M)$ to the likelihood $P(\mathbf{D}|\mathbf{\Theta}_M, M)$ described above as

$$P(\mathbf{\Theta}_M|\mathbf{D}, M)P(\mathbf{D}|M) = P(\mathbf{\Theta}_M, \mathbf{D}|M) = P(\mathbf{D}|\mathbf{\Theta}_M, M)P(\mathbf{\Theta}_M|M) \tag{1.7}$$

where $P(\mathbf{\Theta}_M, \mathbf{D}|M)$ represents the *joint* probability of having an underlying set of parameters $\mathbf{\Theta}_M$ that describe the data and observing the particular set of data **D** we have already collected. Rearranging this equality into a more convenient form gives us **Bayes' Theorem**:

$$P(\mathbf{\Theta}_M|\mathbf{D}, M) = \frac{P(\mathbf{D}|\mathbf{\Theta}_M, M)P(\mathbf{\Theta}_M|M)}{P(\mathbf{D}|M)}, \tag{1.8}$$

where $\boldsymbol{\Theta}_M \equiv$ proposition asserting the truth of the model parameters

$M \equiv$ proposition representing our model or prior information

$\mathbf{D} \equiv$ proposition representing data

In the language of Bayesian statistics, these terms are

$P(\boldsymbol{\Theta}_M|M) \equiv$ prior probability of the model parameters (prior, $\pi(\boldsymbol{\Theta}_M|M)$)

$P(\mathbf{D}|\boldsymbol{\Theta}_M, M) \equiv$ probability of obtaining the data $\mathbf{D}$, if $\boldsymbol{\Theta}_M$ and M are true (also called the likelihood function $\mathcal{L}(\boldsymbol{\Theta}_M)$)

$P(\boldsymbol{\Theta}_M|\mathbf{D}, M) \equiv$ posterior probability of $\boldsymbol{\Theta}_M$ (posterior $\mathcal{P}(\boldsymbol{\Theta})$)

$P(\mathbf{D}|M)$ is called *evidence* which is sometimes ignored in the calculation as a normalization factor. The normalization factor will ensure that

$$\sum_i P(\boldsymbol{\Theta}_M|\mathbf{D}, M) = 1. \tag{1.9}$$

In the case where we have a continuous hypothesis space, the normalization condition will be

$$\int \mathrm{d}\boldsymbol{\Theta}_M P(\boldsymbol{\Theta}_M|\mathbf{D}, M) = 1. \tag{1.10}$$

In summary, we could write our Bayesian's theorem as

$$\boxed{\begin{aligned} P(\text{Hypothesis}|\text{Data, Prior Information}) \quad \propto \quad & P(\text{Data}|\text{Hypothesis, Prior Information}) \\ & \times P(\text{Hypothesis}|\text{Prior Information}) \end{aligned}} \tag{1.11}$$

or

$$\boxed{\text{Posterior} \quad \propto \quad \text{Likelihood} \times \text{Prior}} \tag{1.12}$$

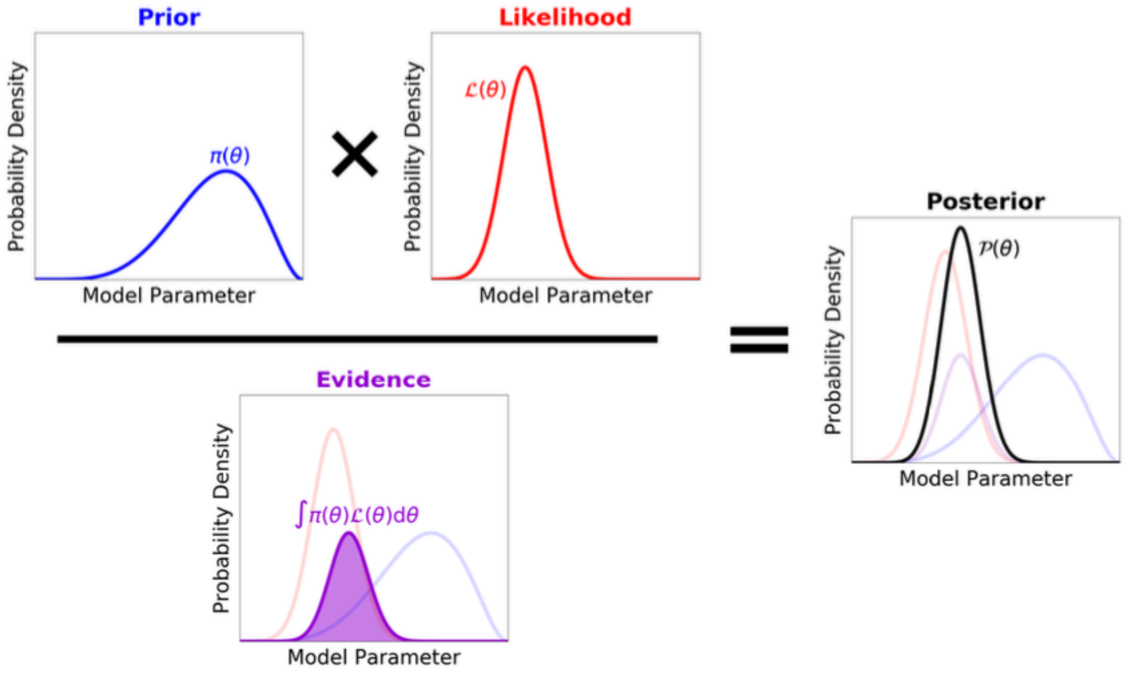


Figure 1.3: An illustration of Bayes' Theorem. The posterior probability $\mathcal{P}(\Theta)$ (black) of our model parameters Θ is based on a combination of our prior beliefs $\pi(\Theta)$ (blue) and the likelihood $\mathcal{L}(\Theta)$ (red), normalized by the overall evidence $\mathcal{Z} = \int \pi(\Theta)\mathcal{L}(\Theta)\mathrm{d}\Theta$ (purple) for our particular model.

Example 1.1 Monty Hall Problem

- Monty shows you three closed doors and tells you that there is a prize behind each door: one prize is a car the other two are less valuable prizes like peanut butter and fake finger nails. The prizes are arranged at random.

- The object of the game is to guess which door has the car. If you guess right, you get to keep the car.

- You pick a door, which we will call Door A. We'll call the other doors B and C.

- Before opening the door you chose, Monty increases the suspense by opening either Door B or C, whichever does not have the car. (If the car is actually behind Door A, Monty can safely open B or C, so he chooses one at random.)

- Then Monty offers you the option to stick with your original choice or switch to the one remaining unopened door.

The question is, should you "stick" or "switch" or does it make no difference? Most people have a strong intuition that it makes no difference. There are two doors left, they reason, so the chance that the car is behind Door A is 50%. But that is wrong. In fact, the chance of winning if you stick with Door A is only 1/3; if you switch, your chances are 2/3.

Choice	**Prior** $\pi(\Theta)$	**Likelihood** $\mathcal{L}(\Theta)$	$\pi(\Theta)\cdot\mathcal{L}(\Theta)$	**Posterior** $\mathcal{P}(\Theta\|D)$
A	1/3	1/2	1/6	1/3
B	1/3	0	0	0
C	1/3	1	1/3	2/3

Table 1.2: A Monty Hall problem we the contestant chose A initially and Monty showed that B is empty.

Filling the priors is easy because we are told that the prizes are arranged at random, which suggests that the car is equally likely to be behind any door. Figuring out the likelihoods takes some thought, but with reasonable care we can be confident that we have it right:

- If the car is actually behind A, Monty could safely open Doors B and C. So the probability that he choose B is 1/2.
- If the car is actually behind B, Monty has to open door C, so the probability that he opens door B is 0.
- If the car is actually behind C, Monty opens B with probability 1.

The sum of third coloumn is 1/2. Dividing through yields $P(A|D) = 1/3$ and $P(C|D) = 2/3$. So you better off switching.

■

1.1.3 Marginalisation

Suppose that we have a proposition B with its negative counterpart $\bar{B}$. From the sum rule, Eq. (1.1),

$$P(A, B|I) + P(A, \overline{B}|I) = P(A|I). \tag{1.13}$$

This is called *marginalisation*. It could be further generalized for an exhaustive and mutually exclusive set of discrete or continuous propositions. For example, in the case of a discrete exhaustive and mutually exclusive proposition space, B_i's,

$$P(A, B_1|I) + P(A, B_2|I) + \ldots + P(A, B_N|I) = 1, \tag{1.14}$$

or a continuous proposition space,

$$\int \mathrm{d}B P(A, B|I) = P(A|I). \tag{1.15}$$

Marginalisation is very powerful device in data analysis because it enables us to deal with *nuisance parameters*; that is, quantities which necessarily enter the analysis but are of no interest. The unwanted background signal present in many experimental measurements, and instrumental parameters which are difficult to calibre, are examples of nuisance parameters.

From marginalisation, we could calculate the evidence as

$$\mathcal{Z} = \int \mathrm{d}\Theta_M \pi(\Theta_M) \mathcal{L}(\Theta_M) \tag{1.16}$$

1.1.4 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 $\mathcal{P}(\Theta) \propto \mathcal{L}(\Theta)\pi(\Theta)$. 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 $\mathbf{\Theta}$ themselves or other quantities $f(\mathbf{\Theta})$ 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:

$$\mathcal{Z} = \int \mathcal{L}(\mathbf{\Theta})\pi(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta} \equiv \int \tilde{\mathcal{P}}(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta} \tag{1.17}$$

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

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

$$\mathcal{P}(\mathbf{\Theta}_{\text{int}}) = \int \mathcal{P}(\mathbf{\Theta}_{\text{int}}, \mathbf{\Theta}_{\text{nuis}})\,\mathrm{d}\mathbf{\Theta}_{\text{nuis}} = \int \mathcal{P}(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta}_{\text{nuis}} \tag{1.18}$$

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

$$\mathbb{E}_{\mathcal{P}}\left[f(\mathbf{\Theta})\right] \equiv \frac{\int f(\mathbf{\Theta})\mathcal{P}(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta}}{\int \mathcal{P}(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta}} = \frac{\int f(\mathbf{\Theta})\tilde{\mathcal{P}}(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta}}{\int \tilde{\mathcal{P}}(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta}} = \int f(\mathbf{\Theta})\mathcal{P}(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta} \tag{1.19}$$

since $\int \mathcal{P}(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta} = 1$ by definition and $\tilde{\mathcal{P}}(\mathbf{\Theta}) \propto \mathcal{P}(\mathbf{\Theta})$. This represents a weighted average of $f(\mathbf{\Theta})$, where at each value $\mathbf{\Theta}$ we weight the resulting $f(\mathbf{\Theta})$ 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 $\mathbb{E}_{\mathcal{P}}[f(\boldsymbol{\Theta})]$ 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 1.4.

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:

$$L_{\mathcal{P}}(\hat{\boldsymbol{\Theta}}) \equiv \mathbb{E}_{\mathcal{P}}\left[L(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta})\right] = \int L(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta})\mathcal{P}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \tag{1.20}$$

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

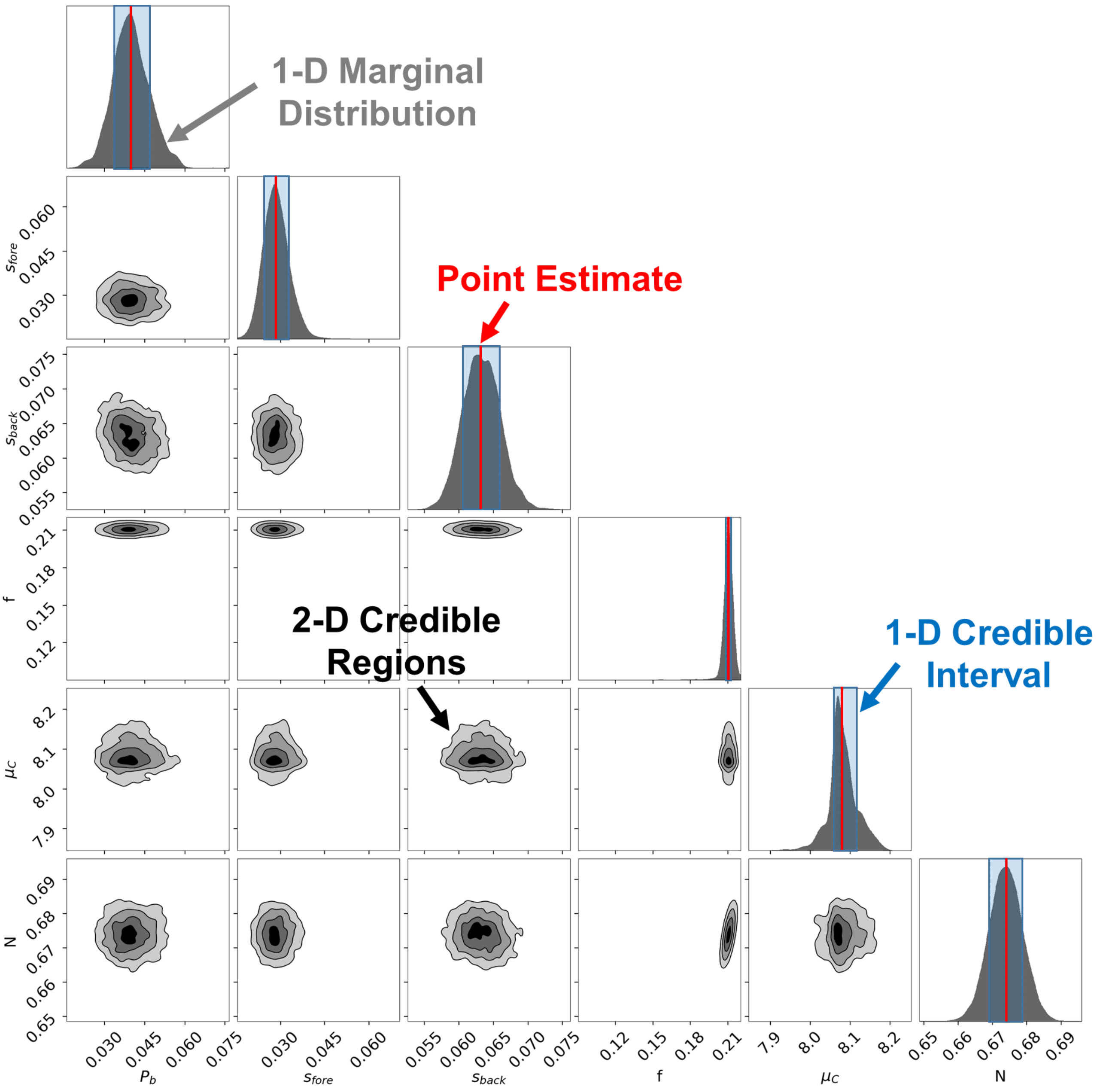


Figure 1.4: A “corner plot” showing an example of how posteriors are used in practice. Each of the top panels shows the 1-D marginalized posterior distribution for each parameter (grey), along with associated median point estimates (red) and 68% credible intervals (blue). Each central panel shows the 10%, 40%, 65%, and 85% credible regions for each 2-D marginalized posterior distribution. See §1.1.4 for additional details.

(unknown) loss:

$$\hat{\Theta} \equiv \underset{\Theta'}{\operatorname{argmin}} \left[L_{\mathcal{P}}(\Theta') \right] \tag{1.21}$$

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

While this strategy can work for any arbitrary loss function, solving for $\hat{\Theta}$ often requires using numerical methods and repeated integration over $\mathcal{P}(\Theta)$. 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 $\hat{\Theta}$ under squared loss is simply the mean.

Quantifying Uncertainty

In many cases we are not just interested in computing a prediction $\hat{\Theta}$ for Θ_*, but also constraining a region $\mathcal{C}(\Theta)$ of possible values within which Θ_* 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 Θ_*?

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

$$\int_{\Theta \in \mathcal{C}_X} \mathcal{P}(\Theta) \mathrm{d}\Theta = \frac{X}{100} \tag{1.22}$$

given

$$\mathcal{C}_X \equiv \{\Theta : \mathcal{P}(\Theta) \geq \mathcal{P}_X\} \tag{1.23}$$

In other words, we want to integrate our posterior over all Θ where the value $\mathcal{P}(\Theta) > \mathcal{P}_X$ 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

$$\int_{-\infty}^{x_p} \mathcal{P}(x) \mathrm{d}x = \frac{p}{100} \tag{1.24}$$

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 $x_{\text{high}} = x_{(1+Y)/2}$. 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.

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 $\mathbf{D}$:

$$P(\tilde{\mathbf{D}}|\mathbf{D}) \equiv \int P(\tilde{\mathbf{D}}|\boldsymbol{\Theta})P(\boldsymbol{\Theta}|\mathbf{D})\mathrm{d}\boldsymbol{\Theta} \equiv \int \tilde{\mathcal{L}}(\boldsymbol{\Theta})\mathcal{P}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} = \mathbb{E}_{\mathcal{P}}\left[\tilde{\mathcal{L}}(\boldsymbol{\Theta})\right] \tag{1.25}$$

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 $\mathcal{P}(\boldsymbol{\Theta})$.

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**:

$$\mathcal{R}^1_2 \equiv \frac{P(M_1|\mathbf{D})}{P(M_2|\mathbf{D})} = \frac{P(\mathbf{D}|M_1)P(M_1)}{P(\mathbf{D}|M_2)P(M_2)} \equiv \frac{\mathcal{Z}_1}{\mathcal{Z}_2}\frac{\pi_1}{\pi_2} \tag{1.26}$$

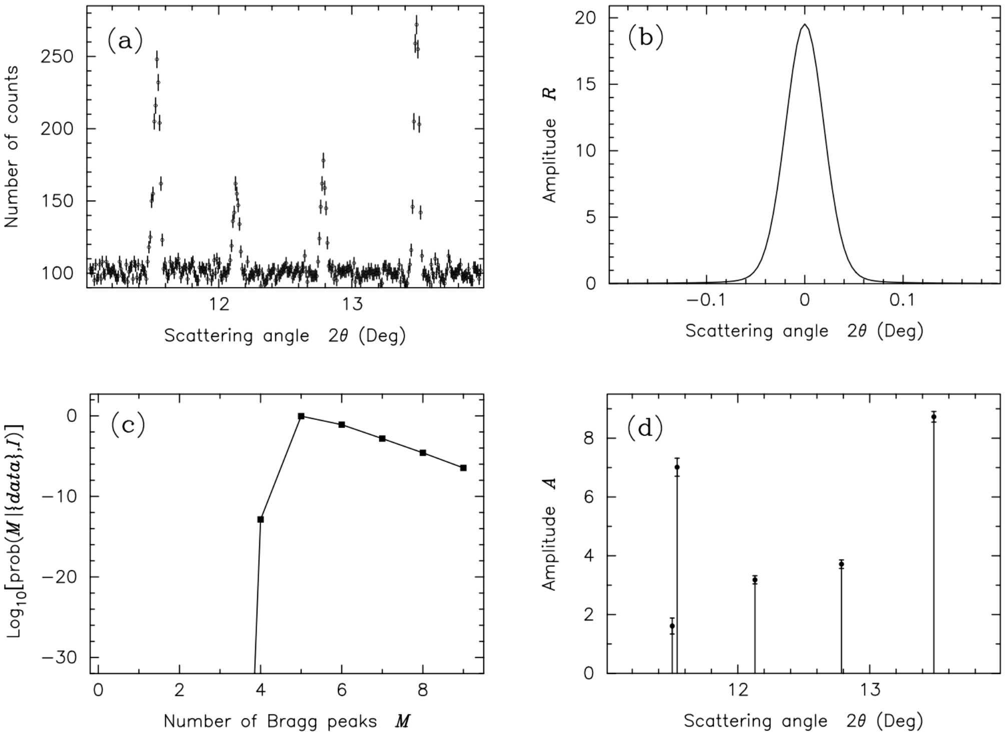


Figure 1.5: (a) Part of the X-ray diffraction data from a zeolite. (b) The calibrated profile, or resolution function, of the Bragg peaks. (c) The logarithm of the posterior pdf for the number of lines. (d) The inferred amplitudes and positions of the Bragg peaks, and their estimated error-bars.

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 $\mathbf{\Theta}_M$, and our previous relative confidence in the model.

Again, note that computing $\mathcal{Z}_M$ requires computing the integral $\int \tilde{\mathcal{P}}(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta}$ of the unnormalized posterior $\tilde{\mathcal{P}}(\mathbf{\Theta})$ over $\mathbf{\Theta}$. 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.

1.2 Maximum Likelihood Estimator

Suppose that we are fitting N data points $(x_i, y_i), i = 1, \ldots, N$, to a model that has M adjustable parameters $\theta_j, j = 1, \ldots, M$. With our notation, we shall use lower-case letters, x, for variables and model parameters while upper-case letters, X, are reserved for propositions. The model predicts a functional relationship between the measured independent and dependent variables,

$$y_{\text{th}}(x) = y_{\text{th}}(x|\theta_1\theta_2\dots\theta_M) \tag{1.27}$$

where the notation indicates the dependence on the parameters explicitly on the right-hand side, following the vertical bar.

What, exactly, do we want to minimize to get fitted values for the θ_j's? The first thing that comes to mind is the familiar least-square fit,

$$\sum_{i=1}^{N}\left[y_i - y_{\text{th}}(x_i|\theta_1\theta_2\dots\theta_M)\right]^2. \tag{1.28}$$

1.2.1 Uncorrelated data points

By minimising Eq. (1.28), we would obtain the optimal set of parameters that fits the data best. However, we have not taken the uncertainty in measurement into account. With the measurement uncertainty, we assume that each data point y_i has a measurement error that is independently random and distributed as a normal (Gaussian) distribution around the "true" model $y(x)$. And suppose that the standard deviation σ_i for the point (x_i, y_i). Hence for the data D with data points $(x_i, y_i, \sigma_i), i = 1, 2, \dots, N$ and parameters $\mathbf{\Theta}_M$ with parameters $\theta_j, j = 1, 2, \dots M$.

$$\mathcal{L}(\mathbf{D}|\mathbf{\Theta}_M) = \prod_{i=1}^{N}\left\{ \exp\left[-\frac{1}{2}\left(\frac{(y_i - y_{\text{th}}(x_i))^2}{\sigma_i}\right)\right]\Delta y\right\}. \tag{1.29}$$

Notice that there is a factor of Δy in each term in the product. As often as not, we take a constant (i.e. Δy) as *non-informative* prior. The most probable model, then, is the one that maximizes Eq. (1.29) or, equivalently, minimizes the negative of its logarithm,

$$\left[\sum_{i=1}^{N}\frac{\left[y_i - y_{\text{th}}(x_i)\right]^2}{2\sigma_i^2}\right] - N\log\Delta y. \tag{1.30}$$

Since N and Δy are all constants, minimising this equation is equivalent to minimising Eq. (1.28). We define chi-square as

$$\chi^2 \equiv \sum_{i=1}^{N} \left(\frac{y_i - y_{\text{th}}(x_i|\theta_1\theta_2 \ldots \theta_M)}{\sigma_i} \right)^2 . \tag{1.31}$$

To whatever extent the measurement errors actually are normally distributed, the quantity χ^2 is correspondingly a sum of N squares of normally distributed quantities, each normalised to unit variance. Once we have adjusted the $a_0 a_1 \ldots a_{M-1}$ to minimise the value of χ^2, the terms in the sum are not all statistically independent. Hence the likelihood function in terms of chi-squared maximum likelihood estimator is given by

$$\mathcal{L}(\mathbf{\Theta}_M) \equiv \mathcal{L}(\mathbf{\Theta}_M|M) = \mathcal{L}_0 \exp\left(-\frac{1}{2}\chi^2\right), \tag{1.32}$$

where $\mathcal{L}_0$ is a constant which shall be disregarded in calculation. We shall define the term call *log-likelihood* function;

$$L \equiv \log \mathcal{L} = \mathcal{L}_0 - \frac{1}{2}\chi^2. \tag{1.33}$$

The log-likelihood function will normally be calculated due to more numerically stable than the likelihood function.

1.2.2 Correlated data points

Everything we have done so far has assumed that the error associated with each datum is independent of the errors for the others, and that the Gaussian describing our knowledge of the error. In general, however, the errors can have different variances, and could be correlated. For correlated data points, the value of chi-square will be

$$\chi^2 = (\mathbf{y} - \mathbf{y}_{\text{th}}(\mathbf{x}))^{\mathsf{T}} \cdot \mathbf{\Sigma}_{\text{D}}^{-1} \cdot (\mathbf{y} - \mathbf{y}_{\text{th}}(\mathbf{x})), \tag{1.34}$$

where $\mathbf{y}$ is the data column vector and $\mathbf{y}_{\text{th}}(\mathbf{x})$ is the model prediction at $\mathbf{x}$;

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \mathbf{y}_{\text{th}}(\mathbf{x}) = \begin{pmatrix} y_{\text{th}}(x_1) \\ y_{\text{th}}(x_2) \\ \vdots \\ y_{\text{th}}(x_N) \end{pmatrix}. \tag{1.35}$$

Σ_{D} is called *data covariance matrix*; which is given by

$$\Sigma_{\text{D}} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \sigma_{N3} & \cdots & \sigma_{NN} \end{pmatrix}. \tag{1.36}$$

If the errors are independent, Σ_{D} will be diagonal, with entries equal to σ_i^2 and the chi-square is reduced to Eq. (1.31).

1.3 Parameter Estimation

The posterior encodes our inference about the parameter in the model, given the data and the relevant background information. Often; however, we wish to summarise this with just two numbers: the best estimate and a measure of its reliability. Since the probability (density) associated with any particular value of the parameter is a measure of how much we believe that it lies in the neighborhood of that point, our best estimate is given by the maximum of the posterior pdf.

1.3.1 One-parameter model

If we denote the quantity of interest by x, with a posterior pdf $\mathcal{P}(\Theta|\mathbf{D}, M)$, then the best estimate of its value θ_0 is given by the condition

$$\left.\frac{\mathrm{d}\mathcal{P}}{\mathrm{d}\theta}\right|_{\theta=\theta_0} = 0. \tag{1.37}$$

We should also check the sign of the second derivative to ensure that θ_0 represents a maximum rather than a minimum (or a point of reflection):

$$\left.\frac{\mathrm{d}^2\mathcal{P}}{\mathrm{d}\theta^2}\right|_{\theta_0} < 0. \tag{1.38}$$

To obtain a measure of the reliability of this best estimate, we need to look at the width or spread of the posterior pdf about θ_0. When considering the behaviour of any function in the neighbourhood of a particular point, it is often helpful to carry out a Taylor series expansion; this is simply a standard tool for locally approximating a complicated function by a low-order polynomial. Rather than dealing directly with the posterior pdf $\mathcal{P}$, which is a 'peaky' and positive function, it is better to work with logarithm L,

$$L = \log\big[\mathcal{P}(\Theta|\mathbf{D}, M)\big], \tag{1.39}$$

since this varies much more slowly with θ. Expanding L about the point $\theta = \theta_0$, we have

$$L = L(\theta_0) + \left.\frac{1}{2}\frac{\mathrm{d}^2 L}{\mathrm{d}\,\theta^2}\right|_{\theta_0} (\theta - \theta_0)^2 + \dots, \tag{1.40}$$

where the best estimate of θ is given by the condition

$$\left.\frac{\mathrm{d}L}{\mathrm{d}\,\theta}\right|_{\theta_0} = 0. \tag{1.41}$$

The first term in the Taylor series, $L(\theta_0)$, is a constant and tells us nothing about the shape of the posterior pdf. The linear term is missing because we are expanding about the maximum. The quadratic term is, therefore, the dominant factor determining the width of the posterior pdf and plays a central role in the reliability analysis. Ignoring all the higher-order contributions, the exponential yields

$$\mathcal{P}(\Theta|\mathbf{D}, M) \approx \mathcal{P}_0 \exp\left[\frac{1}{2}\frac{\mathrm{d}^2 L}{\mathrm{d}\,\theta^2}\bigg|_{\theta_0}(\theta - \theta_0)^2\right], \tag{1.42}$$

where $\mathcal{P}_0$ is a normalisation constant. This is the normal distribution,

$$P(\theta|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\theta - \mu)^2}{2\sigma^2}\right], \tag{1.43}$$

where $\mu \equiv \theta_0$ and

$$\sigma \equiv \left(-\frac{\mathrm{d}^2 L}{\mathrm{d}\,\theta^2}\bigg|_{\theta_0}\right)^{-1/2}. \tag{1.44}$$

Our inference about the quantity of interest is conveyed very precisely, therefore, by the statement

$$\theta = \theta_0 \pm \sigma. \tag{1.45}$$

1.3.2 Multi-parameter model

Similar to the previous section where we give the mean, μ, and the standard deviation, σ, as the best estimate of the parameter θ. We will turn on attention to multi-parameter model with the posterior,

$$\mathcal{P}(\boldsymbol{\Theta}|\mathbf{D}, M), \tag{1.46}$$

where $\boldsymbol{\Theta}$ is the parameter vector;

$$\boldsymbol{\Theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{pmatrix}. \tag{1.47}$$

If we denote the best estimate of $\boldsymbol{\Theta}$ by $\boldsymbol{\Theta}_0$ which gives the maximum value of the posterior. The maximum of the posterior is given by

$$\nabla_{\boldsymbol{\Theta}}\mathcal{P} = 0. \tag{1.48}$$

Or, equivalently, by solving simultaneous equations,

$$\left.\frac{\partial \mathcal{P}}{\partial \theta_i}\right|_{\mathbf{\Theta}_0} = 0, \tag{1.49}$$

where $i = 1, 2, \ldots$ up to the number of parameters to be inferred. In case of a two-parameter model, where we have parameter $x \equiv \theta_1$ and $y \equiv \theta_2$ and the posterior is maximum at $(x_0, y_0) \equiv (\theta_{1,0}, \theta_{2,0})$ given the conditions;

$$\left.\frac{\partial L}{\partial x}\right|_{x_0,y_0} = 0 \quad \text{and} \quad \left.\frac{\partial L}{\partial y}\right|_{x_0,y_0} = 0, \tag{1.50}$$

where $L = \log\left[P(\{x, y\}|\mathbf{D}, M)\right]$.

To obtain a measure of the reliability of the best estimate, we need to look at the spread of the two-dimensional posterior pdf about (x_0, y_0). By using a Taylor's expansion,

$$L = L(x_0, y_0) \quad + \quad \frac{1}{2}\left[\left.\frac{\partial^2 L}{\partial x^2}\right|_{x_0,y_0}(x - x_0)^2 + \left.\frac{\partial^2 L}{\partial y^2}\right|_{x_0,y_0}(y - y_0)^2\right. \tag{1.51}$$

$$\left. +2\left.\frac{\partial^2 L}{\partial x \partial y}\right|_{x_0,y_0}(x - x_0)(y - y_0)\right] + \ldots. \tag{1.52}$$

Similar to the one-parameter model, the posterior is approximately

$$P(\{x, y\}|\mathbf{D}, I) \approx \text{const.} \exp\left[-\frac{1}{2}Q\right], \tag{1.53}$$

where

$$Q = \begin{pmatrix} x - x_0 & y - y_0 \end{pmatrix} \cdot \begin{pmatrix} A & C \\ C & B \end{pmatrix} \cdot \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix}. \tag{1.54}$$

The component of the 2×2 symmetric matrix in the middle are given by the second derivatives of L, evaluated at the maximum (x_0, y_0):

$$A = \left.\frac{\partial^2 L}{\partial x^2}\right|_{x_0,y_0}, B = \left.\frac{\partial^2 L}{\partial y^2}\right|_{x_0,y_0}, C = \left.\frac{\partial^2 L}{\partial x \partial y}\right|_{x_0,y_0} \tag{1.55}$$

The contour of Q in x–y plane; within our quadratic approximation is shown in the Figure. 1.6.

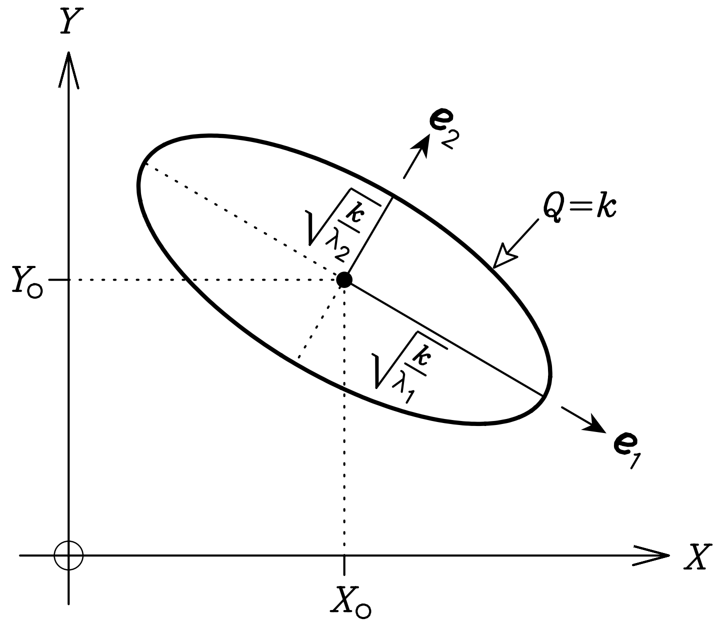


Figure 1.6: The contour in x–y parameter space along with $Q = k$, a constant centred on $(x_\mathrm{O}, y_\mathrm{O})$. The characteristics of which are determined by the eigenvalues λ and the eigenvectors $\boldsymbol{e}$ of the second-derivatives.

Chapter 2

Metropolis-Hastings Algorithm

In recent decades, there has been a significant enhancement in the quality and quantity of available data, attributed to advancements that enable faster and more cost-effective collection and storage. Concurrently, the technology facilitating the accumulation of extensive datasets has resulted in a substantial boost in computational power and resources dedicated to their analysis.

Collectively, these advancements have facilitated the investigation of increasingly intricate models through techniques capable of leveraging enhanced computational capabilities. Consequently, there has been a substantial increase in the number of published works employing **Monte Carlo** methods. These methods utilize a combination of numerical simulation and random number generation to navigate these complex models.

Among the Monte Carlo methods, **Markov Chain Monte Carlo (MCMC)** has gained significant popularity. MCMC methods are attractive due to their straightforward and intuitive approach, enabling the simulation of values from an unknown distribution. Furthermore, these simulated values can be utilized for subsequent analyses, making MCMC methods applicable across a diverse range of domains.

2.1 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 §1.1.4 therefore requires the use of numerical methods.

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:

$$\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right] = \int f(\boldsymbol{\Theta})\mathcal{P}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \approx \sum_{i=1}^{n} f(\boldsymbol{\Theta}_i)\mathcal{P}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i \tag{2.1}$$

where

$$\Delta\boldsymbol{\Theta}_i = \boldsymbol{\Theta}_{j+1} - \boldsymbol{\Theta}_j \tag{2.2}$$

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

$$\boldsymbol{\Theta}_i = \frac{\boldsymbol{\Theta}_{j+1} + \boldsymbol{\Theta}_j}{2} \tag{2.3}$$

is just defined to be the mid-point between $\boldsymbol{\Theta}_j$ and $\boldsymbol{\Theta}_{j+1}$.[1] As shown in Figure 2.1, 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*

$$\Delta\boldsymbol{\Theta}_i = \prod_{j=1}^{d} \Delta\Theta_{i,j} \tag{2.4}$$

where $\Delta\Theta_{i,j}$ is the width of the ith cuboid in the jth dimension. See Figure 2.1 for a visual representation of this procedure.

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:

$$\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right] = \frac{\int f(\boldsymbol{\Theta})\mathcal{P}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}}{\int \mathcal{P}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}} = \frac{\int f(\boldsymbol{\Theta})\tilde{\mathcal{P}}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}}{\int \tilde{\mathcal{P}}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}} \approx \frac{\sum_{i=1}^{n} f(\boldsymbol{\Theta}_i)\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i}{\sum_{i=1}^{n} \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i} \tag{2.5}$$

[1] Choosing $\boldsymbol{\Theta}_i$ to be one of the end-points gives consistent behavior (see §2.1.3) as the number of grid points $n \to \infty$ but generally leads to larger biases for finite n.

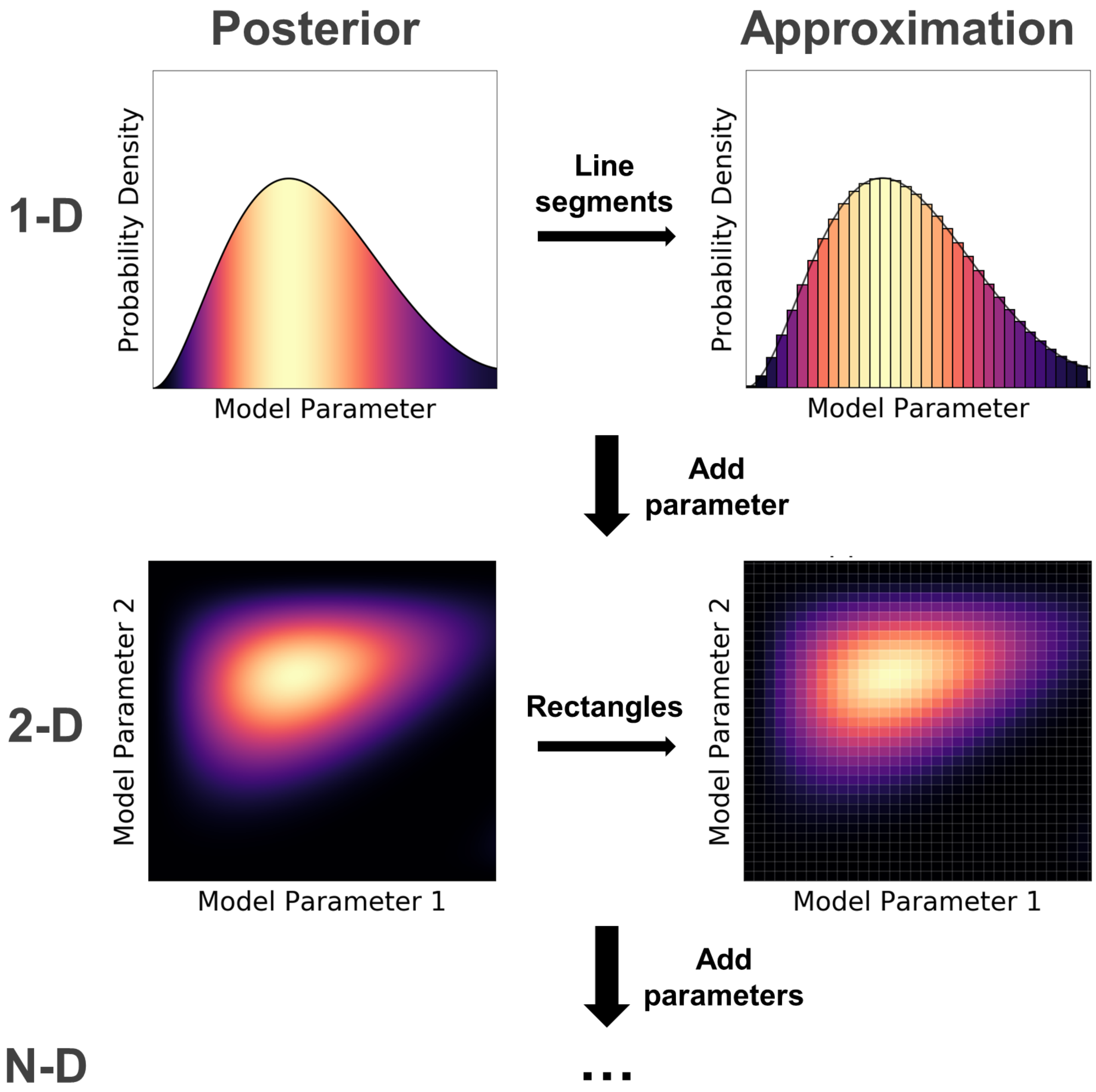


Figure 2.1: An illustration of how to approximate posterior integrals using a discrete grid of points. We break up the posterior into contiguous regions defined by a position $\boldsymbol{\Theta}_i$ (e.g., an endpoint or midpoint) with corresponding posterior density $\mathcal{P}(\boldsymbol{\Theta}_i)$ 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 §2.1 for additional details.

Note the denominator is now an estimate for the evidence:

$$\mathcal{Z} = \int \tilde{\mathcal{P}}(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \approx \sum_{i=1}^{n} \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i) \Delta\boldsymbol{\Theta}_j \tag{2.6}$$

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 $\mathcal{Z}$.

2.1.1 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

$$n \sim \prod_{j=1}^{d} k = k^d \tag{2.7}$$

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).

2.1.2 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 (2.5) 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:

$$\hat{f}_{\rm mean} \equiv \frac{\sum_{i=1}^{n} w_i f_i}{\sum_{i=1}^{n} w_i} \tag{2.8}$$

Indeed, if we define

$$f_i \equiv f(\boldsymbol{\Theta}_i), \quad w_i \equiv \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i \tag{2.9}$$

then the connection between the weighted sample mean in equation (2.8) and the expectation value from our grid in equation (2.5) becomes explicit:

$$\mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right] \approx \frac{\sum_{i=1}^{n} f(\boldsymbol{\Theta}_i)\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i}{\sum_{i=1}^{n} \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i} \equiv \frac{\sum_{i=1}^{n} w_i f_i}{\sum_{i=1}^{n} w_i} \tag{2.10}$$

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:

$$n_{\rm eff} \equiv \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2} \tag{2.11}$$

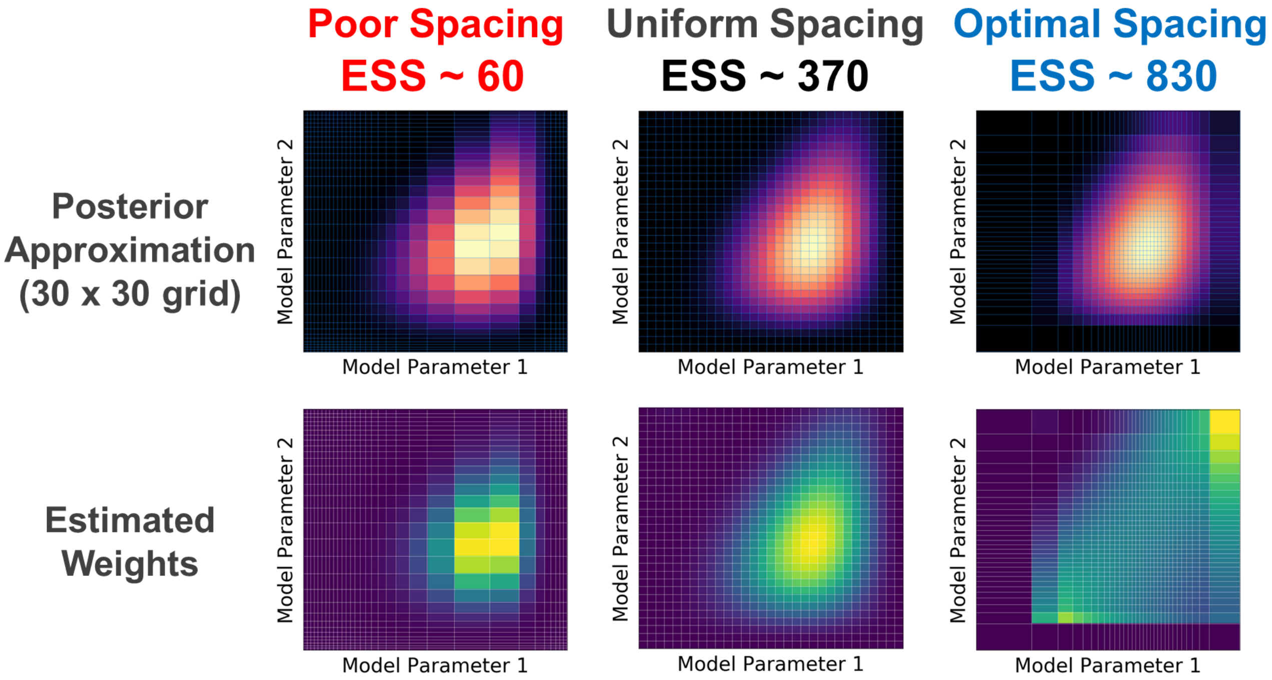


Figure 2.2: An example of how changing the spacing (volume elements) of the grid can dramatically affect its associated estimate of posterior integrals. On a toy 2-D posterior $\mathcal{P}(\mathbf{\Theta})$, simply changing the spacing of the associated 2-D 30×30 grid dramatically affects the effective sample size (ESS) (see §2.1.2). 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 §2.1.2 for additional details.

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

$$n_{\rm eff}^{\rm best} = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2} = \frac{(nw)^2}{\sum_{i=1}^{n} w^2} = \frac{n^2 w^2}{n w^2} = n \tag{2.12}$$

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):

$$n_{\rm eff}^{\rm worst} = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2} = \frac{(w)^2}{w^2} = 1 \tag{2.13}$$

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 2.2.

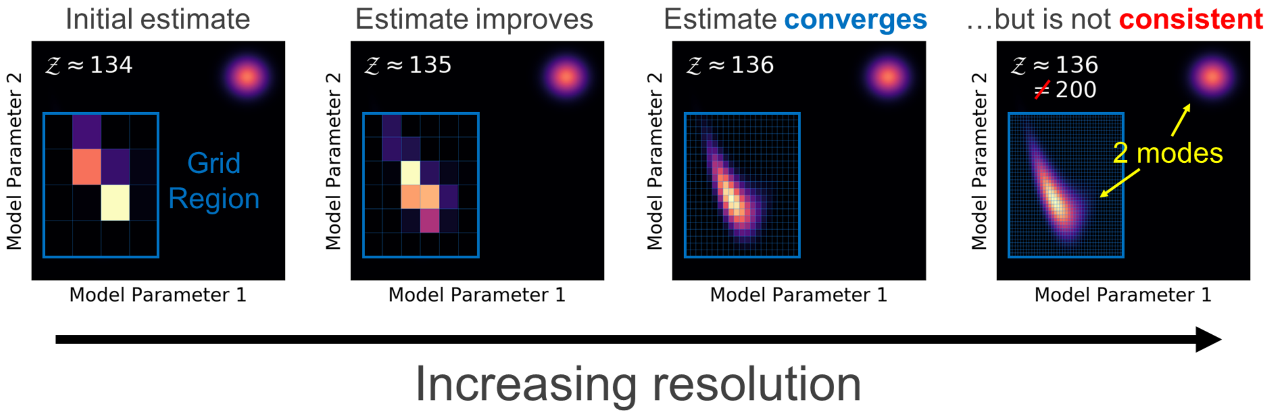


Figure 2.3: An illustration of how grid-based estimates can be *convergent* (i.e. converge to a single value as the number of grid points increases) but not *consistent* (i.e. the value it converges to is not the correct answer). Our toy 2-D unnormalized posterior $\tilde{\mathcal{P}}(\boldsymbol{\Theta})$ 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 §2.1.3 for additional details.

2.1.3 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 $n \to \infty$:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} f(\boldsymbol{\Theta}_i)\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i}{\sum_{i=1}^{n} \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i} = C \tag{2.14}$$

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

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} f(\boldsymbol{\Theta}_i)\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i}{\sum_{i=1}^{n} \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i} = \mathbb{E}_{\mathcal{P}}\left[f(\boldsymbol{\Theta})\right] \tag{2.15}$$

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 2.3. This fundamental problem is not shared by Monte Carlo methods.

2.2 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 $\{\tilde{w}_1, \dots, \tilde{w}_n\}$ associated with each sample are constant. 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 $\{\boldsymbol{\Theta}_1 \to \dots \to \boldsymbol{\Theta}_n\}$ 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

$$\rho(\boldsymbol{\Theta}) \equiv \frac{m(\boldsymbol{\Theta})}{n} \tag{2.16}$$

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

$$\int_{\boldsymbol{\Theta}\in\delta_{\boldsymbol{\Theta}}} \mathcal{P}(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \approx \int_{\boldsymbol{\Theta}\in\delta_{\boldsymbol{\Theta}}} \rho(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \approx n^{-1}\sum_{j=1}^{n} \mathbb{1}\left[\boldsymbol{\Theta}_j \in \delta_{\boldsymbol{\Theta}}\right] \tag{2.17}$$

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 $\delta_{\mathbf{\Theta}}$ and normalizing by the total number of samples n. A schematic illustration of this concept is shown in Figure 2.4.

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

$$\mathcal{Z} = \int \frac{\tilde{\mathcal{P}}(\mathbf{\Theta})}{\rho(\mathbf{\Theta})} \rho(\mathbf{\Theta}) \mathrm{d}\mathbf{\Theta} \equiv \mathbb{E}_{\rho} \left[\tilde{\mathcal{P}}(\mathbf{\Theta}) / \rho(\mathbf{\Theta}) \right] \approx n^{-1} \sum_{i=1}^{n} \frac{\tilde{\mathcal{P}}(\mathbf{\Theta}_i)}{\rho(\mathbf{\Theta}_i)} \tag{2.18}$$

This is just the average of the ratio between $\tilde{\mathcal{P}}(\mathbf{\Theta}_i)$ and $\rho(\mathbf{\Theta}_i)$ 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

$$\mathbb{E}_{\mathcal{P}} \left[f(\mathbf{\Theta}) \right] \approx \frac{n^{-1} \sum_{i=1}^{n} f_i \tilde{w}_i}{n^{-1} \sum_{i=1}^{n} \tilde{w}_i} = \frac{n^{-1} \sum_{i=1}^{n} f_i}{n^{-1} \sum_{i=1}^{n} 1} = n^{-1} \sum_{i=1}^{n} f_i \tag{2.19}$$

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

To summarize, the idea behind MCMC is to simulate a series of values $\{\mathbf{\Theta}_1 \to \cdots \to \mathbf{\Theta}_n\}$ in a way that their density $\rho(\mathbf{\Theta})$ after a given amount of time follows the underlying posterior $\mathcal{P}(\mathbf{\Theta})$. We can then estimate the posterior within any particular region $\delta_{\mathbf{\Theta}}$ 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.

2.3 Metropolis-Hastings Algorithm

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 $\mathbf{\Theta}_i \to \mathbf{\Theta}_{i+1}$ such that the distribution of the final samples

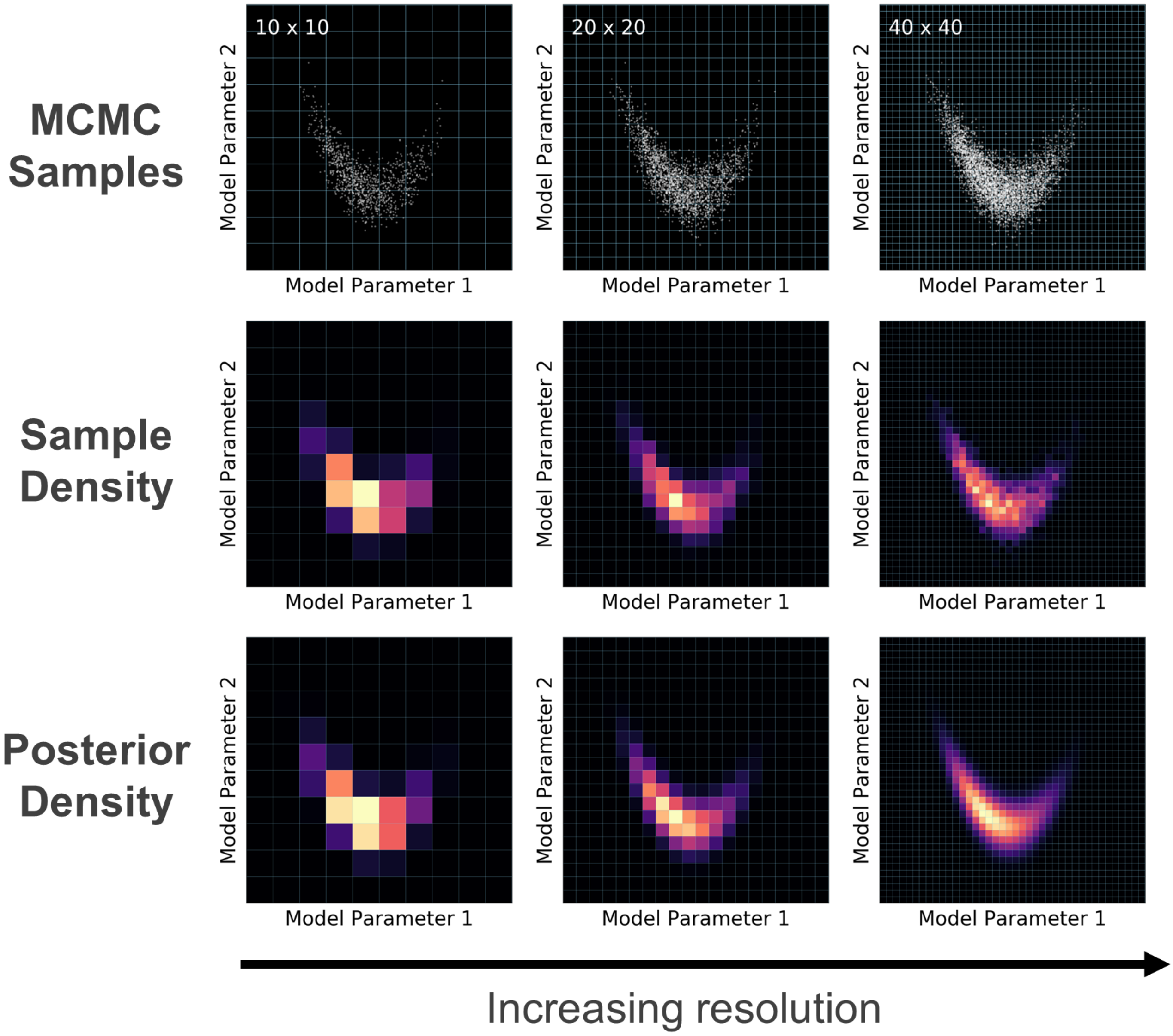


Figure 2.4: A schematic illustration of Markov Chain Monte Carlo (MCMC). MCMC tries to create a chain of n (correlated) samples $\{\boldsymbol{\Theta}_1 \to \cdots \to \boldsymbol{\Theta}_n\}$ (top) such that the number of samples m in some particular volume δ gives a relative density m/n (middle) comparable to the posterior $\mathcal{P}(\boldsymbol{\Theta})$ integrated over the same volume (bottom). See §2.2 for additional details.

$\rho(\boldsymbol{\Theta})$ 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 §2.1.3.

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:

$$P(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)P(\boldsymbol{\Theta}_i) = P(\boldsymbol{\Theta}_{i+1}, \boldsymbol{\Theta}_i) = P(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{i+1})P(\boldsymbol{\Theta}_{i+1}) \tag{2.20}$$

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

$$\frac{P(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)}{P(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{i+1})} = \frac{P(\boldsymbol{\Theta}_{i+1})}{P(\boldsymbol{\Theta}_i)} = \frac{\mathcal{P}(\boldsymbol{\Theta}_{i+1})}{\mathcal{P}(\boldsymbol{\Theta}_i)} \tag{2.21}$$

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 $\boldsymbol{\Theta}_i \to \boldsymbol{\Theta}'_{i+1}$ based on a **proposal distribution** $\mathcal{Q}(\boldsymbol{\Theta}'_{i+1}|\boldsymbol{\Theta}_i)$ similar in nature to the $\mathcal{Q}(\boldsymbol{\Theta})$. 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:

$$P(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i) \equiv \mathcal{Q}(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)T(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i) \tag{2.22}$$

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:

$$\frac{T(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)}{T(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{i+1})} = \frac{\mathcal{P}(\boldsymbol{\Theta}_{i+1})}{\mathcal{P}(\boldsymbol{\Theta}_i)} \frac{\mathcal{Q}(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{i+1})}{\mathcal{Q}(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)} \tag{2.23}$$

It is straightforward to show that the **Metropolis criterion**

$$T(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i) \equiv \min\left[1, \frac{\mathcal{P}(\boldsymbol{\Theta}_{i+1})}{\mathcal{P}(\boldsymbol{\Theta}_i)} \frac{\mathcal{Q}(\boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{i+1})}{\mathcal{Q}(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)}\right] \tag{2.24}$$

satisfies this constraint.

Generating samples following this approach can be done using the **Metropolis-Hastings (MH) Algorithm**:

1. *Propose* a new position $\boldsymbol{\Theta}_i \to \boldsymbol{\Theta}'_{i+1}$ 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 $[0, 1]$.
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 2.5 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

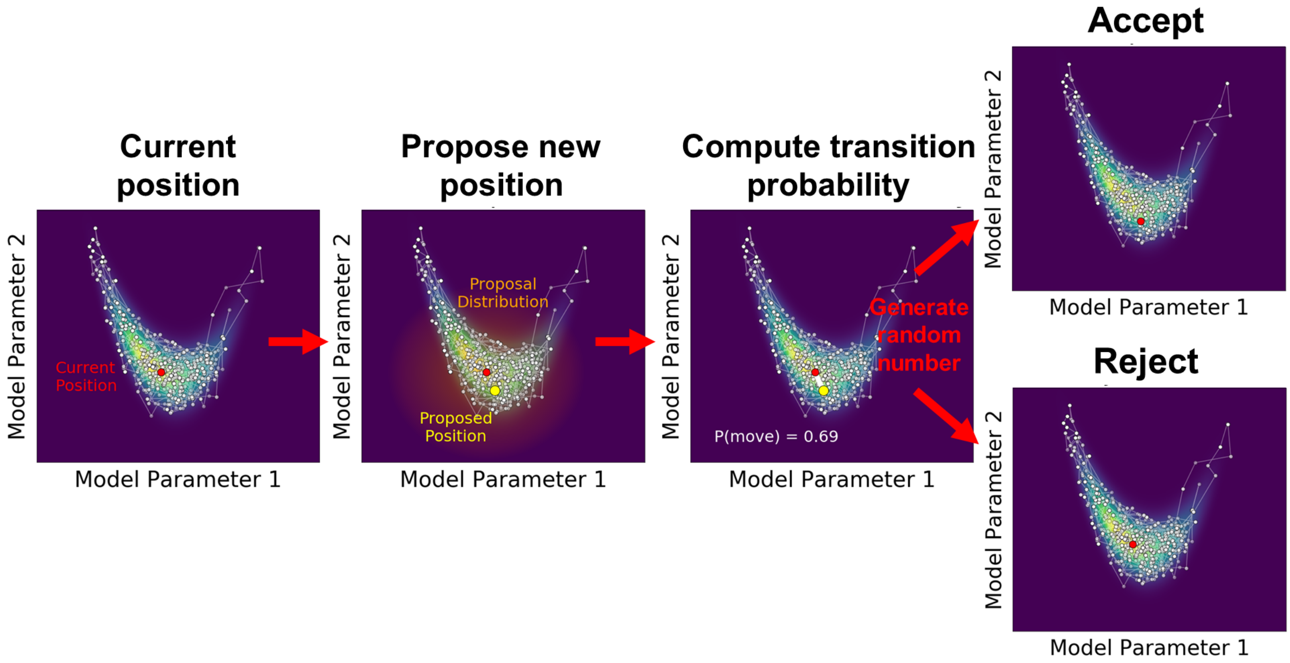


Figure 2.5: A schematic illustration of the Metropolis-Hastings algorithm. At a given iteration i, we have generated a chain of samples $\{\boldsymbol{\Theta}_1 \to \cdots \to \boldsymbol{\Theta}_i\}$ (white) up to the current position $\boldsymbol{\Theta}_i$ (red) whose behavior follows the underlying 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 §2.3 for additional details.

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 Θ_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.

2.4 Monitoring Convergence - Gelman-Rubin Method

Operationally, effective convergence of Markov chain simulation has been reached when inferences for quantities of interest do not depend on the starting point of the simulations. This suggests monitoring convergence by comparing inferences made from several independently sampled sequences with different starting points. Before considering methods of comparing inferences, we briefly discuss the standard method for constructing inferences under the assumption that convergence has been approximately reached. It is standard practice to discard observations within an initial transient phase. Most methods for inference are then based on the assumption that what remains can be treated as if the starting points had been drawn from the target distribution.

We present the method of Gelman and Rubin using our general perspective of comparison of inferences. The method presupposes that m chains have been simulated in parallel, each with different starting points which are overdispersed with respect to the target distribution. A number of methods have been proposed for generating initial values for MCMC samplers. Gelman and Rubin proposed using a simple mode-finding algorithm to locate regions of high density and sampling from a mixture of t-distributions located at these modes to generate suitable starting values.

Given any individual sequence, and if approximate convergence has been reached, an assumption is made that inferences about any quantity of interest is made by computing the sample mean and variance from the simulated draws. Thus, the m chains yield m possible inferences; to answer the question of whether these inferences are similar enough to indicate approximate convergence, Gelman and Rubin suggested comparing these to the inference made by mixing together the mn draws from all the sequences. Consider a scalar summary—that is, a random variable—θ , that has mean μ and variance σ^2 under the target distribution, and suppose that we have some unbiased estimator $\hat{\mu}$ for μ. Letting θ_{jt} denote the tth of the n iterations of θ in chain j, we take $\hat{\mu} = \bar{\theta}_{..}$, and calculate the between-sequence variance B/n, and the within-sequence variance W, defined by

$$\begin{aligned} B/n &= \frac{1}{m-1}\sum_{j=1}^{m}\left(\bar{\theta}_{j.}-\bar{\theta}_{..}\right)^2 \qquad (2.25) \\ W &= \frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{t=1}^{n}\left(\bar{\theta}_{jt}-\bar{\theta}_{j.}\right)^2 \qquad (2.26) \end{aligned}$$

Having observed these estimates, we can estimate σ^2 by a weighted average of B and W,

$$\hat{\sigma}^2 = \frac{n-1}{n}W + \frac{B}{n}, \qquad (2.27)$$

which would be an unbiased estimate of the true variance σ^2 if the starting points of the sequences were drawn from the target distribution, but overestimates σ^2 if the starting distribution is appropriately overdispersed. The comparison of pooled and within-chain inferences is expressed as a variance ratio,

$$R = \frac{\hat{\sigma}^2}{\sigma^2} \qquad (2.28)$$

which is called the scale reduction factor (Strictly speaking, the term “scale reduction factor” applies to $\sqrt{R}$). Because the denominator of R is not itself known, it must be estimated from the data

$$\hat{R} = \frac{m+1}{m}\frac{\hat{\sigma}^2}{W} - \frac{n-1}{mn}, \qquad (2.29)$$

which is called the potential scale reduction factor, or PSRF, and can be interpreted as a convergence diagnostic as follows. If $\hat{R}$ is large, this suggests that the estimate of the variance $\hat{\sigma}^2$ can be further decreased by more simulations. Alternatively, if the PSRF is close to 1, we can conclude that each of the m sets of n simulated observations is close to the target distribution.