Bayesian Statistics

Lectures for the MRes in Big Data and Machine Learning in Physics

Boris Leistedt
Based on course notes by Alan Heavens, Imperial College London

November 13, 2025

Contents

List of Examples

1 Books and Syllabus

1.1 Recommended Books

- D. Silvia & J. Skilling: Data Analysis: a Bayesian Tutorial (CUP). *Nice small book for the basics.*
- P. Saha: Principles of Data Analysis. (Capella Archive)
  `https://www.physik.uzh.ch/ psaha/pda/`
  Similarly, a good, clear, small volume. Free online version as well as a physical book.
- T. Loredo: Bayesian Inference in the Physical Sciences
  `http://www.astro.cornell.edu/staff/loredo/bayes/`
- D. Mackay: Information Theory, Inference and Learning Algorithms. (CUP)
  `http://www.inference.phy.cam`
  More on the information theory basis of the subject.
- A. Gelman et al: Bayesian Data Analysis (CRC Press) *Comprehensive.*

1.2 Course Syllabus

The course begins with an **Introduction** covering the fundamental differences between Bayesian and frequentist interpretations of probability. We discuss how scientific questions are often inverse problems, where we wish to infer properties of a system from observed data rather than predict data from known parameters. The foundation of Bayesian inference, Bayes theorem, is introduced as the key tool for updating our beliefs in light of new evidence.

In the section on **Parameter Inference**, we develop the machinery for inferring model parameters from data. The likelihood function, which quantifies how probable our observed data are given specific parameter values, is central to this discussion. We examine the role of priors in encoding our initial state of knowledge, including the treatment of location parameters (parameters that shift distributions) and scale parameters (parameters that affect the spread of distributions). The concept of 'noninformative' priors is discussed critically, acknowledging both their utility and limitations. We also introduce conjugate priors, which are prior distributions that combine mathematically conveniently with certain likelihood functions to produce posteriors in the same distributional family.

The **Posterior** section focuses on extracting information from the posterior distribution. Marginalisation allows us to focus on parameters of interest by integrating over nuisance parameters—those parameters necessary for the model but not of direct scientific interest. We distinguish between conditional and marginal errors, emphasizing that conditional errors (obtained by fixing other parameters) typically underestimate the true uncertainty. The profile likelihood method, which maximizes the likelihood over nuisance parameters for each value of the parameter of interest, is discussed as an alternative approach with connections to frequentist methods.

Sampling methods are introduced as practical tools for working with complex posterior distributions that cannot be evaluated analytically. We explain how samples from a distribution can be used to represent that distribution numerically, allowing us to compute marginal distributions, moments, and credible intervals. The theoretical foundations include detailed balance, which ensures that a sampling algorithm targets the correct distribution, and Markov processes, which describe how samples are generated sequentially with each new sample depending only on the previous state.

The **MCMC** (Markov Chain Monte Carlo) section presents the Metropolis-Hastings algorithm for low-dimensional problems. This algorithm generates samples by proposing moves in parameter space and accepting or rejecting them according to a criterion that ensures detailed balance. The choice of proposal distribution significantly affects the efficiency of the sampler. We examine sample correlations, quantified by the autocorrelation function, which arise because successive samples in a Markov chain are not independent. The effective sample size accounts for these correlations to give a measure of the equivalent number

of independent samples. The burn-in period, during which the chain moves from its arbitrary initial position to the high-probability regions of the posterior, must be discarded. Convergence tests, particularly the Gelman-Rubin diagnostic, help assess whether chains have adequately explored the target distribution.

For **Higher-Dimensional Problems**, standard MCMC methods become inefficient due to the difficulty of proposing moves that are likely to be accepted in high dimensions. Hamiltonian Monte Carlo (HMC) addresses this by using gradient information to propose distant moves that follow the geometry of the posterior distribution. Gibbs sampling offers an alternative strategy: when conditional distributions for individual parameters (or blocks of parameters) are available, we can sample from them in turn, effectively decomposing a high-dimensional problem into a sequence of lower-dimensional ones.

Multi-level Models, or Bayesian Hierarchical Models, extend the basic framework to situations where the data have natural grouping or structure. In these models, latent parameters (also called hyperparameters) govern the distribution of the parameters at lower levels of the hierarchy. This approach allows information to be shared across groups, leading to improved inference compared to analyzing each group independently or pooling all data without accounting for group structure.

Model Comparison addresses the question of which model best explains the data. The Bayesian evidence (or marginal likelihood) provides a natural tool for model comparison, with the ratio of evidences (Bayes factor) quantifying the relative support for competing models. For nested models, the Savage-Dickey density ratio provides a convenient shortcut for computing Bayes factors. We also discuss information criteria such as AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and DIC (Deviance Information Criterion), which approximate model comparison while accounting for model complexity.

Likelihood-Free Inference (also called Simulation-Based Inference or implicit likelihood methods) tackles problems where the likelihood function cannot be evaluated, but we can simulate data from the forward model. Rejection sampling forms the basis of Approximate Bayesian Computation (ABC), where we accept parameters if they generate data sufficiently similar to our observations. Kernel density estimation (KDE) provides a non-parametric method for estimating probability densities from samples, useful for constructing approximate posteriors.

The topic of **Extreme Data Compression** introduces the MOPED (Multiple Optimized Parameter Estimation and Data compression) algorithm, which shows that high-dimensional data can often be compressed to a number of summary statistics equal to the number of parameters being inferred, without loss of information. This dramatic compression can greatly speed up inference for problems with large datasets.

Finally, we address **Complications** that arise in real data analysis. Selection effects occur when our ability to observe data depends on the values of the data themselves. Truncation occurs when data outside some range are completely unobserved, while censoring occurs when we know that data exist beyond some threshold but do not know their exact values. Both effects must be properly modeled in the likelihood to avoid biased inferences.

1.3 Learning Objectives

By the end of this course, students will be able to apply Bayesian reasoning to scientific problems, construct likelihood functions and priors, derive posterior distributions, implement MCMC sampling, build hierarchical models, perform model comparison, and handle complex data scenarios including missing data and selection effects.

1.4 Mathematical Notation

This course uses standard mathematical notation for probability and statistics. Familiarity with the following conventions will aid understanding:

Random variables and parameters:

- x, y, θ, μ, σ denote scalar quantities (single numbers)
- $\mathbf{x}$, $\mathbf{d}$, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ denote vectors (sets of numbers) using bold font
- $\mathbf{X}$, $\mathbf{C}$, $\boldsymbol{\Sigma}$ denote matrices using bold capital letters
- $\mathbf{I}$ denotes the identity matrix

Probability notation:

- $p(x)$ denotes the probability mass function (for discrete x) or probability density function (for continuous x)
- $p(x|y)$ denotes conditional probability: the probability of x given that y has occurred or is known
- $p(x, y)$ denotes joint probability: the probability that both x and y occur
- $p(\mathbf{d}|\boldsymbol{\theta}, M)$ denotes the probability of data $\mathbf{d}$ given parameters $\boldsymbol{\theta}$ and model M
- $\mathcal{L}(\boldsymbol{\theta})$ or $\mathcal{L}(\mathbf{d}|\boldsymbol{\theta})$ denotes the likelihood function (data treated as fixed, parameters as variables)
- $\pi(\boldsymbol{\theta})$ or $p(\boldsymbol{\theta}|M)$ denotes the prior distribution on parameters
- $\sim$ means "is distributed as," e.g., $x \sim \mathcal{N}(0, 1)$ means "x is distributed as a standard normal"

Statistical operations:

- $\mathbb{E}[x]$ or $\langle x \rangle$ denotes the expectation (mean) of x
- $\mathrm{Var}[x]$ or σ^2 denotes the variance of x
- $\mathrm{Cov}[x, y]$ denotes the covariance between x and y
- $\int f(x)dx$ denotes integration over all possible values of x
- $\sum_i f(x_i)$ denotes summation over discrete index i
- $\propto$ means "proportional to" (equality up to a normalization constant)

Special symbols:

- $\delta(x)$ denotes the Dirac delta function (infinite at $x = 0$, zero elsewhere, integrates to 1)
- $\delta^D(x - a)$ denotes the Dirac delta function centered at $x = a$
- $\ln x$ or $\log x$ denotes the natural logarithm (base e)
- $\log_{10} x$ denotes the base-10 logarithm (when explicitly needed)
- ∇f denotes the gradient (vector of partial derivatives) of f
- $\partial f/\partial x$ denotes the partial derivative of f with respect to x
- $\arg\max_x f(x)$ denotes the value of x that maximizes $f(x)$

Abbreviations:

- MLE: Maximum Likelihood Estimate
- MAP: Maximum A Posteriori estimate
- MCMC: Markov Chain Monte Carlo
- PDF: Probability Density Function
- CDF: Cumulative Distribution Function
- i.i.d.: independent and identically distributed

Gaussian (Normal) Distribution Notation:

Throughout this course, we use the semicolon notation $\mathcal{N}(x; \mu, \sigma^2)$ to denote the Gaussian probability density function, where x is the random variable, μ is the mean, and σ^2 is the variance. This notation emphasizes that x is the argument of the density function, while μ and σ^2 are parameters.

1D Gaussian:

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

Properties:

- Mean: $\mathbb{E}[x] = \mu$
- Variance: $\text{Var}[x] = \sigma^2$
- Standard deviation: $\sigma = \sqrt{\sigma^2}$
- Precision (inverse variance): $\lambda = 1/\sigma^2$
- The Gaussian is symmetric about μ and integrates to 1: $\int_{-\infty}^{\infty} \mathcal{N}(x; \mu, \sigma^2) dx = 1$
- Linear transformation: If $x \sim \mathcal{N}(\mu, \sigma^2)$, then $ax + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
- Sum of independent Gaussians: If $x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $x_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then $x_1 + x_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Multivariate Gaussian:

For a d-dimensional random vector $\mathbf{x} = (x_1, \ldots, x_d)^T$:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{2}$$

where $\boldsymbol{\mu}$ is the d-dimensional mean vector, $\boldsymbol{\Sigma}$ is the $d \times d$ covariance matrix, and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

Properties:

- Mean vector: $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$
- Covariance matrix: $\text{Cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T] = \boldsymbol{\Sigma}$
- Diagonal elements: $\Sigma_{ii} = \text{Var}[x_i]$
- Off-diagonal elements: $\Sigma_{ij} = \text{Cov}[x_i, x_j]$ for $i \neq j$
- Linear transformation: If $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{Ax} + \mathbf{b} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$
- Marginalization: If $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ is partitioned with $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

  $$\boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix} \tag{3}$$

  then the marginal distribution is $\mathbf{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$
- Conditioning: For the same partitioned Gaussian, the conditional distribution is

  $$\mathbf{x}_1|\mathbf{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) \tag{4}$$

- Product of Gaussians: The product of two Gaussian densities (in the same variable) is proportional to a Gaussian:

  $$\mathcal{N}(x; \mu_1, \sigma_1^2) \cdot \mathcal{N}(x; \mu_2, \sigma_2^2) \propto \mathcal{N}(x; \mu_3, \sigma_3^2) \tag{5}$$

  where $\sigma_3^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1}$ and $\mu_3 = \sigma_3^2\left(\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}\right)$

These Gaussian properties are fundamental to Bayesian inference, particularly for conjugate priors, Kalman filtering, and Gaussian processes.

A more comprehensive notation reference is provided in Appendix A.

1.5 Graphical Models for Probabilistic Reasoning

Note: This material is non-examinable, but provides a very useful tool for visualizing probabilistic models and will be used throughout the course.

Directed Acyclic Graphs (DAGs) provide a visual language for representing probabilistic models. In these diagrams, nodes represent variables and directed edges (arrows) represent probabilistic dependencies. The graph is acyclic, meaning you cannot follow arrows in a loop back to where you started.

The notation conventions for DAGs are:

- **Circles (or unfilled nodes)** represent *unknown/latent variables* or *parameters* that we wish to infer. These are unobserved quantities with uncertainty.
- **Shaded circles (or filled nodes)** represent *observed variables* or *data*. These are quantities we have measured or know.
- **Points or small dots** (or sometimes double circles) represent *fixed/deterministic quantities* or *hyperparameters* whose values are set and not inferred.
- **Plates** (rectangular boxes around groups of nodes) represent *repetition* or *replication*. A plate labeled with $i = 1, \ldots, N$ (or simply N) indicates that the contents are repeated N times, such as N independent observations.
- **Arrows** indicate *direct probabilistic dependence*: an arrow from A to B means that B depends on A, i.e., B is sampled from a distribution conditioned on A.

The graph structure encodes the factorization of the joint probability distribution: each variable is conditionally independent of all non-descendants given its parents (the nodes with arrows pointing to it).

These diagrams will help us visualize model structure, understand the flow of information from data to parameters, and communicate assumptions clearly. Throughout the course, we will use DAGs to represent hierarchical models, parameter inference problems, and complex dependencies between variables. Examples of DAGs that appear in this course include:

- **Virus testing** (section 2, Figure 1): simple diagnostic test with latent disease status
- **Coin toss** (section 4, Figure 4): parameter inference with Beta-Binomial conjugacy
- **Supernova distance** (section 5, Figure 6): nuisance parameter marginalization
- **Line fitting with errors** (section 6, Figure 8): Gibbs sampling with latent variables
- **Hierarchical models** (section 9, Figures 11, 10): multi-level structure with hyperparameters
- **Model comparison** (section 10, Figure 12): nested model structures
- **Censored data** (section 12, Figure 14): selection effects and missing data
- **Variational inference** (section 13, Figure 15): Bayesian linear regression

2 Bayesian Inference

Bayesian and frequentist views of probability differ in fundamental ways. The frequentist view is usually expressed in terms of relative occurrences of events in multiply-repeated experiments, such as the fraction of heads thrown in the repeated toss of a coin. In Bayesian statistics, probability is sometimes interpreted as a state of knowledge. It is better matched to answering scientific questions, since this notion of probability encapsulates what is often desired when doing science. To see this, one must first recognize that most scientific questions are inverse problems - what can be learned about the world from the data that has been collected?

2.1 Inverse Problems

Data analysis problems in physics are typically inverse problems: given some observed data, the goal is to infer something about the underlying physical process that generated those data. This stands in contrast to forward modeling (also called generative modeling), where observational outcomes are predicted given a known physical process and its parameters. Inverse problems are generally much harder than forward problems because they require working backwards from effects to causes, and this mapping is often not unique.

There are two principal classes of inverse problems in Bayesian statistics. First, **parameter inference** addresses questions such as: given a model with unknown parameters, what do the data tell us about the values of those parameters? Second, **model comparison** asks: given multiple competing models that could explain the data, which model is best supported by the observations? Both types of questions are naturally framed in the Bayesian framework.

2.2 What Questions Can Bayesian Inference Answer?

Bayesian inference provides a systematic framework for answering scientific questions in the presence of uncertainty. Consider some concrete examples of the types of questions that might be asked.

For **parameter inference**, a model structure is assumed and the goal is to determine the values of its parameters given observed data. For instance: if there is a set of (x, y) pairs with measurement errors and a linear relationship $y = mx+c$ is assumed, what are the slope m and intercept c, and what are the uncertainties in these parameters? Or, if 5 X-ray photons have been detected from a source at a known distance in the laboratory, what is the power output of the source and its uncertainty? As a more complex example, given LIGO gravitational wave observations, what are the masses and other properties of the inspiralling compact objects that generated the signal? In each case, not just point estimates but full probability distributions over the parameters are desired, properly accounting for all sources of uncertainty.

For **model comparison**, the goal is to determine which of several competing models is best supported by the data. Do the observed planetary motions support General Relativity or Newtonian gravity? Is the standard cosmological model (ΛCDM) more probable than specified alternative models given the current observational data? Do Large Hadron Collider (LHC) data support the existence of the Higgs boson, or do they favor a model without it? These questions cannot be answered by simply fitting parameters—a principled way to compare models of different complexity and structure is needed. Bayesian model comparison provides this through the calculation of model evidences and Bayes factors.

2.3 Probability: Frequency vs. Degree of Belief

The interpretation of probability is central to understanding Bayesian statistics. In the frequentist view, probability describes the relative frequency of outcomes in infinitely long sequences of repeated trials—for example, the limiting fraction of heads in an infinite sequence of coin tosses. This interpretation works well for repeatable random experiments but becomes problematic when making statements about unique events or fixed parameters.

The Bayesian view, by contrast, interprets probability as expressing a degree of belief or state of knowledge about a proposition. In this framework, probabilities can be assigned to any logical proposition—a statement that could be true or false. The conditional probability $p(A|B)$ represents the degree to which the truth of

logical proposition B implies that proposition A is also true. This interpretation allows discussion of the probability that a scientific hypothesis is correct given the data that has been observed.

The Bayesian interpretation naturally expresses what is often desired in science. For example, one might ask: given the Planck satellite's observations of the cosmic microwave background (CMB), what is the probability that the density parameter of cold dark matter lies between 0.3 and 0.4? This question makes sense in the Bayesian framework, where probabilities can be assigned to parameter values. In the frequentist framework, by contrast, the density parameter is either in this range or it is not—it is a fixed (though unknown) value, not a random variable, and so cannot be assigned a probability.

2.4 Fundamental Probability Rules

Bayesian inference rests on a small number of fundamental probability rules, which are used throughout this course.

The **sum rule** states that probabilities of mutually exclusive, exhaustive outcomes must sum to unity: $p(x) + p(\sim x) = 1$, where $\sim x$ means "not x".

The **product rule** relates joint and conditional probabilities: $p(x, y) = p(x|y)p(y)$, where $p(x|y)$ is the conditional probability of x given y (read as "the probability of x given y"), and $p(x, y)$ is the joint probability that both x and y occur.

The **marginalisation rule** allows us to obtain the probability of one variable by summing or integrating over all possible values of other variables. For discrete variables: $p(x) = \sum_k p(x, y_k)$, summing over all possible discrete values y_k. For continuous variables: $p(x) = \int p(x, y)dy$. Here $p(x, y)$ is a probability density function (pdf), where $p(x, y) \geq 0$ and $p(x, y)dxdy$ represents the probability that x and y occur in an infinitesimal interval $dxdy$ around the values x, y. Note that a probability density can be greater than 1—it is a density, not a probability itself.

Since the joint probability is symmetric, $p(x, y) = p(y, x)$, we can write $p(x|y)p(y) = p(y|x)p(x)$. Rearranging this fundamental equality gives us **Bayes' theorem**:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}. \tag{6}$$

This simple equation is the foundation of all Bayesian inference. It tells us how to update our beliefs about y (the prior $p(y)$) in light of observing x, using the likelihood $p(x|y)$ and the marginal probability $p(x)$ (which serves as a normalization constant).

2.5 The Asymmetry of Conditional Probability

A fundamental error in reasoning with probabilities is to confuse $p(x|y)$ with $p(y|x)$. These conditional probabilities are generally very different, and confusing them leads to serious mistakes in inference.

The rain and umbrella example perfectly illustrates why conditional probabilities aren't symmetric. While about 75% of people carry umbrellas when it's raining ($P(\text{Umbrella}|\text{Raining}) \approx 0.75$), only about 30% of people carrying umbrellas are doing so because it's currently raining ($P(\text{Raining}|\text{Umbrella}) \approx 0.30$). This dramatic difference occurs because umbrellas serve multiple purposes—people carry them for potential rain later, sun protection, or out of habit—not just for current rainfall. In contrast, when it is actually raining, the response is more predictable: most people grab an umbrella. This asymmetry shows why seeing rain is a strong predictor of umbrellas, but seeing an umbrella is only a weak predictor of rain. You can't look out your window, spot someone with an umbrella, and confidently conclude it's raining—but if you see rain, you can reasonably assume most people outside have umbrellas.

One might think that no one would make such an elementary mistake, but confusion between $p(x|y)$ and $p(y|x)$ is surprisingly common, with serious consequences in fields from medical diagnosis to legal reasoning. The next section provides a medical testing example that illustrates the practical importance of getting this distinction right.

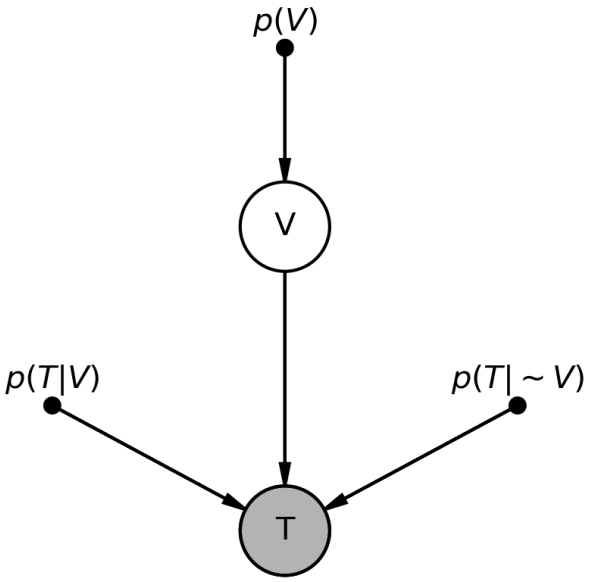

Figure 1: DAG for virus diagnostic testing model. The latent virus status V (prevalence 1%) determines test results T with sensitivity 80% and false positive rate 10%.

Example: Virus Testing: Conditional Probability and Base Rate Fallacy

Consider a medical screening scenario that illustrates the crucial distinction between $p(T|V)$ and $p(V|T)$. A diagnostic test for a virus gives a positive result (T) in infected patients (where V denotes "has the virus") with probability 0.8—this is the test's *sensitivity* or true positive rate. Suppose the prevalence of the virus in the population is 1%, so $p(V) = 0.01$.
Now imagine you take the test and receive a positive result. What is the probability that you actually have the virus? Many people intuitively think the answer is 80% (the sensitivity of the test), but this confuses $p(T|V)$ with $p(V|T)$. We want $p(V|T)$—the probability of having the virus given a positive test—which we can calculate using Bayes' theorem:

$$p(V|T) = \frac{p(T|V)p(V)}{p(T)} = \frac{p(T|V)p(V)}{p(T|V)p(V) + p(T| \sim V)p(\sim V)}. \tag{7}$$

In the denominator, we have used marginalisation to write $p(T) = p(T, V) + p(T, \sim V) = p(T|V)p(V) + p(T| \sim V)p(\sim V)$ (applying the product rule to both terms).
Substituting the numerical values:

$$p(V|T) = \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.1 \times 0.99} = \frac{0.008}{0.008 + 0.099} = \frac{0.008}{0.107} \approx 0.075. \tag{8}$$

Thus, even with a positive test result, there is only a 7.5% probability that you have the virus—and still a 92.5% chance that you do not! This counterintuitive result arises because the disease is rare: false positives among the large uninfected population ($0.1 \times 0.99 = 0.099$) outnumber true positives from the small infected population ($0.8 \times 0.01 = 0.008$) by more than 10 to 1. This example shows why the base rate (prior probability) matters enormously in diagnostic testing, and why medical professionals must be trained to reason correctly with conditional probabilities.

2.6 Bayes' Theorem for Parameter Inference

In parameter inference, we work with three key quantities: the observed data $\mathbf{d}$, a model M that describes how those data arise, and the model parameters $\boldsymbol{\theta}$ whose values we wish to determine. The first and most important rule when approaching any inference problem is to write down precisely what we want to know.

In most cases, what we seek is the probability distribution for the parameters given the data and assuming the model, denoted $p(\boldsymbol{\theta}|\mathbf{d}, M)$. This distribution, called the **posterior**, encapsulates everything the data tell us about the parameter values within the framework of the chosen model. The posterior is computed using Bayes' theorem:

$$p(\boldsymbol{\theta}|\mathbf{d}, M) = \frac{p(\mathbf{d}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)}{p(\mathbf{d}|M)}. \tag{9}$$

The three terms on the right-hand side have specific names and interpretations. The term $p(\mathbf{d}|\boldsymbol{\theta}, M)$ is the **likelihood**, commonly denoted $\mathcal{L}(\mathbf{d}|\boldsymbol{\theta})$, which quantifies how probable the observed data are if the parameters take specific values $\boldsymbol{\theta}$. The term $p(\boldsymbol{\theta}|M)$ is the **prior**, denoted $\pi(\boldsymbol{\theta})$, which encodes our state of knowledge about the parameters before observing the data. Finally, $p(\mathbf{d}|M)$ is the **Bayesian evidence** (or marginal likelihood), which plays a crucial role in model comparison but serves only to normalize the posterior in parameter inference, ensuring that it integrates to unity over all possible parameter values.

For clarity, when focusing solely on parameter inference with a fixed model, we often suppress the explicit dependence on M and write Bayes' theorem in the simplified form:

$$p(\boldsymbol{\theta}|\mathbf{d}) = \frac{\mathcal{L}(\mathbf{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\mathbf{d})}. \tag{10}$$

This formula will be used throughout the remainder of this section, with the understanding that all probabilities are implicitly conditioned on the chosen model. We will restore the model dependence explicitly when we turn to the problem of model comparison.

In the context of parameter inference (i.e. for a given fixed model M), the Evidence serves only to make the posterior a properly normalised probability distribution as a function of the parameters $\boldsymbol{\theta}$. For continuous parameters (re-introducing M),

$$p(\mathbf{d}|M) = \int p(\mathbf{d}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{11}$$

where the integral may be multidimensional (multiple parameters).

2.7 Practical Problem Analysis

Setup is the most important and useful step when tackling almost any problem. It is about comprehensively answering the following questions:

- What are the data, $\mathbf{d}$?
- What is the model for the data? In other words, what is the likelihood function $\mathcal{L}(\boldsymbol{\theta})$?
- What is the model considered? How is it structured? What are the parameters of interest? What are the relationships between the parameters (known and unknown)?
- What is the prior on the parameters $\pi(\boldsymbol{\theta})$?
- Are there any other pieces of information available? It can be useful to write down all of the probability distributions and relationships known or accessible. Typically this involves writing down a list of conditional probability distributions.

3 The Likelihood

3.1 The Likelihood Function and the Sampling Distribution

It is important to pause here to think about $\mathcal{L}$. We can view this distribution two ways. If we fix $\boldsymbol{\theta}$ (as is rather implied by the expression), then we have the distribution of the data for given $\boldsymbol{\theta}$. This is a proper probability distribution that integrates to unity when integrated over all possible data $\mathbf{d}$. Used this way it is properly called the Sampling Distribution.

In Bayesian inference, though, the data are fixed (that is what we have), and this term is treated as a function of $\boldsymbol{\theta}$. In this context, it is called the Likelihood, and is not a proper probability distribution, in the sense that integrating it over $\boldsymbol{\theta}$ at fixed $\mathbf{d}$ does not give unity. Only the full posterior does this.

In practice, real data often contains outliers that can strongly affect inference with standard Gaussian likelihoods. subsection 14.5 (Robust Statistics) discusses robust likelihoods using heavy-tailed distributions like the Student-t that down-weight extreme observations.

3.2 Example: Gaussian Likelihood and Maximum Likelihood Estimation

Example: Gaussian Likelihood and Maximum Likelihood Estimation

Problem: A scientist performs repeated measurements of a physical quantity and wants to estimate its true value. Each measurement is subject to random Gaussian noise with known standard deviation. Given a set of independent measurements, how should we combine them to obtain the best estimate of the true value? What is the uncertainty in this estimate? This is one of the most fundamental problems in data analysis, arising in contexts from laboratory experiments to astronomical observations.

Solution:

Suppose we have N independent measurements $\{d_1, d_2, \ldots, d_N\}$ of a quantity whose true value is μ. Each measurement is corrupted by independent Gaussian noise with known variance σ^2:

$$d_i = \mu + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{12}$$

The sampling distribution for each measurement is:

$$p(d_i|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(d_i - \mu)^2}{2\sigma^2}\right) = \mathcal{N}(d_i; \mu, \sigma^2) \tag{13}$$

Since the measurements are independent, the joint sampling distribution (the probability of observing all the data given the parameters) is the product of individual probabilities:

$$p(\mathbf{d}|\mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(d_i - \mu)^2}{2\sigma^2}\right) \tag{14}$$

Taking the logarithm (which is monotonic and thus preserves the location of extrema) gives the log-likelihood:

$$\ln \mathcal{L}(\mu, \sigma) = -\frac{N}{2} \ln(2\pi) - N \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (d_i - \mu)^2 \tag{15}$$

Maximum Likelihood Estimate for the Mean:

To find the maximum likelihood estimate (MLE) of μ, we differentiate the log-likelihood with respect to μ and set the derivative to zero:

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (d_i - \mu) = 0 \tag{16}$$

Solving for μ yields the familiar sample mean:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N} d_i = \bar{d} \tag{17}$$

This result confirms the intuitive notion that the best estimate of the true value is simply the average of all measurements.

Maximum Likelihood Estimate for the Variance:

If the variance σ^2 is also unknown, we can find its MLE by differentiating with respect to σ:

$$\frac{\partial \ln \mathcal{L}}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(d_i - \mu)^2 = 0 \tag{18}$$

Solving yields:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N}(d_i - \hat{\mu})^2 \tag{19}$$

This is the sample variance. Note that the MLE of the variance uses N in the denominator, which is known to be biased. The unbiased estimator uses $N-1$ instead, accounting for the fact that one degree of freedom was used to estimate μ. From a Bayesian perspective with uniform priors, the MAP estimate coincides with the MLE, but the full posterior provides proper uncertainty quantification without needing to apply ad-hoc bias corrections.

Geometric Interpretation:

The MLE for μ can be understood geometrically: it minimizes the sum of squared residuals $\sum_{i=1}^{N}(d_i - \mu)^2$. This is equivalent to finding the point μ that is closest to all data points in a least-squares sense. The Gaussian assumption makes this the maximum likelihood solution, but least-squares is optimal only when errors are truly Gaussian. For non-Gaussian noise (e.g., heavy-tailed distributions with outliers), other likelihood functions may be more appropriate.

Connection to Bayesian Inference and Posterior Derivation:

The likelihood function derived here forms the foundation for Bayesian parameter estimation. In the Bayesian framework, we combine this likelihood with a prior $\pi(\mu)$ to obtain the posterior using Bayes' theorem:

$$p(\mu|\mathbf{d}, \sigma) = \frac{p(\mathbf{d}|\mu, \sigma)\pi(\mu)}{p(\mathbf{d}|\sigma)} \tag{20}$$

Let us derive the posterior distribution for the simplest case: a uniform (noninformative) prior on μ. This illustrates the basic mechanics of Bayesian updating when we have no prior knowledge.

Step 1: Write down the likelihood

From our earlier derivation, the likelihood for N independent Gaussian measurements is:

$$p(\mathbf{d}|\mu, \sigma) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i - \mu)^2\right) \tag{21}$$

Step 2: Complete the square in μ

Expanding the sum of squares and completing the square (details omitted), we obtain:

$$\sum_{i=1}^{N}(d_i - \mu)^2 = \sum_{i=1}^{N} d_i^2 - N\bar{d}^2 + N(\mu - \bar{d})^2 \tag{22}$$

where $\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$ is the sample mean.

Step 3: Apply Bayes' theorem with uniform prior

A uniform prior $\pi(\mu) = \text{const}$ is improper (does not integrate to unity) but is often used as a noninformative prior when we have no prior knowledge about μ. With a uniform prior, the posterior is proportional to the likelihood:

$$p(\mu|\mathbf{d}, \sigma) \propto p(\mathbf{d}|\mu, \sigma) \propto \exp\left(-\frac{N(\mu - \bar{d})^2}{2\sigma^2}\right) \tag{23}$$

Step 4: Recognize the Gaussian form
This is the kernel of a Gaussian distribution in μ with mean $\bar{d}$ and variance σ^2/N. The properly normalized posterior is:

$$p(\mu|\mathbf{d}, \sigma) = \mathcal{N}\left(\mu; \bar{d}, \frac{\sigma^2}{N}\right) \tag{24}$$

Interpretation:
This derivation shows several important features of Bayesian inference:

- The posterior mean is the sample mean $\bar{d}$, which coincides with the MLE when using a uniform prior
- The posterior variance is σ^2/N, meaning the posterior uncertainty decreases as $1/\sqrt{N}$ with more measurements—a manifestation of the central limit theorem
- With a uniform prior, the MAP estimate coincides with the MLE, but the Bayesian approach provides the full posterior distribution, which quantifies our uncertainty about μ rather than just a point estimate
- The posterior is properly normalized (integrates to unity over μ), unlike the likelihood which does not integrate to unity over μ for fixed data

For the case of a Gaussian prior on μ, which demonstrates conjugacy and Bayesian learning, see section 4.

4 The Prior

4.1 Specifying Prior Distributions

The prior distribution $\pi(\boldsymbol{\theta})$ encodes our state of knowledge about parameters before observing new data. Priors can be derived from three main sources:

1. **Previous experiments:** When prior experimental data exist, they provide empirical information about parameter values. The posterior from a previous analysis can serve as the prior for a new dataset, creating a sequential updating framework. In fact, probabilities are sometimes explicitly written to include this prior information I, as $p(\boldsymbol{\theta}|I, M)$, though we typically suppress this notation for clarity.

2. **Theoretical models:** Physical theories, conservation laws, or domain knowledge can constrain parameter ranges or suggest functional forms. For example, in physics we might know that a mass must be positive, or that a probability must lie in $[0, 1]$. Theoretical models can also specify relationships between parameters that inform prior construction.

3. **Analysis choices:** In the absence of previous experiments or strong theoretical constraints, we may choose priors based on analytical considerations:

   - **Uniform priors**: $\pi(\theta) = \text{const}$ are typical for *location parameters*, such as the mean of a distribution
   - **Jeffreys priors**: $\pi(\theta) \propto 1/\theta$ are often applied to *scale parameters* where the parameter is positive. Each decade is equally likely, making the prior uniform in $\ln \theta$. For a rigorous information-theoretic foundation of Jeffreys priors, see subsubsection 4.3.4 (Information Theory) and subsection 14.6 (Maximum Entropy Priors).
   - **Conjugate priors**: Priors chosen so that the posterior has the same functional form as the prior, simplifying analytical calculations (see section 4)

The choice of prior matters for two distinct reasons:

- **Impact on results:** The prior directly affects the posterior distribution, particularly when data are sparse or weakly informative. For parameter inference with abundant data, the likelihood dominates and the prior becomes less important. However, for model comparison (see section 10), the prior remains critically important regardless of data quantity.

- **Analytical simplification:** Certain prior choices (particularly conjugate priors) enable closed-form posterior calculations, avoiding the need for numerical methods like MCMC. This can dramatically reduce computational cost and improve interpretability.

For a Gaussian likelihood, one might reasonably choose a uniform prior for the mean (a location parameter) and a Jeffreys prior for the standard deviation (a scale parameter). Note that we sometimes assume a uniform prior over an infinite range, which is an *improper prior*—it cannot be normalized to integrate to 1. Provided it yields a proper posterior, this is acceptable for parameter inference. (For model comparison, we must use proper priors).

4.2 Uninformative Priors and the Curse of Dimensionality

Using previous data to define our state of knowledge is fine, but the very first dataset that was used to determine our state of knowledge will have had to have a prior with no previous data to go on. For such situations, we often try to choose an 'uninformative' prior, which is not as easy as it sounds (and its meaning may not be particularly well-defined).

A uniform prior may seem natural, but it is worth thinking a bit more. Consider this problem: imagine a parameter space with N dimensions, represented by cartesian coordinates $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$, with each

coordinate restricted to the range $(-\frac{1}{2}, \frac{1}{2})$. We adopt a uniform prior over this N-dimensional hypercube:

$$\pi(\boldsymbol{\theta}) = \begin{cases} 1 & \text{if } |\theta_i| \leq \frac{1}{2} \text{ for all } i = 1, \ldots, N \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

This prior assigns equal probability density to all points within the hypercube. The total volume of this hypercube is $V_{\text{cube}} = 1^N = 1$.

Now consider the N-sphere of radius $r = \frac{1}{2}$ inscribed within this hypercube (the largest sphere that fits inside). The volume of an N-sphere of radius r is given by:

$$V_{\text{sphere}}(r) = \frac{\pi^{N/2}}{\Gamma(1 + N/2)} r^N \tag{26}$$

For $r = \frac{1}{2}$, this becomes:

$$V_{\text{sphere}}\left(\frac{1}{2}\right) = \frac{\pi^{N/2}}{\Gamma(1 + N/2)} \left(\frac{1}{2}\right)^N = \frac{\pi^{N/2}}{2^N \Gamma(1 + N/2)} \tag{27}$$

Under the uniform prior, the probability of a randomly drawn parameter vector $\boldsymbol{\theta}$ falling inside the inscribed N-sphere is simply the ratio of the sphere volume to the hypercube volume:

$$p(\boldsymbol{\theta} \in \text{sphere}) = \frac{V_{\text{sphere}}}{V_{\text{cube}}} = \frac{\pi^{N/2}}{2^N \Gamma(1 + N/2)} \tag{28}$$

This probability decreases rapidly as N increases, illustrating the curse of dimensionality: in high dimensions, most of the volume of a hypercube is concentrated in its corners, far from the center. The ratio of the volume of the inscribed N-sphere to the volume of the hypercube goes to zero exponentially fast (see Figure 2). For the simplest case of $N = 2$, this reduces to the familiar $\pi/4$. This counterintuitive behavior demonstrates that uniform priors can be highly informative in high-dimensional spaces, as they implicitly favor parameter values at the extremes of the prior range rather than near the center.

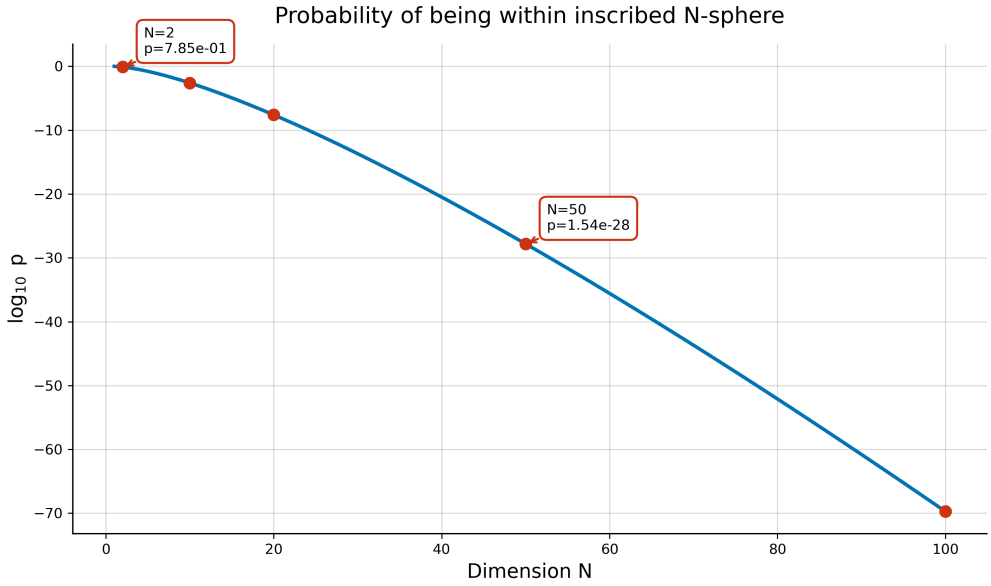

Figure 2: Probability of being within an inscribed N-sphere as a function of dimension N. The plot shows $\log_{10} p$ vs N, demonstrating the curse of dimensionality: as dimensions increase, nearly all the probability mass concentrates in the corners of the hypercube, far from the center.

An apparently uninformative prior may be highly informative when viewed a different way.

4.3 Jeffreys Prior (unassessed)

The Jeffreys prior provides a principled approach to constructing "uninformative" or "reference" priors that are invariant under reparameterization. This is a crucial property: our inferences about physical quantities should not depend on arbitrary choices of how we parameterize them.

4.3.1 Motivation: Reparameterization Invariance

Consider a parameter θ with prior $\pi(\theta)$. If we reparameterize using $\phi = f(\theta)$ for some invertible function f, the induced prior on ϕ is obtained via the change of variables formula:

$$\pi(\phi) = \pi(\theta) \left| \frac{d\theta}{d\phi} \right| \tag{29}$$

A uniform prior $\pi(\theta) = \text{const}$ on θ does not remain uniform under reparameterization. For example, if $\phi = \ln \theta$, then:

$$\pi(\phi) = \pi(\theta) \frac{d\theta}{d\phi} = \text{const} \cdot e^{\phi} \tag{30}$$

which is exponentially weighted in ϕ. This demonstrates that "uninformativeness" is not preserved under reparameterization.

Jeffreys (1946) proposed that an uninformative prior should satisfy a reparameterization invariance property: if $\pi_J(\theta)$ is the Jeffreys prior for θ, and we transform to $\phi = f(\theta)$, then the induced prior on ϕ should equal the Jeffreys prior for ϕ computed directly.

4.3.2 Derivation for a Single Parameter

The Jeffreys prior is defined as:

$$\pi_J(\theta) \propto \sqrt{I(\theta)} \tag{31}$$

where $I(\theta)$ is the Fisher information:

$$I(\theta) = -\mathbb{E}\left[\frac{d^2 \ln p(d|\theta)}{d\theta^2} \right] = \mathbb{E}\left[\left(\frac{d \ln p(d|\theta)}{d\theta} \right)^2 \right] \tag{32}$$

where the expectation is over the data distribution $p(d|\theta)$.

To verify reparameterization invariance, consider the transformation $\phi = f(\theta)$. The Fisher information in terms of ϕ is:

$$I(\phi) = I(\theta) \left(\frac{d\theta}{d\phi} \right)^2 \tag{33}$$

Under the change of variables, the Jeffreys prior transforms as:

$$\pi_J(\phi) = \pi_J(\theta) \left| \frac{d\theta}{d\phi} \right| = \sqrt{I(\theta)} \left| \frac{d\theta}{d\phi} \right| = \sqrt{I(\theta) \left(\frac{d\theta}{d\phi} \right)^2} = \sqrt{I(\phi)} \tag{34}$$

Thus, the induced prior on ϕ is precisely the Jeffreys prior computed directly for ϕ, establishing reparameterization invariance.

4.3.3 Example: Location and Scale Parameters

Location parameter: For a location parameter in models where $p(d|\mu) = f(d - \mu)$ (e.g., Gaussian mean with fixed variance), the Fisher information is constant: $I(\mu) = \text{const}$. The Jeffreys prior is therefore:

$$\pi_J(\mu) \propto 1 \tag{35}$$

which is the uniform prior.

Scale parameter: For a scale parameter $\sigma > 0$ in models where $p(d|\sigma) = \frac{1}{\sigma} f(d/\sigma)$ (e.g., Gaussian standard deviation with fixed mean), the Fisher information scales as $I(\sigma) \propto 1/\sigma^2$. The Jeffreys prior is:

$$\pi_J(\sigma) \propto \frac{1}{\sigma} \tag{36}$$

This is the prior mentioned earlier that is uniform in $\ln \sigma$, making each decade of σ equally probable. It reflects the intuition that for a scale parameter, ratios matter more than differences: the difference between $\sigma = 1$ and $\sigma = 2$ should be weighted the same as the difference between $\sigma = 10$ and $\sigma = 20$.

4.3.4 Fisher Information Matrix (unassessed)

Information theory provides a rigorous mathematical framework for quantifying the information content of data and for constructing uninformative priors. The central object in this framework is the Fisher information matrix, which measures how much information an experiment provides about model parameters and establishes fundamental limits on parameter estimation precision.

The Fisher information matrix quantifies the curvature of the log-likelihood function. For a model with parameters $\boldsymbol{\theta}$ and data $\mathbf{d}$ drawn from the sampling distribution $p(\mathbf{d}|\boldsymbol{\theta})$, the Fisher information matrix has elements:

$$\mathcal{F}_{ij}(\boldsymbol{\theta}) = -\left\langle \frac{\partial^2 \ln p(\mathbf{d}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} \right\rangle = \left\langle \frac{\partial \ln p(\mathbf{d}|\boldsymbol{\theta})}{\partial\theta_i} \frac{\partial \ln p(\mathbf{d}|\boldsymbol{\theta})}{\partial\theta_j} \right\rangle \tag{37}$$

where the angle brackets denote expectation over the data distribution $p(\mathbf{d}|\boldsymbol{\theta})$. The equality between the two expressions follows from differentiating the normalization condition $\int p(\mathbf{d}|\boldsymbol{\theta}) d\mathbf{d} = 1$ and applying integration by parts, assuming appropriate regularity conditions. The Fisher information matrix is symmetric and positive semi-definite, encoding the local geometry of the likelihood function near the true parameter values.

The Cramér-Rao bound establishes a fundamental limit on the precision of parameter estimation. For any unbiased estimator $\hat{\boldsymbol{\theta}}$ of the parameters (meaning $\mathbb{E}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$), the covariance matrix of the estimator satisfies:

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) \geq \mathcal{F}^{-1} \tag{38}$$

where the inequality means that the difference $\mathrm{Cov}(\hat{\boldsymbol{\theta}}) - \mathcal{F}^{-1}$ is positive semi-definite. This bound tells us that no unbiased estimator can achieve smaller uncertainties than those specified by the inverse Fisher matrix, establishing a fundamental limit imposed by the information content of the data. Maximum likelihood estimators are asymptotically efficient, meaning they achieve this bound in the limit of large sample size.

In the Bayesian context, the Fisher information connects to the posterior distribution through the Bernstein-von Mises theorem. For regular models with sufficiently concentrated priors and large datasets, the posterior distribution becomes approximately Gaussian centered at the maximum likelihood estimate, with covariance given by the inverse Fisher information: $p(\boldsymbol{\theta}|\mathbf{d}) \approx \mathcal{N}(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}, \mathcal{F}^{-1})$. This Gaussian approximation forms the basis of the Laplace approximation (subsection 14.7) and provides intuition for why Bayesian and frequentist approaches often yield similar results for large datasets.

The Fisher information matrix plays a central role in optimal data compression schemes such as MOPED (section 13), where the compression vectors are constructed from the derivatives of the mean with respect to parameters, weighted by the inverse covariance—a structure closely related to the Fisher matrix. In experimental design, the Fisher information quantifies how informative different experimental configurations would be, enabling principled selection of measurements that maximize expected information gain about parameters of interest (subsection 14.9).

4.3.5 Multivariate Jeffreys Prior

Using the Fisher information matrix $\mathcal{F}$ defined in the previous section, the multivariate Jeffreys prior for parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is:

$$\pi_J(\boldsymbol{\theta}) \propto \sqrt{\det \mathcal{F}(\boldsymbol{\theta})} \tag{39}$$

This prior represents maximal ignorance about the parameters in the sense that it treats all parameter values equally after accounting for the natural metric induced by the likelihood geometry. The determinant of the Fisher matrix can be interpreted as the volume element of the parameter space under the Fisher metric, and the Jeffreys prior assigns uniform probability density per unit Fisher volume.

4.3.6 Limitations and Practical Considerations

While the Jeffreys prior has appealing theoretical properties, it has important limitations:

- **Improper priors:** The Jeffreys prior is often improper (does not integrate to 1). For example, $\pi(\sigma) \propto 1/\sigma$ on $(0, \infty)$ is improper. This is acceptable for parameter inference if the posterior is proper, but precludes use in model comparison.
- **Multiparameter non-invariance:** For multiple parameters, the Jeffreys prior is not invariant under arbitrary subspace transformations. Different orderings of marginalization can yield different priors.
- **Information dependence:** The Jeffreys prior depends on the likelihood function and thus on the experimental setup, which some view as contradicting the notion of a "prior" that should precede data collection.
- **Computational challenges:** Computing the Fisher information requires evaluating expectations over the data distribution, which may be analytically or computationally difficult for complex models.

Despite these limitations, the Jeffreys prior remains a valuable tool for constructing reference priors, particularly for scale parameters where it captures the natural invariance structure of the problem.

4.4 Sequential Updating of Priors

If we now obtain some more information, perhaps from a new experiment, then we can use Bayes' theorem to update our state of knowledge of the parameters. The posterior of the last experiment becomes the prior for the next one. This is fine, but it begs the question of what prior did the very first experiment use? This is where the 'uninformative' priors come in.

For this to be a consistent process, we should verify that the order and manner in which we analyze data does not affect the final posterior. Given two datasets $\mathbf{d}_1$ and $\mathbf{d}_2$, there are several possible analysis scenarios:

1. **Analyzing $\mathbf{d}_1$ alone:** Update the prior $\pi(\boldsymbol{\theta})$ with only the first dataset to obtain $p(\boldsymbol{\theta}|\mathbf{d}_1)$
2. **Analyzing $\mathbf{d}_2$ alone:** Update the prior $\pi(\boldsymbol{\theta})$ with only the second dataset to obtain $p(\boldsymbol{\theta}|\mathbf{d}_2)$
3. **Analyzing $\mathbf{d}_1$ before $\mathbf{d}_2$:** First obtain $p(\boldsymbol{\theta}|\mathbf{d}_1)$, then use this as the prior for analyzing $\mathbf{d}_2$ to obtain $p(\boldsymbol{\theta}|\mathbf{d}_2, \mathbf{d}_1)$
4. **Analyzing $\mathbf{d}_2$ before $\mathbf{d}_1$:** First obtain $p(\boldsymbol{\theta}|\mathbf{d}_2)$, then use this as the prior for analyzing $\mathbf{d}_1$ to obtain $p(\boldsymbol{\theta}|\mathbf{d}_1, \mathbf{d}_2)$
5. **Analyzing $\mathbf{d}_1$ and $\mathbf{d}_2$ together:** Combine both datasets and update the prior simultaneously to obtain $p(\boldsymbol{\theta}|\mathbf{d}_1, \mathbf{d}_2)$

A key requirement for consistency is that scenarios 3, 4, and 5 must yield the same final posterior distribution. We now demonstrate this equivalence, showing that sequential updating (scenarios 3 and 4) produces identical results to joint analysis (scenario 5), and that the order of sequential updates does not matter.

Let's do the analysis in two stages, firstly analysing $\mathbf{d}_1$. Let's be explicit about the prior information I (defined to be the state of knowledge before the first experiment is done) in Bayes' theorem applied to the first dataset:

$$p(\boldsymbol{\theta}|\mathbf{d}_1, I) = \frac{p(\mathbf{d}_1|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)}{p(\mathbf{d}_1|I)}. \tag{40}$$

Now we analyse the second data set. It's similar, but the data are different, $\mathbf{d}_1 \to \mathbf{d}_2$ of course, and we have some extra information from the first experiment, so we should update I to include $\mathbf{d}_1$:

$$I \to \mathbf{d}_1, I. \tag{41}$$

So, Bayes' theorem applied to the second dataset gives a posterior

$$p(\boldsymbol{\theta}|\mathbf{d}_2, \mathbf{d}_1, I) = \frac{p(\mathbf{d}_2|\boldsymbol{\theta}, \mathbf{d}_1, I)p(\boldsymbol{\theta}|\mathbf{d}_1, I)}{p(\mathbf{d}_2|\mathbf{d}_1, I)}. \tag{42}$$

We now notice that the new prior in this expression is just the old posterior probability from equation (4.1), i.e. we have updated our prior state of knowledge from the original prior, instead using the posterior from the first dataset.

We can also the rules of probability to write the new likelihood as

$$p(\mathbf{d}_2|\mathbf{d}_1, \boldsymbol{\theta}, I) = \frac{p(\mathbf{d}_2, \mathbf{d}_1|\boldsymbol{\theta}, I)}{p(\mathbf{d}_1|\boldsymbol{\theta}, I)}. \tag{43}$$

Substituting this into equation (4.3) in the old posterior probability along with the expression for the posterior after analysing $\mathbf{d}_1$ (equation 4.1 gives

$$p(\boldsymbol{\theta}|\mathbf{d}_1, \mathbf{d}_2, \boldsymbol{\theta}, I) = \frac{1}{p(\mathbf{d}_2|\mathbf{d}_1, I)} \times \frac{p(\mathbf{d}_2, \mathbf{d}_1|\boldsymbol{\theta}, I)}{\cancel{p(\mathbf{d}_1|\boldsymbol{\theta}, I)}} \times \frac{\cancel{p(\mathbf{d}_1|\boldsymbol{\theta}, I)}p(\boldsymbol{\theta}|I)}{p(\mathbf{d}_1|I)} \tag{44}$$

So we see that the posterior after the second dataset is

$$p(\boldsymbol{\theta}|\mathbf{d}_2, \mathbf{d}_1, I) = \frac{p(\mathbf{d}_2, \mathbf{d}_1|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)}{p(\mathbf{d}_2, \mathbf{d}_1|I)}, \tag{45}$$

where we have used the product rule in the denominator. This has the same form as equation (4.1), the outcome from the initial experiment, but now with the new data incorporated, i.e. the result we would write down if we analysed the data together ($\mathbf{d} \to \{\mathbf{d}_1, \mathbf{d}_2\}$). So, analysing separately and updating the prior after the first dataset gives the same answer as analysing the combined dataset with the original prior.

Bayes' theorem gives us a natural and self-consistent way of improving our statistical inferences as our state of knowledge increases.

4.5 Conjugate Priors

Sometimes a prior is chosen for mathematical convenience, where, when combined with a given form for the likelihood, the posterior can be calculated analytically and has the same mathematical form as the prior. This property—called conjugacy—offers both computational advantages (avoiding numerical integration or sampling) and interpretive benefits (the prior and posterior have the same functional form, making sequential updating transparent).

Examples of conjugate pairs include:

- **Gaussian likelihood with Gaussian prior:** If the likelihood is Gaussian with known variance, and the mean μ is the parameter of interest, then a conjugate prior is also a Gaussian. The conjugate prior can have any mean and variance, so it is flexible enough to encode various degrees of prior knowledge.
- **Gaussian likelihood with known mean:** If the mean μ is known but the variance σ^2 is unknown, an inverse gamma distribution is a conjugate prior for σ^2.
- **Binomial likelihood:** For coin flips or Bernoulli trials, the Beta distribution is conjugate to the Binomial likelihood (demonstrated in the coin toss example below).

Note that there is nothing inherently special about conjugate priors; they are just convenient mathematically, but they may be flexible enough to specify sensible location and scale constraints. When they are not flexible enough or inappropriate for the problem, non-conjugate priors should be used, accepting the need for numerical methods.

We now present a detailed worked example of Gaussian-Gaussian conjugacy, extending the Gaussian likelihood example from section 3 by incorporating an informative prior.

Example: Gaussian Likelihood with Gaussian Prior: Complete Conjugacy Derivation

Problem: Building on the Gaussian likelihood example from section 3, suppose we now have prior knowledge about the mean parameter μ from previous experiments or theoretical considerations. Rather than using a uniform (noninformative) prior, we encode this prior knowledge as a Gaussian distribution with mean μ_0 and variance σ_0^2. How do we combine this informative prior with new data to obtain the posterior distribution? This scenario is fundamental in Bayesian inference and demonstrates the power of conjugacy: when both the prior and likelihood are Gaussian, the posterior is also Gaussian, and we can derive its parameters analytically.

Solution:

As in section 3, we have N independent measurements $\{d_1, d_2, \ldots, d_N\}$ with known measurement variance σ^2. The likelihood is:

$$p(\mathbf{d}|\mu,\sigma) = \prod_{i=1}^{N} \mathcal{N}(d_i;\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i-\mu)^2\right) \tag{46}$$

We now specify a Gaussian prior on μ:

$$\pi(\mu) = \mathcal{N}(\mu;\mu_0,\sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \tag{47}$$

Step 1: Apply Bayes' theorem

The posterior is proportional to the product of likelihood and prior:

$$p(\mu|\mathbf{d},\sigma) \propto p(\mathbf{d}|\mu,\sigma)\pi(\mu) \tag{48}$$

Taking logarithms (to work with the exponents more easily):

$$\ln p(\mu|\mathbf{d},\sigma) = \text{const} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i-\mu)^2 - \frac{(\mu-\mu_0)^2}{2\sigma_0^2} \tag{49}$$

where "const" absorbs all terms that do not depend on μ.

Step 2: Express likelihood as Gaussian in μ

From section 3, completing the square in the likelihood gives:

$$p(\mathbf{d}|\mu,\sigma) \propto \mathcal{N}\left(\mu;\bar{d},\frac{\sigma^2}{N}\right) \tag{50}$$

where $\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$ is the sample mean. The likelihood is a Gaussian in μ with mean $\bar{d}$ and variance σ^2/N.

Step 3: Apply product of Gaussians identity

The posterior is proportional to the product of two Gaussians (likelihood and prior), both in μ:

$$p(\mu|\mathbf{d},\sigma) \propto \mathcal{N}\left(\mu;\bar{d},\frac{\sigma^2}{N}\right) \cdot \mathcal{N}(\mu;\mu_0,\sigma_0^2) \tag{51}$$

Using the product of Gaussians identity from section 1, this product is proportional to a Gaussian:

$$p(\mu|\mathbf{d},\sigma) \propto \mathcal{N}(\mu;\mu_N,\sigma_N^2) \tag{52}$$

where the posterior precision (inverse variance) is the sum of the individual precisions:

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \tag{53}$$

and the posterior mean is the precision-weighted average:

$$\mu_N = \sigma_N^2 \left(\frac{N\bar{d}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) = \frac{\frac{N}{\sigma^2}\bar{d} + \frac{1}{\sigma_0^2}\mu_0}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}} \tag{54}$$

Step 4: Write the posterior
The posterior distribution is:

$$p(\mu|\mathbf{d}, \sigma) = \mathcal{N}(\mu; \mu_N, \sigma_N^2) \tag{55}$$

with parameters given above.
Interpretation:
This result reveals several profound insights:

- **Precision addition:** The posterior precision is the sum of the data precision (N/σ^2) and the prior precision ($1/\sigma_0^2$). Information combines additively when expressed as precisions.
- **Weighted average:** The posterior mean μ_N is a precision-weighted average of the sample mean $\bar{d}$ and the prior mean μ_0. Sources with higher precision (lower variance) receive more weight.
- **Data vs. prior dominance:**
  - If $N\sigma_0^2 \gg \sigma^2$ (many data or tight prior), then $\mu_N \approx \bar{d}$ (data dominates)
  - If $N\sigma_0^2 \ll \sigma^2$ (few data or loose prior), then $\mu_N \approx \mu_0$ (prior dominates)
- **Variance reduction:** The posterior variance σ_N^2 is always smaller than both the prior variance σ_0^2 and the data variance σ^2/N, reflecting increased certainty from combining information.
- **Conjugacy:** The Gaussian prior and Gaussian likelihood combine to yield a Gaussian posterior with updated parameters. This makes sequential updating trivial: the posterior from one dataset becomes the prior for the next.
- **Limiting cases:**
  - Uniform prior limit: As $\sigma_0^2 \to \infty$ (infinitely diffuse prior), we recover the result from section 3: $\mu_N \to \bar{d}$ and $\sigma_N^2 \to \sigma^2/N$
  - Strong prior limit: As $\sigma_0^2 \to 0$ (infinitely precise prior), $\mu_N \to \mu_0$ and $\sigma_N^2 \to 0$ (data cannot overcome an infinitely strong prior belief)

Gaussian likelihood with known mean: If on the other hand the mean μ is known, but the variance σ^2 is unknown, an inverse gamma distribution is a conjugate prior for $x = \sigma^2$:

$$f(x; \alpha, \beta) \equiv \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\beta/x} \tag{56}$$

Exercise: show that the posterior is also an inverse gamma distribution, with parameters updated as follows $\alpha \to \alpha + n/2$, and $\beta \to \beta + \sum_{i=1}^{n}(x_i - \mu)^2/2$ for n data.

5 The Posterior

The posterior is the natural outcome of a Bayesian inference problem. It encapsulates our current state of knowledge of the model parameters. It may be very high dimensional, if there are many parameters, and we may want to put it into a more digestible form. It is common to marginalise over all but two parameters, and plot marginal posteriors as a function of each pair of parameters. These are often plotted in 'corner plots'.

Figure 3 shows a corner plot for a three-parameter model with correlated posterior distributions. The diagonal panels display the one-dimensional marginal posterior distributions for each parameter, obtained by integrating over all other parameters. The 68% credible intervals are indicated by the shaded green regions. The off-diagonal panels show the two-dimensional marginal posteriors for each pair of parameters, with contours indicating the 68% and 95% credible regions. These credible regions are highest posterior density (HPD) regions, where the posterior probability density is highest. The correlations between parameters are clearly visible in the elliptical shapes of the 2D contours—for instance, θ_1 and θ_2 are positively correlated, while θ_1 and θ_3 are negatively correlated.

5.1 Example: Beta-Binomial Conjugacy and Prior Dominance

In parameter inference problems, as more data are collected, the likelihood gets progressively more peaked around the true parameter values. For a sufficiently narrow likelihood, the prior is almost constant over the relevant range, and it becomes unimportant (the height of the prior there is irrelevant as it is normalised away by the evidence in the denominator). Let us explore this through a detailed coin-flipping example (adapted from Sivia & Skilling) that demonstrates both the mechanics of Bayesian updating and how data dominates the prior as sample size increases. The probabilistic graphical model is shown in Figure 4.

Example: Coin Toss: Beta-Binomial Conjugacy and Prior Dominance

Problem: Consider a coin-flipping experiment designed to determine whether a coin is fair. The parameter of interest is θ, the probability of obtaining heads on a single toss, which can take any value between 0 and 1. We perform a sequence of tosses and observe the outcomes. We want to: (1) derive the complete posterior distribution using Bayes' theorem, identifying all components; (2) understand how the posterior evolves as more data are collected; and (3) explore how the choice of prior affects inference, particularly comparing uninformative and informative priors.

Solution:

To understand the posterior distribution completely, we derive it step by step and identify all probability distributions involved. We apply Bayes' theorem:

$$p(\theta|\text{data}) = \frac{p(\text{data}|\theta)\pi(\theta)}{p(\text{data})} \tag{57}$$

The three components are:

1. The Prior: We assume a uniform (flat) prior over the interval $[0, 1]$:

$$\pi(\theta) = \begin{cases} 1 & \text{if } 0 \leq \theta \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{58}$$

This is a special case of the Beta distribution with parameters $\alpha = 1$ and $\beta = 1$, denoted $\text{Beta}(1, 1)$. The general Beta distribution is:

$$\text{Beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{59}$$

For $\alpha = \beta = 1$, this reduces to the uniform distribution.

2. The Likelihood: For n independent coin tosses with k heads observed, the likelihood follows a binomial distribution:

$$p(k|\theta, n) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{60}$$

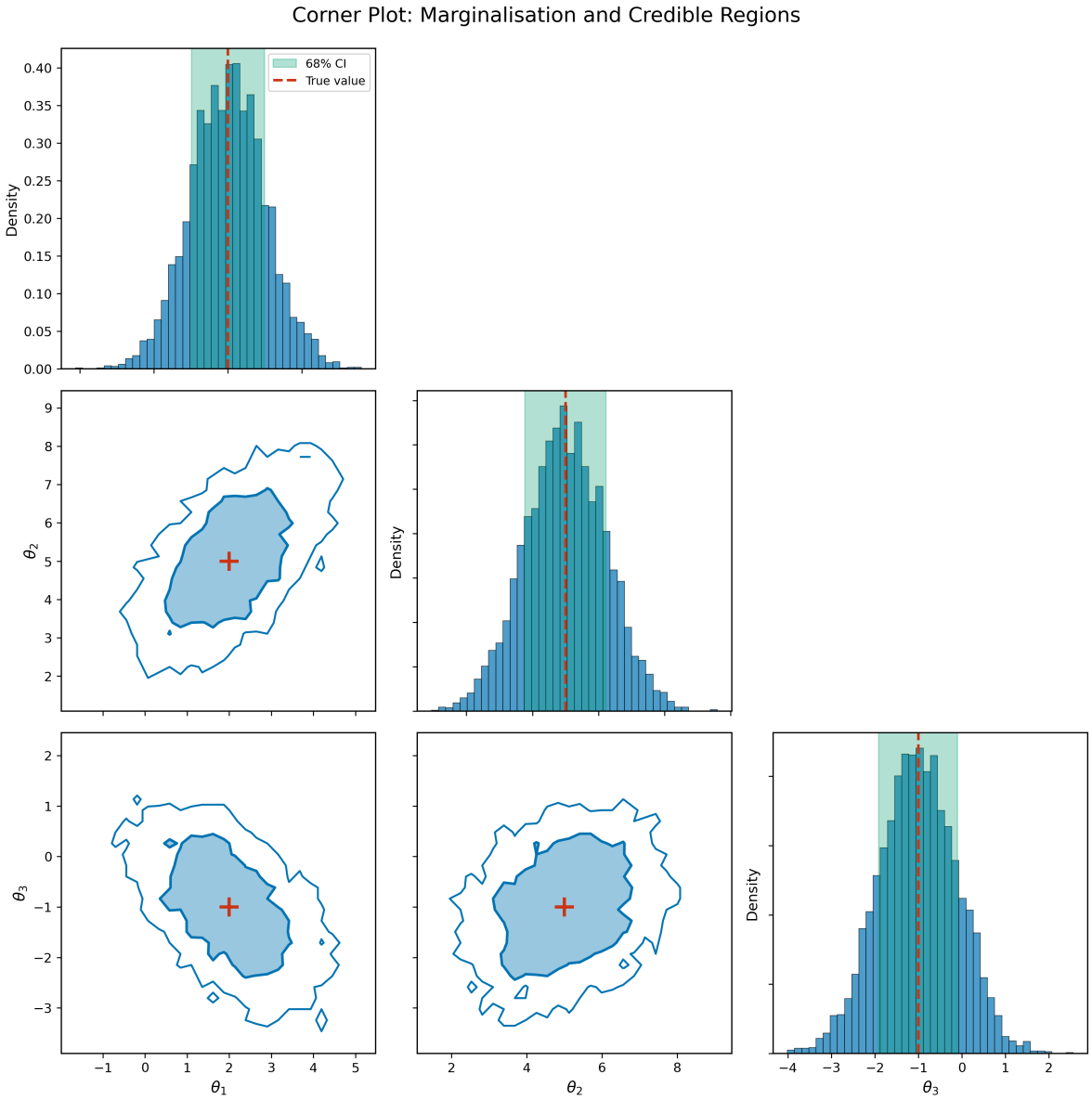

Figure 3: Corner plot illustrating marginalisation and credible regions for a three-parameter model. The diagonal shows 1D marginal posterior distributions with 68% credible intervals (green shaded regions). The off-diagonal panels display 2D marginal posteriors with 68% (dark blue) and 95% (light blue) credible contours. The red cross marks the true parameter values. The elliptical shapes of the contours reveal correlations between parameters.

Figure 4: DAG for coin toss model

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient. This represents the probability of observing exactly k heads in n tosses, given that the probability of heads on each toss is θ.

3. The Evidence: The denominator is computed by marginalizing over all possible values of θ:

$$p(k|n) = \int_0^1 p(k|\theta, n)\pi(\theta)d\theta = \int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k} \cdot 1 \, d\theta \tag{61}$$

This integral can be evaluated using the definition of the Beta function $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. With $\alpha = k+1$ and $\beta = n-k+1$:

$$p(k|n) = \binom{n}{k}B(k+1, n-k+1) = \binom{n}{k}\frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} = \binom{n}{k}\frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1} \tag{62}$$

4. The Posterior: Combining these elements:

$$p(\theta|k, n) = \frac{\binom{n}{k}\theta^k(1-\theta)^{n-k} \cdot 1}{\frac{1}{n+1}} \tag{63}$$

$$= (n+1)\binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{64}$$

$$= \frac{(n+1)!}{k!(n-k)!}\theta^k(1-\theta)^{n-k} \tag{65}$$

$$= \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n-k+1)}\theta^k(1-\theta)^{n-k} \tag{66}$$

This is precisely the Beta distribution $\text{Beta}(\theta|k+1, n-k+1)$. The posterior has updated the parameters of the Beta prior: $\alpha = 1 \to k+1$ and $\beta = 1 \to n-k+1$.

5. Sequential Updating and Prior Dominance:
After the first toss, suppose we observe a head (H). Using Bayes' theorem, the posterior is:

$$p(\theta|H) \propto p(H|\theta)\pi(\theta) = \theta \times 1 = \theta \tag{67}$$

This posterior is linear in θ, favoring higher values of θ but still allowing for the possibility that the coin is biased toward tails.
As we continue tossing the coin, we accumulate more data. Suppose the sequence begins HHTT... and eventually we observe 40 heads in 64 tosses. For this specific case, the posterior is:

$$p(\theta|k=40, n=64) = \text{Beta}(\theta|41, 25) = \frac{\Gamma(66)}{\Gamma(41)\Gamma(25)}\theta^{40}(1-\theta)^{24} \tag{68}$$

The mean of this Beta distribution is $\frac{\alpha}{\alpha+\beta} = \frac{41}{66} \approx 0.621$ and the mode (the peak) is $\frac{\alpha-1}{\alpha+\beta-2} = \frac{40}{64} = 0.625$, which matches the observed frequency of heads. The variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \approx 0.0035$, indicating a relatively narrow distribution concentrated around the observed frequency.

6. Comparison with Informative Priors:
Figure 5 illustrates posteriors for different scenarios comparing uninformative and informative priors with varying amounts of data. The key insights are:

- **Small sample size:** With only a few tosses, the choice of prior matters considerably. An informative prior strongly influences the posterior, and different priors yield substantially different posteriors.
- **Large sample size:** After many tosses, the likelihood dominates, and the posterior is largely determined by the data rather than the prior. If we had started with a different prior—say, one strongly biased toward $\theta = 0.5$ (a highly informative prior favoring a fair coin)—we would need more data to overcome this prior belief. But eventually, with sufficient data, even a highly informative prior would be overwhelmed by the evidence, and the posteriors from different priors would converge to similar distributions centered on the observed frequency.
- **Data dominance principle:** This demonstrates a key principle of Bayesian inference: as more data are collected, the posterior becomes increasingly concentrated around the true parameter value, and the influence of the prior diminishes. The prior is eventually "washed out" by the data.

5.2 Marginalisation and Nuisance Parameters

In many inference problems, we are interested in only a subset of the model parameters. The remaining parameters—called nuisance parameters—must be included in the model because they affect the data, but we do not care about their values for our scientific question. Marginalization is the operation that allows us to "integrate out" these nuisance parameters, obtaining the posterior distribution for only the parameters of interest while properly accounting for uncertainty in the nuisance parameters.

The Marginalisation Operation:
Marginalising over all n parameters except θ_1 and θ_2 is accomplished by integration:

$$p(\theta_1, \theta_2|\mathbf{d}) = \int p(\theta_1, \ldots, \theta_n|\mathbf{d}) d\theta_3 \ldots d\theta_n \tag{69}$$

More generally, suppose we partition the parameters into those of interest $\boldsymbol{\phi}$ (the target parameters) and nuisance parameters $\boldsymbol{\psi}$, so that $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\psi})$. The joint posterior is:

$$p(\boldsymbol{\phi}, \boldsymbol{\psi}|\mathbf{d}) = \frac{p(\mathbf{d}|\boldsymbol{\phi}, \boldsymbol{\psi})\pi(\boldsymbol{\phi}, \boldsymbol{\psi})}{p(\mathbf{d})} \tag{70}$$

To obtain the marginal posterior for the parameters of interest, we integrate over the nuisance parameters:

$$p(\boldsymbol{\phi}|\mathbf{d}) = \int p(\boldsymbol{\phi}, \boldsymbol{\psi}|\mathbf{d}) d\boldsymbol{\psi} \tag{71}$$

This marginalisation automatically accounts for the uncertainty in the nuisance parameters, producing the correct uncertainties for $\boldsymbol{\phi}$.

Analytical vs Numerical Marginalization:
Marginalization can sometimes be performed analytically, particularly when:

- The posterior has an analytical form (often the case with conjugate priors)
- The integral over nuisance parameters can be evaluated in closed form
- The model structure allows separation of parameters

However, most real-world problems require numerical marginalization through methods like:

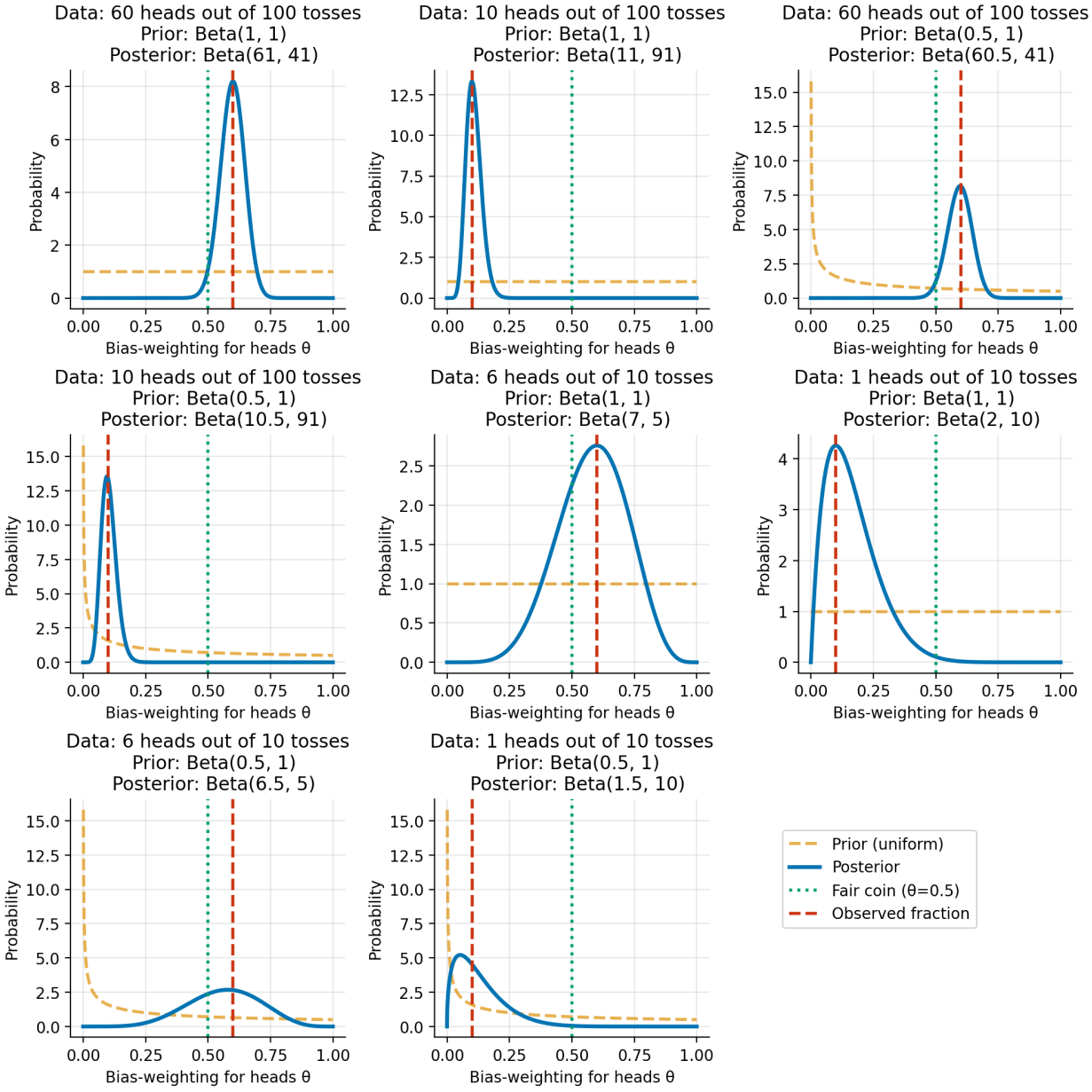

Figure 5: Coin toss analysis showing different scenarios for the data and the prior distributions, yielding different posterior distributions. As the number of observations increases, the posterior becomes increasingly dominated by the data, and the choice of prior becomes less important.

- **Numerical integration** for low-dimensional nuisance parameter spaces
- **Monte Carlo sampling** (MCMC, section 6) where marginalization is achieved by simply ignoring samples of nuisance parameters when computing summaries
- **Variational inference** (subsection 14.2) for approximate marginalization in high dimensions

We now present three examples illustrating analytical marginalization in simple but important scenarios.

Example: Gaussian Measurements: Marginalizing Over Unknown Variance

Problem: In Sections 3.2 and 4.4, we analyzed Gaussian measurements $\{d_1, \ldots, d_N\}$ assuming the measurement variance σ^2 was known. In practice, σ^2 is often unknown and must be inferred alongside the mean μ. If we are primarily interested in μ, how do we properly account for our uncertainty in σ^2? This requires marginalizing over σ^2 as a nuisance parameter.

Solution:

Starting from the Gaussian likelihood (section 3), we now treat both μ and σ^2 as unknown. With a uniform prior on μ and a scale-invariant Jeffreys prior $\pi(\sigma^2) \propto 1/\sigma^2$, the joint posterior is:

$$p(\mu, \sigma^2|\mathbf{d}) \propto (\sigma^2)^{-(N/2+1)} \exp\left(-\frac{S(\mu)}{2\sigma^2}\right) \tag{72}$$

where $S(\mu) = \sum_{i=1}^{N}(d_i - \mu)^2$.

Marginalizing over σ^2:

$$p(\mu|\mathbf{d}) = \int_0^\infty p(\mu, \sigma^2|\mathbf{d})d\sigma^2 \propto \int_0^\infty (\sigma^2)^{-(N/2+1)} \exp\left(-\frac{S(\mu)}{2\sigma^2}\right) d\sigma^2 \tag{73}$$

This integral evaluates to a Student's t-distribution:

$$p(\mu|\mathbf{d}) \propto \left[1 + \frac{(\mu - \bar{d})^2}{s^2/(N-1)}\right]^{-N/2} \tag{74}$$

where $\bar{d} = \frac{1}{N}\sum d_i$ is the sample mean and $s^2 = \frac{1}{N-1}\sum(d_i - \bar{d})^2$ is the sample variance.

Interpretation: Marginalizing over the unknown variance produces heavier tails than the Gaussian posterior from section 3—the t-distribution properly accounts for uncertainty in σ^2. As $N \to \infty$, uncertainty in σ^2 decreases and the t-distribution approaches the Gaussian limit. This demonstrates that ignoring nuisance parameters by fixing them at point estimates underestimates uncertainty.

Example: Coin Toss: Marginalizing Over Hyperparameters

Problem: In section 5, we analyzed coin tosses using a Beta(α, β) prior on the success probability θ. Suppose we are uncertain about the hyperparameters α and β themselves—perhaps we know the coin comes from a manufacturer with variable quality control, but we don't know the exact distribution of bias. How do we account for this higher-level uncertainty when inferring θ?

Solution:

We build a hierarchical model by placing a prior $\pi(\alpha, \beta)$ on the hyperparameters:

$$k|\theta, n \sim \text{Binomial}(n, \theta) \tag{75}$$

$$\theta|\alpha, \beta \sim \text{Beta}(\alpha, \beta) \tag{76}$$

$$\alpha, \beta \sim \pi(\alpha, \beta) \tag{77}$$

The joint posterior is:

$$p(\theta, \alpha, \beta|k, n) \propto \binom{n}{k}\theta^k(1-\theta)^{n-k} \cdot \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} \cdot \pi(\alpha, \beta) \tag{78}$$

To obtain the marginal posterior for θ, we integrate over the hyperparameters:

$$p(\theta|k,n) = \int\int p(\theta,\alpha,\beta|k,n)d\alpha\, d\beta \tag{79}$$

Interpretation: For most choices of $\pi(\alpha,\beta)$, this integral requires numerical evaluation (typically via MCMC, section 6). Marginalization propagates our uncertainty about the hyperparameters into our inference about θ, yielding wider credible intervals than the fixed-hyperparameter case in section 5. This is the foundation of hierarchical Bayesian modeling (section 9): uncertainty at each level of the hierarchy is properly accounted for through marginalization.

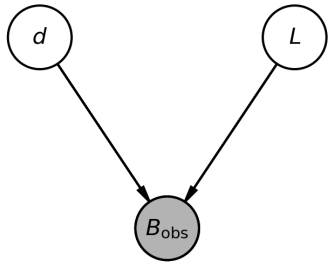

Figure 6: DAG for supernova distance measurement. Distance d (parameter of interest) and luminosity L (nuisance parameter) determine the observed brightness B_{obs}.

Example: Supernova: Nuisance Parameter Marginalization

Problem: Astronomers observe a supernova and measure its apparent brightness. They want to determine the distance to the supernova, which is crucial for cosmological measurements. However, the observed brightness depends on both the distance and the intrinsic luminosity of the supernova. Since each supernova explosion has a different intrinsic luminosity that is not directly observable, we face a classical inference challenge: how to determine the distance when another unknown parameter (the luminosity) affects our measurements. The distance is our parameter of interest, while the luminosity is a nuisance parameter that we must account for but are not primarily interested in. This example demonstrates how to properly marginalize over nuisance parameters to obtain correct uncertainties for the parameter of interest.

Solution:

The physical relationship between observed brightness, distance, and luminosity follows the inverse-square law. The model is:

$$B_{\text{obs}} = \frac{L}{4\pi d^2} + \epsilon, \quad \epsilon \sim \mathcal{N}(0,\sigma^2) \tag{80}$$

where d is the parameter of interest and L is a nuisance parameter. The likelihood is:

$$p(B_{\text{obs}}|d,L) = \mathcal{N}\left(B_{\text{obs}}; \frac{L}{4\pi d^2}, \sigma^2\right) \tag{81}$$

With a uniform prior on d and a prior $\pi(L)$ on the luminosity (perhaps from similar supernovae), the joint posterior is:

$$p(d,L|B_{\text{obs}}) \propto \exp\left(-\frac{1}{2\sigma^2}\left(B_{\text{obs}} - \frac{L}{4\pi d^2}\right)^2\right)\pi(L) \tag{82}$$

The marginal posterior for distance is obtained by integrating over all possible luminosities:

$$p(d|B_{\text{obs}}) = \int_0^\infty p(d, L|B_{\text{obs}})dL \tag{83}$$

This marginal posterior properly accounts for our uncertainty in the intrinsic luminosity L, yielding larger uncertainties in d than would be obtained if we fixed L at a single value. Other examples of nuisance parameters include detector gain (which may have a prior set by laboratory calibration measurements) or calibration parameters that affect the relationship between observed and true quantities.

5.3 Conditional vs Marginal Uncertainties

If all parameters are kept fixed (typically at the maximum posterior values), and the posterior distribution computed as a function of the remaining parameter, this is a conditional distribution, with an associated conditional error (e.g. standard deviation). For example, if we fix $\theta_2, \ldots, \theta_n$ at specific values $\theta_2^*, \ldots, \theta_n^*$ (often the maximum posterior values), the conditional distribution for θ_1 is:

$$p(\theta_1|\theta_2^*, \ldots, \theta_n^*, \mathbf{d}) \propto p(\theta_1, \theta_2^*, \ldots, \theta_n^*|\mathbf{d}) \tag{84}$$

This should be contrasted with the marginal distribution:

$$p(\theta_1|\mathbf{d}) = \int p(\theta_1, \theta_2, \ldots, \theta_n|\mathbf{d})d\theta_2 \ldots d\theta_n \tag{85}$$

The conditional distribution is rarely relevant for reporting parameter uncertainties, since it does not reflect the additional uncertainty that arises from incomplete knowledge of the other parameters. The marginal distribution properly accounts for this uncertainty by integrating over all possible values of the nuisance parameters, weighted by their posterior probability.

To see why conditional errors underestimate the true uncertainty, consider the variance. The marginal variance can be decomposed as:

$$\text{Var}[\theta_1|\mathbf{d}] = \mathbb{E}_{\theta_2,\ldots,\theta_n}[\text{Var}[\theta_1|\theta_2, \ldots, \theta_n, \mathbf{d}]] + \text{Var}_{\theta_2,\ldots,\theta_n}[\mathbb{E}[\theta_1|\theta_2, \ldots, \theta_n, \mathbf{d}]] \tag{86}$$

where the expectation and variance on the right-hand side are taken over the marginal posterior of $\theta_2, \ldots, \theta_n$. This is the law of total variance. The first term represents the average conditional variance, while the second term represents the additional variance due to uncertainty in the other parameters. Since both terms are non-negative, we always have:

$$\text{Var}[\theta_1|\mathbf{d}] \geq \text{Var}[\theta_1|\theta_2^*, \ldots, \theta_n^*, \mathbf{d}] \tag{87}$$

Thus, the marginal standard deviation (uncertainty) is always at least as large as the conditional standard deviation, with equality only when θ_1 is independent of the other parameters in the posterior.

5.4 Summarizing the Posterior Distribution

Once the posterior distribution $p(\boldsymbol{\theta}|\mathbf{d})$ has been computed, we typically need to summarize it for reporting and decision-making. The appropriate summary depends on the shape of the posterior and the intended use. This section describes the main approaches to posterior summarization: point estimates, uncertainty quantification through credible regions, and when simple summaries are sufficient versus when the full distribution must be preserved.

5.4.1 Point Estimates

Several point estimates can be extracted from the posterior, each optimal under different loss functions:

Posterior mean: The expected value of the parameters under the posterior:

$$\bar{\boldsymbol{\theta}} = \mathbb{E}[\boldsymbol{\theta}|\mathbf{d}] = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta}|\mathbf{d})d\boldsymbol{\theta} \tag{88}$$

This minimizes the expected squared error and is optimal for symmetric loss functions. It is the natural summary for Gaussian posteriors.

Maximum a posteriori (MAP) estimate: The mode of the posterior:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{d}) \tag{89}$$

With uniform priors, this reduces to the maximum likelihood estimate (MLE). The MAP estimate can be misleading for multimodal or skewed distributions, as it identifies only a single peak and ignores uncertainty.

Posterior median: For each parameter θ_i, the value that divides the marginal posterior into equal probability masses:

$$\int_{-\infty}^{\theta_{i,\text{med}}} p(\theta_i|\mathbf{d})d\theta_i = 0.5 \tag{90}$$

The median is robust to outliers and skewness, and is optimal for absolute error loss.

5.4.2 Credible Regions and Intervals

Point estimates alone do not convey uncertainty. Credible regions quantify the range of parameter values consistent with the data.

Definition: A $X\%$ credible region is any volume Ω in parameter space such that

$$\int_{\Omega} p(\boldsymbol{\theta}|\mathbf{d})d\boldsymbol{\theta} = \frac{X}{100}. \tag{91}$$

There is considerable freedom in choosing Ω. Two common choices are:

- **Highest Posterior Density (HPD) regions:** The smallest region containing $X\%$ of the posterior probability. All points inside have higher posterior density than points outside. For unimodal posteriors this is sensible, but for multimodal posteriors the HPD region may consist of several disconnected islands.
- **Central credible intervals:** For a single parameter, the central $(1-\alpha)$ credible interval $[a, b]$ satisfies:

$$\int_a^b p(\theta|\mathbf{d})d\theta = 1 - \alpha \tag{92}$$

with equal tail probabilities: $\int_{-\infty}^{a} p(\theta|\mathbf{d})d\theta = \int_b^{\infty} p(\theta|\mathbf{d})d\theta = \alpha/2$.

In 2D, credible regions are often shown as contour plots for $X = 68.3, 95.5, 99.7$ (corresponding to the probabilities enclosed in a Gaussian distribution by $\pm 1\sigma, 2\sigma, 3\sigma$), though the choice is arbitrary. Always identify what the contours represent when using them.

Credible regions are not confidence regions: The Bayesian credible region is not the same as the frequentist confidence interval. A 95% credible region means there is a 95% probability that the true parameter lies within it, given the data and model. A 95% confidence interval means that in 95% of repeated experiments, the interval would contain the true value—a statement about long-run frequency, not about probability given the observed data. The Bayesian interpretation directly addresses scientific questions about parameter values.

5.4.3 Variance and Covariance

Posterior variance and covariance: For continuous parameters, the posterior covariance matrix is:

$$\text{Cov}[\boldsymbol{\theta}|\mathbf{d}] = \mathbb{E}[(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T|\mathbf{d}] \tag{93}$$

The diagonal elements give the marginal variances, while off-diagonal elements capture correlations between parameters. Strong correlations indicate parameter degeneracies that can only be understood by examining the joint distribution.

5.4.4 Gaussian Posteriors and Chi-Squared Contours

With multivariate Gaussian posteriors, the contour levels that contain $X\%$ of the posterior can be calculated using the chi-squared distribution. For Gaussian likelihoods with independent measurement errors, the chi-squared statistic is:

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(d_i - m_i(\boldsymbol{\theta}))^2}{\sigma_i^2} \tag{94}$$

where d_i are observed data, $m_i(\boldsymbol{\theta})$ are model predictions, and σ_i are measurement uncertainties. For correlated errors with covariance matrix $\mathbf{C}$:

$$\chi^2(\boldsymbol{\theta}) = (\mathbf{d} - \mathbf{m}(\boldsymbol{\theta}))^T \mathbf{C}^{-1} (\mathbf{d} - \mathbf{m}(\boldsymbol{\theta})) \tag{95}$$

For a Gaussian likelihood, the log-likelihood is $\ln \mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2}\chi^2(\boldsymbol{\theta}) + \text{constant}$. The maximum likelihood estimate corresponds to the minimum value χ^2_{min}, and we define $\Delta\chi^2 \equiv \chi^2 - \chi^2_{\text{min}}$ to measure deviations from the best fit.

With uniform priors, the relative posterior probability is $\exp(-\Delta\chi^2/2)$. Standard references like Numerical Recipes provide tables for choosing $\Delta\chi^2$ contour levels that contain specified probabilities for different numbers of parameters. For example, for two parameters jointly, $\Delta\chi^2 = 2.30$ gives a 68.3% credible region, while for a single parameter, $\Delta\chi^2 = 1$ gives 68.3%.

Important: In the non-Gaussian case, it is better to find numerically the HPD regions that contain $X\%$ of the posterior, rather than adopting Gaussian chi-squared contour levels, which may be highly misleading.

5.4.5 When Are Summaries Sufficient?

For approximately Gaussian posteriors, the mean and covariance matrix provide a complete summary. However, simple summaries can be inadequate or misleading for:

- **Multimodal posteriors**: Point estimates can be meaningless, falling between modes or highlighting one mode while ignoring others. The full distribution must be presented, showing all modes and their relative probabilities.
- **Highly skewed posteriors**: The mean may lie in a low-probability region. The median and HPD intervals are more appropriate, but visualizing the full distribution is still recommended.
- **Complex parameter spaces**: Strong correlations and degeneracies require visualization. Corner plots (Figure 3) showing all pairwise marginal distributions are essential for understanding parameter relationships.
- **Decision-making**: Different loss functions require different summaries. For example, the mean is optimal for squared-error loss, the median for absolute-error loss, and the mode for zero-one loss. The appropriate summary depends on the consequences of errors.

In such cases, presenting samples from the posterior (e.g., from MCMC, section 6) or visualization tools like corner plots are preferable to simple numerical summaries. The full posterior should always be preserved for further analysis, even if summaries are reported for publication. Modern tools make it straightforward to share full MCMC chains or grid-based posterior representations alongside published results.

5.5 Grid-Based Posterior Computation

When the posterior distribution does not follow a closed-form analytical expression—which is the case for most real-world problems—numerical methods are required to compute and characterize the posterior. For problems with low to moderate dimensionality ($d \lesssim 3$ parameters), evaluating the posterior on a discrete grid in parameter space provides a direct and intuitive approach.

5.5.1 The Grid Method

In grid-based computation, we discretize the parameter space by defining a regular grid of points and evaluate the unnormalized posterior at each grid point. The approach consists of the following steps:

1. **Define parameter ranges:** For each parameter θ_j, specify a range $[\theta_{j,\min}, \theta_{j,\max}]$ wide enough to capture all significant posterior mass.
2. **Create the grid:** Discretize each parameter dimension with n equally-spaced points. For d parameters, this yields n^d total grid points.
3. **Evaluate the unnormalized posterior:** At each grid point $\boldsymbol{\theta}_i$, compute:
$$L_i = p(\mathbf{d}|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i) \tag{96}$$
4. **Normalize:** The posterior probability at grid point i becomes:
$$p_i = \frac{L_i}{\sum_{k=1}^{n^d} L_k} \tag{97}$$
5. **Compute summaries:** Use the discrete distribution $\{p_i\}$ to calculate marginal distributions, credible intervals, and other posterior summaries.

5.5.2 Advantages of Grid Methods

Grid-based approaches offer several benefits for appropriate problems:

- **Conceptual simplicity:** The method is intuitive and provides complete visualization of the posterior landscape.
- **Systematic exploration:** The entire specified parameter space is explored uniformly, ensuring no modes or features are missed.
- **Exact marginalization:** Marginal distributions are computed exactly (up to discretization) by summing over grid dimensions.
- **Deterministic results:** Unlike sampling methods, there are no convergence issues or random variations in results.
- **Parallel computation:** Likelihood evaluations are independent and easily parallelized.

5.5.3 The Curse of Dimensionality

Despite their simplicity, grid-based methods face a fundamental limitation: the **curse of dimensionality**. The total number of grid points scales as n^d, growing exponentially with the number of parameters.

For example, with $n = 100$ grid points per dimension:

- $d = 1$: 100 evaluations
- $d = 2$: $10,000$ evaluations
- $d = 3$: $1,000,000$ evaluations
- $d = 5$: 10^{10} evaluations (computationally challenging)
- $d = 10$: 10^{20} evaluations (computationally infeasible)

The computational cost explodes exponentially, making grid methods impractical for dimensionality $d \gtrsim 5$.

5.5.4 Additional Limitations

Beyond computational scaling, grid methods suffer from several inefficiencies:

Wasted computation: Most grid points typically lie in regions of negligible posterior probability, yet computational effort is spent equally everywhere.

Volume concentration: In high dimensions, probability mass concentrates in surprising ways. For example, in a d-dimensional unit hypercube, most volume lies near the boundaries rather than the center, making uniform exploration strategies inefficient.

The typical set: For smooth distributions in high dimensions, most probability mass concentrates not at the mode, but in a thin "typical set" surrounding it. This counter-intuitive geometry means that grid methods spend most effort in low-probability regions.

5.5.5 When to Use Grid Methods

Grid-based computation is most effective for:

- Problems with $d \leq 3$ parameters where full visualization is valuable
- Initial exploration to understand posterior structure
- Cases where complete characterization of multimodal posteriors is essential
- Validation of sampling-based methods for small problems

5.5.6 The Need for Sampling Methods

As dimensionality increases beyond $d \sim 3-5$, the exponential scaling of grid methods necessitates alternative approaches. **Monte Carlo sampling methods** (section 6) address these limitations by:

- **Adaptive exploration:** Sampling methods concentrate computational effort on high-probability regions rather than exploring uniformly.
- **Dimension-independent scaling:** Monte Carlo estimation error scales as $1/\sqrt{N_{\text{samples}}}$, independent of dimensionality, making high-dimensional inference tractable.
- **Efficient marginalization:** Marginal distributions are obtained by simply ignoring irrelevant parameters in the sample collection.

This fundamental shift from systematic grid evaluation to adaptive sampling enables Bayesian inference for realistic problems with dozens, hundreds, or even thousands of parameters.

5.6 Sampling Representation

This is to use a completely different representation of the posterior $p(\boldsymbol{\theta})$: a large number of samples drawn from the distribution, with (expected) density that is proportional to $p(\boldsymbol{\theta})$. This is usually in an ordered list, called a 'chain', of values of the parameters $\boldsymbol{\theta}$. The samples may also have a weight associated with them, and are constructed such that the expected weighted number density is proportional to the posterior. Note that we don't need to calculate the constant of proportionality (which can be expensive to do), since in parameter inference problems, the relative probability of parameters is given by the ratio of p.

The reason why the list is ordered is that the algorithms for generating the chain typically produce correlated samples, so the ordering is important (one might, for example, want to 'thin' the chain by selecting only separated samples, thus reducing the correlations. If the samples are correlated, then the 'effective sample size' is smaller than the length of the chain. subsection 14.1 provides formal definitions of autocorrelation, integrated autocorrelation time, and effective sample size.

The samples effectively replace the continuous density p by a (weighted) sum of Dirac delta functions:

$$p(\boldsymbol{\theta}) \simeq \frac{\sum_{s=1}^{S} w_s \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_s)}{\sum_{s=1}^{S} w_s}. \tag{98}$$

This is clearly crude for p itself, but for integrated quantities, it makes sense. e.g. an estimate of the expectation value is

$$\hat{\mu} = \langle \boldsymbol{\theta} \rangle = \int p(\boldsymbol{\theta}) \boldsymbol{\theta} d\boldsymbol{\theta} \simeq \int \frac{\sum_{s=1}^{S} w_s \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_s)}{\sum_{s=1}^{S} w_s} \boldsymbol{\theta} d\boldsymbol{\theta} = \frac{\sum_{s=1}^{S} w_s \boldsymbol{\theta}_s}{\sum_{s=1}^{S} w_s}. \tag{99}$$

This is generalised to any function $f(\boldsymbol{\theta})$,

$$\langle f(\boldsymbol{\theta}) \rangle = \int p(\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{\sum_{s=1}^{S} w_s f(\boldsymbol{\theta}_s)}{\sum_{s=1}^{S} w_s}. \tag{100}$$

This is Monte Carlo integration. For example, one might want the covariance of the distribution, whose elements are estimated by

$$\hat{\Sigma}_{ij} = \frac{\sum_{s=1}^{S} w_s (\theta_{i,s} - \hat{\mu}_i)(\theta_{j,s} - \hat{\mu}_j)}{\sum_{s=1}^{S} w_s}. \tag{101}$$

6 Sampling methods

There are several generic methods for generating samples. We will concentrate on three of the most common ones, highlighting when each of them can usefully be applied. They are:

- Metropolis-Hastings
- Gibbs Sampling
- Hamiltonian (or Hybrid) Monte Carlo (HMC)

Modern implementations like Stan use advanced variants such as the No-U-Turn Sampler (NUTS), which automatically tunes HMC trajectory length. See subsection 14.12 (Advanced HMC Variants). For high-dimensional problems where MCMC is too slow, alternatives include Variational Inference (subsection 14.2) and the Laplace Approximation (subsection 14.7).

First, though, some general remarks.

6.1 MCMC Basics

These are all examples of MCMC (Markov Chain Monte Carlo), where random steps are taken in parameter space, according to a proposal distribution. The goal is always to give a chain of samples of the target distribution (usually the posterior or the likelihood), with an expected number density proportional to the posterior. The target distribution need not be normalised, but it needs to be everywhere positive, and normalisable (i.e. the integral is finite).

6.1.1 Markov processes

Markov processes are sequential processes for which the new element depends only on the previous element, and not on any previous ones. In MCMC, the next point in the chain depends only on the parameters (and the target value) of the previous point. For a rigorous theoretical foundation of why MCMC algorithms work, including detailed balance, ergodicity, and mixing properties, see subsection 14.1 (Markov Chain Theory for MCMC).

The general algorithm is as follows:

- Choose a starting point $\boldsymbol{\theta}_0$. No general rule here, but (see later) there are advantages in having a 'dispersed' starting point, which is not near the peak of the target distribution (see Convergence Tests later). A random point drawn from a prior distribution is common.
- Subsequent points $\boldsymbol{\theta}_{s+1}$ are generated from $\boldsymbol{\theta}_s$ by generating a trial point through some random process, and which is either accepted or rejected (depending on the algorithm)
- If the trial point is accepted, it becomes the next point in the chain. If it is rejected, the previous sample is repeated in the chain (or equivalently, its weight is increased from 1 to 2 (and can go higher if subsequent trials are also rejected).
- The chain is stopped at some point. There is no magic answer as to when to stop, but the main idea is convergence, which we will cover later.

6.1.2 Detailed balance

If the sampling procedure satisfies detailed balance, the expected number density to be proportional to the target distribution $p(\boldsymbol{\theta})$, which is what we desire. We don't want the target distribution ρ to evolve as the chain develops, so it is a stationary distribution. In Bayesian inference problems, the target is sometimes the posterior, sometimes the likelihood, and it can be something different again.

Let us assume there is a discrete set of parameters (the argument generalises to continuous parameters), labelled by an index (it can still be a label in a multi-dimensional parameter space). As we move from one

sample to the next in the chain, there is a probability that the state shifts from i to j given by P_{ij}. The MCMC chain satisfies detailed balance if

$$\rho_i P_{ij} = \rho_j P_{ji}. \tag{102}$$

One can think of the left hand side as being the flux of probability flowing from i to j, and the r.h.s. from j to i. If they balance, the chain is stationary.

Detailed balance is a stronger condition than that required to give a stationary distribution (which can be achieved via a more complicated route).

Proof: if we have samples drawn from a density distribution ρ_i, then after a transition, the probability distribution changes to an expected value ρ_j given by

$$\sum_i \rho_i P_{ij} \tag{103}$$

including all the routes to populate j from the other states i. If detailed balance is satisfied, this is $\sum_i \rho_j P_{ji} = \rho_j \sum_i P_{ji} = \rho_j$, since, in the last step, the state j must end up in some i, so the sum of probabilities is 1. So the expected density stays as ρ and does not change.

6.2 Metropolis-Hastings

This is perhaps the most common form of MCMC, and is suitable for relatively low-dimensional problems (perhaps up to 5 or 10). We define a proposal distribution to generate a new proposed sample, which is either accepted or rejected.

$$q(\boldsymbol{\theta}'|\boldsymbol{\theta}) \tag{104}$$

= probability of a proposed sample at $\boldsymbol{\theta}'$ from a previous state $\boldsymbol{\theta}$. Typically this is a function of $\boldsymbol{\theta}' - \boldsymbol{\theta}$, but it doesn't have to be, and a common choice is a gaussian centred on the previous sample in the chain.

The algorithm specifies that the point is accepted with probability

$$\alpha = \min\left[1, \frac{\rho(\boldsymbol{\theta}')q(\boldsymbol{\theta}'|\boldsymbol{\theta})}{\rho(\boldsymbol{\theta})q(\boldsymbol{\theta}|\boldsymbol{\theta}')}\right]. \tag{105}$$

Let us see if this satisfies detailed balance. Let $\boldsymbol{\theta}$ be labelled by i, $\boldsymbol{\theta}'$ by j. For concreteness, let us assume that

$$\rho_j q_{ji} \leq \rho_i q_{ij} \tag{106}$$

The probability of an accepted transition from i to j is

$$P_{ij} = q_{ij} \min\left[1, \frac{\rho_j q_{ji}}{\rho_i q_{ij}}\right] = \frac{\rho_j q_{ji}}{\rho_i} \tag{107}$$

where the first term is the probability that the transition is proposed, and the second is the probability that it is accepted. The reverse probability is

$$P_{ji} = q_{ji} \min\left[1, \frac{\rho_i q_{ij}}{\rho_j q_{ji}}\right] = q_{ji} \tag{108}$$

since the proposed sample is accepted with probability 1 in this case. Hence the detailed balance relation is satisfied with Metropolis-Hastings. Note that if q is symmetric, (i.e. $q_{ij} = q_{ji}$), the acceptance probability is simplified, and the algorithm is called Metropolis.

Remember! If the proposed sample is rejected, the previous sample is repeated in the chain (or equivalently, its weight is increased from 1 to 2 (and to 3 if the next proposed sample is also rejected, and so on).

6.2.1 Choice of proposal distribution

The performance of Metropolis-Hastings depends critically on the choice of proposal distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$. The most common choice is a Gaussian centered on the current state:

$$q(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}'; \boldsymbol{\theta}, \boldsymbol{\Sigma}) \tag{109}$$

where $\boldsymbol{\Sigma}$ is the proposal covariance matrix. This is symmetric, so the acceptance probability simplifies to the Metropolis form: $\alpha = \min[1, \rho(\boldsymbol{\theta}')/\rho(\boldsymbol{\theta})]$.

The key trade-off is between exploration and acceptance rate:

Proposal too narrow (small $\boldsymbol{\Sigma}$): The chain takes small steps, leading to high acceptance rates ($\alpha \approx 1$) but slow exploration of parameter space. The chain exhibits strong autocorrelation and requires many samples to effectively cover the posterior.

Proposal too wide (large $\boldsymbol{\Sigma}$): The chain proposes large jumps that frequently land in low-probability regions, resulting in low acceptance rates. The chain becomes stuck, repeatedly rejecting proposals and duplicating the current state.

Optimal tuning: For high-dimensional problems, theoretical results suggest that an acceptance rate of approximately 23% is optimal. For low-dimensional problems (1-2 parameters), a higher acceptance rate of around 50% is often more efficient. In practice, adaptive MCMC schemes adjust $\boldsymbol{\Sigma}$ during a tuning phase to achieve target acceptance rates.

The proposal covariance should ideally match the posterior covariance structure. If parameters have very different scales or are strongly correlated, using $\boldsymbol{\Sigma} \propto \text{Cov}(\boldsymbol{\theta}|\mathbf{d})^{-1}$ (when approximately known from pilot runs) can dramatically improve efficiency. For multimodal posteriors, more sophisticated proposals such as mixture models or adaptive schemes may be necessary.

6.2.2 Burn-in

For convergence tests (see later) it is often necessary to run two or more chains, with 'dispersed' starting points (i.e. not similar). Each chain may take some time to find the region(s) where the target distribution is high. This exploratory phase is called burn-in and these samples are discarded. There is no golden rule about how many samples to throw away, but one common and useful technique is to find the first sample which is within some factor (say 0.1) of the highest value of the target in the entire chain, and discard all the previous samples.

6.3 Sample Marginalization

This is trivial to do. Each sample has values for all of the parameters. If we want the distribution of θ_1 say, then we simply ignore the values of $\theta_i, i > 1$ in the chain, and plot the distribution of θ_1. A potentially conceptually hard multidimensional integral is solved very easily.

6.4 Correlated samples

Some sampling algorithms will produce correlated samples from the posterior (in fact this is normal). If nearby samples in the chain are correlated, the effective number of independent samples is smaller than the total number of samples. We can quantify this with the autocorrelation function, estimated by

$$\hat{C}_\Delta \equiv \frac{1}{S-\Delta} \sum_{s=1}^{S-\Delta} \frac{(\theta_s - \hat{\mu})(\theta_{s+\Delta} - \hat{\mu})}{\hat{\Sigma}} \tag{110}$$

where $\hat{\mu}$ is the estimate of the mean parameter (in practice, just the weighted average), and $\hat{\Sigma}$ is the estimated variance. We compute this for every parameter in the problem. Note that $\hat{C}_0 = 1$, and ideally we'd like $\hat{C}_\Delta$ to be zero otherwise.

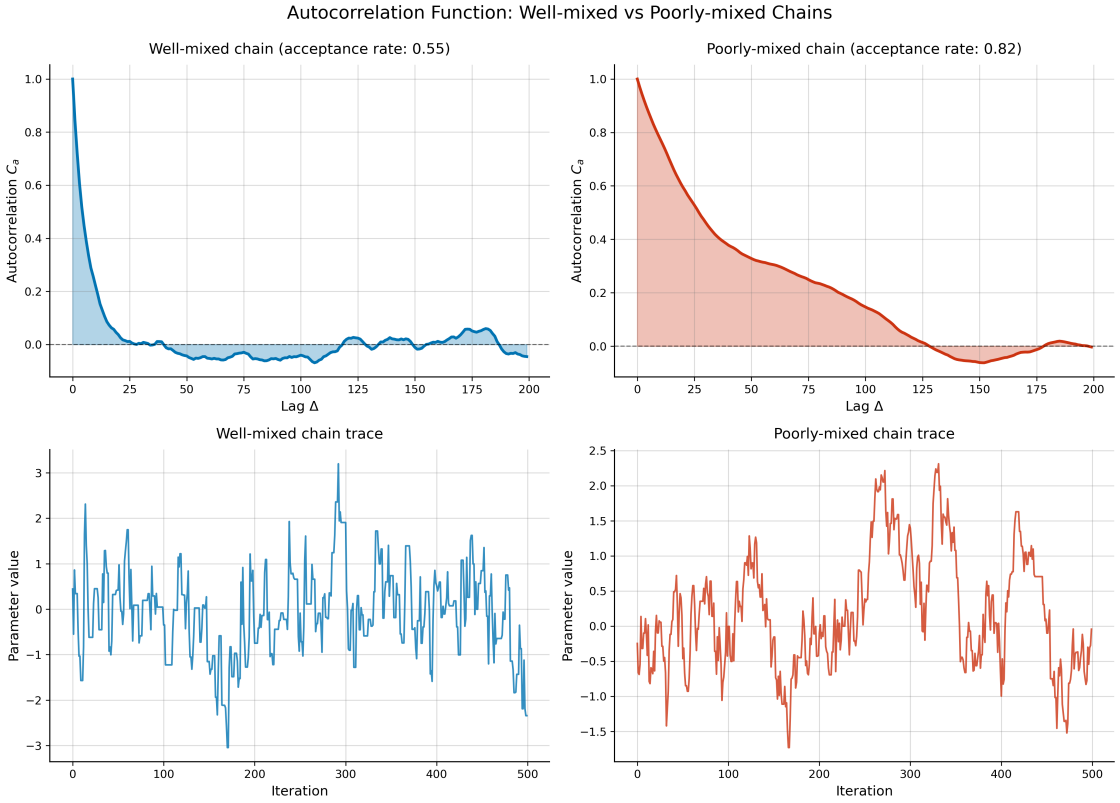

Figure 7: MCMC Chain Diagnostics (section 6): Top panels show autocorrelation functions for well-mixed (left, blue) and poorly-mixed (right, red) chains. The well-mixed chain quickly decorrelates, while the poorly-mixed chain shows persistent autocorrelation, indicating inefficient sampling. Bottom panels show the corresponding chain traces over 500 iterations, with the well-mixed chain exploring the parameter space efficiently compared to the slowly-moving poorly-mixed chain.

6.4.1 Effective sample size

The effective number of independent samples will be smaller than S if the chain is correlated. One definition of the effective sample size is

$$S_{\text{eff}} \equiv \frac{N}{1 + 2\sum_{\Delta=1}^{\Delta_0 - 1} \hat{C}_\Delta}, \tag{111}$$

where N is the total number of samples, and Δ_0 is the point where $\hat{C}_\Delta$ crosses zero for the first time.

6.5 Gibbs sampling

Gibbs sampling is a powerful MCMC technique that exploits the conditional structure of the posterior distribution. Unlike Metropolis-Hastings, which proposes moves in the full parameter space and accepts or rejects them, Gibbs sampling updates each parameter (or block of parameters) by sampling from its conditional distribution given all other parameters. When these conditional distributions can be sampled efficiently, Gibbs sampling has a 100% acceptance rate and can be very effective.

6.5.1 Algorithm

For parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$, the basic Gibbs sampling algorithm iterates through the following steps:

$$\theta_1^{(s+1)} \sim p(\theta_1 | \theta_2^{(s)}, \theta_3^{(s)}, \ldots, \theta_n^{(s)}, \mathbf{d}) \tag{112}$$

$$\theta_2^{(s+1)} \sim p(\theta_2 | \theta_1^{(s+1)}, \theta_3^{(s)}, \ldots, \theta_n^{(s)}, \mathbf{d}) \tag{113}$$

$$\theta_3^{(s+1)} \sim p(\theta_3 | \theta_1^{(s+1)}, \theta_2^{(s+1)}, \theta_4^{(s)}, \ldots, \theta_n^{(s)}, \mathbf{d}) \tag{114}$$

$$\vdots \tag{115}$$

$$\theta_n^{(s+1)} \sim p(\theta_n | \theta_1^{(s+1)}, \theta_2^{(s+1)}, \ldots, \theta_{n-1}^{(s+1)}, \mathbf{d}) \tag{116}$$

Note that each conditional distribution uses the most recently updated values of the other parameters. The order of updates can be randomized or systematically varied to improve mixing.

6.5.2 Proof of Correctness

To prove that Gibbs sampling correctly samples from the target distribution $p(\boldsymbol{\theta}|\mathbf{d})$, we need to show that the target distribution is the stationary distribution of the Markov chain defined by the Gibbs sampler. We demonstrate this by proving that the target distribution satisfies the **detailed balance condition**.

Detailed Balance Condition: A Markov chain with transition kernel $T(\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta})$ has stationary distribution $\pi(\boldsymbol{\theta})$ if:

$$\pi(\boldsymbol{\theta}) T(\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta}) = \pi(\boldsymbol{\theta}') T(\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}') \tag{117}$$

For Gibbs sampling, we need to show that $p(\boldsymbol{\theta}|\mathbf{d})$ satisfies detailed balance under the Gibbs transition kernel.

Setup: Consider the Gibbs sampler for two parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$ (the proof generalizes straightforwardly to n parameters). A complete Gibbs step consists of:

1. Update θ_1: $\theta_1' \sim p(\theta_1 | \theta_2, \mathbf{d})$
2. Update θ_2: $\theta_2' \sim p(\theta_2 | \theta_1', \mathbf{d})$

The transition from (θ_1, θ_2) to (θ_1', θ_2') has transition density:

$$T((\theta_1', \theta_2') \leftarrow (\theta_1, \theta_2)) = p(\theta_1' | \theta_2, \mathbf{d}) \cdot p(\theta_2' | \theta_1', \mathbf{d}) \tag{118}$$

Proof: We need to verify:

$$p(\theta_1, \theta_2 | \mathbf{d}) \cdot T((\theta_1', \theta_2') \leftarrow (\theta_1, \theta_2)) = p(\theta_1', \theta_2' | \mathbf{d}) \cdot T((\theta_1, \theta_2) \leftarrow (\theta_1', \theta_2')) \tag{119}$$

Substituting the transition kernels:

$$p(\theta_1, \theta_2|\mathbf{d}) \cdot p(\theta_1'|\theta_2, \mathbf{d}) \cdot p(\theta_2'|\theta_1', \mathbf{d}) \tag{120}$$
$$= p(\theta_1', \theta_2'|\mathbf{d}) \cdot p(\theta_1|\theta_2', \mathbf{d}) \cdot p(\theta_2|\theta_1, \mathbf{d}) \tag{121}$$

Using the factorization of the joint distribution:

$$p(\theta_1, \theta_2|\mathbf{d}) = p(\theta_1|\theta_2, \mathbf{d}) \cdot p(\theta_2|\mathbf{d}) \tag{122}$$

The left-hand side becomes:

$$p(\theta_1|\theta_2, \mathbf{d}) \cdot p(\theta_2|\mathbf{d}) \cdot p(\theta_1'|\theta_2, \mathbf{d}) \cdot p(\theta_2'|\theta_1', \mathbf{d}) \tag{123}$$

Similarly, the right-hand side becomes:

$$p(\theta_1'|\theta_2', \mathbf{d}) \cdot p(\theta_2'|\mathbf{d}) \cdot p(\theta_1|\theta_2', \mathbf{d}) \cdot p(\theta_2|\theta_1, \mathbf{d}) \tag{124}$$

Rearranging terms, we need to show:

$$p(\theta_2|\mathbf{d}) \cdot p(\theta_1'|\theta_2, \mathbf{d}) = p(\theta_2'|\mathbf{d}) \cdot p(\theta_1|\theta_2', \mathbf{d}) \tag{125}$$

Using the chain rule again:

$$p(\theta_2|\mathbf{d}) \cdot p(\theta_1'|\theta_2, \mathbf{d}) = p(\theta_1', \theta_2|\mathbf{d}) \tag{126}$$
$$p(\theta_2'|\mathbf{d}) \cdot p(\theta_1|\theta_2', \mathbf{d}) = p(\theta_1, \theta_2'|\mathbf{d}) \tag{127}$$

However, this approach shows the conditions under which detailed balance holds, but we can provide a more direct proof using **reversibility**.

Alternative Proof via Reversibility:

For Gibbs sampling, we can prove correctness more directly by showing that each conditional update preserves the target distribution. Consider updating θ_1 while keeping θ_2 fixed.

Starting from any distribution $\pi(\theta_1, \theta_2)$, after updating θ_1 according to $p(\theta_1|\theta_2, \mathbf{d})$, the marginal distribution of θ_2 remains unchanged:

$$\int \pi(\theta_1, \theta_2) p(\theta_1'|\theta_2, \mathbf{d}) d\theta_1 = \pi(\theta_2) \int p(\theta_1'|\theta_2, \mathbf{d}) d\theta_1' = \pi(\theta_2) \tag{128}$$

If we start with $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{d})$, then:

$$\pi(\theta_2) = p(\theta_2|\mathbf{d}) = \int p(\theta_1, \theta_2|\mathbf{d}) d\theta_1 \tag{129}$$

After the Gibbs update of θ_1, the joint distribution becomes:

$$p(\theta_1'|\theta_2, \mathbf{d}) \cdot p(\theta_2|\mathbf{d}) = p(\theta_1', \theta_2|\mathbf{d}) \tag{130}$$

This shows that each Gibbs step preserves the target distribution, proving that $p(\boldsymbol{\theta}|\mathbf{d})$ is indeed the stationary distribution of the Gibbs sampler.

Ergodicity: For the Markov chain to converge to the stationary distribution from any starting point, we also need irreducibility and aperiodicity. Gibbs sampling is typically irreducible when the target distribution has connected support and all conditional distributions are absolutely continuous with respect to Lebesgue measure. Aperiodicity is automatically satisfied when the parameter space is continuous.

This proof extends straightforwardly to the n-parameter case and to block Gibbs sampling, where entire blocks of parameters are updated jointly from their conditional distributions.

6.5.3 Block Gibbs sampling

For high-dimensional problems, parameters can be partitioned into blocks $\boldsymbol{\theta} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_K)$ where the conditional distribution of each block given the others is tractable. The algorithm then samples entire blocks:

$$\boldsymbol{\phi}_1^{(s+1)} \sim p(\boldsymbol{\phi}_1 | \boldsymbol{\phi}_2^{(s)}, \ldots, \boldsymbol{\phi}_K^{(s)}, \mathbf{d}) \tag{131}$$

$$\boldsymbol{\phi}_2^{(s+1)} \sim p(\boldsymbol{\phi}_2 | \boldsymbol{\phi}_1^{(s+1)}, \boldsymbol{\phi}_3^{(s)}, \ldots, \boldsymbol{\phi}_K^{(s)}, \mathbf{d}) \tag{132}$$

$$\vdots \tag{133}$$

Block Gibbs sampling is particularly effective when:

- Parameters within a block are strongly correlated (updating them jointly reduces autocorrelation)
- The conditional distribution of a block has a standard form (e.g., multivariate Gaussian, Dirichlet)
- Some parameters are conditionally independent given others

The choice of blocking structure can dramatically affect convergence. Grouping highly correlated parameters into blocks typically improves efficiency, though this must be balanced against the computational cost of sampling from higher-dimensional conditional distributions.

6.5.4 When Gibbs sampling works well

Gibbs sampling can be applied to very high-dimensional problems (even millions of parameters in some applications). It is particularly useful for:

- **Conjugate priors**: When the prior and likelihood are conjugate, conditional distributions are often standard distributions that can be sampled directly
- **Hierarchical models** (section 9): Natural conditional structure often emerges in hierarchical models
- **Missing data problems**: Latent variables can be sampled as additional parameters
- **Mixture models**: Component assignments can be sampled conditionally on mixture parameters

When conditional distributions cannot be sampled directly, one can often employ rejection sampling, slice sampling, or embed a Metropolis-Hastings step within the Gibbs sampler (Metropolis-within-Gibbs).

6.5.5 Limitations and convergence issues

Despite its advantages, Gibbs sampling has important limitations:

Strong correlations and degeneracies: When parameters are highly correlated in the posterior, Gibbs sampling can be extremely slow to converge. Consider two parameters θ_1 and θ_2 with strong posterior correlation. Gibbs sampling moves along the conditional distributions, which are perpendicular to the parameter axes. If the posterior has strong correlation (e.g., an elongated ellipse), the sampler takes many small steps to move along the principal direction of the posterior, resulting in slow exploration and high autocorrelation.

Exact degeneracies: In cases of exact parameter degeneracy (e.g., $\theta_1 + \theta_2 = c$ for some constant c), the conditional distributions may be improper or the sampler may fail to explore the degenerate manifold efficiently. Such situations require reparameterization or alternative sampling strategies.

Unknown conditional distributions: Gibbs sampling requires knowing the conditional distributions analytically or being able to sample from them efficiently. When this is not possible, one must resort to Metropolis-Hastings or Hamiltonian Monte Carlo.

Systematic scan effects: The order in which parameters are updated can affect mixing, especially with strong dependencies. Random or adaptive ordering schemes can help mitigate this.

For posteriors with strong correlations, Hamiltonian Monte Carlo (section 6) or reparameterization to decorrelate parameters often provides superior performance.

Example: Gibbs Sampling: Linear Regression with Errors in Both Variables

Consider a parameter inference problem that illustrates the power of Gibbs sampling for problems with latent variables. Suppose there is a set of data pairs $(\hat{x}, \hat{y})$ (for simplicity, consider just one pair) representing noisy measurements of true values (x, y) that are related by a linear model $y = mx$. The model parameters are the slope m, and the measurement errors are assumed to be Gaussian and independent in both coordinates. A directed acyclic graph (DAG) representing this problem is shown in Figure 8.

Applying Rule 1, the quantity of interest is the posterior distribution of the slope given the observed data:

$$p(m|\hat{x}, \hat{y}) \tag{134}$$

(strictly speaking, this is also conditional on knowing the error distributions for $\hat{x}$ and $\hat{y}$, but this dependence is suppressed for clarity).

The key challenge is that there are extra unknowns in this problem: the true (unobserved) values x and y are latent variables. The model connects these true variables via $y = mx$, not the observed quantities $\hat{x}$ and $\hat{y}$. The latent variables x and y are nuisance parameters that must ultimately be marginalized over to obtain the posterior for m.

6.5.6 Analytical solution

Assume the sampling distribution of the observed quantities is known:

$$p(\hat{x}, \hat{y}|x, y) = p(\hat{x}|x)p(\hat{y}|y) \tag{135}$$

where the equality holds because the measurement errors are independent. Applying Bayes' theorem:

$$p(m|\hat{x}, \hat{y}) = \frac{p(\hat{x}, \hat{y}|m)p(m)}{p(\hat{x}, \hat{y})} \propto p(\hat{x}, \hat{y}|m)p(m) \tag{136}$$

Introduce the latent variables x and y, writing the likelihood as a marginal integral:

$$p(m|\hat{x}, \hat{y}) \propto \int p(\hat{x}, \hat{y}, x, y|m)p(m)dx\,dy \tag{137}$$

Using the product rule:

$$p(m|\hat{x}, \hat{y}) \propto \int p(\hat{x}, \hat{y}|x, y, m)p(x, y|m)p(m)dx\,dy \tag{138}$$

The measurement errors are independent of the model parameter m, so $p(\hat{x}, \hat{y}|x, y, m) = p(\hat{x}, \hat{y}|x, y)$. The product rule gives $p(x, y|m) = p(y|x, m)p(x|m)$. Since the model is deterministic, $p(y|x, m) = \delta(y - mx)$. Assuming the prior on x is independent of m gives $p(x|m) = p(x)$. Combining these:

$$p(m|\hat{x}, \hat{y}) \propto \int p(\hat{x}, \hat{y}|x, y)p(y|x, m)p(x)p(m)dx\,dy \tag{139}$$

$$\propto \int p(\hat{x}, \hat{y}|x, y)\delta(y - mx)p(x)p(m)dx\,dy \tag{140}$$

The integration over y is trivial due to the Dirac delta function:

$$p(m|\hat{x}, \hat{y}) \propto \int p(\hat{x}, \hat{y}|x, mx)p(x)p(m)dx \tag{141}$$

Assuming the errors in $\hat{x}$ and $\hat{y}$ are independent Gaussians with unit variance, and adopting uniform priors for x and m:

$$p(m|\hat{x}, \hat{y}) \propto \int e^{-\frac{1}{2}(\hat{x}-x)^2} e^{-\frac{1}{2}(\hat{y}-mx)^2} dx \tag{142}$$

Completing the square and integrating yields:

$$p(m|\hat{x}, \hat{y}) \propto \frac{1}{\sqrt{1+m^2}} \exp\left(-\frac{(-m\hat{x}+\hat{y})^2}{2(1+m^2)}\right) \tag{143}$$

6.5.7 Results

The posterior has been marginalized analytically over x. If desired, the joint distribution of x and m can also be investigated:

$$p(x, m|\hat{x}, \hat{y}) \propto p(\hat{x}, \hat{y}|x, mx)p(x)p(m) \propto e^{-\frac{1}{2}(\hat{x}-x)^2} e^{-\frac{1}{2}(\hat{y}-mx)^2} \tag{144}$$

6.5.8 Gibbs Sampling

The same problem can be solved using Gibbs sampling. At fixed x, the conditional distribution of m given x is (suppressing the explicit conditioning on the data $\hat{x}, \hat{y}$ for clarity):

$$p(m|x) \propto \exp\left[-\frac{(\hat{y} - mx)^2}{2}\right] \propto \exp\left[-\frac{x^2}{2}\left(m - \frac{\hat{y}}{x}\right)^2\right] \tag{145}$$

which is a Gaussian distribution in m:

$$p(m|x, \hat{x}, \hat{y}) \sim \mathcal{N}\left(\frac{\hat{y}}{x}, \frac{1}{x^2}\right) \tag{146}$$

Similarly, the conditional distribution of x given m is:

$$p(x|m, \hat{x}, \hat{y}) \propto \exp\left[-\frac{(\hat{x} - x)^2}{2} - \frac{(\hat{y} - mx)^2}{2}\right] \tag{147}$$

which, after completing the square, is also a Gaussian distribution (in x):

$$p(x|m, \hat{x}, \hat{y}) \sim \mathcal{N}\left(\frac{\hat{x} + \hat{y}m}{1 + m^2}, \frac{1}{1 + m^2}\right) \tag{148}$$

The Gibbs sampling algorithm alternates between sampling from these two conditional distributions to generate samples from $p(m, x|\hat{x}, \hat{y})$. Marginalization over x is achieved trivially by ignoring the sampled values of x and constructing a histogram or kernel density estimate from the m values alone. This problem could equally well be solved using Metropolis-Hastings or Hamiltonian Monte Carlo, but Gibbs sampling is particularly efficient here because the conditional distributions are known analytically and can be sampled directly.

6.6 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) is a powerful MCMC technique that exploits gradient information to efficiently explore high-dimensional posterior distributions. Unlike random-walk Metropolis-Hastings, HMC uses the gradient of the log-posterior to guide proposals toward regions of high probability, enabling distant moves that maintain high acceptance rates. The primary requirement is that derivatives of the target distribution with respect to the model parameters must be computable, either analytically or through automatic differentiation.

The fundamental insight of HMC is to treat the negative log-posterior as a potential energy function and introduce auxiliary momentum variables to define Hamiltonian dynamics on an extended phase space. Proposals are generated by simulating these dynamics, allowing the sampler to move efficiently along the typical set of the posterior while avoiding the random walk behavior that plagues standard MCMC in high dimensions.

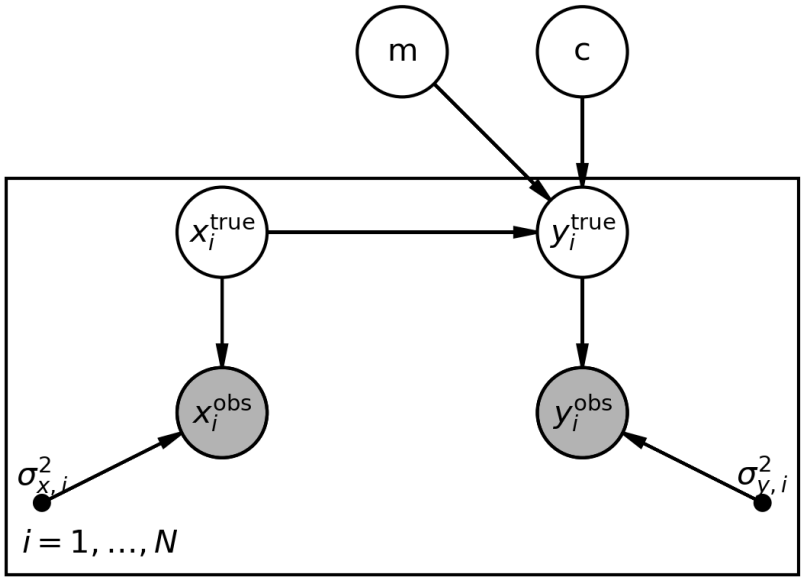

Figure 8: DAG for line fitting with errors in both coordinates. The parameter m (slope), latent variables x and y (true values), and measurement error variances determine the observed data $\hat{x}$ and $\hat{y}$ (shaded circles). The diamond shape indicates that y is a deterministic function $y = mx$.

6.6.1 Hamiltonian dynamics

HMC defines a potential energy corresponding to the negative log-posterior:

$$U(\boldsymbol{\theta}) = -\ln p(\boldsymbol{\theta}|\mathbf{d}) \tag{149}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ represents the position in parameter space. Auxiliary momentum variables $\mathbf{u} = (u_1, \ldots, u_n)$ are introduced with a kinetic energy:

$$K(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T\mathbf{M}^{-1}\mathbf{u} \tag{150}$$

where $\mathbf{M}$ is a mass matrix (often taken to be the identity). The Hamiltonian (total energy) is:

$$H(\boldsymbol{\theta}, \mathbf{u}) = U(\boldsymbol{\theta}) + K(\mathbf{u}) \tag{151}$$

This construction defines a joint distribution in the extended $2n$-dimensional phase space:

$$\pi(\boldsymbol{\theta}, \mathbf{u}) \propto \exp[-H(\boldsymbol{\theta}, \mathbf{u})] = p(\boldsymbol{\theta}|\mathbf{d}) \cdot \mathcal{N}(\mathbf{u}; \mathbf{0}, \mathbf{M}) \tag{152}$$

Since the momentum variables are independent of the parameters, marginalizing over $\mathbf{u}$ recovers the target posterior $p(\boldsymbol{\theta}|\mathbf{d})$.

Proposals are generated by simulating Hamiltonian dynamics via Hamilton's equations:

$$\frac{d\theta_i}{dt} = \frac{\partial H}{\partial u_i} = [\mathbf{M}^{-1}\mathbf{u}]_i \tag{153}$$

$$\frac{du_i}{dt} = -\frac{\partial H}{\partial \theta_i} = \frac{\partial \ln p(\boldsymbol{\theta}|\mathbf{d})}{\partial \theta_i} \tag{154}$$

These equations must be solved numerically using a symplectic integrator that preserves volume in phase space and is time-reversible, both required for detailed balance. The standard choice is the leapfrog integrator.

6.6.2 Leapfrog integrator

The leapfrog algorithm is a symplectic integrator that is time-reversible and volume-preserving, properties essential for satisfying detailed balance:

$$u_i\left(t+\frac{\epsilon}{2}\right) = u_i(t) - \frac{\epsilon}{2}\left(\frac{\partial U}{\partial \theta_i}\right)_{\boldsymbol{\theta}(t)} \tag{155}$$

$$\theta_i(t+\epsilon) = \theta_i(t) + \epsilon \cdot [\mathbf{M}^{-1}\mathbf{u}(t+\epsilon/2)]_i \tag{156}$$

$$u_i(t+\epsilon) = u_i\left(t+\frac{\epsilon}{2}\right) - \frac{\epsilon}{2}\left(\frac{\partial U}{\partial \theta_i}\right)_{\boldsymbol{\theta}(t+\epsilon)} \tag{157}$$

where ϵ is the step size. The half-step momentum update, followed by a full-step position update, followed by another half-step momentum update, ensures time-reversibility.

6.6.3 HMC algorithm

The complete HMC algorithm proceeds as follows:

Algorithm: Hamiltonian Monte Carlo

Input: Initial state $\boldsymbol{\theta}^{(0)}$, number of samples N_{samples}, trajectory length L, step size ϵ
Output: Chain of samples $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N_{\text{samples}})}\}$

for $i = 1$ **to** N_{samples} **do**
  Sample momentum: $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$
  Set $(\boldsymbol{\theta}^*, \mathbf{u}^*) \leftarrow (\boldsymbol{\theta}^{(i-1)}, \mathbf{u})$
  for $j = 1$ **to** L **do**
    Apply leapfrog step with step size ϵ:
      $\mathbf{u}^* \leftarrow \mathbf{u}^* - (\epsilon/2)\nabla U(\boldsymbol{\theta}^*)$
      $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}^* + \epsilon \mathbf{M}^{-1}\mathbf{u}^*$
      $\mathbf{u}^* \leftarrow \mathbf{u}^* - (\epsilon/2)\nabla U(\boldsymbol{\theta}^*)$
  end for
  Compute acceptance probability:
    $\alpha = \min\left\{1, \exp[-(H(\boldsymbol{\theta}^*, \mathbf{u}^*) - H(\boldsymbol{\theta}^{(i-1)}, \mathbf{u}))]\right\}$
  Draw $r \sim \text{Uniform}(0, 1)$
  if $r < \alpha$ **then**
    $\boldsymbol{\theta}^{(i)} \leftarrow \boldsymbol{\theta}^*$ (accept)
  else
    $\boldsymbol{\theta}^{(i)} \leftarrow \boldsymbol{\theta}^{(i-1)}$ (reject)
  end if
end for

6.6.4 Practical considerations

The key tuning parameters are the step size ϵ and the number of leapfrog steps L. Smaller step sizes yield more accurate Hamiltonian dynamics, reducing energy errors and increasing acceptance rates, but require more gradient evaluations per sample. The product $L\epsilon$ determines the trajectory length: longer trajectories explore the posterior more efficiently but are more computationally expensive. In practice, an acceptance rate around 65-70% is often optimal, though this depends on the problem dimensionality.

Automatic differentiation frameworks such as JAX, PyTorch, and Stan enable HMC without manual gradient derivation. Modern implementations employ adaptive schemes (e.g., the No-U-Turn Sampler in subsection 14.12) that automatically tune ϵ and L during a warmup phase, making HMC applicable to a wide range of problems without extensive manual tuning.

The efficiency of HMC stems from the conservation of energy along Hamiltonian trajectories. If the numerical integration were exact, the Hamiltonian H would be perfectly conserved, the trajectory would

lie on a constant-energy surface, and all proposals would be accepted. Numerical errors introduce energy fluctuations, which the Metropolis-Hastings acceptance step corrects, ensuring detailed balance is satisfied despite discretization errors.

7 Importance Sampling

Importance sampling provides an alternative to Markov Chain Monte Carlo for computing expectations under a target distribution. Rather than drawing samples directly from the posterior $p(\boldsymbol{\theta}|\mathbf{d})$, importance sampling draws samples from a proposal distribution $q(\boldsymbol{\theta})$ and reweights them to correct for the mismatch between the proposal and the target.

7.1 The Importance Sampling Principle

Suppose we wish to estimate the expectation $\mathbb{E}_p[f(\boldsymbol{\theta})]$ where $p(\boldsymbol{\theta})$ is the target distribution (typically the posterior), but we cannot sample from p directly or sampling from p is computationally expensive. The key observation is that we can rewrite the expectation as:

$$\mathbb{E}_p[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \int f(\boldsymbol{\theta})\frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}q(\boldsymbol{\theta})d\boldsymbol{\theta} = \mathbb{E}_q\left[f(\boldsymbol{\theta})w(\boldsymbol{\theta})\right] \tag{158}$$

where $q(\boldsymbol{\theta})$ is a proposal distribution from which we can easily sample, and the importance weight is:

$$w(\boldsymbol{\theta}) = \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \tag{159}$$

Given N samples $\boldsymbol{\theta}_i \sim q(\boldsymbol{\theta})$, the Monte Carlo estimator is:

$$\mathbb{E}_p[f(\boldsymbol{\theta})] \approx \frac{1}{N}\sum_{i=1}^{N} w_i f(\boldsymbol{\theta}_i) \tag{160}$$

where $w_i = p(\boldsymbol{\theta}_i)/q(\boldsymbol{\theta}_i)$.

7.2 Self-Normalized Importance Sampling

In Bayesian inference, the posterior $p(\boldsymbol{\theta}|\mathbf{d}) = p(\mathbf{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/p(\mathbf{d})$ involves the evidence $p(\mathbf{d})$, which is typically unknown or expensive to compute. Fortunately, we can use self-normalized importance sampling, which requires only unnormalized weights:

$$\mathbb{E}_p[f(\boldsymbol{\theta})] \approx \frac{\sum_{i=1}^{N} w_i f(\boldsymbol{\theta}_i)}{\sum_{i=1}^{N} w_i} \tag{161}$$

where now $w_i = p(\mathbf{d}|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)/q(\boldsymbol{\theta}_i)$, and the unknown constant $p(\mathbf{d})$ cancels in the ratio.

This self-normalized estimator is particularly valuable in Bayesian inference because it eliminates the need to compute the evidence, which is often the most computationally challenging aspect of posterior inference.

7.3 Conditions for Success

The performance of importance sampling depends critically on the choice of proposal distribution $q(\boldsymbol{\theta})$. The key requirement is that q must have support wherever $p(\boldsymbol{\theta})f(\boldsymbol{\theta}) \neq 0$. More specifically:

Good proposal distributions: The ideal proposal has $q(\boldsymbol{\theta}) \propto |f(\boldsymbol{\theta})|p(\boldsymbol{\theta})$, which yields zero variance in the estimate. While this ideal is generally unattainable (it requires knowing the answer in advance), a good proposal should:

- Have heavier tails than the target distribution p, ensuring all important regions of p are explored
- Concentrate mass in regions where $|f(\boldsymbol{\theta})|p(\boldsymbol{\theta})$ is large

- Be computationally inexpensive to sample from

Effective sample size: The quality of importance sampling is quantified by the effective sample size:

$$N_{\text{eff}} = \frac{(\sum_{i=1}^{N} w_i)^2}{\sum_{i=1}^{N} w_i^2} \tag{162}$$

where $N_{\text{eff}} \in [1, N]$. When all weights are equal ($w_i = \text{const}$), we have $N_{\text{eff}} = N$, indicating that the proposal perfectly matches the target (up to normalization). When a few samples dominate the weight sum, $N_{\text{eff}} \ll N$, indicating poor performance with highly variable weights.

7.4 Conditions for Failure

Importance sampling can fail catastrophically in several well-characterized scenarios:

Proposal too narrow: If q has lighter tails than p, rare but important regions of p are never sampled. Even a single missed high-weight sample can lead to arbitrarily large errors. The variance of the importance sampling estimator is:

$$\text{Var}[\hat{\mu}] = \frac{1}{N} \int f(\boldsymbol{\theta})^2 \frac{p(\boldsymbol{\theta})^2}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} - \mu^2 \tag{163}$$

This variance is infinite if $p(\boldsymbol{\theta})^2/q(\boldsymbol{\theta})$ is not integrable, which occurs when q decays faster than p^2 in the tails.

High dimensionality: In high dimensions, the volume of the typical set of p becomes exponentially different from that of q unless q is very carefully chosen. Most samples from q have negligible weight, and the effective sample size decreases exponentially with dimension. This is the curse of dimensionality for importance sampling: in d dimensions, the volume ratio between concentric shells scales as $(1+\delta)^d - 1 \approx d\delta$ for small δ, meaning that exponentially many samples are required to adequately sample the typical set.

Weight degeneracy: When a few samples dominate the total weight ($\sum_i w_i \approx w_{\max}$), the estimate is determined by a single sample, leading to high variance. This occurs when:

$$\max_i w_i \gg \bar{w} = \frac{1}{N} \sum_{j=1}^{N} w_j \tag{164}$$

Weight degeneracy is diagnosed by computing the effective sample size. If $N_{\text{eff}}/N < 0.5$, the estimate is unreliable.

7.5 Diagnostics and Best Practices

Diagnostic: Always compute and report N_{eff}. A common heuristic is:

- $N_{\text{eff}}/N > 0.5$: good performance
- $0.1 < N_{\text{eff}}/N < 0.5$: acceptable but caution warranted
- $N_{\text{eff}}/N < 0.1$: unreliable, redesign q or use alternative methods
- $N_{\text{eff}} < 100$: unreliable regardless of N

If importance sampling performs poorly, consider:

- **Adaptive importance sampling:** Iteratively refine q based on previous samples
- **Sequential importance sampling:** For time-series data, propagate weighted particles forward
- **MCMC:** When dimensionality is high or the posterior is complex, MCMC methods (section 6) are more robust

7.6 Connection to Other Methods

Importance sampling appears in several contexts:

- **Model comparison:** Importance weights appear in evidence estimation and model averaging
- **Likelihood-free inference:** Importance sampling forms the basis for sequential Monte Carlo ABC methods (section 12)
- **Pareto smoothed importance sampling (PSIS):** Used in leave-one-out cross-validation (PSIS-LOO) for model assessment (subsection 14.10)

For time-series data with state evolution $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and observation model $p(\mathbf{y}_t|\mathbf{x}_t)$, sequential Monte Carlo (particle filters) maintains weighted samples that are resampled and propagated forward, providing a sequential importance sampling framework essential for real-time inference in experimental physics.

8 Convergence tests

It is vital to know that the chain has enough points in it to represent well the target distribution. It will never be perfect, but asymptotically it approaches the right distribution if the detailed balance condition holds. How do we know? A standard technique is the Gelman-Rubin test (1992). Here, two or more chains that begin at 'dispersed' starting points are compared, after their burn-ins are removed. The idea is that if the chains have converged, then their means and variances should agree, except for fluctuations due to there being a finite number of samples. The test is applied separately to each parameter of the model.

The algorithm is

1. Calculate the mean of each chain:
$$\bar{x}_c = \frac{1}{S}\sum_{s=1}^{S} x_{c,s} \tag{165}$$
2. Calculate the variance of each chain:
$$\sigma_c^2 = \frac{1}{S-1}\sum_{s=1}^{S}(x_{c,s} - \bar{x}_c)^2 \tag{166}$$
3. Calculate the mean of all the chains (i.e., the best combined estimate for the mean of the distribution):
$$\bar{x} = \frac{1}{C}\sum_{c=1}^{N_c}\frac{1}{S}\sum_{s=1}^{S} x_{c,s} = \frac{1}{C}\sum_{c=1}^{C}\bar{x}_c \tag{167}$$
4. Calculate the average of the individual chains' variances:
$$\sigma^2_{\text{chains}} = \frac{1}{C}\sum_{i=1}^{C}\sigma_c^2 \tag{168}$$
5. Estimate the variance of the chains' means:
$$\sigma^2_{\text{means}} = \frac{1}{C-1}\sum_{i=1}^{C}(\bar{x}_c - \bar{x})^2 \tag{169}$$
6. The estimated posterior variance is a weighted average of σ^2_{means} and σ^2_{chains}:
$$\hat{V} = \frac{S-1}{S}\sigma^2_{\text{chains}} + \frac{C+1}{C}\sigma^2_{\text{means}} \tag{170}$$

7. We calculate the ratio

$$\hat{R} = \frac{\hat{V}}{\sigma^2_{\text{chains}}} = 1 - \frac{1}{S} + \frac{C+1}{C}\frac{\sigma^2_{\text{means}}}{\sigma^2_{\text{chains}}}. \tag{171}$$

The test statistic $\hat{R}$ can be used to assess convergence. If the chains are well mixed and have all sampled the target distribution then $\sigma^2_{\text{chains}} \simeq \sigma^2_{\text{means}}$ and $\hat{R} \simeq 1$. Whereas if the chains have sampled different parts of the target distribution then their individual variances will be less than the variance between the estimates of the chains and $\hat{R} > 1$. The common heuristic approach is to regard the chains as converged if $\hat{R} \lesssim 1.2$ (we often look for 1.03 or less). The use of means and variances in calculating $\hat{R}$ means it is most appropriate to target densities that are close to be normal and do not have heavy tails. The statistic is also useful in general, even though its distribution under correct sampling is then more difficult to calculate.

9 Hierarchical Models

In many practical situations, the likelihood can be difficult to evaluate, since it may be hard to write a direct expression down for the sampling distribution. But we can often make progress by analysing problems as a multilevel system, or Bayesian Hierarchical Model. We have in fact already seen one of these - fitting a straight line to data with errors in x and y.

A good starting point is to draw a diagram that represents what you would need to do to generate the data. This Directed Acyclic Graph (DAG) is a representation of the data model. Fig. 18 shows the DAG for the straight line fitting problem. x and y are latent variables that are not measured, but which are part of the data model. The general way to deal with these is to introduce them into the probabilities, and marginalise over them. Very often, HMC techniques are used to sample jointly from the parameters of interest and the latent variables simultaneously.

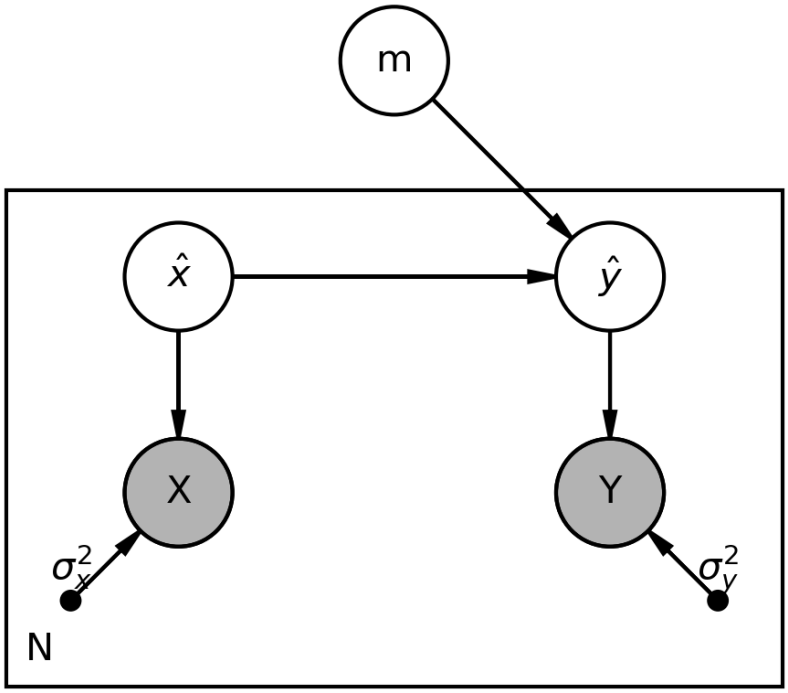

Figure 9: DAG for straight-line fitting with errors on x and y. Circles indicate values drawn from a probability distribution, grey shaded are the measured data. The diamond indicates that y is a deterministic function of the inputs. The rectangle ('plate') with N inside says there are N repetitions of the data. The variances are fixed parameters.

A typical example of a BHM is when we have a population of objects, and we use the collection of individual objects to infer something about the population, whose properties may be specified by one or more population parameters $\boldsymbol{\theta}$.

9.1 Ordinary vs Hierarchical Bayes

In ordinary Bayesian inference, we have a direct relationship between our model parameters and the data. For example, if we're measuring the mean height of students in a class, we use the actual height measurements to directly infer the population mean. The mathematical expression for this straightforward approach is:

$$p(\boldsymbol{\theta}|\mathbf{d}) \propto p(\mathbf{d}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{172}$$

Hierarchical Bayesian models introduce additional layers of structure by recognizing that our parameters might themselves be drawn from higher-level distributions with their own parameters (called hyperparameters). Think of it this way: instead of just asking "What's the average height in this class?", we might ask "What's the average height in this class, given that this class belongs to a school, and this school belongs to a region, and different regions might have different typical heights?"

In hierarchical models, we introduce latent variables $\boldsymbol{\phi}$ that aren't directly observed but help structure our problem. For instance, these might represent the true but unmeasured values that our data are noisy observations of, or they might represent group-level parameters in a multi-level analysis. The mathematical form becomes:

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{d}) \propto p(\mathbf{d}|\boldsymbol{\theta}, \boldsymbol{\phi})p(\boldsymbol{\phi}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{173}$$

where $\boldsymbol{\phi}$ represents these latent variables or parameters. Often we're not directly interested in these latent variables themselves, but rather in how they help us better understand our main parameters $\boldsymbol{\theta}$. We can obtain the distribution of our main parameters by integrating out (marginalizing over) the latent variables:

$$p(\boldsymbol{\theta}|\mathbf{d}) \propto \int p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{d}) d\boldsymbol{\phi} \tag{174}$$

This is a two-level system, but it can be extended to more levels. To make this concrete, let us look at an example (done in lectures) of a population of pairs x, y where the population is drawn from a gaussian distribution $y_i \sim \mathcal{N}(\bar{y}_i, \sigma^2)$, where the mean grows linearly with x, $\bar{y}_i = mx_i + c$, and the variance of the population around this line is fixed at σ^2. m, c, σ^2 are the parameters of the problem. In this case, for simplicity, we take x_i as fixed and known. subsection 14.8 (Empirical Bayes) discusses a pragmatic approach for estimating hyperparameters from data, while subsection 14.5 (Robust Statistics) covers hierarchical models with outliers. The DAG for this model is shown in Figure 10.

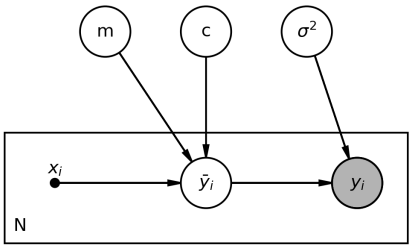

Figure 10: DAG for population linear regression model where $y_i \sim \mathcal{N}(mx_i + c, \sigma^2)$. The parameters m, c, and σ^2 determine the linear relationship, while x_i values are fixed and known.

Example: Hierarchical Model: Radon in Minnesota

Radon is a carcinogen and levels of radon in houses in the US have been studied, with a famous dataset collected and analysed in Gelman et al.'s BDA book.

The data are noisy radon measurements, made in different counties in the US, and on different floors (the radon levels will be higher nearer to the ground). The idea is to pool data from many house measurements, to assess the radon risk in a county c, and to extrapolate to living areas if the measurements were taken in the basement.

The data model is constructed at multiple hierarchical levels. First, we assume that the expected radon level is a linear function of the floor level f:

$$\mu = a_c + b_c f, \tag{175}$$

where f takes the value 0 for basement measurements and 1 for living space measurements. The

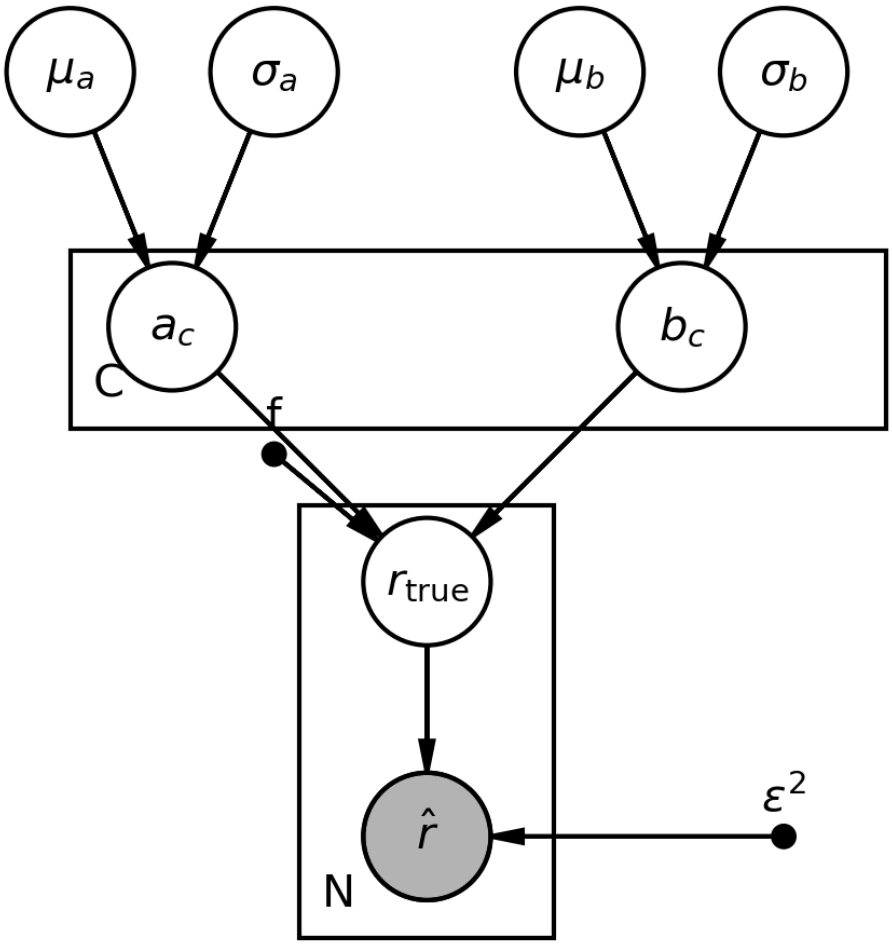

Figure 11: Hierarchical model DAG for radon measurements. Population-level hyperparameters μ_a, σ_a, μ_b, σ_b (top) govern the distributions of county-specific parameters a_c and b_c (middle). The true radon level r_{true} is determined deterministically by a_c, b_c, and the floor level f. The observed radon measurement $\hat{r}$ (shaded) is drawn from a Gaussian distribution centered on r_{true} with measurement error scale ε. The dashed box represents replication over C counties, while the solid box represents replication over N houses per county.

subscript c indicates that both the intercept a_c and the slope b_c depend on the county.
At the measurement level, we assume that observed radon measurements $d(f, c)$ are subject to Gaussian noise with unknown variance ϵ^2:

$$d(f, c) = a_c + b_c f + n, \tag{176}$$

where $n \sim \mathcal{N}(0, \epsilon^2)$ represents the measurement error, which we wish to infer along with other parameters.
The hierarchical structure enters because the coefficients a_c and b_c are not fixed across counties but vary according to population distributions. We model these county-level parameters as drawn from Gaussian distributions with universal (population-level) hyperparameters: $a_c \sim \mathcal{N}(\mu_a, \sigma_a^2)$ and $b_c \sim \mathcal{N}(\mu_b, \sigma_b^2)$, where the hyperparameters μ_a, σ_a, μ_b, and σ_b are themselves unknown and must be inferred. This hierarchical structure allows information to be shared across counties: data from one county inform our beliefs about the population distributions, which in turn improve our inferences about other counties, particularly those with sparse data.
The Directed Acyclic Graph (DAG) for this model is shown in Figure 11. It clearly displays the hierarchical structure, with variability present at multiple levels: population hyperparameters, county-level parameters, and individual measurements. The outer plate (labeled C) indicates that the county-specific parameters a_c and b_c are replicated for each of the C counties. The inner plate (labeled N) indicates that the measurement process is replicated for N houses within each county.
As usual, we analyse the problem systematically by identifying the key components. First, *what do we want to know?* We seek the posterior distributions of multiple quantities: the baseline risk levels for each county (a_c), the additional risk associated with basement measurements in each county (b_c), the measurement variability ϵ (representing either true house-to-house variation or measurement device error), and the population-level parameters μ_a, σ_a, μ_b, and σ_b that characterize variation across counties. All of these inferences are conditioned on the observed data.
Second, *what are the data?* The observations consist of radon measurements $\hat{r}$ taken at various floor levels in houses across different counties.
Third, *what is the model?* The model structure is captured by the DAG mentioned above, showing the hierarchical relationships between population parameters, county parameters, true radon levels, and observed measurements.
Fourth, *what are the parameters?* The complete parameter set includes population hyperparameters μ_a, σ_a, μ_b, σ_b; county-specific parameters a_c, b_c (for each county c); and the measurement error scale ϵ.
Fifth, *what is the likelihood?* At the lowest level of the hierarchy (the measurement level), the likelihood is $\hat{r} \sim \mathcal{N}(r_{\text{true}}, \epsilon^2)$, where $r_{\text{true}} = a_c + b_c f$ is the true radon level.
The inference is typically carried out by sampling from the joint posterior distribution of all unknowns using MCMC methods such as Hamiltonian Monte Carlo, which can efficiently handle the high-dimensional parameter space and complex dependencies in this hierarchical model.

10 Model Comparison

Model comparison addresses a higher-level question than parameter inference: which theoretical framework or model is preferred given the data, regardless of specific parameter values? The models under consideration may be completely different (e.g., comparing Big Bang with Steady State cosmology), or they may be variants of the same fundamental idea. For instance, comparing a simple cosmological model where the Universe is assumed to be flat with a more general model where curvature is allowed to vary represents a comparison between nested models, where adding an extra parameter defines a new, more complex model.

The central question in model comparison is essentially: "Do the data favour a more complex model?" This question cannot be answered using the likelihood alone, as the likelihood will always increase (or at least not decrease) when more parameters are allowed to vary. A principled framework for model comparison must therefore account for both goodness of fit and model complexity.

10.1 Bayesian Evidence

Consider two competing models denoted by M and M', with data vector $\mathbf{d}$ and parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ of length n and n', respectively. Following Rule 1, we write down what we want to know: the probability of each model given the data, $p(M|\mathbf{d})$. Applying Bayes' theorem:

$$p(M|\mathbf{d}) = \frac{p(\mathbf{d}|M)\pi(M)}{p(\mathbf{d})} \tag{177}$$

The key quantity here is the Bayesian evidence (or marginal likelihood), defined as:

$$p(\mathbf{d}|M) = \int d\boldsymbol{\theta}\, p(\mathbf{d}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M) \tag{178}$$

likelihood * prior

Computing the Bayesian evidence is challenging, as it requires a multidimensional integral over the entire parameter space. Nested Sampling (subsection 14.3) is specifically designed for this purpose and is implemented in packages like MultiNest and dynesty. For a model with no free parameters, the integral reduces trivially to $p(\mathbf{d}|M)$, which is simply the sampling distribution.

The relative probabilities of two models is given by:

$$\frac{p(M'|\mathbf{d})}{p(M|\mathbf{d})} = \frac{\pi(M')}{\pi(M)} \frac{\int d\boldsymbol{\theta}'\, p(\mathbf{d}|\boldsymbol{\theta}', M')\pi(\boldsymbol{\theta}'|M')}{\int d\boldsymbol{\theta}\, p(\mathbf{d}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)} \tag{179}$$

With equal priors on the models ($\pi(M') = \pi(M)$), this ratio simplifies to the ratio of evidences, known as the Bayes Factor:

$$B \equiv \frac{\int d\boldsymbol{\theta}'\, p(\mathbf{d}|\boldsymbol{\theta}', M')\pi(\boldsymbol{\theta}'|M')}{\int d\boldsymbol{\theta}\, p(\mathbf{d}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)} \tag{180}$$

10.1.1 Nested models

An important special case occurs when model M' is nested within model M, meaning that M' is a simpler model with fewer parameters ($n' < n$). Specifically, the n' parameters of M' are a subset of the n parameters of M, with the remaining $p \equiv n - n'$ parameters fixed to fiducial values in M'. A crucial feature of nested models is that the more complicated model M will inevitably achieve a likelihood at least as high as that of M'. However, the evidence may still favour the simpler model if the fit is nearly as good, because the evidence automatically penalizes the larger prior volume required by the additional parameters.

Assuming uniform (and hence separable) priors over ranges $\Delta\theta_i$ for each parameter, the prior is $p(\boldsymbol{\theta}|M) = (\Delta\theta_1 \dots \Delta\theta_n)^{-1}$. The Bayes factor becomes:

$$B = \frac{\int d\boldsymbol{\theta}'\, p(\mathbf{d}|\boldsymbol{\theta}', M')}{\int d\boldsymbol{\theta}\, p(\mathbf{d}|\boldsymbol{\theta}, M)} \frac{\Delta\theta_1 \dots \Delta\theta_n}{\Delta\theta'_1 \dots \Delta\theta'_{n'}} \tag{181}$$

For this expression to be meaningful, the prior ranges must be large enough to contain essentially all of the likelihood; otherwise, the position of the boundaries would artificially influence the Bayes factor. In the nested case, the ratio of prior hypervolumes simplifies to:

$$\frac{\Delta\theta_1 \dots \Delta\theta_n}{\Delta\theta'_1 \dots \Delta\theta'_{n'}} = \Delta\theta_{n'+1} \dots \Delta\theta_{n'+p} \tag{182}$$

where p is the number of extra parameters in the more complicated model.

10.1.2 Computational challenges and alternatives

The evidence requires a multidimensional integral over the likelihood and prior, which can be computationally expensive or intractable for high-dimensional parameter spaces. While detailed algorithms are beyond the scope of this course, the most widely used method is nested sampling (implemented in packages such as MultiNest and PolyChord), which samples the likelihood in an efficient manner by exploring nested iso-likelihood contours.

Classical approximations such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) may be unreliable in many contexts. These criteria are based on the best-fit χ^2 and do not properly account for the full parameter space that would yield the data with high probability. Moreover, they do not include the prior and are not fully Bayesian. Information criteria like WAIC and LOO-CV (subsection 14.10) provide more principled alternatives for model assessment.

Rather than selecting a single "best" model, Bayesian Model Averaging (subsection 10.4) offers an approach that combines predictions from multiple models weighted by their posterior probabilities. This approach naturally accounts for model uncertainty. For practical model assessment using cross-validation techniques, see subsection 14.10.

10.2 Savage-Dickey Ratio

For nested models where model M_0 is a special case of M_1, the Savage-Dickey Density Ratio (SDDR) provides a remarkable simplification for computing the Bayes factor. Consider the case where model M_0 has parameters ψ, while model M_1 has parameters (ψ, ϕ), with M_0 corresponding to $\phi = \phi_0$ (a fixed value). Assuming the conditional priors are continuous such that $\lim_{\phi\to\phi_0} \pi_1(\psi|\phi) = \pi_0(\psi)$, or equivalently $\pi_1(\psi|\phi = \phi_0) = \pi_0(\psi)$, we can derive a simple expression for the Bayes factor.

The Bayes factor comparing M_0 to M_1 is:

$$B_{01} \equiv \frac{p(x|M_0)}{p(x|M_1)} = \frac{\int p_0(x|\psi)\pi_0(\psi)d\psi}{\int p_1(x|\psi,\phi)\pi_1(\psi,\phi)d\psi d\phi} \tag{183}$$

Using the continuity assumption, this can be rewritten as:

$$B_{01} = \frac{\int p_1(x|\psi,\phi=\phi_0)\pi_1(\psi,\phi=\phi_0)d\psi}{\int p_1(x|\psi,\phi)\pi_1(\psi,\phi)d\psi d\phi} = \frac{p_1(x|\phi=\phi_0)}{p_1(x)} \tag{184}$$

where the last equality follows from $p(A|B,C)p(B|C) = p(A,B|C)$ and $\int p(A,B|C)dB = p(A|C)$ applied to the numerator, with analogous manipulations in the denominator.

Applying Bayes' theorem to the conditional distribution:

$$p_1(x|\phi=\phi_0) = \frac{p_1(\phi=\phi_0|x)p_1(x)}{\pi_1(\phi=\phi_0)} \tag{185}$$

Substituting this into the expression for B_{01} yields the Savage-Dickey Density Ratio:

$$B_{01} = \frac{p_1(\phi=\phi_0|x)}{\pi_1(\phi=\phi_0)} \tag{186}$$

This remarkably simple result expresses the Bayes factor as the ratio of posterior to prior density at the point $\phi = \phi_0$. Despite its simplicity, practical application requires care, as the numerator is a posterior

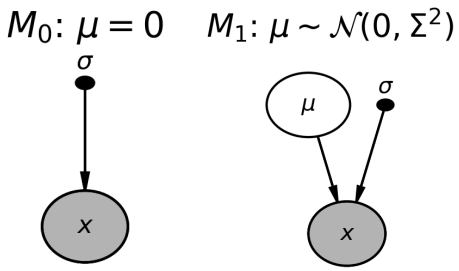

Figure 12: DAG for nested Gaussian model comparison (section 10). Model M_0 (left) fixes $\mu = 0$, while model M_1 (right) allows μ to vary with prior $\mathcal{N}(0, \Sigma^2)$.

density that must be estimated from MCMC samples. The denominator is straightforward if the prior has a simple functional form. The numerator can be estimated from posterior samples in model M_1 using kernel density estimation or, more simply, by computing the fraction $f(\Delta\phi, \phi_0)$ of samples within a small range $\pm\Delta\phi$ of ϕ_0, giving $p_1(\phi_0|x) \simeq f(\Delta\phi, \phi_0)/(2\Delta\phi)$.

Example: Model Comparison: Nested Gaussian Models

Problem: A scientist measures a quantity and obtains a single observation. They must decide between two competing hypotheses: either the true value is exactly zero (a simple model with no free parameters), or the true value could be any number (a more complex model with one free parameter). The data consist of a single measurement with known Gaussian measurement noise. This is a classic model comparison problem where we must balance the goodness-of-fit of the more flexible model against the simplicity of the constrained model. The question is: how strong must the evidence be to prefer the more complex model over the simpler one?

Solution:

We formalize this as a comparison between nested Gaussian models. Let M_0 be $x \sim \mathcal{N}(0, \sigma^2)$, and M_1 be $x \sim \mathcal{N}(\mu, \sigma^2)$, where the prior on μ is gaussian with variance Σ^2. Let the measurement be $x = \lambda\sigma$. In this gaussian example, we can evaluate the Bayesian evidence integrals analytically.

$$p_0(x|M_0) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)} \tag{187}$$

and

$$p_1(x|\mu, M_1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \tag{188}$$

Hence

$$B_{01} = \frac{p_0(x|M_0)}{\int_{-\infty}^{\infty} p_1(x|\mu, M_1) p_1(\mu|M_1) d\mu} \tag{189}$$

i.e.,

$$B_{01} = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}}{\frac{1}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}\Sigma} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/(2\sigma^2)} e^{-\mu^2/(2\Sigma^2)} d\mu} \tag{190}$$

so

$$B_{01} = \sqrt{1 + \frac{\Sigma^2}{\sigma^2}} \exp\left[-\frac{\lambda^2}{2(1 + \frac{\sigma^2}{\Sigma^2})}\right] \tag{191}$$

If $\lambda \gg 1$, then $B_{01} \ll 1$ and M_1 is favoured. If $\lambda \simeq 1$ and $\sigma \ll \Sigma$, then M_0 is favoured (Occam's razor). If likelihood is much broader than prior, $\sigma \gg \Sigma$ then $B_{01} \simeq 1$ and nothing has been learned. This diagram is very interesting and instructive, and somewhat counter-intuitive. To favour the more complicated model with high probability (say 10 times the probability of the simple model), then the

deviation from the simple model parameter value needs to be at least about 3σ. So a 3σ 'result' is really not very significant in a model comparison context, since a probability of 10% is not particularly small.

Example: Analytic Evidence: Gaussian Likelihood with Gaussian Prior

Problem: Building on the Gaussian likelihood example from section 4, we now compute the Bayesian evidence for a model with a Gaussian prior on the mean parameter μ. This allows us to compare different models with different prior specifications or to assess the overall support for the model given the data. Computing the evidence analytically demonstrates the key principle that the evidence automatically penalizes models with larger prior volumes (implementing Occam's razor) while rewarding models that predict the data well.

Solution:

Recall from section 4 that we have N independent measurements $\{d_1, d_2, \ldots, d_N\}$ with known measurement variance σ^2. The likelihood is:

$$p(\mathbf{d}|\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i-\mu)^2\right) \tag{192}$$

With a Gaussian prior on μ:

$$\pi(\mu) = \mathcal{N}(\mu;\mu_0,\sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \tag{193}$$

The evidence is the marginal likelihood, obtained by integrating over the parameter μ:

$$p(\mathbf{d}|\sigma,M) = \int_{-\infty}^{\infty} p(\mathbf{d}|\mu,\sigma)\pi(\mu)d\mu \tag{194}$$

Using the result from section 3 that the likelihood can be written as:

$$p(\mathbf{d}|\mu,\sigma) \propto \exp\left(-\frac{N(\mu-\bar{d})^2}{2\sigma^2}\right) \tag{195}$$

where $\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$ is the sample mean.
The evidence integral becomes:

$$p(\mathbf{d}|\sigma,M) = \int_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{N(\mu-\bar{d})^2}{2\sigma^2}\right)\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)d\mu \tag{196}$$

This integral is a product of two Gaussians in μ. Using the Gaussian product identity from section 1, the product of two Gaussians integrates to give another Gaussian normalization. The result is:

$$p(\mathbf{d}|\sigma,M) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \frac{1}{\sqrt{2\pi\sigma_0^2}}\sqrt{2\pi\sigma_N^2}\exp\left(-\frac{(\bar{d}-\mu_0)^2}{2(\sigma^2/N+\sigma_0^2)}\right) \tag{197}$$

where $\sigma_N^2 = \left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}$ is the posterior variance from section 4.
Simplifying:

$$p(\mathbf{d}|\sigma,M) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \frac{\sqrt{\sigma_N^2}}{\sqrt{\sigma_0^2}}\exp\left(-\frac{(\bar{d}-\mu_0)^2}{2(\sigma^2/N+\sigma_0^2)}\right) \tag{198}$$

Interpretation:

The evidence can be understood as the likelihood of the data under the prior predictive distribution. The exponential term penalizes models where the observed mean $\bar{d}$ is far from the prior mean μ_0,

measured in units of the combined uncertainty $\sqrt{\sigma^2/N + \sigma_0^2}$. The prefactor $\sqrt{\sigma_N^2}/\sqrt{\sigma_0^2} < 1$ implements Occam's razor: models with broader priors (larger σ_0^2) are penalized because they spread their prior probability over a larger parameter space, making any specific prediction less probable.
To compare two models with different prior widths $\sigma_{0,1}^2$ and $\sigma_{0,2}^2$, the Bayes factor is:

$$B_{12} = \frac{p(\mathbf{d}|M_1)}{p(\mathbf{d}|M_2)} = \frac{\sqrt{\sigma_{N,1}^2}}{\sqrt{\sigma_{N,2}^2}} \frac{\sqrt{\sigma_{0,2}^2}}{\sqrt{\sigma_{0,1}^2}} \exp\left(-\frac{1}{2}\left[\frac{(\bar{d}-\mu_{0,1})^2}{\sigma^2/N + \sigma_{0,1}^2} - \frac{(\bar{d}-\mu_{0,2})^2}{\sigma^2/N + \sigma_{0,2}^2}\right]\right) \tag{199}$$

This demonstrates the automatic trade-off between model complexity (prior volume) and goodness of fit that is inherent in the Bayesian evidence.

Example: Analytic Evidence: Beta-Binomial Coin Toss

Problem: For the coin toss problem from section 5, we compute the Bayesian evidence to compare different models or assess how well the model predicts the data. This example demonstrates evidence calculation for discrete data with a Beta prior, complementing the continuous Gaussian case above. The evidence was already computed as an intermediate step in section 5, but here we emphasize its interpretation in the context of model comparison.
Solution:
Recall from section 5 that for n coin tosses with k heads observed, the likelihood is:

$$p(k|\theta, n) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{200}$$

With a Beta prior:

$$\pi(\theta) = \text{Beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{201}$$

The evidence is:

$$p(k|n, M) = \int_0^1 p(k|\theta, n)\pi(\theta)d\theta = \binom{n}{k} \int_0^1 \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} d\theta \tag{202}$$

Using the Beta function identity $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$:

$$p(k|n, M) = \binom{n}{k} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(k+\alpha)\Gamma(n-k+\beta)}{\Gamma(n+\alpha+\beta)} \tag{203}$$

For the uniform prior case $\alpha = \beta = 1$, this simplifies to:

$$p(k|n, M_{\text{uniform}}) = \binom{n}{k} \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} = \binom{n}{k} \frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1} \tag{204}$$

This is the result we obtained in section 5.
Model Comparison: To compare a uniform prior $\text{Beta}(1,1)$ against an informative prior $\text{Beta}(\alpha, \beta)$ that encodes prior belief about fairness, the Bayes factor is:

$$B = \frac{p(k|n, M_{\text{informative}})}{p(k|n, M_{\text{uniform}})} = (n+1)\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(k+\alpha)\Gamma(n-k+\beta)}{\Gamma(n+\alpha+\beta)} \tag{205}$$

Interpretation:
The evidence $p(k|n)$ represents the prior predictive probability of observing exactly k heads in n tosses, averaging over all possible values of θ weighted by the prior. For a uniform prior, this gives

equal probability $\frac{1}{n+1}$ to all possible outcomes $k = 0, 1, \ldots, n$, reflecting maximum uncertainty about θ.
An informative prior that concentrates probability near $\theta = 0.5$ (e.g., Beta(10, 10)) will assign higher evidence to outcomes near $k = n/2$ and lower evidence to extreme outcomes. If the data show $k \approx n/2$, the informative prior will have higher evidence; if k is far from $n/2$, the uniform prior may have higher evidence because it did not commit strongly to any particular value of θ.
This demonstrates the fundamental principle of Bayesian model comparison: models that make more specific predictions are rewarded when those predictions are confirmed, but penalized when they are not. The evidence automatically implements this trade-off without requiring ad-hoc complexity penalties.

10.3 Occam's Razor

Occam's razor, the principle that "entities should not be multiplied without necessity," finds its most rigorous mathematical expression in Bayesian model comparison. The Bayesian evidence automatically implements this principle by balancing goodness of fit against model complexity without requiring ad-hoc penalty terms.

10.3.1 The Mathematical Foundation

Consider two models: a simple model M_s with parameter space Θ_s and a complex model M_c with parameter space Θ_c, where $\Theta_s \subset \Theta_c$ (nested models). The evidence for each model is:

$$p(\mathbf{d}|M_s) = \int_{\Theta_s} p(\mathbf{d}|\boldsymbol{\theta})\pi_s(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{206}$$

$$p(\mathbf{d}|M_c) = \int_{\Theta_c} p(\mathbf{d}|\boldsymbol{\theta})\pi_c(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{207}$$

The key insight is that the prior $\pi_c(\boldsymbol{\theta})$ for the complex model must integrate to unity over the larger space Θ_c, while $\pi_s(\boldsymbol{\theta})$ integrates to unity over the smaller space Θ_s. This creates an automatic **complexity penalty**.

To see this explicitly, consider the case where both models make identical predictions in the region Θ_s, but the complex model allows additional parameter variation. If we use uniform priors:

$$\pi_s(\boldsymbol{\theta}) = \frac{1}{V_s} \quad \text{for } \boldsymbol{\theta} \in \Theta_s \tag{208}$$

$$\pi_c(\boldsymbol{\theta}) = \frac{1}{V_c} \quad \text{for } \boldsymbol{\theta} \in \Theta_c \tag{209}$$

where V_s and V_c are the volumes of the respective parameter spaces, with $V_c > V_s$.

Since the complex model spreads its prior probability over a larger volume, it assigns lower prior probability to any specific parameter value: $\pi_c(\boldsymbol{\theta}) = \frac{1}{V_c} < \frac{1}{V_s} = \pi_s(\boldsymbol{\theta})$ in the region Θ_s.

10.3.2 Derivation of the Complexity Penalty

For a concrete example, suppose the data strongly constrain the parameters to lie in a small region $\Delta\Theta \subset \Theta_s$ where the likelihood is approximately constant: $p(\mathbf{d}|\boldsymbol{\theta}) \approx L_{\max}$. Then:

$$p(\mathbf{d}|M_s) \approx L_{\max} \int_{\Delta\Theta} \pi_s(\boldsymbol{\theta})d\boldsymbol{\theta} \approx L_{\max} \frac{|\Delta\Theta|}{V_s} \tag{210}$$

$$p(\mathbf{d}|M_c) \approx L_{\max} \int_{\Delta\Theta} \pi_c(\boldsymbol{\theta})d\boldsymbol{\theta} \approx L_{\max} \frac{|\Delta\Theta|}{V_c} \tag{211}$$

The Bayes factor is:

$$B = \frac{p(\mathbf{d}|M_s)}{p(\mathbf{d}|M_c)} \approx \frac{V_c}{V_s} \tag{212}$$

This shows that the simple model is preferred by a factor equal to the ratio of parameter space volumes—a direct manifestation of Occam's razor. The complex model is penalized proportionally to how much additional parameter space it explores.

10.3.3 Information-Theoretic Interpretation

From an information-theoretic perspective, models that make more specific predictions (smaller parameter space) are taking a greater "risk" by ruling out more possible data sets. When these risky predictions succeed, they are rewarded with higher evidence. Conversely, models that hedge their bets with broader parameter ranges are penalized for their lack of specificity.

The complexity penalty can be understood as the **information cost** of encoding additional parameters. In the limit of weak data (low signal-to-noise), the evidence reduces to the prior predictive distribution, and complex models are heavily penalized. As data become more informative, the likelihood begins to dominate, and the penalty for complexity diminishes.

10.3.4 The Data-Dependent Trade-off

Figure 13 illustrates how Occam's razor operates in practice. The preference between simple and complex models depends critically on where the data fall:

- **Data within simple model range**: The simple model is strongly preferred because it "predicted" this region with higher prior probability.
- **Data outside simple model range**: The complex model is preferred because it can accommodate the data while the simple model cannot.
- **Borderline cases**: The trade-off depends on the precise balance between the likelihood concentration and the prior volume ratio.

This mechanism ensures that models are only made more complex when the data actually require that complexity. It provides a principled, automatic implementation of Occam's razor that requires no arbitrary tuning parameters or external complexity measures.

10.4 Bayesian Model Averaging

Model comparison via Bayes factors, as discussed above, identifies which model is most strongly supported by the data. However, selecting a single "best" model discards information about model uncertainty and can lead to overconfident predictions. Bayesian Model Averaging (BMA) provides a principled framework for incorporating model uncertainty into predictions by averaging over all candidate models, weighted by their posterior probabilities.

10.4.1 Framework

Consider a set of K candidate models $\mathcal{M} = \{M_1, \ldots, M_K\}$, where each model M_k specifies a likelihood $p(\mathbf{d}|\boldsymbol{\theta}_k, M_k)$ and prior $\pi(\boldsymbol{\theta}_k|M_k)$ over its parameters $\boldsymbol{\theta}_k$. Different models may have different numbers of parameters and completely different parameterizations.

The posterior probability of model M_k given observed data $\mathbf{d}$ is:

$$p(M_k|\mathbf{d}) = \frac{p(\mathbf{d}|M_k)\pi(M_k)}{\sum_{j=1}^{K} p(\mathbf{d}|M_j)\pi(M_j)} \tag{213}$$

where $p(\mathbf{d}|M_k)$ is the marginal likelihood (evidence) for model M_k:

$$p(\mathbf{d}|M_k) = \int p(\mathbf{d}|\boldsymbol{\theta}_k, M_k)\pi(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k \tag{214}$$

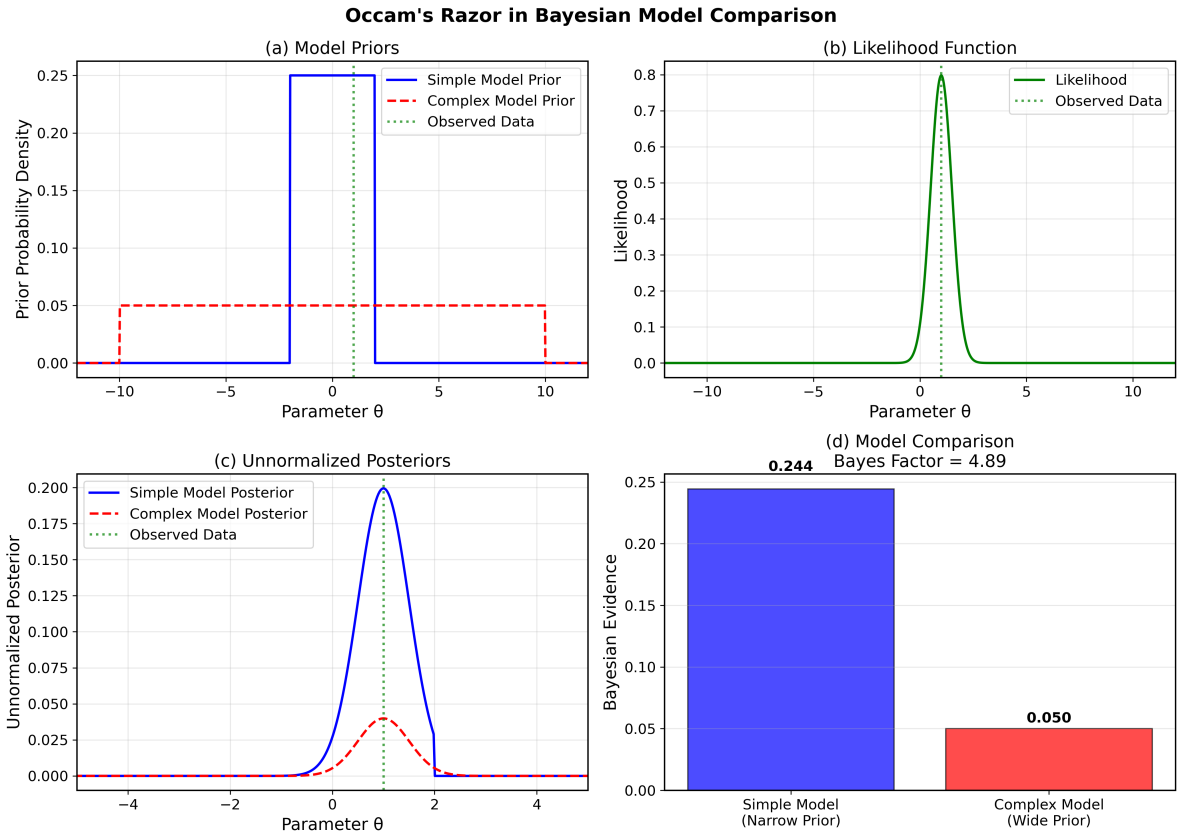

Figure 13: Illustration of Occam's Razor in Bayesian model comparison. (a) Two models with different prior ranges: simple model (narrow blue) vs. complex model (wide red dashed). (b) Likelihood function peaked at observed data. (c) Resulting unnormalized posteriors. (d) Evidence comparison showing automatic complexity penalty. The simple model is preferred despite identical likelihood values because it assigned higher prior probability to the observed data region.

and $\pi(M_k)$ is the prior probability assigned to model M_k. If we have no prior preference among models, we set $\pi(M_k) = 1/K$.

To make predictions for new data $\tilde{\mathbf{y}}$, we must account for both parameter uncertainty within each model and uncertainty about which model is correct. The full posterior predictive distribution is:

$$p(\tilde{\mathbf{y}}|\mathbf{d}) = \sum_{k=1}^{K} p(\tilde{\mathbf{y}}|M_k, \mathbf{d}) p(M_k|\mathbf{d}) \tag{215}$$

where the predictive distribution for each model is:

$$p(\tilde{\mathbf{y}}|M_k, \mathbf{d}) = \int p(\tilde{\mathbf{y}}|\boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k|\mathbf{d}, M_k) d\boldsymbol{\theta}_k \tag{216}$$

This is a mixture distribution where each model contributes to the prediction with weight equal to its posterior probability. Models with higher evidence (better fit while accounting for complexity) receive more weight.

10.4.2 Derivation and Properties

The BMA posterior predictive distribution follows from the law of total probability:

$$p(\tilde{\mathbf{y}}|\mathbf{d}) = \sum_{k=1}^{K} p(\tilde{\mathbf{y}}, M_k|\mathbf{d}) \tag{217}$$

$$= \sum_{k=1}^{K} p(\tilde{\mathbf{y}}|M_k, \mathbf{d}) p(M_k|\mathbf{d}) \tag{218}$$

$$= \sum_{k=1}^{K} \left[\int p(\tilde{\mathbf{y}}|\boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k|\mathbf{d}, M_k) d\boldsymbol{\theta}_k \right] p(M_k|\mathbf{d}) \tag{219}$$

The posterior mean prediction under BMA is:

$$\mathbb{E}[\tilde{\mathbf{y}}|\mathbf{d}] = \sum_{k=1}^{K} \mathbb{E}[\tilde{\mathbf{y}}|M_k, \mathbf{d}] p(M_k|\mathbf{d}) \tag{220}$$

a weighted average of predictions from each model. The posterior variance decomposes as:

$$\mathrm{Var}[\tilde{\mathbf{y}}|\mathbf{d}] = \underbrace{\sum_{k=1}^{K} \mathrm{Var}[\tilde{\mathbf{y}}|M_k, \mathbf{d}] p(M_k|\mathbf{d})}_{\text{within-model}} + \underbrace{\sum_{k=1}^{K} (\mathbb{E}[\tilde{\mathbf{y}}|M_k, \mathbf{d}] - \mathbb{E}[\tilde{\mathbf{y}}|\mathbf{d}])^2 p(M_k|\mathbf{d})}_{\text{between-model}} \tag{221}$$

BMA naturally accounts for two sources of uncertainty: uncertainty about parameters within each model, and uncertainty about which model is correct. Standard single-model inference captures only the first term and thus systematically underestimates predictive uncertainty when model selection uncertainty is non-negligible.

10.4.3 Practical Considerations

Computing BMA requires evaluating the evidence $p(\mathbf{d}|M_k)$ for each model. Methods include nested sampling (subsection 14.3), bridge sampling, thermodynamic integration, or approximate methods like BIC (though this is less accurate).

When one model has overwhelming posterior support ($p(M_k|\mathbf{d}) \approx 1$), BMA reduces to inference under that single model. However, when multiple models have comparable support, BMA provides more honest uncertainty quantification than model selection followed by inference conditional on the selected model.

10.4.4 Continuous Model Spaces

The discrete model averaging framework extends to continuous model spaces. For instance, in polynomial regression, we might be uncertain whether the true model is linear, quadratic, cubic, or higher order. This leads to a variable-dimension inference problem.

Reversible Jump MCMC (RJMCMC) and transdimensional samplers construct a Markov chain that moves between models of different dimensionality. The chain visits model M_k with frequency proportional to $p(M_k|\mathbf{d})$, and within each model visits parameter values according to $p(\boldsymbol{\theta}_k|\mathbf{d}, M_k)$. The combined chain yields samples from the joint distribution over models and parameters, from which BMA predictions can be constructed.

10.4.5 Implications and Limitations

BMA addresses the fundamental problem that model selection ignores model uncertainty. By averaging rather than selecting, predictions are more robust and uncertainty is more accurately quantified. This is particularly important for decision-making under uncertainty, where underestimating predictive variance can lead to poor decisions.

However, BMA has limitations. First, it requires specifying a finite set of candidate models, and predictions are conditional on this set. If the true data-generating process is not well-approximated by any candidate model, BMA provides no protection. Second, BMA predictions can be difficult to interpret when different models make qualitatively different predictions. Third, computational cost grows linearly with the number of models, making BMA with very large model spaces challenging.

Despite these limitations, BMA represents a principled Bayesian approach to model uncertainty and is widely used in fields ranging from climate science to econometrics, where accounting for structural uncertainty is essential for honest quantification of predictive uncertainty. For practical model assessment using cross-validation techniques that do not require evidence calculation, see subsection 14.10.

11 Selection Effects

A common challenge in experimental science arises when measurements do not always produce a definitive result. This occurs when signals fall below detection thresholds or, conversely, exceed the dynamic range of the instrument. How should such incomplete data be incorporated into Bayesian inference? The answer requires careful consideration of the data generation process, and directed acyclic graphs (DAGs) provide a useful framework for structuring this analysis.

We distinguish between two types of incomplete data:

- **Censoring**: The experiment indicates that a measurement was attempted, but no detection was made. We know how many non-detections occurred, but not their values.
- **Truncation**: The experiment returns only detected measurements. We do not know how many objects failed to produce detections, nor do we even know that they exist.

An astronomical example illustrates the distinction. Suppose we compile a catalogue of optically bright stars in a patch of sky and attempt to measure their radio flux. For some stars, the radio telescope fails to detect emission, and we record that the flux lies below the detection limit. This is censoring: we know which stars were observed and which failed to produce detections.

In contrast, if we conduct only the radio survey without prior knowledge of the star positions, we observe only those stars with detectable radio emission. The non-emitting stars remain completely unknown. This is truncation: the number of undetected objects is itself uncertain.

We now examine a concrete example of censored data:

Example: Selection Effects: Censored Data

An experiment measures the mass of N identical objects, each with true mass μ. The measurement errors are independent and Gaussian distributed with zero mean and known variance σ^2. For $M \leq N$ objects, the experiment successfully detects the mass (designated $I = 1$), returning a measured value x_d. For the remaining $N - M$ objects, the experiment determines that the measured value falls below a threshold $x_{\min} = 3\sigma$, indicating insufficient confidence, and reports only a non-detection ($I = 0$) without providing a numerical estimate.

The Bayesian approach proceeds as follows. We begin by constructing a DAG that represents the data generation process, as illustrated in Fig. 25. The parameter μ is drawn from a prior distribution and generates N independent measurements x, each corrupted by Gaussian noise of variance σ^2. The detection criterion determines whether each measurement is reported as a detection ($x_d = x$ if $x \geq x_{\min}$) or as a non-detection ($I = 0$ if $x < x_{\min}$).

Following the first principle of Bayesian inference, we write down precisely what we wish to determine: the posterior probability distribution of μ given the M detected measurements x_d and the $N - M$ non-detections. Applying Bayes' theorem with prior $\pi(\mu)$:

$$p(\mu|x_d, I) \propto p(x_d, I|\mu)\pi(\mu) \tag{222}$$

To evaluate the likelihood, we introduce the latent (unobserved) true values x and marginal over them. Since the observed value equals the true value for detections ($x_d = x$ when detected), we treat detections and non-detections separately. For notational simplicity, we represent non-detections by setting $x_d = 0$, allowing us to suppress the indicator I and work solely with the data vector x_d:

$$p(\mu|x_d) \propto \pi(\mu) \int p(x_d, x|\mu)dx \tag{223}$$

$$\propto \pi(\mu) \prod_{i=1}^{N} \int p(x_{d,i}|x_i, \mu)p(x_i|\mu)dx_i \tag{224}$$

The sample partitions into detections and non-detections. For detected objects, the conditional distribution $p(x_d|x) = \delta^D(x - x_d)$ (Dirac delta function) makes the integral trivial. For non-detected objects, the true value x can lie anywhere below the threshold, with $p(x_{d,i} = 0|x_i) = 1$ for $x_i < x_{\min}$.

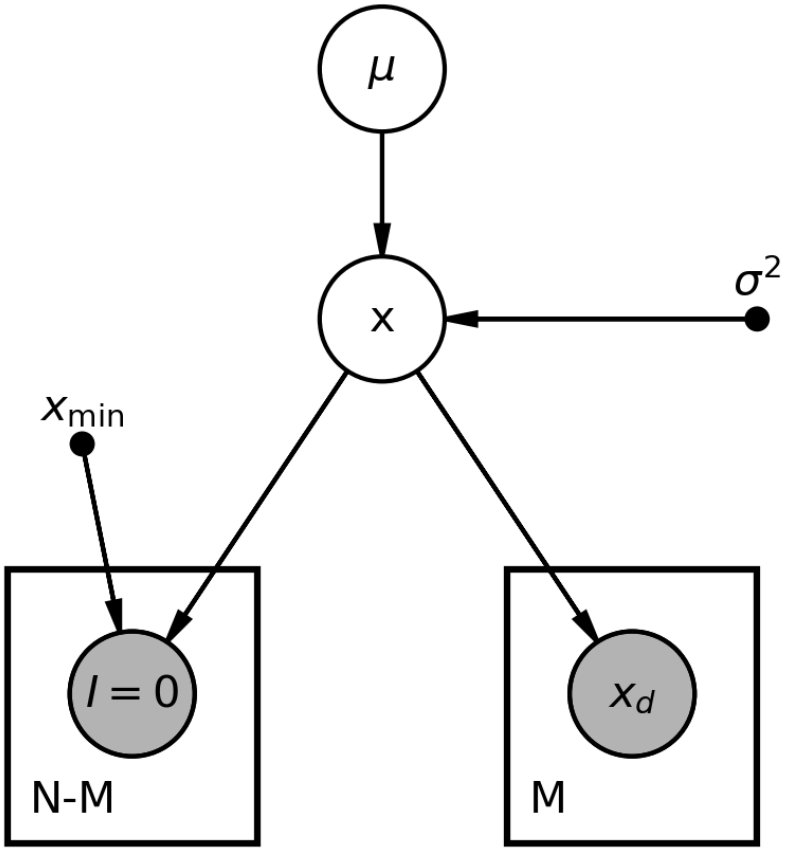

Figure 14: DAG for the censored data model. The parameter μ generates true values x with Gaussian noise σ^2. The threshold $x_{\min}$ determines whether a measurement results in a detection (x_d) or non-detection ($I = 0$). The plates indicate M detected objects and $N - M$ non-detected objects.

The contribution from undetected objects therefore requires integrating the sampling distribution over all sub-threshold values:

$$\int_{-\infty}^{x_{\min}} \mathcal{N}(x_i; \mu, \sigma^2) dx_i \equiv \Phi(x_{\min} - \mu), \tag{225}$$

where $\Phi(x) \equiv \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}\sigma}\right)\right]$ and erf is the error function.
The posterior distribution therefore takes the form

$$p(\mu|x_d) \propto \pi(\mu)\Phi^{N-M}(x_{\min} - \mu)\binom{N}{M}\prod_{i=1}^{M}\mathcal{N}(x_{d,i}; \mu, \sigma^2). \tag{226}$$

The binomial coefficient $\binom{N}{M}$ accounts for the number of ways to select M detections from N measurements. Since N and M are fixed by the data, this factor is constant and can be absorbed into the normalization. To evaluate the posterior numerically, we need only specify a prior for μ. As μ is a location parameter, a uniform prior is typically appropriate.

11.1 Truncation

We now consider a modified scenario where the total number of measurements attempted is unknown. The experiment returns only the M detected values, with no information about how many objects failed to produce detections. The data generation process remains the same, but the parameter space now includes the unknown total number of objects $N \geq M$.

Since our primary interest lies in μ rather than N, we treat N as a nuisance parameter and marginalize over it. The joint posterior distribution for both μ and N is

$$p(\mu, N|x_d) \propto p(x_d|\mu, N)\pi(\mu)\pi(N). \tag{227}$$

The mathematical analysis follows the censored case, but the combinatorial factor $\binom{N}{M}$ now depends on the unknown parameter N and cannot be absorbed into the normalization. Marginalizing over the discrete parameter N (which must satisfy $N \geq M$) yields the posterior for μ:

$$p(\mu|x_d) \propto \pi(\mu)\sum_{N=M}^{\infty}\pi(N)\Phi^{N-M}(x_{\min} - \mu)\binom{N}{M}\prod_{i=1}^{M}\mathcal{N}(x_{d,i}; \mu, \sigma^2). \tag{228}$$

An appropriate prior for N is the Jeffreys prior $\pi(N) \propto 1/N$, reflecting the fact that N is a scale parameter. This discrete sum converges provided the detection probability $\Phi(x_{\min} - \mu)$ is sufficiently small, ensuring that the likelihood decreases rapidly with increasing N.

12 Simulation-based Inference

Likelihood-free inference (LFI), also known as implicit likelihood or simulation-based inference (SBI), represents a fundamentally different approach to Bayesian parameter inference. This methodology is particularly suited to cases where the likelihood is either computationally prohibitive or impossible to evaluate analytically. The approach requires a simulator that can generate synthetic data for given parameter values, typically implemented as a computer program with adjustable model parameters. The fundamental principle is to run a large number of simulations with parameters drawn from a prior distribution, retaining only those simulations whose outputs match the observed data. The distribution of parameters corresponding to these matching simulations provides an approximation to the posterior distribution.

Several fundamental challenges arise in this approach. First, when the data are continuous rather than discrete, the probability of exactly reproducing the observed data is precisely zero (a set of measure zero), necessitating the introduction of some tolerance criterion. Second, even with a non-zero tolerance, the probability of matching all components of a high-dimensional dataset becomes vanishingly small as dimensionality increases. For example, simulating the Universe and expecting to reproduce the detailed structure of the Milky Way, its neighbour Andromeda, and all dwarf galaxies of the Local Group would have negligible probability of success.

Consequently, rather than demanding a perfect match, the method typically requires only that certain summary statistics are reproduced within an acceptable tolerance. Common examples of such summary statistics include correlation functions and power spectra, which capture essential features of the data in a lower-dimensional representation.

12.1 Approximate Bayesian Computation

Consider a simple case: a model with one parameter θ and one data point d (denoted α and $\tilde{\alpha}$ in Fig. 23). We draw θ from a prior distribution $\pi(\theta)$ and run a simulation with that parameter value, generating a synthetic data point. By repeating this process many times, we obtain samples from the joint distribution $p(\theta, d)$, as illustrated in Fig. 23.

The simplest approach to obtaining the posterior is to select those simulations whose outputs lie close to the observed data $d = d_m$, within some tolerance ϵ. This method is known as Approximate Bayesian Computation (ABC). The distribution of retained θ values approximates the posterior $p(\theta|d_m)$, with the approximation improving as $\epsilon \to 0$. However, as ϵ decreases, fewer simulations satisfy the acceptance criterion, increasing sampling noise. To obtain sufficient accepted samples, a very large number of simulations must be run, rendering the method computationally expensive. See Fig. 23.

It is worth noting that one can also obtain an estimate of the likelihood function (or more precisely, the sampling distribution) by selecting simulations at approximately fixed values of θ. For this reason, the approach is sometimes referred to as implicit likelihood rather than likelihood-free inference, since the likelihood is implicitly represented in the simulation outputs.

As an alternative to the rejection-based ABC procedure, the distribution of simulated points can be approximated by fitting a continuous function using machine learning techniques, broadly categorized as kernel density estimation (KDE). The DELFI package implements this approach. Once the distribution is fitted, the posterior can be evaluated by computing the approximated probability density at $d = d_m$ as a function of θ.

12.2 Neural Density Estimation

While ABC provides a conceptually simple approach, it suffers from poor computational efficiency, particularly in high-dimensional settings. Modern practice therefore favors fitting a parametric function $q_\varphi(\theta, d)$ with learnable parameters φ to approximate either the joint distribution $p(\theta, d)$, the likelihood function $p(d|\theta)$, or the posterior distribution $p(\theta|d)$ directly. This is typically achieved using neural networks. These methods are known as neural likelihood estimation (NLE) and neural posterior estimation (NPE), respectively, with several variants having been developed. Neural density estimation is closely connected to broader developments in modern machine learning; subsection 14.2 discusses variational inference with neural networks, while subsection 14.11 covers Bayesian neural networks and normalizing flows.

12.2.1 Kullback-Leibler (KL) divergence

The KL divergence is a measure of how different two distributions p and q are:

$$D_{KL}(p||q) \equiv -\int dx\, p(x) \ln\left[\frac{q(x)}{p(x)}\right] \tag{229}$$

The KL divergence satisfies $D_{KL} \geq 0$, with equality holding if and only if $p = q$ almost everywhere. This non-negativity can be proven using the inequality $\ln(x) \leq x - 1$ for $x > 0$:

$$D_{KL}(p||q) \equiv -\int dx\, p(x) \ln\left[\frac{q(x)}{p(x)}\right] \tag{230}$$

$$\geq \int dx\, p(x) \left[\frac{q(x)}{p(x)} - 1\right] \tag{231}$$

$$= \int dx\, q(x) - \int dx\, p(x) = 0 \tag{232}$$

where the final equality follows from the normalization of probability density functions p and q, which integrate to unity.

12.2.2 Neural Density Estimation

Taking x to represent both parameters and data, $\boldsymbol{\theta}, \mathbf{d}$, and given N samples drawn from $p(\boldsymbol{\theta}, \mathbf{d})$, the KL divergence between the true distribution p and the approximation q_φ can be estimated as

$$D_{KL}(p||q) \equiv -\int dx\, p(\boldsymbol{\theta}, \mathbf{d}) \ln\left[\frac{q_\varphi(\boldsymbol{\theta}, \mathbf{d})}{p(\boldsymbol{\theta}, \mathbf{d})}\right] \tag{233}$$

$$\simeq -\frac{1}{N}\sum_{i=1}^{N} [\ln q_\varphi(\boldsymbol{\theta}_i, \mathbf{d}_i) - \ln p(\boldsymbol{\theta}_i, \mathbf{d}_i)]. \tag{234}$$

When optimizing the parameters φ of q, the second term in the sum does not depend on φ and can therefore be treated as a constant. Consequently, minimizing the KL divergence reduces to the optimization problem

$$\varphi = \arg\min_\varphi \left[-\sum_{i=1}^{N} \ln q_\varphi(\boldsymbol{\theta}_i, \mathbf{d}_i)\right]. \tag{235}$$

Neural networks are particularly well suited to this type of optimization problem. Common choices for the parametric family q_φ include Gaussian mixture models and normalizing flows.

12.2.3 Neural Likelihood and Posterior Estimation

The same approach can be applied to construct surrogates for the likelihood function $p(\mathbf{d}|\boldsymbol{\theta})$ or the posterior distribution $p(\boldsymbol{\theta}|\mathbf{d})$ directly. For the likelihood, we observe that

$$D_{KL}(p||q) \equiv -\int dx\, p(\boldsymbol{\theta}, \mathbf{d}) \ln\left[\frac{q_\varphi(\mathbf{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})}\right] \tag{236}$$

$$\simeq -\frac{1}{N}\sum_{i=1}^{N} [\ln q_\varphi(\mathbf{d}_i|\boldsymbol{\theta}_i) - \ln p(\mathbf{d}_i|\boldsymbol{\theta}_i)]. \tag{237}$$

yielding the optimization problem for neural likelihood estimation:

$$\varphi = \arg\min_\varphi \left[-\sum_{i=1}^{N} \ln q_\varphi(\mathbf{d}_i|\boldsymbol{\theta}_i)\right]. \tag{238}$$

An analogous derivation applies to neural posterior estimation. Each approach presents distinct advantages and disadvantages, and several variants exist, including sequential neural likelihood estimation (SNLE),

sequential neural posterior estimation (SNPE), and neural ratio estimation (NRE), which are beyond the scope of this discussion. It is important to note that both NLE and NPE yield functions that depend on both the parameters $\boldsymbol{\theta}$ and the data $\mathbf{d}$.

13 Data Compression

While the methods described above function adequately in low-dimensional settings, they rapidly become computationally intractable as dimensionality increases, due to the exponentially decreasing density of points in the vicinity of the observed data. In typical physics experiments, if there are N data points and M model parameters, the joint distribution resides in an $(N+M)$-dimensional space, which can be prohibitively large. Even when summary statistics are employed, their number often remains too large for efficient inference, necessitating aggressive dimensionality reduction to a manageable set of informative statistics.

Consequently, substantial data compression is typically required.

For a model with M parameters, the maximum compression of summary statistics that avoids information loss (in the limit of the assumptions stated below) reduces the data to M numbers. The question is whether this compression can be achieved in a manner that preserves the information content relevant for parameter inference. The MOPED (Massively Optimized Parameter Estimation and Data compression) algorithm provides one such approach. The Fisher information matrix (subsubsection 4.3.4) provides a theoretical foundation for optimal data compression schemes such as MOPED. For guidance on designing experiments that maximize expected information gain about parameters, see subsection 14.9 on Bayesian Experimental Design.

The MOPED algorithm relies on several key assumptions:

- The data follow a Gaussian sampling distribution;
- The parameter dependence enters through the mean $\boldsymbol{\mu}(\boldsymbol{\theta})$ of the data vector $\mathbf{d}$;
- The data covariance matrix Σ is independent of the model parameters $\boldsymbol{\theta}$;
- The likelihood function can be approximated by a first-order Taylor expansion in the parameters, meaning that derivatives of $\boldsymbol{\mu}$ with respect to $\boldsymbol{\theta}$ beyond the gradient are negligible.

Even when these assumptions are not satisfied exactly, the resulting data compression often retains nearly all the information relevant for constraining the model parameters, making MOPED a remarkably robust compression scheme in practice.

13.1 MOPED Derivation

The MOPED algorithm was originally derived for a different purpose by Heavens et al. (2000, MNRAS, 317, 965). The derivation presented here, following Alsing & Wandelt (MNRAS, 2018, 476, 60), provides a more transparent approach.

Under the assumption of Gaussian data with parameter-independent covariance, the log-likelihood takes the form

$$\ln p(\mathbf{d}|\boldsymbol{\theta}) = \text{constant} - \frac{1}{2}[\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta})]^T \Sigma^{-1} [\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta})] \tag{239}$$

where $\mathbf{d}$ is the N-dimensional data vector, $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the model prediction for the mean as a function of the M-dimensional parameter vector $\boldsymbol{\theta}$, and Σ is the known $N \times N$ covariance matrix.

Expanding $\boldsymbol{\mu}$ to first order in a Taylor series about some fiducial parameter point $\boldsymbol{\theta}_*$, we obtain

$$\ln p(\mathbf{d}|\boldsymbol{\theta}) = \text{cst.} - \frac{1}{2}\left[\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta}_*) - \frac{\partial \boldsymbol{\mu}}{\partial \theta_\alpha}\tilde{\theta}_\alpha\right]^T \Sigma^{-1} \left[\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta}_*) - \frac{\partial \boldsymbol{\mu}}{\partial \theta_\beta}\tilde{\theta}_\beta\right] \tag{240}$$

where $\tilde{\theta}_\alpha \equiv \theta_\alpha - \theta_{*\alpha}$ represents the parameter deviation from the fiducial values, and we employ the Einstein summation convention (summation over repeated indices α and β from 1 to M). The partial derivatives

$\partial\boldsymbol{\mu}/\partial\theta_\alpha$ are N-dimensional vectors evaluated at $\boldsymbol{\theta}_*$. Expanding the quadratic form:

$$\ln p(\mathbf{d}|\boldsymbol{\theta}) = \text{cst.} - \frac{1}{2}[\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta}_*)]^T \Sigma^{-1} [\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta}_*)] \tag{241}$$

$$+ \left(\frac{\partial\boldsymbol{\mu}}{\partial\theta_\alpha}\right)^T \Sigma^{-1} [\mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta}_*)]\tilde{\theta}_\alpha \tag{242}$$

$$- \frac{1}{2}\left(\frac{\partial\boldsymbol{\mu}}{\partial\theta_\alpha}\right)^T \Sigma^{-1} \left(\frac{\partial\boldsymbol{\mu}}{\partial\theta_\beta}\right) \tilde{\theta}_\alpha\tilde{\theta}_\beta. \tag{243}$$

where the two cross terms arising from the expansion have been combined. The first term is simply $\ln p(\mathbf{d}|\boldsymbol{\theta}_*)$, which is independent of $\boldsymbol{\theta}$ and can therefore be absorbed into the constant when considering the parameter dependence of the likelihood.

The key observation is that the data $\mathbf{d}$ enter the likelihood only through the M compressed data values

$$y_\alpha \equiv \mathbf{b}_\alpha^T(\mathbf{d} - \boldsymbol{\mu}_*) \tag{244}$$

where the MOPED vectors are defined as

$$\mathbf{b}_\alpha = \Sigma^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_\alpha}. \tag{245}$$

Here, each $\mathbf{b}_\alpha$ is an N-dimensional vector that weights the data according to both the sensitivity of the model to parameter θ_α and the inverse covariance structure of the noise.

This is the central result: instead of retaining all N components of the original data vector $\mathbf{d}$, we need only the M compressed values y_α, where typically $M \ll N$. Under the stated assumptions, this compression is lossless—the likelihood function computed from $\{y_\alpha\}$ is identical to that computed from the full data vector $\mathbf{d}$, meaning no information about the parameters has been discarded.

Alternatively, the MOPED-compressed data can be converted directly to point estimates of the parameters under the same linear approximation. Maximizing the log-likelihood with respect to $\boldsymbol{\theta}$ yields:

$$0 = \frac{\partial \ln p(\mathbf{d}|\boldsymbol{\theta})}{\partial\theta_\gamma} = y_\gamma - (\mathbf{b}_\alpha^T\Sigma\mathbf{b}_\gamma)\tilde{\theta}_\alpha \tag{246}$$

where we have used $\partial\tilde{\theta}_\beta/\partial\theta_\gamma = \delta_{\beta\gamma}$ (the Kronecker delta). Solving for the parameter deviations gives the maximum likelihood estimate:

$$\tilde{\boldsymbol{\theta}} = D^{-1}\mathbf{y} \tag{247}$$

where the matrix D has elements

$$D_{\alpha\beta} = \mathbf{b}_\alpha^T\Sigma\mathbf{b}_\beta = \left(\frac{\partial\boldsymbol{\mu}}{\partial\theta_\alpha}\right)^T \Sigma^{-1}\left(\frac{\partial\boldsymbol{\mu}}{\partial\theta_\beta}\right). \tag{248}$$

The maximum likelihood parameter estimate is thus

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_* + D^{-1}\mathbf{y}. \tag{249}$$

For simulation-based inference, either $\mathbf{y}$ or $\hat{\boldsymbol{\theta}}$ can serve as compressed data summaries. Both are statistics (deterministic functions of the data $\mathbf{d}$) that, under ideal conditions, preserve all the information content of the original N-dimensional dataset while requiring storage and processing of only M numbers, achieving compression factors that can exceed 10^6 in realistic applications.

13.2 Examples of Intractable Likelihoods

To understand when likelihood-free inference becomes necessary, it is instructive to consider specific examples where the likelihood function is either computationally prohibitive or analytically intractable.

13.2.1 Cosmic Structure Formation

In cosmology, predicting the distribution of matter in the Universe requires simulating the gravitational evolution of billions of particles from the early Universe to the present day. The model parameters $\boldsymbol{\theta}$ include fundamental cosmological quantities such as the dark matter density, the amplitude of primordial fluctuations, and the Hubble constant. The observed data $\mathbf{d}$ consist of measurements such as the positions and properties of galaxies, the cosmic microwave background temperature fluctuations, or gravitational lensing maps.

The likelihood $p(\mathbf{d}|\boldsymbol{\theta})$ is intractable because:

- The forward model requires solving the coupled gravitational N-body problem for $\sim 10^9$ particles, followed by complex baryonic physics (gas cooling, star formation, supernova feedback), implemented through computationally expensive hydrodynamic simulations (Springel, 2005, MNRAS, 364, 1105; Vogelsberger et al., 2014, MNRAS, 444, 1518).
- The mapping from the simulated density field to observable quantities (galaxy positions, luminosities, spectra) involves stochastic processes that cannot be captured by a simple analytic probability distribution.
- Even if one could run the simulation, the high dimensionality of the data ($N \sim 10^6$ galaxy positions and properties) and the complex correlations between data points make it impossible to write down an analytic form for $p(\mathbf{d}|\boldsymbol{\theta})$.

However, one can readily generate synthetic data by running the simulation with different parameter values. Summary statistics such as the two-point correlation function, the power spectrum, or compressed representations via MOPED or neural networks make the inference tractable. Simulation-based inference has been successfully applied to cosmological parameter estimation using galaxy clustering (Alsing et al., 2019, MNRAS, 488, 4440), weak gravitational lensing (Jeffrey et al., 2021, MNRAS, 501, 954), and the Lyman-alpha forest (Pedersen et al., 2021, JCAP, 05, 033).

13.2.2 Selection Effects with Unknown Selection Function

Consider the censored and truncated data examples from section 11. In those cases, we could write down the likelihood explicitly because the selection function (the criterion determining which objects are detected) was known analytically: objects with measured values below a threshold $x_{\min}$ were not detected.

Now suppose the selection function depends on multiple properties in a complex, unknown way. For example:

- An astronomical survey detects galaxies based on their brightness, color, size, and morphology, with detection probability varying in a complicated manner across this multi-dimensional space.
- The detection probability is determined by instrumental effects (e.g., atmospheric conditions, detector sensitivity variations) that are themselves uncertain and must be simulated.
- The survey strategy varies across the sky, introducing spatial variations in the selection function.

In this scenario, the likelihood $p(\mathbf{d}_{\text{detected}}|\boldsymbol{\theta})$ cannot be written down analytically. However, if we have a sufficiently realistic simulator of the survey (including the detection process), we can:

1. Draw model parameters $\boldsymbol{\theta}$ from the prior.
2. Simulate the true population of galaxies given $\boldsymbol{\theta}$.
3. Run the observation simulator, applying the realistic (but analytically intractable) selection effects.
4. Obtain synthetic detected samples that can be compared with the real data.

This is precisely the regime where likelihood-free inference is required. The likelihood is implicit in the simulator but cannot be evaluated directly. Such complex selection functions arise in large astronomical surveys including SDSS (Strauss et al., 2002, AJ, 124, 1810), the Dark Energy Survey (Zuntz et al., 2018, MNRAS, 481, 1149), and LSST (Ivezić et al., 2019, ApJ, 873, 111), where simulation-based inference methods have proven essential for robust cosmological parameter constraints.

13.2.3 Systems Biology and Epidemiology

In modeling the spread of infectious diseases, the parameters $\boldsymbol{\theta}$ might include transmission rates, recovery rates, and population mobility patterns. The data $\mathbf{d}$ consist of daily counts of infections, hospitalizations, and deaths across different regions.

The likelihood is intractable because:

- The forward model is a stochastic agent-based simulation where millions of individuals interact according to complex contact networks (Eubank et al., 2004, Nature, 429, 180; Ferguson et al., 2006, Nature, 442, 448).
- The mapping from model states (individual infection status) to observables (reported case counts) involves reporting delays, testing biases, and other stochastic effects.
- The system exhibits chaotic dynamics, where small changes in initial conditions lead to divergent trajectories, making it impossible to approximate the likelihood as a simple Gaussian.

Despite the intractability of the likelihood, simulators can generate synthetic epidemic curves for any choice of parameters, enabling likelihood-free inference using summary statistics such as peak timing, total case counts, or epidemic growth rates. ABC and neural simulation-based inference have been successfully applied to estimate parameters of influenza (McKinley et al., 2009, PLoS Comp. Bio., 5, e1000456), HIV (Ratmann et al., 2012, Mol. Biol. Evol., 29, 1917), and COVID-19 transmission models (Parag et al., 2020, PLoS Comp. Bio., 16, e1008409).

13.3 Alternatives to MOPED

When the assumptions underlying MOPED are violated—particularly when the information about parameters resides in the covariance structure $\Sigma(\boldsymbol{\theta})$ rather than the mean $\boldsymbol{\mu}(\boldsymbol{\theta})$, or when the data are non-Gaussian—alternative compression schemes based on neural networks become necessary.

The Information Maximizing Neural Network (IMNN; Charnock et al. 2018, PRD, 97, 3004) trains a neural network to construct summary statistics that maximize the Fisher information about the parameters. This approach is more flexible than MOPED and can capture nonlinear dependencies. Graph Neural Networks (Graph NN; Makinen et al. 2022, arxiv 2207.05202) provide another powerful alternative, particularly suited to data with natural graph structure such as point clouds or networks.

14 Additional Topics (unassessed)

This section covers important topics that complement the main course material. These are either advanced extensions or practical techniques that are increasingly important in modern physics research.

14.1 Markov Chain Theory for MCMC (unassessed)

Markov Chain Monte Carlo (MCMC) methods provide the computational foundation for modern Bayesian inference, particularly when analytical solutions are intractable. Understanding why MCMC algorithms work and how to diagnose their failure requires familiarity with the fundamental mathematical properties of Markov chains. This understanding connects directly to the practical MCMC algorithms presented in section 6, providing both theoretical justification and diagnostic tools.

A Markov chain is a sequence of random variables $\{\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots\}$ characterized by the Markov property: the probability of transitioning to the next state depends only on the current state, not on the history of previous states. Mathematically, this property is expressed as:

$$p(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t-1)}, \ldots, \boldsymbol{\theta}^{(0)}) = p(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \tag{250}$$

The transition from one state to the next is governed by a transition kernel $T(\boldsymbol{\theta}'|\boldsymbol{\theta}) = p(\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}'|\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta})$, which specifies the probability distribution over potential next states given the current state. For MCMC to successfully explore the posterior distribution $p(\boldsymbol{\theta}|\mathbf{d})$, the transition kernel must satisfy specific mathematical conditions.

The central requirement is detailed balance, also known as reversibility. A Markov chain satisfies detailed balance with respect to a target distribution $\pi(\boldsymbol{\theta})$ when:

$$\pi(\boldsymbol{\theta})T(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}')T(\boldsymbol{\theta}|\boldsymbol{\theta}') \tag{251}$$

This condition states that the probability flux from state $\boldsymbol{\theta}$ to state $\boldsymbol{\theta}'$ equals the flux in the reverse direction, ensuring microscopic reversibility. While detailed balance is sufficient for π to be a stationary distribution of the chain, it is not strictly necessary; however, it provides a straightforward design principle for constructing MCMC algorithms. If we design a transition kernel satisfying detailed balance with respect to the posterior $p(\boldsymbol{\theta}|\mathbf{d})$, samples from the chain will eventually be distributed according to the posterior.

A distribution $\pi(\boldsymbol{\theta})$ is stationary (or invariant) under transition kernel T if it satisfies:

$$\pi(\boldsymbol{\theta}') = \int \pi(\boldsymbol{\theta})T(\boldsymbol{\theta}'|\boldsymbol{\theta})d\boldsymbol{\theta} \tag{252}$$

Stationarity means that if the chain is distributed according to π at time t, it remains distributed according to π at all subsequent times. Detailed balance implies stationarity, which can be verified by integrating both sides of the detailed balance equation over $\boldsymbol{\theta}$. However, stationarity alone does not guarantee that the chain will reach this distribution from an arbitrary starting point.

For practical MCMC application, the chain must not only have the posterior as its stationary distribution but must also converge to it from any starting point. This convergence property is encapsulated in ergodicity, which requires two key properties: irreducibility and aperiodicity. Irreducibility demands that every state can be reached from every other state, possibly through multiple transitions. Formally, for any sets A and B with $\pi(A) > 0$ and $\pi(B) > 0$, there must exist an integer n such that:

$$\int_A T^{(n)}(\boldsymbol{\theta}'|\boldsymbol{\theta})d\boldsymbol{\theta}' > 0 \quad \text{for some } \boldsymbol{\theta} \in B \tag{253}$$

where $T^{(n)}$ denotes the n-step transition kernel obtained by composing T with itself n times. In practical terms, irreducibility ensures that the chain can reach all regions of the posterior with non-zero probability. This property can fail if the proposal distribution is too narrow relative to the posterior or if the posterior contains disconnected modes that the chain cannot traverse.

Aperiodicity requires that the chain does not cycle through states with a fixed period. Formally, a chain is aperiodic if for all $\boldsymbol{\theta}$ in the support of π:

$$\gcd\{n : T^{(n)}(\boldsymbol{\theta}|\boldsymbol{\theta}) > 0\} = 1 \tag{254}$$

Most MCMC algorithms, including Metropolis-Hastings with continuous proposals, automatically satisfy aperiodicity.

The ergodic theorem establishes the fundamental justification for using MCMC samples to compute posterior expectations. If a Markov chain is irreducible, aperiodic, and has stationary distribution π, then for any starting point $\boldsymbol{\theta}^{(0)}$ and any integrable function f:

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} f(\boldsymbol{\theta}^{(t)}) = \int f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \quad \text{almost surely} \tag{255}$$

This theorem guarantees that time averages along the Markov chain converge to ensemble averages under the posterior distribution, providing the mathematical foundation for using MCMC samples to approximate posterior expectations.

Even when a chain is ergodic, the rate of convergence to the stationary distribution varies dramatically depending on the properties of the transition kernel and the target distribution. The mixing time quantifies the number of iterations required for the chain to approximately reach its stationary distribution from a given initial state. Poor mixing occurs when the posterior contains well-separated modes (causing the chain to become trapped in a single mode), when strong correlations exist between parameters (resulting in slow exploration), or when the proposal distribution is poorly matched to the posterior shape (leading to either excessive rejections or inefficiently small steps).

Consecutive samples in a Markov chain are correlated, with the degree of correlation decaying as the separation between samples increases. The autocorrelation function at lag k is defined as:

$$\rho_k = \frac{\mathbb{E}[(f(\boldsymbol{\theta}^{(t)}) - \mu)(f(\boldsymbol{\theta}^{(t+k)}) - \mu)]}{\mathrm{Var}[f(\boldsymbol{\theta}^{(t)})]} \tag{256}$$

where $\mu = \mathbb{E}[f(\boldsymbol{\theta})]$ under the stationary distribution and the expectation is taken over the stationary distribution. This correlation structure reduces the effective information content of the chain. The effective sample size accounts for these correlations:

$$N_{\text{eff}} = \frac{N}{1 + 2\sum_{k=1}^{\infty} \rho_k} \approx \frac{N}{\tau_{\text{int}}} \tag{257}$$

where τ_{int} is the integrated autocorrelation time and N is the total number of samples. This quantity indicates how many independent samples the N correlated samples are equivalent to for the purpose of estimating expectations.

The Metropolis-Hastings algorithm (section 6) achieves detailed balance through its acceptance probability. Given a proposal distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$, the algorithm accepts a proposed move from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ with probability:

$$\alpha(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \min\left(1, \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right) \tag{258}$$

This acceptance rule can be verified to satisfy detailed balance by considering the probability flux in both directions. The combined proposal-and-acceptance mechanism ensures that the chain has the target distribution π as its stationary distribution regardless of the choice of proposal distribution q, provided q has sufficient support.

Gibbs sampling (section 6) constructs transitions by sampling each parameter from its conditional distribution given all other parameters. Detailed balance is automatically satisfied because the conditional distributions are derived from the joint distribution. Specifically, for a move that updates component i from θ_i to θ_i' while holding $\boldsymbol{\theta}_{-i}$ fixed:

$$\pi(\theta_i, \boldsymbol{\theta}_{-i})p(\theta_i'|\boldsymbol{\theta}_{-i}) = \pi(\theta_i'|\boldsymbol{\theta}_{-i})\pi(\boldsymbol{\theta}_{-i}) = \pi(\theta_i', \boldsymbol{\theta}_{-i}) \tag{259}$$

This identity follows from the definition of conditional probability and the product rule. Hamiltonian Monte Carlo (section 6) uses gradient information to propose distant moves while maintaining high acceptance rates, thereby improving mixing efficiency for posteriors with strong parameter correlations.

Understanding Markov chain theory provides essential diagnostic tools for assessing MCMC performance. Trace plots visualize the parameter values as a function of iteration number, revealing non-convergence (trending behavior) or poor mixing (high autocorrelation). The Gelman-Rubin statistic compares within-chain variance to between-chain variance across multiple chains initialized from dispersed starting points; values substantially exceeding unity indicate non-convergence. Autocorrelation plots display ρ_k as a function of lag k, quantifying how quickly samples decorrelate. The effective sample size provides a single summary statistic measuring the information content of the chain after accounting for autocorrelation. These diagnostics, grounded in Markov chain theory, enable practitioners to assess whether MCMC has produced reliable samples from the posterior distribution.

14.2 Variational Inference (unassessed)

Variational Inference (VI) provides an alternative to Markov Chain Monte Carlo for approximating the posterior distribution, trading exactness for computational speed. While MCMC methods draw samples from the posterior through iterative stochastic processes, variational inference transforms the inference problem into an optimization problem. This transformation becomes essential for very high-dimensional problems where MCMC is computationally prohibitive, particularly in modern applications involving neural networks and large-scale datasets.

The central idea is to approximate the intractable posterior distribution $p(\boldsymbol{\theta}|\mathbf{d})$ with a simpler, tractable distribution $q_\phi(\boldsymbol{\theta})$ from a chosen family of distributions parametrized by variational parameters ϕ. The optimal approximation is found by minimizing the Kullback-Leibler (KL) divergence from the approximating distribution to the true posterior:

$$\phi^* = \arg\min_\phi D_{KL}(q_\phi(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{d})) \tag{260}$$

Since the KL divergence involves the unknown posterior distribution, we cannot optimize it directly. However, minimizing this KL divergence is equivalent to maximizing the Evidence Lower BOund (ELBO), which depends only on quantities we can compute:

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi}[\ln p(\mathbf{d}|\boldsymbol{\theta})] - D_{KL}(q_\phi(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) \tag{261}$$

The ELBO consists of two terms: the expected log-likelihood under the variational distribution (which measures how well the approximation explains the data) and the KL divergence between the variational distribution and the prior (which regularizes the approximation toward the prior). Maximizing the ELBO is a tractable optimization problem that can be solved using gradient-based methods.

The simplest and most commonly used variational family is the mean-field approximation, which assumes that the posterior factorizes across parameters: $q_\phi(\boldsymbol{\theta}) = \prod_{i=1}^n q_i(\theta_i)$. This assumption of independence dramatically simplifies computation but potentially underestimates posterior correlations. Under the mean-field approximation, the optimal factors can be found through coordinate ascent. The optimal distribution for parameter j given fixed distributions for all other parameters satisfies:

$$\ln q_j^*(\theta_j) = \mathbb{E}_{q_{-j}}[\ln p(\mathbf{d}, \boldsymbol{\theta})] + \text{const} \tag{262}$$

where the expectation is taken over all factors except q_j, and the constant ensures proper normalization. This equation shows that the optimal factor for each parameter depends on the expected log-joint distribution under the current approximation for all other parameters, creating a system of coupled equations that can be solved iteratively.

Example: Variational Inference for Bayesian Linear Regression

Problem: A researcher has measurements of input-output pairs and wants to learn a linear relationship between them while quantifying uncertainty in the regression weights. Traditional MCMC methods would require expensive sampling, especially as the dimensionality of the problem grows. The question is: can we obtain an approximate posterior distribution more efficiently by formulating inference as an optimization problem rather than a sampling problem? This example demonstrates

how variational inference provides a fast deterministic alternative to MCMC by finding the best approximation to the posterior within a chosen family of distributions.

Solution:

Consider a linear regression model $y_i = \mathbf{w}^T\mathbf{x}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with known noise variance. The model specification is:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \mathcal{N}(\mathbf{y}; \mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}) \tag{263}$$

$$\pi(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \tau^2\mathbf{I}) \tag{264}$$

For a mean-field variational family, assume $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2))$, a Gaussian with diagonal covariance. The variational parameters are $\phi = \{\boldsymbol{\mu}, \boldsymbol{\sigma}^2\}$.

The ELBO can be written as:

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \mathbb{E}_q[\ln p(\mathbf{y}|\mathbf{X}, \mathbf{w})] - D_{KL}(q(\mathbf{w})||\pi(\mathbf{w})) \tag{265}$$

The expected log-likelihood term is:

$$\mathbb{E}_q[\ln p(\mathbf{y}|\mathbf{X}, \mathbf{w})] = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\mathbb{E}_q[(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})] \tag{266}$$

$$= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}\|^2 + \text{tr}(\mathbf{X}^T\mathbf{X}\text{diag}(\boldsymbol{\sigma}^2))\right] \tag{267}$$

The KL divergence between two Gaussians is:

$$D_{KL}(q||\pi) = \frac{1}{2}\left[\sum_j \left(\frac{\sigma_j^2}{\tau^2} + \frac{\mu_j^2}{\tau^2} - \ln\frac{\sigma_j^2}{\tau^2} - 1\right)\right] \tag{268}$$

Combining these and maximizing with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ yields the optimal variational parameters:

$$\boldsymbol{\mu}^* = (\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{269}$$

$$\text{diag}(\boldsymbol{\sigma}^{2*}) = \sigma^2(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I})^{-1} \tag{270}$$

These are exactly the posterior mean and covariance for Bayesian linear regression. In this Gaussian case, the mean-field approximation with Gaussian factors is exact because the true posterior is Gaussian. However, VI with more restrictive families (e.g., diagonal covariance when the true posterior has correlations) will underestimate posterior uncertainty.

The key insight is that VI converts inference into an optimization problem. For non-conjugate models where the posterior is not Gaussian, VI provides a fast approximate solution, though it may underestimate uncertainty by finding a mode rather than fully exploring the posterior distribution.

Variational inference connects to multiple aspects of Bayesian computation discussed throughout this course. It provides an alternative to MCMC (section 6) that is particularly valuable for high-dimensional problems where sampling becomes prohibitively expensive. Modern variational inference increasingly employs neural networks as flexible variational families, connecting directly to neural density estimation methods in likelihood-free inference (section 12). While variational inference offers substantial computational advantages over sampling-based methods, practitioners must recognize its fundamental trade-off: by restricting the approximation to a chosen family of distributions, VI may underestimate posterior uncertainty, particularly when the true posterior exhibits strong correlations or multimodality that the variational family cannot capture. This limitation makes variational inference most appropriate when computational constraints preclude full MCMC or when approximate solutions with quantified error bounds suffice for the application at hand.

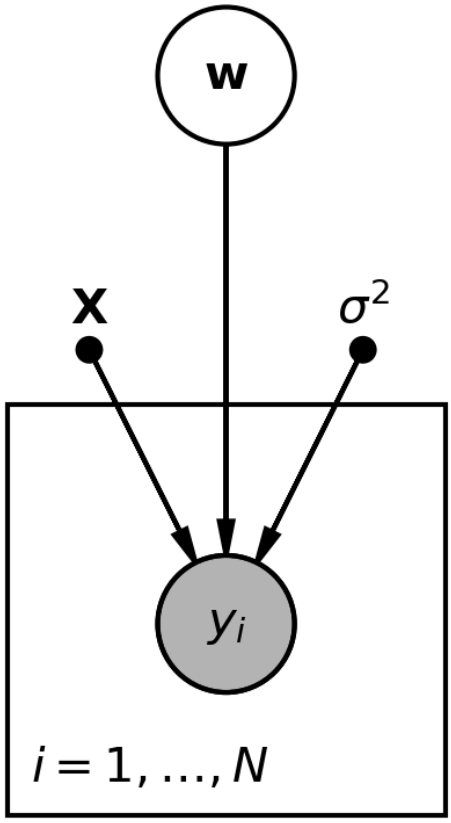

Figure 15: DAG for Bayesian linear regression (subsection 14.2). Weights $\mathbf{w}$ with prior $\mathcal{N}(\mathbf{0}, \tau^2\mathbf{I})$ combine with features $\mathbf{X}$ to generate observations y_i with noise variance σ^2.

14.3 Nested Sampling (unassessed)

Nested sampling provides a specialized algorithm for computing the Bayesian evidence $p(\mathbf{d}|M)$, addressing one of the most challenging computational problems in Bayesian inference. While MCMC methods efficiently explore the posterior distribution, they do not directly compute the evidence integral required for model comparison via Bayes factors. Nested sampling attacks this problem through a clever transformation that converts the multi-dimensional evidence integral into a one-dimensional integral that can be approximated through iterative refinement.

The key insight of nested sampling is to transform the evidence integral using the concept of prior mass. Define $X(\lambda)$ as the fraction of the prior with likelihood exceeding λ:

$$X(\lambda) = \int_{\mathcal{L}(\boldsymbol{\theta})>\lambda} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{271}$$

This prior mass decreases monotonically from $X(0) = 1$ to $X(\infty) = 0$ as the likelihood threshold λ increases. Through this transformation, the evidence can be expressed as a one-dimensional integral over prior mass:

$$\mathcal{Z} = p(\mathbf{d}|M) = \int_0^1 \mathcal{L}(X) dX \tag{272}$$

where $\mathcal{L}(X)$ denotes the likelihood value corresponding to prior mass X. This reformulation replaces integration over the potentially high-dimensional parameter space with integration over the single variable

X.

The nested sampling algorithm exploits this transformation by maintaining a set of "live points" sampled from the prior and iteratively constraining them to regions of increasing likelihood. At each iteration, the algorithm identifies the live point with the lowest likelihood value, records its contribution to the evidence integral, and replaces it with a new point drawn from the prior subject to the constraint that its likelihood exceeds the discarded point's likelihood. This process gradually contracts the live points toward the posterior mode while building up an estimate of the evidence through numerical integration.

The evidence estimate is obtained by approximating the one-dimensional integral as a weighted sum:

$$\ln \mathcal{Z} \approx \ln \sum_{i=1}^{N} w_i \mathcal{L}_i \tag{273}$$

where $w_i = X_{i-1} - X_{i+1}$ are weights representing the prior mass interval associated with each discarded point, and $\mathcal{L}_i$ is the likelihood value at that point. The logarithmic formulation ensures numerical stability when dealing with very small evidence values.

Nested sampling proves essential for model comparison (section 10) when Bayes factors are required, providing a direct method for evidence calculation that MCMC does not offer. By targeting the evidence rather than the posterior distribution, nested sampling complements rather than replaces MCMC, offering a fundamentally different approach to Bayesian computation. Modern implementations including MultiNest, PolyChord, and dynesty have made nested sampling accessible for practical applications across physics and astronomy, particularly for problems involving model selection among competing theories.

14.4 Posterior Predictive Checks (unassessed)

Posterior predictive checks provide a principled framework for assessing whether a Bayesian model adequately captures the structure of observed data. While parameter estimation and model comparison focus on selecting among competing models, posterior predictive checks address the more fundamental question of whether any model under consideration is appropriate for the data at hand. This diagnostic approach compares real data to simulated data generated from the posterior predictive distribution, revealing systematic model inadequacies that parameter inference alone cannot detect.

The posterior predictive distribution describes our predictions for new, unobserved data $\tilde{\mathbf{y}}$ given the observed data $\mathbf{d}$. This distribution is obtained by marginalizing over the posterior uncertainty in the parameters:

$$p(\tilde{\mathbf{y}}|\mathbf{d}) = \int p(\tilde{\mathbf{y}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{d}) d\boldsymbol{\theta} \tag{274}$$

This integral weights each possible prediction $p(\tilde{\mathbf{y}}|\boldsymbol{\theta})$ by the posterior probability of that parameter value, naturally incorporating parameter uncertainty into predictions. In practice, we approximate this integral using posterior samples obtained from MCMC or other inference methods:

$$p(\tilde{\mathbf{y}}|\mathbf{d}) \approx \frac{1}{S} \sum_{s=1}^{S} p(\tilde{\mathbf{y}}|\boldsymbol{\theta}_s) \tag{275}$$

where $\{\boldsymbol{\theta}_s\}_{s=1}^{S}$ are samples from the posterior distribution. Generating data from this approximation is straightforward: for each posterior sample $\boldsymbol{\theta}_s$, we simulate a dataset $\tilde{\mathbf{y}}_s$ from the likelihood $p(\tilde{\mathbf{y}}|\boldsymbol{\theta}_s)$.

To quantify model adequacy, we define test statistics $T(\mathbf{y})$ that capture relevant features of the data, such as the mean, variance, skewness, or more complex summary statistics tailored to the scientific problem. We compute the test statistic for the observed data, $T(\mathbf{d})$, and for each posterior predictive sample, $T(\tilde{\mathbf{y}}_s)$. If the model is adequate, the observed test statistic should appear typical among the posterior predictive values. The posterior predictive p-value quantifies this comparison:

$$p_{\text{pp}} = P(T(\tilde{\mathbf{y}}) \geq T(\mathbf{d})|\mathbf{d}) \tag{276}$$

This p-value represents the probability that a replicated dataset would produce a test statistic at least as extreme as that observed in the real data. Values near 0 or 1 indicate that the observed data is atypical

under the model, suggesting systematic model deficiencies. Unlike classical p-values, posterior predictive p-values properly account for parameter uncertainty and do not require asymptotic approximations.

Posterior predictive checks complement model comparison by operating at a different level of inference. Model comparison distinguishes among competing models, identifying which best explains the data. Posterior predictive checks assess whether any model under consideration adequately describes the data's structure, potentially revealing that all candidate models fail to capture important features. This diagnostic is particularly valuable when validating models constructed for parameter inference, ensuring that estimates and uncertainties derived from inadequate models do not mislead scientific conclusions.

14.5 Robust Statistics (unassessed)

Real experimental data frequently contains outliers arising from measurement errors, instrument malfunctions, or contamination from background processes. Standard Gaussian likelihoods are exquisitely sensitive to such outliers because they assign exponentially decreasing probability to observations far from the mean. A single extreme outlier can dramatically distort parameter estimates when using Gaussian models, pulling the inferred mean toward the outlier and inflating uncertainty estimates. Robust statistical approaches address this fragility by employing likelihood functions that automatically down-weight extreme observations, allowing inference to proceed reliably even in the presence of contaminated data.

The most common robust approach replaces the Gaussian likelihood with a Student-t distribution, which has heavier tails that assign higher probability to extreme values. The Student-t likelihood is given by:

$$p(y_i|\mu,\sigma,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}\sigma}\left[1+\frac{1}{\nu}\left(\frac{y_i-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}} \tag{277}$$

where μ is the location parameter, σ is the scale parameter, and ν is the degrees of freedom parameter that controls the tail heaviness. As $\nu \to \infty$, the Student-t distribution converges to a Gaussian, recovering the standard non-robust analysis. Small values of ν (such as $\nu = 4$) produce heavy tails that accommodate outliers without distorting the location and scale estimates for the bulk of the data. The degrees of freedom parameter can either be fixed based on prior knowledge about outlier frequency or treated as an unknown parameter to be inferred from the data.

An alternative robust approach employs mixture models that explicitly represent outliers as draws from a separate distribution. In this framework, each observation is modeled as originating from either an inlier distribution or an outlier distribution:

$$p(y_i|\boldsymbol{\theta}) = (1-\epsilon)\mathcal{N}(y_i;\mu,\sigma^2) + \epsilon\mathcal{N}(y_i;0,\sigma^2_{\text{out}}) \tag{278}$$

where ϵ is the outlier probability, the first Gaussian describes typical data with parameters of scientific interest (μ,σ), and the second Gaussian with large variance σ^2_{out} captures outliers. This formulation is illustrated in Figure 16.

The mixture model approach offers two advantages over heavy-tailed likelihoods. First, it provides interpretable estimates of the outlier fraction ϵ, quantifying the degree of data contamination. Second, it yields posterior probabilities $p(\text{outlier}_i|\mathbf{d})$ for each observation, allowing researchers to identify which specific data points are likely outliers. This diagnostic information can guide decisions about data cleaning, instrument calibration, or background subtraction.

Robust statistical methods extend the likelihood framework discussed in section 3 by recognizing that the Gaussian assumption often fails for real data. Implementation within hierarchical models (section 9) is straightforward, simply replacing Gaussian likelihoods with Student-t or mixture likelihoods. For experimental physics, where outliers from cosmic ray hits, electrical noise, or environmental disturbances are common, robust methods are not merely technical niceties but essential components of reliable inference. By explicitly modeling outlier-generating processes, Bayesian robust statistics provides both improved parameter estimates and valuable diagnostic information about data quality.

14.6 Maximum Entropy Priors (unassessed)

The maximum entropy principle provides a systematic framework for constructing prior distributions that are maximally non-committal with respect to missing information while respecting known constraints. This

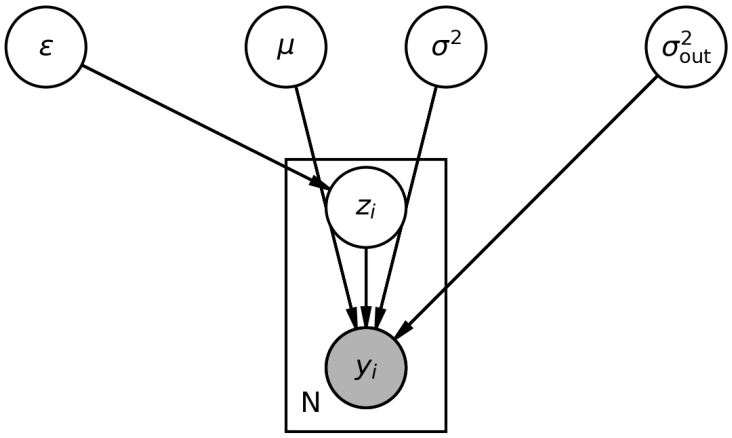

Figure 16: DAG for mixture model with outliers. Each observation y_i is either an inlier or outlier with probability controlled by ϵ, with different Gaussian distributions for each component.

approach resolves the ambiguity inherent in specifying "uninformative" priors by operationalizing the concept of minimal assumption: among all distributions satisfying our constraints, we choose the one that maximizes entropy, thereby avoiding the inadvertent introduction of information we do not possess. This principle connects Bayesian inference to information theory and statistical mechanics, offering both philosophical justification and practical guidance for prior construction.

Shannon entropy quantifies the uncertainty or information content of a probability distribution. For a discrete distribution with probabilities p_i, the entropy is defined as:

$$H = -\sum_i p_i \ln p_i \tag{279}$$

This quantity is maximized when all outcomes are equally probable, reflecting maximum uncertainty. For continuous distributions $p(x)$, the differential entropy generalizes this concept:

$$H = -\int p(x) \ln p(x) dx \tag{280}$$

Higher entropy corresponds to distributions that are more spread out and less concentrated, representing greater prior uncertainty about the parameter value.

The maximum entropy principle states that when our knowledge of a parameter is summarized by certain constraints (such as known moments or support), we should choose the prior that maximizes entropy subject to those constraints. This choice is uniquely determined by the constraints and makes no additional assumptions beyond what is specified. Different constraints lead to different maximum entropy distributions, providing a principled derivation of familiar distributional forms.

When only normalization is required (the distribution must integrate to unity), the maximum entropy distribution is uniform over the allowed range, reflecting complete ignorance within that range. If we additionally know the mean μ of a positive variable but nothing else, the maximum entropy distribution is exponential: $p(x) \propto e^{-\lambda x}$, where λ is determined by the mean constraint. This exponential form arises naturally from maximizing entropy subject to a fixed first moment.

For parameters on the real line with known mean and variance, the maximum entropy distribution is Gaussian. This result provides information-theoretic justification for the ubiquity of Gaussian distributions: when we know only the first two moments, the Gaussian makes the fewest additional assumptions about the

distribution's form. This principle also extends to multivariate parameters, justifying multivariate Gaussian priors when only the covariance structure is specified.

For positive scale parameters where we wish to specify the geometric mean rather than the arithmetic mean, the maximum entropy distribution is the Jeffreys prior $p(x) \propto 1/x$. This connection illuminates why Jeffreys priors, derived from transformation invariance considerations (section 4), also emerge from information-theoretic principles. Both approaches converge on the same functional forms, providing complementary justifications for these reference priors.

The maximum entropy framework offers both practical and conceptual benefits for prior selection. Practically, it provides explicit formulas for constructing priors when certain constraints are known, eliminating guesswork in prior specification. Conceptually, it grounds the notion of "uninformative" or "objective" priors in information theory rather than relying on vague intuitions about lack of knowledge. This connection to statistical mechanics, where maximum entropy principles govern equilibrium distributions, further demonstrates the deep mathematical unity underlying seemingly disparate areas of physics and statistics. When domain knowledge provides genuine constraints on parameter values, the maximum entropy principle translates that knowledge into mathematically precise prior distributions that introduce no spurious information beyond the stated constraints.

14.7 Laplace Approximation (unassessed)

The Laplace approximation provides a fast deterministic method for approximate Bayesian inference by fitting a Gaussian distribution to the posterior at its mode. While MCMC methods explore the full posterior through iterative sampling, the Laplace approximation bypasses sampling entirely, using only local curvature information at the maximum a posteriori (MAP) estimate. This dramatic reduction in computational cost comes at the expense of accuracy: the Gaussian approximation is valid only when the posterior is approximately unimodal and symmetric. Nevertheless, for rapid exploratory analysis or when computational resources preclude full MCMC, the Laplace approximation offers a valuable compromise between speed and rigor.

The approximation begins by finding the posterior mode $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{d})$ through numerical optimization. Around this mode, we expand the log-posterior in a second-order Taylor series:

$$\ln p(\boldsymbol{\theta}|\mathbf{d}) \approx \ln p(\boldsymbol{\theta}^*|\mathbf{d}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{H} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \tag{281}$$

where $\mathbf{H}$ is the Hessian matrix of second derivatives of the negative log-posterior, evaluated at the mode:

$$H_{ij} = -\left.\frac{\partial^2 \ln p(\boldsymbol{\theta}|\mathbf{d})}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}^*} \tag{282}$$

The first-derivative term vanishes because we are expanding around a critical point. Exponentiating this quadratic approximation to the log-posterior yields a Gaussian approximation to the posterior itself:

$$p(\boldsymbol{\theta}|\mathbf{d}) \approx \mathcal{N}(\boldsymbol{\theta}^*, \mathbf{H}^{-1}) \tag{283}$$

The posterior mean is approximated by the MAP estimate $\boldsymbol{\theta}^*$, and the posterior covariance is the inverse Hessian $\mathbf{H}^{-1}$. This inverse Hessian quantifies the local curvature of the log-posterior: sharper curvature (larger eigenvalues of $\mathbf{H}$) corresponds to smaller posterior uncertainty, while flatter directions yield larger posterior variance.

Beyond providing a posterior approximation, the Laplace approximation also yields an estimate of the evidence (marginal likelihood) through integration of the Gaussian approximation:

$$\ln p(\mathbf{d}|M) \approx \ln p(\mathbf{d}|\boldsymbol{\theta}^*) + \ln \pi(\boldsymbol{\theta}^*) + \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{H}| \tag{284}$$

where n is the dimensionality of the parameter space. This formula decomposes the evidence into the likelihood at the mode, the prior density at the mode, and a term proportional to the volume of the posterior (measured by $\det(\mathbf{H}^{-1})$). This evidence approximation enables rapid model comparison without the computational expense of nested sampling or thermodynamic integration.

The Laplace approximation offers a fast alternative to MCMC (section 6) when the Gaussian assumption is reasonable, typically when the posterior is dominated by a single mode with approximately symmetric tails. The Hessian matrix connects directly to the Fisher information matrix discussed earlier, with the Hessian of the log-posterior approaching the Fisher information in the limit of large data. For model comparison (section 10), the Laplace evidence approximation provides quick approximate Bayes factors without requiring specialized algorithms. However, practitioners must recognize the approximation's limitations: multimodal posteriors, heavy tails, or strong parameter correlations can render the Gaussian approximation inadequate, leading to underestimated uncertainties and unreliable inferences.

14.8 Empirical Bayes (unassessed)

Empirical Bayes methods address a practical dilemma that arises in hierarchical modeling: how should we specify hyperparameters for the prior distribution when we lack strong prior knowledge about their values? The fully Bayesian approach places a hyperprior over these hyperparameters and marginalizes over the resulting uncertainty. Empirical Bayes takes a more pragmatic path, estimating hyperparameters directly from the data and then proceeding with inference conditional on these estimates. This compromise between full Bayesian inference and point estimation sacrifices some theoretical purity for computational simplicity and often delivers excellent practical performance, particularly when the data provide strong information about hyperparameter values.

The empirical Bayes framework considers a two-level hierarchical model with three distributional components:

$$p(\mathbf{d}|\boldsymbol{\theta}) \quad \text{(likelihood)} \tag{285}$$

$$p(\boldsymbol{\theta}|\boldsymbol{\eta}) \quad \text{(prior with hyperparameters)} \tag{286}$$

$$p(\boldsymbol{\eta}) \quad \text{(hyperprior)} \tag{287}$$

where $\boldsymbol{\theta}$ represents the parameters of scientific interest, and $\boldsymbol{\eta}$ represents hyperparameters that govern the prior distribution over $\boldsymbol{\theta}$. The fully Bayesian treatment would specify $p(\boldsymbol{\eta})$ and integrate over both $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ simultaneously. This hierarchical structure is illustrated in Figure 17.

Empirical Bayes departs from full Bayesian inference by treating the hyperparameters $\boldsymbol{\eta}$ as unknown fixed quantities to be estimated rather than random variables to be marginalized. The standard estimation procedure, known as Type-II maximum likelihood, selects hyperparameters that maximize the marginal likelihood:

$$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} p(\mathbf{d}|\boldsymbol{\eta}) = \arg\max_{\boldsymbol{\eta}} \int p(\mathbf{d}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})d\boldsymbol{\theta} \tag{288}$$

This marginal likelihood is obtained by integrating out the model parameters $\boldsymbol{\theta}$ at each candidate value of $\boldsymbol{\eta}$, effectively asking which hyperparameter values make the observed data most probable when we account for uncertainty in $\boldsymbol{\theta}$. Once $\hat{\boldsymbol{\eta}}$ is determined, we proceed with standard Bayesian inference using the posterior $p(\boldsymbol{\theta}|\mathbf{d}, \hat{\boldsymbol{\eta}})$, treating the estimated hyperparameters as fixed.

The fully Bayesian alternative avoids point estimation of hyperparameters by jointly inferring $(\boldsymbol{\theta}, \boldsymbol{\eta})$ and marginalizing over hyperparameter uncertainty:

$$p(\boldsymbol{\theta}|\mathbf{d}) = \int p(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathbf{d})d\boldsymbol{\eta} \tag{289}$$

This approach properly propagates uncertainty in hyperparameters into the final inferences about $\boldsymbol{\theta}$, yielding posterior credible intervals that reflect both parameter uncertainty and hyperparameter uncertainty. However, when the data strongly constrain the hyperparameters, the empirical Bayes approximation becomes increasingly accurate because the posterior $p(\boldsymbol{\eta}|\mathbf{d})$ concentrates sharply around $\hat{\boldsymbol{\eta}}$, making the marginalization nearly equivalent to conditioning on the point estimate.

Empirical Bayes extends the hierarchical modeling framework (section 9) by providing a practical approach when hyperprior knowledge is weak or when computational constraints make full hierarchical inference prohibitive. This methodology appears ubiquitously in modern statistical practice, including Gaussian process regression (where kernel hyperparameters are estimated from data), regularized regression (where

regularization strengths are tuned via cross-validation, a form of empirical Bayes), and random effects models (where variance components are estimated from the data). While empirical Bayes sacrifices the full accounting of uncertainty that characterizes pure Bayesian inference, it offers substantial computational advantages and often performs remarkably well when the data provide sufficient information to reliably estimate hyperparameters.

14.9 Bayesian Experimental Design (unassessed)

Bayesian experimental design addresses a fundamental question that arises before data collection begins: among all possible experiments we could conduct, which will provide the most information about the parameters of interest? This question becomes particularly urgent when experiments are expensive, time-consuming, or ethically constrained, making it essential to extract maximum information from each measurement. Rather than choosing experimental designs based on intuition or convention, Bayesian experimental design formalizes the concept of informativeness through utility functions and selects designs that maximize expected utility with respect to the prior distribution over parameters.

The framework begins by defining a utility function $u(\mathbf{d}, \xi)$ that quantifies how valuable a particular dataset $\mathbf{d}$ would be for inference, where ξ parametrizes the experimental design (such as measurement locations, sample sizes, or instrument settings). Since the data have not yet been collected, we cannot evaluate the utility directly. Instead, we compute the expected utility by averaging over all possible datasets that could arise from design ξ:

$$U(\xi) = \mathbb{E}_{\mathbf{d}|\xi}[u(\mathbf{d}, \xi)] \tag{290}$$

This expectation is taken over the prior predictive distribution $p(\mathbf{d}|\xi) = \int p(\mathbf{d}|\boldsymbol{\theta}, \xi)p(\boldsymbol{\theta})d\boldsymbol{\theta}$, which weights possible datasets by their probability under the current state of knowledge encoded in the prior. The optimal experimental design maximizes this expected utility.

Different utility functions formalize different notions of what makes an experiment valuable. The most common choice measures information gain through the Kullback-Leibler divergence from the prior to the posterior: $u = D_{KL}(p(\boldsymbol{\theta}|\mathbf{d})||p(\boldsymbol{\theta}))$. This utility quantifies how much the data would update our beliefs about $\boldsymbol{\theta}$, favoring experiments that are expected to substantially revise the prior. Equivalently, we can use the Shannon information, which measures the reduction in entropy: $u = H[p(\boldsymbol{\theta})] - H[p(\boldsymbol{\theta}|\mathbf{d})]$. Both formulations lead to the same optimal design because the prior entropy $H[p(\boldsymbol{\theta})]$ does not depend on the experimental design. A third common utility uses the expected posterior variance: $u = -\text{trace}[\text{Cov}(\boldsymbol{\theta}|\mathbf{d})]$, favoring designs that yield precise parameter estimates.

The information gain criterion leads to the expected information gain, also known as expected KL divergence:

$$I(\xi) = \int p(\mathbf{d}|\xi) \left[\int p(\boldsymbol{\theta}|\mathbf{d}, \xi) \ln \frac{p(\boldsymbol{\theta}|\mathbf{d}, \xi)}{p(\boldsymbol{\theta})} d\boldsymbol{\theta} \right] d\mathbf{d} \tag{291}$$

This double integral first computes the KL divergence between posterior and prior for each possible dataset (the inner integral), then averages this divergence over all possible datasets weighted by their prior predictive probability (the outer integral). The optimal design is:

$$\xi^* = \arg\max_{\xi} I(\xi) \tag{292}$$

Computing this expected information gain requires solving nested integrals: for each candidate design ξ, we must integrate over possible datasets, and for each dataset, we must compute the posterior and its KL divergence from the prior. This computational burden is substantial but often worthwhile when experimental resources are severely limited.

Bayesian experimental design leverages concepts from information theory, particularly the KL divergence and Shannon entropy, to formalize the intuitive notion that good experiments are those that substantially update our beliefs. The methodology directly answers the question "what measurement should I make next?" by explicitly optimizing for information gain rather than relying on ad hoc heuristics. For expensive experiments in physics, such as particle collider runs, gravitational wave detector configurations, or astronomical survey strategies, Bayesian experimental design can dramatically improve the scientific return on investment by ensuring that limited resources are allocated to the most informative measurements. The approach naturally handles sequential experimental design, where each experiment is optimized based on the posterior

from previous data, enabling adaptive strategies that respond to emerging information as an investigation progresses.

14.10 Model Assessment (unassessed)

Model assessment addresses the question of how well a Bayesian model will perform on new, unseen data, providing a crucial check against overfitting and a practical criterion for model comparison. While Bayes factors compare models based on their ability to explain observed data while penalizing complexity, model assessment focuses directly on predictive performance through cross-validation and information criteria. This predictive perspective is often more aligned with scientific goals: we typically care less about which model best explains the specific dataset at hand than about which model will generalize most reliably to future observations.

Leave-one-out cross-validation (LOO-CV) operationalizes predictive performance by iteratively holding out each data point and assessing how well the model trained on the remaining data predicts the held-out observation. For each data point i, we compute the log predictive density:

$$\text{lpd}_i = \ln p(y_i|\mathbf{y}_{-i}) = \ln \int p(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_{-i})d\boldsymbol{\theta} \tag{293}$$

This quantity measures how probable observation y_i is under the posterior distribution obtained from all other observations $\mathbf{y}_{-i}$. Models that generalize well assign high probability to held-out data, yielding large (less negative) log predictive densities. The total LOO score aggregates these individual contributions:

$$\text{LOO} = \sum_{i=1}^{n} \ln p(y_i|\mathbf{y}_{-i}) \tag{294}$$

Higher LOO scores indicate better out-of-sample predictive performance. This approach naturally guards against overfitting because complex models that fit noise in the training data will perform poorly when predicting held-out observations.

The computational cost of LOO-CV appears prohibitive at first glance: it requires fitting the model n times, once for each held-out observation. However, Pareto smoothed importance sampling (PSIS-LOO) approximates LOO-CV using only the posterior samples from the full dataset. The key observation is that the leave-one-out posterior can be expressed as a weighted version of the full posterior:

$$p(\boldsymbol{\theta}|\mathbf{y}_{-i}) \propto \frac{p(\boldsymbol{\theta}|\mathbf{y})}{p(y_i|\boldsymbol{\theta})} \tag{295}$$

This relationship enables importance sampling: we can reweight posterior samples from $p(\boldsymbol{\theta}|\mathbf{y})$ to approximate expectations under $p(\boldsymbol{\theta}|\mathbf{y}_{-i})$ using importance weights proportional to $1/p(y_i|\boldsymbol{\theta})$. Pareto smoothed importance sampling stabilizes these weights by fitting a generalized Pareto distribution to the tail of the importance weight distribution, preventing individual extreme weights from dominating the approximation. This stabilization makes PSIS-LOO remarkably accurate while requiring no additional model fitting beyond the original posterior computation.

The Widely Applicable Information Criterion (WAIC) provides an alternative approximation to LOO-CV based on information-theoretic considerations. WAIC takes the form:

$$\text{WAIC} = -2(\text{lppd} - p_{\text{WAIC}}) \tag{296}$$

where the log pointwise predictive density (lppd) measures in-sample fit:

$$\text{lppd} = \sum_{i=1}^{n} \ln \mathbb{E}_{\text{post}}[p(y_i|\boldsymbol{\theta})] \tag{297}$$

and the effective number of parameters p_{WAIC} provides a bias correction:

$$p_{\text{WAIC}} = \sum_{i=1}^{n} \text{Var}_{\text{post}}[\ln p(y_i|\boldsymbol{\theta})] \tag{298}$$

The effective number of parameters captures model flexibility: models that adapt strongly to individual data points exhibit high pointwise variance in log-likelihood, resulting in larger p_{WAIC} and thus larger (worse) WAIC values. This penalty for flexibility guards against overfitting in a manner analogous to classical information criteria like AIC and BIC, but WAIC properly accounts for Bayesian uncertainty in parameters through the posterior variance.

Model assessment through LOO-CV and WAIC provides an alternative to Bayes factors (section 10) that emphasizes predictive performance rather than evidential support. This predictive focus is often more practical than evidence calculation because it directly addresses the question scientists care about: which model will perform best for future applications? LOO-CV and WAIC also require less computational effort than evidence calculation via nested sampling or thermodynamic integration, making them particularly attractive for routine model comparison in data analysis workflows. Both metrics are implemented efficiently in modern Bayesian software packages, enabling practitioners to assess predictive performance as a standard component of Bayesian analysis.

14.11 Machine Learning Links (unassessed)

The intersection of Bayesian statistics and modern machine learning represents one of the most active areas of contemporary research, with techniques from each field enriching the other. Traditional machine learning often produces point predictions without quantifying uncertainty, while Bayesian methods naturally provide probability distributions over predictions. Conversely, machine learning's flexible function approximators, particularly neural networks, enable Bayesian inference in settings where classical parametric models prove inadequate. This synergy has produced powerful hybrid approaches that combine the representational flexibility of deep learning with the principled uncertainty quantification of Bayesian inference.

Bayesian neural networks extend conventional neural networks by treating the network weights $\mathbf{w}$ as random variables with a posterior distribution rather than point estimates obtained through optimization. Predictions are made by marginalizing over weight uncertainty:

$$p(y|\mathbf{x}, \mathbf{d}) = \int p(y|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathbf{d}) d\mathbf{w} \tag{299}$$

where $\mathbf{x}$ is the input, y is the prediction, and $\mathbf{d}$ represents the training data. This marginalization naturally produces predictive uncertainties that reflect both aleatoric uncertainty (irreducible noise in the data) and epistemic uncertainty (uncertainty about the model weights). Since exact posterior inference over millions or billions of network weights is intractable, practical implementations employ approximations including variational inference with mean-field or structured approximations, dropout interpreted as approximate Bayesian inference where stochastic weight dropping approximates sampling from the posterior, and stochastic gradient Hamiltonian Monte Carlo that scales MCMC to high-dimensional parameter spaces through mini-batch gradients and momentum-based proposals.

Normalizing flows provide a flexible framework for learning complex probability distributions through invertible neural network transformations. The core idea is to transform a simple base distribution (typically a standard Gaussian) through a sequence of invertible mappings parametrized by neural networks:

$$\mathbf{z} = f_\phi(\boldsymbol{\theta}), \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{300}$$

The density of the transformed variable $\boldsymbol{\theta}$ is obtained via the change of variables formula:

$$p(\boldsymbol{\theta}) = p(\mathbf{z}) \left| \det \frac{\partial f_\phi}{\partial \boldsymbol{\theta}} \right| \tag{301}$$

The invertibility constraint ensures that we can both sample from $p(\boldsymbol{\theta})$ (by sampling $\mathbf{z}$ and applying f_ϕ^{-1}) and evaluate densities (using the Jacobian determinant). Normalizing flows find application across Bayesian inference: they can parametrize flexible approximate posteriors in variational inference, enabling better approximations than mean-field Gaussians; they provide the foundation for neural density estimation in likelihood-free inference (section 12), learning the likelihood or posterior directly from simulations; and they serve as powerful generative models for complex data distributions, particularly in image and text generation.

Score-based models and diffusion processes represent another connection between machine learning and Bayesian inference through the lens of denoising. The score of a distribution, defined as the gradient of the log-density $\nabla_{\mathbf{x}} \ln p(\mathbf{x})$, provides complete information about the distribution up to a normalization constant. Score matching learns this gradient field from data without requiring computation of intractable partition functions:

$$\nabla_{\mathbf{x}} \ln p(\mathbf{x}) = -\mathbb{E}[\nabla_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}_{\text{clean}}\|^2] \tag{302}$$

This connection between scores and denoising enables training neural networks to estimate probability distributions through denoising objectives, with applications to both generative modeling and approximate posterior inference.

These machine learning techniques represent modern extensions of concepts explored in likelihood-free inference (section 12), where neural networks learn summary statistics, density estimators, or discriminators to enable inference when likelihoods are intractable. The variational inference framework discussed earlier connects directly to deep learning optimization, with neural networks parametrizing flexible variational families and stochastic gradient descent providing scalable optimization. Beyond methodological connections, Bayesian approaches to machine learning address the crucial challenge of uncertainty quantification in deployed ML systems, where understanding when a model is uncertain about its predictions proves essential for safe and reliable decision-making. The convergence of Bayesian statistics and machine learning continues to generate new methods that leverage the strengths of both paradigms, expanding the frontiers of what can be learned from data.

14.12 Advanced HMC Variants (unassessed)

Hamiltonian Monte Carlo (section 6) dramatically improved MCMC efficiency by using gradient information to propose distant moves with high acceptance probabilities. However, basic HMC requires manual tuning of two critical parameters: the step size for leapfrog integration and the number of leapfrog steps to take before proposing a new state. Poor choices for these parameters severely degrade performance, yet optimal values depend on the posterior geometry and are difficult to determine a priori. Advanced HMC variants address these limitations through adaptive schemes and geometric insights, making HMC more robust and efficient across diverse inference problems.

The No-U-Turn Sampler (NUTS) eliminates the need to manually specify trajectory length by automatically detecting when the Hamiltonian trajectory begins to double back on itself. The algorithm builds trajectories in both forward and backward directions from the current state, extending the path until the trajectory starts to reverse direction. This reversal is detected by the U-turn criterion: the algorithm stops when $(\boldsymbol{\theta}_+ - \boldsymbol{\theta}_-) \cdot \mathbf{p}_+ < 0$ or $(\boldsymbol{\theta}_+ - \boldsymbol{\theta}_-) \cdot \mathbf{p}_- < 0$, where $\boldsymbol{\theta}_+$ and $\boldsymbol{\theta}_-$ are the forward and backward trajectory endpoints, and $\mathbf{p}_+$ and $\mathbf{p}_-$ are the corresponding momentum variables. This criterion ensures that trajectories explore efficiently without wasting computation on paths that return toward their starting points. NUTS adaptively adjusts trajectory length to the local geometry of the posterior, automatically taking longer trajectories in regions where the posterior is elongated and shorter trajectories near modes where curvature is high. This adaptive behavior eliminates manual tuning of the number of leapfrog steps, making HMC substantially more user-friendly. NUTS has become the default sampler in modern Bayesian software including Stan, PyMC, and NumPyro, enabling practitioners to apply HMC without requiring expert knowledge of algorithm tuning.

Riemannian Manifold HMC extends basic HMC by adapting the mass matrix to the local geometry of the posterior at each point in parameter space. Standard HMC uses a fixed mass matrix that remains constant throughout sampling, which can be inefficient when the posterior exhibits strong curvature that varies across the parameter space. Riemannian HMC employs a position-dependent mass matrix based on the Fisher information matrix:

$$M(\boldsymbol{\theta}) = \mathcal{F}(\boldsymbol{\theta})^{-1} \tag{303}$$

This choice naturally adapts the sampler's behavior to the local posterior geometry: in regions where the Fisher information is large (indicating strong curvature and tight constraints), the mass matrix assigns small masses that enable rapid exploration along these directions, while regions with weak curvature receive larger masses that prevent overshooting. The position-dependent metric defines a Riemannian manifold structure on the parameter space, with the sampler following geodesics on this curved manifold rather than

straight lines in Euclidean space. This geometric perspective yields superior exploration of posteriors with complex curvature, particularly in high dimensions where parameter correlations and varying scales create challenging geometry. However, Riemannian HMC incurs substantial computational overhead: it requires computing the Fisher information matrix and its inverse at each position, and the leapfrog integrator must account for the position-dependent metric. This expense limits Riemannian HMC to problems where the improved exploration justifies the computational cost, typically models with strong parameter-dependent curvature that defeats standard HMC.

These advanced HMC variants represent the state of the art in gradient-based MCMC, offering substantially improved performance over basic Metropolis-Hastings while maintaining theoretical guarantees of convergence to the correct posterior distribution. NUTS has emerged as the default choice for applied Bayesian inference, balancing efficiency and ease of use. Riemannian HMC remains a specialized tool for particularly challenging geometries where its computational cost proves worthwhile. Together, these methods have made HMC accessible to practitioners across science and engineering, enabling routine Bayesian inference for models that would have been computationally intractable with earlier MCMC algorithms.

A Mathematical Notation and Probability Distributions

A.1 Probability Notation Conventions

Throughout this document, we use a compact notation for probability that is standard in the Bayesian statistics literature but may be unfamiliar to some readers. This appendix clarifies these conventions and their precise mathematical meaning.

A.1.1 Random Variables vs Values

The notation $p(x)$ is shorthand for $P(X = x)$, where:

- X is a **random variable** (a function from the sample space to real numbers)
- x is a **specific value** that the random variable can take
- $P(X = x)$ is the probability that random variable X equals the value x

For continuous random variables, $p(x)$ represents a probability density function (PDF), so $p(x)dx$ is the probability that X lies in the infinitesimal interval $[x, x + dx]$. The full notation would be $f_X(x)$ where f_X is the PDF of random variable X.

For discrete random variables, $p(x)$ represents a probability mass function (PMF), where $p(x) = P(X = x)$ is the actual probability that the discrete random variable X takes the specific value x. Unlike continuous distributions, discrete probabilities sum to 1: $\sum_x p(x) = 1$, and individual probabilities can range from 0 to 1.

A.1.2 Examples of the Notation

- $p(\theta)$ means $P(\Theta = \theta)$ or $f_\Theta(\theta)$ - the probability/density that parameter Θ equals value θ
- $p(d)$ means $P(D = d)$ or $f_D(d)$ - the probability/density that data D equals observed value d
- $p(x, y)$ means $P(X = x, Y = y)$ or $f_{X,Y}(x, y)$ - the joint probability/density

A.1.3 Conditional Probability Notation

For conditional distributions, the same convention applies:

- $p(x|y)$ is shorthand for $P(X = x|Y = y)$ or $f_{X|Y}(x|y)$
- This represents the probability/density that $X = x$ given that $Y = y$
- The vertical bar | denotes conditioning: "given that"

A.1.4 Multi-dimensional Case

For vectors, we use bold notation:

- $p(\boldsymbol{\theta})$ means the joint density $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ where $\boldsymbol{\Theta} = (\Theta_1, \Theta_2, \ldots, \Theta_k)$
- $p(\mathbf{d}|\boldsymbol{\theta})$ is the conditional density $f_{\mathbf{D}|\boldsymbol{\Theta}}(\mathbf{d}|\boldsymbol{\theta})$

A.2 Likelihood Function vs Sampling Distribution

A crucial distinction in Bayesian inference is between the likelihood function and the sampling distribution. These represent the same mathematical expression viewed from different perspectives.

A.2.1 The Sampling Distribution

The **sampling distribution** $p(\mathbf{d}|\boldsymbol{\theta})$ is a probability distribution over possible data $\mathbf{d}$ for fixed parameters $\boldsymbol{\theta}$:

- **Fixed**: Parameters $\boldsymbol{\theta}$ are held constant
- **Variable**: Data $\mathbf{d}$ varies over all possible outcomes
- **Normalization**: $\int p(\mathbf{d}|\boldsymbol{\theta})d\mathbf{d} = 1$
- **Interpretation**: "If the true parameters were $\boldsymbol{\theta}$, what data might we observe?"

This answers the forward question: given a model with parameters $\boldsymbol{\theta}$, what is the probability of observing different datasets?

A.2.2 The Likelihood Function

The **likelihood function** $\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}|\mathbf{d})$ is the same expression viewed as a function of parameters $\boldsymbol{\theta}$ for fixed observed data $\mathbf{d}$:

- **Fixed**: Observed data $\mathbf{d}$ is held constant
- **Variable**: Parameters $\boldsymbol{\theta}$ vary over their possible values
- **Not normalized**: $\int \mathcal{L}(\boldsymbol{\theta})d\boldsymbol{\theta} \neq 1$ in general
- **Interpretation**: "Given the observed data $\mathbf{d}$, how likely are different parameter values?"

This answers the inverse question: given observed data, which parameter values are more or less plausible?

A.2.3 Mathematical Relationship

The key insight is that these are the same function:

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{d}) = p(\mathbf{d}|\boldsymbol{\theta}) \tag{304}$$

The difference lies in which argument we treat as the variable:

- **Sampling distribution**: $p(\mathbf{d}|\boldsymbol{\theta})$ with $\boldsymbol{\theta}$ fixed, $\mathbf{d}$ variable
- **Likelihood function**: $\mathcal{L}(\boldsymbol{\theta}|\mathbf{d})$ with $\mathbf{d}$ fixed, $\boldsymbol{\theta}$ variable

A.2.4 Example: Normal Distribution

Consider data $d_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \ldots, N$.

As a sampling distribution (for fixed μ, σ):

$$p(\mathbf{d}|\mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(d_i - \mu)^2}{2\sigma^2}\right) \tag{305}$$

This integrates to 1 over all possible datasets $\mathbf{d}$.

As a likelihood function (for observed $\mathbf{d}$):

$$\mathcal{L}(\mu, \sigma|\mathbf{d}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(d_i - \mu)^2}{2\sigma^2}\right) \tag{306}$$

This does not integrate to 1 over all possible (μ, σ) values.

A.3 Implications for Bayesian Inference

This distinction is crucial for understanding Bayes' theorem:

$$p(\boldsymbol{\theta}|\mathbf{d}) = \frac{p(\mathbf{d}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{d})} = \frac{\mathcal{L}(\boldsymbol{\theta}|\mathbf{d})\, p(\boldsymbol{\theta})}{p(\mathbf{d})} \tag{307}$$

The likelihood $\mathcal{L}(\boldsymbol{\theta}|\mathbf{d})$ tells us how the observed data constrains our beliefs about the parameters, while the prior $p(\boldsymbol{\theta})$ encodes our initial beliefs. The posterior $p(\boldsymbol{\theta}|\mathbf{d})$ combines both sources of information.

The evidence $p(\mathbf{d})$ serves as a normalization constant in parameter inference, ensuring the posterior integrates to 1:

$$p(\mathbf{d}) = \int \mathcal{L}(\boldsymbol{\theta}|\mathbf{d})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{308}$$

This integral marginalizes over all possible parameter values, weighted by their prior probability and likelihood given the data.

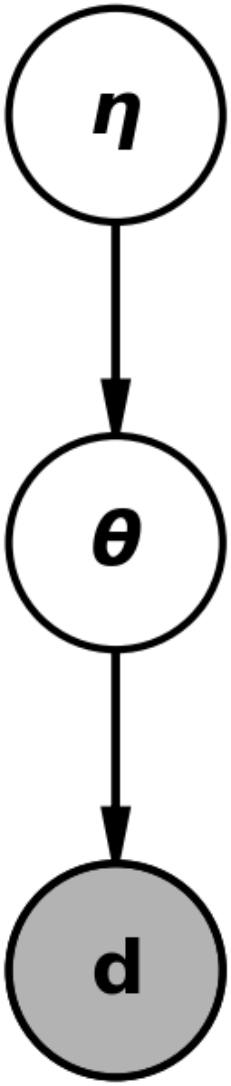

Figure 17: DAG for empirical Bayes two-level hierarchy. Hyperparameters $\boldsymbol{\eta}$ are estimated from data via $\hat{\boldsymbol{\eta}} = \arg\max p(\mathbf{d}|\boldsymbol{\eta})$.