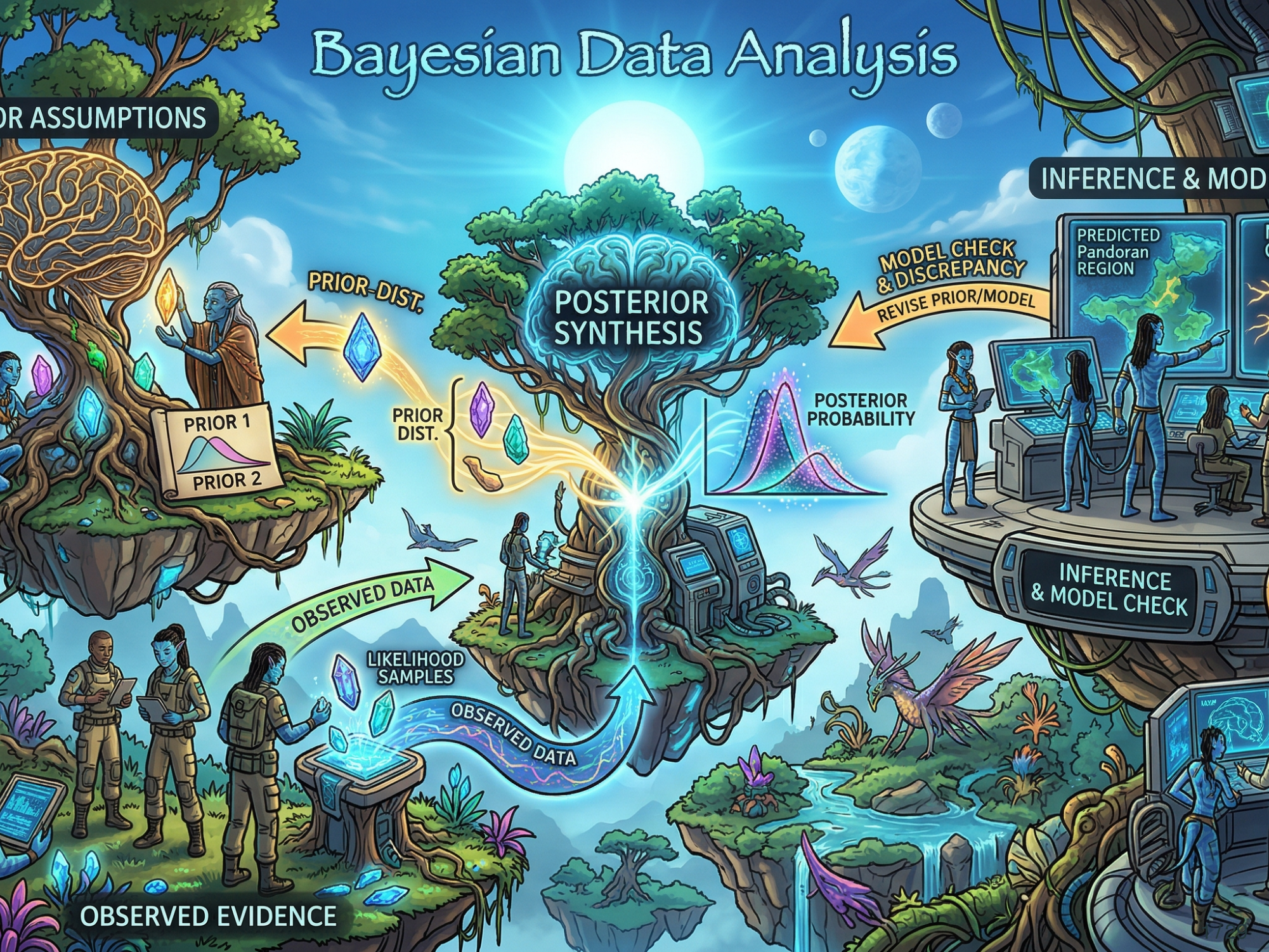

Bayesian Data Analysis
OR ASSUMPTIONS
INFERENCE & MOD
PRIOR-DIST.
POSTERIOR SYNTHESIS
MODEL CHECK & DISCREPANCY
REVISE PRIOR/MODEL
PREDICTED Pandoran REGION
PRIOR 1
PRIOR 2
PRIOR DIST.
POSTERIOR PROBABILITY
INFERENCE & MODEL CHECK
OBSERVED DATA
LIKELIHOOD SAMPLES
OBSERVED DATA
OBSERVED EVIDENCE

# Learning Objectives

- Review the foundational concepts of Bayesian inference
- Review conjugate prior distributions and the exponential family
- Review and discuss approaches to constructing noninformative priors
- Recap computational methods for posterior inference

# Statistical Inference

**Definition**

**Statistical inference** concerns drawing conclusions about a quantity that is not observed.

- Such a quantity of interest is usually referred to as an estimand
- Estimands can be potentially observable (e.g., average height of all humans)
- Or hypothetical (e.g., average change in weight if every human's height increases by 1 inch)

# Estimands and Parametric Models

- In this course, we focus on parametric models where the target estimand is equivalent to certain model parameter(s) under a specified model
- We are usually interested in not only a single estimand, but multiple aspects of the data generating process
  - We need to characterize the whole "system"
  - This often requires inference on multiple parameters simultaneously
- This leads to the challenge of multivariate inference, which is a central theme of this course

# Frequentist vs Bayesian Inference

## Bayesian Perspective

- Conditional: What do we know based on the data I have?
- Optimistic: Find the best prior and model to extract maximum information
- Generative: Requires a generative model mapping parameters to data

## Frequentist Perspective

- Unconditional: What can we say under repeated use of the model?
- Pessimistic: Protect from worst-case scenarios
- Marginal: Some popular frameworks avoid specifying full data distributions

*"There is no philosophy of a data analysis. There are philosophies about inferences. That they aren't the same doesn't make things less valid." (@dan_p_simpson, X, 2024)*

# Bayesian Inference: Overview

**Goal**

Make inference about the parameters and structures of a statistical model using data and quantify relevant uncertainty with probabilities.

Bayesian inference proceeds in three steps:

1. Define a probability model for all components (observables and unobservables): $p(y|\theta)p(\theta)$
2. Find the posterior distribution conditional on observed data:

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

3. Evaluate the goodness of fit of the model

# Bayesian inference

- Every unobserved quantity (parameters, missing data, etc.) is treated as a random variable. For now denote them all by $\theta$.
- $p(\theta)$ is the prior distribution of $\theta$.
- The likelihood $p(y|\theta)$ is the distribution of the observables $y$ given the parameters.
- $p(y) = \int_\Theta p(\theta)p(y|\theta)d\theta$ is the marginal distribution of $y$.
- Usually the most important quantities of interest are based on the posterior:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta)$$

- The posterior distribution is on expectation less variable than the prior distribution:

$$\mathbb{E}(\theta) = \mathbb{E}(\mathbb{E}(\theta|y))$$

$$\mathrm{var}(\theta) = \mathbb{E}(\mathrm{var}(\theta|y)) + \mathrm{var}(\mathbb{E}(\theta|y))$$

# Posterior Inference: Expectations

## Posterior expectation

For a function $g(\theta)$ of the parameters, we are usually interested in:

$$\mathbb{E}(g(\theta)|y) = \int_{\Theta} g(\theta)p(\theta|y)d\theta$$

Some common examples of posterior expectations include:

- Posterior mean: $g(\theta) = \theta$
- Posterior probability of event $A$: $g(\theta) = \mathbf{1}_A(\theta)$
- Posterior credible intervals (the reverse problem): the $(1-\alpha)$ equal-tailed interval $[q_{\alpha/2},\ q_{1-\alpha/2}]$, where the posterior quantile $q_p$ satisfies

$$\int_{-\infty}^{q_p} p(\theta \mid y)\, d\theta = p$$

# Marginal and Predictive Inferences

## Marginal Posteriors

When $\theta$ is multivariate, we may be interested in only some components:

$$p(\theta_1|y) = \int_{\Theta_2} p(\theta_1, \theta_2|y) d\theta_2$$

## Posterior Predictive Distribution

For an unobserved variable $z$ whose value we need to predict:

$$p(z|y) = \int_{\Theta} p(z|\theta, y) p(\theta|y) d\theta$$

In many cases, $p(z|\theta, y)$ simplifies to $p(z|\theta)$ when $z$ is independent of observed data given $\theta$.

# Binomial example

- Let $y \sim Bin(n, \theta)$ and the prior is $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$.
- The likelihood is $p(y|\theta) \propto \theta^{y}(1-\theta)^{n-y}$.
- So the posterior is

$$p(\theta|y) \propto \theta^{y+a-1}(1-\theta)^{n-y+b-1}$$

  i.e., $\theta|y \sim Beta(Y+a, n-Y+b)$.
- This is a simple example of conjugate prior distributions.
- **Definition** Let $\mathcal{F} = \{p(y|\theta), \theta \in \Theta\}$ be a family of sampling distributions. A class $\mathcal{P}$ is said to be a conjugate family for $\mathcal{F}$ if for all $p \in \mathcal{F}$ and $p(\theta) \in \mathcal{P}$ then $p(\theta|y) \in \mathcal{P}$.
- Conjugate prior distributions are usually easy to understand due to their analytical forms, thus they are usually used in simple problems or as building blocks for complex problems.

# Exponential Family: Definition

## General Form

For $n$ i.i.d observations $y_i$ from $p(\cdot|\theta)$, the exponential family has the form:

$$p(y_i|\theta) = f(y_i)g(\theta)e^{\phi(\theta)^T u(y_i)}$$

The key components of exponential family distributions are:

- $\phi(\theta)$: the natural parameter (often same dimension as $\theta$)
- $u(y_i)$: sufficient statistic

Distributions in the exponential family have natural conjugate priors.

# Exponential Family: Likelihood and Sufficiency

- For $y = (y_1, \ldots, y_n)$, the likelihood for $\theta$ is:

$$\ell(\theta) = p(y|\theta) \propto g(\theta)^n e^{\phi(\theta)^T t(y)}$$

  where $t(y) = \sum_i u(y_i)$ is the sufficient statistic

- Only $t(y)$ depends on data in the exponential term. All information about $\theta$ is captured here

# Exponential Family: Conjugate Prior and Posterior

- If the prior has the form:

$$p(\theta) = C(\eta, v) g(\theta)^{\eta} e^{\phi(\theta)^T v}$$

  then the posterior is of the same conjugate form

- The posterior predictive distribution for $z = (z_1, \ldots, z_m) \sim_{iid} p(z|\theta)$ is:

$$p(z|y) = \Big( \prod_{i=1}^{m} f(z_i) \Big) \frac{C(\eta + n, v + t(y))}{C(\eta + n + m, v + t(y) + t(z))}$$

- This closed-form expression is a key advantage of conjugate families

# Normal Model with Known Variance

- Consider a random sample of size $n$ from $N(\theta, \sigma^2)$ with known variance
- The sampling distribution can be reduced to: $\bar{y}|\theta \sim N(\theta, \sigma^2/n)$
- The conjugate prior for $\theta$ is normal: $\theta \sim N(\mu_0, \tau_0^2)$
- The posterior is

$$\begin{aligned} p(\theta|y) &\propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)\exp\left(-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2\right) \\ &= \exp\left(-\frac{1}{2}\left[\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)\theta^2 - 2\left(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{y}}{\sigma^2}\right)\theta\right] + \text{const}\right) \end{aligned}$$

- Completing the square shows the posterior of $\theta$ is $N(\mu_n, \tau_n^2)$ where:

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}, \quad \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}$$

# Normal Model with Unknown Mean and Variance

- When $\sigma^2$ is unknown, let $\sigma^2 \sim$ Inv-Gamma$(\alpha, \beta)$ and $\theta|\sigma^2 \sim N(\theta_0, \sigma^2/\lambda_0)$
- This gives a Normal-Inverse-Gamma (NIG) distribution on $(\theta, \sigma^2)$ with parameters $(\mu_0, \lambda_0, \alpha, \beta)$
- Using $\sum_i (y_i - \theta)^2 = \underbrace{\sum_i (y_i - \bar{y})^2}_{S} + n(\bar{y} - \theta)^2$, the posterior is

$$p(\theta, \sigma^2|y) \propto (\sigma^2)^{-(\alpha + n/2 + 3/2)} \exp\left(-\frac{1}{\sigma^2}\left[\beta + \frac{S}{2} + \frac{n(\bar{y} - \theta)^2 + \lambda_0(\theta - \mu_0)^2}{2}\right]\right)$$

which is also NIW with updated parameters after completing the squares:

$$\mu_n = \frac{\lambda_0 \mu_0 + n\bar{y}}{\lambda_0 + n}, \quad \lambda_n = \lambda_0 + n, \quad \alpha_n = \alpha + n/2$$

$$\beta_n = \beta + \frac{1}{2}\sum_i (y_i - \bar{y})^2 + \frac{n\lambda_0}{\lambda_0 + n}\frac{(\bar{y} - \mu_0)^2}{2}$$

# PARAMETERIZATION

- Notice that there are several different parameterizations of the normal conjugate family in the literature, e.g.,
  - It is common to use precision ($1$/variance) instead of variance in normal models.
  - Note that
    $$\sigma^2 \sim \text{InvGamma}(\text{shape} = \alpha, \text{scale} = \beta)$$
    is equivalent to
    $$1/\sigma^2 \sim \text{Gamma}(\text{shape} = \alpha, \text{rate} = \beta)$$
    or
    $$1/\sigma^2 \sim \text{Gamma}(\text{shape} = \alpha, \text{scale} = 1/\beta)$$
  - Another common parameterization you will see in literature is $\text{InvGamma}(\alpha/2, \beta/2)$.
  - Scaled inverse-$\chi^2$ parameterization is also used, e.g., extensively in BDA3. The distribution Inv-$\chi^2(\nu_0, \sigma_0^2)$ is equivalent to
    $$\text{InvGamma}(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}).$$
    Here $\nu_0$ is the degree of freedom and $\sigma_0^2$ is the scale parameter.
- Also note that independent priors on $\mu$ and $\sigma^2$ leads to conditional conjugacy, which is still useful in Gibbs samplers.

# Noninformative prior distributions

- Where does the prior distribution come from in general? One school of thought is to seek objectivity by choosing a prior distribution that plays a minimal role in the posterior distribution.
- Such priors are usually described as vague, flat, diffuse, noninformative, etc.
- One possible approach to construct such a noninformative prior is by investigating the posterior distribution, for example, in the normal example,

$$\theta|y \sim N\left(\frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\right)$$

  which is approximately $N(\bar{y}, \sigma^2/n)$ if $\tau_0 \to \infty$.
- As $\tau_0 = \infty$, the prior becomes $p(\theta) \propto 1$, which is not a valid distribution. In such case, where $p(\theta)$ does not integrate to $1$, we call it improper.
- Sometimes (not always!) improper priors lead to proper posterior density, i.e., $\int p(\theta|y)d\theta$ is finite for any $y$.

# Jeffreys prior

- Another approach to define noninformative prior was introduced by Harold Jeffreys. The idea is to consider a prior that expresses the same beliefs after one-to-one transformations of the parameter.
- Jeffreys prior is given by

$$p(\theta) \propto |I(\theta)|^{1/2}$$

where $I(\theta)$ is the Fisher Information matrix consisting of

$$I_{ij}(\theta) = \mathbb{E}\Big( - \frac{\partial^2 \log p(y|\theta)}{\partial \theta_i \partial \theta_j} \Big)$$

- Note that there is also a class of prior called reference priors, which chooses a prior that maximizes some measure of distance or divergence to the posterior. For one dimensional $\theta$, it is equivalent to Jeffreys prior.

# Jeffreys prior

- To see how Jeffreys prior remains invariant upon transformations of $\theta$, let us consider one-dimensional $\theta$ here for simplicity.
- Consider a one-to-one transformation $\phi = g(\theta)$, for any $p(\theta)$, the equivalent prior for $\phi$ is $p(\phi) = p(\theta)|\frac{d\theta}{d\phi}|$.
- First notice that under some mild regularity conditions,

$$I(\theta) = \mathbb{E}\big( - \frac{d^2 \log p(y|\theta)}{d\theta^2}\big) = \mathbb{E}\big((\frac{d \log p(y|\theta)}{d\theta})^2\big)$$

- So for $\phi = g(\theta)$, let $p(y|\theta)$ and $q(y|\phi)$ denote the two sampling distributions and shorten $\log p(y|\theta)$ into $\ell(\theta)$ and $\log q(y|\phi)$ into $\ell(\phi)$, we have

$$\begin{aligned} I(\phi) &= \int \Big(\frac{d}{d\phi}\ell(\phi)\Big)^2 q(y|\phi)dy \\ &= \int \Big(\frac{d}{d\phi}\ell(\phi)\Big)^2 p(y|\theta)dy \\ &= \int \Big(\frac{d}{d\theta}\ell(\theta)|_{\theta=g^{-1}(\phi)}\Big)^2 |\frac{d\theta}{d\phi}|^2 p(y|\theta)dy \\ &= I(\theta)|\frac{d\theta}{d\phi}|^2 \end{aligned}$$

- So Jeffreys prior for $\phi$ is $p(\phi) \propto |I(\phi)|^{1/2} = I(\theta)^{1/2}|\frac{d\theta}{d\phi}|$, which is the same as the prior implied by the Jeffreys prior on $\theta$.

# Jeffreys prior in more than one dimensions

- Consider $y \sim N(\mu, \sigma^2)$, we can find the separate Jeffreys priors for the two parameters by

$$I(\mu) = \frac{1}{\sigma^2} \propto 1 \quad I(\sigma) \propto \frac{1}{\sigma^2}$$

  so $p(\mu) \propto 1$ and $p(\sigma) \propto (\frac{1}{\sigma^2})^{1/2}$.
- However, if we consider the two parameters together,

$$I(\mu, \sigma) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}$$

  so $p(\mu, \sigma) \propto (\frac{1}{\sigma^4})^{1/2}$, which is not the same as the product of independent Jeffreys prior.
- Notice that if we further consider $y_i \sim N(\mu_i, \sigma^2)$ for $i = 1, ..., m$, Jeffreys prior for the joint parameter vector is

$$p(\mu_1, ..., \mu_m, \sigma) \propto (\sigma^{-2(m+1)})^{1/2}$$

  which implies the marginal prior $p(\sigma) \to 0$ as $m \to \infty$.
- In general, Jeffery's principle can be tricky when considering more than one parameters.

# Weakly Informative Priors

- Pure noninformativeness is sometimes questionable in practice
- A density that is flat in one parameterization may be highly informative for another
- When likelihood is strong, the posterior is not sensitive to a wide range of "flat-ish" priors
- When likelihood is weak, prior information is usually required to make valid inference from the scientific perspective
- There is a growing literature on weakly informative priors, which are often a better choice than pure noninformativeness (see Canvas references)

# Simulation

- A natural way of making inference about the posterior distribution is to obtain samples of $p(\theta|y)$.
- Suppose a sample $\theta^{(1)}, ..., \theta^{(M)}$ of $p(\theta|y)$ is available then we can approximate

$$\mathbb{E}(g(\theta)|y) \approx \frac{1}{M} \sum_{m=1}^{M} g(\theta^{(m)})$$

- In some cases, such samples are easy to obtain from using direct sampling. Consider $k$-dimensional $\theta = (\theta_1, ..., \theta_k)$, if we can factorize it into blocks, e.g.,

$$p(\theta) = p(\theta_1)p(\theta_2|\theta_1) \cdots p(\theta_k|\theta_1, ..., \theta_{k-1})$$

  where each distribution can be sampled directly, then we can sample each $\theta_i$ recursively.
- When direct sampling is not available, we need other tools to draw the samples from a target distribution $p(\theta|y)$.

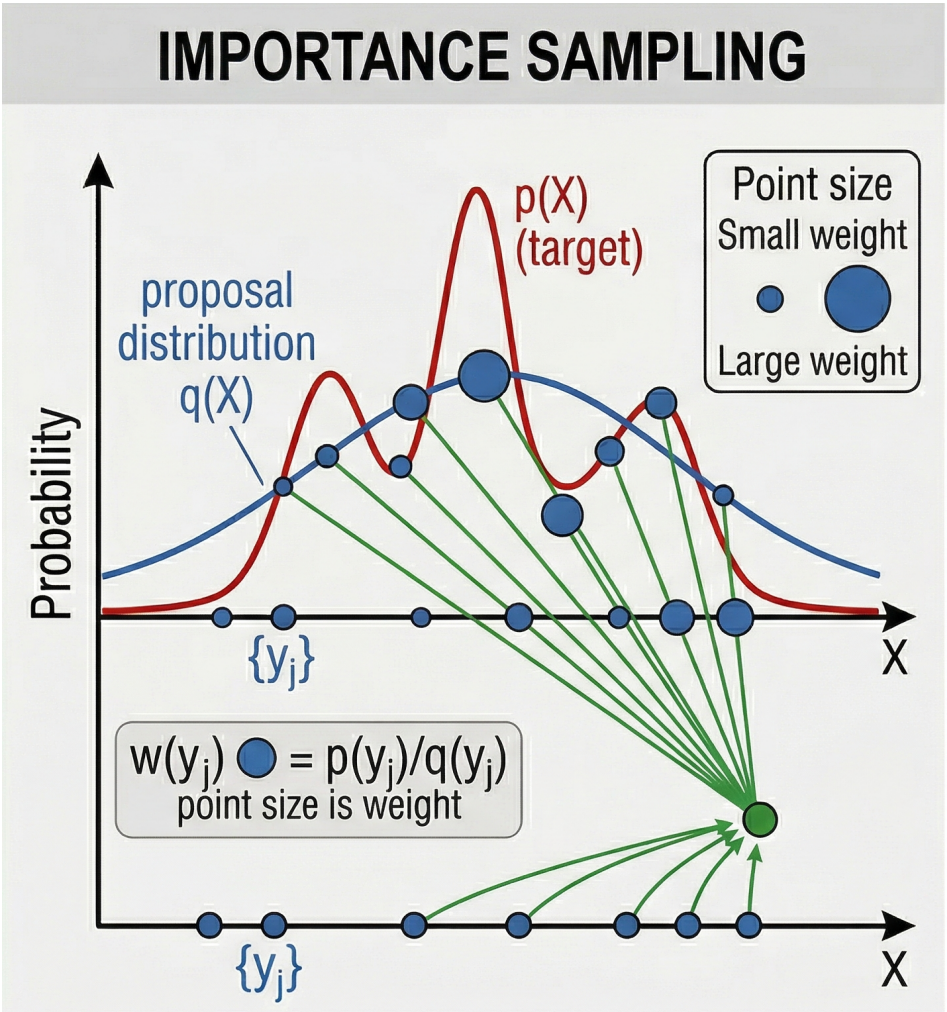
# The normalizing constant problem

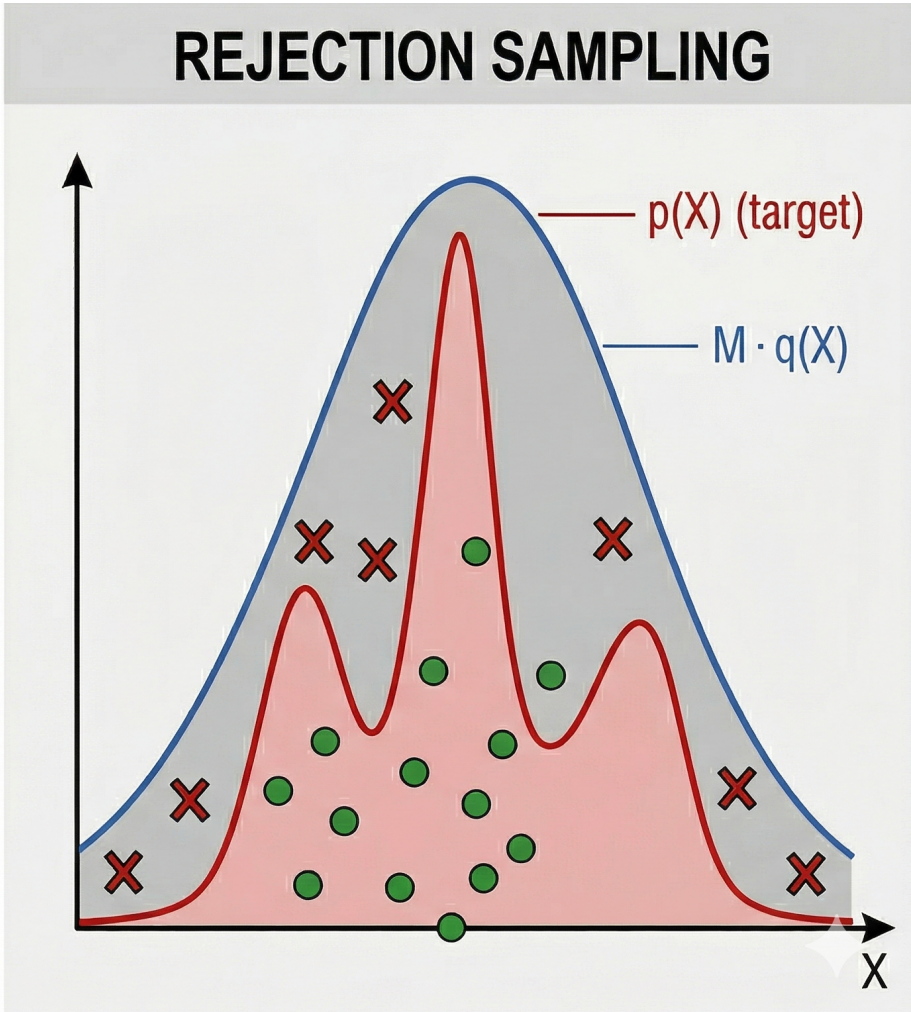
- Often times we only know $p(\theta|y)$ up to a proportional constant, i.e.,

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \propto p(y|\theta)p(\theta)$$

- The normalizing constant $\int p(y|\theta)p(\theta)d\theta$ is usually intractable, i.e., prohibitively expensive to compute.
- However, for low-dimensional $\theta$ in simple problems, sometimes it is fine to do grid approximation for the normalizing constant.
- That is, evaluate $p(y|\theta)p(\theta)$ on a grid of $s = 1, ..., S$ of $\theta$, and compute the normalizing constant with summation.
- Samples of $\theta \sim p(\theta|y)$ can be drawn by sampling from the discrete distribution of $S$ points, or using inverse CDF sampling, i.e., generating $u \sim Unif[0, 1]$ and let $\theta = F^{-1}(u)$.
- The range and finesse grid needs to be chosen carefully to avoid missing important areas in the posterior.

# Importance and rejection sampling



# Importance sampling

- Sometimes we do not need to evaluate the posterior $p(\theta|y)$ but only some posterior expectation $\mathbb{E}(g(\theta)|y) = \int_\Theta g(\theta)p(\theta|y)d\theta$.
- We can use a density $h(\theta)$ that is easy to sample from and adjust the difference,

$$\mathbb{E}(g(\theta)|y) = \int_\Theta \frac{g(\theta)}{h(\theta)} p(\theta|y)h(\theta)d\theta \approx \frac{1}{M}\sum_{m=1}^{M} g(\theta^{(m)})w(\theta^{(m)})$$

  where $\theta^{(m)} \sim H(\theta)$ and $w(\theta) = \frac{p(\theta|y)}{h(\theta)}$ is called the importance weight.
- If we know $p(\theta|y)$ up to a proportional constant, we can compute weights that are $w^\star(\theta) = Cp(\theta|y)/h(\theta)$ for some unknown $C$. Then

$$\mathbb{E}(g(\theta)|y) \approx \frac{\sum_{m=1}^{M} g(\theta^{(m)})w^\star(\theta^{(m)})}{\sum_{m=1}^{M} w^\star(\theta^{(m)})}$$

- This works because

$$\frac{1}{M}\sum_{m=1}^{M} w^\star(\theta^{(m)}) \approx \mathbb{E}_{h(\theta)}(w^\star(\theta)) = \int C\frac{p(\theta|y)}{h(\theta)}h(\theta)d\theta = C$$

- For $\theta$ where $h(\theta)$ is much smaller than $p(\theta|y)$, $w(\theta)$ is large, but it is difficult to get those samples, thus it is usually better to choose $h(\theta)$ with a heavier tail.

# Rejection sampling

- Following the similar idea, we can obtain samples from the posterior by sampling from a different but known distribution and filter the samples.
- A rejection scheme to sample from the posterior distribution can be set in the following way. Let $p^\star(\theta|y)$ be the unnormalized posterior.
- Step 1: Obtain a sample from a density $h(\theta)$. This needs to be as good an approximation to $p(\theta|y)$ as possible. The ratio $p^\star(\theta|y)/h(\theta)$ needs to be bounded above. Let $M$ be the upper bound.
- As an example, we can take $M = \sup_{\theta\in\Theta} \ell(\theta)$, i.e., the MLE's likelihood.
- Then with $h(\theta) = p(\theta)$, we can see it provides a valid upper bound:

$$M \geqslant \ell(\theta) = p(y|\theta) = \frac{p^\star(\theta|y)}{p(\theta)} = \frac{p^\star(\theta|y)}{h(\theta)}$$

- Step 2: Accept the sample as a draw from $p^\star(\theta|y)$ with probability

$$\frac{p^\star(\theta|y)}{Mh(\theta)}$$

- As with importance sampling, the efficiency of this scheme depends on the ability to find a good proposal distribution $h(\theta)$.

# Markov Chain Monte Carlo (MCMC)

- Rejection and importance sampling suffer when $\theta$ is high-dimensional
- MCMC idea: Build a Markov Chain that is easy to simulate and has equilibrium distribution equal to $p(\theta|y)$
- Requires choosing appropriate transition kernels $q(\theta, \phi)$ that are:
  - Homogeneous, irreducible, and aperiodic
  - Easy to sample from
  - Produce limiting distribution equal to $p(\theta|y)$
- We will examine MCMC in more detail later; you should have been already familiar with two basic techniques from previous courses

# Gibbs Sampling

**Core Idea**

Iteratively sample from the full conditional distributions of the parameters.

Suppose $\theta$ consists of $k$ blocks $\theta_1, \ldots, \theta_k$. Denote $\theta_{-i}$ as all blocks excluding $\theta_i$.

The Gibbs sampling algorithm works as follows:

1. Initialize $\theta$
2. Sample $\theta_1^{(m+1)}$ from $p(\theta_1 | \theta_2^{(m)}, \ldots, \theta_k^{(m)}, y)$
3. Sample $\theta_2^{(m+1)}$ from $p(\theta_2 | \theta_1^{(m+1)}, \theta_3^{(m)}, \ldots, \theta_k^{(m)}, y)$
4. Continue for all components, cycle through until convergence

The key advantage of Gibbs sampling is that it only requires knowledge of the full conditionals, not the joint distribution.

# Metropolis-Hastings Sampler

## Algorithm

Choose an initial $\theta$, then at each iteration $m$:

1. Sample a proposal $\theta^\star$ from $q(\theta^\star|\theta^{(m)})$
2. Compute the acceptance probability:

$$\alpha = \min\left(1, \frac{p(\theta^\star|y)}{p(\theta^{(m)}|y)} \cdot \frac{q(\theta^{(m)}|\theta^\star)}{q(\theta^\star|\theta^{(m)})}\right)$$

3. Accept $\theta^{(m+1)} = \theta^\star$ with probability $\alpha$; otherwise $\theta^{(m+1)} = \theta^{(m)}$

Some special cases of the Metropolis-Hastings algorithm are worth noting:

- Metropolis: Symmetric proposal, e.g., $\theta^\star \sim N(\theta^{(m)}, \sigma^2)$
- Independent MH: Proposal independent of $\theta^{(m)}$, e.g., $\theta^\star \sim N(\mu, \sigma^2)$

Unlike rejection sampling, MCMC generates dependent samples — this must be accounted for in convergence diagnostics.