Lecture 11: Bayesian Modeling

CS109A: Introduction to Data Science

Harvard University

Course:	CS109A: Introduction to Data Science
Lecture:	Lecture 11
Instructors:	Pavlos Protopapas, Kevin Rader, Chris Gumb
Topics:	Bayesian inference, Bayes' Rule, Prior and Posterior distributions, Normal-Normal model, Conjugate priors, Connection to Ridge and Lasso regression

Contents

1 Introduction to Bayesian Thinking

Lecture Overview

This lecture introduces **Bayesian modeling**, a fundamentally different approach to statistical inference that treats parameters as random variables with probability distributions. Instead of finding a single "best estimate," Bayesian methods provide a complete probability distribution representing our uncertainty about parameters.

Key Learning Objectives:

- Understand the difference between Frequentist and Bayesian perspectives
- Master Bayes' Rule and its application to inference
- Learn about prior distributions and their role in modeling
- Understand the Normal-Normal conjugate model
- Discover the deep connection between Bayesian inference and regularization (Ridge/Lasso)

1.1 Two Paradigms of Statistics

Statistics has two major schools of thought about how to approach inference: the **Frequentist** approach and the **Bayesian** approach. Both are valid and useful, but they have fundamentally different philosophies.

Definition: Frequentist Statistics

In the frequentist approach:

- **Parameters are fixed but unknown constants.** There exists one "true" value of θ that we're trying to estimate.
- **Data is random.** Each time we collect data, we get a random sample from some population.
- **Probability means long-run frequency.** When we say "95% confidence," we mean that if we repeated the experiment infinitely many times, 95% of our intervals would contain the true parameter.
- **Inference is about the sampling distribution.** We reason about what would happen across many hypothetical repetitions of our experiment.

Definition: Bayesian Statistics

In the Bayesian approach:

- **Parameters are random variables.** We express our uncertainty about θ using probability distributions.
- **Data is fixed (once observed).** After we collect our data, it's no longer random—it's what we actually observed.
- **Probability means degree of belief.** When we say "95% probability," we mean we're 95% confident that θ is in a certain range.
- **Inference is about updating beliefs.** We start with prior beliefs and update them based on observed data.

Example: Treasure Hunt Analogy

Imagine you're searching for buried treasure.

Frequentist Approach:

- There is ONE exact location where the treasure is buried (fixed, unknown parameter)
- You collect clues (data) and estimate the location
- Your estimate is a single point: "The treasure is at coordinates (100, 200)"

Bayesian Approach:

- You start with a "probability map" showing where you think the treasure might be
- As you gather clues (data), you **update your map**
- Your result is the entire updated map: "60% chance it's in region A, 30% in region B, 10% in region C"

1.2 Why Learn Bayesian Methods?

Bayesian modeling offers several advantages:

1. **Intuitive interpretation:** "There's a 95% probability that θ is between 2 and 5" is more natural than the frequentist confidence interval interpretation.
2. **Incorporating prior knowledge:** You can formally include expert opinion, historical data, or physical constraints into your model.

3. **Flexibility:** Bayesian methods can handle complex hierarchical models where data is measured at multiple levels.
4. **Sequential updating:** As new data arrives, you can update your beliefs continuously without starting from scratch.
5. **Connection to regularization:** Ridge and Lasso regression have elegant Bayesian interpretations.
6. **Meta-analysis:** Combining results from multiple studies is natural in the Bayesian framework.

Important Note

When to Choose Each Approach

Frequentist methods are often preferred when:

- You need quick, computationally efficient solutions
- Classical approaches are well-established in your field
- You want to avoid specifying prior beliefs

Bayesian methods are often preferred when:

- You have relevant prior information to incorporate
- You need probabilistic statements about parameters
- Your model is hierarchical or complex
- Data arrives sequentially

In practice, **both methods often give similar results** when the prior is uninformative and the sample size is large.

2 Bayes' Rule: The Foundation

Everything in Bayesian inference flows from a single formula: **Bayes' Rule**.

2.1 The Formula

Definition: Bayes' Rule

For parameters θ and data X:

$$P(\theta|X) = \frac{P(X|\theta) \cdot P(\theta)}{P(X)} \tag{1}$$

Or in terms of probability density functions (PDFs):

$$f(\theta|X) = \frac{f(X|\theta) \cdot f(\theta)}{f(X)} \tag{2}$$

Let's understand each term:

- **P**$(\theta|$**X**$)$ — **Posterior Probability**
  - The probability distribution of θ **after** seeing the data X
  - This is what we want to find—our updated belief about θ
  - Read as: "the probability of theta **given** the data"
- **P**$($**X**$|\theta)$ — **Likelihood**
  - The probability of observing data X **if** the parameter were θ
  - This is the same likelihood we've seen in MLE!
  - Read as: "the probability of the data given theta"
- **P**(θ) — **Prior Probability**
  - The probability distribution of θ **before** seeing any data
  - Represents our initial belief or background knowledge
  - Read as: "the prior probability of theta"
- **P**$($**X**$)$ — **Evidence (Marginal Likelihood)**
  - The total probability of observing the data across all possible values of θ
  - Computed as $P(X) = \int P(X|\theta)P(\theta)d\theta$
  - Acts as a **normalizing constant** to ensure probabilities sum to 1

2.2 The Proportionality Form

Since $P(X)$ doesn't depend on θ, we often write:

Key Summary

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \tag{3}$$

$$f(\theta|X) \propto f(X|\theta) \cdot f(\theta) \tag{4}$$

The posterior is **proportional to** the likelihood times the prior. We can always normalize later to get a proper probability distribution.

Example: Intuitive Interpretation

Think of Bayes' Rule as a belief-updating mechanism:

1. **Start with your prior belief** $P(\theta)$: What you thought before seeing any data
2. **Observe data and compute likelihood** $P(X|\theta)$: How likely is this data under different values of θ?
3. **Update your belief** $P(\theta|X)$: Combine prior and likelihood to get your new, informed belief

Values of θ that:

- Were believed likely (high prior) AND make the data likely (high likelihood) → stay high in posterior
- Were believed unlikely (low prior) OR make the data unlikely (low likelihood) → low in posterior

3 A Discrete Example: Coin Selection

Let's work through a complete example to solidify these concepts.

3.1 Problem Setup

You have three coins in your pocket:

- **Coin A:** Biased toward tails with $P(\text{heads}) = 0.1$
- **Coin B:** Fair coin with $P(\text{heads}) = 0.5$
- **Coin C:** Biased toward heads with $P(\text{heads}) = 0.9$

You randomly pull out one coin and flip it 4 times, getting **3 heads and 1 tail**.

Question: What's the probability that you drew each coin?

3.2 Step-by-Step Solution

Step 1: Define the Prior Distribution

Before flipping the coin, you have no information about which coin was selected. Since you drew randomly from three coins:

$$P(\theta = 0.1) = P(\theta = 0.5) = P(\theta = 0.9) = \frac{1}{3} \tag{5}$$

This is a **discrete uniform prior** over the three possible values.

Step 2: Define the Likelihood (Data Model)

The number of heads in 4 flips follows a **Binomial distribution**:

$$P(X = k|\theta) = \binom{4}{k}\theta^k(1-\theta)^{4-k} \tag{6}$$

We observed $X = 3$ (3 heads out of 4 flips).

Step 3: Calculate the Likelihood for Each Coin

For each possible value of θ, calculate $P(X = 3|\theta)$:

$$P(X = 3|\theta = 0.1) = \binom{4}{3}(0.1)^3(0.9)^1 = 4 \times 0.001 \times 0.9 = \mathbf{0.0036} \tag{7}$$

$$P(X = 3|\theta = 0.5) = \binom{4}{3}(0.5)^3(0.5)^1 = 4 \times 0.125 \times 0.5 = \mathbf{0.25} \tag{8}$$

$$P(X = 3|\theta = 0.9) = \binom{4}{3}(0.9)^3(0.1)^1 = 4 \times 0.729 \times 0.1 = \mathbf{0.2916} \tag{9}$$

Observation: Coin C ($\theta = 0.9$) makes the data most likely, which makes sense—if a coin is biased toward heads, we're more likely to see 3 heads!

Step 4: Apply Bayes' Rule

Calculate the unnormalized posterior for each value:

$$P(\theta = 0.1|X = 3) \propto 0.0036 \times \frac{1}{3} = 0.0012 \tag{10}$$

$$P(\theta = 0.5|X = 3) \propto 0.25 \times \frac{1}{3} = 0.0833 \tag{11}$$

$$P(\theta = 0.9|X = 3) \propto 0.2916 \times \frac{1}{3} = 0.0972 \tag{12}$$

Step 5: Normalize to Get the Posterior

Sum the unnormalized values: $0.0012 + 0.0833 + 0.0972 = 0.1817$

Divide each by this sum:

$$P(\theta = 0.1|X = 3) = \frac{0.0012}{0.1817} \approx \mathbf{0.7\%} \tag{13}$$

$$P(\theta = 0.5|X = 3) = \frac{0.0833}{0.1817} \approx \mathbf{45.8\%} \tag{14}$$

$$P(\theta = 0.9|X = 3) = \frac{0.0972}{0.1817} \approx \mathbf{53.5\%} \tag{15}$$

Key Summary

Results:

Coin	Prior	Likelihood	Posterior
A ($\theta = 0.1$)	33.3%	0.0036	**0.7%**
B ($\theta = 0.5$)	33.3%	0.2500	**45.8%**
C ($\theta = 0.9$)	33.3%	0.2916	**53.5%**

Interpretation: Before seeing the data, each coin was equally likely (33.3% each). After observing 3 heads in 4 flips, we now believe Coin C (biased toward heads) is most likely (53.5%), Coin B (fair) is also plausible (45.8%), and Coin A (biased toward tails) is nearly ruled out (0.7%).

3.3 The Effect of Sample Size

Important Note

What happens if we flip more coins?

If instead of 4 flips with 3 heads, we had **40 flips with 30 heads** (same proportion: 75%):

- The likelihood for Coin A ($\theta = 0.1$) would be **astronomically small** (essentially 0)
- The likelihoods for Coins B and C would shift even more toward C
- The posterior would show: A $\approx$ 0%, B $\approx$ 35%, C $\approx$ 65%

Key Insight: With more data, the posterior becomes more concentrated. The data "speaks louder" than the prior as sample size increases.

4 Choosing Prior Distributions

One of the most important (and sometimes controversial) aspects of Bayesian modeling is choosing the **prior distribution**. There are three main approaches:

4.1 Informative Priors

Definition: Informative Prior

A prior distribution that encodes specific knowledge or beliefs about the parameter before seeing data.

When to use:

- You have expert knowledge about reasonable parameter values
- You have data from previous studies
- Physical or logical constraints exist on the parameter

Example: Temperature Prediction

Predicting tomorrow's noon temperature in Boston during October:

Prior Information:

- Historical average for this date: 60°F
- Standard deviation of day-to-day variation: 5°F

Informative Prior:

$$\mu \sim \text{Normal}(\mu_0 = 60, \sigma_0^2 = 25) \tag{16}$$

This says: "I believe the temperature will be around 60°F, probably within 50–70°F (95% of the time within ±2 standard deviations)."

4.2 Uninformative (Vague) Priors

Definition: Uninformative Prior

A prior distribution designed to have minimal influence on the posterior, letting the data dominate the inference.

When to use:

- You have no prior knowledge about the parameter
- You want to be "objective" and let data speak for itself
- You're concerned about prior sensitivity

Example: Success Probability

Estimating the success probability p of a new medical treatment:

Uninformative Prior:

$$p \sim \text{Uniform}(0, 1) \tag{17}$$

This says: "I have no idea what p is. Any value between 0 and 1 is equally plausible before seeing the data."

4.3 Conjugate Priors

Definition: Conjugate Prior

A prior distribution that, when combined with a particular likelihood function, produces a posterior distribution of the **same family** as the prior.

Why use conjugate priors?

- **Mathematical convenience:** The posterior has a known closed-form solution
- **Interpretability:** Prior parameters often have intuitive meanings
- **Computational efficiency:** No need for numerical integration or simulation

Key Summary

Common Conjugate Prior-Likelihood Pairs:

Likelihood	Parameter	Conjugate Prior	Posterior
Binomial	p (success probability)	Beta	Beta
Poisson	λ (rate)	Gamma	Gamma
Normal (known σ^2)	μ (mean)	Normal	Normal
Normal	$1/\sigma^2$ (precision)	Gamma	Gamma
Exponential	λ (rate)	Gamma	Gamma

5 The Normal-Normal Model

The most important conjugate model for continuous data is the **Normal-Normal model**, where both the data likelihood and the prior on the mean are normal distributions.

5.1 Model Setup

Likelihood (Data Model):

$$X_1, X_2, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2) \quad \text{(i.i.d.)} \tag{18}$$

- μ is the unknown population mean (the parameter we want to estimate)
- σ^2 is the known variance (we assume this is fixed and known for simplicity)

Prior Distribution:

$$\mu \sim \text{Normal}(\mu_0, \sigma_0^2) \tag{19}$$

- μ_0 is the **prior mean** — our best guess for μ before seeing data
- σ_0^2 is the **prior variance** — how uncertain we are about this guess
- These are called **hyperparameters** (parameters of the prior distribution)

5.2 The Posterior Distribution

Because Normal-Normal is a conjugate pair, the posterior is also Normal:

$$\mu | X \sim \text{Normal}(\mu_n, \sigma_n^2) \tag{20}$$

where the **posterior mean** is:

$$\mu_n = \frac{\sigma^2 \cdot \mu_0 + n \cdot \sigma_0^2 \cdot \bar{X}}{\sigma^2 + n \cdot \sigma_0^2} \tag{21}$$

and the **posterior variance** is:

$$\sigma_n^2 = \frac{\sigma^2 \cdot \sigma_0^2}{\sigma^2 + n \cdot \sigma_0^2} \tag{22}$$

5.3 Understanding the Posterior

Let's dissect these formulas to understand what they're telling us:

Key Information

Posterior Mean as a Weighted Average:
The posterior mean μ_n can be rewritten as:

$$\mu_n = w \cdot \mu_0 + (1 - w) \cdot \bar{X} \tag{23}$$

where $w = \frac{\sigma^2}{\sigma^2 + n \cdot \sigma_0^2}$

This is a **weighted average** of:

- μ_0: the prior mean (what we believed before)
- $\bar{X}$: the sample mean (what the data says)

The weight w depends on:

- **Sample size** n: More data → smaller w → more weight on $\bar{X}$
- **Prior uncertainty** σ_0^2: Larger σ_0^2 → smaller w → more weight on $\bar{X}$
- **Data variance** σ^2: Larger σ^2 → larger w → more weight on μ_0

Important Note

Limiting Cases:

1. **As** $n \to \infty$ (lots of data):
   - $\mu_n \to \bar{X}$ (posterior mean approaches sample mean)
   - $\sigma_n^2 \to 0$ (posterior variance shrinks to zero)
   - **Data dominates**: With enough data, the prior becomes irrelevant
2. **As** $\sigma_0^2 \to 0$ (very confident prior):
   - $\mu_n \to \mu_0$ (posterior stays at prior mean)
   - **Prior dominates**: A very strong prior can't be overridden by data
3. **As** $\sigma_0^2 \to \infty$ (uninformative prior):
   - $\mu_n \to \bar{X}$ (posterior mean equals sample mean)
   - $\sigma_n^2 \to \sigma^2/n$ (same as frequentist standard error!)
   - **Data dominates**: Bayesian results match frequentist results

Example: Numerical Example

Setup:

- Prior: $\mu \sim \text{Normal}(\mu_0 = 100, \sigma_0^2 = 25)$
- Data: $n = 10$ observations with $\bar{X} = 110$, known $\sigma^2 = 100$

Calculate Posterior:

$$\mu_n = \frac{100 \cdot 100 + 10 \cdot 25 \cdot 110}{100 + 10 \cdot 25} = \frac{10000 + 27500}{350} = \frac{37500}{350} = 107.14 \tag{24}$$

$$\sigma_n^2 = \frac{100 \cdot 25}{100 + 250} = \frac{2500}{350} = 7.14 \tag{25}$$

Result: $\mu | X \sim \text{Normal}(107.14, 7.14)$

Interpretation:

- Prior mean was 100, data mean was 110
- Posterior mean (107.14) is pulled toward the data but not all the way
- Posterior standard deviation is $\sqrt{7.14} \approx 2.67$
- 95% Credible Interval: $107.14 \pm 1.96 \times 2.67 = [101.9, 112.4]$

6 Bayesian Point and Interval Estimation

Once we have the posterior distribution, we can extract useful summaries.

6.1 Point Estimates

The posterior is a full distribution, but sometimes we need a single "best guess":

- **Posterior Mean**: $E[\theta|X]$
  - The expected value of the posterior distribution
  - Minimizes squared error loss
  - Most commonly used
- **Posterior Mode (MAP)**: $\arg\max_\theta f(\theta|X)$
  - The value where the posterior is highest
  - Called **Maximum A Posteriori (MAP)** estimate
  - Connects to Ridge and Lasso (more on this later!)
- **Posterior Median**: The 50th percentile
  - The value that splits the posterior in half
  - Minimizes absolute error loss

For a Normal posterior, all three are equal (because the Normal is symmetric).

6.2 Credible Intervals

Definition: Credible Interval

A $100(1-\alpha)\%$ **credible interval** is a range $[a, b]$ such that:

$$P(a \leq \theta \leq b|X) = 1 - \alpha \tag{26}$$

For a Normal posterior $\mu|X \sim \text{Normal}(\mu_n, \sigma_n^2)$:

$$95\% \text{ Credible Interval} = \mu_n \pm 1.96 \times \sigma_n \tag{27}$$

Critical: Credible vs. Confidence Intervals

This is one of the most important distinctions in statistics!

Bayesian 95% Credible Interval:

- **Interpretation:** "Given the data we observed, there is a 95% probability that θ lies in this interval."
- The parameter θ is random, the interval is fixed (once calculated)
- **This is the intuitive interpretation most people want!**

Frequentist 95% Confidence Interval:

- **Interpretation:** "If we repeated this experiment many times and computed a confidence interval each time, 95% of those intervals would contain the true θ."

- The parameter θ is fixed, the interval is random (varies across experiments)
- **Warning:** You CANNOT say "there's a 95% probability θ is in this interval"

7 Bayesian Linear Regression

We can apply Bayesian principles to linear regression by placing prior distributions on the regression coefficients.

7.1 Model Setup

Likelihood (Same as OLS):

$$y_i \sim \text{Normal}(\beta_0 + \beta_1 x_i, \sigma^2) \tag{28}$$

Unknown Parameters: $\beta_0, \beta_1, \sigma^2$

Prior Distributions (Conjugate):

$$\begin{align} \beta_0 &\sim \text{Normal}(\mu_{\beta_0}, \sigma^2_{\beta_0}) \tag{29} \\ \beta_1 &\sim \text{Normal}(\mu_{\beta_1}, \sigma^2_{\beta_1}) \tag{30} \\ 1/\sigma^2 &\sim \text{Gamma}(a_0, \lambda_0) \tag{31} \end{align}$$

7.2 What Bayesian Regression Provides

Instead of single point estimates, we get **entire posterior distributions** for each parameter:

- $\beta_0 | X, y$ has a posterior distribution
- $\beta_1 | X, y$ has a posterior distribution
- $\sigma^2 | X, y$ has a posterior distribution

This allows us to make probabilistic statements like:

- "There's a 97% probability that $\beta_1 > 0$"
- "The 95% credible interval for β_1 is [2.1, 4.8]"
- "Given the data, there's less than 5% chance that $|\beta_1| > 10$"

8 Connection to Ridge and Lasso

One of the most beautiful insights in statistics is the connection between Bayesian inference and regularization methods like Ridge and Lasso.

8.1 Recall: Regularized Loss Functions

Ordinary Least Squares (OLS):

$$\text{Loss}_{OLS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{32}$$

Ridge Regression (L2 penalty):

$$\text{Loss}_{Ridge} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{33}$$

Lasso Regression (L1 penalty):

$$\text{Loss}_{Lasso} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{34}$$

8.2 The MAP Connection

In Bayesian inference, the **MAP (Maximum A Posteriori)** estimate is:

$$\hat{\beta}_{MAP} = \arg\max_{\beta} f(\beta|X) = \arg\max_{\beta}[f(X|\beta) \cdot f(\beta)] \tag{35}$$

Taking the logarithm (which preserves the maximum):

$$\hat{\beta}_{MAP} = \arg\max_{\beta}[\log f(X|\beta) + \log f(\beta)] \tag{36}$$

Equivalently, minimizing the negative:

$$\hat{\beta}_{MAP} = \arg\min_{\beta}[-\log f(X|\beta) - \log f(\beta)] \tag{37}$$

Key observation:

- $-\log f(X|\beta)$ is proportional to the **sum of squared errors** (for Normal likelihood)
- $-\log f(\beta)$ is the **penalty term** from the prior!

8.3 Ridge = Normal Prior

Definition: Ridge Regression as Bayesian MAP

If we place a **Normal prior centered at zero** on each β_j:

$$\beta_j \sim \text{Normal}(0, \tau^2) \tag{38}$$

Then:

$$-\log f(\beta_j) = -\log\left(\frac{1}{\sqrt{2\pi\tau^2}} e^{-\beta_j^2/2\tau^2}\right) = \text{constant} + \frac{\beta_j^2}{2\tau^2} \tag{39}$$

This is proportional to β_j^2, the **L2 penalty**!

Key Summary

Ridge Regression is Bayesian MAP with Normal Prior

- The penalty $\lambda \sum \beta_j^2$ comes from assuming $\beta_j \sim N(0, \tau^2)$
- The regularization parameter λ is inversely related to prior variance τ^2
- Larger λ = smaller prior variance = stronger belief that β_j should be near zero

Intuition: The Normal prior says "I believe coefficients are probably small (near zero) with a smooth, bell-shaped probability." This pulls estimates toward zero but doesn't make them exactly zero.

8.4 Lasso = Laplace Prior

Definition: Lasso Regression as Bayesian MAP

If we place a **Laplace (Double Exponential) prior** on each β_j:

$$f(\beta_j) = \frac{1}{2b} e^{-|\beta_j|/b} \tag{40}$$

Then:

$$-\log f(\beta_j) = \text{constant} + \frac{|\beta_j|}{b} \tag{41}$$

This is proportional to $|\beta_j|$, the **L1 penalty**!

Key Summary

Lasso Regression is Bayesian MAP with Laplace Prior

- The penalty $\lambda \sum |\beta_j|$ comes from assuming a Laplace prior on β_j
- The Laplace distribution has a **sharp peak at zero**
- This creates **sparse solutions** where some $\beta_j = 0$ exactly

Intuition: The Laplace prior says "I strongly believe most coefficients should be exactly zero." The sharp peak at zero gives high probability mass to $\beta_j = 0$, encouraging sparsity (feature selection).

Important Note

Visual Comparison: Normal vs. Laplace Priors

Normal Distribution (Ridge):

- Smooth, bell-shaped curve
- Gradual decrease away from zero
- Coefficients are pulled toward zero but rarely exactly zero

Laplace Distribution (Lasso):

- Sharp, peaked curve at zero
- Rapid decrease away from zero
- High probability mass at exactly zero → sparse solutions

8.5 Summary: Regularization as Bayesian Inference

Key Information

Method	**Penalty**	**Bayesian Prior**	**Effect**
OLS	None	Flat/Improper	No shrinkage
Ridge	$\lambda \sum \beta_j^2$	Normal$(0, \tau^2)$	Shrink toward zero
Lasso	$\lambda \sum \lvert\beta_j\rvert$	Laplace$(0, b)$	Shrink + Sparsity

Key Insight: The choice of regularization is implicitly a choice of prior belief about the coefficients!

9 Computational Methods: MCMC

When conjugate priors don't apply or models become complex, we need computational methods to approximate the posterior distribution.

9.1 The Problem

The posterior is:

$$f(\theta|X) = \frac{f(X|\theta)f(\theta)}{f(X)} = \frac{f(X|\theta)f(\theta)}{\int f(X|\theta')f(\theta')d\theta'} \tag{42}$$

The integral in the denominator is often **intractable** (impossible to compute analytically).

9.2 The Solution: Simulation

Instead of computing the posterior exactly, we **draw samples** from it:

$$\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)} \sim f(\theta|X) \tag{43}$$

With enough samples, we can approximate any property of the posterior:

- Posterior mean: $E[\theta|X] \approx \frac{1}{N}\sum_{i=1}^{N} \theta^{(i)}$
- Posterior variance: $\text{Var}[\theta|X] \approx \frac{1}{N-1}\sum_{i=1}^{N}(\theta^{(i)} - \bar{\theta})^2$
- Credible interval: Sort samples and find the 2.5th and 97.5th percentiles

9.3 MCMC: Markov Chain Monte Carlo

Definition: MCMC

Markov Chain Monte Carlo is a class of algorithms that generate samples from a target distribution by constructing a Markov chain whose stationary distribution is the target.

Intuition: Imagine a random walker exploring a landscape (the posterior distribution). The walker:

- Tends to stay in high-probability regions (peaks)
- Occasionally visits low-probability regions (valleys)
- After walking long enough, the proportion of time spent in each region matches the posterior probability

9.4 Gibbs Sampling

For multi-parameter models, **Gibbs Sampling** is a popular MCMC algorithm:

Idea: Instead of sampling all parameters at once, sample each parameter **one at a time**, conditional on the current values of the others.

Algorithm (for 3 parameters $\theta_1, \theta_2, \theta_3$):

1. Initialize: $\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}$

2. For $t = 1, 2, \ldots, N$:
   - Sample $\theta_1^{(t)}$ from $f(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, X)$
   - Sample $\theta_2^{(t)}$ from $f(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, X)$
   - Sample $\theta_3^{(t)}$ from $f(\theta_3|\theta_1^{(t)}, \theta_2^{(t)}, X)$
3. Discard early samples (burn-in period) to remove initialization effects
4. Use remaining samples as draws from the joint posterior

Important Note

Burn-in Period

The initial samples depend heavily on where we started (the initialization). We discard the first several thousand samples (the "burn-in" period) to ensure the remaining samples are representative of the posterior.

10 Quiz Review: Categorical Variables

Before diving into Bayesian modeling, let's review an important concept from the quiz about interpreting categorical (dummy) variables in regression.

10.1 Setup

Consider predicting hours spent on homework based on course registration (4 groups):

- AC 209 (baseline/reference group)
- CS 1090
- CSCI E-109
- STAT 109

The model produces:

$$\hat{y} = 11.0 - 2.0 \cdot x_{CS1090} + 3.5 \cdot x_{CSCIE109} + 5.0 \cdot x_{STAT109} \tag{44}$$

10.2 Interpretation

- **Intercept (11.0):** The predicted hours for the **baseline group** (AC 209). This is also the sample mean for AC 209 students.
- **Coefficient for CS 1090 (-2.0):** CS 1090 students spend, on average, **2 fewer hours** than AC 209 students. Predicted hours: $11.0 - 2.0 = 9.0$
- **Coefficient for CSCI E-109 (+3.5):** CSCI E-109 students spend **3.5 more hours** than AC 209. Predicted hours: $11.0 + 3.5 = 14.5$
- **Coefficient for STAT 109 (+5.0):** STAT 109 students spend **5 more hours** than AC 209. Predicted hours: $11.0 + 5.0 = 16.0$

Important Note

Common Mistake: "Controlling for other variables"
When the only predictor is a categorical variable (like course registration), do NOT say "controlling for other variables" in your interpretation. There are no other variables to control for! Each coefficient simply represents the difference from the baseline group.

10.3 Changing the Baseline Group

If CS 1090 were the baseline instead of AC 209:

- New intercept: $11.0 - 2.0 = 9.0$ (mean for CS 1090)
- New coefficient for AC 209: +2.0 (now positive, since AC 209 is 2 hours above CS 1090)
- Other coefficients adjust accordingly

11 Quiz Review: Validation Metrics

11.1 The Question

When using Ridge regression with cross-validation, which metric should we use to evaluate model performance on the validation set?

Option A: MSE (Mean Squared Error)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{45}$$

Option B: Ridge loss with penalty

$$\text{Ridge Loss} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{46}$$

11.2 The Answer: Option A (MSE)

Key Summary

Why use MSE without the penalty for validation?

1. **Training vs. Validation have different purposes:**
   - **Training:** Find coefficients β that balance fit and complexity (use penalty)
   - **Validation:** Assess how well the model predicts new data (use natural error metric)
2. **The penalty is part of the optimization process, not the evaluation:**
   - The penalty term $\lambda\sum\beta_j^2$ is a regularization tool to prevent overfitting
   - It's not a measure of predictive accuracy
3. **We want to compare models fairly:**
   - Different λ values lead to different penalties
   - Using penalized loss would unfairly favor models with larger λ
   - MSE measures true prediction error, which is what we care about

12 Chapter Summary

Key Summary

Key Concepts from Lecture 11:

1. Bayesian vs. Frequentist Thinking

- Frequentist: Parameters are fixed, data is random
- Bayesian: Parameters have distributions (representing uncertainty), data is fixed once observed

2. Bayes' Rule

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \tag{47}$$

$$f(\theta|X) \propto f(X|\theta) \cdot f(\theta) \tag{48}$$

3. Types of Priors

- **Informative:** Encode specific prior knowledge
- **Uninformative:** Let data dominate
- **Conjugate:** Mathematical convenience (posterior same family as prior)

4. Normal-Normal Model

- Posterior mean is weighted average of prior mean and sample mean
- With more data, posterior concentrates around sample mean
- With stronger prior, posterior stays near prior mean

5. Credible vs. Confidence Intervals

- Credible: "95% probability parameter is in this interval" (Bayesian, intuitive)
- Confidence: "95% of intervals from repeated experiments contain true value" (Frequentist)

6. Regularization as Bayesian MAP

- **Ridge = Normal prior** on coefficients (shrinkage)
- **Lasso = Laplace prior** on coefficients (shrinkage + sparsity)

7. MCMC for Complex Models

- Sample from posterior when analytical solutions don't exist
- Gibbs sampling: Sample parameters one at a time
- Discard burn-in samples, use remaining samples for inference

13 Key Formulas Reference

Key Information

Bayes' Rule:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \quad \text{or} \quad f(\theta|X) \propto f(X|\theta)f(\theta) \tag{49}$$

Normal-Normal Posterior:

$$\mu_n = \frac{\sigma^2\mu_0 + n\sigma_0^2\bar{X}}{\sigma^2 + n\sigma_0^2}, \quad \sigma_n^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2} \tag{50}$$

95% Credible Interval (Normal):

$$[\mu_n - 1.96\sigma_n, \quad \mu_n + 1.96\sigma_n] \tag{51}$$

MAP Estimation:

$$\hat{\theta}_{MAP} = \arg\max_\theta f(\theta|X) = \arg\max_\theta [f(X|\theta)f(\theta)] \tag{52}$$

Ridge = Normal Prior:

$$\beta_j \sim N(0, \tau^2) \quad \Rightarrow \quad \text{Penalty} = \lambda \sum \beta_j^2 \tag{53}$$

Lasso = Laplace Prior:

$$\beta_j \sim \text{Laplace}(0, b) \quad \Rightarrow \quad \text{Penalty} = \lambda \sum |\beta_j| \tag{54}$$