Lecture #11: Bayesian Modeling

aka STAT109A, AC209A, CSCIE-109A

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Gumb

Lecture Outline: Bayes

- Review
- Bayes Inference
    - Choosing a prior
- Bayesian Estimators
- Bayesian Regression
- Simulating a Posterior

CS109A
GAME Time

Q13. When training a regression model with Ridge regularization which metric do we typically use to evaluate performance on the validation set?

Options

A. $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

B. $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}|\beta_j|^2$

C. $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$

D. $\frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|}$

Q14. Open-Ended Question

Consider a linear regression model to predict the mount of time spent working on HW1 (in hours) for students in Data Science based on which course they are enrolled in (4 options: ac_209, cs_1090, csci_e109, or stat_109).

The following estimated regression model equation was calculated:

$$\hat{y} = 11.0 - 2.0x_{cs_{1090}} + 3.5x_{csci_{e109}} + 5.0x_{stat_{109}}$$

i. Which group spent the most time on HW1, on average?

A. ac_209
B. cs_1090
C. csci_e109
D. stat_109

ii. Interpret the values 11.0 and 5.0 in this model.

iii. Pat is enrolled as a $csci_{e109}$ student in the extension school. Use this model to predict the amount of time Pat spent working on HW1.

iv. Write out the estimated regression model if the variable $x_{cs_{1090}}$ was removed and replaced with $x_{ac_{209}}$.

Lecture Outline: Bayes

- Review
- **Bayes Inference**
    - Choosing a prior
- Bayesian Estimators
- Bayesian Regression
- Simulating a Posterior

Bayes Rule

- Bayes' rule (formula) provides a way to go from $P(B \mid A)$ to $P(A \mid B)$ (they are in general not equal…)
- If A and B are two events whose probabilities are not 0 or 1:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

Bayes Rule, priors and posteriors

- Note: this calculation is based off the fact that your chance of being pregnant before taking the test was assumed to be 30%.
    - This is called the **prior probability**.
    - This may not actually be 30%. Maybe you believe you have more like a 50% chance.
- This probability was updated to be 97.65% after testing positive based on the test.
    - This is called the **posterior probability**.
- This change from prior to posterior is essentially *updating* the probability given evidence.
- This can be applied to theory (parameters) and data…

Bayes Rule, for distributions!

- We just saw the simplest form of Bayes' Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- What is Bayes' Rule effectively doing?

- How would this be useful for statistical inference?
  *Think: parameters (θ) and data (X).

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

Bayes Rule/Inference, for continuous RVs

- This can be rewritten for a set of parameters, θ, treating it as a continuous random variable, in terms of PDFs:

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

- Let's break this down:

X : the vector (or matrix) of data: $X_1, X_2, \ldots, X_n$

θ : the vector of parameters (or just a scalar).

$f(\mathbf{X}|\theta)$: the likelihood of X_i's

$f(\mathbf{X})$: the marginal pdf of X_i's (just a normalizing constant)

$f(\theta)$: the *prior distribution* of θ

$f(\theta|\mathbf{X})$: the *posterior distribution* of θ

Bayesian Inference: from prior to posterior

- The prior distribution, $f(\theta)$, often is based on a known distribution with it's own set of parameters. These are called *hyperparameters*.
- The marginal PDF of X is the distribution of X ignoring θ. How do you solve for a marginal PDF based on a joint distribution?

$$f(X) = \int_{\theta} f(x, \theta) d\theta = \int_{\theta} f(X|\theta) f(\theta) d\theta$$

- By definition, this marginal PDF of X will not involve θ. Thus, it can be though of as a multiplicative normalizing constant with respect to θ.
- So we can write the posterior dist. as proportional to:

$$f(\theta|X) = \frac{f(X|\theta) f(\theta)}{f(X)} \propto f(X|\theta) f(\theta)$$

Bayesian Inference, a very simple example

- You own 3 coins: a fair one (with $p = 0.50$ of landing heads) and two biased coins (one with $p = 0.10$ and the other $p = 0.90$). You reach into your pocket and select one coin at random to flip.
- You flip it 4 times and see 3 heads and one tail.
- Intuitively, which coin(s) do you feel are plausible to have been the one chosen? What if you had to pick just one?
- What is the posterior distribution for p?

$P(p = 0.10 \mid X) = 0.007$, $P(p = 0.50 \mid X) = 0.458$, $P(p = 0.90 \mid X) = 0.535$

- Now which coin do you believe was chosen? Are you certain?
- What would happen if $n = 4$, $k = 2$? What about if $n = 40$ and $k = 30$?
- Note: this parameters space is discrete, which is rarely the case in practice.

Bayesian Perspective

- So how is this Bayesian approach different from the Frequentist approach (which typically only uses the likelihood function)?
- It also relies on a prior distribution. So an analyst has to place some *a priori* probability on the distribution of the parameter.
- This adds some extra uncertainty into the approach. Different analysts can come up at the same problem with different priors, and thus get different results ☹
- But this is really no different than different Frequentists making different assumptions on the data (independence, specific properties of the underlying distribution of the X_i's, etc…)

Bayesian Probability of θ

- The other difference from a Frequentist's approach is now we have distribution(s) of the parameter(s) (both the prior and the posterior distributions).
- So what is this probability distribution really measuring?
- A Frequentist's "definition" of probability: the long run expected **frequency** of an occurrence of a random variable if an experiment is performed an infinite number of times. Can only be applied to random things.
- A Bayesian's "definition" of probability: a measure or description of belief or plausibility…and can be applied to any unknown quantities ☺ Random entities **or** unknown latent variables/parameters.
- Sounds a whole lot like a Frequentist's use of the word *confidence* in a Confidence Interval!

Bayesian's Prior and Posterior

- A Bayesian's prior distribution, $f(\theta)$, captures one's prior belief or experience of the parameter. This belief should be updated based on what? The data!!! $X_1,\ldots,X_n$
- And the posterior distribution, $f(\theta \,/\, X_1,\ldots,X_n)$, can be thought of exactly this way: as a measure of belief on the parameter given the data seen in the sample.
- And how should this belief be updated? Weighted based on the likelihood!
- So more likely values of θ will have more bearing on the posterior, given the data we see.
- So once the data is fixed at what is actually measured, then the posterior will be weighted towards values of θ that agree with those measurement.

Bayes Approaches to Frequentist Ideas

- Bayesian inferences on the parameters, θ, can then be based solely on the posterior distribution. Which makes life simple!
- The posterior is not exactly a sampling distribution though. Why not?
- But the posterior is a measure of uncertainty of the parameter, and can be used to examine the uncertainty of an estimator.
- The posterior can also be used to calculate Bayesian analogues to Frequentist inferential techniques: estimates and their intervals (and hypothesis tests)!

Which is better: Bayes or Frequentist?

- So which should we use: the Bayesian approach or the Frequentist approach?
- It depends on the setting. And depends on who you are doing the work for.
- Frequentist approaches are classical approaches, and were developed first because they were easy to solve.
- Bayesian approaches usually are more computationally intensive, and only recently (10+ years) have taken off.
- In practice in modern times, both approaches are often used for the same data and both analyses are presented.
- Both often give quite similar results.
- At the very least, we first have to define what an estimator is in the Bayesian paradigm…

Why Bayesian Modeling?

- What are some advantages and use cases for Bayesian modeling (and inference)?
  - Allows for greater flexibility in the model (and for more complex models)
  - Allows for *expert opinion* or prior information/data to be used in the model.
  - Allows for combining various sources of data or studies together:
    - Example: meta-analyses where the raw data is not available, but summaries are!
  - Allows for data to be measured at various different *levels*:
    - Example: hierarchical models, that could have measurements at the state, county, and individual levels!
      - We will see these later in the course!

An example: Bayesian Normal-Normal Model

- Let $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$ and where σ^2 is known (maybe from a previous study). Let's put a prior on $\mu \sim N(\mu_0, \sigma_0^2)$.
- What are the parameter(s) and the hyperparameters?
- Write down the prior:
- Write down the likelihood:
- Write down the normalizing constant (the denominator):

Normal-Normal Model: Posterior Result

- So the posterior distribution is:

$$\mu|X \sim N\left(\frac{\sigma^2\mu_0 + n\sigma_0^2\bar{X}}{\sigma^2 + n\sigma_0^2}, \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}\right)$$

- So what?
- The posterior distribution for the mean of a normal distribution, given the data, only depends on the sample data in terms of the sample mean. The posterior of μ is normally dist. (if we start with a prior that is normally dist.).
- What is the posterior mean estimator (the mean of this distribution)?
- The posterior mean of μ is a weighted average of the prior mean, μ_0, and n-times the sample mean. So what happens to the effect of the prior on the posterior (and the estimator) as n increases?
  - The variance of the posterior decreases as n increases.

Lecture Outline: Bayes

- Review
- Bayes Inference
  - **Choosing a prior**
- Bayesian Estimators
- Bayesian Regression
- Simulating a Posterior

Choosing a Prior

- Recall, in a Bayes statistical modeling problem, the posterior distribution of θ is calculated from:

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

- The posterior distribution not only depends on the likelihood, $f(X|\theta)$but also on the prior, $f(\theta)$.
- How does an analyst then choose a prior? One of 3 ways:
  1) Based on previous studies
  2) One that puts as little info on θ as possible
  3) One that makes the posterior easy to compute

Previous Studies

- The most scientifically sound way to come up with a prior is from previous studies. For example:
- You're interested in predicting the temperature at noon tomorrow. What might be a reasonable prior?
  - A Normal distribution with μ_0 = today's temperature at noon and σ_0^2 = the variance from day-to-day noon temperatures (over the last 30 days).
- You're interested in modeling the probability, p, that a patient will be cured from a disease based on a treatment. What might be a reasonable prior for p?
  - A distribution whose mean will be the "standard of care" cure rate.

Uninformative Priors

- Another approach, which is useful if no previous study is available, is to make sure the prior has as little effect on the posterior as possible.
- This is called an *uninformative prior*, but really it's just a minimally informative one.
- For predicting the temperature at noon tomorrow, what might be a reasonable uninformative prior?
  - A Uniform distribution with min at the record low and max at the record high at noon for that date.
- For the proportion cured by a new treatment example, what might be a reasonable uninformative prior?
  - A Uniform distribution between 0 and 1

Conjugate Priors

- The third approach, which is mathematically useful, is to choose a prior that leads to a closed form distribution for the posterior.
- One type with this property is called a conjugate prior.
- A *conjugate prior* is a prior distribution that results in a posterior distribution of the same family (so if your prior is a Normal, then your posterior is also a Normal). The conjugacy of a prior also depends on the distribution of the X_i's.
- We've seen one of these in this lecture:
    - Normal(prior)-Normal(data/likelihood)-Normal(posterior)
- What is the μ and σ^2 are both unknown?
    - $\mu \sim Normal$ is still the conjugate prior
    - $(1/\sigma^2) \sim Gamma$ is the conjugate prior ($1/\sigma^2$ is called the *precision*).

Conjugate Priors: A Short List

- Besides the normal distribution, there are countless other distributions we could model our data with (aka, the likelihood).
- Each of these distributions have their own parameters and conjugate priors:
- Normal:
    - $\mu \sim Normal$
    - $(1/\sigma^2) \sim Gamma$
- Exponential:
    - $\lambda \sim Gamma$
- Binomial:
    - $p \sim Beta$
- Poisson:
    - $\lambda \sim Gamma$

Lecture Outline: Bayes

- Review
- Bayes Inference
  - Choosing a prior
- **Bayesian Estimators**
- Bayesian Regression
- Simulating a Posterior

Interpreting the posterior distribution

- Recall our posterior distribution:

$$\mu|X = N\left(\frac{\sigma^2\mu_0 + n\sigma_0^2\bar{X}}{\sigma^2 + n\sigma_0^2}, \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}\right)$$

- How can we summarize this distribution to aid in interpreting what it means?
- Just use standard summaries! Think measures of center & spread!
    - Measures of center: Means, medians, modes
    - Measures of spread: how wide is the middle 95% of the distribution?
- These are then called:
    - The posterior mean, the posterior mode, and the 95% Credible Intervals.

Posterior estimators for the normal-normal model

$$\mu|X = N\left(\frac{\sigma^2\mu_0 + n\sigma_0^2\bar{X}}{\sigma^2 + n\sigma_0^2}, \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}\right)$$

- What is the Bayes' posterior mean estimator for this distribution?

- What is the posterior mode estimator (sometimes called the max *a posteriori* estimator, or MAP)?

- What is the 95% credible interval for μ?

Lecture Outline: Bayes

- Review
- Bayes Inference
  - Choosing a prior
- Bayesian Estimators
- **Bayesian Regression**
- Simulating a Posterior

Towards Bayesian Regression

- What is the probabilistic model for linear regression (let's keep it simple for now)?

- What are the parameters?

- What prior distributions should we put on these parameters?
- What are their conjugate distributions?

Bayesian Linear Regression: Conjugate Priors

- Commonly used [conjugate] prior distributions in simple linear regression:
  - $\beta_0 \sim Normal(\mu_0, \sigma_0^2)$
  - $\beta_1 \sim Normal(\mu_1, \sigma_1^2)$
  - $(1/\sigma^2) \sim Gamma(a_0, \lambda_0)$
- What are the ramifications for choosing different values for their hyperparameters?
  - What would be reasonable for μ_1 and σ_1^2? What would be good choices for a minimally informative prior? What would be good choices for a *null* association between X and Y before collecting data?

Bayesian Linear Regression: Posterior Distributions

- The posterior distribution of the parameters become:
  - $(\beta_0, \beta_1 | \sigma^2, X, y) \sim MVN(\quad)$
  - $(1/\sigma^2)|X, y \sim Gamma(\quad)$
- Note #1: $\beta_0, \beta_1 \, | \sigma^2$ is the **conditional** posterior distribution of β_0, β_1 given σ^2.
- Note #2: this can be generalized so that the prior on β_0, β_1 can have a correlation/covariance matrix.
- Note #3: this result can be extended to multiple regression. See many online sources for that (like wikipedia).
- This is called the [MV]Normal-Gamma joint distribution.

Ridge and LASSO: a review

- What is the loss function in linear regression?

$$\mathcal{L}_{OLS} = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{1,i} + \cdots)\right)^2$$

- What is the loss function in Ridge regression?

$$\mathcal{L}_{Ridge} = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{1,i} + \cdots)\right)^2 + \lambda \sum_{j=1}^{p} (\beta_j)^2$$

- What is the loss function in LASSO regression?

$$\mathcal{L}_{LASSO} = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{1,i} + \cdots)\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Q: How can we add these penalized terms probabilistically?
    - A: with carefully chosen priors for our β coefficients!

Ridge and LASSO: the Bayesian perspective

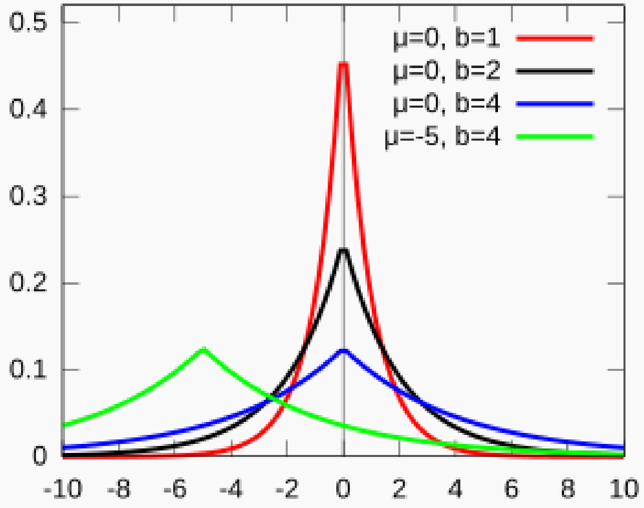
- In LASSO and Ridge, what value are we shrinking our β estimates to?
- What does this mean for the prior distributions for β?
  - The priors should be centered at zero!
- How can we control the amount of shrinkage (aka, λ)?
  - If we put more *weight* on out prior, then we are shrinking more towards zero!
  - This is equivalent to putting more prior point mass at zero!
- What distribution, in terms of log-pdf, has a quadratic effect of distance from a mean of zero?
  - The Normal!
- What distribution, in terms of log-pdf, has a linear effect of distance from a mean of zero?
  - The Exponential! <- we just need to allow for negative values!!!!

An aside: the Laplace Distribution

- The Laplace distribution, sometimes called the *double exponential distribution*, is comprised of two exponential distributions glued back-to-back, at location μ (most often $\mu = 0$) and with scale parameter b (you can think of $b = \frac{1}{\lambda}$, where λ is the rate parameter of the Expo).
  - Note: the Expo PDF is scaled by ½ since it is mirrored across $x = \mu$.
- Let $X \sim \text{Laplace}(\mu, b)$. Then it's PDF is:

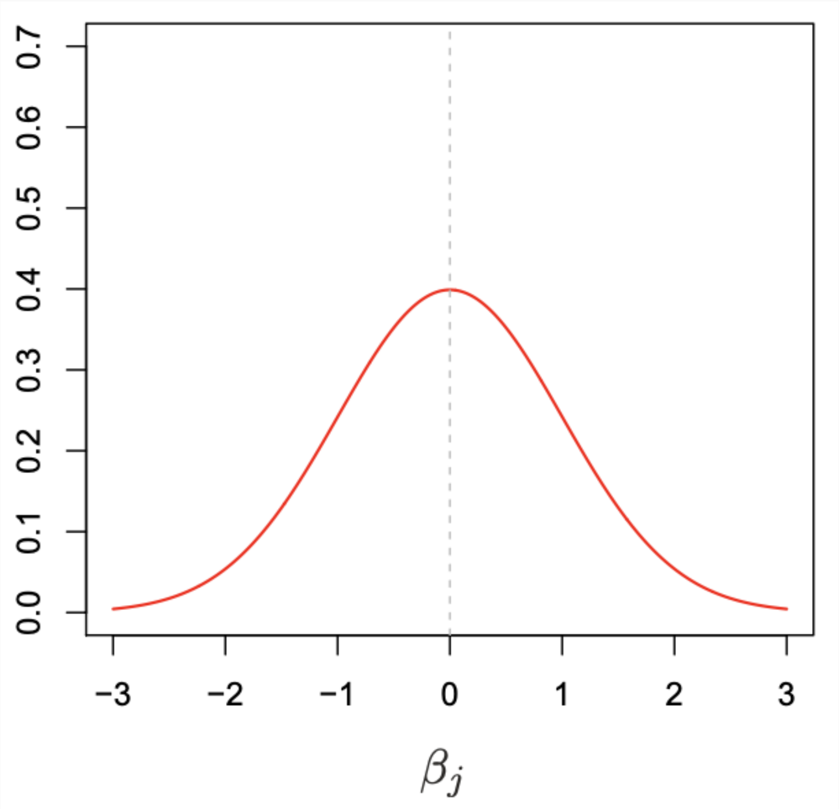
$$f(x) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

- Note: a Laplace r.v. is the difference between two independent exponential random variables:
  - $X, Y \sim Expo(\lambda) \to (X - Y) \sim \text{Laplace}\left(\mu = 0, b = \frac{1}{\lambda}\right)$

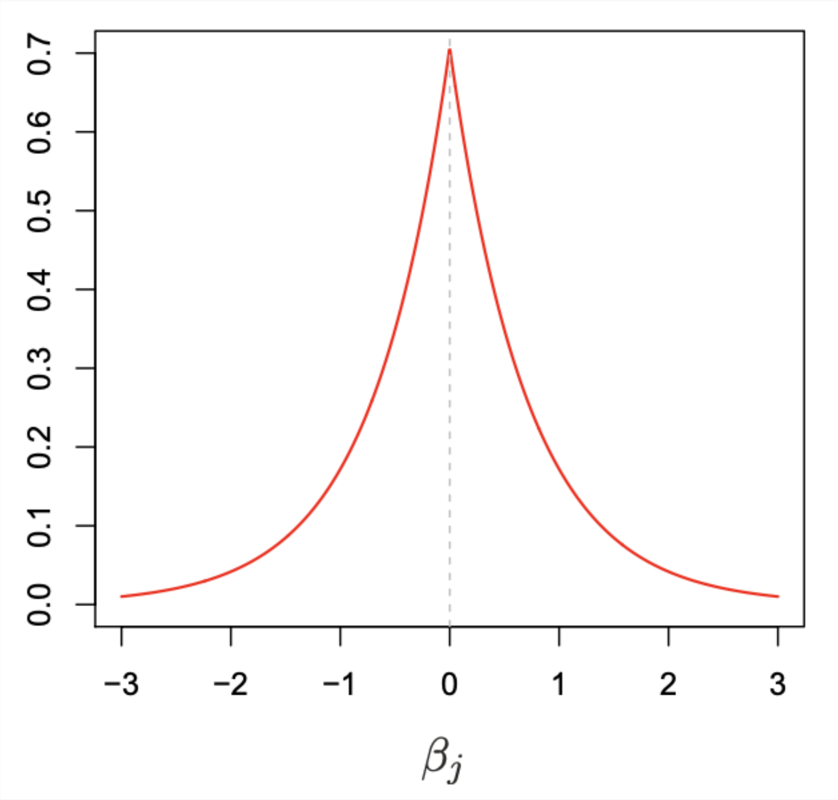


Ridge and LASSO: the Bayesian perspective

The Normal PDF as a prior:
$\beta \sim N(\mu = 0, \sigma^2 = 1)$



The Laplace PDF as a prior:
$\beta \sim \text{Laplace}(\mu = 0, b = 1)$



Ridge and LASSO: the Bayesian math

$$f(\theta|X) = \frac{f(X|\theta)f(\theta)}{f(X)} \propto f(X|\theta)f(\theta)$$

Ridge prior: $\beta \sim N(\mu = 0, \sigma^2 = ?)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

LASSO prior: $\beta \sim \text{Laplace}(\mu = 0, b = ?)$

$$f(x) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

Let's put it together:

Lecture Outline: Bayes

- Review
- Bayes Inference
    - Choosing a prior
- Bayesian Estimators
- Bayesian Regression
- **Simulating a Posterior**

Bayesian Computation Methods

- When a posterior distribution has a well-known distribution (for example, when you are using a conjugate prior), then inferences are easy to develop.
  - You can get closed form solutions for specific estimators for θ: like the posterior mean, posterior median, or posterior mode
- If the posterior is not well known, or if there are a lot of parameters, then inferences on the posterior are not so easy
- We need a numerical/computational way to calculate inferences (estimators, credibility intervals, and hypothesis tests).
- Simulating from the posterior may be the way to go!

Bayesian Computation Methods

- When a posterior distribution has a well-known distribution (for example, when you are using a conjugate prior), then inferences are easy to develop.
  - You can get closed form solutions for specific estimators for θ: like the posterior mean, posterior median, or posterior mode
- If the posterior is not well known, or if there are a lot of parameters, then inferences on the posterior are not so easy
- We need a numerical/computational way to calculate inferences (estimators, credibility intervals, and hypothesis tests).
- Simulating from the posterior may be the way to go!

Simulating from a Posterior

- If the Posterior distribution has a closed form, but is not a well-known distribution (Normal-Gamma anyone?), then it may be easier to simulate directly from the distribution to get estimates.
- If the posterior can be broken into conditional and marginal distributions, this makes life easier.
- To simulate in this situation:
  1. Collect *nsims* values of θ_1 from the marginal posterior distribution of $f(\theta_1/X)$
  2. For each simulated value of θ_1, select a value for θ_2 from the conditional posterior distribution of $f(\theta_2/\theta_1,X)$

Estimates from a Simulated Posterior

- Calculating the posterior mean:
  - Simply just need to find the empirical estimate of the mean of the simulated distribution of observations of θ:
- Calculating the posterior mode, :
  - Not so simple ☹. Why?
  - Need to fit an entire empirical distribution/curve, and then find the mode of that (aka: bump-hunting).
- How to calculate the credible interval?
  - It's simple: just calculate the desired quantiles from the empirical/simulated distribution.
  - Note: this is equivalent to extracting the Confidence Interval from the empirical bootstrap distribution!

Normal-Normal Model Simulation

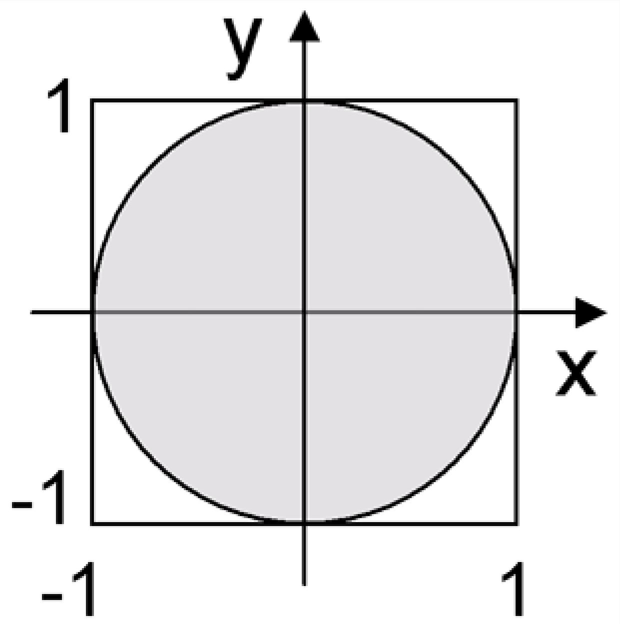
- What is the joint posterior dist. for (μ,σ^2) when X_i ~ Normal, and the priors are $\mu|\sigma^2$ ~Normal and $1/\sigma^2$ ~Gamma?
- It's Normal-Gamma of course! What parameters?
  - $\mu|\sigma^2, X \sim N[m_n=(p_0 m_0+n\bar{x})/(p_0+n), \sigma^2/(p_0+n)]$
  - $1/\sigma^2|X \sim Gamma[(v_0+n)/2, SS_n/2 = (SS_0+SS+(np_0)*(\bar{x}-m_0)^2/(p_0+n)/2]$
- How to simulate the distribution? First sample a σ^2 from the Gamma posterior, then a $\mu|\sigma^2$ from the Normal posterior.
- How do we make inferences from this distribution?
- Calculate some estimates or the credible interval

Markov Chain Monte Carlo

- *Markov Chain Monte Carlo* (MCMC) methods are algorithms to simulate from a probability distribution (like a posterior) that is usually a joint distribution.
- There are MANY examples, and a whole course could be dedicated to these algorithms. Here are a few:
  - Adaptive Rejection: sample from a distribution based on throwing darts at it (essentially).
  - Gibbs Sampling: sample from a distribution based on all of the conditional distributions.
  - Metropolis-Hastings: based on a random walk through a proposal joint density, and a method for rejecting possible moves.
  - And many more…
- We will illustrate the idea here, and then implement them in a future lecture/section (maybe).

Markov Chain Monte Carlo

- Throw a dart at the (*x*,*y*) plane based on proposal distribution (simplest one to think of is the uniform).
- If the point is within the distribution, accept *x* as a random observation. Otherwise, reject it and throw again.
- Simple example: collect a random sample of points from the unit circle:

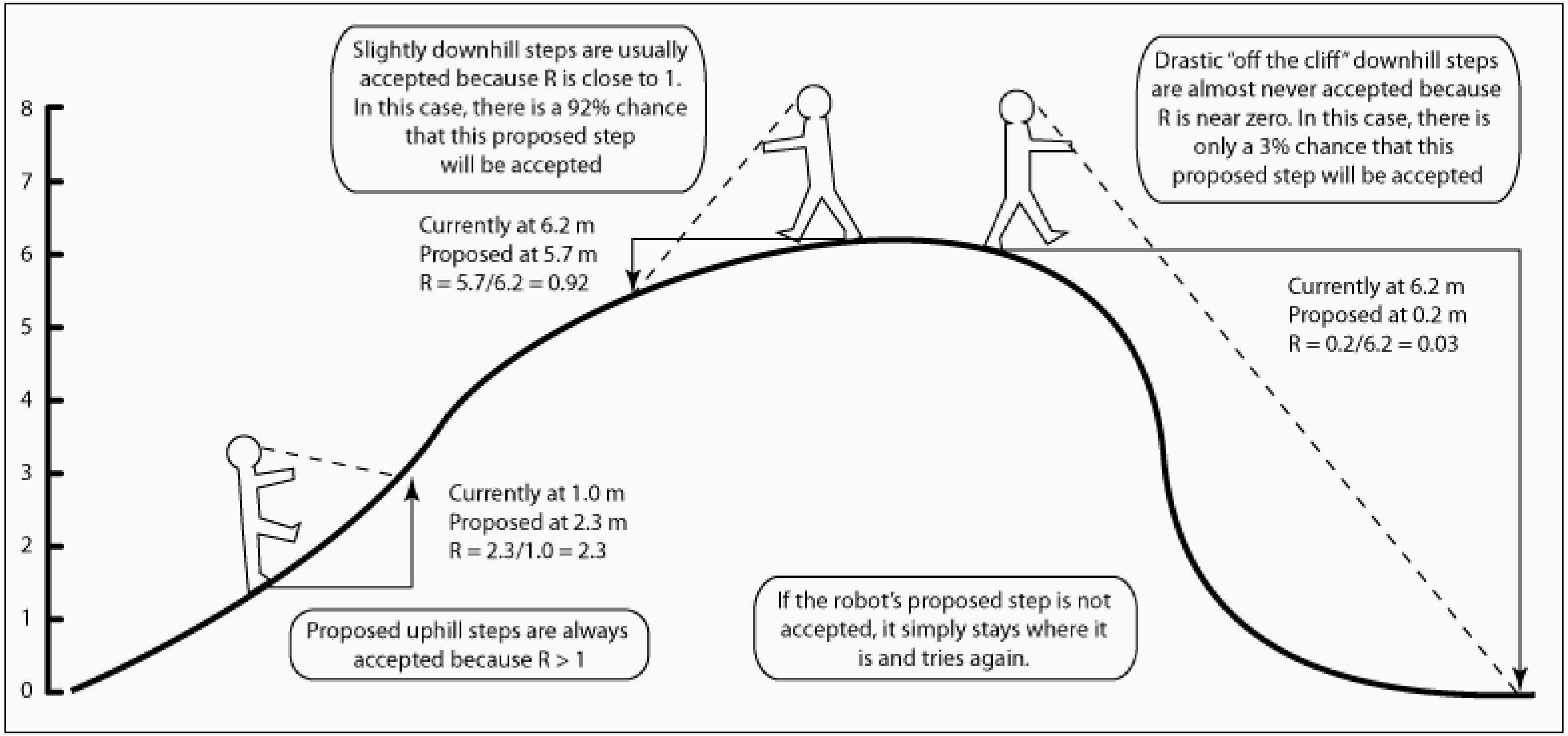


Gibbs Sampling

- If all of the conditional distributions of the variables are known, then we can use Gibbs sampling. So for 3 parameters, we know $f(\theta_1|\theta_2,\theta_3)$, $f(\theta_2|\theta_1,\theta_3)$, and $f(\theta_3|\theta_1,\theta_2)$.
- The algorithm is as follows:
  0) Select an initial set of values of θ_1,θ_2 , and θ_3.
  1) Choose a new θ_1^* from $f(\theta_1|\theta_2,\theta_3)$.
  2) Choose a new θ_2^* from $f(\theta_2|\theta_1^*,\theta_3)$.
  3) Choose a new θ_3^* from $f(\theta_3|\theta_1^*,\theta_2^*)$.
- Repeat steps 1-3 until…
- After a burn in period (~100 iterations), then keep the remaining results of each iteration as a joint realization of $(\theta_1,\theta_2,\theta_3)$ until you reach the desired number of realizations.
- See any issues with this method? It will be an issue with the next method (metro-hasty) as well.

Metropolis-Hastings Illustration

- Start at an initial point, x. Propose a next point x^*, and accept it as the new spot at the rate of $f(x^*)/f(x)$ (if $f(x^*) \geq f(x)$, then always accept the move). Otherwise, stay in the same spot.
- Like a random walk through the distribution…



Memes of the day