9/19

Bayesian Computational Methods I: Introduction to Markov Chain Monte Carlo (MCMC))

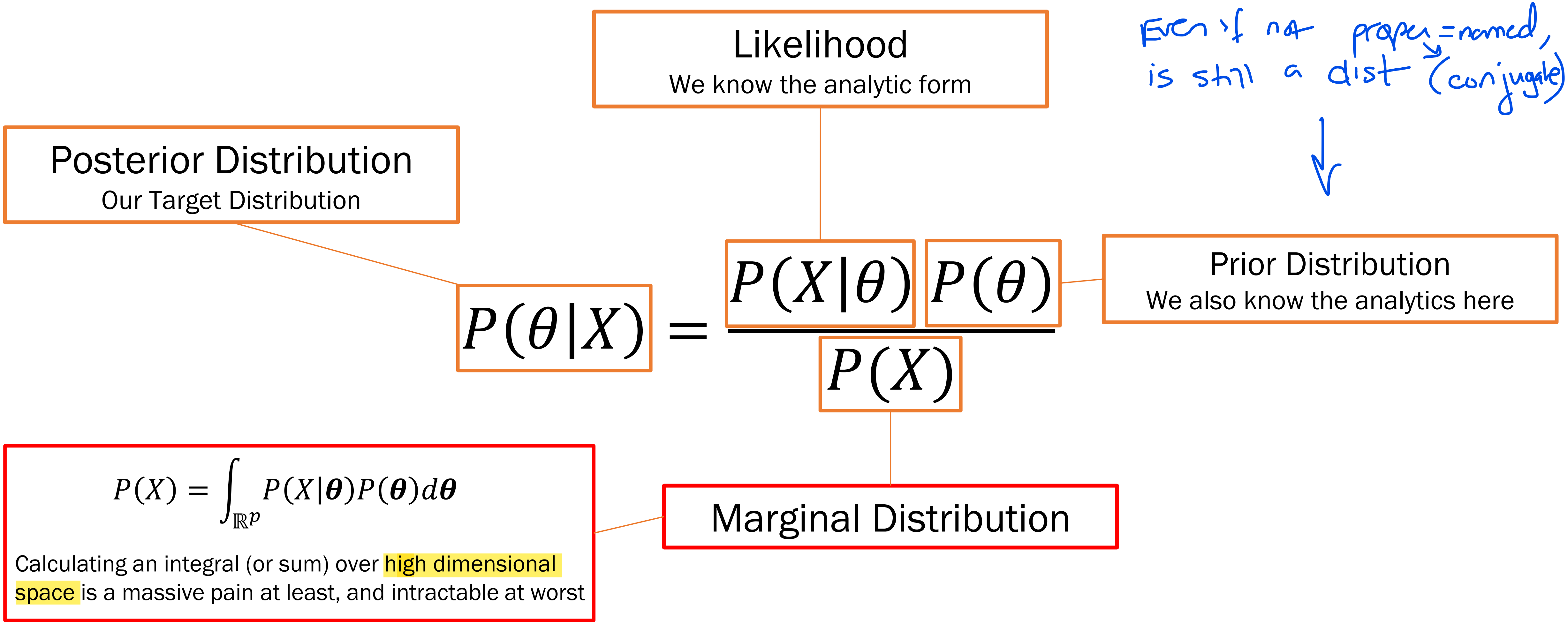
DS6040 Fall 2024
Teague R. Henry

University of Virginia | School of Data Science

Outline

- Posterior Review
- Monte Carlo Sampling
- Markov Chain Monte Carlo Sampling
  - Gibbs Samplers
  - Random Walk Metropolis/-Hastings
- Hamiltonian MCMC
- NUTS – No U-Turn Sampler

Review



Likelihood
We know the analytic form

Even if not proper=named, is still a dist (conjugate)

Posterior Distribution
Our Target Distribution

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

Prior Distribution
We also know the analytics here

$$P(X) = \int_{\mathbb{R}^p} P(X|\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta}$$

Calculating an integral (or sum) over high dimensional space is a massive pain at least, and intractable at worst

Marginal Distribution

Issues with Bayes Theorem

Unless you use conjugate priors, the posterior distribution for your model will not be a proper, well defined, distribution.

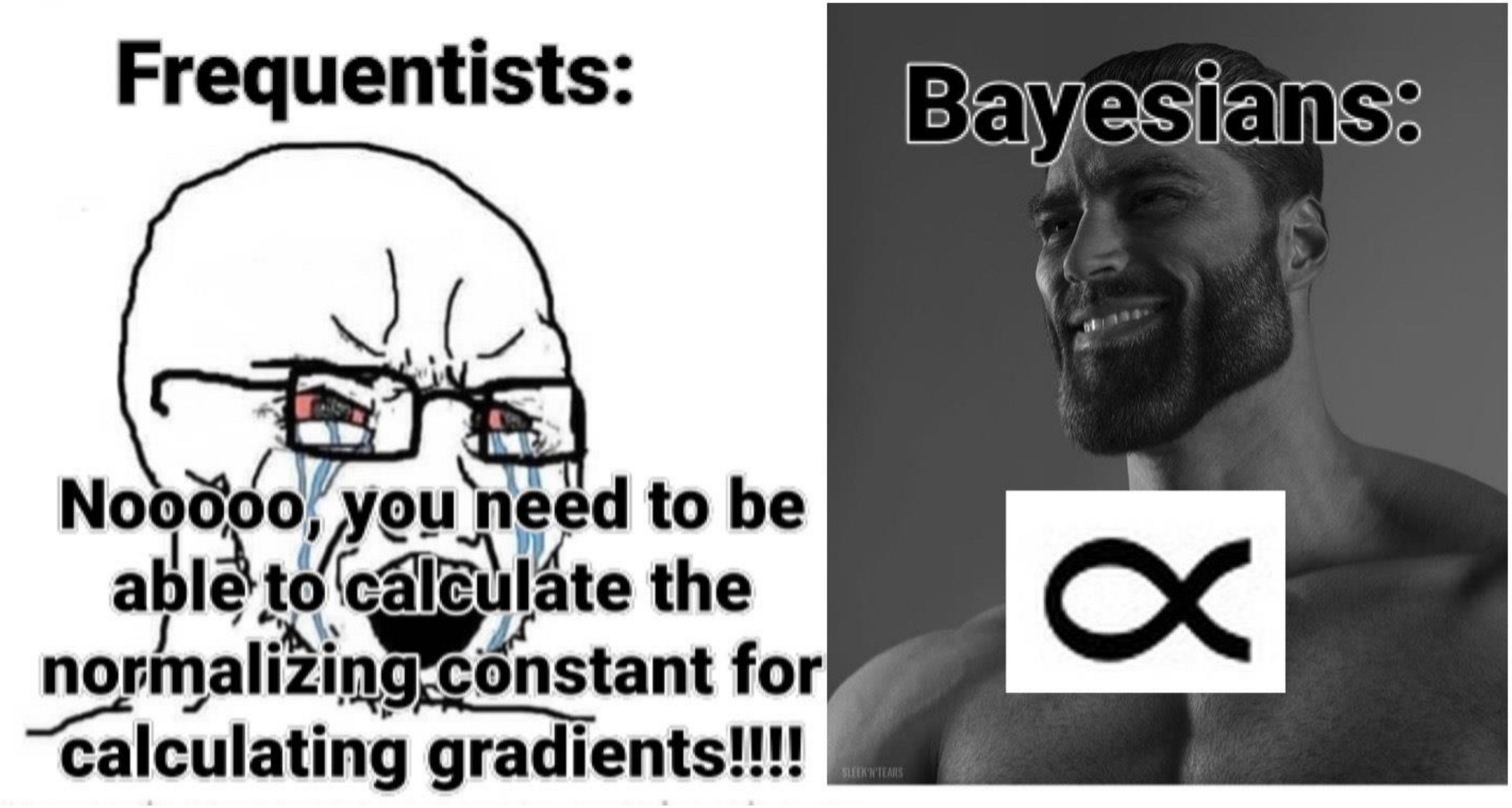
- Which means no nice analytics, no nice properties, and no reading the expected value and other summaries directly off.

More broadly, the issue with non-conjugate priors is that we can't easily guess what $P(X)$ is.

- This means we can't correctly normalize the posterior, which in turn means we can't get accurate estimates of posterior probabilities.
- We can still compare unnormalized values, but only comparisons work.

So, we need a way of estimating the posterior that takes care of the marginal.

But First a Meme


Frequentists:
Nooooo, you need to be able to calculate the normalizing constant for calculating gradients!!!!
Bayesians:
∝
SLEEK'N'TEARS

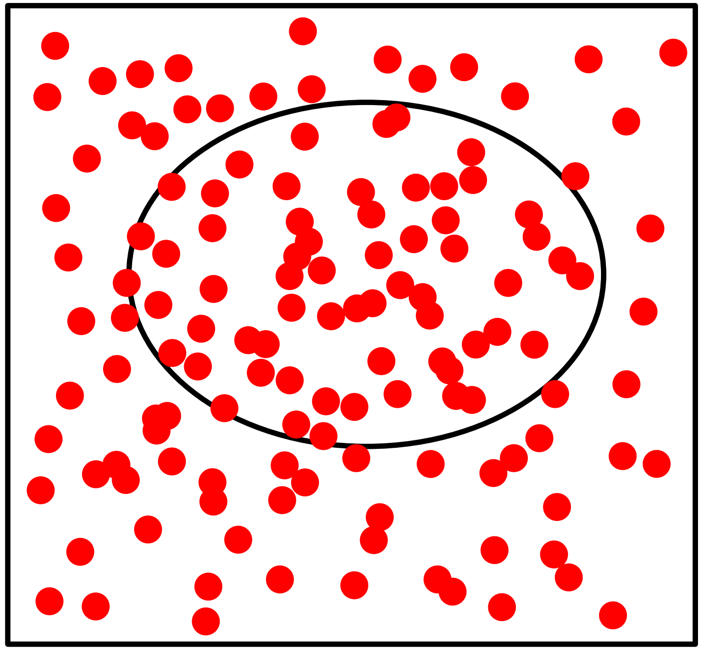
Sampling – Not Just for Data Collection

You want to determine the area of the oval, how do you do that without measuring its lengths?

Core Idea: Sampling can be used to estimate quantities.

- Put the oval in a rectangle of known volume
- Throw a known number of dots randomly on the shape
- Calculate the proportion of dots inside the circle.
- Area of rectangle × proportion ≈ Area of Circle (about)

With increasing number of samples comes increasing accuracy of approximation (towards the true value)

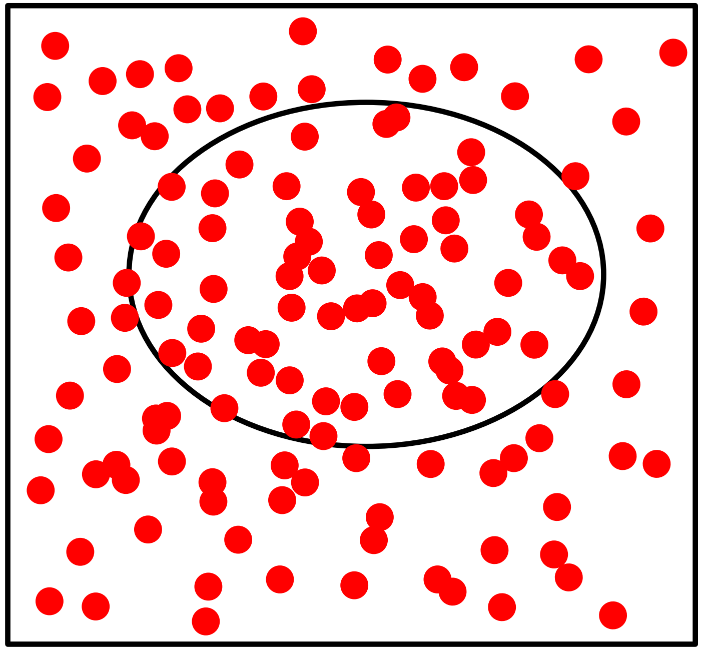


(Instead of measuring, approx posterior)

Sampling – Approximating the Posterior

Sampling in Bayesian Estimation –

- You can't easily calculate the marginal distribution (i.e. no conjugate priors)
- Goal - Sample from the posterior distribution.



Two Key Issues:

- Parameter space is very large for even small models
- Sampling from probability distributions needs to involve probabilities. - Want shape of it

Sampling – Another Metaphor

You've made a series of questionable decisions and are lost in the woods, on a pitch-black moonless night. You need to get to the highest hill, but all you have is a set of throwable sensors that emit a noise based on how high they are.

- Assume you can't tell the slope of the landscape wherever you are standing (No gradient information!) – no up/down info
- Assume you can always navigate to the sensor, and you can always tell which is the loudest sensor (compared to any others you've thrown.)
- Assume the wolves that are closing in are all hearing impaired.

How do you get to the highest hill?

Sampling – Another Metaphor

One possible method:

1. Throw your sensors randomly around you.
2. Pick up all in order of loudness, ending on the loudest sensor.
3. Repeat.

What’s the main issue with this method?

- Hint: What if you are already on the highest hill?

Easy fix: Step 1 is drop a sensor at your feet, then throw the rest randomly.

Sampling – Another Metaphor

Sampling is incredibly powerful, and the core is very simple. Quite literally all you need is:

1. The ability to evaluate your target function at any location.
   - The sensors giving noise based on height
   - Calculating the likelihood of data given a specific set of parameters
   - Evaluating the cost function at a given set of parameters.

- likelihood of data at any set params also ?

2. The ability to generate random numbers.
   - Technically, any distribution would do as you can transform distributions into other forms.

All sampling methods, from MCMC to particle filters, are riffs off of this basic design. (eval target, then choose to move based on criteria so get closer to where want to be

The Gibbs Sampler

The Gibbs Sampler is one of the simplest Bayesian estimation methods.

- Instead of randomly throwing out sensors, we use the structure of the probability distributions to guide our search.
- Requirements: The conditional distribution of a parameter given all other parameters must be a proper distribution (i.e. we have to be able to sample from that distribution).

  ↳ known, named
- Note: If the priors are not conjugate, it is still often the case that the conditional distributions of parameters are proper. If so, Gibbs can radically improve computation time.

The Gibbs Sampler

Example: We want $P(\theta_1, \theta_2 | X)$. But we didn't use fully conjugate priors.

- This means we can't skip directly to the posterior

But: We know that $P(\theta_1 | X, \theta_2)$ and $P(\theta_2 | X, \theta_1)$ are both proper (named) probability distributions.

- E.g. we can sample directly from each.

Intuition: $P(\theta_1 | X)$ is just $P(\theta_1 | X, \theta_2)$ over all values of θ_2.

Gibbs Sampler:

$\theta_{1[0]}$ and $\theta_{2[0]}$ are starting values.

1. Let i be the iteration number
2. Sample $\theta_{1[i]}$ from $P\left(\theta_1 \middle| X, \theta_{2[i-1]}\right)$
3. Sample $\theta_{2[i]}$ from $P\left(\theta_2 \middle| X, \theta_{1[i]}\right)$
4. Update $i = i + 1$

The Gibbs Sampler - Example

$$X_i \sim N(\mu, \sigma^2)$$ -data

$$P(\mu) \propto c$$

$$P(\sigma^2) \propto \frac{1}{\sigma^2}$$

Normal, unknown mean and variance, with Jeffreys priors on both. (25 obs)

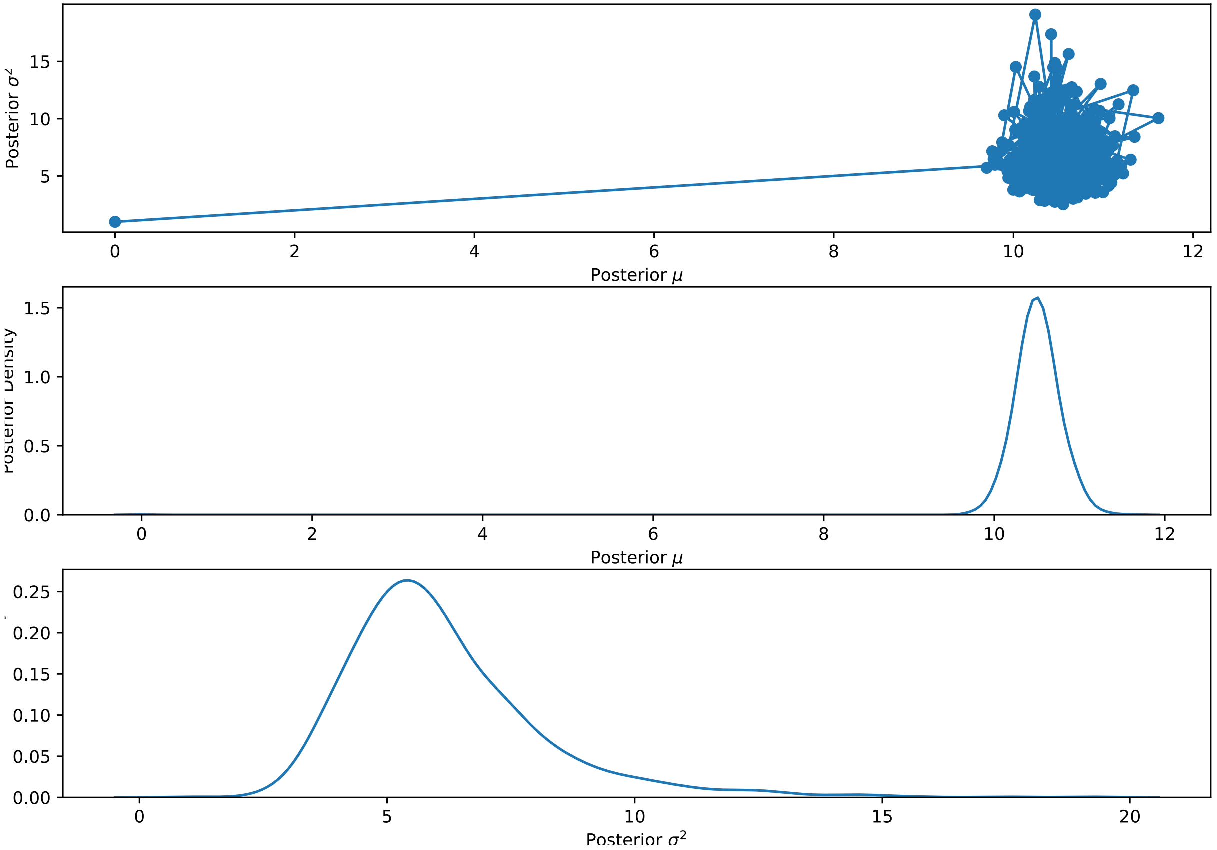
Fortunately,

sampling dist

$$\mu | \sigma^2, X \sim N(\bar{x}, \frac{\sigma^2}{n})$$

$$\sigma^2 | \mu, X \sim InverseGamma(\frac{n}{2}, \frac{\sum (x_i - \mu)^2}{2})$$

Gibbs → go back + forth btw them



The Gibbs Sampler

Gibbs Sampling simplifies estimation by only requiring that the conditional densities are proper distributions

- i.e. we can sample directly from them…

Gibbs tends to perform well, given the restrictive assumptions needed. But:

- For categorical variables, or extremely high dimensional variables with pathological distributions, Gibbs can take forever to converge.
- Gibbs can't sample what it cannot get to…

Probability Matrix for a 2 Dimensional State Variable

	A = 1	A = 0
B = 1	.5	0
B = 0	0	.5

Starting values: $A_0 = 1, B_0 = 1$

- Sample $A_1|B_0$? = 0
- Sample $B_1|A_1$? c 0.5

You can't get from $(1, 1)$ to $(0,0)$ by sampling from 1 variable at a time…

Into the Unknown

What if even the conditional distributions are not proper?

Remember:

$$P(\theta|X) \overset{\substack{\text{propto} \\ \downarrow}}{\propto} P(X|\theta)P(\theta)$$

We know the posterior up to its normalizing constant...

Thought experiment:

- $P(\theta = 1|X) = \frac{P(X|\theta = 1)P(\theta=1)}{P(X)}$
- $P(\theta = 0|X) = \frac{P(X|\theta = 0)P(\theta=0)}{P(X)}$

Notice a common term??? marginal $P(X)$

will compare
there and
since $P(X)$
not change,
we can tell
which prob
is more likely

Random Walk Metropolis

Consider the pool game Marco-Polo

- Player is blindfolded, trying to find other players
- Player says “Marco,” other players say “Polo”
- Player moves in the direction of loudest “Polo”

Metropolis Marco-Polo–

- Opponent stays still
- Before you move, you point in a random direction
- Opponent says “Polo” with volume proportional to how far away your pointing is
- You chose to remain in the same location or move a couple of steps in the direction of your finger.

Random Walk Metropolis Sampler:

- You can evaluate $P(X|\theta)P(\theta)$ for any value of θ

1. Generate a proposal θ^* from a proposal distribution dependent on current $\hat{\theta}$ estimate.
2. Calculate $a = \frac{P(X|\theta^*)P(\theta^*)}{P\left(X\middle|\hat{\theta}\right)P(\hat{\theta})}$ = acceptance ratio
3. Sample r~Uniform(0,1)
4. If $r \leq a$, $\hat{\theta} = \theta^*$, else $\hat{\theta} = \hat{\theta}$ (only move if have better guess)

↓ switch to new est

↓ if not, don't

Metropolis Sampler - Example

Normal, unknown mean (10) and variance (5), with Jeffreys priors on both. (25 obs)

current val

proposal dist - will work eventually, but want better performance

1. Draw $\mu^* \sim \mathrm{N}(\hat{\mu}, .5)$

tuning param

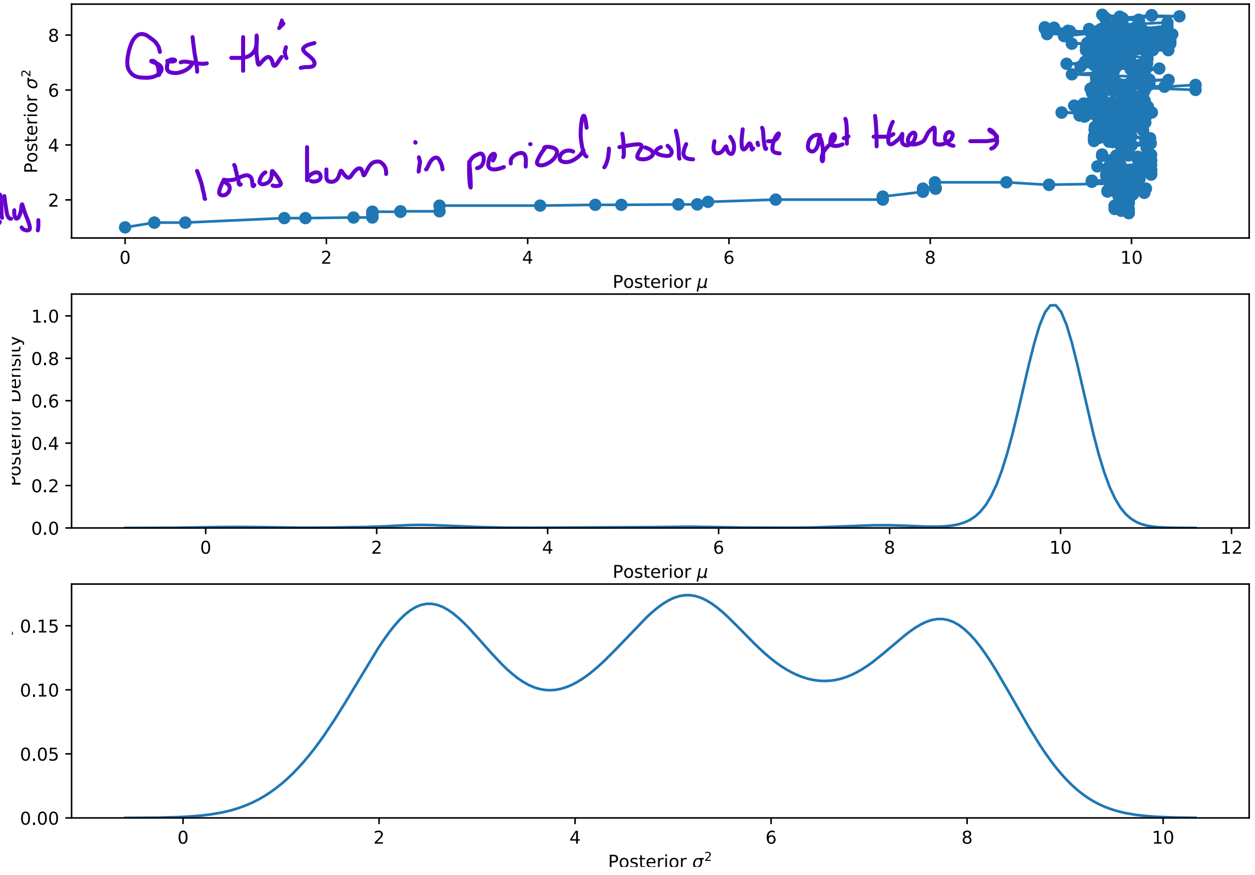
2. $$a_\mu = \frac{P\left(\bar{x} \middle| \mu^*, \widehat{\sigma^2}\right)c}{P\left(\bar{x} \middle| \hat{\mu}, \widehat{\sigma^2}\right)c}$$

3. Draw $r \sim Unif(0,1)$

r.v. you drew, every time do a step

   - If $r < a_\mu, \hat{\mu} = \mu^*$, else, $\hat{\mu} = \hat{\mu}$

Same for σ^{2*}, just with a smaller proposal density, and use of Jeffreys.



Get this

lotsa burn in period, took while get there →

Metropolis Sampler - Extensions

Metropolis Acceptance Ratio –

$$a = \frac{P(X|\theta^*)P(\theta^*)}{P(X|\hat{\theta})P(\hat{\theta})}$$

$*$ = your guess

$\wedge$ = what currently standing on

Only works when the proposal distributions are symmetric.

Metropolis-Hastings Acceptance Ratio –

$$a = \frac{P(X|\theta^*)P(\theta^*)\, g(\theta^*|\hat{\theta})}{P(X|\hat{\theta})P(\hat{\theta})g(\hat{\theta}|\theta^*)}$$

in practice, use this

Where $g(\theta^*|\hat{\theta})$ is the probability of proposing θ^* when “standing” at $\hat{\theta}$

Metropolis Sampler – Within Gibbs

You don’t need to sample your whole joint posterior in one shot (in fact, terrible idea).

Guideline: It tends to be easier to sample from smaller dimensional, conditional distributions, than it is to sample from high dimensional distributions

So, you don’t have nicely behaved conditionals (so no Gibbs) but you have many many parameters (so no one shot Metropolis-Hastings.).

Metropolis-in-Gibbs – Use Metropolis/Metropolis-Hastings to update each parameter conditional on the previous value of all other parameters.

- Instead of sampling directly from nicely behaved conditionals…

Metropolis Sampler – Considerations

* how far @ each step = proposal dist

Metropolis/Metropolis-Hastings is akin to navigating a city blindfolded, pointing random directions, and only moving when you hear you would be going in the right direction.

- It's better than randomly teleporting around the city until you get to your destination (rejection sampling).
- It's not quite as good as being told a how far to move vertically/horizontally and alternating (Gibbs)
- It's not nearly as good as just having the map (Conjugate priors)

<u>But you will eventually (most likely) get there.</u>

DISTRICT
Jacob Javits Center
Queens-Midtown Tunnel
Madison Square Garden Penn Station
Empire State Building
CHELSEA
GRAMERCY
STUYVESANT TOWN
GREENWICH VILLAGE
WEST VILLAGE
EAST VILLAGE
Williamsburg Bridge
SOHO
LITTLE ITALY
LOWER EAST SIDE
Holland Tunnel
TRIBECA
CHINA TOWN
Manhattan Bridge
Brooklyn Bridge
BROOKLYN
FINANCIAL DISTRICT
BATTERY PARK CITY
Stock Exchange
Battery Park
Ellis Island
Liberty Island
N

Metropolis Sampler – Considerations

Metropolis/Metropolis-Hastings is the brute force sampler, the monkeys banging on typewriters sampler.

Issues:

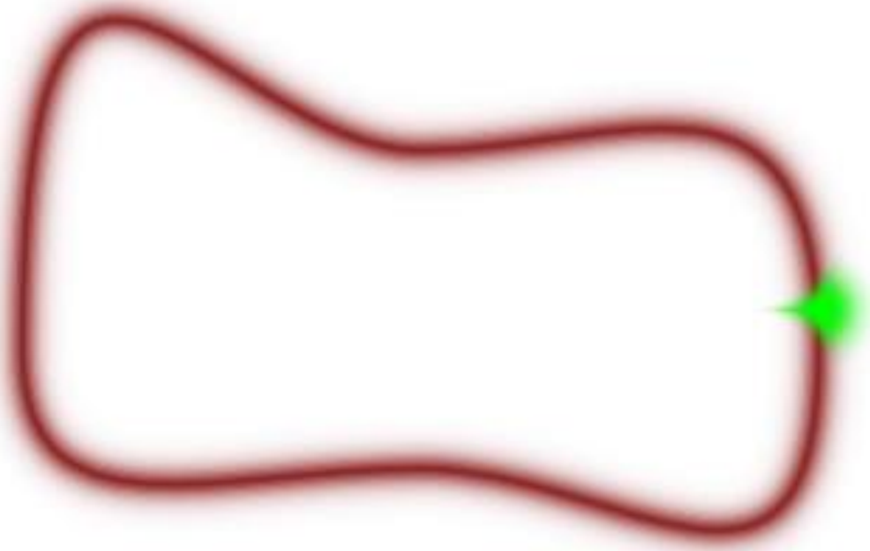
- What your proposal density is matters quite a bit
  - You don't want to make huge leaps and miss, and you don't want to crawl and not move…
- Starting locations – matter a lot
  - E.g. Mean is 1 billion, you start at 0. Proposals like 1000, 100, 10 are almost indistinguishable from 0, so why would you move?
- Slow as all get out
  - You have to first reach the posterior, then you have to sample from it

Hamiltonian Monte Carlo

Big issue: In high dimensions probability densities are concentrated around the mode, but due to this concentration, don't make much contribution to the expected value. What we want is the "typical set."

HMC – how works + sensors → Can tell if going up/down

- Uses gradient information about the target posterior (a topological map) - use info about slope to guide guesses
- Models our parameter estimates as "physical" particles moving along the posterior landscape
  - Give our guesses momentum so they keep going in the same "direction" ✪ - less likely to go backward bc know slope
- Instead of falling and getting trapped into high density areas, HMC samples from the posteriors typical area



From Fig 10 of "A Conceptual Introduction to Hamiltonian Monte Carlo" by Michael Betancourt ✪

• accessible technical

No U-Turn Sampler – NUTS

HMC is sensitive to two tuning parameters set by the analyst

- The main one is the number of steps the "physical" simulation runs for any given update.

If tuning is set poorly, the HMC might loop back to where you started

- In essence, it was too enthusiastic following the gradients.

NUTS automatically tunes the HMC to avoid those sorts of issues. – in Stan

– auto tuned; not handle categorical – can't apply NUTS here

Summary – Get Conceptual ideas

Samplers attempt to sample from the posterior distribution

- Gibbs – When you have conjugate conditional distributions – known forms
- Random Walk Metropolis Hastings – For, well, almost anything -(works for) -might take forever
- Hamiltonian MC – A much better version of MH-MCMC — gradient/slope info
  - For high dimensional continuous parameter sets
  - Not much good for discrete parameters like mixtures (no gradient information)
- NUTS – A auto-tuned Hamiltonian MC method – auto tuned -we will use
  - Integrated into PyMC3 and STAN
    → python

Issues arise with samplers because of how the posteriors probability landscape is structured (hills and valleys, flat plains, sharp corners all cause issues).

Next Week – Characterization and Computation

Diagnosing your Sampler – Sometimes fail

- Multiple sampling chains
- Convergence/divergence statistics
- Eyeballing your traceplots

What's the deal with categorical variables?

- Handling combinatorics
- Mixture models and label switching

HW3 will involve samplers! – Gibbs (EC = metrop hastings

- Some simple implementation, and a bit of diagnosing.