Bayesian approaches to parameter estimation

Epid 814 - Marisa Eisenberg

Bayesian approaches to parameter estimation

- Bayes' Theorem, rewritten for inference problems:

$$P(p \mid z) = P(params \mid data) = \frac{P(z \mid p) \cdot P(p)}{P(z)}$$

- Allows one to account for prior information about the parameters
    - E.g. previous studies in a similar population
- Update parameter information based on new data

Bayesian approaches to parameter estimation

Likelihood

Prior distribution

$$P(p \mid z) = P(params \mid data) = \frac{P(z \mid p) \cdot P(p)}{P(z)}$$

Normalizing constant
(can be difficult to calculate!)

$$P(z) = \int_p P(z, p) dp$$

Denominator term - P(z)

- The denominator term:

$$P(z) = \int_p P(z,p)dp$$

- Probability of seeing the data z from the model, over all parameter space
- Often doesn't have a closed form solution—evaluating numerically can also be difficult
  - E.g. if p is a three dimensional, then if we took 1000 grid points in each direction, the grid representing the function to be integrated has $1000^3 = 10^9$ points

Maximum *a posteriori* (MAP) estimation

- Instead of working with the full term, just use the numerator:

$$P(p|z) = \frac{P(z|p) \cdot P(p)}{P(z)}$$

- The denominator is a constant, so the numerator is proportional to the posterior we are trying to estimate

- Then the ***p*** which yields max($P(z|p) \cdot P(p)$) is the same ***p*** that maximizes $P(p|z)$

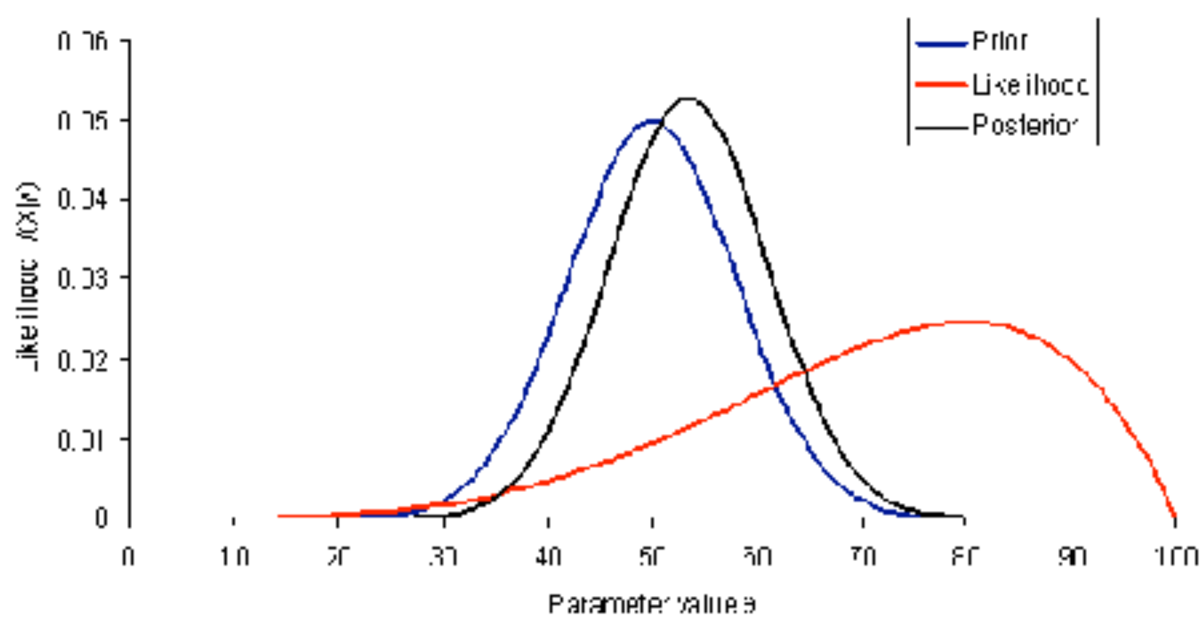
- If we only need a point estimate, MAP gets around needing to estimate $P(z)$

Bayesian Parameter Estimation

- Can think of Bayesian estimation as a map, where we update the prior to a new posterior based on data

$$P(p) \longrightarrow \boxed{\times \frac{P(z|p)}{P(z)}} \longrightarrow P(p|z)$$

Prior — Likelihood/P(z) — Posterior

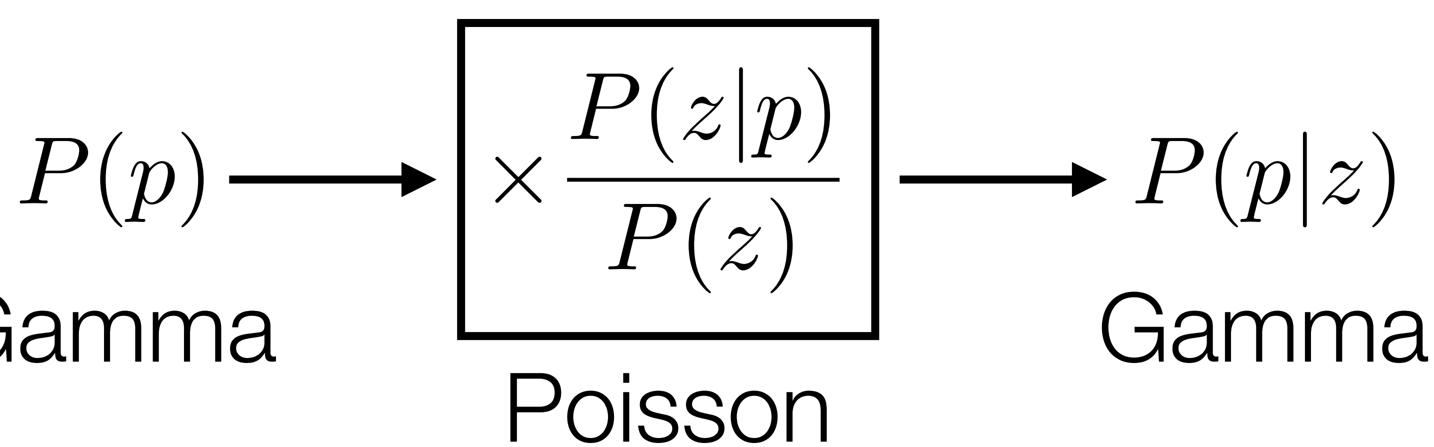


Conjugate Priors

- For a likelihood distribution, there may be a distribution family for our prior, which makes the posterior and prior come from the same type of distribution

- This is called a **conjugate prior** for that likelihood

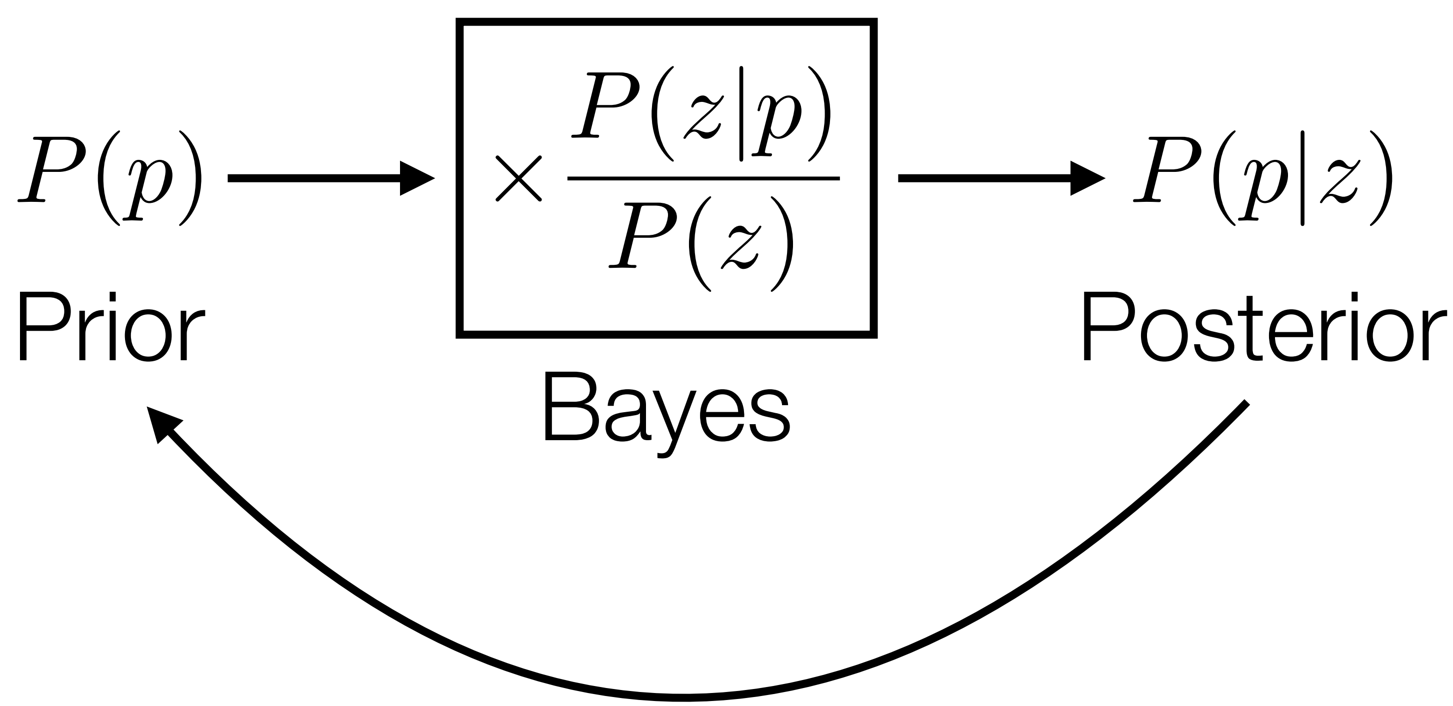
- For example, a gamma distribution is the conjugate prior for a Poisson likelihood.



Why conjugate priors?

- If we have a conjugate prior, we can calculate the posterior directly from the likelihood and the prior—handles the issue with calculating the denominator P(z)

- Also makes it easier to repeat Bayesian estimation—making the posterior the prior and updating as new data comes in



Conjugate prior example: coin flip

- Let z be the data—i.e. the coin flip outcome, z = 1 if it's heads, z = 0 if it's tails

- Let θ be the probability the coin shows heads

- Likelihood: Bernoulli distribution

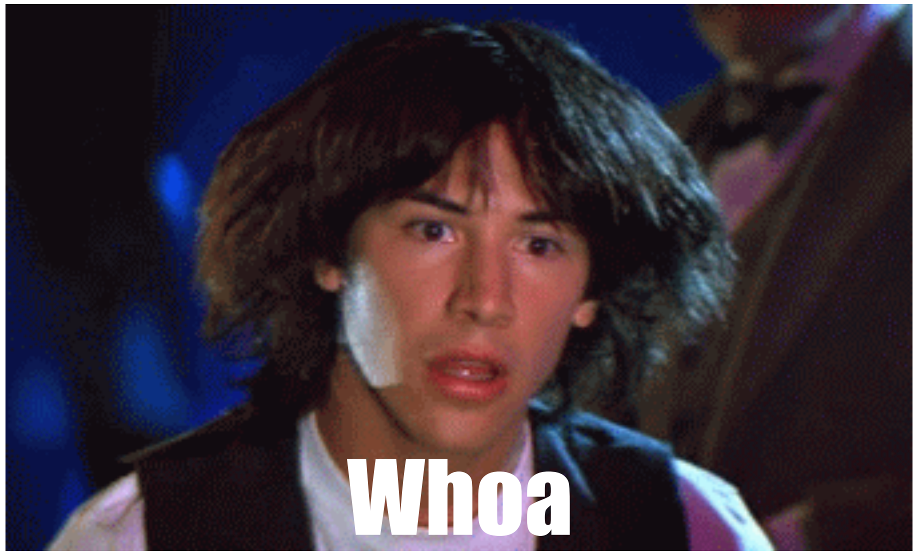
$$P(z|\theta) = \theta^z (1 - \theta)^{1-z}$$

Conjugate prior example: coin flip

- **Conjugate prior**: beta distribution

$$P(\theta|\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}$$

- α and β are **hyperparameters** - shape parameters that describe the distribution of the model parameters



How does the posterior work out to be a beta distribution as well?

$$P(\theta|z) = \frac{P(z|\theta)P(\theta|\alpha,\beta)}{P(z)}$$

$$= \frac{\theta^z(1-\theta)^{1-z}\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}}{P(z)}$$

$$= \frac{\theta^z(1-\theta)^{1-z}\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}}{\int_0^1 P(z,\theta)d\theta}$$

$$= \frac{\theta^z(1-\theta)^{1-z}\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}}{\int_0^1 \theta^z(1-\theta)^{1-z}d\theta}$$

Etc.—but you can see it will work out to be beta distributed

Coin flip example - Posterior

- **Beta distributed** with posterior hyperparameters:

$$\alpha_{post} = \alpha + z \qquad \beta_{post} = \beta + 1 - z$$

- If we take multiple data points, this works out to be:

$$\alpha_{post} = \alpha + \sum_{i=1}^{n} z_i \qquad \beta_{post} = \beta + n - \sum_{i=1}^{n} z_i$$

Sampling Methods

- What if we want priors that aren't conjugate? Or what if our likelihood is more complicated and it isn't clear what the conjugate prior is?
- Now we need some way to get the posterior, even though the denominator term is annoying
- Sampling-based methods—in particular, **Markov chain Monte Carlo (MCMC)**

Markov Chain Monte Carlo (MCMC)

- **MCMC** is a method for sampling from a distribution
- **Markov chain**: a type of (discrete) Markov process
  - Markov: memoryless, i.e. what happens at the next step only depends on the current step
- **Monte Carlo methods** are a class of algorithms that use sampling/randomness—often used to solve deterministic problems (such as approximating an integral)

Markov Chain Monte Carlo (MCMC)

- **Main idea**: make a Markov chain that converges to the distribution we're trying to sample from (the posterior)
  - The Markov chain will have some transient dynamics (burn-in), and then reach an equilibrium distribution which is the one we're trying to approximate

Markov Chain Monte Carlo (MCMC)

- Many MCMC methods are based on random walks
  - Set up walk to spend more time in higher probability regions
- Typically don't need the actual distribution for this, just something proportional—so we can get the relative probability density at two points
  - So we don't need to calculate P(z)! We can just use the numerator

Example: Metropolis Algorithm

- Idea is to 'walk' randomly through parameter space, spending more time in places that are higher probability—that way, the overall distribution draws more from higher probability spots

- Setup—we need

  - A function $f(p)$ proportional to the distribution we want to sample, in our case $f(p) = P(z|p) \cdot P(p)$

  - A proposal distribution (how we choose the next point from the current one) - more on this in a minute

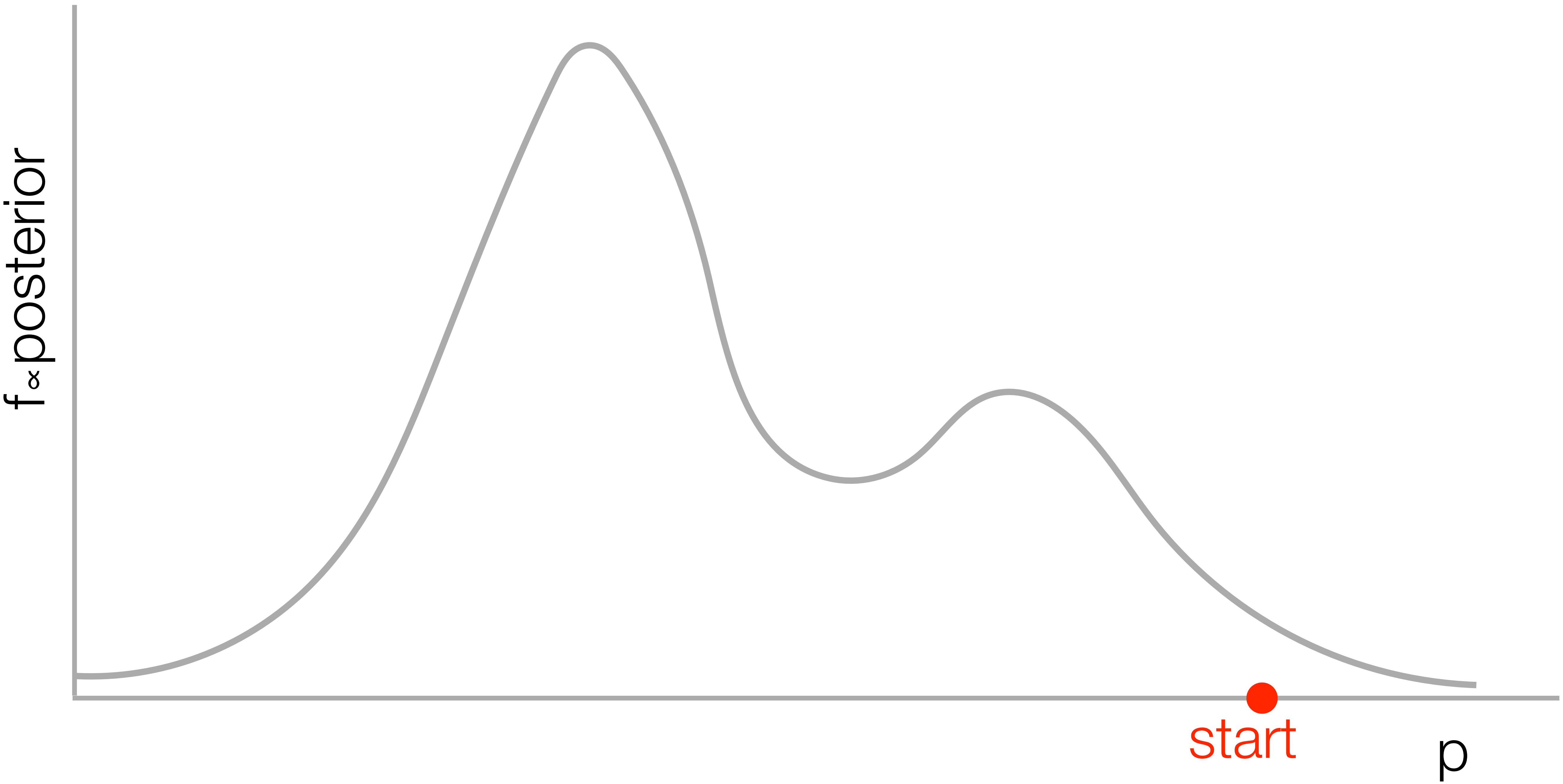
Metropolis Algorithm

- Start at some point in parameter space
- For each iteration
    - Propose a new random point p_{next} based on the current point p_{curr} (using the proposal distribution)
    - Calculate the **acceptance ratio**, $\alpha = f(p_{next})/f(p_{curr})$
        - If $\alpha \geq 1$, the new point is as good or better—accept
        - If $\alpha < 1$, accept with probability α

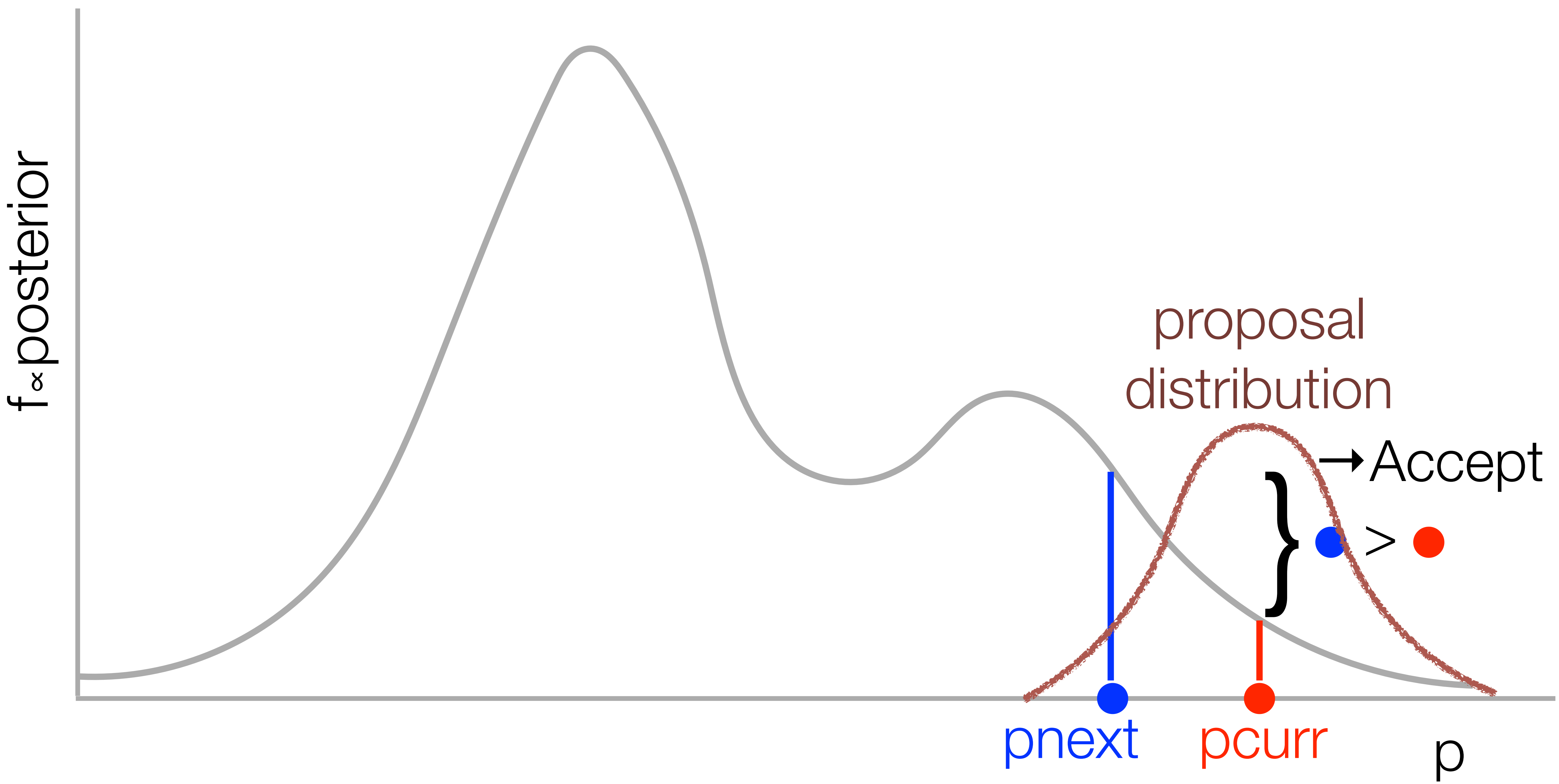
Proposal Distribution

- A distribution that lets us choose our next point randomly from our current one
- For Metropolis algorithm, must be symmetric
- Common to choose a normal distribution centered on current point
- Width (SD) of normal = proposal width
    - Choice of proposal width can strongly affect how the Markov chain behaves, how well it converges, mixes, etc.

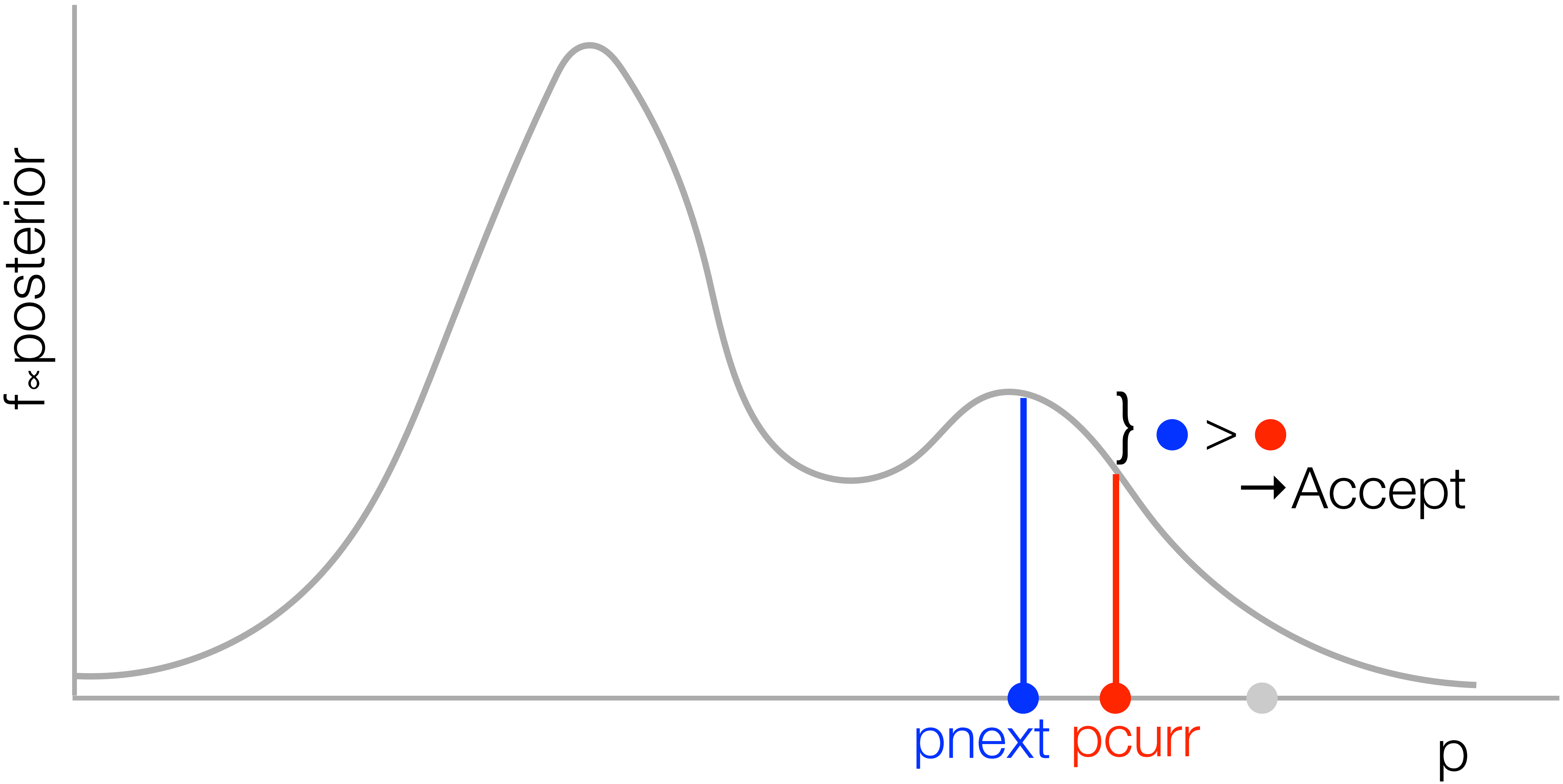
What does the metropolis algorithm do?



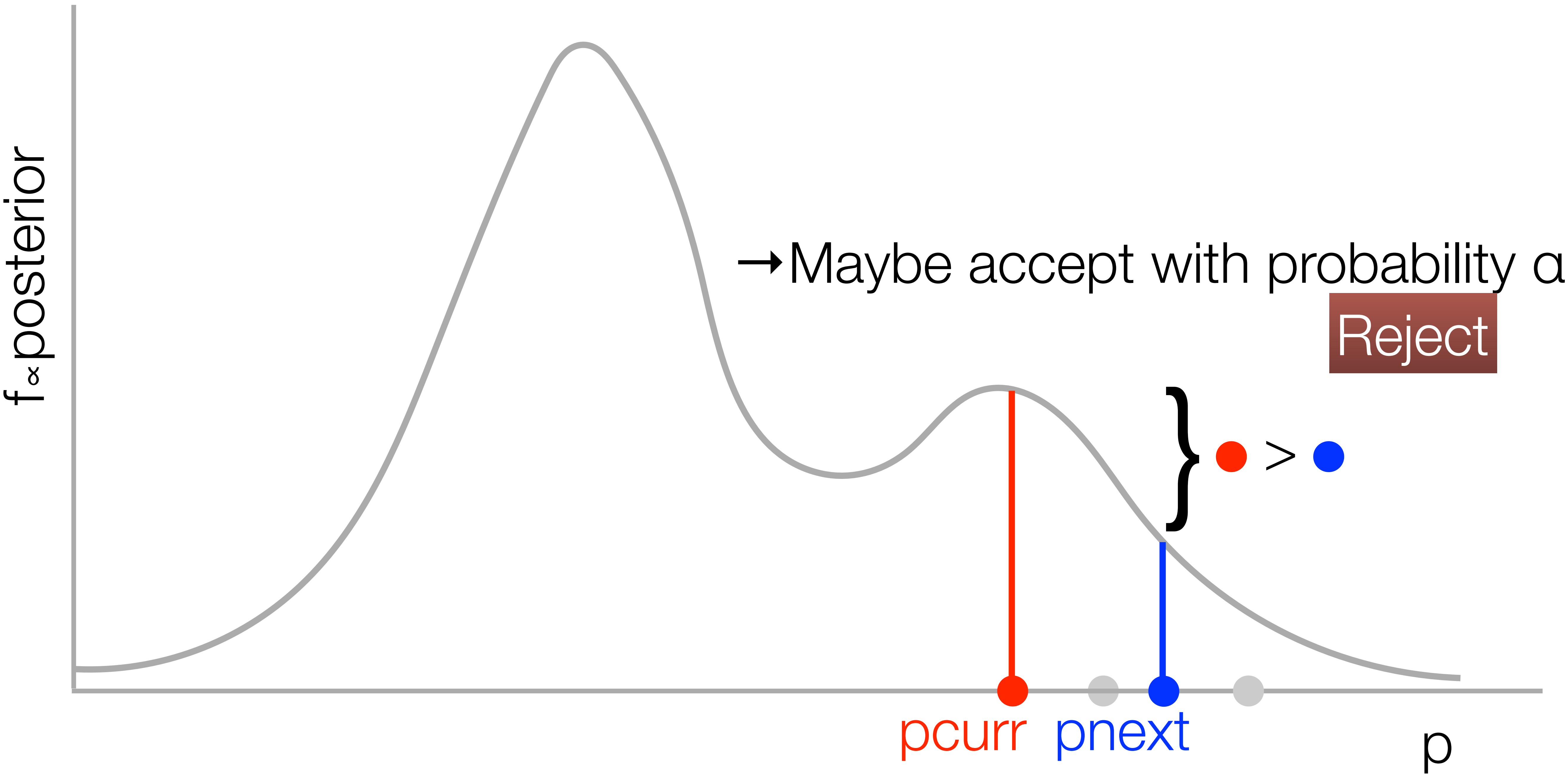
What does the metropolis algorithm do?



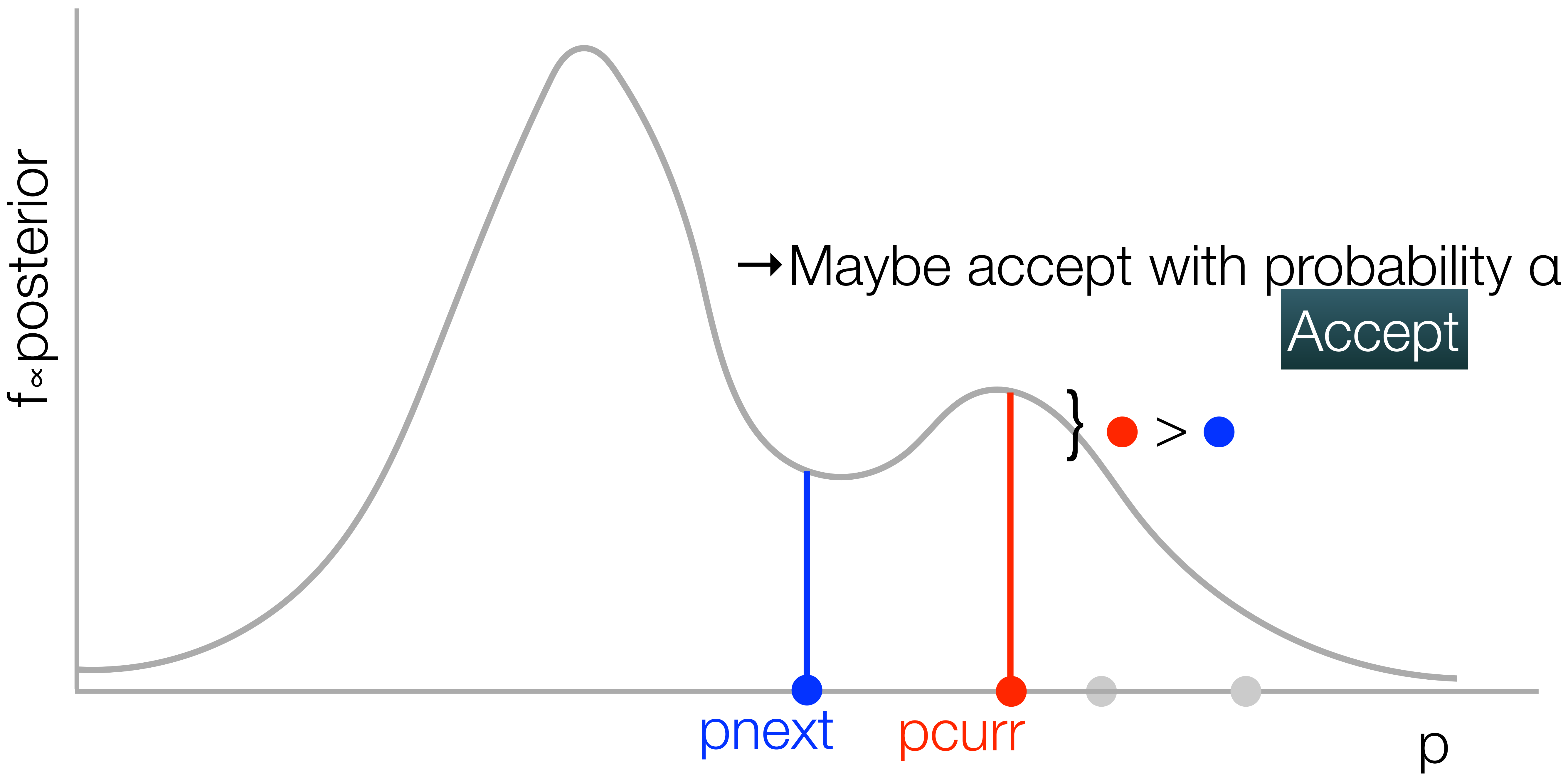
What does the metropolis algorithm do?



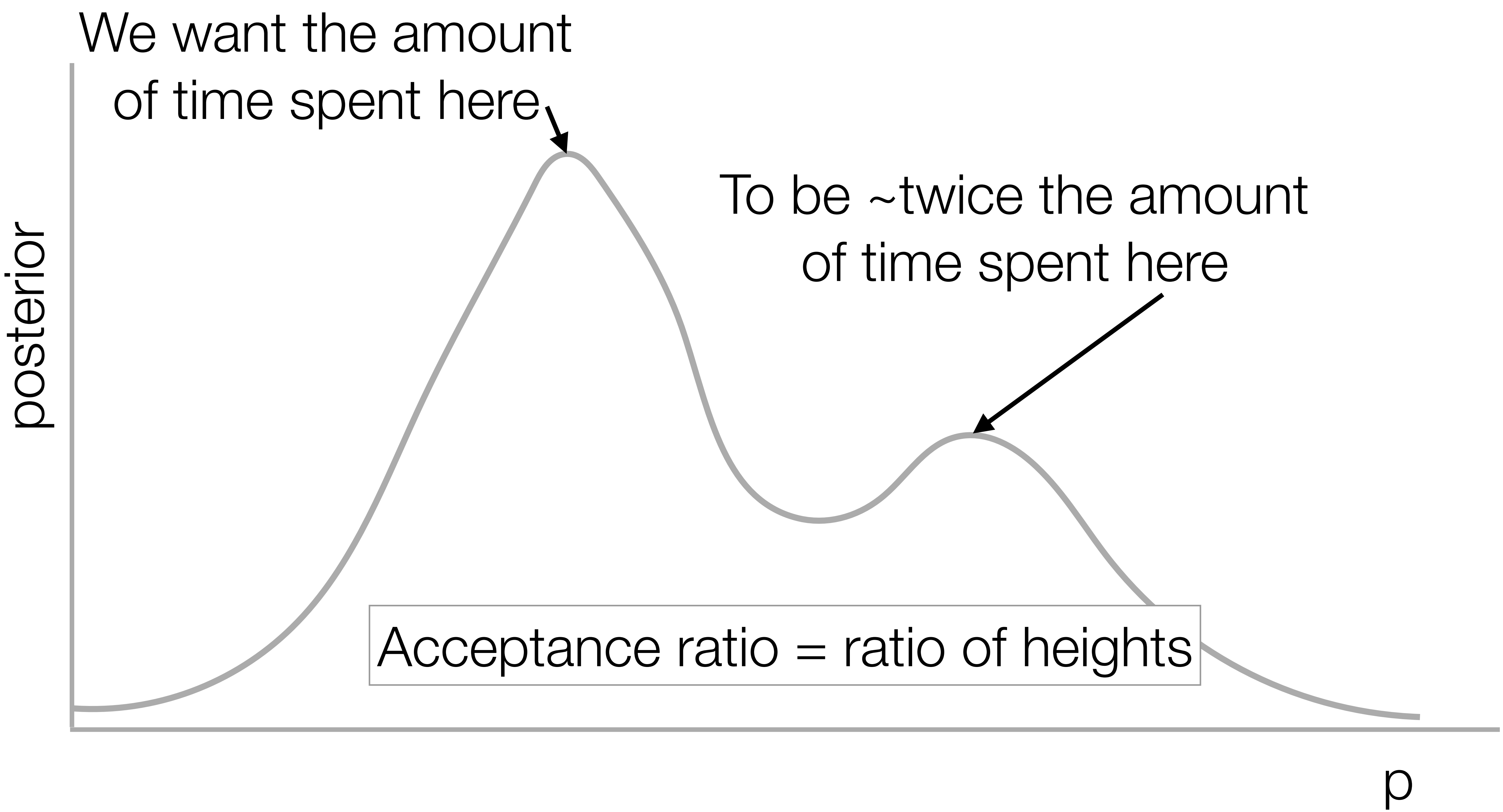
What does the metropolis algorithm do?



What does the metropolis algorithm do?



Why does this recover the posterior distribution? Key is the acceptance ratio α

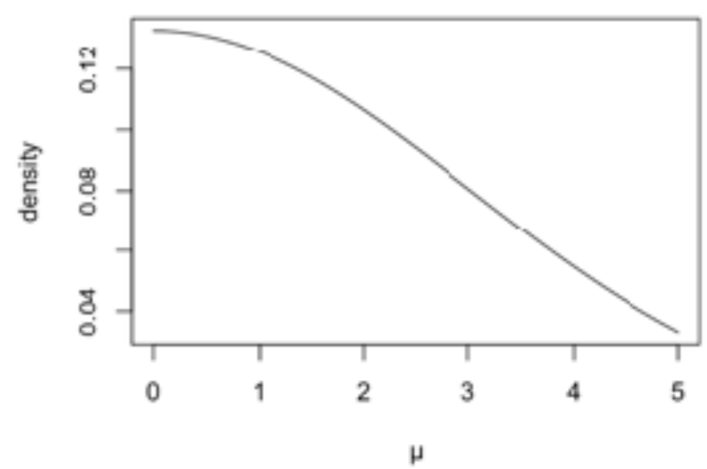


Why does this recover the posterior distribution?

- The acceptance ratio $\alpha = f(p_{next})/f(p_{curr})$
- Note it is equal to $P(p_{next}|z)/P(p_{curr}|z)$ since the denominators cancel
- Suppose we're at the peak
    - If $f(p_{curr}) = 2\ f(p_{next})$, then $\alpha = 1/2$, i.e. we accept with 1/2 probability
- Overall, will mean the number of samples we take from a region will be proportional to the height of the distribution

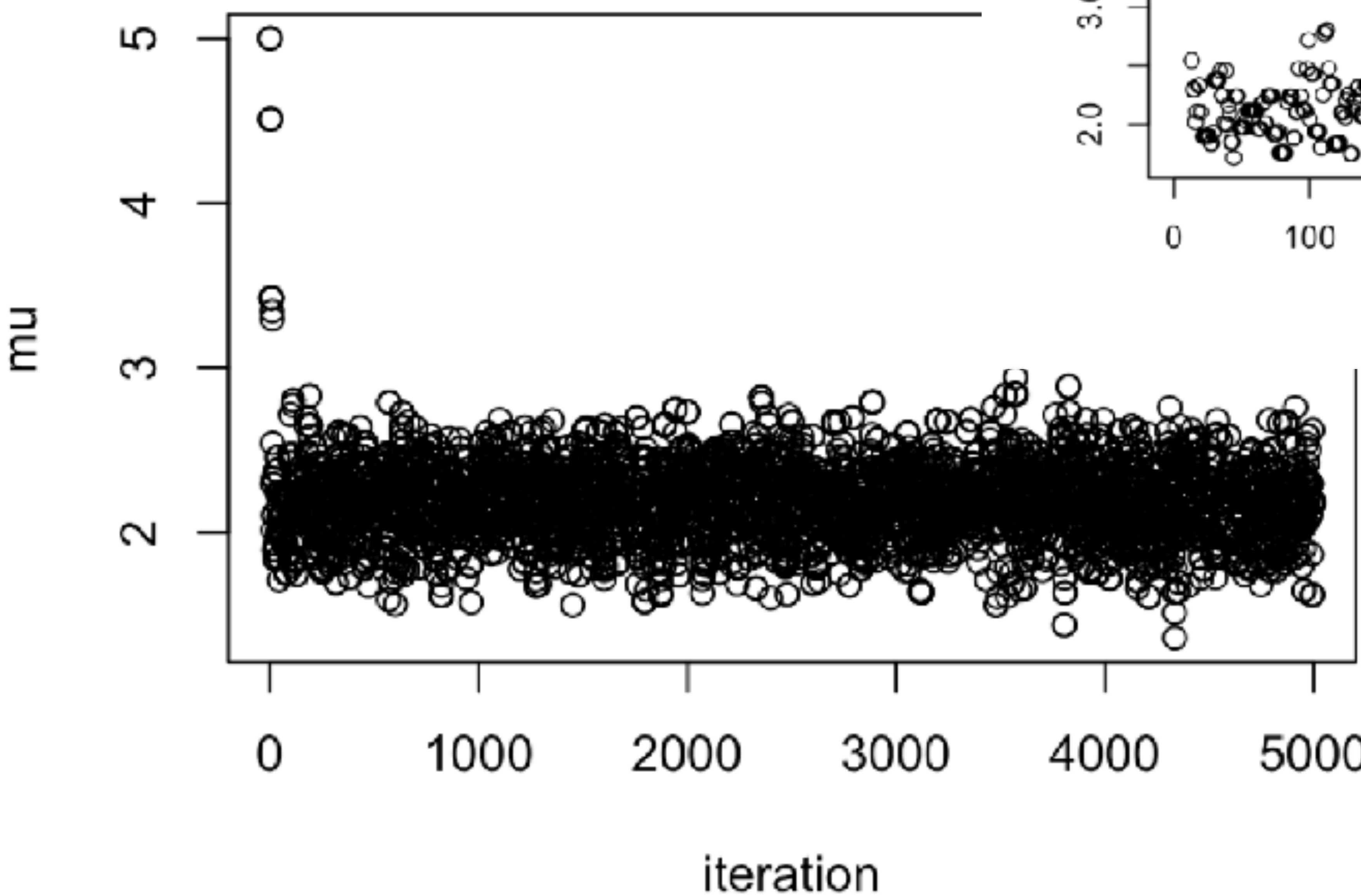
Example

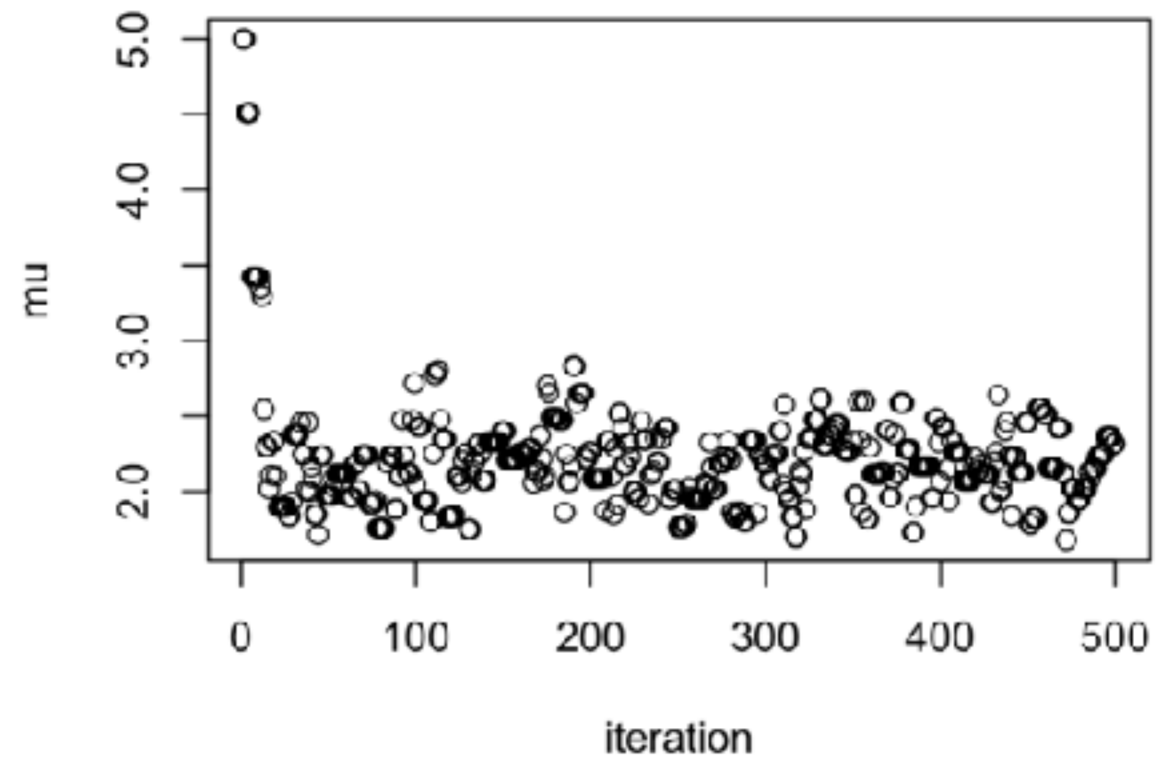
- Model: normal distribution $\mathcal{N}(\mu, \sigma)$
  - Suppose σ is known, μ to be estimated
  - Likelihood: $P(z_i \mid \mu, 1) = f(z_i \mid \mu, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}}$ $\quad P(z \mid \mu) = \prod_{i=1}^{n} f(z_i \mid \mu, 1)$
- Prior: $\mu \sim \mathcal{N}(0, 3)$



- Suppose we have 20 data points

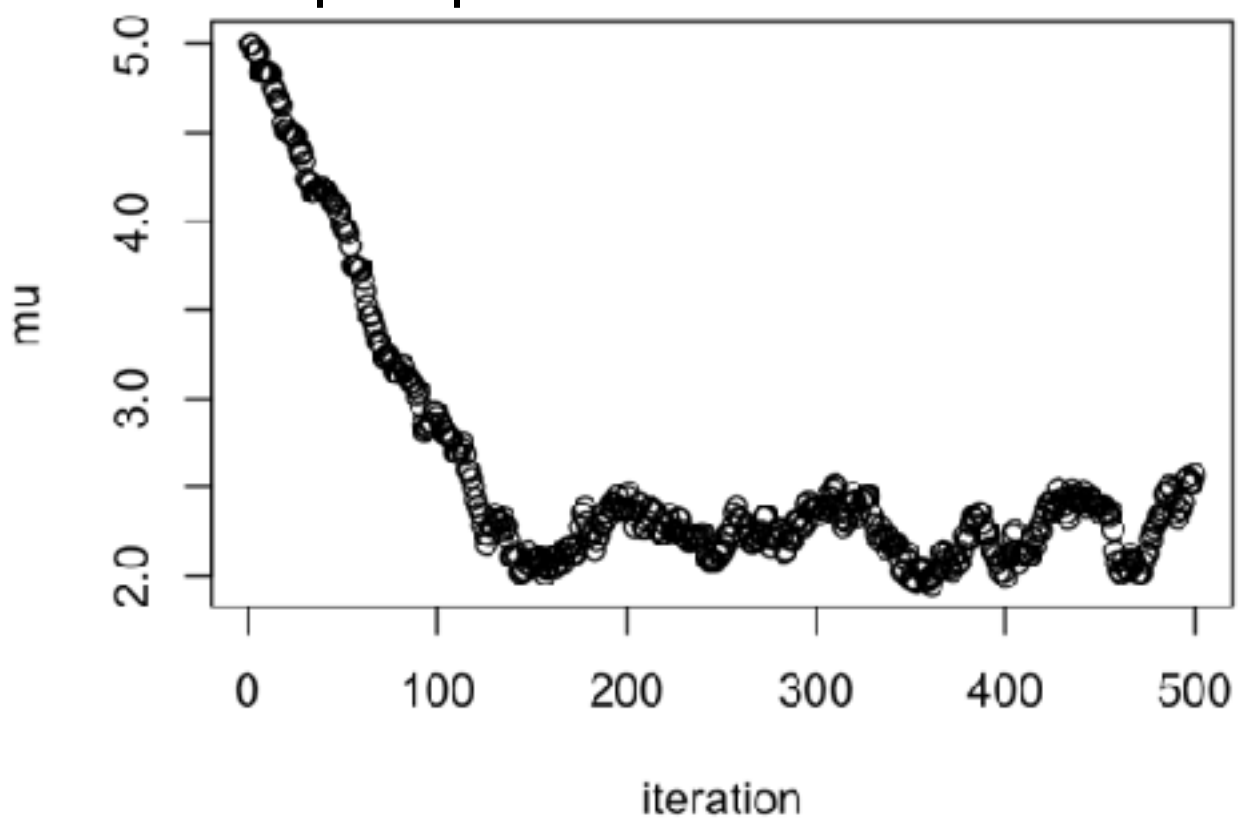
Example - proposal width: SD = 0.5



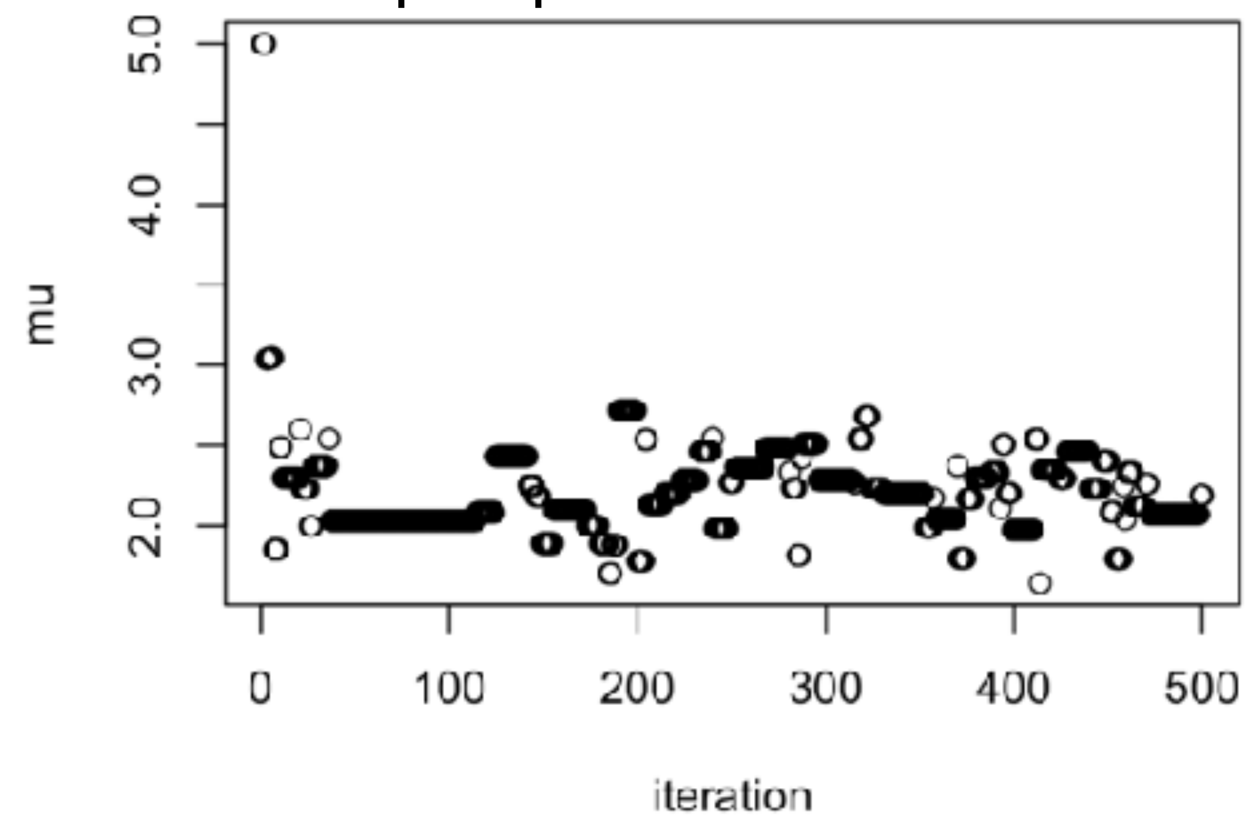


Goldilocks problem:
What happens if we change the proposal width?

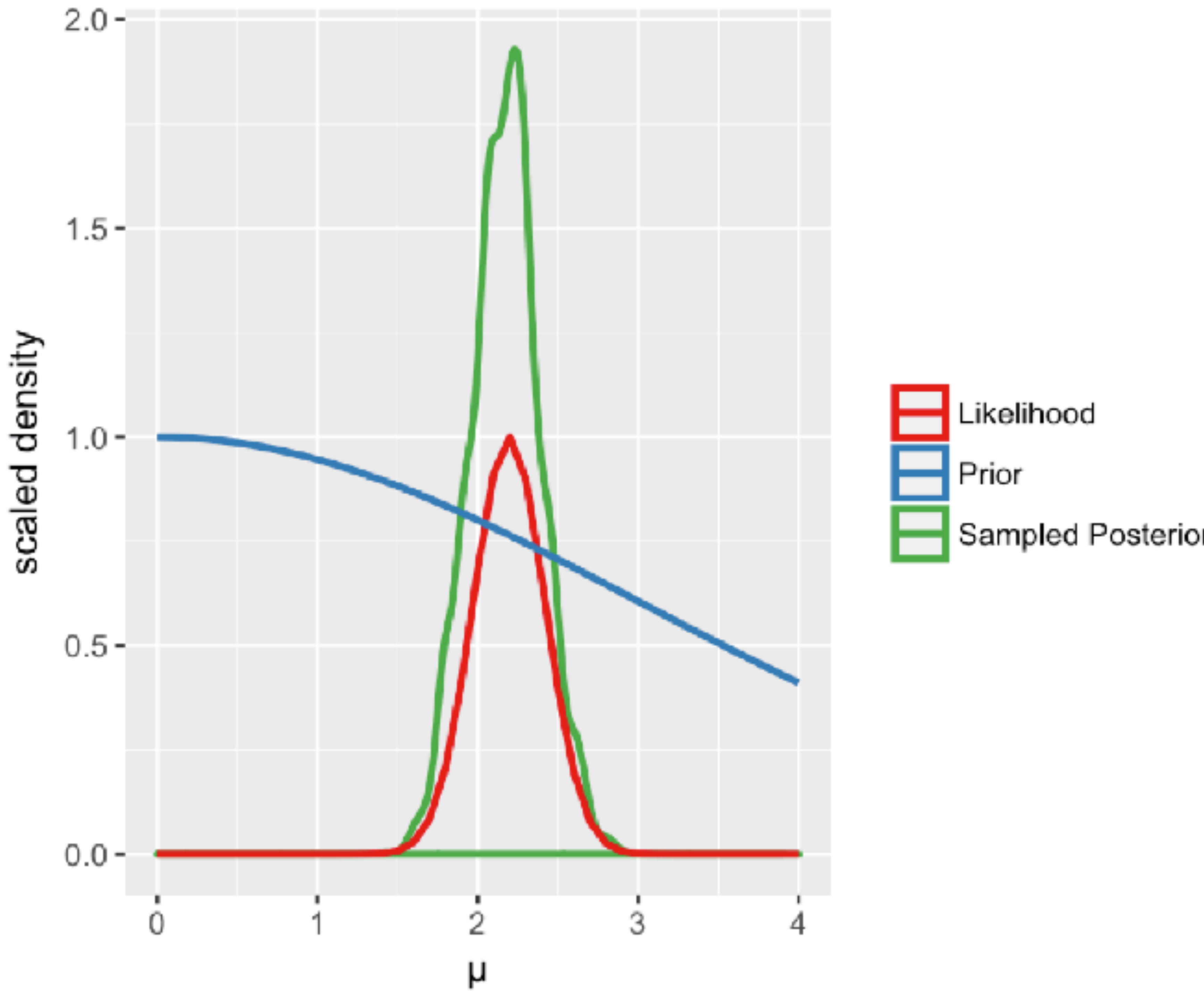
proposal SD = 0.05



proposal SD = 2



Example: prior, likelihood, and posterior (all scaled)

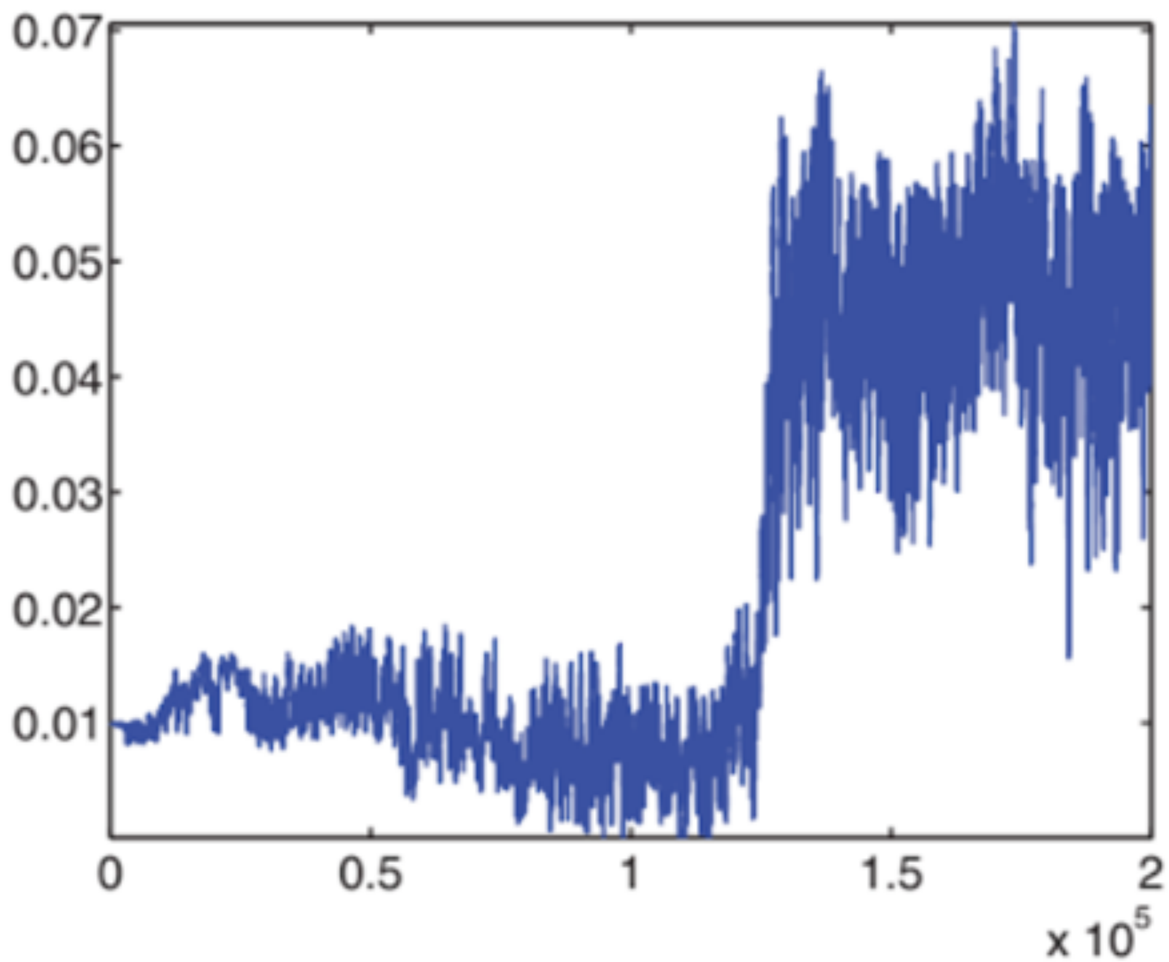


Assessing convergence

- MCMC methods will let us sample the posterior once they've converged to their equilibrium distribution
- How to know once we've reached equilibrium?
  - Visual evaluation of **burn-in**
  - Acceptance ratio
  - Autocorrelation of elements in chain *k* iterations apart
- Careful about MCMC results reported in literature

Assessing convergence

- Often done visually
- Although, this can be misleading:



Chain shifts after 130,000 iterations due to a local min in sum of squares (Example from R. Smith, *Uncertainty Quantification*)

Metropolis & Metropolis-Hastings Caveats

- Assessing convergence—how long is burn-in?
- Correlated samples
- How to choose a proposal width? (~size of next jump)

Next time

- Code our own Metropolis sampler
- Then, next week:
    - More on priors—uninformed priors (e.g. Jeffreys), identifiability issues
    - Work with sampling packages & more realistic models!