Lecture 16: Bayesian Approach in Action

Instructor: Sergei V. Kalinin

Bayesian paradigm in science

Posterior: probability of the model given the data

Likelihood: probability of the data given the model

Prior: probability of the model

$$P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model})\,P(\text{model})}{P(\text{data})}$$

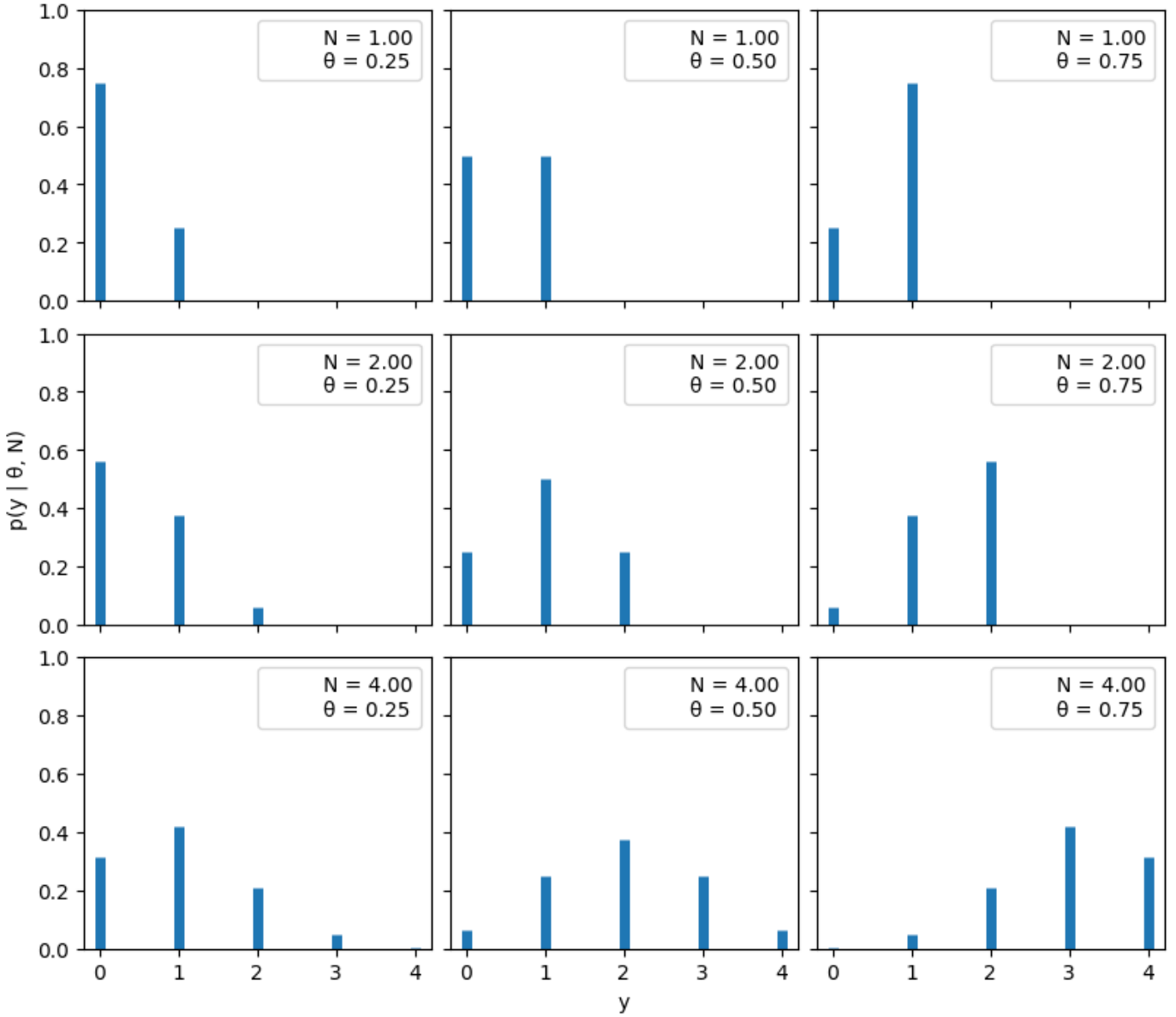
Evidence [can typically be absorbed into the normalization of the posterior]

Coin Toss

- **Probability of tails:** p
- **Probability of heads:** 1-p

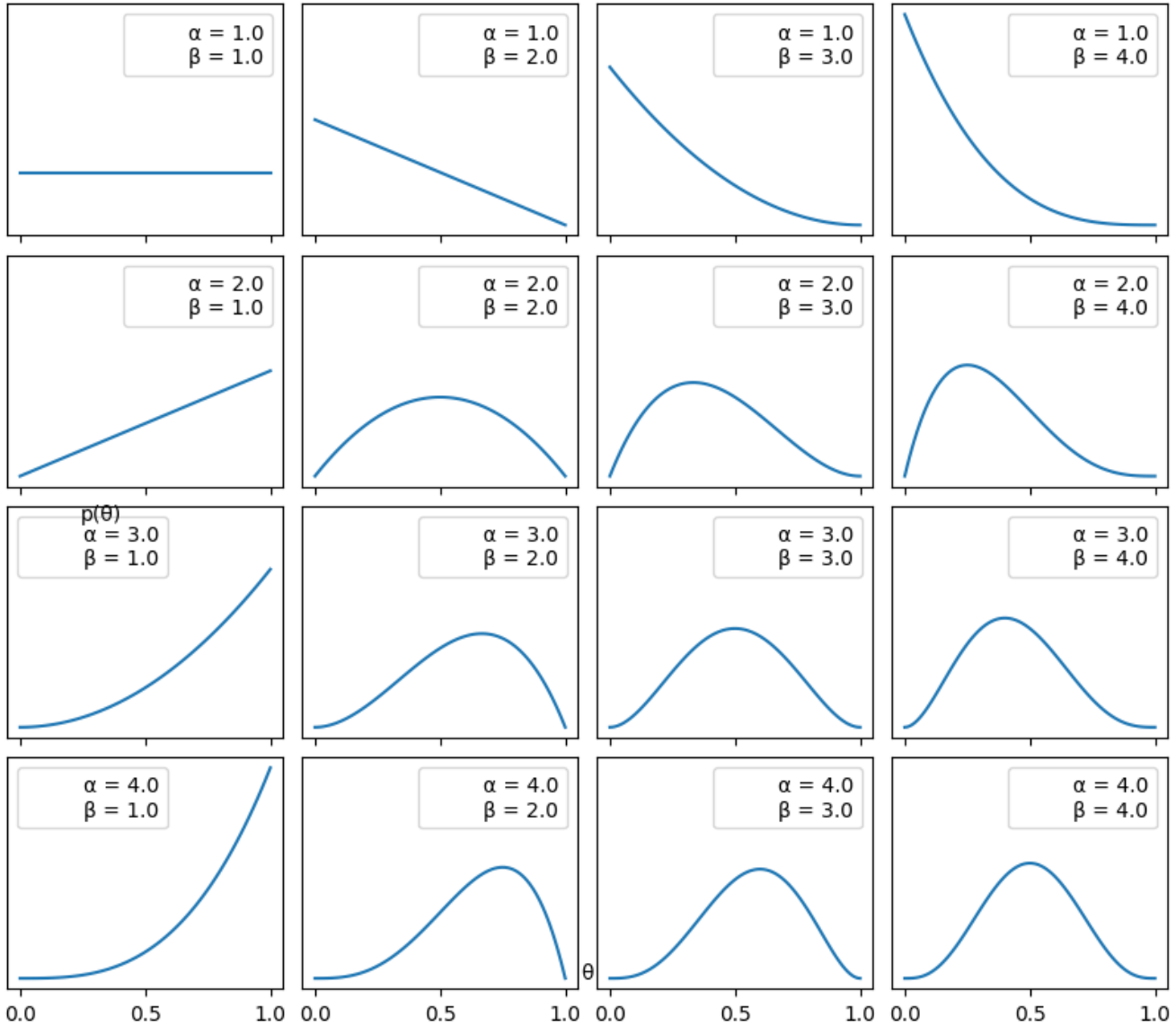
- Probability of getting n tails out of N tosses of coin:

$$C_N^n p^n (1-p)^{N-n}$$



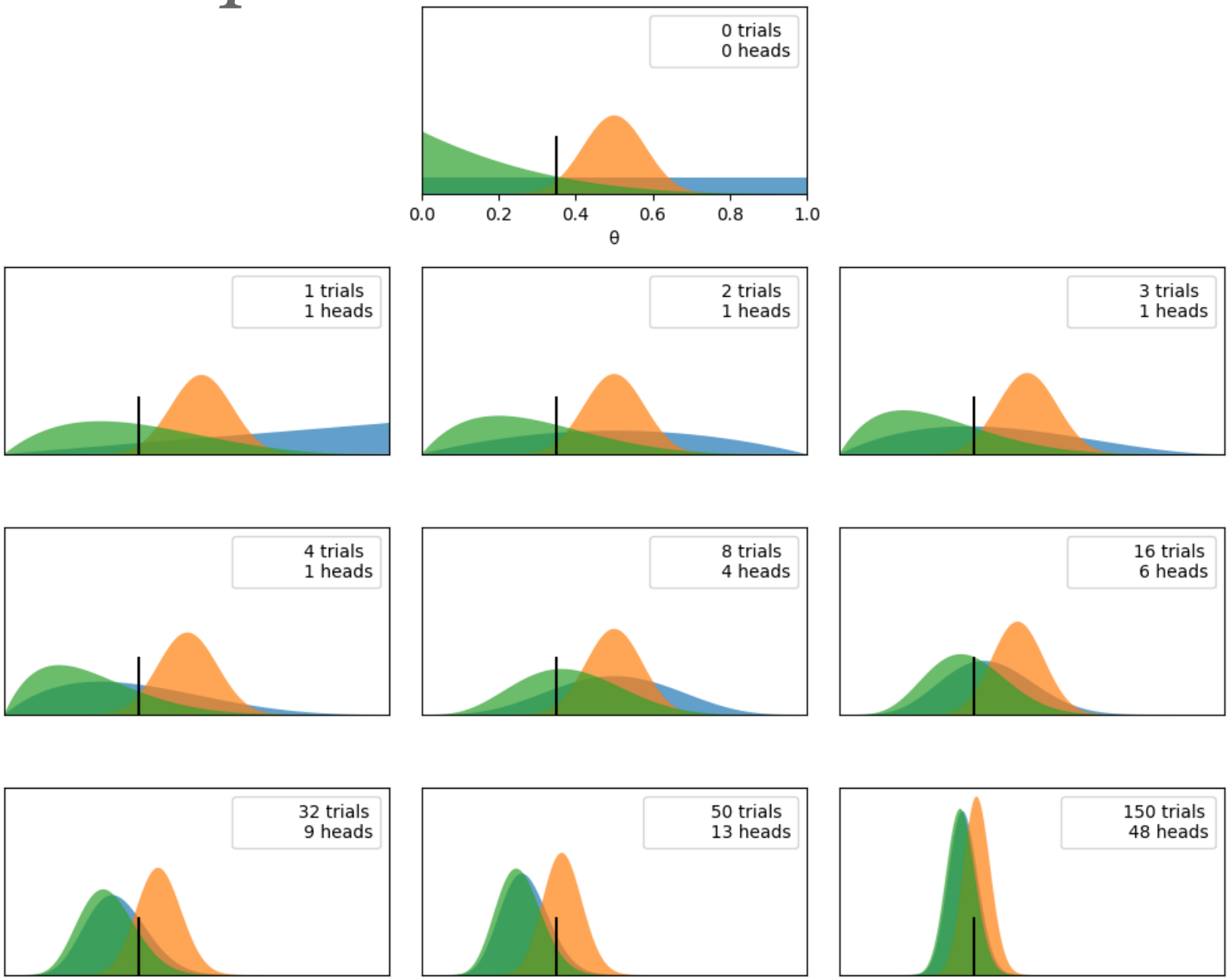
Beta distribution: conjugate to binomial

$$
\begin{aligned}
f(x;\alpha,\beta) &= \text{constant} \cdot x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}
\end{aligned}
$$



https://en.wikipedia.org/wiki/Beta_distribution

Can we learn p from several coin tosses?

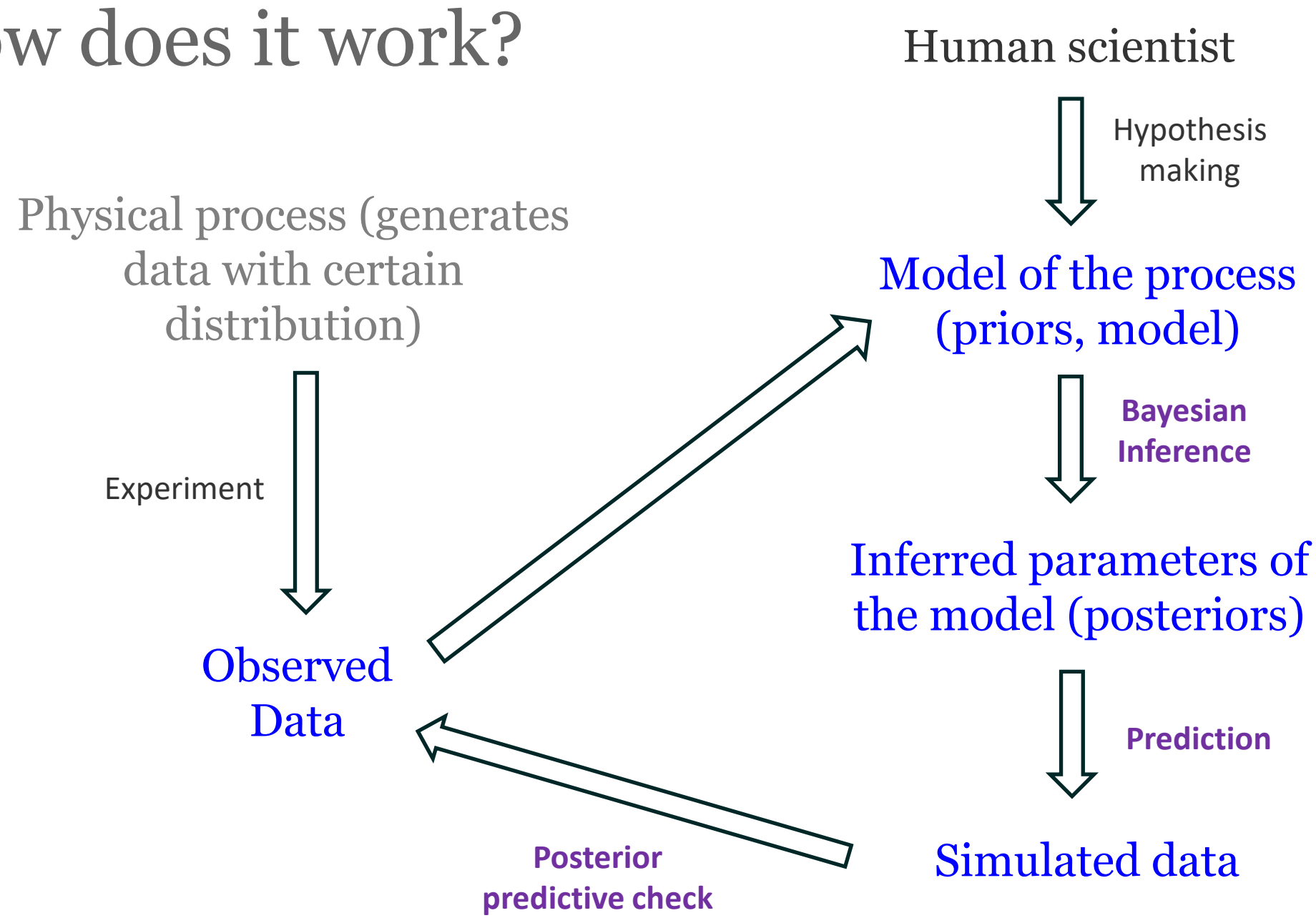


- Given some data and some assumptions on how this data could have been generated, we design a model by combining building blocks known as probability distributions. Most of the time these models are crude approximations, but most of the time that's all we need.

- We use Bayes' theorem to add data to our models and derive the logical consequences of combining the data and our assumptions. We say we are conditioning the model on our data.

- We evaluate the model, and its predictions, under different criteria, including the data, our expertise on the subject, and sometimes by comparing it to other models.

How does it work?



Main Elements of Bayesian Models

- **Prior Distribution** – use probability to quantify uncertainty about unknown quantities (parameters)
- **Likelihood** – relates all variables into a "full probability model"
- **Posterior Distribution** – result of using data to update information about unknown quantities (parameters)

Why has Bayes not been popular?

(1) Without Markov Chain Monte Carlo, it wasn't practical.

(2) Some people distrust prior distributions, thinking that science should be objective (as if that were possible).

Bayes is becoming much more common, due to MCMC.

From Jeff Witmer, Introduction to Bayesian Statistics, slides 24 February 2016

Conjugate distributions

Conjugate Prior: A prior distribution is said to be conjugate to a likelihood function if the resulting posterior distribution is in the same family as the prior. In other words, if you start with a certain type of distribution as your prior, and after observing data and updating your beliefs (via Bayes' theorem), your posterior is still of that same type, then the prior is a conjugate prior for that likelihood function.

• **Computational Convenience**: Using conjugate priors can greatly simplify the mathematical computation required to find the posterior distribution. This can be especially useful in situations where you're continually updating your beliefs with new data; with conjugate priors, you can easily update your posterior without complex integrals or advanced sampling methods.

•**Analytical Solutions**: Many standard problems in Bayesian statistics can be solved analytically using conjugate priors, leading to exact posterior distributions.

Conjugate distributions

1. **Beta distribution is conjugate to the Binomial likelihood**: This means that if you have a Binomial likelihood (e.g., flipping coins) and a Beta-distributed prior on the probability of heads, the resulting posterior distribution after observing some data will also be a Beta distribution.

2. **Gamma distribution is conjugate to the Poisson likelihood**: If you're observing the number of events occurring in fixed intervals of time or space (modeled by a Poisson distribution) and have a Gamma-distributed prior on the rate parameter, the posterior will also be Gamma-distributed.

3. **Normal distribution is conjugate to itself**: If both the likelihood and the prior are normally distributed, then the posterior will also be normally distributed.

Priors are the key!

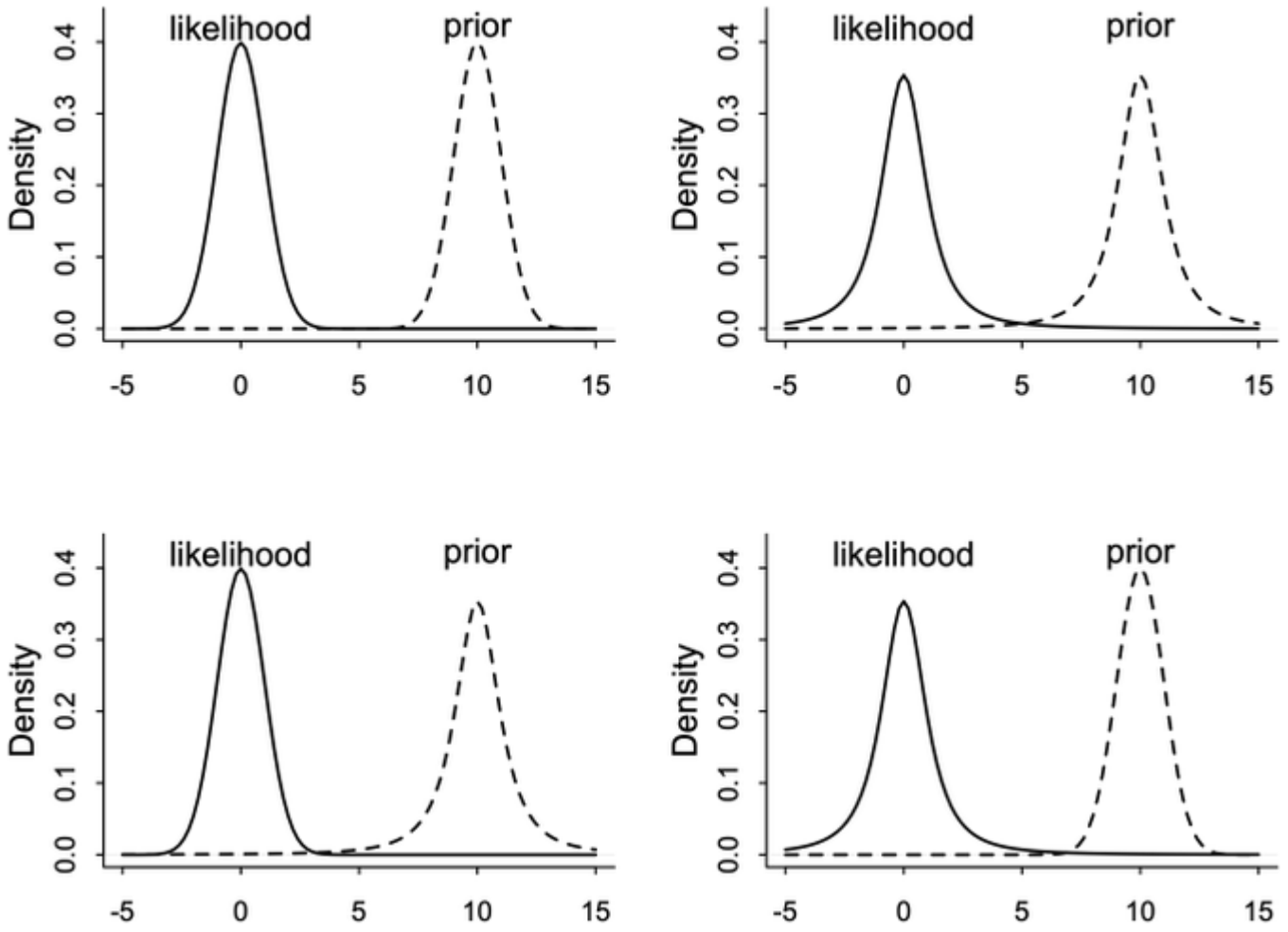
- Most Bayesian analysts assume “uninformative priors”
  - no strong assumptions about the parameter estimates other than the shape of their distributions
- When we use uninformative priors and analyze the data using both a traditional approach and a Bayesian approach, the resulting parameter estimates are the same (for all practical purposes) = *no strong rationale for Bayesian*
  - Uninformative priors means the results are strongly determined by the current experiment’s data

When to use strong priors

- When you are willing to specify informative priors
- When there is no existing analysis for your design

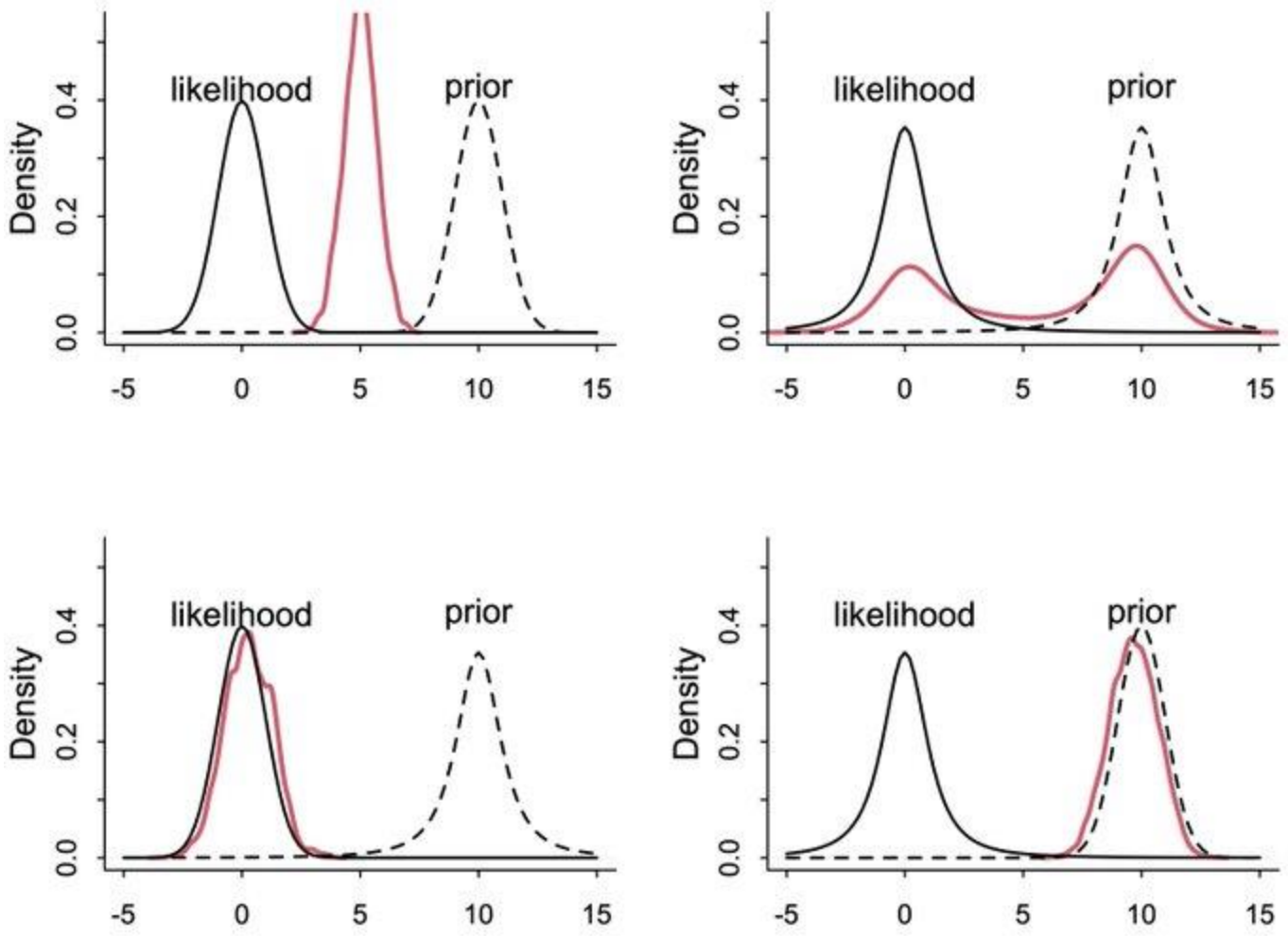
- Narrower priors will have a bigger effect on the posterior estimation
- The effect will be larger if the new data is limited or highly variable
  - This situation indicates that the new data are equivocal and offer little new

McElreath Quartet



https://twitter.com/rlmcelreath/status/1701165075493470644

McElreath Quartet



https://twitter.com/rlmcelreath/status/1701165075493470644

Probability of data is generally intractable

- Prior information $p(\theta)$ on parameters θ
- Likelihood of data given parameter values $f(x|\,\theta)$

Problem for physics/theory

$$p(\theta|x) = \frac{f(x|\theta)\,p(\theta)}{f(x)}$$

Problem for computation

$$f(x) = \int_{-\infty}^{\infty} f(x|\theta)p(\theta)d\theta$$

To use Bayesian methods, we need to be able to evaluate the denominator, which is the integral of the numerator over the parameter space. In general, this integral is very hard to evaluate.

Colab

Solution: Markov Chain Monte Carlo

We don't need to evaluate any integral, *we just sample from the distribution many times* (e.g., 50K times) and find (estimate) the posterior mean, middle 95%, etc., from that.

Metropolis-Hastings:

- An algorithm that generates a sequence
  $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots\}$ from a Markov Chain whose stationary distribution is $\pi(\theta)$ (i.e., the posterior distribution)
- Fast computers and recognition of this algorithm has allowed Bayesian estimation to develop.

From Jeff Witmer, Introduction to Bayesian Statistics, slides 24 February 2016

Metropolis-Hastings algorithm

We don't need to evaluate any integral, *we just sample from the distribution many times* (e.g., 50K times) and find (estimate) the posterior mean, middle 95%, etc., from that.

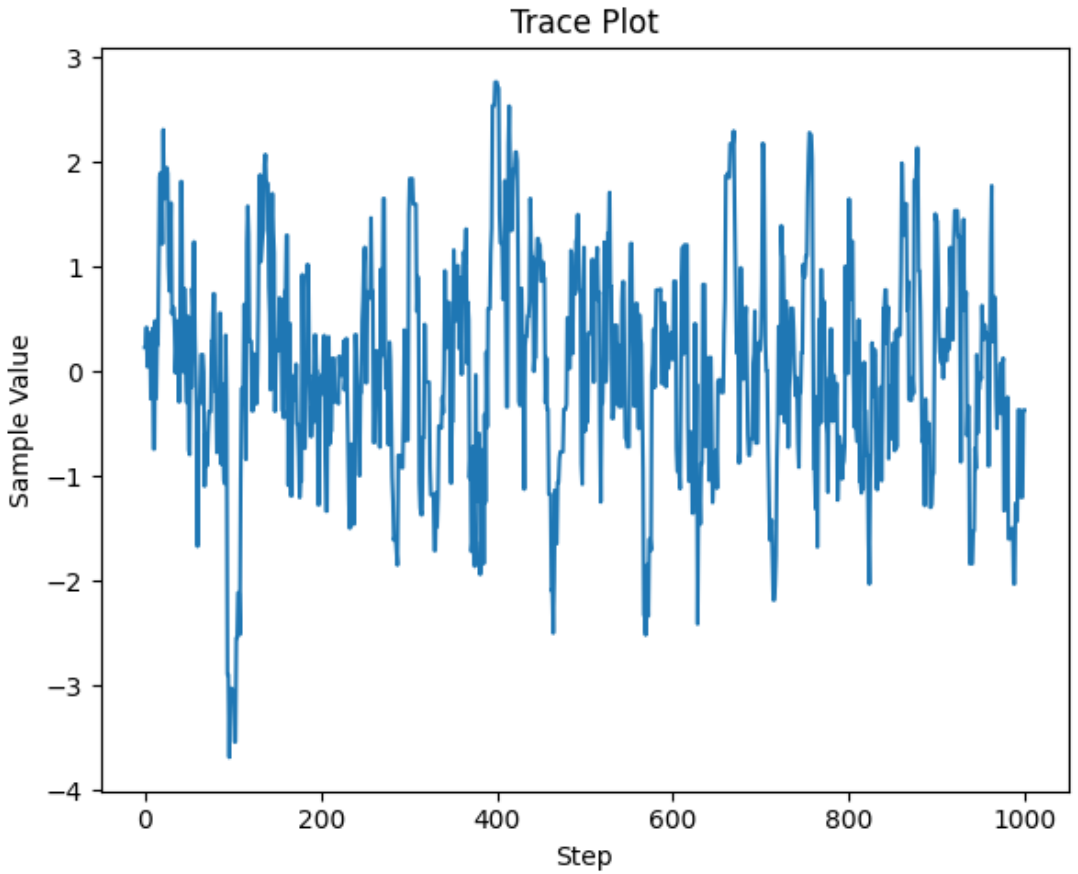
- Initial value $\theta^{(0)}$ to start the Markov Chain
- Propose new value
- Accepted value: θ'

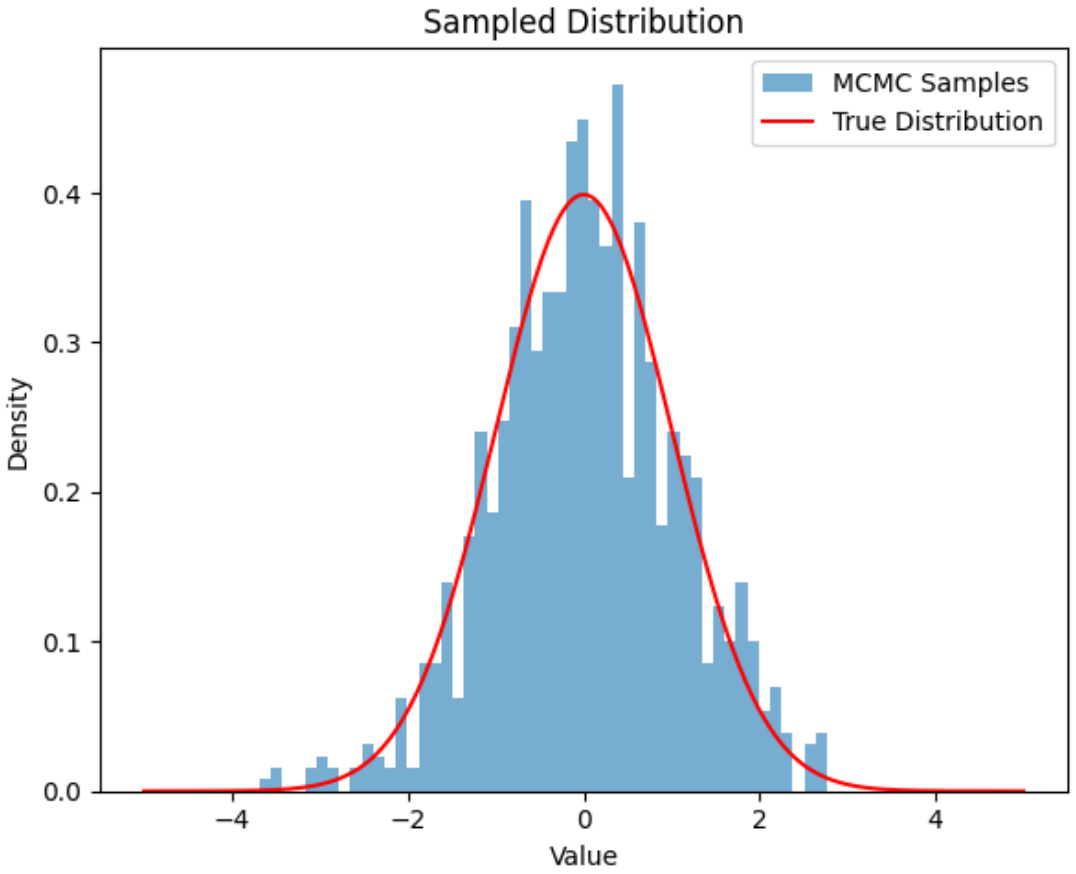
$$\theta^{(1)} = \begin{cases} \theta' \text{ with probability } \alpha \\ \theta^{(0)} \text{ with probability } 1-\alpha \end{cases} \qquad \text{where } \alpha = \min\left(1, \frac{\pi(\theta')}{\pi(\theta^{(0)})}\right)$$

MCMC Solvers:

- BUGS – Bayes Using Gibbs Sampling
- JAGS – Just Another Gibbs Sampler
- Stan – uses Hamiltonian Monte Carlo
- NUTS – No U-Turn Sampler

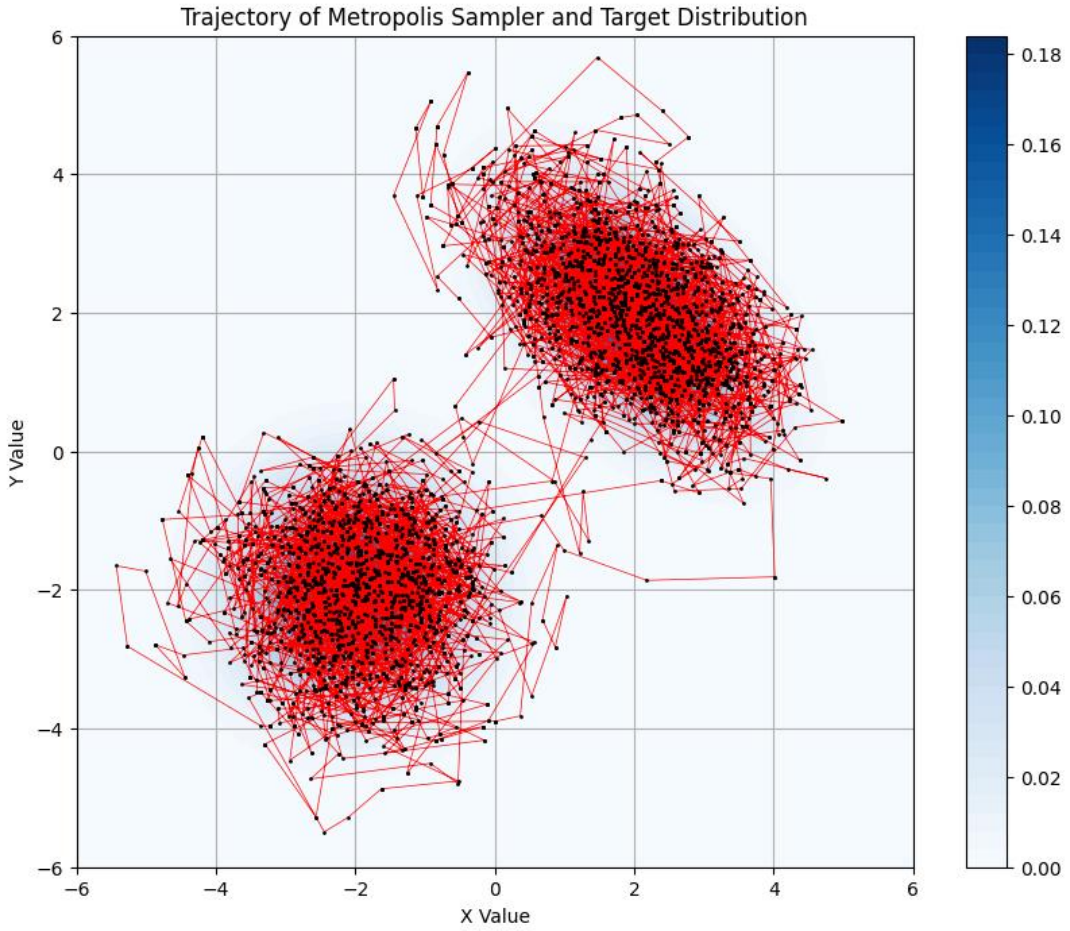
Sampling 1D Gaussian

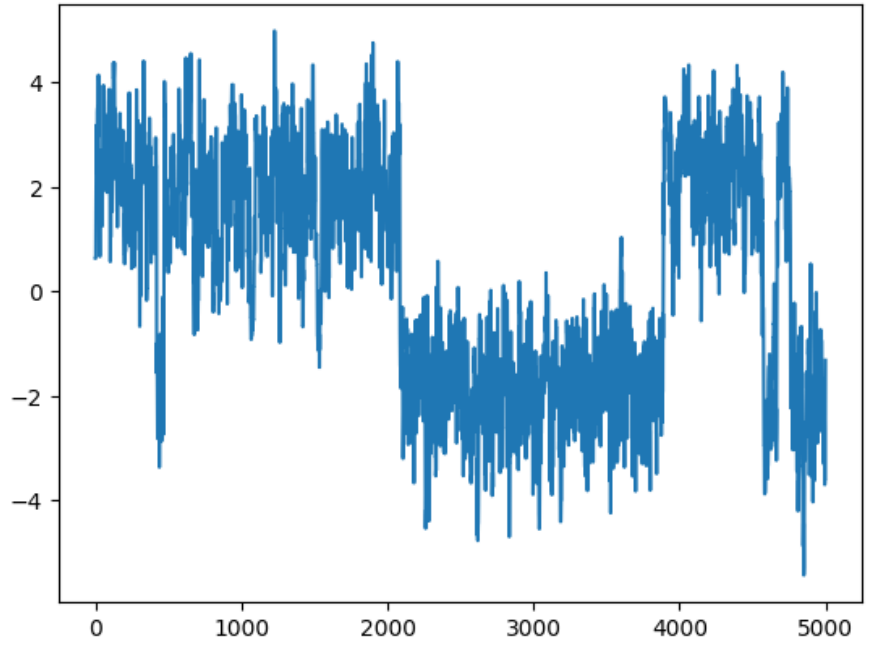




We can calculate any function of the posterior by summing over the trace

Sampling 2D Gaussian





MCMC Solvers:

- BUGS – Bayes Using Gibbs Sampling
- JAGS – Just Another Gibbs Sampler
- Stan – uses Hamiltonian Monte Carlo
- NUTS – No U-Turn Sampler