Characterising a posterior distribution

Model Fitting and Inference for Infectious Disease Dynamics short course

Recap

Last session we saw that the posterior distribution of θ, given observed data D, is

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Constant}}$$

Our aim is to characterize the posterior.

Recap

Last session we saw that the posterior distribution of θ, given observed data D, is

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

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

Our aim is to characterize the posterior.

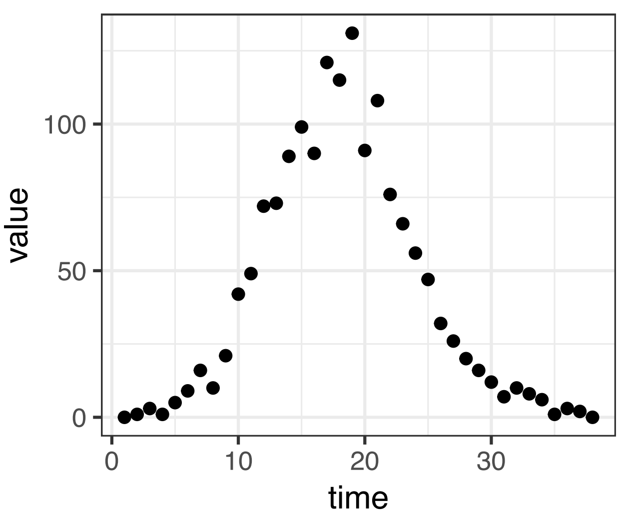
Recap

$p(\theta|D) \propto p(D|\theta)p(\theta)$ Posterior $\propto$ Likelihood $\times$ Prior

The posterior is a probability distribution that tells us what parameter values are credible given the data we have observed and our pre-existing (prior) beliefs about the parameters.

This allows us to answer questions like: given some case data and a model, plus some (potentially vague) prior beliefs, which values of R_0 are plausible?

Data and model

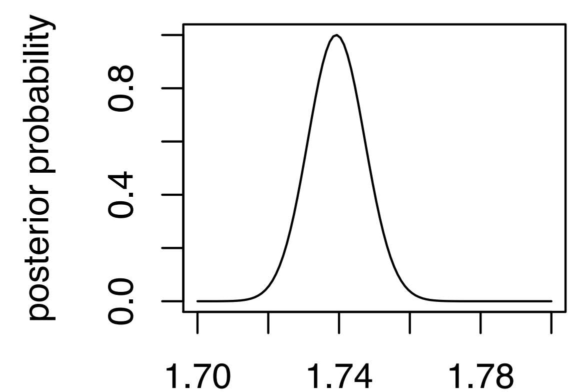


$$\begin{cases} \frac{\mathrm{d}S}{\mathrm{d}t} = -\beta S \frac{I}{N} \\ \frac{\mathrm{d}I}{\mathrm{d}t} = \beta S \frac{I}{N} - \nu I \\ \frac{\mathrm{d}R}{\mathrm{d}t} = \nu I \end{cases}$$

$R_0 = \beta/\nu$

+ Prior beliefs

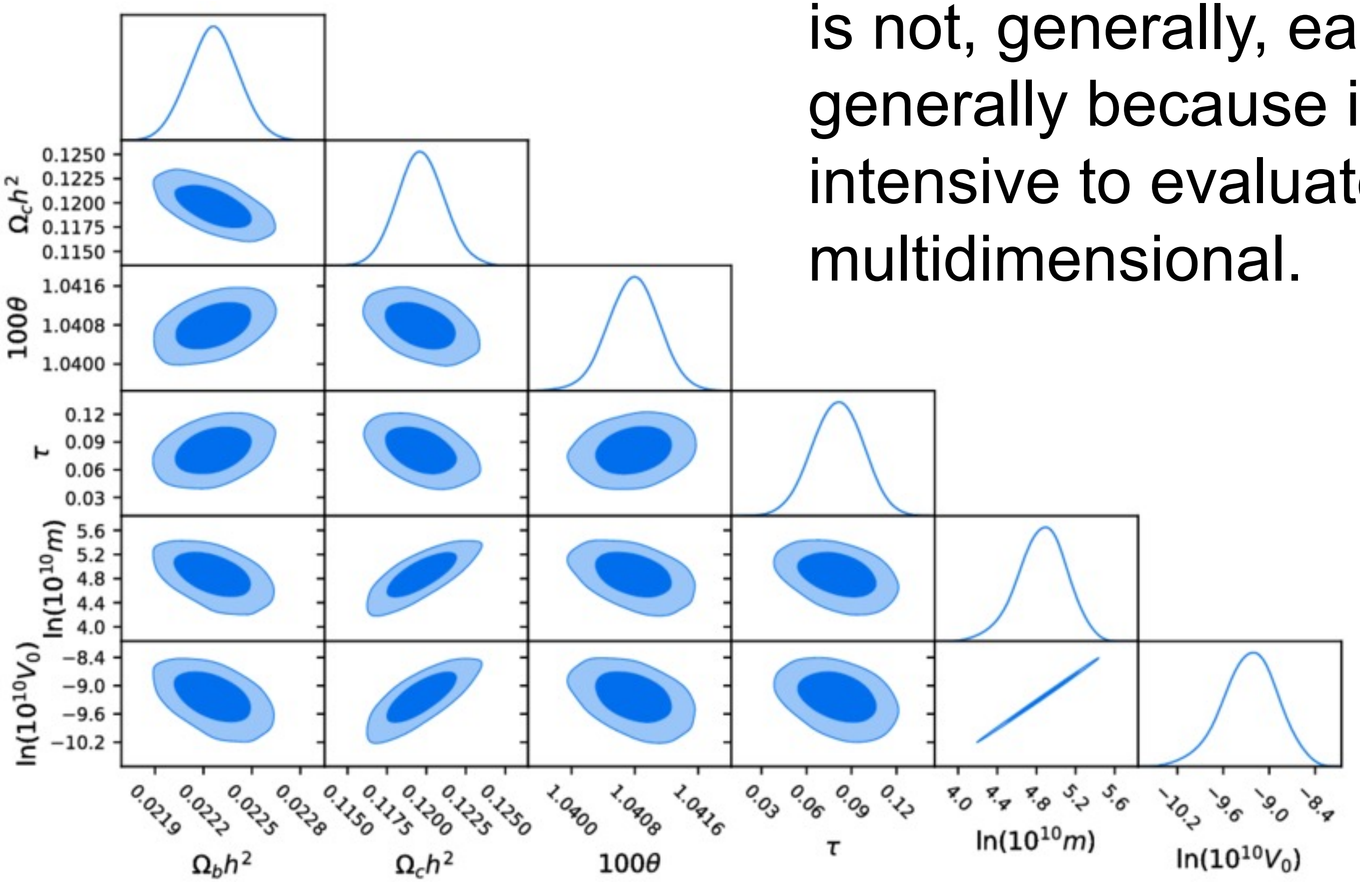
$\boldsymbol{R_0 \sim U(1,2)}$



Posterior for R_0

The problem

The posterior, $p(\theta|D) \propto p(D|\theta)p(\theta)$ is not, generally, easy to "solve" for, generally because it is complicated, intensive to evaluate, and multidimensional.



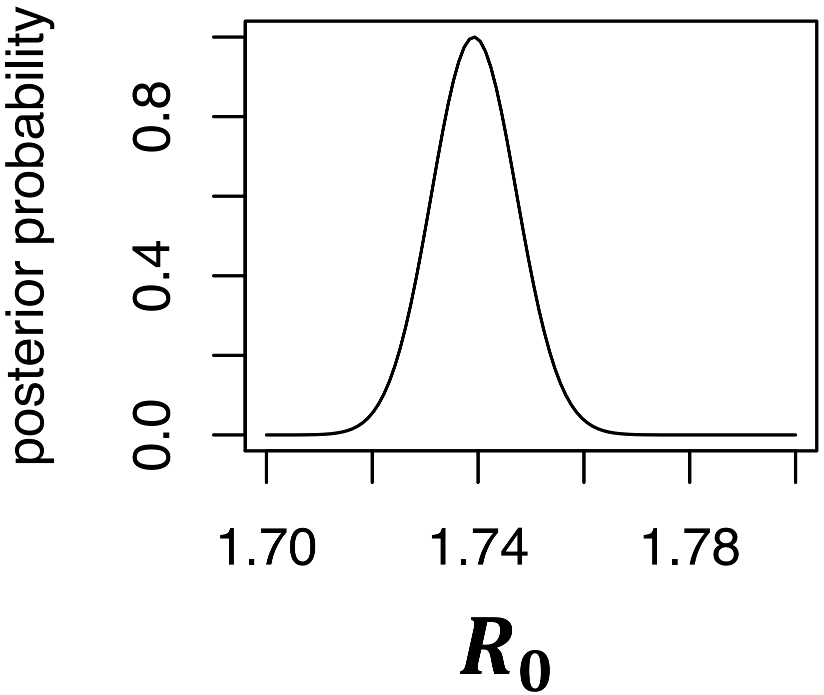
Iqbal et al. (2018) *Astrophysics and Space Science* 10.1007/s10509-018-3446-3.

The problem

The posterior, $p(\theta|D) \propto p(D|\theta)p(\theta)$ is not, generally, easy to "solve" for, generally because it is complicated, intensive to evaluate, and multidimensional.

So how do we characterize the posterior, *i.e.*:

- find the mean, median, mode of R_0?
- visualize R_0 in plots?
- give "credible intervals" for R_0?
- use fitted R_0 to make predictions?

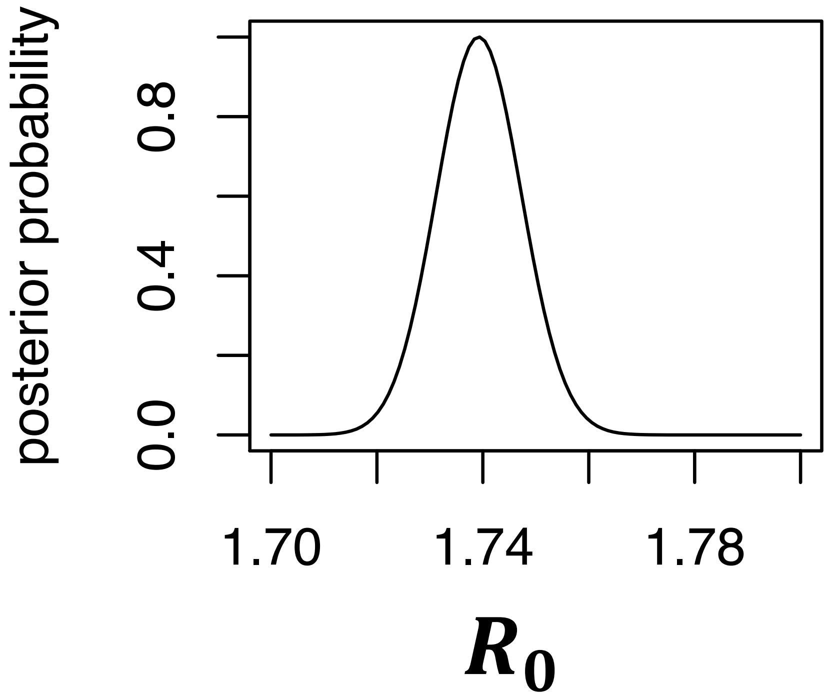


Methods suitable in low dimensions

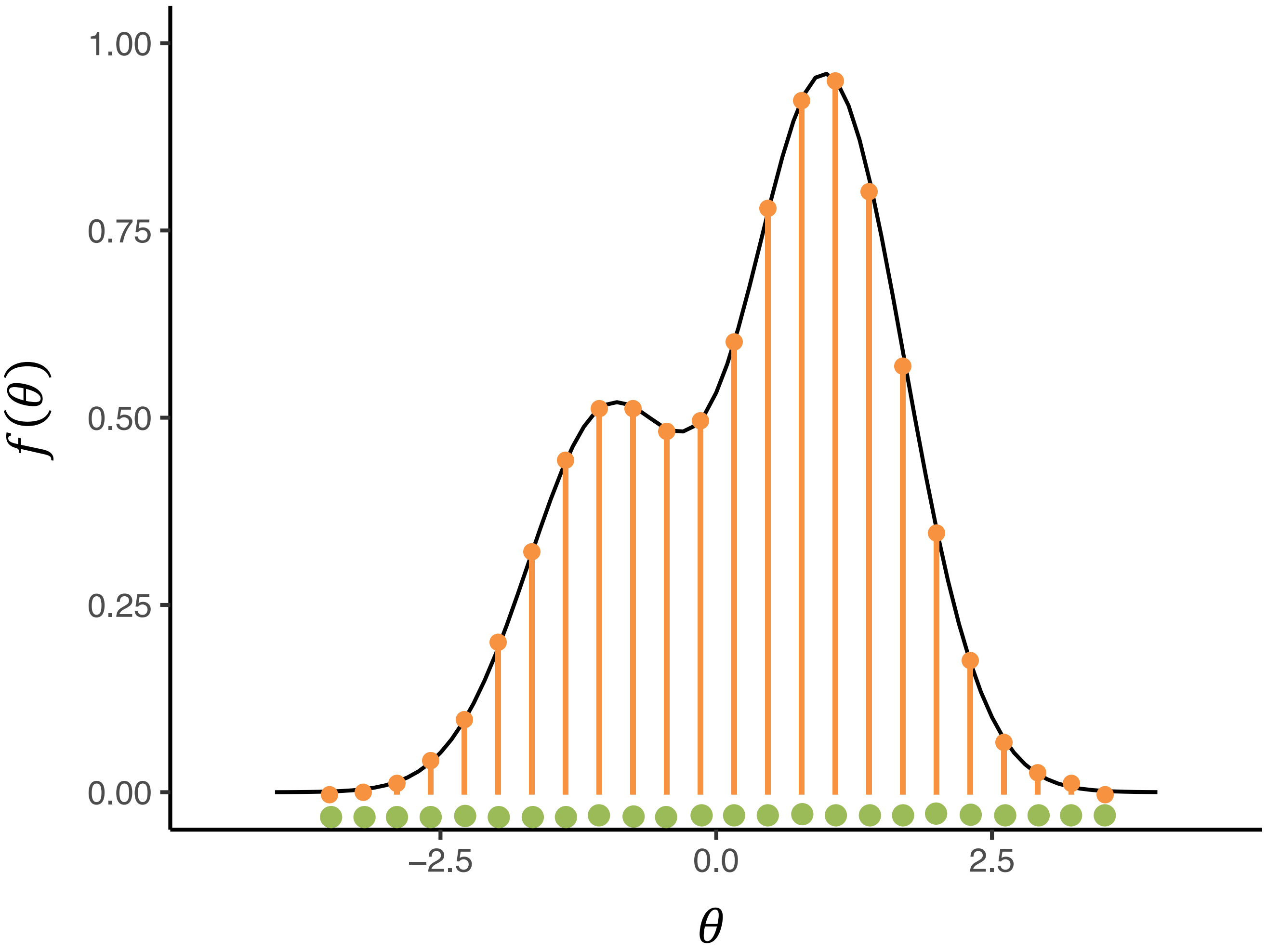
When the posterior $p(\theta|D)$ has relatively few dimensions (i.e. $\theta \in \mathbb{R}^d$ with $d = 1$ or 2) there are “simpler” methods than MCMC that may give equally good results.

We will start by exploring two such methods:

- grid approximation
- rejection sampling

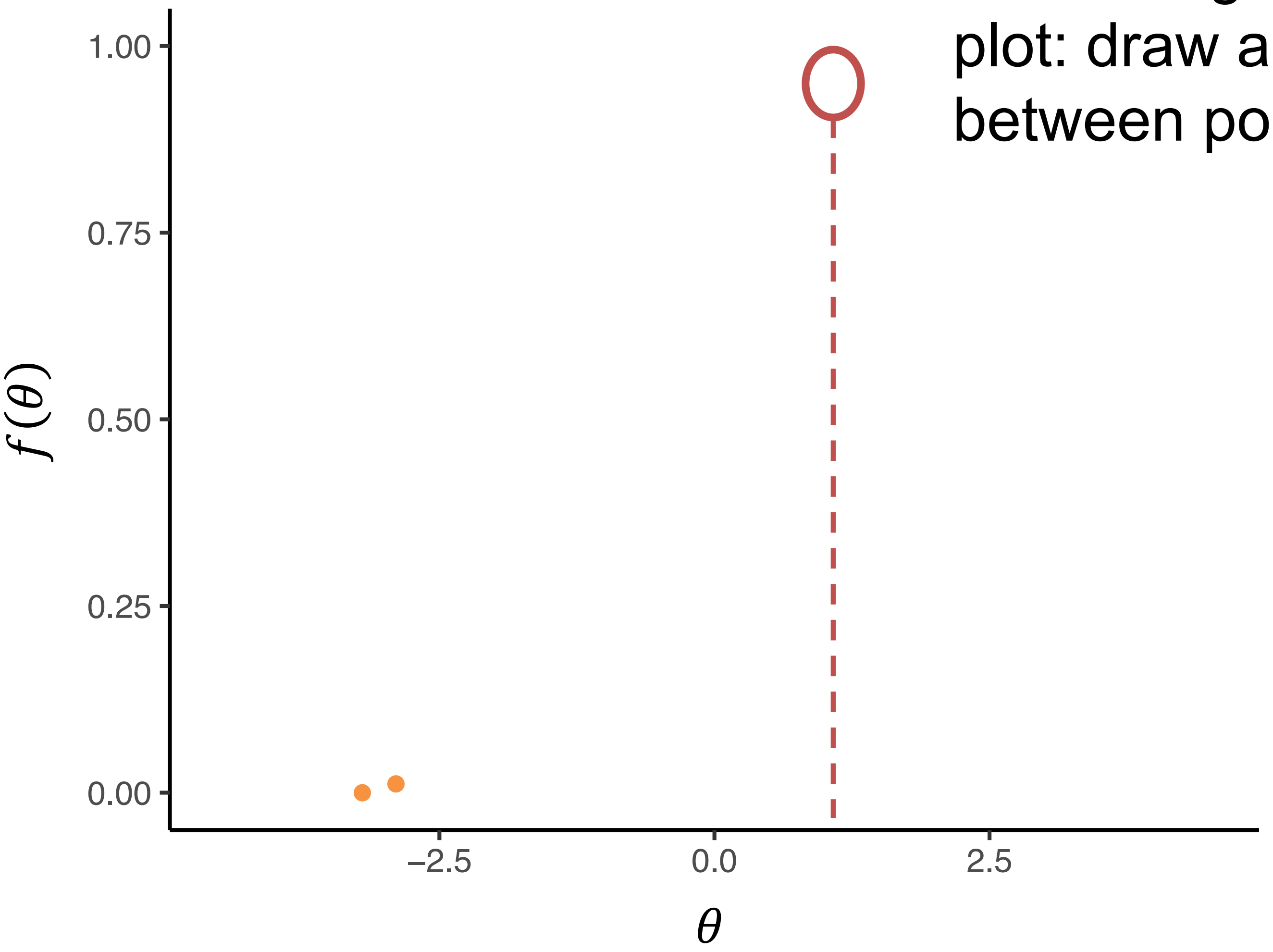


Method 1: Grid approximation

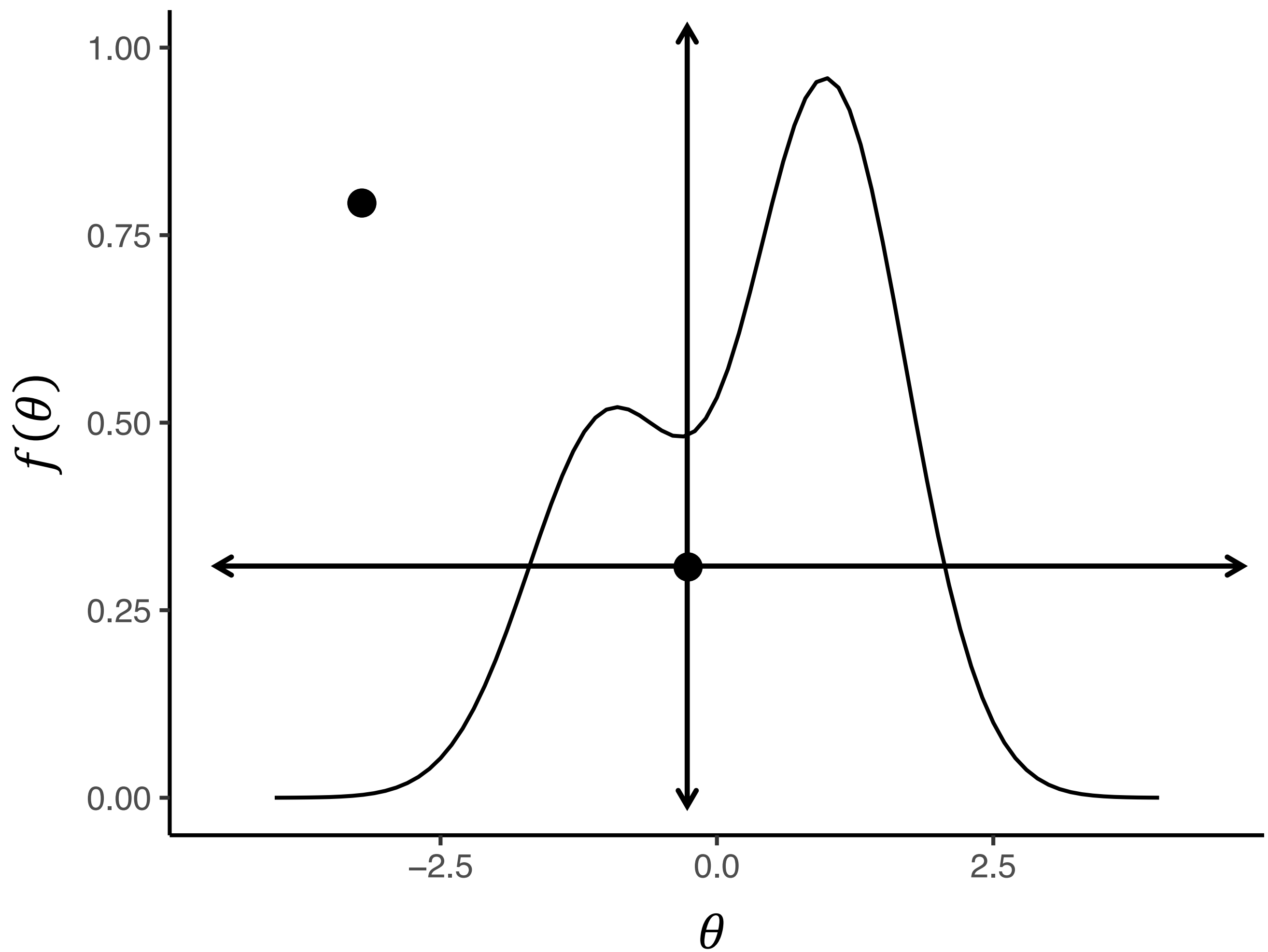


Method 1: Grid approximation

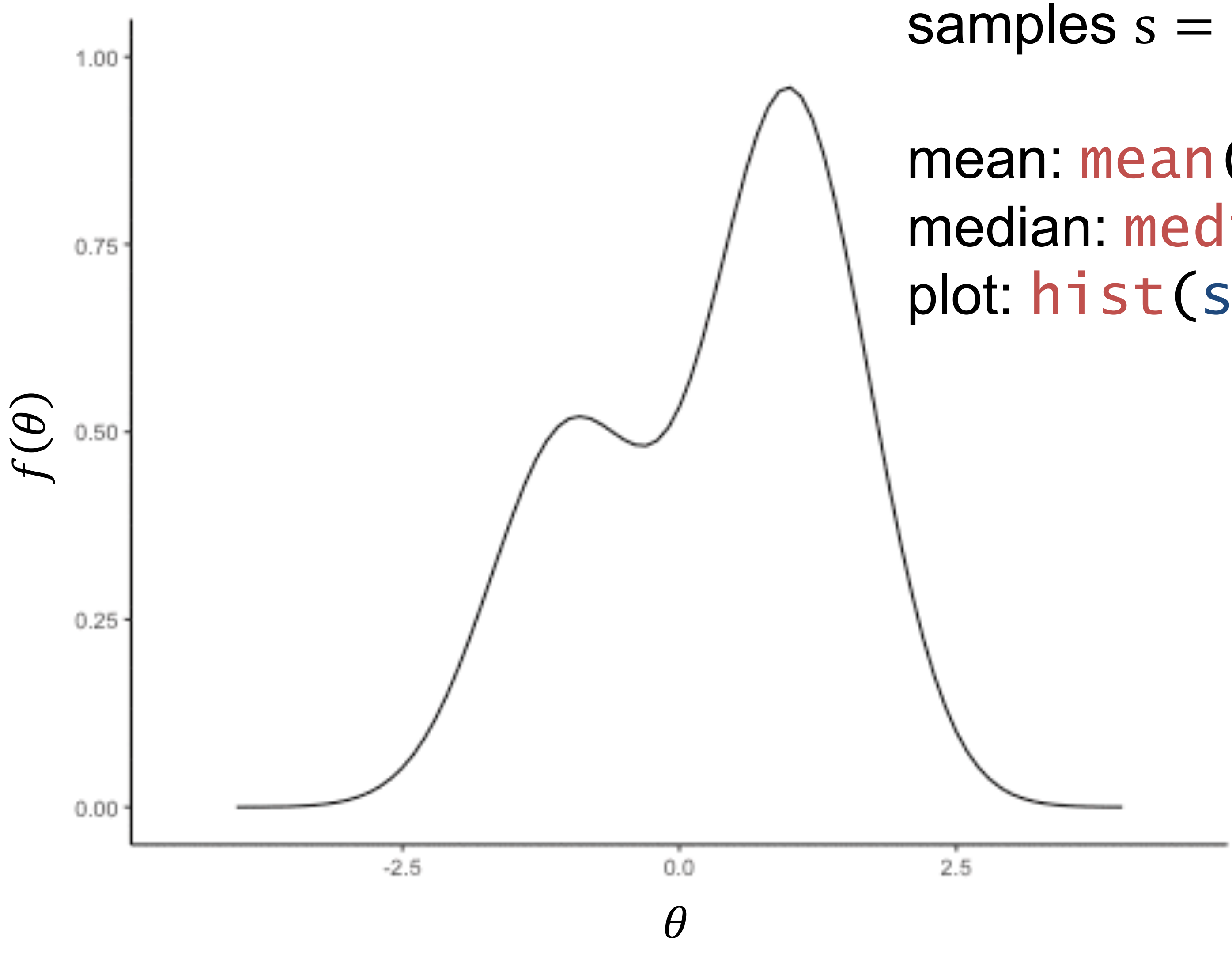
mode: “highest” point
plot: draw a line
between points



Method 2: Rejection sampling



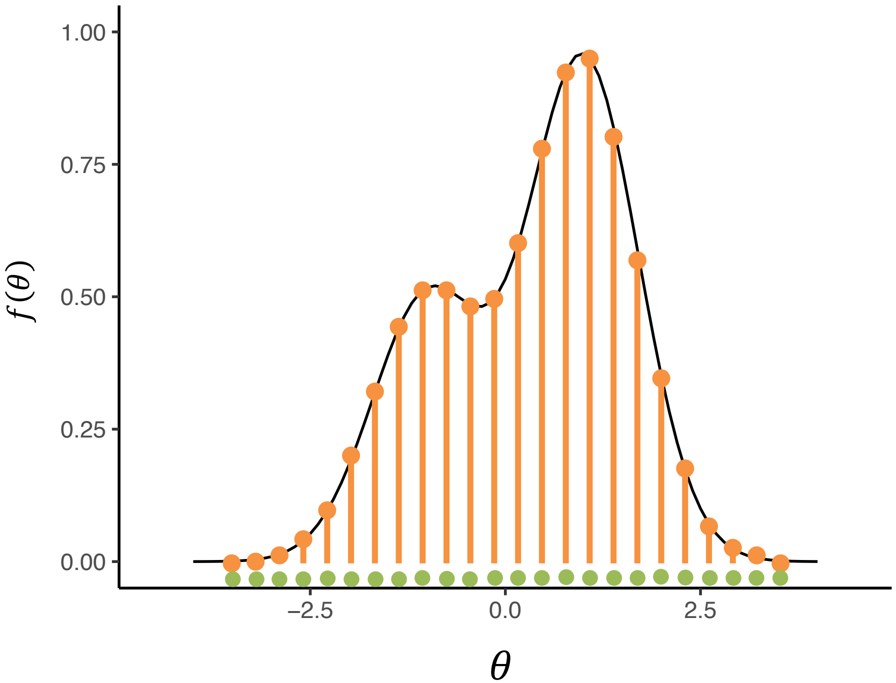
Method 2: Rejection sampling

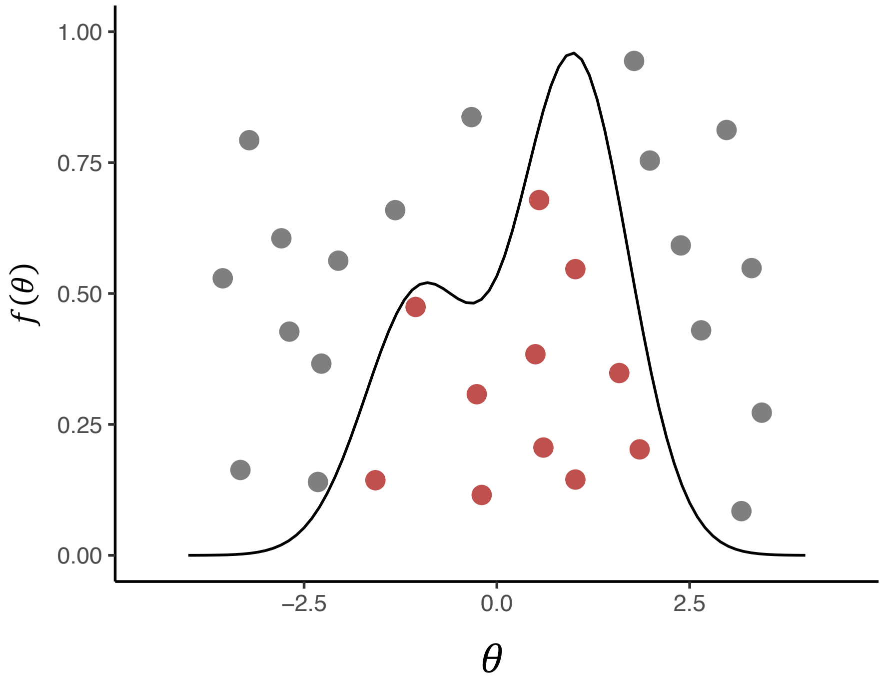


samples $s = \{\theta_1, \theta_2, \ldots\}$

mean: `mean(s)`
median: `median(s)`
plot: `hist(s)`

Practical, part 1



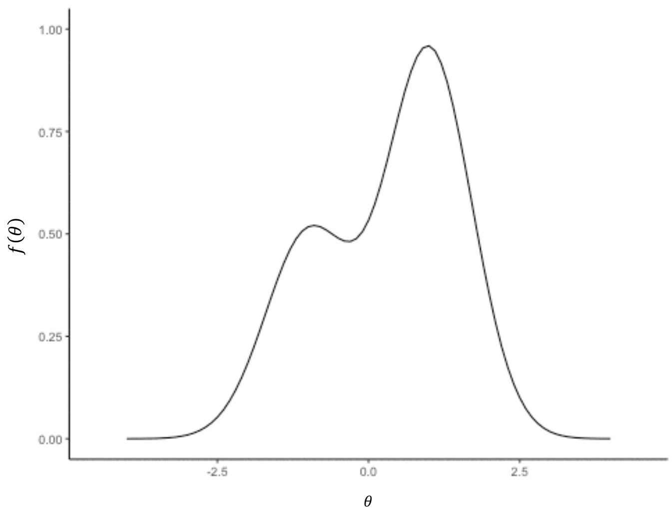


Start the practical:
"Grid approximation" and "Rejection sampling"
in the MCMC session

Practical 1

1. How much do the summary statistics change if you perform the sampling again?
2. If you decrease the number of attempts from 1000 to 100, would you expect the summary statistics to change more each time sampling is performed or less? What does this tell you about reliable sampling?
3. What are the advantages and disadvantages of grid approximation versus rejection sampling?

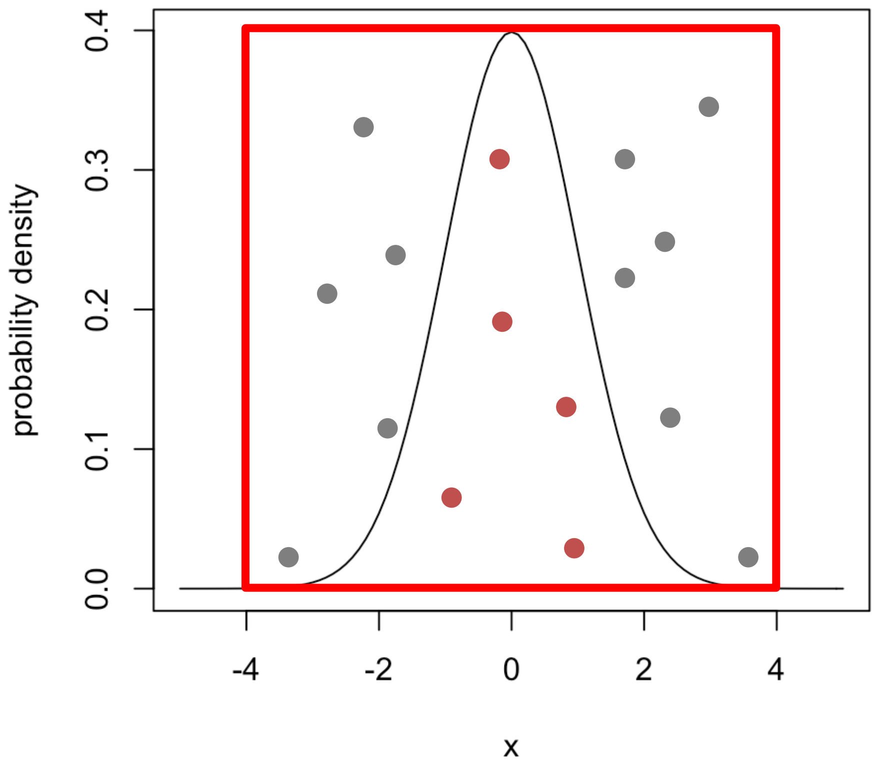
Issues with grid / rejection methods



Need to specify the limits of the distribution

As dimensions increase:
Curse of dimensionality

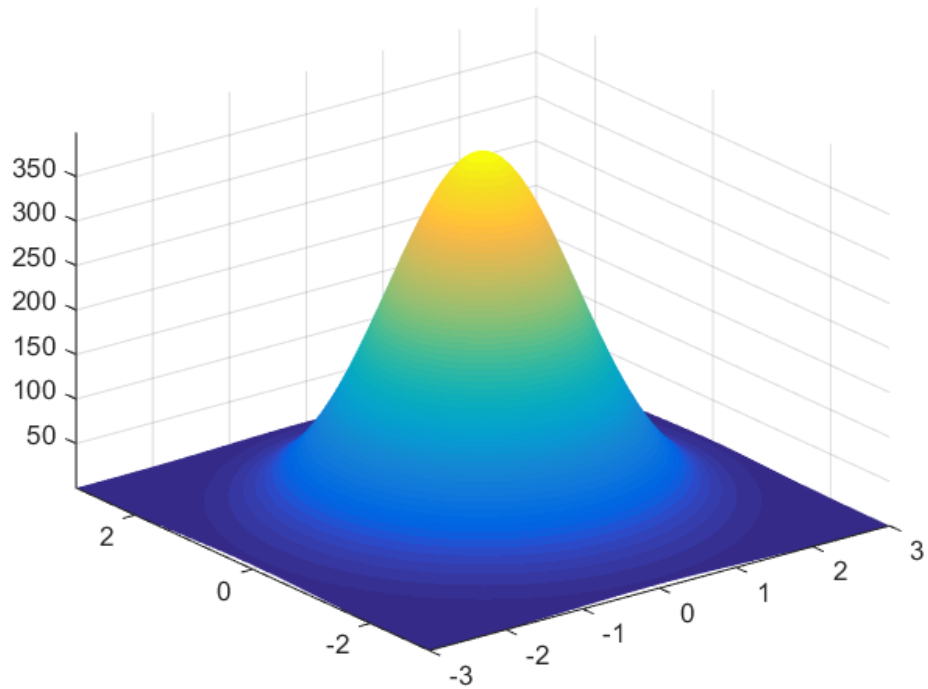
Curse of dimensionality



Throwing darts at the standard normal distribution between -4*SD and 4*SD:

Fraction 0.31 will “hit”

Curse of dimensionality



2D normal distribution:
$0.31^2 = 0.098$

10D normal distribution:
$0.31^{10} = 0.0000091$

(to get 1000 samples, you need to throw 110 million darts... and that's when limits are known)

What can we do with posterior samples?

We've been focusing on ways of *reporting* parameter estimates...

samples $s = \{\theta_1, \theta_2, \dots\}$

```
mean(s) # if s is vector of R0, this gives mean R0
```

But we can use them to make predictions, test interventions, etc

```
for theta in s {
    run model with R0 = theta and 0% vaccination
    run model with R0 = theta and 50% vaccination
    record results
}
```

Markov Chain Monte Carlo

Model Fitting and Inference for Infectious Disease Dynamics short course

Markov Chain Monte Carlo (MCMC)

Markov chain: stochastic sequence of states in which the next state depends only upon the current state

$$\theta_{t+1} \sim \mathbb{D}(\theta_t)$$

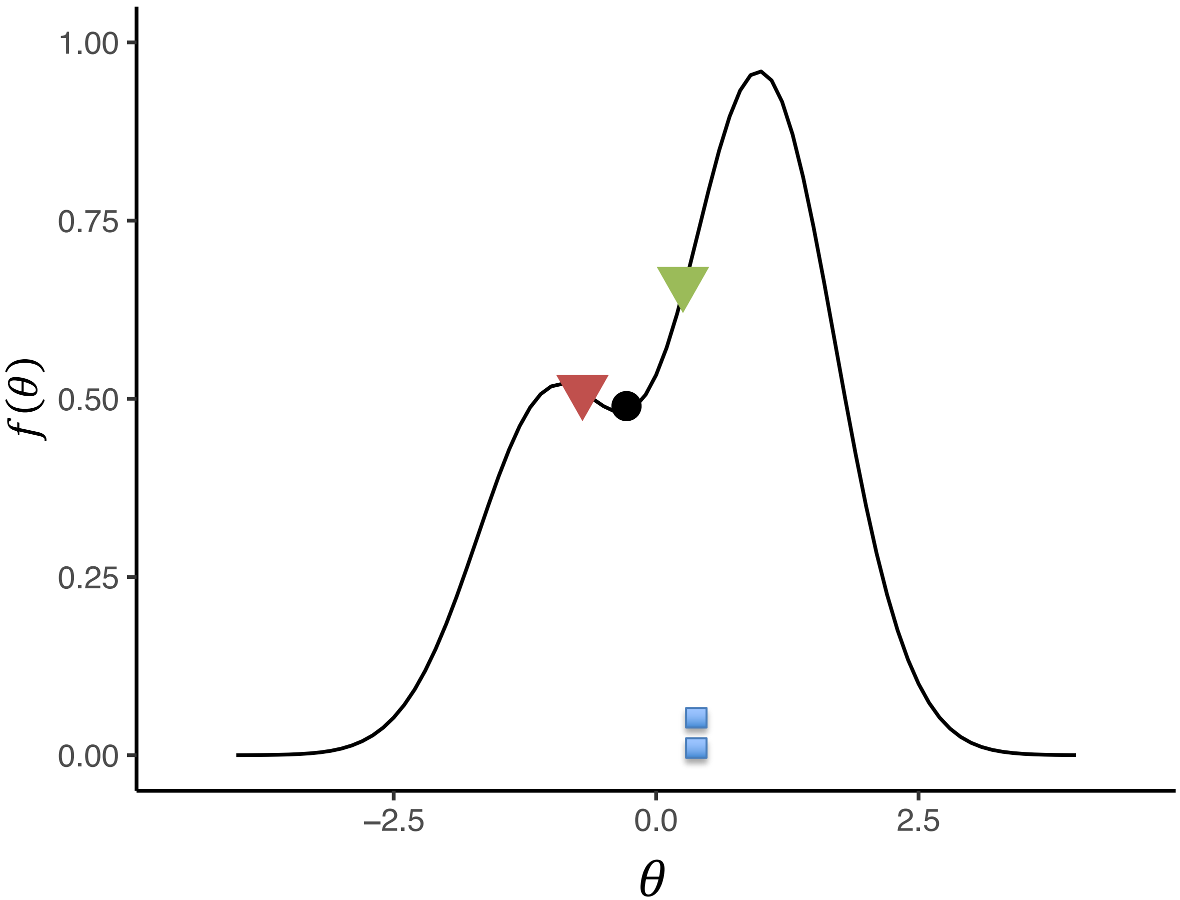
Monte Carlo: a famous casino. Also a class of algorithms in which random sampling is used to solve problems.

Metropolis-Hastings algorithm: a particular way of using MCMC to sample from a distribution

MCMC: Outline

- What the algorithm is
- Practical: Implementing Metropolis-Hastings MCMC
- Why it works.

MCMC algorithm



1. PROPOSE
2. MOVE OR STAY ("acceptance")
3. SAVE LOCATION

Choose a starting point, $\theta = \theta_0$

PROPOSE

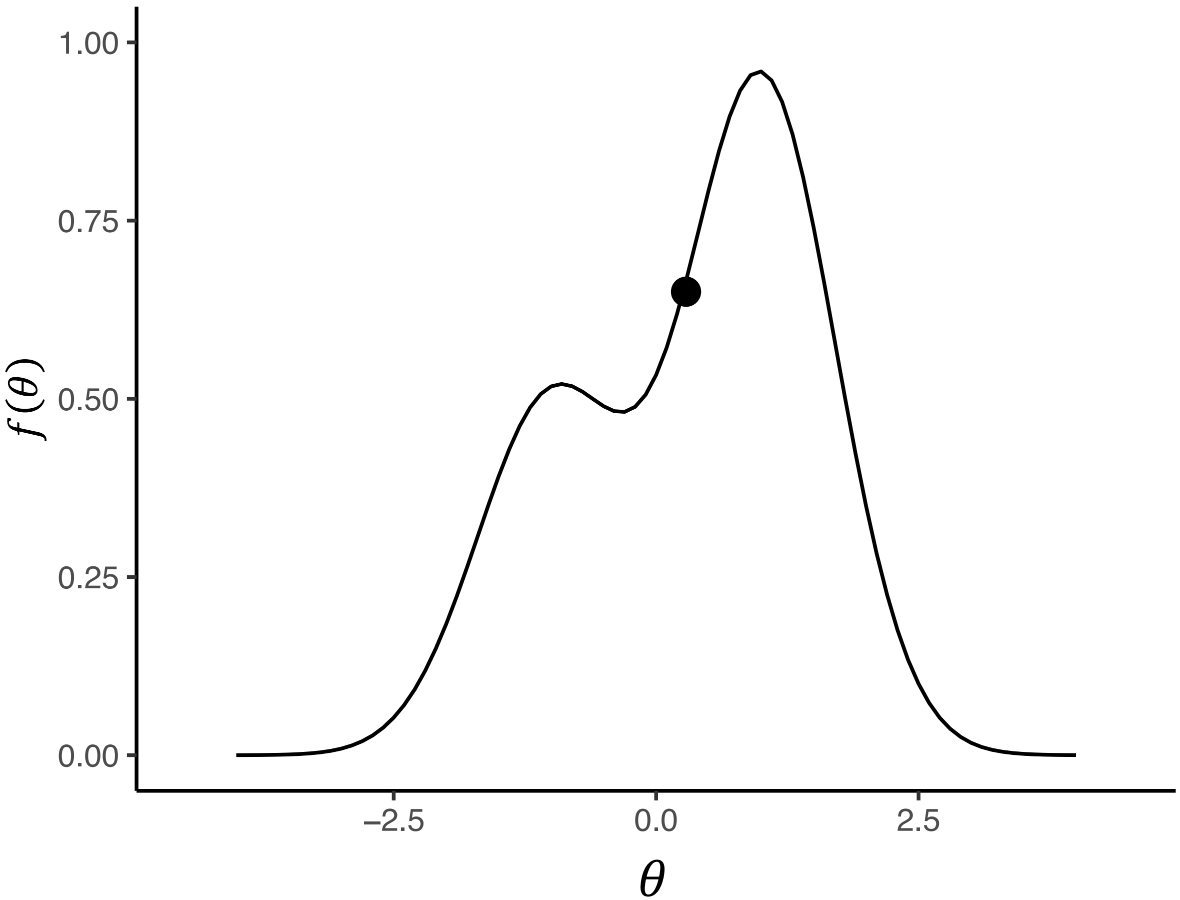
$$\theta' = \theta + \varepsilon \mid \varepsilon \sim Q$$

MOVE OR STAY

If $f(\theta') > f(\theta)$, definitely move.

If $f(\theta') < f(\theta)$, move with probability $f(\theta')/f(\theta)$, otherwise stay.

MCMC algorithm



1. PROPOSE
2. MOVE OR STAY ("acceptance")
3. SAVE LOCATION

Choose a starting point, $\theta = \theta_0$

PROPOSE

$$\theta' = \theta + \varepsilon \mid \varepsilon \sim Q$$

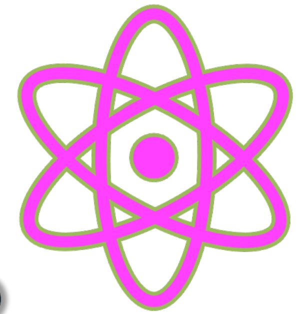
MOVE OR STAY

$$\theta \rightarrow \begin{cases} \theta' & \Pr(a) & Accept \\ \theta & \Pr(1-a) & Reject \end{cases}$$

$$a = \min\left(1, \frac{f(\theta')}{f(\theta)}\right)$$

SAVE LOCATION

$$\theta_t \leftarrow \theta$$



“Metropolis” acceptance ratio!

Choose a starting point, $\theta = \theta_0$

PROPOSE

$$\theta' = \theta + \varepsilon \mid \varepsilon \sim Q$$

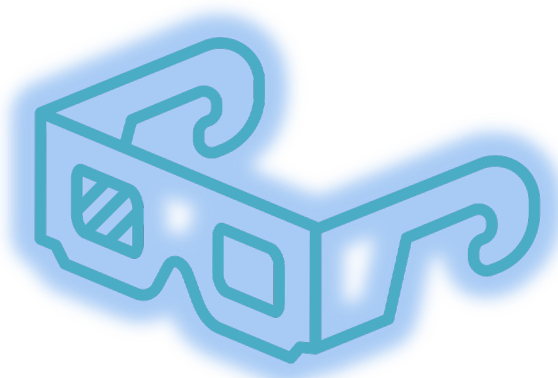
MOVE OR STAY

$$\theta \rightarrow \begin{cases} \theta' & \Pr(a) & Accept \\ \theta & \Pr(1-a) & Reject \end{cases}$$

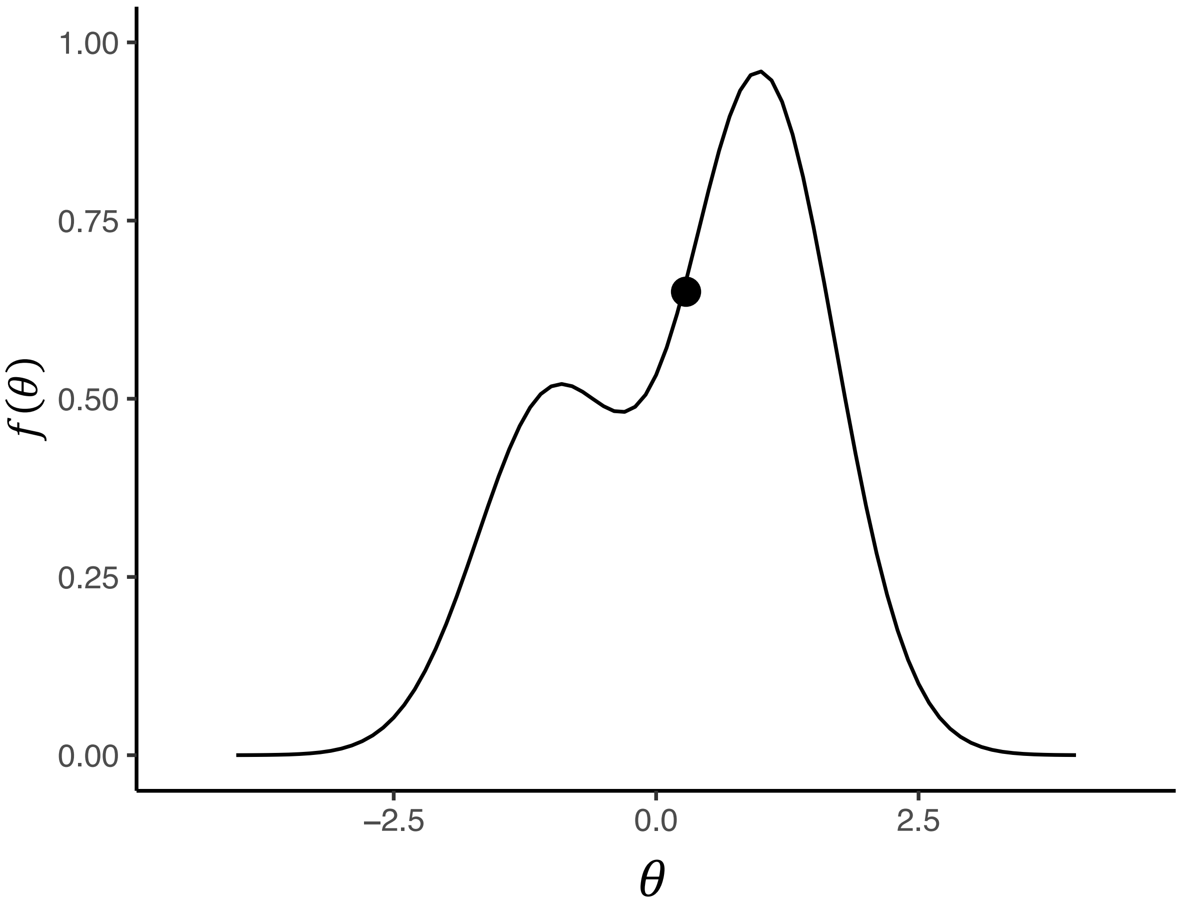
$$a = \min\left(1, \frac{f(\theta')}{f(\theta)}\right)$$

SAVE LOCATION

$$\theta_t \leftarrow \theta$$



MCMC algorithm



1. PROPOSE
2. MOVE OR STAY ("acceptance")
3. SAVE LOCATION

Choose a starting point, $\theta = \theta_0$

PROPOSE

$$\theta' = \theta + \varepsilon \mid \varepsilon \sim Q$$

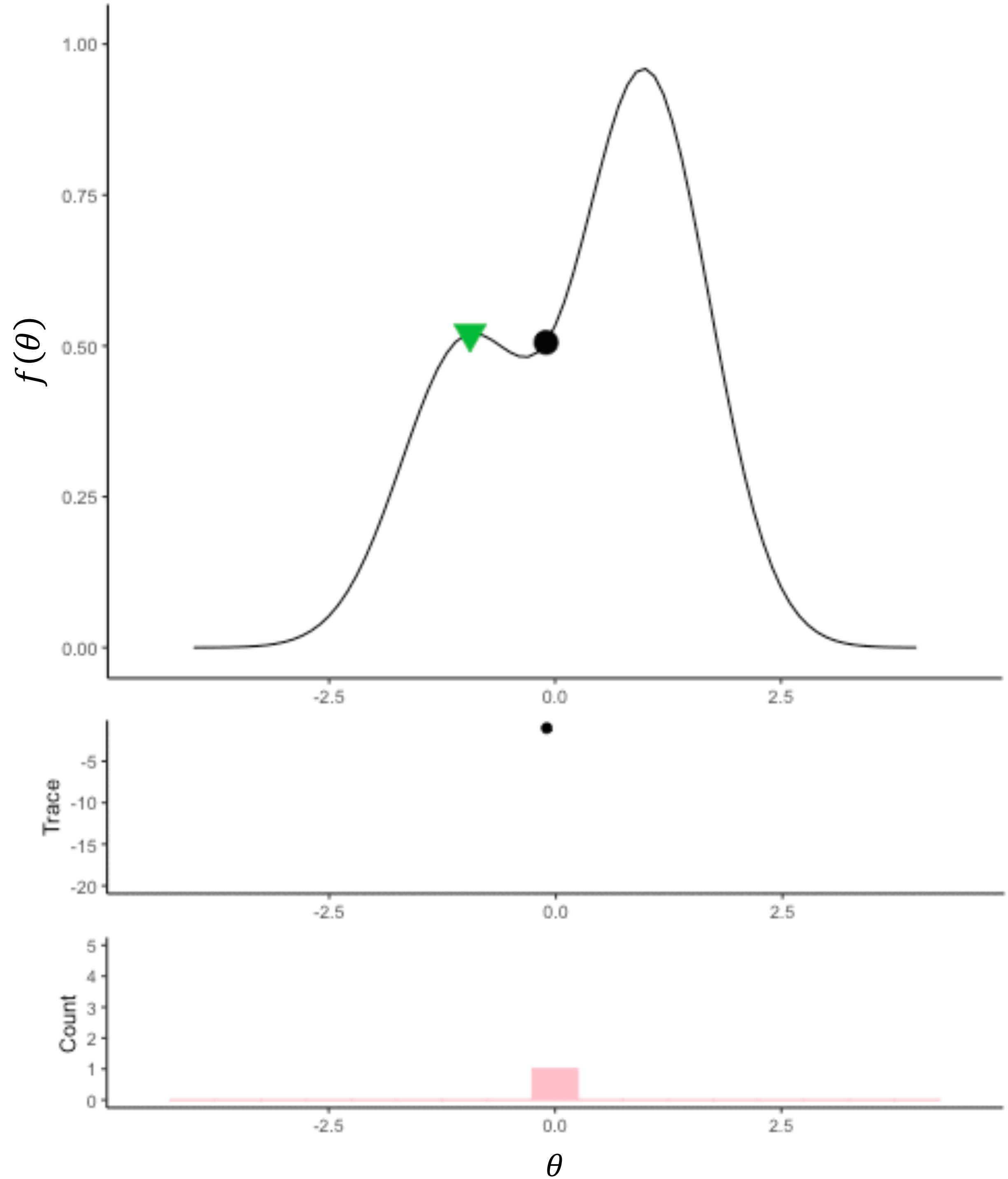
MOVE OR STAY

$$\theta \to \begin{cases} \theta' & \Pr(a) & Accept \\ \theta & \Pr(1-a) & Reject \end{cases}$$

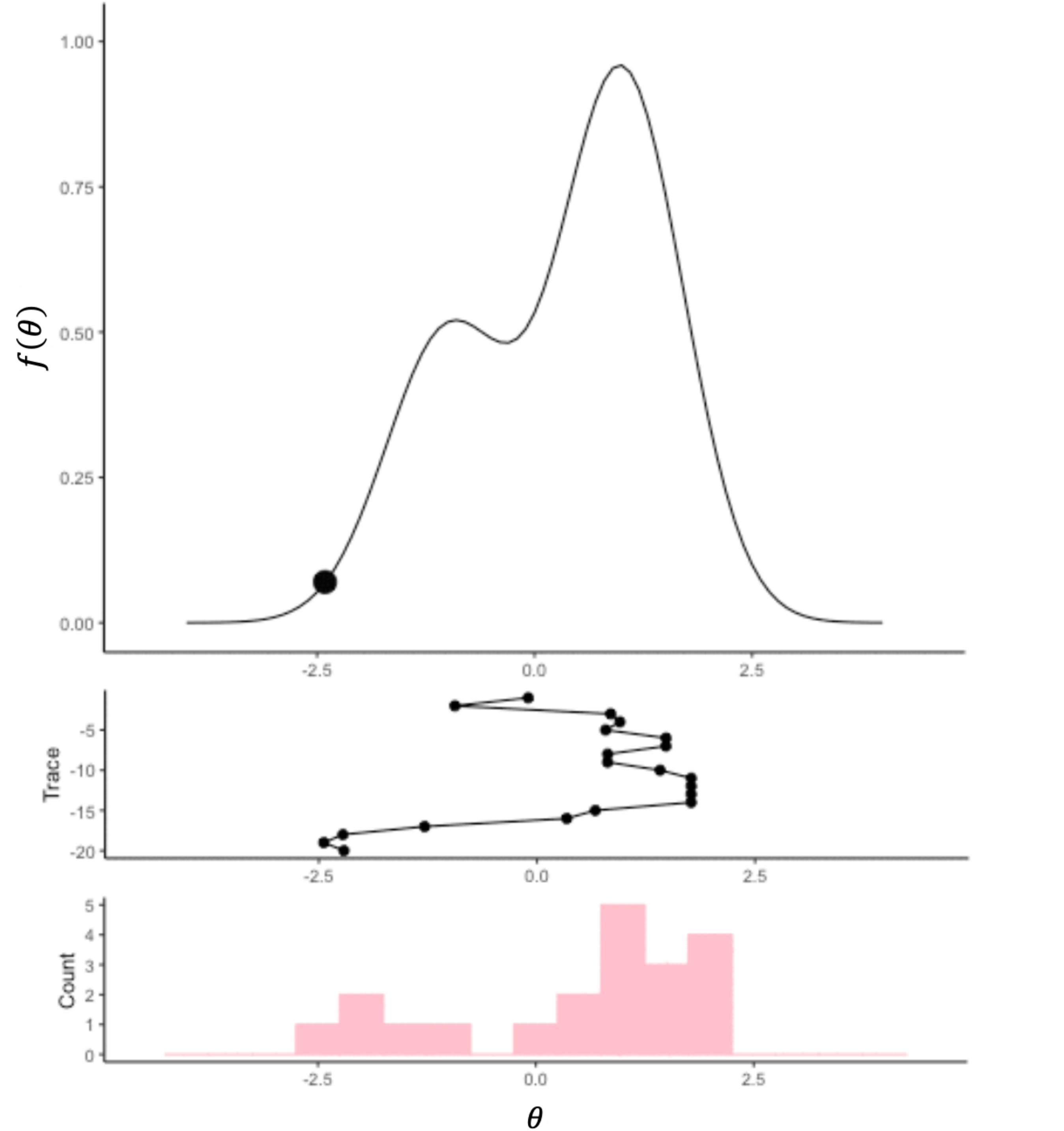
$$a = \min\left(1, \frac{f(\theta')}{f(\theta)}\right)$$

SAVE LOCATION

$$\theta_t \leftarrow \theta$$

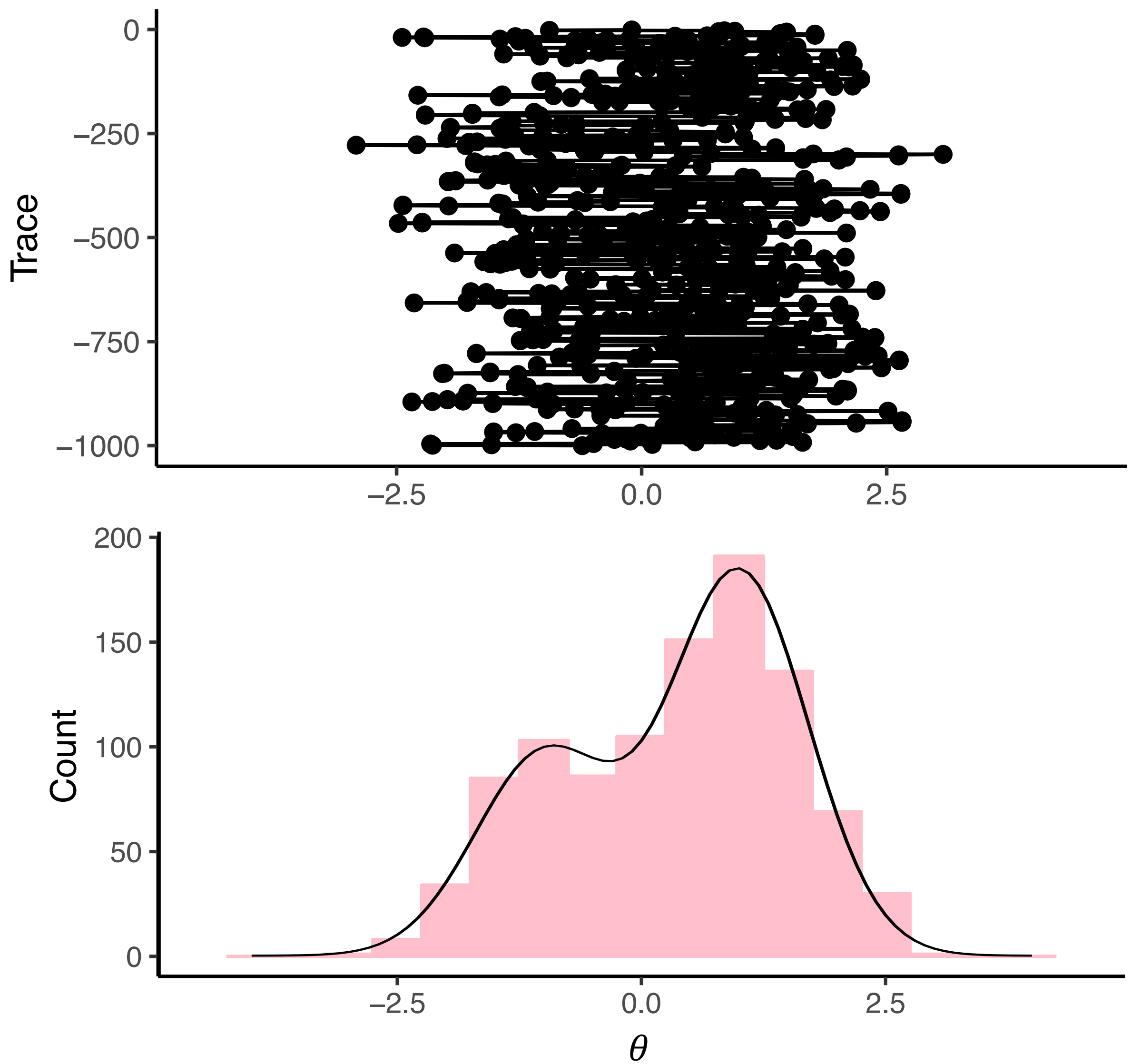


1.00
0.75
0.50
0.25
0.00
$f(\theta)$
-2.5
0.0
2.5
-5
-10
-15
-20
Trace
5
4
3
2
1
0
Count
θ

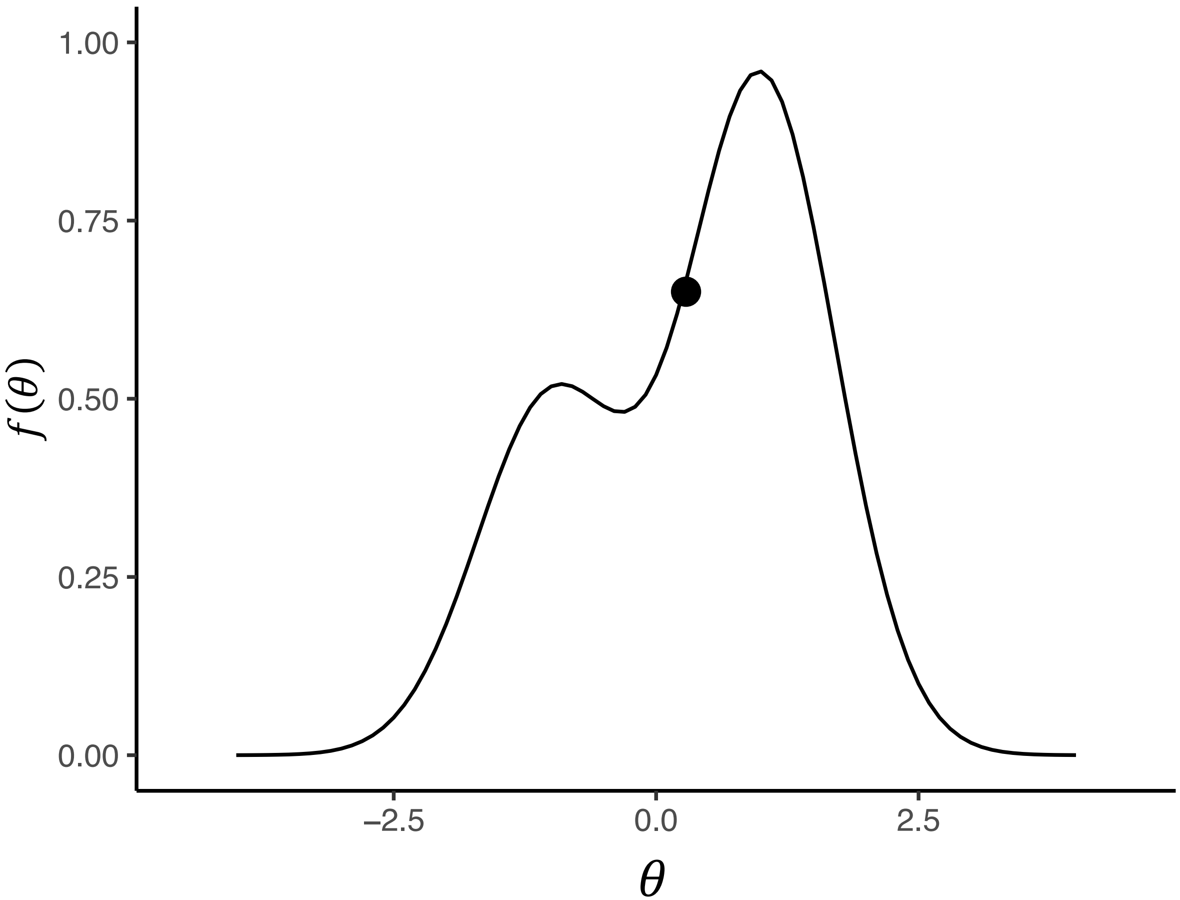


f(θ)
1.00
0.75
0.50
0.25
0.00
-2.5
0.0
2.5
Trace
-5
-10
-15
-20
-2.5
0.0
2.5
Count
5
4
3
2
1
0
-2.5
0.0
2.5
θ

After enough iterations…



Practical, part 2



1. PROPOSE
2. MOVE OR STAY (“acceptance”)
3. SAVE LOCATION

Choose a starting point, $\theta = \theta_0$

PROPOSE

$$\theta' = \theta + \varepsilon \mid \varepsilon \sim Q$$

MOVE OR STAY

$$\theta \to \begin{cases} \theta' & \Pr(a) & \mathit{Accept} \\ \theta & \Pr(1-a) & \mathit{Reject} \end{cases}$$

$$a = \min\left(1, \frac{f(\theta')}{f(\theta)}\right)$$

SAVE LOCATION

$$\theta_t \leftarrow \theta$$

Why does MCMC with Metropolis-Hastings converge to the target distribution?

THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

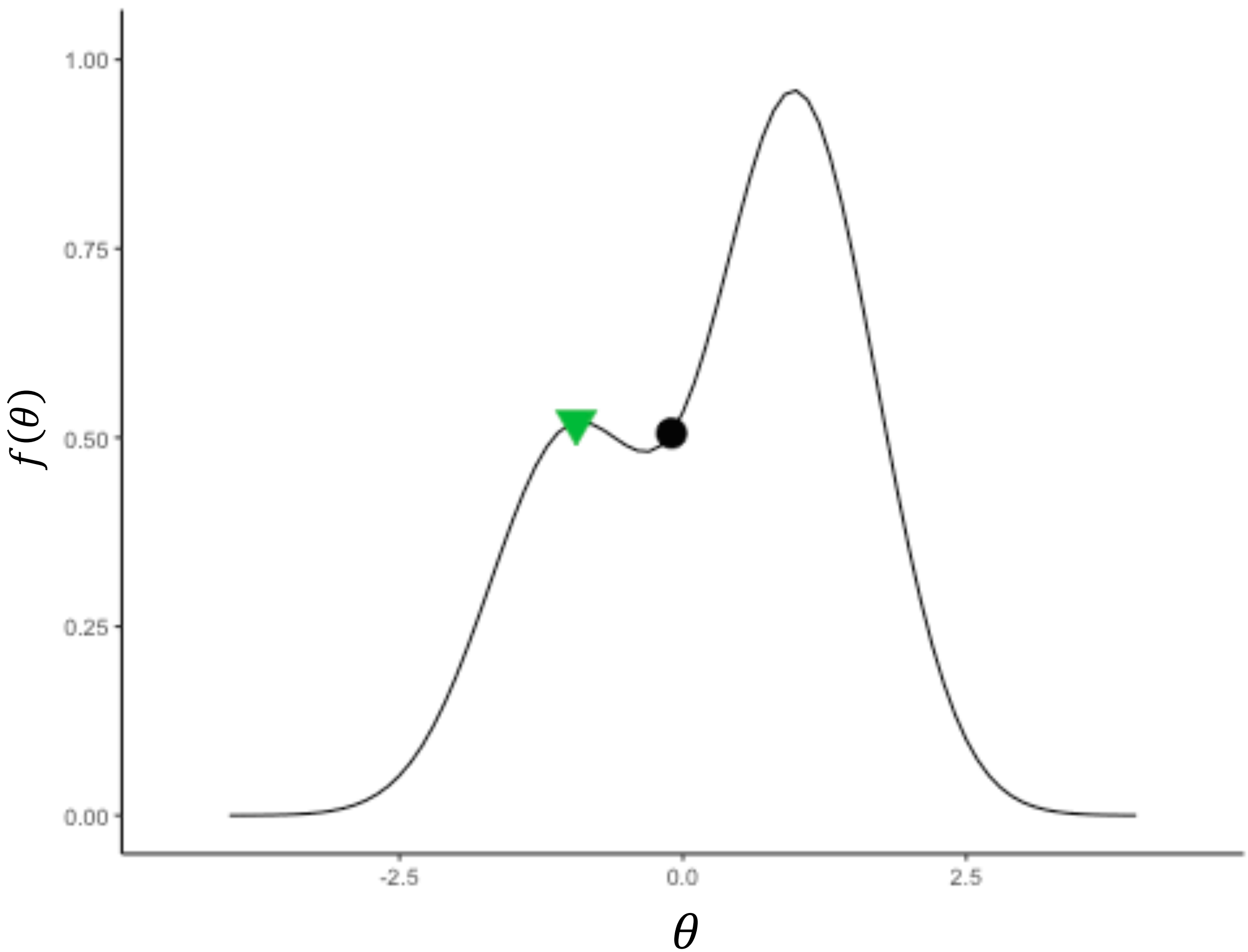
AND

EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

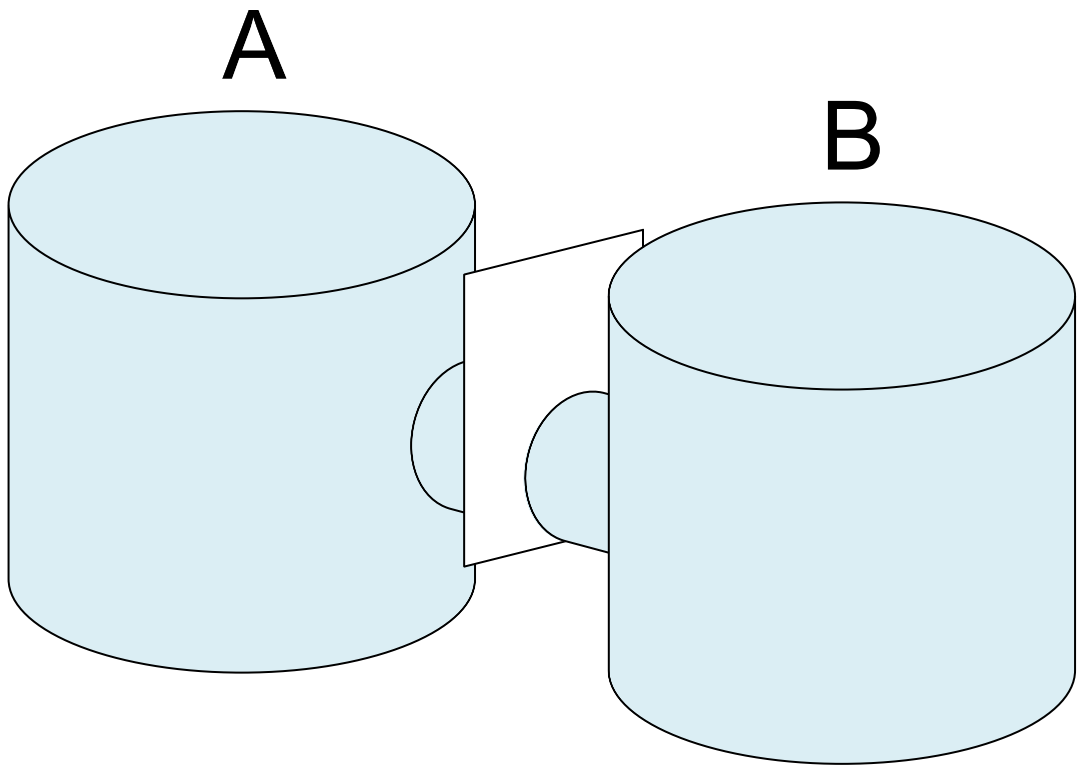
(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

A sloppy hill-climbing algorithm

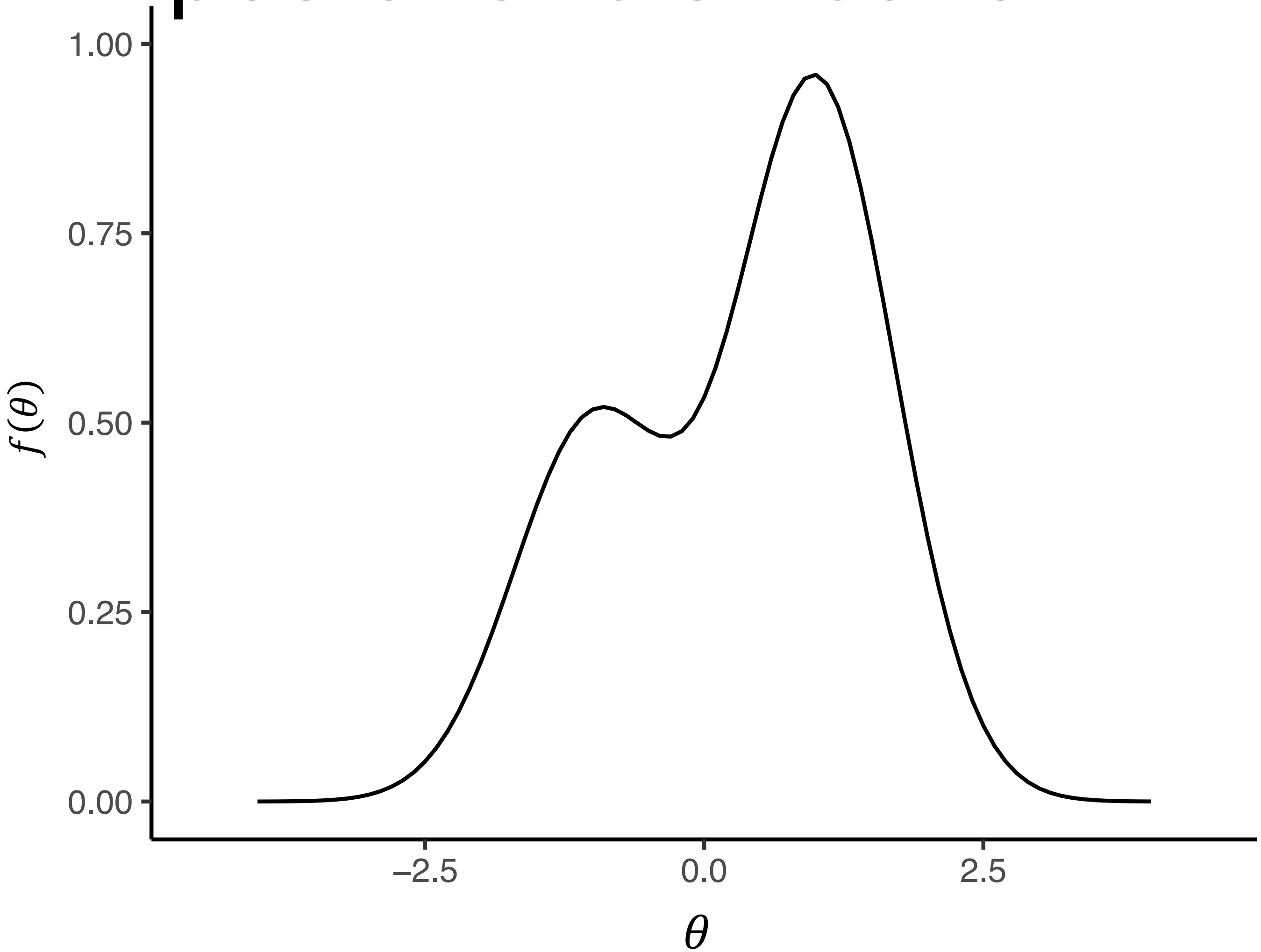


A thought experiment…

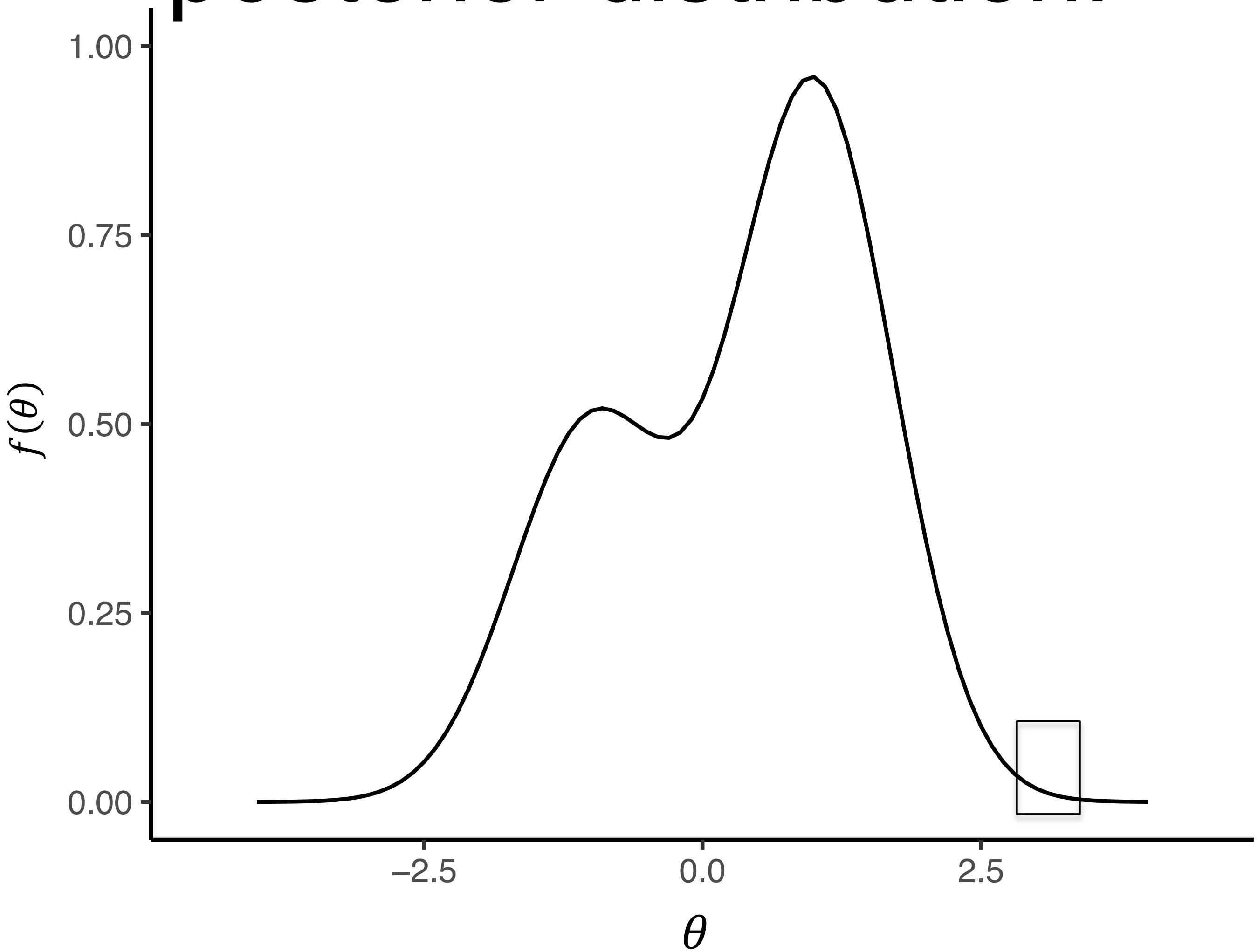


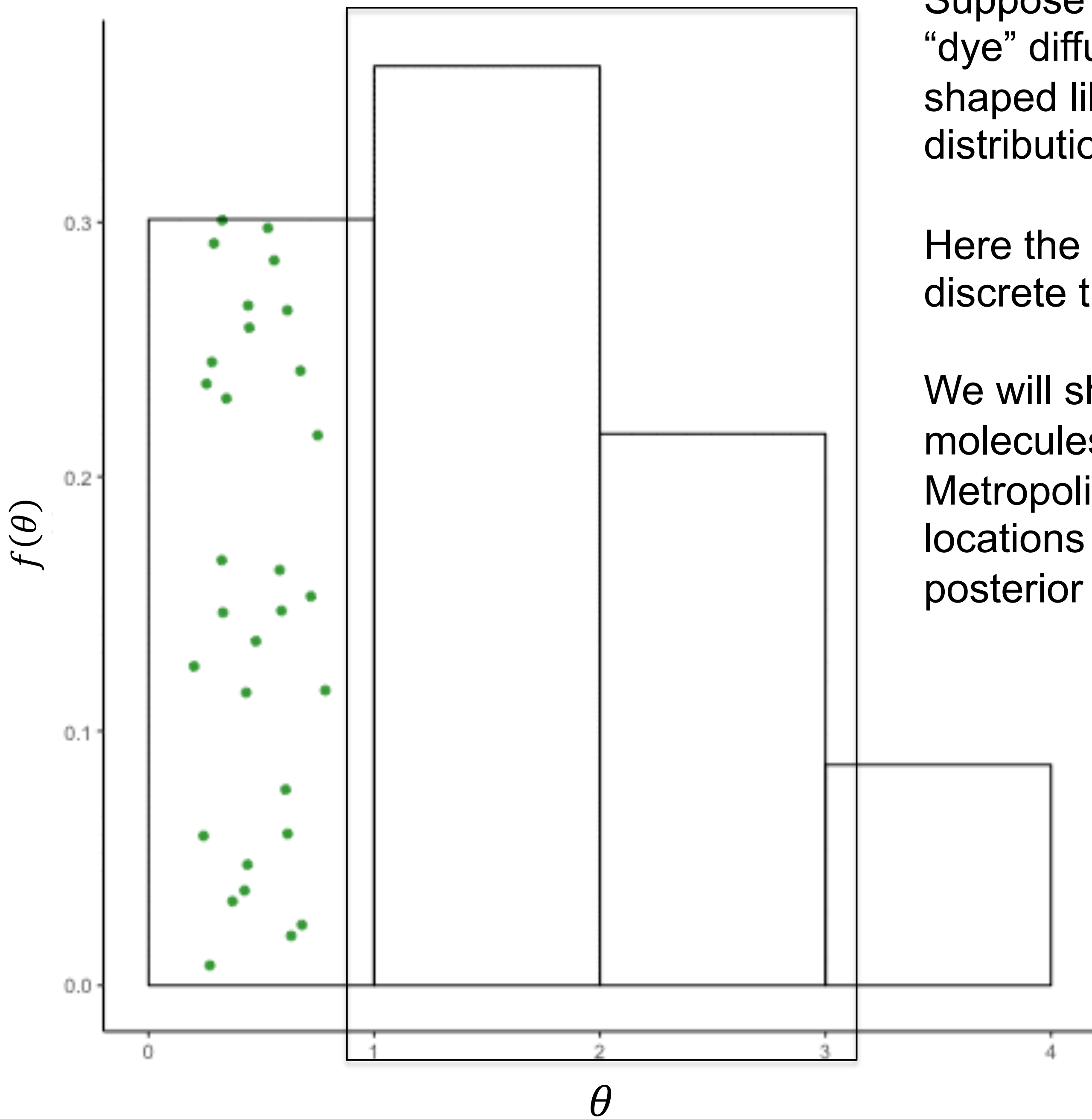
How does the dye “know” to stop flowing?

Let’s zoom in on the posterior distribution:



Let's zoom in on the posterior distribution:

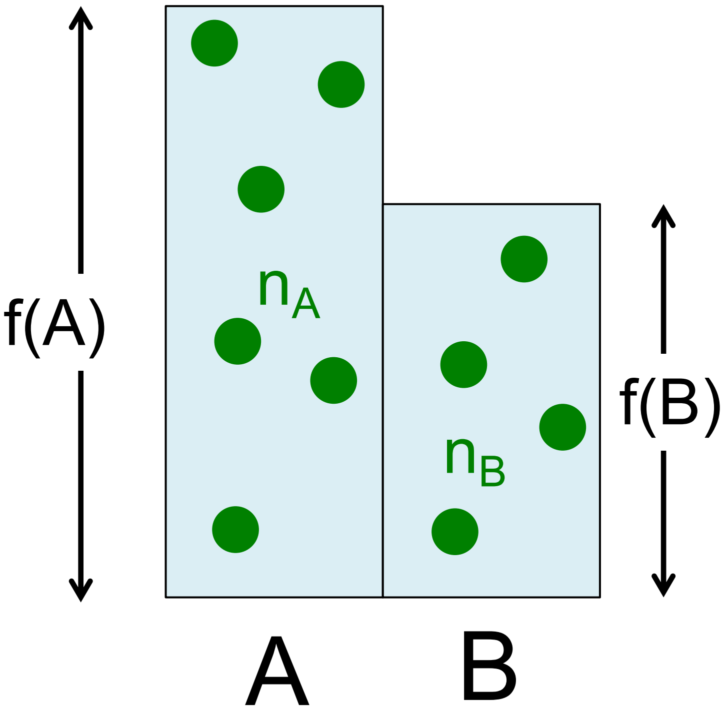




Suppose we have molecules of “dye” diffusing through a container shaped like the posterior distribution

Here the molecules are moving in discrete time steps

We will show that if these molecules move according to the Metropolis acceptance ratio, their locations converge to “match” the posterior distribution



assume $f(A) > f(B)$

$$\Pr(\mathrm{A} \rightarrow \mathrm{B}) = q_{A\rightarrow B} \cdot \frac{f(B)}{f(A)}$$

$$\Pr(\mathrm{B} \rightarrow \mathrm{A}) = q_{B\rightarrow A} \cdot 1$$

$$q_{A\rightarrow B} = q_{B\rightarrow A} = q$$

$$\Pr(A \rightarrow B) = q\frac{f(B)}{f(A)}$$

$$\Pr(B \rightarrow A) = q$$

Net movement from A → B

$$n_A \Pr(\mathrm{A} \rightarrow \mathrm{B}) - n_B \Pr(\mathrm{B} \rightarrow \mathrm{A})$$

$$= q\left(n_A \frac{f(B)}{f(A)} - n_B\right)$$

there will be flow A→B if:

$$n_A \frac{f(B)}{f(A)} - n_B > 0$$

$$\frac{n_A}{n_B} > \frac{f(A)}{f(B)}$$

there will be flow B→A if:

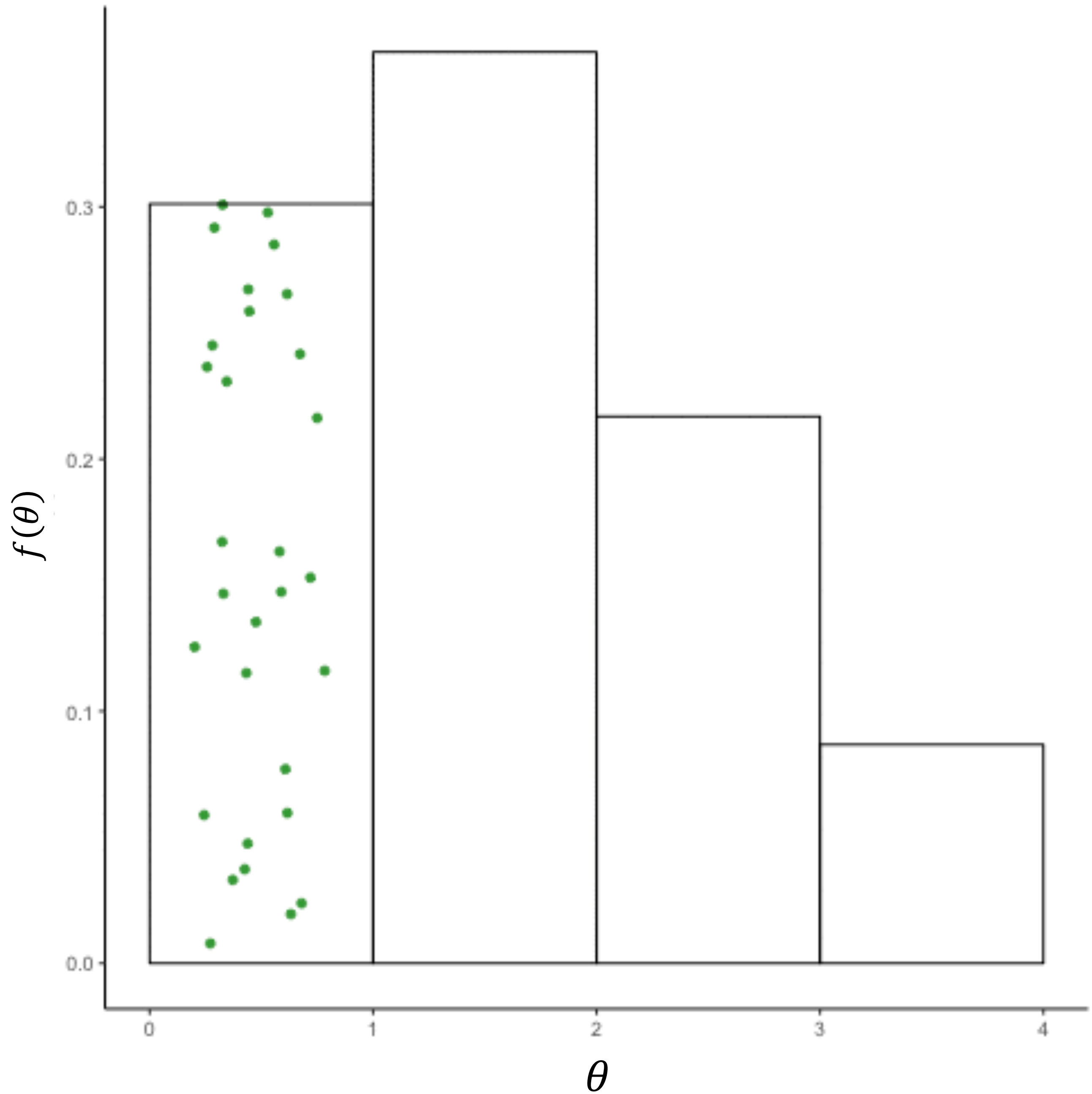
$$n_A \frac{f(B)}{f(A)} - n_B < 0$$

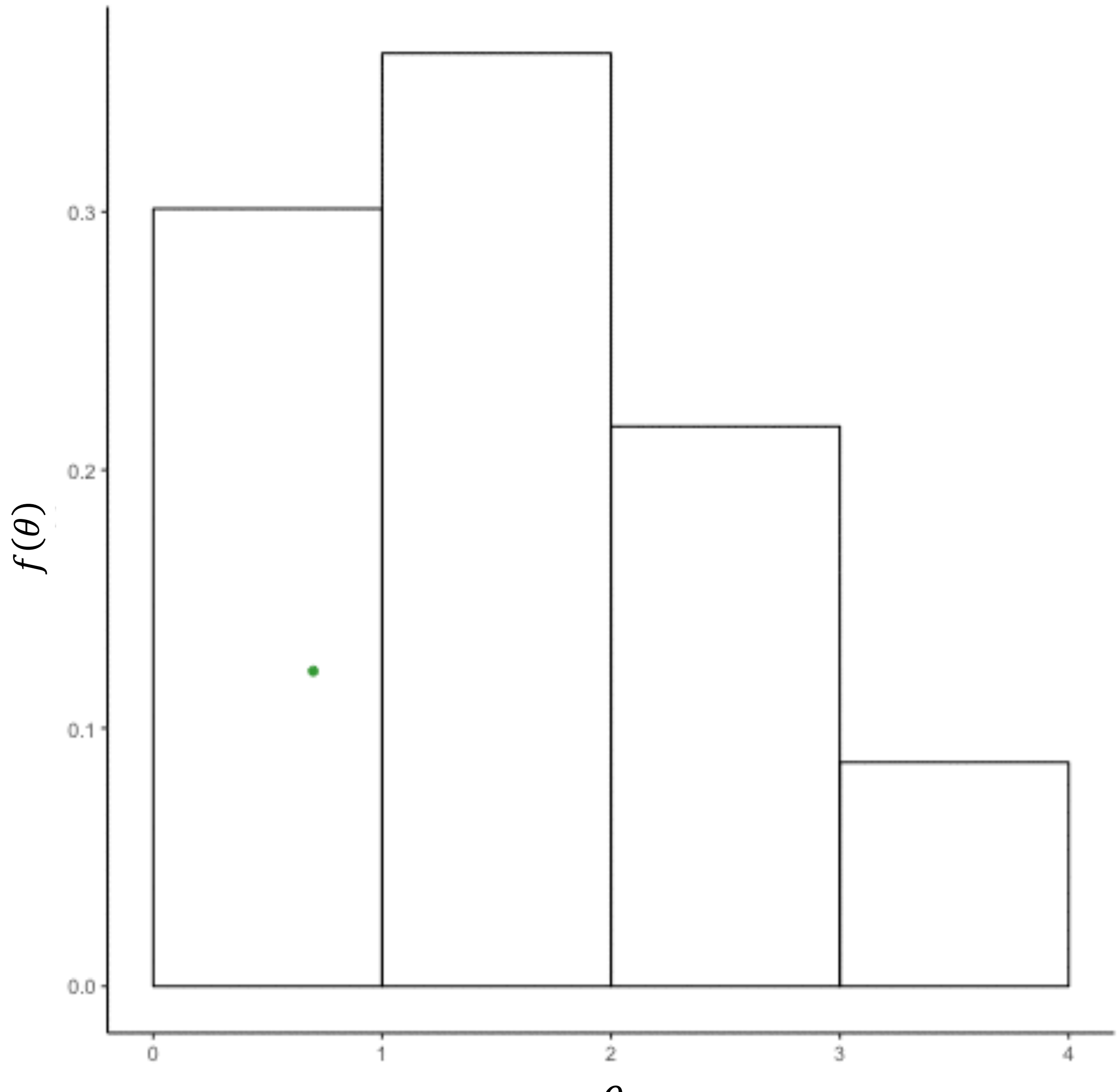
$$\frac{n_A}{n_B} < \frac{f(A)}{f(B)}$$

there will be no net flow A</>B if:

$$n_A \frac{f(B)}{f(A)} - n_B = 0$$

$$\frac{n_A}{n_B} = \frac{f(A)}{f(B)}$$


0.0
0.1
0.2
0.3
$f(\theta)$
0
1
2
3
4
θ



0.3
0.2
0.1
0.0
$f(\theta)$
0
1
2
3
4
θ

Requirements

Symmetry of proposal distribution

“detailed balance” $q_{A\rightarrow B} = q_{B\rightarrow A}$

if $q_{A\rightarrow B} \neq q_{B\rightarrow A}$, use acceptance ratio

$$A = \min\left(1, \frac{f(\theta')}{f(\theta)} \frac{q_{\theta'\rightarrow\theta}}{q_{\theta\rightarrow\theta'}}\right)$$

(Hastings 1970)

Connectedness of distribution

ensures “ergodicity”

Requirements

…$q_{B\to A}$, use acceptance ratio

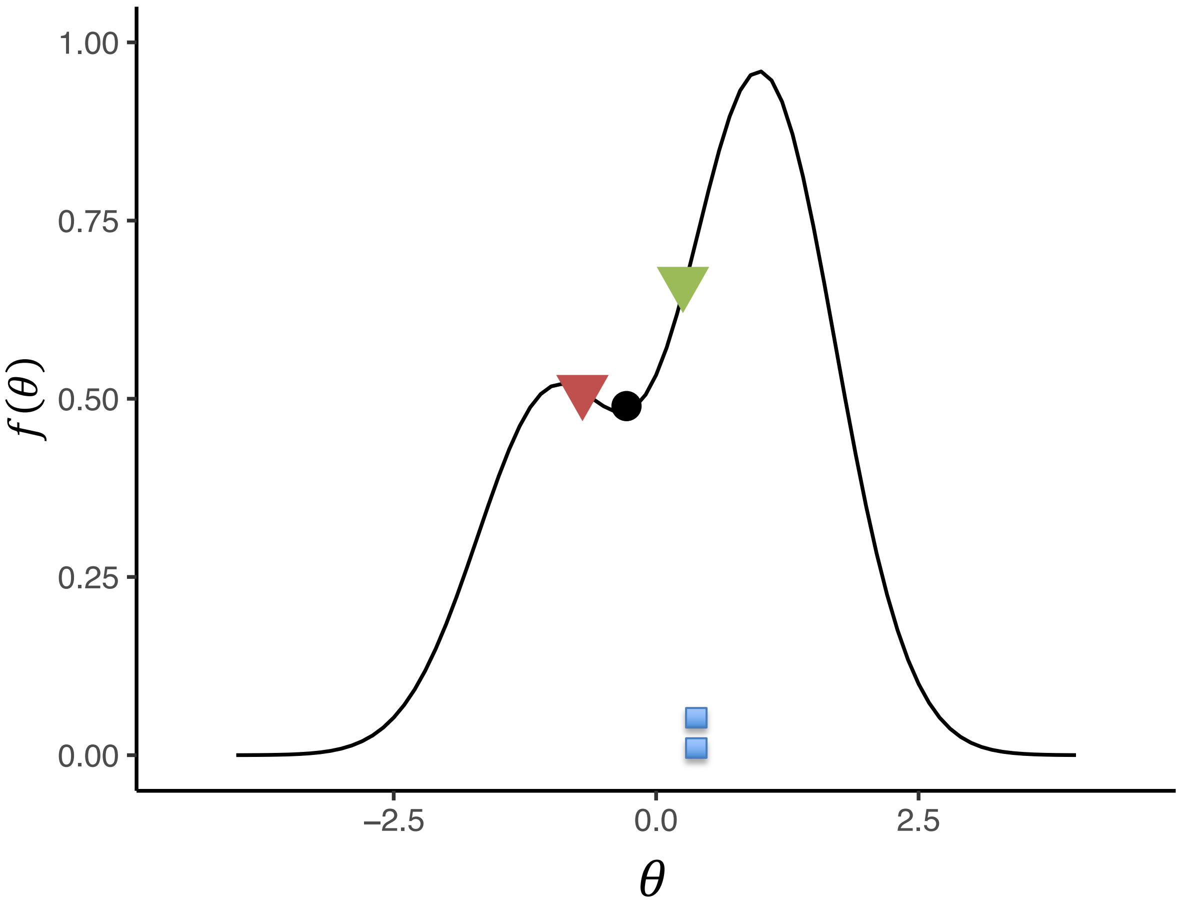
$$A = \min\left(1, \frac{f(\theta')}{f(\theta)} \frac{q_{\theta'\to\theta}}{q_{\theta\to\theta'}}\right)$$

Connectedness of distribution

ensures “ergodicity”

Metropolis-Hastings

MCMC algorithm



Choose a starting point, $\theta = \theta_0$

PROPOSE

$$\theta' = \theta + \varepsilon \mid \varepsilon \sim Q$$

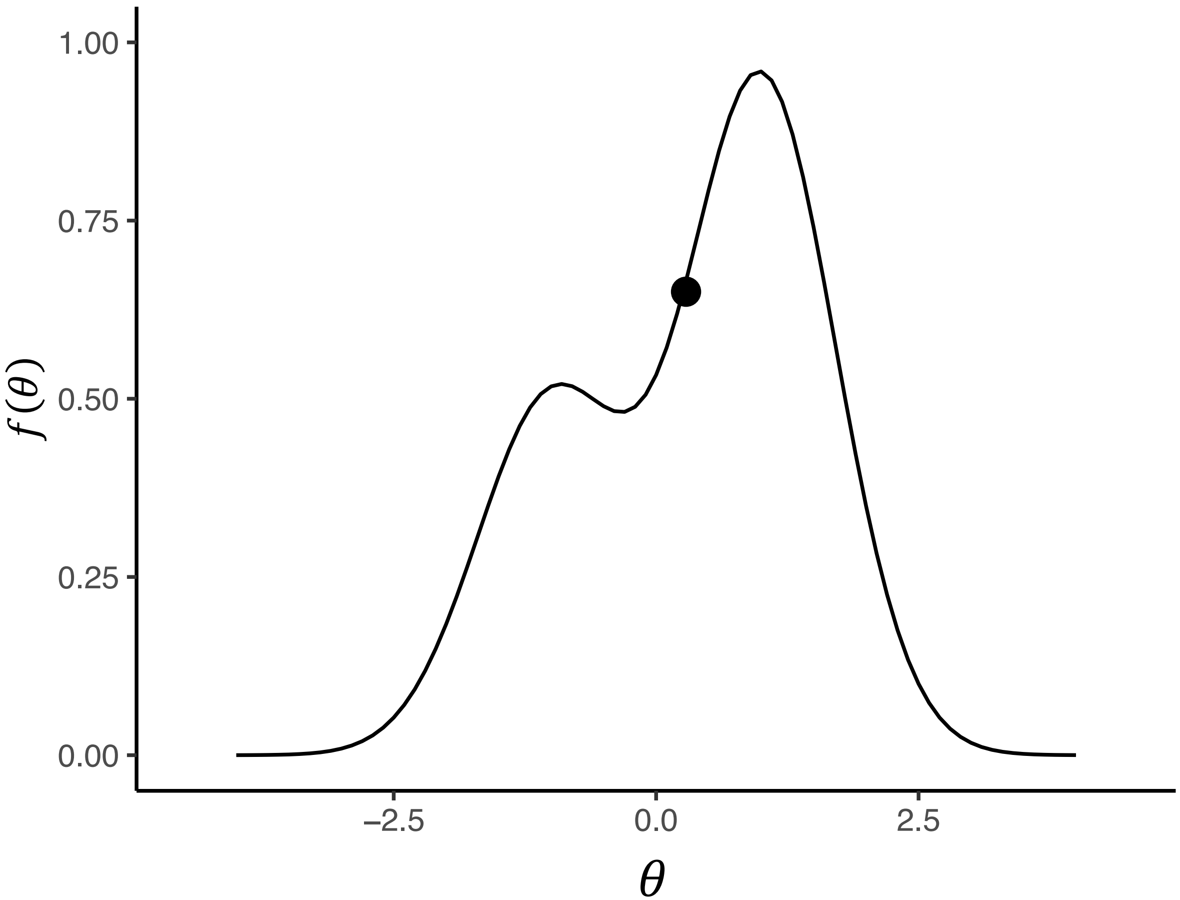
MOVE OR STAY

If $f(\theta') > f(\theta)$, definitely move.

If $f(\theta') < f(\theta)$, move with probability $f(\theta')/f(\theta)$, otherwise stay.

1. PROPOSE
2. MOVE OR STAY ("acceptance")
3. SAVE LOCATION

MCMC algorithm



1. PROPOSE
2. MOVE OR STAY (“acceptance”)
3. SAVE LOCATION

Choose a starting point, $\theta = \theta_0$

PROPOSE

$$\theta' = \theta + \varepsilon \mid \varepsilon \sim Q$$

MOVE OR STAY

$$\theta \to \begin{cases} \theta' & \Pr(a) & Accept \\ \theta & \Pr(1-a) & Reject \end{cases}$$

$$a = \min\left(1, \frac{f(\theta')}{f(\theta)}\right)$$

SAVE LOCATION

$$\theta_t \leftarrow \theta$$