Introduction to statistical inference

Aarya Patil
LSST-Discovery Alliance Catalyst Fellow,
Max-Planck-Institut für Astronomie

Bayesian and Frequentist Statistics

probability of a data set given the null hypothesis

purely driven by the data

prior information

probability of a hypothesis given a particular data set

Fornacon-Wood et al. 2022

Bayesian and Frequentist Statistics

probability of a data set given the null hypothesis

purely driven by the data

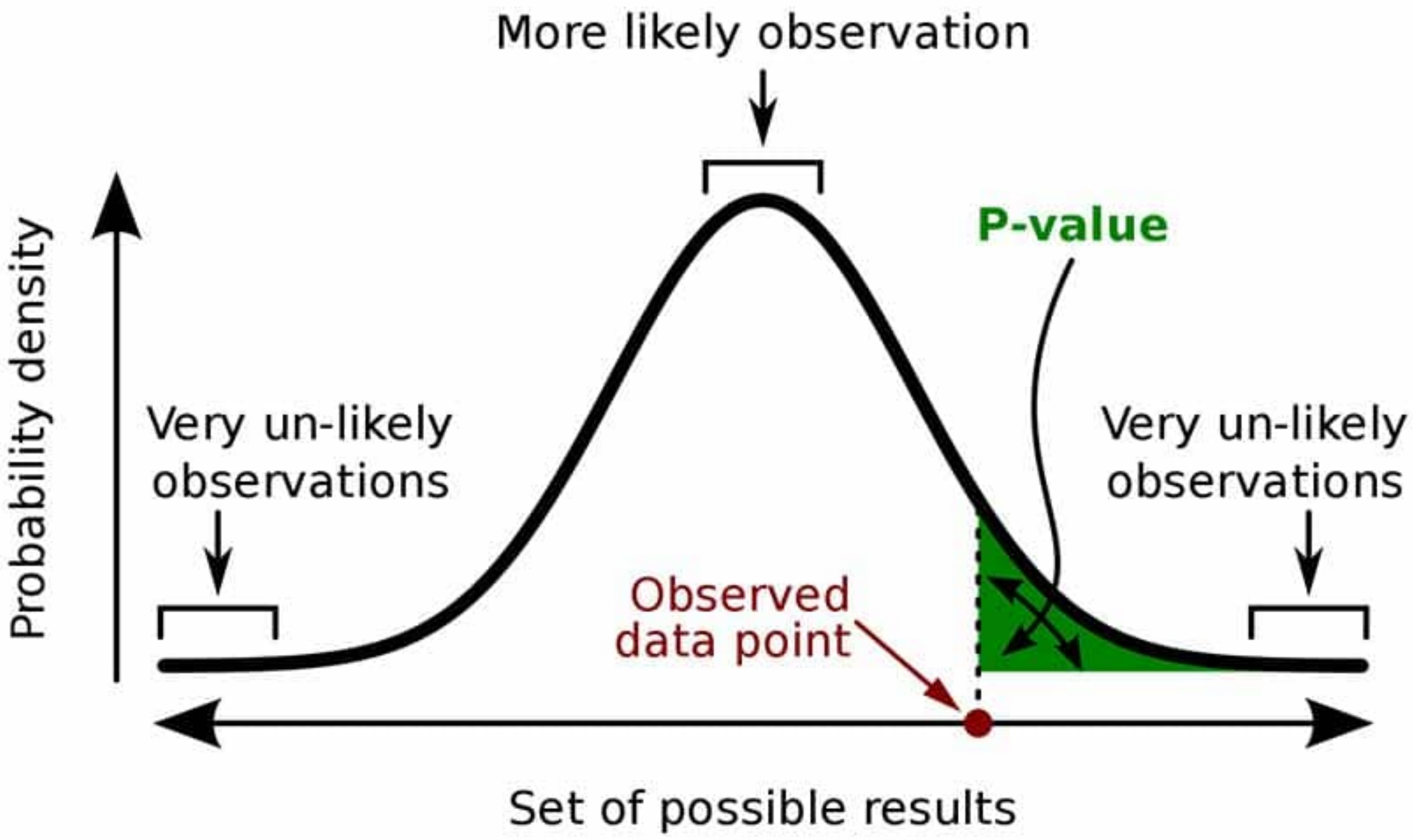
prior information

probability of a hypothesis given a particular data set

***p*-value**

Fornacon-Wood et al. 2022

Bayesian and Frequentist Statistics



p-value

probability of obtaining another data set at least as extreme as the one collected

Fornacon-Wood et al. 2022

Bayesian Statistics

Bayes' theorem

Bayesian Statistics

Bayes' theorem

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

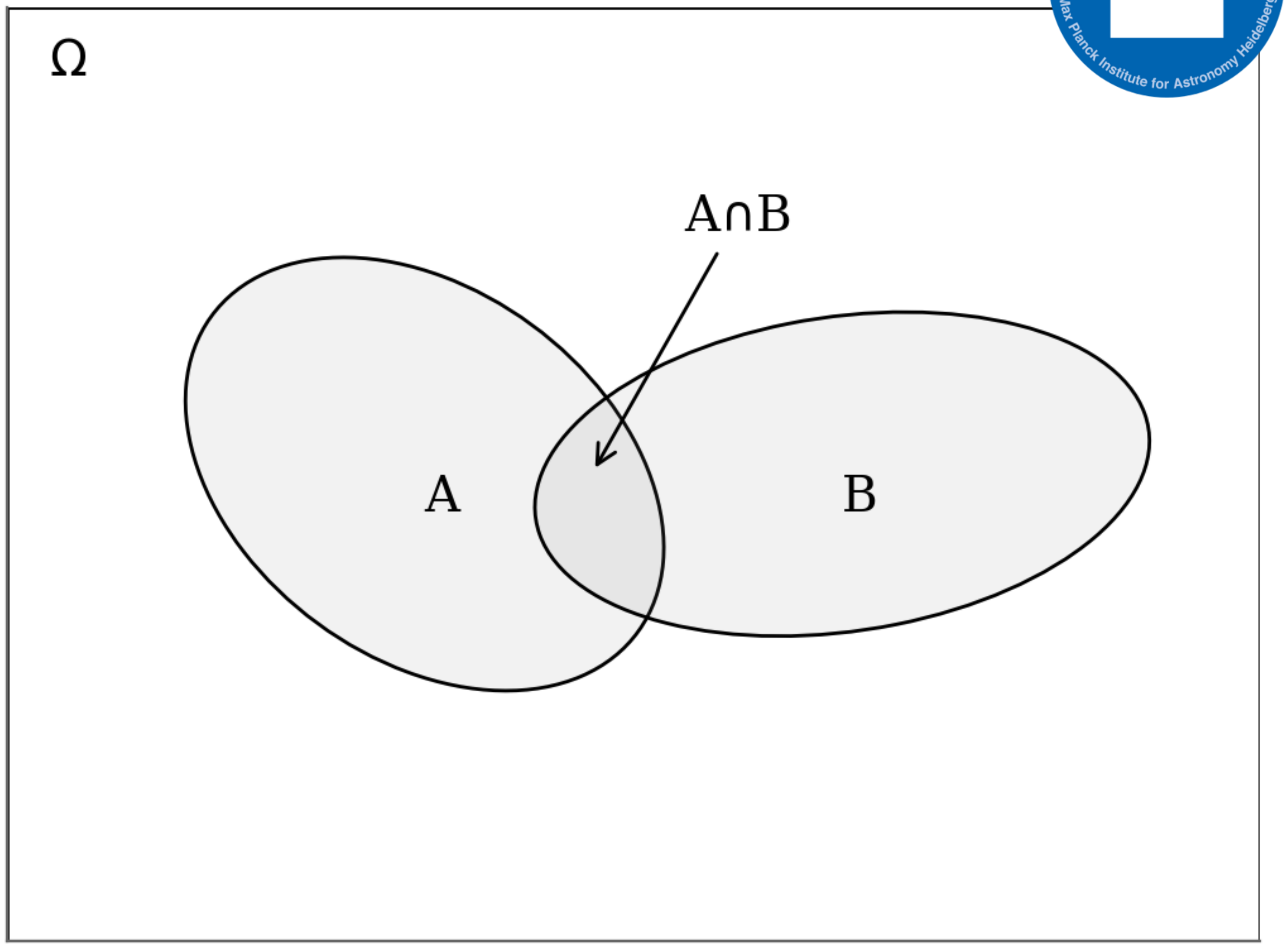
$$[P(B) \neq 0]$$

Bayesian Statistics

Bayes' theorem

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

$$[P(B) \neq 0]$$



$$P(A, B \mid \Omega) = P(A \mid B, \Omega)\, P(B \mid \Omega)$$
$$= P(B \mid A, \Omega)\, P(A \mid \Omega)$$

Data modeling

1. Parameter estimation

2. Model comparison

3. Prediction

Zero-parameter models

How to interpret test results?

A test for COVID-19 gives either a positive or a negative result, and is 98% reliable.

You test positive. What is the probability that you have the disease?

- 98%?
- <98%?
- >98%?

How to interpret test results?

A test for COVID-19 gives either a positive or a negative result, and is 98% reliable.

Probability of testing positive in the absence of COVID-19 is 0.01.

You test positive. What is the probability that you have the disease?

- 98%?
- 99%?
- 97%?
- other?

How to interpret test results?

A test for COVID-19 gives either a positive or a negative result, and is 98% reliable.

Probability of testing positive in the absence of COVID-19 is 0.01.

Among people showing no symptoms, 1 in 200 have COVID-19.

You test positive. What is the probability that you have the disease?

- 98%?
- 99%?
- 99.5%?
- other?

Hypothesis testing

Result D	Is Model M true? M denotes if a person has COVID-19	
	Yes	No
positive	true positive $P(\mathrm{D} \mid \mathrm{M})$	false positive $P(\mathrm{D} \mid \mathrm{M}')$
negative	false negative $P(\mathrm{D}' \mid \mathrm{M})$	true negative $P(\mathrm{D}' \mid \mathrm{M}')$

Hypothesis testing

$$P(\mathrm{M} \mid \mathrm{D}) = \frac{1}{1 + \frac{1}{R}}$$

$$R = \frac{P(\mathrm{D} \mid \mathrm{M})\, P(\mathrm{M})}{P(\mathrm{D} \mid \mathrm{M}')\, P(\mathrm{M}')}$$

“Posterior odds ratio”

Hypothesis testing

$$P(\mathrm{M} \mid \mathrm{D}) = \frac{1}{1 + \frac{1}{R}}$$

$$R = \frac{P(\mathrm{D} \mid \mathrm{M})\, P(\mathrm{M})}{P(\mathrm{D} \mid \mathrm{M}')\, P(\mathrm{M}')}$$

"Posterior odds ratio"

$$\boxed{P(\mathrm{M} \mid \mathrm{D}) = \frac{P(\mathrm{D} \mid \mathrm{M})\, P(\mathrm{M})}{P(\mathrm{D})}}$$

Exercise:

Derive using Bayes' theorem

Exercise

Python notebook

Thinking in terms of frequencies

Parametric models

Posterior

Likelihood

Prior

$$P(\boldsymbol{\Theta} \mid \mathbf{D}, \mathrm{M}) = \frac{P(\mathbf{D} \mid \boldsymbol{\Theta}, \mathrm{M})\, P(\boldsymbol{\Theta} \mid \mathrm{M})}{P(\mathbf{D} \mid \mathrm{M})}$$

Evidence

Why is sampling from $P(\mathbf{x})$ difficult?

Let's assume we can evaluate a function $P^*(\mathbf{x})$ such that

$$P(\mathbf{x}) = P^*(\mathbf{x})/Z$$

Difficulties:

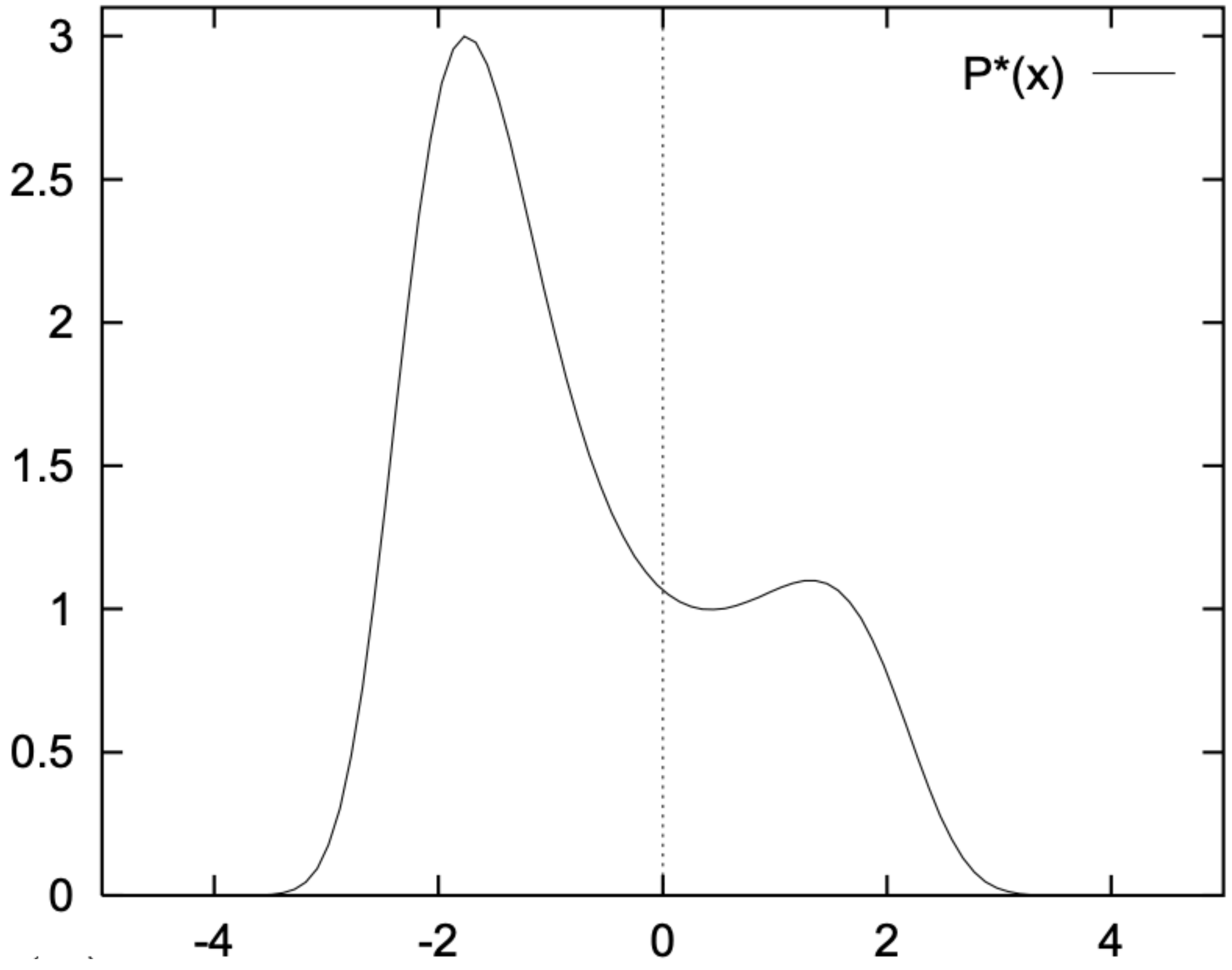
1. Normalizing constant $$Z = \int d^N\mathbf{x}\, P^*(\mathbf{x})$$

2. High-dimensional spaces

MacKay 1998

Why is sampling from $P(\mathbf{x})$ difficult?

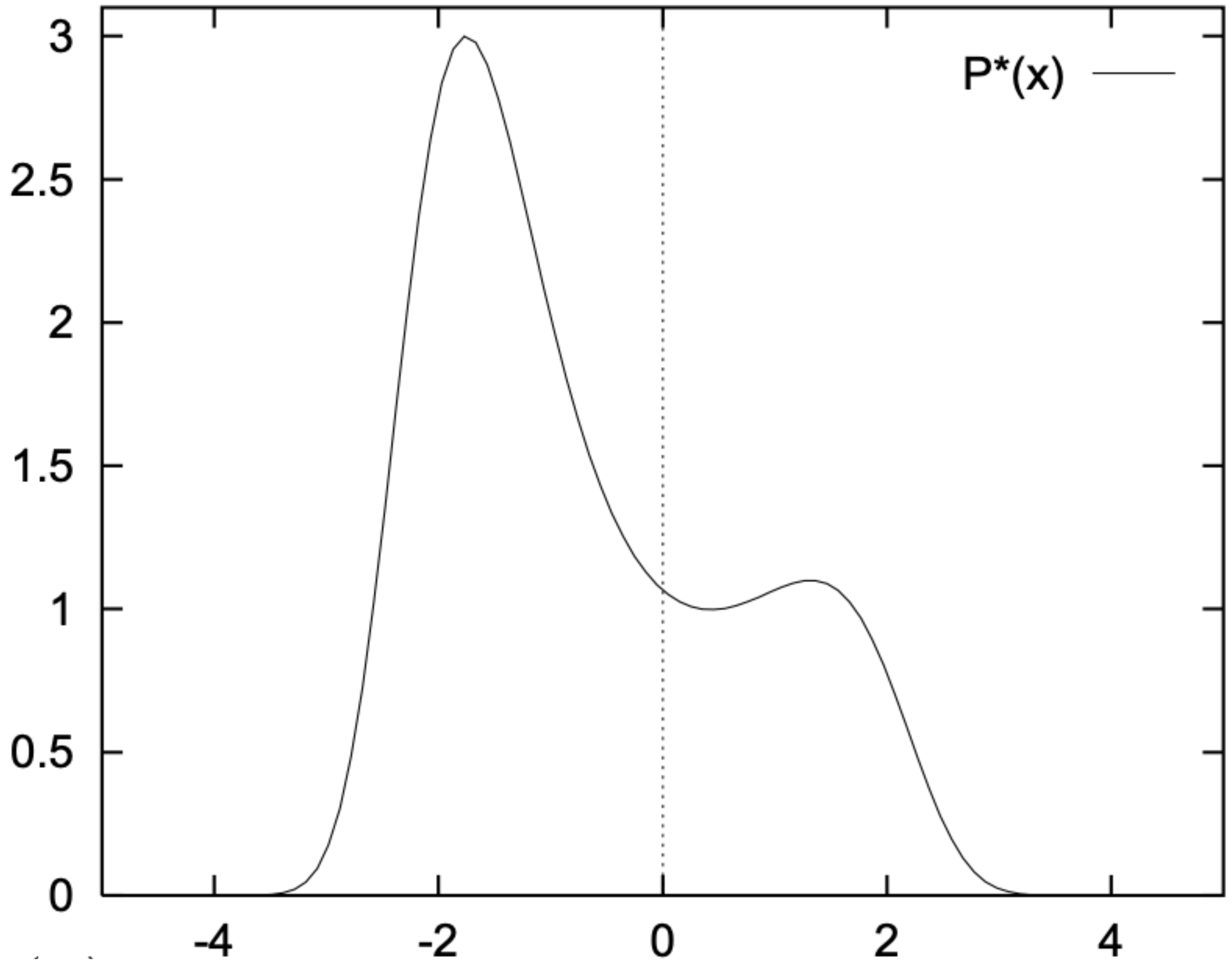
$$P^*(x) = \exp\left[0.4(x-0.4)^2 - 0.08x^4\right],\ \ x \in (-\infty, +\infty)$$



MacKay 1998

Why is sampling from $P(\mathbf{x})$ difficult?

$$P^*(x) = \exp\left[0.4(x-0.4)^2 - 0.08x^4\right], \;\; x \in (-\infty, +\infty)$$



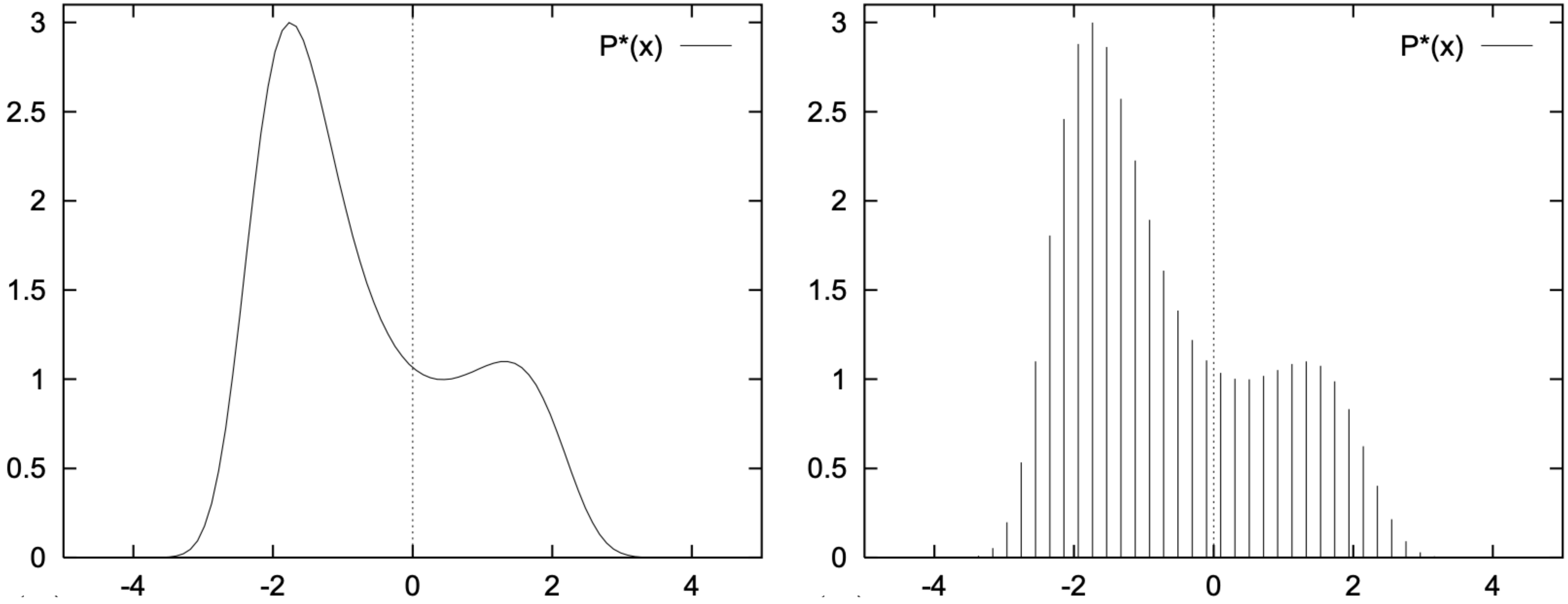
Exercise:

How would you describe this distribution?

MacKay 1998

Why is sampling from $P(\mathbf{x})$ difficult?

$$P^*(x) = \exp\left[0.4(x-0.4)^2 - 0.08x^4\right],\ \ x \in (-\infty, +\infty)$$

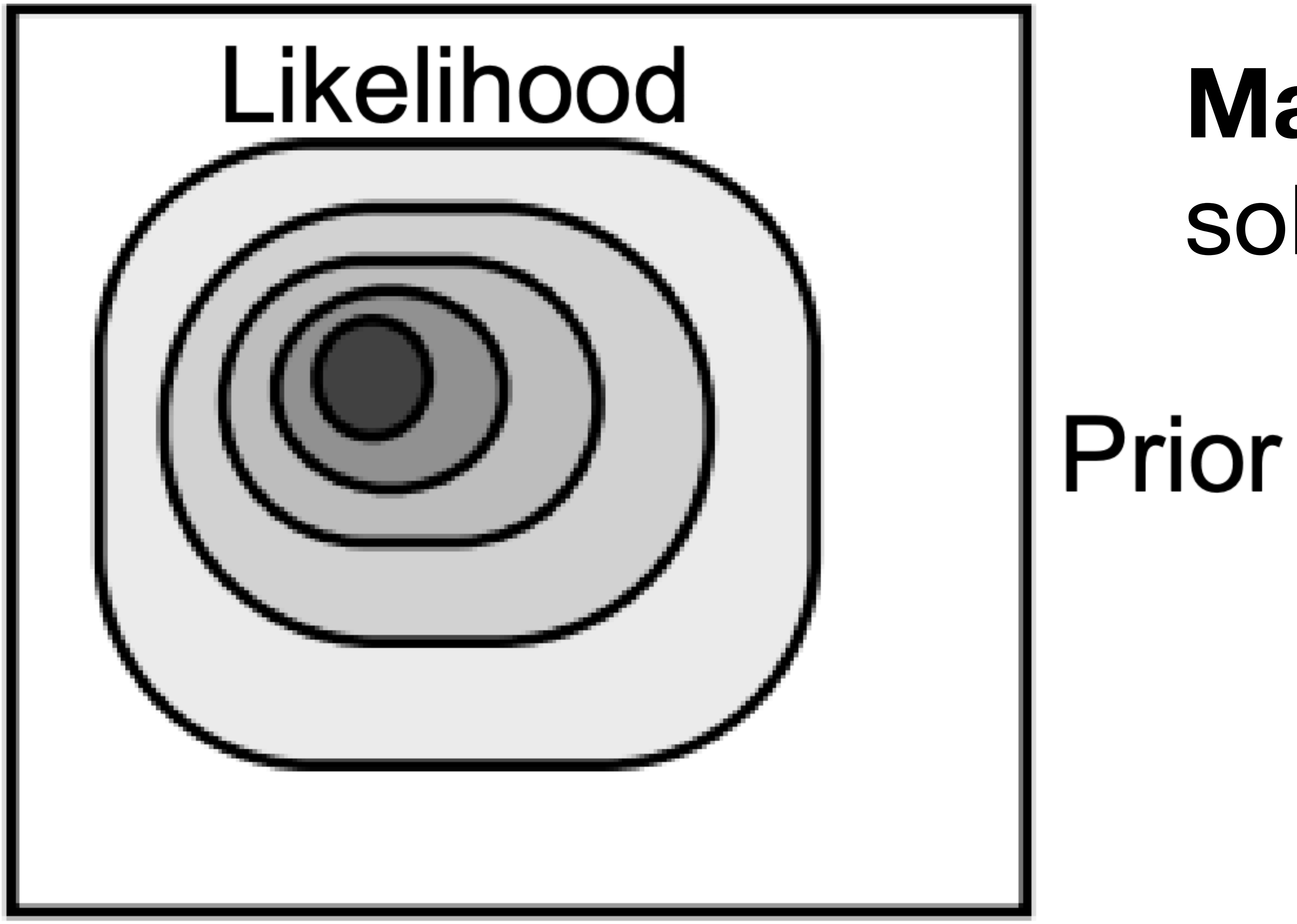


$$Z = \sum_i p_i^*$$

$$p_i = p_i^*/Z$$

MacKay 1998

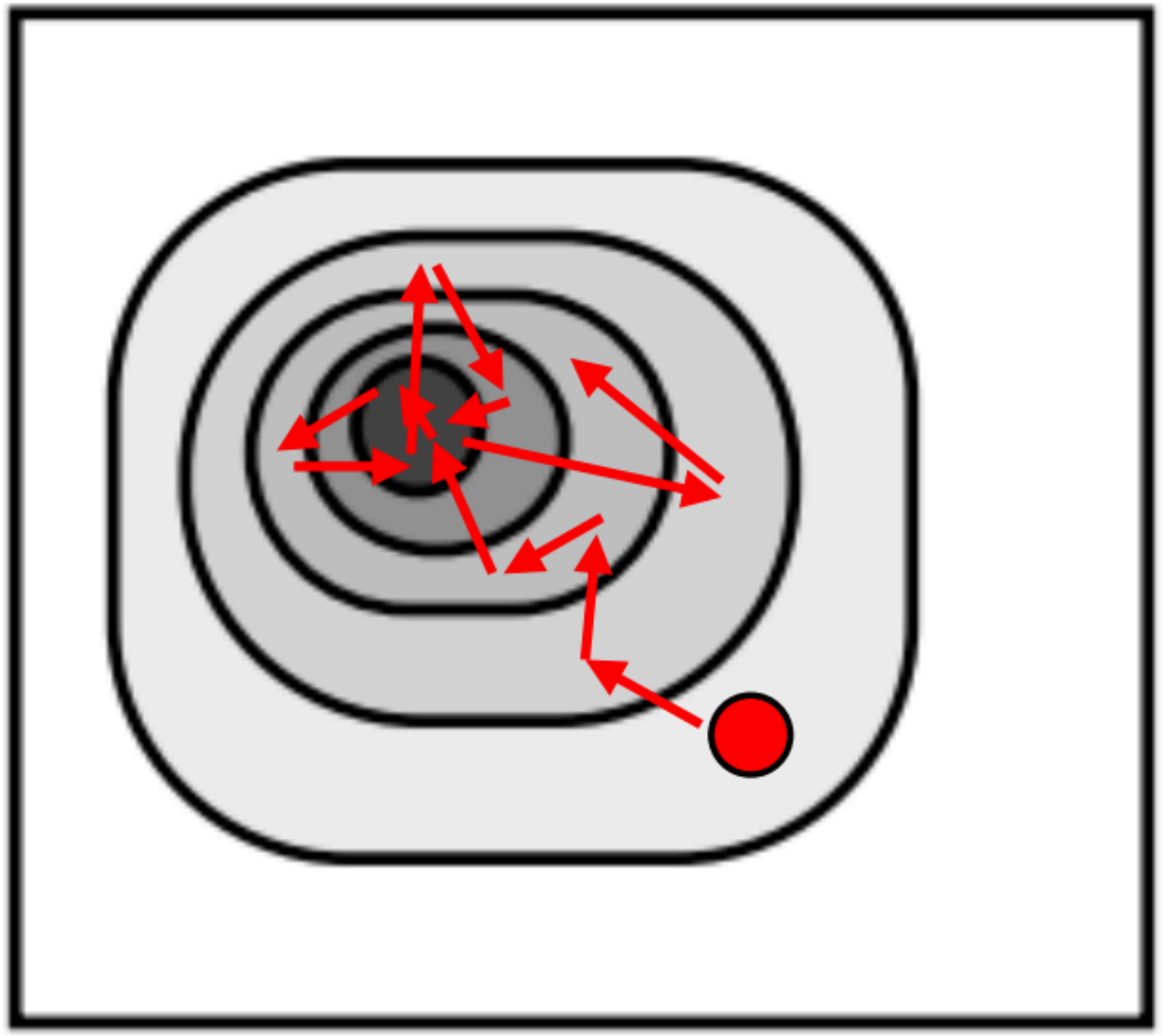
Sampling a distribution



Markov Chain Monte Carlo (MCMC)
solving a difficult problem once

Adapted from 2021 talk by Josh Speagle

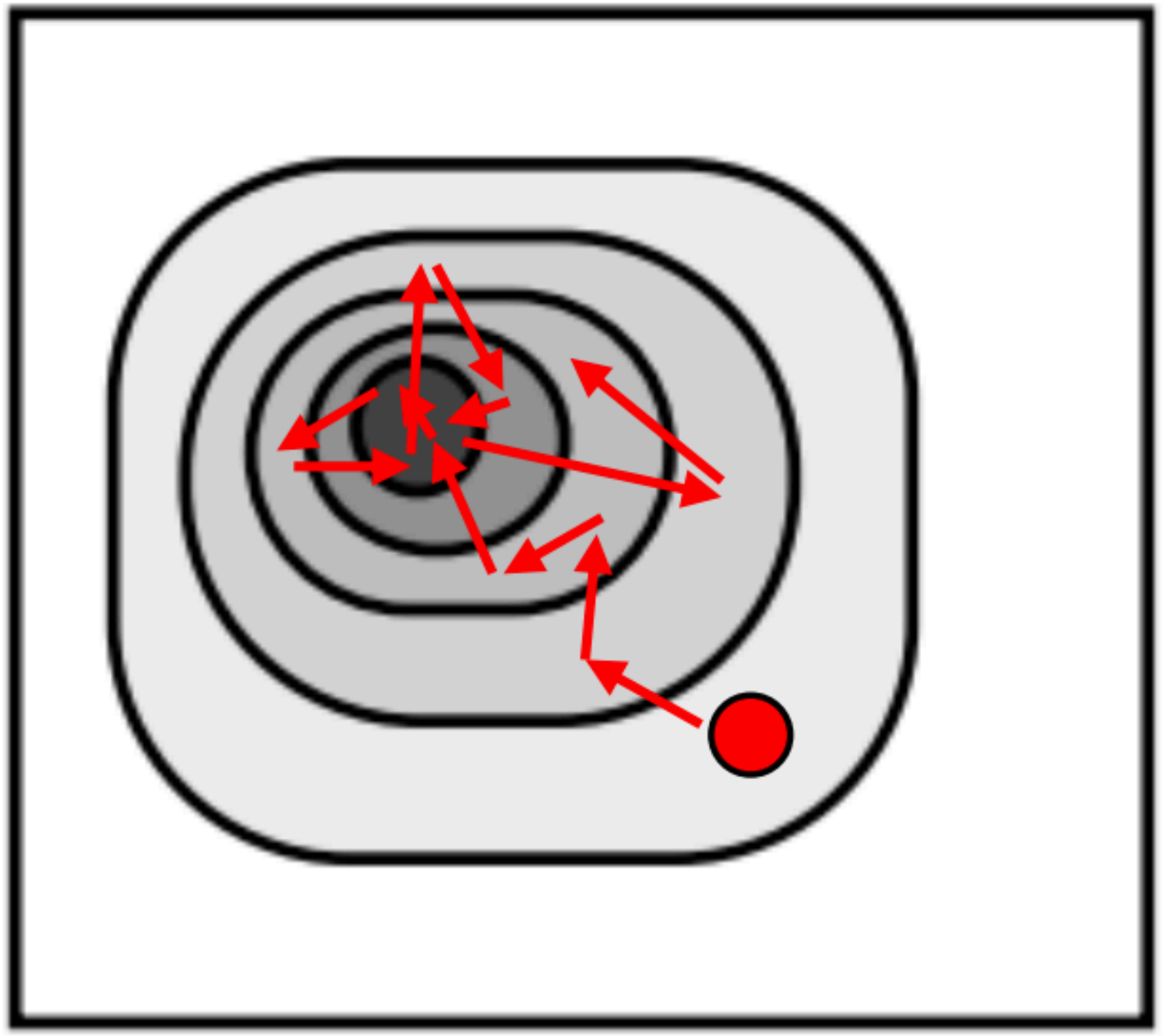
Sampling a distribution



Markov Chain Monte Carlo (MCMC)
solving a difficult problem once

Adapted from 2021 talk by Josh Speagle

Sampling a distribution

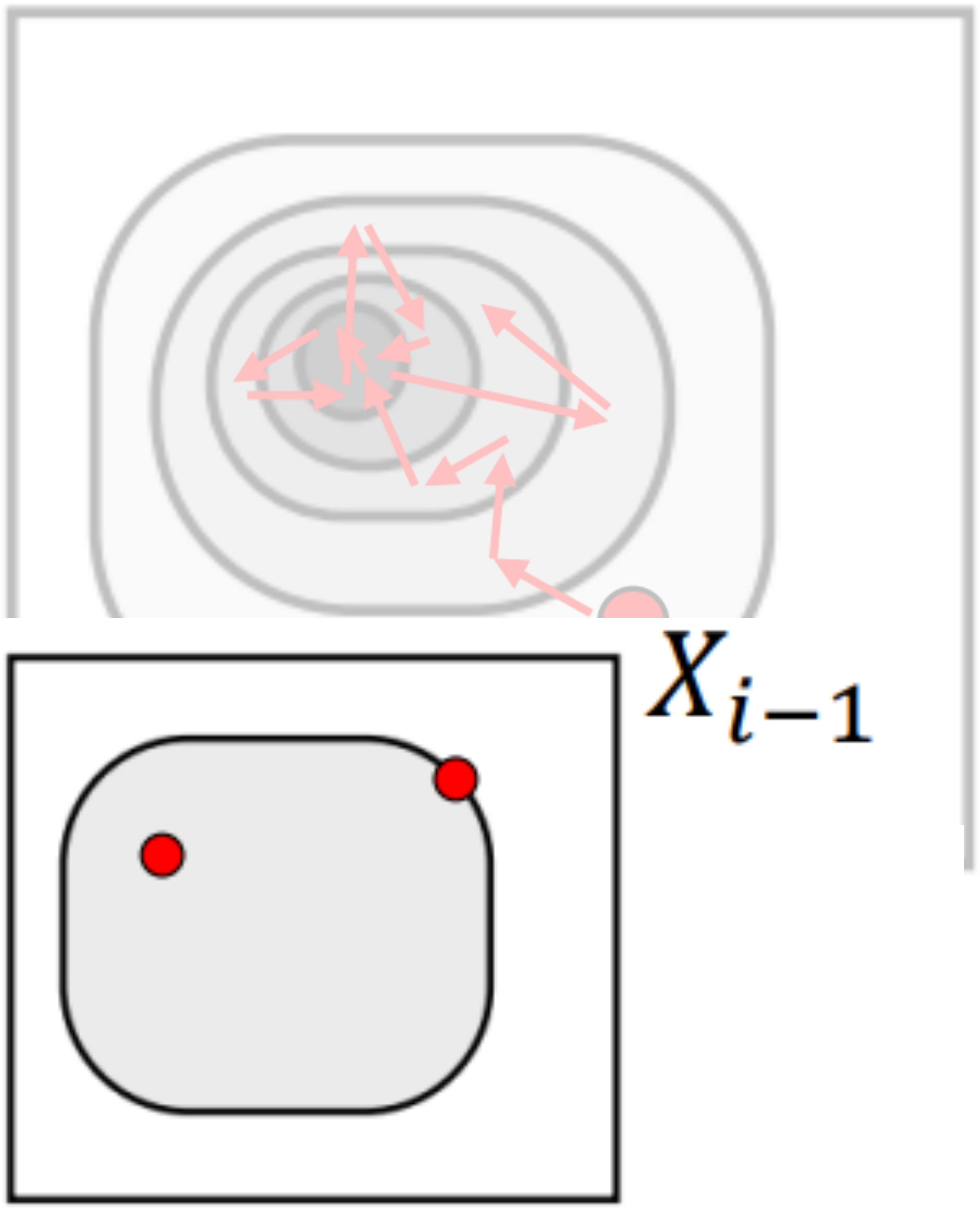


Markov Chain Monte Carlo (MCMC)
solving a difficult problem once

Nested Sampling
solving an easier problem several times

Adapted from 2021 talk by Josh Speagle

Sampling a distribution



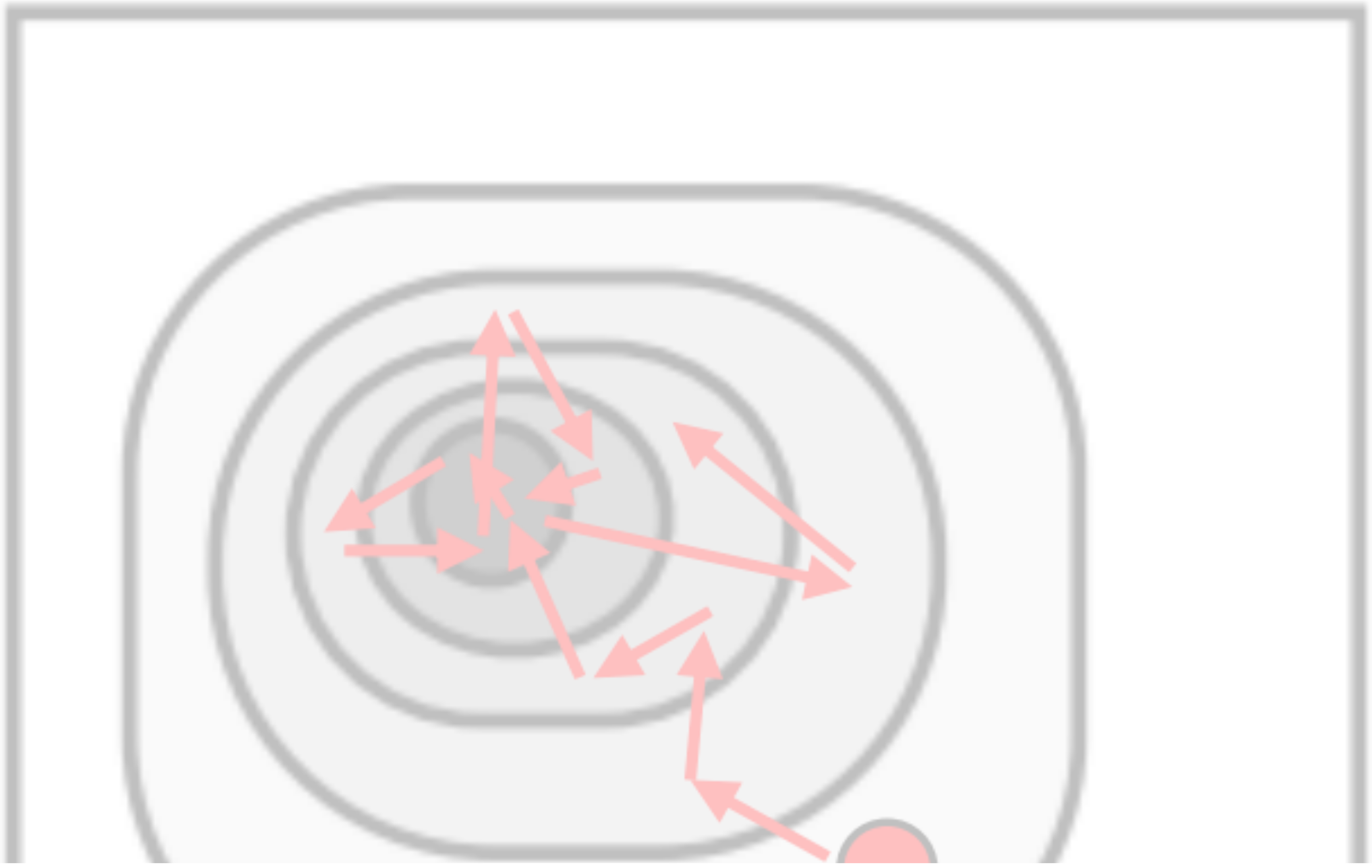
Markov Chain Monte Carlo (MCMC)
solving a difficult problem once

Nested Sampling
solving an easier problem several times

Sampling uniformly within bound
$P(\mathbf{D} \mid \mathbf{\Theta}, \mathrm{M}) > \gamma$ easier

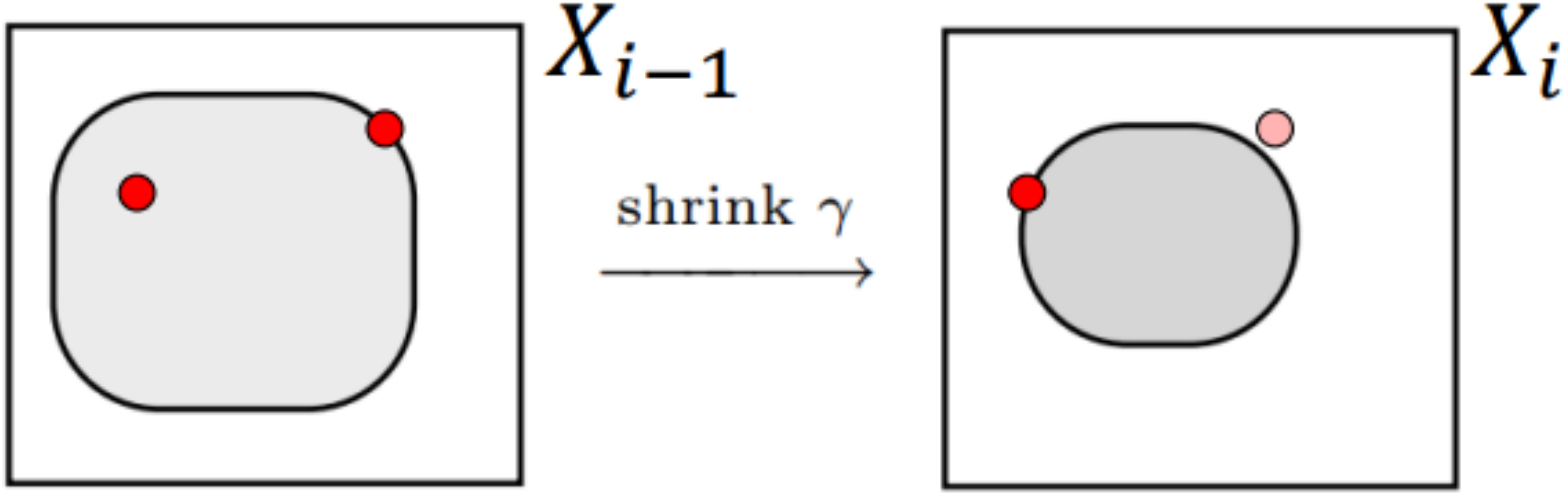
Adapted from 2021 talk by Josh Speagle

Sampling a distribution



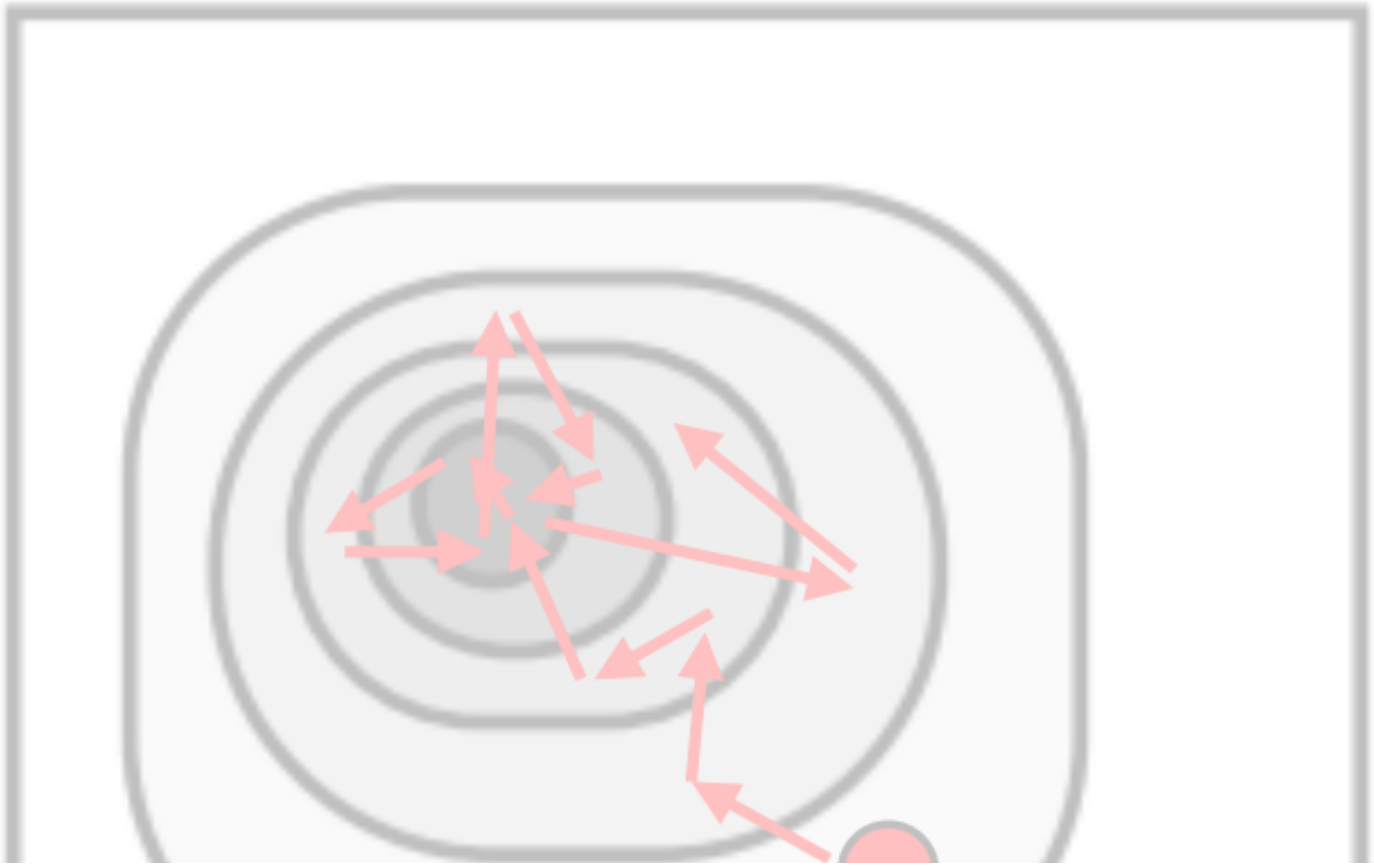
Markov chain Monte Carlo (MCMC)
solving a difficult problem once

Nested Sampling
solving an easier problem several times



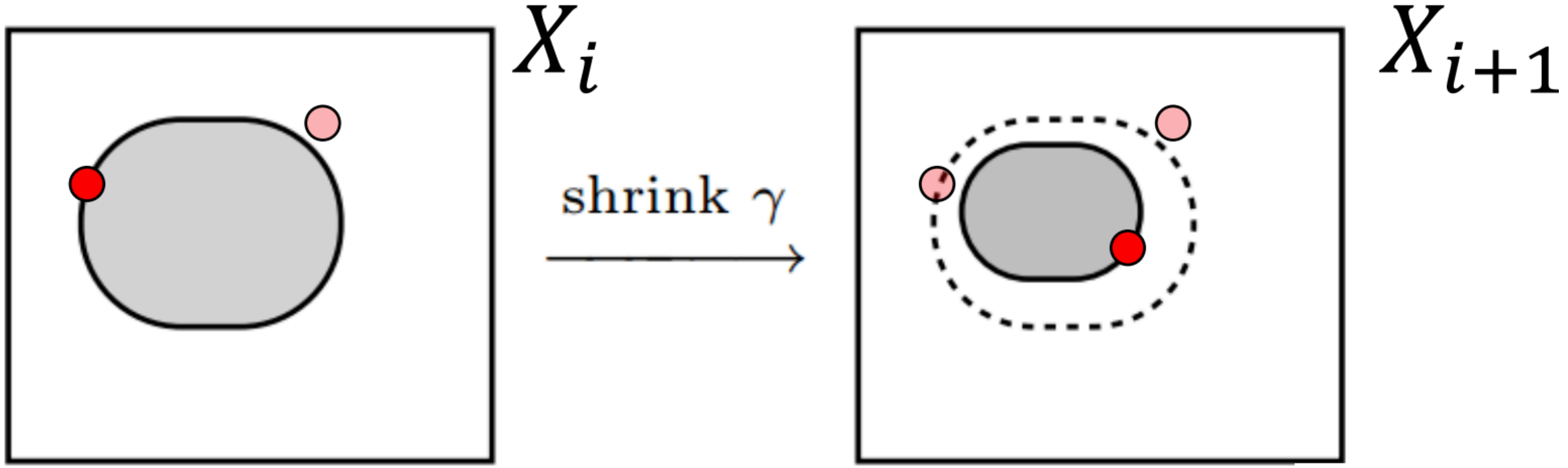
Adapted from 2021 talk by Josh Speagle

Sampling a distribution



Markov chain Monte Carlo (MCMC)
solving a difficult problem once

Nested Sampling
solving an easier problem several times



Adapted from 2021 talk by Josh Speagle

Integrating the posterior

$$\overset{\textbf{Posterior}}{P(\boldsymbol{\Theta} \mid \mathbf{D}, \mathrm{M})} = \frac{\overset{\textbf{Likelihood}}{P(\mathbf{D} \mid \boldsymbol{\Theta}, \mathrm{M})} \; \overset{\textbf{Prior}}{P(\boldsymbol{\Theta} \mid \mathrm{M})}}{\underset{\textbf{Evidence}}{P(\mathbf{D} \mid \mathrm{M})}}$$

Integrating the posterior

Likelihood

Prior

Posterior

$$P(\mathbf{\Theta} \,|\, \mathbf{D}, \mathrm{M}) = \frac{L(\mathbf{\Theta}) \qquad \pi(\mathbf{\Theta})}{Z}$$

Evidence

$$\equiv \int_{\Omega_{\mathbf{\Theta}}} L(\mathbf{\Theta})\pi(\mathbf{\Theta})d\mathbf{\Theta}$$

Linear regression

Python notebook