Statistical Techniques for Data Science & Robotics

Week 11

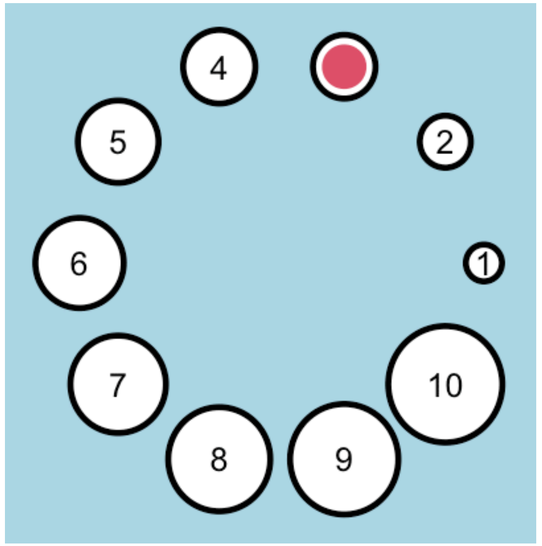
Objectives for today

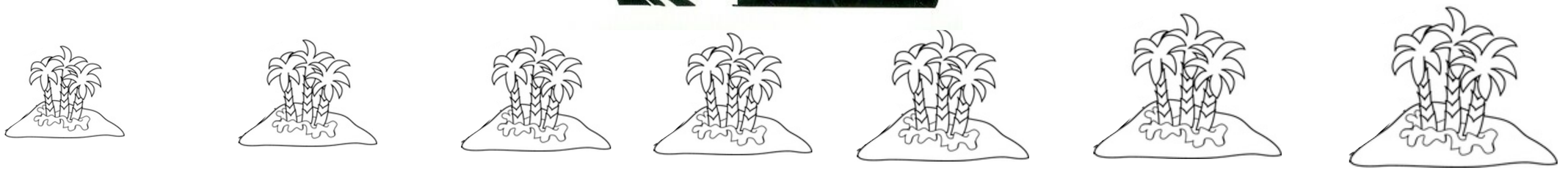
- Markov Chains
- Metropolis-Hastings algorithm
  - Random walk Metropolis
  - Metropolis-Hastings

King Markov

- **King Markov was a benevolent autocrat of an island kingdom, a circular archipelago, with 10 islands.**

Source: https://speakerdeck.com/rmcelreath/statistical-rethinking-2023-lecture-08

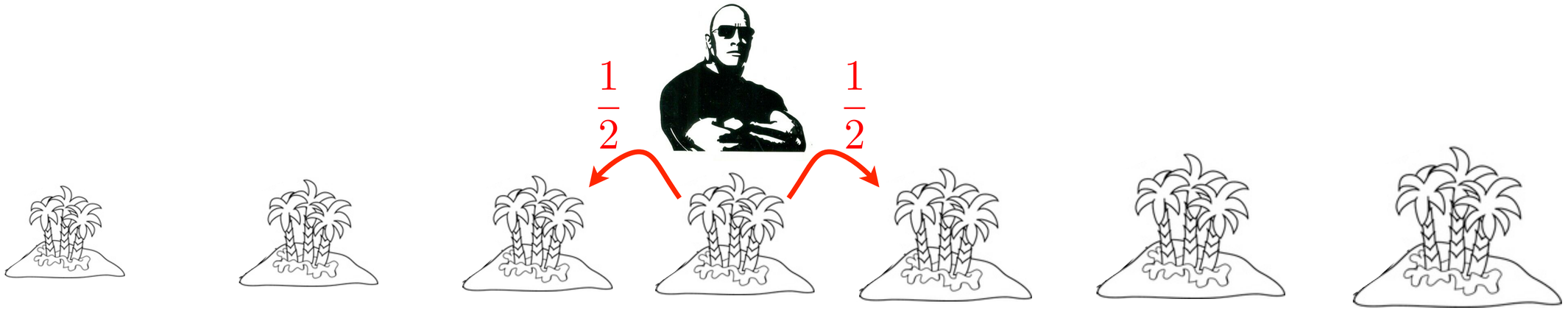




The Metropolis Archipelago

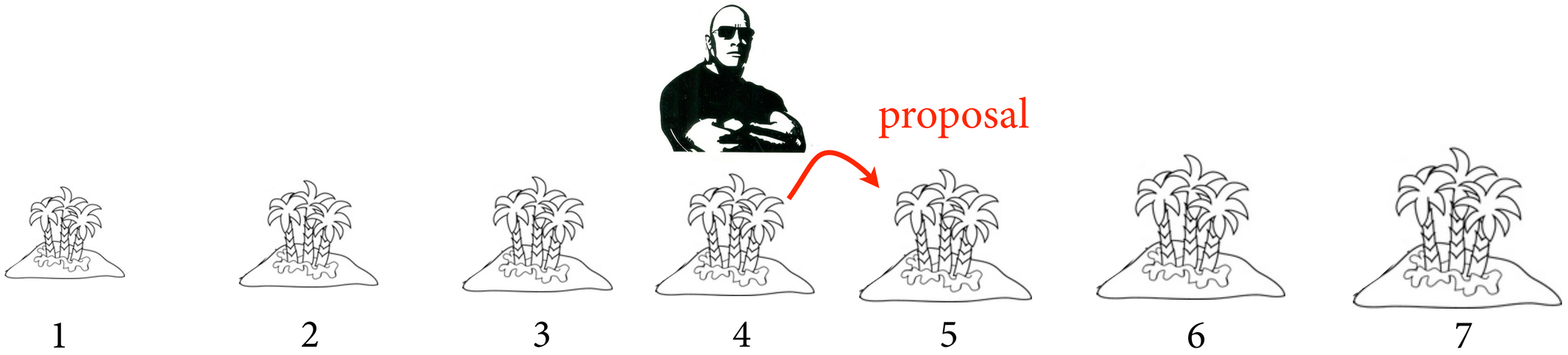
Contract: King Markov must visit each island in proportion to its population size

Here's how he does it...



(1) Flip a coin to choose island on left or right.
Call it the “proposal” island.

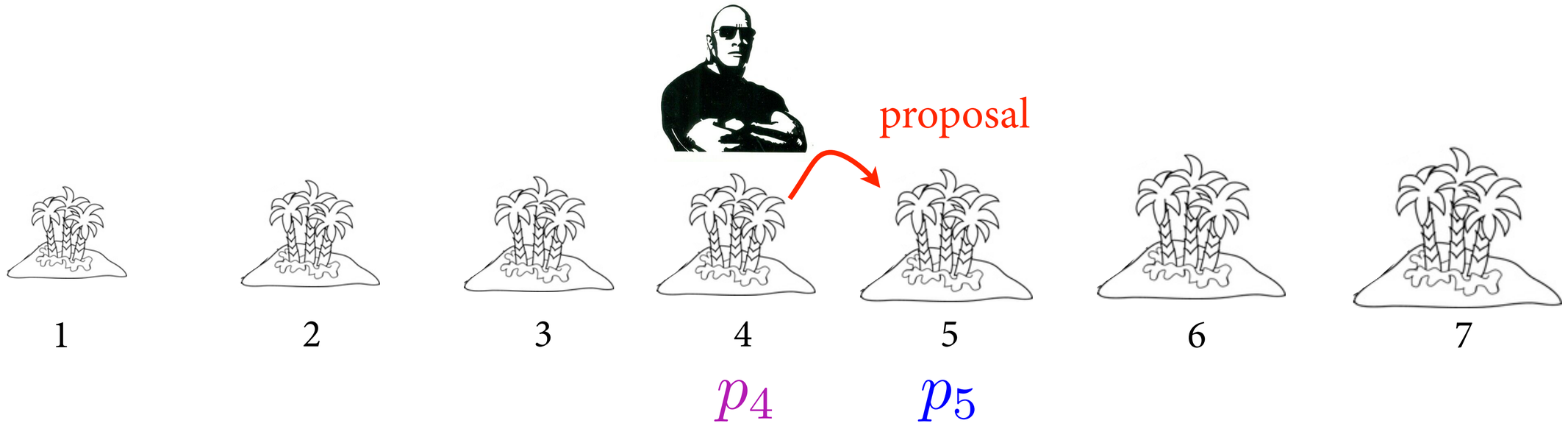
(1) Flip a coin to choose island on left or right.
Call it the "proposal" island.



(2) Find population of proposal island.

(1) Flip a coin to choose island on left or right.
Call it the "proposal" island.

(2) Find population of proposal island.

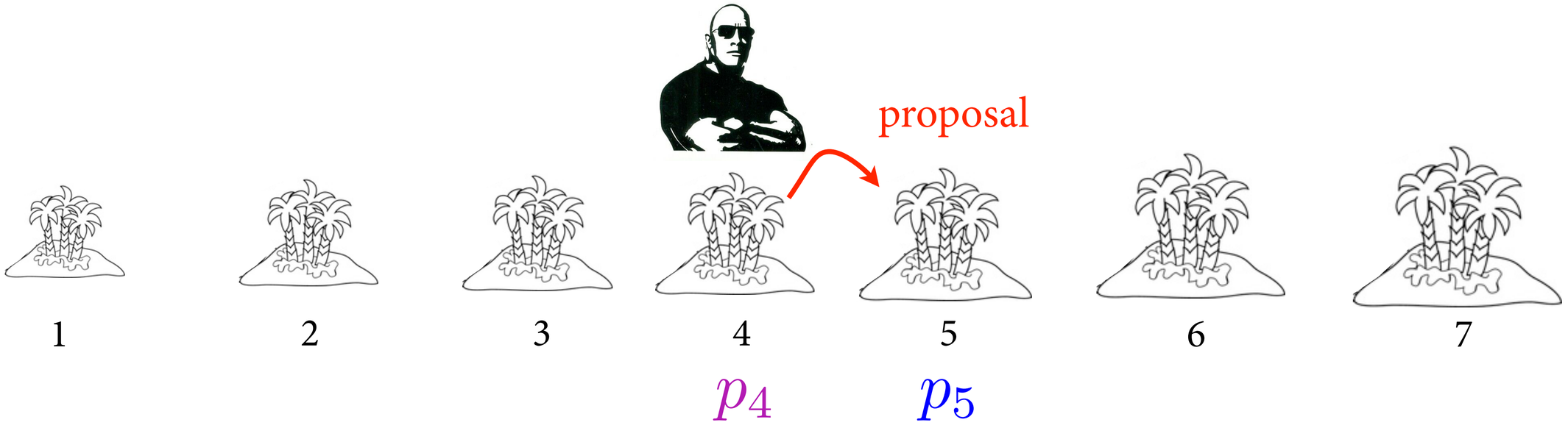


(3) Find population of current island.

(1) Flip a coin to choose island on left or right.
Call it the "proposal" island.

(2) Find population of proposal island.

(3) Find population of current island.



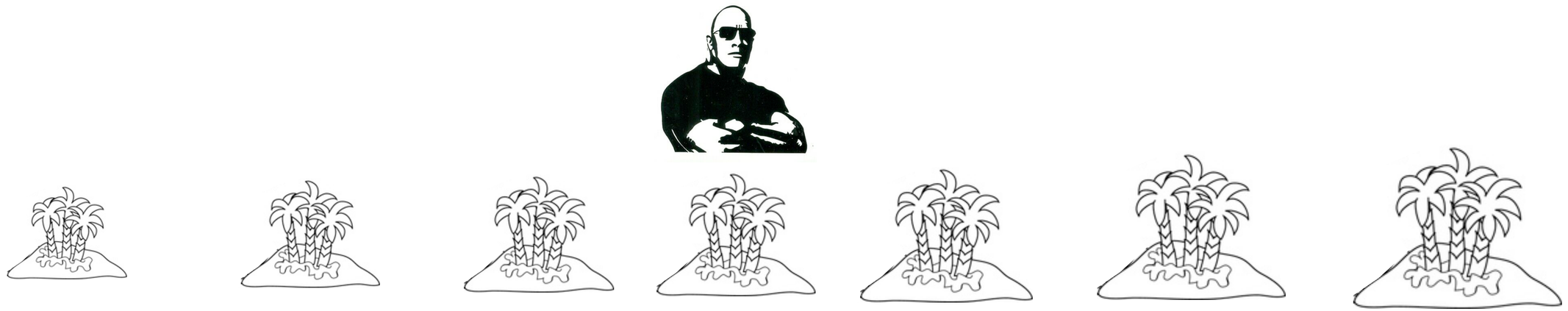
(4) Move to proposal, with probability = $\frac{p_5}{p_4}$

(1) Flip a coin to choose island on left or right.
Call it the "proposal" island.

(2) Find population of proposal island.

(3) Find population of current island.

(4) Move to proposal, with probability = $\frac{p_5}{p_4}$



(5) Repeat from (1)

(1) Flip a coin to choose island on left or right.
Call it the "proposal" island.

(2) Find population of proposal island.

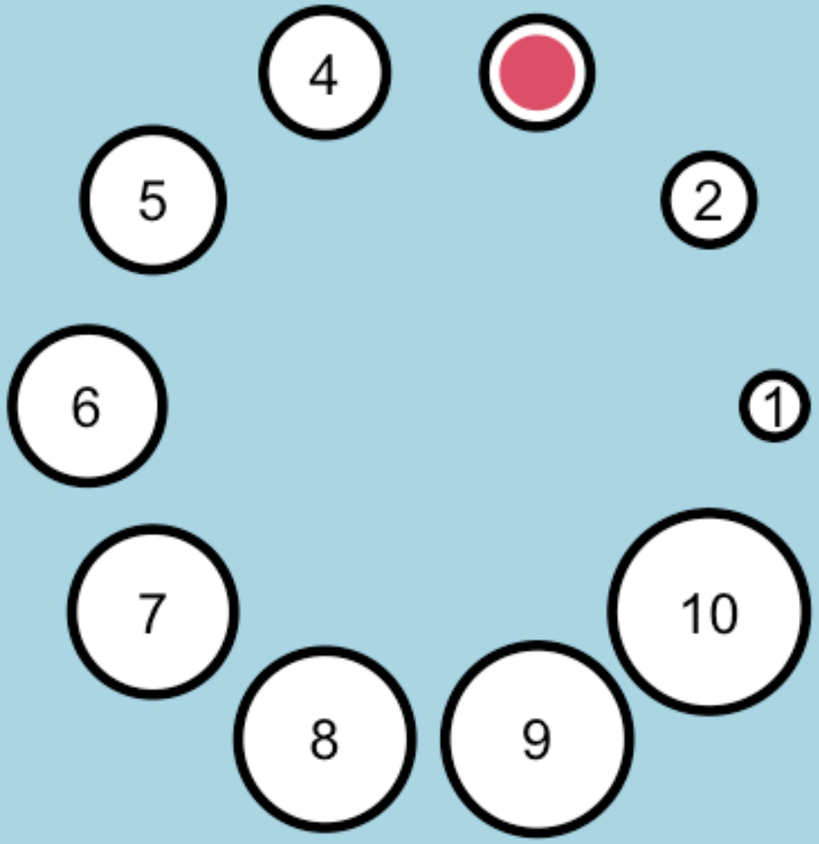
(3) Find population of current island.

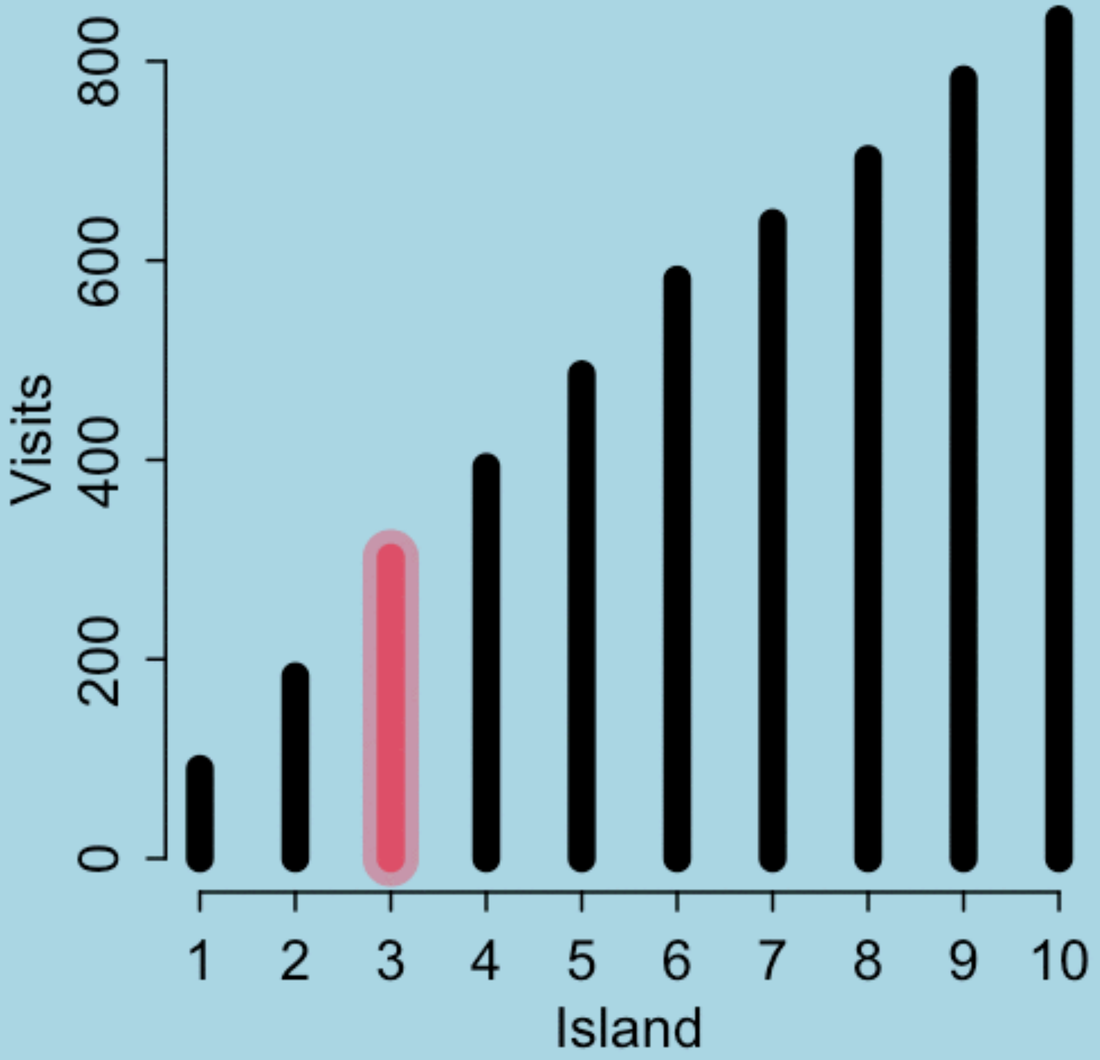
(4) Move to proposal, with probability = $\frac{p_5}{p_4}$

(5) Repeat from (1)

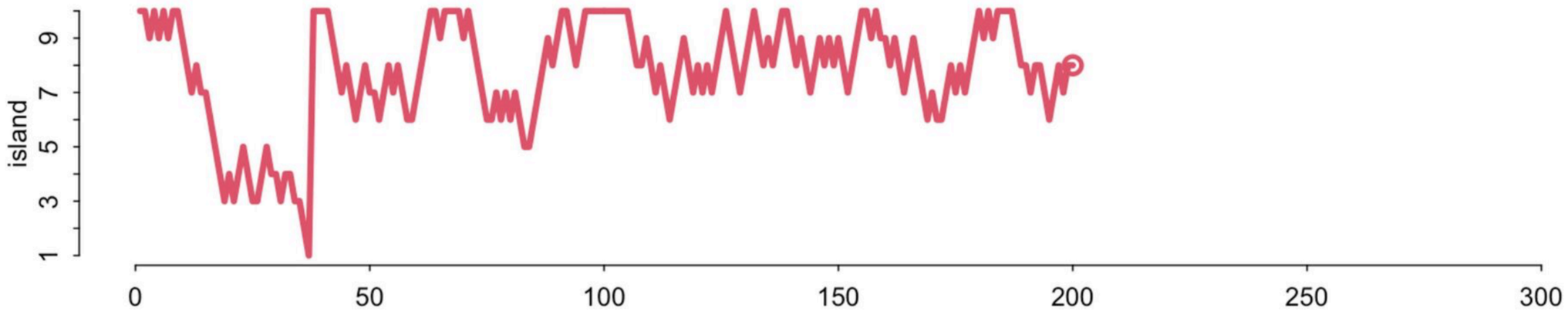
1 2 3 4 5 6 7

This procedure ensures visiting each island in proportion to its population, *in the long run*.


4
5
2
6
1
7
10
8
9


Visits
0
200
400
600
800
1
2
3
4
5
6
7
8
9
10
Island

“Markov chain Monte Carlo”



Metropolis algorithm: Simple version of *Markov chain Monte Carlo* (MCMC)

Easy to write, very general, often inefficient

A tale of Good King Markov

Markov Chains

Recap: Markov chains

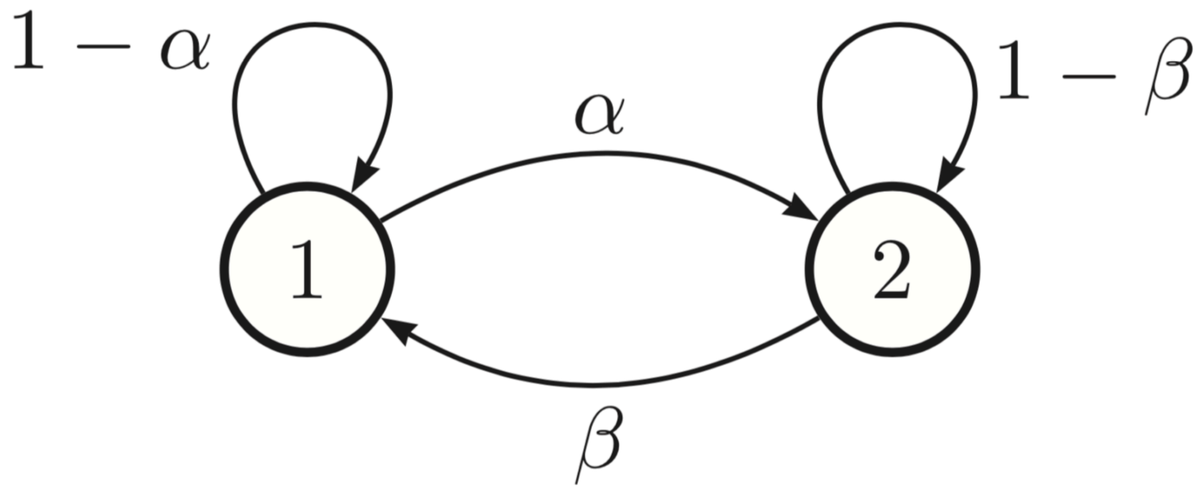
- **Markov Chains are probabilistic models for series that rely on Markov assumption**

- **Markov Assumption** for the first order Markov Chain can be written as X_t captures all relevant information to describe X_{t+1}

$$X_{t+1} | X_t \perp\!\!\!\perp X_{t-1}, X_{t-2}, \ldots, X_1$$

- **in other words:**
- $P(X_{t+1} | X_t, X_{t-1}, \ldots, X_1) = P(X_{t+1} | X_t)$

Transition Matrix



$$A = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

- **Quick check: The transition matrix after n steps looks like = …**

Markov chains. Example. Demo

- http://setosa.io/blog/2014/07/26/markov-chains/

Stationary Distribution

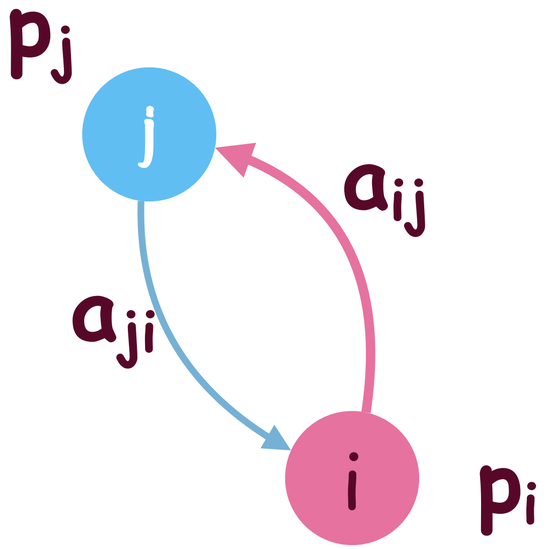
- **MCMC methods rely on the existence of stationary distribution for a Markov Chain.**

- $p^\top = p^\top A$
  - p is a stationary disrtibution
  - A is a transition matrix

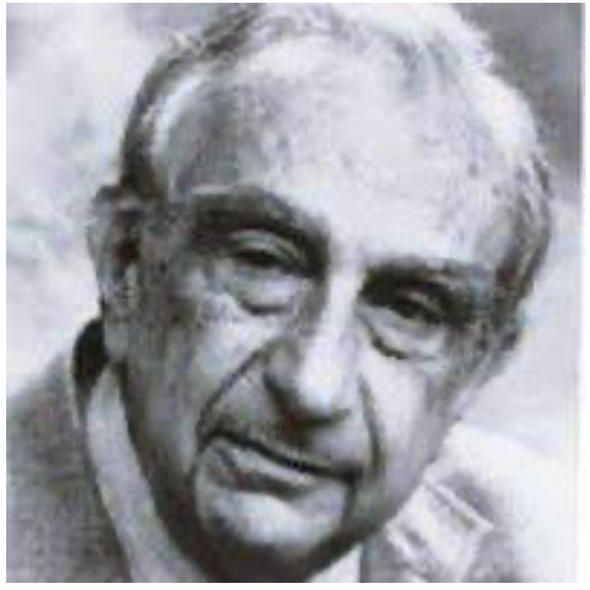
- **we can run a Markov Chain for many steps and it may converge to stationary distr.**
  - $p_0^\top A^\infty = p^\top$
  - p_0 is some random initialization

Conditions for a stationary distribution

- **necessary and sufficient conditions for stationary distribution to exist:**
  - Markov chain is irreducible (it is possible to achieve any state from any state)
  - a detailed balance condition: for all i,j:
  - $\pi_i A_{ij} = \pi_j A_{ji}$



Metropolis-Hastings



Metropolis: Random walk

- **Random Walk:** $q(x^{(j)} \mid x) = q(x \mid x^{(j)})$

$$\rho(x^*, x^{(j)}) = \min\left(1, \frac{f(x^*)}{f(x^{(j)})}\right)$$

Metropolis-Hastings

- You have no access to CDF
- You cannot use Accept-Reject (too slow)
- But if you still can calculate p.d.f. up to a proportionality constant

    - Then you can use MCMC:
        - Metropolis Sampling
        - Metropolis-Hastings Sampling
        - Gibbs Sampling
        - Hamiltonian Monte Carlo

Metropolis-Hastings: Method

- **Given some current value x(j) sample the next value x* using a proposal distribution $q(x)$:**
  - sample x* comes from $q(x \mid x^{(j)})$
- **Calculate acceptance probability:**
  - $$A(x^*, x^{(j)}) = \min\left(1, \frac{f(x^*)}{f(x^{(j)})} \frac{q(x^{(j)} \mid x^*)}{q(x^* \mid x^{(j)})}\right)$$
- **Set x(j+1) <- x* with probability**

$$A(x^*, x^{(j)})$$

- **Otherwise, set x(j+1) <- x(j)**

Metropolis-Hasting algorithm. discrete distribution

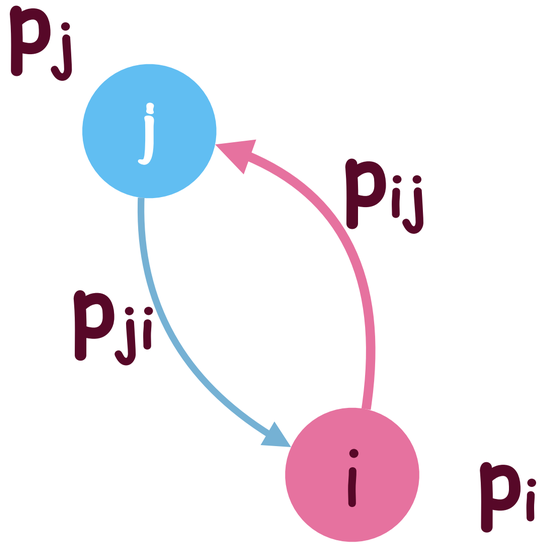
- **The Metropolis-Hasting algorithm is a general method to design a Markov chain whose stationary distribution is a given target distribution p.**
  - **Start with a connected undirected graph G on the set of states.**
- **In general, let r be the maximum degree of any vertex of G.**

Metropolis-Hasting algorithm. discrete distribution

- The transitions of the Markov chain are defined as follows:
- At state i select neighbour j with probability $\frac{1}{r}$.
- Since the degree of i may be less than r, with some probability no edge (transition) is selected and the walk remains at i.
- If a neighbour j is selected and $p_j \geq p_i$, go to j.
- If $p_j < p_i$, then

  go to j with probability $\frac{p_j}{p_i}$ and

  stay at i with probability $(1 - \frac{p_j}{p_i})$



Metropolis-Hasting algorithm

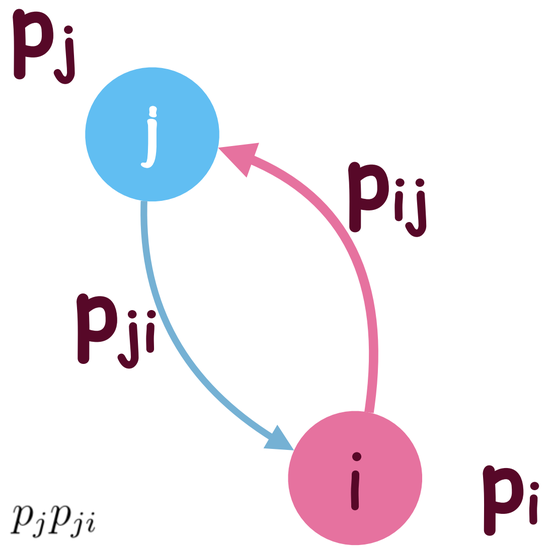
- The transitions of the Markov chain are defined as follows:
- At state i select neighbour j with probability $\frac{1}{r}$.
- Since the degree of i may be less than r, with some probability no edge (transition) is selected and the walk remains at i.
- If a neighbour j is selected and $p_j \geq p_i$, go to j.
- If $p_j < p_i$, then

  go to j with probability $\frac{p_j}{p_i}$ and

  stay at i with probability $(1 - \frac{p_j}{p_i})$

$$p_{ij} = \frac{1}{r} \min\left(1, \frac{p_j}{p_i}\right)$$

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij}$$

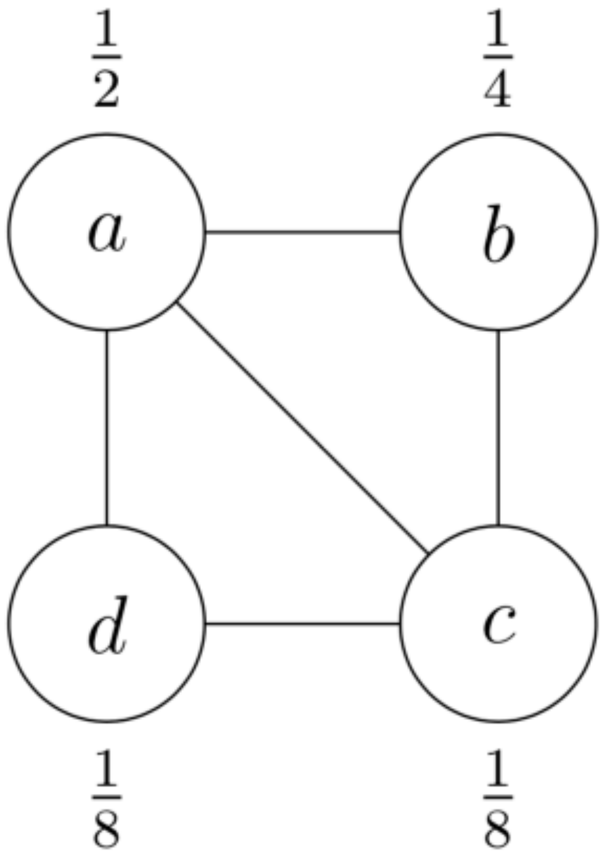


$$p_i p_{ij} = \frac{p_i}{r} \min\left(1, \frac{p_j}{p_i}\right) = \frac{1}{r} \min(p_i, p_j) = \frac{p_j}{r} \min\left(1, \frac{p_i}{p_j}\right) = p_j p_{ji}$$

Example

- **Goals:**
  - Construct the transition matrix using M-H
  - Check the detailed balance property

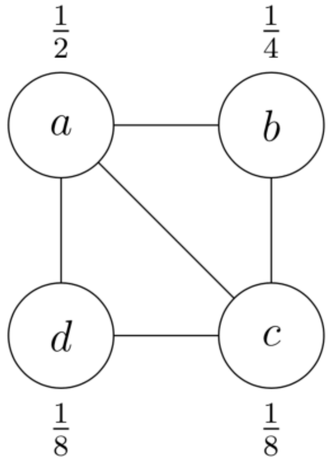
$$p(a) = \frac{1}{2}$$
$$p(b) = \frac{1}{4}$$
$$p(c) = \frac{1}{8}$$
$$p(d) = \frac{1}{8}$$



Example

- **r=3**

$p(a) = \frac{1}{2}$
$p(b) = \frac{1}{4}$
$p(c) = \frac{1}{8}$
$p(d) = \frac{1}{8}$



$a \to b \quad \frac{1}{3}\ \frac{1}{4}\ \frac{2}{1} = \frac{1}{6}$
$a \to c \quad \frac{1}{3}\ \frac{1}{8}\ \frac{2}{1} = \frac{1}{12}$
$a \to d \quad \frac{1}{3}\ \frac{1}{8}\ \frac{2}{1} = \frac{1}{12}$
$a \to a \quad 1-\frac{1}{6}-\frac{1}{12}-\frac{1}{12}=\frac{2}{3}$

$b \to a \quad \frac{1}{3}$
$b \to c \quad \frac{1}{3}\ \frac{1}{8}\ \frac{4}{1} = \frac{1}{6}$
$b \to b \quad 1-\frac{1}{3}-\frac{1}{6} = \frac{1}{2}$

$c \to a \quad \frac{1}{3}$
$c \to b \quad \frac{1}{3}$
$c \to d \quad \frac{1}{3}$
$c \to c \quad 1-\frac{1}{3}-\frac{1}{3}-\frac{1}{3} = 0$

$d \to a \quad \frac{1}{3}$
$d \to c \quad \frac{1}{3}$
$d \to d \quad 1-\frac{1}{3}-\frac{1}{3} = \frac{1}{3}$

$$p(a) = p(a)p(a \to a) + p(b)p(b \to a) + p(c)p(c \to a) + p(d)p(d \to a)$$
$$= \frac{1}{2}\ \frac{2}{3} + \frac{1}{4}\ \frac{1}{3} + \frac{1}{8}\ \frac{1}{3} + \frac{1}{8}\ \frac{1}{3} = \frac{1}{2}$$

Check the property

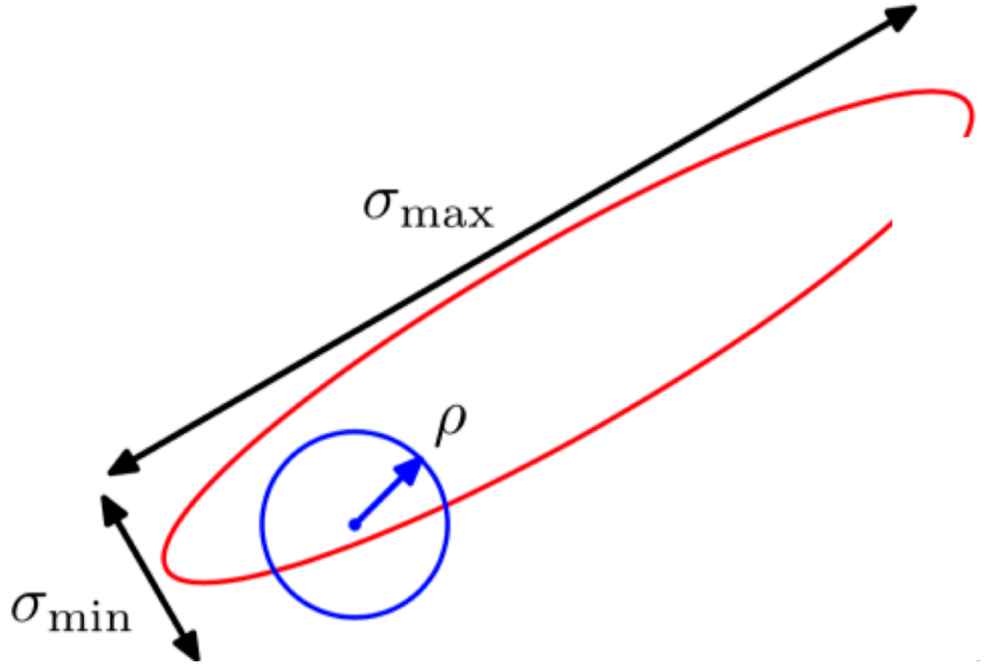
$$p(a) = p(a)p(a \to a) + p(b)p(b \to a) + p(c)p(c \to a) + p(d)p(d \to a)$$
$$= \frac{1}{2}\ \frac{2}{3} + \frac{1}{4}\ \frac{1}{3} + \frac{1}{8}\ \frac{1}{3} + \frac{1}{8}\ \frac{1}{3} = \frac{1}{2}$$

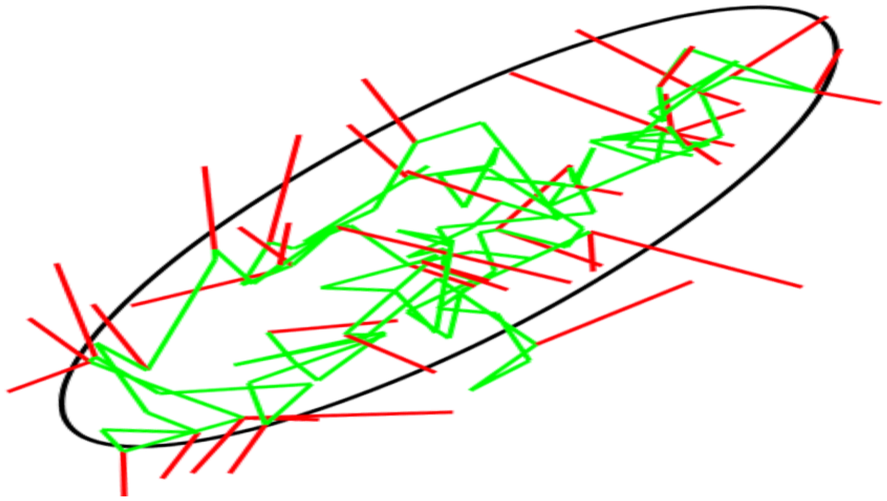
$$p(b) = p(a)p(a \to b) + p(b)p(b \to b) + p(c)p(c \to b)$$
$$= \frac{1}{2}\ \frac{1}{6} + \frac{1}{4}\ \frac{1}{2} + \frac{1}{8}\ \frac{1}{3} = \frac{1}{4}$$

$$p(c) = p(a)p(a \to c) + p(b)p(b \to c) + p(c)p(c \to c) + p(d)p(d \to c)$$
$$= \frac{1}{2}\ \frac{1}{12} + \frac{1}{4}\ \frac{1}{6} + \frac{1}{8}\ 0 + \frac{1}{8}\ \frac{1}{3} = \frac{1}{8}$$

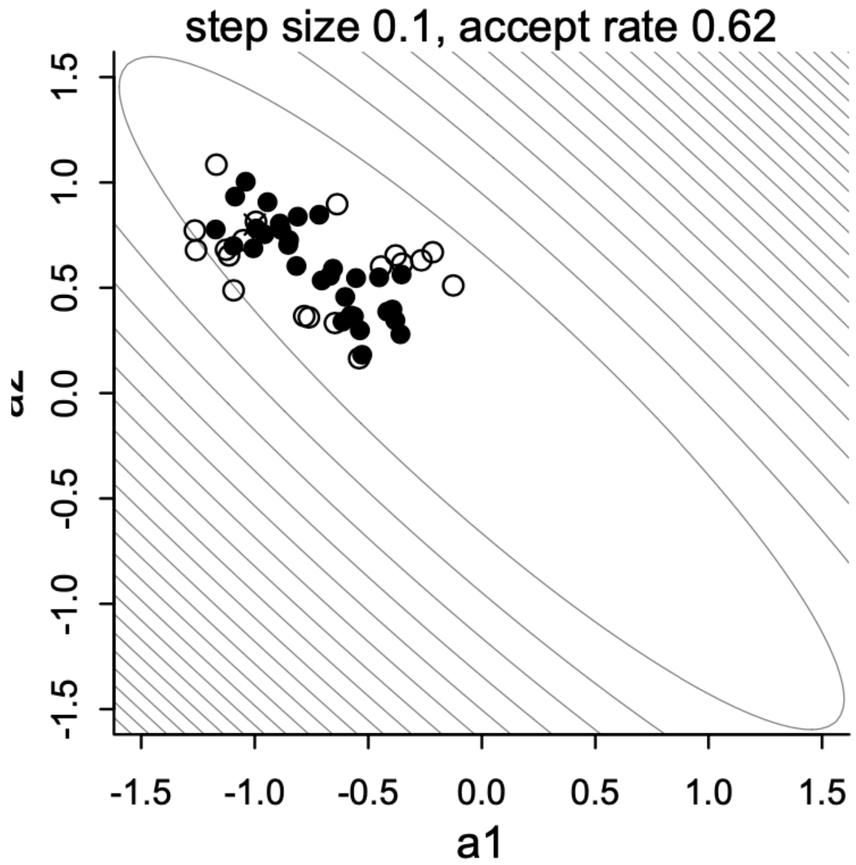
$$p(d) = p(a)p(a \to d) + p(c)p(c \to d) + p(d)p(d \to d)$$
$$= \frac{1}{2}\ \frac{1}{12} + \frac{1}{8}\ \frac{1}{3} + \frac{1}{8}\ \frac{1}{3} = \frac{1}{8}$$

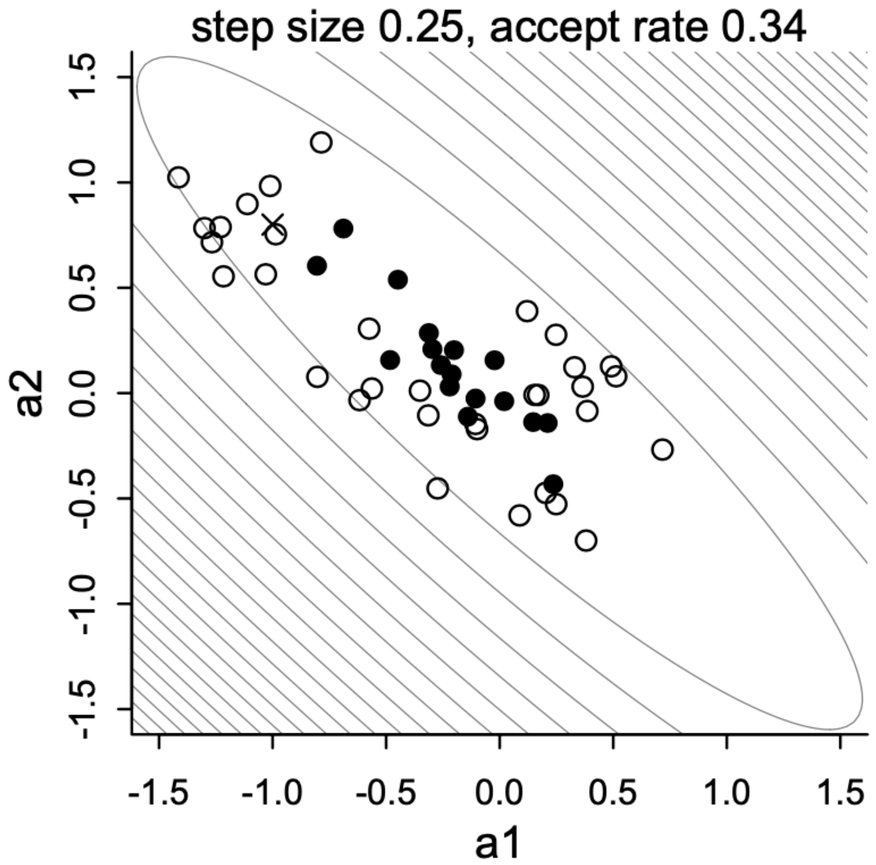
Visualization of MH: continuous case





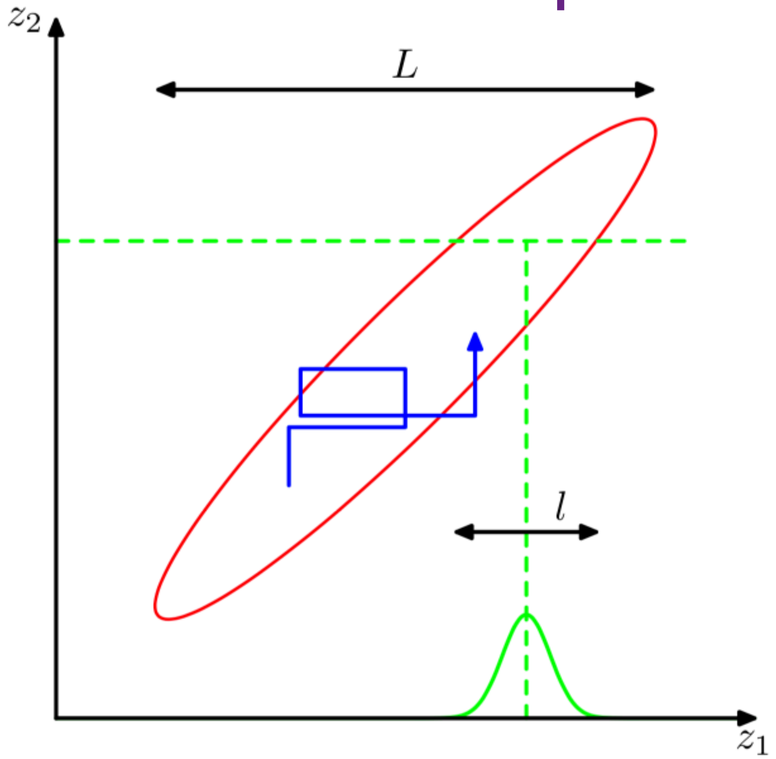
Metropolis chains under high correlation

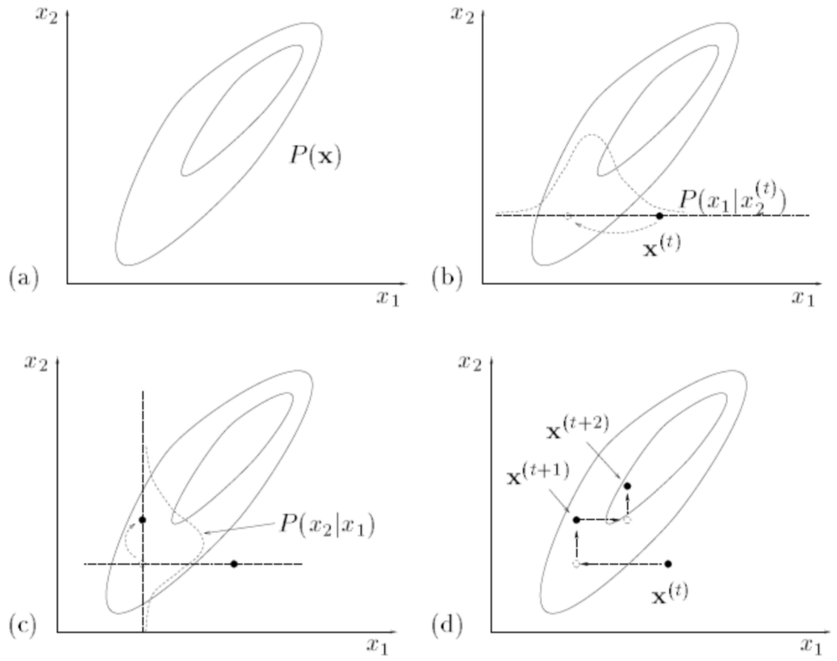




Gibbs Sampling

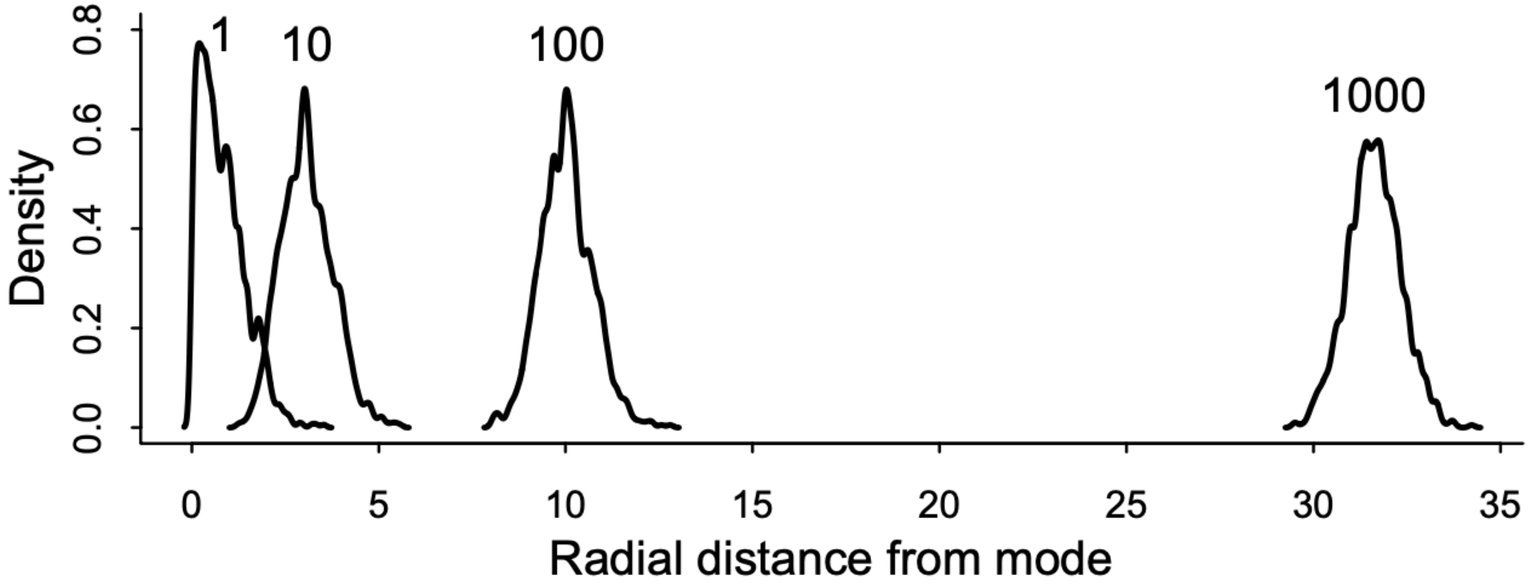
- The idea in Gibbs sampling is that, rather than probabilistically picking the next state all at once, you make a separate probabilistic choice for one of the d dimensions, where each choice depends on the other d − 1 dimensions.

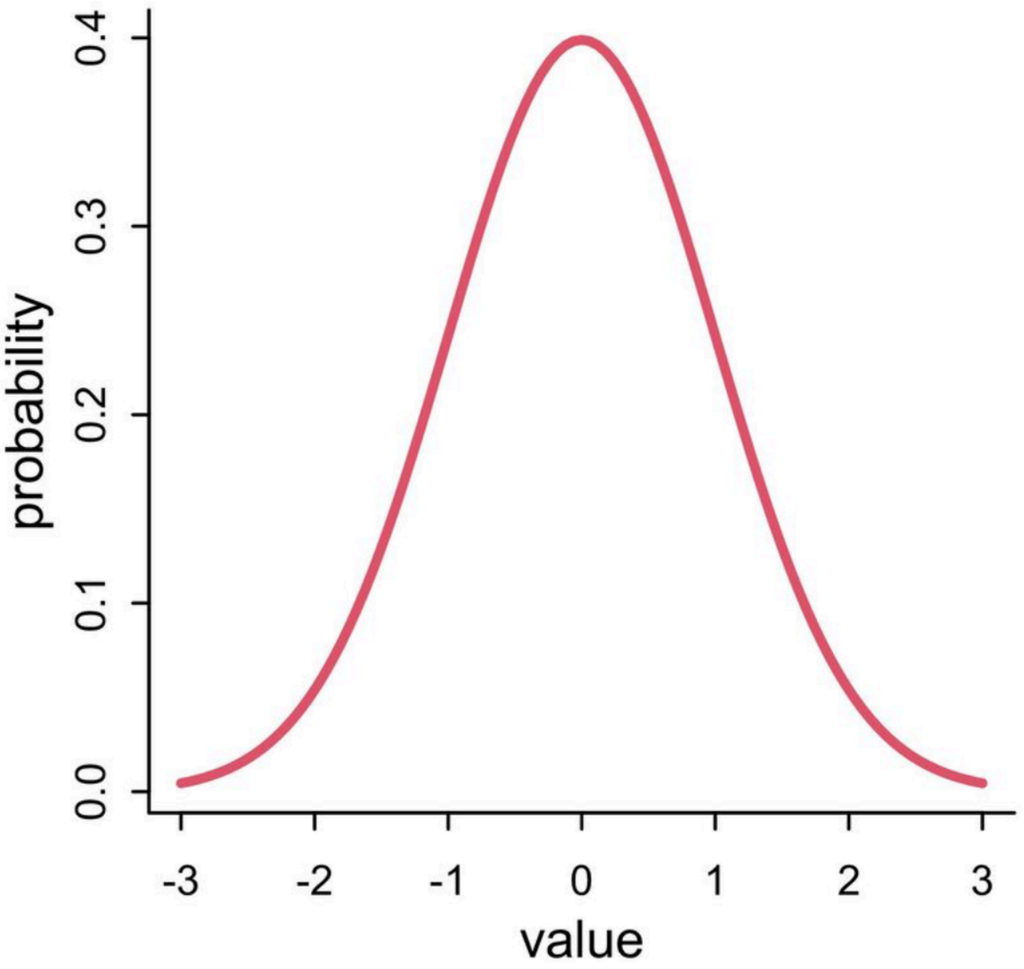


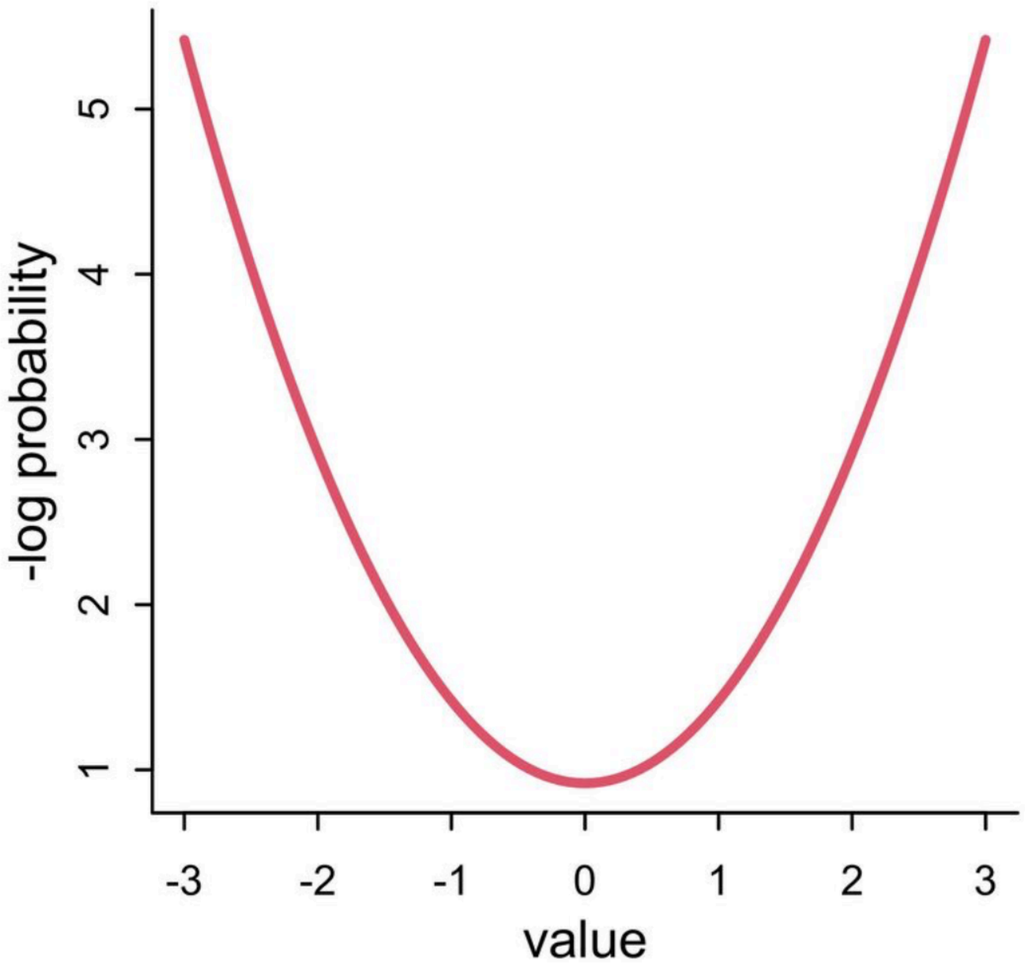


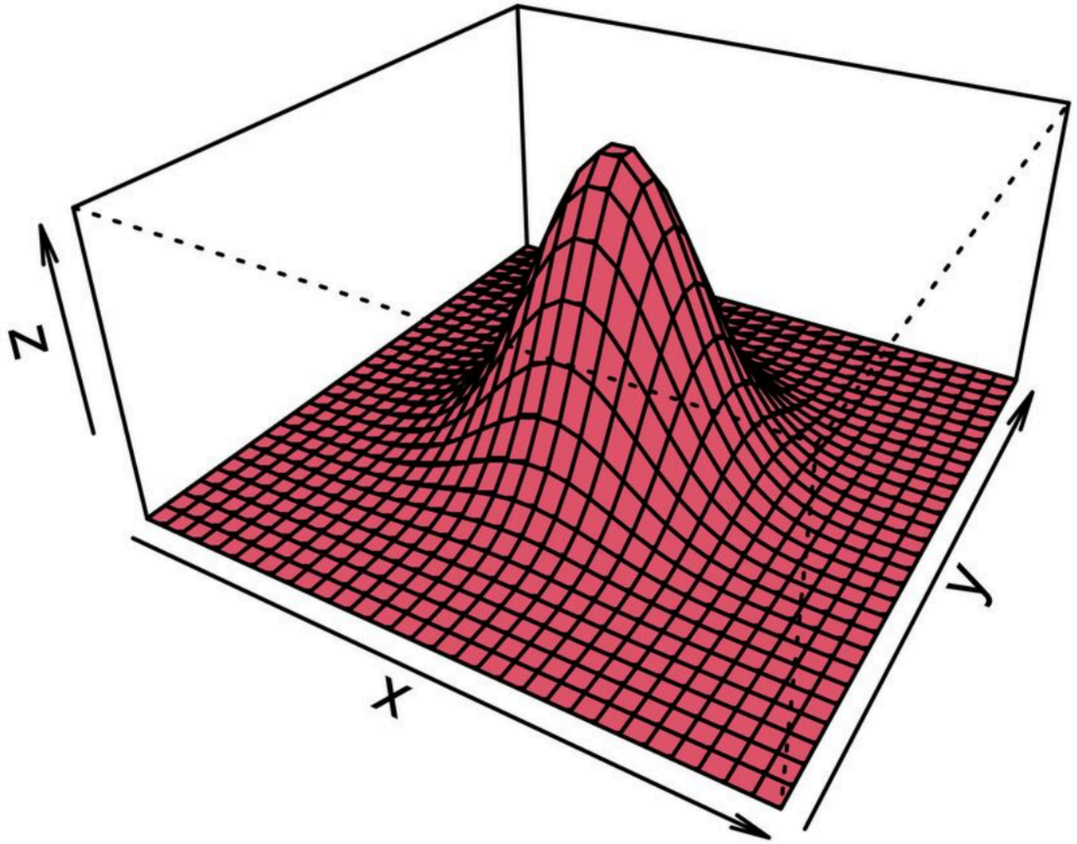
High-dimensional problems. Concentration of measure

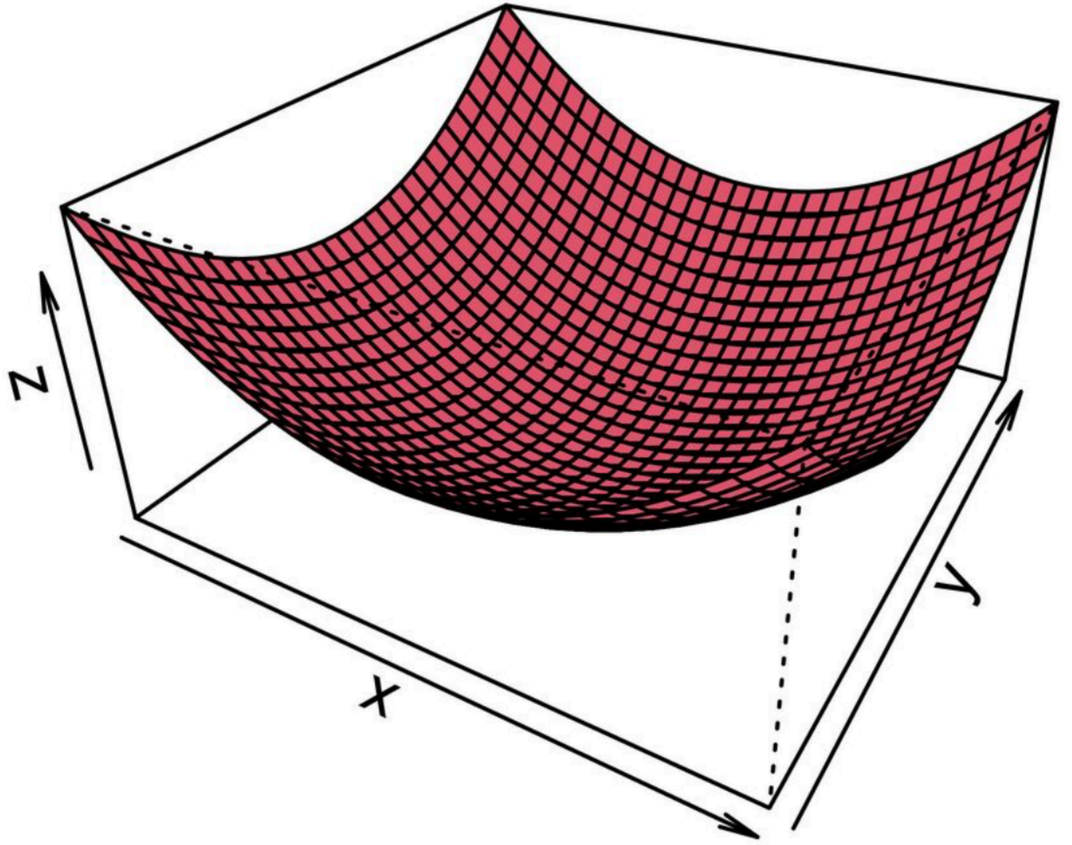
- **The most of the probability mass of a high-dimension distribution is always very far from the mode of the distribution.**
- **=> The combination of parameter values that maximizes posterior probability, the mode, is not actually in a region of parameter values that are highly plausible**




probability
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
value


-log probability
1
2
3
4
5
-3
-2
-1
0
1
2
3
value


z
x
y

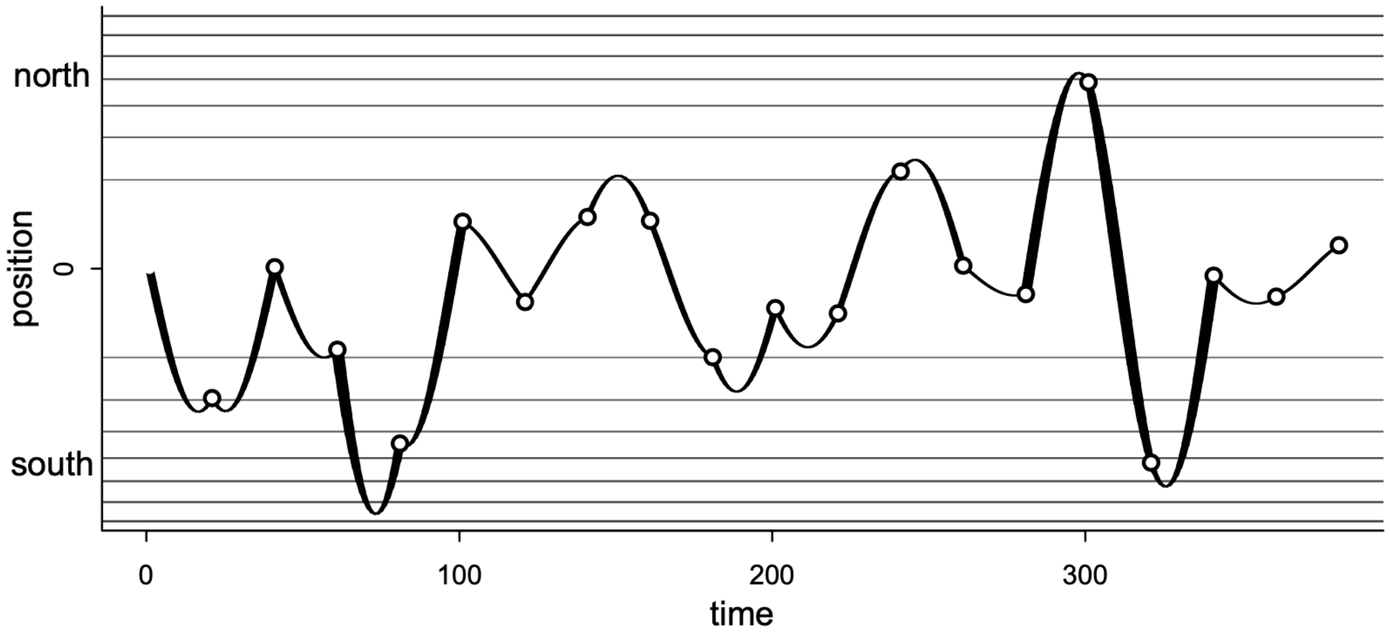

z
x
y


Low probability
High probability

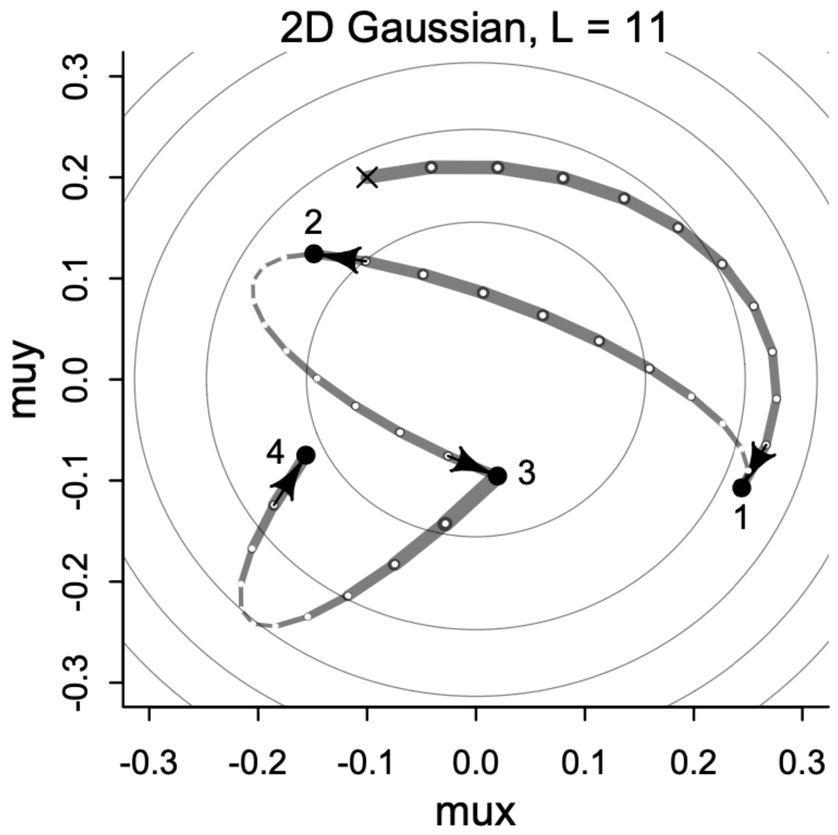
Hamiltonian Monte Carlo (Hybrid Monte Carlo, HMC)

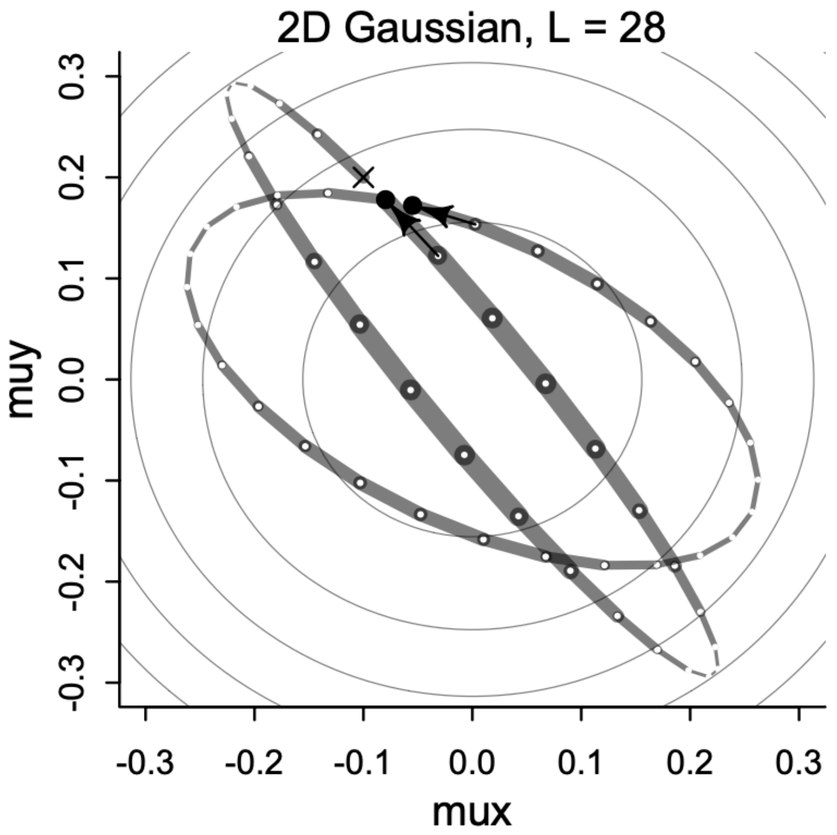
- **HMC really does run a physics simulation, pretending the vector of parameters gives the position of a little frictionless particle.**

- **The log-posterior provides a surface for this particle to glide across.**

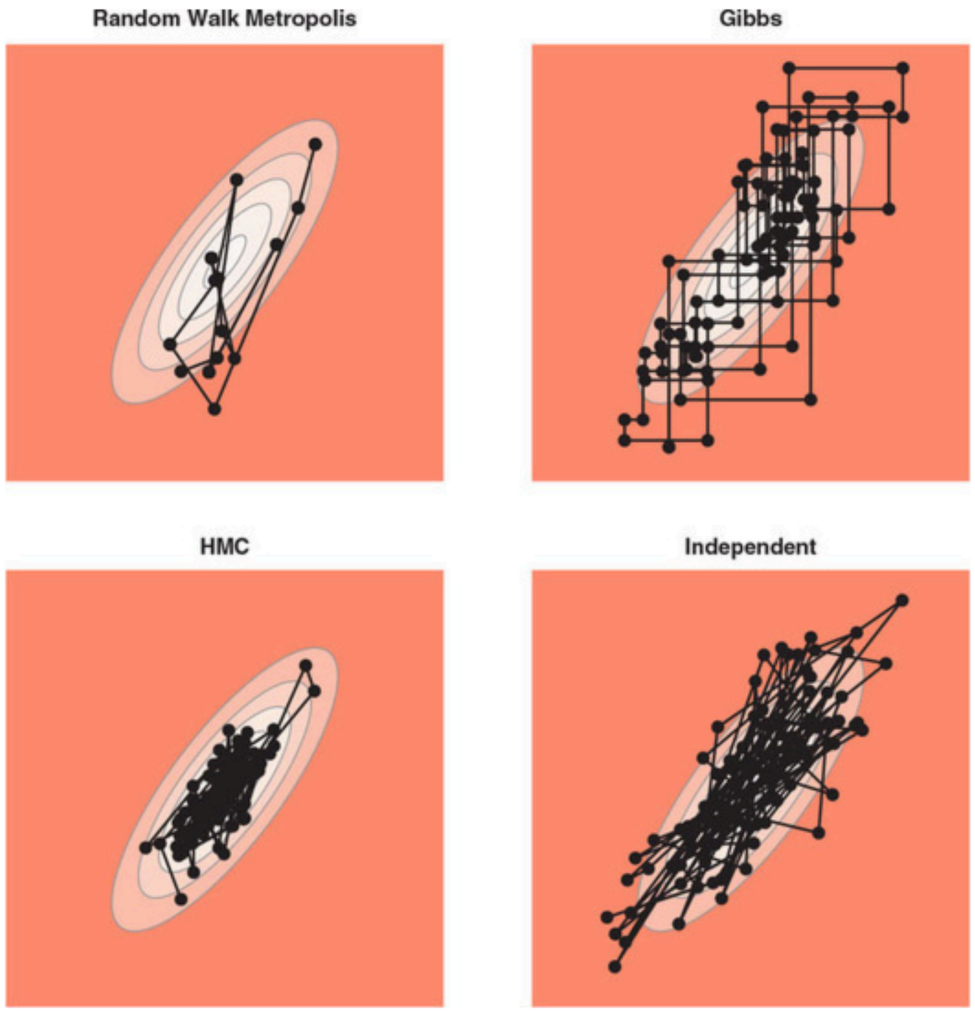


U-turns


2D Gaussian, L = 11
muy
mux
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
1
2
3
4


2D Gaussian, L = 28
muy
mux
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3

Different Sampling methods



- Lambert, Ben. A Students Guide to Bayesian Statistics SAGE Publications. 2018

Assignment: at home (optional)

- **Using Metropolis-Hastings**
  - Sample from f(x):

$$f(x) \propto \exp(-(x-3)^2/18) I(1 < x < 6)$$

  - Let q(x | x(j)) = N(3,1) ; x(0) = 2
  - Using samples, calculate the mean

References

- Probabilistic inference using Markov chain Monte Carlo methods, Radford M. Neal, Technical report: CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993. http://www.cs.toronto.edu/~radford/review.abstract.html
- Various figures and more came from (see also references therein):
  Advances in Markov chain Monte Carlo methods. Iain Murray. 2007. http://www.cs.toronto.edu/~murray/pub/07thesis/
- Information theory, inference, and learning algorithms. David MacKay, 2003. http://www.inference.phy.cam.ac.uk/mackay/itila/
- Pattern recognition and machine learning. Christopher M. Bishop. 2006. http://research.microsoft.com/~cmbishop/PRML/
- Gibbs sampling for graphical models: http://mathstat.helsinki.fi/openbugs/ http://www-ice.iarc.fr/~martyn/software/jags/
- Neural networks and other flexible models: http://www.cs.utoronto.ca/~radford/fbm.software.html
- CODA: http://www-fis.iarc.fr/coda/