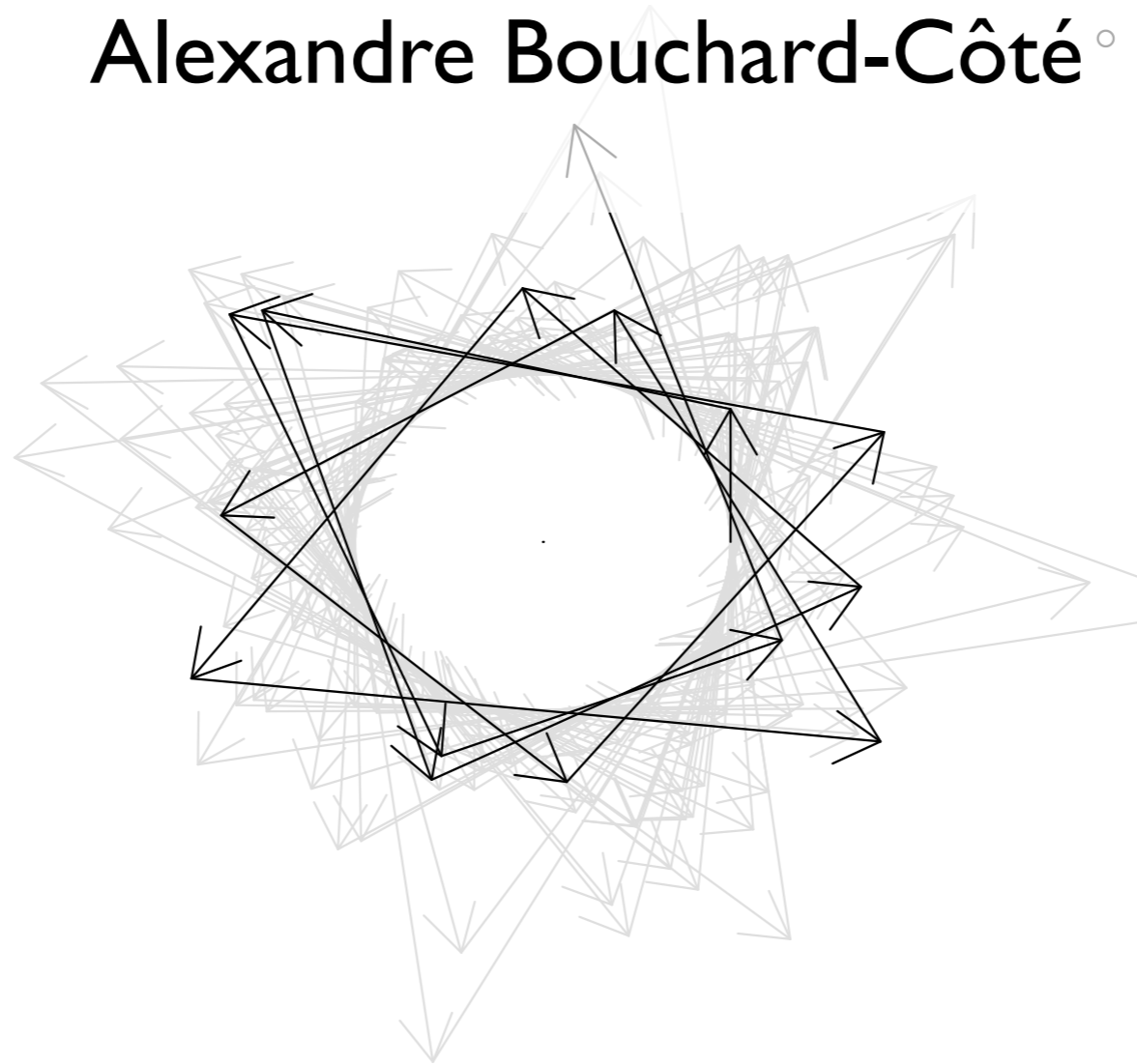


# Monte Carlo methods

Alexandre Bouchard-Côté<sup>o</sup>



o

# Project

- Due: April 26 (send code + pdf by email)
- By next Monday (March 26): send informal plan for project by email (extension possible)
- See Syllabus page on website for more info
- Encouraged to combine with your research and/or other classes

# Typical projects

- **methodology:** e.g. develop a sampler for a new datatype; or, extend an existing one to work around a practical issue
- **analysis:** using a mix of theory and experiments, benchmark a new sampler (e.g. ask me about new work on discrete state space); or, compare the scalability popular methods that have not been compared before
- **application:** e.g. using Monte Carlo on a novel type of data while demonstrating state-of-the-art practices (no just running MCMC with defaults)

# Exercise

- Construct an irreducible discrete Markov chain
- Compute a Monte Carlo average with test function = indicator on one of the states
- Try to make an educated analytical guess for the numerical value of asymptotic variance
- Approximate numerically the asymptotic variance

**Metropolis-Hastings,  
continued**

# MCMC: main ideas

- We have LLNs and CLTs for Markov chains
  - Question: how to characterize the limits? (we cannot do it with the law of an arbitrary  $X_i$  as in iid case)
  - Answer: use the stationary law instead
- **We can design and simulate Markov chains with a prescribed stationary distribution  $\pi$** 
  - Even if we do not know the normalization of  $\pi$

# MH: algorithmic description of the new kernel (denoted $K$ or $T$ )

- 1 Sample  $X^* \sim q(\cdot | X^{(t-1)})$ .
- 2 Compute

Do we need the normalization of  $\pi$ ?  $q$ ?

$$\alpha(X^* | X^{(t-1)}) = \min \left( 1, \frac{\pi(X^*) q(X^{(t-1)} | X^*)}{\pi(X^{(t-1)}) q(X^* | X^{(t-1)})} \right)$$

- 3 Sample  $U \sim \mathcal{U}_{[0,1]}$ . If  $U \leq \alpha(X^* | X^{(t-1)})$ , set  $X^{(t)} = X^*$ , otherwise set  $X^{(t)} = X^{(t-1)}$ .

# MCMC: a naive way to use MH (many alternatives exist!)

- Pick arbitrary initial  $X_1$
- Simulate a Markov chain  $X_1, X_2, X_3, \dots$
- Use the samples to compute the MC average

$$\frac{1}{t} \sum_{i=1}^t \phi \left( X^{(i)} \right)$$

- Note: there will be duplicates in this sum (why?)
- By the LLN for Markov chains, this will get arbitrarily close to the integral of interest



# Examples

- Most frequent choice (simple but usually not best/most efficient!): pick  $q(x, x') = \text{normal density}$ ...
  - centered at  $x$ : random walk metropolis
    - What is the acceptance probability?
  - biased by gradient: Langevin (later)
- Restriction to a neighborhood:  
 $q(x, x') = \pi(x') 1[x' \in N(x)] / Z(x)$
- If  $x' \in N(x) \Leftrightarrow x \in N(x')$  this is (generalized) Gibbs sampling.
  - What is the acceptance probability?
  - Why 'generalized'? For Gibbs  $Z(x) = Z(x')$ ; but this is not always true in general (why?)

## 1. PESKUN'S THEOREM

Let  $X$  be a discrete random variable following distribution  $\pi$ , and let  $P$  be the transition matrix of a Markov chain with  $\pi$  as its invariant distribution. We call  $P$  reversible if

$$\pi(x)P(x, y) = \pi(y)P(y, x).$$

Following Peskun (1973), we define  $P_2 \geq P_1$  for any two transition matrices if each of the off-diagonal elements of  $P_2$  is greater than or equal to the corresponding off-diagonal elements of  $P_1$ . The following lemma is Theorem 2.1.1 of Peskun (1973).

**LEMMA 1.1.** *Suppose each of the irreducible transition matrices  $P_1$  and  $P_2$  is reversible for the same invariant probability distribution  $\pi$ . If  $P_2 \geq P_1$  then, for any  $f$ ,*

$$v(f, \pi, P_1) \geq v(f, \pi, P_2), \tag{1}$$

where

$$v(f, \pi, P) = \lim_{N \rightarrow \infty} N \text{ var}(\hat{I}_N),$$

and  $\hat{I}_N = \sum_{t=1}^N f\{X^{(t)}\}/N$  is an estimator of  $I = E_{\pi}(f)$  using  $N$  consecutive samples from the Markov chains. Kemeny & Snell (1969, p. 84) gave an expression for  $v(f, \pi, P)$  in terms of  $f$ ,  $P$  and  $\pi$ .

# Example, continued

- Let's look at the details of a Gibbs sampler
- How to sample from Markov Random Field?

# Motivation

**Task:** given some images (a 2D array of pixels), segment it into clusters of pixels

In general, there is an unknown number of clusters, so we will apply nonparametric priors, but for now, assume there are only 'background' and 'people' clusters



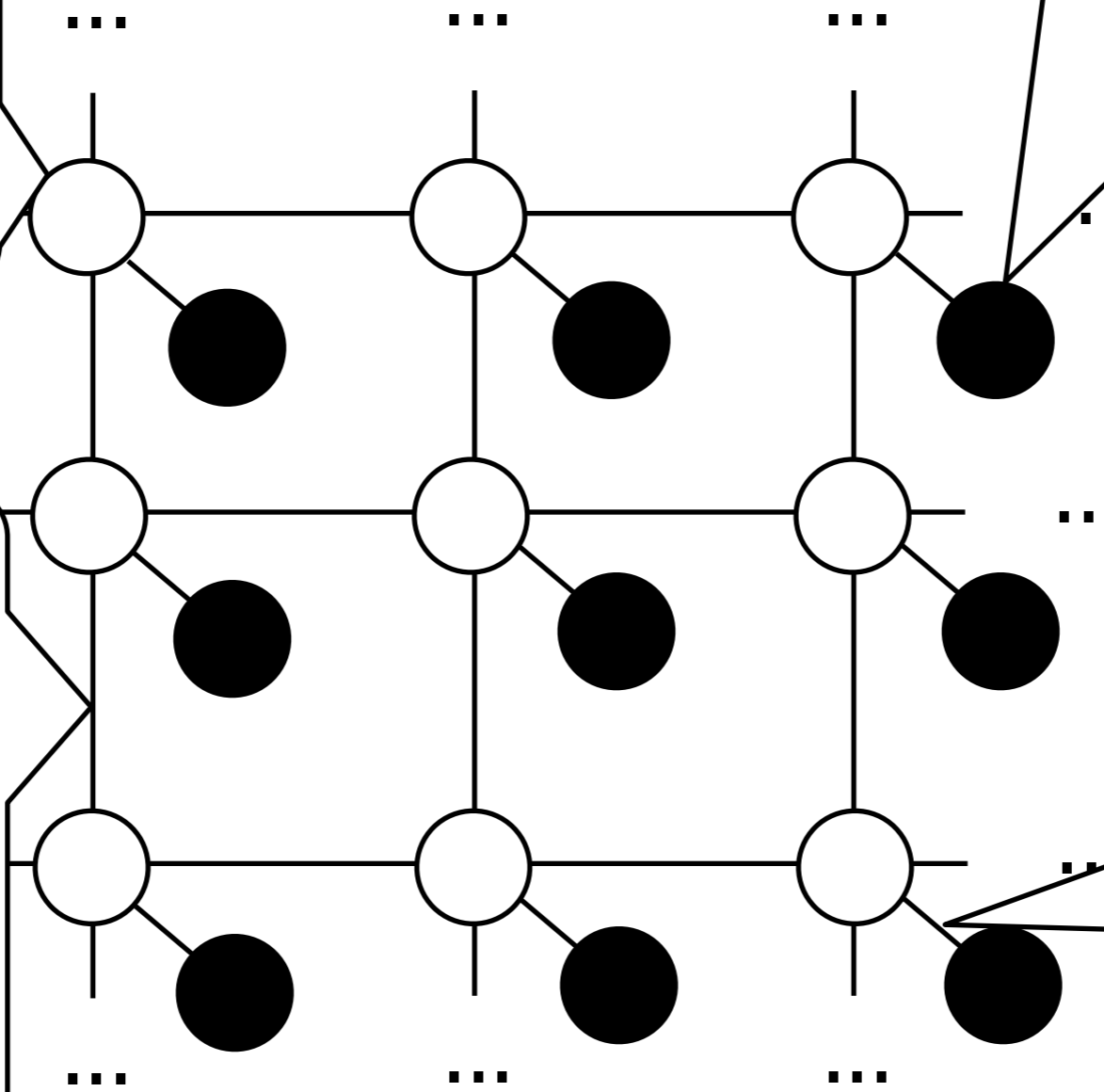
# Model for image segmentation

Is this pixel part of 'background' (B) or 'people' (P) ?

RGB value of the pixel

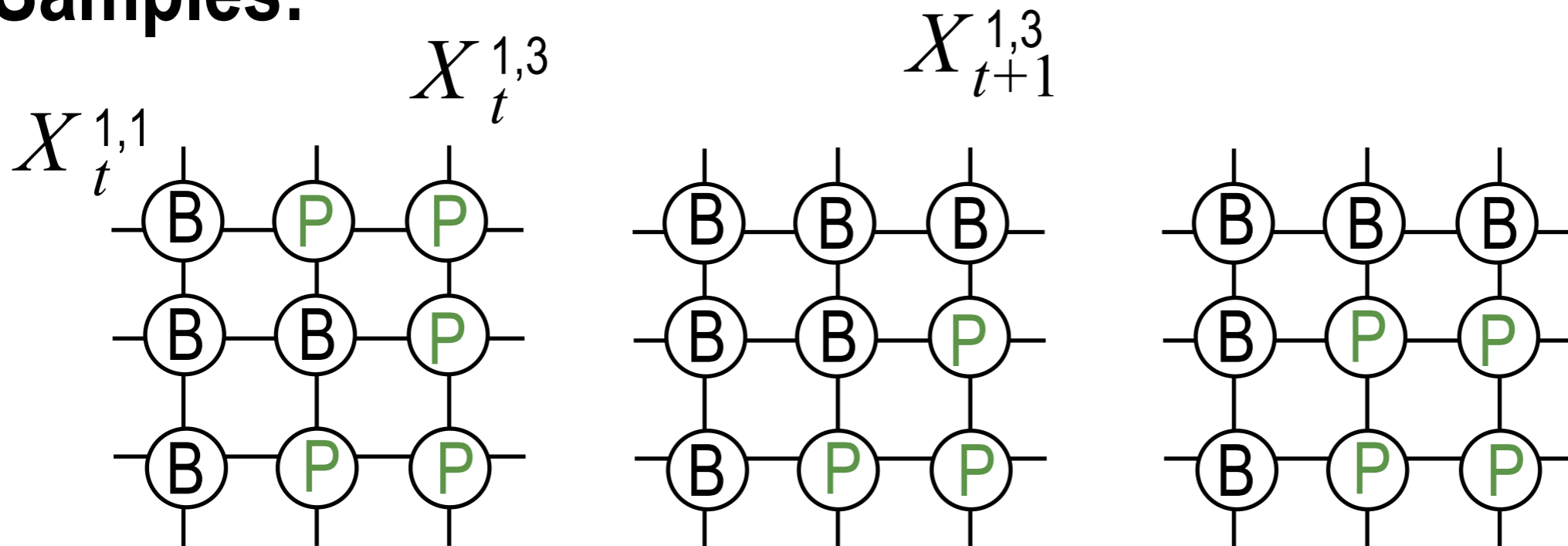
Potentials to encourage adjacent cluster indicators to have the same value, i.e. if  $x \neq x'$   
 $f(x, x) > f(x, x')$

For each cluster, there will be a different distribution over pixel colors



# Computing the posterior

Samples:

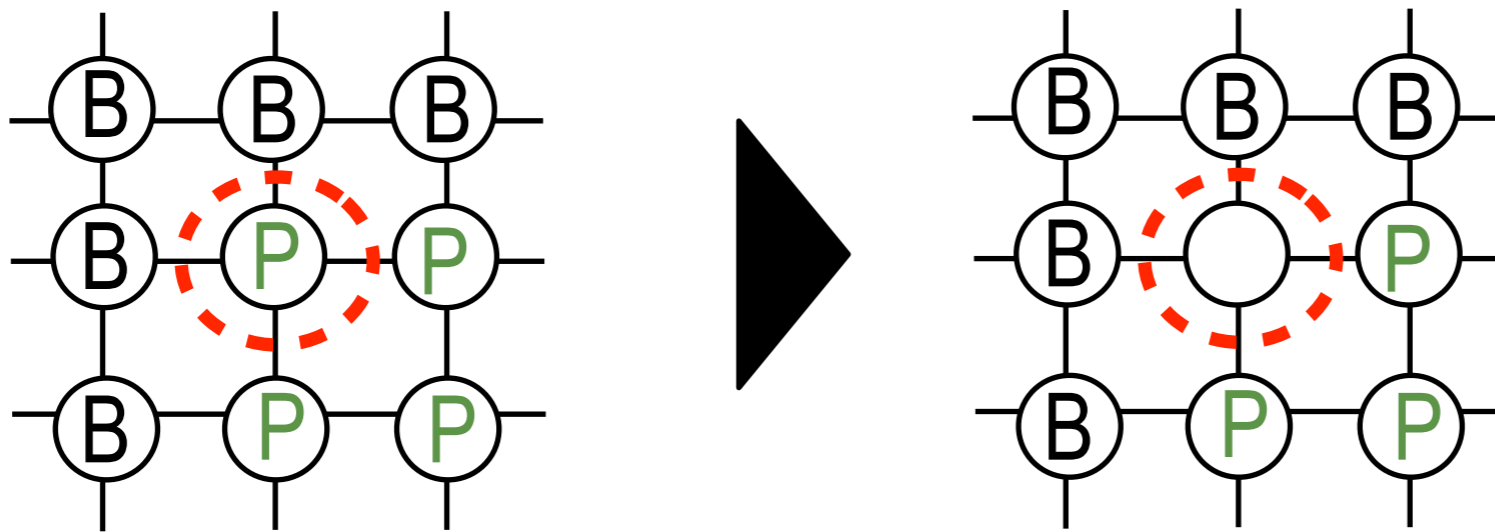


Monte Carlo estimator: for  $S$  samples

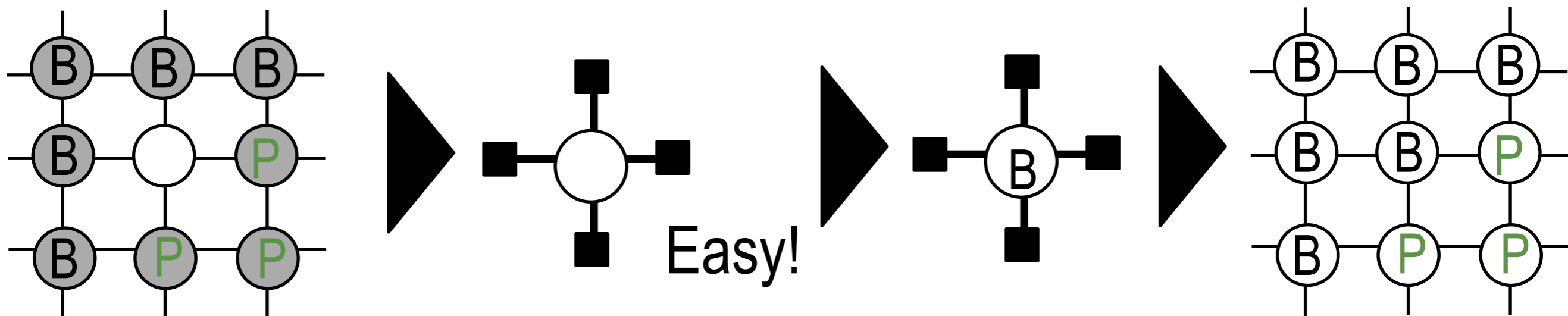
$$\mathbb{E}f(X) \approx \frac{1}{S} \sum_{t=1}^S f(X_t)$$

# 'Naive' Gibbs sampling

**Loop:** pick one node  $(i,j)$  at random, erase the contents of the guessed values in  $(i,j)$ , freeze the value of the other nodes

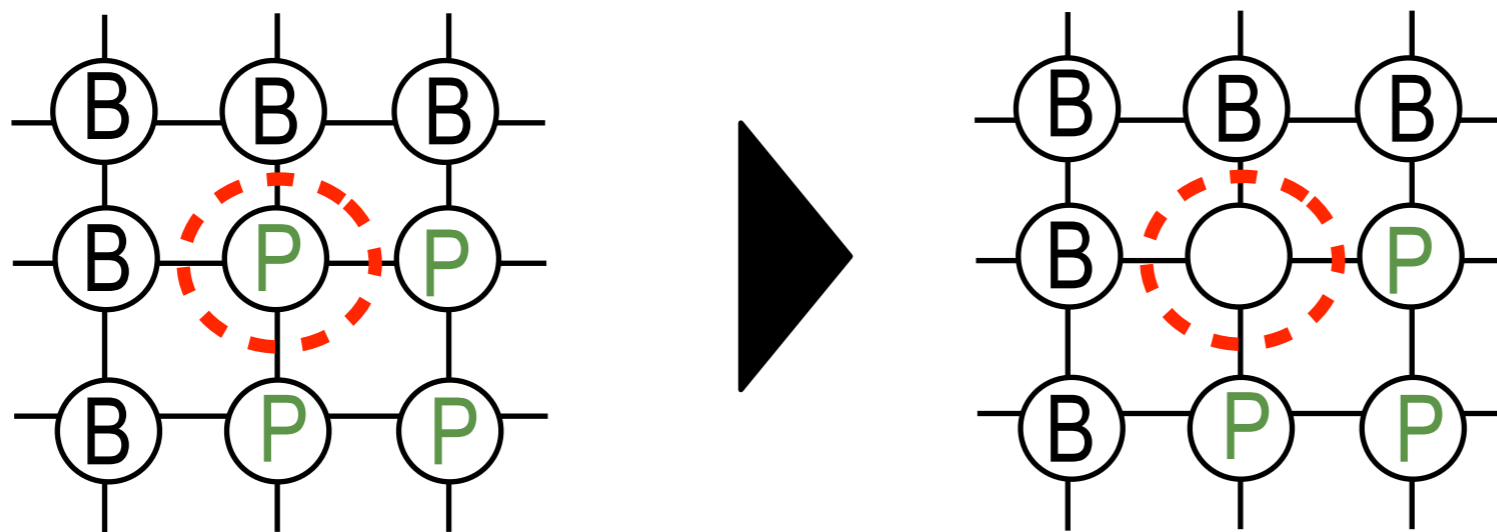


**Then:** resample a value for the node  $(i,j)$  conditioning on all the others, and write this to the current state at  $(i,j)$



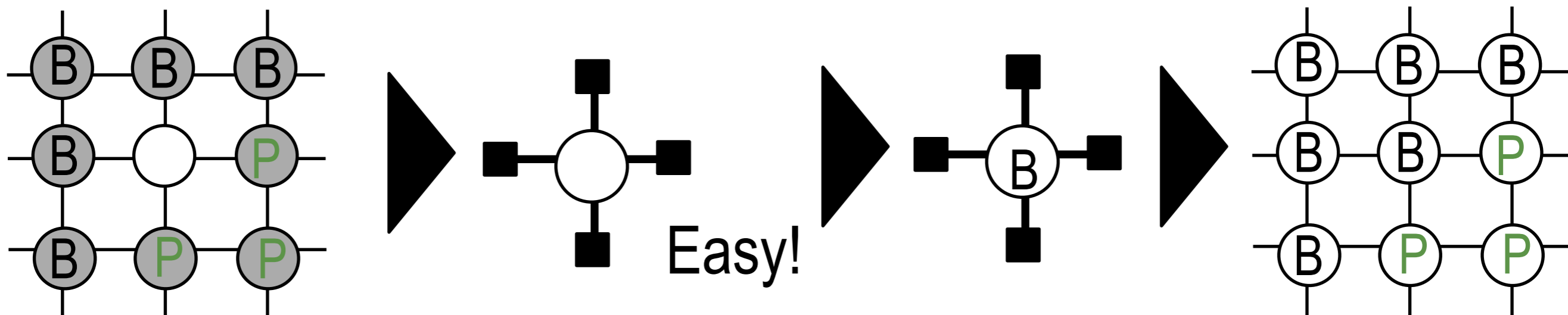
# Starting with a simpler version

**Loop:** pick one node  $(i,j)$  at ~~random~~, erase the contents of the guessed values in  $(i,j)$ , freeze the value of the other nodes



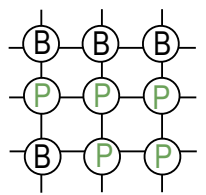
Fix sampling to  
node  $(2,2)$   
Will relax this  
later

**Then:** resample a value for the node  $(i,j)$  conditioning on all the others, and write this to the current state at  $(i,j)$



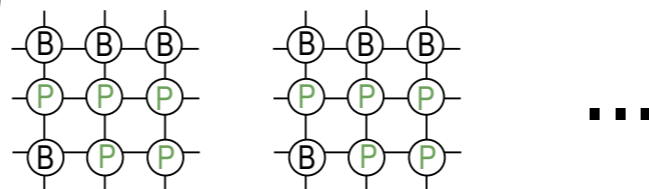


# Transition matrix of the Gibbs sampler: $2^9 \times 2^9$ matrix

$$T = \begin{bmatrix} 0.1 & 0.01 & \dots \\ 0.01 & & \\ \dots & & \end{bmatrix}$$


A diagram showing a 3x3 grid of nodes. The top row contains three white circles labeled 'B'. The middle row contains three green circles labeled 'P'. The bottom row contains a white circle labeled 'B', a green circle labeled 'P', and another green circle labeled 'P'. Each node is connected to its four immediate neighbors (up, down, left, right) by thin black lines.

Way too large to represent in memory but we will compute entries on the fly



# Neighborhood example, continued

- Interesting point: we not need  $x \in N(x)$ 
  - In fact, for discrete space removing  $x$  from  $N(x)$  provably decrease the asymptotic variance (Peskun, 1973)
  - Example: an MCMC sampler with asymptotic variance lower than iid sampling
- Trade-off
  - computation can go from  $O(d)$  per sample to  $O(1)$
  - asymptotic variance typically increases-most serious in highly correlated situations

Often need several kernels to get irreducibility (and hence a CLT)

**Solution 1:** mixing kernels. Suppose we have one Gibbs kernel for each variable  $T^{(1)}, \dots, T^{(9)}$ . Then the mixture of them is also reversible (by linearity)

$$T = \sum_{k=1}^9 \alpha_k T^{(k)}$$

**Solution 2:** alternating kernels deterministically (ie. using the first, then second, etc).

$$T_{x,y} = \sum_{x_1} \cdots \sum_{x_9} T_{x,x_1}^{(1)} T_{x_1,x_2}^{(2)} \cdots T_{x_8,x'}^{(9)}$$

**Often works better:** shuffle then alternate

# Stopping criteria

- Many diagnostic exist
- All have limitations
- Some are dubious
- Best approach is CLT (with same caveats as IS):  
for a 95% confidence interval, use

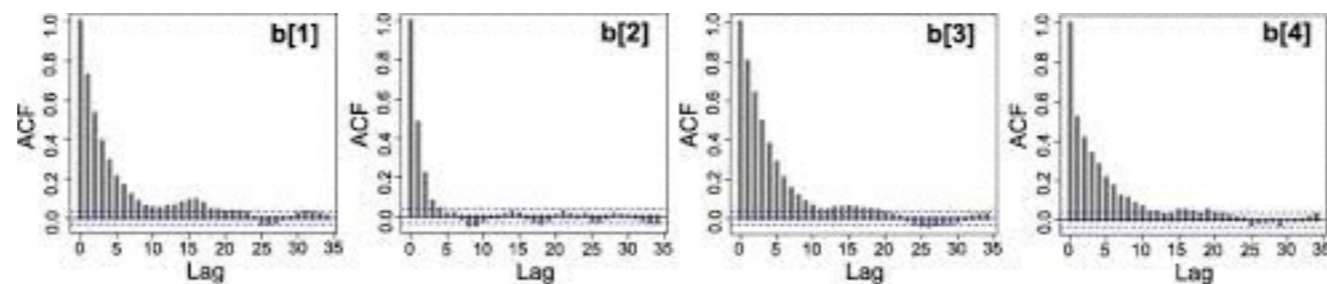
$$I_n \pm 1.96 \sqrt{\sigma_{\text{asympt}}^2 / n}$$

- The asymptotic variance is:

$$\sigma^2(\phi) = \mathbb{V}_\pi[\phi(X_1)] + 2 \sum_{k=2}^{\infty} \text{Cov}_\pi[\phi(X_1), \phi(X_k)].$$

# Estimation of the asymptotic variance

- Direct method: estimate the auto-correlations (ACF)



- Can be done quickly with FFT
- But estimator has infinite variance! Need to truncate. Still unstable in many practical scenarios.

# ESS for MC

- Idea is similar as for IS, but still tied to a test function:

$$\frac{\text{ESS}(N)}{N} \rightarrow \frac{\sigma_{\text{asymptotic}}^2}{\sigma_{\text{iid}}^2}$$

- Can estimate using ACF as in last slide
- Better method: batch estimators.
  - Segment the MCMC trace into chunks of length  $\sqrt{n}$
  - Assume sampler is good enough so that behaviour across blocks is nearly iid
- Standard metric in MCMC literature to compare samplers: ESS per second or ESS per operation

# Analysis of MH

# Plan

- First, analyse one kernel at the time to show it is  $\pi$ -invariant
- Then, show mixture/alternation is also  $\pi$ -invariant AND irreducible
- Conclude LLN holds
- In discrete case, also get CLT, in infinite space, need more (geometric ergodicity)



# Important, overlooked <sup>Def 47</sup> condition on proposal $q$

- Mutual absolute continuity condition:

$$\int_A \pi(dx) q(x, B) > 0 \Leftrightarrow \int_B \pi(dx) q(x, A) > 0$$

- For example, in a discrete state space where the target has full support, this means:

$$q(x, y) > 0 \Leftrightarrow q(y, x) > 0$$

- This can be tricky in combinatorial spaces  
(more on that soon)

# Invariance of a single kernel

- **Lemma.** The transition kernel of the Metropolis-Hastings algorithm is given by

$$K(y | x) \equiv K(x, y) = \alpha(y | x)q(y | x) + (1 - \alpha(x | x))\delta_x(y)$$

where  $\delta_x$  denotes the Dirac mass at  $x$ .

- *Proof.* We have

$$K(x, y) = \int q(x^* | x) \{ \alpha(x^* | x) \delta_{x^*}(y) + (1 - \alpha(x^* | x)) \delta_x(y) \} dx^*$$

# Invariance of a single kernel

- **Lemma.** The transition kernel of the Metropolis-Hastings algorithm is given by

$$K(y | x) \equiv K(x, y) = \alpha(y | x)q(y | x) + (1 - \alpha(x | x))\delta_x(y)$$

where  $\delta_x$  denotes the Dirac mass at  $x$ .

- *Proof.* We have

$$\begin{aligned} K(x, y) &= \int q(x^* | x) \{ \alpha(x^* | x) \delta_{x^*}(y) + (1 - \alpha(x^* | x)) \delta_x(y) \} dx^* \\ &= q(y | x) \alpha(y | x) + \left\{ \int q(x^* | x) (1 - \alpha(x^* | x)) dx^* \right\} \delta_x(y) \end{aligned}$$

# Invariance of a single kernel

- **Lemma.** The transition kernel of the Metropolis-Hastings algorithm is given by

$$K(y | x) \equiv K(x, y) = \alpha(y | x)q(y | x) + (1 - a(x))\delta_x(y)$$

where  $\delta_x$  denotes the Dirac mass at  $x$ .

- *Proof.* We have

$$\begin{aligned} K(x, y) &= \int q(x^* | x) \{ \alpha(x^* | x) \delta_{x^*}(y) + (1 - \alpha(x^* | x)) \delta_x(y) \} dx^* \\ &= q(y | x) \alpha(y | x) + \left\{ \int q(x^* | x) (1 - \alpha(x^* | x)) dx^* \right\} \delta_x(y) \\ &= q(y | x) \alpha(y | x) + \left\{ 1 - \int q(x^* | x) \alpha(x^* | x) dx^* \right\} \delta_x(y) \\ &= q(y | x) \alpha(y | x) + \{ 1 - a(x) \} \delta_x(y) \end{aligned}$$

# Invariance of a single kernel

- **Proposition.** The Metropolis-Hastings kernel  $K$  is  $\pi$ -reversible and thus admit  $\pi$  as invariant distribution.
- *Proof.* For any  $x, y \in \mathbb{X}$ , with  $x \neq y$

$$\pi(x)K(x, y)$$

$$= \pi(y)K(x, y)$$

# Invariance of a single kernel

- **Proposition.** The Metropolis-Hastings kernel  $K$  is  $\pi$ -reversible and thus admit  $\pi$  as invariant distribution.
- *Proof.* For any  $x, y \in \mathbb{X}$ , with  $x \neq y$

$$\pi(x)K(x, y) = \pi(x)q(y | x)\alpha(y | x)$$

$$= \pi(y)K(x, y)$$

# Invariance of a single kernel

- **Proposition.** The Metropolis-Hastings kernel  $K$  is  $\pi$ -reversible and thus admit  $\pi$  as invariant distribution.
- *Proof.* For any  $x, y \in \mathbb{X}$ , with  $x \neq y$

$$\begin{aligned}\pi(x)K(x, y) &= \pi(x)q(y | x)\alpha(y | x) \\ &= \pi(x)q(y | x)\min\left(1, \frac{\pi(y)q(x | y)}{\pi(x)q(y | x)}\right)\end{aligned}$$

$$= \pi(y)K(x, y)$$

# Invariance of a single kernel

- **Proposition.** The Metropolis-Hastings kernel  $K$  is  $\pi$ -reversible and thus admit  $\pi$  as invariant distribution.
- *Proof.* For any  $x, y \in \mathbb{X}$ , with  $x \neq y$

$$\begin{aligned}\pi(x)K(x, y) &= \pi(x)q(y | x)\alpha(y | x) \\ &= \pi(x)q(y | x)\min\left(1, \frac{\pi(y)q(x | y)}{\pi(x)q(y | x)}\right) \\ &= \min(\pi(x)q(y | x), \pi(y)q(x | y)) \\ &= \pi(y)K(x, y)\end{aligned}$$



# Invariance of a single kernel

- **Proposition.** The Metropolis-Hastings kernel  $K$  is  $\pi$ -reversible and thus admit  $\pi$  as invariant distribution.
- *Proof.* For any  $x, y \in \mathbb{X}$ , with  $x \neq y$

$$\begin{aligned}\pi(x)K(x, y) &= \pi(x)q(y | x)\alpha(y | x) \\ &= \pi(x)q(y | x)\min\left(1, \frac{\pi(y)q(x | y)}{\pi(x)q(y | x)}\right) \\ &= \min(\pi(x)q(y | x), \pi(y)q(x | y)) \\ &= \pi(y)q(x | y)\min\left(\frac{\pi(x)q(y | x)}{\pi(y)q(x | y)}, 1\right) \\ &= \pi(y)K(x, y)\end{aligned}$$

# Exercise

- If we have a collections of  $\pi$ -invariant kernels..
- then their mixture is  $\pi$ -invariant as long as the mixture coefficients do *not* depend on the states
- similarly for deterministic alternations
- hence, mixtures of alternations are also  $\pi$ -invariant

# Exercise

- Derive (theoretically, for now) a *valid* MCMC algorithm for one of the following problems:
  - sampling uniformly from perfect bi-partite graph matchings
  - sampling uniformly from unrooted bifurcating phylogenetic trees
  - sampling uniformly from multiple sequence alignments
  - sampling uniformly from another combinatorial structure of your choice

# Ingenious MCMC constructions

# Terminology

**Collapsed sampler:** analytically marginalize some of the variables, and run MCMC on the *reduced* state space

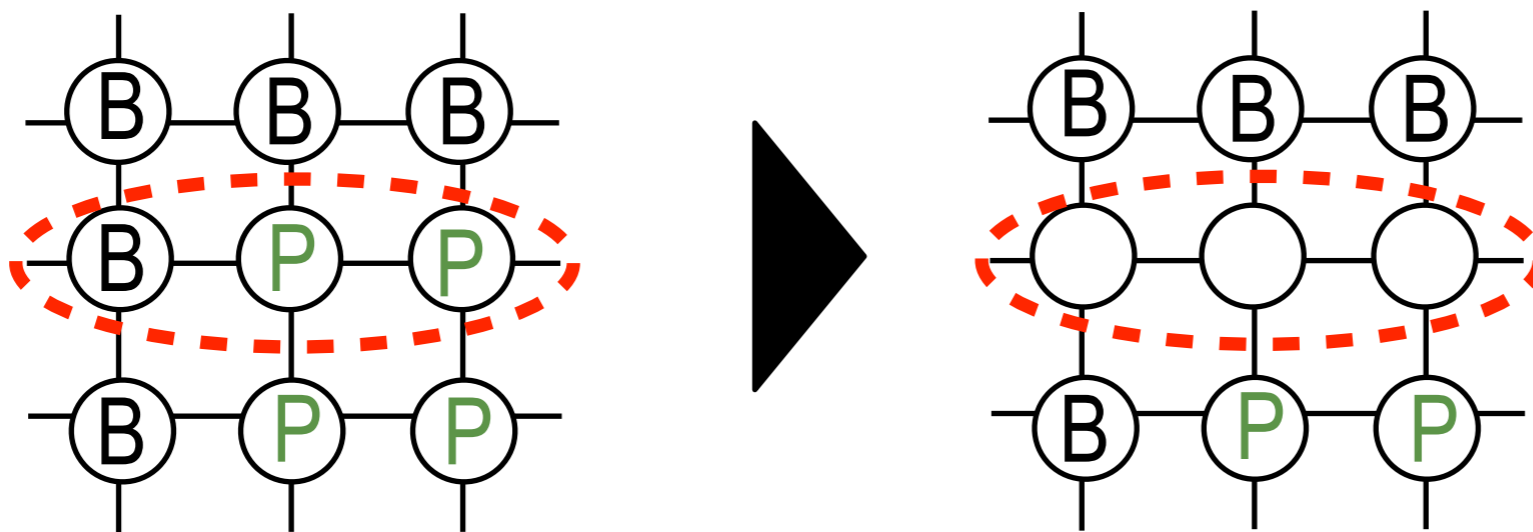
**Example:** *HMM global parameter inference while summing over latent dynamic states*

**Auxiliary variable:** *augment* the state space to facilitate sampling

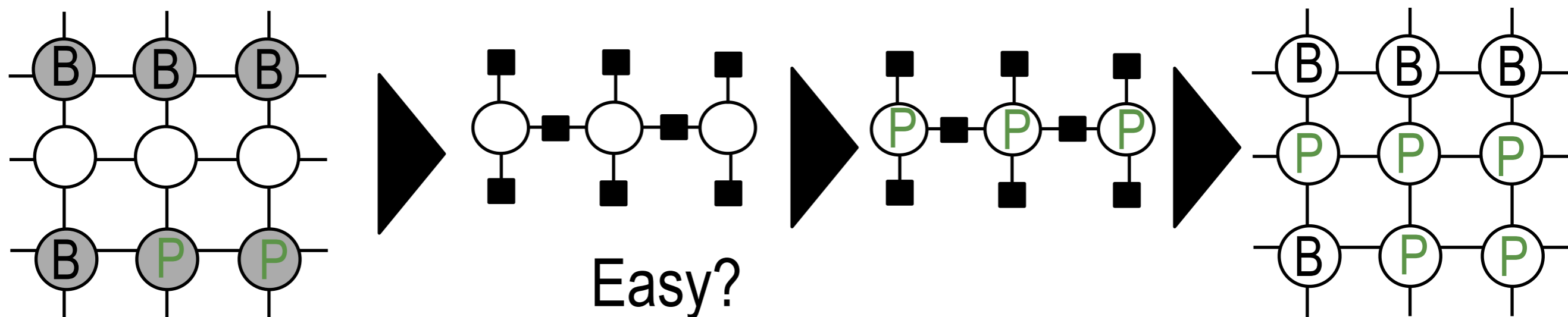
**Example:** *slice sampling*

# Another collapsed example: Collapsed Gibbs samplers

**Loop:** pick a subset of nodes  $N$  at random, erase the contents of the guessed values in  $N$ , freeze the value of the nodes not in  $N$



**Then:** resample a value for the nodes in  $N$  conditioning on all the others, and write this to the current state at  $N$



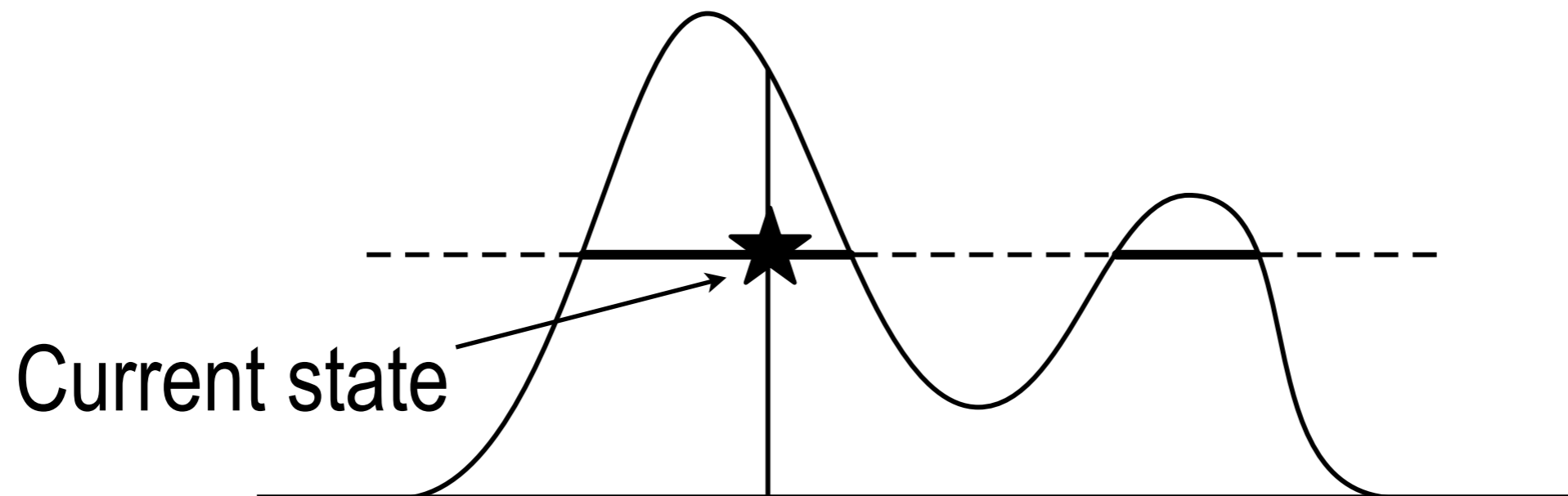
Easy?

# Slice sampling

**Goal:** sampling from a r.v.  $X$  with density  $f(x)/Z$ , where  $Z$  is difficult to compute

**Intuition:** use a MCMC defined on the 2D space defined as the graph of the density

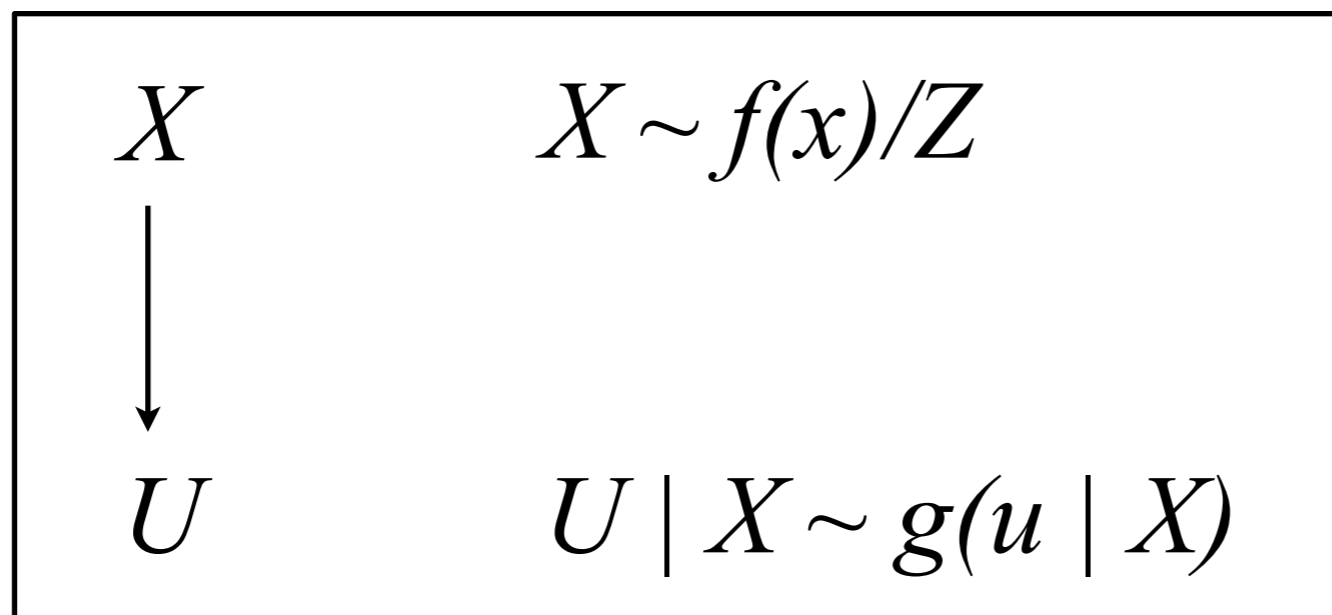
**Moves:** sample uniformly vertically or horizontally



# Slice sampling

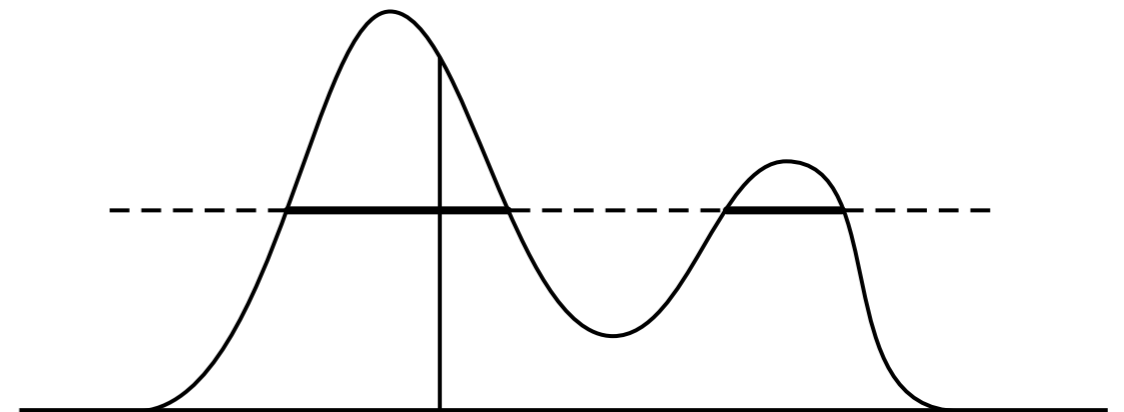
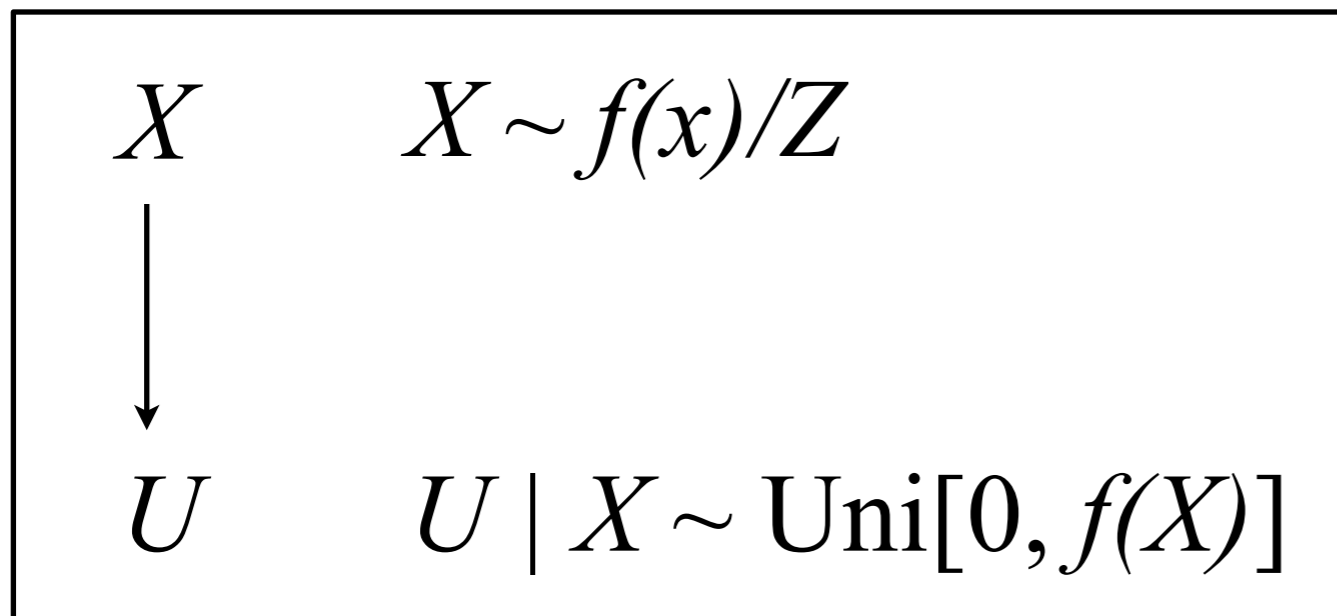
**Goal:** sampling from a r.v.  $X$  with density  $f(x)/Z$ , where  $Z$  is difficult to compute

**General auxiliary variable construction:** adding a new random variable  $U$  with the following graphical model does not change the marginal distribution of  $X$ , no matter what is the conditional density  $g$  of  $U | X$





# Slice sampler



**Vertical move:**  $U | X \sim \text{Uni}[0, f(X)]$

**Horizontal move:**  $X | U \sim \text{Uni}\{x : f(x) \geq U\}$

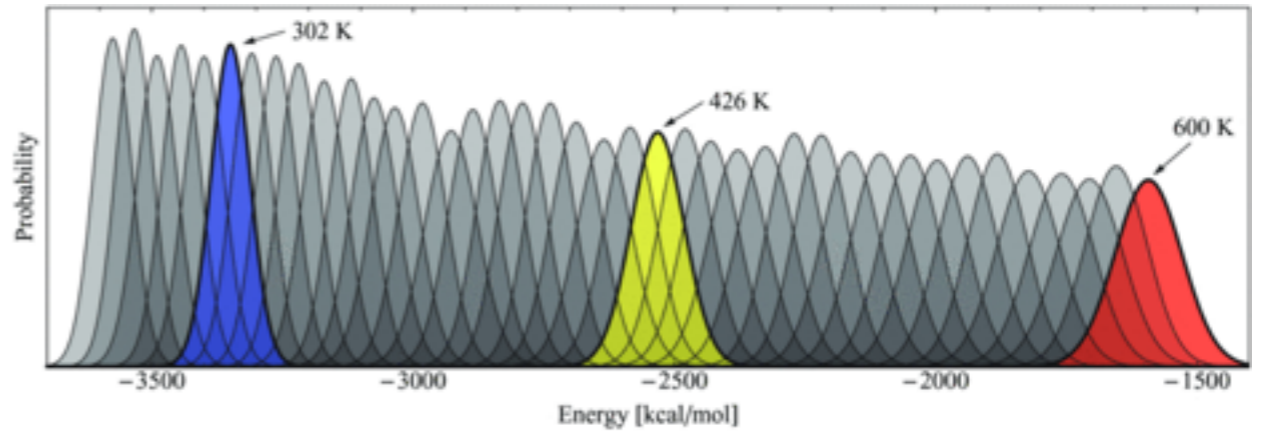
**Note:** Efficient alternatives to the horizontal move exist  
(state-of-the-art: doubling+shrinking procedure, Neal, 2003)

**See** <https://github.com/UBC-Stat-ML/blangSDK/blob/master/src/main/java/blang/mcmc/RealSliceSampler.java>

# Annealing and tempering

- Key idea: using sequences of distributions
  - Denoted, for  $t = 0 \dots l$ ,  $\pi_t$
- The case  $t = 0$  should be easy
  - ideally, such that we can get exact sample in poly-time
- The case  $t = l$  should coincide with the target of interest

# Sequences of $\pi$



- Examples
- Naive: exponentiate the whole target
  - Problem: we don't want non-normalizable targets
  - Solution: Exponentiate only likelihood
- Other issues
  - restrictions in the likelihood
  - computation: interpolate number of datapoints [Project]
- Automatic creation of sequences of distributions in Blang:  
[https://www.stat.ubc.ca/~bouchard/blang/Inference\\_and\\_runtime.html](https://www.stat.ubc.ca/~bouchard/blang/Inference_and_runtime.html)
- Sparsity considerations (changing  $t$  should be  $O(1)$ )

# Annealing

- Make temperature random
- Extract samples when  $t = 1$
- Problem?