Statistical modeling with stochastic processes

Alexandre Bouchard-Côté

Lecture 3, Monday March 7

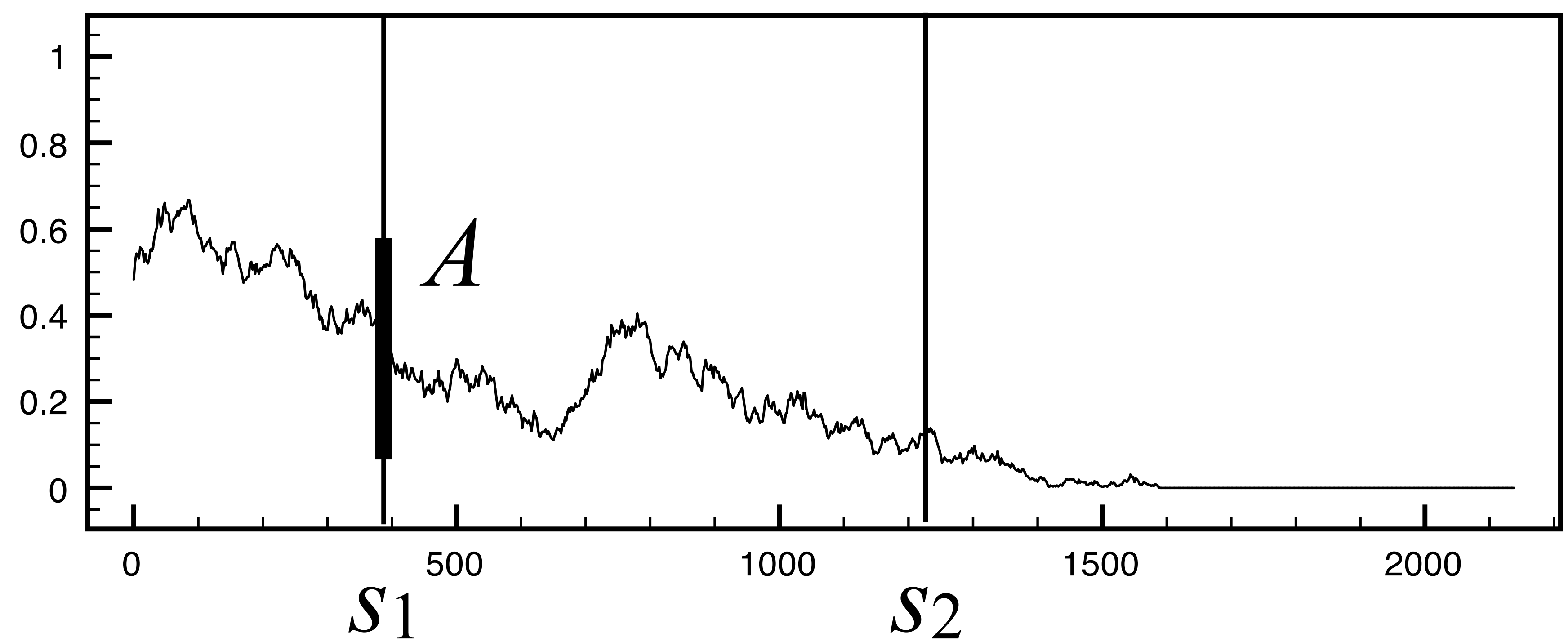
Plan for today

- Exact inference review
- Approximate inference, part I: MCMC
    - Gibbs
    - Metropolis-Hastings
    - Overview of theoretical results available
    - Tricks of the trade

Review

Why do we know the marginals? By definition!

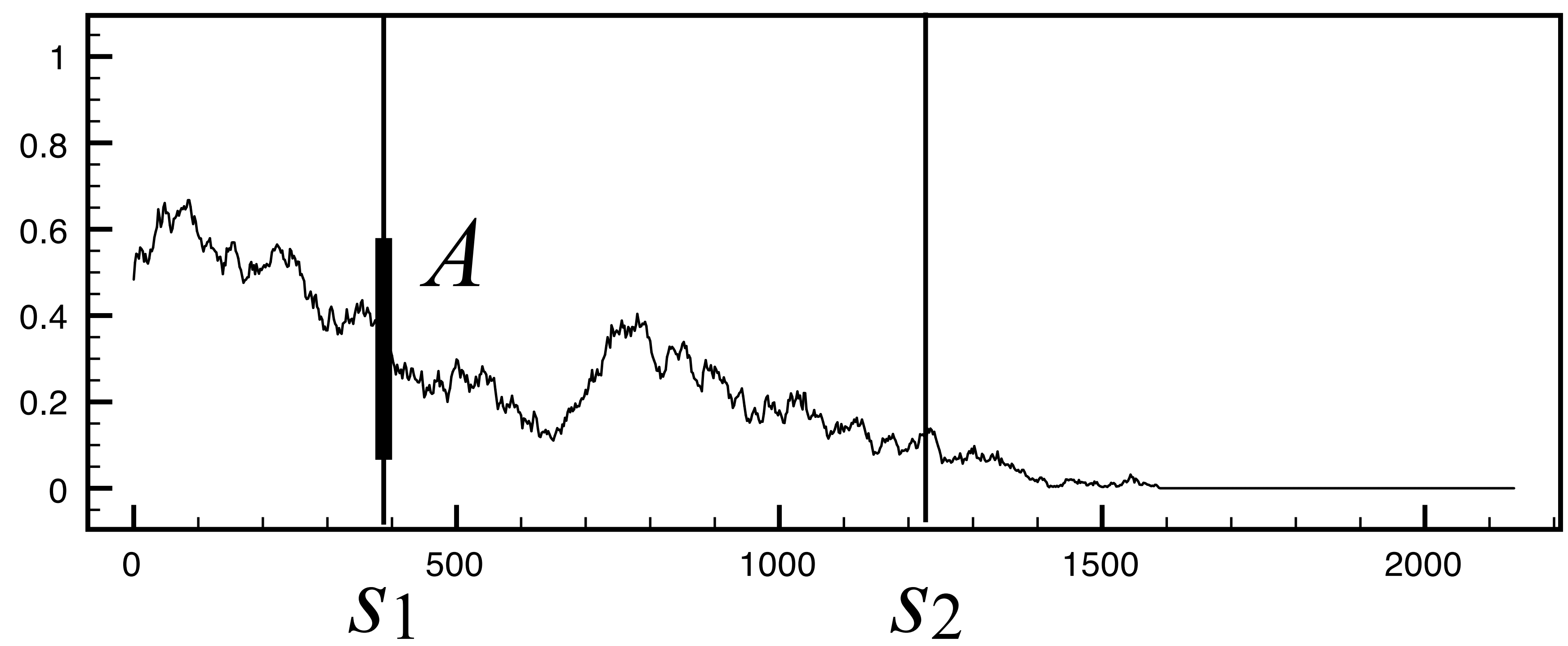
What are the bare minimum conditions for λ to be marginals of Y_s ? I.e. we want $\lambda_s(A) = P(Y_s \in A)$, etc



Why do we know the marginals? By definition!

What are the bare minimum conditions for λ to be marginals of Y_s ? I.e. we want $\lambda_s(A) = P(Y_s \in A)$, etc

$$\boxed{\begin{array}{ll} \lambda_{s_1}(A) = \lambda_{s_1,s_2}(A, \mathbf{R}) & \text{[marginalization]} \\ \lambda_{s_1,s_2}(A_1, A_2) = \lambda_{s_2,s_1}(A_2, A_1) & \text{[perm]} \end{array}}$$



Why do we know the marginals? By definition!

What are the bare minimum conditions for λ to be marginals of Y_s ?

$$\boxed{\begin{array}{ll} \lambda_{s_1}(A) = \lambda_{s_1,s_2}(A, \mathbf{R}) & \text{[marginalization]} \\ \lambda_{s_1,s_2}(A_1, A_2) = \lambda_{s_2,s_1}(A_2, A_1) & \text{[perm]} \end{array}}$$

Kolmogorov: if these *consistency* conditions hold for any finite number of variables (not just a pair), then there is a joint stochastic process with these marginals.

The Bayesian choice

Task: given an observed random variable Y, what value should we guess for a related random variable X which is unobserved?

Criterion: if we make a guess x and the real value is $x*$, we pay a cost of $L(x,x*)$ --- this is called a *loss function*.

In the Bayesian framework: you should answer

$$\boxed{\operatorname{argmin}_x \mathbb{E}(L(x, X)|Y)}$$

Directed Graphical Models

Example: $X \longrightarrow Y \longrightarrow Z$

Interpretation: the collection of all distributions that can be factorized as

$$p(x,y,z) = p_1(x)\ \ p_2(y|x)\ \ p_3(z|y)$$

for some non-negative p_i s such that for each w:

$$\int\ p_i(v|w)\ m(\mathrm{d}v) = 1$$

Undirected Graphical Models

Example: $X - \blacksquare - Y - \blacksquare - Z$

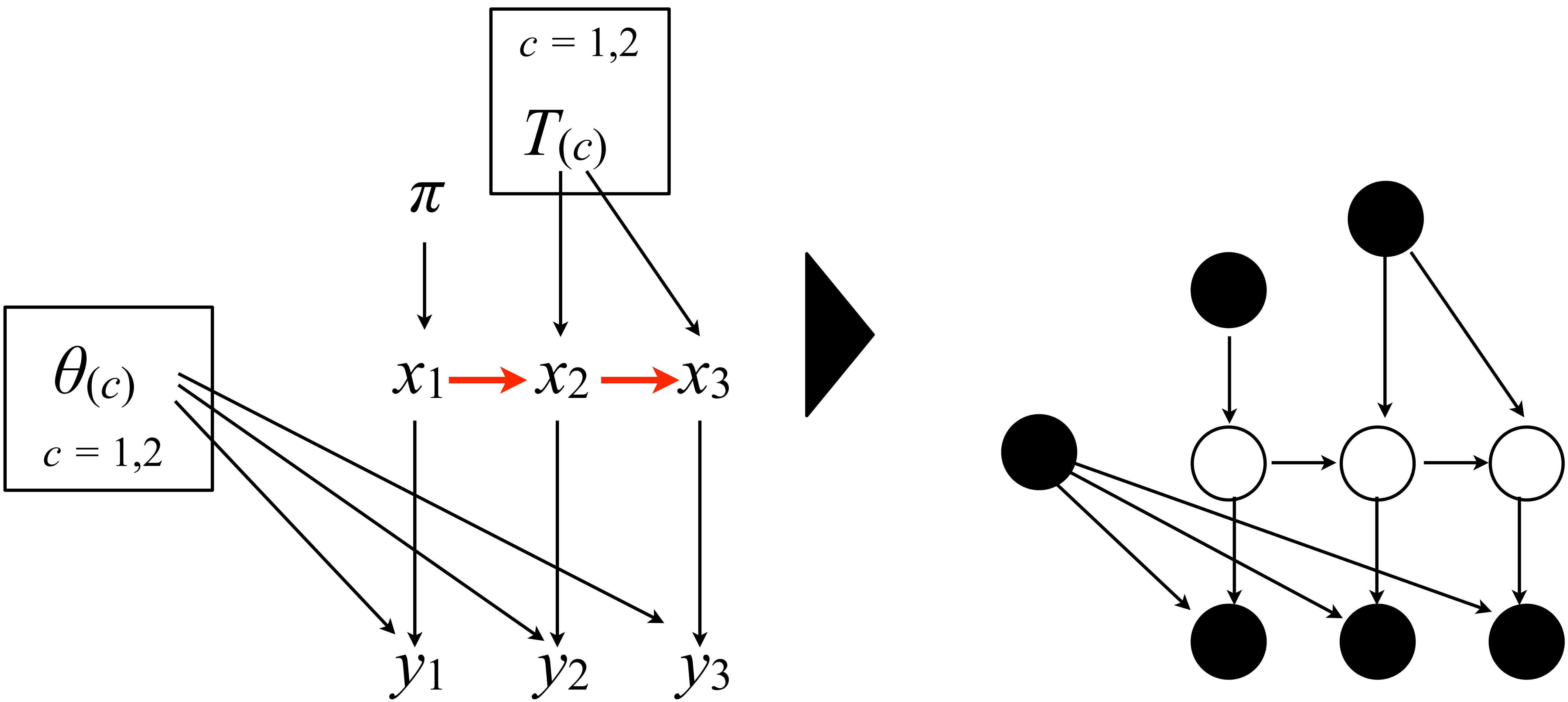
Interpretation: the collection of all distributions such that their density that can be factorized as

$$p(x,y,z) = f_1(x,y)\ f_2(y,z)$$

for some non-negative f_i

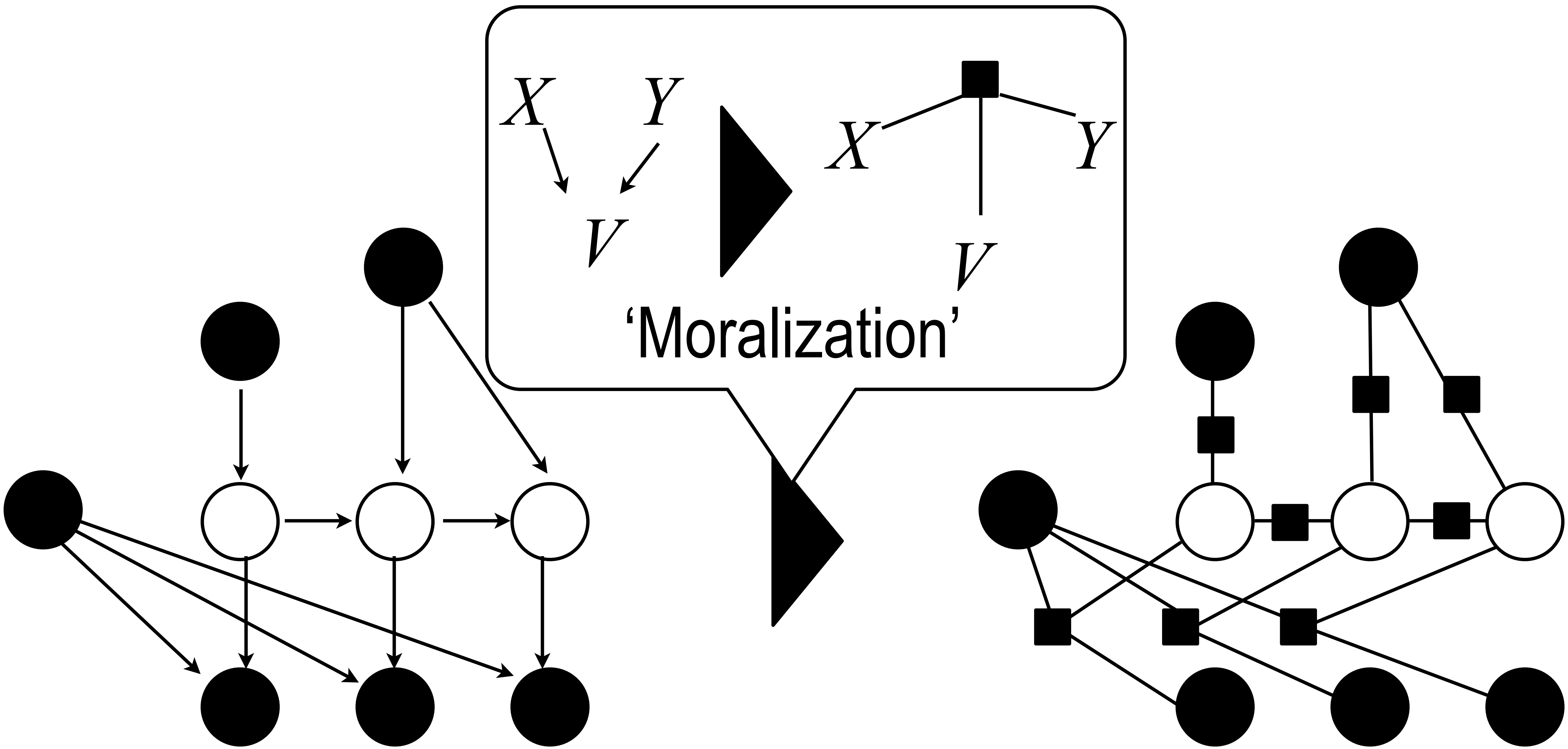
Exact inference and dynamic programming

Suppose: parameters are known, so we condition on them



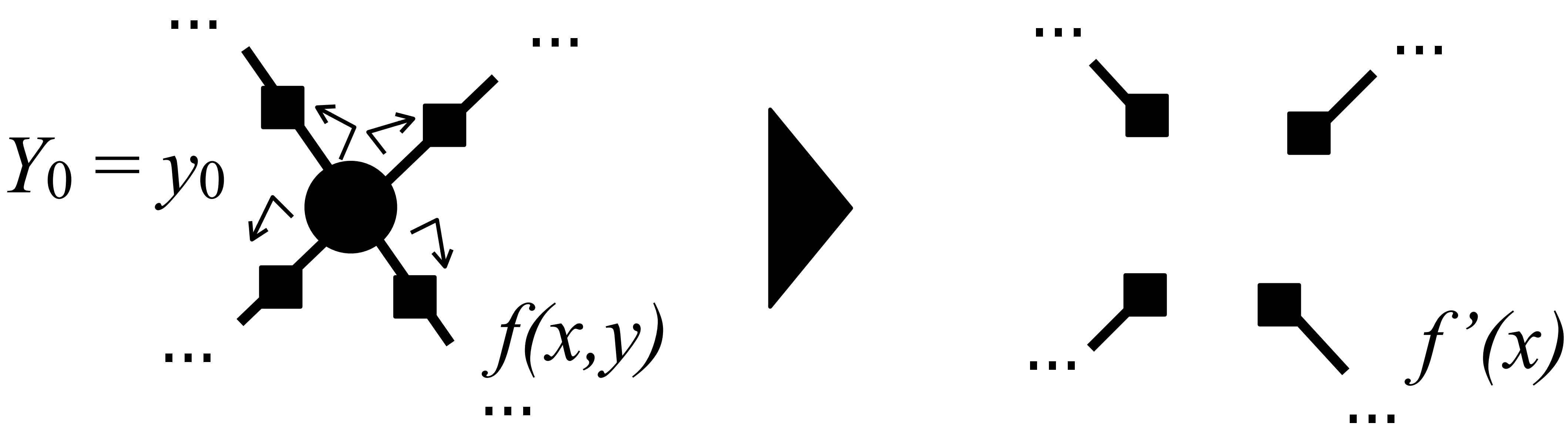
Exact inference and dynamic programming

Next step: turning the directed model into an undirected one



Exact inference and dynamic programming

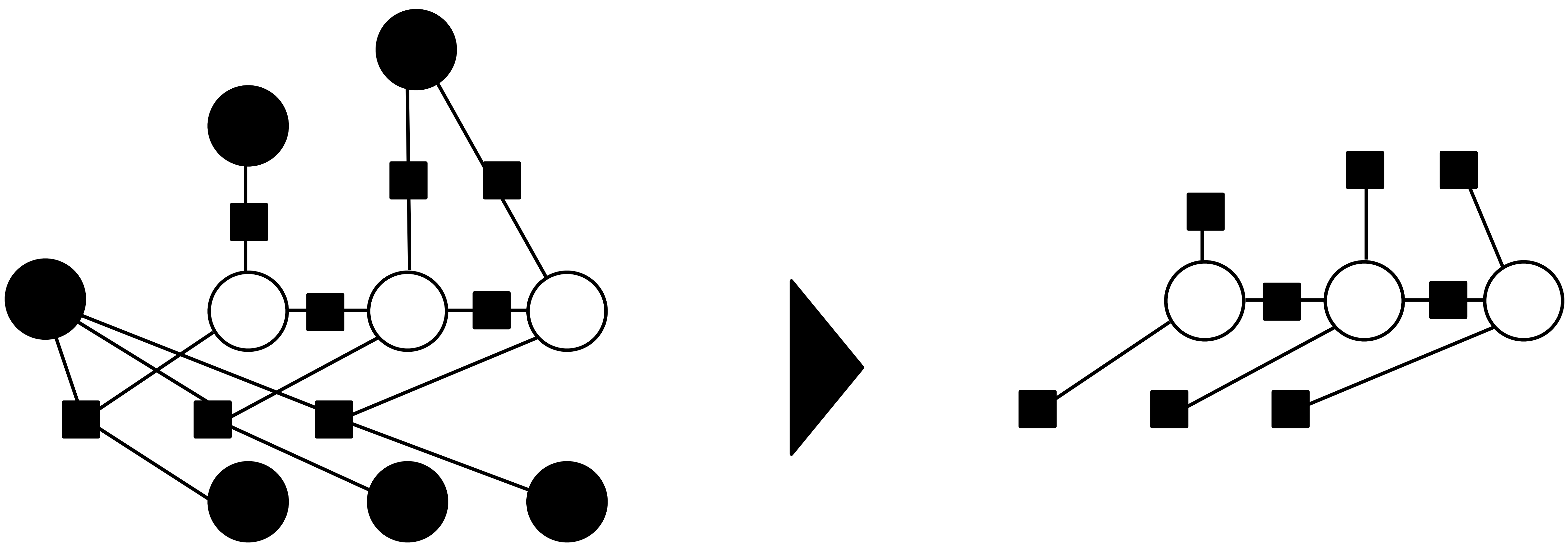
Simplifying undirected models:



$$f'(x) = f(x, y_0)$$

Exact inference and dynamic programming

Simplifications:



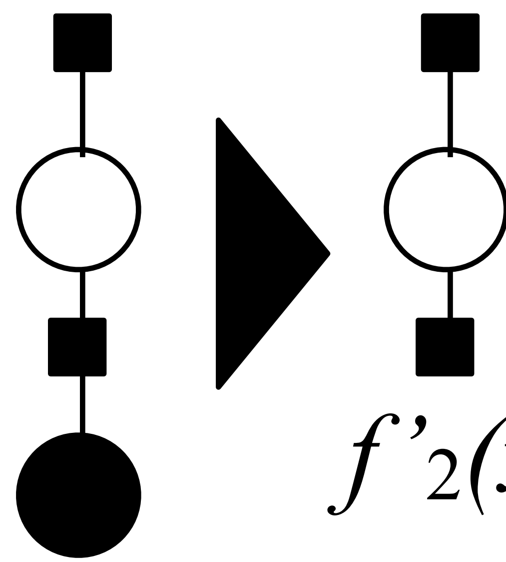
Exact inference and dynamic programming

Consequence of simplification: renormalization needed

Example:

$f_1(x) = p_1(x)$

$f_2(y|x) = p_2(y|x)$



$f_1(x)$

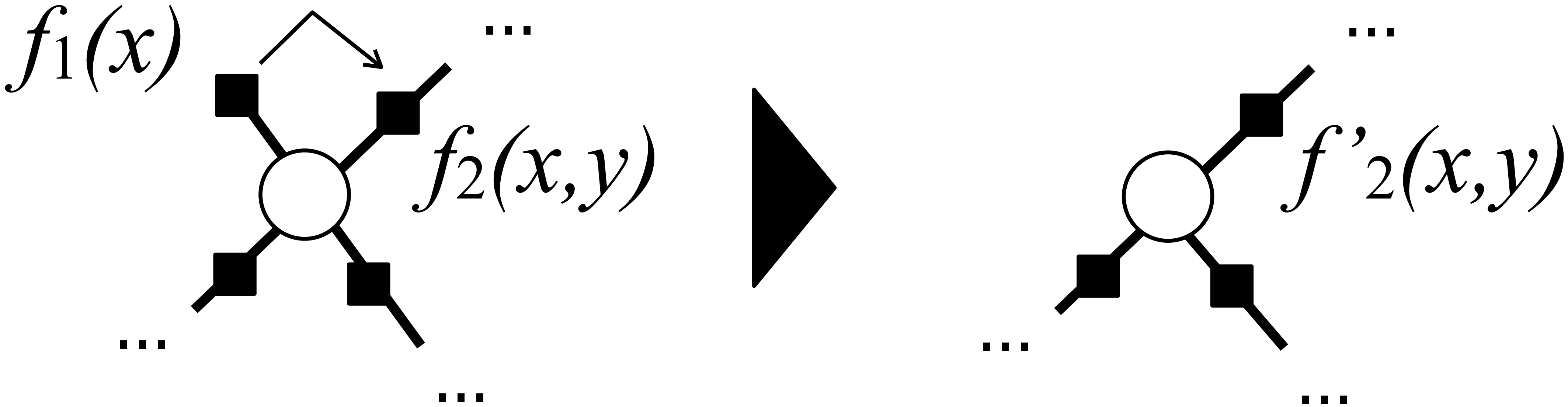
$f'_2(x) = p_2(y_0|x)$

$$P(X = x | Y = y_0) = \frac{f_1(x) f'_2(x)}{\sum_{x'} f_1(x') f'_2(x')}$$

$$= \frac{f_1(x) f'_2(x)}{Z}$$

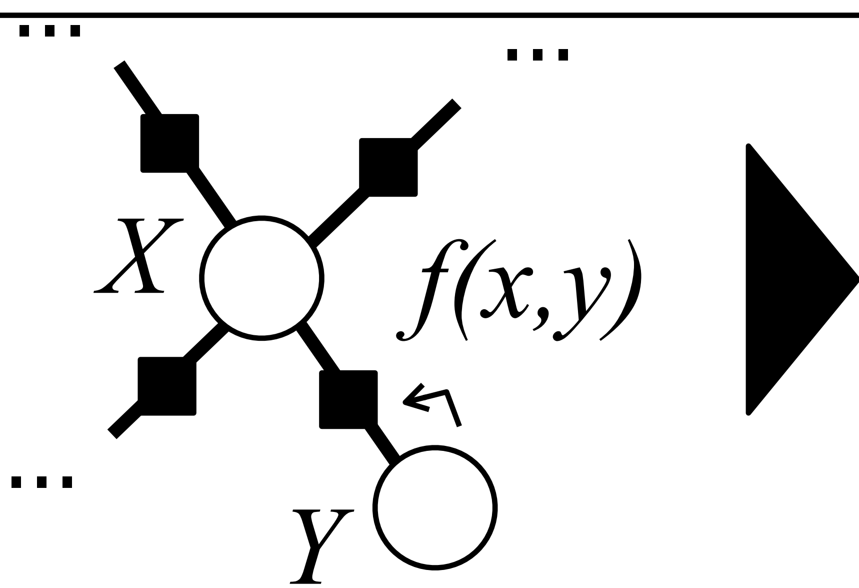
Bayes rule: can interpret Z as $P(Y = y_0)$

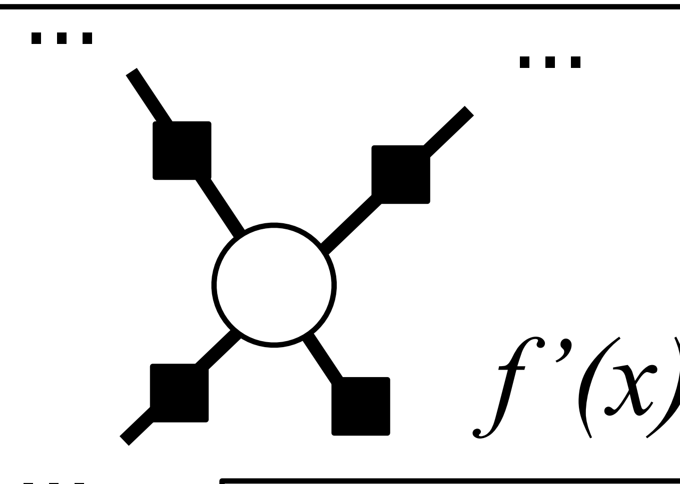
Further simplifications



Pointwise multiplication

$$f'_2(x,y) = f_1(x)\, f_2(x,y)$$



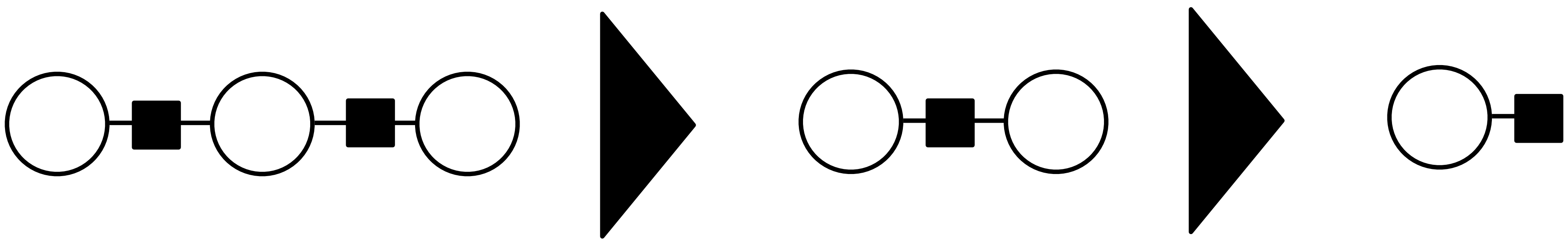


$$f'(x) = \sum_y f(x,y)$$

Marginalization

Efficient inference: elimination algorithm

Consequence: for chains, efficient computation of Z and one-node or two-nodes marginals for tree-shaped undirected graphical models



4 x 4 operations 4 x 4 operations

Much less operations than naive enumeration!

In general: if a *chain* has length T and N states, computing Z takes $T\,N^2$ operations instead of N^T

For tree-shaped models: same story!

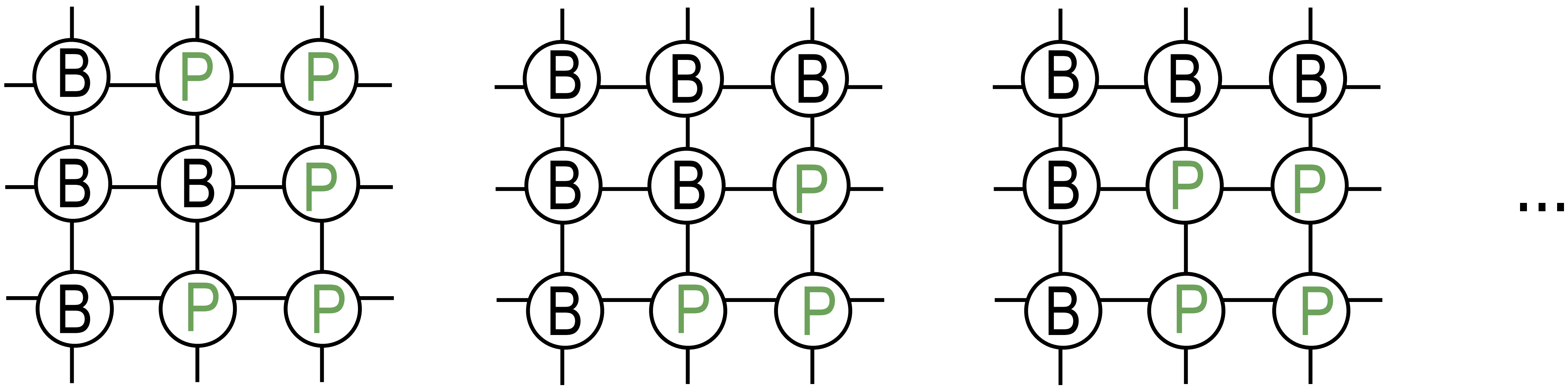
For non-tree models: we need to figure out something else...

MCMC

MCMC methods

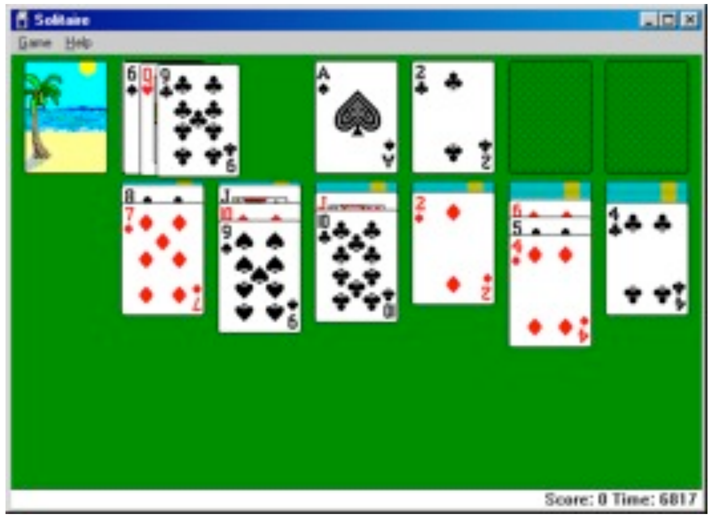
What it does: Same as the elimination algorithm (normalization and posterior), but not limited to trees.

Output: a list of samples, i.e. the model with values for the hidden nodes filled in (imputed)



A bit of history

MC: Usually credited to Stanislaw Ulam, during the Manhattan project.

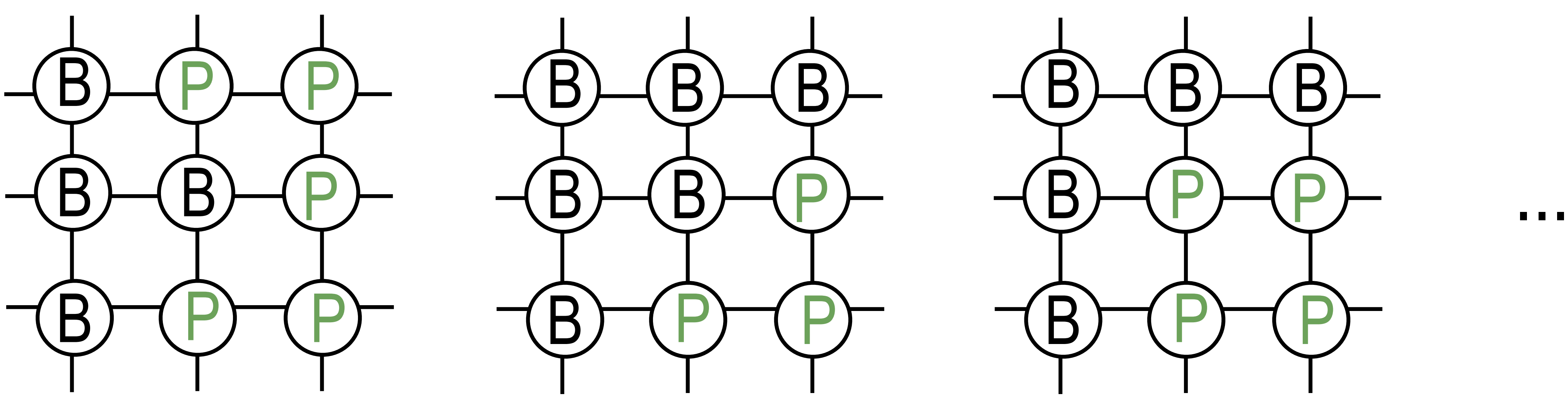


MCMC: Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E.
They ran their chain for 48 iterations on a computer called MANIAC (it took five hours still)

MCMC methods: how does it work?

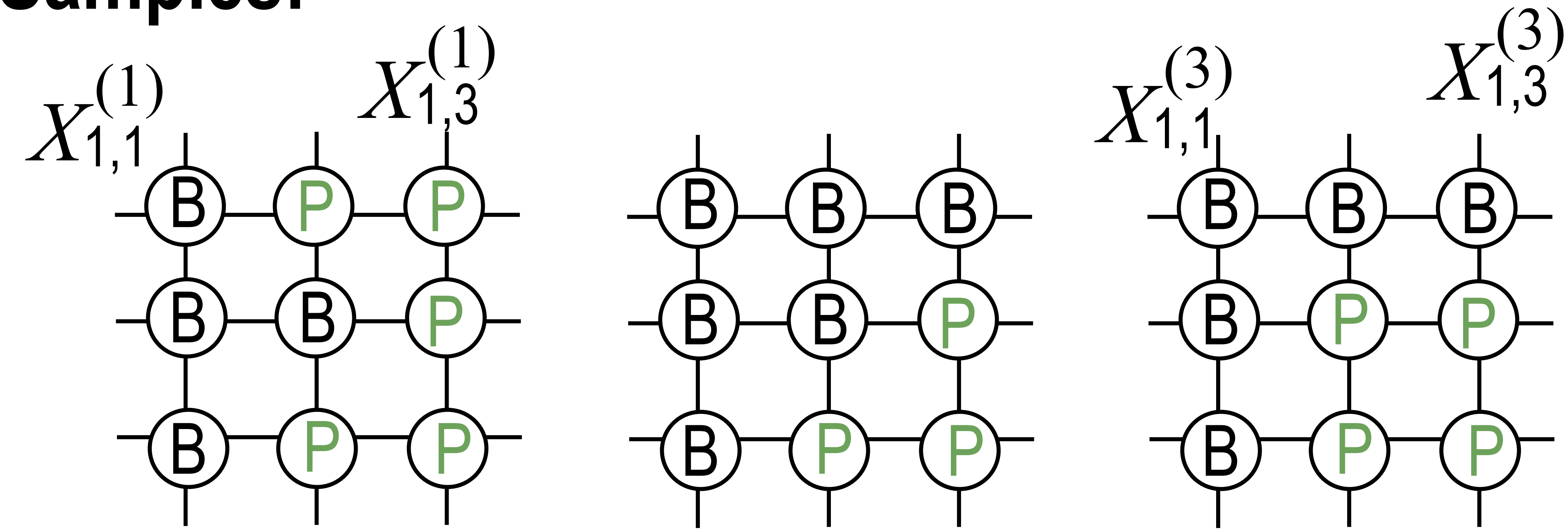
Things to discuss:

- **How to compute posterior expectations from these samples (e.g. Bayes estimator)**
- How to create the samples so that they are approximately distributed according to the posterior?
- How to compute Z from these samples



Computing the posterior

Samples:



Monte Carlo estimator: for S samples, compute

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

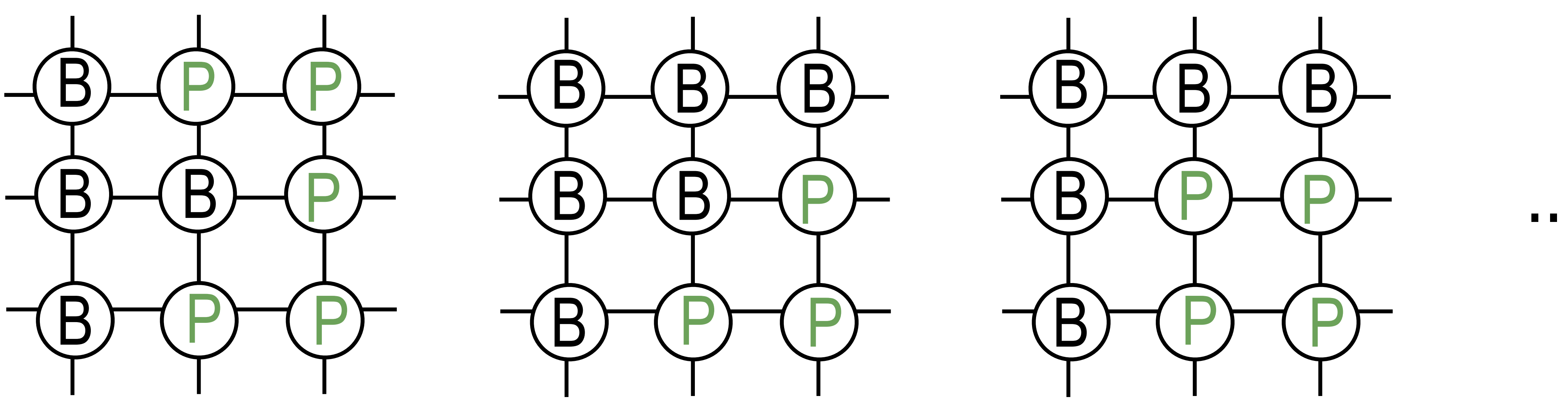
In discrete models, f is generally a vector of indicator functions on variables and values e.g.

$$f_{1,3;\mathrm{B}}(X^{(3)}) = 1$$

MCMC methods: how does it work?

Things to discuss: (note assume for now state is discrete)

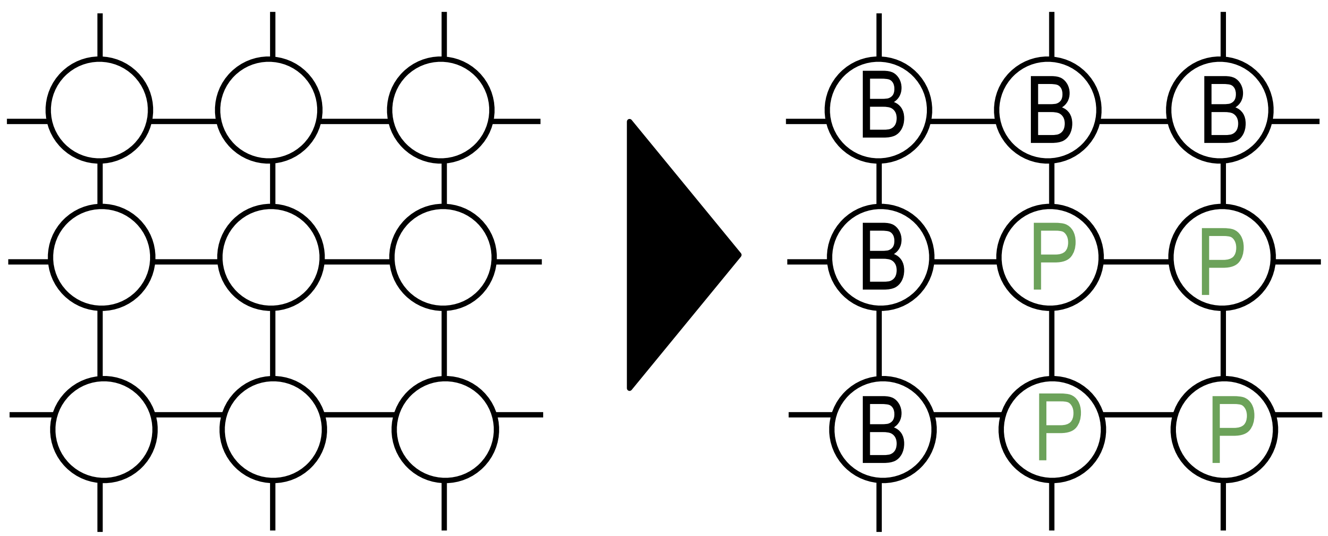
- How to compute posterior expectations from these samples (e.g. Bayes estimator)
- **How to create the samples so that they are approximately distributed according to the posterior?**
- How to compute Z from these samples



Let's start by an easy special case: 'Naive' Gibbs sampling

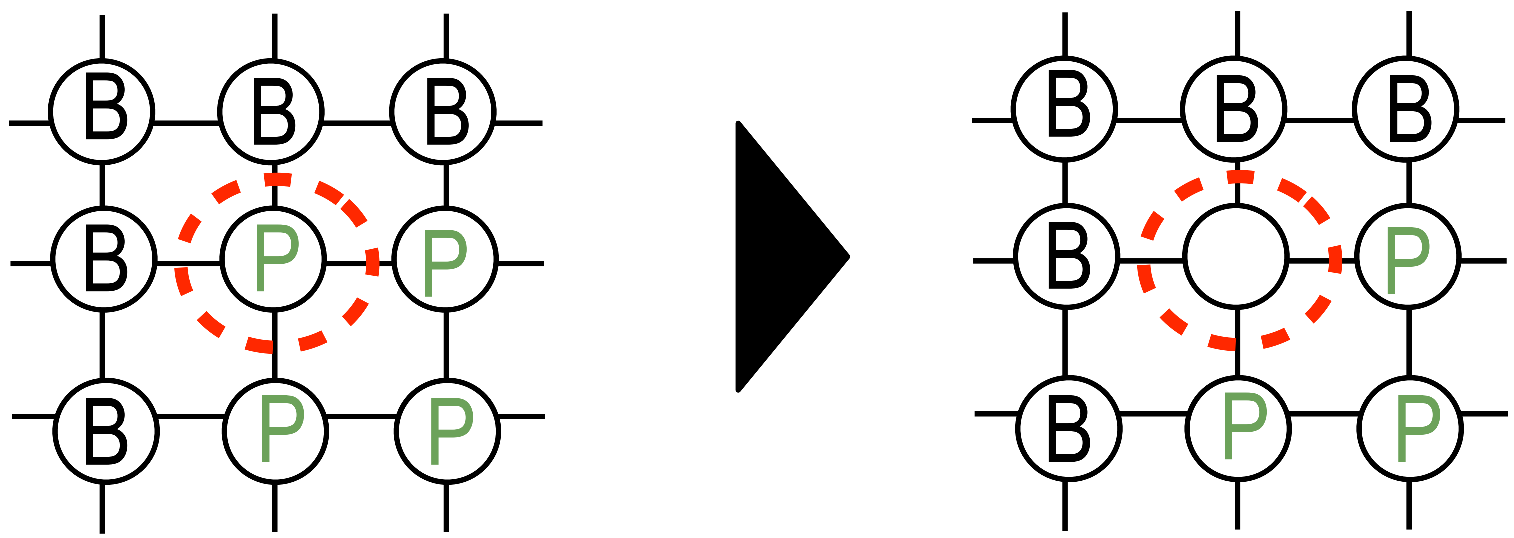
Idea: at each iteration, maintain a guess for all the hidden nodes

Initialization: guess arbitrary values for the hidden nodes

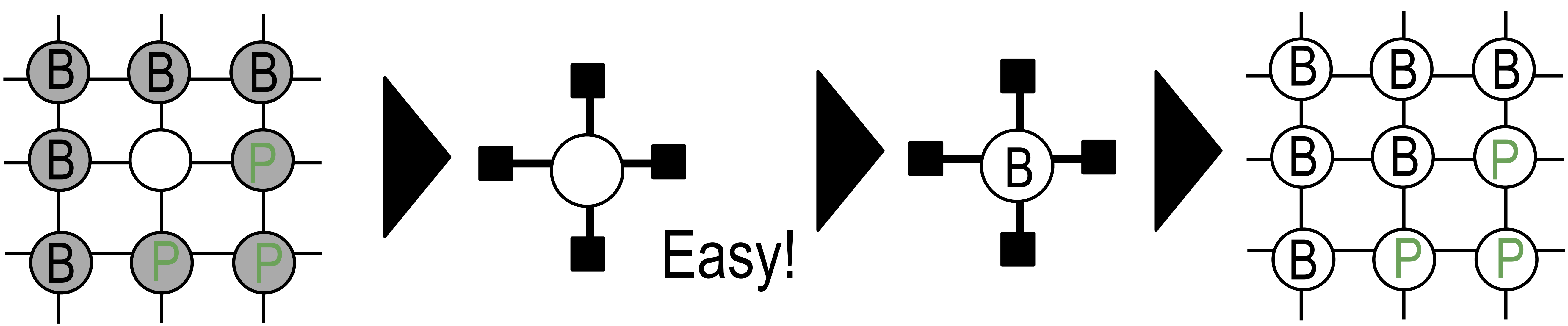


Let's start by an easy special case: 'Naive' Gibbs sampling

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

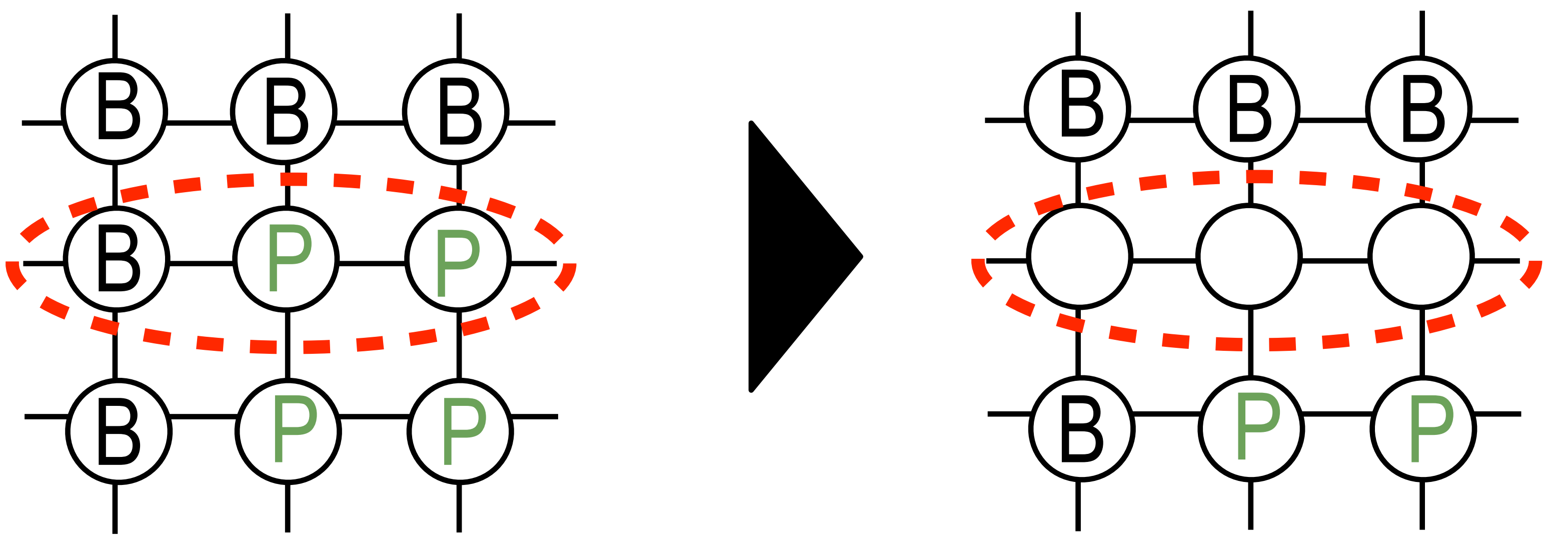


Then: resample a value for the node *(i,j)* conditioning on all the others, and write this to the current state at *(i,j)*

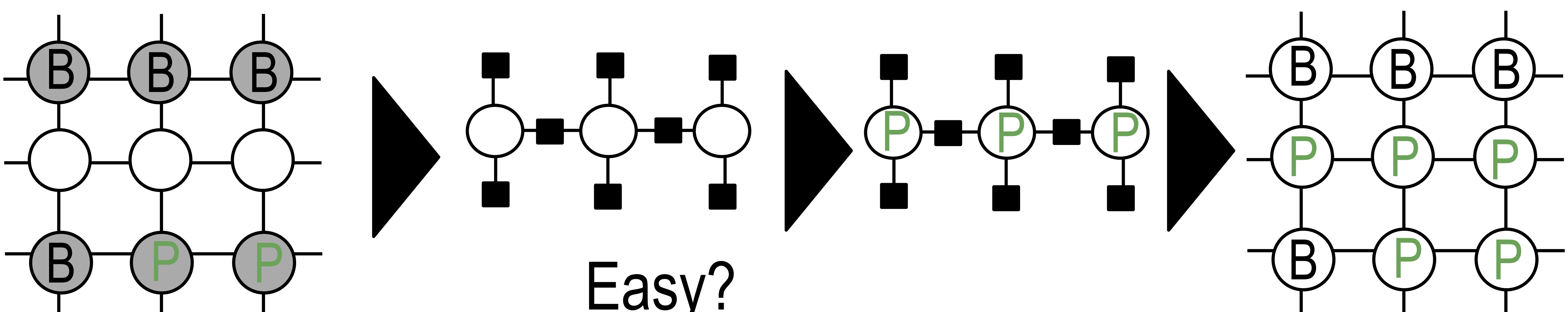


Better Gibbs samplers

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



Next: Metropolis Hastings

Why does it work?

Theoretical framework: The goal is to approximate

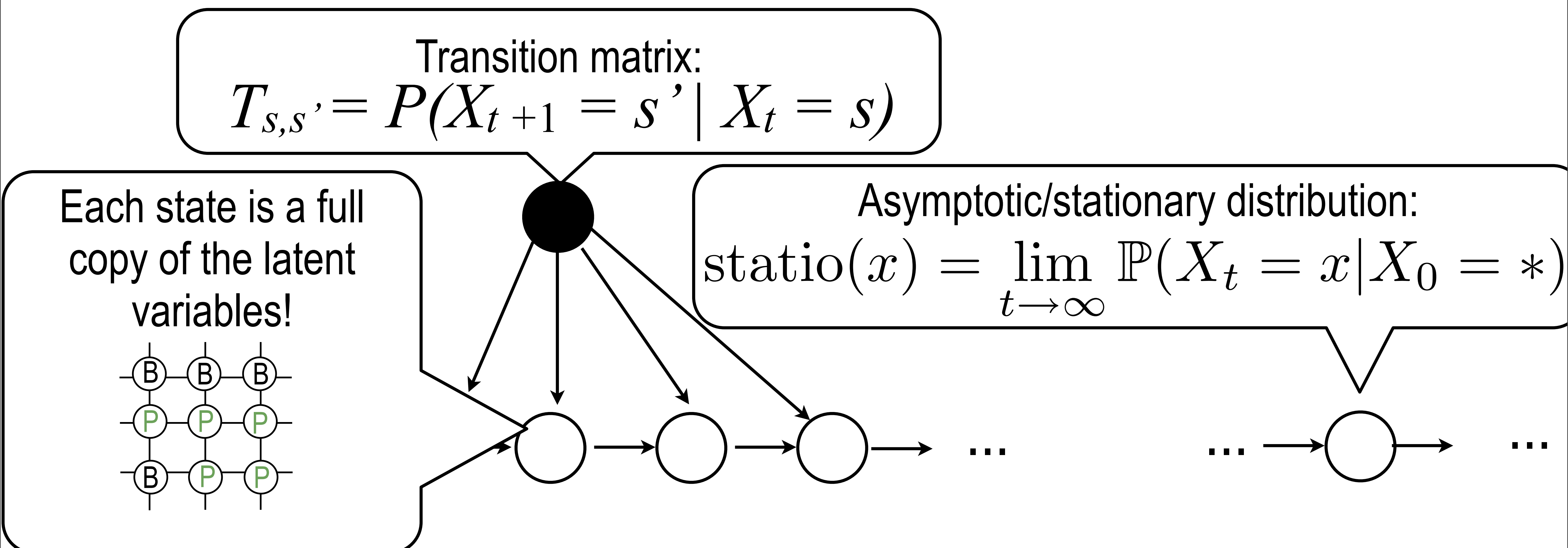
$$\text{target}(x) = \mathbb{P}(X = x | \text{obs}, \text{ params})$$

Method: build a giant *Markov chain* T converging to target(x)

This construction is called a Metropolis-Hastings chain and Gibbs sampling is a special case of it.

Next: Metropolis Hastings

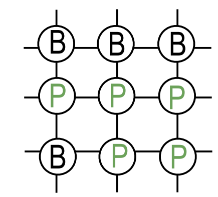
Markov chain:

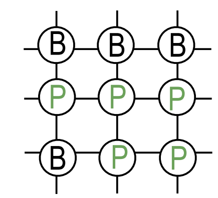


Metropolis Hastings

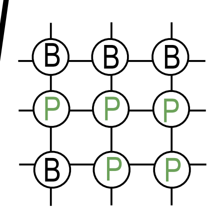
Why it's huge: $2^9 \times 2^9$ matrix

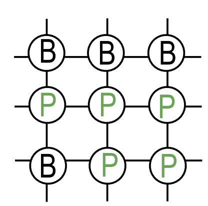
$$T = \begin{bmatrix} 0.1 & 0.01 & \dots \\ 0.01 & & \\ \dots & & \end{bmatrix}$$





...





...

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

Question

How to build T such that:

$$\text{statio}(x) = \text{target}(x)$$

First step: finding a better expression for $\text{statio}(x)$

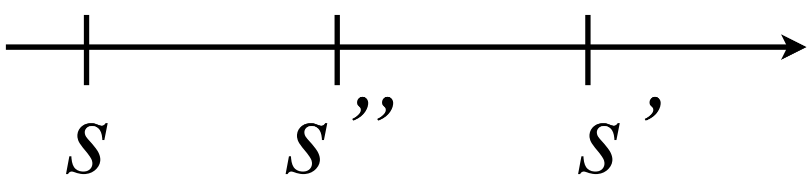
$$\boxed{\begin{aligned} \text{statio}(x) &= \lim_{t \to \infty} \mathbb{P}(X_t = x | X = 0 = *) \\ \text{target}(x) &= \mathbb{P}(X = x | \text{obs, params}) \end{aligned}}$$

Finding a better expression for $\mathrm{statio}(x)$

One step transition: $T_{s,s'} = P(X_{t+1} = s' \mid X_t = s)$

Two steps transition:

$$\mathbb{P}(X_{t+2} = s' | X_t = s) = \left(\sum_{s''} T_{s,s''} T_{s'',s'} \right)_{s,s'}$$



$$= \left(T^2\right)_{s,s'}$$

n-steps transition: T^n

Note: this is a special case of an important principle: Chapman–Kolmogorov equation

Finding a better expression for $\mathrm{statio}(x)$

Definition ('infinite steps' transition): $T^\infty = \lim_{n\to\infty} T^n$

What (matrix-valued) equation should the infinite transition satisfy?

Finding a better expression for $\text{statio}(x)$

Definition ('infinite steps' transition): $T^\infty = \lim_{n \to \infty} T^n$

What (matrix-valued) equation should the infinite transition satisfy?

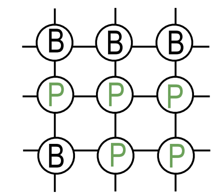
$$T^\infty = T^\infty T$$

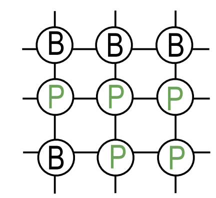
Finding a better expression for $\mathrm{statio}(x)$

Definition ('infinite steps' transition): $T^\infty = \lim_{n \to \infty} T^n$

Hope:

$$T^\infty = \begin{bmatrix} \text{———} \; \pi \; \text{———} \\ \text{———} \; \pi \; \text{———} \\ \vdots \\ \text{———} \; \pi \; \text{———} \end{bmatrix}$$





...

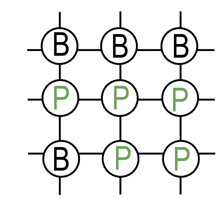
That would mean that no matter what state we use to initialize the sampler, the distribution over the n-th state converges to a distribution called the stationary distribution $\pi(x) = \mathrm{statio}(x) = \mathrm{target}(x)$

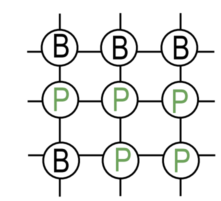
Finding a better expression for $\text{statio}(x)$

Definition ('infinite steps' transition): $T^\infty = \lim_{n \to \infty} T^n$

Hope:

$$T^\infty = \begin{bmatrix} \rule{3cm}{0.4pt} \; \pi \; \rule{3cm}{0.4pt} \\ \rule{3cm}{0.4pt} \; \pi \; \rule{3cm}{0.4pt} \\ \vdots \\ \rule{3cm}{0.4pt} \; \pi \; \rule{3cm}{0.4pt} \end{bmatrix}$$





...

When this is the case (will see later the conditions):

$$\pi(x) = \sum_y \pi(y) T_{y,x} \qquad \text{or} \qquad \pi = \pi T$$

Building T such that statio(x) = target(x)

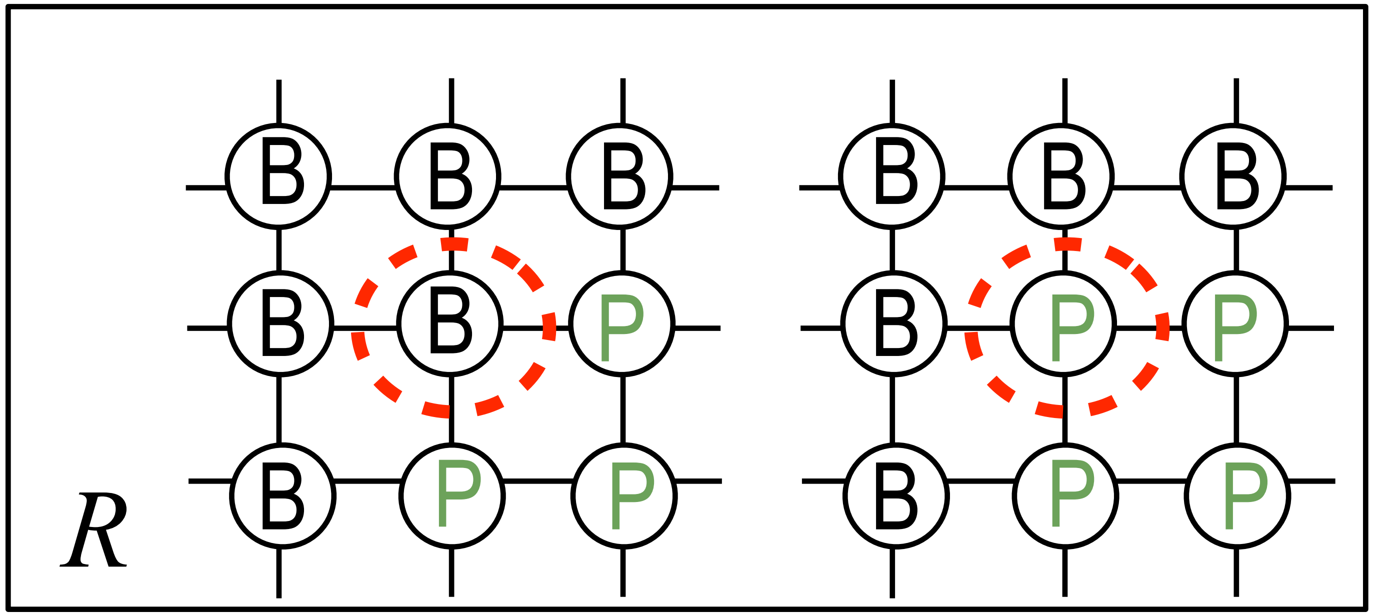
From previous result, want T such that:

$$\mathrm{target}(x) = \sum_y \mathrm{target}(y) T_{y,x}$$

Next: Let's see if Gibbs satisfies this equation!

Definition: Let R denote the set of states reachable by the current Gibbs move

E.g.: in previous Ising example,it has two elements

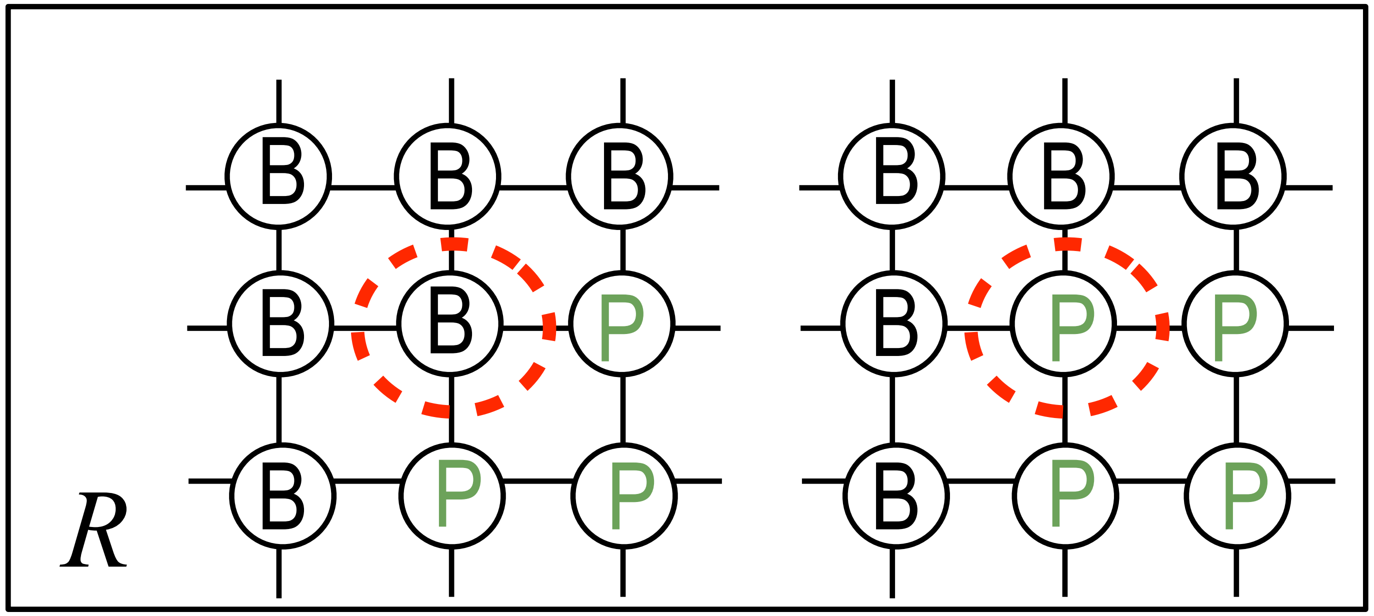


Building T such that statio(x) = target(x)

Goal: Let's see if Gibbs satisfies this equation

$$\text{target}(x) = \sum_y \text{target}(y) T_{y,x} \quad (1)$$

First: Let's find what is $T_{y,x}$



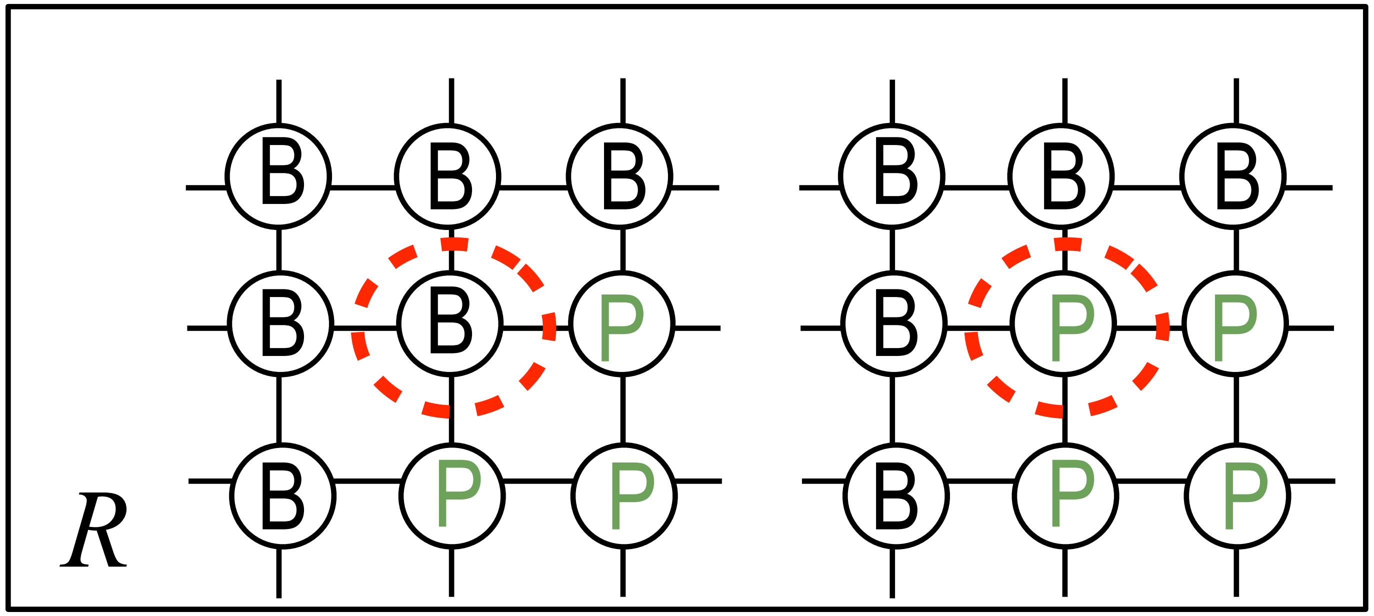
Building T such that statio(x) = target(x)

Goal: Let's see if Gibbs satisfies this equation

$$\text{target}(x) = \sum_y \text{target}(y) T_{y,x} \quad (1)$$

First: Let's find what is $T_{y,x}$

$$T_{y,x} = \frac{\mathbf{1}[x, y \in R]\text{target}(x)}{\sum_{x'} \mathbf{1}[x', y \in R]\text{target}(x')}$$



Building T such that statio(x) = target(x)

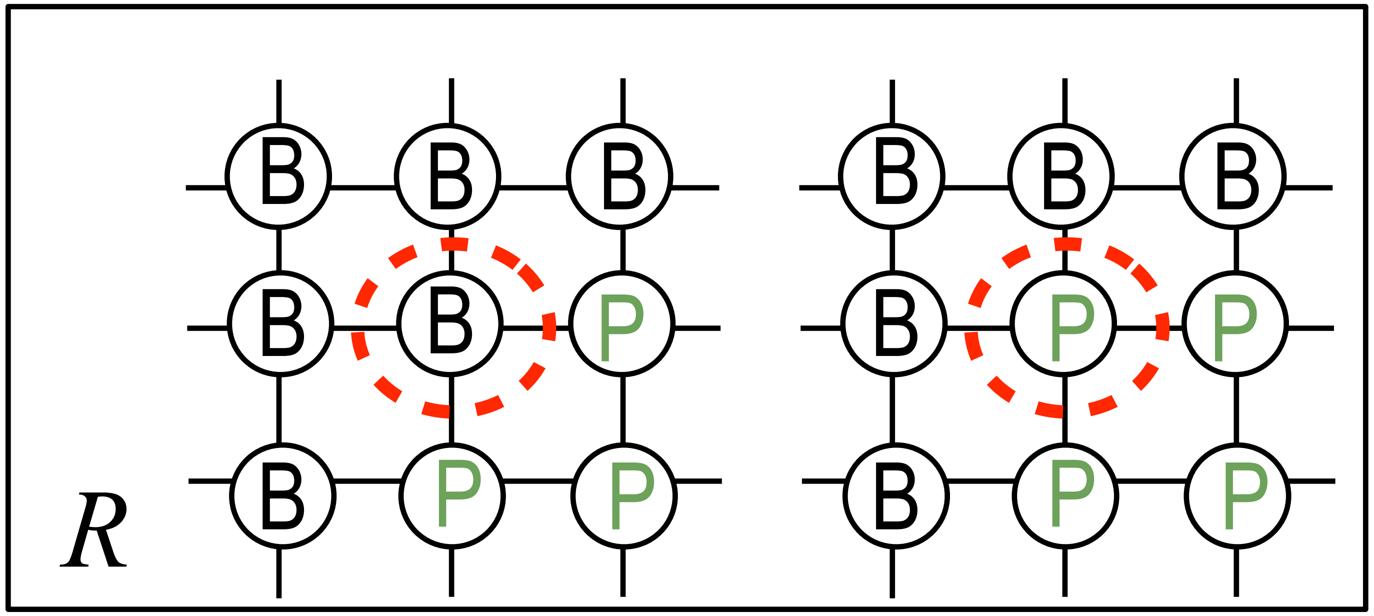
Goal: Let's see if Gibbs satisfies this equation

$$\mathrm{target}(x) = \sum_y \mathrm{target}(y) T_{y,x} \qquad (1)$$

First: Let's find what is $T_{y,x}$

$$T_{y,x} = \frac{\mathbf{1}[x, y \in R]\mathrm{target}(x)}{\sum_{x'} \mathbf{1}[x', y \in R]\mathrm{target}(x')} \qquad (2)$$

Finally: plug-in (2) in (1) and check it works



Gibbs is not always applicable

Example: non-conjugate prior; in which case even a single node has no analytic posterior expression

Generalization: instead of requiring T be proportional to the target distribution, use arbitrary proposal q and correct the discrepancy between q and the target distribution

Terminology: Metropolis-Hastings

Metropolis-Hastings meta-algorithm

Metropolis-Hastings(target(x), $q(x_{\text{next}}|x_{\text{cur}})$, $f(x)$ **)**

Initialize x_0 arbitrarily

$F = 0$; $N = 0$

For t = 1... S

1. Propose a new state x_{prop} according to $q(\ -\ |\ x_{t\text{-}1})$
2. Compute:

$$A(x_{t-1} \to x_{\text{prop}}) = \min\left\{1, \frac{\text{target}(x_{\text{prop}})q(x_{t-1}|x_{\text{prop}})}{\text{target}(x_{t-1})q(x_{\text{prop}}|x_{t-1})}\right\}$$

3. Set x_t to x_{prop} with probability $A(x_{t\text{-}1} \to x_{\text{prop}})$, otherwise set x_t to $x_{t\text{-}1}$
4. $F = F + f(x_t)$, $N = N + 1$

Return F/N

$\approx \mathbb{E}[f(X)]$ for $X \sim$ target

Why Metropolis-Hastings works

From previous result, want T such that:

$$\text{target}(x) = \sum_y \text{target}(y) T_{y,x}$$

Sufficient condition (by summing over y on both sides):

$$\text{target}(x) T_{x,y} = \text{target}(y) T_{y,x}$$

This is called *detailed balance* or *reversibility* condition

Why Metropolis-Hastings works

Goal: checking detailed balance for the MH kernel T

$$\text{target}(x)T_{x,y} = \text{target}(y)T_{y,x}$$

First: what is $T_{x,y}$? When $x = y$, the result trivially holds, so let's assume that $x \neq y$

When $x \neq y$, $T_{x,y}$ is equal to the probability that
(1) y is proposed by $q(\text{-}|x)$ times
(2) the probability that it is accepted:

$$T_{x,y} = q(y|x)A(x \to y)$$

Why Metropolis-Hastings works

Final step: using the form of $T_{x,y}$ for $x \neq y$ to check detailed balance for the MH kernel T

Goal: $$\text{target}(x)T_{x,y} = \text{target}(y)T_{y,x}$$

Known:

$$A(x_{t-1} \to x_{\text{prop}}) = \min\left\{1, \frac{\text{target}(x_{\text{prop}})q(x_{t-1}|x_{\text{prop}})}{\text{target}(x_{t-1})q(x_{\text{prop}}|x_{t-1})}\right\}$$

$$T_{x,y} = q(y|x)A(x \to y)$$

Notes on Metropolis-Hastings

Critical: the target and proposal densities always appear as ratios, so if they are only known up to a normalization Z, the normalizations cancel out

$$A(x_{t-1} \to x_{\text{prop}}) = \min\left\{1, \frac{\text{target}(x_{\text{prop}})q(x_{t-1}|x_{\text{prop}})}{\text{target}(x_{t-1})q(x_{\text{prop}}|x_{t-1})}\right\}$$

Practical note: should be computed in log space and exponentiated only after taking ratio (difference of logs)

Special cases: when q is symmetric (e.g. isotropic normal), the q's cancel out as well. When $q(\text{-}|x_{\text{cur}})$ is independent of x_{cur}, it's called an *independence chain* (still has dependence because of A)

Useful theoretical results

Definition ('infinite steps' transition): $T^\infty = \lim_{n \to \infty} T^n$

When does this limit make sense (when does it exists and is stochastic)?

Hope:

$$T^\infty = \begin{bmatrix} \text{———}\ \pi\ \text{———} \\ \text{———}\ \pi\ \text{———} \\ \vdots \\ \text{———}\ \pi\ \text{———} \end{bmatrix} \quad \ldots$$

When does the limit has this form?

Counter-example 1

$$T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Limit T^{∞} not even defined!

Problem: a waltz between states

Definition: A state s (or chain) has period k if any return to state s must occur in multiples of k steps. The chain is aperiodic if one (all) states have period 1.

Easy to avoid: add epsilon self-transitions

Counter-example 2

$$T = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

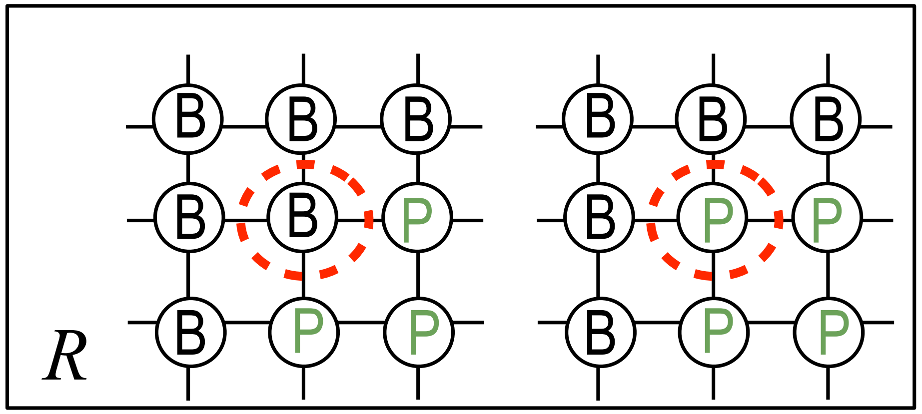
Exercise: the asymptotic distribution depends on the starting state. In fact: $T^\infty = T$

Problem: some pairs of states cannot reach each other

Definition: An *irreducible* chain is a chain where there is a path between each pair of states (for each x, y there is an integer n such that $(T^n)_{x,y} > 0$)

Are the MH and Gibbs kernels we have introduced earlier irreducible?

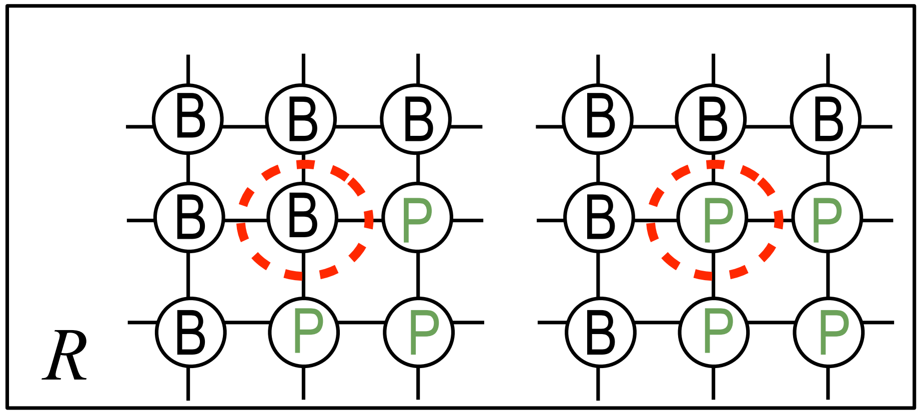
Example: is this irreducible?



Are the MH and Gibbs kernels we have introduced earlier irreducible?

Example: is this irreducible?

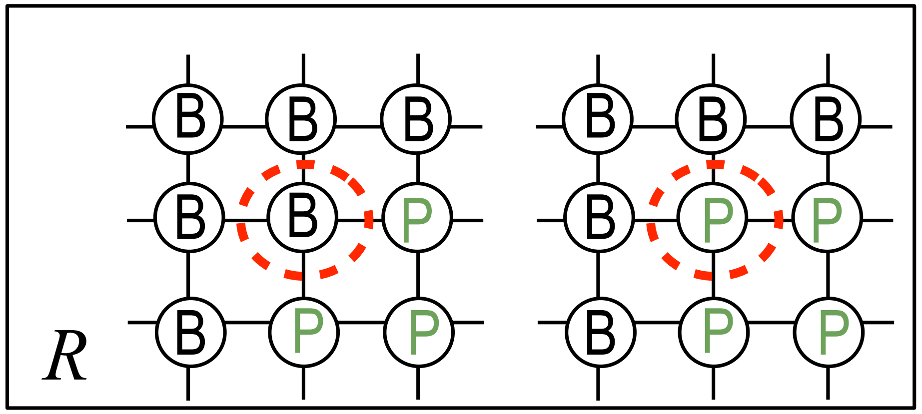
$$T_{y,x} = \frac{\mathbf{1}[x, y \in R]\mathrm{target}(x)}{\sum_{x'} \mathbf{1}[x', y \in R]\mathrm{target}(x')}$$



Are the MH and Gibbs kernels we have introduced earlier irreducible?

Example: is this irreducible?

$$T_{y,x} = \frac{\mathbf{1}[x, y \in R]\text{target}(x)}{\sum_{x'} \mathbf{1}[x', y \in R]\text{target}(x')}$$



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

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

Are the MH and Gibbs kernels we have introduced earlier irreducible?

Solution 1: mixing kernels. Suppose we have one Gibbs kernel for each variable $T^{(1)}$, ..., $T^{(9)}$. 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).

$$T_{x,y} = \sum_{x_1} \cdots \sum_{x_9} T^{(1)}_{x,x_1} T^{(2)}_{x_1,x_2} \cdots T^{(9)}_{x_8,x'}$$

Often works better: shuffle then alternate

Existence of π such that $\pi = \pi T$

Suppose: (still assuming discrete state space)

1. T is irreducible
2. T is aperiodic

Consequence: There is a unique probability distribution π such that $\pi = \pi T$

Proofs: Consequence of Perron–Frobenius theorem (T^n is positive for n large enough, and π is then the eigenvector corresponding to the unique eigenvalue of highest modulus). --- **Note:** can be used to debug samplers

More general arguments exist

Convergence theorem 1

Suppose: (still assuming discrete state space)

1. T is irreducible
2. T is aperiodic

Consequence: There is a unique probability distribution π such that $\pi = \pi T$; moreover, for all x,

$$\lim_{n \to \infty} T^n_{x,y} = \pi(y)$$

i.e.:

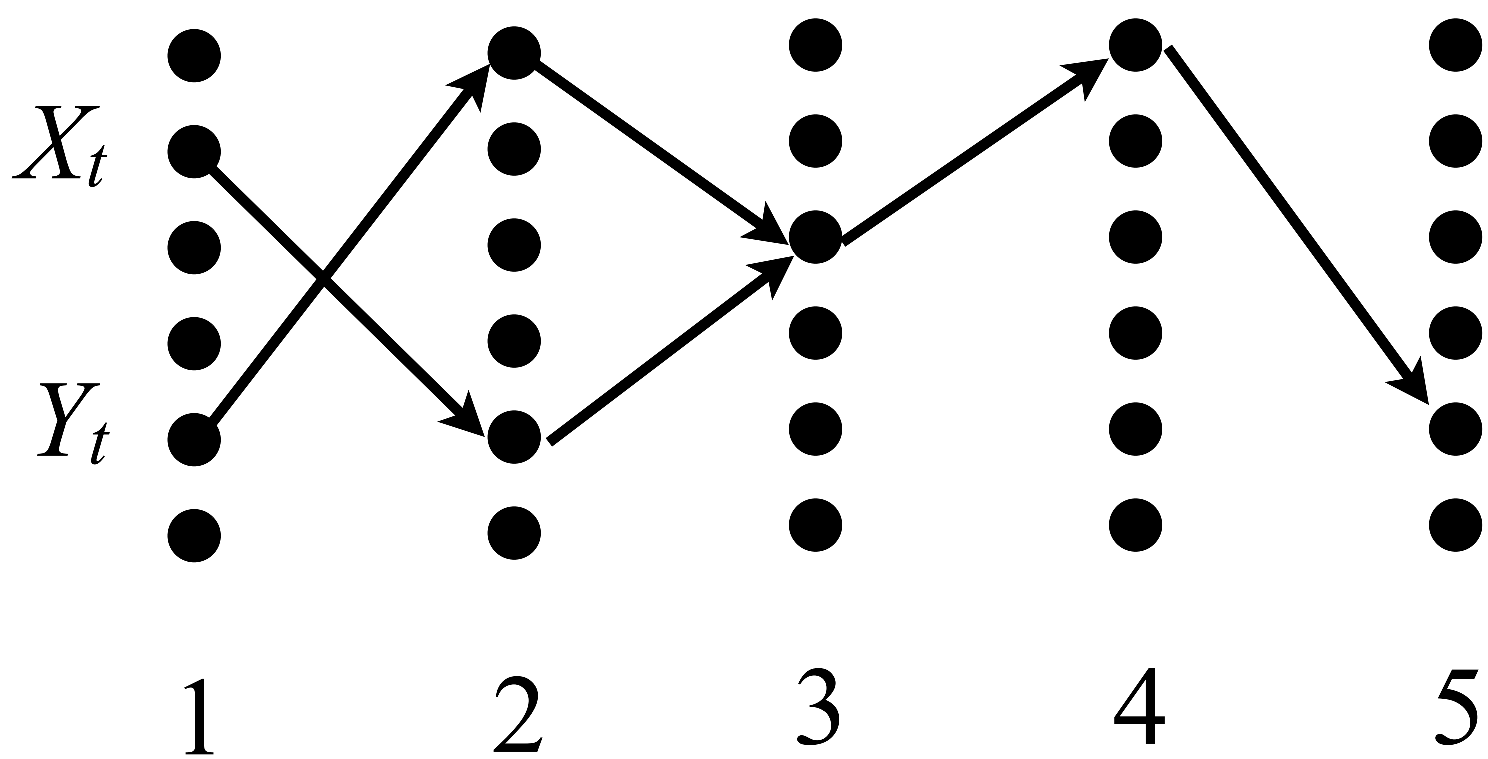
$$T^\infty = \begin{bmatrix} \text{———} \ \pi \ \text{———} \\ \text{———} \ \pi \ \text{———} \\ \vdots \\ \text{———} \ \pi \ \text{———} \end{bmatrix}$$

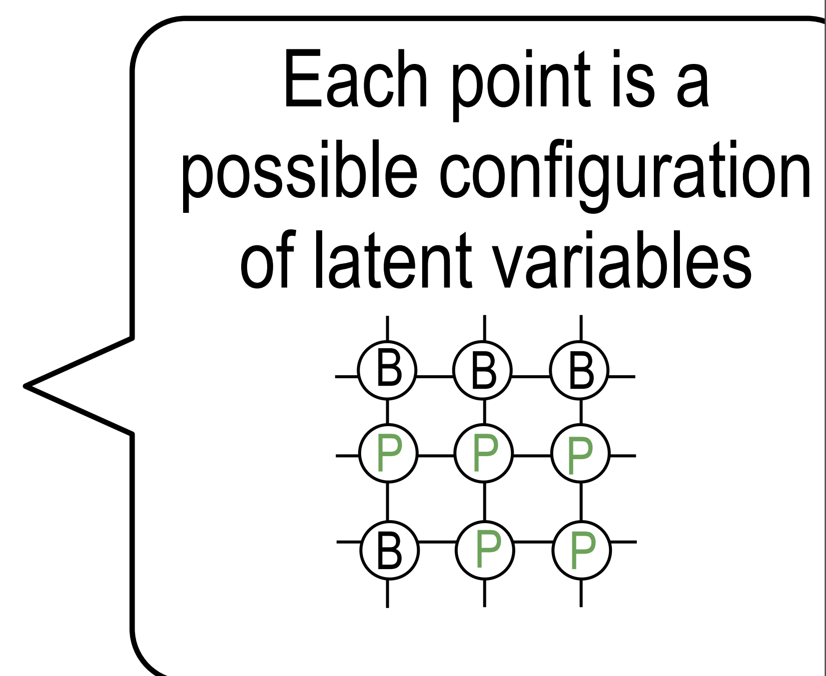
Proof: coupling argument

Idea: simulate a *pair* of chains (X_t, Y_t) such that the marginal transitions are given by T:

$$P(X_t = x' | X_{t-1} = x, Y_{t-1} = y) = T_{xx'}$$

Joint distribution: simulate *independent* transitions if $x \neq y$, and *identical* transitions if $x = y$.



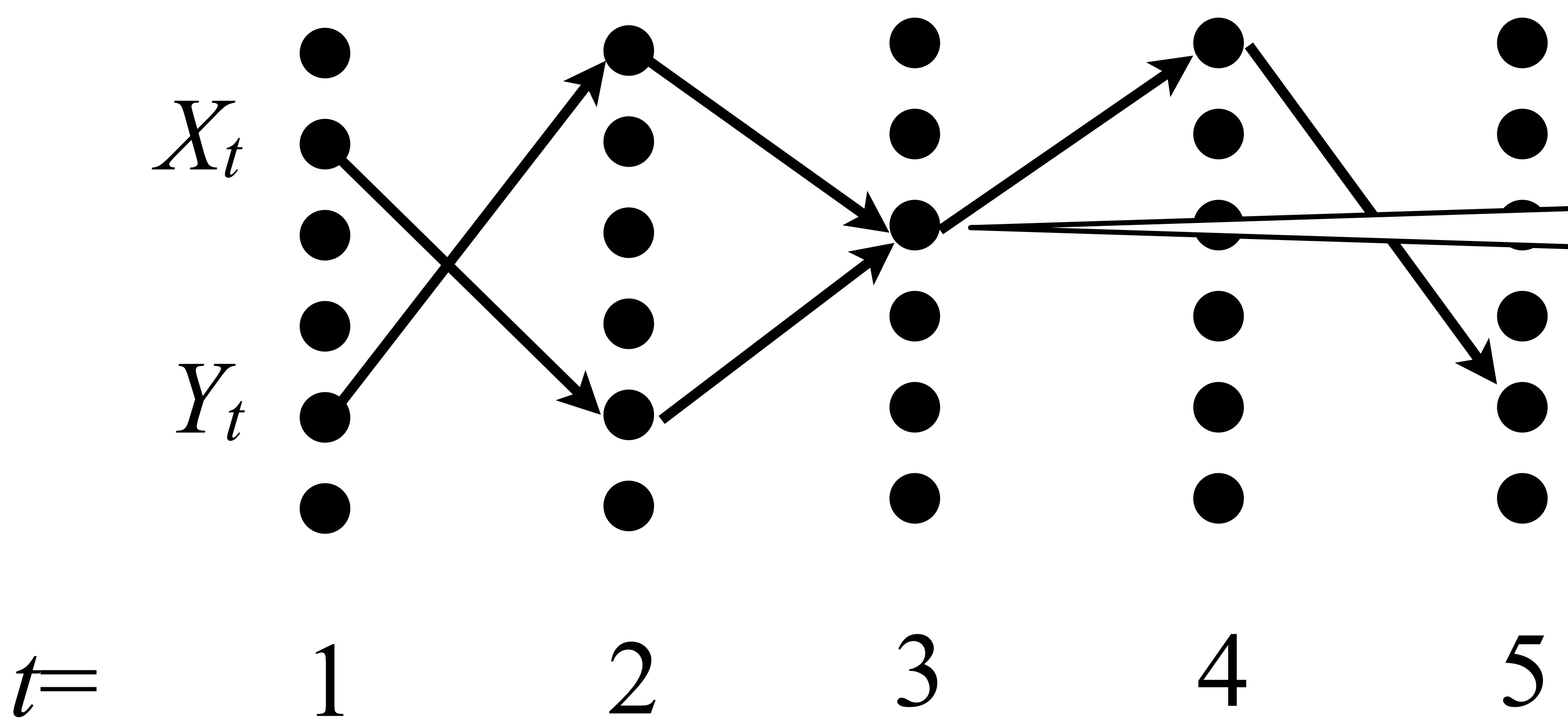


Proof: coupling argument

Initial distributions: $X_0 \sim \pi$ and $Y_0 \sim$ arbitrary distribution

Note: $X_t \sim \pi$ for all t since $\pi = \pi T$

Goal: showing that $\lim_{n\to\infty} \sum_y \left|\mathbb{P}(X_n = y) - \mathbb{P}(Y_n = y)\right| = 0$



Notation: the hitting time H where X_t=Y_t for the first time.
e.g. H=3 here

This is not exactly what we need though...

Recall: the goal is to compute an expectation (with respect to the posterior distribution), not to sample!

Connexion: the law of large numbers

Vanilla version: If X_t are iid π and f is finite, then

$$\lim_{n \to \infty} \frac{1}{S} \sum_{t=1}^{S} f(X_t) = \sum_{x} f(x)\pi(x)$$

Misconception: to have the same conclusion hold for MCMC, we need to *burn-in* and/or *thin* the chain

Burn-in and thinning

Burn-in:

$$\frac{1}{S-b}\sum_{t=b}^{S} f(X_t)$$

Thinning:

$$\frac{1}{S/n}\sum_{t\in\{n,2n,\ldots,S\}} f(X_t)$$

Burn-in and thinning are unnecessary

The law of large numbers for Markov chains: If X_t is an irreducible Markov chain with stationary distribution π and f is finite, then

$$\lim_{n\to\infty} \frac{1}{S} \sum_{t=1}^{S} f(X_t) = \sum_{x} f(x)\pi(x)$$

Note 1: Aperiodicity not needed for this result

Note 2: For small S, burning-in might improve the estimator, but might as well maximize during burn-in

Note 2: Thinning to reduce auto-correlation is not a good idea and can be harmful (only reasons to do it is to save memory writes or memory---but most of the time only finite dimensional sufficient statistics need to be stored)