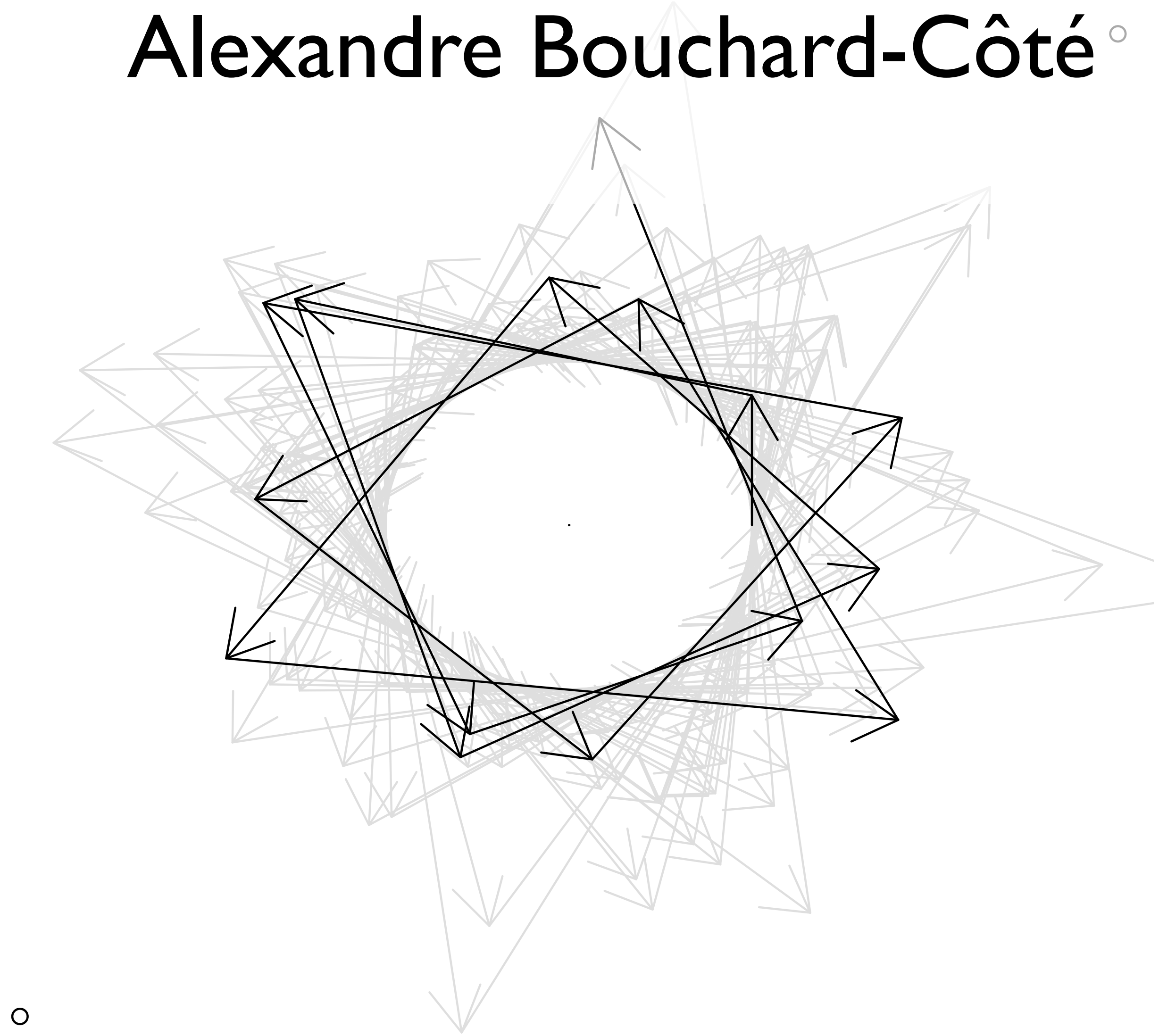
Monte Carlo methods

Alexandre Bouchard-Côté



Markov chain Monte Carlo, continued

Motivation

- Methods we have seen so far (Simple MC, RS, IS)...
  - do not scale well in *d* (except for a few special cases)
  - often work poorly in combinatorial spaces

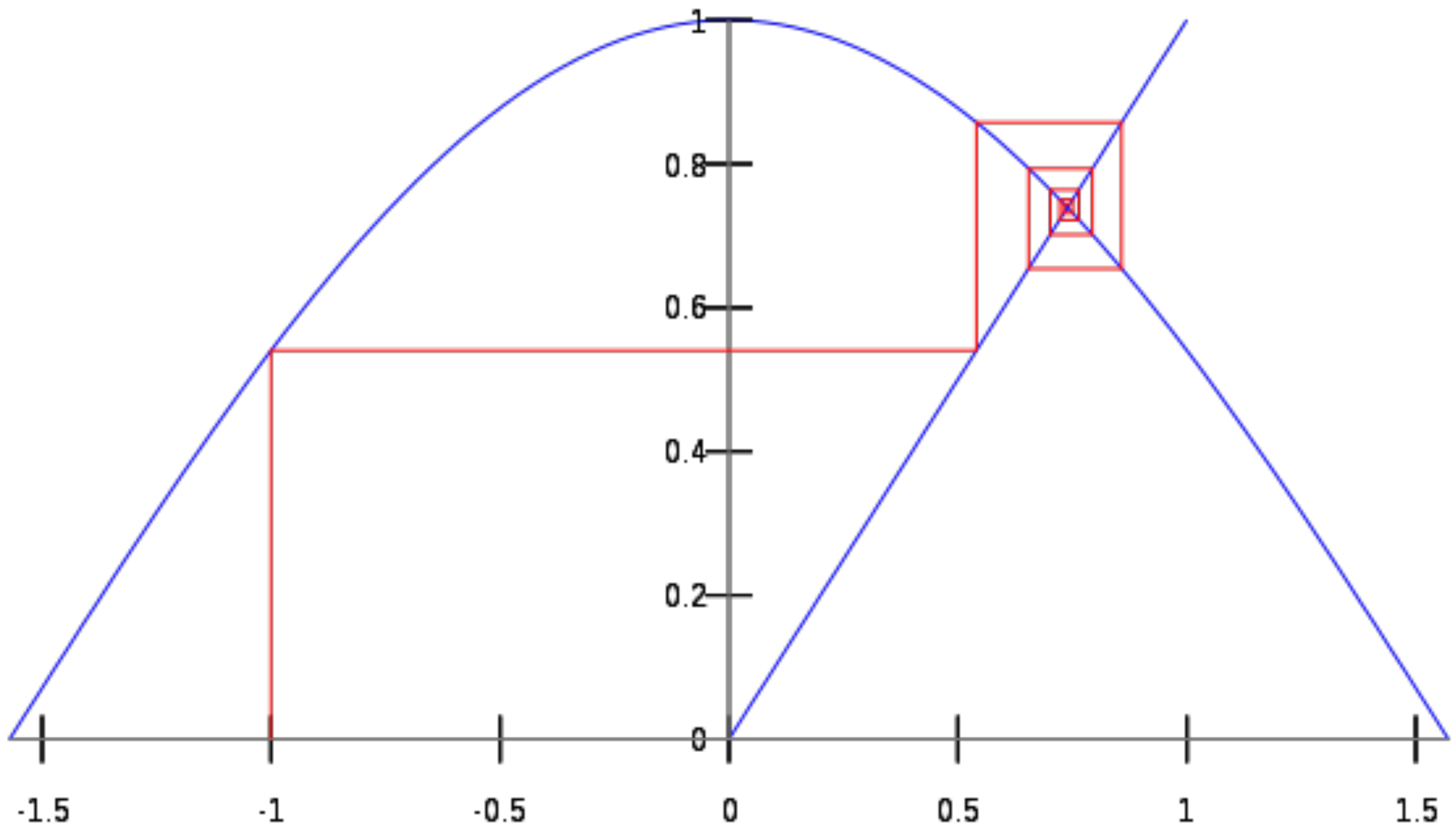
MCMC: main ideas

- **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 π
  - Even if we do not know the normalization of π

Towards MC LLN&CLT: Finite MC review

Stationarity/invariance

Fixed points of the transition kernels



Def 28

- **Definition:** A distribution π is said to be *invariant* or *stationary* for a Markov kernel, K, if $\pi K = \pi$.
- If there exists t such that $X_t \sim \pi$ where π is a stationary distribution, then $X_{t+s} \sim \pi K^s = \pi$ for all $s \in \mathbb{N}$. (Note that this tells us nothing about the correlation between the states or their joint distribution.)
- *Example*: For any $\theta \in [0, 1]$

$$K_\theta = \begin{pmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{pmatrix}$$

admits

$$\pi = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

as invariant distribution.

Def 29

- **Definition:** A Markov kernel K is π-reversible if

$$\forall x, y \in \mathbb{X} : \; \pi_x K_{xy} = \pi_y K_{yx}.$$

- **Lemma**: If K is π-reversible then K is π-invariant.
- **Proof**. Indeed we have

$$\sum_{x \in \mathbb{X}} \pi_x K_{xy} = \sum_{x \in \mathbb{X}} \pi_y K_{yx} = \pi_y,$$

i.e . $\left(\pi K\right)_y = \pi_y$

- Reversibility means that the statistics of the time-reversed version of the process match those of the process in the forward distribution, K_θ is π-reversible as $\pi_1 K_{\theta,12} = \frac{1}{2}\left(1-\theta\right) = \pi_2 K_{\theta,21}$.

Example 30

- Let $P = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$.

  We have $\pi P = \pi$ for $\pi = (1/2, 1/3, 1/6)$.
- P cannot be π reversible as

$$1 \to 3 \to 2 \to 1$$

  is a possible sequence whereas

$$1 \to 2 \to 3 \to 1$$

  is not (as $P_{2,3} = 0$).
- Detailed balance does not hold as $\pi_2 P_{23} = 0 \neq \pi_3 P_{32}$.

Example 31

- All finite Markov chains have at least one stationary distribution but not all stationary distributions are also limiting distributions.
- **Example**

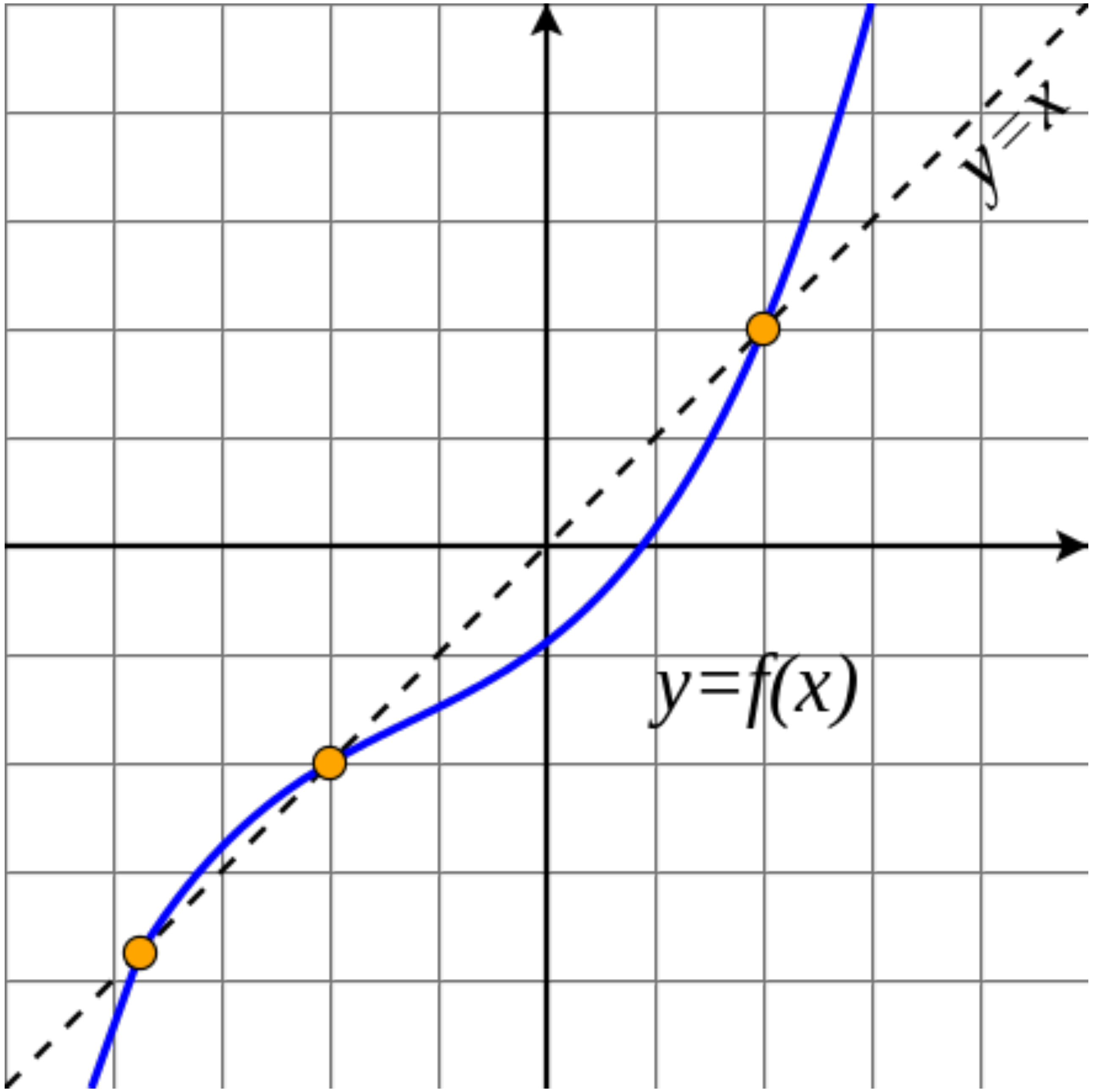
$$P = \begin{pmatrix} 0.4 & 0.6 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0.4 & 0.6 \\ 0 & 0 & 0.2 & 0.8 \end{pmatrix}$$

Two left eigenvectors of eigenvalue 1:

$$\begin{aligned} \pi_1 &= (1/4, 3/4, 0, 0)\,, \\ \pi_2 &= (0, 0, 1/4, 3/4) \end{aligned}$$

depending on initial state we get a different stationary distribution.

Intuition



Def 32

- **Definition:** A Markov chain is said to be irreducible if all the states communicate with each other, that is $\forall x, y \in \mathbb{X}$ $\inf \left\{ t : K_{xy}^{t} > 0 \right\} < \infty$.

- **Definition:** An irreducible Markov chain is aperiodic if there exists $x \in \mathbb{X}$ such that

$$\gcd \{ s \geq 1 : K_{xx}^{s} > 0 \} = 1$$

where gcd denotes the greatest common divisor.

- *Example*: $K_\theta = \left(\begin{array}{ll} \theta & 1-\theta \\ 1-\theta & \theta \end{array} \right)$ is irreducible if $\theta \in [0,1)$ and aperiodic if $\theta \in (0,1)$. If $\theta = 0$, the gcd is 2.

Prop 33

- **Proposition**: If a finite state-space Markov chain is irreducible then it has a unique stationary distribution and

$$\widehat{I}_n := \frac{1}{n} \sum_{t=1}^{n} \phi\left(X_t\right) \to I := \sum_{x \in \mathbb{X}} \phi\left(x\right) \pi(x). \quad \text{a.s.}$$

- **Proposition**: If a finite state-space Markov chain is irreducible and aperiodic, then there exists $0 \leq \alpha < 1$ such that

$$\frac{1}{2} \left|\mathbb{P}(\, X_t = x |\, X_1) - \pi(x)\right| \leq \alpha^t .$$

- *Remark*: Aperiodicity is not required for the averages to converge to the expectation; e.g. take K_0.

This result (convergence of marginals) is not as useful to us directly (but it is as an intermediate result in proofs)

A few more definitions

- *Stationarity*: marginal distributions are all equal
  - I.e. chain is initialized at a stationary distribution

LLN for MC: intuition

Mixing

Prop 35

- How fast is the chain forgetting about its initalization.
- Many definitions. One of them is *rho-mixing*:

$$\rho_n = \sup_{f,g \in L_2(\pi)} \left| \mathrm{cor}\big(f(X_k), g(X_{k+n})\big) \right| \to 0$$

- Following result from Roberts and Rosenthal 1997 (Thm 2.1) is useful:
  - Every *reversible*, *geometrically ergodic*, *stationary* Markov chain is rho-mixing
  - We will define geometric ergodicity later, for now just use it holds for irreducible finite chains
- Also (Bradley, 1986, Thm 4.2): if the chain is stationary then convergence is at an exponential rate

Prop 36

Consider a stationary chain:

$$\lim_{n\to\infty} n\mathbb{V}_\pi\left(\widehat{I}_n\right) = \mathbb{V}_\pi\left(\phi\left(X_1\right)\right) + 2\sum_{k=1}^{\infty} \underbrace{\mathbb{C}\mathsf{ov}_\pi\left(\phi\left(X_1\right), \phi\left(X_{k+1}\right)\right)}_{:=C(k)}$$

Proof: We have $\mathbb{E}_\pi\left(\widehat{I}_n\right) = I$ and

$$n\mathbb{V}_\pi\left(\widehat{I}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \underbrace{\mathbb{C}\mathsf{ov}_\pi\left(\phi\left(X_i\right), \phi\left(X_j\right)\right)}_{=C(i-j)}$$

$$= \frac{1}{n}\sum_{k=-n+1}^{n-1} C\left(k\right) \times \underbrace{\left(\#\ \text{pairs} : i - j = k\right)}_{=n-|k|}$$

$$= \sum_{k=-n+1}^{n-1}\left(1 - \frac{|k|}{n}\right) C\left(k\right) = \sum_{k=-\infty}^{\infty} \max\left(0, 1 - \frac{|k|}{n}\right) C\left(k\right)$$

CLT for MC: intuition

Why we need a CLT

- As before with IS, we want:
  - to determine when we have enough samples
  - to compare the running time of competing methods

Exercise 34

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

Prop 37

CLT (discrete version)

- For irreducible Markov chains with stationary distribution π:

$$\lim \sqrt{t}\left[\frac{1}{t}\sum_{i=1}^{t}\phi\left(X_i\right)-\int_{\mathbb{X}}\phi\left(x\right)\pi\left(x\right)dx\right]\overset{D}{\to}\mathcal{N}\left(0,\sigma^2\left(\phi\right)\right)$$

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

Metropolis-Hastings

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 π**
    - Even if we do not know the normalization of π

Metropolis-Hastings (MH)

- Idea: start with a transition probability q(x'|x) called the proposal
  - This defines a Markov chain, but it is not π-invariant
- Transform it to get a new Markov chain with transition probability K which is π-invariant
  - Surprisingly, all we have to do is move some mass towards self-transition!

Def 38

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

Do we need the normalization of π? *q*?

1. Sample $X^\star \sim q\left(\cdot \middle| X^{(t-1)}\right)$.
2. Compute

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

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

Def 39

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

- Pick arbitrary initial X_1
- Simulate a Markov chain $X_1, X_2, X_3, \ldots$
- 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

Optional tweaks

- Remove a prefix of the sequence (burn-in)
  - Not necessary by LLN
  - Heuristic can be useful when initialization is poor (but there are ways to get good initialization. Hint: don’t use MAP!)
- Take one out of every *k* samples (thinning)
  - Again, not necessary by LLN
  - Only good reason to do it is for memory/storage reasons
  - But often, used because of misunderstanding of theory

Example 40

Examples

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

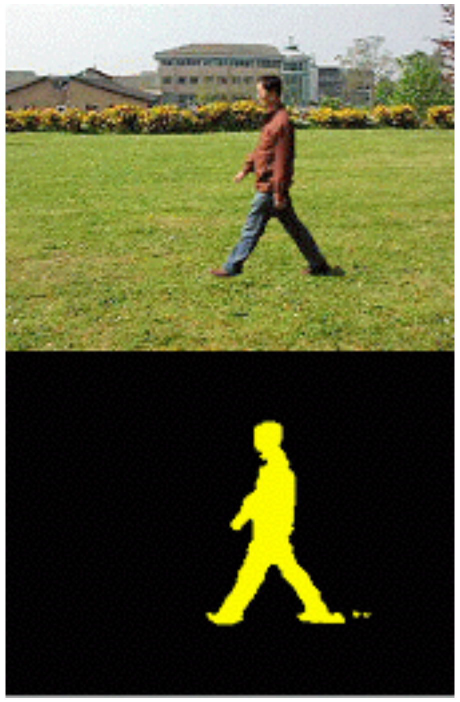
Example, continued

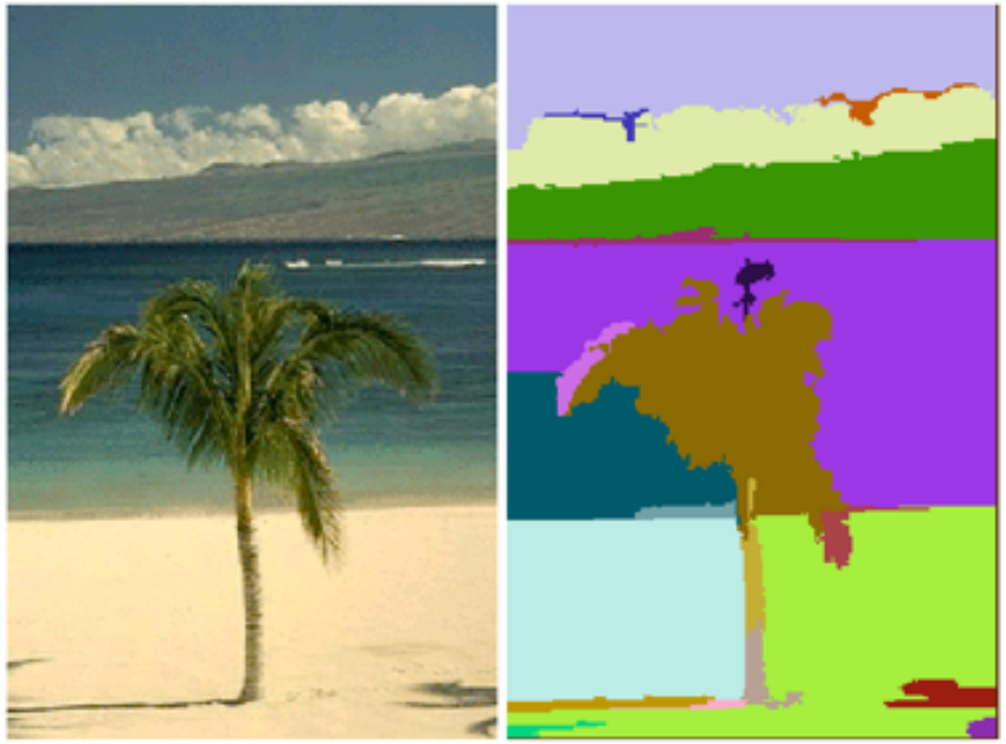
- Let’s look at the details of a Gibbs sampler
- How to sample from Markov Random Field?

Motivation

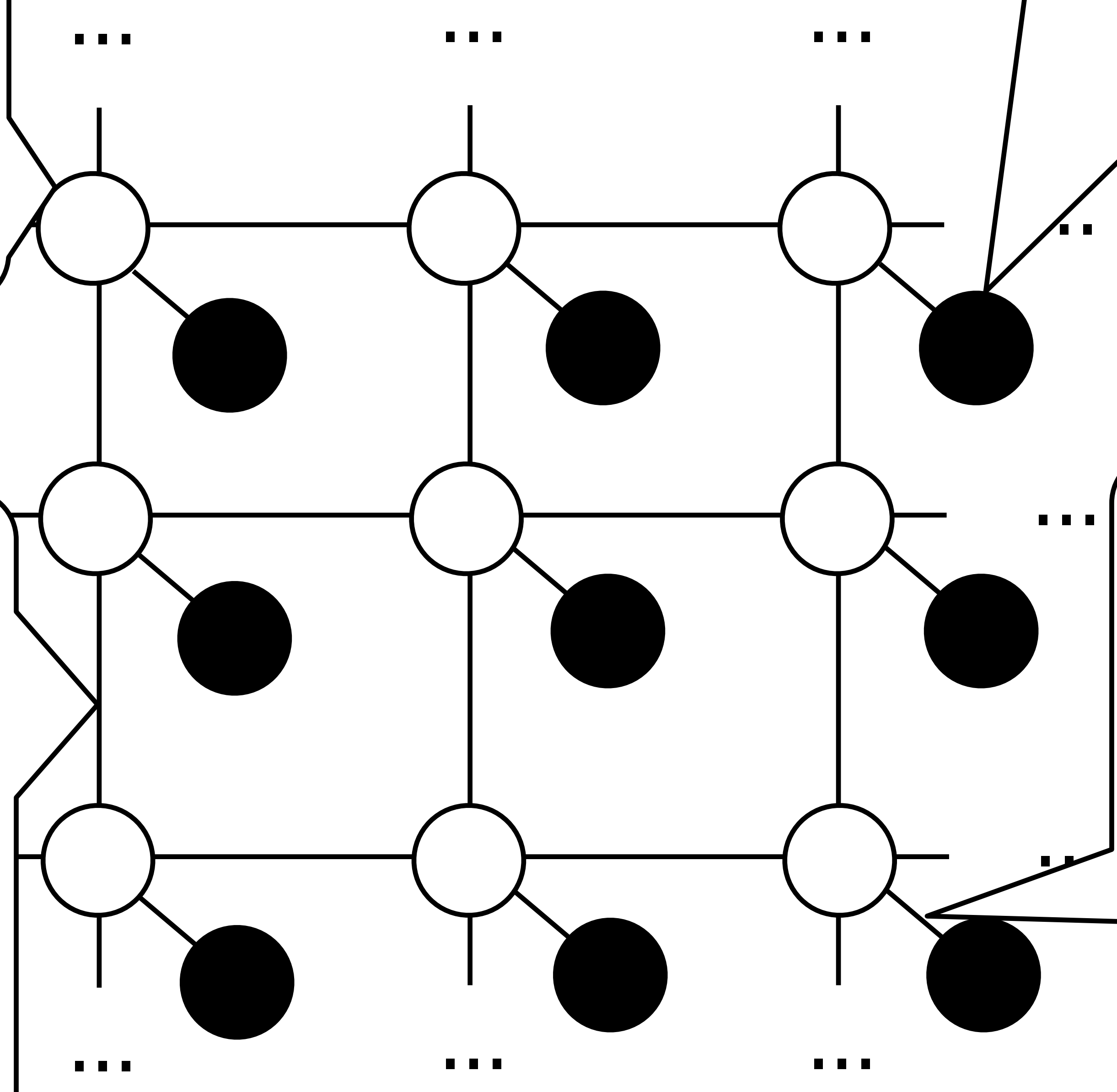
Task: given some images (a 2D array of pixels), segment it into clusters of pixels

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) ?

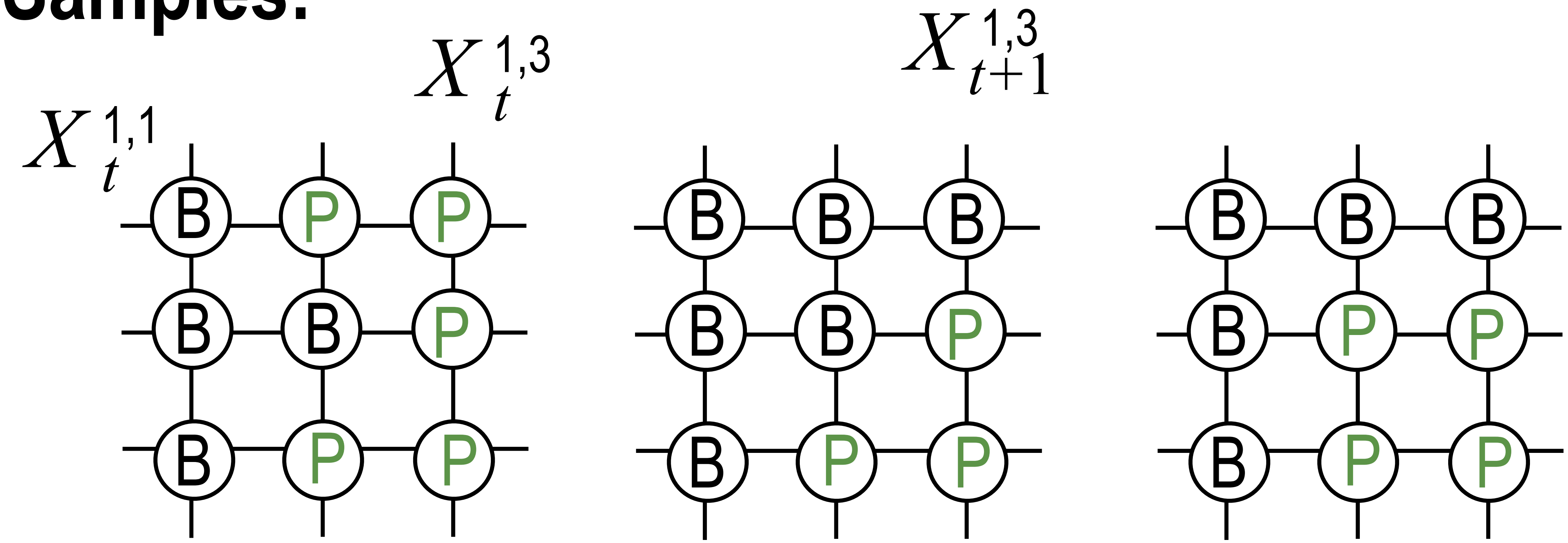
RGB value of the pixel

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:

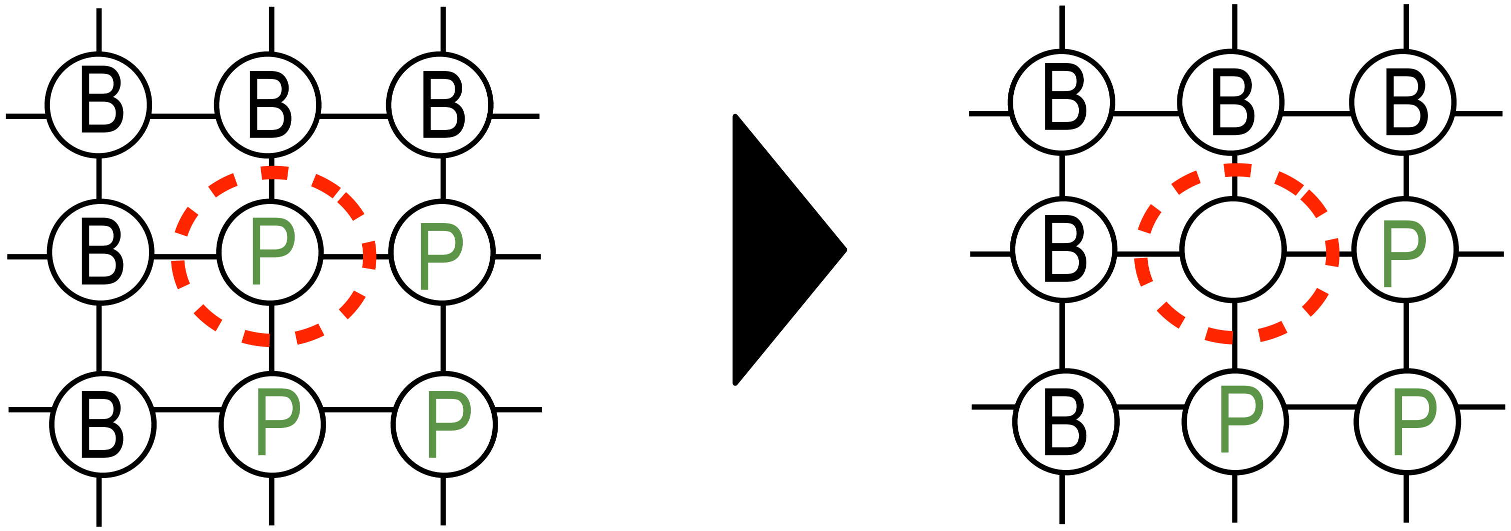


Monte Carlo estimator: for S samples

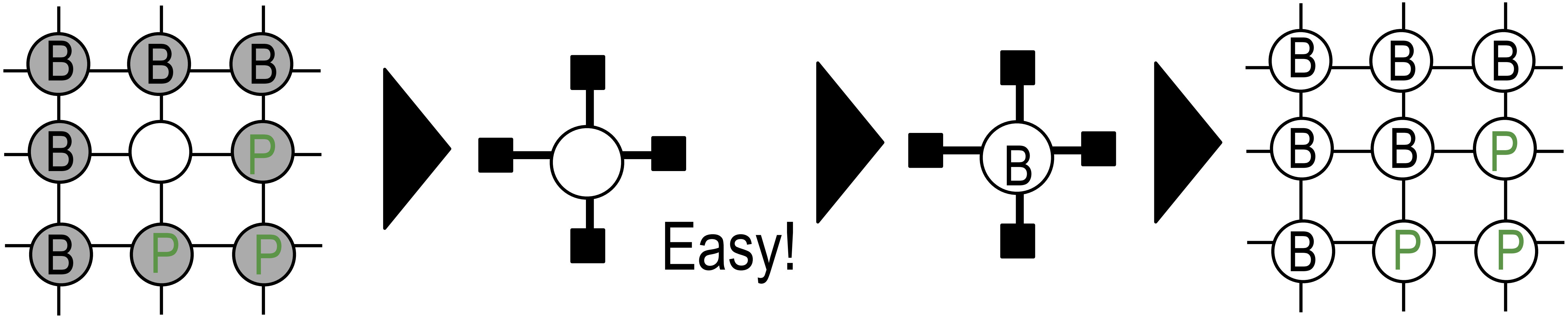
$$\mathbb{E}f(X) \approx \frac{1}{S}\sum_{t=1}^{S} f(X_t)$$

‘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

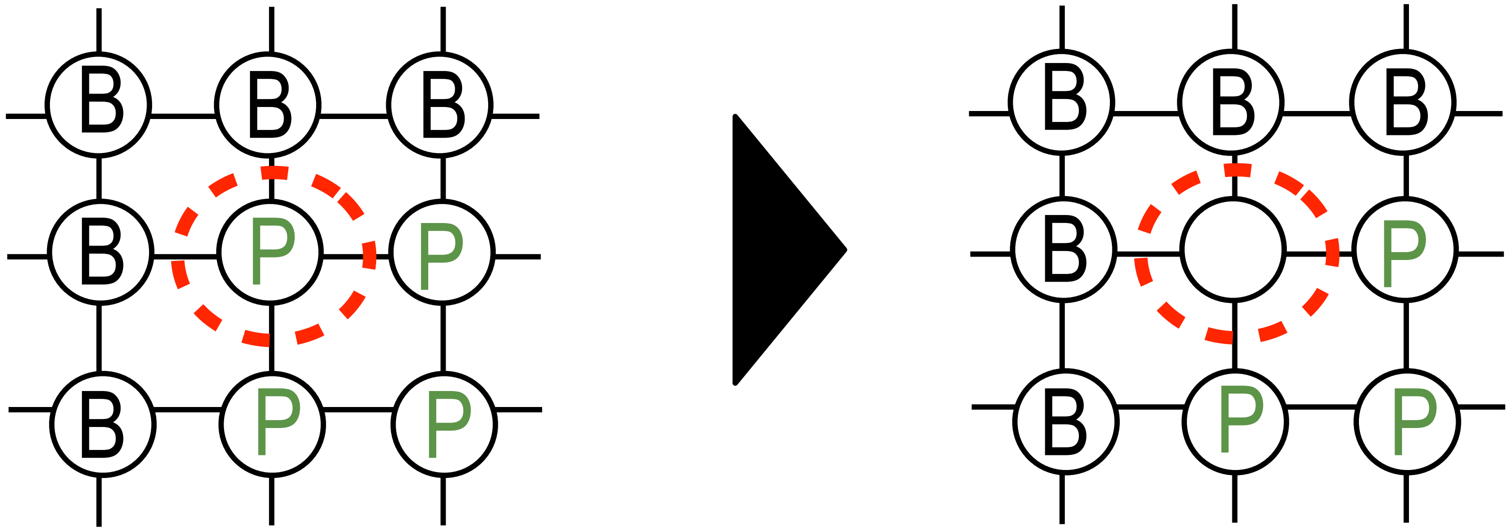


Then: resample a value for the node *(i,j)* conditioning on all the others, and write this to the current state at *(i,j)*



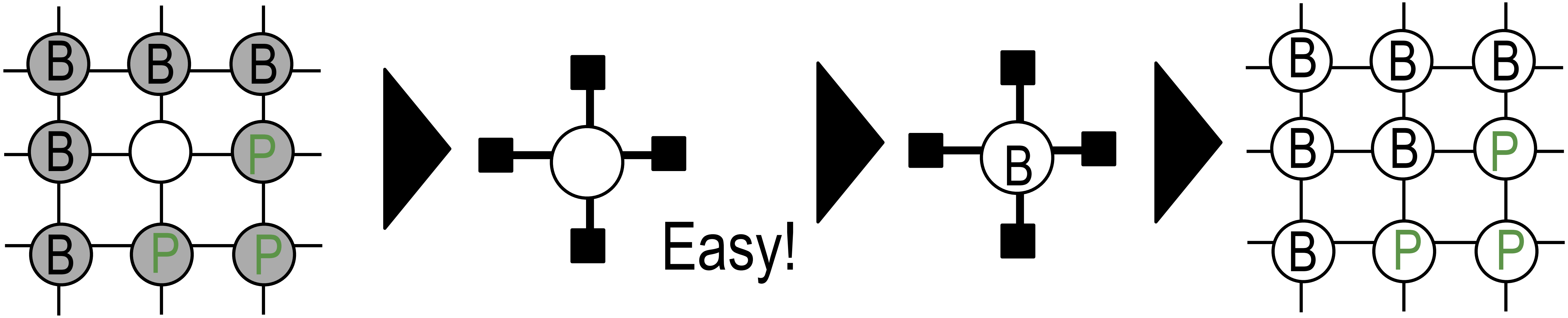
Starting with a simpler version

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

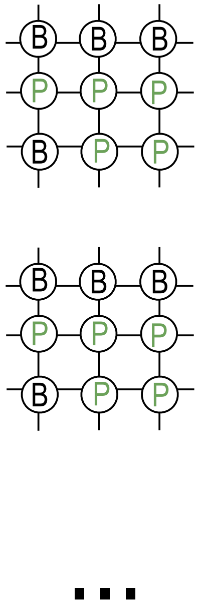
Then: resample a value for the node *(i,j)* conditioning on all the others, and write this to the current state at *(i,j)*

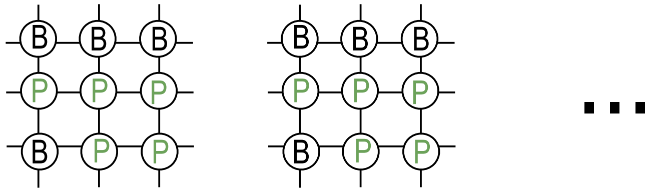


First example: Gibbs construction

Transition matrix of the Gibbs sampler: 2^9 x 2^9 matrix

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





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

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

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

Exercise

Exercise 42

- 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

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

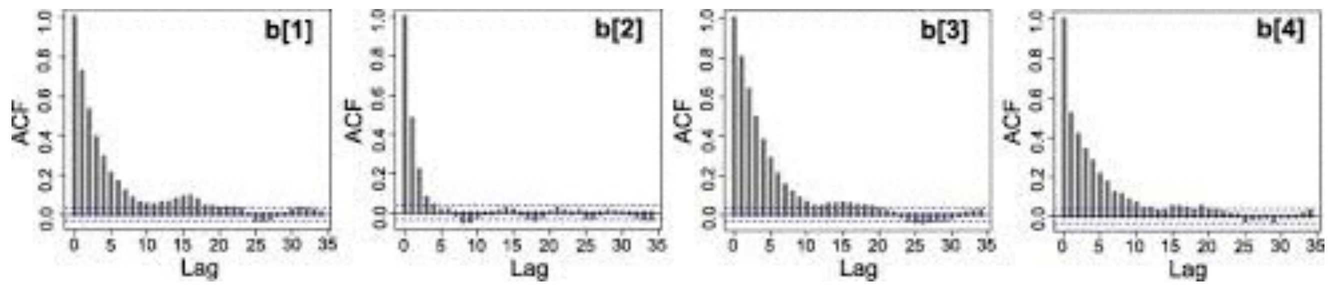
$$I_n \pm 1.96\sqrt{\sigma^2_{\text{asympt}}/n}$$

- The asymptotic variance is:

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

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 (factoid: should be positive for reversible processes, so a heuristic is to truncate when negative-can still be unstable).

ESS for MC

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

$$\frac{\mathrm{ESS(N)}}{N} \to \frac{\sigma^2_{\mathrm{asymptotic}}}{\sigma^2_{\mathrm{iid}}}$$

- 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

Analysis of MH

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

Prop 44

Invariance of a single kernel

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

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

  where δ_x denotes the Dirac mass at x.

- *Proof*. We have

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

Invariance of a single kernel

- **Proposition**. The Metropolis-Hastings kernel K is $\pi-$reversible and thus admit π as invariant distribution.

- *Proof.* For any $x, y \in \mathbb{X}$, with $x \neq y$

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

Exercise

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

Important, overlooked condition on proposal *q*

Def 47

- Mutual absolute continuity condition:

$$\int_A \pi(\mathrm{d}x) q(x, B) > 0 \Leftrightarrow \int_B \pi(\mathrm{d}x) q(x, A) > 0$$

- For example, for discrete state space where the target is positive:

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

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