Def. Metropolis-Hastings Algorithm

- Start with a candidate transition matrix K
- To ensure unique stationary distribution, it is sufficient to have
  - $K_{xx} > 0$, $\forall x \in \mathcal{X}^n$ [aperiodic]
  - for any $x, y \in \mathcal{X}^n$, $\exists$ a path $(x_1 = x, x_2, \cdots, x_m = y)$ of positive probability transitions [irreducibility]

$$K_{x_i x_{i+1}} > 0, \quad \forall i \in \{1, \cdots, m-1\}$$

what we want

$$(*) \quad Q_{xy} \cdot P(x) = Q_{yx} P(y)$$

what we have

$$K_{xy} P(x) \underset{\text{w.l.o.g}}{>} K_{yx} P(y)$$

the main trick is to remove some "probability mass" from the larger one.

Define: $R_{xy} \triangleq \min\left\{ 1, \dfrac{P(y) K_{yx}}{P(x) K_{xy}} \right\}$ for each x, y

$$Q_{xy} = K_{xy} \cdot R_{xy}$$

$$Q_{xx} \underset{\substack{\text{s.t.} \\ \text{probability sums to one}}}{=} 1 - \sum_{y \neq x} Q_{xy}$$

with this rejection sampling $(1 - R_{xy})$

claim: $(Q, P(x))$ satisfy $(*)$.

proof>
$$\begin{aligned} p(x) \cdot Q_{xy} &= P(x) \cdot R_{xy} \cdot K_{xy} \\ &= P(y) K_{yx} \qquad (\text{if } R_{xy} < 1 \ \& \ R_{yx} = 1) \\ &= P(y) R_{yx} \cdot K_{yx}. \end{aligned}$$

* Remark: as $P(x)$ is only used in $R_{xy} = \min\left\{1, \frac{P(y)K_{yx}}{P(x)K_{xy}}\right\}$

we only need $\frac{P(y)}{P(x)} = \prod_{(i,j)\in E} \frac{f_{ij}(y_i,y_j)}{f_{ij}(x_i,x_j)}$ ← takes $O(|E|)$ computation
└ does not require Z.

* Remark: Do we need to store K and $Q \in \mathbb{R}^{|\mathcal{X}|^n \times |\mathcal{X}|^n}$

- We can choose K to be simple, such as $K = \frac{1}{|\mathcal{X}|^n}\cdot \mathbb{1}\mathbb{1}^T$

Step1 - At time t first generate candidate $x^{t+1/2}$ from $K(x^t, x^{t+1/2})$

Step2 - Accept $x^{t+1/2}$ with probability $R_{x^t, x^{t+1/2}}$: $x^{t+1} = x^{t+1/2}$

Step3 - Otherwise reject and keep current state : $x^{t+1} = x^t$

* Theorem: Metropolis-Hastings Algorithm finds ℓ_1-projection of K onto the space of reversible Markov Chains with stationary distribution $P(x)$.

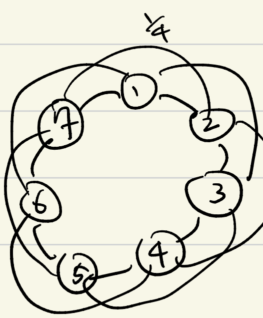
$$Q_{MH} = \underset{\tilde{Q} \text{ s.t. } P^T\tilde{Q} = P^T}{\arg\min} \sum_x \sum_{y \neq x} \left| P(x)\cdot K_{xy} - P(x)\tilde{Q}_{xy} \right|$$

* Are we done? - the art is in choosing K.

if the spread of K is too large, then acceptance rate is low

if the spread of K is too narrow, then mixing time can be large

example> $K = \frac{1}{|\mathcal{X}|^n}\mathbb{1}\mathbb{1}^T$, $R_{xy} = \min\left(1, \prod_{(i,j)\in E} \frac{f_{ij}(y_i,y_j)}{f_{ij}(x_i,x_j)}\right)$



· all pairs first sampled with equal probability (as per K)

· but many candidates might be unlikely and be rejected.

* Def. Gibbs Sampling.

$Q \begin{cases} \text{Step 1. sample } i \in \{1, \cdots, n\} \text{ uniformly at random.} \\ \text{Step 2. set } \tilde{X}_{-i} = X^{(t)}_{-i} \quad (-i = [n] \setminus \{i\}) \\ \text{Step 3. sample } \tilde{X}_i \text{ from } P(\tilde{X}_i \mid X^{(t)}_{-i}) \end{cases}$

step 4. $X^{(t+1)} \leftarrow \tilde{X}$

* Remark. $P(\tilde{X}_i \mid X^{(t)}_{-i}) \propto \prod_{j \in \partial i} f_{ij}(\tilde{X}_i, X^{(t)}_j)$ is efficient to compute

* claim. $(Q, P(x))$ satisfy $(*)$.

proof. for $\tilde{x}$ that differ at only i-th coordinate from x,

$$P(x) \cdot Q_{x\tilde{x}} = P(x) \cdot \frac{1}{n} \cdot P(\tilde{x}_i \mid X_{-i})$$

$$\text{Bayes} \rightarrow = P(X_i \mid X_{-i}) P(X_{-i}) \cdot \frac{1}{n} P(\tilde{x}_i \mid x_{-i})$$

$$= \underbrace{P(X_i \mid X_{-i}) \frac{1}{n}}_{Q_{\tilde{x}x}} \underbrace{P(X_{-i}) P(\tilde{x}_i \mid X_{-i})}_{P(\tilde{x})}$$

Otherwise $Q_{x\tilde{x}} = 0$ if x & $\tilde{x}$ differ more than one coordinate.

* the resulting dynamics of the Markov chain is called Glauber Dynamics.

examples> Glauber dynamics for proper k-coloring

Proper Coloring: given a graph G, and K-colors $\{1, 2, \dots, k\}$.
a coloring is an assignment of colors to nodes
$X = [X_1, \dots, X_n]$, $X_i \in \{1, \dots, k\}$.
a proper coloring is a coloring where all pairs of nodes connected by an edge have different colors.

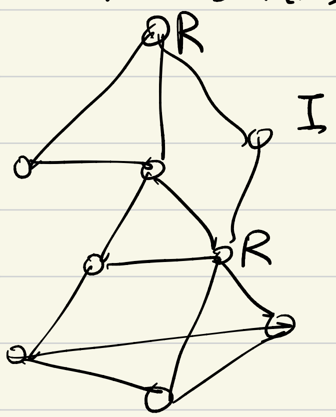
Graphical Model: $P(X) = \frac{1}{Z_K} \prod_{(i,j) \in E} \mathbb{I}(X_i \neq X_j)$

Z_K ← # of proper colorings

we propose Gibbs Sampling to solve, for example, $\mathbb{P}(X_1 = X_2), \dots$

Gibbs Sampling: Initialize with $X_0 \in [K]^n$ that is proper

$K=3$ $\{R, G, B\}$



Repeat
Sample $I \in [n]$ uniformly at random
Sample a color $X_I^{(t+1)}$ uniformly s.t.
$X_I^{(t+1)} \neq X_{\partial I}^{(t)}$

Consider Metropolis Hastings with $K = \frac{1}{|\mathcal{X}|^n} \mathbb{1}\mathbb{1}^T$,
each time a random configuration is proposed and most likely rejected. It is very slow as rejection rate is high.

example > Ising models.

$$P(x) = \frac{1}{Z} \exp\{ \underset{\substack{\uparrow \\ \text{inverse temperature}}}{\beta} \sum_{ij \in E} \theta_{ij} X_i X_j \}$$

$X_i \in \{\pm 1\}$

$\beta \uparrow$ clustered.

$\beta \downarrow$ random

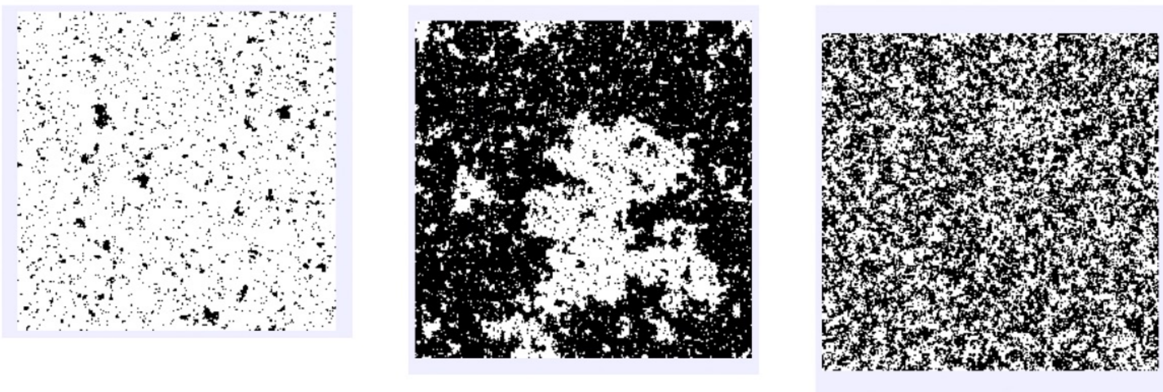


FIGURE 3.2. Glauber dynamics for the Ising model on the 250 × 250 torus viewed at times $t = 1{,}000$, 16,500, and 1,000 at low, critical, and high temperature, respectively. Simulations and graphics courtesy of Raissa D'Souza.

* Mixing Time of a Markov Chain.

Def. <u>ε-mixing time</u> of a Markov Chain Q

is the smallest time $T_{mix}(\varepsilon)$ such that for all $t > T_{mix}(\varepsilon)$ and for all initial state $\xi^{(0)}$ (a distribution over $\mathcal{X}^n$)

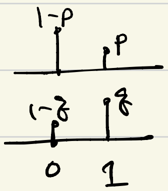
$$\left| \left(\xi^{(0)}\right)^T \cdot \underbrace{Q^t}_{Q \cdot Q \cdots Q} - \pi^T \right|_{TV} \le \varepsilon$$

π^T: stationary distribution

where $|p - \xi|_{TV} = \frac{1}{2} \sum_{x \in \mathcal{X}^n} |p(x) - \xi(x)|$ is the <u>total variation distance</u>.

$$= \sum_{x \in \mathcal{X}^n} \underbrace{[P(x) - \xi(x)]^+}_{\max\{0,\, P(x) - \xi(x)\}}$$

ex>



Bern(p)

Bern(ξ)

$|p - \xi|_{TV} = |p - \xi|$

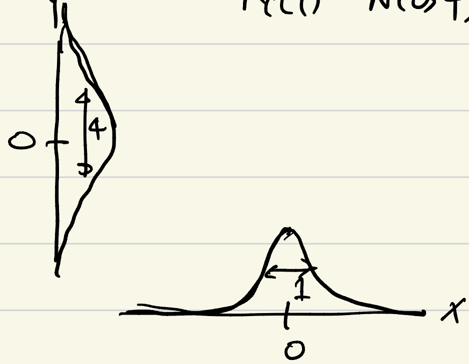
* Bounding mixing time via <u>Coupling</u>

Def. a coupling of two random variables X and Y with distributions $P_X(x)$ and $P_Y(y)$, is a construction of a joint probability distribution over (X, Y), i.e., $P_{XY}(x, y)$ such that the marginals are preserved:

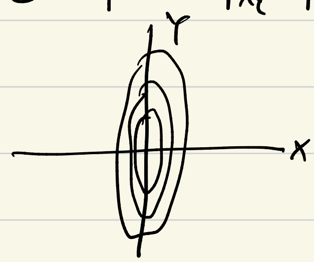
$$\begin{cases} \sum_{y} P_{XY}(x, y) = P_X(x) \\ \sum_{x} P_{XY}(x, y) = P_Y(y) \end{cases}$$

example > $P_X(x) \sim N(0,1)$

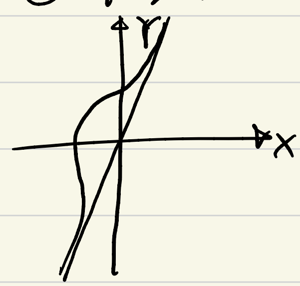
$P_Y(y) \sim N(0,4)$



① independent: $P_{XY} = P_X \cdot P_Y$



② $Y = 2X$



example> $P_X \sim \mathrm{Bern}(p)$

$P_Y \sim \mathrm{Bern}(q)$

$$P_Y \begin{bmatrix} 1-q \\ q \end{bmatrix}$$

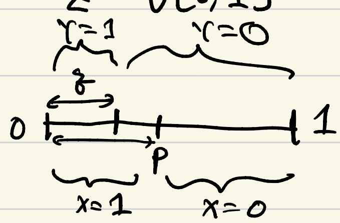
$$P_X \,[1-p,\ p]$$

① indep.

$$P_{XY} \begin{array}{c} \\ \\ \end{array}\begin{array}{cc} \overset{X}{0} & 1 \\ \end{array}$$
$$\begin{array}{cc|c} & & Y \\ \bar{p}\bar{q} & p\bar{q} & 0 \\ \bar{p}q & pq & 1 \end{array}$$

② construct coupling from $U[0,1]$

$Z \sim U[0,1]$



Def. <u>optimal coupling</u> for $X, Y \in \mathcal{X}$

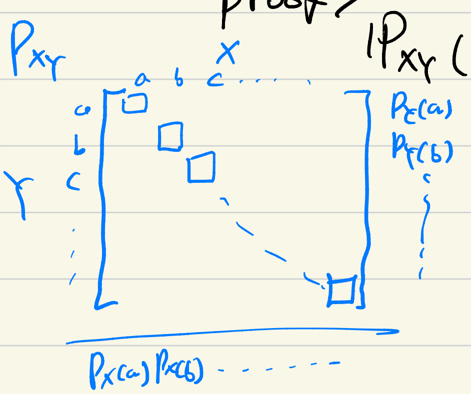
$$\min_{\text{couplings of } P_X, P_Y} \mathbb{P}_{XY}(X \neq Y)$$

* Coupling lemma:

for two random variables X, Y (either continuous or discrete) in the same domain,

$$|P_X - P_Y|_{TV} = \min_{\text{couplings of } P_X, P_Y} \mathbb{P}_{XY}(X \neq Y)$$

proof>



$$\mathbb{P}_{XY}(X \neq Y) = 1 - \sum_x \overbrace{P_{XY}(x,x)}^{\leq \min\{P_X(x),\, P_Y(x)\}}$$

$$\geq \sum_x \left\{ P_X(x) - \min\{P_X(x), P_Y(x)\} \right\}$$

$$= \sum_x \max\{0,\ P_X(x) - P_Y(x)\}$$

$$= |P_X - P_Y|_{TV}$$

* further, exists $P_{XY}(x,y)$ s.t. minimum is achieved.

* Corollary: $\| P_X - P_Y \|_{TV} \leqq \mathbb{P}_{XY}(X \neq Y)$ for any coupling.

• any coupling can be used to upper bound TV-distance of two distributions.

ex2. optimal coupling of Bernoulli distributions.

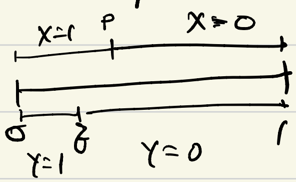
$P_X \sim Bern(p)$ $\quad p > q.$

$P_Y \sim Bern(q)$

$$P_{XY} = \begin{array}{c|cc|c} & \multicolumn{2}{c}{X} & \\ & 0 & 1 & \\ \hline Y \; 0 & 1-p & p-q & \\ Y \; 1 & 0 & q & \end{array}$$

$TV = |p-q|$

You can sample $\xi \sim U[0,1]$



* Coupling for bounding $T_{mix}(\varepsilon)$ of Gibbs Sampling.

Strategy: let X_t, Y_t be the random state after t transitions as per Q, started with X_0, Y_0, respectively.

$$\| P_{X_t} - \pi \|_{TV} \leqq \max_{P_{X_0}, P_{Y_0}} \| P_{X_t} - P_{Y_t} \|_{TV}$$

$$\overset{\text{coupling lemma}}{\leqq} \max_{P_{X_0}, P_{Y_0}} \mathbb{P}(X_t \neq Y_t).$$

for any coupling.

We will analyze a coupling that is simple enough for analysis, and yet gives a tight upper bound.

Proposed coupling of two Gibbs sampling chains. for $x, y \in \{0,1\}^n$

couple two processes closely, while preserving marginal distribution.

X_0 $\quad\quad$ Y_0

Step1. draw $I \in \{1, \cdots, n\}$ random index

Step2. optimal coupling of $P(X_I^{(t+1)} | X_{-I}^{(t)})$ $\quad$ $P(Y_I^{(t+1)} | Y_{-I}^{(t)})$