Given a description of $P(x)$ that we want to sample from

$$P(x) = \frac{1}{Z} \prod_{(i,j)\in E} f_{ij}(x_i, x_j)$$

Construct a Markov Chain $Q : \mathcal{X}^n \to \mathcal{X}^n$ s.t. stationary distribution $\pi = P(x)$

Metropolis-Hastings

select $K : \mathcal{X}^n \to \mathcal{X}^n$

create $R \in \mathbb{R}^{|\mathcal{X}|^n \times |\mathcal{X}|^n}$, $R_{xy} \triangleq \min\left\{1, \frac{\overset{\text{given}}{P(y)}\, \overset{\text{chosen}}{K_{yx}}}{P(x)\, K_{xy}}\right\}$

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

Detailed Balanced equation.

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

implies that stationary dist. of Q is $P(x)$.

Remark 1. M.H. is efficient.

R_{xy} only depends on the ratio of $\frac{P(y)}{P(x)} = \frac{\overbrace{\prod_{(i,j)\in E} f_{ij}(y_i, y_j)}^{O(|E|)}}{\prod_{(i,j)\in E} f_{ij}(x_i, x_j)}$

does not require Z.

Remark 2. We do not need to store $K, R \in \mathbb{R}^{|\mathcal{X}|^n \times |\mathcal{X}|^n}$

- first we choose K that is compact. $K = \frac{1}{|\mathcal{X}|^n} \cdot \mathbb{1}\mathbb{1}^T$

Step 1. at time t, generate Candidate $x^{(t+\frac{1}{2})}$ from $K_{x^{(t)},\cdot}$

step 2. ACCEPT $x^{(t+\frac{1}{2})}$ with probability $R_{x^{(t)}x^{(t+\frac{1}{2})}}$

$$x^{(t+1)} \leftarrow x^{(t+\frac{1}{2})}$$

Otherwise, REJECT

$$x^{(t+1)} \leftarrow x^{(t)}$$

* Theorem: Metropolis-Hastings Algorithm finds L_1-projection of K onto the space of transition matrices that have stationary distribution $P(x)$.

$$Q_{MH} = \underset{\tilde{Q}}{\arg\min} \sum_x \sum_{y \neq x} \underbrace{\left| P(x) K_{xy} - P(x) \tilde{Q}_{xy} \right|}$$

s.t. $P^T \tilde{Q} = P^T$

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

  if K is too spread out (like $K = \frac{1}{|\mathcal{X}|^n} \mathbb{1}\mathbb{1}^T$)

  $\rightarrow$ rejection rate is large

  if K is too narrow

  $\rightarrow$ takes long time to explore, mixing time $\uparrow$.

· Def. Gibbs Sampling $\qquad P(x)$ defined on $\mathcal{X}^n$

$x^{(t)} = x_1^{(t)}, x_2^{(t)}, \dots, x_I^{(t)}, \dots, x_n^{(t)}$

Repeat

step 1: sample $I \in \{1, \dots, n\}$ uniformly random

step 2: set $\tilde{x}_{-I} = x_{-I}^{(t)}$

$\tilde{x}_{-I}$: $\{1, \dots, n\} \setminus \{I\}$

step 3: sample $\tilde{x}_I$ from $P(\tilde{x}_I \mid x_{-I}^{(t)})$

$\tilde{x}_I \in \mathcal{X}$

knowledge of $P(x)$

step 4: $x^{(t+1)} \leftarrow \tilde{x}$

*Remark: $P(\tilde{x}_I \mid x_{-I}^{(t)}) \propto \prod_{j \in \partial I} f_{Ij}(\tilde{x}_I, x_j^{(t)}) \leftarrow O(\text{degree} I)$

$|\mathcal{X}|$ many evaluations.

Claim: $(\overset{\text{Gibbs}}{\downarrow}{Q}, P(x))$ satisfy <u>detailed balance equation.</u>

$$Q_{xy}\,P(x) = Q_{yx}\,P(y).$$

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

x, y differ in coordinate i

$$\overset{\text{Bayes}}{=} P(x_i \mid X_{-i})\,P(X_{-i})\,\frac{1}{n}\cdot P(y_i \mid X_{-i})$$

$$= \underbrace{P(y_i \mid \overset{Y_{-i}}{\overset{\shortparallel}{X_{-i}}})\,P(\overset{Y_{-i}}{\overset{\shortparallel}{X_{-i}}})}_{P(Y)} \cdot \underbrace{\frac{1}{n} P(X_i \mid \overset{Y_{-i}}{\overset{\shortparallel}{X_{-i}}})}_{Q_{yx}}$$

$$= P(Y)\,Q_{yx}$$

* The resulting dynamics Glauber Dynamics.

example> Glauber Dynamics for <u>Proper K-coloring</u>

proper coloring: given a G, K-colors $\{1, \dots, K\}$

coloring is $X = (X_1, \dots, X_n) \in \{1, \dots K\}^n$

proper coloring X s.t.

$$X_i \neq X_j, \quad \forall (i,j) \in E$$

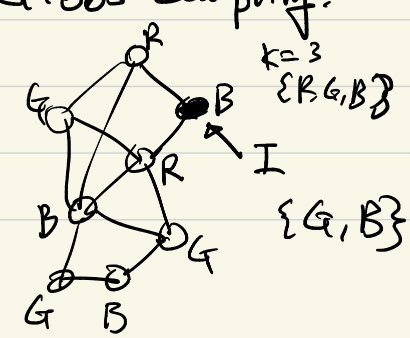
Graphical Model:

$$P(X) = \frac{1}{Z_K} \prod_{(i,j)\in E} \mathbb{I}(X_i \neq X_j) \quad \leftarrow \text{lots of Zeros.}$$

Z_K: # possible proper coloring

to solve inference problems for K-coloring. eg> $\mathbb{P}(X_1 = X_3)$?

Gibbs Sampling:

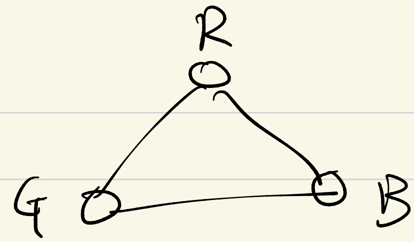


Initialize with a proper color $X^{(0)}$ ⇐ easy if $K \geq d_{max}$

Repeat

✓ sample $I \in [n]$ uniform.

✓ sample $P_{X_I \mid X_{-I}^{(t)}}$ = uniform over colors that are not taken by its neighbors.



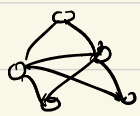
comparison > M.H.

$K = \frac{1}{|\chi|^n} \mathbb{1}\mathbb{1}^T$ → explore all colorings, and most will be rejected.

If inference task compute Z

Sampling. $\frac{1}{N}\sum^{N} \mathbb{I}(x_i=0)$

If you have a black box for marginalization. $P(X_i)$



$$Z = \sum_X f(x)$$

$$\boxed{P(x_i) = \frac{1}{Z}\sum_{x_{-i}} f(x_i, x_{-i})}$$

$X^{(1)} \sim P(x)$ after mixing time

N

$X^{(N)}$

$N \cong \frac{1}{\varepsilon^2}$ $\varepsilon \sim$ accuracy.

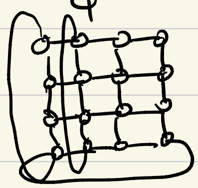
example > Ising models.

$x_i \in \{\pm 1\}$

$$P(x) = \frac{1}{Z} \prod_{(i,j)\in G} \exp(\beta \cdot \theta_{ij} x_i x_j)$$

inverse temperature

G



$\beta = 0$, uniform Random

$\beta \uparrow$, cluster.

$\beta \uparrow$

$\beta_{critical}$

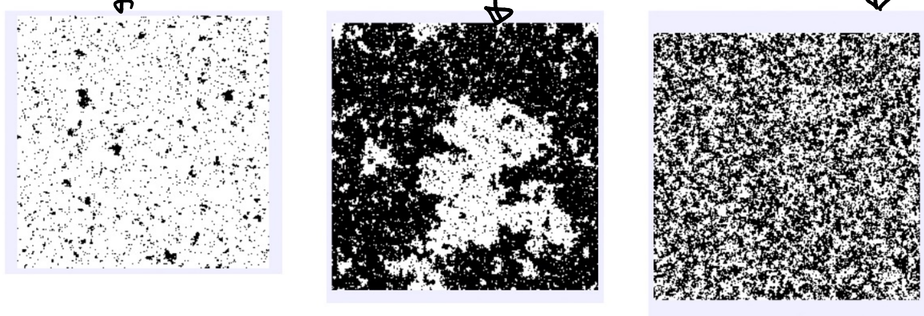


FIGURE 3.2. Glauber dynamics for the Ising model on the 250 × 250 torus viewed at times $t = 1{,}000$, 16,500, and 1,000 at low, critical, and high temperature, respectively. Simulations and graphics courtesy of Raissa D'Souza.

* Spread of Innovations.
[Montanari, Saberi 2010]

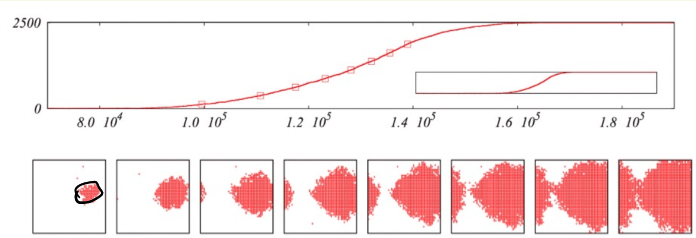


Fig. 4. Diffusion of the risk-dominant strategy in a small-world network with $d = 2$, $k = 3$, $n = 50^2$ and mostly short-range connections, namely $r = 5$. We use $\beta = 0.75$ and $h = 0.5$. (*Upper*) Evolution of the number of nodes adopting the new strategy (in the inset, same curve with time axis starting at $t = 0$). (*Lower*) Configurations at times such that the number of nodes adopting the risk-dominant strategy is, from left to right, 125, 375, 625, 875, 1125, 1375, 1625, 1875 (indicated by squares in *Upper*).

Incentive

$$\begin{array}{c c} & \begin{array}{cc} A & S \end{array} \\ \begin{array}{c} A \\ S \end{array} & \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix} \end{array}$$

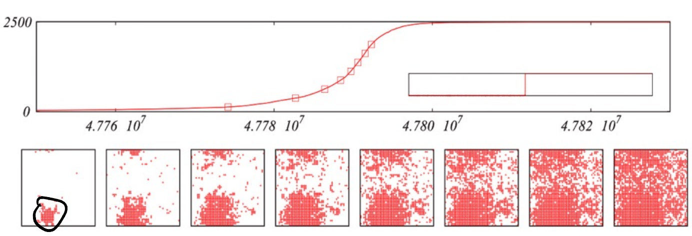
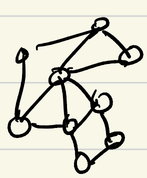


Fig. 5. As in Fig. 4 but for a small-world network with a larger number of long-range connections, namely $r = 2$.

G



Model as Glauber Dynamics.

* Mixing Time.

Def. ε-mixing time $T_{mix}(\varepsilon)$ of a Markov Chain Q.

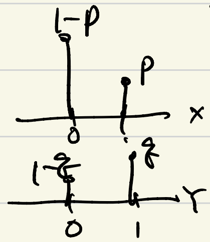
is the smallest time $T_{mix}(\varepsilon)$ such that for all $t > T_{mix}$ and for all initializations $q^{(0)}$, where $(q^{(t)})^T = (q^{(0)})^T Q^t$ is the dist. at time t.

$$|q^{(t)} - \underset{\substack{\uparrow \\ \text{Stationary dist. of } Q.}}{\pi}|_{TV} \le \varepsilon$$

Total variation distance: $|p - q|_{TV} = \frac{1}{2} \sum_x |p(x) - q(x)| = \sum_x \underset{\substack{\uparrow \\ \max\{0,\, p(x)-q(x)\}}}{[p(x) - q(x)]^+}$

ex> $p \sim Bern(p)$

$q \sim Bern(q)$



$|p - q|_{TV} = |p - q|$

Def. a Coupling of two random variables X and Y with marginal distributions $P_X(X)$ & $P_Y(Y)$ is a joint probability distribution over (X, Y)

$$P_{XY}(X, Y)$$

s.t. marginals are preserved.

$$\begin{cases} \sum_X P_{XY}(X, Y) = P_Y(Y) \\ \sum_Y P_{XY}(X, Y) = P_X(X) \end{cases}$$

examples>

$P_X \sim Bern(p)$

$P_Y \sim Bern(q)$

① Independent.

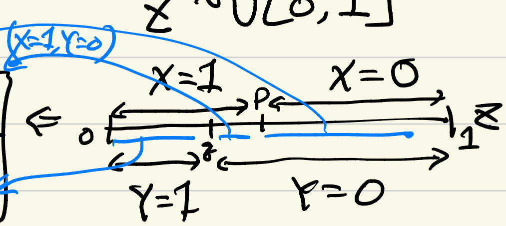
$$\begin{array}{cccc} & P_Y & & P_{XY} \\ 0 & \begin{bmatrix} 1-q \\ q \end{bmatrix} & \xleftarrow{\text{sum}} & \begin{bmatrix} (1-p)(1-q) & p(1-q) \\ (1-p)q & pq \end{bmatrix} \\ 1 & & & \end{array}$$

$$P_X \; \begin{array}{cc} [\,1-p & p\,] \\ 0 & 1 \end{array}$$

② optimal coupling

$Z \sim U[0,1]$

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



$p > q$