Lecture 5: Sampling from a chain

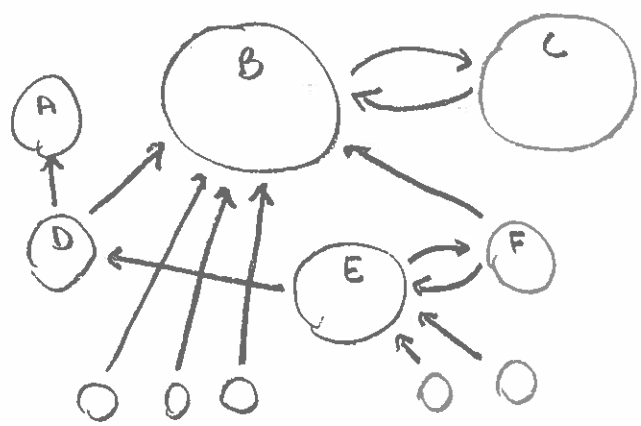
- ☐ Markov chain averaging
- ☐ Metropolis-Hastings samplers
- ☐ Ising single-flip sampler
- ☐ Gibbs samplers
- ☐ Swendsen-Wang and Wolff samplers

Main idea: We can run a Markov chain $\mathbb{P}$ and compute time averages $\frac{1}{T}\sum_{t=0}^{T-1} f(X_t)$ to estimate spatial averages $\sum_{i \in \mathcal{X}} \pi(i) f(i)$ with respect to a target distribution π.

- This week, $\mathcal{X}$ is finite or countable. Next week, uncountably infinite $\mathcal{X}$.
- Ex: PageRank transition matrix

$$\mathbb{P}_{x,y} = \begin{cases} \alpha/d_x + (1-\alpha)V_y, & x \text{ has } d_x \text{ outgoing links including link } y, \\ (1-\alpha)V_y, & \text{otherwise.} \end{cases}$$

We can estimate the PageRank of website i using $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{1}\{X_t = i\}$.



Proposition (Markov chain averaging)

Consider a state $x \in \mathcal{X}$ and define $t_1 = \min\{t > 0 : X_t = x\}$.

Assume that:

a) $\mathbb{E}[t_1 \mid X_0 = x] < \infty$ ("positive recurrence").

b) X_0 is started from an arbitrary distribution q s.t. $\mathbb{P}\{t_1 < \infty \mid X_0 \sim q\} = 1$ ("irreducibility").

Then there exists a unique stationary measure π s.t.

a) $\frac{1}{T} \sum_{t=0}^{T-1} f(X_t) \xrightarrow{T \to \infty} \sum_{i \in \mathcal{X}} \pi(i) f(i)$ for all bounded functions f ("ergodicity").

b) $\pi^T \mathbb{P} = \pi^T$ ("stationarity").

<u>Pf</u> Set $t_0 = 0$, $t_{i+1} = \min\{t > t_i : X_t = x\}$ for $i = 0, 1, \dots$

Then the random variables $\left(\sum_{s=t_i}^{t_{i+1}-1} f(X_s) \right)_{i=1,2,\dots}$ are independent + identically distributed with mean $\mathbb{E}\left[\sum_{s=0}^{t_1 - 1} f(X_s) \mid X_0 = x \right]$, so the SLLN implies

$$\frac{1}{N} \sum_{s=0}^{t_{N+1}-1} f(X_s) \xrightarrow{N \to \infty} \mathbb{E}\left[\sum_{s=0}^{t_1 - 1} f(X_s) \mid X_0 = x \right]. \qquad (*)$$

Similarly,

$$\frac{1}{N} t_{N+1} \xrightarrow{N \to \infty} \mu = \mathbb{E}[t_1 \mid X_0 = x]. \qquad (**)$$

Combining $(*)$ with $(**)$,

$$\frac{1}{T}\sum_{t=0}^{T-1} f(X_t) \xrightarrow{T\to\infty} \frac{1}{\mu}\mathbb{E}\left[\sum_{s=0}^{t_1-1} f(X_s) \,\middle|\, X_0 = x\right].$$

Next set

$$\pi(i) = \frac{1}{\mu}\mathbb{E}\left[\sum_{s=0}^{t_1-1} \mathbb{1}\{X_s = i\} \,\middle|\, X_0 = x\right] = \frac{1}{\mu}\sum_{s=0}^{\infty} \mathbb{P}\{X_s = i,\ t_1 > s \mid X_0 = x\}$$

so that

$$\frac{1}{T}\sum_{t=0}^{T-1} f(X_t) \xrightarrow{T\to\infty} \sum_{i\in\mathcal{X}} \pi(i) f(i) \qquad (\text{"ergodicity"}).$$

Last, observe

$$\sum_{j\in\mathcal{X}} \pi(j)\, p(j,i)$$

$$= \sum_{j\in\mathcal{X}} \left(\frac{1}{\mu}\sum_{s=0}^{\infty} \mathbb{P}\{X_s = j,\ t_1 > s \mid X_0 = x\}\right) p(j,i)$$

$$= \frac{1}{\mu}\sum_{s=1}^{\infty} \mathbb{P}\{X_s = i,\ t_1 \geq s \mid X_0 = x\}$$

$$= \begin{cases} \frac{1}{\mu}\sum_{s=0}^{\infty} \mathbb{P}\{X_s = i,\ t_1 > s \mid X_0 = x\} = \pi(i), & i \neq x \\ \frac{1}{\mu}\sum_{s=1}^{\infty} \mathbb{P}\{t_1 = s \mid X_0 = x\} = \frac{1}{\mu} = \pi(x), & i = x \end{cases}$$

$\square$

$$\Rightarrow \pi^T \mathbb{P} = \pi^T \qquad (\text{"stationarity"}).$$

- Let's use this idea in reverse: Start with a target distribution π and design a sampler $\mathbb{P}$ satisfying $\pi^T \mathbb{P} = \pi^T$.
- Ideally we should verify positive recurrence + irreducibility, but we usually just check $\pi^T \mathbb{P} = \pi^T$.

Detailed balance Simplest criterion is "detailed balance" / "Microscopic reversibility"

$\pi(i) p(i,j) = \pi(j) p(j,i), \quad \forall \; i,j \in \mathcal{X}.$

$\Rightarrow$ Assuming $X_0 \sim \pi$, this means running the video in forward and reverse are both eqvally valid samples from the Markov chain $(X_t)_{t \geq 0}$

$$\Rightarrow \sum_{i \in \mathcal{X}} \pi(i) p(i,j) = \sum_{i \in \mathcal{X}} \pi(j) p(j,i) = \pi(j) \quad \forall \; j \in \mathcal{X}$$

So detailed balance implies stationarity $\pi^T \mathbb{P} = \pi^T$.

Metropolis-Hastings Let's design $\mathbb{P}$ by proposing a transition $x \to y$ with probability $T(x,y)$ and accepting it with probability

$\min\left\{1, \dfrac{\pi(y) T(y,x)}{\pi(x) T(x,y)}\right\}$. Otherwise, stay put $x \to x$.

Q: Does MH satisfy stationarity?

A: $\pi(x) p(x,y) = \pi(x) T(x,y) \min\left\{1, \dfrac{\pi(y) T(y,x)}{\pi(x) T(x,y)}\right\}$

$= \min\left\{\pi(x) T(x,y), \pi(y) T(y,x)\right\}$ = symmetric function of x and y, $\forall \; x \neq y$.

$\Rightarrow$ Detailed balance $\pi(x) p(x,y) = \pi(y) p(y,x)$ is satisfied

$\Rightarrow \pi^T \mathbb{P} = \pi^T$.

Metropolis-Hastings sampler

Input: target pmf $\pi(x)$, known up to a normalizing constant; transition rule $T(x,y)$; starting point x; terminal time T

Output: Approximate samples $(X_t)_{t=0,1,\dots,T}$ from π.

1. Set $X_0 = x$.
2. For $t = 1, 2, \dots, T$:

   a) Generate a proposal $Y_t \sim T(X_{t-1}, \cdot)$.

   b) Sample $U_t \sim \mathrm{Unif}(0,1)$.

   c) $$\text{Set } X_t = \begin{cases} y_t, & \text{if } U_t \le \dfrac{\pi(y_t)\,T(y_t, X_{t-1})}{\pi(X_{t-1})\,T(X_{t-1}, y_t)}, \\ X_{t-1}, & \text{otherwise.} \end{cases}$$

Ising single-flip sampler

$$\pi(x) = \frac{1}{Z_\beta} \exp\Big(\beta \sum_{i\sim j} x_i x_j\Big), \quad Z_\beta = \sum_y \exp\Big(\beta \sum_{i\sim j} y_i y_j\Big).$$

- Choose vertex i uniformly at random.
- Propose $x_i \to -x_i$.
- Accept if $$U_t \le \frac{\exp\left(-\beta \sum_{j\sim i} x_i x_j\right)}{\exp\left(\beta \sum_{j\sim i} x_i x_j\right)} = \exp\Big(-2\beta \sum_{j\sim i} x_i x_j\Big).$$

$$\begin{matrix} + & + & - & - \\ + & - & - & - \\ - & - & + & - \\ - & - & - & - \end{matrix} \quad \longrightarrow \quad \begin{matrix} + & + & - & - \\ + & + & - & - \\ - & - & + & - \\ - & - & - & - \end{matrix}$$

"The Metropolis algorithm aided with Hastings' (1978) generalization is so general that one always tends to feel unsatisfactory in settling down on any specific proposal chain."

– Jun Liu

Gibbs sampler

Idea: Assuming $X = (x_1, x_2)$, transition $X \to y = (y_1, x_2)$ where $y_1 \sim \pi(\cdot | x_2)$.

Q: Does Gibbs satisfy stationarity?

A: $\sum_{x} \pi(x_1, x_2)\, p((x_1, x_2), (y_1, y_2))$

$$= \sum_{x_1} \pi(x_1, y_2)\, \pi(y_1 | y_2) = \pi(y_2)\pi(y_1 | y_2) = \pi(y_1, y_2),$$

So π is a stationary measure, $\pi^T = \pi^T \mathbb{P}$.

– We can combine conditional samplers for various coordinates or groups of coordinates.

Gibbs sampler

Input: target pmf $\pi(x)$, known up to a normalizing constant; procedures for sampling $x_1 \sim \pi(\cdot | x_2)$, $x_2 \sim \pi(\cdot | x_1)$; starting point x; terminal time T

Output: Approximate samples $(X_t)_{t=0,\dots,T}$ from π.

1. Set $X_0 = x$.
2. For $t = 1, 2, \dots, T$:

a) If t is odd, set $X_{t,2} = X_{t-1,2}$ and sample $X_{t,1} \sim \pi(\cdot \mid X_{t,2})$.

b) If t is even, set $X_{t,1} = X_{t-1,1}$ and sample $X_{t,2} \sim \pi(\cdot \mid X_{t,1})$.

⇒ Always accept :)

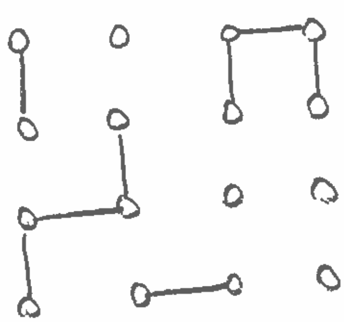
⇒ But it's hard to sample $\pi(\cdot \mid X_2)$ and $\pi(\cdot \mid X_1)$.

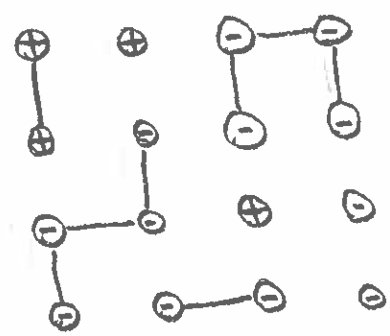
⇒ Let's show how to do this for the Ising model.

<u>Random cluster model</u> $\omega : E \to \{0,1\}$ identifies open $(\omega(e)=1)$ and closed $(\omega(e)=0)$ edges in a graph.

- Random cluster pmf $\pi(\omega) = \frac{1}{Z_p} 2^{c(\omega)} \prod_{e \in E} p^{\omega(e)} (1-p)^{1-\omega(e)}$,

where $c(\omega) = \#$ vertex clusters connected by open edges.





$c(\omega) = 8$

- Random cluster + Ising pmf $\pi(\omega, x) = \frac{1}{Z_p} \prod_{e \in E} p^{\omega(e)} (1-p)^{1-\omega(e)}$

if $x_i = x_j$ for all i,j in the same cluster.

⇒ Independently assign each cluster Unif$\{+1,-1\}$ spins.

Q: Sampling $x \mid \omega$ is easy, but how do we sample $\omega \mid x$?

A: $$\pi(x) = \sum_{\omega} \pi(\omega, x) = \frac{1}{Z_p}(1-p)^{\sum_{i\sim j} \mathbb{1}\{x_i \neq x_j\}}$$

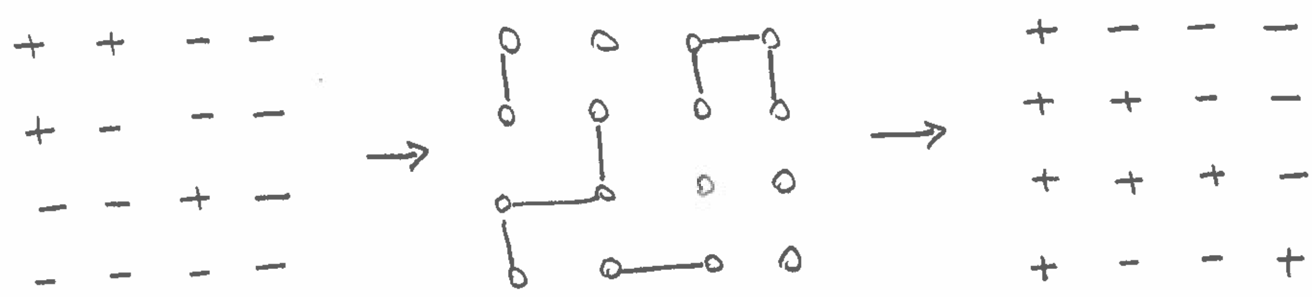
$$= \frac{1}{Z_p}(1-p)^{\sum_{i\sim j}(1-x_i x_j)/2} = \frac{(1-p)^{\#E/2}}{Z_p} \exp\left(\beta \sum_{i\sim j} x_i x_j\right),$$

where $\beta = -\frac{1}{2}\log(1-p)$, $p = 1 - e^{-2\beta}$,

$$\Rightarrow \pi(w \mid x) = \begin{cases} \prod_{e_{ij} : x_i = x_j} p^{w(e_{ij})}(1-p)^{1-w(e_{ij})}, & w(e_{ij}) = 0 \text{ if } x_i \neq x_j \\ 0, & \text{otherwise.} \end{cases}$$

Swendsen-Wang sampler

1. Sample initial state x_0.
2. For $t = 1, 2, \dots, T$:
   a) For each $i \sim j$:
      i. If $x_{t-1,i} \neq x_{t-1,j}$, set $\omega_t(e_{ij}) = 0$.
      ii. If $x_{t-1,i} = x_{t-1,j}$, set
      $$\omega_t(e_{ij}) = \begin{cases} 1, & \text{with prob } 1 - e^{-2\beta}, \\ 0, & \text{with prob } e^{-2\beta}. \end{cases}$$
   b) For each connected component C, sample $x_C \sim \text{Unif}\{+1, -1\}$ and set $x_{t,i} = x_C$ for each $i \in C$.



Idea: why not combine Metropolis + Gibbs?

- Sample the random clusters, given the spins (Gibbs).
- Randomly sample a vertex, identify the open connected component, flip all the spins (Metropolis).

Wolff sampler

1. Sample initial state X_0.
2. For $t = 1, 2, \ldots, T$:
   a) Sample a random vertex i and set $C = C^{new} = \{i\}$.
   b) While C^{new} is non-empty:
      i. Set $C^{old} = C^{new}$, $C^{new} = \emptyset$.
      ii. For each edge e_{ij} with $i \in C^{old}$, $j \notin C$, and $X_{t-1,i} = X_{t-1,j}$, with probability $1 - e^{-2\beta}$ set $C = C \cup \{j\}$ and $C^{new} = C^{new} \cup \{j\}$.

   c) Set
$$X_{t,i} = \begin{cases} X_{t-1,i}, & i \notin C, \\ -X_{t-1,i}, & i \in C. \end{cases}$$