CS 7535 **Markov Chain Monte Carlo Methods** **Fall 2017**

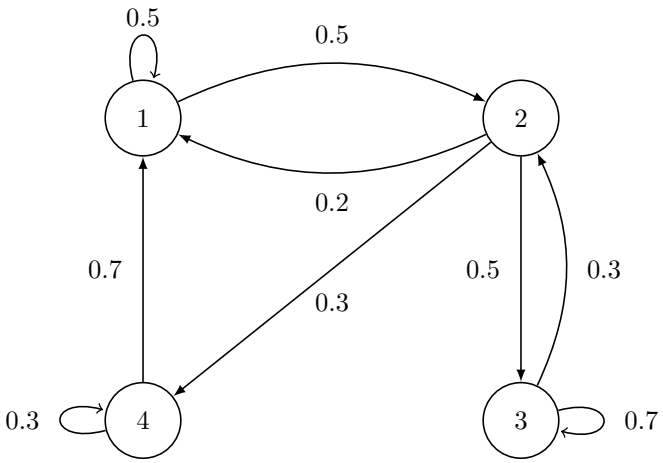
# Lecture 5: September 5

*Lecturer: Prof. Eric Vigoda* *Scribes: Alex Mueller (amueller35)*



## 5.1 Markov Chains

Example:




Transition matrix for the above Markov Chain: $P = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.2 & 0 & 0.5 & 0 \\ 0 & 0.3 & 0.7 & 0 \\ 0.7 & 0 & 0 & 0.3 \end{bmatrix}$

Let $\Omega$ be the set of all possible states and $X_t \in \Omega$ be a random variable representing the state at discrete time $t \geq 0$. $P$ is an $N \times N$ transition matrix representing the Markov Chain. $P$ is stochastic, meaning all of the elements in a given row sum to 1.

For a pair of state $i, j \in \Omega$, $Pr(X_{t+1} = j | X_t = i) = P(i, j)$:

$$Pr(X_{t+2} = j | X_t = 1) = \sum_{k \in \Omega} Pr(X_{t+1} = k | X_t = i) Pr(X_{t+2} = j | X_{t+1} = k) \tag{5.1}$$

$$= \sum_{k} P(i, k) P(k, j) \tag{5.2}$$

$$= P^2(i, j) \tag{5.3}$$

Furthermore,

$$Pr(X_{t+\ell} = j | X_t = i) = P^{\ell}(i, j) \tag{5.4}$$

Let $\mu_t$ be an $N$-length vector representing the distribution of the states at time $t$ such that the sum of all element sum to 1. In other words, element $i$ of $\mu_t$ is the probability of being at state $i$ at time $t$.

$$X_0 \sim \mu_0 \quad \text{(distribution at time 0)} \tag{5.5}$$
$$X_1 \sim \mu_1 = \mu_0 * P \tag{5.6}$$
$$X_t \sim \mu_t = \mu_0 * P^t \tag{5.7}$$

As $t \to \infty$,

$$\lim_{t\to\infty} P^t = \begin{bmatrix} -\pi- \\ -\pi- \\ -\pi- \\ -\pi- \end{bmatrix} \tag{5.8}$$

where $\pi$ is an eigenvector with eigenvalue 1. In other words, no matter what state you started in, each row will end up with the same distribution.

When is there a stationary distribution $\pi$? If the system is ergodic, than there is a unique stationary distribution.

Ergodic means that:

- $\forall\ i, j\ \exists\ t$ such that $P^t(i, j) > 0$
- Another way to look at ergodicity is that if we look at a graph for $P^t$, it will be fully-connected
- Ergodic implies irreducible and aperiodic (and vice versa)
- Irreducible: $\forall\ i, j\ \exists$ t such that $P^t(i, j) > 0$ (graph defined by P is 1 strongly connected component)
- Aperiodic: for state $i \in \Omega$, let $T_i = \{t : P^t(i, i) > 0\}$. Aperiodic means that $\forall$ i $gcd(T_i) = 1$, which means there is no periodic structure.

**Theorem 5.1** *For a finite ergodic Markov chain, $\exists$ a unique stationary distribution on $\pi$ and $\forall$ $X_0$, $\lim_{t\to\infty} X_t \sim \pi$. For all $i, j \in \Omega$, $\lim_{t\to\infty} P^t(i, j) = \pi(j)$. What is $\pi$? If $P$ is symmetric (i.e. $P(i, j) = P(j, i)$), then $\pi = Uniform(\Omega)$ and $\pi P = \pi$.*

$(\pi P)(i) = \pi(i) = \frac{1}{N}$ where $N = |\Omega|$...

$$(\pi P)(i) = \sum_k \pi(k) P(k, i) = \sum_k \frac{1}{N} P(k, i) \tag{5.9}$$
$$= \frac{1}{N} \sum_k P(k, i) = \frac{1}{N} \sum_k P(i, k) \quad \text{by symmetry of P} \tag{5.10}$$
$$= \frac{1}{N}(1) = \frac{1}{N} \tag{5.11}$$

$P$ is reversible with respect to $\pi$ if $\forall i, j$, $\pi(i)P(i,j) = \pi(j)P(j,i)$.

$$\begin{aligned}(\pi P)(i) &= \sum_k \pi(k)P(k,i) = \sum_k \pi(i)P(i,k) \\ &= \pi(i)\sum_k P(i,k) = \pi(i)(1) = \pi(i)\end{aligned}$$

Typically, the only way we can see what $\pi$ is if $P$ is symmetric or reversible with respect to $\pi$.

For a $d$-regular (every node has exactly d neighbors), undirected graph $G = (V, E)$, in a random walk at state $i$, there is a probability of $\frac{1}{2}$ of staying at state $i$ and a probability of $\frac{1}{2}$ to move to one of its neighboring $d$ states, and for edge $(i, j) \in E$, the probability $P(i, j)$ of moving along this edge is $\frac{1}{d}$ (also, because it is undirected, $P(i, j) = \frac{1}{d} = P(j, i)$). Thus, $\pi = \text{Uniform}(V)$.

For a non-regular graph, $\pi(i) = \frac{deg(i)}{z}$ where $z = \sum_i deg(i) = 2m$. And in this case,

$$\pi(i)P(i,j) = \frac{deg(i)}{z} \cdot \frac{1}{2} \cdot \frac{1}{deg(i)} = \frac{1}{2z} = \pi P(j,i)$$

## 5.2 MCMC: Markov Chain Monte Carlo

Let $G = (V, E)$ be an undirected graph and $\Omega$ is the set of all matchings in G. The goal is to sample uniformly from $\Omega$, so we can design a Markov chain model.
From $X_t \in \Omega$,

1. Choose an edge $e \in E$ uniformly at random.
2. $X' = X_t \oplus e = \begin{cases} X_t \cup e & \text{if} \quad e \notin X_t \\ X_t \setminus e & \text{if} \quad e \in X_t \end{cases}$
3. If $X' \in \Omega$ with probability $\frac{1}{2}$ set $X_{t+1} = X'$. Else $X_{t+1} = X_t$.

A random walk will eventually lead to a matching.

For $m \in \Omega$, $P(m, m) \geq \frac{1}{2}$, $G$ is aperiodic. $\forall m, m' \in \Omega$, $P^t(m, m') > 0$, $G$ is irreducible. If $G$ is ergodic and symmetric, then $\pi = \text{Uniform}(\Omega)$. How fast does $X_t \to \pi$? We don't need an exact sample from $\pi$, but we want a sample from a distribution that is close to $\pi$.

## 5.3 Mixing Time

Recall the definition for mixing time,

$$T_{mix}(\epsilon) = \max_{X_0} \min\{t : d_{TV}(P^t(X_0, \cdot), \pi) \leq \epsilon\} \tag{5.12}$$

For $X_0, \epsilon > 0$, $T_{mix}^{X_0} = \min\{t : d_{TV}(P^t(X_0, \cdot), \pi) \leq \epsilon\}$.

$$T_{mix}(\epsilon) = \max_{X_0} T_{mix}^{X_0}(\epsilon)$$

$$T_{mix}(\epsilon) = T_{mix}(\frac{1}{4}) \log(\frac{1}{\epsilon})$$

We will later prove the mixing time for this method is $O(mn)$.

## 5.4 Ising Model

The Ising Model is used by physicists to model ferromagnetic solids. Let $\Omega = \{+1, -1\}$ where $+1$ and $-1$ are possible electron states. For $\sigma \in \Omega$, energy is calculated by using the Hamiltonian $H(\sigma) =$ number of monochromatic edges. For $\beta > 0$, $\beta = \frac{1}{T}$ (where $\beta$ is the inverse temperature) and $\omega(\sigma) = e^{-\beta H(\sigma)} = e^{\beta * \#\text{monochromatic edges}}$.

The Gibbs distribution is

$$\mu(\sigma) = \frac{\omega(\sigma)}{Z} \tag{5.13}$$

where $Z = \sum_{\sigma \in \Omega} \omega(\sigma)$ is the normalizing factor. We want to design a Markov chain with the above $\mu(\sigma)$ Gibbs distribution.

A Markov chain can also be implemented via a Metropolis chain. To do this, take $X_t \in \Omega$,

1. Choose $v \in V$ uniformly at random and $s \in \{+1, -1\}$ uniformly at random.
2. Set $X'(w) = X_t(w)$ for $w \neq v$.
3. Go to $X_{t+1} = X'$ with probability from the Metropolis filter $\min\{1, \frac{\omega(X')}{\omega(X_t)}\}$, else set $X_{t+1} = X_t$

The Gibbs distribution and Metropoplis chain can be used to sample from and approximate a distribution.

## Miscellaneous notes

- Jerrum's book is available online on his website.
- The Levin, Wilmer, Peres book is available on Prof. Vigoda's website.

**CS 7535 Markov Chain Monte Carlo Methods** **Fall 2017**

# Lecture 6: September 7

*Lecturer: Prof. Eric Vigoda* *Scribes: Tianhang Zhu (tzhu71)*



## 6.5 Coupling:

### 6.5.1 Definition:

For a finite space $\Omega$, where $\mu$ and $\nu$ are two distributions on $\Omega$, a coupling is a joint distribution $w$ on $\Omega \times \Omega$, such that $\forall i \ \sum_{j\in\Omega} w(i,j) = \mu(i)$ and $\forall j \ \sum_{i\in\Omega} w(i,j) = \nu(j)$.

For distributions $\mu = (\frac{1}{2}, \frac{1}{4}, 0, \frac{1}{4})$ and $\nu = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0)$ Here are two examples of couplings of $\mu, \nu$:

$$w_1 = \begin{bmatrix} \frac{1}{3} & \frac{1}{12} & \frac{1}{12} & 0 \\ 0 & \frac{1}{8} & \frac{1}{8} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \frac{1}{8} & \frac{1}{8} & 0 \end{bmatrix}$$

$$w_1 = \begin{bmatrix} \frac{1}{3} & \frac{1}{12} & \frac{1}{12} & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 \end{bmatrix}$$

### 6.5.2 Coupling Lemma:

(a) for any coupling $w$ of $\mu$ and $\nu$, $d_{tv}(\mu,\nu) \le P(\sigma,\tau)$, where $\sigma$ and $\tau$ forms a sample that is drawn from the coupling.

(b) there exists a coupling where $d_{tv}(\mu,\nu) = P(\sigma,\tau)$, which is called the optimal coupling

**Proof:**

(a)

$$P(\sigma = \tau) = \sum_{\eta\in\sigma} w(\eta,\eta) \tag{6.14}$$

$$\le \sum_{\eta} min(\mu(\eta), \nu(\eta)) \tag{6.15}$$

$$P(\sigma \neq \tau) \geq 1 - \sum_{\eta \in \sigma} w(\eta, \eta) \tag{6.16}$$

$$\geq \sum_{\eta} \mu(\eta) - min(\mu(\eta), \nu(\eta)) \tag{6.17}$$

because for cases where $\mu(\eta) \leq \nu(\eta)$, above equation will be 0

$$= \sum_{\eta : \mu(\eta) > \nu(\eta)} (\mu(\eta) - \nu(\eta)) \tag{6.18}$$

$$= \max_{S \subset \Omega} \mu(S) - \nu(S) \tag{6.19}$$

$$= d_{tv}(\mu, \nu) \tag{6.20}$$

$$\tag{6.21}$$

(b)

To obtain equality, we need $w(\eta, \eta) = \min(\mu(\eta), \nu(\eta))$

### 6.5.3 Coupling to bound MCMC mixing time

Say we want to bound the mixing time of a Markov chain on $\Omega$ with P, how we will use coupling to bound the mixing time?

1. make two copies of the chain $X_t$ and $Y_t$
2. design their transition probability so that $X_t, Y_t$ form a coupling
3. make sure that once they agree they will agree later
4. then we can bound the mixing time of the chain by $P(\sigma \neq \tau)$ using the property of the coupling

To form a coupling with $X_t$ and $Y_t$, we need that:

$\forall i, j, k, l \in \Omega$:

$$P(X_{t+1} = k \mid X_t = i, Y_t = j) = P(i, k) \in P$$
$$P(Y_{t+1} = l \mid X_t = i, Y_t = j) = P(j, l) \in P$$

plus if $X_t = Y_t$ then $X_{t+1} = Y_{t+1}$, so once they agree they agree afterwards

$\forall i, j \in \Omega$:

$$T_{couple}^{i,j} = min\{t : P(X_t \neq Y_t \mid X_0 = i, Y_0 = j) \leq \frac{1}{4}\}$$

$T_{couple}^{i,j}$ is the mixing time bound when the chains start at i, j, Let

$$T_{couple} = \max_{i,j} T_{couple}^{i,j}$$

the total mixing time is then bounded by the max of all starting point, i.e., $T_{mix} \leq T_{couple}$

## Examples:

**Random walk on hypercube:**

Vertices are n bit strings. Edges are E(x,y) are two vertices, x and y differ in one coordinate. From $X_t \in V$(set of all vertices):

1. with probability $\frac{1}{2}$ stay at current state

2. with $\frac{1}{2}$ probability choose a random neighbor, essentially picking a position of current string and flip the bit

Another equivalent way of describing this is:

1. pick coordinate $i \in \{1...n\}$ and $b \in \{0, 1\}$ u.a.r

2. Set $X_{t+1}(i) = b$ and $X_t(j)$

To form a valid coupling: if $X_t = Y_t$: choose i, b, u.a.r if $X_t \neq Y_t$: choose the same i and same b

$$D_t = \{j : X_t(j) \neq Y_t(j)\}$$

$D_t$ is the set of disagreeing bits between $X_t$ and $Y_t$ at step t

$$A_t = \{j : X_t(j) = Y_t(j)\}$$

$A_t$ is the set of agreeing bits between $X_t$ and $Y_t$ at step t

For example: $X_t = 1100110$ and $Y_t = 0101001$

Pick a position i at random,

if $i \in A_t$, then

$$D_{t+1} = D_t$$

if $i \in D_t$, then

$$\begin{aligned} D_{t+1} &= D_t \setminus \{i\} \\ E(|D_{t+1}|) &= |D_t| * (1 - \frac{1}{n}) \\ E(|D_t|) &= |D_0| * (1 - \frac{1}{n})^t \\ &\leq n * \exp \frac{-t}{n} \end{aligned}$$

$|D_0|$ is bounded by n

$$P(X_t \neq Y_t) \leq E(|D_t|) \leq \frac{1}{4} \quad \text{with } t = n * \ln(4 * n)$$

because if $X_t = Y_t$, $P(X_t \neq Y_t) = 0$ and $E(|D_t|) = 0$ and if $X_t \neq Y_t$, $P(X_t \neq Y_t) = 1$ and $E(|D_t|) \geq 1$

**Card Shuffling**

Top-to-random shuffling: $\Omega$ is all permutations of n cards. Each state has n neighbours and transition is:

1. take top card

2. place in random position

It's aperiodic because it has self loop. It's irreducible because from a state $\mu$ you can go to any other state using $O(n^2)$ time. Thus it's ergodic and the stationary distribution $\pi$ is unique.

P is double-stochastic $\iff$ $\pi$ is uniform.

We study its inverse chain:

1. pick a random card

2. put it to the top

Let $X_t$ and $Y_t$ be state at time t for two copies of this chain. At each time step, pick the same card to put on t op.

$$
\begin{aligned}
T &= \text{time to choose any card} \\
T_i &= \text{time to get the ith card to be identical between } X_t \text{ and } Y_t \\
T &= \sum_i T_i \quad \text{where } T_i \text{ is Geometric}(p_i) \text{ where} \\
p_i &= \frac{n-i+1}{n} \\
E(T_i) &= \frac{1}{p_i} \\
&= \frac{n}{n-i+1} \\
E(T) &= \sum_{i=1}^{n} E(T_i) \\
&= \sum \frac{n}{n-i+1} \\
&= n * (\frac{1}{n-1} + \frac{1}{n-2} + ...) \\
&\leq n * (1 + \log n) \\
Pr(T > 4E[T]) &\leq \frac{1}{4} \\
\text{Hence for} \quad T &= 4n(1 + \ln n) \\
Pr(X_T \neq Y_T) &\leq \frac{1}{4}
\end{aligned}
$$

Therefore the mixing time is $T_{mix} = O(n \log n)$