Markov Chain Monte Carlo

Lecturers: Dana Randall, Eric Vigoda Scribes: Matthew Fahrbach, Chunxing Yin

March 29 and 31, 2016

1 Finite Markov Chains

A finite Markov chain is a random walk among the elements of a finite state space Ω in the following manner: when at $x \in \Omega$, the next position is chosen according to a fixed transition probability distribution $P(x, \cdot)$. More precisely, a sequence of random variables $(X_0, X_1, \dots,)$ is a Markov chain with state space Ω and transition matrix P if for all $x, y \in \Omega$ and $t \geq 0$, we have the memoryless property that

$$\mathsf{P}(X_{t+1} = y | X_t = x, X_{t-1} = x_{t-1}, \dots, X_0 = x_0) = \mathsf{P}(X_{t+1} = y | X_t = x) = P(x, y),$$

if both conditional probabilities are well-defined. Therefore, a $|\Omega| \times |\Omega|$ matrix P suffices to describe the transitions, since all that matters at each time step is the last state. The x-th row of P is the distribution $P(x, \cdot)$, thus P is stochastic, giving $\sum_{y \in \Omega} P(x, y) = 1$ for all $x \in \Omega$. Additionally, by the definition of matrix multiplication, we have the property that $P^t(x, y) = \sum_{z \in \Omega} P^{t-1}(x, z) P(z, y)$, where $P^t(x, y)$ the the probability of moving in t steps from x to y. For any initial distribution μ_0 on the state space, the distribution at time t is given by $\mu_t = \mu_0 P^t$ for all $t \geq 0$. Typically Ω is an exponentially large state space, so many matrix computations on P are often intractable.

Definition 1.1. *A distribution π is a* stationary distribution *if $\pi(x) = \sum_{y \in \Omega} \pi(y) P(y, x)$ for all $x \in \Omega$.*

All Markov chains have a stationary distribution, but it is not always unique. Note that π is a left eigenvector of P with left eigenvalue 1. The following definitions and lemma formalize useful properties for designing Markov chains.

Definition 1.2. *A Markov chain is* ergodic *if it is*

1. irreducible*, i.e., there is a t such that $P^t(x, y) > 0$ for all $x, y \in \Omega$ (connected), and*
2. aperiodic*, i.e.,* $\gcd\{t : P^t(x, y) > 0\} = 1$ *for all $x, y \in \Omega$ (non-bipartite).*

Definition 1.3. *The* lazy version *of a Markov chain with transition matrix P has the transition matrix $(P + I)/2$.*

It is often convenient to analyze lazy versions of Markov chains, because laziness guarantees aperiodicity and certain spectral properties, which we will soon encounter.

Lemma 1.4. *Any finite, ergodic Markov chain converges to a unique stationary distribution π, i.e., for all $x, y \in \Omega$, we have that $\lim_{t \to \infty} P^t(x, y) = \pi(y)$.*

Lemma 1.5. *Let M be an ergodic Markov chain on a finite state space Ω with transition matrix P. If $\pi' : \Omega \to [0,1]$ is any function satisfying the detailed balance equation:*

$$\pi'(x)P(x,y) = \pi'(y)P(y,x),$$

and if it also satisfies $\sum_{x\in\Omega} \pi'(x) = 1$, then π' is the unique stationary distribution of M.

Proof. Sum both sides over all $x \in \Omega$:

$$\sum_{x\Omega} \pi'(x)P(x,y) = \sum_{x\in\Omega} \pi'(y)P(y,x) = \pi'(y) \sum_{x\in\Omega} P(y,x) = \pi'(y),$$

since P is stochastic. □

Definition 1.6. *Any Markov chain that satisfies the detailed balance equation for some π' is called* time-reversible.

Example 1.7 (Random Walk on the Hypercube). Consider a random walk on the the vertices of the hypercube $(0,1)^n$. Clearly $|\Omega| = 2^n$. For any adjacent vertices $x \sim y$, set $P(x,y) = 1/n$. Start the walk from $(0,0,\ldots,0)$. This chain is periodic, because the parity of the number of 1 bits in a state changes with each step. To remedy this, introduce a self loop at each vertex, and set $P(x,x) = 1/2$ and $P(x,y) = 1/2n$ for all $x \neq y$. Now that the chain is ergodic, the detailed balance equation implies that $\pi(x) = \pi(y)$ for all $x, y \in \Omega$, so π is the uniform distribution. We will revisit this example when we discuss coupling. ♦

Example 1.8 (Sampling Independent Sets). Suppose want to sample independent sets of $G = (V,E)$ uniformly. Our goals are to connect the state space, define transition probabilities so π is uniform, and (eventually) show that it converges to π quickly. Connect independent sets $I, J \in \mathcal{I}$ if $|I \triangle J| = 1$, i.e., I and J differ by one vertex. From any independent set, uniformly choose a neighbor and transition there. The state space is connected, because every state can reach $\emptyset \in \mathcal{I}$, but the stationary distribution may not be uniform since the degree of each vertex isn't always the same. Now consider the following chain. If $I \sim J$, then $P(I,J) = 1/2n$, so $\pi(I) = \pi(J)$ by detailed balance and π is uniform.

$\underline{MC_{\text{independent set}}}$

Starting at the vertex $I_0 = \emptyset \in \mathcal{I}$, repeat:

1. Choose a vertex and bit $(v,b) \in V \times \{0,1\}$ uniformly at random.
2. If $b = 0$ let $J = I_t \setminus \{v\}$, otherwise let $J = I_t \cup \{v\}$.
3. If $J \in \mathcal{I}$ let $I_{t+1} = J$, otherwise let $I_{t+1} = I_t$.

Now suppose we wish to sample independent sets from the weighted distribution

$$\pi(I) = \lambda^{|I|}/Z$$

for any $\lambda \in \mathbb{R}_{>0}$, where $Z = \sum_{I'\in\Omega} \lambda^{|I'|}$ is the normalizing constant. When $\lambda < 1$ the stationary distribution favors small independent sets, and when $\lambda > 1$ the stationary distribution favors large independent sets. The unbiased case is $\lambda = 1$, for which we just designed a chain. To change the distribution from which the chain samples, consider this modified version of $MC_{\text{independent set}}$ for $\lambda < 1$.

$MC_{\text{independent set } (\lambda < 1)}$

Starting at the vertex $I_0 = \emptyset \in \mathcal{I}$, repeat:

1. Choose a vertex and bit $(v, b) \in V \times \{0, 1\}$ uniformly at random.
2. If $b = 0$ let $J = I_t \setminus \{v\}$, otherwise let $J = I_t \cup \{v\}$.
3. If $J \in \mathcal{I}$ and $b = 0$, let $I_{t+1} = J$;
4. Else if $J \in \mathcal{I}$ and $b = 1$, let $I_{t+1} = J$ with probability λ;
5. Else let $I_{t+1} = I_t$.

Suppose that $J = I \cup \{v\}$ for some $v \notin I$. Then $P(I, J) = 1/2n$ and $P(J, I) = \lambda/2n$. The stationary distributions satisfy $\pi(I) = \lambda\pi(J)$, so the detailed balance equation is satisfied:

$$\pi(I)P(I, J) = \pi(I)\frac{1}{2n} = \pi(J)\frac{\lambda}{2n} = \pi(J)P(J, I).$$

Observe that we never actually needed to compute Z, which is an intractable task in general; instead, we only need relative probabilities. When $\lambda > 1$, we use essentially the same idea. Sampling a configuration with probability proportional to its size is an example of Boltzmann sampling, which is analogous to sampling from a Boltzmann distribution in statistical mechanics. ♦

The Metropolis-Hastings algorithm [11] is a simple, yet tremendously robust idea that tells us how to assign the transition probabilities of any Markov chain so that it will converge to any distribution.

The Metropolis Algorithm

Starting at x, repeat:

1. Pick a neighbor y of $x \in \Omega$ uniformly with probability $1/2\Delta$, where Δ is the maximum degree in the graph G.
2. Move to y with probability $\min(1, \pi(y)/\pi(x))$.
3. With all remaining probability stay at x.

Using the detailed balance equation, it is easy to verify that if the state space is connected, then π must be the stationary distribution.

Exercise 1.1. In the Metropolis algorithm, instead of choosing a neighbor y of $x \in \Omega$ uniformly, consider what happens when y is chosen from a different probability distribution.

2 Mixing Time

The time a Markov chain takes to converge to its stationary distribution, known as the *mixing time* of the chain, is measured in terms of the variation distance between the distribution at time t and the stationary distribution.

Definition 2.1. *Given two distributions π and ν, the* total variation distance *is*

$$\|\mu, \nu\|_{tv} = \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \nu(x)| = \max_{S\subseteq\Omega} \sum_{x\in S} \mu(x) - \nu(x) = \max_{S\subseteq\Omega} \mu(S) - \nu(S).$$

Definition 2.2. *Letting $P^t(x, y)$ denote the t-step probability of going from x to y, define*

$$\tau_x(\varepsilon) = \min\{t : \|P^{t'}(x, \cdot), \pi\|_{tv} \leq \varepsilon, \forall t' \geq t\}$$

for any $\varepsilon > 0$. The mixing time $\tau(\varepsilon)$ *is $\tau(\varepsilon) = \max_{x\in\Omega} \tau_x(\varepsilon)$.*

In practice, we often let $\varepsilon = 1/4$ so that if $\|P^t(x, \cdot), \pi\|_{\text{tv}} \leq 1/2$ then $\|P^{kt}(x, \cdot), \pi\|_{\text{tv}} \leq 1/2^k$. This is known as probability amplification.

Definition 2.3. *A Markov chain is* rapidly mixing *if its mixing time $\tau(\varepsilon)$ is $O(\text{poly}(n, \log \varepsilon^{-1}))$, where n is the size of each configuration in the state space.*

Definition 2.4. *A Markov chain is* slowly (or torpidly) mixing *if the mixing time $\tau(\varepsilon)$ is $\Omega(\exp(\text{poly}(n, \log \varepsilon^{-1})))$, where n is the size of each configuration in the state space.*

It is well-known from probability theory that the eigenvalue gap of the transition matrix of a Markov chain provides a good bound on its mixing time. Since P is stochastic and π is the largest eigenvector with eigenvalue 1, we have $1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{|\Omega|-1} > -1$. Let $\text{Gap}(P) = 1 - \max\{|\lambda_1|, |\lambda_{|\Omega|-1}|\}$ denote the spectral gap. The following result relates the spectral gap with the mixing time [14]. It is worth noting that if P is the transition of a lazy Markov chain, all the eigenvalues are nonnegative.

Theorem 2.5. *Let $\pi_* = \min_{x\in\Omega}$. For all $\varepsilon > 0$ we have*

(a) $\tau(\varepsilon) \leq \frac{1}{Gap(P)} \log\left(\frac{1}{\pi_* \varepsilon}\right)$.

(b) $\tau(\varepsilon) \geq \frac{1-Gap(P)}{2Gap(P)} \log\left(\frac{1}{2\varepsilon}\right)$.

This approach to mixing is extremely for Markov chains with very well structured transition matrices, such as card shuffling applications and walks on the symmetric group, but it tends to be less useful for the more complicated state spaces that arise in computer science. In particular, for most algorithmic applications the size of the state space is exponentially large and we do not have a compact, mathematical representation of the adjacency matrix, so it is far too difficult to determine the eigenvalues of the transition matrix.

3 Coupling

One of the most popular methods for bounding mixing time is coupling, both because of its elegance and its simplicity.

Definition 3.1. *A* coupling *is a Markov chain on $\Omega \times \Omega$ defining a stochastic process $(X_t, Y_t)_{t=0}^{\infty}$ with the properties:*

1. *Each of the processes X_t and Y_t is a faithful copy of $\mathcal{M}$ (given initial states $X_0 = x$ and $Y_0 = y$).*
2. *If $X_t = Y_t$, then $X_{t+1} = Y_{t+1}$.*

Condition 1 ensures that each process, viewed in isolation, simulates the original chain, yet the coupling updates them simultaneously so that they will tend to coalesce, or move closer together, according to some notion of distance. Once the pair of configurations agree, condition 2 guarantees they agree from that time forward. The coupling (or expected coalescence) time can prived a good bound on the mixing time of the chain $\mathcal{M}$ if it is a carefully chosen coupling.

Definition 3.2. *For initial states $x, y \in \Omega$, let*

$$T^{x,y} = \min\{t : X_t = Y_t | X_0 = x, Y_0 = y\}$$

and define the coupling time *to be* $T = \max_{x,y} \mathsf{E}(T^{x,y})$.

The following result relates the mixing time and the coupling time [1].

Theorem 3.3. *For any $\varepsilon > 0$, we have $t(\varepsilon) \leq \lceil Te \ln \varepsilon^{-1} \rceil$.*

Remark 3.4. Holley [5] provided an alternate and short proof of the FKG inequality using a Markov chain and path coupling.

Theorem 3.5 (FKG Inequality). *Let X be a finite distributive lattice, and μ a nonnegative function on it, that is assumed to satisfy the lattice condition $\mu(x \wedge y)\mu(x \vee y) \geq \mu(x)\mu(y)$ for all x, y in the lattice X. Then for any two monotonically increasing functions f and g on X, the following positive correlation inequality holds:*

$$\left(\sum_{x \in X} f(x)g(x)\mu(x)\right)\left(\sum_{x \in X} \mu(x)\right) \geq \left(\sum_{x \in X} f(x)\mu(x)\right)\left(\sum_{x \in X} g(x)\mu(x)\right).$$

Example 3.6 (Coupling on the Hypercube). Consider the following Markov chain.

$\underline{MC_{\text{cube}}}$

Starting at the vertex $X_0 = (0, 0, \ldots, 0)$, repeat:

1. Pick $(i, b) \in \{1, 2, \ldots, n\}\{0, 1\}$.
2. Let X_{t+1} be X_t with the i-th bith changed to b.

Letting $\varphi(\cdot, \cdot)$ be the Hamming distance, the transition matrix of MC_{cube} is

$$P(X, Y) = \begin{cases} 1/2n & \text{if } \phi(X, Y) = 1, \\ 1/2 & \text{if } X = Y, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to check that this chain is ergodic and symmetric, hence the stationary distribution is uniform. To couple, we start with any two vertices X_0 and Y_0 on the hypercube and update them simultaneously by choosing the same pair (i, b). The two configurations will coalesce as soon as we have updated each index at least once. By the coupon collector's problem, this takes $\Theta(n \log n)$ steps in expectation, so Theorem 3.3 we have $\tau(\varepsilon) = O(n \log n \log \varepsilon^{-1})$. This upper bound is tight, as an exact analysis reveals that $\Theta(n \log n \log \varepsilon^{-1})$. ♦

Example 3.7 (Intuition for Coupling). Consider shuffling a deck of cards by choosing a random card and bringing it to the top of the deck. Assume we have two decks: the left deck is perfectly shuffled (mixed) and we want to shuffle the right deck. For the coupling, choose a random card i in the right deck and bring it to the top. Then find i in the left deck and bring it to the top. Both moves are faithful copies of the shuffling chain, and once each card has been selected in the right deck, it will be coupled with the already-mixed left deck. Again by the coupon collector problem, this coupling takes $O(n \log n)$ in expectation. ♦

The *path coupling* technique, introduced by Bubley and Dyer [3], gives us a way to measure the expected change in distance between two arbitrary configurations in a Markov chain by only considering pairs of configurations that are close. The following version of the path coupling theorem given in [4] is convenient.

Theorem 3.8 (Bubley and Dyer [3]). *Let φ be an integer valued metric defined on $\Omega \times \Omega$ which takes values in $\{0, \ldots, B\}$. Let U be a subset of $\Omega \times \Omega$ such that for all $(x_t, y_t) \in \Omega \times \Omega$ there exists a path $x_t = z_0, z_1, \ldots, z_r = y_t$ between x_t and y_t such that $(z_i, z_{i+1}) \in U$ for $0 \leq i < r$ and $\sum_{i=0}^{r-1} \varphi(z_i, z_{i+1}) = \varphi(x_t, y_t)$. Let be a Markov chain on Ω with transition matrix P. Consider any random function $f : \Omega \to \Omega$ such that $\mathsf{P}[f(x) = y] = P(x, y)$ for all $x, y \in \Omega$, and define a coupling of the Markov chain by $(x_t, y_t) \to (x_{t+1}, y_{t+1}) = (f(x_t), f(y_t))$.*

1. *If there exists $\beta < 1$ such that $\mathsf{E}[\varphi(x_{t+1}, y_{t+1})] \leq \beta \varphi(x_t, y_t)$ for all $(x_t, y_t) \in U$, then the mixing time satisfies*

$$\tau(\varepsilon) \leq \frac{\ln(B\varepsilon^{-1})}{1 - \beta}.$$

2. *If $\beta = 1$ (so $\mathsf{E}[\Delta\varphi(x_t, y_t)] \leq 0$ for all $x_t, y_t \in U$, let $\alpha > 0$ satisfy $\mathsf{P}[\varphi(x_{t+1}, y_{t+1}) \neq \varphi(x_t, y_t)] \geq \alpha$ for all t such that $x_t \neq y_t$. The mixing time of then satisfies*

$$\tau(\varepsilon) \leq \left\lceil \frac{eB^2}{a} \right\rceil \lceil \ln(\varepsilon^{-1}) \rceil.$$

Example 3.9 (k-Coloring [13]). We demonstrate the technique of path coupling on a Markov chain MC_{col} for sampling k-colorings of a graph G when $k \geq 2d$, where d is the maximum degree of G. Vigoda [15] showed that MC_{col} mixes rapidly when $k \geq 11d/6$.

$\underline{MC_{\text{col}}}$

Starting at t_0, repeat:

1. With probability $1/2$ do nothing.
2. Pick $(v, c) \in V \times \{1, \ldots, \}$.
3. If v can be recolored with color c, recolor it; otherwise do nothing.

We can easily check that MC_{col} converges to the uniform distribution over k-colorings. To apply path coupling, for any $x, y \in \Omega$ let $x = z_0, z_1, \ldots, z_\ell = y$ be the shortest path such that z_i and z_{i+1} are colorings that differ at a single vertex, and set $\varphi(x, y) = \ell$. When the number of colors used is much larger than d, it is simple to verify that $\varphi(x, y) \leq B = 2n$. Let U be the set of pairs of colorings at distance 1. To apply the path coupling theorem, we need to consider $\mathsf{E}[\Delta\varphi(r, s)]$ for any $(r, s) \in U$. Suppose w is the vertex that is colored differently in r and s, and assume WLOG that $w \in r$ is red and $w \in s$ is blue. We propose the following coupling that updates X_t and Y_t. Choose the same vertex v in each step. If $v \notin \Gamma(w)$, the neighborhood of w, then choose the same color for v each both X_t and Y_t. If $v \in \Gamma(w)$, then couple the color choices (red, blue) and (blue, red) for X_t and Y_t, with the remaining colors being the same in each of X_t and Y_t. Consider the three possible cases:

- $\underline{\text{Case 1: } v = w\text{:}}$ If $v = w$ and the color chosen is c, then r and s will become the same if no neighbor of w is colored c. Otherwise the move is rejected. Therefore, the expected change in distance is $\mathsf{E}[\Delta\varphi(r, s)] = -(k - |\Gamma(w)|)/(kn) \leq (d - k)/(kn)$.

- Case 2: $v \in \Gamma(w)$: If $v \in \Gamma(w)$, the distance between r and s will remain unchanged unless the pair of colors chosen is (blue, red), in which case the distance increases by 1. Therefore, $\mathsf{E}[\Delta\varphi(r,s)] = |\Gamma(w)|/(kn) \leq d/(kn)$.
- Case 3: $v \neq w$ and $v \notin \Gamma(w)$: If w has distance at least 2 from v in the graph G, then any proposed move will be accepted by both processes in the coupling, or rejected by both processes. In either case the expected change in the distance is 0.

Putting these pieces together, we find

$$\mathsf{E}[\Delta\varphi(r,s)] \leq \frac{1}{kn}((d-k)+d) = \frac{2d-k}{kn}.$$

When $k \geq 2d+1$, this gives us the bound

$$\tau(\varepsilon) \leq \frac{\ln(n\varepsilon^{-1})}{1-\frac{1}{kn}},$$

which lets us conclude $O(n\log(n\varepsilon^{-1}))$ mixing. When $k = 2d$, we have $\mathsf{E}[\Delta\varphi(r,s)] = 0$. Since there is no bias towards either increasing or decreasing distance, the colorings r and s will coalesce in $O(n^2 \log \varepsilon^{-1})$ steps. ♦

Remark 3.10 (Coupling from the past). Among MCMC algorithms, *coupling from the past*, invented by Propp and Wilson [12], is a method for *perfectly* sampling from the stationary distribution of a Markov chain. It is especially practical when the coupling produces *monotone* systems, since only the bottom and top configurations need to coalesce.

4 Matchings in Bipartite Graph

Let $G = (U, V, E)$ be a bipartite graph, and $|U| = |V| = n$. A *matching* in G is a subset $M \subseteq E$ consisting pairwise vertex-disjoint edges. A matching M is *perfect* if it covers all vertices in U and V. We define the adjacency matrix A of G as follows:

$$A_{ij} = \begin{cases} 1 & \text{if } (u_i, v_j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

The adjacency matrix representing a perfect matching is a permutation matrix, which contains exactly one 1 in each row and each column. So we can use the adjacency matrix to compute number of perfect matchings in the graph.

Definition 4.1. *Let A be an $n \times n$ matrix, the* permanent *of A is defined as $perm(A) = \sum_{\sigma \in S_\sigma} \prod_{i=1}^{n} A_{i,\sigma(i)}$, where S_σ is the set of all permutations of $1, 2, \ldots, n$.*

Theorem 4.2. *Let $G = (U, V, E)$ be a bipartite graph, and let A be its adjacency matrix. Number of perfect matchings in G is equal to $perm(A)$.*

Proof. Given a permutation σ and bipartite graph $G = (U, V, E)$, we match vertex u_i to $v_{\sigma(i)}$. If this is a perfect matching, then $(u_i, v_{\sigma(i)}) \in E$ for all i. By the definition of adjacency matrix, $\prod_{i=1}^{n} A_{i,\sigma(i)} = 1$, and so $\text{perm}(A) = \sum_{\sigma \in S_\sigma} 1 = \#$ perfect matchings. □

Theorem 4.3 (Valiant). *It is #P-complete to compute the permanent of 0-1 matrices.*

Thus we want to design a uniform sampler to generate a random perfect matching in order to approximately count the number of matchings in the graph. Suppose $G = (V, E)$ is an arbitrary graph. Let $\mathcal{P}$ denote the set of perfect mathings in G, and let $\mathcal{M}$ denote the set of all matchings. We give a Markov Chain on all matchings whose stationary distribution is uniform.

$\underline{MC_{\text{all matchings}}}$

From $X_t \in \mathcal{M}$:

1. Choose an edge $e = (u, v) \in E$ uniformly at random.
2. There are three mutually exclusive cases.
   (a) If u, v are unmatched, then $X' = X_t \cup \{e\}$.
   (b) If $e \in X_t$, then $X' = X_t \setminus \{e\}$.
   (c) If u is unmatched but v is matched by edge e' (or vice versa), then $X' = X_t \cup \{e\} \setminus \{e'\}$.
   (d) If none of above cases happen, then $X' = X_t$.
3. With probability $\frac{1}{2}$, set $X_{t+1} = X'$, otherwise $X_{t+1} = X_t$.

In fact this Markov Chain also works for non-bipartite graph. But for convenience of analyzing the canonical paths, we only focus on bipartite graph here. Note that this chain is irreducible since it is connected, and it is aperiodic due to self-loops introduced at step 3. Before analyzing this chain, we want to state some spectral properties related to the mixing time.

Suppose P is the transition matrix of a Markov Chain, and π is its stationary distribution. Then we know that $\pi^\intercal P = \pi$, and π is an eigenvector of P. Moreover, if π is the uniform distribution, the corresponding eigenvalue λ_0 is 1.

Theorem 4.4 (Perron-Fronbenius). *Sort the other eigenvalues of P in non-increasing order, then they are in the range $1 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n-1} > -1$.*

Recall the definition of spectral gap in Section 2. We have the following results.

Theorem 4.5. *Suppose Ω is the state space, and let $\pi_{min} = \min_{x \in \Omega} \pi(x)$. Then the mixing time of the Markov Chain $\tau(\varepsilon)$ is bounded by $\Omega(\frac{1}{Gap(P)}) = \tau(\varepsilon) = O(\frac{1}{Gap(P)} \log \frac{1}{\pi_{min}})$.*

Definition 4.6. *For $S \subseteq \Omega$, we define the* conductance *of S to be*

$$\Phi(S) = \mathsf{P}[X_{t+1} \notin S | X_t \in S, X_t \sim \pi] = \frac{\sum_{x \in S, y \in \overline{S}} \pi(x) P(x, y)}{\pi(S)}.$$

The conductance of the graph *is $\Phi = \min_{S : \pi(S) \leq \frac{1}{2}} \Phi(S)$.*

Theorem 4.7 (Cheeger's Inequality). *We have $\frac{\Phi^2}{2} \leq Gap(P) \leq \Phi$.*

Based on this theorem, we obtain similar bounds on the mixing time with conductance.

Theorem 4.8 (Lovász [10]). *We have*

$$\Omega\left(\frac{1}{\Phi}\right) = \tau(\varepsilon) = O\left(\frac{1}{\Phi^2} \log \frac{1}{\pi_{mix}}\right).$$

5 Canonical Paths

Definition 5.1. *For any pair $I, F \in \Omega$, define a* canonical path γ_{IF} *from I to F in ($\mathcal{M}$,E) to be $\gamma_{IF} = (I = z_0, z_1, \ldots, z_l = F)$ through adjacent states in the MC.*

For a transition $T \in E$, denote $cp(T) = \{(I, F) : T \in \gamma_{IF}\}$ the set of pairs $(I, F) \in \Omega$ whose canonical paths use transition T. Let $c^* = \max_T |cp(T)|$.

Definition 5.2. *Let $\Gamma = \{\gamma_{IF} | I, F \in \Omega\}$ be the set of all canonical paths. We define the* congestion *of the chain to be*

$$\rho(\Gamma) = \max_{T=(u,v)} \frac{1}{\pi(u)P(u,v)} \sum_{(x,y)\in cp(T)} \pi(x)\pi(y)|\gamma_{xy}|.$$

Theorem 5.3 (Jerrum [6]). *The mixing time of the lazy Markov chain is bounded by*

$$\tau_x(\epsilon) \leq 2\rho(2\log\varepsilon^{-1} + \log\pi(x)^{-1}).$$

Example 5.4 (Canonical path for matching [6]). Given two matchings I and F, we need to connect I and F by canonical path γ_{IF}. Along the path, we will have to insert or remove at least the edges in the symmetric difference $I \oplus F$. The set $I \oplus F$ decomposes into a collection of alternating paths and even-length cycles, and we will process the components one at a time.

Given an ordering of all vertices, we can then fix the order of components in $I \oplus F$ by smallest labeled vertex in each component. Within each component, we identify a "start vertex": in the case of a cycle this will be the smallest vertex, and in the case a path be this will be the smaller endpoint. Then we orient each path away from its start vertex, and each cycle so that the edge in I incident to the start vertex is oriented away from the start vertex. To get from I to F, we now process the components of $(V, I \oplus F)$ in the order of $P_1, P_2, \ldots, P_m$.

- In each cycle, we first remove the edge in I incident to the start vertex using transition (b) defined in the MC; then with a sequence of (c) transitions following the cycle's orientation, we replace I by F edges; finally we perform an (a) transition to add the edge in F incident to the start vertex.
- In each path, if the start vertex is incident to an F edge, we use (c) transitions along the path and finish by an (a) transition in case the path has odd length. If the start vertex is incident to an I edge, we start with a (b) transition, then use (c) transitions along the path, and finish with another (a) transition in case the path has odd length.

We illustrate the process using Figure 1 from [6], and conclude our description for canonical paths. ♦

To bound the mixing time of the chain, we want to bound above the congestion. Fix a transition $T : m \to m'$, we want to show that $cp(T) \leq |\mathcal{M}|\text{poly}(n)$. Consider the mapping $\eta_T : cp(T) \to \mathcal{M}$, if we can show the mapping is injective, then we can conclude that $cp(T) \leq |\mathcal{M}|$. By Theorem 5.3 we will be able to upper bound the mixing time. We say that T is *troublesome* if T is a (c) transition and the current component is a cycle. We denote by e_{IFT} the edge in I that is adjacent to the start vertex of the cycle begin processed by T. Hence we define the mapping as

$$\eta_T(I, F) = \begin{cases} (I \oplus F \oplus (M \cup M')) \setminus \{e_{IFt}\} & \text{if } t \text{ is troublesome,} \\ I \oplus F \oplus (M \cup M') & \text{otherwise.} \end{cases}$$

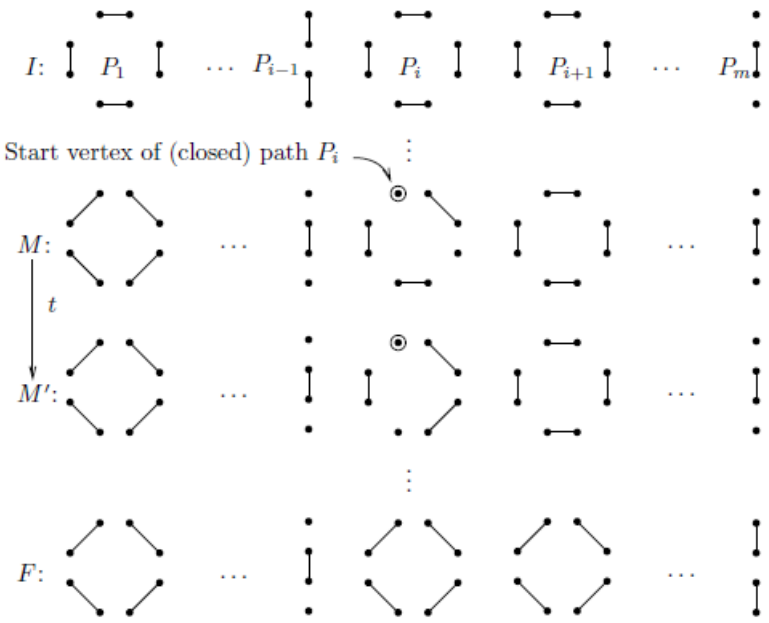

FIGURE 1: A step in a canonical path.

Claim 5.5. *For every transition T, the function $\eta_T : cp(T) \to \mathcal{M}$ is injective.*

Proof. Let $T = (M, M')$ be a transition, and $(I, F) \in cp(T)$, we wish to show that the pair (I, F) can be uniquely reconstructed from a knowledge only of T and $\eta_T(I, F)$. I t is immediate from the definition of η_T that $I \oplus F$ can be recovered from $E = \eta_T(I, F)$ using

$$I \oplus F = \begin{cases} (E \oplus (M \cup M')) \cup (e_{IFT}) & \text{if } T \text{ is } \textit{troublesome}, \\ E \oplus (M \cup M') & \text{otherwise.} \end{cases}$$

Given $I \oplus F$, we can at once infer the sequence of paths $P_1, P_2, \ldots, P_m$ that have to be processed along the canonical path, and the transition T tells us which of these is the current one, say P_i. The partition of $I \oplus F$ is as follows:

- I agrees with E on paths $P_1, \ldots, P_{i-1}$, and with M on paths $P_{i+1}, \ldots, P_m$.
- On current path P_i, the matching I agrees with E on already processed part, and with M on the rest.

Finally, the reconstruction of I and F is completed by noting that $I \cap F = M \setminus (I \oplus F)$, which is immediate from the definition of the paths. Hence I and F can be uniquely recovered from $E = \eta_T$, so η_T is injective. □

6 Markov Chain on Perfect Matchings

Definition 6.1. *Given G = (V,E), a matching N(u,v) is* near-perfect *if it contains only two unmatched vertices u and v, and N(u,v) is a perfect matching on* $G \backslash \{u, v\}$.

Let $\mathcal{N} = \bigcup_{u,v} N(u, v)$, and we can define the same Markov Chain on $\Omega = \mathcal{P} \cup \mathcal{N}$ (perfect matchings and near-perfect matchings), but we only move to X' if it is a perfect or near-perfect matching. Note that the chain is still ergodic and the stationary distribution is uniform.

Claim 6.2. *If G is dense, i.e., deg(v)>* $\frac{n}{2}$ *for all* $v \in V$, *then* $\frac{|\mathcal{P}|}{|\Omega|} \geq \frac{1}{n^2}$. *Hence* $\pi(\mathcal{P}) \geq \frac{1}{n^2}$.

For $I, F \in \Omega$, the canonical path γ_{IF} from I to F can be defined as follows:

- If $I, F \in \mathcal{P}$, γ_{IF} is the same as we defined in Section 5.
- If $I \in \mathcal{N}$, $F \in \mathcal{P}$, then $I \oplus F$ consists of one path and a few cycles. Then we first process the path, and get to a state I', which is in $\mathcal{P}$, and identify the path $\gamma_{I'F}$.
- If $I \in \mathcal{P}$, $F \in \mathcal{N}$, then $I \oplus F$ still consists of one path and cycles. Then we process the cycles first, and process the path at the end.
- If $I, F \in \mathcal{N}$, choose a perfect match $M \in \mathcal{P}$ uniformly at random, and define the canonical path $I \to M \to F$.

Unfortunately, the number of near-perfect matchings could be much larger than the number of perfect matchings, which can be seen from the following example.

Example 6.3 (Number of near-perfect matchings). Consider the "box graph" consisting of $n + 2$ vertices that form $n/4$ boxes. Fix an order of the vertices $(v_0, v_1, v_2, ..., v_{n+1})$. Each 4 vertices $(v_{4k+1}, v_{4k+2}, v_{4k+3}, v_{4k+4})$ form a cycle, and we connect two consecutive cycles using a single edge. Connect v_0 to the first cycle, and connect v_{n+1} to the last cycle. It is easy to see that there is only 1 perfect matching in this graph. However, the number of near perfect matchings is $2^{n/4}$. ♦

> **Exercise 6.1.** Show that in any dense bipartite graph, the ratio of the number of near-perfect matchings to the number of perfect matchings is bounded by a polynomial in n.

Definition 6.4 (Jerrum and Sinclair [7]). *An almost uniform generator for set N is a probabilistic algorithm that when presented with G and a positive real bias* ϵ, *outputs an element of N such that the probability of each element appearing approximates* $|N|^{-1}$ *within ratio* $1 + \varepsilon$.

Theorem 6.5 (Broder [2]). *Suppose that for all dense bipartite graph G, there exists a fully polynomial almost uniform generator (FPAUG) for* $\mathcal{M} \cup \mathcal{N}$. *Then there exists an FPRAS for* $|\mathcal{M}|$ *for all such graphs G.*

Since $|\mathcal{N}|/|\mathcal{M}|$ is bounded in dense bipartite graphs (this claim may not be true in general graphs), then we can use the Markov Chain as a FPAUG to sample perfect matchings. If it ends up with a near-perfect matching, we need to re-run the chain, but this probability is small in dense bipartite graph. We note that Jerrum et al. [8] show how to reduce counting to sampling in general.

On the other hand, we want to modify the Markov chain so that it would prefer moving to perfect matchings. We can achieve this by assigning each matching M a *weight* $w(M)$. If $M \in \mathcal{P}$, then $w(M) = 1$. If $M = N(u, v) \in \mathcal{N}$, then $w(M) = |\mathcal{P}|/|N(u, v)|$. Note that $w(M) \ll 1$ for $M \in \mathcal{N}$. Then we modify step 3 of the Markov chain accordingly:

$$X_{t+1} = \begin{cases} X' & \text{with probability } \min\{1, \frac{w(X')}{w(X_t)}\}, \\ X_t & \text{otherwise.} \end{cases}$$

Letting $Z = \sum_{M \in \Omega} w(M)$, we can see that $\pi(\mathcal{P}) = |\mathcal{P}|/Z$, and

$$\pi(N(u, v)) = \frac{|\mathcal{P}|}{|N(u, v)|} \cdot \frac{|N(u, v)|}{Z} = \frac{|\mathcal{P}|}{Z}.$$

Note that computing Z exactly is a hard problem. For if it could be done efficiently, one could compute $Z = Z(\lambda)$ at a sequence of distinct values of λ, and then extract the coefficients of $Z(\lambda)$ by interpolating the computed values. The following discussion will focus on bipartite graph.

To prove this chain is rapid mixing, we find $w'(M) \approx w(M)$ that gives fast mixing. Suppose we are sampling the matchings with stationary distribution $\pi_{w'}$ with corresponding weights w'. Similar to the previous analysis, we have $\pi_{w'}(\mathcal{P}) = |\mathcal{P}|/Z_{w'}$ and $\pi_{w'}(N(u, v)) = w'(u, v)|N(u, v)|/Z_{w'}$. Thus

$$\frac{\pi_{w'}(\mathcal{P})}{\pi_{w'}(N(u, v))} = \frac{|\mathcal{P}|}{w'(u, v)|N(u, v)|} = \frac{w(u, v)}{w'(u, v)}$$

with $w(u, v) = w'(u, v)\pi(\mathcal{P})/\pi(N(u, v))$. Hence given a bipartite graph $G = (V, E)$, we can compute $w'(u, v)$ iteratively using the following algorithm.

Algorithm: Calibrarte w.

For $i = 0 \to \lceil n^2 \ln n \rceil$:

1. Set $\lambda_i = (1 - 1/n)^i$.
2. Set
$$\lambda(u, v) = \begin{cases} 1 & \text{if } (u, v) \in E, \\ \lambda_i & \text{otherwise.} \end{cases}$$
3. Compute $\lambda(M) = \prod_{e \in M} \lambda(e)$.
4. Run $MC_{\text{near-perfect matching}}$ with λ_i and w_{i-1} from last iteration to update w_i.

To conclude, we list some open problems:

- Is there a Markov Chain on all matchings with $O(n^3)$ mixing time?
- Is there a Markov Chain on perfect matchings and near-perfect matchings for general graphs?

References

[1] D. Aldous. *Séminaire de Probabilités XVII 1981/82: Proceedings*, chapter Random walks on finite groups and rapidly mixing markov chains, pages 243–297. Springer Berlin Heidelberg, Berlin, Heidelberg, 1983.

[2] A. Z. Broder. How hard is it to marry at random? (on the approximation of the permanent). In *Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*, STOC '86, pages 50–58, New York, NY, USA, 1986. ACM.

[3] R. Bubley and M. Dyer. Path coupling: A technique for proving rapid mixing in markov chains. In *In FOCS 97: Proceedings of the 38th Annual Symposium on Foundations of Computer Science (FOCS)*, page 223, 1997.

[4] M. Dyer and C. Greenhill. A more rapidly mixing markov chain for graph colorings. *Random Structures & Algorithms*, 13(3-4):285–317, 1998.

[5] R. Holley. Remarks on the fkg inequalities. *Comm. Math. Phys.*, 36(3):227–231, 1974.

[6] M. Jerrum. Chapter 5: canonical paths and matchings.

[7] M. Jerrum and A. Sinclair. Approximating the permanent. *SIAM J. Comput.*, 18(6):1149–1178, December 1989.

[8] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform. *Theor. Comput. Sci.*, 43(2-3):169–188, July 1986.

[9] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov chains and mixing times*. Providence, R.I. American Mathematical Society, 2009. With a chapter on coupling from the past by James G. Propp and David B. Wilson.

[10] L. Lovász and R. Kannan. Faster mixing via average conductance. In *Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing*, STOC '99, pages 282–287, New York, NY, USA, 1999. ACM.

[11] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[12] J.G. Propp and D.B. Wilson. Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9:223–252, 1996.

[13] D. Randall. Rapidly mixing markov chains with applications in computer science and physics. *Computing in Science and Engineering.*, 8(2):30–41, March 2006.

[14] A. Sinclair. *Algorithms for Random Generation and Counting: A Markov Chain Approach*. Birkhäuser Verlag, Basel, Switzerland, 1993.

[15] E. Vigoda. Improved bounds for sampling colorings. *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 0:51, 1999.