Advanced Algorithms XII (Fall 2020)

Instructor: Chihao Zhang
Scribed by: Zhiyang Xun & Runzhe Wu

Last modified on Nov 27, 2020

We will step into Markov chain Monte Carlo (MCMC) methods, which are a class of algorithms for sampling from probability distributions. In this lecture, we will discuss some fundamental properties of Markov Chains.

1 The Power of Sampling

Sampling from a probability distribution has massive applications in theoretical computer science.

In many circumstances, we want to uniformly sample some combinatorial structures, say, an independent set in a graph G. The most straightforward method is to assign a random bit for each vertex to decide whether to pick it; then, test whether this vertex set is an independent set in G. The method is often called *rejection sampling*. Note that in many graphs, among all vertex subsets, only an exponentially small proportion is an independent set. Thus, this sampler needs to run exponential times to successfully sample an independent set, which is by no means efficient. Another example is to uniformly generate a proper coloring of a graph. With the help of MCMC, both problems can be solved efficiently under certain conditions.

In the two problems mentioned above, the problem of sampling uniformly from a set is computationally equivalent to giving an $(1 \pm \epsilon)$-approximation to the size of the set [1]. Thus efficient samplers for independent sets and proper colorings can be turned into efficient approximation algorithms for counting the number of independent sets or proper colorings respectively.

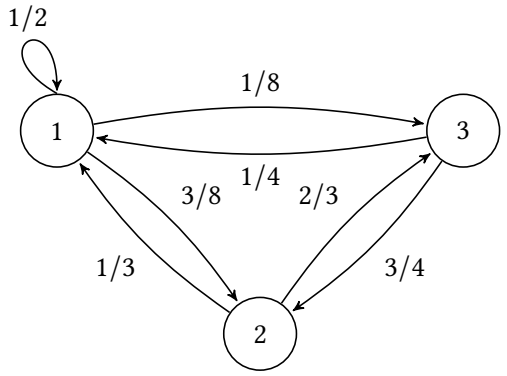
Another intriguing application of MCMC approach is to draw samples from a probability density function $f(\boldsymbol{x})$ in Euclidean spaces. In this model, we are able to query the value of $f(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^d$, or sometimes query the value of $\nabla f(\mathbf{x})$.

An important special case is that when $f(\boldsymbol{x})$ only values from $\{0, 1\}$. In this case, $f(\boldsymbol{x})$ actually defines a region in $\mathbb{R}^d$. If we are able to uniformly sample points in the region efficiently, we can estimate the volume of this region. In fact, when the region is convex, the MCMC is provably efficient for this task.

2 Markov Chain and Stationary Distribution

2.1 Introduction to Markov Chain

We can view a discrete Markov chain as a random walk on a graph. Let us see an example first:



This Markov chain has three states. For each state, it walks to other states with respective constant probability and the sum of these probabilities always equals to 1 for sure. This transition graph can be represented by the matrix

$$P = (P_{ij}) = \begin{bmatrix} 1/2 & 3/8 & 1/8 \\ 1/3 & 0 & 2/3 \\ 1/4 & 3/4 & 0 \end{bmatrix}.$$

The states of a Markov chain can be written as a sequence of random variables

$$X_0, X_1, \ldots, X_t, \ldots$$

In this lecture, we always assume our Markov chain is time-homogeneous, that is, the probability of any state transition is independent of time. Thus for every $t \geq 0$,

$$P_{ij} = \mathbf{Pr}\left[X_{t+1} = j \mid X_t = i\right].$$

At any time t, the distribution of X_t can be expressed as a vector μ_t, where

$$\mu_t(i) \triangleq \mathbf{Pr}\left[X_t = i\right].$$

Since

$$\mu_{t+1}(j) = \sum_i \mu_t(i) \cdot P_{ij},$$

we can see that

$$\mu_t^T P = \mu_{t+1}^T.$$

As a result, we have

$$\mu_t^T = \mu_0^T P^t.$$

With this formula, we can calculate the distribution vector of all time as long as we have μ_0 and P.

2.2 Stationary Distribution

Definition 1. *A distribution π is a stationary distribution if it remains unchanged in the Markov chain as time progresses, i.e.,*

$$\pi^T P = \pi^T$$

where P is the transition matrix of the Markov chain.

The idea of MCMC is to design a Markov Chain of which the stationary distribution is the given distribution. Therefore, we hope that as time progresses, our distribution function will gradually converge to a stationary distribution. If so, we can start from any state and simulate the Markov chain for a number of steps to get the stationary distribution. However, this does not always hold.

To better understand stationary distribution, we raise three questions:

Q1. Does each Markov chain have a stationary distribution?

Q2. If a Markov chain has a stationary distribution, is it unique?

Q3. If the chain has a unique stationary distribution, does μ_t always converge to it?

To answer the first question, we use the following theorem.

Theorem 2 (Perron-Frobenius Theorem). *Each nonnegative matrix A has a nonnegative real eigenvalue with spectral radius $\rho(A) = a$, and a has a corresponding nonnegative eigenvector.*

Now let us consider the transition matrix P for a Markov chain. Clearly it satisfies

$$P \cdot \vec{1} = \vec{1}.$$

Thus, P has an eigenvalue 1. Since every eigenvalue of P is no larger than the row sum, 1 is the largest eigenvalue. Also, P^T shares the same characteristic polynomial with P, which implies the eigenvalues of P^T and P are the same. As a result, $\rho(P^T)$ also equals to 1. According to Perron-Frobenius theorem, there exists a nonnegative eigenvector π such that

$$P^T \pi = \pi,$$

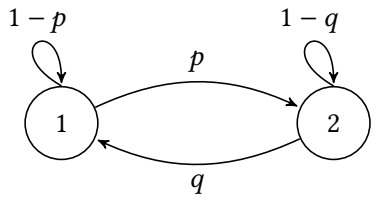
which is equivalent to

$$\pi^T P = \pi^T.$$

It follows that $\frac{\pi}{\|\pi\|_1}$ is a stationary distribution of P.

Theorem 3. *Each Markov chain has a stationary distribution.*

Now we consider Q2 and Q3. Consider the following simple Markov chain:



Clearly, the transition matrix of this Markov chain is

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

It is easy to verify that

$$\pi = \left(\frac{q}{p+q}, \frac{p}{p+q} \right)^T$$

is a stationary distribution of P. We are going to check whether starting from any μ_0, the distribution μ_t will always converge to π, i.e.,

$$\left\|\mu_0^T P^t - \pi^T\right\| \to 0.$$

In our example, the distribution has only two dimensions and the sum of the two components equals to 1, so we only need to check whether the first dimension converges, i.e.,

$$\left|\mu_0^T P^t(1) - \pi(1)\right| \to 0.$$

Now we define

$$\begin{aligned}
\Delta_t &\triangleq |\mu_t(1) - \pi(1)| \\
&= \left|\mu_{t-1}^T \cdot P(1) - \pi(1)\right| \\
&= \left|(1-p)\cdot\mu_{t-1}(1) + q\cdot(1-\mu_{t-1}(1)) - \frac{q}{p+q}\right| \\
&= \left|(1-p-q)\cdot\mu_{t-1}(1) + q\cdot\left(1-\frac{1}{p+q}\right)\right| \\
&= |1-p-q|\cdot\Delta_{t-1}
\end{aligned}$$

Therefore, we can see that $\Delta_t \to 0$ except in the two bad cases:

1. $p = q = 0$,
2. $p = q = 1$.

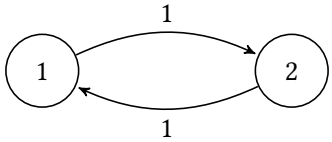
First we deal with the case when $p = q = 0$. The Markov chain will look like this:

The random walk graph is disconnected, so it can be partitioned into two irrelevant components. Since each component is still a Markov chain, each of them has its own stationary distribution. Notice that any convex combination of these small distributions is a stationary distribution for the whole Markov chain. It immediately follows that in this case the stationary distribution is not unique. It gives a negative answer to the second question.

This observation brings us to define a new property:

Definition 4. *A Markov chain is reducible if and only if there exists a state which cannot reach all states.*

When $p = q = 1$, the Markov chain looks like this:



This random walk graph is bipartite. It is easy to construct a μ_0 such that μ_t will jump periodically between 'left' and 'right'. Therefore, the answer to the third question is no. This phenomenon is captured by the following notion:

Definition 5. *A Markov chain is periodic if there exists some state v such that*

$$\gcd\{|c| \mid c \in C_v\} > 1,$$

where C_v denotes the set of the cycles that contains v.

We say a Markov chain is irreducible if it is not reducible, and it is aperiodic if it is not periodic.

3 Fundamental Theorem of Markov Chains

Lemma 6. *Given an irreducible, aperiodic, and finite Markov chain P, then it holds that*

1. $\forall x, y, \exists t: P^t(x, y) > 0$.
2. $\exists t, \forall x, y: P^t(x, y) > 0$.

Proof. It is clear that 2 implies 1. However, the proof of the latter's is based on the former's.

1. Being irreducible simply implies the existence of a path from x to y for any states x, y. Hence, starting at x, after some time t, one will eventually arrive at y with positive probability.
2. It suffices to prove that for each pair of states x, y, there exists a time threshold $t_{x,y}$ such that for every $t > t_{x,y}$, $P^t(x, y) > 0$ holds. Then, we can simply pick the largest $t_{x,y}$ over all pairs.

   By 1, starting at x, one will finally reach y in some time, say, t_0. Since it is an aperiodic Markov chain, the greatest common divisor of the lengths of all cycles passed through y will be exactly 1. Let's say the lengths are $l_1, l_2, \ldots, l_m$. Hence, $x_1 l_1 + x_2 l_2 + \cdots, x_m l_m = t$ will always have nonnegative solution $(x_1, x_2, \ldots, x_n)$ (x_i intuitively means the times to go through the i-th cycle) when t is sufficiently large, say, when t is larger than t'. That basically means after time $t_{x,y} \triangleq t_0 + t'$, $P^t(x, y) > 0$ always holds.

□

Theorem 7 (Fundamental Theorem of Markov Chains). *If a finite Markov chain $P \in \mathbb{R}^{n \times n}$ is irreducible and aperiodic, then it has a unique stationary distribution $\pi \in \mathbb{R}^n$. Moreover, for any distribution $\mu \in \mathbb{R}^n$,*

$$\lim_{t \to \infty} \mu^T P^t = \pi^t.$$

Proof. By lemma 6, let t_0 denote the time threshold such that for any $t > t_0$, $P^t(x, y) > 0$ holds for all x, y.

Set

$$\Pi = \begin{bmatrix} \pi^T \\ \pi^T \\ \vdots \\ \pi^T \end{bmatrix}.$$

Let $P^\infty \triangleq \lim_{t \to \infty} P^t$.

It is equivalent to showing $P^\infty = \Pi$ because for any distribution $\mu \in \mathbb{R}^n$, $\mu^T P^\infty$ can be considered as a convex combination of P^∞'s rows. Since by assumption the result of the convex combination is always π^T, the only possibility is that each row of P^∞ is exactly π^T.

To this end, it suffices to prove that $\lim_{k\to\infty}(P^{t_0})^k = \Pi$, because if that holds, then for any $0 < t < t_0$, $\lim_{k\to\infty} P^t \cdot (P^{t_0})^k = P^t\Pi = \Pi$.

As all entries of P^{t_0} are positive, we can decompose it into

$$P^{t_0} = \delta\Pi + (1-\delta)Q$$

where $0 < \delta \leq 1$ and all entries of Q is positive as well. In addition to it, Q is stochastic as well, i.e., it is also a transition matrix of some Markov chain.

Let $\theta \triangleq 1 - \delta$, we claim

$$\left(P^{t_0}\right)^k = (1-\theta^k)\Pi + \theta^k Q^k. \tag{1}$$

The proof of (1) is straightforward by induction. When $k = 1$, it holds trivially. Generally, by induction hypothesis,

$$\begin{aligned}
\left(P^{t_0}\right)^{k+1} &= \left[(1-\theta^k)\Pi + \theta^k Q^k\right] P^{t_0} \\
&= (1-\theta^k)\Pi + \theta^k Q^k P^{t_0} \\
&= (1-\theta^k)\Pi + \theta^k Q^k \left[(1-\theta)\Pi + \theta Q\right] \\
&= (1-\theta^k)\Pi + \theta^k\Pi - \theta^{k+1}\Pi + \theta^{k+1}Q^{k+1} \\
&= (1-\theta^{k+1})\Pi + \theta^{k+1}Q^{k+1}.
\end{aligned}$$

Note that $0 \leq \theta < 1$. Hence, when $k \to \infty$, $\left(p^{t_0}\right)^k \to \Pi$. That is, $\lim_{t\to\infty} P^t = \Pi$.

□

4 Reversibility

Definition 8. *A Markov chain is reversible if its stationary distribution π and transition probability matrix P satisfy the following* detailed balance condition*:*

$$\forall x, y : \pi(x)P(x,y) = \pi(y)P(y,x). \tag{2}$$

In fact, any probability distribution π satisfying (2) is a stationary distribution because for any state y,

$$(\pi^T P)(y) = \sum_x \pi(x)P(x,y) = \sum_x \pi(y)P(y,x) = \pi(y),$$

which implies $\pi^T P = \pi^T$.

The reversibility makes it easier to design a Markov chain P so that its stationary distribution is a certain given distribution. We only need to design P so that the detailed balance condition meets. Another reason for us to study reversible chains is that we can apply the powerful spectral tools to analyze the behavior of a chain.

4.1 Spectral Decomposition Theorem for Reversible Chains

We first introduce the Spectral Decomposition Theorem for symmetric matrices. It's proof can be found on linear algebra textbooks.

Theorem 9 (Spectral Decomposition Theorem). *Every $n \times n$ real symmetric matrix has n eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ together with their respective orthonormal eigenvectors $v_1, v_2, \cdots, v_n$.*

In other words, A can be decomposed as

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^T = Q \Lambda Q^T$$

where Q is an orthogonal matrix whose columns are the eigenvectors of A, and Λ is a diagonal matrix whose entries are the eigenvalues of A.

Theorem 10 (Spectral Decomposition Theorem for Reversible Chains). *Given a reversible Markov chain $P \in \mathbb{R}^n$ with a stationary distribution π, define the "weighted" inner product $\langle x, y \rangle_\pi \triangleq x^T D_\pi y = \sum_{i=1}^{n} x(i)y(i)\pi(i)$ where*

$$D_\pi = \begin{bmatrix} \pi(1) & & & \\ & \pi(2) & & \\ & & \ddots & \\ & & & \pi(n) \end{bmatrix}.$$

Then P has n eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ together with n respective orthonormal eigenvectors $v_1, v_2, \cdots, v_n$ with respect to the inner product $\langle \cdot, \cdot \rangle_\pi$.

Furthermore, we can write P as

$$P = \sum_{i=1}^{n} \lambda_i v_i v_i^T D_\pi.$$

Proof. Consider the matrix $Q \triangleq D_\pi^{1/2} P D_\pi^{-1/2}$. It is symmetric, for

$$\begin{aligned} Q(x, y) &= \pi(x)^{1/2} p(x, y) \pi(y)^{-1/2} \\ &= \pi(x)^{1/2} p(x, y)^{1/2} p(x, y)^{1/2} p(y, x)^{1/2} p(y, x)^{-1/2} \pi(y)^{-1/2} \\ &= \pi(x)^{-1/2} p(y, x) \pi(y)^{1/2} \\ &= Q(y, x). \end{aligned}$$

Applying theorem 9, we obtain its eigenvalues $\{\mu_i\}_{i=1}^n$ and orthonormal eigenvectors $\{w_i\}_{i=1}^n$, and thus

$$Q = \sum_{i=1}^{n} \mu_i w_i w_i^T. \tag{3}$$

Multiplying $D_\pi^{-1/2}$ and $D_\pi^{1/2}$ on each side of (3) respectively, we get

$$P = \sum_{i=1}^{n} \mu_i D_\pi^{-1/2} w_i w_i^T D_\pi^{1/2}. \tag{4}$$

Let $v_i \triangleq D_\pi^{-1/2} w_i$ and $\lambda_i \triangleq \mu_i$. Then, we can see

$$\langle v_i, v_j \rangle_\pi = w_i^T D_\pi^{-1/2} D_\pi D_\pi^{-1/2} w_j = \langle w_i, w_j \rangle = [i = j].$$

Therefore, the $\{v_i\}_{i=1}^n$ and $\{\lambda_i\}_{i=1}^n$ is exactly what we want. By those, P can be written as

$$P = \sum_{i=1}^{n} \lambda_i v_i v_i^T D_\pi.$$

□

Without loss of generality, we assume $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. We following theorem states some properties of the spectrum.

Theorem 11. *By the above definitions, we have the following properties for* $\{\lambda_i\}_{i=1}^n$*:*

1. $\lambda_1 = 1$.
2. $\lambda_n \geq -1$, *and* $\lambda_n = -1$ *iff* P *is bipartite (i.e.* P *is periodic).*
3. $\lambda_2 = 1$ *iff* P *is not connected (i.e.* P *is reducible)*

Theorem 12. *Under the previous setting,* $P^t = \sum_{i=1}^n \lambda_i^t v_i v_i^T D_\pi$.

Proof. When $t = 1$, it is exactly the result of theorem 10. For general cases, when $t > 1$, by induction hypothesis,

$$\begin{aligned} P^{t+1} &= \left(\sum_{i=1}^n \lambda_i^t v_i v_i^T D_\pi\right)\left(\sum_{i=1}^n \lambda_i v_i v_i^T D_\pi\right) \\ &= \sum_{i=1}^n \lambda_i^t v_i v_i^T D_\pi \cdot \lambda_i v_i v_i^T D_\pi \\ &= \sum_{i=1}^n \lambda_i^{t+1} v_i \left(v_i^T D_\pi v_i\right) v_i^T D_\pi \\ &= \sum_{i=1}^n \lambda_i^{t+1} v_i v_i^T D_\pi \end{aligned}$$

□

Therefore, we can write

$$P^t = \sum_{i=1}^n \lambda_i^t v_i v_i^T D_\pi = \Pi + \sum_{i=2}^n \lambda_i^t v_i v_i^T D_\pi$$

since $\lambda_1 = 1$ and $v_1 = \mathbf{1}$. By theorem 11, we find that P^t converges to Π iff $\lambda_2 < 1$ and $\lambda_n > -1$, which is exactly equivalent to P being aperiodic and irreducible. This matches the result of theorem 7.

Moreover, the convergence rate of P^t is obviously related to $|1 - \lambda_2|$ and $|1 - |\lambda_n||$. A common trick is to add loops into the graph so that all eigenvalues will become nonnegative, in which case we merely need to consider $|1 - \lambda_2|$.

References

[1] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani, *Random generation of combinatorial structures from a uniform distribution*, Theoretical computer science, 43 (1986), pp. 169–188. 1