13 Sampling Problems, Markov Chains

13.1 Sampling Problems

We've seen many decision problems and optimization problems. In the following lectures, we will introduce sampling problems and a powerful tool for sampling.

Let μ be the probability distribution on Ω given by $(x) \propto w(x)$, i.e.,

$$\mu(x) = \frac{w(x)}{Z}$$

for all $x \in \Omega$, where $Z = \sum_{x \in \Omega} w(x)$ is the normalizing factor. A sampling problem asks to sample a random element X from Ω according to such a distribution μ.

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 Markov chains, both problems can be solved efficiently under certain conditions.

13.2 Markov Chain Basics

We now introduce a powerful tool for sampling (approximately).

Definition 13.1. Discrete Markov chain
Suppose there is a sequence of random variables

$$X_0, X_1, \ldots, X_t, X_{t+1}, \ldots$$

where Range$(X_i) \subseteq \Omega$ for some countable set Ω. Then $\{X_n\}$ is a *discrete Markov chain* if $\forall\, t \geq 0$ and $\forall\, a_0, a_1, \ldots, a_{t+1} \in \Omega$,

$$\Pr[X_{t+1} = a_{t+1} \mid X_t = a_t, X_{t-1} = a_{t-1}, \ldots, X_0 = a_0] = \Pr[X_{t+1} = a_{t+1} \mid X_t = a_t].$$

This property is called *Markov property*, or *lack of memory*.

If for all $i, j \in \Omega$, there exists a constant $p_{i,j}$ such that

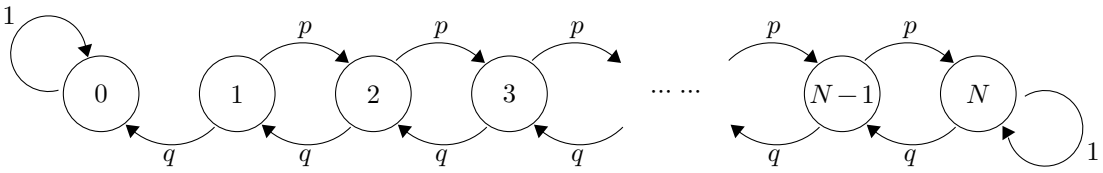
$$\forall\, t \geq 0, \qquad \Pr[X_{t+1} = j \mid X_t = i] = p_{i,j},$$

the Markov chain is called a *time-homogeneous* Markov chain.

Example 13.1.
Gambler's ruin
Consider a gambler who starts with an initial fortune of 1 and then on each successive gamble either wins 1 or loses 1 independent of the past with probabilities p and $q = 1 - p$ respectively. The gamble ends when the gambler reaches the total fortune of N (the gambler *wins*) or gets ruined (the gambler *loses*).
Let X_n be the total fortune after the n-th gamble. Then $X_0 = 1$ and for all $t \geq 0$,

$$\Pr[X_{t+1} = j \mid X_t = i] = \begin{cases} 1, & \text{if } i = j = 0; \\ 1, & \text{if } i = j = N; \\ p, & \text{if } 1 \leq i \leq N-1 \text{ and } j = i+1; \\ q, & \text{if } 1 \leq i \leq N-1 \text{ and } j = i-1. \end{cases}$$

We can also use a state-transition graph or an automaton to describe the Markov chain:



Example 13.2.
Random walk on $\mathbb{Z}$
The set of state Ω is $\mathbb{Z}$, and the transition probability is given by

$$p_{i,j} = \Pr[X_{t+1} = j \mid X_t = i] = \begin{cases} 1/2, & \text{if } |i-j| = 1; \\ 0, & \text{otherwise}. \end{cases}$$

Remark 13.1.
In this course, we only consider Markov chains whose state space Ω is a finite set. We can use an $|\Omega| \times |\Omega|$ matrix $\mathbf{P} \in [0,1]^{\Omega \times \Omega}$ to denote transition probabilities, where $\mathbf{P}(i,j) = p_{i,j}$.

Let μ_t be the distribution of X_t, i.e. $X_t \sim \mu_t$. Then $\forall\, i \in \Omega$,

$$\begin{aligned} \mu_{t+1}(j) &= \Pr[X_{t+1} = j] = \sum_{i \in \Omega} \Pr[X_{t+1} = j \wedge X_t = i] \\ &= \sum_{i \in \Omega} \Pr[X_t = i] \cdot \Pr[X_{t+1} = j \mid X_t = i] \\ &= \sum_{i \in \Omega} \mu_t(i) \cdot p_{i,j} \end{aligned}$$

Suppose that $\Omega = \{0, 1, \dots, N\}$, then we can use a column vector to denote μ_t, where

$$\mu_t = \begin{pmatrix} \mu_t(0), \\ \mu_t(1), \\ \vdots, \\ \mu_t(N)) \end{pmatrix}.$$

Thus ${\mu_{t+1}}^\mathsf{T} = {\mu_t}^\mathsf{T}\mathbf{P}$, and by induction, it follows directly that

$$\forall\, t \geq 0, \qquad {\mu_t}^\mathsf{T} = {\mu_0}^\mathsf{T}\mathbf{P}^t\,.$$

Suppose the initial state follows the distribution μ_0, then after implementing a Markov chain transition t steps, we will obtain a state whose distribution is $\mu_0^\mathsf{T}\mathbf{P}^t$.

A natural question is, does $\mathbf{P}^t$ converge as $t \to \infty$?

13.3 Stationary Distribution

Let $\mathbf{P}$ be a Markov chain transition matrix. It is clear that for all $i \in \Omega$,

$$\sum_{j\in\Omega} p_{i,j} = \sum_{j\in\Omega} \Pr[X_{t+1} = j \mid X_t = i] = 1\,,$$

namely, all row sums are equal to 1. This kind of matrix is called *stochastic matrix*.

Definition 13.2. Stochastic Matrix
A *stochastic matrix* $\mathbf{P}$ is a square matrix whose rows are probability vectors, i.e. $\mathbf{P} \in [0,1]^{\Omega\times\Omega}$ and $\forall i, \sum_{j\in S} \mathbf{P}(i,j) = 1$.

If $\mu_0^\mathsf{T}\mathbf{P}^t$ converges to some distribution ν^T as $t \to \infty$, then it holds that

$$\nu^\mathsf{T}\mathbf{P} = \lim_{t\to\infty} \mu_0^\mathsf{T}\mathbf{P}^{t+1} = \nu^\mathsf{T}\,.$$

Such distribution is called a *stationary distribution*.

Definition 13.3. Stationary distribution
Let $\{X_n\}$ be a Markov chain with transition matrix $\mathbf{P}$. Suppose π is a distribution such that

$$\pi^\mathsf{T}\mathbf{P} = \pi^\mathsf{T}\,.$$

Then π is called a *stationary distribution* of Markov chain $\{X_n\}$.

Remark 13.2.
Note that the definition of the stationary distribution does not depends on the convergence of the Markov chain. We will see the connection between them later.

We now give some examples here.

Suppose $G = (V, E)$ is an undirected graph with n vertices and m edges. Let $d_i = \deg(i)$ denote the degree of vertex i. Consider the following random walk on G:

$$\mathbf{P}(i,j) = \begin{cases} 1/d_i, & \text{if } i \sim j; \\ 0, & \text{if } i \nsim j\,. \end{cases}$$

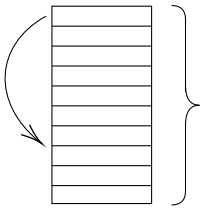
Then what is the stationary distribution π?

Suppose G is a regular graph, i.e., d_i is a constant d for all i. Then $\pi = (1/n, 1/n, \dots, 1/n)^\mathsf{T}$. If G is not a regular graph, we claim that the stationary distribution is

$$\pi = \left(\frac{d_1}{\sum d_k}, \frac{d_2}{\sum d_k}, \dots, \frac{d_n}{\sum d_k} \right)^\mathsf{T}.$$

Note that $\sum d_k = 2m$. So $\pi = (d_1/2m, d_2/2m, \dots, d_n/2m)^\mathsf{T}$. It is easy to verify that π is indeed a stationary distribution. We will leave it as an exercise here.

Another example is card shuffling. Consider a naïve "top-to-random" card shuffle:



Suppose we have n cards, everytime we take the top card of the deck and insert it into the deck at one of the n distinct possible places uniformly at random.

Let i, j be two permutations on $[n]$. W.l.o.g. assume that $i = (1, 2, \dots, n)$. Then $\mathbf{P}(i, j) > 0$ iff there exists k s.t. $j = (2, 3, \dots, k, 1, k+1, \dots, n)$.

Performing the shuffle repeatedly is a Markov chain. It is not difficult to verify that the uniform distribution $(1/n!, 1/n!, \dots, 1/n!)^\mathsf{T}$ over all permutations is a stationary distribution. We leave it as an exercise again.

For stationary distributions, we have the following three questions: under which condition can we prove

1. the existence of a stationary distribution π?
2. the uniqueness of π?
3. the convergence of the Markov chain?

We first answer the first question. The stationary distribution is equivalence to a nonnegative left eigenvector with eigenvalue 1. Since $\mathbf{P}$ is a stochastic matrix, it is clear that it has a right eigenvalue 1 with eigenvector $\mathbf{1} = (1, 1, \dots, 1)^\mathsf{T}$. Can we find a left eigenvector with the same eigenvalue?

Definition 13.4. Spectral radius

Let A be a $n \times n$ nonnegative matrix. Then the *spectral radius* of A, denoted by $\rho(A)$, is the maximum norm of its eigenvalues. Namely,

$$\rho(A) = \max \{|\lambda| : \det(\lambda I - A) = 0\}.$$

Proposition 13.1.

Let $A = (a_{i,j}) \in \mathbb{R}_{\geq 0}^{n \times n}$ be a nonnegative matrix, i.e., $a_{i,j} \geq 0$. Then

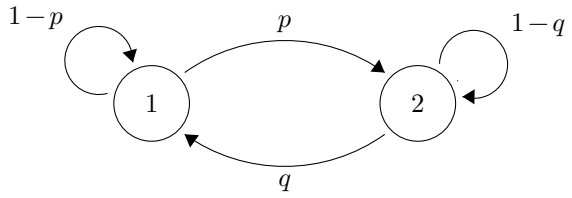
$$\min_{1 \leq i \leq n} \sum_{j=1}^{n} a_{i,j} \leq \rho(A) \leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} a_{i,j}.$$

Theorem 13.2. Perron-Frobenius Theorem

Let $A = (a_{i,j}) \in \mathbb{R}_{\geq 0}^{n \times n}$ be a nonnegative matrix with spectral radius $\rho(A) = \alpha$. Then α is a eigenvalue of A, and has both left and right nonnegative eigenvectors.

The celebrated Perron-Frobenius Theorem answers the first question. Let $\mathbf{P}$ be a stochastic matrix. Then Proposition 13.1 implies that $\rho(\mathbf{P}) = 1$. So $\mathbf{P}$ has eigenvalue 1 and both left and right nonnegative eigenvectors, that is, there exists $\pi \geq \mathbf{0}$ s.t. $\pi^{\mathsf{T}}\mathbf{P} = \pi^{\mathsf{T}}$. Normalizing π we can obtain a distribution vector.

For the second and the third question, let's consider the following Markov chain.



Then $\pi = \left(\frac{q}{p+q}, \frac{p}{p+q}\right)^{\mathsf{T}}$ is a stationary distribution. Now let $\Delta_t \triangleq |\mu_t(1) - \pi(1)|$. We have

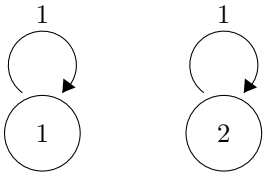
$$\begin{aligned}
\Delta_t &= \left|(\mu_{t-1}{}^{\mathsf{T}}\mathbf{P})(1) - \pi(1)\right| \\
&= \left|(1-p)\cdot\mu_{t-1}(1) + q\cdot(1-\mu_{t-1}(1)) - \frac{q}{p+q}\right| \\
&= \left|(1-p-q)\cdot\mu_{t-1}(1) + q\cdot\left(1-\frac{1}{p+q}\right)\right| \\
&= |1-p-q|\cdot\Delta_{t-1}\,.
\end{aligned}$$

So $\Delta_t \to 0$ except for

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

Case 1: $p = q = 0$

In this case, we call the Markov chain *reducible* and the stationary distributions may not be unique.

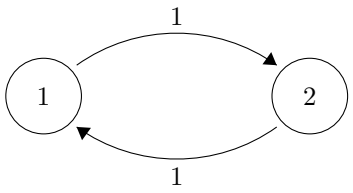


Definition 13.5. Reducibility
We say that j is *accessible* from i iff $\exists t > 0$, s.t. $\Pr[X_t = j \mid X_0 = i] > 0$. Moreover, i *communicates* with j if i is accessible from j and j is accessible from i. We also define an equivalence relation $i \simeq j$: $i \simeq j$ iff i communicates with j. Then a Markov chain is *irreducible* iff the number of equivalent classes is 1. In other words, the state graph is strongly connected. Otherwise, the Markov chain is called *reducible*.

If a Markov chain is irreducible, then its stationary distribution is unique. Otherwise its stationary distributions may not be unique.

Case 2: $p = q = 1$

In this case, $X_t = X_0$ if t is even and X_t is the other state if t is odd. Then the Markov chain is called a *periodic* chain.



For all $i \in \Omega$, let $d_i \triangleq \gcd\{u : \mathbf{P}^u(i,i) > 0\}$. Namely, d_i is the greatest common divisor of the length of all loops starting from i and ending at i. Then we have the following lemma.

Lemma 13.3.
If i and j communicate with each other, then $d_i = d_j$.

Proof. Suppose that $\mathbf{P}^{n_1}(i,j) > 0$, $\mathbf{P}^{n_2}(j,i) > 0$ and $\mathbf{P}^n(j,j) > 0$. Note that $d_i \mid (n_1 + n_2)$ and $d_i \mid (n_1 + n_2 + n)$. Thus $d_i \mid n$. It is easy to see that for all n that $\mathbf{P}^n(j,j) > 0$, $d_i \mid n$. So $d_i \mid d_j$, and vice versa. □

Definition 13.6. Periodicity
A Markov chain is *aperiodic* if $d_i = 1$ for all i, and is *periodic* otherwise.

Clearly, periodic chains do not converge.

In fact, these two cases are the only cases that the answer to the second or the third question is "No".

Theorem 13.4. Fundamental Theorem of Markov Chains
If a finite Markov chain $\{X_n\}$ with transition matrix $\mathbf{P}$ is irreducible and aperiodic, then there exists a unique stationary distribution π, and

$$\forall \mu, \qquad \lim_{t\to\infty} \mu^\mathsf{T} \mathbf{P}^t \to \pi^\mathsf{T}.$$