**AI2613 随机过程** 2020–2021 春季学期

Lecture 6 – Examples and Applications of Markov Chains (III)

2021 年 3 月 29 日

*Lecturer*: 张驰豪 *Scribe*: 杨宽

Today we are going to talk about some algorithmic applications of Markov chains. For simplicity, we assume that all Markov chains we discuss today are (I) and (A).

# 1 (Time-)Reversible Markov Chains and Metropolis Algorithm

**Definition 1** ((Time-)Reversible Markov Chains). A Markov chain is called *reversible* (or *time-reversible*) if there exists a distribution $\pi$ s.t.

$$\forall\, x, y \in \mathcal{S}, \qquad \pi(x)\cdot \mathbf{P}(x,y) = \pi(y)\cdot \mathbf{P}(y,x)\,.$$

The equation above is called the *detailed balance condition*.

**Proposition 1**. *If $\pi$ exists, then $\pi$ is the stationary distribution of* $\mathbf{P}$.

*Proof.* We now verify that $\pi$ is the stationary distribution:

$$\begin{aligned}(\pi^{\mathsf{T}}\mathbf{P})(y) &= \sum_{x\in\mathcal{S}} \pi(x)\cdot\mathbf{P}(x,y)\\ &= \sum_{x\in\mathcal{S}} \pi(y)\cdot\mathbf{P}(y,x)\\ &= \pi(y)\cdot\sum_{x\in\mathcal{S}}\mathbf{P}(y,x) = \pi(y)\,.\end{aligned}$$

□

*Remark.* Checking the detailed balance condition is usually the simplest way to verify that a particular distribution is stationary. Furthermore, the detailed balance condition implies that for all $x_0, x_1, \ldots, x_n$,

$$\pi(x_0)\mathbf{P}(x_0,x_1)\cdots\mathbf{P}(x_{n-1},x_n) = \pi(x_n)\mathbf{P}(x_n,x_{n-1})\cdots\mathbf{P}(x_1,x_0)\,,$$

namely,

$$\mathbf{Pr}_{X_0\sim\pi}[X_0=x_0, X_1=x_1,\ldots,X_n=x_n] = \mathbf{Pr}_{X_0\sim\pi}[X_0=x_n, X_1=x_{n-1},\ldots,X_n=x_0]\,.$$

Thus, if a Markov chain $\{X_t\}$ satisfies the detailed balance condition and starts from the stationary distribution, then the distribution of $(X_0, X_1, \dots, X_n)$ is the same as the distribution of $(X_n, X_{n-1}, \dots, X_0)$, and that's why the chains satisfying the detailed balance condition are called *time-reversible*.

**Example 2** (Random walks on graphs). Recall the simple random walk on graphs that we mentioned in the second lecture.
Given an undirected graph $G = (V, E)$, we define the random walk as follows.
Let $X_0, X_1, \dots X_t, \dots \in V$, and for each $X_i$, pick a neighbor $u$ of $X_i$ uniformly at random and let $X_{i+1} = u$.
We showed in the second lecture that the stationary distribution of this Markov chain is

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

We now verify that this Markov chain is time-reversible:

$$\pi(i) \cdot \mathbf{P}(i,j) = \frac{d_i}{\sum d_k} \cdot \frac{\mathbb{1}_{[i \sim j]}}{d_i} = \frac{\mathbb{1}_{[i \sim j]}}{\sum d_k} = \frac{d_j}{\sum d_k} \cdot \frac{\mathbb{1}_{[i \sim j]}}{d_j} = \pi(j) \cdot \mathbf{P}(j,i),$$

where we use $\sim$ to denote the relation of adjacency.

**Question**. Given an arbitrary distribution $\mu$, can we design a random walk on the graph s.t. its stationary distribution is $\mu$?

**Example 3** (Metropolis Algorithm). Let $\Delta = \max_{i \in V} \deg(i)$. Then for all $i \in V$, our algorithm (random walk) moving from $i$ has two steps:

1. for every neighbor $j$ of $i$, propose to move to $j$ with probability $1/\Delta$;
2. accept with probability $\min\{\frac{\mu(j)}{\mu(i)}, 1\}$.

Formally, we define the entries in the transition matrix $\mathbf{P}$ as follows:

$$\mathbf{P}(i,j) = \begin{cases} 0, & \text{if } i \not\sim j \text{ and } i \neq j; \\ \frac{1}{\Delta} \min\{\frac{\mu(j)}{\mu(i)}, 1\}, & \text{if } i \sim j; \\ 1 - \sum_{i \sim j} \mathbf{P}(i,j), & \text{if } i = j. \end{cases}$$

We now verify that $\mu$ is indeed the stationary distribution of $\mathbf{P}$. If $i = j$ or $i \not\sim j$, it is clear that $\mu(i)\mathbf{P}(i,j) = \mu(j)\mathbf{P}(j,i)$. So we assume that $i \sim j$, and w.l.o.g. we further assume that $\mu(j) \geq \mu(i)$. Since

$$\mu(i)\mathbf{P}(i,j) = \mu(i) \cdot \frac{1}{\Delta} \cdot \min\left\{\frac{\mu(j)}{\mu(i)}, 1\right\}$$

we obtain that

$$\mu(j)\mathbf{P}(j,i) = \mu(j) \cdot \frac{1}{\Delta} \cdot \frac{\mu(i)}{\mu(j)} = \frac{\mu(i)}{\Delta}$$

and

$$\mu(i)\mathbf{P}(i,j) = \mu(i) \cdot \frac{1}{\Delta} = \frac{\mu(i)}{\Delta}\,.$$

**Question**. However, if the distribution $\mu$ is already known, why do we design a Markov chain instead of sampling directly?

In fact, for most algorithmic applications, the desired distribution $\mu$ is unknown or hard to compute, but it is often much easier to calculate $\mu(i)/\mu(j)$. The key is that computing $\mu(i)$ directly costs $\Theta(|\mathcal{S}|)$ times of calculation while the state distribution of a Markov chain may be sufficiently close to stationary within $o(|\mathcal{S}|)$ runs.

**Example 4**. Let $\mathcal{S} = [n] = \{1, 2, \ldots, n\}$. For all $i \in \mathcal{S}$, a weight $w(i)$ is given (by an oracle). Our goal is to sample $i \in \mathcal{S}$ with the distribution $\mu$ satisfying

$$\mu(i) \propto w(i)\,,$$

namely,

$$\mu(i) = \frac{w(i)}{\sum_k w(k)}\,.$$

Then $\mu(i)/\mu(j)$ is easy to compute and the Markov chain is possible to *mix rapidly*.

## 2 Simulated Annealing

Given a set $\mathcal{S}$, for all $x \in \mathcal{S}$, $x$ has a weight $w(x)$. Our goal is to find an element $x$ to minimize $w(x)$.

**Example 5** (Maximum Independent Set). Given a graph $G = (V, E)$, a set $I$ of vertices is called an *independent set* iff

$$\forall\, i, j \in I, \qquad i \not\sim j \text{ (i.e., } (i,j) \notin E)\,.$$

The problem of maximum independent set is to find an independent set $I$ of maximum size.
In other words, given an independent set $I$, let $w(I) = |V \setminus I|$. Our goal is to find an independent set $S$ minimizing $w(S)$.

*Remark.* We will show in the Algorithm course that the problem of maximum independent set is NP-hard.

Intuitively, we can define a distribution $\mu$ over $\mathcal{S}$, where the elements of small weights have large probability density, and then we can sample from the distribution. If the distribution we find has sufficiently good properties, the element sampled from the distribution can become a sufficiently good approximation.
However, for most situations we cannot ganrantee the time of convergence. To formalize this problem, we first give an example.
We introduce a parameter $T > 0$ (which is called the *temperature* in the simulated annealing method). For any $T$, let

$$\mu_T(x) \sim \mathrm{e}^{-w(x)/T}.$$

Then

$$\mu_T(x) = \frac{\mathrm{e}^{-w(x)/T}}{\sum_y \mathrm{e}^{-w(y)/T}}.$$

Clearly, for any $T > 0$, elements of larger weights will have smaller probability density. Moreover, the smaller the parameter $T$ is, the more the probability density will be concentrated on the most weighted elements.
Our goal is to find an element $x$ minimizing $w(x)$. Let $\mathcal{S}^* = \{x \in \mathcal{S} : w(x) = \min_{y \in \mathcal{S}} w(y)\}$ be the set of all $x$ to optimize $w(x)$. Namely $\mathcal{S}^* = \arg\min_{x \in \mathcal{S}} w(x)$. So in other words our goal is to find an element in $\mathcal{S}^*$.
We also let $\mu^*$ be the uniform distribution over $\mathcal{S}^*$. Then $\mu^*$ puts positive probability only on globally optimal solutions of our optimization problem. If we can sample from $\mu^*$, then our problem is solved. However, usually it is impossible to sample from $\mu^*$ directly, because the state graph of the corresponding Markov chain may not be connected at all. We will see an example of independent sets later. Now, the following fact indicates that we can sample from $\mu_T$ for sufficiently small $T$ instead of sampling from $\mu^*$ directly.

**Fact 2.** *As $T \downarrow 0$, $\mu_T \xrightarrow{D} \mu^*$* (convergence in distritubion), *that is, for all $x$,*

$$\lim_{T \to 0} \mu_T(x) = \mu^*(x).$$

We do not give a rigorous proof here, but we provide some intuitions. Fix an $x \in \mathcal{S}^*$. For all $y \notin \mathcal{S}^*$,

$$\frac{\mu_T(y)}{\mu_T(x)} = \frac{\mathrm{e}^{w(y)/T}}{\mathrm{e}^{w(x)/T}} = \mathrm{e}^{\frac{w(x)-w(y)}{T}}.$$

Since $w(x) < w(y)$, we obtain that $\mu_T(y)/\mu_T(x) \to 0$ as $T \to 0$.

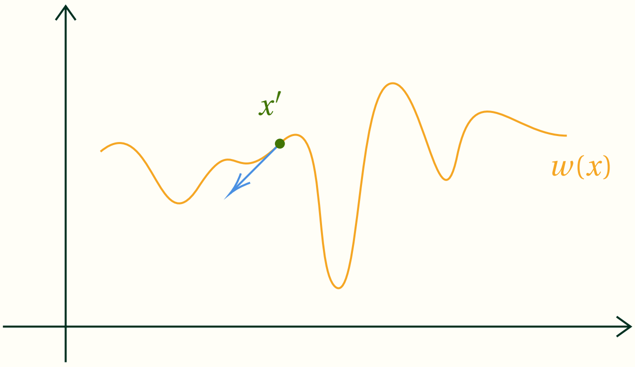
Now let's see the example of maximum independent sets.

Let $\mathcal{S} = \{x \in \{0,1\}^V : x \text{ is an independent set}\}$, and let $x \sim y$ iff $\sum_{i=1}^{|V|} \left|x(i) - y(i)\right| = 1$. Then for any $T > 0$, we define a Markov chain whose stationary distribution is $\mu_T$ via the Metropolis Algorithm:

$$\mathbf{P}_T(x,y) = \begin{cases} \frac{1}{n+1} \min\{\mathrm{e}^{(w(x)-w(y))/T}, 1\}, & \text{if } x \sim y; \\ 1 - \sum_z \frac{1}{n+1} \min\{\mathrm{e}^{(w(x)-w(z))/T}, 1\}, & \text{if } x = y. \end{cases}$$

Moreover, in this problem the set consisting of all maximum independent sets (i.e., sets with positive probability in $\mu^*$) is not connected. In fact, even every *maximal* independent set is an isolated vertex. So we consider sampling from $\mu_T$ instead of $\mu^*$.

However if $T$ is small enough (such as $T = 0$), the algorithm has a problem: at temperature $T = 0$, the process will never make an uphill move. Thus, running at temperature 0 is a descent method, which will get stuck in local minima, and therefore will not approach global minima. For example, in the following figure, if the algorithm samples $x'$ at some step, then with very high probability (even probability 1 if $T = 0$) it will go left next. However the global minimum is to the right of $x'$.



So what is the simulated annealing method? Intuitively if we choose a great $T$ at beginning, and then gradually lower the temperature, we can get the chain to converge in distribution to $\mu^*$. We must lower the temperature slowly enough so that the chain can always "catch up" and remain close to the stationary distribution for the current temperature.

Therefore, simulated annealing is not just another "descent method", since we allow ourselves positive probability of taking steps that increase the weight $w$. This feature of the procedure prevents it from getting stuck in local minima.

Specifically, a simulated annealing procedure can be described as follows.

Choose a "*cooling schedule*" $T_0, T_1, \ldots$; the schedules we will discuss later will have the property that $T_n \downarrow 0$ as $n \to \infty$. Choose the initial state $X_0$ according to an arbitrary distribution $\nu_0$. Let the succession of states $X_0, X_1, X_2, \ldots$ form a *time-inhomogeneous* Markov chain with probability transition matrices $\mathbf{P}_{T_0}, \mathbf{P}_{T_1}, \mathbf{P}_{T_2}, \ldots$, so that

$$\mathbf{Pr}\left[X_{n+1} = j \mid X_n = i\right] = \mathbf{P}_{T_n}(i, j),$$

and

$$X_n \sim \nu_n \qquad \text{where } {\nu_n}^\top = {\nu_0}^\top \cdot \mathbf{P}_{T_0} \cdot \mathbf{P}_{T_1} \cdot \cdots \cdot \mathbf{P}_{T_{n-1}}.$$

Then the following theorem shows that $\nu_n$ will converge to $\mu^*$ if we choose a cooling schedule decreasing "slowly enough".

**Definition 6.** Define the *radius* $r$ of the state graph $G$ of the Markov chain by

$$r = \min_{i \in \mathcal{S}_c} \max_{j \in \mathcal{S}} \operatorname{dist}_G(i, j)$$

where $\mathcal{S}_c = \{i \in \mathcal{S} : w(j) > w(i) \text{ for some } j \sim i\}$ be the set of all vertices that are not local maxima of $w$.

Define $L$ by the largest "*local fluctuation*" of $w(\cdot)$, that is,

$$L = \max_{i \in \mathcal{S}} \max_{j \sim i} \left| w(j) - w(i) \right|.$$

**Theorem 3** (Simulated Annealing). *For any cooling schedule $T_0, T_1, \ldots$ satisfying*

1. $T_n \downarrow 0$
2. $\sum_k \exp\left(-rL/T_{kr-1}\right) = \infty$,

*we have*

$$\forall \nu, \qquad \nu^\top \cdot \mathbf{P}^{(n)} \xrightarrow{D} (\mu^*)^\top \textit{ as } n \to \infty,$$

*where*

$$\mathbf{P}^{(n)} \triangleq \prod_{k=0}^{n-1} \mathbf{P}_{T_k}.$$

*Remark.* Taking $\gamma > rL$ and

$$T_n = \frac{\gamma}{\log n}$$

for $n > 1$, it is easy to check that the conditions of the theorem hold.

*Remark.* The theorem tells us a simulated annealing procedure will converge to the optimal value, but it doesn't claim any results to the convergence time.

# 3 Mixing Times

Now we are ready to discuss the speed of convergence (or formally, *mixing*) of Markov chains. Recall the proof of the fundamental theorem of Markov chains. In the last lecture, we used coupling to prove the theorem. We constructed two sequence of random variables

$$\begin{tikzcd}
\mu_0 & \mu_1 & & & \mu_t & & \\
\wr & \wr & & & \wr & & \\
X_0 \arrow[r] & X_1 \arrow[r] & X_2 \arrow[r] & \cdots \arrow[r] & X_t \arrow[r] & X_{t+1} \arrow[r] & \cdots \\
Y_0 \arrow[r] & Y_1 \arrow[r] & Y_2 \arrow[r] & \cdots \arrow[r] & Y_t \arrow[r] & Y_{t+1} \arrow[r] & \cdots \\
\wr & \wr & & & \wr & & \\
\pi & \pi & & & \pi & &
\end{tikzcd}$$

and a coupling $\mathcal{C}_t$ such that once $X_t = Y_t$ then $X_t$ and $Y_t$ move together forever. The coupling lemma establishes the connection between the distance of distributions and the discrepancy of random variables. To show that $\|\mu_t - \pi\|_{\mathrm{TV}} \to 0$, it is sufficient to compute $\mathbf{Pr}[X_t \neq Y_t]$.

Property $(\mathrm{I}) + (\mathrm{A})$ ganrantees that there exists $t > 0$, s.t. $\forall\, i, j$, $\mathbf{P}^t(i,j) > 0$. So $\mathbf{Pr}[X_t = Y_t] > 0$. Let

$$\theta \triangleq \mathbf{Pr}[X_t = Y_t] \geq \min_{i,j} \left(\mathbf{P}^t(i,j)\right)^2.$$

Applying the Markov property, we have

$$\mathbf{Pr}[X_{kt} \neq Y_{kt}] \leq (1-\theta)^k$$

and thus $\mathbf{Pr}[X_n = Y_n] \to 1$ as $n \to \infty$.

Intuitively, the proof yields that there exists constants $\alpha \in (0,1)$ and $C > 0$ s.t.

$$\max_{x \in \mathcal{S}} \left\|\mathbf{P}^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \leq C \cdot \alpha^t.$$

Clearly the smaller $\alpha$ is, the faster the Markov chain converges. Note that $\alpha$ depends only on the coupling we designed, so if we have a smart way to design a coupling which makes $\alpha$ very small, then we can bound the speed of convergence.

It is useful to introduce a parameter which measures the time of convergence. We use the *mixing time* $\tau_{\mathsf{mix}}(\varepsilon)$ to formalize the time required by a Markov chain for the distance to stationarity to be small.

**Definition 7** (Mixing Time). Let $\{X_t\}$ be a Markov chain with transition matrix $\mathbf{P}$ and stationary distribution $\pi$. Then the *mixing time* of the Markov chain is given by

$$\tau_{\mathsf{mix}}(\varepsilon) = \max_{\mu_0} \min\{t : \|\mu_t - \pi\|_{\mathrm{TV}} \le \varepsilon\},$$

where $\mu_t{}^{\top} = \mu_0{}^{\top} \cdot \mathbf{P}^t$.

**Example 8** (Random walk on $n$-dimensional hypercube). An $n$-dimensional hypercube is the set $\{0,1\}^n$ where

$$x \sim y \iff \sum_{i=1}^{n} |x(i) - y(i)| = 1.$$

Let $\{X_t\}$ be the random walk on the $n$-dimensional hypercube:

$$X_{t+1} = \begin{cases} X_t, & \text{with probability } 1/2; \\ \text{one of } X_t\text{'s neighbor u.a.r.}, & \text{with probability } 1/2. \end{cases}$$

To bound the mixing time of this Markov chain, it is sufficient to design a coupling such that for all $x_0$ and $y_0$, $\mathbf{Pr}[X_t = Y_t \mid X_0 = x_0, Y_0 = y_0]$ is large enough for some $t > 0$. Note that every coupling gives a bound of $\mathbf{Pr}[X_t = Y_t \mid X_0 = x_0, Y_0 = y_0]$ and thus gives a bound of mixing time. Our goal is to find a coupling which gives the bound as small as possible.

Actually, the random walk on hypercubes has another description: for any $X_t$, we pick a position $i \in [n]$ and a value $c \in \{0,1\}$ independently and uniformly at random, and then let $X_{t+1} = X_t^{i \leftarrow c}$. We claim that the Markov chain is the same as the one defined above.

The advantage of this definition is that it induces a natural coupling directly. For any $t > 0$, let the transfer of $X_t$ and $Y_t$ share the same randomness. Specifically, pick a position $i \in [n]$ and a value $c \in \{0,1\}$ independently and uniformly at random, and then let $X_{t+1} = X_t^{i \leftarrow c}$ and $Y_{t+1} = Y_t^{i \leftarrow c}$. As long as we choose some position $i$, the values of $X_t$ and $Y_t$ at this position will be the same from now on. So

$$\mathbf{Pr}[X_t = Y_t] \ge \mathbf{Pr}[\text{pick all positions in the first } t \text{ choices}].$$

The probability on the right side is a well-known problem called *coupon collector*. Now we claim that for $t > n\ln n + cn$, $\mathbf{Pr}[X_t \ne Y_t] < \mathrm{e}^{-c}$. Thus the mixing time of the random walk is bounded by

$$\tau_{\mathsf{mix}}(\varepsilon) \le \min\{t : \mathbf{Pr}[X_t \ne Y_t] < \varepsilon\} \le n\ln n + n\ln(1/\varepsilon).$$

Finally we introduce the problem of coupon collector.

**Example 9** (Coupon Collector). Suppose there are $n$ distinct types of coupons and someone would like to collect a complete set of all types of coupons. Each purchase gives an object from all $n$ types of coupons independently and uniformly at random. Let $X$ be the random variable that denotes the number of purchases until collecting all $n$ types of coupons.

**Lemma 4.** *The expectation of $X$ with $n$ distinct types of coupons is $n \ln n + O(n)$. Furthermore, the probability that the collecting process does not end after $n \ln n + cn$ times of purchases is at most $\mathrm{e}^{-c}$.*

The complete proof is left as an exercise.
*Hints.* The method to prove the expectation has been mentioned before (in Lecture 4), and the following two lemmas are useful tools in the proof.

**Lemma 5** (Harmonic Number). *The limit of $H_n - \ln n$ exists, where*

$$H_n \triangleq 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$$

*is the* harmonic number.
*In fact, the limit is denoted by $\gamma$, i.e.,*

$$\gamma \triangleq \lim_{n\to\infty} H_n - \ln n = \lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{k} - \ln n\,.$$

*Then $\gamma \approx 0.577215665$ is called the* Euler's constant *(or* Euler-Mascheroni constant*).*

**Lemma 6** (Union Bound). *Let $A_1, A_2, \ldots, A_n$ are $n$ events. Then*

$$\mathbf{Pr}[A_1 \cup A_2 \cup \cdots \cup A_n] \le \sum_{i=1}^{n} \mathbf{Pr}[A_i]\,.$$