ADVANCED ALGORITHMS (IX)

CHIHAO ZHANG

Today we begin to introduce Markov chains, a powerful and widely used tool in the design of algorithms. Some of you might have met them in the probability class. In this course, I will develop the theory mainly based on the spectral methods we learnt before.

1. Basic Concepts

I assume that you know what a Markov chain is. We use Ω to denote the state space and in most cases today, Ω is simply $[n]$. We only consider the case that Ω is finite and the chain is *time-homogeneous*, which means the transition matrix is the same at each step. We let $P \in \mathbb{R}^{n\times n}$ be the transition matrix, namely $P(i,j)$ is the probability that one moves from state i to state j. The matrix P is a *stochastic matrix* in the sense that each entry of P is nonnegative, and the sum of entries in each row of P is one. Sometimes we just use P to denote the chain. For every $t \geq 1$, you can verify that P^t is the transition matrix of the chain in t-steps, namely $P^t(i,j)$ is the probability that one moves from state i to state j in t steps. Sometimes, it is convenient to talk about the transition graph $\mathcal{G}_P = (\Omega, \mathcal{E}_P)$ where an edge $(i,j) \in \mathcal{E}_P$ iff $P(i,j) > 0$. Moreover, an edge (i,j) is associated with a weight $P(i,j)$.

A distribution π on Ω is called a *stationary distribution* if $\pi^T P = \pi^T$ holds. Note that for any stochastic matrix P, a stationary distribution always exists (since the matrix $P - I$ is singular).

In this section, we use a two-state Markov chain as an example to introduce some important concepts. Let the transition matrix be $P = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix}$ where $p, q \in [0,1]$ are two reals.

It is not hard to verify that $\pi = \left(\frac{q}{p+q}, \frac{p}{p+q}\right)^T$ is a stationary distribution of P. We are interested in the following question: Does $\mu^T P^t$ converge to π^T when $t \to \infty$ for any distribution μ?

At least for our two-state chain P, the question is not hard to answer. Assume we start from any initial distribution μ and let $\mu_i = \mu^T P^i$ be the distribution after i steps. Define $\Delta_i \triangleq \left|\mu_i(0) - \pi(0)\right|$ as the distance between μ_i and π. Then μ_i converges to π if and only if Δ_i converges to zero. We can directly compute

$$\Delta_{i+1} = \left|\mu_{i+1}(0) - \frac{q}{p+q}\right| = \left|\mu_i(0)\cdot(1-p) + (1-\mu_i(0))\cdot q - \frac{q}{p+q}\right| = |1-p-q|\cdot\Delta_i.$$

Therefore, unless $\mu = \pi$, $\lim_{t\to\infty}\Delta_t = 0$ if and only if $|1-p-q| \neq 1$.

In fact, there are two cases that can make $|1-p-q| = 1$ happen: $p = q = 1$ or $p = q = 0$. These two cases prevent the chain from converging for different reasons.

(1) When $p = q = 0$, the chain is called *reducible*, which means the state space is disconnected. If we start at state 0 in the beginning, then we will always stay there (and the same holds for state 1).
(2) When $p = q = 1$, the chain is called *periodic*. In our example, we will alternate between state 0 and state 1.

We now formally define (the opposite of) these two concepts. We call a Markov chain *irreducible*, if for every $x, y \in \Omega$, $P^t(x,y) > 0$ for some t. For every $x \in \Omega$, let $C(x) = \left\{t \geq 1 \,:\, P^t(x,x) \geq 0\right\}$. Then we call a Markov chain *aperiodic* if for every $x \in \Omega$, $\gcd C(x) = 1$.

We remark that for an irreducible chain, when the transition matrix P satisfies $P(x,y) > 0 \iff P(y,x) > 0$, the aperiodic condition is equivalent to that the transition graph $\mathcal{G}_P$ is not bipartite.

The following proposition describes an important property of chains that are both irreducible and aperiodic. The proof is not hard and I leave it as an exercise.

Proposition 1. *If P is the transition matrix of an irreducible and aperiodic chain, then there exists some $t \geq 1$ such that $P^t(x,y) > 0$ holds for every $x, y \in \Omega$.*

In our two-state example, we know that both irreducibility and aperiodicity are necessary for the chain to converge to stationary distribution. We now prove that they are also sufficient.

Theorem 2. *If a finite time-homogeneous chain P is irreducible and aperiodic, then it has a unique stationary distribution π. Moreover, for any initial distribution μ, it holds that*

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

Theorem 2 can be proved in many ways. In fact, I will introduce three proofs in this course. The proofs are based on different tools that can be used to analyze Markov chains. We will also see some further development of these tools.

Today we prove theorem 2 by decomposing P into the sum a good matrix and a bad one. We then show that the bad one vanishes during the iteration.

Proof of Theorem 2. We assume the state space of the chain is $[n]$. Let $\Pi = \begin{bmatrix} \pi^T \\ \vdots \\ \pi^T \end{bmatrix}$ be the $n \times n$ matrix whose rows are π^T. Then $\mu^T \Pi = \pi^T$ holds for any distribution μ. The matrix Π is our *good* matrix. It follows from proposition 1 that for some $t \geq 1$, each entry of the matrix P^t is positive. Therefore, we can find some constant $\delta > 0$ such that each entry of $P^t - \delta\Pi$ is nonnegative. Let $Q = \frac{P^t - \delta\Pi}{1-\delta}$ (equivalently $P^t = \delta\Pi + (1-\delta)Q$), then Q is stochastic.

We now show by induction that $P^{tk} = (1-\theta^k)\Pi + \theta^k Q^k$, where $\theta = 1 - \delta$. The base case $k = 1$ is just our definition of Q. Assuming it is true for smaller k, then

$$P^{t(k+1)} = P^{tk}P^t = \left((1-\theta^k)\Pi + \theta^k Q^k\right) P^t = (1-\theta^k)\Pi + \theta^k Q^k \left((1-\theta)\Pi + \theta Q\right) = (1-\theta^{k+1})\Pi + \theta^{k+1}Q^{k+1},$$

where in the third equality, we used the fact that $\Pi P = \Pi$ and in the last equality, we used the fact that $Q\Pi = \Pi$.

Therefore, for every $0 \leq j < t$, we have

$$P^{tk+j} = (1-\theta^k)\Pi + \theta^k Q^k P^j.$$

The above tends to Π when $k \to \infty$ since $0 < \theta < 1$. □

2. Spectral Method

Consider a transition matrix P. We would like to apply spectral methods developed in previous lectures to analyze P. However, the spectral decomposition theorem requires the matrix to be symmetric in order to guarantee its eigenvalues to be real. This is not true in general for P, but we can still apply these tools for at least a large family of Markov chains.

Let $\pi \in \mathbb{R}^\Omega$ be a distribution over the space and $P \in \mathbb{R}^{\Omega\times\Omega}$ be a stochastic matrix. If the following *detailed balance condition*

$$\pi(x)P(x,y) = \pi(y)P(y,x), \tag{1}$$

holds for every $x, y \in \Omega$, then π is a stationary distribution of P. To see this, we note that

$$\pi^T P(x) = \sum_{y\in\Omega} \pi(y)P(y,x) = \sum_{y\in\Omega} \pi(x)P(x,y) = \pi(x).$$

It is worth to note that the detailed balance condition is only a sufficient condition for π to be a stationary distribution, but not a necessary one. We call those Markov chains whose stationary distribution satisfies eq. (1) the *(time) reversible* chains. The condition 1 implies that the transition matrix of a reversible chain is *symmetric* in the following sense:

Let π be the stationary distribution of a reversible Markov chain P on the space $[n]$. We define an inner product $\langle\cdot,\cdot\rangle_\pi$ such that for every $x, y \in \mathbb{R}^{[n]}$,

$$\langle x, y\rangle_\pi \triangleq \sum_{i=1}^n \pi(i)x(i)y(i).$$

If we let $D_\pi = \operatorname{diag}(\pi(1), \ldots, \pi(n))$ be the diagonal matrix whose i-th entry on the diagonal is $\pi(i)$, then the above inner product can be written as $\langle x, y\rangle_\pi = y^T D_\pi x$. In other words, we are working on a Hilbert space $\mathbb{R}^n$ endowed with the inner product $\langle\cdot,\cdot\rangle_\pi$, and P is considered to be symmetric here. We have the following spectral decomposition theorem:

Theorem 3. *Let $P \in \mathbb{R}^{n\times n}$ be reversible with respect to π, then the Hilbert space $(\mathbb{R}^n, \langle\cdot,\cdot\rangle_\pi)$ has an orthonormal basis $\{v_i\}_{i\in[n]}$ corresponding to real eigenvalues $\{\lambda_i\}_{i\in[n]}$.*

Proof. We prove the theorem by reducing it to the spectral decomposition theorem with respect to the ordinary inner product $\langle\cdot,\cdot\rangle$ (See notes of Lecture 5). Since P is reversible with respect to π, if we define a matrix $Q \triangleq D_\pi^{\frac{1}{2}} P D_\pi^{-\frac{1}{2}}$, then Q is symmetric. To see this, we have

$$Q(x,y) = \pi(x)^{\frac{1}{2}} P(x,y) \pi(y)^{-\frac{1}{2}} = \pi(x)^{-\frac{1}{2}} \pi(x) P(x,y) \pi(y)^{-\frac{1}{2}} = \pi(x)^{-\frac{1}{2}} \pi(y) P(y,x) \pi(y)^{-\frac{1}{2}} = Q(y,x).$$

Then we can apply spectral decomposition theorem on Q, and let $\{w_i\}_{i\in[n]}$ and $\{\mu_i\}_{i\in[n]}$ be corresponding orthonormal eigenvectors and eigenvalues. We have

$$Q = D_\pi^{\frac{1}{2}} P D_\pi^{-\frac{1}{2}} = \sum_{i=1}^n \mu_i w_i w_i^T.$$

The above implies

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

For every $i \in [n]$, we let $v_i = D_\pi^{-\frac{1}{2}} w_i$ and $\lambda_i = \mu_i$. We can verify that

$$\langle v_i, v_j\rangle_\pi = v_j^T D_\pi v_i = w_j D_\pi^{-\frac{1}{2}} D_\pi D_\pi^{-\frac{1}{2}} w_i = \langle w_i, w_j\rangle,$$

and therefore $\{v_i\}_{i\in[n]}$ is an orthonormal basis with respect to $\langle\cdot,\cdot\rangle_\pi$. Moreover, it holds that

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

and this implies that λ_i is the eigenvalue of v_i for every $i \in [n]$. □

Consider the random walk on a d-regular graph G with adjacency matrix A. The transition matrix of this Markov chain is exactly $\frac{A}{d}$. In general, the transition matrix P can be viewed as a weighted adjacency matrix of the transition graph $\mathcal{G}_P$. So the following properties of P is quite similar to those of normalized Laplacians that we are familiar with:

Proposition 4. *Let P be the transition matrix of a reversible Markov chain on $[n]$ with stationary distribution π. Let $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ be its eigenvalues. Then*

1. *$\lambda_n = 1$;*
2. *$\lambda_1 \ge -1$ and $\lambda_1 = -1$ if and only if one of components of $\mathcal{G}_P$ is bipartite;*
3. *$\lambda_{n-1} = 1$ if and only if P is reducible.*

You can compare the statement of above proposition with Proposition 6 of Lecture 5. We leave the proof of this proposition as an exercise. It is easy to see that we can take $v_n = \mathbf{1}$ since P is stochastic.

It is instructive to compute P^t using the spectral decomposition, which gives

$$P^t = \left(\sum_{i=1}^n \lambda_i v_i v_i^T D_\pi\right)^t = \sum_{i=1}^n \lambda_i^t v_i v_i^T D_\pi.$$

Therefore, it follows from proposition 4 and $v_n = \mathbf{1}$ that when P is irreducible ($\lambda_{n-1} < 1$) and aperiodic ($\lambda_1 > -1$), we have $\lim_{t\to\infty} P^t = \mathbf{1}\mathbf{1}^T D_\pi = \Pi$. This again justify theorem 2 for reversible chains.

3. Remark

More details on Markov chains can be found in the monograph [LP17].

References

[LP17] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017. 3