[AI2613 Lecture 2]: Concentration Inequalities, Discrete Markov Chain

March 12, 2022

In this lecture, we first introduce the balls-into-bins model, which is a common structure arising in probabilistic analysis. Then we look at the "concentration inequalities", namely a set of inequalities that provide bounds on how a random variable deviates from its expectation. Finally, we start our journey on finite Markov chains.

1 Balls-into-Bins

Balls-into-bins is a simple random process in which a person throws m balls into n bins uniformly at random. Many interesting questions can be asked about the process.

1.1 Birthday Paradox

Birthday paradox refers to the seemly counter-intuitive fact that some students in the class are very likely to share the same birthday. Viewing bins as dates and balls as students, the event that two students have the same birthday can be modeled as the event that some bin contains more than one ball.

Note that each ball is thrown independently. Condition on there is no collision after the $k-1$ balls are thrown, the probability that no collision occurs after throwing the k^{th} ball is $\frac{n-k+1}{n}$. Hence,

$$\begin{aligned}
\mathbf{Pr}\left[\text{no same birthday}\right] &= \prod_{k=1}^{m} \frac{n-k+1}{n} \\
&= \prod_{k=1}^{m-1} \left(1-\frac{k}{n}\right) \\
&\le \exp\left\{-\frac{\sum_{k=1}^{m-1} k}{n}\right\} \quad (\text{by } 1+x \le e^x) \\
&= \exp\left\{-\frac{m(m-1)}{2n}\right\}. \qquad (1)
\end{aligned}$$

For $m = O(\sqrt{n})$, the probability can be arbitrarily close to 0.

When n is sufficiently large, Equation (1) is quite tight because $\frac{k}{n} \le \frac{m}{n} = O(\frac{1}{\sqrt{n}}) \to 0$ and $1 + x \le e^x$ is quite tight when x is small.

1.2 Coupon Collector

The coupon collector problem asks the following question: If each box of a brand of cereals contains a coupon, randomly chosen from n different types of coupons, what is the number of boxes one needs to buy to collect all n

coupons? In the language of balls-into-bins, it asks how many balls one needs to throw until each of the n bins contains at least one ball.

The expectation can be easily calculated using the linearity of expectations. Let X_i be the number of balls to throw to get the i-th distinct type of coupon while exactly $i-1$ distinct types of coupons are already in had. Then the number of draws X to collect all coupons satisfies

$$X = \sum_{i=1}^{n-1} X_i.$$

By the linearity of expectations:

$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{E}[X_i].$$

It is clear that $X_i \sim \texttt{Gem}(\frac{n-i+1}{n})$ and therefore $\mathbf{E}[X_i] = \frac{n}{n-i+1}$. As a result,

$$\mathbf{E}[X] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n \cdot H(n),$$

where $H(n)$ is the harmonic number satisfying $\lim_{n\to\infty} H_(n) = \log n + \gamma$ for $\gamma = 0.577\ldots$.

γ is called the Euler constant.

2 Concentration Inequalities

In addition to the expectation, we are often interested in how a random variable deviates from certain fixed value. Concentration inequalities are inequalities of this form.

2.1 Markov's Inequality

Theorem 1 (Markov's Inequality) . *For any non-negative random variable X and $a > 0$,*

$$\mathbf{Pr}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}.$$

Proof. Since X is non-negative, we have

$$\mathbf{E}[X] \geq a \cdot \mathbf{Pr}[X \geq a] + 0 \cdot \mathbf{Pr}[X < a].$$

This is equivalent to

$$\mathbf{Pr}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}.$$

□

Example 1 (Concentration for Coupon Collector) . *Recall that X is the number of balls we need. Apply Markov's inequality, for $c > 0$ we have*

$$\mathbf{Pr}[X \geq c] \leq \frac{\mathbf{E}[X]}{c} = \frac{nH_n}{c}.$$

Thus, the probability that we need to draw the coupon for more than $100 \cdot nH_n$ times is less than 0.01.

2.2 Chebyshev's Inequality

A common trick to improve concentration is to consider $\mathbf{E}[f(X)]$ instead $\mathbf{E}[X]$ for some increasing function $f : \mathbb{R} \to \mathbb{R}$ since

$$\mathbf{Pr}[X \geq a] = \mathbf{Pr}[f(X) \geq f(a)].$$

Concentration inequalities give a sense that how the random variable deviate from its expectation. Then the probability we care about is actually $\mathbf{Pr}[|X - \mathbf{E}[X]| \geq a]$ for some postive constant a. Choosing the increasing function $f(x) = x^2$, we get the following Chebyshev's inequality.

Theorem 2 (Chebyshev's Inequality) *. For any random variable with bounded* $\mathbf{E}[X]$ *and* $a \geq 0$*, it holds that*

$$\mathbf{Pr}[|X - \mathbf{E}[X]| \geq a] \leq \frac{\mathbf{Var}[X]}{a^2}$$

Proof. Let $Y = |X - \mathbf{E}[X]|$, then clearly $Y \geq 0$. Therefore

$$\begin{aligned}\mathbf{Pr}[|X - \mathbf{E}[X]| \geq a] = \mathbf{Pr}[Y \geq a] = \mathbf{Pr}\left[Y^2 \geq a^2\right] &\leq \frac{\mathbf{E}\left[Y^2\right]}{a^2} \\ &= \frac{\mathbf{E}\left[(X - \mathbf{E}[X])^2\right]}{a^2} = \frac{\mathbf{Var}[X]}{a^2}.\end{aligned}$$

□

Example 2 (Coupon Collector Revisited) *We apply Chebyshev's inequality to the coupon collector problem. Assuming the notation before, we have*

$$\mathbf{Pr}[X \geq nH_n + t] \leq \mathbf{Pr}[|X - \mathbf{E}[X]| \geq t] \leq \frac{\mathbf{Var}[X]}{t^2}.$$

Recall that the variable X_i indicates the number of draws to get a new coupon while there are i coupons in hands. For distinct i and j, X_i and X_j are independent. Then

$$\mathbf{Var}[X] = \mathbf{Var}\left[\sum_{i=0}^{n-1} X_i\right] = \sum_{i=0}^{n-1} \mathbf{Var}[X_i].$$

For $i \in \{0, 1, \ldots, n-1\}$, $X_i \sim \mathtt{Geom}\left(\frac{n-i}{n}\right)$, so we have

$$\mathbf{Var}[X_i] = \frac{1 - \frac{n-i}{n}}{\left(\frac{n-i}{n}\right)^2} = \frac{i \cdot n}{(n-i)^2} \leq \frac{n^2}{(n-i)^2}.$$

It remains to bound $\sum_{i=0}^{n-1} \frac{1}{(n-i)^2} = \sum_{i=1}^{n} \frac{1}{i^2}$. Note that

$$\sum_{i=1}^{n} \frac{1}{i^2} \leq 1 + \int_1^{\infty} \frac{\mathrm{d}x}{x^2} = 2.$$

Therefore, we have $\mathbf{Var}[X] \leq 2n^2$ and $\mathbf{Pr}[X \geq nH_n + t] \leq \frac{2n^2}{t^2}$. The probability that we need to draw the coupon for more than $\sqrt{200}n + nH_n$ times is less than 0.01.

The bound obtained by Chebyshev's inequality is much tighter than that via Markov's inequality where in order to obtain the same confidence, one needs to choose $t = \Theta(n \log n)$.

2.3 Vanilla Chernoff Bound

If we apply Markov inequality to

$$\mathbf{Pr}\left[f(X) \geq f(t)\right]$$

with $f(x) = e^{\alpha x}$ where $\alpha > 0$, then the bound amounts to bound $\mathbf{E}\left[e^{\alpha X}\right]$ which is the *moment generating function* of X.

When the random variable X can be written as the sum of independent Bernoulli variables, its moment generating function is easy to estimate and we obtain sharp concentration bounds.

Theorem 3 (Chernoff Bound) . *Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \sim \mathrm{Ber}(p_i)$ for each $i = 1, 2, \ldots, n$. Let $X = \sum_{i=1}^n X_i$ and denote $\mu \triangleq \mathbf{E}[X] = \sum_{i=1}^n p_i$, we have*

$$\mathbf{Pr}\left[X \geq (1+\delta)\mu\right] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$$

If $0 < \delta < 1$, then we have

$$\mathbf{Pr}\left[X \leq (1-\delta)\mu\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}$$

Proof. We only prove the upper tail bound and the proof of lower tail bound is similar. For every $\alpha > 0$, we have

$$\mathbf{Pr}\left[X \geq (1+\delta)\mu\right] = \mathbf{Pr}\left[e^{\alpha X} \geq e^{\alpha(1+\delta)\mu}\right] \leq \frac{\mathbf{E}\left[e^{\alpha X}\right]}{e^{\alpha(1+\delta)\mu}}.$$

Therefore, we need to estimate the moment generating function $\mathbf{E}\left[e^{\alpha X}\right]$. Since $X = \sum_{i=1}^n X_i$ is the sum of independent Bernoulli variables, we have

$$\mathbf{E}\left[e^{\alpha X}\right] = \mathbf{E}\left[e^{\alpha \sum_{i=1}^n X_i}\right] = \mathbf{E}\left[\prod_{i=1}^n e^{\alpha X_i}\right] = \prod_{i=1}^n \mathbf{E}\left[e^{\alpha X_i}\right].$$

Since $X_i \sim \mathrm{Ber}(p_i)$, we can compute $\mathbf{E}\left[e^{\alpha X_i}\right]$ directly:

$$\mathbf{E}\left[e^{\alpha X_i}\right] = p_i e^{\alpha} + (1-p_i) = 1 + (e^{\alpha}-1)p_i \leq e^{((e^{\alpha}-1)p_i)}.$$

Therefore,

$$\mathbf{E}\left[e^{\alpha X}\right] \leq \prod_{i=1}^n e^{((e^{\alpha}-1)p_i)} = e^{\left((e^{\alpha}-1)\sum_{i=1}^n p_i\right)} = e^{((e^{\alpha}-1)\mu)}.$$

Therefore,

$$\mathbf{Pr}\left[X \leq (1+\delta)\mu\right] \leq \frac{\mathbf{E}\left[e^{\alpha x}\right]}{e^{\alpha(1+\delta)\mu}} \leq \left(\frac{e^{(e^{\alpha}-1)}}{e^{(\alpha(1+\delta))}}\right)^{\mu}$$

Note that above holds for any $\alpha > 0$. Therefore, we can choose α so as to minimize $\frac{e^{(e^{\alpha}-1)}}{e^{(\alpha(1+\delta))}}$. To this end, we let $\left(\frac{e^{(e^{\alpha}-1)}}{e^{(\alpha(1+\delta))}}\right)' = 0$. This gives $\alpha = \log(1+\delta)$. Therefore

$$\mathbf{Pr}\left[X \leq (1+\delta)\mu\right] \leq \left(\frac{e^{(e^{\alpha}-1)}}{e^{(\alpha(1+\delta))}}\right)^{\mu} = \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}.$$

□

The following form of Chernoff bound is more convenient to use (but weaker):

Corollary 4 *For any* $0 < \delta < 1$,

$$\mathbf{Pr}\left[X \geq (1+\delta)\mu\right] \leq \exp\left\{\left(-\frac{\delta^2}{3}\mu\right)\right\}$$
$$\mathbf{Pr}\left[X \leq (1-\delta)\mu\right] \leq \exp\left\{\left(-\frac{\delta^2}{2}\mu\right)\right\}$$

Proof. We only prove the upper tail. It suffices to verify that for $0 < \delta < 1$, we have

$$\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}} \leq \exp\left\{\left(-\frac{\delta^2}{3}\right)\right\}$$

Taking logarithm of both sides, this is equivalent to

$$\delta - (1+\delta)\ln(1+\delta) \leq -\frac{\delta^2}{3}$$

Let $f(\delta) = \delta - (1+\delta)\ln(1+\delta) + \frac{\delta^2}{3}$ and note that

$$f'(\delta) = -\ln(1+\delta) + \frac{2}{3}\delta, \quad f''(\delta) = -\frac{1}{1+\delta} + \frac{2}{3}.$$

Then for $0 < \delta < 1/2$, $f''(\delta) < 0$, and for $1/2 < \delta < 1$, $f''(\delta) > 0$. Therefore, $f'(\delta)$ first decreases and then increases in $[0,1]$. Also note that $f'(0) = 0$, $f'(1) < 0$ and $f'(\delta) \leq 0$ when $0 \leq \delta \leq 1$. Therefore $f(\delta) \leq f(0) = 0$. □

Example 3 (Tossing p-coins) *. Consider a p-coin that we get a head with probability p when tossing it. If we toss a p-coin n times, the average number of heads is pn.We want to determine the value δ such that with high probability (say 99%), the total number of heads is in the interval of $[(1-\delta)pn, (1+\delta)pn]$. We use Chernoff bound to determine δ.*

Let X denote the total number of heads, and $X_i \sim \mathrm{Ber}(p)$ be the indicator of whether the i-th toss gives a head. Then by Chernoff bound, we have

$$\mathbf{Pr}\left[|X - pn| \geq \delta \cdot pn\right] \leq 2\exp\left\{\left(-\frac{\delta^2}{3} \cdot pn\right)\right\} \leq 0.01$$

So if p is a constant, it suffices to choose

$$\delta = \Omega\left(\frac{1}{\sqrt{n}}\right).$$

3 Discrete Markov Chain

3.1 Markov Chain

Definition 5 (Discrete Markov Chain) *Suppose there is a sequence of random variables*

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

where the $\mathrm{Ran}(X_t) \subseteq \Omega$ *for some countable* Ω*. Then we call* $\{X_t\}$ *a discrete Marcov chain if* $\forall t \geq 1$ *the distribution of* X_t *is only related to* X_{t-1}*, that is* $\forall a_0, a_1, \ldots, a_t \in \Omega$*,*

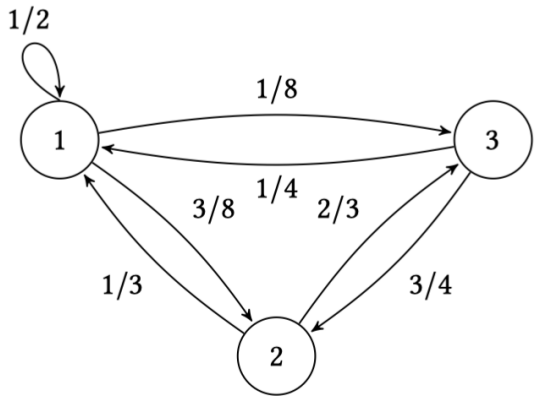
$$\mathbf{Pr}\left[X_t = a_t | X_{t-1} = a_{t-1}, \ldots, X_1 = a_1, X_0 = a_0\right] = \mathbf{Pr}\left[X_t = a_t | X_{t-1} = a_{t-1}\right].$$

Example 4 (Random Walk on $\mathbb{Z}$) *. Consider the random walk on* $\mathbb{Z}$*. One starts at* 0 *and in each round, he tosses a fair coin to determine the direction of moving: with probability 50% to the left and 50% to the right. If we use* X_t *to denote his position at time* t*, then we have* $X_0 = 0$ *and for every* $t > 0$*,* $X_t = X_{t-1} + 1$ *with probability 50% and* $X_t = X_{t-1} - 1$ *with probability 50%. This is a simple Markov chain, since the position at time* t *only depends on the position at time* $t - 1$*.*

In this lecture, we consider the situation that the state space $\Omega = [n]$ is finite. Then a (time-homogeneous) Markov chain can be characterized by a $n \times n$ matrix $P = \left(p_{ij}\right)_{i,j\in[n]}$ where $p_{ij} = \mathbf{Pr}\left[X_{t+1} = j \mid X_t = i\right]$ for all $t \geq 0$.

In general, a Markov chain can be equivalently viewed as a random walk on a weighted directed graph where the edge weight from i to j means the probability of moving to vertex j when one is standing at vertex i.

Example 5 (Finite State Random Walk) *The following three vertex directed graph corresponds to the Markov chain with transition matrix* $P = (p_{ij}) = \begin{bmatrix} 1/2 & 3/8 & 1/8 \\ 1/3 & 0 & 2/3 \\ 1/4 & 3/4 & 0 \end{bmatrix}$*. We sometimes call the graph the* transition graph *of* P*.*



At any time $t \geq 0$, we use μ_t to denote the distribution of X_t meaning

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

By the law of total probability, $\mu_{t+1}(j) = \sum_i \mu_t(i) \cdot p_{ij}$, we have $\mu_t^\intercal P = \mu_{t+1}^\intercal$. As a result, we have $\mu_t^\intercal = \mu_0^\intercal P^t$. This is a useful formula as we can compute

the distribution at any time given the initial distribution and the transition matrix.

Sometimes, we will simply denote the transition matrix P as the Markov chain for convenience.

3.2 Stationary Distribution

Definition 6 (Stationary Distribution) *. A distribution π is a stationary distribution of P if it remains unchanged in the Markov chain as time progresses, i.e.,*

$$\pi^\top P = \pi^\top.$$

One of the major algorithmic applications of Markov chains is the *Markov chain Monte Carlo (MCMC)* method. It is a general method for designing an algorithm to sample from a certain distribution π. The idea of MCMC is

- First design a Markov Chain of which the stationary distribution is the desired π;
- Simulate the chain from a certain initial distribution for a number of steps and output the state.

Therefore, we hope that the distribution μ_t is close to π when t is large enough.

Example 6 (Card Shuffling) *Consider a naive "top-to-random" card shuffle: Suppose we have n cards, every time 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. Thus, there are $n!$ possible permutations and $p_{ij} > 0$ only if the i^{th} permutation can come to the j^{th} through one step "top-to-random" shuffle.*

Performing the shuffle repeatedly is a Markov chain. It is not difficult to verify that the uniform distribution $\left(\frac{1}{n!}, \frac{1}{n!}, \ldots, \frac{1}{n!}\right)^T$ over all $n!$ permutations is a stationary distribution.

One of the main purposes of the course is to understand the MCMC method. Therefore, the following four basic questions regarding stationary distributions are important.

- Does each Markov chain have a stationary distribution?
- If a Markov chain has a stationary distribution, is it unique?
- If the chain has a unique stationary distribution, does μ_t always converge to it from any μ_0?
- If μ_t always converges to the stationary distribution, what is the rate of convergence?

4 Fundamental Theorem of Markov Chains

4.1 The Existence of Stationary Distribution

We will show that, for every finite Markov chain P, there exists some π such that $\pi^\top P = \pi^\top$. Observe that this is equivalent to "1 is an eigenvalue of $P^\top$ with a nonnegative eigenvector ($P^\top \pi = \pi$)".

We use the following lemma and theorem in linear algebra.

Lemma 7 *Every eigenvalue of nonnegative matrix P is no larger than the maximum row sum of P.*

Proof. Let λ be a eigenvalue of P and x is the corresponding eigenvector. We have

$$\|\lambda x\|_\infty = \|Px\|_\infty \le \|P\|_\infty \cdot \|x\|_\infty.$$

Note that $\|\lambda x\|_\infty = |\lambda| \|x\|_\infty$ and $\|x\|_\infty > 0$. Thus, we have $\lambda \le |\lambda| \le \|P\|_\infty$, that is λ is no larger than the maximum row sum of nonnegative matrix P. □

Theorem 8 (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.*

Let $A = (a_{ij})_{i\in[n], j\in[m]}$. We say A is nonnegative (resp. positive) if every $a_{ij} \ge 0$ (resp. > 0).

We will prove the Perron-Frobenius theorem in Section 4.3.

Since P is a stochastic matrix, we have

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

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

$$P^\top \pi = \pi,$$

which is equivalent to

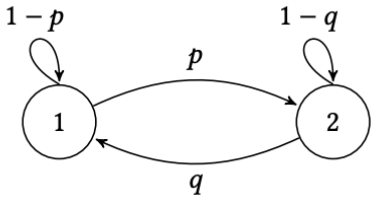
$$\pi^\top P = \pi^\top.$$

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

4.2 Uniqueness and Convergence

Consider the following Markov chain with two states. 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

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

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.,

$$\lim_{t\to\infty} \left\|\mu_0^{\mathsf{T}} P^t - \pi^{\mathsf{T}}\right\| = 0.$$

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.,

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

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 cases:

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

In fact, the two cases prevent convergence for different reasons.

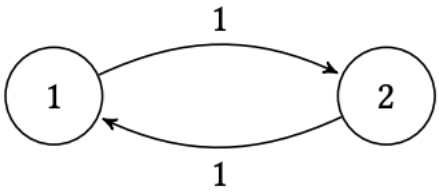
Let us first consider the case when $p = q = 0$. The Markov chain looks like: The transition graph is disconnected, so it can be partitioned into two disjoint components. Since each component is still a Markov chain, each of them has its own stationary distribution. Notice that any convex combination 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 motivates us to define the following:

Definition 9 *(Irreducibility). A finite Markov chain is* irreducible *if its transition graph is strongly connected.*

If the transition graph of P is not strongly connected, we say P is *reducible*.

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



is bipartite. It is easy to see that $(\frac{1}{2}, \frac{1}{2})$ is the unique stationary distribution of it. However, for $\mu_0 = (1, 0)$, one can see that μ_t ocsillates between "left" and "right". Therefore, the answer to the third question is no.

This phenomenon is captured by the following notion:

Definition 10 *(Aperiodicity). A Markov chain is* aperiodic *if for any state v, it holds that*

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

where C_v denotes the set of the directed cycles containing v in the transition graph.

Otherwise, we say the chain *periodic*.

We have the following important theorem.

Theorem 11 *(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 $\mu \in \mathbb{R}^n$,*

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

4.3 Proof of Perron-Frobenius Theorem

Most proofs in the section are from [Mey00]. We first prove the Perron-Frobenius theorem for positive matrices. Then we use this theorem and Lemma 13 to prove Theorem 8.

In the following statement, we use $|\cdot|$ to denote a matrix or vector of absolute values, i.e., $|A|$ is the matrix with entries $|a_{ij}|$. We say a vector or matrix is larger than $\mathbf{0}$ if all its entries are larger than 0 and denote it by $A > \mathbf{0}$. We define the operation $\geq$, $\leq$ and $<$ for vectors and matrices similarly.

Theorem 12 (Perron-Frobenius Theorem for Positive Matrices) *Each positive matrix $A > 0$ has a positive real eigenvalue $\rho(A)$, and $\rho(A)$ has a corresponding positive eigenvector.*

Proof. We first prove that $\rho(A) > 0$. If $\rho(A) = 0$, then all the eigenvalues of A is 0 which is equivalent to that A is nilpotent. This is impossible since every $a_{ij} > 0$. Thus $\rho(A) > 0$ for positive matrix A.

Assume that λ is the eigenvalue of A that $|\lambda| = \rho(A)$. Then we have

$$|\lambda||x| = |\lambda x| = |Ax| \leq |A||x| = A|x|.$$

Then we show that $|\lambda||x| < A|x|$ is impossible. Let $z = A|x|$ and $y = z - \rho(A)|x|$. Assume that $y \neq \mathbf{0}$, We have that $Ay > \mathbf{0}$. There must exist some $\epsilon > 0$ such that $Ay > \epsilon\rho(A) \cdot z$ or equivalently, $\frac{A}{(1+\epsilon)\rho(A)} z > z$. Successively multiply both sides of $\frac{A}{(1+\epsilon)\rho(A)} z > z$ by $\frac{A}{(1+\epsilon)\rho(A)}$ and we have

$$\left(\frac{A}{(1+\epsilon)\rho(A)}\right)^k z > \cdots > \frac{A}{(1+\epsilon)\rho(A)} z > z, \quad \text{for } k = 1, 2, \ldots.$$

Note that $\lim_{k\to\infty} \left(\frac{A}{(1+\epsilon)\rho(A)}\right)^k \to \mathbf{0}$ because $\rho\left(\frac{A}{(1+\epsilon)\rho(A)}\right) = \frac{\rho(A)}{(1+\epsilon)\rho(A)} < 1$. Then, in the limit, $z < \mathbf{0}$. This conflicts the fact that $z > \mathbf{0}$. The assumption that $y \neq \mathbf{0}$ is invalid

Thus we have $y = \mathbf{0}$ which means $\rho(A)$ is a positive eigenvalue of A and $|x|$ is the corresponding eigenvector. Since $\rho(A)|x| = A|x| > 0$, we have $|x| > 0$. □

Lemma 13 *For $A, B \in \mathbb{C}^{n\times n}$, if $|A| \leq B$, then $\rho(A) \leq \rho(B)$.*

Proof. By spectral radius formula, we have that for any sub-multiplicative norm $\|\cdot\|$, $\rho(A) = \lim_{k\to\infty} \left\|A^k\right\|^{\frac{1}{k}}$ and $\rho(B) = \lim_{k\to\infty} \left\|B^k\right\|^{\frac{1}{k}}$.

Note that since $|A| \leq B$, we have $|A|^k \leq B^k$ for $k \in \mathbb{N} \setminus \{0\}$. Then $\left\|A^k\right\|_\infty \leq \left\||A|^k\right\|_\infty \leq \left\|B^k\right\|_\infty$ and sequentially $\left\|A^k\right\|_\infty^{\frac{1}{k}} \leq \left\|B^k\right\|_\infty^{\frac{1}{k}}$. Thus, $\rho(A) \leq \rho(B)$. □

Theorem 14 *(Theorem 8 restated). Each nonnegative matrix A has a nonnegative real eigenvalue with spectral radius $\rho(A) = a$, and a has a corresponding nonnegative eigenvector.*

Proof. Construct a matrix sequence $\{A_k\}_{k=1}^\infty$ by letting $A_k = A + \frac{\mathbf{E}}{k}$ where $\mathbf{E}$ is the matrix of all 1's. Let $a_k = \rho(A_k) > 0$ and $x_k > \mathbf{0}$ is the corresponding eigenvector.[1] Without loss of generality, let $\|x_k\|_1 = 1$. Since $\{x_k\}_{k=1}^\infty$ is

[1] The existance of such x_k is guaranteed by Theorem 12.

bounded, by BolzanoWeierstrass theorem, there exists a subsequence of $\{x_k\}_{k=1}^{\infty}$ in $\mathbb{R}^n$ that is convergent. Denote this convergent subsequence by $\{x_{k_i}\}_{i=1}^{\infty}$ and $\{x_{k_i}\}_{i=1}^{\infty} \to z$ where $z \geq 0$ and $z \neq 0$ (for each x_{k_i} satisfies that $\|x_{k_i}\|_1 = 1$). Since $\{A_k\}_{k=1}^{\infty}$ is monotone decreasing, by Lemma 13, we have that $a_1 \geq \cdots \geq a_k \geq a$. Sequence $\{a_k\}_{k=1}^{\infty}$ is nonincreasing and bounded, so $\lim_{k\to\infty} a_k \to a^*$ exists and $\lim_{i\to\infty} a_{k_i} \to a^* \geq a$. Then we have

$$Az = \lim_{i\to\infty} A_{k_i} x_{k_i} = \lim_{i\to\infty} a_{k_i} x_{k_i} = a^* z.$$

Thus, a^* is an eigenvalue of A and $a^* \leq a$. Then we have $a^* = a$. So A has a nonnegative real eigenvalue a and z is the corresponding nonnegative eigenvetor. □

References

[Mey00] Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 71. SIAM, 2000. 10