[AI2613 Lecture 2] Discrete Markov Chains, Coupling

March 15, 2023

1 Discrete Markov Chain

1.1 Markov Chain

Definition 1 (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 Markov 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 1 (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 $t \geq 0$.

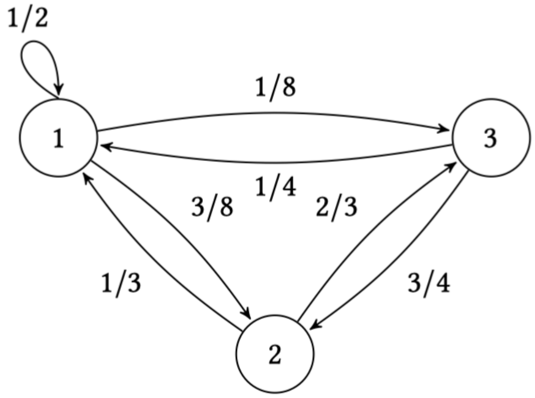
In general, a Markov chain can be equivalently viewed as a random walk on a weighted directed graph where the edge weight from i to j means the probability of moving to vertex j when one is standing at vertex i.

Example 2 (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 $\mu_t^{\mathsf{T}} P = \mu_{t+1}^{\mathsf{T}}$. As a result, we have $\mu_t^{\mathsf{T}} = \mu_0^{\mathsf{T}} P^t$. 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.

1.2 Stationary Distribution

Definition 2 (Stationary Distribution)**.** . *A distribution π is a stationary distribution of P if it remains unchanged in the Markov chain as time progresses, i.e.,*

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

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 3 (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 $\left(\frac{1}{n!}, \frac{1}{n!}, \dots, \frac{1}{n!}\right)^T$ 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?

2 Fundamental Theorem of Markov Chains

2.1 The Existence of Stationary Distribution

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

We use the following lemma and theorem in linear algebra.

Lemma 3. *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 4 (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 2.3.

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^{T} shares the same characteristic polynomial with P, which implies the eigenvalues of P^{T}

and P are the same. As a result, $\rho(P^{\mathsf{T}})$ also equals to 1. According to Perron-Frobenius theorem, there exists a nonnegative eigenvector π such that

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

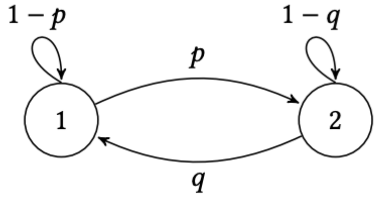
which is equivalent to

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

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

2.2 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

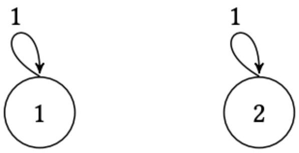
$$\begin{aligned}
\Delta_t &\triangleq \left|\mu_t(1) - \pi(1)\right| \\
&= \left|\mu_{t-1}^T \cdot P(1) - \pi(1)\right| \\
&= \left|(1-p)\cdot\mu_{t-1}(1) + q\cdot(1-\mu_{t-1}(1)) - \frac{q}{p+q}\right| \\
&= \left|(1-p-q)\cdot\mu_{t-1}(1) + q\cdot\left(1-\frac{1}{p+q}\right)\right| \\
&= |1-p-q|\cdot\Delta_{t-1}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

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.



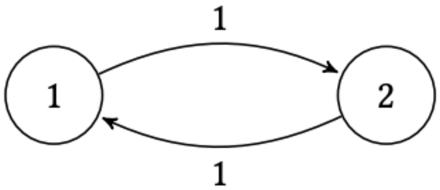
This observation motivates us to define the following:

Definition 5. *(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 6. *(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 call the chain *periodic*.

We have the following important theorem.

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

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

2.3 *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 9 to prove Theorem 4.

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 8 (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 9. *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 10. *(Theorem 4 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 bounded, by Bolzano–Weierstrass theorem, there exists a subsequence of $\{x_k\}_{k=1}^{\infty}$ in $\mathbb{R}^n$ that is convergent. Denote this convergent subsequence by $\left\{x_{k_i}\right\}_{i=1}^{\infty}$ and $\left\{x_{k_i}\right\}_{i=1}^{\infty} \to z$ where $z \geq 0$ and $z \neq 0$ (for each x_{k_i} satisfies that $\left\|x_{k_i}\right\|_1 = 1$). Since $\{A_k\}_{k=1}^{\infty}$ is monotone decreasing, by Lemma 9, 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. □

3 *Coupling*

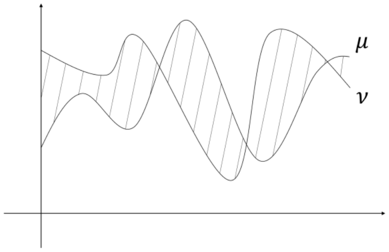
To measure how close the two distributions are, we need to define a distance between them.

[1] The existance of such x_k is guaranteed by Theorem 8.

Definition 11 (Total Variation Distance)**.** . *The total variation distance between two distributions μ and ν on a countable state space Ω is given by*

$$D_{\mathrm{TV}}(\mu,\nu)=\frac{1}{2}\sum_{x\in\Omega}\left|\mu(x)-\nu(x)\right|.$$

We can look at the following figure of two distributions on the sample space. The total variation distance is half the area enclosed by the two curves.



This figure gives us the intuition of the following proposition which states that the total variation distance can be equivalently viewed in another way.

Proposition 12. *We define $\mu(A)=\sum\limits_{x\in A}\mu(x)$, $\nu(A)=\sum_{x\in A}\nu(x)$, then we have*

$$D_{\mathrm{TV}}(\mu,\nu)=\max_{A\subseteq\Omega}\left|\mu(A)-\nu(A)\right|.$$

Our main tool to bound the distance between two distributions is the *coupling*. This is a useful technique in analysis of probabilities. A coupling of two distributions is simply a joint distribution of them.

Definition 13 (Coupling)**.** . *Let μ and ν be two distributions on the same space Ω. Let ω be a distribution on the space $\Omega\times\Omega$. If $(X,Y)\sim\omega$ satisfies $X\sim\mu$ and $Y\sim\nu$, then ω is called a coupling of μ and ν.*

We now give a toy example about how to construct different couplings on two fixed distributions. There are two coins: the first coin has probability $\frac{1}{2}$ for head in a toss and $\frac{1}{2}$ for tail, and the second coin has probability $\frac{1}{3}$ and $\frac{2}{3}$ respectively. We now construct two couplings as follows.

The table defines a joint distribition and the sum of a certain row/column equal to the corresponding marginal probability. It is clear that both table are couplings of the two coins. Among all the possible couplings, sometimes we are interested in the one who is "mostly coupled".

In other words, the marginal probabilities of the disjoint distribution ω are μ and ν respectively. A special case is when x and y are independently. However, in many applications, we want x and y to be correlated while keeping their respect marginal probabilities correct.

prob y x	HEAD	TAIL
HEAD	1/3	1/6
TAIL	0	1/2

prob y x	HEAD	TAIL
HEAD	1/6	1/3
TAIL	1/6	1/3

Lemma 14 (Coupling Lemma)**.** . *Let μ and ν be two distributions on a sample space Ω. Then for any coupling ω of μ and ν it holds that,*

$$\mathbf{Pr}_{(X,Y)\sim\omega}[X \neq Y] \geq D_{\mathrm{TV}}(\mu,\nu).$$

And furthermore, there exists a coupling ω^* of μ and ν such that

$$\mathbf{Pr}_{(X,Y)\sim\omega^*}[X \neq Y] = D_{\mathrm{TV}}(\mu,\nu).$$

Proof. For finite Ω, designing a coupling is equivalent to filling a $\Omega \times \Omega$ matrix in the way that the marginals are correct.

Clearly we have

$$\begin{aligned}\mathbf{Pr}[X = Y] &= \sum_{t\in\Omega} \mathbf{Pr}[X = Y = t] \\ &\leq \sum_{t\in\Omega} \min\{\mu(t), \nu(t)\}.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbf{Pr}[X \neq Y] &\geq 1 - \sum_{t\in\Omega} \min(\mu(t), \nu(t)) \\ &= \sum_{t\in\Omega} (\mu(t) - \min\{\mu(t), \nu(t)\}) \\ &= \max_{A\subseteq\Omega} \{\mu(A) - \nu(A)\} \\ &= D_{\mathrm{TV}}(\mu,\nu).\end{aligned}$$

To construct ω^* achieving the equality, for every $t \in \Omega$, we let $\mathbf{Pr}_{(X,Y)\sim\omega^*}[X = Y = t] = \min\{\mu(t), \nu(t)\}$. □

The coupling lemma provides a way to upper bound the distance between two distributions: For any two distributions μ and ν and any coupling ω of μ and ν, an upper bound for $\mathbf{Pr}_{(X,Y)\sim\omega}[X \neq Y]$ is an upper bound for $D_{TV}(\mu,\nu)$. This is a quite useful approach to bound the total variation distance.

References

[Mey00] Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 71. SIAM, 2000. 6