MARKOV CHAIN AND MIXING TIME NOTES

CHENDONG SONG

1. Introduction and basic technique

Definition 1.1. A sequence of random variables $(X_0, X_1, ...)$ is a Markov chain with state space Ω and transition matrix P satisfying the Markov property, i.e. if for all $x, y \in \Omega$, all $t > 1$, and all events $H_{t-1} = \bigcap_{s=0}^{t-1} \{X_s = x_s\}$ satisfying $\mathbf{P}(H_{t-1} \cap \{X_t = x\}) > 0$ we have

$$\mathbb{P}\{X_{t+1} = y | H_{t-1} \cap \{X_t = x\}\} = \mathbb{P}\{X_{t+1} = y | X_t = x\} = P(x, y)$$

A chain P is called **irreducible** if for any two states $x, y \in \Omega$, there exists an integer t (depending on x, y) such that $P^t(x, y) > 0$. Let $\mathcal{T}(x) := \{t \geq 1 : P^t(x, x) > 0\}$ be the set of times when it is possible for the chain to return to starting position x. The **period of state x** is defined to be the great common divisor of $\mathcal{T}(x)$. The **periodic of chain** is the common period of all states. The chain is **aperiodic** if all states have period 1. If a chain is not aperiodic, we call it **periodic**.

If a probability distribution π satisfy $\pi = \pi P$, i.e. $\sum_{x \in \Omega} P(x, y)\pi(x) = \pi(y)$, then we call π satisfying a **stationary distribution**.

Theorem 1.2. *Let P be a transition matrix of an irreducible Markov chain, then there exists a unique probability distribution π.*

Example 1.3 (Random Walk on $\mathbb{Z}$)**.** Let Ω be $\mathbb{Z}$ with transition matrix $P(k, k-1) = P(k, k+1) = 1/2$. That is, a random walker starting from some point, each move to left or right with probability 1/2.

Proposition 1.4. *Let (X_t) be simple random walk on $\mathbb{Z}$, and recall that*

$$\tau_0 = \min\{t \geq 0 : X_t = 0\}$$

is the first time the walk hits zero. Then

$$\mathbb{P}_k\{\tau_0 > r\} \leq \frac{6k}{\sqrt{r}}$$

for any integers $k, r > 0$.

Proof. Step 1(Reflection principle)

$$\mathbb{P}_k[\tau_0 \leq r, X_r = j] = \mathbb{P}_k[X_r = -j].$$

Step 2

$$\mathbb{P}_k\{\tau_0 > r\} = \mathbb{P}_0\{-k \leq X_r \leq k\}$$

Proof:

$$\mathbf{P}_k\{X_r > 0\} = \mathbf{P}_k\{X_r > 0, \tau_0 \leq r\} + \mathbf{P}_k\{\tau_0 > r\}$$

By Step 1,

$$\mathbf{P}_k\{X_r > 0\} = \mathbf{P}_k\{X_r < 0\} + \mathbf{P}_k\{\tau_0 > r\}.$$

By symmetry of the walk, $\mathbf{P}_k\{X_r < 0\} = \mathbf{P}_k\{X_r > 2k\}$, and so

$$\begin{aligned}\mathbf{P}_k\{\tau_0 > r\} &= \mathbf{P}_k\{X_r > 0\} - \mathbf{P}_k\{X_r > 2k\} \\ &= \mathbf{P}_k\{0 < X_r \leq 2k\} = \mathbf{P}_0\{-k < X_r \leq k\}.\end{aligned}$$

Step 3

$$\mathbb{P}_0\{X_t = k\} \leq \frac{3}{\sqrt{t}}$$

.

Proof: Using Stiring formula

$$\mathbb{P}_0\{X_{2r} = 2k\} \leq \begin{pmatrix} 2r \\ r \end{pmatrix} 2^{-2r} = \frac{(2r)!}{(r!)^2 2^{2r}} \leq \frac{2}{\sqrt{\pi}}\frac{1}{\sqrt{r}}$$

□

Example 1.5 (coupon collection)**.** Consider a collector attempting to collect a complete set of coupons. Assume that each new coupon is chosen uniformly and independently from the set of n possible types, and let τ be the random number of coupons collected when the set first contains every type.

Proposition 1.6. $\mathbf{E}(\tau) = n\sum_{k=1}^{n}\frac{1}{k} \sim n\log n$. *For any* $c > 0$. $\mathbf{P}\{\tau > n\log n + cn\} \leq e^{-c}$.

Proof. Let τ_k be the total number of coupons accumulated when the collection first contains k coupons. Note that $\tau_k - \tau_{k-1}$ has geometric distribution $G(\frac{n-k+1}{n})$, so its expectation is $\frac{n}{n-k+1}$. Therefore,

$$\mathbf{E}(\tau) = \sum_{k=1}^{n}\mathbf{E}(\tau_k - \tau_{k-1}) = n\sum_{k=1}^{n}\frac{1}{n-k+1} = n\sum_{k=1}^{n}\frac{1}{k}.$$

Also let A_i be the event that i-th coupon does not appear. Observe that

$$\mathbf{P}\{\tau > n\log n + cn\} \leq \sum_{i=1}^{n}\mathbf{P}(A_i) \leq n(1-\frac{1}{n})^{n\log n + cn} \leq e^{-c}.$$

□

From the theorem 1.2, we know that a Markov chain will finally converge to a stable distribution. Furthermore, we are also interested in quantifying the speed of convergence. So we introduce an appropriate metric for measuring the distance between distributions.

Definition 1.7. The **total variation distance** between to probability distribution μ and ν on Ω is defined by

$$||\mu - \nu||_{TV} = \max_{A\subset\Omega}|\mu(A) - \nu(A)| = \frac{1}{2}\sum_{x\in\Omega}|\mu(x) - \nu(x)|$$

For a Markov chain (X_t) starting from x, we are interested in the distance between the distribution at time t and the stationary distribution π, that is $||P^t(x,\cdot) - \pi||_{TV}$. And we consider the worst case to be

$$d(t) := \max_{x\in\Omega}||P^t(x,\cdot) - \pi||_{TV}. \tag{1}$$

Since every Markov chain will eventually converge to its stationary distribution, d(t) will ultimately tend to 0 as t tends to infinity. And we're also interested in the time required to

make the distance small. Here we introduce the *mixing time*

$$t_{mix}(\epsilon) := \min\{t : d(t) \leq \epsilon\}. \tag{2}$$

Another distance is

$$\bar{d}(t) := \max_{x,y\in\Omega} ||P^t(x,\cdot) - P^t(y,\cdot)||_{TV} \tag{3}$$

This two distance have the relationship $d(t) \leq \bar{d}(t) \leq 2d(t)$.

To bound the distance, here we introduce the coupling method. A coupling (X, Y) is a probability distribution defined on the product space $\Omega \times \Omega$ with marginal distribution μ and ν. That is, $\mathbb{P}(X = x) = \mu(x)$ and $\mathbb{P}(Y = y) = \nu(y)$.

Proposition 1.8. *Let μ and ν be two probability distribution on Ω. Then*

$$||\mu - \nu|| = \inf\{\mathbb{P}(X \neq Y) : (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\} \tag{4}$$

Proof. On the one hand, for a coupling (X, Y),

$$\mu(A) - \nu(A) = \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) \leq \mathbb{P}(X \in A, Y \notin A) \leq P(X \neq Y).$$

Therefore, it immediately follows that

$$||\mu - \nu|| \leq \inf\{\mathbb{P}(X \neq Y) : (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}$$

On the other hand, we should construct a coupling satisfy (4). To meet this requirement, there should be some correlation between X and Y. Here we firstly consider

$$p = \sum_{x\in\Omega} \mu(x) \wedge \nu(x) = 1 - ||\mu - \nu||_{TV}$$

and flip a coin with probability of heads equal to p. And

- If the coin comes up head, then choose a random value Z according to the probability distribution
$$\gamma(x) = \frac{\mu(x) \wedge \nu(x)}{p}$$
and set $X = Y = Z$.
- If the coin comes up tail, then choose X and Y independently with probability distribution
$$\gamma_X(x) = \frac{\mu(x) - \nu(x)}{||\mu - \nu||_{TV}} \mathbb{1}_{\{\mu(x)>\nu(x)\}}$$
$$\gamma_Y(x) = \frac{\nu(x) - \mu(x)}{||\mu - \nu||_{TV}} \mathbb{1}_{\{\nu(x)>\mu(x)\}}$$
Clearly X and Y follow the distribution μ and ν. Also note that $X \neq Y$ if and only if the coin toss is tail. So in this case
$$\mathbb{P}(X \neq Y) = 1 - p = ||\mu - \nu||_{TV}$$

□

Coupling is a very useful method to bound the total variance. We define **coupling of Markov chains** to be two process (X_t, Y_t) with the same transition matrix P, although they may have different starting distributions. Moreover, two chains will stay together "Tietie" at all the time after their first visit. More precisely, if $X_s = Y_s$, then $X_t = Y_t$ for all $t \geq s$.

Let (X_t, Y_t) be a coupling for which $X_0 = x, Y_0 = y$. Let τ_{couple} be the first time the chains meet, which says $\tau_{couple} = \min\{t : X_t = Y_t\}$

then

$$||P^t(x,\cdot) - P^t(y,\cdot)||_{TV} \leq \mathbb{P}\{\tau_{couple} > t\}.$$

So

$$d(t) \leq \mathbb{P}\{\tau_{couple} > t\}.$$

2. CUTOFF PHENOMENON

A finite Markov chain is said to exhibit cutoff if its distance from the stationary measure drops abruptly, over a negligible time period known as the cutoff window, from near its maximum to near 0.

Definition 2.1. A Markov chain has a cutoff if for all $\epsilon > 0$

$$\lim_{n\to\infty} \frac{t_{mix}^{(n)}(1-\epsilon)}{t_{mix}^{(n)}(\epsilon)} = 1. \tag{5}$$

The rate of convergence in (6) is addressed by the following: A sequence $w_n = o\left(t_{mix}^{(n)}\left(\frac{1}{4}\right)\right)$ is called a cutoff window for the family of chains X_n if for any $\epsilon > 0$ there exists some $c_\epsilon > 0$ such that for all n,

$$t_{mix}^{(n)}(\epsilon) - t_{mix}^{(n)}(1-\epsilon) \leq c_\epsilon w_n.$$

We say a family of chains has a pre-cutoff if it satisfies the weaker condition

$$\sup_{0<\epsilon<1/2} \limsup_{n\to\infty} \frac{t_{mix}^{(n)}(1-\epsilon)}{t_{mix}^{(n)}(\epsilon)} < \infty. \tag{6}$$

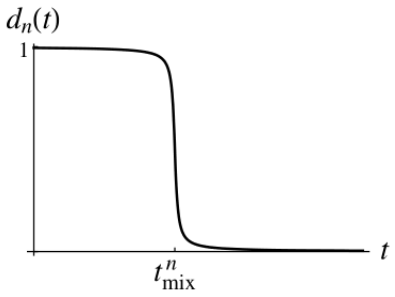

FIGURE 1. Cutoff phenomenon

Clearly if a Markov chain has cutoff, it must have pre-cutoff. It's still an open question to determine which kinds of Markov chains have cutoff. However, there're a plenty of studies focusing on the cutoff phenomenon of some specific families of chains. The following are some models with cutoff.

Example 2.2. *(Biased random walk on a line segment)*[1] Let $p \in (1/2, 1)$ and $q = 1-p$ and bias $\beta = (p-q)/2 > 0$. Consider the lazy nearest-neighbor random walk with bias β on the interval $\Omega = \{0, 1, \cdots, n\}$. When at an interior vertex, the walk remains in its current position with probability 1/2, moves to the right with probability p/2, and moves to the left with probability q/2. When at an end-vertex, the walk remains in place with probability 1/2 and moves to the adjacent interior vertex with probability 1/2.

This lazy random walk with bias β has a cutoff at $\beta^{-1}n$ with a window size of order $\sqrt{n}$.

Example 2.3. *(Random walk on hypercube)*[1] The *n-dimensional hypercube* is a graph whose vertices are the binary $n-$tuples $\{0,1\}^n$. Two vertices are connected by an edge when they differ in exactly one coordinate. The *lazy random walk on hypercube* moves from a vertex $(x^1, x^2, \cdots, x^n)$ by randomly choosing a coordinate $j \in \{1, 2, \cdots, n\}$ uniformly at random and setting the new state equal to $(x^1, \cdots, x^{j-1}, 1-x^j, x^{j+1}, \cdots, x^n$ with probability $1/2$ and remains the same with probability $1/2$.

The lazy random walk on hypercube has a cutoff at $1/2n \log n$ with a window size of n.

Proposition 2.4 ($n \log n$ bound)**.** *In lazy random walk on hypercube, the total variance have the upper bound*

$$d(n \log n + cn) \leq \mathbb{P}\{\tau > n \log n + cn\} \leq e^{-c} \tag{7}$$

Proof. Consider two random walk (X_t, Y_t) starting at different states. We generate an algorithm to couple the two random walk: First we pick among n coordinates uniformly at random, and then we refresh the bit at this coordinate with the same random fair bit. And from this time on, the chosen coordinate of two chains always remain the same. Individually the two chains move on as a lazy random walk, but their same coordinates accumulate along the time. This is a typical model of coupon collection. It has average mixing time $n \log n$ and can decrease fast when $t > n \log n$. Immediately we can deduce the mixing time

$$t_{mix}(\epsilon) \leq n \log n + \log(1/\epsilon)n. \tag{8}$$

□

Proposition 2.5 ($1/2n \log n$ bound)**.** *In lazy random walk on hypercube, the mixing time have the upper bound*

$$t_{mix}(\epsilon) \leq 1/2n \log n + cn. \tag{9}$$

Proof. Let $X_t = \left(X_t^1, \ldots, X_t^n\right)$ be the position of the random walk at time t, and let $W_t = W(X_t) = \sum_{i=1}^n X_t^i$ be the *Hamming weight* of X_t. Then (W_t) is a lazy version of the Ehrenfest urn chain whose transition matrix is

$$P(j,k) = \begin{cases} \frac{n-j}{n} & \text{if } k = j+1, \\ \frac{j}{n} & \text{if } k = j-1 \\ 0 & \text{otherwise.} \end{cases}. \tag{10}$$

We write π_W for the stationary distribution of (W_t), which is $B(n, 1/2)$

The study of (X_t) can be reduced to the study of (W_t) because of the following identity:

$$\left\|\mathbb{P}_{\mathbf{1}}\left\{\boldsymbol{X}_t \in \cdot\right\} - \pi\right\|_{\mathrm{TV}} = \left\|\mathbb{P}_n\left\{W_t \in \cdot\right\} - \pi_W\right\|_{\mathrm{TV}}.$$

By symmetry,

$$d(t) = \left\|\mathbf{P}_{\mathbf{1}}\left\{\boldsymbol{X}_t \in \cdot\right\} - \pi\right\|_{\mathrm{TV}}$$

We now construct a coupling (W_t, Z_t) of the lazy Ehrenfest chain started from w with the lazy Ehrenfest chain started from z. Provided that the two particles have not yet collided, at each move, a fair coin is tossed to determine which of the two particles moves; the chosen particle makes a transition according to the transition matrix (10), while the other particle remains in its current position. The particles move together once they have met for the first time.

Suppose, without loss of generality, that $z \geq w$. Since the particles never cross each other, $Z_t \geq W_t$ for all t. Consequently, if $D_t = |Z_t - W_t|$, then $D_t = Z_t - W_t \geq 0$. Let $\tau :=$

$\min\{t \geq 0 : Z_t = W_t\}$. Conditioning that $(Z_t, W_t) = (z_t, w_t)$, where $z_t \neq w_t$

$$D_{t+1} - D_t = \begin{cases} 1 & \text{with probability } (1/2)(1 - z_t/n) + (1/2)w_t/n, \\ -1 & \text{with probability } (1/2)z_t/n + (1/2)(1 - w_t/n). \end{cases} \tag{11}$$

From (11) we see that on the event $\{t < \tau\}$,

$$\mathbb{E}_{z,w}[D_{t+1} - D_t \mid Z_t = z_t, W_t = w_t] = -\frac{(z_t - w_t)}{n} = -\frac{D_t}{n}$$

Therefore,

$$\mathbb{E}[D_{t+1}] \leq (1 - \frac{1}{n})\mathbb{E}[D_t]$$

By induction,

$$\mathbb{E}_{z,w}[D_t] \leq \left(1 - \frac{1}{n}\right)^t (z - w) \leq ne^{-t/n}$$

Also the process D_t is likely to move downwards as it is to move upwards. Thus, the process D_t can be coupled with a simple random walk S_t so that $S_0 = D_0$ and $D_t \leq S_t$. Therefore, by proposition 1.4,

$$\mathbb{P}_k[\tau > u] \leq \frac{6k}{\sqrt{u}}. \tag{12}$$

Therefore,

$$\mathbb{P}_{z,w}[\tau > s + u | D_s] = \mathbb{P}_{D_s}[\tau > u] \leq \frac{6D_s}{\sqrt{u}}.$$

Taking expectations,

$$\mathbb{P}_{z,w}[\tau > s + u] \leq \frac{6ne^{-s/n}}{\sqrt{u}}.$$

Take $s = 1/2n \log n + \alpha n$ and $u = \alpha n$ we have

$$d(1/2n \log n + \alpha n) \leq \frac{6}{\sqrt{\alpha}}.$$

□

Example 2.6. *(Stratified random walk on the hypercube from* [2]*)* Let $\Omega = \{0,1\}^n \backslash \{0\}$ and consider the Markov chain X_t defined as follows: if the current state is x and if $x(i)$ denotes the bit at the $i-$th coordinate of x , then the walk proceeds by choosing uniformly at random an ordered pair (i, j) of distinct coordinates and replace $x(j)$ by $x(j) + x(i)$ (mod 2).

The chain X_t has a cut-off at time $\frac{3}{2}n \log n$ with window time n.

Example 2.7. *(Cyclic dynamics on hypercube from* [3]*)* Consider a hypercube on 3 states: $\Omega = Q^{V_n}$, where $Q = \{1, 2, 3\}$ and $V_n = \{1, \ldots, n\}$. At each time step, the vertex $v \in V_n$ is uniformly chosen. Assume that v has color i in Q. Then, we reassign the color of vertex v as i with probability $1 - p$ and $i + 1$ with probability p, where $0 < p < 1$.

The cyclic dynamics defined on Σ_n with probability $0 < p < 1$ exhibit cutoff at mixing time

$$t(n) = \frac{1}{3p} n \log n$$

with a window of size $O(n)$.

My work is to find some Markov chains that have pre-cutoff but don't have cutoff. In [4], Hubert gives a product chain with pre-cutoff but without cutoff.

References

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

[2] Ben-Hamou, Anna, and Yuval Peres. "Cutoff for a stratified random walk on the hypercube." (2018): 1-10.

[3] Lim, Keunwoo. "Cutoff Phenomenon for Cyclic Dynamics on Hypercube." arXiv preprint arXiv:2010.01756 (2020).

[4] Lacoin, Hubert. "A product chain without cutoff." (2015): 1-9.