[AI2613 Lecture 3] Proof of FTMC, Mixing Time

March 15, 2023

1 *Fundamental Theorem of Markov Chains*

Recall the fundamental theorem of Markov chains for *finite* chains we introduced in the last lecture.

Theorem 1 (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 $\mu \in \mathbb{R}^n$,*

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

Today we give a proof of the theorem. To this end, we first study the properties of the transition matrix P of an irreducible and aperiodic chain. Then we introduce the notion of *coupling*, a powerful technique to analyze stochastic processes.

Claim 2. *Let $P \in \mathbb{R}^{n\times n}$ be an irreducible and aperiodic Markov chain. It holds that*

$$\exists\, t^* : \forall\, i, j \in [n] : \quad P^{t^*}(i,j) > 0\,.$$

We use Lemma 3 to prove Claim 2.

Lemma 3. *Let $c_1, c_2, \ldots, c_s$ be a group of positive integers satisfying $\gcd(c_1,\ldots,c_s) = 1$. For any sufficiently large integer b, there exists $y_1, y_2, \ldots, y_s \in \mathbb{N}$ such that*

$$c_1 y_1 + c_2 y_2 + \cdots c_s y_s = b.$$

Proof. By Bézout's identity there exists $x_1, x_2, \ldots, x_s \in \mathbb{Z}$ such that

$$c_1 x_1 + c_2 x_2 + \cdots c_s x_s = 1\,.$$

We apply induction on s. The case $s = 1$ trivially holds. Assume $s \geq 2$ and the lemma holds for smaller s. Let $g = \gcd(c_1,\ldots,c_{s-1})$. By induction hypothesis, we know that

$$\frac{a_1}{g}\cdot x_1 + \frac{a_2}{g}\cdot x_2 + \cdots + \frac{a_{s-1}}{g}\cdot x_{s-1} = b' \iff a_1\cdot x_1 + a_2\cdot x_2 + \cdots + a_{s-1}x_{s-1} = g\cdot b'$$

has non-negative solutions for sufficiently large b'. Therefore, we only need to prove that the equation

$$g\cdot b' + a_s \cdot x_s = b \tag{1}$$

has nonegative solution (b', x_s) with sufficiently large b' when b is sufficiently large. In other words, we need to prove for any $b_0 > 0$, eq. (1) has nonegative solution with $b' > b_0$ for any sufficiently large b.

That is, there exists some $b_0 > 0$ such that for any $b > b_0$, the diophantine equation $c_1 y_1 + c_2 y_2 + \cdots + c_s y_s = b$ always has non-negative solutions

Annotator:

本节的任务是证明 FTMC，工具是 Coupling.

回忆：FTMC 是三推一，("一"包含了存在唯一和收敛)

这里先证一个引理：$\exists t^*$，走 t^* 步后从 $\forall i$ 到 $\forall j$ 概率都 >0，

条件是 IR + AP：

IR → irreducible　　AP → Aperiodic.

证明思路为：

step 1: IR $\Rightarrow \forall i,j, \exists t^*, P(i,j)>0$

(这是IR的自然定义)

step 2: IR+AP $\Rightarrow \exists t^*, \forall i,j, P_{ij}>0$.

这里的直观理解为：AP 使得 $t>t'$ 时 $\forall$ 点可以通过其自环，凑出 $\forall$ 个整数步从 i 出发回到 i (t' 为该点自环性质所能凑)，因此，需要 t^* 步从 i 到 j，只需消磨 $t'-L(i,j)$ 步在自环中，再走 $L(i,j)$ 步从 i 到 j 即为可行方案，故 t^* 时 $P_{ij}>0$，由有限性，可以在所有点的 t' 已知后取一个 $t^* \geq \max\{t'_k\}$，即可满足定理.

Note that $\gcd(g, a_s) = 1$, we can find integers (y, x) such that

$$g \cdot y + a_s \cdot x = 1 \iff g \cdot (by) + a_s \cdot (bx) = b.$$

Noting that for any $k \in \mathbb{Z}_{\geq 0}$, we have $g \cdot (by + ka_s) + a_s \cdot (bx - kg) = b$. We need $by + ka_s > b_0$ and $bx - kg \geq 0$, which are equivalent to

$$\frac{bx}{g} \geq k > \frac{b_0 - by}{a_s}.$$

We can always find such an integer k if $b \geq g(b_0 + a_s)$.

□

Proof of Claim 2. The property of irreducibility implies that

$$\forall\, i, j : \exists\, t : \quad P^t(i, j) > 0\,.$$

Suppose that there are s loops of length $c_1, c_2, \ldots, c_s$ starting from and ending at state i. Then by aperiodicity we have

$$\gcd(c_1, c_2, \ldots, c_s) = 1\,.$$

For any sufficiently large m and any pair of states (i, j), by Lemma 3 and irreducibility, there exists a path from i to j with exactly m steps. Thus, there exist $t^* > 0$ such that for any state pair (i, j), $P^{t^*}(i, j) > 0$. Furthermore, for any $t > t^*$, $P^t(i, j) > 0$ for any $i, j \in \Omega$.

□

1.1 Proof of FTMC

Proof. We already know that P has a stationary distribution π. What we would like to show is that for all starting distribution μ_0, it holds that

$$\lim_{t \to \infty} D_{\mathrm{TV}}(\mu_t, \pi) = 0\,,$$

where $\mu_t^{\mathrm{T}} = \mu_0^{\mathrm{T}} P^t$.

Suppose that $\{X_t\}$ and $\{Y_t\}$ are two identical Markov chains starting from different distribution, where $Y_0 \sim \pi$ while X_0 is generated from an arbitrary distribution μ_0.

Now we have two sequence of random variables:

$$\begin{array}{ccccccccccccc}
\mu_0 & & \mu_1 & & & & & & \mu_t & & & & \\
\wr & & \wr & & & & & & \wr & & & & \\
X_0 & \to & X_1 & \to & X_2 & \to & \cdots & \to & X_t & \to & X_{t+1} & \to & \cdots \\
\\
Y_0 & \to & Y_1 & \to & Y_2 & \to & \cdots & \to & Y_t & \to & Y_{t+1} & \to & \cdots \\
\wr & & \wr & & & & & & \wr & & & & \\
\pi & & \pi & & & & & & \pi & & & &
\end{array}$$

由于存在性与特征值法相证，这里证明一收敛性，这两个问题与由一个式子覆盖

即 $\mu_0 \to \underset{t\to\infty}{MC} \to \mu_t$ 后 μ_t 与稳态解无距离无限近

证明方法是构成 r.v. X_0、Y_0，放入 MC 中运行得 X_t、Y_t，且令 $Y_0 = \pi$，这样 $\forall t, Y_t = \pi$，而 $X_0 = \mu_0$。

定义 Coupling 为 X、Y 独立运行 MC 直至某时刻 $X_{t'} = Y_{t'}$，之后两者合并运行 MC

这样就能调用 Coupling 定理来证 $D_{TV} \to 0$

The coupling lemma establishes the connection between the distance of distributions and the discrepancy of random variables. To show that $D_{\mathrm{TV}}(\mu_t, \pi) \to 0$, it is sufficient to construct a coupling ω_t of μ_t and π and then compute $\mathbf{Pr}_{(X_t,Y_t)\sim\omega_t}[X_t \neq Y_t]$.

Here we give a simple coupling. Let $(X_t, Y_t) \sim \omega_t$ and we construct ω_{t+1}. If $X_t = Y_t$ for some $t \geq 0$, then let $X_{t'} = Y_{t'}$ for all $t' > t$, otherwise X_{t+1} and Y_{t+1} are independent. Namely, $\{X_t\}$ and $\{Y_t\}$ are two independent Markov chains until X_t and Y_t reach the same state for some $t \geq 0$, and once they meet together then they move together forever. The coupling lemma tells us that $D_{\mathrm{TV}}(\mu_t, \pi) \leq \mathbf{Pr}_{(X_t,Y_t)\sim\omega_t}[X_t \neq Y_t]$.

Let t^* be the same t^* with Claim 2. Let α be a positive constant such that $P^{t^*}(i,j) \geq \alpha > 0$ for any state pair (i,j). Define event B as $\{\exists t < t^*, X_t = Y_t\}$. We have that

$$\mathbf{Pr}[X_{t^*} = Y_{t^*}] = \mathbf{Pr}[X_{t^*} = Y_{t^*} \wedge B] + \mathbf{Pr}\left[X_{t^*} = Y_{t^*} \wedge \bar{B}\right] \tag{2}$$

Suppose $\{X'_t\}$ and $\{Y'_t\}$ are two independent Markov chains with transition matrix P and $X'_0 \sim \mu_0$ and $Y'_0 \sim \pi$. The only difference between $(\{X'_t\}, \{Y'_t\})$ and $(\{X_t\}, \{Y_t\})$ is that $\{X'_t\}$ and $\{Y'_t\}$ are independent all the time. Then

$$\begin{aligned}\mathbf{Pr}\left[X_{t^*} = Y_{t^*} = 1 \wedge \bar{B}\right] &= \mathbf{Pr}\left[X'_{t^*} = Y'_{t^*} = 1 \wedge \bar{B}\right] \\ &= \mathbf{Pr}[X'_{t^*} = 1] \cdot \mathbf{Pr}[Y'_{t^*} = 1] \\ &\quad - \sum_{t=0}^{t^*-1} \sum_{z\in[n]} \mathbf{Pr}[X'_t = Y'_t = z \wedge \forall s < t, X'_s \neq Y'_s] \cdot \mathbf{Pr}[X'_{t^*} = 1 \mid X'_t = z] \cdot \mathbf{Pr}[Y'_{t^*} = 1 \mid Y'_t = z].\end{aligned}$$

Note that

$$\begin{aligned}\mathbf{Pr}[X_{t^*} = Y_{t^*} \wedge B] &\geq \mathbf{Pr}[X_{t^*} = Y_{t^*} = 1 \wedge B] \\ &= \sum_{t=0}^{t^*-1} \sum_{z\in[n]} \mathbf{Pr}[X_t = Y_t = z \wedge \forall s < t, X_s \neq Y_s] \cdot \mathbf{Pr}[X_{t^*} = 1 \mid X_t = z] \\ &= \sum_{t=0}^{t^*-1} \sum_{z\in[n]} \mathbf{Pr}[X'_t = Y_t = z \wedge \forall s < t, X'_s \neq Y'_s] \cdot \mathbf{Pr}[X'_{t^*} = 1 \mid X'_t = z].\end{aligned}$$

Thus, Equation (2) $\geq \mathbf{Pr}[X'_{t^*} = 1] \cdot \mathbf{Pr}[Y'_{t^*} = 1] \geq \alpha^2$.

By the coupling and the Markov property, we have

$$\begin{aligned}\mathbf{Pr}[X_{2t^*} \neq Y_{2t^*}] &= \mathbf{Pr}[X_{2t^*} \neq Y_{2t^*} | X_{t^*} = Y_{t^*}] \mathbf{Pr}[X_{t^*} = Y_{t^*}] \\ &\quad + \mathbf{Pr}[X_{2t^*} \neq Y_{2t^*} | X_{t^*} \neq Y_{t^*}] \mathbf{Pr}[X_{t^*} \neq Y_{t^*}] \\ &\leq \mathbf{Pr}[X_{2t^*} \neq Y_{2t^*} | X_{t^*} \neq Y_{t^*}] \mathbf{Pr}[X_{t^*} \neq Y_{t^*}] \\ &\leq (1-\alpha^2)^2.\end{aligned}$$

Then we have $\mathbf{Pr}[X_{kt^*} \neq Y_{kt^*}] \leq (1-\alpha^2)^k$ by recursion. It yields that

$$\mathbf{Pr}[X_t \neq Y_t] = \sum_{x_0,y_0\in[n]} \mu_0(x_0) \cdot \pi(y_0) \cdot \mathbf{Pr}[X_t \neq Y_t | X_0 = x_0, Y_0 = y_0] \to 0$$

as $t \to \infty$. □

2 Mixing Time

We are ready to study the convergence rate of Markov chains. We start with the notion of mixing time. For any $\varepsilon > 0$, the mixing time of a Markov chain P up to error ε is the minimum step t such that if we run the Markov chain from any initial distribution, its total variation distance to the stationary distribution is at most ε. Formally,

$$\tau_{\text{mix}}(\varepsilon) := \min_{t} \max_{\mu_0} D_{\text{TV}}(\mu_t, \pi) \le \varepsilon.$$

Recalling in our proof of FTMC using the coupling argument, we obtain the following inequality

$$D_{\text{TV}}(\mu_t, \pi) \le \mathbf{Pr}_{(X_t, Y_t)\sim\omega_t}[X_t \neq Y_t].$$

Therefore, if we can construct a coupling ω_t such that for two arbitrary initial distributions, $\mathbf{Pr}_{(X_t,Y_t)\sim\omega_t}[X_t \neq Y_t] \le \varepsilon$, then $\tau_{\text{mix}}(\varepsilon) \le t$.

Example 1 (Random walk on hypercube). *Consider the random walk on the n-cube. The state space $\Omega = \{0,1\}^n$, and there is an edge between two state x and y iff $\|x - y\|_1 = 1$. We start from a point $X_0 \in \Omega$. In each step,*

- *With probability $\frac{1}{2}$ do nothing.*
- *Otherwise, pick $i \in [n]$ uniformly at random and flip $X(i)$.*

It's equivalent to the following process:

- *Pick $i \in [n], b \in \{0,1\}$ uniformly at random.*
- *Change $X(i)$ to b.*

Now we analyze the mixing time of the process using coupling. We apply the following simple coupling rule:

- *We couple two walks X_t and Y_t by choosing the same i, b in every step.*

  Once a position $i \in [n]$ has been picked, $X_t(i)$ and $Y_t(i)$ will be the same forever. Therefore, the problem again reduces to the coupon collector problem.

  For $t \ge n\log n + cn$, the probability that the i^{th} dimension is not chosen is

$$\left(1 - \frac{1}{n}\right)^{n\log n + cn} \le \frac{e^{-c}}{n}.$$

Then the probability that there exists at least one dimension which is not chosen is no larger than e^{-c}. We want this value to be less than ϵ. Then we choose $c > \log\frac{1}{\epsilon}$. Thus,

$$\tau_{\text{mix}}(\varepsilon) \le n\log\frac{n}{\varepsilon}.$$

下求收敛速度.

我们定义 mixing time 作为收敛速度的指标.

τ_{mix} 意为使不同初始分布 μ_0 之中最大的 $D_{TV}(\mu_t, \pi)$ 最小的 t.

❗ 这里是证收敛速度的关键:

如果能证明对 $\forall X, Y, X_0, Y_0$, 有 $\Pr_{(X,Y)\sim\omega_t}[X_t \neq Y_t] \le \varepsilon$, 则可取 $Y_0 \sim \pi$, X 保持任意, 则有 $\Pr[X_t \neq Y] \le \varepsilon$.

又 $D_{TV} \le \Pr$, 即得证在时间 t 时收敛的结论. 对 t 的数量级进行计算, 即得收敛速度.

❗ 和 convergence 不同之处在于 convergence 关心 $\Pr[X_t \neq Y] \to 0$ $(t \to \infty)$ 是否成立, 由 $t \to \infty$, 其证明的放缩更为随意.

因为对称性, 确定 Y 为平均分布, 设计了一个更简洁且满足的 MC.

设计 coupling

Example 1 是一个用 coupling 求收敛速度的例题;

基本流程为:

① 确定 π, 并设 MC 使其收敛于 π (或题目给 MC)

② 构建 coupling ω, 使在 $\forall X_0, Y_0$ 下, $\Pr[X_t \neq Y] \le \varepsilon$, 求选择的 $t(\varepsilon)$

③ $\tau_{mix} \le t(\varepsilon)$

发现在这种 coupling 下使 $\le \varepsilon$ 的 t 可研究

类似抽卡模型

证毕

方二: $\Pr[\exists i, X_t(i) \neq Y_t(i)] \le \sum_{i=1}^{n} [X_t(i) \neq Y_t(i)]$ ($P(A\cup B) \le P(A)+P(B)$)

$= n \cdot \left(1 - \frac{1}{n}\right)^t \le \varepsilon$ (抽 t 次在此都不同)

化简得相同 t 数量级

Let's modify the process a bit by changing $\frac{1}{2}$ into $\frac{1}{n+1}$, i.e. w.p. $\frac{1}{n+1}$ do nothing, to make the lazy walk more active. Note that we add the lazy move in order to make the chain aperiodic.

Now in this case, we describe another coupling of X_t, Y_t. Without loss of generality, we can reorder the entries of two vectors so that all disagreeing entries come first. Namely there exists an index k such that $X_t(i) \neq Y_t(i)$ if $1 \leq i \leq k$, and $X_t(i) = Y_t(i)$ for $i > k$. Our coupling is as follows:

- *If $k = 0$, Y acts the same as X.*
- *If $k = 1$, Y acts the same as X except when X flips the first entry, Y does nothing and vice versa.*
- *For $k > 2$, we distinguish between whether X flip indices in $[k]$:*
  - *If X did nothing or flipped one of $i > k$: Y acts the same.*
  - *If X flipped $1 \leq i \leq k$: Y flips $(i \bmod k) + 1$, i.e. $1 \mapsto 2, 2 \mapsto 3, \cdots, k-1 \mapsto k, k \mapsto 1$.*

It's clear that the above is indeed a coupling. In fact, this coupling acts like a doubled speed coupon collector, since in the case $k > 2$ we can always collect two coupons at a time when lady luck is smiling. It is therefore conceivable that

$$\tau_{mix} \leq \frac{1}{2} n \log n + O(n).$$

Example 2 (Shuffling cards). *. Given a deck of n cards, consider the following rule of shuffling*

- *pick a card uniformly at random;*
- *put the card on the top.*

The shuffling rule can be viewed as a random walk on all $n!$ permutations of the n cards and it is easy to verify that the uniform distribution is the stationary distribution. Let us design a coupling for this Markov chain. That is, let X_t and Y_t be decks of cards, and we construct X_{t+1} and Y_{t+1} by

- *picking the same random card and put it on the top.*

This is clearly a coupling, and once some card, say $\heartsuit K$ has been picked, then $\heartsuit K$ in two decks will be always at the same location. Therefore, if we ask in how many rounds T, $X_T = Y_T$, then the question is equivalent to the coupon collector problem again. So we have,

$$\tau_{\text{mix}}(\varepsilon) \leq n \log \frac{n}{\varepsilon}.$$

Note that we are picking the "same card", not the card at the same location. That is, we draw a random card from X_t, say $\heartsuit K$, and then we pick $\heartsuit K$ in Y_t as well.

切换题目后元未变，但MC变了。

由新MC设计一个性能更优的Coupling，使为所得的 $t = \frac{1}{\varepsilon}$ 抓，

* 设计Coupling的一个检查点是X,Y间有联系，但外人看来二者按各自分布抽样（联合分布换个描述）

Example 2 是洗牌例子，元们为均匀分布，MC已定，

我们设计了如左的Coupling方法，即Y看X抽样时抽的什么牌，把相同的牌抽出置顶 能。

这样coupling，X,Y变相同速度很快，换句话说 $P_r[X_t \neq Y_t] \leq \varepsilon$ 的 t 很小，即收敛速度很快（同样为抽卡模型）

值得注意一种直观理解。

Coupling Argument 告诉我们只要当一种方法使 $\forall X_0 Y_0$，代入Markov 后 $X_t = Y_t$ 概率很大，则 $X_t \sim M_t$，$Y_t \sim V_t$，M_t 与 V_t 分布间距离很小。

大概率你能M(这么洗牌)已经足够接近均匀分布了

而收敛速度中运用的抽卡模型又提示我们：只需每张卡都抽中过一次。