Lecture 1-2

*Lecturer: Michael Choi* *Scribe: Michael Choi*

# 1 Goal of this lecture

In this lecture, we review essential concepts and introduce basic definitions and techniques in Markov chains mixing time. Most of the material in this lecture is taken from Levin et al. (2009).

# 2 Random mapping representation of Markov chain

A Markov chain is a process which moves among the elements of a set $\mathcal{X}$ in the following manner: when at $x \in \mathcal{X}$, the next position is chosen according to a fixed probability distribution $P(x, \cdot)$ depending only on $x$. More precisely, a sequence of random variables $X = (X_0, X_1, X_2, \ldots)$ is a Markov chain on state space $\mathcal{X}$ with transition matrix $P$ if for all $x, y \in \mathcal{X}$, all events $H_{n-1} = \cap_{s=0}^{n-1}\{X_s = x_s\}$ satisfying $\mathbb{P}(H_{n-1} \cap \{X_n = x\}) > 0$, we have

$$\mathbb{P}(X_{n+1} = y | H_{n-1} \cap \{X_n = x\}) = \mathbb{P}(X_{n+1} = y | X_n = x) = P(x, y).$$

This property is called the **Markov property**. The $x$-th row of $P$ is the distribution $P(x, \cdot)$, and $P$ is stochastic, that is, its entries are all non-negative and

$$\sum_{y \in \mathcal{X}} P(x, y) = 1.$$

A **random mapping representation** of $P$ is a function $f : \mathcal{X} \times \Lambda \to \mathcal{X}$, along with a $\Lambda$-valued random variable $Z$, satisfying $\mathbb{P}(f(x, Z) = y) = P(x, y)$ for all $x, y \in \mathcal{X}$. Our first result below says that we can consider $X$ as a random mapping representation with appropriate $f$ and $Z$:

**Theorem 1** *Every Markov chain on $\mathcal{X}$ has a random mapping representation.*

**Proof:** Take $Z$ to be uniform random variable on $(0, 1)$. For any $i, j \in \mathcal{X}$, set

$$F_{i,j} = \sum_{m=1}^{j} P(i, m)\,.$$

Define

$$f(i, z) := j \qquad \text{when} \quad F_{i,j-1} < z \leq F_{i,j}\,.$$

We have

$$\mathbb{P}(f(i, Z) = j) = \mathbb{P}(F_{i,j-1} < Z \leq F_{i,j}) = P(i, j)\,.$$

□

## 3 Ergodic theorem of Markov chain

$X$ is said to be **irreducible** if for any two states $x, y \in \mathcal{X}$, there exists an integer $n$, possibly depending on $x$ and $y$, such that $P^n(x, y) > 0$. Let $\mathcal{T}(x) := \{n \geq 1;\ P^n(x, x) > 0\}$ be the set of times that is possible to return to the starting state $x$. The period of $x$ is the greatest common divisor of $\mathcal{T}(x)$. $X$ is said to be **aperiodic** if all states have period 1.

We will make use of the following fact to prove the ergodic theorem of Markov chain:

**Fact 1** *Let $X$ be an irreducible and aperiodic Markov chain on a finite state space $\mathcal{X}$. Then there exists an integer $r_0$ such that for all $r \geq r_0$ and for all $x, y \in \mathcal{X}$,*

$$P^r(x, y) > 0.$$

$\pi$ is said to be a **stationary distribution** of $X$ if $\pi P = \pi$. In the following, we would like to quantify the speed of convergence of $P^n$ to $\pi$. Before we do that, we first introduce the total variation metric that will be used to quantify the distance between distributions:

**Definition 1** *The total variation distance between two probabilities $\mu$ and $\nu$ on $\mathcal{X}$ is defined as*

$$||\mu - \nu||_{TV} := \frac{1}{2} \sum_{s \in \mathcal{X}} |\mu(s) - \nu(s)| .$$

We can verify that the above definition defines a metric, and we have the following equivalent characterizations:

**Proposition 2** *1.*

$$||\mu - \nu||_{TV} = \max_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)| .$$

*2.*

$$||\mu - \nu||_{TV} = \max_{\mu(x) \geq \nu(x)} \mu(x) - \nu(x) .$$

*3.*

$$||\mu - \nu||_{TV} = \inf\{\mathbb{P}(X \neq Y);\ (X, Y) \text{ is a coupling of } (\mu, \nu)\} .$$

**Theorem 3 (Convergence theorem)** *Suppose that $P$ is irreducible and aperiodic with stationary distribution $\pi$. Then there exist constants $\alpha \in (0, 1)$ and $C > 0$ such that*

$$\max_{x \in \mathcal{X}} ||P^n(x, \cdot) - \pi||_{TV} \leq C\alpha^n.$$

**Proof:** Since $P$ is irreducible and aperiodic, there exists $r$ such that $P^r$ has strictly positive entries by Fact 1. Let $\Pi$ be the matrix with constant row given by $\pi$. For sufficiently small $\delta > 0$, we have

$$P^n(x, y) \geq \delta\pi(y)$$

for all $x, y \in \mathcal{X}$. Let $\theta = 1 - \delta$. The equation

$$P^r = (1 - \theta)\Pi + \theta Q$$

defines a stochastic matrix $Q$.

Note that for any stochastic matrix $M$, $M\Pi = \Pi$. For $M$ such that $\pi M = \pi$, we have $\Pi M = \Pi$.

Next, we prove by induction on $k \geq 1$ that

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

Clearly, when $k = 1$, the statement is true. Now, we consider

$$\begin{aligned} P^{r(k+1)} = P^{rk}P^r &= (1-\theta^k)\Pi P^r + \theta^k Q^k P^r \\ &= (1-\theta^k)\Pi + \theta^k Q^k\left((1-\theta)\Pi + \theta Q\right) \\ &= (1-\theta^k)\Pi + \theta^k(1-\theta)\Pi + \theta^{k+1}Q^{k+1} \\ &= (1-\theta^{k+1})\Pi + \theta^{k+1}Q^{k+1}. \end{aligned}$$

Multiplying $P^j$ with $0 \leq j \leq r$ gives

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

Now, we sum the absolute values of the elements and divided by 2. On the left, we have $||P^{rk+j}(x,\cdot) - \pi||_{TV}$, thus

$$||P^{rk+j}(x,\cdot) - \pi||_{TV} \leq \theta^k ||Q^k P^j(x,\cdot) - \pi||_{TV} \leq \theta^k = \theta^{-j/r}\theta^{\frac{rk+j}{r}} \leq C\alpha^{rk+j},$$

where we take $C = \theta^{-1}$ and $\alpha = \theta^{1/r}$. □

**Remark 1** *There are many proofs for the convergence theorem. Another classic proof relies on the notion of coupling. We run two independent Markov chains $(X_n)$ and $(Y_n)$, where $X_0 \sim \delta_x$ is the Dirac mass at $x$ and $Y_0 \sim \pi$. Let $\tau_{couple}$ be the coalescence time of the two chains, that is,*

$$\tau_{couple} := \inf\{n;\ X_s = Y_s \text{ for } s \geq n\}.$$

*Then we have*

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

*Then it remains to prove that $\tau_{couple}$ is finite almost surely.*

# 4 Markov chain mixing time

We are interested in

$$\begin{aligned} d(n) &:= \max_{x\in\mathcal{X}} ||P^n(x,\cdot) - \pi||_{TV} \\ \overline{d}(n) &:= \max_{x,y\in\mathcal{X}} ||P^n(x,\cdot) - P^n(y,\cdot)||_{TV} \end{aligned}$$

**Lemma 4** *1.*

$$d(n) \leq \overline{d}(n) \leq 2d(n).$$

*2. The function $\overline{d}$ is submultiplicative, that is, $\overline{d}(m+n) \leq \overline{d}(m)\overline{d}(n)$.*

**Proof:** We first prove item 1. Using triangle inequality, it is easy to see that $\overline{d}(n) \leq 2d(n)$. Next, for arbitrary $x \in \mathcal{X}$ and set $A$, we have

$$\begin{aligned} |P^n(x,A) - \pi(A)| &= \left| \sum_y \pi(y) \left(P^n(x,A) - P^n(y,A)\right) \right| \\ &\leq \sum_y \pi(y) ||P^n(x,\cdot) - P^n(y,\cdot)||_{TV} \leq \overline{d}(n). \end{aligned}$$

Maximizing the left hand side over $x$ and $A$ yields $d(n) \leq \overline{d}(n)$.

Next, we prove item 2. Fix $x, y \in \mathcal{X}$, and let $(X_m, Y_m)$ be the optimal coupling between $P^m(x,\cdot)$ and $P^m(y,\cdot)$ such that

$$||P^m(x,\cdot) - P^m(y,\cdot)||_{TV} = \mathbb{P}(X_m \neq Y_m).$$

Since

$$P^{m+n}(x,w) = \mathbb{E}_x(P^n(X_m, w)),$$

summing over $w \in A$ leads to

$$\begin{aligned} P^{m+n}(x,A) - P^{m+n}(y,A) &= \mathbb{E}_{(x,y)} \left(P^n(X_m,w) - P^n(Y_m,w)\right) \\ &\leq \mathbb{E}_{(x,y)} \left(\overline{d}(n) \mathbb{1}_{\{X_m \neq Y_m\}}\right) \\ &= \overline{d}(n)\overline{d}(m), \end{aligned}$$

where $\mathbb{1}_A$ is the indicator function of the set $A$. □

It is useful to introduce a parameter which measures the time required by a Markov chain for the distance to stationarity to be small. The **mixing time** is defined to be, for $\epsilon > 0$,

$$t_{mix}(\epsilon) := \inf\{n \geq 0;\ d(n) \leq \epsilon\}.$$

A commonly used mixing time parameter is that we take $\epsilon = 1/4$, and define

$$t_{mix} := t_{mix}(1/4).$$

If $l$ is a positive integer, then

$$d(lt_{mix}(\epsilon)) \leq \overline{d}(lt_{mix}(\epsilon)) \leq \overline{d}(t_{mix}(\epsilon))^l \leq (2\epsilon)^l.$$

Taking $\epsilon = 1/4$ yields

$$d(lt_{mix}) \leq 2^{-l}, \quad t_{mix}(\epsilon) \leq \lceil \log_2(\epsilon^{-1}) \rceil t_{mix}.$$

# 5 Spectral bounds of mixing time for reversible Markov chains

In Theorem 3, the constants $C$ and $\alpha$ cannot be readily computed to give bounds on $t_{mix}$. In this section, we will prove computable spectral bounds on $t_{mix}$ for reversible Markov chains.

The transition matrix $P$ is said to be **reversible** with respect to the stationary distribution $\pi$ if for all $x, y \in \mathcal{X}$, the detailed balance condition is satisfied, i.e. $\pi(x)P(x,y) = \pi(y)P(y,x)$. Denote by $\langle \cdot, \cdot \rangle$ the usual inner product on $\mathbb{R}^{\mathcal{X}}$, given by $\langle f, g \rangle = \sum_{x \in \mathcal{X}} f(x)g(x)$. We will also need another inner product, denoted by $\langle \cdot, \cdot \rangle_\pi$ and defined by

$$\langle f, g \rangle_\pi := \sum_{x \in \mathcal{X}} f(x)g(x)\pi(x) \tag{1}$$

We write $\ell^2(\pi)$ for the vector space $\mathbb{R}^{\mathcal{X}}$ equipped with the inner product (1). We begin with a useful fact on the eigenvalues of a general (not necessarily reversible) Markov chain:

**Fact 2** *Let $P$ be the transition matrix of a finite Markov chain.*

1. *If $\lambda$ is an eigenvalue of $P$, then $|\lambda| \leq 1$.*
2. *If $P$ is ergodic (i.e. irreducible and aperiodic), then the eigenvalue 1 has both algebraic and geometric multiplicity of 1, and $-1$ is not an eigenvalue of $P$.*

**Proof:** We first prove item 1. Let $||f||_\infty := \max_{x \in \mathcal{X}} |f(x)|$, then $||Pf||_\infty \leq ||f||_\infty$. Now, we take $f$ to be the eigenfunction of $\lambda$.

Item 2 follows directly from the Perron-Frobenius theorem. □

**Theorem 5 (spectral decomposition of reversible Markov chain)** *Let $P$ be reversible with respect to $\pi$.*

1. *The inner product space $\left(\mathbb{R}^{\mathcal{X}}, \langle \cdot, \cdot \rangle_\pi\right)$ has an orthonormal basis of real-valued eigenfunctions $\{f_j\}_{j=1}^{|\mathcal{X}|}$ corresponding to real eigenvalues $\{\lambda_j\}$.*
2. *The matrix $P$ can be decomposed as*

$$\frac{P^t(x,y)}{\pi(y)} = \sum_{j=1}^{|\mathcal{X}|} f_j(x)f_j(y)\lambda_j^t.$$

3. *The eigenfunction $f_1$ corresponding to the eigenvalue 1 can be taken to be the constant vector $\mathbf{1}$, in which case*

$$\frac{P^t(x,y)}{\pi(y)} = 1 + \sum_{j=2}^{|\mathcal{X}|} f_j(x)f_j(y)\lambda_j^t.$$

**Proof:** We first prove item 1. We write $D_\pi$ to be the diagonal matrix with diagonal entries $\pi$, that is, $D_\pi(x,x) = \pi(x)$ for $x \in \mathcal{X}$, and define $A := D_\pi^{1/2} P D_\pi^{-1/2}$, i.e.

$A(x,y) = \pi(x)^{1/2}\pi(y)^{-1/2}P(x,y)$ for all $x, y \in \mathcal{X}$. Reversibility of $P$ implies $A$ is symmetric, and the spectral theorem for symmetric matrices guarantees that the inner product space $\left(\mathbb{R}^{\mathcal{X}}, \langle\cdot,\cdot\rangle_\pi\right)$ has an orthonormal basis of real-valued eigenfunctions $\{f_j\}_{j=1}^{|\mathcal{X}|}$ corresponding to real eigenvalues $\{\lambda_j\}$.

If $f_j := D_\pi^{-\frac{1}{2}}\varphi_j$, then $f_j$ is an eigenfunction of $P$ with eigenvalue $\lambda_j$:

$$Pf_j = PD_\pi^{-\frac{1}{2}}\varphi_j = D_\pi^{-\frac{1}{2}}\left(D_\pi^{\frac{1}{2}}PD_\pi^{-\frac{1}{2}}\right)\varphi_j = D_\pi^{-\frac{1}{2}}A\varphi_j = D_\pi^{-\frac{1}{2}}\lambda_j\varphi_j = \lambda_j f_j.$$

Although the eigenfunctions $\{f_j\}$ are not necessarily orthonormal with respect to the usual inner product, they are orthonormal with respect to the inner product $\langle\cdot,\cdot\rangle_\pi$ (The first equality follows since $\{\varphi_j\}$ is orthonormal with respect to the usual inner product.) This proves item 1.

Next, we prove item 2. Let $\delta_y$ be the function

$$\delta_y(x) = \begin{cases} 1 & \text{if } y = x, \\ 0 & \text{if } y \neq x. \end{cases}$$

Considering $\left(\mathbb{R}^{\mathcal{X}}, \langle\cdot,\cdot\rangle_\pi\right)$ with its orthonormal basis of eigenfunctions $\{f_j\}_{j=1}^{|\mathcal{X}|}$, the function $\delta_y$ can be written via basis decomposition as

$$\delta_y = \sum_{j=1}^{|\mathcal{X}|} \langle \delta_y, f_j\rangle_\pi f_j = \sum_{j=1}^{|\mathcal{X}|} f_j(y)\pi(y) f_j.$$

Since $P^t f_j = \lambda_j^t f_j$ and $P^t(x,y) = \left(P^t\delta_y\right)(x)$,

$$P^t(x,y) = \sum_{j=1}^{|\mathcal{X}|} f_j(y)\pi(y)\lambda_j^t f_j(x).$$

Divide by $\pi(y)$ completes the proof of item 2. Item 3 follows from item 2 and $f_1 = \mathbf{1}$. □

Define

$$\lambda_\star := \max\{|\lambda| : \lambda \text{ is an eigenvalue of } P, \lambda \neq 1\}.$$

Some people call $\lambda_\star$ the **second largest eigenvalue in modulus (SLEM)**. The difference $\gamma_\star := 1 - \lambda_\star$ is called the **absolute spectral gap**. Fact 2 implies that if $P$ is aperiodic and irreducible, then $\gamma_\star > 0$. For a reversible ergodic transition matrix $P$, we label the eigenvalues of $P$ in decreasing order:

$$1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|\mathcal{X}|} > -1$$

The **spectral gap** of a reversible chain is defined by $\gamma := 1 - \lambda_2$. The **relaxation time** $t_{rel}$ of a reversible Markov chain with absolute spectral gap $\gamma_\star$ is defined to be

$$t_{rel} := \frac{1}{\gamma_\star}.$$

We are now ready to relate mixing time with relaxation time:

**Theorem 6 (Spectral upper bound of mixing time)** *Let $P$ be the transition matrix of a reversible and ergodic Markov chain on state space $\mathcal{X}$, and let $\pi_{min} := \min_{x\in\mathcal{X}} \pi(x)$. Then*

$$t_{mix}(\epsilon) \le t_{rel} \log\left(\frac{1}{\epsilon \pi_{min}}\right).$$

**Proof:** By Theorem 5 item 3 and the Cauchy-Schwartz inequality, we have

$$\left|\frac{P^t(x,y)}{\pi(y)} - 1\right| \le \sum_{j=2}^{|\mathcal{X}|} |f_j(x) f_j(y)|\, \lambda_\star^t \le \lambda_\star^t \left[\sum_{j=2}^{|\mathcal{X}|} f_j^2(x) \sum_{j=2}^{|\mathcal{X}|} f_j^2(y)\right]^{1/2}. \tag{2}$$

Using the orthonormality of $\{f_j\}$ shows that

$$\pi(x) = \langle \delta_x, \delta_x \rangle_\pi = \left\langle \sum_{j=1}^{|\mathcal{X}|} f_j(x)\pi(x) f_j, \sum_{j=1}^{|\mathcal{X}|} f_j(x)\pi(x) f_j \right\rangle_\pi = \pi(x)^2 \sum_{j=1}^{|\mathcal{X}|} f_j(x)^2.$$

Consequently, $\sum_{j=2}^{|\mathcal{X}|} f_j(x)^2 \le \pi(x)^{-1}$. This bound and (2) imply that

$$\left|\frac{P^t(x,y)}{\pi(y)} - 1\right| \le \frac{\lambda_\star^t}{\sqrt{\pi(x)\pi(y)}} \le \frac{\lambda_\star^t}{\pi_{min}} = \frac{(1-\gamma_\star)^t}{\pi_{min}} \le \frac{e^{-\gamma_\star t}}{\pi_{min}}.$$

□

**Theorem 7 (Spectral lower bound of mixing time)** *Let $P$ be the transition matrix of a reversible ergodic Markov chain on state space $\mathcal{X}$, and suppose that $\lambda \neq 1$ is an eigenvalue of $P$. Then*

$$t_{mix}(\epsilon) \ge \left(\frac{1}{1-|\lambda|} - 1\right) \log\left(\frac{1}{2\epsilon}\right).$$

*In particular,*

$$t_{mix}(\epsilon) \ge (t_{rel} - 1) \log\left(\frac{1}{2\epsilon}\right).$$

**Proof:** We may assume that $\lambda \neq 0$. Suppose that $Pf = \lambda f$ with $\lambda \neq 1$. Since $\mathbb{E}_\pi(f) = \langle \mathbf{1}, f \rangle_\pi = 0$, it follows that

$$\left|\lambda^t f(x)\right| = \left|P^t f(x)\right| = \left|\sum_{y\in\mathcal{X}} \left[P^t(x,y) f(y) - \pi(y) f(y)\right]\right| \le \|f\|_\infty 2d(t).$$

Taking $x$ with $|f(x)| = \|f\|_\infty$ yields

$$|\lambda|^t \le 2d(t).$$

Therefore, $|\lambda|^{t \min(\varepsilon)} \le 2\varepsilon$, whence

$$t_{mix}(\varepsilon)\left(\frac{1}{|\lambda|} - 1\right) \ge t_{mix}(\epsilon) \log\left(\frac{1}{|\lambda|}\right) \ge \log\left(\frac{1}{2\epsilon}\right).$$

□

## References

D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, Providence, RI, 2009.