Lectures 16–17: Markov chains and mixing times CSE 525, Spring 2019
Instructor: James R. Lee

1 General Markov chains

Consider a finite state space Ω and a *transition kernel* $P : \Omega \times \Omega \to [0,1]$ such that for every $x \in \Omega$, $\sum_{y\in\Omega} P(x,y) = 1$. The *Markov chain* corresponding to the kernel P is the sequence of random variables $\{X_0, X_1, X_2, \ldots\}$ such that for every $t \geqslant 0$, we have $\mathbb{P}[X_{t+1} = y \mid X_t = x] = P(x,y)$. Note that we also have to specify a distribution for the initial state X_0.

Corresponding to every such process, one can consider the (weighted) directed graph $D = (\Omega, A)$ with $A = \{(x,y) : P(x,y) > 0\}$ and edge weights $w(x,y) = P(x,y)$. Then the random process $\{X_t\}$ corresponds precisely to random walk on D: At every time step, one moves from the current vertex x to a neighbor y with probability $P(x,y)$.

Convergence to stationarity. For every $t \geqslant 0$, let $P^t(x,y) = \mathbb{P}[X_t = x \mid X_0 = y]$. The Markov chain described by P is said to be *irreducible* if for every $x, y \in \Omega$, there is some t such that $P^t(x,y) > 0$; in other words, there is always some way to reach any state form any other. This corresponds precisely to the digraph D being strongly connected. The chain is *aperiodic* if for every $x, y \in \Omega$,

$$\gcd(\{t : P^t(x,y) > 0\}) = 1\,.$$

Theorem 1.1 (Fundamental Theorem of Markov Chains)**.** *If P is irreducible and aperiodic, then there is a unique probability measure $\pi : \Omega \to [0,1]$ such that for every $x, y \in \Omega$, we have*

$$P^t(x,y) \to \pi(y) \quad \textit{as} \quad t \to \infty\,.$$

In other words, the Markov chain "forgets" where it started and converges to a unique limiting distribution. This is referred to as the *stationary measure π.*

Reversibility. A Markov chain is said to be *reversible with respect to the measure μ* if for every $x, y \in \Omega$, we have $\mu(x)P(x,y) = \mu(y)P(y,x)$. (These are called the "detailed balance conditions.") The chain is said to be *reversible* if it is reversible with respect to some probability measure. Note that reversible chains correspond precisely to random walks on (weighted) *undirected graphs.*

Also, if P is irreducible and aperiodic—and hence has a unique stationary measure π by Theorem 1.1—then actually $\pi = \mu$. To see this, note that by the detailed balance conditions: For every $y \in \Omega$, we have

$$\sum_{x\in\Omega} \mu(x)P(x,y) = \sum_{x\in\Omega} \mu(y)P(y,x) = \mu(y)\sum_{x\in\Omega} P(y,x) = \mu(y)\,. \tag{1.1}$$

The right-hand side can be interpreted as the probability of going to y in one step started from the measure μ. Now Theorem 1.1 implies that if we start from distribution μ, then we converge to π; on the other hand, (1.1) says that if we start distributed according to μ, then we stay that way under the chain. Thus $\mu = \pi$. This provides a nice local way to check that some measure is the stationary measure of the chain.

Remark 1.2. If P is irreducible, but not necessarily aperiodic, then there is still a unique stationary distribution, i.e. a probability π such that for every $x \in \Omega$, $\sum_{y\in\Omega} P(x,y)\pi(y) = \pi(x)$. But it may not be the case that the chain converges to π from some starting states.

For instance, if the chain is given by a directed graph with two nodes $\Omega = \{x, y\}$ and arcs (x, y) and (y, x), then $\pi = (1/2, 1/2)$ is the unique stationary measure, but the chain does not converge to π when starting in either state x or y (because of periodicity).

For our purposes, aperiodicity is a rather weak obstruction to mixing. Given any chain P and number $\alpha \in (0, 1)$, we can consider the chain $P' = \alpha I + (1 - \alpha)P$. If P is irreducible, then so is P'. Moreover, for any such α, the chain P' is aperiodic (even if P was not). When measuring convergence to equilibrium, this α "self loop" probability does not slow down the chain too much.

1.1 The Fundamental Theorem

Let us sketch a proof of Theorem 1.1. We want to begin with a *stationary measure* π for P, i.e., a probability distribution π on Ω (interpreted as a row vector) such that $\pi P = \pi$. Suppose $\{X_t\}$ is the Markov chain with transition law P and for $x \in \Omega$, define

$$\tau_x^+ := \min\{t > 0 : X_t = x\}.$$

We do not prove the following lemma (see, e.g., Section 1.5.3 in the Levin-Peres-Wilmer book).

Lemma 1.3. *Suppose that P is irreducible and aperioidic. If we define*

$$\pi(x) = \frac{1}{\mathbb{E}[\tau_x^+ \mid X_t = 0]},$$

then π is a probability distribution satisfying $\pi P = \pi$.

Now let us argue that $P^t(x, y) \to \pi(y)$ for every $x, y \in \Omega$. Let Π denote the $|\Omega| \times |\Omega|$ matrix where every row is π.

Fact 1.4. *It holds that $\Pi P = \Pi$ and $Q\Pi = \Pi$ for* every *row-stochastic matrix Q.*

Using the fact that P is irreducible and aperiodic, choose r such that $P^r(x, y) > 0$ for every $x, y \in \Omega$. Let $\theta < 1$ be such that

$$P^r(x, y) \geqslant (1 - \theta)\pi(y) \qquad \forall x, y \in \Omega.$$

Then we can write

$$P^r = (1 - \theta)\Pi + \theta Q, \tag{1.2}$$

where Q is stochastic. Now we claim that for every $k \geqslant 1$:

$$P^{kr} = (1 - \theta^k)\Pi + \theta^k Q^k. \tag{1.3}$$

If this holds, then for $s < r$, we have

$$P^{kr+s} = P^{kr} P^s = (1 - \theta^k)\Pi + \theta^k Q^k P^s,$$

and thus $P^t \to \Pi$ as $t \to \infty$, completing the proof of Theorem 1.1.

We prove (1.3) by induction on k. The case $k = 1$ is (1.2). In the general case, write

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

where in the second line we use (1.2). Now (1.4) implies that $\Pi P^r = \Pi$ and $Q^k \Pi = \Pi$, hence

$$P^{r(k+1)} = \left[(1 - \theta^k) + \theta^k(1 - \theta)\right]\Pi + \theta^{k+1} Q^{k+1} = (1 - \theta^{k+1})\Pi + \theta^{k+1} Q^{k+1},$$

completing the proof.

1.2 Eigenvalues and mixing

It will be useful to give a more quantitative proof of Theorem 1.1 in the reversible case. To do this, we again think of P as an $\Omega \times \Omega$ matrix. If we also think about a probability measure $\mu \in \mathbb{R}^\Omega$ as a row vector, then μP denotes the distribution that arises by starting at μ and taking one step of the chain associated to P.

If P is reversible with respect to π, then (1.1) implies that $\pi P = \pi$, i.e. π is a (left) eigenvector with eigenvalue 1. We now analyze the other eigenvalues of P.

Real eigenvalues. Note that P is not necessarily a symmetric matrix, but we can prove that P is similar to a symmetric matrix. Let D denote the diagonal matrix with $D_{xx} = \pi(x)$. Then

$$(\sqrt{D^{-1}}P\sqrt{D})_{xy} = \langle e_x, e_y\sqrt{D^{-1}}P\sqrt{D}\rangle = \langle \sqrt{D^{-1}}e_x, e_y P\sqrt{D}\rangle = \sqrt{\frac{\pi(x)}{\pi(y)}}P(x,y)\,.$$

But by (1.1), this is equal to $\sqrt{\frac{\pi(y)}{\pi(x)}}P(y,x)$. Thus $\sqrt{D^{-1}}P\sqrt{D}$ is a real, symmetric matrix and hence has real eigenvalues. This implies that P also has real eigenvalues.

All eigenvalues in $[-1,1]$. Now note that for any $v \in \mathbb{R}^\Omega$, we have

$$\|vP\|_1 \leqslant \||v|P\|_1 = \|v\|_1\,, \tag{1.4}$$

where $|v|$ denotes the vector whose entries are the absolute value of the corresponding entries in v. This is simply because P is an averaging operator.

Now suppose that $vP = \lambda v$. Then using (1.4)

$$|\lambda| \cdot \|v\|_1 = \|vP\|_1 \leqslant \|v\|_1\,,$$

implying that $|\lambda| \leqslant 1$.

Unique eigenvector with eigenvalue 1. Suppose now that $v = vP$ and consider the corresponding Laplacian matrix $L = D - PD$ (using our notation for "edge Laplacians," this is $\frac{1}{2}\sum_{x,y}\pi(x)P(x,y)L_{\{x,y\}}$). One can check that this matrix is symmetric since PD is symmetric by the detailed balance conditions. As we saw in Lectures 14-15, for any vector w we have

$$wLw^T = \frac{1}{2}\sum_{x,y}\pi(x)P(x,y)\,(w_x - w_y)^2\,.$$

This is easiest to see by writing $L = \sum_{\{x,y\}} c_{xy}L_{xy}$ where $c_{xy} := \pi(x)P(x,y)$ and L_{xy} is the (unweighted) Laplacian corresponding to the graph with a single edge $\{x,y\}$. (The factor 1/2 is due to the fact that we are summing over all pairs x, y vs. all edges $\{x,y\}$.)

Let $w = vD^{-1}$. Then $vP = v \implies wL = 0 \implies wLw^T = 0$. Thus $w_x = w_y$ whenever $P(x,y) > 0$. But since the chain P is irreducible, we can connect every pair x, y by a chain of such implications, implying that $w = \alpha(1,1,\ldots,1)$ is a multiple of the all-ones vector. But this implies that $v = Dw$ is a multiple of π. Since P is an averaging operator, it preserves the ℓ_1 norm, hence $\alpha = 1$ and $v = \pi$.

Not a bipartite graph. Now we claim that if P is aperiodic, -1 cannot be an eigenvalue of P. Suppose, for the sake of contradiction, that $vP = -v$ for some $v \neq 0$. Again, let $|v|$ denote the vector whose entries are the absolute values of the corresponding entries in v. Then

$$\|v\|_2^2 = \left||v|Pv^T\right| \leqslant |v|P|v|^T \leqslant \|v\|_2^2\,,$$

where the last inequality follows from the fact that all the eigenvalues of P lie in $[-1,1]$. We conclude that $|v|P=|v|$, implying that $|v|=\pi$.

Finally, observe that $vP=-v$ implies that for every x, one has $v_x=-(Pv)_x$, hence

$$\pi(x)\mathrm{sgn}(v_x)=v_x=-(vP)_x=-\sum_y P(y,x)v_y=-\sum_y P(y,x)\pi(y)\mathrm{sgn}(v_y)\,.$$

But by the detailed balance conditions, we have $\pi(x)=\sum_y P(y,x)\pi(y)$. Hence it must be that $\mathrm{sgn}(v_x)=-\mathrm{sgn}(v_y)$ whenever $P(y,x)>0$.

Thus if we set $L=\{x : v_x<0\}$ and $R=\{x : v_x>0\}$, then $P(x,y)>0$ implies x and y are on different sides of the bipartition. (Note that this is a bipartition since $|v|=\pi$ implies that $v_x\neq 0$ for any $x\in\Omega$.) But this implies that for x,y on the same side of the bipartition, we have $P^t(x,y)=0$ when t is odd, contradicting the fact that P was assumed aperiodic.

Convergence to stationarity (spectral argument). Let us fix an inner product in which the matrix P is self-adjoint:

$$\langle u,v\rangle_{L^2(\pi)}=\sum_{x\in\Omega}\pi(x)u_xv_x,$$

and let $\|u\|_{L^2(\pi)}=\sqrt{\langle u,u\rangle_{L^2(\pi)}}$ denote the corresponding Euclidean norm.

Consider any vector $w\in\mathbb{R}^\Omega$. Let $\lambda_1=1,\lambda_2,\ldots,\lambda_n$ denote the (left) eigenvalues of P arranged so that $1=|\lambda_1|\geqslant|\lambda_2|\geqslant\cdots\geqslant|\lambda_n|$. Since P is self-adjoint with respect to $L^2(\pi)$, we can choose an $L^2(\pi)$-orthonormal basis $v^{(1)},v^{(2)},\ldots,v^{(n)}$ of corresponding eigenvectors with $v^{(1)}$ a multiple of π.

Recall that from the above reasoning, we have $|\lambda_i|<1$ for $i>1$. Write $w=\sum_{i=1}^n\alpha_i v^{(i)}$, and note that for any $t\geqslant 0$,

$$wP^t=\alpha_1v^{(1)}+\sum_{i\geqslant 2}\lambda_i^t\alpha_iv^{(i)}\,.$$

In particular, we have

$$\left\|wP^t-\alpha_1v^{(1)}\right\|_{L^2(\pi)}^2=\sum_{i\geqslant 2}\lambda_i^{2t}|\alpha_i|^2\leqslant\lambda_2^{2t}\|w\|_{L^2(\pi)}^2\,. \tag{1.5}$$

Since $|\lambda_2|<1$, this implies that $\|wP^t-\alpha_1v^{(1)}\|_{L^2(\pi)}\to 0$ as $t\to\infty$, showing that $wP^t\to\alpha_1v^{(1)}$, where we recall that $v^{(1)}$ is a multiple of π.

Note that if w has all non-negative entries, then since P is an averaging operator, we have $\|wP^t\|_1=\|w\|_1$, hence $wP^t\to\|w\|_1\pi$. Finally, observe that if $v=e_x$ then this implies $e_xP^t\to\pi$, which is exactly the claim of Theorem 1.1 (in the reversible case). For later use, we note the following consequence of (1.5): If $x\in\Omega$, then

$$\left\|e_xP^t-\pi\right\|_{L^2(\pi)}^2\leqslant\lambda_2^{2t}\pi(x). \tag{1.6}$$

1.3 Mixing times

Now we have seen that any irreducible, aperiodic Markov chain P on a finite state space Ω converges to a unique stationary measure π. We are not only concerned with convergence, but also the rate of convergence—we would like to be able to sample efficiently from π.

To this end, we first introduce a metric on the space of probability measures on Ω: For any two measures μ and ν on Ω, the *total variation distance* is defined by

$$d_{TV}(\mu,\nu)\stackrel{\text{def}}{=}\frac{1}{2}\|\mu-\nu\|_1=\frac{1}{2}\sum_{x\in\Omega}|\mu(x)-\nu(x)|\,.$$

As an exercise, one can also show that $d_{TV}(\mu,\nu) = \max_{A\subseteq\Omega} |\mu(A)-\nu(A)|$.

For simplicity of notation, let us define $p_t^{(x)}$ to be the distribution given by $P^t e_x$ (i.e., the distribution of the chain started at x after t steps). For any $t \geqslant 0$ and $x \in \Omega$, define the quantity $\Delta_x(t) = d_{TV}(\pi, p_t^{(x)})$, and we set $\Delta(t) = \max_{x\in\Omega} \Delta_x(t)$. For $\varepsilon > 0$, we denote

$$\tau(\varepsilon) = \min\{t : \Delta(t) \leqslant \varepsilon\}\,.$$

In words, this is the first time t such that, starting from any initial state, the measure of the chain after t steps is within ε of the stationary measure. Finally, by convention, one takes $\tau_{\text{mix}} = \tau(1/2e)$ as the *mixing time* of the Markov chain P. Note that the precise value of ε is not so important; as the following lemma shows, once we have obtained the mixing time, further convergence to stationarity happens very fast.

Lemma 1.5. *For every $t \geqslant 0$, we have*

$$\Delta(t) \leqslant \exp\left(-\left\lfloor \frac{t}{\tau_{\text{mix}}} \right\rfloor\right).$$

In particular, for every $\varepsilon > 0$, it holds that $\tau(\varepsilon) \leqslant \tau_{\text{mix}} \lceil \ln(1/\varepsilon) \rceil$.

We will not prove this, but it can be done using the coupling characterization of total variation distance that appears in Homework #6. Finally, we can use our proof of Theorem 1.1 in the reversible case to give an upper bound on τ_{mix} in terms of the spectral gap of the chain.

Theorem 1.6. *Let P be a reversible and irreducible, aperiodic Markov chain on the state space Ω. Suppose that P has eigenvalues $1 = \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$, and let $\lambda(P) = \max\{|\lambda_2|, |\lambda_n|\}$. Then*

$$\tau_{\text{mix}} \leqslant \left\lceil \frac{1+\ln(1/\pi_{\min})}{1-\lambda(P)} \right\rceil,$$

where $\pi_{\min} := \min\{\pi(x) : x \in \Omega\}$.

Proof. Consider $x \in \Omega$ and $\varepsilon > 0$. Recall that

$$d_{TV}(p_t^{(x)}, \pi) = \frac{1}{2}\sum_{y\in\Omega} \left|P^t(x,y) - \pi(y)\right| = \frac{1}{2}\sum_{y\in\Omega} \pi(y)\left|\frac{P^t(x,y)}{\pi(y)} - 1\right| \leqslant \frac{1}{2}\left(\sum_{y\in\Omega}\pi(y)\left|\frac{P^t(x,y)}{\pi(y)}-1\right|^2\right)^{1/2},$$

where the last line uses $\mathbb{E}[X^2] \geqslant (\mathbb{E}[X])^2$. Observe that

$$\sum_{y\in\Omega}\pi(y)\left|\frac{P^t(x,y)}{\pi(y)}-1\right|^2 \leqslant \frac{1}{\pi_{\min}}\sum_{y\in\Omega}\pi(y)^2\left|\frac{P^t(x,y)}{\pi(y)}-1\right|^2 = \frac{1}{\pi_{\min}}\sum_{y\in\Omega}\pi(y)\left|P^t(x,y)-\pi(y)\right|^2$$

$$= \frac{1}{\pi_{\min}}\left\|e_x P^t - \pi\right\|_{L^2(\pi)}^2.$$

Combining the preceding two inequalities with (1.6) gives

$$d_{TV}(p_t^{(x)}, \pi)^2 \leqslant \frac{1}{\pi_{\min}}\left\|e_x P^t - \pi\right\|_{L^2(\pi)}^2 \leqslant \frac{\pi(x)}{\pi_{\min}}\lambda(P)^{2t}$$

Now setting $t = \lceil \frac{1}{1-\lambda(P)} \ln(1/(\varepsilon\pi_{\min}))\rceil$ and using the fact that $(1-\delta)^{1/\delta} \leqslant e^{-1}$ for $\delta > 0$ yields

$$d_{TV}(p_t^{(x)}, \pi)^2 \leqslant \frac{\varepsilon^2}{4},$$

implying $d_{TV}(p_t^{(x)}, \pi) \leqslant \varepsilon/2$. Setting $\varepsilon = 1/e$ and recalling the definition of τ_{mix} yields the desired result. □

Finally, one should note that this bound is essentially tight up to the $O(\log(1/\pi_{\min}))$ factor.

Theorem 1.7. *Under the assumption of Theorem 1.6, we have*

$$\tau_{\text{mix}} \geqslant \frac{1}{1-\lambda(P)} - 1\,.$$

Proof. Let v be a (left) eigenvector of P with eigenvalue $\lambda = \lambda(P) \neq 1$. In that case, since π is also an eigenvector of P, we see that v is orthogonal to the stationary measure π, i.e. $\sum_{y\in\Omega} \pi(y)v_y = 0$. It follows that for $t \geqslant 0$ and any $x \in \Omega$,

$$|\lambda^t v_x| = |(vP^t)_x| = \left|\sum_y P^t(x,y)v_y - \pi(y)v_y\right| \leqslant \|v\|_\infty \sum_y |P^t(x,y) - \pi(y)| = 2\|v\|_\infty\, d_{TV}(p_t^{(x)}, \pi)\,.$$

Now choose x so that $|v_x| = \|v\|_\infty$, yielding

$$d_{TV}(p_t^{(x)}, \pi) \geqslant \frac{1}{2}\lambda(P)^t\,.$$

Therefore $\lambda(P)^{\tau_{\text{mix}}} \leqslant 1/e$, implying that

$$\tau_{\text{mix}} \geqslant \frac{-1}{\log(1-(1-\lambda(P)))} \geqslant \frac{1}{1-\lambda(P)} - 1\,,$$

where in the final line we have used that $\log(1-a) \geqslant 1 + \frac{1}{a-1}$ for all $a \in [0,1]$. ◻

So we see that up to a $\log(1/\pi_{\min})$ factor, the spectral gap $1-\lambda(P)$ controls the mixing time of the chain: If we set $\tau_{\text{rel}} := \frac{1}{1-\lambda(P)}$ (commonly called the "relaxation time" of the chain), then

$$\tau_{\text{rel}} - 1 \leqslant \tau_{\text{mix}} \leqslant O(\log(1/\pi_{\min}))\,\tau_{\text{rel}}.$$

1.4 Some Markov chains

One famous state space is the set of all permutations of n objects (for $n = 52$). In this case, $|\Omega| = n!$. Here are some shuffles:

1. **Random transposition.** At every step, we choose two uniformly random positions i and j (with replacement) and swap the cards at positions i and j.

2. **Top to random.** We take the top card and insert it at one of the n positions in the deck uniformly at random.

3. **Riffle shuffle.** We split the deck into two parts L and R uniformly at random, and then take a uniformly random interleaving of L and R.

And here's a combinatorial example: Let $G = (V,E)$ be a graph with degree at most Δ, and suppose we have q colors with $q \geqslant \Delta + 1$ (so we are assured that G is q-colorable). Let Ω be the set of all q-colorings on G. Here is a natural Markov chain: Suppose we have a proper coloring $\chi : V \to [q]$. We choose a uniformly random $v \in V$ and a uniformly random color $c \in [q]$. If no neighbor of v in χ has color c, then we color v with c. Otherwise, we stay at the current coloring.

This example demonstrates the complex structure of Markov chains on combinatorial state spaces. For what values of q (depending on Δ) is the chain irreducible? It turns out that if $q \geqslant \Delta + 2$, then the chain is always irreducible, and the stationary measure is uniform on proper q-colorings. A huge open problem in MCMC (Markov chain Monte Carlo) is to resolve the following conjecture.

Conjecture 1.8. *For all $q \geqslant \Delta + 2$, this Markov chain has mixing time $O(n \log n)$, where $n = |V|$.*

The best bound (due to Vigoda, 1999) is that this holds for $q \geqslant \frac{11}{6}\Delta$.[1]

[1] This year (2019), a group from MIT has improved this bound slightly.