Statistics C206B Lecture 3 Notes

Daniel Raban

January 25, 2022

1 Markov Chains and Mixing Times

1.1 Markov chains

1.1.1 Irreducibility and stationarity

We will review some Markov chain theory, including topics such as mixing times, which will be useful for future arguments. The reference book is by Levin, Peres, and Wilmer.

We will work with chains with finite state space in discrete time. This means that we have a finite state space Ω and a transition kernel (a matrix) P with entries $P(x, y)$ being the probability of going from x to y in one time step.

This forms a sequence of random variables $X_0, X_1, \dots$, where X_0 is the initial state and X_t is the state at time t. The chain evolves by

$$\mathbb{P}(X_{t+1} = y \mid X_t = x) = P(x, y)$$

Definition 1.1. A chain is **irreducible** if for any $x, y \in \Omega$, there is some t such that $P^t(x, y) > 0$.

Example 1.1 (Random walk on a graph)**.** Consider a graph $G = (V, E)$.

For $x, y \in V$,

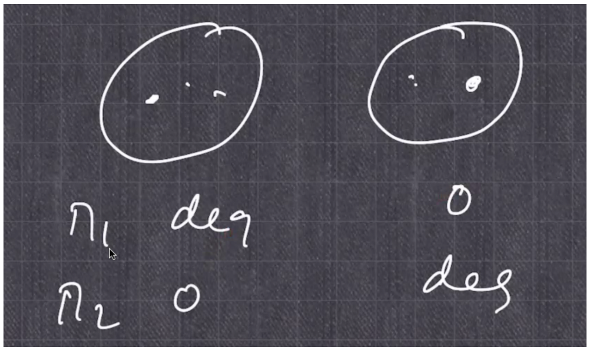
$$P(x, y) = \begin{cases} 0 & x \not\sim y \\ \frac{1}{\deg x} & x \sim y. \end{cases}$$

Definition 1.2. A **stationary/invariant measure** is a measure π such that if $X_0 \sim \pi$, then $X_1 \sim \pi$. That is, $\pi P = \pi$.

A stationary measure can be thought of as a vector $(\pi(1), \pi(2), \dots, \pi(n))$.

Example 1.2. If G is a graph, what is π for the random walk on G? Then $\pi(x) \propto \deg(x)$ is stationary.

But can there be other stationary measures? π is unique if the Markov chain is irreducible. If the graph is not connected, multiple connected components may have their own stationary measures.



$\pi(x)$ has an expression terms of return times. If $\tau_x = \inf\{t > 0 : x_t = x\}$, then

$$\pi(x) = \frac{1}{\mathbb{E}_x[\tau_x]}.$$

1.1.2 Aperiodicity

For any $x \in \Omega$, let $S_x = \{t : P^t(x,x) > 0\}$ be the set of return times to x.

Definition 1.3. A Markov chain is **aperiodic** if for any x, $\gcd\{t : t \in S_x\} = 1$.

Example 1.3. For the random walk on $\mathbb{Z}$, $P^t(x,x) > 0$ iff t is even, so the chain is periodic.

1.1.3 Reversibility

The intuition is that a reversible chain looks the same when you run it forwards and backwards.

Definition 1.4. A Markov chain is said to be **reversible** if there exists a probability measure π such that for any $x, y \in \Omega$,

$$\pi(x)p(x,y) = \pi(y)p(y,x).$$

These are called the **detailed balance equations**. A probability measure satisfying detailed balance must be invariant. This is sometimes an easier condition to check.

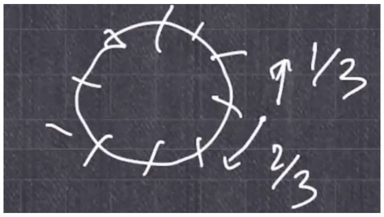
Why is this called reversible? Start a Markov chain from the invariant measure π and then run it for 1 step. Reversibility says that $(X_0, X_1) \overset{d}{=} (X_1, X_0)$.

Example 1.4. Random walks on graphs satisfy detailed balance. To check the equation, we want

$$\deg(x)p(x,y) = \deg(y)p(y,x)$$

Note that $P(x,y) = 0$ iff $P(y,x) = 0$ because $x \sim y$ iff $y \sim x$. If these are not zero, both sides are $\deg x \deg y$ (multiplied by the same constant factor).

Example 1.5. Here is an irreversible chain. Take a circle with n points labeled on it, and travel clockwise with probability 2/3 and counterclockwise with probability 1/3.



The uniform distribution is stationary, but π does not satisfy detailed balance.

There are extra tools we can use to study reversible chains.

Here is a very important example.

Example 1.6 (Coupon collector)**.** We have n types of coupons, and somebody wants to collect all n types. At every discrete time, they independently collect a random coupon. Then we can form the Markov chain X_i being the number of distinct coupon types collected by time i.

Suppose $X_i = k$. The probability of collecting a new coupon is $1 - \frac{k}{n}$. The waiting time between new coupons has Geometric$(1 - \frac{k}{n})$ distribution. The average waiting times are $\frac{n}{n-k}$, so the total expected time to collect all the coupons is $\sum_{k=1}^{n} \frac{n}{n-k} \approx n \log n$.

1.2 Mixing times of Markov chains

1.2.1 Measuring distance to stationarity

From now on, we will only discuss irreducible and aperiodic Markov chains.

For a finite Markov chain, starting from any state, the distribution at time t "converges" to the stationary measure π as $t \to \infty$. We have two considerations

1. Define the appropriate notion of convergence.

2. Find the rate of convergence.

We will use the total variation distance.

Definition 1.5. The **total variation distance** between two measure in a common space Ω is

$$d_{\mathrm{TV}}(\mu,\nu) = \sup_{A \subseteq \Omega} \mu(A) - \nu(A).$$

It turns out this is a metric. Here are two equivalent definitions:

1. L^1 distance:

   $$\begin{aligned} d_{\mathrm{TV}}(\mu,\nu) &= \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \nu(x)| \\ &= \sum_{x\in\Omega} \left| \left(\frac{\mu(x)}{\nu(x)} - 1 \right) \right| \nu(x), \end{aligned}$$

   so we can think of the distance as an L^1 distance or an average of how the density of μ with respect to ν differs from 1.

2. Couplings:

   Definition 1.6. A **coupling** of μ and ν is a measure γ on $\Omega \times \Omega$ such that if $(X,Y) \sim \gamma$, then $X \sim \mu$ and $Y \sim \nu$.

   $$d_{\mathrm{TV}}(\mu,\nu) = \inf_{\text{coupling } (X,Y) \text{ of } \mu,\nu} \mathbb{P}(X \neq Y).$$

For a Markov chain, let

$$d(t) = \sup_{x\in\Omega} d_{\mathrm{TV}}(P^t(x,\cdot),\pi).$$

Convergence to equilibrium can mean making $d(t)$ small.

We can also define

$$\overline{d}(t) = \sup_{x,y\in\Omega} d_{\mathrm{TV}}(P^t(x,\cdot),P^t(y,\cdot)).$$

These are comparable:

$$d(t) \le \overline{d}(t) \le 2d(t).$$

The advantage of $\overline{d}$ is that it is submultiplicative:

$$\overline{d}(t+s) \le \overline{d}(t)\overline{d}(s).$$

Definition 1.7. The **mixing time** is

$$t_{\mathrm{mix}}(\varepsilon) = \inf\{t : d(t) \le \varepsilon\}.$$

We denote $t_{\mathrm{mix}} = t_{\mathrm{mix}}(1/4)$.

1.2.2 Upper bounding mixing times

Here is a general method to bound t_{mix}: Construct a Markovian coupling

$$\begin{aligned} x &= X_0, X_1, \dots \\ y &= Y_0, Y_1, \dots \end{aligned}$$

and let the time to coalescence be $\tau_{\text{coal}} = \inf\{t : X_t = Y_t\}$. Given this, we can get a bound on $\overline{d}$:

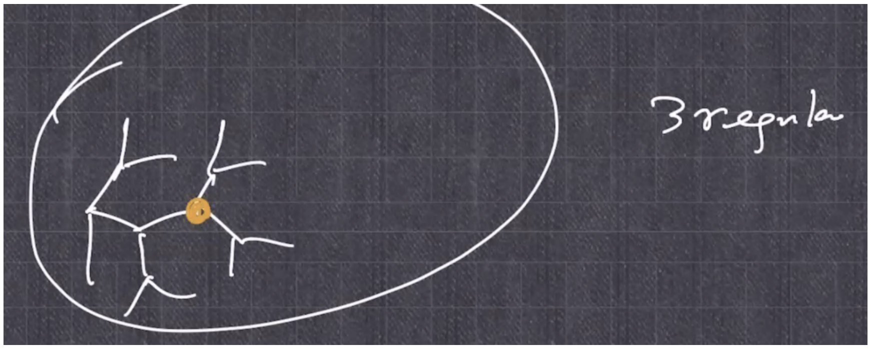
$$\overline{d}(t) \leq \max_{x,y} \mathbb{P}(\tau_{\text{coal}} > t).$$

Example 1.7 (Random walk on the hypercube)**.** Start with state space $\{0,1\}^n$. If we are at $x \in \Omega$, pick a random index and replace it by a new Ber(1/2) random variable. We can use the coupon collecting problem to bound the coalescence time of a Markovian coupling for the random walk on the hypercube. This is a strategy to upper bound t_{mix}.

1.2.3 Lower bounding mixing times

Here are approaches to lower bound the mixing time:

1. For small times, you don't reach enough sites: Suppose we have a 3 regular graph

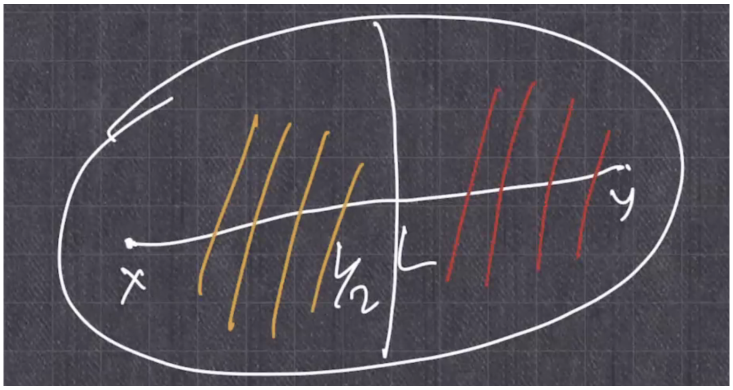


In time t, how many sites does the random walk have access to? This is essentially the volume of a t-ball around the starting point, which is $\approx 2^t$. The stationary measure is uniform on a regular graph. Then for some set A,

$$\pi(A) - \underbrace{P^t(x, A)}_{=0} \geq \frac{n - 2^t}{n} - 0 = 1 - \frac{2^t}{n}.$$

This vaguely indicates that the graph should mix in $\log n$ time.

2. Diameter lower bound: Suppose $d(x, y) = L$, which is the diameter of the chain (where we define diameter in terms of the natural graph structure of the Markov chain). The Markov chain jumps by distance at most 1 in 1 step. If you run the Markov chain for $L/2$ steps, then the support of the two measures are disjoint



So

$$\overline{d}(L/2 - 1) = 1,$$

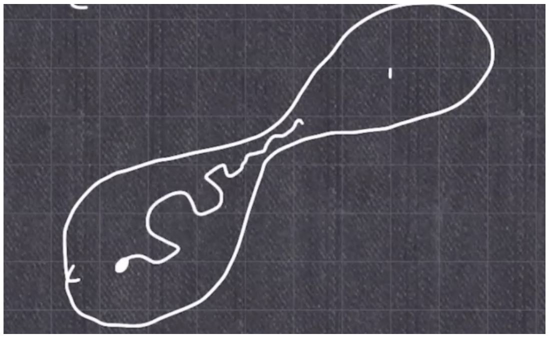
which tells us that

$$d(L/2 - 1) \geq 1/2.$$

This gives a lower bound on the mixing time:

$$t_{\text{mix}} = t_{\text{mix}}(1/4) \geq \frac{L}{2} - 1.$$

3. Bottlenecks: Suppose the Markov chain looks like this.



Then the Markov chain should take a long time to mix across the bottleneck. We can quantify the expansion of a set $S \subseteq \Omega$ by

$$\theta(S, S^c) = \sum_{x \in S, y \in S^c} \pi(x) P(x, y),$$

the probability that if we start in S, we go to S' in one step. Then we can define

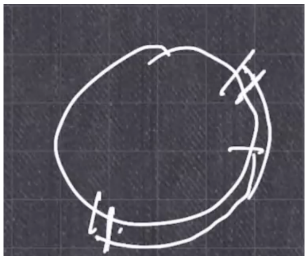
$$\phi(S) = \frac{\theta(S, S^c)}{\pi(S)},$$

$$\phi_* \inf_{\pi(S) \leq 1/2} \phi(S).$$

ϕ_* measures the expansion of the chain. If ϕ_* is small, then there is a bottleneck. In particular,

$$t_{\text{mix}}(1/4) \geq \frac{1}{4\phi_*}.$$

Example 1.8. For a random walk on an n-cycle, $\phi_* \approx 1/n$. The worst case set is a half-circle, which gives $Q(S, S^c) = \frac{2}{n}$.



So $t_{\text{mix}} \geq n$.

But this is not tight. If we start moving on the cycle, we need to move $n/4$ in one direction to get to the other side. This is a gambler's ruin problem, and we can calculate that the time to do this is $\sim n^2$.

4. Projection: Construct a function $\Omega \to \mathbb{R}$ such that $\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f] > r\sigma$, where $\max(\text{Var}_\mu(f), \text{Var}_\nu(f)) = \sigma^2$. Then

   $$d_{\text{TV}}(\mu, \nu) \geq 1 - \frac{8}{\gamma^2}.$$

   There is a slightly more complicated version of this, due to WIlson, which gives very sharp lower bounds.