NICE SUMMER SCHOOL ON RANDOM WALKS AND COMPLEX NETWORKS: PART 1 OF THE COURSE ON RANDOM WALKS AND MARKOV CHAINS

LOUIGI ADDARIO-BERRY, ANNA BEN HAMOU, AND PERLA SOUSI

Abstract. Sections 1 and 2 consist of background material for the course on random walks and Markov chains. Participants should be comfortable with this material, including the exercises (perhaps with the exception of those marked "harder") at the start of the course. Section 3 contains the material covered in lectures 1-5 of the course.

1. Definitions, hitting times, total variation distance, reversibility.

The first chapter of these notes should be familiar to anyone who has taken a first course on Markov chains. We will briefly run through some of the basic definitions and results that we will need later on. For simplicity of notation we take $\mathbb{N} = \{0, 1, 2, \ldots\}$.

1.1. Stochastic processes, Markov processes, Markov chains: definitions.

Given a measurable space $(Y, \mathcal{G})$ a Y-valued random variable is a $(\mathcal{F}/\mathcal{G})$-measurable function $X : \Omega \to Y$ from some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to Y.

For example: a standard Gaussian is an $\mathbb{R}$-valued random variable; if $U_1, \ldots, U_n$ are $\mathbb{R}$-valued random variables defined on a common probability space then $(U_1, \ldots, U_n)$ is an $\mathbb{R}^n$-valued random variable; the random graph $G_{n,p}$ is a random variable taking values in the set of graphs with vertex set labeled by $[n] := \{1, \ldots, n\}$ (in the set of graphs "on $[n]$" for short).

A *Y-valued stochastic process* is a collection of Y-valued random variables $(X_i, i \in I)$ with I some index set defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In other words, for each $i \in I$, $X_i : \Omega \to Y$ is a $(\mathcal{F}/\mathcal{G})$-measurable map.

Example 1. *Here are some basic examples of stochastic processes.*

- *A sequence of independent real random variables $(X_n, n \geq 0)$ is an $\mathbb{R}$-valued stochastic process indexed by $\mathbb{N} := \{0, 1, 2, \ldots\}$.*
- *A simple random walk $(S_n, n \geq 0)$ is an $\mathbb{Z}$-valued stochastic process indexed by $\mathbb{N}$.*
- *The random graph $G_{n,p}$ may be viewed a single random object. It may also be identified with a $\{0, 1\}$-valued stochastic process $(1_{n,p}(e), e \in E(K_n))$ indexed by the edges of the complete graph K_n; here $1_{n,p}(e) = 1$ if e is an edge of $G_{n,p}$ and $1_{n,p}(e) = 0$ otherwise.*

Date: July 11, 2019.

- *It is common to couple the random graphs $G_{n,p}$ as follows. Let $(U_e, e \in E(K_n))$ be independent Uniform$[0,1]$ random variables. Then, for $p \in [0,1]$, let $G_{n,p}$ be the graph on $[n]$ with edge set $\{e \in E(K_n) : U_e \leq p\}$. Then $(G_{n,p}, p \in [0,1])$ is a stochastic process indexed by $[0,1]$ and taking values in the set of graphs on $[n]$.*

If I is a discrete set then we say the process is a *discrete-time process*. We call I the *domain of definition* of the process, and call $(Y, \mathcal{G})$ its *state space*; we will sometimes abuse notation and refer to Y as the state space.

In these notes, we will usually have $I = \mathbb{N}$, $I = \mathbb{Z}$, or else $I \subset \mathbb{R}$ some finite or infinite interval. In these cases, the *filtration generated by X* is the increasing sequence of σ-algebras $(\mathcal{F}_i, i \in I)$, where $\mathcal{F}_i = \sigma(X_j, j \in I, j \leq i)$. Informally, $\mathcal{F}_i$ contains "all information about the process up to time i. For $i \in I$ we also let $\mathcal{F}_{\geq i} = \sigma(X_j, j \in I, j \geq i)$.

A discrete time stochastic process X with state space $(Y, \mathcal{G})$ is a *Markov process* if for all $i, j \in I$ with $j < i$, and all $B \in \mathcal{G}$,

$$\mathbf{P}\{X_i \in B | \mathcal{F}_j\} \stackrel{\text{a.s.}}{=} \mathbf{P}\{X_i \in B | X_j\},$$

and this property is called the *Markov property*. Informally, the Markov property states that conditional on X_j, the future (relative to time j) is independent of the past.

We say a Markov process $(X_i, i \in I)$ with state space $(Y, \mathcal{G})$ is *time-homogeneous* if for all $B \in \mathcal{G}$ and $i, j \in I$,

$$\mathbf{P}\{X_i \in B | X_j\} \stackrel{\text{a.s.}}{=} \mathbf{P}\{X_{i+t} \in B | X_{j+t}\},$$

for all t for which $i + t \in I$ and $j + t \in I$. *In these notes, Markov processes are time-homogeneous by default.*

A *Markov chain* is a (time-homogeneous) Markov process $(X_i, i \in \mathbb{N})$ with finite or countable state space. These notes are almost exclusively concerned with Markov chains; in this case the Markov property simplifies to the statement that for any $i \geq 0$ and any sequence $v_0, \ldots, v_{i+1}$ of elements of V,

$$\mathbf{P}\{X_{i+1} = v_{i+1} | X_0 = v_0, \ldots, X_i = v_i\} = \mathbf{P}\{X_1 = v_{i+1} | X_0 = v_i\}$$

whenever both conditionings are non-degenerate.

Let $X = (X_i, i \in \mathbb{N})$ be a Markov chain with state space V. We can see from above that the distribution of X is completely specified by two pieces of information: the distribution of X_0, which we call the *initial distribution* of X; and the *transition matrix*

$$P = P(X) = (p_{u,v})_{u,v \in V},$$

where $p_{u,v} = \mathbf{P}\{X_1 = v | X_0 = u\}$. If a Markov chain has transition matrix P and initial distribution λ, then we say that it is Markov(λ, P). We sometimes write $\mathbb{P}_{\lambda,P}(\cdot)$ for the probability measure associated to a chain with initial distribution λ and transition matrix P; we will also write $\mathbb{P}_\lambda(\cdot)$ or $\mathbb{P}(\cdot)$ when the transition matrix P and/or initial distribution can be gleaned from context or do not need to be explicitly described. Also, if the initial distribution is a Dirac measure δ_v at v, we abuse notation and write $\mathbb{P}_v(\cdot)$ instead of $\mathbb{P}_{\delta_v}(\cdot)$.

1.2. **Stopping times, the strong Markov property.** Let $X = (X_i, i \in I)$ be a Markov process, with associated filtration $(\mathcal{F}_i, i \in I)$. A random variable T taking values in I is called a *stopping time* for X if for all $i \in I$, the event that $T \leq i$ is measurable with respect to $\mathcal{F}_i$. The idea is that if we are told to stop when T occurs then by watching the Markov chain evolve we will know when to stop. For example, the first day in June that it rains is a "real-world example" of a stopping time, whereas the last day in June that it rains is not.

We write $\mathcal{F}_T$ for the *stopped σ-algebra*, defined as

$$\mathcal{F}_T = \{E \in \mathcal{F} : \forall\, i \in I,\ E \cap \{T \leq i\} \in \mathcal{F}_i\}.$$

Exercise 1.1 (Strong Markov property)**.** *Let $X = (X_i, i \in \mathbb{N})$ be a Markov chain with state-space $(Y, \mathcal{G})$. Fix a stopping time T for X with $T < \infty$ almost surely. Then for all $i \in \mathbb{N}$ and all $B \in \mathcal{G}$,*

$$\mathbf{P}\left\{X_{T+i} \in B \mid \mathcal{F}_T\right\} = \mathbf{P}_{X_T}\left\{X_i \in B\right\}\ .$$

One of the most basic and important special classes of of stopping times are *hitting times*. For a discrete chain $X = (X_i, i \in \mathbb{N})$ with state space V, and $A \subset V$, we write

$$H^A = \inf\{i \in \mathbb{N} : X_i \in A\}, \quad H^A_{>0} = \inf\{i > 0 : X_i \in A\}\,. \tag{1}$$

with $H^A = \infty$ if the chain never visits A. (You should perhaps verify that H^A is a stopping time.) If A consists of a single state, $A = \{a\}$, we often abuse notation and write H^a in place of $H^{\{a\}}$.

Exercises. In the below exercises, $X = (X_n, n \in \mathbb{N})$ is a Markov chain with state space V and transition matrix P. Given a vector $\lambda = (\lambda(v) : v \in V)$ we write $\|\lambda\|_1 = \sum_{v \in V} \lambda(v)$. We say that λ is *invariant* for P if $\lambda P = \lambda$.

Exercise 1.2. *We say that a transition matrix P is* irreducible *if for all $u, v \in V$ there is $n \in \mathbb{N}$ such that the u, v entry of P^n is non-zero. In this exercise assume P is irreducible. For $x, y \in V$ write $\nu_x(y) = \mathbf{E}_x\left\{\#\{i < H^x_{>0} : X_i = y\}\right\}$.*

(i) *Show that $\|\nu_x\|_1 = \mathbf{E}_x\left\{H^x_{>0}\right\}$.*

(ii) *Show that the vector $\nu_x = (\nu_x(v) : v \in V)$ is invariant for P and that $\nu_x(x) = 1$.*

(iii) *Show that if $\lambda = (\lambda(y) : y \in V)$ is any invariant vector with $\lambda(x) = 1$ then $\lambda \geq \nu_x$.*

(iv) *Show that if there exists π invariant for P with $\|\pi\|_1 < \infty$ then $\mathbf{E}_x\left\{H^x_{>0}\right\} < \infty$ for all x, and so $\mathbf{P}_x\left\{H^y < \infty\right\} = 1$ for all x, y.*

Exercise 1.3. *For $v \in V$ write $N_v = \#\{i \in \mathbb{N} : X_i = v\}$ and $h^v_v = \mathbf{P}_v(H^v_{>0} < \infty)$. Then for all $k \geq 0$, $\mathbf{P}_v(N_v > k) = (h^v_v)^k$, and so*

$$\mathbb{E}_v(N_v) = \frac{1}{1 - h^v_v}.$$

1.3. **Total variation distance and coupling random variables.** Given that X and Y are two random elements of some (countable) set V, we define the *total variation distance between X and Y* to be

$$d_{\mathrm{TV}}(X,Y) := \sup_{A\subset V} |\mathbf{P}(X\in A) - \mathbf{P}(Y\in A)|.$$

By considering the set $B = \{v \in V : \mathbf{P}(X = v) \geq \mathbf{P}(Y = v)\}$, one can obtain the following equivalent formulation.

Proposition 2.

$$d_{\mathrm{TV}}(X,Y) = \frac{1}{2}\sum_{v\in V} |\mathbf{P}(X=v) - \mathbf{P}(Y=v)| = \sum_{v\in B}(\mathbf{P}(X=v) - \mathbf{P}(Y=v)).$$

Proof. Let B be the subset of all $v \in V$ for which $\mathbf{P}(X = v) \geq \mathbf{P}(Y = v)$. Then for any set $C \subset B$, $\mathbf{P}(X \in C) - \mathbf{P}(Y \in C) \geq 0$, and this inequality is reversed if $C \subset B^c$. Thus, for any $A \subset V$,

$$\mathbf{P}(X\in A)-\mathbf{P}(Y\in A) \leq \mathbf{P}(X\in A\cap B)-\mathbf{P}(Y\in A\cap B) \leq \mathbf{P}(X\in B)-\mathbf{P}(Y\in B),$$

and likewise

$$\mathbf{P}(Y\in A)-\mathbf{P}(X\in A) \leq \mathbf{P}(Y\in A\cap B^c)-\mathbf{P}(X\in A\cap B^c) \leq \mathbf{P}(Y\in B^c)-\mathbf{P}(X\in B^c).$$

But $0 \leq \mathbf{P}(X \in B) - \mathbf{P}(Y \in B) = \mathbf{P}(Y \in B^c) - \mathbf{P}(X \in B^c)$ so

$$\begin{aligned}
|\mathbf{P}(X\in A) - \mathbf{P}(Y\in A)| &\leq \mathbf{P}(X\in B) - \mathbf{P}(Y\in B)\\
&= \frac{1}{2}(\mathbf{P}(X\in B) - \mathbf{P}(Y\in B)) + \frac{1}{2}(\mathbf{P}(Y\in B^c) - \mathbf{P}(X\in B^c))\\
&= \frac{1}{2}\sum_{v\in V} |\mathbf{P}(X=v) - \mathbf{P}(Y=v)|.
\end{aligned}$$

Since A was arbitrary it follows that

$$d_{\mathrm{TV}}(X,Y) \leq \frac{1}{2}\sum_{v\in V} |\mathbf{P}(X=v) - \mathbf{P}(Y=v)|.$$

But we can achieve this bound by simply taking $A = B$, and so we in fact have equality. □

If μ is the distribution of X and ν is the distribution of Y, then we may equivalently write

$$d_{\mathrm{TV}}(X,Y) = \frac{1}{2}\sum_{v\in V} |\mu(v) - \nu(v)| = \sup_{A\subset V} |\mu(A) - \nu(A)| =: \|\mu - \nu\|_{\mathrm{TV}},$$

so $d_{\mathrm{TV}}(X,Y)$ and $\|\mu - \nu\|_{\mathrm{TV}}$ are two pieces of notation for essentially the same thing. We will also call this quantity the total variation distance between μ and ν, and, in general, will view the total variation distance as relating to either the variables X, Y or to their distributions, whichever happens to be more convenient in context.

Exercise 1.4. *Prove that* $\| \cdot - \cdot \|_{\mathrm{TV}}$ *is a metric on the set of all measures* μ *on* V *with* $\mu(V) < \infty$.

We have seen two different formulas for $\|\mu - \nu\|_{\mathrm{TV}}$. We now introduce a third, and to do so we must introduce the (fundamental) notion of a *coupling* between two random variables. Consider distributions $\mu : V \to \mathbb{R}$ and $\nu : V \to \mathbb{R}$. A *coupling* of μ and ν is a random variable (X, Y) taking values in $V \times V$, such that X has distribution μ and Y has distribution ν. In other words, *any way of defining* X *and* Y *on a common probability space* yields a coupling between μ and ν. We will also refer to the *distribution* $q : V \times V \to \mathbb{R}$ of the pair (X, Y) as the coupling (again, choosing whether to speak of (X, Y) or of its distribution depending on context).

For example, suppose μ is the distribution of of a Bernoulli(p) random variable, so $\mu(1) = p$ and $\mu(0) = 1 - p$, and suppose ν is also this distribution. Here are two valid couplings of μ and ν. First, we could take $q(1, 1) = p$, $q(0, 0) = 1 - p$ and $q(1, 0) = q(0, 1) = 0$. This corresponds to letting X be Bernoulli(p) and letting $Y = X$. Second, we could take $q(1, 1) = p^2$, $q(0, 0) = (1 - p)^2$, and $q(1, 0) = q(0, 1) = p(1 - p)$. This corresponds to letting X and Y be *independent* with distributions Bernoulli(p). The first of these couplings captures the fact that X and Y have the same distribution by making them always identical. This hints at a relation between total variation distance and couplings, a relation that is made explicit by the following proposition.

Proposition 3. *If* $\mu : V \to \mathbb{R}$ *and* $\nu : V \to \mathbb{R}$ *are two distributions then*

$$\|\mu - \nu\|_{\mathrm{TV}} = \inf\{\mathbf{P}(X \neq Y) : (X, Y) \text{ a coupling between } \mu \text{ and } \nu\}.$$

Exercise 1.5. *Let* X *be Bernoulli*(p) *and* Y *be Bernoulli*(p'). *Find* $d_{\mathrm{TV}}(X, Y)$ *and construct a coupling such that* $d_{\mathrm{TV}}(X, Y) = \mathbf{P}(X \neq Y)$.

Exercise 1.6. *Prove Proposition 3.*

1.4. **Invariant distributions, time reversal.** In this section we suppose that P is an irreducible transition matrix. Recall from above that a measure λ is *invariant* for P if $\lambda P = \lambda$. Using ideas similar to those developed in Exercise 1.2, it is not too difficult to show that if λ and μ are two invariant measures for P then $\lambda = c\mu$ for some $c \geq 0$. If $\|\lambda\|_1 < \infty$, it then follows that there is an unique vector π with $\pi P = \pi$ and $\|\pi\|_1 = 1$; π is called the *invariant* or *stationary distribution* of the chain.

A Markov chain with stationary distribution π and transition matrix P is called *reversible* if for all $u, v \in V$ we have $\pi(u)p_{u,v} = \pi(v)p_{v,u}$. This is called reversibility because it is equivalent to saying that for all $u, v \in V$,

$$\mathbf{P}_\pi \{X_0 = u, X_1 = v\} = \mathbf{P}_\pi \{X_1 = u, X_0 = v\} \ ,$$

or, more informally, that when the chain is in stationarity we cannot tell whether it is running forwards or backwards. One important special class of reversible Markov chains is *simple random walk on a graph*, defined as follows. Given a graph $G = (V, E)$,

for $v \in V$ write $\deg(v) = \#\{w \in V : vw \in E\}$. Then the transition matrix P of the simple random walk on G has entries given by

$$p_{vw} = \begin{cases} \frac{1}{\deg(v)} & \text{if } vw \in E \\ 0 & \text{otherwise.} \end{cases}$$

Exercise 1.7. (i) *Given a chain with transition matrix P, show that if π is a distribution such that $\pi(v)p_{vw} = \pi(w)p_{wv}$ for all v, w, then the chain is reversible and π is the stationary distribution.*

(ii) *Show that the stationary distribution for simple random walk on a finite connected graph $G = (V, E)$ is given by $\pi(v) = \deg(v)/2|E|$ for all $v \in V$.*

(iii) *Suppose that we are given a finite set V and edge weights $c = \{c_{\{v,w\}} : v, w \in V\}$ such that $c_{\{v,w\}} \geq 0$ for all $v, w \in V$ and $\sum_{x \in V} c_{\{v,x\}} > 0$ for all $v \in V$. Then the* weighted *simple random walk with weights c has a transition matrix with entries given by $p_{v,w} = c_{\{v,w\}} / \sum_{x \in V} c_{\{v,x\}}$. Show that any finite reversible chain can be represented as a weighted simple random walk. (Note: since $\{v, w\}$ is a set we have $c_{\{v,w\}} = c_{\{w,v\}}$.)*

Exercise 1.8. *Fix a reversible Markov chain $(X_n, n \geq 0)$ and v, w in the state space of the chain. Then for any path $Q = (v_0, v_1, \ldots, v_k)$ with $v_0 = v$, $v_k = w$, we have*

$$\begin{aligned} \mathbf{P}_v \{(X_n, 0 \leq n \leq H^w) = Q \mid H^w < H^v_{>0}\} \\ = \mathbf{P}_w \{(X_n, 0 \leq n \leq H^w) = Q^r \mid H^v < H^w_{>0}\} , \end{aligned}$$

where $Q^r = (v_k, \ldots, v_0)$ is the reversal of Q.

2. Convergence to equilibrium and the mixing time

2.1. **Coupling Markov chains.** In this section we recall the concept of coupling two Markov chains, and use it to prove that many chains converge to their equilibrium distribution.

If $(X_n)_{n \geq 0}$ is Markov(λ, P) with state space V, and $(Y_n)_{n \geq 0}$ is Markov(λ', P') with state space V', then a coupling of the two chains is simply a random sequence $(U_n, W_n)_{n \geq 0}$ of elements of $V \times V'$ such that if we only look at the first coordinate $(U_n)_{n \geq 0}$, we see a chain which is Markov(λ, P), and if we only look at the second coordinate $(W_n)_{n \geq 0}$ then we see a chain which is Markov(λ', P'). For example, we can always simply take $(X_n)_{n \geq 0}$ to be Markov(λ, P), independently take $(Y_n)_{n \geq 0}$ to be Markov(λ', P'), and consider the sequence $(X_n, Y_n)_{n \geq 0}$.

Rather than letting the two chains be completely independent, we can instead have one chain *follow the other*. In other words, let $(X_n)_{n \geq 0}$ and $(Y_n)_{n \geq 0}$ be independent as before. Then let $T = \inf\{n \geq 0 : X_n = Y_n = v\}$ be the first time that X_n and Y_n are both at v, and let $U_n = X_n$ for $n < T$ and $U_n = Y_n$ for $n \geq T$. Finally, let $W_n = Y_n$ for all n. In other words, the two chains behave independently until the first time they meet at v, at which point the X-chain sticks to the Y-chain and follows it.

Exercise 2.1. *$(U_n, W_n)_{n\geq 0}$ is a coupling of the X-chain and the Y-chain.*

This last coupling actually gives us a way to control the total variation distance between X_n and Y_n, since once the two chains meet at v they stick together. If we write $\lambda^{(n)} = \lambda P^n$ for the distribution of X_n, and likewise write $\gamma^{(n)} = \gamma P^n$ for the distribution of Y_n, then by Proposition 3,

$$\|\lambda^{(n)} - \gamma^{(n)}\|_{\mathrm{TV}} \leq \mathbf{P}(U_n \neq W_n) = \mathbf{P}(n < T).$$

In particular, we have the following corollary.

Corollary 4. *If $(X_n)_{n\geq 0}$ is Markov(λ, P), $(Y_n)_{n\geq 0}$ is Markov(γ, P), both with state space V, and there is $v \in V$ such that the stopping time $T = \inf\{n \geq 0 : X_n = Y_n = v\}$ satisfies $\mathbf{P}(T < \infty) = 1$, then $d_{\mathrm{TV}}(X_n, Y_n) \to 0$ as $n \to \infty$.*

This follows immediately from the bound just before the corollary, since if $\mathbf{P}(T < \infty) = 1$ then $\mathbf{P}(T > n) \to 0$ as $n \to \infty$.

The next theorem is the first, fundamental result of the theory of Markov chains. In the proof we will use the following basic fact, stated as an exercise. A transition matrix P is *aperiodic* if for all v we have $\gcd\{n \geq 1 : p_{vv}^{(n)} > 0\} = 1$.

Exercise 2.2. *If P is an aperiodic irreducible transition matrix then there exists $n \geq 0$ such that for all u and v, and all $m \geq n$, we have $p_{uv}^{(m)} > 0$.*

Theorem 5 (Fundamental theorem of Markov chains)**.** *If $(X_n)_{n\geq 0}$ is Markov(λ, P) with state space V and P is irreducible and aperiodic, and has invariant distribution π, then $\|\lambda^{(n)} - \pi\|_{\mathrm{TV}} \to 0$ as $n \to \infty$.*

Proof. Let $(Y_n)_{n\geq 0}$ be Markov(π, P), and independent of $(X_n)_{n\geq 0}$. Since π is the invariant distribution, $\pi^{(n)} = \pi P^n = \pi$. Thus, in view of the above corollary, it suffices to show that there is $v \in V$ such that $\mathbf{P}(T < \infty) = 1$.

For this we use the first, "independent" coupling. Recall that $(X_n, Y_n)_{n\geq 0}$ is Markov$(\hat{\lambda}, \hat{P})$, with initial distribution $\hat{\lambda}_{u,v} = \lambda_u \pi(v)$ and transition probabilities $\hat{p}_{(u,v),(x,y)} = p_{ux} p_{vy}$. Fix any two pairs $(u, v), (x, y) \in V \times V$. Since P is aperiodic, for all sufficiently large n, both $p_{ux}^{(n)}$ and $p_{vy}^{(n)}$ are positive, and so

$$\hat{p}_{(u,v),(x,y)}^{(n)} = p_{ux}^{(n)} p_{vy}^{(n)} > 0.$$

In other words, in the paired chain, it is possible to get from anywhere in $V \times V$ to anywhere else, so the chain is irreducible. Next, let $\hat{\pi}$ be defined by $\hat{\pi}_{(u,v)} = \pi(u)\pi(v)$. Then

$$\sum_{(u,v)\in V\times V} \hat{\pi}_{(u,v)} = \sum_{u\in V} \pi(u) \sum_{v\in V} \pi(v) = \sum_{u\in V} \pi(u) \cdot 1 = 1,$$

so we have defined a distribution. And, for all $(u, v) \in V \times V$,

$$\begin{aligned}\sum_{(x,y)\in V\times V} \hat{\pi}_{(x,y)}\hat{p}_{(x,y),(u,v)} &= \sum_{x\in V}\sum_{y\in V} \pi(x)\pi(y)p_{xu}p_{yv} \\ &= (\sum_{x\in V} \pi(x)p_{xu}) \cdot (\sum_{y\in V} \pi(y)p_{yv}) \\ &= \pi(u)\pi(v) \\ &= \hat{\pi}_{(u,v)},\end{aligned}$$

so $\hat{\pi}$ is invariant for $\hat{P}$. Since $\hat{P}$ is also irreducible, it follows by Exercise 1.2 that $(X_n, Y_n)_{n\geq 0}$ is recurrent, so with probability one, any given state of $V \times V$ is eventually visited. In particular, for *any* $v \in V$, $\mathbf{P}(T < \infty) = 1$, which is what we needed to prove. □

The study of Markov chain mixing times considers the rate of convergence in the above theorem. A major aim of these notes is to discuss a technique for studying mixing times that is now well-known to experts but has not yet become fully accessible to non-specialists. We will need to develop some more refined tools in order to present these methods, but we first give some simple bounds that can be proved more easily.

Exercise 2.3. *Prove the following, quantitative version of Theorem 5. Under the conditions of Theorem 5, if the state space is finite then there exist constants $\alpha \in (0, 1)$ and $C > 0$ such that for all n and λ,*

$$\|\lambda^{(n)} - \pi\|_{\mathrm{TV}} \leq C\alpha^n .$$

2.2. **Mixing times: bounding the speed of convergence.** Next, fix an irreducible Markov chain $X = (X_n, n \geq 0)$ with state space V and stationary distribution π, and for $n \geq 0$ write

$$d(n) = \max_{v\in V} \|\mathbf{P}_v \{X_n \in \cdot\} - \pi\|_{\mathrm{TV}};$$

in words, $d(n)$ is the *worst case total variation distance from stationarity at time n*. For $\epsilon > 0$ we define

$$t_{\mathrm{MIX}}(\epsilon) = \min\{n \geq 0 : d(n) \leq \epsilon\} .$$

And call $t_{\mathrm{MIX}}(\epsilon)$ the ϵ-mixing time of the chain. Exercise 2.3 implies that the precise choice of ϵ is relatively unimportant; it is relatively standard to write $t_{\mathrm{MIX}} = t_{\mathrm{MIX}}(1/4)$ and call t_{MIX} "the" (total variation) mixing time of the chain.

Exercise 2.4. *Show that for all n, and any probability distribution λ on V,*

$$\|\mathbf{P}_\lambda \{X_n \in \cdot\} - \pi\|_{\mathrm{TV}} \leq d(n)$$

By Exercise 2.3, if X has finite state space and is irreducible and aperiodic then $d(n) \to 0$ exponentially quickly as $n \to \infty$. The subject of mixing times is in large part concerned with proving more precise bounds on the manner in which $d(n)$

tends to zero. When bounding $d(n)$, it is often useful to consider the following, related quantity: let

$$\overline{d}(n) = \max_{u,v \in V} \|\mathbf{P}_v \{X_n \in \cdot\} - \mathbf{P}_u \{X_n \in \cdot\}\|_{\mathrm{TV}} . \tag{2}$$

It is immediate by the triangle inequality for the total variation distance that $\overline{d}(n) \leq 2d(n)$.

Exercise 2.5. (i) *Show that*

$$\|\mathbf{P}_v \{X_n \in \cdot\} - \pi\|_{\mathrm{TV}} = \max_{A \subset V} \left| \sum_{w \in V} \pi(w) \left(\mathbf{P}_v \{X_n \in A\} - \mathbf{P}_w \{X_n \in A\}\right) \right|$$

(ii) *Use (i), the triangle inequality and convexity to establish that for all n, $d(n) \leq \overline{d}(n)$.*

Exercise 2.6. (i) *(Levin-Peres-Wilmer, Exercise 4.3) Show that for any two distributions μ, ν on V, we have $\|\mu P - \nu P\|_{\mathrm{TV}} \leq \|\mu - \nu\|_{\mathrm{TV}}$.*

(ii) *Show that both $d(\cdot)$ and $\overline{d}(\cdot)$ are non-increasing.*

(iii) *(Harder.) Show that for all $n, m \in \mathbb{N}$, $\overline{d}(m+n) \leq \overline{d}(m)\overline{d}(n)$, but that this property need not hold for $d(\cdot)$.*

Exercise 2.7 (Strong stationary times, harder.)**.**
A stopping time τ is called a strong stationary time *if for all $v, w \in V$, and $n \geq 0$,*

$$\mathbf{P}_v \{\tau = n, X_\tau = w\} = \mathbf{P}_v \{\tau = n\} \cdot \pi(w) .$$

or in other words if X_τ has distribution π and is independent of τ.

(i) *Fix $v \in V$. Show that if τ is a strong stationary time then for all $n \geq 0$,*

$$d(n) \leq \max_{v \in V} \mathbf{P}_v \{\tau > n\} ,$$

and so $t_{\mathrm{MIX}}(1/4) \leq 4 \max_{v \in V} \mathbf{E}_v \{\tau\}$.

(ii) *Show that strong stationary times always exist for aperiodic irreducible finite chains.*

3. Lectures 1-5 (in 4 parts)

3.1. Various notions of mixing, stationary and strong stationary times, the Green's function identity. A) Basics.

Our Markov chains usually will be denoted $\mathrm{X} = (X_t, t \geq 0)$; in my part of the course the Markov chains are all discrete, so $t \in \mathbb{N} = \{0, 1, 2, \ldots\}$. In later parts of the course continuous time Markov chains will also be considered but not for now. The state space is always assumed finite.

The transition matrix is P, so P^t describes time-t transition probabilities; $P^t(x, y) = \mathbf{P}\{X_t = y \mid X_0 = x\}$.

The Markov chain is *irreducible* if for all u, v there exists t such that $P^t(u, v) > 0$, and is *aperiodic* if for all v we have $\gcd\{t \in \mathbb{N} : P^t(v, v) > 0\} = 1$.

The *ergodic theorem*, or *fundamental theorem of Markov chains*, says that if X (or P) is irreducible and aperiodic then there exists a unique probability distribution π such that $\pi P = \pi$ (a *stationary distribution*) and, moreover, for any states u and v,

$$P^t(u, v) \to \pi(v)$$

as $t \to \infty$. The theory of mixing times itself is about *quantifying the ergodic theorem*; understanding the rate at which convergence occurs, under various definitions of "rate".

B) Mixing and sampling.

One of the initial motivations for studying mixing times comes from theoretical CS. Let $A = (a_{ij})_{i,j=1}^n$ be a $\{0, 1\}$ matrix.

- *Determinant of A:*

$$\sum_{\sigma \text{ permutation}} \text{sign}(\sigma) \prod_{i=1}^{n} a_{i\sigma(i)};$$

  computable in polynomial time.
- *Permanent of A:*

$$\sum_{\sigma \text{ permutation}} \prod_{i=1}^{n} a_{i\sigma(i)};$$

  $\#P$-complete (as hard as counting any NP structures).

  Jerrum and Sinclair (1989): If A is *dense* has at least $n/2$ ones in any row then the permanent of A can be approximated to within a multiplicative $1 + o(1)$ factor in polynomial time.

  Technique: reduce to approximately sampling from the stationary distribution in an appropriately defined Markov chain.
- *Spanning trees.* Suppose A is the adjacency matrix of graph G, $a_{ij} = \mathbf{1}_{ij \in E}$. If A is dense then the number of spanning trees is super-exponentially large in n. The natural Markov chain on G, namely a random walk, gives a way to do without exhaustively generating all trees.
- *The web graph.* Basically the only way to investigate its properties is through following links and seeing where they take us. (You may not do this yourself but this is how Google decides what to show you.)

C) Various definitions of mixing. We said mixing time was about quantifying the convergence to stationarity in the ergodic theorem; the question is how to measure distance to stationarity. Given the way that we stated the ergodic theorem, a natural distance would seem to be

$$\widehat{\text{dist}}_\infty(P^t(u, \cdot), \pi(\cdot)) := \sup_v |P^t(u, v) - \pi(v)|.$$

This is not the best definition. To see why, think chain with n states (where n is large) where the stationary distribution is uniform, $\pi(v) = 1/n$ for all v. then for $\text{dist}^{(0)}$ to be small it suffices that $P^n(u, v)$ is small for all u and v - but this doesn't mean that the distribution of X_n is close to π. The next exercise is one of many possible exercises which exhibits the fact that this is not the best definition.

Exercise 3.1. *There exists $C > 0$ such that for all $n \geq 1$ the following holds. Let P be lazy simple random walk on a cycle of length n, so $P(i, i+1) = P(i, i-1) = 1/4$ and $P(i, i) = 1/2$ for $1 \leq i \leq n$; here $i+1$ and $i-1$ should be understood* $\bmod n$. *Then*

(a) $\mathrm{dist}_1(P^{\lfloor n/4 \rfloor}, \pi) \leq C/n^{1/2}$.
(b) *For all $1 \leq i \leq n$ there is a set $S \subset \{1, \ldots, n\}$ with $|S| \geq n/2$ such that $P^{\lfloor n/4 \rfloor}(i, S) = 0$.*

So how should we measure the distance to stationarity? Here are some options.

- Using the $\ell_p(\pi)$-norms

$$\|\mu\|_p := \left(\sum_{x \in \Omega} |\mu(x)|^p \pi(x) \right)^{1/p} = \left(\mathbb{E}_{X \sim \pi}(\mu(x)^p) \right)^{1/p} .$$

It's natural to use the stationary measure π as the reference measure since that's what we're measuring distance to. The corresponding distance to stationarity is

$$\widehat{\mathrm{dist}}_p(\mu, \nu) = \left(\sum_{v \in V} \pi(v) |\mu(v) - \nu(v)|^p \right)^{1/p} .$$

The next exercise makes the link with the above notation $\widehat{\mathrm{dist}}_\infty$.

Exercise 3.2. *For any probability measures μ, ν on V,*

$$\lim_{p \to \infty} \widehat{\mathrm{dist}}_p(\mu, \nu) = \sup_{v \in V} |\mu(v) - \nu(v)|$$

- The above distances still have the problem that $\widehat{\mathrm{dist}}_p(\mu, \nu)$ is small as soon as $\sup_v \max(\mu(v), \nu(v))$ is small. To deal with this it is reasonably natural to normalize by π, defining

$$\mathrm{dist}_p(\mu, \nu) = \left(\sum_{v \in V} \pi(v) \left(\frac{|\mu(v) - \nu(v)|}{\pi(v)} \right)^p \right)^{1/p} .$$

When $p = 1$, this gives the *total variation distance*

$$\mathrm{dist}_{\mathrm{TV}}(\mu, \pi) := \frac{1}{2} \mathrm{dist}_1(\mu, \pi) = \frac{1}{2} \sum_{v \in V} \pi(v) \left(\frac{|\mu(v) - \pi(v)|}{\pi(v)} \right) .$$

Total variation distance is very commonly used for measuring distance to stationarity, due in large part to the link with couplings given in Proposition 3. Recall that the *ϵ-total variation mixing time* is defined as

$$t_{\mathrm{mix}}(\epsilon) = \min\{t \geq 0 : d(t) \leq \epsilon\},$$

where

$$\begin{aligned} d(t) &= \max_u \operatorname{dist}_{\mathrm{TV}}(P^t(u,\cdot),\pi(\cdot)) \\ &= \max_u \frac{1}{2}\sum_v |\mathbf{P}_u\{X_t = v\} - \pi(v)| \\ &= \max_u \sum_{\{v:\mathbf{P}_u\{X_t=v\}>\pi(v)\}} (\mathbf{P}_u\{X_t = v\} - \pi(v)). \end{aligned}$$

- You may have seen an ergodic theorem summarized as "time averages equal space averages", which says that the long-run proportion of time spent at state v is given by $\pi(v)$ for all v. A corresponding notion of convergence is given by defining

$$\nu_u^t(v) = \frac{1}{t}\sum_{s=0}^{t-1} P^s(u,v) = \frac{1}{t}\mathbb{E}_u\#\{0 \le s < t : X_s = v\}.$$

The *Cesaro mixing time* is

$$t_{\mathrm{Ces}}(\epsilon) = \min(t : d_{\mathrm{Ces}}(t) \le \epsilon),$$

where $d_{\mathrm{Ces}}(t) = \max_u \operatorname{dist}_{\mathrm{TV}}(\nu_u^t(\cdot),\pi(\cdot))$. The next exercise relates Cesaro mixing to total variation mixing. The following exercise provides another link between time averages and space averages.

Exercise 3.3. *Show that* $t_{\mathrm{Ces}}(1/4) \le 6t_{\mathrm{mix}}(1/8)$ *and that* $t_{\mathrm{Ces}}(1/2^k) \le kt_{\mathrm{Ces}}(1/4)$ *for all* $k \ge 1$.

The second assertion of the exercise is false. To see this, fix $\epsilon \in (0, 1/2]$ and $M \ge 2$ and consider a two-state chain with states u, v and transition probabilities given by $P(u,u) = \epsilon = 1 - P(u,v)$ and $P(v,v) = \epsilon/M = 1 - P(v,u)$. This chain has stationary distribution π given by $\pi(u) = 1/(M+1)$, $\pi(v) = M/(M+1)$. Also, it is lazy so $P^t(u,u)$ is non-increasing and $P^t(v,v)$ is non-increasing.

It can be verified that

$$P^t(u,v) = \frac{M}{M+1}\left(1 - \left(1 - \epsilon\Big(\frac{M+1}{M}\Big)\right)^t\right)$$

and (necessarily)

$$P^t(u,u) = \frac{1}{M+1}\left(1 + M\left(1 - \epsilon\Big(\frac{M+1}{M}\Big)\right)^t\right)$$

For any t we have

$$\left(1 - \epsilon\Big(\frac{M+1}{M}\Big)\right)^t \le \exp\left(-\epsilon t\Big(\frac{M+1}{M}\Big)\right) < e^{-\epsilon t}$$

so if $t \ge 2.1/\epsilon$ then

$$P^t(u,u) - \pi(u) \le e^{-2.1} < 1/8.$$

It is similarly straightforward to see that $P^t(v,v) - \pi(v) < 1/8$. It follows that $t_{\text{mix}}(1/8) \le \lceil 2.1/\epsilon \rceil < 3/\epsilon$, so by Exercise 3.3, $t_{\text{Ces}}(1/4) < 24/\epsilon$.

On the other hand,

$$\left(1 - \epsilon\Big(\frac{M+1}{M}\Big)\right)^t \ge 1 - t\epsilon\frac{M+1}{M}$$

so

$$P^t(u,u) - \pi(u) = P^t(u,u) - \frac{1}{M+1} \ge \frac{M}{M+1} - t\epsilon \ge \frac{2}{3} - t\epsilon.$$

In particular, if $t \le 1/(3\epsilon)$ then $P^t(u,u) - \pi(u) \ge \frac{1}{3}$. We'll assume for simplicity that $1/(3\epsilon)$ is an integer.

If $t = mt_{\text{Ces}}(1/4) \le m \cdot 24/\epsilon$ then

$$\begin{aligned} t\nu_u^t(u) &= \sum_{s=0}^{t} P^s(u,u) \\ &\ge \sum_{s=0}^{1/(3\epsilon)} P^s(u,u) + \sum_{s=1/(3\epsilon)}^{t} P^s(u,u) \\ &\ge \sum_{s=0}^{1/(3\epsilon)} (P^s(u,u) - \pi(u)) + t\pi(u) \\ &\ge t\pi(u) + \frac{1}{9\epsilon} \\ &\ge t\pi(u) + \frac{t}{216m}. \end{aligned}$$

It follows that

$$d_{TV}(\nu_u^{mt_{\text{Ces}}(1/4)}(\cdot), \pi(\cdot)) = \nu_u^{mt_{\text{Ces}}(1/4)}(u) - \pi(u) \ge \frac{1}{216m},$$

which shows that $t_{\text{Ces}}(\delta)$ must in fact grow linearly in $1/\delta$; in particular, if m is large enough that $2^{-m} < 1/(216m)$ then this bound rules out the inequality

$$t_{\text{Ces}}(1/2^k) \le kt_{\text{Ces}}(1/4).$$

Exercise 3.4. *Fix a stopping time τ and write $G_\tau(u,v)$ for the expected number of visits v before time τ, starting from u :*

$$G_\tau(u,v) = \mathbf{E}\#\{0 \le t < \tau : X_t = v\} = \sum_{t \ge 0} \mathbf{P}_u\left\{X_t = v, t < \tau\right\}.$$

(a) **[A Green's function identity.]** *Show that if $\mathbf{P}_u\{X_\tau = u\} = 1$ then for all v,*

$$G_\tau(u,v) = \pi(v) \cdot \mathbf{E}_u\{\tau\},$$

by showing that $G_\tau(u,\cdot)$ is stationary for P.

(b) *Consider a random walk on a connected weighted graph $G = (V, E)$ with edge weights $(c_e, e \in E)$. Write $c_v = \sum_{e \ni v} c_e$ and let $c = \sum_v c_v = 2\sum_e c_e$.*
With $H^v_+ = \min(t \geq 1 : X_t = v)$, show that $\mathbf{E}_v\left\{H^v_+\right\} = c/c_v$.

(c) **[The commute-time identity for non-reversible chains.]** *Prove that for any distinct states u and v*

$$\mathbf{E}_u\left\{\#\{0 \leq t \leq H^v : X_t = u\}\right\} = \pi(u)(\mathbf{E}_u\left\{H^v\right\} + \mathbf{E}_v\left\{H^u\right\}).$$

(d) **[The edge-commute inequality.]** *With the same setup as (b), show that $\mathbf{E}_u\left\{H^v\right\} + \mathbf{E}_v\left\{H^u\right\} \leq c/c_{uv}$. (Hint: Apply (b) to the Markov chain $(Z_t)_{t\geq 0}$ given by $Z_t = (X_t, X_{t+1})$, using the time $\tau = \min(t \geq 1 : X_t = v, X_{t+1=u})$.)*

- The *separation distance* is for those with FOMO (fear of missing out); if the separation distance from stationarity is small then there are no sites which are substantially "under-weighted". Formally,

$$t_{\text{sep}}(\epsilon) = \min(t : d_{\text{sep}}(t) \leq \epsilon),$$

where

$$d_{\text{sep}}(t) = \max_{u,v}\left(1 - \frac{P^t(u,v)}{\pi(v)}\right).$$

The next exercise asks you to show that separation and total variation mixing times are equivalent up to polynomial factors. Recall the definition of $\overline{d}$ from (2).

Exercise 3.5. (a) *Show that $\overline{d}(t) \leq s(t)$ and that $s(2t) \leq 1-(1-\overline{d}(t))^2$.*
(b) *Show that $\overline{d}(t) \leq \exp(-\lfloor t/t_{\text{mix}}(1/e)\rfloor)$.*
(c) *Show that $t_{\text{mix}}(1/e) \leq t_{\text{sep}}(1/e) \leq 4t_{\text{mix}}(1/e)$.*

D) Mixing and stationary times.

We now discuss the connection between stationary times and mixing times. Mixing times are *deterministic* times at which the random walk is guaranteed to *approximate* the invariant distribution. Stationary times are *random* times at which the chain is guaranteed to have *exactly* the stationary distribution. More precisely:

- A stopping time τ is a *stationary* time if for all states v,

$$\mathbf{P}\left\{X_\tau = v\right\} = \pi(v).$$

- A stopping time τ is a *strong stationary* time if for all states v,

$$\mathbf{P}\left\{X_\tau = v, \tau = t\right\} = \pi(v) \cdot \mathbf{P}\left\{\tau = t\right\}.$$

It is easy to see that, for irreducible finite Markov chains, stationary times exist; just let U be a random state with distribution π, and let $\tau = \min(t : X_t = U)$. (Exercise: τ is finite with probability 1.)

Strong stationary times need not exist (consider a random walk on a bipartite graph), but they always exist for aperiodic irreducible chains (see the next exercise).

Exercise 3.6. (a) *Let $t^{(1)} = t_{\text{sep}}(1/e)$. For each state u, construct a stopping time τ_u such that*

$$\mathbf{P}\left\{X_{\tau_u} = v, \tau_u = kt^{(1)} \mid X_0 = u\right\} = e^{-(k-1)}(1-e^{-1})\pi(v).$$

(b) *Use part (a) to show that there exists a strong stationary time τ with*

$$\mathbf{E}[\tau] \leq \frac{t_{\text{sep}(1/e)}}{1-1/e}\,.$$

The next theorem provides a link between stationary times and the Cesaro mixing time.

Theorem 6. *For any stationary time τ and any state u,*

$$\text{dist}_{\text{TV}}(\nu_u^t(v), \pi(v)) \leq \frac{1}{t}\mathbf{E}_u\tau\,.$$

Proof. Note that if τ is stationary then $\tau + s$ is also stationary:

$$\begin{aligned}
\mathbf{P}\{X_{\tau+s} = v\} &= \sum_u \mathbf{P}\{X_\tau = u, X_{\tau+s} = v\} \\
&= \sum_u \mathbf{P}\{X_\tau = u\}\,\mathbf{P}_u\{X_s = v\} \\
&= \sum_u \pi(u)\mathbf{P}_u\{X_s = v\} \\
&= \mathbf{P}_\pi\{X_s = v\} \\
&= \pi(v)\,.
\end{aligned}$$

We can therefore rewrite

$$\begin{aligned}
t\pi(v) &= \sum_{s=0}^{t-1} \mathbf{P}_u\{X_{\tau+s} = v\} \\
&= \sum_{r\geq 0} \mathbf{P}_u\{X_r = v, \tau \leq r < \tau + t\} \\
&\geq \sum_{r=0}^{t-1} \mathbf{P}_u\{X_r = v, \tau \leq r\}\,.
\end{aligned}$$

This gives

$$t\nu_u^t(v) - t\pi(v) \leq \sum_{r=0}^{t-1} \mathbf{P}_u\{X_r = v, \tau > r\}\,,$$

so

$$\begin{aligned} t\mathrm{dist}_{\mathrm{TV}}(\nu_u^t(v), \pi(v)) &= \sum_{\{v:\nu_u^t(v)>\pi(v)\}} (t\nu_u^t(v) - t\pi(v)) \\ &\leq \sum_v \sum_{r=0}^{t-1} \mathbf{P}_u\{X_r = v, \tau > r\} \\ &= \sum_{r=0}^{t-1} \mathbf{P}_u\{\tau > r\} \\ &\leq \mathbf{E}_u\{\tau\}\ . \end{aligned}$$

□

It follows from the above bound that

$$t_{\mathrm{Ces}}(1/4) \leq 4 \max_u \mathbf{E}_u\{\tau\} + 1.$$

We next prove a result similar to the above theorem, for strong stationary times.

Theorem 7. *If τ is a strong stationary time then for any state u,*

$$\mathrm{dist}_{\mathrm{TV}}(P^t(u,\cdot), \pi(\cdot)) \leq \mathbf{P}_u\{\tau > t\}\,.$$

Proof. We have

$$\begin{aligned} P^t(u,v) &= \mathbf{P}_u\{X_t = v\} \\ &= \mathbf{P}_u\{X_t = v, \tau > t\} + \sum_{0\leq s\leq t}\sum_x \mathbf{P}_u\{\tau = s, X_\tau = x, X_t = v\} \\ &\leq \mathbf{P}_u\{X_t = v, \tau > t\} + \sum_{0\leq s\leq t}\sum_x \mathbf{P}_u\{\tau = s\}\,\pi(x)\mathbf{P}_x\{X_{t-s} = v\} \\ &= \mathbf{P}_u\{X_t = v, \tau > t\} + \sum_{0\leq s\leq t} \mathbf{P}_u\{\tau = s\}\,\mathbf{P}_\pi\{X_{t-s} = v\} \\ &= \mathbf{P}_u\{X_t = v, \tau > t\} + \mathbf{P}_u\{\tau \leq t\}\,\pi(v) \\ &= \mathbf{P}_u\{X_t = v, \tau > t\} + (1 - \mathbf{P}_u\{\tau > t\})\pi(v)\,, \end{aligned}$$

so

$$\begin{aligned} |P^t(u,v) - \pi(v)| &= |\mathbf{P}_u\{X_t = v, \tau > t\} - \pi(v)\cdot\mathbf{P}_u\{\tau > t\}\,| \\ &\leq \mathbf{P}_u\{X_t = v, \tau > t\} + \pi(v)\cdot\mathbf{P}_u\{\tau > t\}\ . \end{aligned}$$

It follows that

$$\begin{aligned} 2\mathrm{dist}_{\mathrm{TV}}(P^t(u,\cdot), \pi(\cdot)) &= \sum_v |P^t(u,v) - \pi(v)| \\ &\leq \sum_v (\mathbf{P}_u\{X_t = v, \tau > t\} + \pi(v)\cdot\mathbf{P}_u\{\tau > t\}) \\ &= 2\mathbf{P}\{\tau > t\}\,. \end{aligned}$$

Since u was arbitrary, this completes the proof. □

3.2. **Hitting times and cover times.** We've seen that mixing times are controlled by strong stationary times; in this lecture we restrict our attention to reversible Markov chains. You saw reversible Markov chains in Section 1.4; recall that reversibility means that

$$\forall u, v,\ \pi(u)P(u,v) = \pi(v)P(v,u).$$

Exercise 3.7. *Show that if P is reversible with stationary distribution π, then for all $t \geq 1$ and all u, v,*

$$\pi(u)P^t(u,v) = \pi(v)P^t(v,u),$$

so P^t is also reversible with stationary distribution π.

The goal of the section is to show that mixing times are controlled by *hitting times from a random site*. In this section we always denote $H^u = \min(t \geq 0 : X_t = u)$ for the first hitting time of state u.

We say a chain P is *lazy* if $P(u,u) \geq 1/2$ for every state u. Equivalently, P is lazy if it is possible to write $P = (Q + I)/2$ where Q is some transition matrix and I is the identity matrix. Note that lazy chains are automatically aperiodic.

Theorem 8. *If P is reversible and lazy with stationary distribution π then*

$$t_{\text{mix}}(1/4) \leq 2 \max_u \mathbf{E}_\pi \{H^u\} + 1 \leq 2t_{\text{hit}} + 1.$$

Here $t_{\text{hit}} := \max_{u,v} \mathbf{E}_u \tau_v$ is the worst-case expected hitting time between two states. The theorem clearly fails if the chain is not reversible; a biased random walk on a cycle of length n has mixing time $\Theta(n^2)$ but all hitting times are $\Theta(n)$ in expectation.

The theorem also fails if the chain is reversible but not lazy. For this consider a complete bipartite graph augmented with self-loops of very low weight. If the weights are sufficiently small then the random walk will with high probability visit every site before ever following a loop; but until the walk follows a loop it is behaving periodically so can not mix.

Our proof follows that given in Chapter 10 of Levin-Peres-Wilmer fairly closely. We require three lemmas. They do not all need the full assumptions of the theorem.

Lemma 9. *Assuming reversibility, for all $t \geq 0$ and any state u,*

$$P^{t2+2}(u,u) \leq P^{2t}(u,u).$$

Moreover, if P is lazy, so can be written as $P = (Q + I)/2$, then $P^t(u,u)$ is monotone decreasing for all u.

Proof. Begin by decomposing $\pi(u)P^{2t+2}(u,u)$ according to the locations of the random walk at time t and at time $t+2$:

$$\begin{aligned}
&\pi(u)P^{2t+2}(u,u)\\
&= \sum_{v,w} \pi(u)P^t(u,v)P^2(v,w)P^t(w,u)\\
&= \sum_{v,w} \pi(v)P^t(v,u)P^2(v,w)P^t(w,u)\\
&= \sum_{v,w} [P^t(v,u)(\pi(v)P^2(v,w))^{1/2}] \cdot [(\pi(w)P^2(w,v))^{1/2}P^t(w,u)]
\end{aligned}$$

This rewriting uses that $\pi(u)P(u,v) = \pi(v)P(v,u)$ and that $\pi(v)P(v,w) = \pi(w)P(w,v)$. The terms in the square brackets are related by exchanging v and w, so the final line has the form

$$\sum_{v,w} a_{vw}a_{wv}.$$

By Cauchy-Schwartz, this is bounded by

$$(\sum_{v,w} a_{v,w}^2)^{1/2}(\sum_{v,w} a_{w,v}^2)^{1/2} = \sum_{v,w} a_{v,w}^2.$$

Plugging this in above gives

$$\begin{aligned}
&\pi(u)P^{2t+2}(u,u)\\
&\leq \sum_{v,w} P^t(v,u)^2\pi(v)P^2(v,w)\\
&= \sum_{v} P^t(v,u)^2\pi(v)\\
&= \sum_{v} \pi(u)P^t(u,v)P^t(v,u)\\
&= P^{2t}(u,u)\,.
\end{aligned}$$

To prove the second assertion, simply note that if $P = (Q+I)/2$ then $P^t(u,u) = K^{2t}(u,u)$ for an auxiliary chain K. The chain K can be defined by adding a new "reflector" vertex r_{uv} along each edge uv of the chain Q. Set

$$\begin{aligned}
K(u, r_{uv}) &= Q(u,v)\\
K(r_{u,v}, u) = K(r_{u,v}, v) &= 1/2 \text{ if } u \neq v\\
K(r_{u,u}, u) &= 1
\end{aligned}$$

Then K is reversible and has stationary measure π_K given by

$$\begin{aligned}
\pi_K(u) &= \pi(u)/2 \text{ for } u \text{ in the original chain}\\
\pi_K(r_{uv}) &= \pi(u)Q(u,v).
\end{aligned}$$

Finally, $K^2(u,v) = P(u,v)$, so $P^{t+1}(u,u) = K^{2t+2}(u,u) \leq K^{2t}(u,u) = P^t(u,u)$. □

The second lemma is a surprising identity for hitting times starting from a random site. It does not require reversibility. The sum on the right is not absolutely convergent but makes sense when interpreted as a limit.

Lemma 10. *If P is irreducible and aperiodic and has stationary distribution π, then for any state u, writing $H^u = \min(t \geq 0 : X_t = u)$,*

$$\pi(u)\mathbf{E}_\pi\left\{H^u\right\} = \sum_{t\geq 0}(P^t(u,u) - \pi(u))\,.$$

Proof. We will use the result from Exercise 3.4 (a), applied to the stopping time $\tau = H^u_{\geq m} := \min\{t \geq m : X_t = u\}$. This stopping time has the property that $\mathbf{P}_u\left\{X_\tau = u\right\} = 1$, so that exercise tells us that

$$\begin{aligned}
&\pi(u)\cdot\mathbf{E}_u\left\{\tau\right\}\\
&= G_\tau(u,u)\\
&= \sum_{t\geq 0}\mathbf{P}_u\left\{X_t = u, t < \tau\right\}\\
&= \sum_{t=0}^{m-1}\mathbf{P}_u\left\{X_t = u\right\}\\
&= \sum_{t=0}^{m-1}P^t(u,u)\,,
\end{aligned}$$

since the random walk can not visit u at times $t \in [m, \tau - 1)$ by definition. Also,

$$\begin{aligned}
\mathbf{E}_u\left\{\tau\right\} &= m + \sum_v \mathbf{P}_u\left\{X_m = v\right\}\cdot\mathbf{E}_u\left\{\tau - m \mid X_m = v\right\}\\
&= m + \sum_v \mathbf{P}_u\left\{X_m = v\right\}\mathbf{E}_v\left\{H^u\right\}\\
&= m + \mathbf{E}_{\mu_m}\left\{H^u\right\},
\end{aligned}$$

where μ_m is the distribution of X_m when $X_0 = u$. Multiplying both sides by $\pi(u)$ and combining with the previous identity gives

$$\pi(u)\mathbf{E}_{\mu_u}\left\{H^u\right\} = \sum_{t=0}^{m-1}(P^t(u,u) - \pi(u)).$$

Since μ_m tends to π as $m \to \infty$, taking a limit in m proves the lemma. □

The final lemma is essentially a careful application of Cauchy-Schwartz.

Lemma 11. *Assuming reversibility, for all $m \geq 0$ and any state u,*

$$\mathrm{dist}_{\mathrm{TV}}(P^m(u,\cdot),\pi(\cdot)))^2 \leq \frac{1}{4}\left(\frac{P^{2m}(u,u)}{\pi(u)} - 1\right).$$

Proof. We will show that $(2\mathrm{dist}_{\mathrm{TV}}(P^m(u,\cdot),\pi(\cdot)))^2 \leq \frac{P^{2m}(u,u)}{\pi(u)} - 1$. To do so, write

$$|P^m(u,v) - \pi(v)| = \pi(v)^{1/2} \cdot \frac{|P^m(u,v) - \pi(v)|}{\pi(v)^{1/2}}$$

and apply Cauchy Schwartz to obtain

$$(\sum_v |P^m(u,v) - \pi(v)|)^2 \leq (\sum_v \pi(v)) \cdot \sum_v \frac{(P^m(u,v) - \pi(v))^2}{\pi(v)}$$

Using reversibility to write $\frac{P^m(u,v)^2}{\pi(v)} = \frac{P^m(u,v)P^m(v,u)}{\pi(u)}$, the right-hand side is

$$\sum_v \left(\frac{P^m(u,v)P^m(v,u)}{\pi(u)} - 2P^m(u,v) + \pi(v) \right) = \frac{P^{2m}(u,u)}{\pi(u)} - 1\,.$$

□

Proof of Theorem 8. Fix any state u. Using the hitting time identity (Lemma 10) and monotonicity (Lemma 9), for any $m \geq 1$ we have

$$\begin{aligned}
\mathbf{E}_\pi\left\{H^u\right\} &= \frac{1}{\pi(u)} \sum_{t \geq 0} (P^t(u,u) - \pi(u)) \\
&\geq \frac{1}{\pi(u)} \sum_{t=1}^{2m} (P^t(u,u) - \pi(u)) \\
&\geq \frac{2m}{\pi(u)} (P^{2m}(u,u) - \pi(u))\,;
\end{aligned}$$

dividing though by $2m$ and using the Cauchy-Schwartz bound (Lemma 11) then gives

$$\begin{aligned}
\mathrm{dist}_{\mathrm{TV}}(P^m(u,\cdot),\pi(\cdot)))^2 &\leq \frac{1}{4}\left(\frac{P^{2m}(u,u)}{\pi(u)} - 1 \right) \\
&\leq \frac{\mathbf{E}_\pi\left\{\tau_u\right\}}{8m}.
\end{aligned}$$

If $m \geq 2\max_v \mathbf{E}_\pi\left\{\tau_v\right\}$ then the right-hand side is at most $1/16$. Since u was arbitrary, for such m this yields that

$$\max_u \mathrm{dist}_{\mathrm{TV}}(P^m(u,\cdot),\pi(\cdot))) \leq \frac{1}{4},$$

so $t_{\mathrm{mix}}(1/4) \leq \lceil 2\max_v \mathbf{E}_\pi\left\{\tau_v\right\} \rceil \leq 2\max_v \mathbf{E}_\pi\left\{\tau_v\right\} + 1$. □

Exercise 3.8. [The random target lemma] *Show that in any irreducible chain, for any states u and v, if*

$$\sum_x \pi(x)\mathbf{E}_u\left\{\tau_x\right\} = \sum_x \pi(x)\mathbf{E}_v\left\{\tau_x\right\},$$

or in shorthand, $\mathbf{E}_u\left\{\tau_\pi\right\} = \mathbf{E}_v\left\{\tau_\pi\right\}$.

We close the lecture by briefly discussing the *cover time*

$$\tau_{\text{cov}} := \min(t \geq 0 : (X_s, 0 \leq s \leq t) \text{ has visited all the states}) = \max_v H^v.$$

Writing $t_{\text{cov}} = \max_u \mathbf{E}_u \{\tau_{\text{cov}}\}$, it's clear from the second identity that

$$t_{\text{cov}} \geq \max_{u,v} \mathbf{E}_u \{H^u\}.$$

Consider a random walk on a connected weighted graph $G = (V, E)$ with edge weights $(c_e, e \in E)$. Recall that $c_v = \sum_{e \ni v} c_e$ and let $c = \sum_v c_v = 2 \sum_e c_e$. We next prove an upper bound on the cover time which uses spanning trees of G.

Theorem 12. *It holds that*

$$t_{\text{cov}} \leq c \cdot \min_T \sum_{e \in T} \frac{1}{c_e},$$

where the minimum is over spanning trees T of G.

Proof. Write $n = |V|$. Fix any state v and spanning tree T, and a "contour" path $v = v_0, v_1 \ldots, v_{2n-2}$ of T; such a path traverses each edge once in each direction. Then

$$\begin{aligned} \mathbf{E}_v \{\tau_{\text{cov}}\} &\leq \sum_{j=0}^{2n-3} \mathbf{E}_{v_j} \{H^{v_{j+1}}\} \\ &= \sum_{e=uv \in T} (\mathbf{E}_u \{H^v\} + \mathbf{E}_v \{H^u\}) \\ &\leq \sum_{e \in T} \frac{c}{c_e}, \end{aligned}$$

the last bound holding by the edge-commute inequality (Exercise 3.4 (d)). □

3.3. **Spectral techniques.** This section is based on the books by Aldous and Fill (Section 3.4) and Levin, Peres and Wilmer (Sections 12.1 and 2) In this lecture we restrict attention to reversible chains. For a function $f : V \to \mathbb{R}$ and a probability distribution μ on the state space V, we write

$$\text{Var}_\mu(f) := \mathbf{E}\left[(X - \mathbf{E}X)^2\right]$$

where X is a random variable with distribution μ. Also, write $P^t f$ for the function

$$x \overset{P^t f}{\longmapsto} \sum_{y \in V} P^t(x, y) f(y).$$

The main point of the lecture is to understand the consequences of the spectral representation of reversible, irreducible finite state Markov chains for the mixing time.

The spectral representation is the following: there exists an orthonormal basis of eigenfunctions $(f_i, 1 \leq i \leq |V|)$ with corresponding eigenvalues $1 = \lambda_1 > \lambda_2 \geq$

$\ldots \geq \lambda_{|V|} \geq -1$ such that for any function $f : V \to \mathbb{R}$,

$$P^t f \equiv \sum_{i=1}^{|V|} \langle f, f_i \rangle_\pi f_i \lambda_i^t. \tag{3}$$

Here

$$\langle f, g \rangle_\pi := \sum_{v \in V} f(v) g(v) \pi(v).$$

Moreover, f_1 can be taken to be identically 1, $f_1(v) = 1$ for all v; this statement is part of what we mean by "the spectral representation". The fact that no eigenvalue is greater than 1 is obvious: fix any function $f : V \to \mathbb{R}$, let $\|f\|_\infty := \max_{v \in V} |f(v)|$, and let $u \in V$ be such that $|f(u)| = \|f\|_\infty$. Then

$$|(Pf)(u)| = |\sum_v P(u,v) f(v)| \leq \|f\|_\infty \sum_v P(u,v) = \|f\|_\infty = |f(u)|.$$

The fact that $\lambda_2 < 1$ is also obvious: π is a left eigenfunction with eigenvalue 1; a second such eigenfunction would contradict the uniqueness of π.

The statement that $(f_i, 1 \leq i \leq |V|)$ is an orthonormal basis means that $\langle f_i, f_j \rangle_\pi = \mathbf{1}_{[i=j]}$ for all $1 \leq i, j \leq |V|$. The basis decomposition of a function then states that any function $f : V \to \mathbb{R}$ can be written as

$$f \equiv \sum_{i=1}^{|V|} \langle f, f_j \rangle_\pi f_i. \tag{4}$$

Exercise 3.9. *The functions f_i are "right eigenfunctions": $P^t f_i = \lambda_i^t f_i$. Show that $f_i \pi$, defined by $(f_i \pi)(u) = f_i(u)\pi(u)$ is a left eigenfunction of P with eigenvalue λ_i.*

Before proving (3), we develop the link with mixing times. Write $\lambda_{\max} = \max\{|\lambda_i|, 2 \leq i \leq |V|\}$ for the largest absolute value of a "non-trivial" eigenvalue. The *absolute spectral gap* of the chain is defined to be

$$\gamma := 1 - \lambda_{\max}$$

and the *relaxation time* of the chain is $t_{\text{rel}} := 1/\gamma$.

Theorem 13. *Let P be a reversible irreducible Markov chain with finite state space V. Write $\pi_{\min} = \min_{v \in V} \pi(v)$. Then*

$$\log\left(\frac{1}{2\epsilon}\right)(t_{\text{rel}} - 1) \leq t_{\text{mix}}(\epsilon) \leq \log\left(\frac{1}{2\epsilon \pi_{\min}}\right) t_{\text{rel}}\,.$$

Lemma 14. *For any state $u \in V$,*

$$\pi(u) \sum_{i=1}^{|V|} f_i(u)^2 = 1.$$

Proof. For any $u \in V$, the Dirac delta $\delta_u : V \to \mathbb{R}$,

$$\delta_u(v) = \mathbf{1}_{[v=u]},$$

has the representation

$$\delta_u \equiv \sum_{i=1}^{|V|} f_j(u)\pi(u)f_j.$$

(To see this, simply apply the basis decomposition to δ_u and rearrange.) It follows that

$$\begin{aligned}\pi(u) &= \langle \delta_u, \delta_u \rangle_\pi \\ &= \langle \sum_{i=1}^{|V|} f_i(u)\pi(u)f_i, \sum_{i=1}^{|V|} f_i(u)\pi(u)f_i \rangle_\pi \\ &= \pi(u)^2 \sum_{i=1}^{|V|} f_i(v)^2.\end{aligned}$$

For the last equality, we use that $(f_i, 1 \leq i \leq |V|)$ is an orthonormal basis for $\langle \cdot, \cdot \rangle_\pi$. □

Lemma 15. *The spectral representation (3) is equivalent to the statement that*

$$P^t(u,v) = \pi(v)\left(1 + \sum_{i=2}^{|V|} f_i(u)f_i(v)\lambda_i^t\right). \tag{5}$$

Proof. Note that

$$(P^t\delta_v)(u) = \sum_{w\in V} P^t(u,w)\delta_v(w) = P^t(u,v).$$

Using the spectral representation,

$$(P^t\delta_v)(u) = \sum_{i=1}^{|V|} \langle f_i, \delta_v \rangle_\pi f_i(u)\lambda_i^t \equiv \sum_{i=1}^{|V|} f_i(v)\pi(v)f_i(u)\lambda_i^t = \pi(v)\sum_{i=1}^{|V|} f_i(u)f_i(v)\lambda_i^t\,.$$

Since f_1 is identically 1 and $\lambda_1 = 1$, this establishes (5). For the converse, simply write

$$\begin{aligned} P^t f(u) &= \sum_{v\in V} P^t(u,v) f(v) \\ &= \sum_{v\in V} \left(\sum_{i=1}^{|V|} f_i(u) f_i(v) \lambda_i^t \right) \pi(v) f(v) \\ &= \sum_{i=1}^{|V|} f_i(u) \lambda_i^t \left(\sum_{v\in V} f(v) f_j(v) \pi(v) \right) \\ &= \sum_{i=1}^{|V|} f_i(u) \lambda_i^t \langle f, f_i \rangle_\pi \\ &= \left(\sum_{i=1}^{|V|} \langle f, f_i \rangle_\pi \lambda_i^t f_i \right)(u) . \end{aligned}$$

□

Proof of Theorem 13. By Lemma 15 we have

$$\begin{aligned} |P^t(u,v) - \pi(v)| &\le \pi(v) \cdot \sum_{i=2}^{|V|} |f_i(u) f_i(v) \lambda_i^t| \\ &\le \pi(v) \cdot \sum_{i=2}^{|V|} |f_i(u) f_i(v)| \lambda_{\max}^t \\ &\le \pi(v) \lambda_{\max}^t \cdot \left(\sum_{i=2}^{|V|} f_i(u)^2 \cdot \sum_{i=2}^{|V|} f_i(v)^2 \right)^{1/2} \end{aligned}$$

By Lemma 14 we know $\sum_{i=2}^{|V|} f_i(u)^2 \le 1/\pi(u)$, so the preceding bound gives

$$|P^t(u,v) - \pi(v)| \le \frac{\lambda_{\max}^t \pi(v)}{\sqrt{\pi(u)\pi(v)}} \le \frac{\lambda_{\max}^t \pi(v)}{\pi_{\min}}.$$

Summing over v gives that

$$\mathrm{dist}_{\mathrm{TV}}(P^t(u,\cdot), \pi(\cdot)) \le \frac{\lambda_{\max}^t}{2\pi_{\min}} = \frac{(1-\gamma)^t}{2\pi_{\min}} \le \frac{e^{-\gamma t}}{2\pi_{\min}} ,$$

so if $t \ge \log(1/2\epsilon\pi_{\min})/\gamma$ then $\mathrm{dist}_{\mathrm{TV}}(P^t(u,\cdot), \pi(\cdot)) \le \epsilon$. This gives the upper bound.

For the lower bound, we actually prove that $t_{\mathrm{mix}}(\epsilon) \ge \frac{\log(1/2\epsilon)}{\log \gamma}$. This implies the lower bound since $1/\log\gamma \ge \gamma - 1$; the latter inequality is equivalent to the fact that $e^x \ge 1 + x$ for $x \ge 0$.

Fix any eigenfunction f_j with $j \neq 1$ Since f_j is orthogonal to $f_1 \equiv 1$, we have

$$\sum_{v \in V} \pi(v) f_j(v) = \langle f_1, f_j \rangle_\pi = 0.$$

Thus, for any $t \geq 0$, writing $\|f_j\|_\infty = \max_{v \in V} |f_j(v)|$, and let u be such that $|f_j(u)| = \|f_j\|_\infty$. Then

$$\begin{aligned}
|\lambda_j^t| \|f_j\|_\infty &= |\lambda_j^t f_j(u)| \\
&= |(P^t f_j)(u)| \\
&= \left| \sum_{v \in V} P^t(u,v) f_j(v) \right| \\
&= \left| \sum_{v \in V} (P^t(u,v) f_j(v) - \pi(v) f_j(v)) \right| \\
&\leq \|f_j\|_\infty \cdot \sum_{v \in V} |P^t(u,v) - \pi(v)| \\
&\leq 2 \|f_j\|_\infty \mathrm{dist}_{\mathrm{TV}}(P^t(u,\cdot), \pi(\cdot)),
\end{aligned}$$

so $|\lambda_j|^t \leq 2\mathrm{dist}_{\mathrm{TV}}(P^t(u,\cdot), \pi(\cdot))$. Taking $t = t_{\mathrm{mix}}(\epsilon)$ and maximizing over $j \neq 1$ gives

$$\lambda_{\max}^{t_{\mathrm{mix}}(\epsilon)} \leq 2\epsilon.$$

Taking logs and rearranging gives

$$t_{\mathrm{mix}}(\epsilon) \geq \frac{\log(1/2\epsilon)}{\log(1/\lambda_{\max})} = \frac{\log(1/2\epsilon)}{\log \gamma}$$

□

Exercise 3.10. [The Poincaré inequality] *Fix a reversible chain with state space V and stationary distribution π. Prove that for any function $f : V \to \mathbb{R}$, for all $t \geq 0$,*

$$\mathrm{Var}_\pi(P^t f) \leq \left(1 - \frac{1}{t_{\mathrm{rel}}}\right)^{2t} \mathrm{Var}_\pi(f).$$

Exercise 3.11. *Prove the first assertion of Lemma 9 using the spectral representation of reversible transition matrices.*

Our proof of (3) is not self-contained; it relies on the spectral theorem for symmetric matrices. P is not symmetric, but since it is reversible, if we define $Q(u,v) = (\frac{\pi(u)}{\pi(v)})^{1/2} P(u,v)$ then Q is symmetric. It's useful to rewrite this as $Q = D^{1/2} P D^{-1/2}$, where $D = \mathrm{diag}(\pi)$ is the diagonal matrix with $D(u,u) = \pi(u)$. This makes it obvious that $(\pi)^1/2$ is an eigenvector of Q with eigenvalue 1.

The spectral theorem states that there exists an orthonormal matrix $M = (M(u,v))_{u,v \in V}$ such that

$$Q = M^t \Lambda M,$$

where $\Lambda = \mathrm{diag}(\lambda_i, 1 \leq i \leq n)$ is the diagonal matrix with the eigenvalues λ_i along the diagonals. (The ordering of the λ_i is achieved simply by indexing appropriately.) We can think of the columns of M as eigenfunctions $(g_i, 1 \leq i \leq |V|)$: we have $(Mg_i)_j = \delta_i(j)$,

$$(Qg_i)_j = (M^t \Lambda M g_i)_j = M^t \lambda_i \delta_i(j) = \lambda_i g_i.$$

Orthonormality means $\langle g_i, g_j \rangle = \mathbf{1}_{[i=j]}$, where $\langle \cdot, \cdot \rangle$ is the usual inner product on $\mathbb{R}^{|V|}$. Recall we have $g_1 = (\pi)^{1/2}$.

Defining $f_i = D^{-1/2} g_i$, since $P = D^{-1/2} Q D^{1/2}$, we have

$$P f_i = D^{-1/2} Q D^{1/2} f_i = D^{-1/2} Q g_i = D^{-1/2} \lambda_i g_i = \lambda_i f_i.$$

We then have

$$\langle f_i, f_j \rangle_\pi = \sum_{v \in V} f_i(v) f_j(v) \pi(v) = \langle D^{1/2} f_i, D^{1/2} f_j \rangle = \langle g_i, g_j \rangle,$$

so $(f_i, 1 \leq i \leq |V|)$ are orthogonal for $\langle \cdot, \cdot \rangle_\pi$.

Finally, using that

$$\delta_v \equiv \sum_{i=1}^{|V|} f_i(v) \pi(v) f_i,$$

we have

$$\begin{aligned} P^t(u, v) &= (P^t \delta_v)(u) \\ &= \sum_{i=1}^{|V|} f_i(v) \pi(v) (P^t f_i)(u) \\ &= \sum_{i=1}^{|V|} f_i(v) \pi(v) \lambda_i^t f_i(u) \\ &= \pi(v) \left(1 + \sum_{i=2}^{|V|} f_i(u) f_i(v) \lambda_i^t \right), \end{aligned}$$

which by Lemma 15 is equivalent to (3).

3.4. **The Lovász Local Lemma and sampling uniform spanning trees using random walks.** The current section is adapted from the paper "Uniform sampling through the Lovász local lemma", by Heng Guo, Mark Jerrum and Jingcheng Liu. We first describe the framework.

- Let $X = (X_i, 1 \leq i \leq n)$ be independent random variables, with X_i having distribution μ_i and taking values in some set S_i.
- Fix a finite sequence $(A_1, \ldots, A_m)$ of "bad events". For each $1 \leq \ell \leq m$ there is a set $\mathrm{var}(\ell) \subset [n]$; the random variables $(X_i, i \in \mathrm{var}(\ell)$ determine whether A_ℓ occurs. In other words, one may think of A_ℓ as a function, $A_\ell : \prod_{i=1}^n S_i \to \{0, 1\}$, so that

$$A(x) = A(y)$$

whenever $(x_i, i \in \mathrm{var}(\ell)) = (y_i, i \in \mathrm{var}(\ell))$. (The idea being that $A(x) = 1$ means "A(x)" occurs when $X = x$.)

- The *dependency graph* of $\{A_1, \ldots, A_m\}$ is the graph $D = (V, E)$ with vertices $V = [m]$ and with $\{k, \ell\} \in E$ if and only if $\mathrm{var}(k) \cap \mathrm{var}(\ell) \neq \emptyset$.

Theorem 16 (Asymmetric Lovász local lemma)**.** *Suppose there exists $c \in [0,1)^m$ such that for all $i \in [m]$,*

$$\mathbf{P}\{A_i\} \leq c_i \cdot \prod_{(i,j)\in E} (1 - c_j).$$

Then

$$\mathbf{P}\{\textit{None of the bad events occur}\} \geq \prod_{i=1}^{m} (1 - c_i) > 0.$$

We will not prove the local lemma, but instead focus on how to sample good from "good configurations" (where no bad events occur) in the setting of the LLL.

Example: Uniform spanning trees. Fix a connected graph $G = (V, E)$ and a "root" node $r \in V$. For each $u \neq r$ let X_u be a uniformly random neighbour of u (equivalently, a uniformly random edge incident to u).

A directed cycle in G is a sequence $C = (c_0, c_1, \ldots, c_m)$ of vertices with $c_m = c_0$. For each directed cycle C let A_C be the (bad) event that $X_{c_j} = c_{j+1}$ for $1 \leq j < m$.

Note that no bad events occur precisely if the set of edges $\{(u, X_u), u \neq r\}$ forms the edge set of a spanning tree; in this case, viewing the spanning tree as rooted at r, all edges are oriented toward the root.

Note also that for any spanning tree T of G

$$\mathbf{P}\{\{(u, X_u), u \neq r\} = E(T)\} = \prod_{u \neq r} \frac{1}{\deg(u)};$$

so if $\{(u, X_u), u \neq r\}$ happens to comprise the edges of a tree, then it distributed as a uniform spanning tree of T. If $\{(u, X_u), u \neq r\}$ does not form a tree, the next algorithm will get rid of the cycles for us.

We now describe the "partial rejection sampling" algorithm for resampling bad events.

Partial Rejection Sampling Algorithm

(1) Draw independent samples of all random variables $X_1, \ldots, X_n$ from their respective distributions
(2) While at least one bad event occurs, independently resample all variables in $\bigcup_{i:A_i \text{ occurs}} \mathrm{var}(A_i)$.
(3) Output the current assignment.

We say the dependency graph D is *extremal* if any two events A_i and A_j are either independent or they are disjoint. In other words, D is extremal if $A_i \cap A_j = \emptyset$ for all $(i, j) \in E$.

Theorem 17 (Wilson; Guo and Jerrum and Liu, 2016)**.** *If D is extremal then the output of the PRS Algorithm is a sample from the product distribution: For any $z = (z_i, 1 \leq i \leq n) \in \prod_{i=1}^n S_i$,*

$$\mathbf{P}\left\{\textit{PRS outputs } z\right\} \propto \prod_{i=1}^{n} \mathbf{P}\left\{X_i = z_i\right\}.$$

The theorem implies that, in the spanning tree example, the output of the PRS algorithm is a uniform spanning tree.

We analyze the PRS algorithm using *stack popping*. The idea of stack popping is the following. The algorithm involves resampling random variables; imagine that we have *pre-computed and stored* such samples before running the algorithm. This equips us with a stack $x_i = (x_{i,j}, j \geq 0)$ of possible values of X_i, for each $1 \leq i \leq n$. The initial state of the algorithm is $X_i = x_{i,0}$ for each i.

At the start of the algorithm (step 1), when we look at the stacks "from above" we simply see the initial values $x_{i,0}$ of the random variables. When we resample a bad event A_ℓ, we "pop the stacks" in $\mathrm{var}(A_\ell)$, removing the top element of the stack and throwing it away. Write $j(i,t)$ for the number of times that X_i has been resampled up to step t of the algorithm. Then the current values at step t are $x(t) := (x_{i,j(i,t-1)}, 1 \leq i \leq n)$.

If $A_\ell(x(t))$ occurs and we choose to resample A_ℓ at step k, then we simply increment the number of times variables in $\mathrm{var}(A_\ell)$ have been resampled:

$$j(i,k+1) = \begin{cases} j(i,k) & \text{if } i \notin \mathrm{var}(A_\ell) \\ j(i,k+1) & \text{if } i \in \mathrm{var}(A_\ell). \end{cases}$$

Let $\sigma_x = \max(t : \exists i, A_i(x(t-1)) \text{ occurs})$. Write

$$x^* = ((x_{i,j(i,s)}, 0 \leq s \leq \sigma), 1 \leq i \leq n).$$

We call x^* the *log* of the algorithm; it keeps track of all information associated with running the PRS Algorithm on stacks x. Note that $\sigma_{x^*} = \sigma_x$, since by the definition of σ_x, $A_i(x(\sigma_x)) = A_i(x^*(\sigma_x))$ does not occur for any $1 \leq i \leq n$. We write $\sigma = \sigma_x$ for the rest of the proof.

Next fix sequences

$$x' = ((x'_{i,j(i,s)}, 0 \leq s \leq \sigma), 1 \leq i \leq n)$$

which can be obtained from x by only changing the final value in each vector, and only such that the final assignment is still valid. Formally, we require x' to satisfy that

$$x'_{i,j(i,s)} = x_{i,j(i,s)}$$

whenever $j(i,s) < j(i,\sigma_x)$, and that $A_i(x'(\sigma))$ does not occur for all $1 \leq i \leq n$.

We claim that for all $t < \sigma$,

$$\{i : A_i(x^*(t)) \text{ occurs}\} = \{i : A_i(x(t)) \text{ occurs}\} = \{i : A_i(x'(t)) \text{ occurs}\} \quad (6)$$

The first equality is obvious. We prove the second by contradiction. Suppose there is $t < \sigma$ such that $\{i : A_i(x(t)) \text{ occurs}\} \neq \{i : A_i(x'(t)) \text{ occurs}\}$. Let t_0 be the smallest such step, and fix a bad event A such that $A(x'(t_0))$ occurs but $A(x(t_0))$

does not occur. Then there is $i \in \mathrm{var}(A)$ such that $j(i,t_0) = j(i,\sigma)$, as this is the only way for $x(i,j(i,t_0))$ and $x'(i,j(i,t_0))$ to differ.

Let $I = \{i \in \mathrm{var}(A) : j(i,t_0) = j(i,\sigma)\}$. By assumption, $A(x'(\sigma))$ does not occur, so it can not be that $I = \mathrm{var}(A)$. Each element k of $\mathrm{var}(A) \setminus I$ has $j(k,t_0) < j(k,\sigma)$, so is resampled after time t_0 when running Algorithm 2 on x.

Let $k_{\min}$ be the index of the first element of $\mathrm{var}(A) \setminus I$ to be resampled after time t_0, and say $k_{\min}$ is resampled at time t_1. Let B be the event which causes this resampling, so $B(x(t_1))$ occurs and $k_{\min} \in \mathrm{var}(B)$.

Note that if $k \in \mathrm{var}(B)$ then X_k is resampled at time t_1 so $j(k,t_1) < j(k,\sigma)$ and thus $k \notin I$. Write $J = \mathrm{var}(A) \cap \mathrm{var}(B)$; this is non-empty since in particular $k_{\min} \in J$. Moreover, $J \subset \mathrm{var}(A) \setminus I$.

We have that $j(k,t_0) = j(k,t_1)$ for all $k \in \mathrm{var}(A) \cap \mathrm{var}(B)$, since $k_{\min}$ was the first resampling event to affect $\mathrm{var}(A) \setminus I$ after t_0. It follows that it is well-defined to set

$$y_i = \begin{cases} x(i,j(i,t_0)), i \in \mathrm{var}(A) \\ x(i,j(i,t_1)), i \in \mathrm{var}(B) \\ x(i,j(i,\sigma)), i \notin \mathrm{var}(A) \cup \mathrm{var}(B). \end{cases}$$

We have assigned both value $x(i,j(i,t_0))$ and value $x(i,j(i,t_1))$ to elements of $\mathrm{var}(A) \cap \mathrm{var}(B)$, but that is OK since these values are the same. By construction, $A(y)$ and $B(y)$ both occur, so the events are not disjoint. But also $\mathrm{var}(A) \cap \mathrm{var}(B) \neq \emptyset$ so A and B are not independent. This contradicts the assumption of extremality.

Proof of Theorem 17. Fix points $z = (z_i, 1 \le i \le n)$ and $z' = (z'_i, 1 \le i \le n)$ with $z_i \in S_i$ and $z'_i \subset S_i$ for each $1 \le i \le n$, such that no bad events occur for z or z' (i.e. $A_j(z)$ and $A_j(z')$ do not occur for $1 \le j \le m$). Equation (6) implies that there is a bijection between sequences x yielding a log x^* with $x^*(\sigma_{x^*}) = z$ and sequences yielding log x' with $x'(\sigma_{x'}) = z'$. It follows that

$$\begin{aligned} \mathbf{P}\left\{\text{PRS outputs } z\right\} &= \sum_{x^* : x^*(\sigma_{x^*}) = z} \mathbf{P}\left\{\text{PRS has log} x^*(\sigma_{x^*})\right\} \\ &= \prod_{i=1}^{n} \frac{\mathbf{P}\left\{X_i = z'_i\right\}}{\mathbf{P}\left\{X_i = z'_i\right\}} \sum_{x^* : x^*(\sigma_{x^*}) = z'} \mathbf{P}\left\{\text{PRS has log} x^*(\sigma_{x^*})\right\} \\ &= \prod_{i=1}^{n} \frac{\mathbf{P}\left\{X_i = z'_i\right\}}{\mathbf{P}\left\{X_i = z'_i\right\}} \mathbf{P}\left\{\text{PRS outputs } z'\right\}, \end{aligned}$$

and the result follows. □

Exercise 3.12. *Suppose that instead of resampling all bad events simultaneously, we resample bad events one-at-a-time according to some rule R. Show that for any $t \le \sigma_x$ and $1 \le k \le m$, if $A_k((x_{i,j(i,t)}, i \in \mathrm{var}(k))$ occurs and so the entries of $(x_{i,j(i,t)}, i \in \mathrm{var}(k))$ are resampled, then $(x_{i,j(i,t)}, i \in \mathrm{var}(k))$ is also resampled when we resample according to rule R. This implies that the choice of rule doesn't change the set of events which are resampled, or the outcome of the algorithm.*

Analyzing the uniform spanning tree example: Wilson's algorithm To analyze the running time of the PRS algorithm on the uniform spanning tree example, we exploit the Abelian property of the PRS algorithm given by Exercise 3.12. Instead of resampling all cycles at once, use the stacks to simulate a simple random walk on G. A simple random walk has $\mathbf{P}_u\{X_1 = v\} = \frac{1}{\deg(v)}\mathbf{1}_{[uv\in E]}$, in agreement with the random variables on the stacks.

So fix a starting node u, and let $(X_i, 0 \le i \le H^r)$ be random walk run starting from u and run until it hits r, where the values of the random walk steps are taken from the stack. Each time the random walk "finds a bad event" (its trace forms a cycle), pop the bad event from the stack and delete the cycle from the trace. This is "loop-erased random walk. Its result is a path from u to r containing no loops. Once this path P_u has been formed, choose another starting node v and run loop-erased random walk from v to P_u, again drawing the random walk steps from the stack. This builds a path P_v from v to P_u. Repeat in this manner until the union of the paths covers all the vertices; this is (an alternative description of) the spanning tree built by the PRS algorithm.

Proposition 18. *The expected number of random samples queried by the PRS algorithm for building a uniform spanning tree of G rooted at r is at most*

$$\sum_{u\neq r} \pi(u)(\mathbf{E}_u H^r + \mathbf{E}_r H^u) \le 2t_{\text{hit}}$$

Proof. Note that the number of random variables queried by the PRS algorithm is just $\sum_{u\neq r} j(u,\sigma)$, where σ is the halting time of the algorithm. By the Abelian property, $j(u,\sigma)$ is the same as the number of times X_u is resampled in the loop-erased random walk construction started from u at time 0. The latter value is just the number of returns to u before H^r. By Exercise 3.4 (c), it follows that

$$\mathbf{E}[j(u,\sigma)] = \mathbf{E}_u\{\#\{0 \le t \le \tau_r : X_t = u\}\} = \pi(u)(\mathbf{E}_u H^r + \mathbf{E}_r H^u).$$

The result follows. □