NOTES FOR THE SUMMER SCHOOL

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.

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.*
- *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 $[n]$.*

Date: May 20, 2019 (typeset May 27, 2019).

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

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

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 $\lambda \ge \nu_x$.*
(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 $k \ge 0$, $\mathbf{P}_v(N_v > k) = (h^v_v)^k$, 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*

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

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

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

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. Show that strong stationary times also exist without the aperiodicity assumption if we are allowed additional randomness in our decision of when to stop.*