18.615 Notes

Lecturer: Jimmy He

Andrew Liu

Fall 2023

My notes for 18.615, "Stochastic Processes", at the Massachusetts Institute of Technology during the Fall 2023 semester. The instructor for this course was Jimmy He (https://he-jimmy.github.io/). All credit for these notes goes to Prof. He and the packet of lecture notes that he prepared for the class. Unlike most of my other classes, the lecture notes he provided were very in depth, so I don't think my notes offer much insight beyond what is already provided (for review purposes). I mainly forced myself to type up these notes anyways as a way to review/study for exams.

Last updated on Thursday 30th November, 2023.

Contents

1 September 7, 2023

1.1 Syllabus

Prerequisites:

- know probability, random variables, expectation, distribution function, etc. don't need to know about measure theory.
- calculus and lin alg. vectors, matrices, eigenvectors, eigenvalues

Grading:

- 6 problem sets (40%). lowest pset is dropped. since the lowest pset is dropped, late homework won't be accepted
- 2 midterms (60%)

1.2 General Outline

Stochastic Processes are a family of random variables indexed by time. Rough outline of things that we'll cover in this class:

1. Markov Chain fundamentals
2. Countable state space markov chains (MC)
3. Martingales, models of fair betting systems
4. Continuous time/space MC

1.3 Intro

Assume everything is discrete time for now.

Definition 1.1
A **stochastic process** is a sequence of r.v.s $X_1, X_1, \ldots$ jointly defined.

Think of the indices $1, 2, \ldots$ as time.

Definition 1.2
A **Markov Chain** is a stochastic process $\{X_i\}$ taking values in $\mathcal{X}$ s.t.

$$\mathbb{P}[X_i = z_i | X_0 = z_0, \ldots, X_{i-1} = z_{i-1}] = \mathbb{P}[X_i = z_i | X_{i-1} = z_{i-1}].$$

We call $\mathcal{X}$ the **state space**.

Intuitively, the probability of any given state only relies on each state at the directly previous timestep. This is called the **Markov property**.

Definition 1.3
We say X_i is **time homogenous** if $\mathbb{P}[X_i = z_i | X_{i-1} = z_{i-1}]$ is independent of i.

In this course, we assume that all markov chains are time homogenous.
Here are some common examples of Markov Chains:

Example 1.4 (Gambler's Ruin)
Let $\mathcal{X} = \mathbb{N}$, and X_i be the amount of money a gambler has at time i, if they bet \$1 during each timestep.

For example, if $X_0 = \$5$, then a valid sequence could be $5, 6, 7, 6, 5, 4, \ldots$.

Example 1.5 (Random Walk)
Let $G = (V, E)$. We move to a neighbor uniformly at random.

Definition 1.6
Let $P^i(x, y) = \mathbb{P}[X_i = y | X_0 = x]$.

This is the probability of moving from x to y in i-steps starting from any point in time. The collection of probabilities $P^1(x, y) = P(x, y)$ is called the **transition probabilities**.

Lemma 1.7
$P^i(x, y)$ is equal to the (x, y)th entry of P^i, where P is a matrix of the transition probabilities.

Proof. Proceed by induction on $i \geq 1$.

Base case $i = 1$ is clear.

Now, using inductive hypothesis and markov property,

$$\begin{aligned}\mathbb{P}[X_{i+1} = y | X_0 = x] &= \sum_z \mathbb{P}[X_{i+1} = y | X_0 = x, X_i = z] \cdot \mathbb{P}[X_i = z | X_0 = x] \\ &= \sum_z P(z,y) P^i(x,z) = P^{i+1}(x,y).\end{aligned}$$

□

Lemma 1.8

Let P be a markov chain and μ be a distribution on $\mathcal{X}$, viewed as a row vector. Then

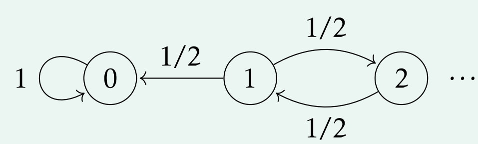
$$(\mu P^i)_x = \mathbb{P}[X_i = x | X_0 \sim \mu].$$

The notation $X_0 \sim \mu$ means that the initial state of the Markov chain is randomly drawn from μ.

1.4 Visual Representation

Draw a directed graph $G = (\mathcal{X}, E)$ where $(x,y) \in E$ if $P(x,y) > 0$, and label (x,y) with $P(x,y)$.

Example 1.9 (Gambler's ruin)



1.5 More definitions

First goal: understand long-term behavior of markov chains.

Definition 1.10

Let P be a MC on $\mathcal{X}$. We say x and y **communicate** and write $x \sim y$ if $\exists i, j > 0$ s.t. $P^i(x,y) > 0$ and $P^j(y,x) > 0$, or $x = y$.

Lemma 1.11 ($\sim$ is an equivalence relation)

1. $x \sim x$
2. $x \sim y \implies y \sim x$
3. $x \sim y, y \sim z \implies x \sim z$

This implies a partition $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_k$, where $x \sim y$ iff $x, y \in \mathcal{X}_i$ for some i.

Definition 1.12

We call these $\mathcal{X}_i$ **communicating classes**. Moreover, we say a class A is **closed** if $P(x,y) = 0 \forall x \in A, y \notin A$.

Definition 1.13

We say a markov chain is **irreducible** if it has exactly one closed class.

Proposition 1.14

Every finite markov chain has a closed class.

Proof. Let A,B be communicating classes. Write $A \to B$ if $\exists x \in A, y \in B$ s.t. $P(x,y) > 0$.

If $A \to B$, and $A \neq B$, then $B \not\to A$. Suppose there were no closed class. Then $\exists$ sequence $A_1 \neq A_2 \neq \ldots$ s.t. $A_1 \to A_2 \to \ldots$, since we can keep picking elements outside of non-closed classes. Given a finite number of elements, there is some i, j such that $A_i = A_j$, contradiction. □

Idea: closed classes are like irreducible Markov Chains.

2 September 12, 2023

2.1 Last lecture review

A time-homogeneous MC is a sequence of r.v.s $X_1, X_2, \dots$ taking values in $\mathcal{X}$, s.t.

$$\mathbb{P}[X_{i+1} = z_{i+1} | X_0 = z_0, \dots, X_i = z_i] = \mathbb{P}[X_1 = z_{i+1} | X_0 = z_i].$$

We also introduced the notation

$$P^i(x,y) = \mathbb{P}[X_i = y | X_0 = x],$$

and said $x \sim y$ (i.e., x **communicates** with y) if $\exists i, j > 0$ s.t.

$$P^i(x,y) > 0 \text{ and } P^i(y,x) > 0, \text{ or } x = y.$$

Since $\sim$ is an equivalence relation, this implies a partition

$$\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \dots, \cup \mathcal{X}_u$$

s.t. $x \sim y$ if and only if $x, y \in \mathcal{X}_i$ for some i. We call each partite set a **communicating class**.

We say that P is **irreducible** if there is exactly one class, i.e., $x \sim y \forall x, y \in \mathcal{X}$. We also say that a class $A \subseteq \mathcal{X}$ is **closed** if $\forall x \in A$, $y \notin A$, $P(x,y) = 0$.

2.2 Periodicity

Definition 2.1 (Periodicity)
Let P be a markov chain on $\mathcal{X}$. Let the **period** of x be

$$\text{per}(x) = \gcd\{i | P^i(x,x) > 0\}.$$

Proposition 2.2
$x \sim y \implies \text{per}(x) = \text{per}(y)$.

Proof. If $x = y$, then we're done, so consider $x \neq y$. Then, $\exists i, j > 0$ s.t. $P^i(x,y) > 0$

and $P^i(y,x) > 0$.

Now, if $P^k(x,x) > 0$, then $P^{k+i+j}(y,y) > 0$, since we can travel from y to x to x to y with non-zero probability. Also, $P^{i+j}(y,y) > 0$. Therefore, $\text{per}(y)|\gcd(k+i+j, i+j) \implies \text{per}(y)|k$. Since this is true for all k for which $P^k(x,x) > 0$, $\text{per}(y)|\text{per}(x)$. Since x and y are interchangeable, $\text{per}(x)|\text{per}(y)$, thus $\text{per}(x) = \text{per}(y)$. □

Definition 2.3

If P is irreducible, it's period is $\text{per}(x)$ for any x. We say that P is **aperiodic** if its period is 1.

Proposition 2.4

Let P be an irreducible MC with period k; then there exists a partition

$$\mathcal{X} = C_1 \cup \ldots \cup C_k$$

s.t. $P(x,y) > 0$ only if $x \in C_i, y \in C_{i+1}$ for some i.

Proof. In the hw. For an example, consider $P = C_k$ with transition probabilities all 1. □

2.3 Stationary Distribution

Definition 2.5

P MC on $\mathcal{X}$. A distribution μ on $\mathcal{X}$ is **stationary** if $\mu P = \mu$.

This is equivalent to:

$$\mathbb{P}[X_i = x | X_i \sim \mu] = \mu(x),$$

which is also equivalent to

$$\sum_x \mu(x)P(x,y) = \mu(y) \forall y.$$

In general, they may or may not exist, and they may not be unique. For example, a random walk on $\mathbb{Z}$ has no stationary distribution. Also, MCs with multiple classes may have multiple stationary distributions.

Notation: stationary distributions will always be π.

Theorem 2.6
If $|\mathcal{X}| < \infty$, $\exists$ a stationary distribution.

We can easily show that there exists a solution to $\mu P = \mu$. In particular, note that $[1, 1, \ldots, 1]^T$ is a right eigenvector for P. Since left and right eigenvectors come in pairs, there exists left eigenvector μ that satisfies $\mu P = \mu$.

The hard part of the proof is to show that there exists a solution that represents an actual distribution, i.e., nonnegative values summing to 1. First, some definitions:

Definition 2.7
Define the **return time**, or **hitting time**, as

$$\tau_x^+ = \min\{i > 0 | X_i = x\}.$$

If we never hit x, $\tau_x^+ = \infty$.

Proposition 2.8
If P is irreducible and $|\mathcal{X}| < \infty$, then $\mathbb{E}[\tau_x^+] < \infty$.

Proof. Since P irreducible, $|\mathcal{X}| < \infty$, there exists $u \in \mathbb{N}$, $\varepsilon > 0$, s.t. $\forall x, y \in \mathcal{X}$, $\exists i \leq u$ s.t. $P^i(x, y) > \varepsilon$.

Then, no matter what X_i is, there is an ε chance to hit x between X_j and X_{j+i}. So,

$$\mathbb{P}[\tau_x^+ > kr] \leq (1-\varepsilon)\mathbb{P}[\tau_x^+ > k(r-1)] \leq (1-\varepsilon)^r.$$

Thus,

$$\begin{aligned} \mathbb{E}[\tau_x^+] &= \sum_{i \geq 0} \mathbb{P}[\tau_x^+ > i] \\ &\leq \sum_{r > 0} k\mathbb{P}[\tau_x^+ > kr] \leq \sum_{r > 0} (1-\varepsilon)^k k < \infty. \end{aligned}$$

□

Now, proof of the main theorem:

Proof. Pick $z \in \mathcal{X}$ in a closed class. Let $\pi(x) = \mathbb{E}[N_x]/\mathbb{E}[\tau_x^+]$, where N_x is the number of visits to x until we return to z. It turns out that this is a stationary distribution.

First we show that π is a distribution. Clearly, $\pi(x) \geq 0 \forall x \in \mathcal{X}$. Also, $\sum_x N_x = \tau_z^+$, which implies that $\sum_x \pi(x) = 1$, so π is a distribution.

Now, we show $\pi P = \pi$. It suffices to show

$$\sum_x \mathbb{E}[N_x] P(x,y) = \mathbb{E}[N_y] \forall y.$$

We know $N_x = \sum_{i \geq 0} \mathbb{1}(X_i = x, \tau_z^+ > i)$, which implies

$$\begin{aligned}\sum_x \mathbb{E}[N_x] P(x,y) &= \sum_x \sum_{i \geq 0} P(x,y) \mathbb{P}[X_i = x, \tau_z^+ > i] \\ &= \sum_x \sum_{i \geq 0} \mathbb{P}(X_{i+1} = y | X_i = x, \tau_z^+ > i) \cdot \mathbb{P}(X_i = x, \tau_z^+ > i).\end{aligned}$$

We can make this substitution since $\{\tau_z^+ > i\} = \{X_1 \neq z, \ldots, X_i \neq z\}$, which only depends on events in the past; by the Markov property, conditioning on past events does not affect the current probability. Now, by the law of total probability, our sum simplifies

$$\begin{aligned}&\sum_{i \geq 0} \mathbb{P}[X_{i+1} = y | \tau_z^+ \geq i+1] \\ &= \mathbb{E}[N_y] - \mathbb{P}[X_0 = y, \tau_z^+ > 0] + \sum_{i=1}^{\infty} \mathbb{P}[X_i = y, \tau_z^+ = i] \\ &= \mathbb{E}[N_y] - \mathbb{P}[X_0 = y, \tau_z^+ > 0] + \mathbb{P}(X_{\tau_z^+} = y).\end{aligned}$$

If $y = z$, the two right hand terms are equal to 1; otherwise, they are both equal to 0. Either way, the sum collapses to $\mathbb{E}[N_y]$, so we are done. □

Theorem 2.9

If P is irreducible, $|\mathcal{X}| < \infty$, there is at most one stationary distribution.

This is still true without assuming $|\mathcal{X}| < \infty$, but the proof is more difficult without this assumption, so we assume it to be true here.

Proof. Let π_1, π_2 be stationary. By HW, $\pi_1(x), \pi_2(x) > 0 \forall x$. Choose z s.t. $\pi_1(z)/\pi_2(z)$ is minimized, which is well defined since we have a finite list of positive probabilities.

$$\frac{\pi_1(z)}{\pi_2(z)} = \frac{\sum_x \frac{\pi_1(x)}{\pi_2(x)} \pi_2(x) P(x,z)}{\sum_x \pi_2(x) P(x,z)}.$$

Note that the right hand side is a weighted average of $\pi_1(x)/\pi_2(x)$ over all x. Therefore, if $\pi_1(x)/\pi_2(x) > \pi_1(z)/\pi_2(z)$ for any x with $P(x,z) > 0$, we get a contradiction, since the RHS would necessary exceed the LHS. This implies $\pi_1(x)/\pi_2(x) = \pi_1(z)/\pi_2(z) \forall x$ with $P(x,z) > 0$. We can replace P with P^i to force this to hold for all x, implying that π_1/π_2 is a constant. Since their elements must both sum to 1, this means that they're the same distribution, so we're done. □

Corollary 2.10

The unique stationary distribution for irreducible P, $|\mathcal{X}| < \infty$ is given by $\pi(x) = 1/\mathbb{E}[\tau_x^+]$.

Proof. We showed that $\pi(x) = \mathbb{E}[N_x]/\mathbb{E}[\tau_x^+]$ works for any $z \in \chi$. Since we can choose $z = x$, this gives $\pi(x) = 1/\mathbb{E}[\tau_x^+]$. □

3 September 14, 2023

3.1 Last lecture review

Defined $\text{per}(x) = \gcd\{i : P^i(x,x) > 0\}$. If $x \sim y$, then $\text{per}(x) = \text{per}(y)$. Period of irreducible P is the period of any $x \in \mathcal{X}$.

We say π is stationary if $\pi P = \pi$, which is the same as saying

$$\sum_x \pi(x) P(x,y) = \pi(y),$$

which is the same as saying $X_0 \sim \pi \implies X_1 \sim \pi$. Here, $\sim$ means "distributed as", and not communication (slightly confusing).

Theorem 3.1

If $\mathcal{X}$ finite, there exists stationary distribution π.

Theorem 3.2

If P irreducible, $|\mathcal{X}| < \infty$, π is unique.

Corollary 3.3

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

3.2 Convergence Theorem

Definition 3.4

P is **reversible** wrt μ if

$$\mu(x)P(x,y) = \mu(y)P(y,x) \quad \forall x,y \in \mathcal{X}.$$

Proposition 3.5

If P is reversible wrt μ, then $\mu P = \mu$, i.e., μ is stationary.

Warning: the converse of this proposition is false.

Example 3.6 (Birth and death chain)

$\mathcal{X} = \{0, 1, \ldots, n\}$. $P(x,y) = p_x$ if $y = x+1$, $P(x,y) = q_x$ if $y = x-1$, or r_x if $y = x$.

Assuming all probabilities positive, this is an irreducible markov chain. Let's try to find π such that P is reversible wrt π.

Consider $\mu(0)P(0,y) = \mu(y)P(y,0)$. If $y = 0$, both sides are the same; if $y > 1$, both sides are 0. So, we need $\mu(0)P(0,1) = \mu(y)P(1,0) \implies \mu(0)p_0 = \mu(1)q_1$. In general, $\mu(i)p_i = \mu(i+1)q_{i+1}$, which implies that

$$\mu(x) = \mu(x-1)\frac{p_{x-1}}{q_x} = \mu(0)\prod_{i=1}^{x}\frac{p_{i-1}}{q_i}.$$

So, the unique stationary distribution is

$$\pi(x) = \frac{\prod_{i=1}^{x}\frac{p_{i-1}}{q_i}}{\sum_x \prod_{i=1}^{x}\frac{p_{i-1}}{q_i}}$$

If $p_x = p \forall x$ and $q_x = q \forall x$, this simplifies:

$$\pi(x) = \frac{(p/q)^x(1-p/q)}{1-(p/q)^{n+1}}$$

This also tells us that

$$\mathbb{E}(\tau_x^+) = \frac{1-(p/q)^n}{(p/q)^x(1-p/q)}.$$

Recall: for discrete random variables, we say $X_i \xrightarrow[n\to\infty]{d} X$ if $\mathbb{P}[X_i = x] \xrightarrow[n\to\infty]{} \mathbb{P}[X = x]$ (for continuous r.v.s we need to use the full cdf). Similarly, $X_i \to \pi$ if $\mathbb{P}[X_i = x] \to \pi(x)$.

Theorem 3.7 (Convergence Theorem)
P irreducible, aperiodic, $|\mathcal{X}| < \infty$. Then, we know there exists unique π stationary, and

$$X_i \xrightarrow[n\to\infty]{d} \pi$$

for any starting distribution $X_0 \sim \mu$. In other words,

$$\lim_{i\to\infty} \mathbb{P}[X_i = x | X_0 \sim \mu] = \pi(x) \forall x.$$

We'll use the following proposition:

Proposition 3.8
If P irreducible, aperiodic, $|\mathcal{X}| < \infty$, then there exists $r > 0$ s.t. $\forall i \geq r$, $P^i(x,y) > 0$ for all $x, y \in \mathcal{X}$.

(no proof)

Now we're ready for the main result.

Proof. Claim: without loss of generality, we can take

$$\mu = \delta_x,$$

where $\delta_x(y) = 1$ if $y = x$ and 0 if $y \neq x$. In other words, it suffices to show that

$P^i(x,y) \to \pi(y) \forall x, y$. We want $\mu P^i(y) \to \pi(y) \forall \mu, y$. This is the same as

$$\lim_{i\to\infty} \mu P^i(y) = \sum_x \mu(x) \lim_{i\to\infty} P^i(x,y) = \pi(y),$$

which is true assuming that we show $P^i(x,y) \to \pi(y)$, call ($\star$).

Let Π be a matrix whose rows are all π. We claim that ($\star$) is equivalent to $P^i \to \Pi$. By the proposition, and the fact that the state space is finite, there exists r and $0 < \theta < 1$ such that

$$P^r(x,y) \geq (1-\theta)\pi(y).$$

Let

$$Q = \frac{1}{\theta}(P^r - (1-\theta)\Pi).$$

We claim that Q is the transition matrix of a MC. Q is a transition matrix because both P^r and Π are transition matrices, i.e., their rows sum to 1, and therefore each row of Q adds to $(1-(1-\theta)\cdot 1)/\theta = 1$. Also, we picked θ so that $Q(x,y) \geq 0$, so Q is a transition matrix.

Now, since $P^r = \theta Q + (1-\theta)\Pi$. Since $\theta < 1$, this means we always have non-zero chance of stepping towards the stationary distribution. Intuitively, this means that if we try hard enough, we'll eventually reach Π. More rigorously:

$$p^{rK} = (1-\theta^k)\Pi + \theta^k Q^k,$$

which we can prove this with induction:

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

All that remains is to show that $Q^k \Pi = \Pi$.

$$Q^k \Pi(x,y) = \sum_z Q^k(x,z)\Pi(z,y) = \sum_z Q^k(x,z)\pi(y) = \pi(y),$$

thus

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

which completes the induction.

Finally, taking the limit, we have $\lim_{k\to\infty} P^{rK} = \Pi$. In general,

$$\lim_{n\to\infty} P^n = \lim_{n\to\infty} P^{r\lfloor n/r\rfloor + (n - r\lfloor n/r\rfloor)} = \Pi.$$

□

4 September 19, 2023

4.1 Ergodic Theorem

Definition 4.1 (Stopping Time)
A **stopping time** for stochastic process X_i is a random variable T on $\mathbb{N} \cup \{\infty\}$ such that $T = i$ can be determined by $X_1, \ldots, X_i$.

For example, τ_z^+ is a stopping time, since the event $\tau_z^+ = i$ occurs only when $X_1, \ldots, X_{i-1} \neq z$ and $X_i = z$.

Proposition 4.2 (Strong Markov property)
Let X_i be a Markov chain and T be a stopping time for X_i. Given $T < \infty$ and $X_T = x$, $(X_{T+i})_{i\in\mathbb{N}}$ is distributed as $(X_i)_{i\in\mathbb{N}}$ starting from $X_0 = x$.

Proof. For T fixed, this is a restatement of the usual Markov property. Also, since T is a stopping time, fixing $T = n$ depends only on $X_0, \ldots, X_n$; therefore, by time homogeneity, this statement is also true conditioned on $T = n$ and $X_n = x$. Since this is true for all n, we are done. □

For Markov chain X_i, let $V_x(n)$ be the number of visits to x before time step n.

Theorem 4.3 (Ergodic Theorem)
Let P be irreducible with $|\mathcal{X}| < \infty$. Then,

$$\frac{V_x(n)}{n} \xrightarrow[n\to\infty]{a.s.} \frac{1}{\mathbb{E}[\tau_x^+]},$$

and for any function $f : \mathcal{X} \to \mathbb{R}$,

$$\frac{\sum_{i=0}^{n-1} f(X_i)}{n} \xrightarrow[n\to\infty]{a.s.} \overline{f},$$

where $\overline{f} = \sum_x \pi(x) f(x)$.

Proof. Fix $z \in \mathcal{X}$. Let T_i be the ith time that z is visited. Then, $T_{i+1} - T_i$ are independent and identically distributed by the Strong Markov property. Since $T_{V_z(n)} \le n$ and $T_{V_z(n)+1} \ge n$,

$$\frac{T_{V_z(n)}}{V_z(n)} \le \frac{n}{V_z(n)} \le \frac{T_{V_z(n)+1}}{V_z(n)}.$$

Let $S_n = 1/(n-1) \cdot \sum_{i=1}^{n-1} (T_{i+1} - T_i)$. By SLLN, $S_n \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[\tau_z^+]$. But note that

$$S_{V_z(n)+1} = \frac{(T_{V_z(n)+1} - T_1)}{V_z(n)} \to \frac{T_{V_z(n)+1}}{V_z(n)} \ge \frac{n}{V_z(n)},$$

while

$$S_{V_z(n)} = \frac{(T_{V_z(n)} - T_1)}{(V_z(n) - 1)} \to \frac{T_{V_z(n)}}{V_z(n)} \le \frac{n}{V_z(n)},$$

so $n/V_z(n) \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[\tau_z^+]$, as desired.

For the second part,

$$\begin{aligned}
\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) &= \frac{1}{n} \sum_{x \in \mathcal{X}} V_x(n) f(x) \\
&= \sum_{x \in \mathcal{X}} \left(\frac{V_x(n)}{n} - \frac{1}{\mathbb{E}[\tau_x^+]} \right) f(x) + \sum_{x \in \mathcal{X}} \frac{f(x)}{\mathbb{E}[\tau_x^+]} \\
&\xrightarrow[n\to\infty]{a.s.} \overline{f},
\end{aligned}$$

by the first part. □

5 September 21, 2023

5.1 Metropolis-Hastings

Definition 5.1 (Ising Model)
Let $G = (V, E)$ be graph. Let $\mathcal{X} = \{-1, 1\}^{|V|}$. The **Ising model** with inverse temperature β is the distribution on $\mathcal{X}$ with

$$\mu(\sigma) = \frac{1}{Z_\beta} \exp\left(\beta \sum_{(v,w)\in E} \sigma(v)\sigma(w)\right),$$

where Z_β is a normalization constant.

Sampling from this distribution is very expensive; to compute the normalization constant, we have to sum over all $2^{|V|}$ possible σ. In general, suppose μ is a distribution on $\mathcal{X}$ which is computationally intractable. Can we find an algorithm to approximately sample from μ? Basic idea: create a Markov chain P whose stationary distribution is μ, and then run the Markov chain for a long time, and hope that we are close to μ.

Definition 5.2 (Metropolis-Hastings)
Let P be a markov chain and μ a distribution on $\mathcal{X}$. Assume $\mu(x) > 0$ for all x. The Metropolis MC wrt P and μ has transition matrix

$$\hat{P}(x,y) = P(x,y)\min\left(1, \frac{\mu(y)P(y,x)}{\mu(x)P(x,y)}\right),$$

whenever $x \neq y$, and $\hat{P}(x,x)$ is defined so that all the rows add to 1.

First note that this is a valid transition matrix, since

$$\sum_{y\neq x} \hat{P}(x,y) \leq \sum_{y\neq x} P(x,y) = 1 - P(x,x) \leq 1,$$

so we can always choose $\hat{P}(x,x)$ so that the rows sum to 1.

Proposition 5.3
Let $\hat{P}$ be the metropolis chain with respect to P and μ. $\hat{P}$ is reversible wrt μ, which implies that μ is stationary.

Proof. We want to show that

$$\mu(x)\hat{P}(x,y) = \mu(y)\hat{P}(y,x).$$

This is true when $x = y$, so assume $x \neq y$. Then, plugging in known values, we want to show

$$\mu(x)P(x,y)\min\left(1, \frac{\mu(y)P(y,x)}{\mu(x)P(x,y)}\right) = \mu(y)P(y,x)\min\left(1, \frac{\mu(x)P(x,y)}{\mu(y)P(y,x)}\right).$$

This is always true, since exactly one of the mins will be 1. □

Lemma 5.4
If P is irreducible and $P(x,y) > 0$ if and only if $P(y,x) > 0$, then $\hat{P}$ is irreducible.

Proof. $\hat{P}(x,y) > 0$ and $\hat{P}(y,x) > 0$ given $P(x,y) > 0$ and $P(y,x) > 0$, meaning that $\hat{P}$ has the same communicating classes. □

5.2 Gibbs Sampling

Note that transition probabilities are much easier to compute now, since $\mu(x)/\mu(y)$ is generally much easier to compute than either of $\mu(x)$ or $\mu(y)$ individually. However, this still does not allow us to sample efficiently from the Ising model, since our MC would have $2^{|V|}$ nodes. We want an easier way to progress through large MCs given transition probabilities.

Definition 5.5 (Gibbs Sampling)
Let $\mathcal{X} = S^n$ for some set S and $n > 0$. Let μ be a distribution on $\mathcal{X}$. The **Gibbs Sampler** associated with μ is the MC starting from $(x_1, \ldots, x_n) \in \mathcal{X}$ and moving randomly:

1. Pick $I \in [n]$ randomly
2. Sample X according to

$$P(X = x) = \frac{\mu(x_1, \ldots, x, \ldots, x_n)}{\sum_y \mu(x_1, \ldots, y, \ldots, x_n)},$$

   where x and y both appear in the Ith coordinate.

3. Move to $(x_1, \ldots, X, \ldots, x_n)$, where X replaces x_i.

In the example of the Ising model, $S = \{-1, 1\}$, and we progress through the MC by randomly sampling a specific node, then flipping/keeping its value.

Proposition 5.6
Let $\hat{P}$ be the Gibbs sampler for μ. Then $\hat{P}$ is reversible wrt μ.

Proof. We want to show that

$$\mu(x_1, \ldots, x_n)\hat{P}((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \mu(y_1, \ldots, y_n)\hat{P}((y_1, \ldots, y_n), (x_1, \ldots, x_n)).$$

If all coordinates are equal, both sides are the same. Otherwise, we only need to consider pairs of states who differ in exactly one coordinate, since otherwise the transition probabilities are zero.

Then WLOG $x_1 \neq y_1$; both the LHS and RHS evaluate to

$$\frac{1}{n} \frac{\mu(x_1, \ldots, x_n)\mu(y_1, \ldots, y_n)}{\sum_z \mu(z, x_2, \ldots, x_n)}.$$

(the $1/n$ comes from the fact that we have choose the first coordinate randomly when we transition). Since LHS=RHS, we are done. □

Example 5.7
Gibbs sampling on the Ising model is called **Glauber dynamics**.

To perform Gibbs sampling on the Ising model:

1. pick vertex $v \in V$ at random

2. $\mu(v_1, \ldots, v, \ldots, v_n)$ is a product of exps. since we only care about ratios of μ between two states, we can ignore the normalization constant and terms that do not involve v. this means we can replace $\sigma(v)$ with either ± 1, equal to 1 with probability

$$\frac{\exp(\beta \sum_{(w,v)\in E} \sigma(w))}{\exp(-\beta \sum_{(w,v)\in E} \sigma(w)) + \exp(\beta \sum_{(w,v)\in E} \sigma(w))}.$$

3. transition to the new state.

6 September 26, 2023

6.1 Total Variation Distance

Definition 6.1 (Total Variation Distance)
Let μ and ν be two distributions on $\mathcal{X}$. The **total variation distance**, $d_{TV}(\mu, \nu)$ is given by

$$d_{TV}(\mu, \nu) = \sup_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)|.$$

We can use μ and ν in the definition of d_{TV} interchangeably with random variables distributed as μ and ν respectively.

Example 6.2
If X and Y are Bernoulli random variables with parameters p, q respectively, then $d_{TV}(X, Y) = |p - q|$.

In this example, $\mathcal{X} = \{0, 1\}$. For all possible $A \subseteq X$, the difference in their probability is at most $|p - q|$.

Proposition 6.3
Total variation distance is a distance metric, along with some other properties:

- $d_{TV}(\mu,\nu) = 0$
- $d_{TV}(\mu,\nu) = d_{TV}(\nu,\mu)$
- Triangle inequality: $d_{TV}(\mu,\nu) \le d_{TV}(\mu,\eta) + d_{TV}(\eta,\nu)$
- $$d_{TV}(\mu,\nu) = \frac{1}{2}\sum_{x\in\mathcal{X}}|\mu(x)-\nu(x)| = \sum_{x:\mu(x)>\nu(x)}\mu(x)-\nu(x).$$
- $X_n \xrightarrow[n\to\infty]{d} X$ if and only if $d_{TV}(X_n, X) \to 0$.

Proof. Proof of third bullet point: think of μ,ν like a bar graph. Shade all area above μ and below ν red, and shade all area above ν and below μ blue. The maximal difference $\mu(A)-\nu(A)$ is achieved by collecting all the blue area, which is the expression on the right side. Since μ and ν are distributions, the red and blue areas are equal; since the middle expression is the sum of both areas, it is also equal to the right hand side. □

The goal of defining a total variation distance is to eventually try to compute

$$d_{TV}(\mu P^i, \pi).$$

As i increases, the total variation distance to the stationary distribution should decrease. To help us understand this more concretely, we define a notion of **coupling**.

Definition 6.4 (Coupling)
A **coupling** of two distributions μ and ν on probability space $\mathcal{X}$ is a joint distribution η on $\mathcal{X}\times\mathcal{X}$ whose marginals are μ and ν respectively.

A coupling of random variables X and Y is a random variable $(\tilde{X},\tilde{Y})$ for which $X\sim\tilde{X}$ and $Y\sim\tilde{Y}$.

Example 6.5
Let X and Y be $\textsc{Bern}(p)$ random variables.

Then, the independent coupling is given by

$$\mathbb{P}[(\tilde{X},\tilde{Y})=(x,y)]=\begin{cases}(1-p)^2 & (x,y)=(0,0)\\ p(1-p) & (x,y)\in\{(0,1),(1,0)\}\\ p^2 & (x,y)=(1,1).\end{cases}$$

Another coupling is to take $\tilde{X}=\tilde{Y}$:

$$\mathbb{P}[(\tilde{X},\tilde{Y})=(x,y)]=\begin{cases}1-p & (x,y)=(0,0)\\ p & (x,y)=(1,1)\\ 0 & (x,y)\in\{(0,1),(1,0)\}.\end{cases}$$

Example 6.6
Consider the finite Gambler's ruin MC with n states. Let $X_0=x$, $Y_0=y$, with $x\le y$. Show that for some i, $\mathbb{P}[X_i=n|X_0=x]\le\mathbb{P}[Y_i=n|Y_0=y]$.

We will use a coupling $(\tilde{X}_i,\tilde{Y}_i)$ with $\tilde{X}_0=x$ and $\tilde{Y}_0=y$ that move left/right in parallel. This is a valid coupling, since the marginal distribution of each variable is the same as their individual distributions. Now,

$$\mathbb{P}[X_i=n]=\mathbb{P}[\tilde{X}_i=n]=\mathbb{P}[\tilde{X}_i=n,\tilde{Y}_i=n]\le\mathbb{P}[\tilde{Y}_i=n]=\mathbb{P}[Y_i=n].$$

Proposition 6.7
$d_{TV}(\mu,\nu)\le\inf\{\mathbb{P}[X\ne Y]:(X,Y)\text{ is a coupling of }\mu,\nu\}$.

Proof.

$$\mu(A)-\nu(A)=\mathbb{P}[X\in A]-\mathbb{P}[Y\in A]\le\mathbb{P}[X\in A,Y\ne A]\le\mathbb{P}[X\ne Y].$$

□

This is always an inequality; there always exists coupling (X,Y) such that $\mathbb{P}[X\ne$

$Y] = d_{TV}(\mu, \nu)$. We will not prove this, but one such example is to take $p \leq q$, $U \sim \textsc{Unif}[0,1]$, $X = \mathbb{1}_{u \leq p}$, $Y = \mathbb{1}_{u \leq q}$.

Theorem 6.8 (Convergence Theorem using Coupling)
P irreducible, aperiodic, $|\mathcal{X}| < \infty$. Then, we know there exists unique π stationary, and

$$X_i \xrightarrow[n\to\infty]{d} \pi$$

for any starting distribution $X_0 \sim \mu$. In other words,

$$\lim_{i\to\infty} \mathbb{P}[X_i = x | X_0 \sim \mu] = \pi(x) \forall x.$$

Proof. This is equivalent to showing that $d_{TV}(\mu P^i, \pi) \to 0$ as $i \to \infty$. Construct coupling (X_i, Y_i) with $X_0 \sim \mu$ and $Y_i \sim \pi$. Then, consider independently X_i' and Y_i' starting from μ, π respectively, and let T be the first time that $X_i' = Y_i'$. Let $X_i = X_i'$ and $Y_i = Y_i'$ for all $i \leq T$ and $X_i = Y_i = X_i'$ for all $i > T$. This pairing has the correct marginal distributions, so it is a coupling.

Now, we have

$$d_{TV}(\mu P^i, \pi) \leq \mathbb{P}[X_i \neq Y_i] = \mathbb{P}[T > i].$$

Note that (X_i, Y_i) is an MC on $\mathcal{X} \times \mathcal{X}$ with (x, x) its only closed class. Therefore, $\mathbb{P}[T > i]$ is the probability that we have not entered this closed class by time i, which approaches 0 as $i \to \infty$, hence done. □

7 September 28, 2023

7.1 Coupling Example

Example 7.1
Consider a lazy random walk on a hypercube, where "lazy" means that each step stays in the same place with $p = 1/2$ and otherwise travels to a uniformly randomly selected neighbor with $p = 1/2$.

The uniform distribution $\pi(x) = 2^{-n}$ is stationary. Let (X_i, Y_i) be two independent copies of the Markov chain, where $X_0 = \vec{0}$ and Y_0 is drawn from π. We can use coupling to generate intuition on how long it takes until $X_i = Y_i$.

Consider Z_i which has coordinate 0 if and only if X_i and Y_i agree in that coordinate. When X_i makes a step, Z_i has $p = 1/2$ of staying the same and $p = 1/2$ of flipping a coordinate; the same is true when Y_i takes a step. Therefore Z_i is equivalent to the original markov chain when taking two steps at a time. Also, $Z_0 \sim \pi$, and finding how long it takes for $X_i = Y_i$ is the same as finding how long it takes until $Z_0 = \vec{0}$. We know $\mathbb{E}[\tau^+_{\vec{0}}] = 2^n$, so it'll take around 2^n steps at most.

Now consider a non-independent coupling. Define (X_i, Y_i) through the following joint process: randomly select a coordinate, and then set that coordinate to be 0 or 1 with equal probability in both X_i and Y_i simultaneously. The marginals X_i and Y_i each follow the original Markov chain, so this is a valid coupling. Also, $X_i = Y_i$ only after every coordinate has been selected at least once, so

$$d_{TV}(\mu P^i, \pi) \leq \mathbb{P}[X_i \neq Y_i] \leq \mathbb{P}[T > i],$$

which turns into the coupon collector problem.

7.2 Lower bound on variation distance

Proposition 7.2

Let P be an irreducible, aperiodic Markov chain and π stationary. Let $A \subseteq \mathcal{X}$ be the set of states which cannot be reached from x in i steps. Then,

$$d_{TV}(P^i(x, \cdot), \pi) \geq \pi(A).$$

Proof.

$$d_{TV}(P^i(x, \cdot), \pi) \geq |\pi(A) - P^i(x, A)| = \pi(A).$$

□

Consider the previous example. After taking $i = n/2$ from $\vec{0}$, the set of reachable states A_i has size at least 2^{n-1}, so

$$d_{TV}(P^i(\vec{0}, \cdot), \pi) \geq \pi(A_i) \geq 2^{n-1}/2^n = 1/2.$$

In other words, after taking $n/2$ steps, we are still "far away" from the stationary distribution.

7.3 Random walk on binary tree

Definition 7.3

A binary tree of **depth** n is a graph with vertices representing binary strings of length at most n, including the empty word ($2^{n+1}-1$ nodes total). Edges exist between nodes such that one can be obtained from the other by adding a 0 or 1.

A lazy random walk on a binary tree remains stationary with $p = 1/2$, or moves to an adjacent node randomly with $p = 1/2$. We want to bound $d_{TV}(P^i(x,\cdot),\pi)$.

Consider the following coupling (X_i, Y_i):

- first, pick one of X_i or Y_i to move, with the other staying. repeat until both are on the same level.
- after the first stage, X_i and Y_i always move or stay together.

Each marginal distribution of both stages is equivalent to the original Markov chain, so this is a valid coupling. Based on the previous lower bound, we can make some heuristics about how long it takes until $X_i = Y_i$:

- the random walk that starts with the lower level will never, at any point, exceed the level of the other walk. Therefore, once the walk who starts with higher level reaches the root, $X_i = Y_i$.
- we will prove in the HW that we can project this coupled Markov chain onto a birth and death chain on the level of each walk.
- starting from root x_0, a random walk can reach at most level i in i steps. Let A be the set of all such vertices. recall that $\pi(v) = \deg(v)/2|E|$, since this is a graph. therefore,
$$\pi(A) \le \frac{3|A|}{2^{n+2}-4}.$$
this implies
$$d_{TV}(P^i(x_0,\cdot),\pi) \ge \pi(A^c) \ge 1 - \frac{3(2^{i+1}-1)}{2^{n+2}-4}.$$
If i is small, this distance is large, so we intuitively need a lot of steps to get close.

8 October 3, 2023

8.1 Random Lattice Walks

Lemma 8.1
Let P be a Markov Chain, and N the number of times that starting from x, it visits x. Then, $\mathbb{P}[N = \infty] = 1$, or $\mathbb{P}[N < \infty] = 1$, which occurs when $\mathbb{P}[\tau_x^+ < \infty] = 1$ or $\mathbb{P}[\tau_x^+ < \infty] < 1$, respectively.

Proof. $\mathbb{P}[\tau_x^+ < \infty]$ represents the probability that we revisit x in a finite amount of time. If this occurs certainly, then we will visit x infinite times; otherwise, N is a geometric sum with parameter < 1, which is finite. □

Example 8.2
Consider a walk on $\mathbb{Z}^d$ with $d = 1$. We want to know whether it will return to 0 infinitely often.

On the number line,

$$\mathbb{E}[N] = \sum_{i=0}^{\infty} \mathbb{P}[X_i = 0] = \sum_{i=0}^{\infty} \frac{1}{2^{2i}} \binom{2i}{i},$$

which sums the probability of seeing an equal number of left and right moves in all sequences of length $2i$. Using Stirling's approximation,

$$\binom{2i}{i} \sim \frac{n^{2n}\sqrt{4\pi n}}{(n/2)^{2n} 2\pi n} = \frac{4^n}{\sqrt{\pi n}},$$

so

$$\mathbb{E}[N] \sim \sum_{i=0}^{\infty} \frac{1}{4^n} \cdot \frac{4^n}{\sqrt{\pi n}} = \infty.$$

In other words, we visit N infinitely often, which implies $\mathbb{P}[\tau_0^+ < \infty] = 1$. On the other hand, we can show that $\mathbb{E}[\tau_0^+] = \infty$. If we let τ_x^y denote the time taken to hit x starting from y, we have

$$\mathbb{E}[\tau_0^+] = \frac{1}{2}\mathbb{E}[\tau_0^1] + \frac{1}{2}\mathbb{E}[\tau_0^{-1}] + 1 = \mathbb{E}[\tau_0^1] + 1,$$

and

$$\mathbb{E}[\tau_0^1] = \frac{1}{2}\mathbb{E}[\tau_0^2] + 1 = \frac{1}{2}\mathbb{E}[\tau_1^2 + \tau_0^1] + 1 = \mathbb{E}[\tau_0^1] + 1,$$

hence $\mathbb{E}[\tau_0^1] = \mathbb{E}[\tau_0^+] = \infty$. This is a counterintuitive result and only possible because our state space is infinite.

Example 8.3

$d = 2, 3$.

omitting other proofs. add later?

Lemma 8.4

Random walks on $\mathbb{Z}^d$ for $d = 1, 2$ return to 0 infinitely often with probability 1. When $d \geq 3$, the walks return to 0 finitely often. $\mathbb{E}[\tau_0^+] = \infty$ for all d.

Definition 8.5

Let P be a Markov chain with countable state space $\mathcal{X}$. $x \in \mathcal{X}$ is **recurrent** if P, starting from x, visits x infinitely often with probability 1. x is **transient** if it only visits x finitely many times with probability 1.

9 October 5, 2023

We'll start to focus more on Markov Chains with countably infinite state spaces, rather than strictly finite state spaces.

Definition 9.1

Let G be a countably infinite graph which is **locally finite**. This means that $\deg(v) < \infty$ for all $v \in \mathcal{X}$. We can define a random walk on G in the same way as the finite case.

Recall from last lecture:

Definition 9.2

A state $x \in \mathcal{X}$ is recurrent if P, starting from x, visits x infinite often with probability 1. x is transient if it only visits x finitely many times with probability 1.

9.1 More on Transience and Recurrence

Definition 9.3

Let P be a Markov chain. Then, **Green's function** $G : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is given by

$$G(x,y) = \sum_{i=0}^{\infty} P^i(x,y),$$

which is equal to the expected number of visits to y starting from x.

In particular, x is recurrent if and only if $G(x,x) = \infty$.

Proposition 9.4

Suppose $x \sim y$. Then, the following are true:

- $G(z,x) < \infty \iff G(z,y) < \infty$
- $G(x,z) < \infty \iff G(y,z) < \infty$

Proof. Since $x \sim y$, there exists r s.t. $P^r(y,x) > 0$. First bullet point:

$$G(z,y)P^r(y,x) = \sum_{i=0}^{\infty} P^i(z,y)P^r(y,x) \le \sum_{i=0}^{\infty} P^{i+r}(z,x) \le G(z,x).$$

Therefore, if $G(z,x) < \infty$, so is $G(z,y)$. This argument is reversible.

The second bullet point follows from the same argument:

$$P^s(x,y)G(y,z) = \sum_{i=0}^{\infty} P^s(x,y)P^i(y,z) \le \sum_{i=0}^{\infty} P^{i+s}(x,z) \le G(x,z).$$

□

Corollary 9.5

Transience and recurrence are class properties.

Recall that a class property is a property that holds for $x \in C$ if and only if it holds for every other element in C.

Proof. If x is transient, $G(x,x) < \infty \iff G(y,x) < \infty \iff G(y,y) < \infty$ for all $y \sim x$ by the previous proposition. Similarly, if x is recurrent, $G(x,x) = \infty \iff G(y,x) = \infty \iff G(y,y) = \infty$ for all $y \sim x$. □

Corollary 9.6

Let P be an irreducible Markov chain. The following are equivalent:

- $G(x,y) < \infty$ for some $x, y \in \mathcal{X}$
- $G(x,y) < \infty$ for all $x, y \in \mathcal{X}$
- There is a transient state
- All states are transient
- $\mathbb{P}[\tau_x^+ = \infty | X_0 = x] > 0$ for some $x \in \mathcal{X}$
- $\mathbb{P}[\tau_x^+ = \infty | X_0 = x] > 0$ for all $x \in \mathcal{X}$

Proof. This is essentially a restatement of the previous proposition:

- $1 \iff 2$ follows directly by the proposition.
- $3 \iff 1$ follows by definition, as does $2 \iff 4$.
- $\mathbb{P}[\tau_x^+ = \infty | X_0 = x] = 1 - \mathbb{P}[\tau_x^+ < \infty | X_0 = x] > 0 \implies \mathbb{P}[\tau_x^+ < \infty | X_0 = x] < 1$, which we showed last lecture was equivalent to x being transient.

□

By the above Corollary, we can now say:

Definition 9.7

An irreducible Markov Chain P is **recurrent** if it has a recurrent state. It is **transient** if it has a transient state.

Proposition 9.8

If $x \in C$ is recurrent, C must be closed.

Proof. Suppose there exists $z \in C$ and $y \notin C$ s.t. $P(z, y) > 0$. Since recurrence is a class property, z must also be recurrent. This is not possible given non-zero possibility of escaping the class. □

9.2 Positive / Null recurrence

Definition 9.9

If $x \in \mathcal{X}$ is recurrent, it is **positive recurrent** if $\mathbb{E}[\tau_x^+] < \infty$. Otherwise, it is null recurrent.

For example:

- Random walks on $\mathbb{Z}^d$ for $d = 1, 2$ returns to 0 infinitely often. On the other hand, we also showed $\mathbb{E}[\tau_x^+] = \infty$, so this is an example of a null recurrent MC.
- Recurrent MCs on finite state spaces are positive recurrent.

Lemma 9.10 (Wald's Lemma)

If Z_i are independent and K is a stopping time wrt Z_i, T_i a function of $Z_0, \ldots, Z_i$ such that T_i are identically distributed, then

$$\mathbb{E}\left(\sum_{i=1}^{K} T_i\right) = \mathbb{E}(K)\mathbb{E}(T_1).$$

We will prove a generalized version of Wald's Lemma later in the class.

Proposition 9.11

Positive/null recurrence are class properties. In particular, z positive recurrent implies $\tau_y^x = \mathbb{E}[\tau_y^+ | X_0 = x] < \infty$ for all $x, y \sim z$.

Proof. Recurrence is a class property, and further recurrent states can only be positive or null recurrent. Therefore, the second part of the proposition implies the first, so it suffices to prove only the second part.

Assume z positive recurrent, which implies x, y recurrent. Now,

$$\mathbb{E}[\tau_z^+] \geq \mathbb{P}[\tau_x < \tau_z^+ | X_0 = z]\mathbb{E}[\tau_z^+ | \tau_x < \tau_z^+, X_0 = z].$$

Also, $\mathbb{E}[\tau_z^+ | \tau_x < \tau_z^+, X_0 = z] \geq \mathbb{E}[\tau_z | X_0 = x] = \mathbb{E}[\tau_z^x]$, since we have to travel from $z \to x \to z$ in the first expectation. Therefore,

$$\mathbb{E}[\tau_z^+] \geq \mathbb{P}[\tau_x < \tau_z^+ | X_0 = z]\mathbb{E}[\tau_z^x].$$

Since $x \sim z$, we have $\mathbb{P}[\tau_x < \tau_z^+ | X_0 = z] > 0$, and thus $\mathbb{E}[\tau_z^x] < \infty$.

Now, we can finish with Wald's Lemma. Let K be the number of visits to z before hitting y, starting from x. After hitting z for the first time, K is geometric with common ratio $\mathbb{P}[\tau_z^+ < \tau_y^z] < 1$, so $\mathbb{E}[K] < \infty$. Define $T_0 = \tau_z^x$ and T_i the time it takes to hit z for the $(i+1)$th time. Define Z_i as the series of steps taken between T_i and T_{i+1}. Clearly, T_i is a function of $Z_0, \ldots, Z_i$, and also $T_{i+1} - T_i$ are independent by the strong Markov property, so we have

$$\mathbb{E}[\tau_y^x] = \mathbb{E}\left[T_0 + \sum_{i=1}^{K-1} T_i\right] = \mathbb{E}[\tau_z^x] + \mathbb{E}[K-1]\mathbb{E}[\tau_z^+] < \infty.$$

□

10 October 12, 2023

Last time, we proved that positive/null recurrence is a class property. Therefore, we may say:

Definition 10.1
Irreducible Markov Chain P is positive recurrent if it has a positive recurrent state. It is null recurrent if it has a null recurrent state.

10.1 Stationary Measures

Definition 10.2
A **measure** on countable set $\mathcal{X}$ is a function $\mu : \mathcal{X} \to \mathbb{R}_{\geq 0}$.

We assume all of our measures are non-zero, i.e., there exists x with $\mu(x) > 0$. Unlike a distribution, we do not require $\sum_{x \in \mathcal{X}} \mu(x) = 1$.

Definition 10.3

A **stationary measure** on Markov Chain P is a measure π s.t. $\pi P = \pi$.

All stationary distributions are stationary measures, and so are all of their scalar multiples. However, even if P is irreducible, stationary measures may not be unique.

Proposition 10.4

Given measure π, P is said to be reversible wrt π if $\pi(x)P(x,y) = \pi(y)P(y,x)$. All reversible measures are also stationary.

Proof. Same proof as for reversible distributions. □

Proposition 10.5

If P has a recurrent state, it also has a stationary measure.

Proof. In the proof of stationary distribution with $|\mathcal{X}| < \infty$, we showed that $\pi(x) = \mathbb{E}(N_x)/\mathbb{E}(\tau_x^+)$ was a stationary distribution, where we defined N_x as the number of visits to x before returning to z. So, $\pi(x) = \mathbb{E}(N_x)$ is a stationary measure, as long as $\mathbb{E}(N_x) < \infty$. Suppose z is a recurrent state. Then, N_x is geometric with common ratio $\mathbb{P}[\tau_x < \tau_z] < 1$, since z is recurrent, so $\mathbb{E}(N_x) < \infty$. □

Proposition 10.6

If P is irreducible and recurrent, then all stationary measures are scalar multiples of each other.

Proof. Let $\mu(x)$ be a stationary measure. We will prove in HW that $\mu(x) > 0$ for all x. Scale μ so that $\mu(z) = 1$ for some $z \in \mathcal{X}$. By the previous proposition, we know that $\pi(x) = \mathbb{E}[N_x]$ is also a stationary measure, and further, $\pi(z) = \mathbb{E}[N_z] = 1 = \mu(z)$.

We now show that $\pi(x) = \mu(x)$ for all x:

$$
\begin{aligned}
\mu(x) &= P(z,x) + \sum_{y_0 \neq z} \mu(y_0)P(y_0,x) \\
&= P(z,x) + \sum_{y_0 \neq z} P(y_0,x)\left(P(z,y_0) + \sum_{y_1 \neq z} \mu(y_1)P(y_1,y_0)\right) \\
&= P(z,x) + \sum_{y_0 \neq z} P(z,y_0)P(y_0,x) + \sum_{y_0,y_1 \neq z} \mu(y_1)P(y_1,y_0)P(y_0,x) \\
&= \dots \\
&= P(z,x) + \sum_{y_0 \neq z} P(z,y_0)P(y_0,x) + \dots + \sum_{y_0,y_1,\dots,y_k \neq z} \mu(y_k)P(y_k,y_{k-1})\dots P(y_0,x) \\
&\geq \sum_{i=1}^{k} \mathbb{P}(X_i = x, \tau_z^+ \geq i | X_0 = z),
\end{aligned}
$$

As $k \to \infty$, this final expression approaches $\mathbb{E}[N_x]$, so $\nu(x) = \mu(x) - \mathbb{E}[N_x] \geq 0$ is another stationary measure. Since we know $\nu(z) = 0$, and P is irreducible, we must have $\nu(x) = 0$ for all x, so $\mu = \pi$ as desired. □

Proposition 10.7

If P is irreducible and has a stationary distribution, it is positive recurrent.

Proof. P must be recurrent, because

$$\sum_x \pi(x)G(x,z) = \sum_{i \geq 0} \sum_x \pi(x)P^i(x,z) = \sum_{i \geq 0} \pi(z) = \infty,$$

implying at least one $x \in \mathcal{X}$ with $G(x,z) = \infty$. By Corollary 9.6, this implies P recurrent.

We show $\pi(x) = 1/\mathbb{E}[\tau_x^+]$, which suffices because we know $\pi(x) > 0$.

$$\begin{aligned}
\pi(x)\mathbb{E}[\tau_x^+] &= \sum_i \mathbb{P}[\tau_x^+ \geq i, X_0 = x | X_0 \sim \pi] \\
&= \mathbb{P}[\tau_x^+ \geq 1, X_0 = x | X_0 \sim \pi] + \sum_{i \geq 2} \mathbb{P}[X_{i-1} \neq x, \ldots, X_1 \neq x, X_0 = x | X_0 \sim \pi] \\
&= \pi(x) + \sum_{i \geq 2} (\mathbb{P}[X_{i-1} \neq x, \ldots, X_1 \neq x | X_0 \sim \pi] - \mathbb{P}[X_{i-1} \neq x, \ldots, X_1 \neq x, X_0 \neq x | X_0 \sim \pi]) \\
&= \pi(x) + \mathbb{P}[X_1 \neq x | X_0 \sim \pi] \\
&\quad + \sum_{i \geq 2} (\mathbb{P}[X_i \neq x, \ldots, X_1 \neq x | X_0 \sim \pi] - \mathbb{P}[X_{i-1} \neq x, \ldots, X_1 \neq x, X_0 \neq x | X_0 \sim \pi]) \\
&= \pi(x) + \mathbb{P}[X_1 \neq x | X_0 \sim \pi] \\
&= \pi(x) + \mathbb{P}[X_0 \neq x | X_0 \sim \pi] = 1,
\end{aligned}$$

where the second to last equality follows by the Markov property, and the last equality follows by the fact that π is a stationary distribution. □

Corollary 10.8
If P is irreducible and positive recurrent, there exists a unique stationary distribution.

11 October 17, 2023

11.1 Convergence theorem on countable MCs

Theorem 11.1
Let P be an irreducible, aperiodic MC, with $\mathcal{X}$ countable.

- If P is positive recurrent and π is its unique stationary distribution, then $d_{TV}(P^i(x,\cdot), \pi) \to 0$ as $i \to \infty$.
- If P is null recurrent, then $P^i(x,y) \to 0$ for all i.

Proof. Let $(X_i, Y_i) \in \mathcal{X} \times \mathcal{X}$ be a Markov Chain with transition matrix $\tilde{P}((x,y),(x',y')) = P(x,x')P(y,y')$. Since P is aperiodic and irreducible, so is $\tilde{P}$. Also, $\tilde{\pi}(x,y) = \pi(x)\pi(y)$

is a stationary distribution, since

$$\begin{aligned}\sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} \tilde{\pi}(x,y)\tilde{P}((x,y),(x',y')) &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} \pi(x)\pi(y)P(x,x')P(y,y') \\ &= \pi(x')\pi(y') = \tilde{\pi}((x',y')).\end{aligned}$$

This implies $\tilde{P}$ positive recurrent, so the first time T that $X_i = Y_i$ is finite almost surely. Therefore, we can construct coupling (X_i, Y_i) with $X_0 = x, Y_0 \sim \pi$ such that they move independently until $i = T$, and then move together thereafter. Then,

$$d_{TV}(P^i(x,\cdot),x) \leq \mathbb{P}[T > i],$$

which goes to 0 as $i \to \infty$, which proves the first part of the theorem.

Now, let μ be a stationary measure. Since P is irreducible, $\mu(x) > 0$ is a class property, so μ is non-zero everywhere. Rescale so that $\mu(y) = 1$.

Define $\tilde{P}$ in the same way as before. If $\tilde{P}$ is transient, $\tilde{G}((x,x),(y,y)) = \sum_{i=0}^{\infty} \tilde{P}^i((x,x),(y,y)) = \sum_{i=0}^{\infty} P^i(x,y)^2 < \infty$, implying $P^i(x,y) = 0$ as $i \to \infty$, so we're done.

Therefore, let $\tilde{P}$ be recurrent. Since P is null recurrent, $\mu(\mathcal{X}) = \infty$, so fix some large M and let $A \subseteq \mathcal{X}$ such that $\mu(A) > M$. Define $\mu_A(z) = \mu(z)/\mu(A)$ if $z \in A$ and 0 otherwise; note that μ_A is a distribution.

Now, use the same coupling as in the first part of the proof, where $X_0 = x$ and $Y_0 \sim \mu_A$. Then, $P^i(x,y) = \mathbb{P}[\tau_{(x,x)} > i]\mathbb{P}[X_i = y|\tau_{(x,x)} > i] + \mathbb{P}[\tau_{(x,x)} \leq i]\mathbb{P}[X_i = y|\tau_{(x,x)} \leq i] \leq \mathbb{P}[\tau_{(x,x)} > i] + \mathbb{P}[Y_i = y]$. Since P is recurrent, $\mathbb{P}[\tau_{(x,x)} > i]$ as $i \to \infty$. Moreover, $P^i[Y_i = y] = \mu_A P^i(y) \leq \mu P^i(y)/\mu(A) \leq 1/M$. Since this holds for all $M > 0$, $\lim_{i\to\infty} P^i(x,y) = 0$, as desired. □

Lemma 11.2

For transient P, the second statement of the above theorem holds.

Proof. If P is transient, $G(x,y) = \sum P^i(x,y) < \infty$, so $P^i(x,y) \to 0$ as $i \to \infty$. □

Example 11.3

Random walks on $\mathbb{Z}^d$ are either transient or null recurrent, since the uniform measure always works. Therefore, the convergence theorem for countable MCs gives $P^i(x,y) \to 0$.

One way to think about this intuitively is that mass escapes to infinity on $\mathbb{Z}^d$.

12 October 19, 2023

12.1 Ergodic theorem on countable MCs

Theorem 12.1

Let P be irreducible. For any starting distribution μ,

- $$\mathbb{P}\left(\frac{V_x(n)}{n} \to \frac{1}{\mathbb{E}[\tau_x^+]}\right) = 1.$$
- If P is positive recurrent, $\pi P = \pi$, and $f : \mathcal{X} \to \mathbb{R}$ is bounded, then
$$\mathbb{P}\left[\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \to \mathbb{E}_\pi(f)\right] = 1.$$
In other words, $V_x(n)/n \xrightarrow[n\to\infty]{a.s.} 1/\mathbb{E}[\tau_x^+]$ and $\sum_{i=0}^{n-1} f(X_i)/n \xrightarrow[n\to\infty]{a.s.} \mathbb{E}_\pi(f)$.

Remember that $V_x(n)$ is the number of visits to x up to but not including time n. This is the exact same as the normal Ergodic theorem.

Proof. **write down the proof later**. □

Corollary 12.2

If P is null recurrent or transient, and $\mathbb{E}[\tau_x^+] = \infty$, then $V_x(n)/n \to 0$ almost surely.

12.2 Expected values

Definition 12.3

Let X,Y by random variables with X in $\mathbb{R}$. Then

$$\mathbb{E}[X|Y] = \sum_y \mathbb{E}[X|Y = y]\mathbb{1}(Y = y).$$

Note that $\mathbb{E}[X|Y]$ is a function (with respect to the random variable Y), and not a single value.

Proposition 12.4

- If $X = f(Y), \mathbb{E}[X|Y] = X$.
- If X,Y are independent, $\mathbb{E}[X|Y] = \mathbb{E}[X]$.
- $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.
- If X is jointly independent of Y, Z, then $\mathbb{E}[XY|Z] = \mathbb{E}[X]\mathbb{E}[Y|Z]$.
- If $Y = f(Z)$, $\mathbb{E}[XY|Z] = \mathbb{E}[X|Z]Y$.
- If $Y = f(Z)$, $\mathbb{E}[\mathbb{E}[X|Y]|Z] = \mathbb{E}[\mathbb{E}[X|Z]|Y] = \mathbb{E}[X|Y]$.

13 October 24, 2023

quiz review!

14 October 31, 2023

Last time:

$$\mathbb{E}[X|Y] = \sum_y \mathbb{E}[X|Y = y]\mathbb{1}_{Y=y}.$$

Also,

$$\mathbb{E}[f(Y)|Y] = f(Y),$$

and

$$\mathbb{E}[X|Y] = \mathbb{E}[X],$$

if X, Y independent, and

$$\mathbb{E}[Xf(Y)|Y] = \mathbb{E}[X|Y]f(Y),$$

and

$$\mathbb{E}[\mathbb{E}[X|f(Y)]|Y] = \mathbb{E}[\mathbb{E}[X|Y]|f(Y)] = \mathbb{E}[X|f(Y)].$$

14.1 Martingales

Definition 14.1
A $\mathbb{R}$-valued stochastic process X_i is a **martingale** if

- $\mathbb{E}(|X_i|) < \infty$
- $\mathbb{E}(X_{i+1}|X_i, \ldots, X_1) = X_i$

Definition 14.2
Y_i is a martingale with respect to X_i if

- $\mathbb{E}[|Y_i|] < \infty$
- Y_i is a function of $X_1, \ldots, X_i$.
- $\mathbb{E}[Y_{i+1}|X_1, \ldots, X_i] = Y_i$.

Proposition 14.3
If Y_i is a martingale wrt X_i, then Y_i is a martingale.

Proof. Since all Y_i are fns of $X_1, \ldots, X_i$, the tower laws imply that

$$\begin{aligned}\mathbb{E}[Y_{i+1}|Y_i, \ldots, Y_1] &= \mathbb{E}[\mathbb{E}[Y_{i+1}|X_1, \ldots, X_i]|Y_1, \ldots, Y_i] \\ &= \mathbb{E}[Y_{i+1}|X_1, \ldots, X_i] = Y_i.\end{aligned}$$

□

Example 14.4
Let X_i be i.i.d with $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[X_i^2] = 1$. If $S_n = X_1 + \ldots + X_n$, then $M_n = S_n^2 - n$ is a martingale wrt X_i.

$\mathbb{E}[|S_i|] < \infty$, and S_i is a fn of $X_1, \ldots, X_i$, so the first two conditions hold. For the third condition,

$$\begin{aligned}\mathbb{E}[M_{i+1}|X_1,\ldots,X_i] &= \mathbb{E}[(X_i+X_i)^2-(i+1)|X_1,\ldots,X_i] \\ &= \mathbb{E}[S_i^2+2S_iX_{i+1}+X_{i+1}^2-i-1|X_1,\ldots,X_i] \\ &= S_i^2-i = M_i.\end{aligned}$$

Lemma 14.5
If X_i, Y_i are independent martingales, then $X_i + Y_i$ is also a martingale.

Proof. If X_i, Y_i are finite, then so is $X_i + Y_i$, so the first condition holds. Also,

$$\mathbb{E}[X_{i+1}+Y_{i+1}|\{X_j\}_{j\le i},\{Y_j\}_{j\le i}] = X_i + Y_i$$

by the linearity of expectation, so the second condition also holds. □

Example 14.6 (Doob martingale)
Let $Y, X_1, X_2, \ldots$ be r.v.s Then, $M_i = \mathbb{E}[Y|X_1,\ldots,X_i]$ is a martingale.

From the tower law,

$$\mathbb{E}[M_{i+1}|X_1,\ldots,X_n] = \mathbb{E}[\mathbb{E}[Y|X_1,\ldots,X_i]|X_1,\ldots,X_i] = \mathbb{E}[Y|X_1,\ldots,X_i] = M_i.$$

Proposition 14.7
Let M_i be a martingale wrt X_i.

1. $\mathbb{E}[M_1] = \mathbb{E}[M_i]\forall i$
2. $\mathbb{E}[M_i|X_1,\ldots,X_j] = M_j$ if $j \le i$.
3. Increments are uncorrelated, i.e.,

$$\mathbb{E}[(M_j - M_i)(M_{j'} - M_{i'})] = 0$$

if $i < j < i' < j'$.

Proof. 1.
$$\mathbb{E}[M_i] = \mathbb{E}[\mathbb{E}[M_{i-1}|X_1,\ldots,X_{i-1}]] = \mathbb{E}[M_{i-1}] = \ldots.$$

2.

$$\begin{aligned}\mathbb{E}[M_i|X_1,\ldots,X_j] &= \mathbb{E}[\mathbb{E}[M_i|X_1,\ldots,X_{i-1}]|X_1,\ldots,X_j] \\ &= \mathbb{E}[M_{i-1}|X_1,\ldots,X_j] \\ &= \vdots \\ &= \mathbb{E}[M_{j+1}|X_1,\ldots,X_j] = M_j.\end{aligned}$$

3. Suffices to assume $j = i+1$ and $j' = i'+1$.

$$\begin{aligned}\mathbb{E}[(M_{i+1}-M_i)(M_{i'+1}-M_{i'})] &= \mathbb{E}[\mathbb{E}[(M_{i+1}-M_i)(M_{i'+1}-M_{i'})]|X_1,\ldots,X_{i+1}] \\ &= \mathbb{E}[(M_{i+1}-M_i)]\mathbb{E}[(M_{i'+1}-M_{i'})|X_1,\ldots,X_{i+1}] = 0.\end{aligned}$$

□

Theorem 14.8 (Martingale convergence theorem)
Let M_i be a martingale with $\mathbb{E}[|M_i|] \le c < \infty$. Then

$$M_\infty = \lim_{i\to\infty} M_i$$

exists a.s, and $\mathbb{E}[M_\infty] < \infty$.

Example 14.9 (Polya's urn)
We have an urn with two types of objects, Reeses and Gumdrops. At each time step, we uniformly pick one item from the urn and replace it with 2 of the same type of object. The urn starts with one of each type of object.

Let R_i, G_i be the number of Reeses and Gumdrops at time i. We want to find

$$\lim_{i\to\infty} \frac{R_i}{i+1},$$

which is the proportion of Reeses in the jar at time i. We will show that

$$\frac{R_i}{i+1} \xrightarrow[n\to\infty]{a.s.} \textsc{Unif}[0,1].$$

Claim 14.10

R_i is uniform on $\{1, \ldots, i\}$.

Proof. We use induction. $i = 1$ works.

$$\mathbb{P}[R_{i+1} = k] = \mathbb{P}[R_i = k]\frac{k}{i+1} + \mathbb{P}[R_i = k-1]\frac{k-1}{i+1} = \frac{1}{i}\frac{i+1-k}{i+1} + \frac{1}{i}\frac{k-1}{i+1} = \frac{1}{i+1}.$$

□

This implies that

$$\lim_{i\to\infty} \mathbb{P}\left(\frac{R_i}{i+1} \in (a,b)\right) = b - a$$

for $0 \leq a < b \leq 1$. Now, let's show a.s. convergence.

Claim 14.11

$$M_i = \frac{R_i}{i+1}$$

is a martingale.

Proof. Moments are finite, so it suffices to check:

$$\begin{aligned}\mathbb{E}\left[\frac{R_{i+1}}{i+2}|R_1, \ldots, R_i\right] &= \mathbb{E}\left[\frac{R_{i+1}}{i+2}|R_i\right] \\ &= \frac{R_i + \mathbb{P}[\text{chooses } R|R_i]}{i+2} \\ &= \frac{R_i + R_i/(i+1)}{i+2} = \frac{R_i}{i+1}.\end{aligned}$$

□

If $M_i \geq 0$, then $\mathbb{E}[|M_i|] = \mathbb{E}[M_i] = \mathbb{E}[M_i]$.

15 November 2, 2023

15.1 Optional stopping theorem

Let the notation "$\wedge$" mean min.

Lemma 15.1

Let M_i be a martingale wrt X_i, and T a stopping time wrt X_i. Then $M_{i\wedge T}$ is a martingale wrt X_i.

Proof.

$$\begin{aligned}\mathbb{E}[M_{(i+1)\wedge T}|X_1,\dots,X_i] &= \mathbb{E}[M_{i+1}\mathbb{1}(T \geq i+1)|X_1,\dots,X_i] + \mathbb{E}[M_T\mathbb{1}(T \leq i)|X_1,\dots,X_i] \\ &= M_i\mathbb{1}(T \geq i+1) + M_T\mathbb{1}(T \leq i) = M_{i\wedge T},\end{aligned}$$

which works since T, M_T are by definition functions of $X_1,\dots,X_i$. □

Theorem 15.2 (Optional Stopping Theorem)

Let M_i be a martingale with respect to X_i, and T a stopping time wrt X_i. $\mathbb{E}[M_T] = \mathbb{E}[M_1]$ if any of the following conditions are satisfied:

- $T \leq c$ almost surely for some $c < \infty$.
- $\mathbb{E}[T] < \infty$ and $|M_{i+1} - M_i| \leq c$ almost surely for some $i < T$ and $c < \infty$
- $|M_{i\wedge T}| \leq c$ almost surely for all i.

The third statement could have $T = \infty$; however, given that the martingale $M_{i\wedge T}$ is bounded by a finite constant, the martingale convergence theorem tells us that its limit exists almost surely, and we can take M_T to be this limit.

Proof. Proof of the first bullet point:

$$\mathbb{E}[M_T] = \mathbb{E}[M_1] + \sum_{i=2}^{c} \mathbb{E}[(M_i - M_{i-1})\mathbb{1}(T \geq i)].$$

We can rewrite

$$\begin{aligned}\mathbb{E}[(M_i - M_{i-1})\mathbb{1}(T \geq i)] &= \mathbb{E}[\mathbb{E}[(M_i - M_{i-1})\mathbb{1}_{T\geq i}|X_1,\dots,X_{i-1}]] \\ &= \mathbb{E}[\mathbb{E}[(M_i - M_{i-1})|X_1,\dots,X_{i-1}]\mathbb{1}_{T\geq i}] = 0.\end{aligned}$$

□

Example 15.3

Let $S_n = X_1 + \ldots + X_n$ with $X_i = \pm 1$ be a martingale. If we start the random walk at 0, what is the probability that we hit a before b given $a < 0 < b$? How long does it take?

This is a classic. We can use the optional stopping theorem with the third criterion, since $|S_{i \wedge T}| \le \max(|a|, b)$. This gives $0 = \mathbb{E}[S_1] = \mathbb{E}[S_T] = pa + (1-p)b$, so the probability of hitting a first is $p = b/(b-a)$.

To compute the expected amount of time this takes, consider the martingale $S_n^2 - n$. Since reaching a, b amounts to hitting a state in a finite, irreducible Markov chain, $\mathbb{E}[T] < \infty$, so the second criterion holds and we can apply the optional stopping theorem. We thus have $\mathbb{E}[S_T^2 - T] = \mathbb{E}[S_1^2 - 1] = 0$, and we know $\mathbb{E}[S_T^2] = b/(b-a) \cdot a^2 + (-a)/(b-a) \cdot b^2 = -ab$, so $\mathbb{E}[T] = -ab$.

Example 15.4

Same walk as before, but biased. Let p, q be the probabilities of moving left,right, respectively.

Let $M_i = (p/q)^{S_i}$. We can show that this is a Martingale. Also, $|M_{i \wedge T}|$ is bounded, so we can apply the optional stopping theorem to get

$$P = \frac{(p/q)^b - 1}{(p/q)^b - (p/q)^a}.$$

We can also show that $S_i - i(p-q)$ is a martingale, which gives

$$\mathbb{E}[T] = \frac{Pa + (1-P)b}{p-q}.$$

Example 15.5

Consider a random walk (X_i, Y_i) on $\mathbb{Z}^2$. How long does it take to travel distance R away from the origin?

Here, we can show that $M_i = X_i^2 + Y_i^2 - i$ is a martingale. Note that this is essentially the same martingale that we used in one dimension. By optional stopping, $\mathbb{E}[X_i^2 + Y_i^2 - T] = 0$, so $\mathbb{E}[T] = \mathbb{E}[X_i^2 + Y_i^2]$, which is somewhere between R^2 and $(R+1)^2$. We can use this same argument to show that this will be the same for walks on $\mathbb{Z}^d$.

16 November 7, 2023

16.1 Harmonic functions

Definition 16.1

Let P be a Markov chain on $\mathcal{X}$. A **harmonic function** is a function $f : \mathcal{X} \to \mathbb{R}$ s.t. $\mathbb{E}[f(X_1)|X_0 = x] = f(x)$.

Proposition 16.2

Let X_i be a Markov Chain and f a harmonic function. $f(X_i)$ is a martingale wrt X_i.

Proof.

$$\mathbb{E}[f(X_{i+1})|X_1, \dots, X_i] = \mathbb{E}[f(X_{i+1})|X_i] = f(X_i).$$

□

Proposition 16.3

If P is irreducible and recurrent, the only bounded harmonic functions are constant.

Proof. If f is a bounded harmonic function, then $f(X_i)$ is a bounded martingale. So, by the martingale convergence theorem, $f(X_i)$ converges to a value almost surely. On the other hand, since P is irreducible, X_i visits every state i.o., so $f(X_i)$ must be the same for every state, otherwise it would take on at least two distinct values i.o. □

Example 16.4

Consider the random biased walk on $\mathbb{Z}$ from last lecture.

Finding the martingale $(p/q)^{S_i}$ can be motivated by harmonic functions. In particular, harmonic functions for this martingale satisfy

$$f(x) = pf(x-1) + qf(x+1),$$

which gives $f(x) = a + b(p/q)^x$ for $a, b \in \mathbb{R}$.

Example 16.5

Consider Markov chain on $\mathbb{Z}$ defined by moving $-2, -1, 0, 1, 2$ with probabilities p_i for $i \in [5]$. How many harmonic functions are there?

We have

$$f(x) = p_0 f(x-2) + p_1 f(x-1) + p_2 f(x) + p_3 f(x+1) + p_4 f(x+2),$$

which has characteristic equation

$$x^2 = p_0 + p_1 x + p_2 x^2 + p_3 x^3 + p_4 x^4.$$

This gives four roots, one of which is always 1, so all harmonic functions are of the form

$$f(x) = a_1 + a_2 \lambda_2^x + a_3 \lambda_3^x + a_4 \lambda_4^x.$$

16.2 Harmonic extensions

Proposition 16.6

Let P be a Markov chain and $S \subseteq \mathcal{X}$. Assume $P(x,x) = 1$ for all $x \in S$ and $X_i \in S$ for some i a.s. Any bounded function $f : S \to \mathbb{R}$ has a unique extension to a harmonic function

$$\tilde{f}(x) = \mathbb{E}[f(X_t) | X_0 = x],$$

where T is the first time that X_i enters S.

Recall that "extending" a function means to replace it with another function whose values are the same on the domain on the original function.

Proof. The function is harmonic, since

$$\begin{aligned}\mathbb{E}[\tilde{f}(X_1) | X_0 = x] &= \mathbb{E}[\mathbb{E}[f(X_T) | X_0 = X_1] | X_0 = x] \\ &= \mathbb{E}[f(X_T) | X_1 = x] = \mathbb{E}[f(X_T) | X_0 = x] = \tilde{f}(x).\end{aligned}$$

To show uniqueness, let g be any harmonic function that extends f. Since we must end up in S a.s. for any starting $x \in \mathcal{X}$, and f is bounded, we must g bounded. Therefore, the martingale $g(X_{T\wedge i})$ is bounded, so we can apply the optional stopping theorem to get $\mathbb{E}[g(X_T)|X_0 = x] = \mathbb{E}[f(X_T)|X_0 = x] = g(x)$, so $g = \tilde{f}$.

Corollary 16.7

The harmonic extension $\tilde{f}$ satisfies

$$\sup_{x\in\mathcal{X}} \tilde{f}(x) = \sup_{x\in S} f(x).$$

Proof. $\tilde{f}$ is an expected value of $f(y)$ for all $y \in S$, so is bounded above by its values. □

Corollary 16.8

Let P be an irreducible Markov chain on a finite state space, and let $A, B \subseteq \mathcal{X}$ be two disjoint subsets. Let T be the first time that X_i enters A or B. Then, $\mathbb{P}[X_T \in A|X_0]$ is harmonic for $\tilde{P}$, which is equal to P except that $\tilde{P}(x,x) = 1$ for $x \in A, B$.

Proof. This is the unique extension for function $f(x) = \mathbb{1}(x \in A)$ over the subset $S = A \cup B$. □

17 November 9, 2023

Something about branching processes.

18 November 14, 2023

18.1 Continuous Time MCs

Now we shift focus from MCs in a discrete state space and discrete time space, to MCs in a discrete state space and a continuous time space. We can intuitively think about MCs in a continuous time space as waiting some time in the current state before jumping to another state, where the waiting time is a real random variable.

We still want the normal Markov property to hold, i.e.,

$$\mathbb{P}[T > t | T > s] = \mathbb{P}[T > t - s],$$

where T is the total amount of time that we have to wait.

Lemma 18.1

If $T \sim \exp(\lambda)$, then $\mathbb{P}[T > t | T > s] = \mathbb{P}[T > t - s]$.

Proof. Recall that $T \sim \exp(\lambda)$ has pdf $\mathbb{P}[T \leq t] = 1 - e^{-\lambda t}$. □

18.2 Poisson processes

Definition 18.2

A continuous time stochastic process is a family of jointly defined r.v.s $\{X_t\}_{t \in \mathbb{R}}$ where all $X_t \in \mathcal{X}$.

Definition 18.3

Let $T_i \sim \textsc{Exp}(\lambda)$ be independent. The **poisson process** of rate λ is the continuous time stochastic process N_t taking values in $\mathbb{N}$ defined by

$$N_t = \max\{i : T_1 + \ldots + T_i \leq t\}.$$

In words, given that we have to wait T_i time per jump, N_t is the stochastic process representing the amount of jumps we take before time t.

Theorem 18.4

Let N_t be a Poisson process of rate λ. Then:

- $M_t = (N_{t+s} - N_s)_{t \geq 0}$ is also a Poisson process of rate λ.
- M_t is independent of N_t for all $t \leq s$.

Proof. For the second bullet point, say we fix $N_s = n$; then, N_t is a function of $T_1, \ldots, T_n$ for all $t \leq s$. On the other hand, M_t relies on timesteps $n+1, \ldots$, so they are independent.

For the first bullet point, we will keep N_s fixed and then show that the amount of time taken to get to N_{t+s} is also a Poisson process with rate λ. Also fix $T_1 = t_1, \ldots, T_n = t_n$. The nonnegative time between s and the end of the first n jumps is $s - (t_1 + \ldots + t_n)$. The positive time between the end of the first $n+1$ jumps and s is $T_1' = T_{n+1} - (s - (t_1 + \ldots + t_n))$. By the memoryless property of T_{n+1}, T_1' is exponential with parameter λ. Now, if we define $T_i' = T_{n+i}$, we have that

$$(N_{t+s} - N_s) = \max\{i : T_1' + \ldots + T_i' \leq t\},$$

since we only care about the number of steps after N_s it takes before reaching time $t + s$, which is exactly t away from the point at which T_1' starts counting from. Since all $\{T_i'\}_{i \geq 1}$ are exponential with parameter λ, this shows that $(N_{t+s} - N_s)_{t \geq 0}$ is too. □

Lemma 18.5

If $X_1, \ldots, X_k \sim \textsc{Exp}(\lambda)$, then $X_1 + \ldots + X_k \sim \Gamma(k, \lambda)$ and has density

$$\frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x}.$$

Proof. There are many ways to prove this, but we'll use induction. Let $Y \sim \Gamma(k, \lambda)$ and $X \sim \textsc{Exp}(\lambda)$. When $k = 1$, $Y \sim \textsc{Exp}(\lambda)$, so the base case holds. Now, assume the lemma holds for all $1, \ldots, k$, and let $Z = X + Y$. Then,

$$\begin{aligned} f_Z(z) &= \int_{y=0}^{z} f_Y(y) f_X(z-y) \mathrm{d}y \\ &= \int_0^z \frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda y} \lambda e^{-\lambda(z-y)} \mathrm{d}y \\ &= \int_0^z \frac{\lambda^{k+1}}{(k-1)!} x^{k-1} e^{-\lambda z} \mathrm{d}y \\ &= \frac{\lambda^{k+1}}{k!} z^k e^{-\lambda z}, \end{aligned}$$

so $Z \sim \Gamma(k+1, \lambda)$, and we are finished. □

Next, we show why we call these processes *poisson* processes.

Proposition 18.6
For all t, $N_t \sim \text{Pois}(\lambda t)$.

Proof. Let $X = T_1 + \ldots + T_n$. By the previous lemma, $X \sim \Gamma(n, \lambda)$, so

$$\begin{aligned}\mathbb{P}[N_t = n] &= \mathbb{P}[X \le t, X + T_{n+1} > t] \\ &= \int_0^t \mathbb{P}[T_{n+1} > t - x] p_X(x) \mathrm{d}x \\ &= \int_0^t e^{-\lambda(t-x)} \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x} \mathrm{d}x \\ &= \frac{(\lambda t)^n}{n!} e^{-\lambda t}.\end{aligned}$$

□

18.3 More definitions for continuous time MCs

Let X_i be a discrete time MC with transition matrix K, and N_t be a Poisson process with rate λ. Then, X_{N_t} can be viewed as a continuous time MC with transitions in exponentially distributed intervals, since N_t (as t increases) jumps only in exponentially distributed intervals.

For continuous time MCs, we define $P^t(x, y) = \mathbb{P}[X_t = y | X_0 = x]$. For this specific MC, we have

$$P^t(x, y) = \sum_n \frac{(\lambda t)^n}{n!} e^{-t\lambda} K^n(x, y),$$

by conditioning on N_t.

Definition 18.7
Let K be an $N \times N$ matrix. The **matrix exponential**

$$\exp(K) = \sum_{i \ge 0} \frac{K^i}{i!}.$$

Proposition 18.8

The matrix exponential satisfies:

- The matrix function $F(t) = \exp(tX)$ is the unique solution to the system of differential equations defined by $\mathrm{d}/\mathrm{d}t F(t) = XF(t)$.
- If $X = ADA^{-1}$, $\exp(X) = A\exp(D)A^{-1}$.
- If X and Y commute, $\exp(X)\exp(Y) = \exp(X+Y)$.

Proof. homework.

add this here. □

From the proposition, $\exp(-\lambda t I) = e^{-t\lambda}$, so

$$P^t = \exp(-\lambda t I)\exp(\lambda t K) = \exp(-\lambda t(I-K)).$$

This leads to the natural extension of a random walk:

Example 18.9

Let G be a graph. The continuous time random walk on G of rate λ is the continuous time MC X_{N_t}, where X_i is a random walk on G and N_t is a poisson process of rate λ.

19 November 16, 2023

19.1 Continuous time Markov Chains (CTMCs)

Definition 19.1

Let $\mathcal{X}$ be a finite state space. A **generator** is a matrix Q whose rows and columns are indexed by $\mathcal{X}$ s.t. $Q(x,y) \geq 0$ if $x \neq y$ and whose rows sum to 0.

Lemma 19.2

If Q is a generator and D is a matrix consisting of the diagonal entries of Q, then the matrix $K = -D^{-1}(Q - D)$ is a transition matrix. If D has any 0s on the diagonal then we interpret the corresponding row in K to be 1 on the diagonal and 0 everywhere else.

Proof. Since diagonal elements are ≤ 0, all entries in K are nonnegative. Further, the xth row of $Q - D$ sums to $-D(x,x)$, so multiplying by D^{-1} forces each row sum to 1. □

Definition 19.3

A **continuous time Markov chain** (CTMC) with state space $\mathcal{X}$ and generator Q is a stochastic process X_t defined inductively as follows. Given $X_s = x$, we wait for $\exp(-Q(x,x))$ distributed amount of time, and then take a step according to $K = -D^{-1}(Q - D)$. If $Q(x,x) = 0$, we stay at x forever. We call K the **emedded discrete time Markov chain**, and the entries of Q the **transition rates**.

Another way to view this definition: given $X_s = x$, we have $T_y \sim \exp(Q(x,y))$ random variable for each y and then move the y with the minimum T_y after T_y time. Remember that for generator Q the rows sum to 0, so $\sum_{z \neq x} Q(x,z) = Q(x,x)$. It can be shown that choosing the waiting for $\exp(-Q(x,x))$ time, and then moving according to K, gives the same distribution as choosing the waiting time and step along the MC at the same time (by taking the minimum waiting time).

Proposition 19.4

A continuous time MC X_t satisfies the Markov property in the following way: conditioned on X_t, future X_s for $s > t$ is distributed as if the MC started from X_t, and is thus independent from past $X_{s'}$ for $s' < t$.

there is some connection between the original CTMC that was introduced last time. what is the generator for that or smth.

Example 19.5
We define another continuous time random walk on graph G with generator $Q(x,y)=1$ if $(x,y)\in E$ and $Q(x,x)=-\deg(x)$.

19.2 Kolmogorov Equations

Theorem 19.6 (Kolmogorov backwards and forwards equations)
Let P^t be a CTMC with generator Q. Then P^t is the unique solution to the equations

$$\frac{d}{dt}P^t = QP^t,$$

and

$$\frac{d}{dt}P^t = P^tQ,$$

with initial condition $P^0 = I$.

Corollary 19.7
Let X_i be a CTMC with generator Q. Then $P^t = \exp(tQ)$.

Proof. $\exp(tQ)$ satisfies the forwards and backwards equations (with the initial condition). By the theorem, this is unique, so it is P^t. □

Now we prove the main theorem.

Proof. First,

$$\frac{\mathrm{d}}{\mathrm{d}t}P^t = \lim_{h\to 0}\frac{P^{t+h}-P^t}{h} = \lim_{h\to 0}P^t\frac{P^h-I}{h} = \lim_{h\to 0}\frac{P^h-I}{h}P^t,$$

where we are allowed to commute by memorylessness. Therefore, it suffices to show that

$$\lim_{h\to 0}\frac{P^h-I}{h} = Q.$$

First we consider diagonal entries. We can stay at state x by taking no jumps, or at least two jumps and landing back at x. Recall each jump is exponential $T\sim$

Exp$(-Q(x,x))$, so the probability that we take at least two timesteps (and land in the same place) is bounded above by $\mathbb{P}[T \le h]^2 = h^2 e^{2Q(x,x)h} \in O(h^2)$. Thus,

$$\frac{P^h(x,x)-1}{h} = \frac{\mathbb{P}[(T>h)]-1+O(h^2)}{h} = \frac{e^{hQ(x,x)}-1}{h} + O(h),$$

which has limit $\lim_{h\to 0} e^{hQ(x,x)}/h = Q(x,x)$.

Next we do non-diagonal entries. We can jump from x to y, $x \neq y$, only by taking at least one jump with correct transition probability. As before, taking at least two jumps is $O(h^2)$, so

$$\frac{P^h(x,y)}{h} = \frac{\mathbb{P}[T<h]K(x,y)+O(h^2)}{h} = \frac{(1-e^{hQ(x,x)}K(x,y))}{h} + O(h).$$

By definition, $K(x,y) = -Q(x,x)^{-1}Q(x,y)$, so the limit is $\lim_{h\to 0}(1-e^{hQ(x,x)})K(x,y)/h = -Q(x,x)K(x,y) = Q(x,y)$. □

20 November 21, 2023

20.1 Stationary distributions on CTMC

Proposition 20.1

For a CTMC, $P^t(x,x) > 0$ for all t and $P^t(x,y) > 0$ for all $t > 0$ if and only if $K^i(x,y) > 0$ for some i, where K is the embedded discrete time chain.

Proof. For any t, $\mathbb{P}[T > t] = e^{tQ(x,x)}$ is the probability that we don't move, so $P^t(x,x) \ge e^{tQ(x,x)} > 0$.

If $K^i(x,y) > 0$, there is some chance to take i steps from x to y within time t, so $P^t(x,y) > 0$. Conversely, if $P^t(x,y) > 0$, then there must exist some path in the MC leading from x to y, so $K^i(x,y) > 0$ for some i. □

Definition 20.2

π is a stationary distribution if $\pi P^t = \pi$ for all t.

Proposition 20.3

The following are equivalent:

- $\pi P^t = \pi$ for all $t > 0$.
- $\pi Q = 0$.
- $\mu K = \mu$, where $\mu(x) = \pi(x)Q(x,x)/\sum_y \pi(y)Q(y,y)$.

Proof. If $\pi P^t = \pi$ for all t, then Kolmogorov gives

$$\frac{d}{dt}\pi P^t = \pi P^t Q = 0 \implies \pi Q = 0.$$

Conversely, if $\pi Q = 0$,

$$\pi P^t = \pi \sum \frac{t^n Q^n}{n!} = \pi,$$

since everything in the sum dies.

For the third bullet point, $\mu K = \mu \iff \mu(I-K) = 0$, which is equivalent to $\mu D^{-1}(D+Q-D) = 0$. But $\mu D^{-1} = \pi$, so $\pi Q = 0$. **huh**. □

Note that this implies that the stationary distribution for a CTMC and its embedded discrete time chain are not the same.

Corollary 20.4

All CTMC on finite state spaces have a stationary distribution. It's unique if the Markov Chain is irreducible.

Proof. The embedded discrete time chain has a stationary distribution, and therefore so does the CTMC. If the CTMC is irreducible, so is K. Then, the mapping from μ to π is unique, since irreducible implies $Q(x,x) > 0$ for all x. □

Definition 20.5

Q is reversible with respect to π if $\pi(x)Q(x,y) = \pi(y)Q(y,x)$.

Lemma 20.6

The following are equivalent:

- Q is reversible with respect to π
- P^t is reversible with respect to π for all t.

Proof. If Q is reversible wrt π, then

$$\pi(x)P^t(x,y) = \sum_i t^i \frac{\pi(x)Q^i(x,y)}{i!} = \sum_i t^i \frac{\pi(y)Q^i(y,x)}{i!} = \pi(y)P^t(y,x).$$

Conversely, if $\pi(x)P^t(x,y) = \pi(y)P^t$, then Kolmogorov gives

$$\frac{d}{dt}\pi(x)P^t(x,y) = \pi(x)P^t(x,y)Q(x,y) = \pi(y)P^t(y,x)Q(y,x) = \frac{d}{dt}\pi(y)P^t(y,x).$$

Setting $t = 0$ gives the desired result. □

finish this up.