18.615: Introduction to Stochastic Processes

Prof. Elchanan Mossel

notes by nambrath

Contents

1 February 3, 2020

Welcome to 18.615, Introduction to Stochastic Processes, taught by Professor Elchanan Mossel.

We will talk about two things in this class, **Markov chains** and **martingales**. Markov chains are used to model the evolution of physical systems, for scientific computing applications, and more:

Predicting the weather.

In each cell of a map, assign parameters (temperature, pressure, etc.). Each cell is randomly assigned new parameters using its parameters and those of its neighbors from the previous time step.

Modeling disease spread.

Represent a population as a graph.

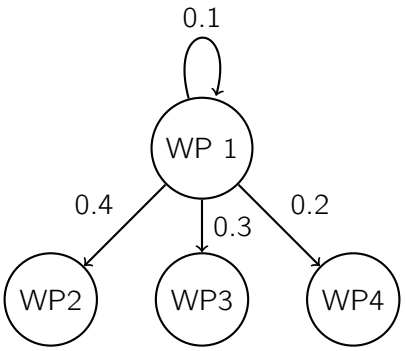
Contact process. Start with 1 infected vertex. At t, each vertex becomes infected with probability $1/10$, if at least one neighbor is infected. Infected vertices don't get better.

S.I.S. Start with 1 infected vertex. A healthy vertex with no infected neighbors stays healthy. A healthy vertex with infected neighbors gets sick with probability $1/10$. An infected vertex becomes healthy with probability $1/20$.

S.I.R. Allows you to account for immunity and recovery.

Page rank is a Markov chain.

Google measures the importance of a web page using a Markov chain. It creates a graph to model the Web:



Each webpage links to others with varying degrees of prominence. Importance is proportional to the amount of time that a process spends at any given page. Not an exact model, but useful.

Martingales represent "fair" processes:

Election betting markets. The betting markets have Trump's odds at -135. So if you bet \$135 now, you'll win \$100 in November. What's P(Trump)?

if he wins, you get \$235 $\rightarrow$ $p \cdot 235 = 135$ $\rightarrow$ $p_0 = \dfrac{135}{235}$

We calculate P(Bernie) at odds of +350: $p = \dfrac{100}{450}$

1.1 What are Markov chains?

Definition. A sequence of random variables $X_0, X_1, \ldots$ is a (finite-space) **Markov chain** if, for all t and all values $x_0, x_1, \ldots, x_{t-1}, x_t, y$ it holds that:

$$\mathbb{P}[\, X_{t+1} = y \mid X_t = x, X_{t-1} = x_{t-1}, \ldots X_0 = x_0 \,]$$

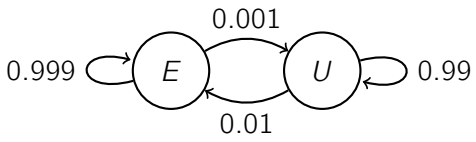
is the same as:

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

If $\mathbb{P}_t(x, y) = \mathbb{P}(x, y)$ for all t, and all x and y, then $X_0, X_1, \ldots$ is a **homogeneous** Markov chain.

Drawing a card from a deck without replacement is *not* a Markov chain.

This simple model of employment is a Markov chain (E for employed, U for unemployed):



1.2 Markov chains and matrices

We can write $\mathbb{P}_t(x, y)$ as a matrix for the job example:

$$\begin{pmatrix} 0.999 & 0.001 \\ 0.010 & 0.990 \end{pmatrix} = \mathbb{P}_t(x, y) = P$$

where the rows are the x states and the columns are the y states.

We can write:

$$\mathbb{P}[X_{10} = E] = \mathbb{P}[X_9 = E] \cdot \mathbb{P}[X_{10} = E \mid X_9 = E] + \mathbb{P}[X_9 = U] \cdot \mathbb{P}[X_{10} = E \mid X_9 = U]$$

This allows us to write:

$$\Big(\mathbb{P}[X_{10} = E] \quad \mathbb{P}[X_{10} = U]\Big) = \Big(\mathbb{P}[X_9 = E] \quad \mathbb{P}[X_9 = U]\Big) \cdot P$$

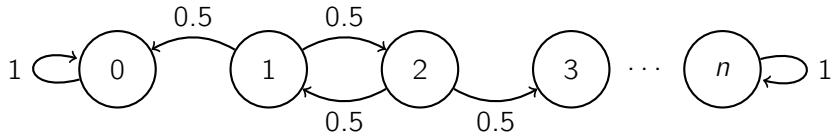
where the row vector on the RHS is multiplying the matrix defined above. We can define μ_t as the row vector given by $(\mathbb{P}[X_t = x])_x$. Then:

$$\begin{aligned} \mu_{t+1} &= \mu_t P \quad \text{for every Markov chain.} \\ &= \mu_0 P_0 P_1 \dots P_{t-1} \end{aligned}$$

If the Markov chain is homogeneous, $\mu_t = \mu_0 P^t$.

1.3 Gambler's ruin

The graph for gambler's ruin looks something like this:



The gambler bets on what the outcome of a fair coin toss will be, and he starts with k money. If he wins he moves up a peg, and if he loses he moves down. The game terminates at some time τ, when the gambler has either lost everything (when he's at 0) or when he's won an amount n.

Here's the matrix:

$$A = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 1/2 & 0 & 1/2 & & \\ 0 & 1/2 & 0 & 1/2 & \\ \vdots & & & \ddots & \\ 0 & & & \dots & 1 \end{bmatrix}$$

Claim: $\mathbb{P}[X_\tau = n \mid X_0 = k] = k/n$.
This is a fair game, so the expected wealth at the end is the expected wealth at the start $= k$. Then: $n \cdot [\mathbb{P}(x_\tau = n \mid x_0 = k)] = k \; to \; P = k/n$.

2 February 5, 2020

Let's pick up where we left off: the probability of getting to n.

2.1 Gambler's ruin

Let τ be the (random) time where $X_\tau = 0$ or n. To show that τ is finite, we use the fact that the chance of the game ending after n steps is 2^{-n} – so almost surely, τ is finite.

Claim: $\mathbb{P}[X_\tau = n \mid X_0 = k] = k/n$.

Proof: Let $a_k = \mathbb{P}[X_\tau = n \mid X_0 = k]$. Then:

$$\begin{aligned}\mathbb{P}[X_\tau = n \mid X_0 = k] = \mathbb{P}[X_1 = k+1, X_\tau = n \mid X_0 = k] + \\ \mathbb{P}[X_1 = k-1, X_\tau = n \mid X_0 = k]\end{aligned}$$

$$\begin{aligned}\mathbb{P}[X_1 = k+1, X_\tau = n \mid X_0 = k] &= \mathbb{P}[X_1 = k+1 \mid X_0 = k] \times \\ &\quad \mathbb{P}[X_\tau = n \mid X_1 = k+1, X_0 = k] \\ \text{(via Markov prop.)} \quad &= \frac{1}{2} \times \mathbb{P}[X_\tau = n \mid X_1 = k+1] \\ \text{(since indexing isn't important)} \quad &= \frac{1}{2} \times \mathbb{P}[X_\tau = n \mid X_0 = k+1] \\ &= \frac{1}{2} a_{k+1}\end{aligned}$$

So we have three equations to solve: $a_0 = 0$, $a_n = 1$, and

$$\mathbb{P}[X_\tau = n \mid X_0 = k] = \frac{1}{2}(a_{k+1} + a_{k-1}) \quad \text{for } 0 < k < n$$

We have to check that $a_k = k/n$ is a unique solution to these equations. Let's use induction:

$$\begin{aligned}a_1 &= \frac{1}{2}(a_0 + a_2) = \frac{1}{2} a_2 \\ a_2 &= \frac{1}{2}(a_1 + a_3) = 2a_1 \to a_3 = 3a_1 \\ &\dots \\ a_k &= k a_1. \qquad \text{If } a_n = n a_1 = 1,\ a_1 = 1/n \implies a_k = k/n\end{aligned}$$

Claim: $E[\tau \mid X_0 = k] = k(n-k)$.

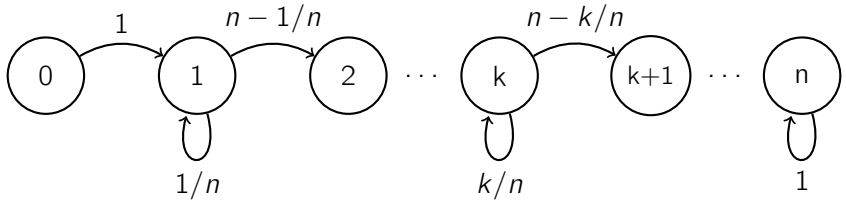
Proof: Let $b_k = E[\tau \mid X_0 = k] = \frac{1}{2}(b_{k+1} + b_{k-1}) + 1$. [1] Solving with $b_0 = 0$ and $b_n = 0$, we get the desired result.

[1] If we have k now, and we take a step (hence the $+1$), we have an equal chance of ending up at $k+1$ and $k-1$. The time left is given by b_{k+1} and b_{k-1} respectively.

2.2 Coupon collector problem

Let's say there are n different types of coupons. You acquire one coupon a day. You are equally likely to get any one of the n types.

Let X_t be the number of types of coupons you have on day t.



$$\mathbb{P}[X_{t+1} = k+1 \mid X_t = k] = \frac{k}{n}\left(\frac{n-k}{n}\right)$$

Let τ be the day you get all n coupons: $E[\tau] = n\sum_{k=1}^{n}\frac{1}{k}$.

> **How?** Let Y_i be the number of days between having $i-1$ and i coupons. Clearly $Y_i = 1$.
>
> Y_2 is a geometric R.V. where $P(\text{success})$ is $\frac{n-1}{n}$, so $E[Y_2] = \frac{n}{n-1}$.
>
> For Y_k, the probability is $\frac{n-k}{n}$, then $E[Y_k] = \frac{n}{n-k}$.
>
> So $\tau = Y_1 + Y_2 + \cdots + Y_n \to E[\tau] = \sum E[Y_k] = n\sum\frac{1}{k} \approx n\log n$.

These two examples – gambler's ruin and the coupon collector – are relatively easy because there are only two possible places to go from each state. *Branching processes* are tougher – we'll model these later!

2.3 Long-term Markov chain behavior

Let $X_0, X_1, \ldots$ be a homogeneous Markov chain.
Let $\pi_0, \pi_1, \ldots$ be distributions of $X_0, X_1, \ldots$ so that $\pi_i = \pi_0 P^i$, where P^i is the transition matrix.

Notation: $\mathbb{P}_x[X_t = y] := \mathbb{P}[X_t = y \mid X_0 = x]$ and similarly $E_x[X_t] := E[X_t \mid X_0 = x]$.

Questions about the limiting behavior of π:

1. Does $\lim_{t\to\infty}\pi_t$ converge?
2. Does $\lim_{t\to\infty}\frac{1}{t+1}\sum_{s=0}^{t}\pi_s$ converge?
   Until t, how much time is spent at previous states? Does it converge?

3. Is there a stationary π? If $\pi_0 = \pi$, then $\pi_t = \pi \ \forall t$.
4. Do the answers here depend on π_0, the starting point?

Limits of the random variable:

1. Is there a random variable such that $\mathbb{P}[\lim_{t\to\infty} X_t = x] = 1$?
   E.g.: with the coupon collector, $\mathbb{P}[\lim_{t\to\infty} X_t = n] = 1$.
2. How many times between 0 and t was the Markov chain at x?
$$\mathbb{P}\Big[\lim_{t\to\infty} \underbrace{\frac{1}{t+1}(\#\, s : X_s = x, 0 \le s \le t)}_{\text{fraction of time spent at } x} = v_x\Big] = 1 \quad \forall x$$
   where v_x is some asymptotic fraction of time.

3 February 10, 2020

Let's revisit our notions of convergence for Markov chains. Let $\pi_s = \pi P^s$.

1. Does $\lim_{t\to\infty} \pi_t$ exist?
2. Does $\lim_{t\to\infty} \frac{1}{t+1} \sum_{s=0}^{t} \pi_s$ exist?
3. Does $\lim_{t\to\infty} X_t$ exist?
4. Let $N_t(a) = \frac{1}{t+1}|\{0 < s \le t \mid X_s = a\}|$. Does $\lim_{t\to\infty} N_t$ exist?

3.1 Some examples.

Let's answer these questions for some Markov chains.

3.1.1 Coin toss

$\mathbb{P}[x_t = 0, 1] = 1/2$. X_t are independent.

1. π_t is the distribution of probabilities over outcomes. $\pi_t = (1/2, 1/2)$ for $t \ge 1$. So $\lim_{t\to\infty} \pi_t = (1/2, 1/2)$.
2. $\lim_{t\to\infty} \frac{1}{t+1} \sum_{s=0}^{t} \pi_s = (1/2, 1/2)$.
3. $\lim_{t\to\infty} X_t$ does not exist.
4. $\lim_{t\to\infty} N_t(0) = 1/2$. More formally $\mathbb{P}[N_t(0) \to 1/2] = 1$ and $\mathbb{P}[N_t(1) \to 1/2] = 1$, i.e. almost surely! This is the **strong law of large numbers**.

3.1.2 Gambler's ruin

$\mathbb{P}[X_{t+1} = X_t \pm 1] = 0.5$ if $1 \leq X_t \leq n-1$, and $X_{t+1} = X_t$ if $X_t = 0, n$.

1. $\lim_{t\to\infty} \pi_t = \left(1 - \frac{k}{n}, 0, 0, \ldots, 0, \frac{k}{n}\right)$ if $X_0 = k$. Note: it depends on k!
2. Calculus fact! If #1 exists, #2 converges to the same limit.
3. $\lim_{t\to\infty} X_t = X \in \{0, n\}$ where $\mathbb{P}[X = n] = \frac{k}{n}$ and $\mathbb{P}[X = 0] = 1 - \frac{k}{n}$.
4. $\lim_{t\to\infty} N_t(a) = 0$ for $0 < a < n$ and does not exist for $a = 0, n$ since

$$\mathbb{P}[\lim N_t(0) = 1] = 1 - \frac{k}{n} \neq 1 \quad \text{and} \quad \mathbb{P}[\lim N_t(n) = 1] = \frac{k}{n} \neq 1$$

   There is no deterministic fraction of time spent at 0 or n.

The main case where #3 and #4 happen simultaneously is in a constant or absorbing Markov chain.

3.1.3 +1 mod k

Fix an integer k and let $X_{t+1} = (X_t + 1) \mod k$.

1. $\lim_{t\to\infty} \pi_t$ doesn't have to exist. The vector cycles and doesn't converge:

$$\begin{aligned} \pi_0 &= (1, 0, 0, \ldots, 0) \\ \pi_1 &= (0, 1, 0, \ldots, 0) \\ &\vdots \\ \pi_k &= (1, 0, 0, \ldots, 0) \end{aligned}$$

2. $\lim_{t\to\infty} \frac{1}{t+1} \sum_{s=0}^{t} \pi_s = (\frac{1}{k}, \ldots, \frac{1}{k})$ since the chain returns to the same state every k steps.

★ Is there a stationary π where $\pi P = \pi$? Yes! $\pi = (\frac{1}{k}, \ldots, \frac{1}{k})$.

3. $\lim_{t\to\infty} X_t$ does not exist.
4. $\mathbb{P}[\lim N_t(a) = \frac{1}{k}] = 1 \ \forall\, a$.

Note: The $(\frac{1}{k}, \ldots, \frac{1}{k})$ vector is the answer to #2, #4, and the bonus!

3.1.4 Binary symmetric channel

This is Claude Shannon's model of noisy communication, described by:

$$P = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix} = (1-2\epsilon)\,\mathbb{1} + \epsilon \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

1. $\lim \pi_t = \begin{pmatrix} 1/2 & 1/2 \end{pmatrix}$ as seen here:

$$\begin{aligned} \pi_{t+1} = \pi_t P &= (1-2\epsilon)\pi_t + 2\epsilon \begin{pmatrix} 1/2 & 1/2 \end{pmatrix} \\ &= (1-2\epsilon)^t \pi_0 + [1-(1-2\epsilon)^t] \begin{pmatrix} 1/2 & 1/2 \end{pmatrix} \\ &\Longrightarrow \begin{pmatrix} 1/2 & 1/2 \end{pmatrix} \quad \text{as t } \to \infty \end{aligned}$$

2. This is the same as #1.

★ We can solve for the stationary π:

$$\begin{pmatrix} \pi_0 & \pi_1 \end{pmatrix} \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix} = \begin{pmatrix} \pi_0 & \pi_1 \end{pmatrix}$$

This reduces to three equations:

$$(1-\epsilon)\pi_0 + \epsilon\pi_1 = \pi_0 \qquad \epsilon\pi_0 + (1-\epsilon)\pi_1 = \pi_1 \qquad \pi_0 + \pi_1 = 1$$

The solution is $\pi = \begin{pmatrix} 1/2 & 1/2 \end{pmatrix}$.

3. $\lim_{t\to\infty} X_t$ does not exist.
4. There is a convergent fraction, but we don't yet know how to show it. It's $N_t(0) = N_t(1) = 1/2$.

3.2 Stationary distributions exist

Theorem.

Let P be a *finite* Markov chain. Then there exists a stationary distribution π such that $\pi P = \pi$.

All the earlier examples were finite Markov chains.

Proof: Let Δ be the set of all distributions on Ω, the state space:

$$\{\, (x_i : i \in \Omega) : \forall\, i, x_i \geq 0, \sum_{i\in\Omega} x_i = 1 \,\}$$

Claim: Δ is *compact*, i.e closed and bounded.
Let $X_n \in A$ where A is a *closed* set. Then for all n and $X_n \to X$: $X \in A$.

Define $F(\pi) = \max_i |\pi(i) - (\pi P)(i)|$. It measures "how stationary" π is. Note: if $F(\pi) = 0$ then π is stationary.

Fact: A continuous F on a compact Δ has a minimum. Let π^* be the minimum. We then need to show $F(\pi^*) = 0$.

Let δ be any distribution (i.e. between 0 and 1) and define $\delta_n = \frac{1}{n} \sum_{i=0}^{n-1} \delta P^i$.
Then $\delta_n - \delta_n P = \frac{1}{n}(\delta - \delta P^n)$. Note that $(\delta - \delta P^n)$ is between 0 and 1.

Thus: $F(\delta_n) = \frac{1}{n} \max_i |\delta(i) - \delta P^n(i)| \leq \frac{1}{n}$.

If it takes values that are arbitrarily close to 0, it takes the value 0 because F is a continuous function and Δ is compact.

Since $\inf F = 0$, $\min F = 0$. □

4 February 12, 2020

TA office hours are Thursday 12-1pm in 2-255, and Friday 11am-12pm in 2-231A. Mossel's office hours are Wednesday from 2:30-4pm in 2-434.

4.1 Unique stationary distributions

Not all Markov chains have unique stationary distributions. For example: with $P = \mathbb{1}$, all distributions are stationary!

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

In words, there is a t such that with positive probability, you can go from $x \to y$ in t steps. This is the same as saying that the transition graph of the chain is *strongly connected*. (A strongly connected graph is a directed graph where there is some path between any two points.)

> Gambler's ruin is *reducible* because you can't go anywhere from 0 or n, i.e. $P^t(x, 0) = 0$ for all t and $x > 0$.
>
> $+1$ mod k is irreducible since it's a cycle.
>
> The binary symmetric channel is irreducible as long as $\epsilon > 0$. For $0 < \epsilon < 1$, $P(x, y) > 0$. For $\epsilon = 1$, $P = \mathbb{1}$ and it's $+1$ mod 2.

4.1.1 Harmonic functions

We've considered the case of $\pi = \pi P$, now let's look at $h = Ph$!

Definition. A function $h : \Omega \to \mathbb{R}$ is **harmonic** if

$$h(x) = \sum_{y \in \Omega} P(x, y) h(y) \qquad \text{for all } x \in \Omega$$

sometimes we say h is "harmonic for P" .

Example: Constant functions are always harmonic.

Example: For $P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, every h is harmonic!

Claim: If P is an irreducible finite Markov chain and h is harmonic, then h is constant.

Proof: Via the "maximum principle". Let $m = \max_x h(x)$, and let x_0 be a coordinate such that $h(x_0) = m$.

If h is harmonic, $h(x_0) = m \implies \sum_y P(x_0, y) h(y) = m$.

Since $h(y) \leq m$ for all y and the sum above is a weighted average, it follows that $h(y) = m$ whenever $P(x_0, y) > 0$. We can apply this logic to all such y, and since the chain is irreducible, eventually each $y \in \Omega$ will be reached. So $h = m$. □

Theorem.

Let P be a finite, *irreducible* Markov chain. Then there exists a unique stationary distribution $\pi : \pi P = \pi$.

Proof: We've already demonstrated existence, so we must show uniqueness. If π is stationary, $\pi = \pi P \implies \pi(P - \mathbb{1}) = 0$. It suffices to show then that $\text{rank}(P - \mathbb{1}) = 1$, i.e. that the kernel is 1-dimensional. With the normalization constraint on distributions, this shows that the solution is unique.

We know that the space of solutions of $(P - \mathbb{1})\, h = 0$ is 1-dimensional because it's just the constant function. This implies that $\text{rank}(P - \mathbb{1}) = 1$.

□

4.1.2 Convergence to π

Theorem.

Let P be a finite, irreducible Markov chain and let μ be any probability measure (any arbitrary distribution). Then:

$$\nu_t := \frac{1}{t}\sum_{s=0}^{t-1} \mu P^s \to \pi$$

where π is the stationary distribution of P.

Proof: Assume by contradiction that $\nu_t \not\xrightarrow[t\to\infty]{} \pi$.

Then it follows that there exists a subsequence $\nu_{t(k)} \to \nu \neq \pi$.

> If you have a sequence of probability measures that doesn't converge to π, a subsequence will converge to something else.
>
> Measures are a compact space, so every sequence has a subsequence that converges to something. If it's not π, it'll be some other ν.

Recall $F : \Delta \to [0,1]$ where $F(\mu) = \max_x |\mu(x) - (\mu P)(x)|$.

We've seen that $F(\nu_{t(k)}) \leq 1/t(k)$.

Since F is continuous, $F(\nu) = \lim_{k\to\infty} F(\nu_{t(k)}) = 0$. So ν is stationary.

But π is the unique stationary distribution and $\pi \neq \nu$! *Contradiction!* $\square$

4.1.3 Reversible Markov chains

How do you find a stationary π? You solve $\pi P = \pi$. Writing it out fully:

$$\pi(x) = \sum_y \pi(y)\, P(y,x) \quad \forall x \text{ and } \sum_x \pi(x) = 1$$

Definition. We say that π satisfies the **detailed balance equations** if $\forall\, x, y$ it holds that

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

If there exists such a π, we say that P is **reversible**.

Claim: If π satisfies the detailed balance equations, then π is stationary.

Proof: $\sum_y \pi(y)\, P(y,x) = \sum_y \pi(x)\, P(x,y) = \pi(x) \sum_y P(x,y) = \pi(x)$
So π is stationary.

Claim: If π satisfies the detailed balance equations, then

$$\mathbb{P}_\pi[X_0 = x_0, \ldots, X_n = x_n] = \mathbb{P}_\pi[X_0 = x_n, \ldots, X_n = x_0]$$

or, written out:

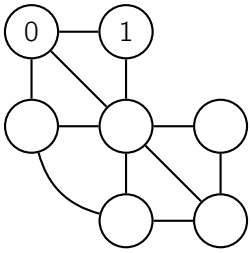
$$\pi(x_0)\, P(x_0, x_1)\, P(x_1, x_2) \ldots P(x_{n-1}, x_n) = \pi(x_n)\, P(x_n, x_{n-1}) \ldots P(x_1, x_0)$$

In short, the arrow of time doesn't exist! A world with no entropy!!!!!

Proof: $\pi(x_0)\, P(x_0, x_1) = \pi(x_1)\, P(x_1, x_0)$ so it comes from there.

4.2 Random walk on a graph

Here's a graph:



Given an undirected graph G, let Ω be the vertices of G. Let

$$P(x, y) = \begin{cases} \frac{1}{\deg(x)} & \text{if } y \text{ is a neighbor of } x \\ 0 & \text{otherwise} \end{cases}$$

This Markov chain is called the **random walk** on G. For example, from vertex 0 you have probability 1/3 to go to 1.

Claim: $\pi(x) = \dfrac{\deg(x)}{\sum_y \deg(y)}$ satisfies the detailed balance equations, i.e. it is stationary.

Proof: $\pi(x)\, P(x, y) = \dfrac{\deg(x)}{\sum_z \deg(z)} \cdot \dfrac{1}{\deg(x)} = \dfrac{1}{\sum_z \deg(z)}$.

This is independent of x and y, so it is the same as $\pi(y)\, P(y, x)$. □

For which graphs is P irreducible? Whenever G is connected.

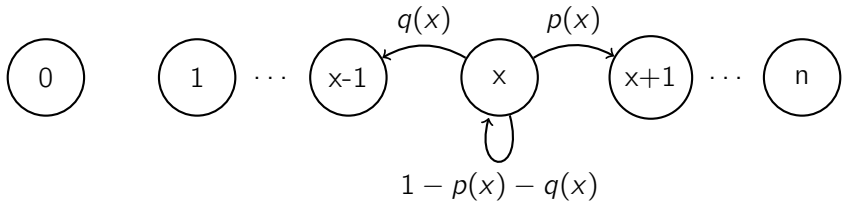
If G is not connected, then G is a union of 2 graphs with no edges between them. If π_1 is stationary for G_1, $\alpha(\pi_1) + (1 - \alpha)\pi_2$ is stationary for all α.

5 February 18, 2020

Let's look at examples of reversible Markov chains.

5.1 Birth and death chains

Let there be states 0,. . . , n with $P(x, y) = 0$ if $|x - y| = 1$.



Define $p(x) = P(x, x + 1)$ where $p(x) > 0$ for $x \leq n - 1$.
Define $q(x) = P(x, x - 1)$ where $q(x) > 0$ for $x \geq 1$.

We want the stationary distribution: $\pi(x)P(x, y) = \pi(y)P(y, x)$. This is true if $|x - y| = 1$ or $x = y$. We need to check:

$$\pi(x)p(x) = \pi(x + 1)q(x + 1)$$

$$\pi(x + 1) = \pi(x)w(x) \qquad \text{where } w(x) = \frac{p(x)}{q(x + 1)}$$

$$\pi(x) = \pi(0) \prod_{y=0}^{x-1} w(y)$$

With the normalization constraint, we get the **explicit equations for π**:

$$\pi(x) = \frac{\prod_{y=0}^{x-1} w(y)}{\sum_z \prod_{y=0}^{z-1} w(y)}$$

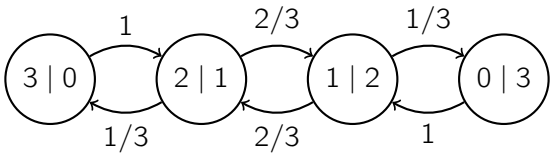
where the empty product has value 1.

Gambler's ruin is a birth and death chain but $q(n) = 0$ and $p(0) = 0$ so the expression for $w(x)$ is meaningless. So it's only sort of a birth and death chain.

The binary symmetric channel chain is an example.

5.1.1 The Ehrenfest chain

Say you have n identical balls in 2 urns. At each step, pick a ball uniformly at random and move it to the other urn. Here is the chain for $n = 3$:



Let x be the number of balls in the left urn. $P(x, x+1) = \frac{n-x}{n}$ and $P(x, x-1) = \frac{x}{n}$.

Claim: The binomial distribution $\pi(x) = 2^{-n}\binom{n}{x}$ satisfies the detailed balance equation.

Proof: We check that $\pi(x)P(x, x+1) = \pi(x+1)P(x+1, x)$. Explicitly:

$$2^{-n}\binom{n}{x}\frac{n-x}{n} = 2^{-n}\binom{n}{x+1}\frac{x+1}{n}$$

This is true! □

The binomial distribution at large n is a normal distribution. This chain is used in physics to model statistical physics processes.

5.2 The ergodic theorem

When and how do Markov chains converge to the stationary distribution? We have seen that if P is irreducible, then for all μ:

$$\frac{1}{t}\sum_{s=0}^{t-1} \mu P^s \to \pi \text{ (stationary)}$$

The ergodic theorem.

Let P be a finite irreducible Markov chain with stationary distribution π. Let $f : \Omega \to \mathbb{R}$. Then, for all x:

$$\mathbb{P}_x\left[\lim_{t\to\infty} \frac{1}{t+1}\sum_{s=0}^{t} f(X_s) = \pi[f]\right] = 1$$

where $\pi[f] = \sum_{x\in\Omega} \pi(x)f(x)$.

Let's parse this. $f(X_s)$ is the value of f at the location of the Markov chain at time t. We take the average value (which is a random variable), and then the limit of this sequence of random variables. That value is almost surely equal to $\pi[f]$, which is the average of f according to π.

Example: Let's apply this to $F(x) = \delta_{x,y}$ for $y \in \Omega$.

$$\mathbb{P}_z\left[\lim_{t\to\infty} \underbrace{\frac{\sum_{s=0}^{t} 1(X_s = y)}{t+1}}_{\text{average of } \{0 \le s \le t \mid X_s = y\}} = \pi(y)\right] = 1$$

where $\pi(y)$ is the probability of y according to the stationary distribution.

5.2.1 Two quick examples

+1 mod k: It satisfies the theorem. The condition and conclusion hold! The chain is finite and irreducible. You spend $1/k$ time at each state.

BSC: The ergodic theorem says $P\left[\lim \frac{1}{t+1}\{0 \le s \le t : X_s = 0\} = \frac{1}{2}\right] = 1$.

5.3 Irreducible aperiodic chains

Recall: P is irreducible if for all x, y there is an $r = r(x, y)$ such that $P^r(x, y) > 0$.

Definition. P is **irreducible and aperiodic** if there exists an r such that $P^r(x, y) > 0$ for all x, y.

Example: +1 mod k is irreducible and *not* aperiodic.

Essentially, in m steps, can you get from all states to all other states? Not in +1 mod k, because after m steps there is only one possible state you can arrive at.

Claim: If $P^r(x, y) > 0$ for all x, y, then $P^{r+1}(x, y) > 0$ for all x, y.

$$P^{r+1} = P \cdot P^r = (\text{row of } P)(\text{column of } P^r) \to \text{strictly positive elements}$$

The rows of P sum to 1, contain non-negative elements, and at least one is strictly positive. The columns of P^r are nonnegative. Hence the result.

Example: BSC is irreducible and aperiodic.

Consider: ± 1 mod k (with probability 1/2) is aperiodic iff k is odd.
+1 mod k with the option to stay in place is aperiodic.

5.4 Total variation distance (T.V.)

Definition. Given 2 probability distributions μ and ν on the same finite space Ω, the **total variation distance** is:

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|$$

Note: $\sum |\mu(x) - \nu(x)|$ is the L_1 distance between ν and μ.

Claim: $\|\mu_1 - \mu_3\|_{TV} \leq \|\mu_1 - \mu_2\|_{TV} + \|\mu_2 - \mu_2\|_{TV}$.

This is the triangle inequality.

Claim: $\|\mu - \nu\|_{TV} = \max_{A \subset \Omega} |\mu(A) - \nu(A)| \leq 1$

Claim: If P is a Markov chain, then $\|\mu P - \nu P\|_{TV} \leq \|\mu - \nu\|_{TV}$

5.5 Convergence to equilibrium

Theorem. Let P be the transition matrix of a finite irreducible aperiodic Markov chain. Then for any distribution μ:

$$\lim_{t \to \infty} \mu P^t = \pi$$

Moreover, there exist constants C and $\alpha < 1$ such that for all distributions μ:

$$\left\| \mu P^t - \pi \right\|_{TV} \leq C \alpha^t$$

Consider the BSC, with $0 < \epsilon < 1$. For any vector $\begin{pmatrix} x & 1-x \end{pmatrix}$ where $0 \leq x \leq 1$,

$$\begin{pmatrix} x & 1-x \end{pmatrix} \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}^t \xrightarrow{t \to \infty} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

Proof: We know there is an r such that $P^r(x, y) > 0$ for all x, y. Let's assume $r = 1$ so $P(x, y) > 0 \quad \forall x, y$.

Then we can write: $P = (1 - \theta)\Pi + \theta Q$.
Π has all rows equal to π for $0 < \theta < 1$.
Q is a Markov matrix so every row sums to 1, and all entries are > 0.

Claim: $\mu\Pi = \pi \quad \forall \mu$, and $A\Pi = \Pi$ for all transition matrices A.

$\mu\Pi$ averages each column of Π, but each column i is just π_i.

$\pi P = \pi \implies \pi Q = \pi$ because $\pi P = \pi$. So $\Pi Q = \Pi$.

Then:

$$\begin{aligned} P^t &= ((1-\theta)\Pi + \theta Q)^t \\ &= \theta^t Q^t + (1-\theta)^t \Pi \qquad (\because \Pi Q = Q\Pi = \pi) \\ \mu P^t &= (1-\theta^t)\pi + \theta^t \mu Q^t \\ \left\|\mu P^t - \pi\right\|_{TV} &= \theta^t \left\|\mu Q^t - \pi\right\|_{TV} \leq \theta^t \end{aligned}$$

In the general case, $P^r(x,y) > 0$. So $\left\|\mu P^{rt} - \pi\right\| \leq \theta^t \quad \forall t$.

$$\begin{aligned} \left\|\mu P^{rt+j} - \pi\right\|_{TV} &= \left\|\mu P^{rt+j} - \pi P^j\right\|_{TV} \\ &\leq \left\|\mu P^{rt} - \pi\right\|_{TV} \\ &\leq \theta^t \end{aligned}$$

□

6 February 19, 2020

Given a "description" of a probability distribution D, how do you sample from it?

6.1 Sampling from distributions

In the 1940s, von Neumann asked: Given access to an *independent and identically distributed* (i.i.d.) sequence Y_i of Bernoulli p coin tosses where $0 < p < 1$, how do you generate a fair coin toss?

The answer is that you look at pairs of unfair tosses, (Y_1, Y_2). If you get (0,1), call it a head. If you get (1,0), call it a tail. Otherwise, discard and repeat. These pairs of unfair tosses can be used as fair tosses.

Now, assume that you have access to (U_i), a sequence of i.i.d uniform random variables.

Exercise: Given $U \sim U[0,1]$, it is easy to sample from a distribution R with a given *cumulative distribution function* (CDF) F.
Hint: generate U and output $F^{-1}(U)$.

Example: Generate uniformly at random a permutation over S_n. For example, how do you really shuffle a deck of cards – by sampling S_{52} uniformly at random?

You have a couple options. You could list all $n!$ permutations. Then generate a uniform number between 1 and $n!$, but this takes $n!$ time. This is virtually impossible.

Here's another option. At position 1, choose 1 of the 52 cards at random. At position 2, choose one of the remaining 51 at random. Etc. This is much faster!

6.1.1 Graph colorings

Example: Given an undirected graph $G = (V, E)$ of maximal degree d on n nodes, find a uniform coloring of the graph with q colors where $q \geq d+1$.

Definition. A legal **coloring** of a graph $G = (V, E)$ with q colors is a map $F : V \to [q]$ where $[q] = \{1, 2, \ldots, q\}$ such that if (u, v) then $f(u) \neq f(v)$. That is, if there is an edge connecting two nodes u and v, they have different colors.

Back to the example. Let $G = (V, E)$, and $\Omega = \{\text{all } q \text{ colorings of } G\}$. Our goal is to sample a uniform element of Ω.

Suggestion: sequentially sample vertex color legally. If stuck, restart. This has unknown chance of success and isn't really uniform. It depends a lot on the starting point.

Slow but sure: choose one of the q^n ($n = |V|$) maps $f : V \to [q]$ at random. If it's legal, you're done. Otherwise retry. This is very slow because you'll have to reject a lot of maps, but it will work.

Why are these sampling questions important? Understanding energy distributions in physical chemistry is reliant on this. Fixing gerrymandering relies on sampling!

6.2 Sampling with the Metropolis chain

Definition. Given a desired distribution π on a state space Ω and a symmetric Markov chain $\psi(x, y)$ on Ω, the **Metropolis chain** is:

$$P(x, y) = \psi(x, y) \times \min\left(1, \frac{\pi(x)}{\pi(y)}\right) \quad \text{for } y \neq x$$
$$P(x, x) = 1 - \sum_{z \neq x} P(x, z)$$

What does this mean? Go according to ψ. If $\pi(y) \geq \pi(x)$, stay at y. If $\pi(y) < \pi(x)$, stay with probability $\pi(y)/\pi(x)$. Otherwise go back to x.

Claim: If P is reversible with respect to π, π is stationary for P.

Proof: As follows:

$$\begin{aligned}
\pi(x)P(x,y) &= \pi(x)\psi(x,y) \times \min\left(1, \frac{\pi(y)}{\pi(x)}\right) \\
&= \min\left(\pi(x)\psi(x,y), \pi(y)\psi(x,y)\right) \\
&= \min\left(\pi(x)\psi(x,y), \pi(y), \psi(y,x)\right) \qquad \because \psi \text{ is symmetric} \\
&= \pi(y)\psi(y,x) \times \min\left(1, \frac{\pi(x)}{\pi(y)}\right)
\end{aligned}$$

Does this work?

- Maybe if P is irreducible and aperiodic.
  (How does the answer depend on ψ and π?)
- You need to be able to run ψ efficiently, and compute $\pi(y)/\pi(x)$ efficiently.
- You want convergence to π to be fast.

Definition. The **general Metropolis chain** given a desired distribution π on a state space Ω and a Markov chain $\psi(x,y)$ on Ω is defined as:

$$\begin{aligned}
P(x,y) &= \psi(x,y) \times \min\left(1, \frac{\pi(y)\psi(y,x)}{\pi(x)\psi(x,y)}\right) \qquad y \neq x \\
P(x,x) &= 1 - \sum_{z \neq x} P(x,z)
\end{aligned}$$

Proof: π is stationary.

$$\begin{aligned}
\pi(x)P(x,y) &= \min\left(\pi(x)\psi(x,y), \pi(y)\psi(y,x)\right) \\
&= \pi(y)\psi(y,x) \times \min\left(1, \frac{\pi(x)\psi(x,y)}{\pi(y)\psi(y,x)}\right) \qquad \text{Neat!}
\end{aligned}$$

6.2.1 Sampling a vertex of an unknown graph

Goal: Sample uniformly random vertex of an unknown graph $G = (V, E)$.

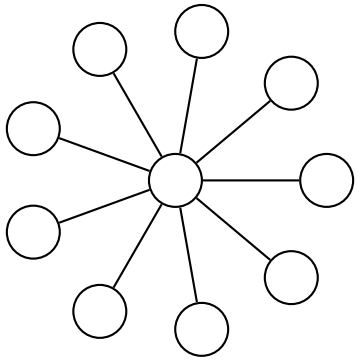
Given $v \in V$. If you are at v, assume that you know its neighbors in the graph. It is natural to take $\psi(x,y) = \frac{1}{\deg(x)}$, a random walk on G.

The stationary distribution of ψ is $\nu(x) = \dfrac{\deg(x)}{\sum_{y \in V} \deg(y)}$

We want to uniformly sample the vertices, so our *desired* π has $\pi(y) = \pi(x)$ for all x, y. What would the Metropolis chain achieve here?

$$P(x,y) = \psi(x,y) \times \min\left(1, \frac{\pi(y)\psi(y,x)}{\pi(x)\psi(x,y)}\right)$$
$$= \psi(x,y) \times \min\left(1, \frac{\deg(x)}{\deg(y)}\right)$$

Example: We can look at a star graph:



What would the Metropolis chain do here? At the center, it would choose a neighbor uniformly at random, since $\deg(x) > \deg(y)$. If it's at one of the tips, it would stay put with probability $1 - 1/n$ for an n-star graph. Otherwise, it would go to the center.

This is better than a standard random walk because you're not passing through the center on every other step.

6.2.2 Sampling graph colorings

Let π be a uniform distribution over all legal graph colorings.
Let ψ be a random walk over all $f : V \to [q]$.

$P(x,y) = \psi(x,y) \times \min\left(1, \frac{\pi(y)}{\pi(x)}\right)$ $\quad$ $\pi(y)$ is 0 if not legal.

The Metropolis chain on graph colorings does the following:

- Start from a legal coloring.
- Pick a vertex v uniformly at random. Pick a color c uniformly at random.
- Try to change the color of v to c.
- If it's legal, keep the change. If it's illegal, discard and pick again.

7 February 24, 2020

A variant on the Metropolis chain is Glauber dynamics.

7.1 Glauber dynamics

Glauber dynamics for coloring sampling looks like this:

- Given a legal coloring x, choose a vertex $v \in V$ at random.
- Choose one of the legal colors at v uniformly at random (including the current color).

Claim: Glauber dynamics is reversible with respect to uniform distributions over colorings.

Proof: We need to check that $\pi(x)P(x,y) = \pi(y)P(y,x)$. Since we're verifying the uniform distribution ($\pi(x) = \pi(y)$), we just need to check that the chain is symmetric.

If x and y are different at more than 1 node, $P(x,y) = P(y,x) = 0$.
If x and y differ only at one node v:

$$P(x,y) = \frac{1}{m} = P(y,x) \quad \text{where } m \text{ is the number of legal ways to color } v$$

So it is reversible! □

Definition. Given $\Omega = S^V$, let $\Omega(x,v) = \{y \mid y \text{ differs from } x \text{ only at } v\}$.
General Glauber dynamics are defined as follows:

- given $x \in \Omega$, pick $v \in V$ uniformly at random
- recolor v to obtain a coloring y with probability $\dfrac{\pi(y)}{\pi(\Omega(x,v))}$

(The denominator in the last term is the sum of π for all states in y.)

Claim: π is reversible for this chain.

7.2 Questions about the Metropolis and Glauber dynamics

1. When is the chain irreducible?
2. Even when it is, how long does it take for the chain to converge?

We'll provide some answers from graph coloring.

Claim: Given a graph $G = (V, E)$ with maximal degree d. If the number of colors q in a coloring of G satisfies:

$$q \geq d + 2$$

then the Metropolis or Glauber chain is irreducible and aperiodic (also called **ergodic**).

Proof: We need to show that if σ, τ are two colorings of G with q colors, then there is a path of positive probability transitions from σ to τ.

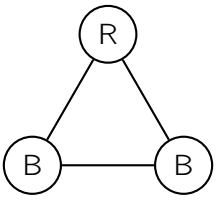
We are going to define a path $\sigma = \sigma_0, \sigma_1, \sigma_2, \ldots, \sigma_n = \tau$, such that $\sigma_i(j) = \tau(j)$ for $j \leq i$. Clearly $\sigma_0 = 0$.

Given $\sigma_{i=1}$, define σ_i as follows. First let $\sigma_{i-1} = \sigma_i$.

If $i, j \in E$ and $j > i$, update $\sigma_i(j)$ to a legal coloring which is different from $\tau(i)$. This is possible since $q \geq d + 2$.

Then set $\sigma_i(i) = \tau(i)$. At the end, $\sigma_n = \tau$. All moves had positive probability! □

Can we do better than this?
No. Consider $d = 2$, $q = 3$. This is a chain that satisfies that:



There is no path from this to every other coloring. It's not irreducible!

More generally, if G is a q-clique (if it has q vertices all connected to each other), then $d = q - 1$ and the chain is not irreducible.

7.3 Coloring a star graph

Let's go back to the star graph. Consider a star graph with n tips v_i and a center node v_0.
Claim: The maximal degree $d = n$. For $q \geq 3$ the chain is ergodic.

This does, however, take a while to converge. Since it's ergodic it will converge to the stationary distribution but in how much time?

Naively you might say that it converges exponentially fast, since we showed that $\left\|\mu P^t - \pi\right\|_{TV} \leq C\alpha^t$. But what if $C \gg 1$ and α is nearly 1? Then it would be really slow.

Theorem. (Slow mixing of star coloring.)

Consider Glauber dynamics or the Metropolis algorithm for the star graph with $q = 3$ and $n + 1$ nodes. Then there exists a state x such that for all t:

$$\left\|\delta_x P^{t-1} - \pi\right\|_{TV} \geq \frac{2}{3} - \frac{4t}{2^n - 2}$$

Note:

1. While $\delta_x P^t \to \pi$, convergence is very slow.
   In particular, for $\left\|\delta_x P^t - \pi\right\|_{TV} \leq 0.01$, we would need $t \geq 2^n/8$.

2. There is an alternate easier way to sample star-coloring.

A good question from the room: how do you find the initial legal coloring for both Glauber dynamics and the Metropolis algorithm? If $q \geq d + 1$, a greedy algorithm will work.

7.3.1 Bottleneck theorem

Bottleneck Theorem.

Suppose P is an irreducible, aperiodic Markov chain with stationary distribution π. Suppose Ω can be partitioned into three sets

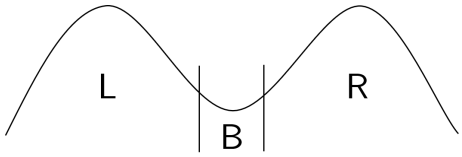
$$\Omega = L \sqcup B \sqcup R \qquad (\sqcup \text{ means disjoint union})$$

such that $P(x, y) = 0$ when $x \in L$ and $y \in R$.

Then there exists $x \in L$ such that for all t,

$$\left\|\delta_x P^{t-1} - \pi\right\|_{TV} \geq \pi(R \sqcup B) - t\frac{\pi(B)}{\pi(L)}$$

The Ω we are describing looks like this:



There is no direct path from L to R without passing through B.

Let's apply the theorem to a star coloring with $q = 3$. Call v_0 the star center and $v_1 \dots v_n$ the tips.

Let $B = \{\sigma \mid \sigma(v_1) = \cdots = \sigma(v_n) \in \{1, 2\}\}$.
Let $L = \{\sigma \mid \sigma(v_0) = 0, \sigma \neq B\}$.
Let $R = \{\sigma \mid \sigma(v_0) \in \{1.2\}, \sigma \neq B\}$.

In words, the L part is all colorings where the center node is colored 0. The R part is all colorings where the center node is colored 1 or 2. The bottleneck is all colorings either where all the tips are colored 1, or all tips are colored 2.

Then we have: $|B| = 4$, $|L| = 2^n - 2$, $|R| = 2 \cdot 2^n - 2$, and $|\Omega| = 3 \cdot 2^n$.

The bottleneck theorem then says that there exists an $x \in L$ such that:

$$\left\|\delta_x P^{t-1} - \pi\right\|_{TV} \geq \pi(R \sqcup B) - t\frac{\pi(B)}{\pi(L)} \geq \frac{2}{3} - \frac{4t}{2^n - 2}$$

□

8 February 26, 2020

Let's start with a proof of the bottleneck theorem.

8.1 Proof of Bottleneck theorem

Note if $x \in L$, to get to R you must go through B.

$$\mathbb{P}[X_t \in R \cup B] \leq \mathbb{P}_x\left(\bigcup_{s=1}^{t}\{X_s \in B\}\right) \leq \sum_{s=1}^{t}\mathbb{P}_x[X_s \in B]$$

The first expression is the probability at t that you're in R or B. The second one is the probability that at $0 < s < t$ you were in B.

A word on notation: we use $\mathbb{P}$ to distinguish from transition probability, P. This is the probability of an event. $\mathbb{P}_x[\dots]$ means $\mathbb{P}[\cdots \mid X_0 = x]$. $\mathbb{P}_\pi[\dots]$ means $\mathbb{P}[\cdots \mid X_0 \sim \pi]$. $\sim$ means "distributed according to".

We choose $x \in L$ randomly according to π.

$$\mathbb{E}_{x\sim\pi}\left[\mathbb{P}_x[X_t \in R\cup B] \mid x \in L\right] \le \mathbb{E}_\pi\left[\sum_{s=1}^{t}\mathbb{P}_x[X_s \in B] \mid x \in L\right]$$

$$= \frac{1}{\pi(L)}\mathbb{E}_\pi\left[\sum_{s=1}^{t}\mathbb{P}_x[X_s \in B]\cdot I_L(x)\right]$$

$I_L(x)$ is an indicator: 1 if $x \in L$, 0 o.w.

$$\le \frac{1}{\pi(L)}\mathbb{E}_\pi\left[\sum_{s=1}^{t}\mathbb{P}_x[X_s \in B]\right]$$

ignoring the indicator

$$= \frac{1}{\pi(L)}\sum_{s=1}^{t}\mathbb{E}_\pi\left[\mathbb{P}_x[X_s \in B]\right]$$

$$= \frac{1}{\pi(L)}\cdot t\,\pi(B)$$

$$\because x\sim\pi \to X_s \sim \pi \to \mathbb{P}(X_s \in B) = \pi(B)$$

$\mathbb{E}_\pi\left[\mathbb{P}_x[X_t \in R\cup B] \mid x \in L\right] \le \mathbb{E}\left[X_t \in R\cup B\right] \le \frac{t\,\pi(B)}{\pi(L)}$.

So there exists x such that $\mathbb{P}_x[X_t \in R\cup B] \le \frac{t\,\pi(B)}{\pi(L)}$.
If the average value of a group of elements is less than something, there must be one element that is also less than that something.

Now let's look at $\left\|\pi - \delta_x P^t\right\|_{TV}$.
$\delta_x P^t$ is the probability distribution of where you might be at t given that you started at x.

$$\left\|\pi - \delta_x P^t\right\|_{TV} \ge \left|\pi(R\cup B) - \delta_x P^t(R\cup B)\right|$$

$$\ge \pi(R\cup B) - \mathbb{P}_x\left[X_t \in R\cup B\right] \ge \pi(R\cup B) - \frac{t\,\pi(B)}{\pi(L)}$$

8.2 Showing quick convergence

There are two ways for us to show quick convergence of a chain: 1) strong stationary times, and 2) coupling. Let's talk about the first.

Definition. τ is a **stopping time** for a sequence of random variables $X_0, X_1, \dots$ if τ is a random variable taking values $\{0, 1, 2, \dots\}$ such that for all $t \in \{0, 1, \dots\} \cup \{\infty\}$ there exists a set A_t such that $\tau = t$ iff $(X_0, \dots, X_t) \in A_t$.

Example: $X_0, X_1, \dots$ are i.i.d coin tosses. For $t \ge 2$, let $\tau = t$ if $X_{t+2} = H$.

Definition. Consider a Markov chain given by $X_t = f(X_{t-1}, Z_t)$ where Z_t are i.i.d random variables. A **randomized stopping time** for X_t is a stopping time for Z_t.

Example: Consider this chain:

$$P = \begin{pmatrix} 0.2 & 0.4 & 0.4 \\ 0.3 & 0.3 & 0.4 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$$

Take Z_t to be i.i.d on $U[0,1]$ interval.

$$f(1, Z_t) = \begin{cases} 1 & \text{if } Z_t \leq 0.2 \\ 2 & \text{if } Z_t \leq 0.6 \\ 3 & \text{otherwise} \end{cases} \quad \dots \quad f(3, Z_t) = \begin{cases} 1 & \text{if } Z_t \leq 0.1 \\ 2 & \text{if } Z_t \leq 0.2 \\ 3 & \text{otherwise} \end{cases}$$

8.2.1 Lazy random walk on hypercube

Definition. This is the **lazy random walk on a hypercube**, $\{0,1\}^n$. Let $Z_t = (I_t, B_t)$ where $I_t \sim U(\{1 \dots n\})$ and $B_t \sim U(\{0,1\})$, i.e. two bits. $U(X)$ means uniformly distributed on X. The walk is defined by

$$f_t(x, z = (i,b)) = y \text{ where } y_i = b \text{ and } y_j = x_j \text{ for } j \neq i$$

Example: Let's consider a 3-bit walk.

$$\begin{aligned} &\text{Start at } (0,0,0). \text{ Let } z = (2,0). \\ &\to (0,0,0). \text{ Let } z = (1,1). \\ &\to (1,0,0). \end{aligned}$$

At each step, you randomly pick a coordinate and assign it a random value.

Let $\tau = \min(t : \{I_1, \dots, I_t\} = \{1, \dots, n\})$ be the **refresh time**. That is, the time at which each coordinate has been updated. τ has the same distribution as a coupon collector! We're drawing uniformly from 1 to n.

8.2.2 Strong stationary time

Definition. The **strong stationary time** for a Markov chain X_t is a randomized stopping time τ such that for all x,

$$\mathbb{P}_x[\tau = t, X_t = y] = \mathbb{P}_x[\tau = t]\, \pi(y)$$

where π is a stationary distribution.

When you stopped, you stopped according to the stationary distribution. The distribution of X_t is π even when we require that τ is a specific t.

Example: The refresh time is a strong stationary time.

In the 3-bit case, if I tell you I stopped after 2 stops, I'm lying.
If I tell you I stopped after 5 stops, that's not really useful information – each bit is equally likely to be 0 or 1. That's the stationary distribution.

Claim: If τ is a strong stationary time, then for any μ:

$$\left\|\mu P^t - \pi\right\|_{TV} \leq \max_x \mathbb{P}_x[\tau > t]$$

The distribution in TV from π is upper-bounded by the chance you haven't stopped. (We'll come back to this.)

Let's apply it to the lazy random walk on a hypercube and refresh time.

For the coupon collector, $\mathbb{P}[\tau \geq n\log n + cn] \leq e^{-c}$. So if $t \geq n\log n + cn$, $\left\|\mu P^t - \pi\right\|_{TV} \leq e^{-c}$. If you run the chain $n\log n + cn$ times, it converges quickly!

Let's show that for the coupon collector, $\mathbb{P}[\tau \geq n\log n + cn] \leq e^{-c}$.

Proof. Let Z_t be the number of coupons not yet collected at time t.

$$\mathbb{P}[\tau > t] = \mathbb{P}[Z_t \geq 1] \leq \mathbb{E}[Z_t] = n\left(1 - \frac{1}{n}\right)^t$$

If $t = n\log n + cn$,

$$n\left(1 - \frac{1}{n}\right)^{n(\log n + c)} \leq n\, e^{-\log n - c} = e^{-c}$$

□

8.2.3 Top to random card shuffle

Take the top card and put it in a random place. How long do you have to do this for it to be shuffled?

This is an ergodic Markov chain. (Think about it!)

Claim: Given that at time t, there are k cards under the original bottom card, each of the $k!$ orderings of the k cards is equally likely.

You can show this inductively.

Claim: Let τ be (the first time the original bottom card is on top + 1). τ is a strong stationary time.

$\mathbb{P}[\tau > t]$. τ is like the coupon collector in reverse.
The probability that the 1st card goes under the bottom is $1/n$.
The probability that the 2nd card goes under the bottom is $2/n$.
etc.

$\tau = X_1 + \cdots + X_{n-1}$ where $X_i \sim \text{Geom}(i/n)$, $X_n = 1$. This is the coupon collector distribution.

So, $\mathbb{P}[\tau \geq n \log n + cn] \leq e^{-c}$.

9 March 2, 2020

The other quick convergence method is coupling.

9.1 Coupling to establish quick convergence

Definition. A **coupling** of two probability distributions μ, ν on the same space Ω is a pair of random variables (X, Y) on Ω^2 such that:

$$\mathbb{P}[X = x] = \mu(x),\ \mathbb{P}[Y = y] = \nu(y) \qquad \text{for all } x, y \in \Omega$$

Example: $\Omega = \{0, 1\}$. $\mu(0) = \mu(1) = 1/2$ and $\nu(0) = 2/3, \nu(1) = 1/3$.
Coupling 1: X is a fair coin toss. Y is an unfair coin toss with probability $2/3$ for 0 and $1/3$ for 1. X and Y are independent.
Coupling 2: With probability $1/3$, $X = Y = 1$. With probability $1/2$, $X = Y = 0$. With probability $1/6$, $X = 1$ and $Y = 0$. X and Y are not independent here!

Definition. A **coupling of Markov chains** with transition matrix P is a process (X_t, Y_t) where each of X_t and Y_t is a Markov chain with transition matrix P such that:

$$\text{if } X_s = Y_s, \text{ then } X_t = Y_t \qquad \text{for all } t \geq s$$

Claim: $\|\mu - \nu\|_{TV} = \inf\left(\mathbb{P}[X \neq Y] \mid (X, Y) \text{ is a coupling of } \mu, \nu\right)$.

What this means is that one way to get the total variation distance is to consider all couplings where $X \sim \mu$ and $Y \sim \nu$, and look for $\mathbb{P}[X \neq Y]$. For example, from Coupling 2 above:

$$\|\mu - \nu\|_{TV} = \frac{1}{2}\left(\frac{1}{6} + \frac{1}{6}\right) = \frac{1}{6} \qquad \text{normal TV definition}$$

$$\|\mu - \nu\|_{TV} = \mathbb{P}[X \neq Y] = \frac{1}{6} \qquad \text{from the claim}$$

Proof: Let's consider the upper bound. $\|\mu - \nu\|_{TV} = \max_A |\mu(A) - \nu(A)|$

$$\begin{aligned}
\text{Fix } A. \quad \mu(A) - \nu(A) &= \mathbb{P}[X \in A] - \mathbb{P}[Y \in A] \\
&\leq \mathbb{P}[X \in A, Y \notin A] \qquad \text{(draw a Venn diagram)} \\
&\leq \mathbb{P}[X \neq Y] \\
\nu(A) - \mu(A) &= \mathbb{P}[Y \in A] - \mathbb{P}[X \in A] \\
&\leq \mathbb{P}[Y \in A, X \notin A] \\
&\leq \mathbb{P}[X \neq Y] \\
\implies |\mu(A) - \nu(A)| &\leq \mathbb{P}[X \neq Y] \qquad \text{cool!}
\end{aligned}$$

Let's look at the lower bound.

$$\min(\mu(x), \nu(x)) = \frac{1}{2}(\mu(x) + \nu(x)) - \frac{1}{2}|\mu(x) - \nu(x)|$$

This makes sense: the average of two numbers minus half the distance between them is the smaller one.

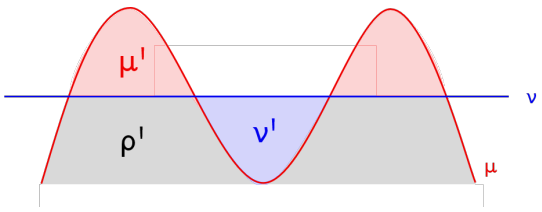
$$\begin{aligned}
a := \sum_{x \in \Omega} \min(\mu(x), \nu(x)) &= \frac{1}{2}\sum_{x \in \Omega}(\mu(x) + \nu(x)) - \frac{1}{2}\sum_{x \in \Omega}|\mu(x) - \nu(x)| \\
&= 1 - \|\mu - \nu\|_{TV}
\end{aligned}$$

where the 1 comes from the fact that half the sum of two measures is 1, and the second term is just the total variation distance.

Now we define three probability distributions:

$$\begin{aligned}
\rho(x) &:= \frac{1}{a}\min(\mu(x), \nu(x)) && \text{part common to } \mu \text{ and } \nu \\
\mu'(x) &:= \frac{1}{1-a}(\mu(x) - a\rho(x)) && \text{only } \mu \\
\nu'(y) &:= \frac{1}{1-a}(\nu(y) - a\rho(y)) && \text{only } \nu
\end{aligned}$$

We can visualize this below:



Now we use the following coupling:

- with probability a, pick x according to ρ and let $X = Y = x$
- with probability $1 - a$, pick $X \sim \mu'$, $Y \sim \nu'$ independently.

Then we have that the marginal of X is μ, and the marginal of Y is ν. So $\mathbb{P}[X \neq Y] = 1 - a = \|\mu - \nu\|_{TV}$ as calculated earlier. $\square$

9.2 Coupling of Markov chains

Claim: Let (X_t, Y_t) be any coupling of Markov chains with some transition matrix P, with $X_0 = x$ and $Y_0 = y$. Then:

$$\left\|\delta_x P^t - \delta_y P^t\right\|_{TV} \leq P_{xy}[T > t]$$

where $T = \min(t : X_t = Y_t)$ (the time when the two chains couple).

If the two chains have met, the distance between them is 0. Otherwise, they are approaching.

Claim: If $\left\|\delta_x P^t - \delta_y P^t\right\|_{TV} \leq \epsilon$ for all x and y, then $\left\|\mu P^t - \pi\right\|_{TV} \leq \epsilon$, where μ is any measure and π is stationary.

Proof:

$$\begin{aligned}\left\|\delta_x P^t - \pi\right\|_{TV} = \left\|\delta_x P^t - \pi P^t\right\|_{TV} &= \left\|\delta_x P^t - \left(\textstyle\sum_y \pi(y)\delta_y\right) P^t\right\|_{TV} \\ &= \left\|\textstyle\sum_y \pi(y)\left(\delta_x P^t - \delta_y P^t\right)\right\|_{TV} \\ &\leq \textstyle\sum_y \pi(y)\left\|\delta_x P^t - \delta_y P^t\right\|_{TV} \leq \epsilon\end{aligned}$$

where the last inequalities comes from the triangle inequality and the given.

$$\begin{aligned}\text{Now: } \left\|\mu P^t - \pi\right\|_{TV} &= \left\|\textstyle\sum_x \mu(x)\left(\delta_x P^t - \pi\right)\right\|_{TV} \\ &\leq \textstyle\sum_x \mu(x)\left\|\delta_x P^t - \pi\right\|_{TV} \leq \epsilon\end{aligned}$$

using the triangle inequality again. □

This is useful since it shows that if you can couple the Markov chain, you can bound the total variation distance to π!

9.2.1 Lazy random walk on a cycle

Consider the cycle $0, 1, \ldots, k-1 \bmod k$.

The "lazy" descriptor means that at a node i, the walk stays put with probability $1/2$, goes to $i+1$ with probability $1/4$, and goes to $i-1$ with probability $1/4$.

Fo the coupling, we need to find an X_0, Y_0 such that each follows the lazy walk rule independently but taken together, they do something interesting.

Let's use:

- with probability $1/2$, X moves and Y stays.
- with probability $1/2$, Y moves and X stays.

In the first case, $Y_{t+1} = Y_t$ and $X_{t+1} = X_t \pm 1 \bmod k$.

Claim: Consider this coupling with $X_0 = x$ and $Y_0 = y$. Then

$$\mathbb{P}[T > t] \leq \frac{k^2}{4t}$$

Proof: Let D_t be the clockwise distance between X_t and Y_t. Then:

$$\mathbb{P}[D_{t+1} = D_t \pm 1] = \frac{1}{2} \qquad T = \min(t : D_t = 0, k)$$

This looks like gambler's ruin! $\mathbb{E}[T] = D_0(k - D_0) \leq \frac{k^2}{4}$.

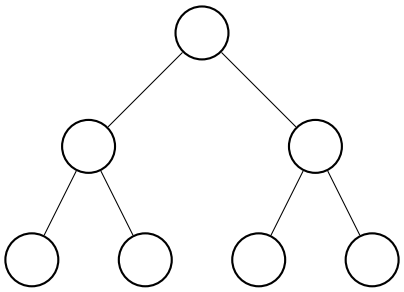
The Markov inequality says that $\mathbb{P}[T > t] \leq \frac{\mathbb{E}[T]}{t}$. So $\mathbb{P}[T > t] = \frac{k^2}{4t}$. $\square$

Corollary: For any μ, $\left\|\mu P^t - \pi\right\|_{TV} \leq \frac{k^2}{4t}$.

Remark: By the Central Limit Theorem, this is the right order of k for the bound, since after ζk^2 stops the walk is likely at most $10\sqrt{\zeta}k$ away from the start. (10 is just some number of standard deviations.)

9.2.2 Lazy random walk on binary tree

Consider a binary tree with k levels. Here we've drawn $k = 2$:



Given that the random walk is at a state X, stay with probability $1/2$. Otherwise move to a uniformly chosen neighbor. For a leaf that's $1/2$ to the parent, for any other node is $1/6$ to the parent or either child.

We'll use a similar coupling for this:

- if X and Y are at *different* levels in the tree, with probability $1/2$ X moves and Y stays (and vice versa)
- if they're at the *same* level, they stay / go up / go down together

Let $T = \min(t : X_t = Y_t)$. WLOG, assume y is at a level $\leq x$.

Claim: $T \leq T' \coloneqq \min(t : Y_t = \text{root})$.
Claim: $\mathbb{E}[T] \leq \mathbb{E}[T'] \leq 6n$.

10 March 4, 2020

So far we have talked about Ω on finite space. Today, Ω is countably infinite! Like integers and rationals.

Let's generalize some definitions:

- A probability distribution over Ω is a vector $(v_i : i \in \Omega)$ such that $v_i \geq 0$ for all $i \in \Omega$ and $\sum_{i \in \Omega} v_i = 1$.
- We say that a probability distribution π on Ω is *stationary* if $\pi P = \pi$, or $(\pi P)(x) = \sum_{y \in \Omega} \pi(y) P(y, x)$.
- P is *irreducible* if for all $x, y \in \Omega$, there exists $t = t(x, y)$ such that $P^t(x, y) > 0$.
- P is *ergodic* if $\gcd(t : P^t(x, x) > 0) = 1$.

Note that the last definition does not depend on x. In finite situations, this is equivalent to the statement: $\exists\, t$ s.t. $P^t(x, y) > 0$ for all x, y.

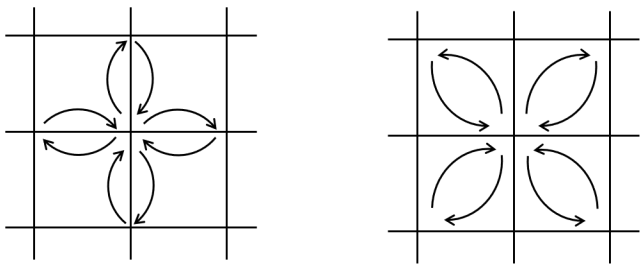
Examples:

$P(i, i) = P(i, i+1) = P(i, i-1) = 1/3$ for all $i \in \mathbb{Z}$.
This is a walk on $\mathbb{Z}$. It is irreducible and aperiodic, but note that there is no t such that $P^t(x, y) > 0$ for all x and y (where $|x - y| \geq t + 1$).

A *random walk on* $\mathbb{Z}$ is: $P(i, i+1) = P(i, i-1) = 1/2$ for all $i \in \mathbb{Z}$.

The *random walk on* $\mathbb{Z}^2$ is (where e_1 and e_2 are basis vectors):
$P(x, x + e_1) = P(x, x - e_1) = P(x, x + e_2) = P(x, x - e_2) = 1/4$.

The *corner random walk on* $\mathbb{Z}^2$ is $P(x, x \pm e_1 \pm e_2) = 1/4$. This is not irreducible!



Left: random walk on $\mathbb{Z}^2$. Right: corner walk on $\mathbb{Z}^2$.

The *biased random walk on* $\mathbb{Z}$ is: $P(i, i+1) = p$, $P(i, i-1) = 1 - p$ with $p > 1/2$.

We can have walks on higher dimensions of $\mathbb{Z}$ in the same way.

10.1 Recurrence and transience

Definition. Define $\tau_x^+ := \min(t \geq 1, X_t = x)$.
We say x is **recurrent** if $\mathbb{P}_x\left[\tau_x^+ < \infty\right] = 1$.
We say x is **transient** if $\mathbb{P}_x\left[\tau_x^+ < \infty\right] < 1$. (or $\mathbb{P}_x\left[\tau_x^+ = \infty\right] > 0$)

We use τ_x to mean that we are allowing $\tau_x = 0$.

Claim: Every x is recurrent for the random walk on $\mathbb{Z}$.

Proof: Write $p = \mathbb{P}_1[\tau_0 = \infty]$. 0 is recurrent iff $p = 1$. p is the probability that starting from any x you visit $x - 1$. We can calculate the probability of getting to 0 after taking one step from 1. There's a half probability that we went from 1 to 0, and a $\frac{1}{2}p^2$ probability that we went to 2 and then hopped backwards twice. So we can write: $p = \frac{1}{2}(1 + p^2)$. Solving, we get $p = 1$. □

Claim: Every x is transient for the biased random walk on $\mathbb{Z}$.

Based on this claim, for Gambler's ruin with a chance of winning $p = 1 - q > 1/2$ and $0 < k \leq n$,

$$\mathbb{P}_k\left[\tau_n < \tau_0\right] = \frac{1 - (q/p)^n}{(1 - q/p)^n}$$

10.2 The Green function

Definition. The **Green function** $G(x, y)$, the expected number of times a chain visits y from x, is:

$$G(x, y) = \sum_{t=0}^{\infty} P^t(x, y) = \mathbb{E}_x\left[\sum_{t=0}^{\infty} 1(X_t = y)\right]$$

For finite irreducible Markov chains, $G(x, y) = 0$.

Theorem.

The following are equivalent for an irreducible Markov chain:

1. There exists x_0 such that $G(x_0, x_0) = \infty$.
2. For all x and y, $G(x, y) = \infty$.
3. There exists x_0 such that $\mathbb{P}_{x_0}\left[\tau_{x_0}^+ < \infty\right] = 1$.
4. For all x and y, $\mathbb{P}_x\left[\tau_y^+ < \infty\right] = 1$.

Definition. A chain satisfies properties 1-4 above is **recurrent**. Otherwise, it is **transient**.

Proof of theorem. Clearly 2 $\implies$ 1. So first we show 1 $\implies$ 2.

Fix x and y. Since P is irreducible, there exist $u, v \geq 1$ such that:

$$P^u(x, x_0) > 0 \qquad P^v(x_0, y) > 0$$

Then:

$$\begin{aligned} G(x,y) = \sum_t P^t(x,y) &\geq \sum_s P^u(x,x_0) P^s(x_0,x_0) P^v(x_0,y) \\ &= P^u(x,x_0) \left(\sum_s P^s(x_0,x_0) \right) P^v(x_0,y) = \infty \end{aligned}$$

Because of the results about P^u, P^v from above, and statement 1.

Now we show 1 $\implies$ 3. The number of returns to x_0 from x_0 is a geometric random variable, with probability of success $\mathbb{P}_x\left[\tau_x^+ < \infty\right]$. So $G(x,x) = \infty$ iff $\mathbb{P}_x\left[\tau_x^+ < \infty\right] = 1$.

4 $\implies$ 3 obviously. We show 3 $\implies$ 4.
Assume by contradiction that $\mathbb{P}_{x_0}\left[\tau_{x_0}^+ < \infty\right] = 1$ and $\mathbb{P}_x\left[\tau_y^+ < \infty\right] < 1$.

Since x_0 is recurrent, $G(x,y) = \infty$ for all x, y (from 1,2).
Moreover, $\mathbb{P}_y\left[\tau_x < \tau_y^+\right] > 0$ since P is irreducible and a path from y to x is possible.

Then $\mathbb{P}_y\left[\tau_y^+ = \infty\right] \geq \mathbb{P}_y\left[\tau_x < \tau_y^+\right] \mathbb{P}_x[\tau_y^+ = \infty] > 0$. So $G(y,y) < \infty$.

But in showing 1 $\iff$ 3, we showed $G(y,y) = \infty$. Contradiction! $\square$

10.2.1 Examples of recurrent walks

Example: The random walk on $\mathbb{Z}$ is recurrent.

$$G(0,0) = \sum_{t=0}^{\infty} P^t(0,0) = \sum_{t=0}^{\infty} P^{2t}(0,0) = \sum_{t=0}^{\infty} 2^{-t} \binom{2t}{t} \geq c \sum_{t=0}^{\infty} \frac{1}{\sqrt{t}} = \infty$$

So it's recurrent!

Example: The corner walk on $\mathbb{Z}^2$ is recurrent.

$$P^t((0,0),(0,0)) = (P^t(0,0))^2 \geq \frac{c^2}{t} \implies G(0,0) \geq \sum \frac{c}{t} = \infty$$

The corner walk on $\mathbb{Z}^3$ is transient! $G(0,0) \approx \sum_t (1/t)^{3/2} < \infty$.

There's a fun story here about the Hungarian mathematician Pólya. He used to get drunk with all his other Hungarian mathematician friends at pubs in Budapest. He'd get into one bar, get wildly drunk, pop out and randomly enter another one on the same street. His wife's question: would he ever get home? His answer: this random walk is recurrent, so yes!

But Budapest is two-dimensional! If he doesn't stick to the same street, will he still come home? Yes, the walk is still recurrent in two dimensions!

But a drunk bird will be free forever.

11 March 9, 2020

Let's talk about the limiting behavior of countably infinite chains.

11.1 Limiting behavior

Let P be an irreducible Markov chain on a countable state space.

What can we say about $lim_{t\to\infty}\mu P^t$?

Claim: If P is transient, then $lim_{t\to\infty}\mu P^t = 0$.
Proof: If P is transient, for all $x, y \in \Omega$, $G(x,y) < \infty$.

$$\infty > G(x,y) = \sum_{t=1}^{\infty} P^t(x,y) \implies P^t(x,y) \to 0 \text{ as } t \to \infty$$

So $\delta_x P^t \to 0$ for all x.
For any μ, $\mu = \sum_x \mu(x)\delta_x$, so $\mu P^t = \sum_x \mu(x)\delta_x P^t \to 0$.

Does a nonzero limit exist if P is recurrent and irreducible? Not always, e.g the random walk on $\mathbb{Z}$.

Claim: For all x, y, $\lim_{t\to\infty} P^t(x,y) = 0$.
Proof: $P^t(x,y)$ is 0 when $|x-y| > t$. Otherwise, when $|x-y| \le t$:

$$P^t(x,y) = 2^{-t}\binom{t}{t/2 - |x-y|/2}$$

$2^{-t}\binom{t}{x} \le \frac{c}{\sqrt{t}} \to 0$ as $t \to \infty$.

11.2 Positive recurrent chains

Definition. A state $x \in \Omega$ is **positive recurrent** if $\mathbb{E}_x\left[\tau_x^+\right] < \infty$.

Note: This implies that $\mathbb{P}_x\left[\tau_x^+ < \infty\right] = 1$, i.e. x is recurrent. Enforcing a finite expected return time is a stronger condition.

If a chain is recurrent but not positive recurrent, $\mathbb{E}\left[\tau_x^+\right] = \infty$ whereas $\mathbb{P}\left[\tau_x^+ < \infty\right] = 1$. For example for $\mathbb{P}\left[\tau = n^2\right] = \frac{\pi}{6n^2}$:

$$\mathbb{P}\left[\tau < \infty\right] = \sum \frac{\pi}{6n^2} = 1 \text{ and } \mathbb{E}[\tau] = \sum \frac{\pi}{6n^2} n^2 = \infty$$

Something similar is true for the random walk on $\mathbb{Z}$.

Example: Biased random walk on $\{0, 1, 2, \dots\}$.
$P(x, x-1) = 3/4$, $P(x, x+1) = 1/4$, and $P(0, 0) = 3/4$.

Here, 0 and all the other states are positive recurrent.

11.3 Stationary distributions for countable Markov chains

Theorem.

An irreducible countable state Markov chain has a stationary distribution π iff all states are *positive-recurrent*. In this case π is the unique stationary distribution and it is given by:

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

The rest of this section is the proof of the theorem. We'll use the concept of *excursions* for the proof: any time the chain leaves x and comes back.

Let $\rho_y(x) = \mathbb{E}_x\left[\sum_{t=1}^{\tau_x^+} 1(X_t = y)\right]$. This is the expected number of times that an excursion starting at x visits y. We have $\rho_x(x) = 1$ (by definition of an excursion) and $\sum_y \rho_y(x) = \mathbb{E}_x[\tau_x^+]$.

Claim 1: For an irreducible recurrent chain, $\rho_y(x) < \infty$ for all x, y.
Proof sketch: Let $\ell_{x,y}(t) = \mathbb{P}_x\left[X_t = y, t \le \tau_x^+\right]$.
Let $f_{x,y}(t) = \mathbb{P}_x\left[X_1 \neq y, \dots, X_{t-1} \neq y, X_t = y\right]$. It is clear that:

$$f_{x,x}(s+t) \ge \ell_{x,y}(s) f_{y,x}(t)$$

The LHS means that starting at x, you go $s+t$ steps without hitting x until the last step. The ℓ term on the RHS means that the chain visited y at the s-th step from x. The f term means that starting from y, the chain went without hitting x until the last step. The RHS is one of the ways that the LHS occurs.

It is also clear that there exists t such that $f_{y,x}(t) > 0$, since the chain is irreducible. Then:

$$\rho_y(x) = \sum_{s=1}^{\infty} \ell_{x,y}(s) \leq \frac{1}{f_{y,x}(t)} \sum_{s=1}^{\infty} f_{x,x}(s+t) \leq \frac{1}{f_{y,x}(t)} < \infty$$

So Claim 1 is true.

Claim 2: The vector ρ satisfies $\rho P = \rho$.
ρ has elements that are finite and ≥ 0. But it may not be normalizable, so it's not a stationary distribution necessarily. The sum of its elements might be infinite.
Proof sketch: For $t \geq 2$, $\ell_{x,y}(t) = \sum_{z \neq x} \ell_{xz}(t-1)P(z,y)$.

$$\begin{aligned}\rho_y(x) &= P(x,y) + \sum_{z \neq x} \sum_{t \geq 2} \ell_{x,z}(t-1)P(z,y) \\ &= \rho_x(x)P(x,y) + \sum_{z \neq x} \rho_x(z)P(z,y)\end{aligned}$$

So $\rho = \rho P$.

So far we have shown that $\rho P = \rho$. If x is positive recurrent, $\sum_y \rho_y(x) = \mathbb{E}_x\left[\tau_x^+\right] < \infty$. So:

$$\frac{\rho}{\mathbb{E}_x\left[\tau_x^+\right]} \text{ is the stationary distribution}$$

We still have to show two new things. First, if P is transient, that there is no stationary distribution:
Suppose π is stationary. Then $\pi(x) = \sum_y \pi(y)P^n(y,x) \to 0$. Contradiction!

We also have to show that if π is stationary, P is positive recurrent and $\pi(x)\mathbb{E}_x[\tau_x^+] = 1$.

Let $\mu(x) = \mathbb{E}_x[\tau_x^+]$. Then $\mu(x)\pi(x) = \sum_{t=1}^{\infty} \mathbb{P}_x\left[\tau_x^+ \geq t\right] \pi(x)$. We can now use a trick! If Z is a random variable taking values 0,1,..., then: $\mathbb{E}[Z] = \sum_{t=1}^{\infty} \mathbb{P}[Z \geq t]$.

So: $\mu(x)\pi(x) = \sum_{t=1}^{\infty} \mathbb{P}_\pi\left[\tau_x^+ \geq t, X_0 = x\right]$.

We can define $a_t = \mathbb{P}_\pi\left[X_0 \neq x, \ldots, X_t \neq x\right]$. Then:

$$\begin{aligned}\mathbb{P}_\pi\left[\tau_x^+ \geq t, X_0 = x\right] &= \mathbb{P}_\pi\left[X_0 = x, X_1 \neq x, X_2 \neq x, \ldots, X_{t-1} \neq x\right] \\ &= \mathbb{P}_\pi\left[X_1 \neq x, \ldots, X_{t-1} \neq x\right] - \mathbb{P}_\pi\left[X_0 \neq x, \ldots, X_{t-1} \neq x\right] \\ &= a_{t-2} - a_{t-1}\end{aligned}$$

Then:

$$\begin{aligned}\mu(x)\pi(x) &= \mathbb{P}_\pi\left[\tau_x^+ \geq 1, X_0 = x\right] + \sum_{t=2}^{\infty}(a_{t-2} - a_{t-1}) \\ &= \pi(x) + \sum_{t=2}^{\infty}(a_{t-2} - a_{t-1}) \\ &= \pi(x) + a_0 - \lim_{t\to\infty} a_t = \pi(x) + (1 - \pi(x)) + 0 = 1\end{aligned}$$

(Since the chain is recurrent, starting from π we will surely hit x, so $\lim_{t\to\infty} a_t = 0$.)

This is what we wanted to show! □

11.3.1 Example: biased random walk

Consider a biased random walk on $\Omega = \{0, 1, 2, \dots\}$ where it goes up 1 with probability p and down with probability q. $q > p$ and $q + p = 1$. At 0, the chain loops with probability q.

Solve $\pi P = \pi$. $\pi(k) = (p/q)^k\pi(0) = (p/q)^k(1 - p/q)$.

Since there exists a stationary distribution and the chain is positive recurrent, we know π is unique.

12 March 11, 2020

Coronavirus has landed. We will be moving to Gradescope for problem set submissions.

12.1 Convergence theorem for positive recurrent chains

Convergence Theorem.

Let P be irreducible and aperiodic. If the chain is positive recurrent, then there is a unique stationary distribution π such that $\pi P = \pi$ and for all $x \in \Omega$,

$$\lim_{t\to\infty} \left\|\delta_x P^t - \pi\right\|_{TV} = 0$$

Proof: We will use coupling. The idea is to start two chains at $x_0, y_0 \in \Omega$ and show that the two chains can be coupled.

Let X_t be the Markov chain started at x_0, and Y_t the chain started at y_0.

Claim 1: The product chain (X_t, Y_t) is irreducible.

Proof sketch: We need to show we can get from $(x, y) \in \Omega^2$ to $(x', y') \in \Omega^2$ with positive probability.

Since P is irreducible and aperiodic, there is a t_x such that for all $t \geq t_x$, $P^t(x, x') > 0$. Similarly, t_y exists such that for all $t \geq t_y$, $P^t(y, y') > 0$.

Take $t = \max(t_x, t_y)$. Then $P((x, y), (x', y')) = P^t(x, x')P^t(y, y') > 0$.

Claim 2: $\pi \times \pi$ is the stationary distribution for the product chain.

Proof sketch: We show that $\pi \times \pi$ satisfies the definition of a stationary distribution:

$$\begin{aligned}(\pi \times \pi)(x, y) &= \pi(x) \cdot \pi(y) \\ &= \left(\sum_z \pi(z)P(z, x)\right)\left(\sum_w \pi(w)P(w, y)\right) \\ &= \sum_{z,w} \pi(z)\pi(w)P(z, x)P(w, y) \\ &= \sum_{z,w} (\pi \times \pi)(z, w)P((z, w), (x, y))\end{aligned}$$

Claim 3: In a positive recurrent irreducible chain, for all x, y, $\mathbb{E}_y[\tau_x^+] < \infty$.

Proof sketch: Because the chain is positive recurrent, $\mathbb{E}_x[\tau_x^+] < \infty$.

Also, $\infty > \mathbb{E}_x[\tau_x^+] \geq \mathbb{P}_x[\tau_y < \tau_x^+]\,\mathbb{E}_y[\tau_x^+]$. i.e, the probability of getting to y before x times the expected steps from y to x.

So $\mathbb{E}_y[\tau_x^+] < \infty$.

We can apply the last claim to the product chain. $T = \min(t : (X_t, Y_t) = (x_0, x_0))$. Then $\mathbb{E}_{(x_0,y_0)}[T] < \infty$.

As in the finite case, this implies:

$$\left\|\delta_{x_0}P^t - \delta_{y_0}P^t\right\|_{TV} \leq \mathbb{P}_{(x_0,y_0)}[T > t] \leq \frac{\mathbb{E}_{(x_0,y_0)}[T]}{t} \to 0 \text{ as } t \to \infty$$

And from this we get the theorem. □

12.2 Queueing example

Suppose that the lifetime of a lamp in days is a random variable X where $\mathbb{P}[X = i] = f_i$. Different lamps have independent lifetimes.

Let Z_n be the Markov chain of residual time (the time left before a lamp

has to be replaced):

$$P(0,j) = f_j \quad j \geq 1 \qquad P(i,i-1) = 1 \quad i \geq 1 \qquad \text{o.w, } P(i,j) = 0$$

If $f_i = 0$, for all $i \geq M$, for some M this is a finite Markov chain.
Otherwise, $\sup(i : f_i > 0) = \infty$ and it is a countable state space.

Q: Is this chain irreducible?
Yes! You always get to 0 and can then get to $\sup(i : f_i \geq 0)$.

Q: Is it aperiodic?
Not always. iff $\gcd(i+1 : f_i > 0) = 1$.

Q: When does Z_n have a stationary distribution?
iff 0 is positive recurrent. This is true iff

$$\mathbb{E}_0[\tau_0^+] < \infty \iff \sum f_i(i+1) < \infty$$

Q: Suppose there is a stationary distribution. What is it?
A: By the proof we did last time,

$$\begin{aligned}
\pi(i) &= \frac{\mathbb{E}_0[\text{time spent at } i \text{ in an excursion}]}{\mathbb{E}_0[\tau_0^+]} \\
&= \frac{\mathbb{E}\left[\sum_{j=1}^{\tau_0^+} 1(X_j = i)\right]}{\sum_i f_i(i+1)} \\
&= \frac{\mathbb{P}[X \geq i]}{\sum_i f_i(i+1)} \\
&= \frac{\sum_{j \geq 1} f_j}{\sum_j f_j(j+1)}
\end{aligned}$$

The denominator in the last three lines is the expected length of the excursion. This is our stationary distribution.

12.3 "Complicated Markov chain"

We'll use $\Omega = \mathbb{Z}^3 \cup \{A, B, C\}$. P is the same as the simple random walk on $\mathbb{Z}^3$ except at $x = 0$:

$$P(0, \pm e_i) = \frac{1}{8} \qquad P(0,A) = P(0,C) = \frac{1}{8}$$

Also, $P(A,B) = P(B,C) = P(C,A) = 1$.

This chain is *not* irreducible. A,B, and C are disconnected from the other states. It is periodic, because $A \to B \to C \to A$.

Q: Is there a stationary distribution?

Yes: $(1/3, 1/3, 1/3)$ on (A, B, C) and 0 elsewhere. It's like $+1$ mod 3.

Q: Does this limit depend on x: $\lim_{t\to\infty} \frac{1}{t+1} \sum_{s=0}^{t} \mathbb{P}_x[X_s = A]$?

The limit should be $\frac{1}{3} \cdot \mathbb{P}_x[\text{stuck at } A, B, C]$. This depends on x! If you're further from 0, you're less likely to get stuck.

Q: What if we removed the random walk on $\mathbb{Z}^3$ and did a random walk on $\mathbb{Z}$ with A, B, and C instead?

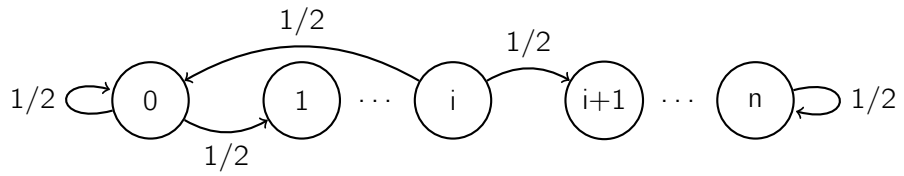
This is still not irreducible, and it's still periodic. The stationary distribution is the same. But now you *will* get to 0 so the limit is $1/3$ on $\mathbb{Z}$ and so it's independent of x. Eventually you will be at A.

12.4 Winning streak example

Toss a fair coin. Let $X_t = 0$ if the last toss is T. Otherwise:

$$X_t = \max(1 \le i \le n : X_t = X_{t-1} = \cdots = X_{t-i+1} = H)$$

Here's an incomplete graph of the chain (you get the gist of it):



It's an irreducible and aperiodic Markov chain.

$\pi(x)P(x, y) = \pi(y)P(y, x)$ does not hold. $P(i, 0) > 0$ but $P(0, i) = 0$ for $i > 1$.

Intuition suggests that $\pi(0) = 1/2$, $\pi(1) = 1/4$, and so on until: $\pi(n - 1) = 2^{-n}$ and $\pi(n) = 2^{-n}$.

(The $n - 1$ value comes from the requirement that the last $n - 1$ tosses were H and before that there was a T. The n value requires that all tosses were H.)

This is the stationary distribution!

Q: How fast is convergence to the stationary distribution?

Via coupling, we can find the answer. Let X_t start at x_0 and Y_t at y_0. We couple them by requiring that both use the same coin tosses.

With probability $1/2$, both go to 0. Otherwise, they both go up one step.

Let $T = \min(t : X_t = Y_t)$. $\mathbb{P}[T \geq t] \leq 2^{-t}$, implying that for all μ, $\left\|\mu P^t - \pi\right\|_{TV} \leq 2^{-t}$.

Similarly, $\mathbb{P}[T > n] = 0$, implying $\|\mu P^n - \pi\|_{TV} = 0$.

This is the last class for a while, because next week all classes are cancelled to allow professors time to virtualize everything, and the subsequent week is spring break. We will then resume digitally (somehow).

13 March 30, 2020

We have resumed digitally (on Zoom). Most of today's lecture was logistics for lectures going forwards and the exam on Wednesday, as well as a review of the homework due today. I've omitted all that from these notes.

Let τ be the number of times to flip a coin until we get the sequence HT. $\mathbb{E}[\tau] = ?$

We will use a Markov chain with the following states: $\{\varnothing, H, T, HT\}$. Define $x_s = \mathbb{E}_s[\tau]$. Then:

$$\begin{aligned} x_\varnothing &= 1 + 0.5(x_H + x_T) \\ x_T &= 1 + 0.5(x_H + x_T) \\ x_H &= 1 + 0.5(x_H + x_{HT}) = 1 + 0.5x_H \\ \implies x_H &= 2,\ x_\varnothing = x_T = 4 \end{aligned}$$

We can do the same thing for the toss sequence TT. We get these equations:

$$\begin{aligned} x_\varnothing &= 1 + 0.5(x_H + x_T) \\ x_T &= 1 + 0.5x_H \\ x_H &= 1 + 0.5(x_H + x_T) \\ \implies x_H &= 6,\ x_T = 4,\ x_\varnothing = 6 \end{aligned}$$

Why can't we use $\mathbb{E}_x[\tau_x^+] = \pi(x)^{-1}$ here?

For one, we are starting from the empty state which is not always equivalent to going from X_0 to X_0. If you're going from $TT \to TT$, the last T helps you, so it's not the same as the empty state. In the HT case, however, it is the same. Also, the chain isn't irreducible or aperiodic, since we can't get to $\varnothing$ from elsewhere.

14 April 6, 2020

Today we're going to talk about conditional expectations.

14.1 Conditional expectations of discrete random variables

Definition. The **conditional probability** that $Y = y$ given $X = x$ is:

$$\mathbb{P}[Y = y \mid X = x] = \frac{\mathbb{P}[Y = y, X = x]}{\mathbb{P}[X = x]}$$

So we can write: $\mathbb{P}[Y = y] = \sum_x \mathbb{P}[X = x]\,\mathbb{P}[Y = y \mid X = x]$.

Claim: We can use this to talk about expectation values:

$$\mathbb{E}[Y \mid X = x] = \frac{\mathbb{E}[Y \cdot 1(X = x)]}{\mathbb{P}[X = x]} = \sum_y y\,\mathbb{P}[Y = y \mid X = x]$$

Proof: The second term above can be written as:

$$\sum_y \frac{y\,\mathbb{P}[Y = y, X = x]}{\mathbb{P}[X = x]} = \sum_y y\,\mathbb{P}[Y = y \mid X = x]$$

Claim: $\mathbb{E}[Y] = \sum_x \mathbb{E}[Y \mid X = x]\,\mathbb{P}[X = x]$.

Proof: We start from the RHS.

$$\begin{aligned}
\sum_x \mathbb{E}[Y \mid X = x]\,\mathbb{P}[X = x] &= \sum_x \sum_y y\,\mathbb{P}[Y = y \mid X = x] \cdot \mathbb{P}[X = x] \\
&= \sum_x \sum_y y\,\mathbb{P}[Y = y, X = x] \\
&= \sum_y y\,\mathbb{P}[Y = y] \\
&= \mathbb{E}[Y]
\end{aligned}$$

Since $\mathbb{E}[Y \mid X = x]$ is an expectation, it satisfies all the usual properties, including linearity: $\mathbb{E}[aY + Z \mid X = x] = a\mathbb{E}[Y \mid X = x] + \mathbb{E}[Z \mid X = x]$

Definition. The **conditional expectation** of Y given X, denoted $\mathbb{E}[Y \mid X]$, is the function of X whose value is $\mathbb{E}[Y \mid X = x]$ if $X = x$.

Note: $\mathbb{E}[Y \mid X]$ is a random variable. Also this is only for discrete RVs.

Claim: $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$.
Proof: The LHS is $\sum_x \mathbb{P}[X = x]\mathbb{E}[Y \mid X = x]$, which is $\mathbb{E}[Y]$ from earlier.

Claim: $\mathbb{E}[Y + Z \mid X] = \mathbb{E}[Y \mid X] + \mathbb{E}[Z \mid X]$.
Proof: For every x, $\mathbb{E}[Y + Z \mid X = x] = \mathbb{E}[Y \mid X = x] + \mathbb{E}[Z \mid X = x]$. So the result follows.

Claim: If Z is independent of Y and X, then $\mathbb{E}[YZ \mid X] = \mathbb{E}[Z]\mathbb{E}[Y \mid X]$.
Proof: The LHS gives us:

$$\frac{\mathbb{E}[YZ \cdot 1(X = x)]}{\mathbb{P}[X = x]} = \mathbb{E}[Z]\mathbb{E}[Y \mid X = x]$$

since Z is independent of X, Y. The result follows.

Claim: If Z is a deterministic function of X, then the conditional expectation $\mathbb{E}[YZ \mid X] = Z\,\mathbb{E}[Y \mid X]$.
Proof: This holds because at a specific x, Z takes a constant value and can pop out of the expectation value. For example, let $X = Y = Z$. Then we have $\mathbb{E}[X^2 \mid X] = X\mathbb{E}[X \mid X] = X^2$.

Some examples:

Coin tosses. Toss a coin n times with success probability p. Let S_k be the number of successes in the first $k < n$ tries. $\mathbb{E}[S_k \mid S_n] =$?

$$\mathbb{E}[S_k \mid S_n] = \mathbb{E}\left[\sum_{i=1}^{k} X_i \middle| S_n\right] = \sum_{i=1}^{k} \mathbb{E}[X_i \mid S_n] = k \cdot \mathbb{E}[X_i \mid S_n] = \frac{k}{n} S_n$$

Dice paradox. If you roll an odd number, you lose and get $Y = 0$ dollars. Otherwise, you win $Y =$ the number of rolls until you get 6.

Roll 2, 4, 1 and you get \$0. Roll 2, 4, 4, 6 and you get \$4. Let X be the indicator that you didn't lose. What is $\mathbb{E}[Y \mid X]$?

$\mathbb{E}[Y \mid X]$ takes two values, since either $X = 0$ or $X = 1$. $\mathbb{E}[Y \mid X = 0] = 0$. What about when $X = 1$?

You might say that if $X = 1$, the only things you've rolled are 2, 4, and 6, following Geom($\frac{1}{3}$). So $\mathbb{E}[Y \mid X = 1] = 3$? No!

Let's do it out formally:

$$\mathbb{P}[X = 1] = \frac{1}{6}\sum_{k=0}^{\infty}\left(\frac{1}{3}\right)^k = \frac{1}{6}\frac{1}{1 - \frac{1}{3}} = \frac{1}{4}$$

$$\mathbb{E}[Y \cdot 1(X = 1)] = \frac{1}{6}\sum_{k=0}^{\infty}\left(\frac{1}{3}\right)^k (k + 1) = \frac{1}{6}\frac{9}{4} = \frac{3}{8}$$

$$\mathbb{E}[Y \mid X = 1] = \frac{3}{8}\left(\frac{1}{4}\right)^{-1} = \frac{3}{2}$$

So then we have $\mathbb{E}[Y \mid X] = \frac{3}{2}X$.

14.2 Conditional expectations of continuous random variables

(X, Y) are continuous with joint density f.

Definition. The **conditional density** of X and Y is

$$f_Y(y \mid X = x) = \frac{f(x, y)}{f_X(x)}$$

Some results that follow:

1. $\mathbb{E}[g(Y) \mid X = x] = \int g(y) f_Y(y \mid X = x) dx$.
2. $f_Y(y) = \int f_Y(y \mid X = x) f_X(x) dx$.
   c.f. the discrete result, $\mathbb{P}[Y = y] = \sum_x \mathbb{P}[Y = y \mid X = x]\mathbb{P}[X = x]$.
3. $\mathbb{E}[Y] = \int \mathbb{E}[Y \mid X = x] f_X(x) dx$.

Example: Suppose X, Y are drawn uniformly from the triangle $\{(x, y) : x \geq 0, y \geq 0, x + y \leq 2\}$. What is $\mathbb{E}[X \mid Y]$?

We can see that $\mathbb{E}[X \mid Y = 2] = 0$, and $\mathbb{E}[X \mid Y = 0] = 1$ (draw out the triangle). Then, if $Y = y$, $X \sim U[0, 2 - y]$. So $\mathbb{E}[X] = 1 - y/2$. Then $\mathbb{E}[X \mid Y] = 1 - Y/2$.

Definition. **Jensen's inequality** states that if ϕ is convex, then

$$\mathbb{E}[\phi(X)] \geq \phi(\mathbb{E}[X])$$

A famous case of this is that $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$. It also applies to conditional probabilities!

$$\mathbb{E}[\phi(Y) \mid X] \geq \phi\left(\mathbb{E}[Y \mid X]\right)$$

So $\mathbb{E}[Y^2 \mid X] \geq (\mathbb{E}[Y \mid X])^2$.

Note: A word on notation. We often condition on $X = (X_1, \ldots, X_n)$. We will write that as $X_{[n]} \equiv (X_1, \ldots, X_n)$.

Next time, martingales!

15 April 8, 2020

Today we'll start martingales.

15.1 Martingales

Definition. Consider two sequences of random variables:

1. $(M_t)_{t\geq 1}$ ("process we care about")
2. $(X_t)_{t\geq 1}$ ("information")

We say that $(M_t)_{t\geq 1}$ is a **martingale** with respect to $(X_t)_{t\geq 1}$ if

1. M_t is a function of $X_{[t]}$ and
2. $\mathbb{E}\left[M_{t+1} \mid X_{[t]}\right] = M_t \quad \forall\, t$

The condition in the equation says that for any possible values x_s of X_s for $s \leq t$,

$$\mathbb{E}\left[M_{t+1} \mid X_1 = x_1, \ldots, X_t = x_t\right] = M_t$$

The intuition here is that the current value is an estimate of the future. Assuming an efficient market, M_t is the value of a stock on day t where X_t is the stock market at day t.

We call M_t a **supermartingale** if instead of equality, we have

$$\mathbb{E}\left[M_{t+1} \mid X_{[t]}\right] \leq M_t \quad \forall\, t$$

or a **submartingale** if:

$$\mathbb{E}\left[M_{t+1} \mid X_{[t]}\right] \geq M_t \quad \forall\, t$$

Claim: If M_t is a supermartingale then $\mathbb{E}[M_{t+1}] \leq \mathbb{E}[M_t]$.
Proof: Take expectations on both sides of the supermartingale definition.

$$\mathbb{E}\left[\mathbb{E}[M_{t+1} \mid X_{[t]}]\right] \leq \mathbb{E}[M_t]$$
$$\mathbb{E}[M_{t+1}] \leq \mathbb{E}[M_t]$$

So if M_t is a submartingale, $\mathbb{E}[M_{t+1}] \geq \mathbb{E}[M_t]$.
If it is a martingale, $\mathbb{E}[M_{t+1}] = \mathbb{E}[M_t]$.

Claim: Let (X_n) be a finite Markov chain with transition matrix P and let $f(x, n)$ be a function of the state x and time n such that:

$$f(x, n) = \sum_y P(x, y) f(y, n+1)$$

Then $M_n = f(X_n, n)$ is a martingale with respect to (X_n).

In this situation, the information is the Markov chain. We have a function f which is sort of like a harmonic function. If we apply f to the n-th state of the Markov chain, that process is a martingale.

Proof: We'll start with the definition of a martingale:

$$\begin{aligned}\mathbb{E}\left[M_{n+1} \mid X_{[n]}\right] &= \mathbb{E}\left[f(X_{n+1}, n+1) \mid X_{[n]}\right] \\ &= \mathbb{E}\left[f(X_{n+1}, n+1) \mid X_n\right] \\ &= \textstyle\sum_y P(X_n, y)\, f(y, n+1) \\ &= f(X_n, n) \\ &= M_n\end{aligned}$$

Note: If h is harmonic, $h(X_n)$ is a martingale: $h(x) = \sum_y P(x,y)h(y)$.

15.1.1 Some examples

Claim: Let $S_n = \sum X_i$ be gambler's ruin where $\mathbb{P}[X_i = 1] = p$ and $\mathbb{P}[X_i = -1] = 1 - p$. Then $M_n = ((1-p)/p)^{S_n}$ is a martingale with regard to X_n.

Proof: We need to check that $h(x) = \left(\frac{1-p}{p}\right)^x$ is harmonic for gambler's ruin. Verify:

$$\begin{aligned}h(x) &= p \cdot h(x+1) + (1-p) \cdot h(x-1) \\ \left(\frac{1-p}{p}\right)^x &= p \cdot \left(\frac{1-p}{p}\right)^{x+1} + (1-p) \cdot @(\frac{1-p}{p})^{x-1} \\ &= (1-p)\left(\frac{1-p}{p}\right)^x + p\left(\frac{1-p}{p}\right)^x \\ &= \left(\frac{1-p}{p}\right)^x\end{aligned}$$

Simple random walk. Let $\mathbb{P}[X_i = 1] = \mathbb{P}[X_i = -1] = 1/2$.
The sequence $M_t = 1$ is a martingale. We can see this by substituting $p = \frac{1}{2}$ into the gambler's ruin result.

Claim: If $p = 1/2$ in gambler's ruin then $M_n = \sum_{i=1}^n X_i$ is a martingale with respect to (X_n).
Proof: $h(x) = x$ is harmonic in this case. $x = \frac{1}{2}(x+1) + \frac{1}{2}(x-1)$.

Claim: If $p = 1/2$ in gambler's ruin then $M_n^2 - n$ is a martingale with respect to (X_n).
Proof: $f(x,n) = x^2 - n$. This is equal to $\frac{1}{2}f(x+1, n+1) + \frac{1}{2}f(x-1, n+1)$. So $f(M_n, n) = M_n^2 - n$ is a martingale.

A simple stock market model. Let $X_i \geq 0$ be independent and positive with $\mathbb{E}[X_i] = 1$. Let $M_n = \Pi_{i=1}^n X_i$. Then M_n is a martingale:

$$\begin{aligned}\mathbb{E}\left[M_{n+1} \mid X_{[n]}\right] &= \mathbb{E}\left[X_{n+1}\Pi_{i=1}^n X_i \mid X_{[n]}\right] \\ &= \Pi_{i=1}^n X_i \mathbb{E}\left[X_{n+1} \mid X_{[n]}\right] \\ &= \Pi_{i=1}^n X_i \mathbb{E}[X_{n+1}] \\ &= \Pi_{i=1}^n X_i = M_n\end{aligned}$$

15.2 Properties of martingales

Claim: If M_n is a (super/sub-)martingale, then for any R.V. Y,

$$\mathbb{E}[Y M_n \mid X_{[n]}] = M_n \mathbb{E}\left[Y \mid X_{[n]}\right]$$

This is true because if we have $X_{[n]}$, then we know M_n.

Claim: If M_n is a martingale and ϕ is a convex function, then $\phi(M_n)$ is a submartingale.
Proof: This is directly from Jensen's inequality:

$$\mathbb{E}\left[\phi(M_{n+1}) \mid X_{[n]}\right] \geq \phi\left(\mathbb{E}[M_{n+1} \mid X_{[n]}]\right) = \phi(M_n)$$

If M_n is a martingale, then M_n^2 is a submartingale.

Claim: If M_n is a martingale,

$$\mathbb{E}[M_{n+1}^2 \mid X_{[n]}] - M_n^2 = \mathbb{E}\left[(M_{n+1} - M_n)^2 \mid X_{[n]}\right]$$

Proof: Start with the RHS:

$$\begin{aligned}&\mathbb{E}[M_{n+1}^2 \mid X_{[n]}] - 2\mathbb{E}[M_{n+1}M_n \mid X_{[n]}] + \mathbb{E}[M_n^2 \mid X_{[n]}] \\ =&\mathbb{E}[M_{n+1}^2 \mid X_{[n]}] - 2M_n \mathbb{E}[M_{n+1} \mid X_{[n]}] + M_n^2 \\ =&\mathbb{E}[M_{n+1}^2 \mid X_{[n]}] - M_n^2\end{aligned}$$

Claim: If M_n is a martingale and $i \leq j \leq k \leq n$ then:

$$\mathbb{E}[M_j(M_n - M_k)] = 0 \quad \text{and} \quad \mathbb{E}[(M_n - M_k)(M_j - M_i)] = 0$$

Proof: We'll prove the first statement because the second one follows. Let's consider a particular expectation:

$$\mathbb{E}[M_j(M_n - M_k) \mid X_{[j]}] = M_j \mathbb{E}[M_n - M_k \mid X_{[j]}]$$

Since $\mathbb{E}[M_{n+1} \mid X_{[n]}] = M_n$, we know $\mathbb{E}[M_n \mid X_{[j]}] = \mathbb{E}[M_k \mid X_{[j]}] = M_j$.

So $\mathbb{E}[M_j(M_n - M_k) \mid X_{[j]}] = 0$. We can take the expectation value of both sides, and we get $\mathbb{E}[M_j(M_n - M_k)] = 0$.

Martingales also obey **orthogonality of increments**:

$$\mathbb{E}[(M_n - M_m)^2] = \sum_{k=m+1}^{n} \mathbb{E}[(M_k - M_{k-1})^2]$$

This comes from expanding out the LHS:

$$\mathbb{E}\left[((M_n - M_{n-1}) + (M_{n-1} - M_{n-2}) + \cdots + (M_{m+1} - M_m))^2\right]$$

Recall if M_n is a martingale then $\mathbb{E}[M_n] = \mathbb{E}[M_m]$ if $n \geq m$. This is often summarized as "you cannot beat an unfavorable game".

Definition. Let M_n be a martingale with respect to X_n. We call H_n a **predictable process** if for each n, H_n is a function of $X_1, \ldots, X_{n-1}$.

Theorem.

If M_n is a martingale with respect to X, H_n is predictable and satisfies $|H_n| \leq c_n$ for all n, then:

$$W_n = \sum_{m=1}^{n} H_m (M_m - M_{m-1}) \text{ is a martingale.}$$

16 April 13, 2020

We begin with a review of the theorem from last time.

16.1 Predictable processes

We'll start with a slightly modified version of the theorem:

Theorem.

If M_n is a supermartingale with respect to X, H_n is predictable and satisfies $0 \leq H_n \leq c_n$ for all n, then:

$$W_n = \sum_{m=1}^{n} H_m (M_m - M_{m-1}) \text{ is a supermartingale.}$$

What is the theorem saying? We can think of $(M_m - M_{m-1})$ as the difference between things today and yesterday, and since it's a supermartingale we know there's a downward trend here. If we use a stock market analogy, H_m is a positive "investment", and the "return" is given by $H_m(M_m - M_{m-1})$. The theorem says that the return is also a supermartingale.

Proof: We want to show that $\mathbb{E}[W_{n+1} \mid X_{[n]}] \leq W_n$. Start with the LHS:

$$\begin{aligned}\mathbb{E}[W_{n+1} \mid X_{[n]}] &= \mathbb{E}\left[\sum_{k=1}^{n+1} H_k(M_k - M_{k-1})\middle| X_{[n]}\right] \\ &= \sum_{k=1}^{n} H_k(M_k - M_{k-1}) + \mathbb{E}\left[H_{n+1}(M_{n+1} - M_n) \mid X_{[n]}\right] \\ &= W_n + H_{n+1}\left(\mathbb{E}\left[M_{n+1} - M_n \mid X_{[n]}\right]\right) \leq W_n\end{aligned}$$

What are we doing here? First we take the sum out of the expectation. H_k and M_k are functions of X, so those come out as well except for the final values. In the last step, $H_{n+1} \geq 0$ because we have required it, and $M_{n+1} - M_n \leq 0$ because it's a supermartingale. So we have our result.

We can rephrase this for martingales and submartingales. The only difference is that for martingales (as we saw last time), we require $|H_n| \leq c_n$ (we can think of this as being allowed to "short" the markets).

16.2 Stopping times for martingales

Definition. T is a **stopping time** if the event $\{T = k\}$ can be determined from $X_1, \ldots, X_k$.

Corollary: If M is a (super-)martingale and T is a stopping time, then $M_{min(T,n)}$ is a (super-)martingale.

So, if we define $N_n \equiv M_{min(T,n)}$, $N_1 = M_1$ and $N_2 = M_{min(T,2)}$ and so on.

Proof: Write $M_{min(T,n)} = M_1 + \sum_{k=2}^{n}(M_k - M_{k-1})1(T \geq k)$. This is a pretty straightforward way of rewriting this, as a sort of telescoping sum with an indicator that enforces that n doesn't go past T.
In order to make use of the theorem we've been looking at, we want $1(T \geq k)$ to be a predictable process. If this is the case, it must be a function of $X_1, \ldots, X_{k-1}$. $1(T \geq k)$ is the complement of $1(T < k)$, which is a function of $X_1, \ldots, X_{k-1}$ by the definition of T. So it is a predictable process, and we can directly apply the theorem.

Question: Suppose (M_n) is a martingale and T is a stopping time. Is $\mathbb{E}[M_T] = \mathbb{E}[M_0]$?

We know that $\mathbb{E}[M_{min(T,n)}] = \mathbb{E}[M_0]$ for all n. But if your stopping rule is that you stop once you've won something, $\mathbb{E}[M_T] = \mathbb{E}[M_0]$ can't be true!

Counter-example: Let $M_n = \sum_{i=1}^{n} X_i$, where X is a simple random walk, and $T = \min(n : M_n = 1)$.

Then $\mathbb{E}[M_n] = 0$, and so $\mathbb{E}[M_{min(T,n)}] = 0$ as well. But $\mathbb{E}[M_t] = 1$, since that's the condition for being at T. How is this possible? Because $\mathbb{P}[T < \infty] = 1$ and $\mathbb{E}[T] = \infty$. So we will surely get to the stopping time but it won't happen for a very very long time.

Theorem.

Suppose M_n is a martingale and T is a stopping time with $\mathbb{P}[T < \infty] = 1$ and $\left|M_{min(T,t)}\right| \leq k$ for all t. Then, $\mathbb{E}[M_T] = \mathbb{E}[M_0]$.

For the simple random walk example, $\mathbb{P}[T < \infty] = 1$ but $\left|M_{min(T,t)}\right|$ can be arbitrarily large. Our conditions for this theorem to hold are that 1) $\mathbb{P}[T < \infty] = 1$, and 2) there exists k such that $\left|M_{min(T,t)}\right| \leq k$ for all t.

Proof: We know $\mathbb{E}[M_{min(T,t)}] = \mathbb{E}[M_0]$ for all t. We want to bound the following expression:

$$\begin{aligned}\left|\mathbb{E}[M_T] - \mathbb{E}[M_{min(T,t)}]\right| &\leq \mathbb{E}\left[\left|M_T - M_{min(T,t)}\right|\right] \\ &= \mathbb{E}[1(T > t)|M_T - M_t|] \\ &\leq 2k\mathbb{P}[T > t]\end{aligned}$$

We get the $2k$ since we've specified bounds on $M_{min(T,t)}$. As $t \to \infty$, this expression goes to 0. What this says is:

$$\mathbb{E}[M_t] = \lim_{t\to\infty} \mathbb{E}[M_{min(T,t)}] = \mathbb{E}[M_0]$$

16.2.1 Applications of stopping times

Unfair gambler's ruin. Consider a version of gambler's ruin with $p(x \to x+1) > 1/2$. Let X_n be the gambler's ruin states.

We showed that $M_n = \theta^{X_n}$ where $\theta = (1-p)/p$ is a martingale.
Let $T = \min(n : M_n \in \{0, 1\})$ be the stopping time. Say the game starts at some c where $a < c < b$. Then $\mathbb{E}[M_T] = \theta^c$, because $M_0 = \theta^c$ (applying the theorem).

Then: $\mathbb{E}[M_T] = p_a\theta^a + (1-p_b)\theta^b$ where p_a is the probability of stopping

at a, similarly for p_b. Setting this equal to θ^c, we get that

$$p_a = \frac{\theta^b - \theta^c}{\theta^b - \theta^a}$$

so there's a sort of geometric behavior here.

Fair gambler's ruin. We'll use $p = 1/2$ here, and say that the game starts at 0, with bounds $a < 0 < b$. Let $M_n = X_n$ and $T = \min(n : X_n \in \{a, b\})$. Then:

$$0 = \mathbb{E}[M_T] = p_a(a) + (1 - p_a)b \implies p_a = \frac{-a}{b - a}$$

What if we use $M_n = X_n^2 - n$ and the same T?

$$0 = \mathbb{E}[M_T] = p_a \cdot a^2 + (1 - p_a)b^2 - \mathbb{E}[T] \implies \mathbb{E}[T] = -ab$$

Can we apply the theorem in this case? Actually no, because $\left|M_{min(T,t)}\right|$ isn't bounded. If n gets very large, there is no longer a bound. So interestingly, the result holds but not because of the theorem. We can work through a proof similar, but not identical, to what we used for the theorem to get this result rigorously.

16.3 Wald's Equation

Theorem.

Suppose $S_n = X_1 + X_2 + \dots$ is a sum of i.i.d random variables with mean μ. Let T be a stopping time with $\mathbb{E}[T] < \infty$. Then $\mathbb{E}[S_T] = \mu\,\mathbb{E}[T]$.

We won't prove this but let's go over what it means. X_i are i.i.d random variables with $\mathbb{E}[X_i] = \mu$. We're interested in $\mathbb{E}\left[\sum_{n=1}^{T} X_n\right]$.

T is set by some rule because it's a stopping time, so the theorem is saying that we can think of this as:

$$\mathbb{E}\left[\sum_{n=1}^{T} X_n\right] = \sum_{n=1}^{\mathbb{E}[T]} \mathbb{E}[X_{[n]}] = \mathbb{E}[T] \cdot \mu$$

which is pretty neat!

Example: Let $X_i = \pm 1$ with probability $1/2$. $T = \min\left(n : \sum_{i=1}^{n} X_n = 1\right)$. $\mathbb{E}[X_i] = 0$ but $\sum_{n=1}^{T} X_n = 1$. What's going on? The problem is that $\mathbb{E}[T] < \infty$ is not true, so we can't use the theorem here.

16.3.1 Coin tosses example

Let $c = c_1 c_2 \dots c_k$ be a pattern of coin tosses. Let T be the first time we observe the pattern. What is $\mathbb{E}[T]$?

Suppose $c = HT$. If we ended up tossing $TTTTTHHT$, we say that the time at which we saw the sequence is 8.

Consider a game where at each round, a new player enters and bets \$1. This player will lose or gain \$1 depending on whether the next toss is c_1. If they lose, they leave. If they win, they bet \$2 that the next toss is c_2. The game stops when the pattern is observed.

This is a fair game, so we can take $M_t = t - p_t$ as a martingale, where p_t is the total payment to all players at the end of round t, and t accounts for all the \$1 bets.

Let T be the stopping time. $\mathbb{E}[M_{\min(t,T)}] = 0$ for all t. So $\mathbb{E}[p_{\min(t,T)}] = \mathbb{E}[\min(t, T)]$ and both sides converge in the limit of large t. Using the theorem, we have that $\mathbb{E}[T] = p_T$. (Note: p_T is not a random variable. It's highly dependent on the sequence.)

What does this mean in practice? Say that the sequence we're looking for HH. The sequence of tosses is $\dots HH$. Up to the first H, nobody gets money because they all lost up until then. The player who came in on the first H wins \$4 (bet \$1, win \$2, bet those \$2, win \$4), and the player who came in on the second H wins \$2. So p_T and thus $\mathbb{E}[T]$ is 6.

If the sequence we're looking for is HT, the player from the H gets \$4 and the player on the T gets \$0. So p_T and thus $\mathbb{E}[T]$ is 4.

17 April 15, 2020

We're going to talk about the convergence of martingales today.

17.1 Convergence Theorem

Theorem. (Convergence Theorem)

Let $M_n \geq 0$ be a supermartingale. Then $M_\infty = \lim_{n\to\infty} M_n$ exists almost surely and $\mathbb{E}[M_\infty] \leq \mathbb{E}[M_0]$.

What this is saying is that $\mathbb{P}[\lim_{n\to\infty} M_n = M_\infty] = 1$. Remember the

requirement for a supermartingale is $\mathbb{E}[M_{n+1} \mid X_{[n]}] \leq M_n$.
Note: even for martingales, we don't necessarily have $\mathbb{E}[M_\infty] = \mathbb{E}[M_0]$. For example, if $M_t = 1 + \sum_{i=1}^{t} X_i$ and $\mathbb{P}[X_i = \pm 1] = 1/2$, with a stopping time $T = \min(t : M_t = 0)$:

$$\mathbb{E}[M_{\min(t,T)}] = 1 \quad \text{but} \quad \mathbb{E}[M_\infty] = \mathbb{E}[M_T] = 0$$

If $a \leq M_n \leq b$ for all n, and M_n is a martingale, then $\mathbb{E}[M_\infty] = \mathbb{E}[M_0]$.

We won't go through the entire proof, but there's an important lemma:

Lemma. (Maximal Inequality)

If $X_n \geq 0$ is a supermartingale and $X > 0$, then:

$$\mathbb{P}[\sup_{t \geq 0} X_t > x] \leq \frac{\mathbb{E}[X_0]}{x}$$

This is a stronger statement than the Markov inequality, which says that if $X \geq 0$, $\mathbb{P}[X \geq a] \leq \mathbb{E}[X]/a$. The lemma is a special case of this, saying that it holds for the maximum of the process over all t.

Lemma Proof: Let T be the stopping time of exceeding x, and use $\mathbb{E}[X_{\min(t,T)}] \leq \mathbb{E}[X_0]$.

$$\begin{aligned} \mathbb{P}[\sup X_t > x] &= \mathbb{P}[X_{\min(t,T)} \geq x] && \text{because of stopping time} \\ &\leq \frac{\mathbb{E}[X_0]}{x} && \text{from Markov inequality} \end{aligned}$$

The theorem is proved by showing that if M_n doesn't converge, there exists a, b where M_n crosses both infinitely often. The proof is a pain so we won't do it.

17.2 Polya's urn

At time 0, there's one white ball in the urn and one black ball.

At time t, draw a ball uniformly at random, put it back and add another of the same color. So there are $t + 2$ balls at that time.

Define M_t as the fraction of black balls in the urn, $M_t \in [0, 1)$. The rules are as follows:

$$\begin{aligned} M_{t+1} &= \frac{(t+2)M_t}{t+3} && \text{with probability } 1 - M_t \\ M_{t+1} &= \frac{(t+2)M_t + 1}{t+3} && \text{with probability } M_t \end{aligned}$$

M_t is a non-homogeneous Markov chain because the update rule depends on t.

Claim: M_t is a martingale (for all starting configurations).

Proof: Because M_t is a Markov chain, $\mathbb{E}[M_{t+1} \mid M_{[t]}] = \mathbb{E}[M_{t+1} \mid M_t]$. So:

$$\begin{aligned}\mathbb{E}[M_{t+1} \mid M_{[t]}] &= \mathbb{E}[M_{t+1} \mid M_t] \\ &= M_t \cdot \frac{(t+2)M_t + 1}{t+3} + (1 - M_t)\frac{(t+2)M_t}{t+3} \\ &= M_t \cdot \frac{1}{t+3} + \frac{(t+2)M_t}{t+3} = M_t\end{aligned}$$

Since $0 \leq M_t \leq 1$ and it's a martingale, the convergence theorem applies and $\mathbb{P}[\lim M_t \to M_\infty] = 1$.

Claim: In the Polya Urn, given that you started with 1 ball of each type, $\lim M_t = M_\infty$ is uniform in $[0, 1]$.

Proof: You can show by induction that $\mathbb{P}[M_t = j/(t+2)] = 1/(t+1)$ for $j = 1, \ldots, t+1$. Every fraction is equally likely.

$$\begin{aligned}\mathbb{P}\left[M_{t+1} = \frac{j}{t+3}\right] &= \mathbb{P}\left[M_t = \frac{j-1}{t+2}\right] \cdot M_t + \mathbb{P}\left[M_t = \frac{j}{t+2}\right] \cdot (1 - M_t) \\ &= \frac{1}{t+1} \cdot M_t + \frac{1}{t+1}(1 - M_t') \\ &= \frac{1}{t+1}\left(\frac{j-1}{t+2} + 1 - \frac{j}{t+2}\right) \\ &= \frac{1}{t+2}\end{aligned}$$

So! $\mathbb{E}[M_\infty] = \mathbb{E}[U[0,1)] = \frac{1}{2} = \mathbb{E}[M_0]$.

17.3 Branching processes

Francis Galton was a British statistician who wanted to model the extinction of family names (particularly his own). He did this by modeling the number of daughters likely to exist in each generation of a family. This is the Galton-Watson process.

Let Y be integer-valued random variables representing the number of daughters. We will make independent copies of Y and label these Y_i^j. This is the number of daughters of individual j in generation i.

Let Z_i be the number of women in generation i. Let Z_0 be some x. Then $Z_{n+1} = \sum_{i=1}^{Z_n} Y_{n+1}^i$.

We want to model the number of daughters in each generation.

Claim: Let $\mathbb{E}[Y] = \mu$ (the expected number of daughters of each individual). Then Z_n/μ^n is a martingale.

Proof: We'll do this in the usual way:

$$\begin{aligned}\mathbb{E}\left[\frac{Z_{n+1}}{\mu^{n+1}}\,\middle|\, Z_{[n]}\right] &= \frac{1}{\mu^{n+1}}\mathbb{E}\left[\sum_{i=1}^{Z_n} Y_{n+1}^i \,\middle|\, Z_{[n]}\right] \\ &= \frac{1}{\mu^{n+1}} Z_n \cdot \mu \\ &= \frac{Z_n}{\mu^n}\end{aligned}$$

Claim: If $\mu < 1$, $\mathbb{P}[Z_n > 0] \to 0$ as $n \to \infty$.

Proof: $\mathbb{P}[Z_n > 0] \leq \mathbb{E}[Z_n]$ by the Markov inequality.

Claim: If $\mu = 1$ and $\mathbb{P}[Y = 1] < 1$, then $\mathbb{P}[Z_n > 0] \to 0$.

Proof: This follows from the limit theorem. $\mu = 1$ and Z_n is a martingale with $Z_n \geq 0$. We can use the convergence theorem, so $Z_n \to Z_\infty$ and $\mathbb{E}[Z_\infty] \leq 1$. Z_∞ cannot be 1, 2, 3, etc. because we cannot promise that the number of women is stable at 2 since $\mathbb{P}[Y = 1] < 1$. This means $Z_\infty = 0$.

18 April 22, 2020

Today we're going to talk about Poisson processes and continuous time Markov chains. But first, we need to talk about some new types of random variables.

18.1 Exponential random variables

Definition. A positive random variable T has **exponential** distribution $\text{Exp}(\lambda)$ if $P(T > t) = e^{-\lambda t}$ for $t \geq 0$.

So the density or pdf, $f_T(t)$, is 0 if $t \leq 0$, and $\lambda e^{-\lambda t}$ if $t \geq 0$. The *moments* of the distribution: $\mathbb{E}[T] = 1/\lambda, \text{Var}[T] = 1/\lambda^2$.

Claim: (Sums) Let T_n be a sum of n i.i.d $\text{Exp}(\lambda)$ random variables.

$$X_i \sim \text{Exp}(\lambda),\ T_n = \sum_{i=1}^{n} X_i \implies f_{T_n}(t) = \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}$$

This is the density of T_n.

Proof: Standard practice for these problems is *convolution* of random variables. How do you get value t out of the sum of random variables?

One variable takes value s, and the rest of them take $t-s$. We can do this via induction. So:

$$\begin{aligned} f_{T_{n+1}}(t) &= \int_0^t f_{T_n}(s) \cdot \lambda e^{-\lambda(t-s)} ds \\ &= \int_0^t \frac{(\lambda s)^{n-1}}{(n-1)!} \cdot \lambda^2 \cdot e^{-\lambda(t-s)} e^{-\lambda s} ds \\ &= \lambda^2 e^{-\lambda t} \int_0^t \frac{(\lambda s)^{n-1}}{(n-1)!} ds \\ &= \lambda^2 e^{-\lambda t} \cdot \frac{(\lambda s)^n}{n!} \frac{1}{\lambda} \Big|_0^t \\ &= \lambda e^{-\lambda t} \frac{(\lambda t)^n}{n!} \end{aligned}$$

Lack of memory property: $\mathbb{P}[T > s+t \mid T > s] = e^{-\lambda t}$ for $T \sim \text{Exp}(\lambda)$, with $s, t > 0$.

What this is saying is that it's like starting the process from the beginning. Shifting time by s doesn't really do anything.

Proof: We can use the definition of conditional probability:

$$\mathbb{P}[T > s+t \mid T > s] = \frac{\mathbb{P}[T > s+t]}{\mathbb{P}[T > s]} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t}$$

18.1.1 Exponential races

Let $S \sim \text{Exp}(\lambda)$ and $T \sim \text{Exp}(\mu)$ be independent.

Maybe S is the chance you die from COVID and T is the chance you die in a car accident. Or, less pessimistically, S is the chance your next WFH snack is salty and T is the chance your next WFH snack is sweet. Which happens first?

Claim: $\min(T, S) \sim \text{Exp}(\lambda + \mu)$.

Proof: Since S, T are independent,

$$\mathbb{P}[\min(T,S) > t] = \mathbb{P}[T > t,\ S > t] = \mathbb{P}[T > t]\mathbb{P}[S > t]$$

Plugging in, $e^{-\mu t} e^{-\lambda t} = e^{-(\lambda+\mu)t}$.

Claim: $P[S < T] = \frac{\lambda}{\lambda+\mu}$. (Does S win?)

Proof: This is similar to the convolution.

$$\begin{aligned} \mathbb{P}[S > T] &= \int_0^\infty \lambda e^{-\lambda s} \mathbb{P}[T > s]\, ds \qquad (f_s \times \text{prob } T \text{ happens later}) \\ &= \int_0^\infty \lambda e^{-\lambda s} e^{-\mu s} ds = \frac{\lambda}{\lambda+\mu} \end{aligned}$$

Claim: Let I be the random variable of which variable is smaller. (I takes values "T" and "S", depending on which one wins.) Then I and $\min(T, S)$ are independent.

$\mathbb{P}[I = \text{"}S\text{"}] = \frac{\lambda}{\lambda+\mu}, \mathbb{P}[I = \text{"}T\text{"}] = \frac{\mu}{\lambda+\mu}$. I is discrete (what do you eat?). $\min(T, S)$ is continuous and distributed according to $\text{Exp}(\lambda+\mu)$ (when do you eat it?). So we want to show that one discrete and one continuous random variable are independent of each other. This is a little tough.

Proof: Let f be the density of $\min(S, T)$ on the set $1(S < T)$.

$$f = f_s(T)\,\mathbb{P}[T > t] = \lambda e^{-\lambda t} e^{-\mu t} = \mathbb{P}[S < T](\lambda + \mu) e^{-(\lambda+\mu)t}$$

So they're independent!

This is sort of the analog of the discrete result that $\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x]\mathbb{P}[Y = y]$ or the continuous result that $f(x, y) = f_X(x) f_Y(y)$.

Note: All of this can be generalized to n independent exponentials.

18.2 Poisson processes

Definition. N has a **Poisson** distribution with mean λ if

$$\mathbb{P}[N = n] = e^{-\lambda} \frac{\lambda^n}{n!}$$

Claim: $\mathbb{E}[N] = \lambda$ and $\text{Var}[N] = \lambda$.

More generally, $\mathbb{E}[N(N-1)\dots(N-k+1)] = \lambda^k$.
This lets you compute $\mathbb{E}[N^3]$, $\mathbb{E}[N^4]$ etc.

Claim: If N_1, N_2 are independent Poisson variables with parameters λ_1, λ_2, then $N_1 + N_2$ is Poisson with parameter $\lambda_1 + \lambda_2$.

Proof: As follows:

$$\begin{aligned}
\mathbb{P}[N_1 + N_2 = n] &= \sum_{k=0}^{n} \mathbb{P}[N_1 = k]\,\mathbb{P}[N_2 = n-k] \\
&= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k}{k!} \cdot \frac{\lambda_2^{n-k}}{(n-k)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \lambda_1^k\, \lambda_2^{n-k} \frac{n!}{k!(n-k)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \binom{n}{k} \lambda_1^k\, \lambda_2^{n-k} \qquad \text{the binomial formula!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1 + \lambda_2)^n
\end{aligned}$$

Definition. The **Poisson process** with intensity λ is defined via independent exponential random variables $\tau_1, \tau_2, \ldots$ with parameter λ. The process is defined in terms of $T_n = \sum_{i=1}^{n} \tau_i$, called the n-th **arrival time** ($T_0 = 0$), and $N(s) = \max(n : T_n \leq s)$ which is the **number of arrivals** by time s.

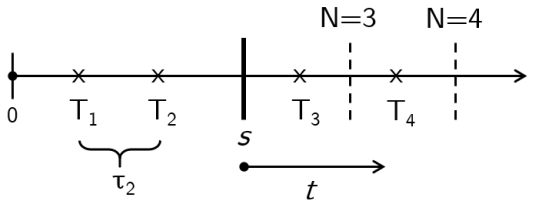
Returning to the meal analogy, T_n is the time of the n-th meal. Each τ_i is the spacing between two meals. $N(s)$, the number of arrivals by time s, is the number of meals eaten by s.

Claim: $N(s)$ is a Poisson random variable with parameter λs.

Proof: Let's consider the probability that $N(s) = k$.

$$\begin{aligned}
\mathbb{P}[N(s) = k] &= \mathbb{P}\left[\tau_1 + \cdots + \tau_k \leq s, \tau_1 + \cdots + \tau_{k+1} > s\right] \\
&= \int_0^s f_{T_k}(t)\, \mathbb{P}[\tau_{k+1} > s - t]\, dt \\
&\quad \text{prob. kth meal at } t \times \text{prob. next meal not until after } s \\
&= \int_0^s \lambda e^{-\lambda t} \frac{(\lambda t)^{k-1}}{(k-1)!} \cdot e^{-\lambda(s-t)} \\
&= \lambda e^{-\lambda s} \int_0^s \frac{(\lambda t)^{k-1}}{(k-1)!} dt \\
&= e^{-\lambda s} \frac{(\lambda s)^k}{k!}
\end{aligned}$$

Claim: $(N(t+s) - N(s))_{t \geq 0}$ is a Poisson process with rate λ independent of $(N(r) : r \leq s)$.



So at time $s + t$, measured from s, $N = \text{Poisson}(\lambda t)$.

Sketch of proof: $(N(r) : r \leq s)$ is a sequence of times $\leq s$ where there are arrivals. The time s will fall between two arrivals:

$$T_1, T_2, \ldots, T_k \leq s \text{ and } T_{k+1} > s$$

The time separating T_k (which falls before our threshold time s) and T_{k+1} (which falls after s) is τ_{k+1}. We know that τ_{k+1} is exponentially distributed, but what we care about is $T_{k+1} - s$, because that's our new τ_1' for the process that begins at s.

The lack of memory property says that the distribution of $T_{k+1} - s$ is Exp(λ).

So this gives us our new τ_1', and we know that all τ_i after that are exponentially distributed from the problem statement. So it's a Poisson process.

Theorem.

If $(N(s) : s \geq 0)$ is a Poisson process with intensity λ, then:

1. $N(0) = 0$.
2. $N(s+t) - N(s) \sim \text{Poisson}(\lambda t)$.
3. $N(t)$ has independent increments*.

Conversely, if $N(s)$ satisfies (1), (2), and (3), it is a Poisson process.

* This means that the number of arrivals in some interval $[t_i, t_{i+1}]$ is Poisson($\lambda(t_{i+1} - t_i)$). It is independent of the number of arrivals in another interval $[t_j, t_{j+1}]$, with $j \neq i$.

18.2.1 Binomial processes

In the binomial process, we look at intervals on $[0, 1]$ of length $1/n$. Each interval is "on" with probability λ/n and "off" otherwise. (Physically, "on" might mean a drop of rain landed there, or in a patch of sky there's a star there, etc.)

Let $N_n(s)$ be the number of subintervals of $[0, s]$ that are on.

Claim: $\lim_{n\to\infty} \mathbb{P}[N_n(s) = k] = \mathbb{P}[\text{Poisson}(\lambda s) = k]$.

The proof isn't too hard:

$$\mathbb{P}[N_n(s) = k] = \binom{sn}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{sn-k}$$

and it goes from there.

We'll end on a claim to begin with next time:

Claim: $\text{Bin}(n, \lambda/n) \to \text{Poisson}(\lambda)$ as $n \to \infty$.

19 April 27, 2020

Picking up where we left off last time...

19.1 Binomial processes

Claim: $\text{Bin}(n, \lambda/n) \to \text{Poisson}(\lambda)$ as $n \to \infty$.

Proof: We'll look at the probability associated with k intervals being "on".

$$\begin{aligned}\mathbb{P}\left[\text{Bin}(n,\lambda/n) = k\right] &= \binom{n}{k}\left(1-\frac{\lambda}{n}\right)^{n-k}\left(\frac{\lambda}{n}\right)^{k} \\ &= \frac{n!}{k!(n-k)!}\cdot\left(\frac{1}{n}\right)^{k}\lambda^{k}\cdot\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\cdot\frac{n(n-1)\ldots(n-k+1)}{n^k}\cdot\left(1-\frac{\lambda}{n}\right)^{n-k} \\ \lim_{n\to\infty} \implies \frac{\lambda^k}{k!}\cdot 1\cdot e^{-\lambda} &= \mathbb{P}\left[\text{Poisson}(\lambda) = k\right]\end{aligned}$$

Claim: Call T the first interval that's "on". T is distributed exponentially.

Proof: The probability that T occurs at some $0 \le X \le 1$ is the probability that everything is "off" until X and then "on".

$$\mathbb{P}[T = X] = \left(1-\frac{\lambda}{n}\right)^{nX}\cdot\frac{\lambda}{n} \longrightarrow \frac{1}{n}\cdot\lambda e^{-\lambda X}(1+O(1))$$

The $O(1)$ is because the nX exponent is a bit of a rounding error. The term $\lambda e^{-\lambda X}$ is the density of $\text{Exp}(\lambda)$. So $T \sim \text{Exp}(\lambda)$.

Claim: The number of successes in $[0, 1/2)$ is independent of the number of successes in $[1/2, 1)$.

We know the whole thing is Poisson, and these two intervals are disjoint. So we can use the independent increments property to show this.

19.2 Compound Poisson processes

Claim: Let $(X_m)_{m=1}^n$ be independent random variables with $\mathbb{P}[X_m = 1] = p_m$ and $\mathbb{P}[X_m = 0] = 1 - p_m$. Let:

$$S_n = \sum_{m=1}^{n} X_m \qquad \lambda = \mathbb{E}[S_n] \qquad Z_n = \text{Poisson}(\lambda)$$

Then $\|S_n - Z_n\|_{TV} \le \sum_{m=1}^n p_m^2$.

We're saying two random variables are close to each other:

If $\|S_n - Z_n\|_{TV} \le \epsilon$, then for all events A, $|\mathbb{P}(S_n \in A) - \mathbb{P}(Z_n \in A)| \le \epsilon$.

Definition. Consider a Poisson process $N(t)$ of rate λ and i.i.d random variables Y_i independent of the process. Then **compound Poisson process** is given by:

$$S(t) = \sum_{i=1}^{N(t)} Y_i$$

This is summing Y over the arrivals. Let each arrival be an order for toilet paper on an online retail store. The volume of the order is i.i.d Y_i. So:
At T_1, $Y_1 = 2$ rolls. $S(1) = 2$.
At T_2, $Y_2 = 1$ rolls. $S(2) = 3$.
At T_3, $Y_3 = 4$ rolls. $S(3) = 7$.

Claim: If $\mathbb{E}[Y_i^2] < \infty$:

1. $\mathbb{E}[S(t)] = \lambda t \cdot \mathbb{E}[Y_i]$. ($\lambda t = \mathbb{E}[\text{arrivals}] = \mathbb{E}[N(t)]$)
2. $\text{Var}[S(t)] = (\lambda t)\,\mathbb{E}[Y_i^2]$
   can be written: $\mathbb{E}[\text{Poisson}(\lambda t)]\,\text{Var}[Y_i] + \text{Var}[\text{Poisson}(\lambda t)]\,\mathbb{E}[Y_i^2]$

Theorem.

Let N (same as $N(t)$ earlier) be a non-negative random variable and let Y_i be i.i.d variables independent of N. Let $S = Y_1 + \cdots + Y_n$. Then:

1. If $\mathbb{E}[|Y_i|] < \infty$ and $\mathbb{E}[N] < \infty$, then $\mathbb{E}[S] = \mathbb{E}[N]\,\mathbb{E}[Y_i]$.
2. If $\mathbb{E}[Y_i^2] < \infty$ and $\mathbb{E}[N^2] < \infty$, then:
   $\text{Var}[S] = \mathbb{E}[N]\,\text{Var}[Y_i] + \text{Var}[N]\,\mathbb{E}[Y_i]^2$.

The first statement is just Wald's equation. The second follows from the **Law of Total Variance**:

$$\text{Var}[s] = \mathbb{E}[\text{Var}[S \mid N]] + \text{Var}[\mathbb{E}[S \mid N]]$$

where $\text{Var}[S \mid N] = \mathbb{E}[S^2 \mid N] - (\mathbb{E}[S \mid N])^2$.

Proof of (2): Using the Law of Total Variance:

$$\begin{aligned}\text{Var}[S] &= \mathbb{E}[N \cdot \text{Var}[Y_i]] + \text{Var}[N \cdot \text{Var}[Y_i]] \\ &= \mathbb{E}[N] \cdot \text{Var}[Y_i] + \text{Var}[N] \cdot \mathbb{E}[Y_i]^2\end{aligned}$$

19.2.1 Thinning and superposition

We can use compound Poisson processes to analyze Poisson thinning and superposition. Consider a compound Poisson process of rate λ with Y_i

taking the values $\{1, \ldots, k\}$. Think of the Y as categories.

Definition. Let $N_a(t) = |\{i : i \le N(t), Y_i = a\}|$. Then $(N_a(t) : 1 \le a \le k)$ are independent Poisson processes with rates $(\lambda \mathbb{P}[Y = a] : 1 \le a \le k)$. This is **Poisson thinning.**

Back to the toilet paper example. Let T_1, T_2, and T_4 be the arrival times of 2-ply orders. Let T_3 be the arrival time of a 1-ply order. We're saying that the orders of 2-ply and 1-ply are independent Poisson processes.

Proof sketch: (For the claim in the definition.)

$$\begin{aligned}\mathbb{P}[N_1(t) = j, N_2(t) = \ell] &= \text{prob. } (j+\ell) \text{ points} \times \text{prob. } j \text{ of 1 and } \ell \text{ of 2} \\ &= \frac{(\lambda t)^{j+\ell}}{(j+\ell)!} e^{-\lambda t} \times \frac{(j+\ell)!}{j!\ell!} p^j (1-p)^\ell \\ &= \frac{(\lambda t p)^j}{j!} e^{-\lambda p t} \times \frac{(\lambda t \cdot (1-p))^\ell}{\ell!} e^{-\lambda(1-p)t}\end{aligned}$$

This is $\mathbb{P}[\text{Poisson}(\lambda p t) = j] \times \mathbb{P}[\text{Poisson}(\lambda(1-p)t) = \ell]$.

Definition. Let $N_1(t), N_2(t), \ldots, N_k(t)$ be independent Poisson processes with rates λ_i. Then $N_1(t) + \cdots + N_k(t)$ is a Poisson process with rate $\sum \lambda_i$. This is **superposition**.

Example: Consider three independent processes with rate 3, 2, and 1. What is the probability that in the first six arrivals, there are 3 from the first, 2 from the second, and 1 from the last.

Look at processes with $\lambda = 6$ when a point is in the first category with probability $1/2$, the second with $1/3$, and the third with $1/6$.

Theorem. (Conditioning)

Let $T_1, T_2, \ldots$ denote the arrival times of a Poisson process with rate λ. Let $U_1, \ldots, U_n$ be i.i.d $U[0, t]$ and let $V_1 < V_2 < \ldots$ be the U_i in increasing order.

Then condition on $N(t) = n$, the vector $(T_1, \ldots, T_n)$ has the same distribution as $(V_1, \ldots, V_n)$.

Example: Consider a Poisson process with $\lambda = 10$. Suppose that up to time 2 there are 2 arrivals. What is the probability that both are in [0,1]?

Let $U_1, U_2 \sim U[0, 2]$. By the theorem, the probability we are looking for is $\mathbb{P}[U_1, U_2 \in [0, 1]] = (1/2)^2 = 1/4$.

Proof: Compute the conditional density of $T_1 = t_1, T_2 = t_2, \ldots$:

$$\frac{1}{e^{-\lambda t}(\lambda t)^n/n!} \times e^{-\lambda(t-t_n)} \prod_{i=1}^{n} \lambda e^{-\lambda(t_i - t_{i-1})} = \frac{n!}{t^n}$$

So the density is uniform over $0 < v_1 < v_2 < \cdots < v_n < t$.

Note: This implies that if $s < t$ and $m \leq n$ then:

$$\mathbb{P}[N(s) = m \mid N(t) = n] = \binom{n}{m} \left(\frac{s}{t}\right)^m \left(\frac{t-s}{t}\right)^{n-m}$$

20 April 29, 2020

Some textbook suggestions if interested: Durrett's *Essentials of Stochastic Processes* and Levin and Peres' *Markov Chains and Mixing Times*.

Today we'll attack continuous time Markov chains.

20.1 Continuous time Markov chains

Definition. Let:

1. $(\phi_k)_{k=0}^{\infty}$ be a discrete-time Markov chain with transition matrix P
2. $(T_i)_{i=0}^{\infty}$ be the arrival times of a Poisson process of rate λ

The **continuous time chain** $(X_t)_{t \geq 0}$ defined by ϕ and (T_i) is defined by letting $X_t = \phi_k$, for $T_k \leq t < T_{k+1}$.

So according to this definition, transitions from $\phi_1 \to \phi_2 \to \ldots$ don't occur at integer time steps ($t = 1, 2, \ldots$) but rather at $t = T_1, T_2, \ldots$ (the arrival times of the Poisson process).

This is a continuous time chain on a discrete space Markov chain. We won't look at continuous time, continuous space chains, but Brownian motion is the most famous (and a very interesting!) example.

Definition. Let $Q(x, y)_{x,y \in \Omega}$ be a *rate matrix*:

$$Q(x, y) \geq 0, \; x \neq y \qquad Q(x, x) = -\sum_{y \neq x} Q(x, y)$$

(so rows of Q sum to 0). Q describes how "quickly" you go from $x \to y$.

The **continuous time** Markov chain $(X_t)_{t \geq 0}$ defined by Q is the following process:

1. Given $X_t = x$, for each $y \neq x$, let $\tau_{x,y}$ denote an independent exponential random variable with parameter $Q(x, y)$.
2. At time $t + \min_{y \neq x} \tau_{x,y}$, the chain moves to state y where y is the winner of this exponential race (i.e. $y = \operatorname{argmin} \tau_{x,y}$).

So this is a model where the infinitesimal rate of moving from $x \to y$ is $Q(x, y)$. More formally, the chance of moving from $x \to y$ in an interval $[t, t + h]$ is $h\,Q(x, y) + O(h)$ (proportional to the width of the interval).

Informally, both definitions are equivalent for $Q = \lambda(P - \mathbb{1})$. The RHS uses information from Def. 1, and the LHS uses Def. 2. Note that this is not a unique representation:

$$Q = \lambda(P - \mathbb{1}) = 2\lambda\left(\frac{P + \mathbb{1}}{2} - \mathbb{1}\right)$$

So now we're looking at a Poisson process with twice the rate, and a different P where half the time you stay put. So it makes sense why this is equivalent to the other one.

Claim: Consider a matrix Q as in Def. 2. Let $\lambda = \max_x(-Q(x, x))$. Then the Markov chain from Def. 2 can be implemented by Def. 1 with a Poisson process of rate λ and:

$$P(x, y) = \frac{Q(x, y)}{\lambda},\ x \neq y \qquad P(x, x) = 1 - \sum_{y \neq x} \frac{Q(x, y)}{\lambda}$$

Proof: The rows of P sum to 1. To show that it's a valid transition matrix, we must also show that $P(x, x) \geq 0$.

$$\sum_{y \neq x} \frac{Q(x, y)}{\lambda} = -\frac{Q(x, x)}{\lambda} \leq 1$$

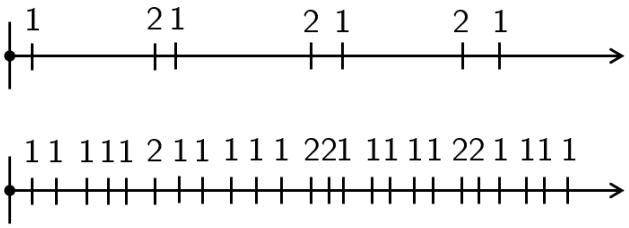
So we're fine.

In definition 2, if we start at x, the next state y is determined by an exponential race between $\tau_{x,y_1}, \tau_{x,y_2}, \tau_{x,y_3}, \ldots$. In definition 1, if we start at some x, the next state y is picked according to P and the transition time is determined by $\text{Exp}(\lambda)$. *This is Poisson thinning!* Definition 2 uses separate processes of rate $Q(x, y_i)$, and Definition 1 combines them all into one process.

Let's use the following Q (which implies the following P):

$$Q = \begin{pmatrix} -1 & 1 \\ 10 & -10 \end{pmatrix} \implies P = \begin{pmatrix} 9/10 & 1/10 \\ 1 & 0 \end{pmatrix}$$

This makes the difference between the Q-approach and the P-approach clear:



On top is what the Q definition would give us.
On the bottom is the P definition.

The Q definition uses 2 separate exponential variables, and the P definition uses one Poisson process and acts with probabilities from P.

Claim: Given P, λ as in Def. 1, the chain is equivalent to the chain of Def. 2 with $Q = \lambda(P - \mathbb{1})$.

(This follows from the discussion above.)

Claim: π is stationary for P iff $\pi Q = 0$.

Proof: π is stationary iff:

$$\pi P = \pi \implies \pi(P - \mathbb{1}) = 0 \implies \pi Q = 0$$

20.2 Heat kernel

Definition. The **heat kernel** of a continuous time chain is:

$$H_t(x, y) \coloneqq \mathbb{P}_x[X_t = y] = \mathbb{P}[X_t = y \mid X_0 = x]$$

Claim: $H_t = e^{tQ}$.

Proof: The definition of a matrix exponential is $e^{tQ} = \sum_{n=0}^{\infty} \frac{1}{n!}(tQ)^n$. We'll proceed using definitions:

$$\begin{aligned} H_t &= \sum_{n=0}^{\infty} \mathbb{P}[N(t) = n]\, P^n \\ &= \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t}\, P^n \\ &= e^{-\lambda t}\, e^{\lambda t P} \\ &= e^{\lambda t(P - \mathbb{1})} = e^{tQ} \end{aligned}$$

Definition. π is **stationary** for the chain with rate matrix Q if $\pi H_t = \pi$ for all t.

Claim: If π is stationary for Q, $\pi Q = 0$.

Proof: Suppose $\pi e^{tQ} = \pi$ for all t. Then:

$$\begin{aligned} \pi(e^{tQ} - \mathbb{1}) &= 0 \\ \pi \frac{e^{tQ} - \mathbb{1}}{t} &= 0 \qquad \text{for all } t > 0 \\ \implies \lim_{t\to 0} \frac{\pi(e^{tQ} - \mathbb{1})}{t} &= 0 \\ \implies \lim_{t\to 0} \pi \left(\sum_{n=1}^{\infty} \frac{(tQ)^n}{n!} \right) \cdot \frac{1}{t} &= \pi Q = 0 \end{aligned}$$

If $\pi Q = 0$, $\pi P = \pi$.

$$\pi H_t = \pi \sum_{n=0}^{\infty} \mathbb{P}[N(t) = n]\, P^n = \pi \sum_{n=0}^{\infty} \mathbb{P}[N(t) = n] = \pi$$

Theorem. (Long-term behavior)

Let P be an irreducible finite Markov chain.
Let $\tilde{P} = 0.5(P + \mathbb{1})$.
Let π be the stationary distribution for P.
Let H_t be the heat kernel for $P - \mathbb{1}$ (or Q). Then for any x:

$$\|\delta_x H_t - \pi\|_{TV} \xrightarrow[t\to\infty]{} 0$$

Moreover,

$$\begin{aligned} \|\delta_x H_k - \pi\|_{TV} &\le \left\|\delta_x \tilde{P}^k - \pi\right\|_{TV} + \mathbb{P}[\text{Poisson}(2k) < k] \\ &\le \left\|\delta_x \tilde{P}^k - \pi\right\|_{TV} + \left(\frac{2}{e}\right)^k \end{aligned}$$

This is like the result for finite, discrete, irreducible, aperiodic chains! Except this has no aperiodic requirement.

Definition. Q is **reversible** with respect to π if

$$\pi(x)Q(x,y) = \pi(y)Q(y,x)$$

Claim: If Q is reversible with respect to π then π is stationary for Q.

Proof: If $\pi(x)Q(x,y) = \pi(y)Q(y,x)$, then $\pi(x)P(x,y) = \pi(y)P(y,x)$ (since $Q = \lambda(P - \mathbb{1})$). So π is stationary for P, and thus π is stationary for Q.

We could do this proof slightly differently! We want to show that $\pi Q = 0$, or that $\sum \pi(x)Q(x,y) = 0$. Since Q is reversible,

$$\sum \pi(x)Q(x,y) = \sum \pi(y)Q(y,x) = \pi(y) \cdot 0 = 0$$

21 May 4, 2020

There are three classes left, including this one! Wednesday is Midterm 2, and on Monday we'll do some fun topics. Today we'll go over some examples and review questions.

A small barbershop. A barber cuts hair at a rate of 3 (an Exp(3) process) and customers arrive at a rate of 2, but will leave if the waiting room is full. The barbershop has 1 chair for haircuts and 2 chairs in the waiting area. Model this as a continuous time MC and find the stationary distribution of this process.

Solution. The state space (possible number of customers in the shop) is $\Omega = \{0, 1, 2, 3\}$. So this gives us:

$$Q = \begin{pmatrix} -2 & 2 & 0 & 0 \\ 3 & -5 & 2 & 0 \\ 0 & 3 & -5 & 2 \\ 0 & 0 & 3 & -3 \end{pmatrix}$$

Now we can find π. We can either solve $\pi Q = 0$ or $\pi_i Q_{ij} = \pi_j Q_{ji}$ (the detailed balance equations) for π, with the additional constraint $\sum_i \pi_i = 1$. Either way, we get:

$$\pi = \frac{1}{65}\begin{pmatrix} 27 & 18 & 12 & 8 \end{pmatrix}$$

A barbershop customer. What is the average amount of time spent in this system for a customer who receives service? What is the fraction of interested customers who actually receive a haircut?

Solution. We can use the memoryless property and π to answer the first question. When 27/65 customers enter the shop, there will be 0 other people inside and they will wait 1 turn to leave with a haircut. When 18/65 people enter, there will be 1 other person and they will wait 2 turns to leave with a haircut. When 12/65 people enter, there will be 2 other people and they will wait 3 turns to leave with a haircut. When 8/65 people enter, they will immediately leave because the shop will be full. The average time for a haircut is 1/3, since the rate of the process is 3. So:

$$A = \frac{1}{27 + 18 + 12}\frac{1}{3}(27 \cdot 1 + 18 \cdot 2 + 12 \cdot 3)$$

For the second question, we have shown that 8/65 people leave without getting a haircut. So 57/65 people receive service.

Towering property. We have seen $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$. Show:

$$\mathbb{E}\left[\mathbb{E}[X \mid Y, Z] \mid Y\right] = \mathbb{E}[X \mid Y]$$

Solution. Write $\mathbb{E}'[W] = \mathbb{E}[W \mid Y = y]$ for a fixed y and for any W.

So the RHS above is $\mathbb{E}[X \mid Y = y] = \mathbb{E}'[X]$. And the LHS:

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}[X \mid Y, Z] \mid Y\right] &= \mathbb{E}\left[\mathbb{E}[X \mid Y = y, Z] \mid Y = y\right] \\ &= \mathbb{E}'\left[\mathbb{E}'[X \mid Z]\right] \\ &= \mathbb{E}'[X]\end{aligned}$$

So LHS = RHS. □

More conditional expectation. Show that:

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X \mid \mathbb{E}[X \mid Y]]$$

Solution. Start with the RHS:

$$\begin{aligned}\mathbb{E}[X \mid \mathbb{E}[X \mid Y]] &= \mathbb{E}[\mathbb{E}[X \mid \mathbb{E}[X \mid Y], Y] \mid \mathbb{E}[X \mid Y]] \\ &= \mathbb{E}[\mathbb{E}[X \mid Y] \mid \mathbb{E}[X \mid Y]] \qquad \text{using previous result} \\ &= \mathbb{E}[X \mid Y]\end{aligned}$$

So LHS = RHS. □

Pandemic supermarket. At most 10 people are allowed to be in a supermarket at any given time. Customers are in the store for Exp(1) time. Suppose a new customer arrives and finds the store full and 2 other people in line. What is their expected wait time?

Solution. Three people need to leave the supermarket before the third in line can go in. The first person leaves after Exp(10) time, because we can treat the entire occupancy of the store as a compound process. The second person also leaves after Exp(10) time because of the memoryless property. So does the third person. So:

$$\mathbb{E}[\text{waiting time}] = \mathbb{E}[\text{Exp}(10) + \text{Exp}(10) + \text{Exp}(10)] = \frac{3}{10}$$

So the new customer can expect to wait for 18 minutes.

Poisson processes. Consider an experiment with 2 independent Poisson processes with rates λ_1, λ_2. Let (T_i) be the arrival times of the first process and (S_i) be the arrival times of the second. $\mathbb{P}[T_n < S_m] = ?$

Solution. Together we have a Poisson process of rate $\lambda_1 + \lambda_2$. $\lambda_1/(\lambda_1 + \lambda_2)$ are of type 1, and $\lambda_2/(\lambda_1 + \lambda_2)$ are of type 2. The probability that we are looking for is the probability that of the first $n + m - 1$ points, at least n are of type 1. So:

$$\mathbb{P} = \sum_{k=n}^{n+m-1} \binom{n+m-1}{k} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n+m-1-k}$$

Geometric exponentials. X_i are i.i.d. exponential variables with parameter λ. N is an independent geometric random variable with parameter p. Show that $\sum_{i=1}^{N} X_i$ is $\text{Exp}(\lambda p)$.

Solution. This is just Poisson thinning! Start with a Poisson process of rate λ. A point is "good" with probability p, and "bad" with parameter λp. The time until the first good point is $\sim \text{Exp}(\lambda p)$, which is the same as $\sum_{i=1}^{N} X_i$ (summing up arrival times until we get a good point).

Fun with binomials. Let X_n take values in $\{0, \ldots, 10\}$ as follows: $X_{n+1} \sim \text{Bin}(10, 0.1X_n)$.

1. Show X_n is a martingale.
2. Show X_n is a Markov Chain.
3. Find $\lim_{n\to\infty} X_n$.

Solution. $\mathbb{E}[X_{n+1} \mid X_{[n]}] = \mathbb{E}[\text{Bin}(10, 0.1X_n)] = 10 \cdot 0.1X_n = X_n$. So X_n is a martingale.

X_n is clearly a Markov chain because its current value on depends on its previous value. If X_n is 0, it stays there with probability 1. If X_n is 10, it stays there with probability 1. From any point in between, it can reach any element of $\{0, \ldots, 10\}$. So this is a bounded martingale, and $\mathbb{E}[X_\infty] = \mathbb{E}[X_0]$. Then:

$$\mathbb{P}[X_\infty = 10] = \frac{X_0}{10} \qquad \mathbb{P}[X_\infty = 0] = \frac{10 - X_0}{10}$$

Submarine failure. A submarine has 3 navigational devices. It can stay at sea if at least 2 are functional. Suppose the failure times are independent exponential random variables with means 1, 1.5, and 3. Compute the average time the submarine spends at sea (if it started with all three working).

Solution. We have three random variables: Exp(1), Exp(2/3), Exp(1/3). The first failure occurs according to an exponential race described by Exp(1+2/3+1/3) = Exp(2). So the first failure will occur at time 1/2.

The time of failure and the particular instrument that failed are independent. So with probability 1/2 the first failed and 2 and 3 remain, with probability 1/3 the second failed and 1 and 3 remain, and with probability 1/6 the third failed and 1 and 2 remain. If the first failed, the second failure will occur according to Exp(1). If the second, Exp(3/4). The third, Exp(5/3). So the total expected time until failure is:

$$\frac{1}{2} + \frac{1}{2} \cdot 1 + \frac{1}{3} \cdot \frac{3}{4} + \frac{1}{6} \cdot \frac{3}{5} = 1.35$$

22 May 11, 2020

We talked about this paper, which discusses optimal strategies and a probabilistic analysis of the game Mafia.

And that's a wrap on this class!