18.615 — Introduction to Stochastic Processes

Class by Elchanan Mossel

Notes by Sanjana Das

Spring 2025

Lecture notes from the MIT class **18.615** (Introduction to Stochastic Processes), taught by Elchanan Mossel. All errors are my own.

Contents

§1 February 4, 2025

§1.1 Logistics

There's a syllabus; we can still change some things as we go. One thing Prof. Mossel plans to do differently is to use the blackboard instead of slides. He also likes students to attend, so the syllabus sounds super aggressive, but he doesn't actually mean it so tough. There are 5 lectures you can not attend and we don't worry about it.

The other administrative issue is that since some of us might have conferences or job interviews or so on, we should take notes of the days of the midterms, and make sure we're in town on those days. It's the day before spring break and May 8.

The other interesting thing about this class is that it tends to have a very interesting mixture of students (we have both freshmen and grad students). Prof. Mossel will try to make it interesting for everyone, to the best we can. From next lecture on, he'll try to find a teaser problem related to each lecture, which we can think about if we're bored.

§1.2 Overview

Traditionally, the main topics for this class are Markov chains and martingales. The reason we're writing them on two boards is there's a zoo of what you can do with these. Markov chains can have various forms — the state space can be either finite or countably infinite or even $\mathbb{R}$ (or some other continuous topological space), and time can either be discrete or continuous. So we're definitely going to spend time on discrete-time finite, and some time on discrete-time countable; we might spend some time on discrete-time $\mathbb{R}$. We're also going to spend time on continuous-time finite and continuous-time countable, and maybe continuous-time $\mathbb{R}$.

There's the values that a Markov chain can take — e.g. 1, 2, ..., 6 (from the roll of a dice), which has to do with the state space. There's also the time — am I rolling the dice every second, or continuously?

With martingales, usually in this class we only do discrete-time martingales. Here it somehow doesn't matter if the state space is discrete or continuous, so we treat all these cases the same. Prof. Mossel is not sure if we'll talk about continuous-time martingales; if we do, maybe we'll talk about Brownian motion, which is in fact a continuous-time martingale and Markov chain and everything.

So that's broadly the plan of the class.

One example of a continuous Markov chain on $\mathbb{R}$ is Poisson processes; this we *will* talk about.

§1.3 Why Markov chains?

We'll start with Markov chains, and that'll be the topic of the next few classes. We should feel free to give feedback in or after class.

First, let's give a motivating pitch for Markov chains. They're a nice model for lots of stuff. They're used to model 'systems' that change with time; in some sense, they're the simplest models. Some examples might be the weather, the magnetization of atoms, the structure of populations of species — these are all from the natural sciences. But these days they're also more and more used in computer science-related things — how people browse (i.e., links followed — this is behind Google Score). And it's also related to LLMs — they in some sense use 'big memory' Markov chains. So these are some motivations for studying Markov chains.

§1.4 Definitions

Now let's go to the math; we'll start with the definition of a Markov chain.

Definition 1.1. A sequence of random variables (RVs) X_0, X_1, ... is a (discrete space, discrete time) Markov chain if for all possible values x_0, x_1, ... and all $t = 1, 2, \ldots$, we have

$$\mathbb{P}[X_t = x_t \mid X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}] = \mathbb{P}[X_t = x_t \mid X_{t-1} = x_{t-1}].$$

(We usually use capital letters for random variables.)

We think of t as a time. And in the condition, we're looking at the probability X_t has a given value at x_t, given the whole history of what's happening before time t — given *everything* we've seen so far, what's the probability $X_t = x_t$? If it were *independent* of what we'd seen before, none of the conditioning would matter. We want it to matter, but just a little bit — we *just* need the information on the last step, and nothing else.

Since this probability on the right-hand side is going to be very important, we usually give it a name:

Notation 1.2. We write $p_t(x_{t-1}, x_t) = \mathbb{P}[X_t = x_t \mid X_{t-1} = x_{t-1}]$; this is called the transition probability at time t from x_{t-1} to x_t.

Sometimes we want the transition rule from step $t-1$ to step t to not actually depend on t:

Definition 1.3. We say the Markov chain is homogeneous if $p_t(x, y) = p(x, y)$ for all x and y.

In other words, it's the same transition rule that determines how I go from where I am at step $t-1$ to where I am at step t (the rule doesn't depend on t).

§1.5 Connection to linear algebra

We're going to use some linear algebra when we talk about this: If the state space (the set of values that the random variables X_0, X_1, ... take) is *finite*, then $p_t(x, y)$ is just a matrix — it's a table with the possible values of x on one dimension and y on the other (we'll write x for rows and y for columns), where in the location (x, y) we have $p_t(x, y)$. So this is just a matrix, which we call P_t.

So we may suspect there's some connection to linear algebra.

Remark 1.4. If the state space is countable, $p_t(x, y)$ is still an infinite matrix — this means you have an infinite (but countable) number of rows and columns.

Here's the first challenge. If you've seen Markov chains before, Prof. Mossel will only talk about finite Markov chains; you can think about what parts of this work for countable Markov chains. (For example, we'll be multiplying matrices; with infinite ones, you have to think about what this means and convergence issues and so on.)

Let's now see the connection to linear algebra. So far we've just used it as a way of representing the data, but the connection is a bit deeper than that.

Notation 1.5. We write μ_t for the row vector $(\mathbb{P}[X_t = x])_x$ (over all x in the state space).

In other words, μ_t is a sequence of numbers representing all the probabilities $\mathbb{P}[X_t = x]$, written as a row vector.

Question 1.6. What's the relationship between μ_t and μ_{t-1}?

Claim 1.7 — We have $\mu_t = \mu_{t-1}P_t$.

Proof. We have to check that this equation is correct; $\mu_{t-1}P_t$ is a row times a matrix, so it's also a row vector; and we want to check its value at some location y. We have

$$(\mu_{t-1}P_t)(y) - \sum_x \mu_{t-1}(x)p_t(x,y).$$

(Somehow the notation for linear algebra is weird, but there's no probability here, this is just linear algebra; this is the usual formula for multiplying a vector by a matrix, just that we're using indices x and y rather than i and j.) And the probabilistic interpretation of this is that it's

$$\sum_x \mathbb{P}[X_{t-1} = x]\mathbb{P}[X_t = y \mid X_{t-1} = x].$$

Now there's probability — we translated this number to the probabilistic interpretation. And by how conditional probability works, this is

$$\sum_x \mathbb{P}[X_t = y,\, X_{t-1} = x].$$

And since we're summing over all possible values of x, this is just equal to

$$\mathbb{P}[X_t = y] = \mu_t(y). \qquad \square$$

In Markov chains, lots of what we do is connecting linear algebra to probability. Sometimes what we do is purely linear algebra, and then we have to understand what it means in terms of probability; and then we get this equation.

We'll do one more thing: What we get from Claim 1.7 is that if the Markov chain is homogeneous (with transition matrix P), then we get $\mu_t = \mu_{t-1}P$; and there's no dependence on time, so we can iterate to get

$$\mu_t = \mu_{t-1}P = \mu_{t-2}P^2 = \cdots = \mu_0 P^t$$

(where μ_0 is the probability at the *beginning* that we're at each of the possible states).

§1.6 Examples

§1.6.1 Gambler's ruin

The first example, which we've probably seen before, is what's called *gambler's ruin*. You sort of have to give it in every probability class because it's historically very important.

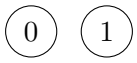
The story is a very complicated gambling game. I go into a casino with k dollars. The casino is actually fair. So we toss a coin; if it's heads I gain 1 dollar, and if it's tails I lose a dollar. As one more point, there's actually two cases where the casino kicks me out — either when I have no money (which makes sense, because they don't want to fund me), or when I have some amount of dollars (where they decide this guy made too much money, I don't trust him and I'll kick him out). So if I get 0 dollars or n dollars, I'm out.

The state space is $\{0, \dots, n\}$. It's homogeneous. Written as a matrix, we have

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1/2 & 0 & 1/2 & & \\ & 1/2 & 0 & 1/2 & \\ & & 1/2 & 0 & 1/2 \end{bmatrix}$$

(we have 0's on the diagonal except at $(0,0)$ and (n,n), and elsewhere we have $(1/2)$'s next to the diagonal on either side).

There's another way of representing this, where our states are vertices of a graph, and we draw directed edges representing movements; and we have a probability on each edge. For example, we draw an edge from 0 to 0 with label 1; and the same at n. Otherwise from every middle vertex we have an edge with $1/2$ up and $1/2$ down. For Markov chains where we don't move too much, this is a much easier representation of what's going on. (But if you jump by a lot, then this will be very mixed up and maybe the matrix is a better representation.)



Let's talk about a few basic results. We'll use τ to denote the (random) time when the game ends — meaning we're either at 0 or n. The usual two claims that people prove about gambler's ruin:

Claim 1.8 — We have $\mathbb{P}[X_\tau = 0 \mid X_0 = k] = 1 - \frac{k}{n}$ and $\mathbb{P}[X_\tau = n \mid X_0 = k] = \frac{k}{n}$.

When I finish the game, I know it's over, so I either have 0 or n. And I want to know, what's the probability I win (have n dollars) or lose (have 0), given that I started with k? And this is the answer.

Proof. (The good way to prove this is using martingales, but let's not do that because we don't know what martingales are yet.)

One way to do this is by solving a system of equations — we'll define

$$a_k = \mathbb{P}[X_\tau = n \mid X_0 = k]$$

as the probability we win (starting with k). And then we'll think about some equations for a_k. If we start with n dollars then we certainly win, so $a_n = 1$; similarly $a_0 = 0$. And then if $k \notin \{0, n\}$, what can I say? Either I'm going to lose a dollar (which is like starting the game at $k-1$ dollars) or win (which is like starting with $k+1$); this means

$$a_k = \frac{1}{2}a_{k-1} + \frac{1}{2}a_{k+1}.$$

(We're just thinking about the process one step forward; with probability $1/2$ I lose a dollar, which is like starting with $k-1$, and so on.) □

(Why isn't this really a proof? One thing that's missing is in principle you might never end at 0 or n. But it's sort of a proof.)

The other important claim, of a similar flavor, is how long does it take to play the game? Remember that τ is the random time when the game ends; so $\mathbb{E}[\tau]$ is how long the game takes. This is also an interesting question — maybe you go to the casino and you're willing to spend some money, but you don't want to spend a week there.

So what's $\mathbb{E}[\tau \mid X_0 = k]$? First, as an intuitive question, when would you expect the game to take longest? (Imagine you want to forget about the real world and spend the maximum amount of time at the casino.) You want to be in the middle — around $n/2$. And in fact:

Claim 1.9 — We have $\mathbb{E}[\tau \mid X_0 = k] = k(n-k)$.

This is maximized exactly at the center.

Proof. We can make a similar system of equations: now we'll define $t_k = \mathbb{E}[\tau \mid X_0 = k]$ (the expected length of the game, given that we start from k). We have $t_n = 0$ (because if we start with n, the game is already over), and similarly $t_0 = 0$. Otherwise, if $k \neq 0, n$, we have to look at cases again. The game isn't over immediately, so there's at least one step; and then either it's the length of the game starting with $k-1$ dollars, or $k+1$. So we get

$$t_k = 1 + \frac{1}{2}t_{k-1} + \frac{1}{2}t_{k+1},$$

and you can solve this equation and get $k(n-k)$ as the solution. □

§1.6.2 Coupon collector

In the past, people used to collect physical objects, in particular coupons. There were in fact cereal boxes with coupons on them, and you could collect the coupons, and you wanted all the coupons. (People might still do this with cards or Pokemons or something.) There's n types of coupons; each day you get a random coupon of one of the n types. If it's a type you already had before, you don't want it; if it's a new type then you're happy that you got a new card, and you put it on your album.

The state space would be $\{0, \dots, n\}$, representing the number of different types of coupons collected so far.

We'll draw the transitions as a directed graph, with nodes $0, \dots, n$. If we're at k, we stay there with probability k/n (if we get a coupon we already had), and move to $k+1$ with the remaining probability $1 - k/n$. At n we just stay there (with probability 1); similarly at 0 we move to 1 with probability 1.

What do you usually ask with this? Here the classical stopping time τ is the random time when all the coupons are collected.

Claim 1.10 — We have $\mathbb{E}[\tau] = 1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1} = n\sum_{k=1}^{n} \frac{1}{k}$.

If you remember $\sum_{k=1}^{n} \frac{1}{k} \approx \log n$, this is about $n \log n$.

Proof. Let T_k be the number of days needed to get the $(k+1)$th coupon, given that you had k coupons. Then T_k is a geometric random variable with probability of success $p = \frac{n-k}{n}$, which means

$$\mathbb{E}[T_k] = \frac{1}{p} = \frac{n}{n-k}.$$

Now, the time to get *all* the coupons is $\tau = T_1 + T_2 + \cdots + T_n$ (the time to get the first coupon, then the second, and so on); so by linearity of expectation,

$$\mathbb{E}[\tau] = \sum_{k} T_n = \sum_{k=0}^{n-1} \frac{n}{n-k}.$$

□

§1.6.3 More examples — the Google Markov chain

We're going to see many more examples of Markov chains, but these are a few.

- Branching processes.
- Google search.

- Contact processes (a model for the spread of disease).

We'll talk about Google search in more detail (based on a class vote). This is a cartoon; we're not actually going to describe the parameters of the Markov chain that Google is actually using, because we can't write a billion parameters on the board.

The main idea is that we can think of the internet (or rather webpages) as a graph $G = (V, E)$. A graph has two pieces — vertices (the basic elements) and edges (how you move between them). Here the vertices are going to be the webpatges; and E consists of the transition probabilities (or directed edges with weights). In our notation from before, we write

$$p'(x, y) = \mathbb{P}[\text{browsing } y \text{ at time } t \mid \text{browse } x \text{ at time } t - 1].$$

The main idea of Google at the time was that first, people spend infinite amounts of time on the internet; they go to a webpage, and then in this webpage there's a bunch of links embedded (ads, references to other places); you look at the font and how many times a link appears on the page; and from that, you can get a model of what people will do next. This gives a model of how people browse the internet — given this webpage, what's the chance you'll look at some other webpage?

Their actual model is

$$p(x, y) = \lambda p'(x, y) + (1 - \lambda) \cdot \text{random page},$$

where λ is close to 1. The point is that people aren't actually on the internet 24 hours a day, so maybe they leave, and when they come back they'll come to somewhere unrelated to where they were before.

Why was this important? At the time, people didn't know how to say if a website is good or popular or not. They had all kinds of heuristics. For example, if a lot of pages point to that page, maybe this is a good page. But that's a very easy way to create money — people just created lots of websites pointing to that page, and no one visited those pages. Google realized this isn't going to work; we want to look at webpages people go to, and then where websites no one goes to point to don't matter, because people have probability 0 of going there.

So there was the true internet and at the start this was good (e.g., CNN would have lots of links to it). But then people realized this was a good way to create money: I create a webpage SellJunk, and then create a billion webpages x, y, z, ... all pointing to this. Then the search engines of the time thought this was a very good page, and everyone went to SellJunk. The idea of Google was we want to follow a longer trajectory.

Remark 1.11. This kind of game still goes on. In science, there are journals — or at least there were — that said we'll publish your paper as long as you pay us enough money and you reference at least 40 other papers in this journal. Because the way people measure impact of papers is a similar idea — how many people link to you, and so on. So this would make our journal super popular, because we have a lot of citations.

§1.7 Long-term behavior

To conclude this class, we'll talk a bit about what we'll study in the next few classes. There's a piece of notation we're going to use (we're not going to prove anything more, but there'll be some new definitions, and next class we'll see some examples).

Something we'll often do to shorten notation:

Notation 1.12. We write $\mathbb{P}_x[X_t = y]$ to denote $\mathbb{P}[X_t = y \mid X_0 = x]$. More generally, we write $\mathbb{P}_x[A] = \mathbb{P}[A \mid X_0 = x]$. Similarly, we write $\mathbb{E}_x[Y] = \mathbb{E}[Y \mid X_0 = x]$.

What this means is we're going to be a bit lazy, and instead of actually writing the conditioning $X_0 = x$, we'll use a subscript x to denote this conditioning.

What's the main goal of the next few lectures?

> **Question 1.13.** Given a Markov chain $X_0, X_1, \ldots$, what happens in the limit?

This is called the *long-term behavior* — what can we say about X_t as $t \to \infty$? Informally, we have this random variable X_t, and we want to understand what happens in the limit as $t \to \infty$.

Now we're going to give a few possible definitions for what this question means — there's not just one interpretation but many.

One interpretation is the following. We'll use ν_t to denote the row vector

$$\mu_t = (\mathbb{P}[X_t = x])_x.$$

Maybe the first question we can ask is, what can we say about this vector μ_t?

> **Question 1.14.** What can we say about $\lim_{t\to\infty} \mu_t$? (Does the limit exist? Does it depend on X_0? And so on.)

So we have a sequence of vectors, and we want to understand whether it has a limit and what this limit is. If there's a limit to μ_t, this means when t is very, very large, if I ask you what the probability is that the Markov chain is at each of the states, you can say it's essentially the limit — it doesn't depend on t — and I can tell you what it is. So that's a very natural question.

Another question:

> **Question 1.15.** What can we say about $\lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \mu_s$?

This measures the expected fraction of time spent at each state — the first coordinate of this vector says, up to time t, how much of the time is spent (in expectation) at state 1? What about state 2? And so on. So this is also a natural question.

> **Question 1.16.** Is there a *stationary distribution* — a row vector π such that $\pi P = \pi$?

What this means is if you start from π and go one step, you're still at π.

Maybe the most sophisticated question we'll mention today:

> **Question 1.17.** What is $\lim_{t\to\infty} \frac{1}{t}(\#\{s \mid X_s = x\})_x$?

So we're looking at a random vector X_s, and some state (e.g., 1). And we're asking, until time t, what fraction of the times in my random sequence were we actually at state 1? Or state 2? And so on. Note that (1), (2), and (3) are all linear algebra — they're some questions about a vector. But this is a question about a random variable — up to time t, what fraction of the time *in the random sequence I'm observing right now* did I spend at state 1, or 2, or so on?

We'll ask whether this limit exists *almost surely*. (The strong law of large numbers is something that happens almost surely — we look at the random quantity of the fraction of times we have heads or tails; and we say that this is a random variable which could take many values, but with probability 1 it takes the value 1/2.) Often this actually becomes a constant — with probability 1, this will converge to some fraction.

The plan for next lecture is we'll start again with these questions, and then after that we'll look at some examples and see how they answer this question.

(The difference between (2) and (4) is that in (2), μ_t is just a vector; and we're averaging those. But (4) is about the *specific* realization of the Markov chain — we're now running the Markov chain, and it could be that we're always lucky, so it could be all-heads or all-tails. But we're asking, is there a phenomenon where even though it's random, almost surely it's 1/2?)

§2 February 6, 2025

Today we'll start taking attendance.

In a Markov chain, we have random variables X_0, X_1, ... We'll be talking about the homogeneous case, where we have a transition matrix P. We'll start with the distribution π_0 defined by $\pi_0(a) = \mathbb{P}[X_0 = a]$; and we saw that the vector π_t defined by $\pi_t(a) = \mathbb{P}[X_t = a]$ is given by

$$\pi_t = P^t \pi_0.$$

§2.1 Definitions for long-term behavior

We want to understand the long-term behavior of this Markov chain. We have vectors π_t denoting the distribution at time t. We'll write

$$\nu_t = \frac{1}{t} \sum_{s=1}^{t} \pi_s$$

to denote the *average* of the vectors π_s (each measuring the distribution of the Markov chain at time s) up to time t. Another thing we didn't define, but appeared in a complicated expression last time — note that π_t and ν_t are vectors. Now we'll define another quantity F_t, which is a *random* vector, as

$$F_t(a) = \frac{1}{t} \#\{1 \le s \le t \mid X_s = a\}.$$

So F_t is a random vector, and to define its ath coordinate, I'm counting all the times between times 1 and t where $X_s = a$.

Let's try again to understand what F_t is. It's a random vector; what's its ath coordinate? I look at what happened when I *actually* ran the chain and ask, how many times up to time t did I have $X_s = a$? And instead of counting, we look at the fraction — what fraction of the time up to t had $X_s = a$? This is random — it depends on the actual randomness we see in the Markov chain (it's not just a fixed vector).

§2.2 Questions

Now that we have all these objects, let's talk about long-term behavior. Last class, we asked a few long-term behavior questions:

> **Question 2.1.** Does $\lim_{t\to\infty} \pi_t$ exist?

This is a question about a sequence of vectors.

> **Question 2.2.** Does $\lim_{t\to\infty} \nu_t$ exist?

Note that π_t was about the distribution *at* time t; ν_t is about the distribution *up to* time t.

> **Question 2.3.** Does $\lim_{t\to\infty} X_t$ exist?

Now this is about a limit of random variables.

Question 2.4. Does $\lim_{t\to\infty} F_t$ exist?

So the first two questions are about vectors, and the last two are about random vectors — the third is about where our chain is, and the fourth about how much time we're spending in each state.

Finally, the last question might not look like one about long-term behavior, but it is.

Question 2.5. Does there exist π (which is a probability distribution) such that $\pi = \pi P$?

This may not look like a limiting question, but it is — if $\pi = \pi P$, then $\pi = \pi P^t$ for all t.

If we didn't have the condition that π is a probability distribution, the 0 vector would always work; so we need *some* condition, and the right condition is that it's a probability distribution.

Do we see any relationships between these — we're asking about four limits. There's in fact some obvious implications and some less obvious implications about the fact that some limit existing implies some other limit exists.

- If $\lim \pi_t$ exists, then $\lim \nu_t$ exists — if you have a limit of a sequence, and you look at *averages* of this sequence, then those also have a limit. In other words, if $\lim \pi_t = \pi$, then we also have $\lim \nu_t = \pi$ (since we're averaging things that converge to π, so this will also converge to π). In short notation, we can say (1) implies (2).

First, what does it mean for $\lim_{t\to\infty} X_t$ to exist? We didn't exactly specify what's going on here. What we like in probability is to say this limit exists, and it's a *constant* (not random). That's the kind of behavior we like in probability. But we didn't say that; we just wrote it in some rough way.

§2.3 Examples

This is a lot of limit notation, so before jumping into theorems, we'll look at some examples.

§2.3.1 Independent coin tosses

The first example is one that we didn't talk about last class.

Example 2.6 (Independent coin tosses)
Suppose that X_t are IID, and are 0 or 1 with probability 1/2.

This is a Markov chain with transition matrix

$$P = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$

(since no matter what happens, the next toss will be heads or tails with probability 1/2). Then we have $\pi_1 = (1/2, 1/2)$ — no matter what row vector you multiply P by, you'll get $(1/2, 1/2)$, i.e.,

$$\pi_1 = (\pi_0(0), \pi_1(1)) \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} = (1/2, 1/2).$$

Similarly, $\pi_t = (1/2, 1/2)$ for all t. This makes sense — it's independent coin tosses. When we independently toss coins, at each time t, we're equally likely to have heads and tails; this means $\pi_t = (1/2, 1/2)$.

Of course if that's the case, then the answer to (1) is yes — $\lim_{t\to\infty} \pi_t = (1/2, 1/2)$. We saw this also implies (2) — we have $\lim_{t\to\infty} \nu_t = (1/2, 1/2)$.

How about (3)? Does $\lim_{t\to\infty} X_t$ exist? First, how should we think about this? If I look at a specific realization of coin tosses, like $0011100011010\ldots$ — does this limit exist? The thing fluctuates between two values, so the limit doesn't exist. A typical sequence of coin tosses is going to fluctuate between 0 and 1. Think about it as looking at the value of a stock — if it keeps going up and down, we wouldn't say the limit exists. So typically the limit doesn't exist — in fact,

$$\mathbb{P}\left[\lim_{t\to\infty} X_t \text{ exists}\right] = 0.$$

(There is a chance it exists — the sequence could be all-1's or all-0's — but that has probability 0.)

Student Question. *Why can't we say this limit is* $(1/2, 1/2)$*?*

Answer. When we write X_t, we mean the actual realization fo this random variable, not the distribution. The *distribution* of X_t is $(1/2, 1/2)$, which is constant and does converge. But X_t is the random variable itself. And if I look at a typical realization, typically you'll see a sequence that does not converge. This is exactly the point that's sort of delicate.

So the answer to (3) is no. How about (4)? Now we're looking at the vectors F_t. It's still a random vector. Let's say we're doing an experiment on our computer, where we're simulating coin tosses; but instead of keeping track of X_t, we keep track of a different quantity. So F_{1000} has two coordinates, the first telling me the fraction of heads in the first 1000 coin tosses, and the second telling me the fraction of tails. What do we know about $\lim_{t\to\infty} F_t$? It converges to $(1/2, 1/2)$ — more precisely,

$$\mathbb{P}\left[\lim_{t\to\infty} F_t = (1/2, 1/2)\right] = 1.$$

What this is saying is $\lim_{t\to\infty} F_t$ is a limit of random quantities; and this limit of random quantities converges to a deterministic vector. This doesn't *always* happen, but it happens with probability 1. This is what we call the *strong law of large numbers* — that says when we toss a fair coin repeatedly, we're going to have half heads and half tails (in the limit) with probability 1.

This was not really an example about Markov chains (or it was a silly Markov chain, with no memory whatsoever); but it's more to make sure we're good with the definitions.

§2.3.2 Cyclic walk

Now we'll see the next example. Prof. Mossel prepared three — gambler's ruin, a cyclic walk (where you walk on a cycle over and over), and the binary symmetric channel.

We'll start with the cyclic walk; this is actually deterministic, but we'll do that. (This is actually the simplest, so it's a good choice.)

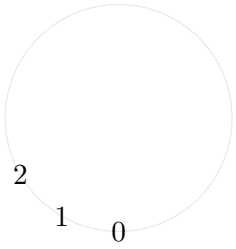
Example 2.7 (Cyclic walk)

Our random variable X_t takes values $0, \ldots, k-1$, where

$$X_{t+1} = X_t + 1 \bmod k$$

(with probability 1).

We're really thinking about the numbers as sitting on a cycle, so everything's mod k.



This is a Markov chain, with matrix consisting of 1's one step right of the diagonal, and 0's elsewhere.

For Question 2.1, the answer is generally no — for example, if $\pi_0 = (1, 0, \ldots, 0)$, then $\pi_1 = (0, 1, 0, \ldots)$, and $\pi_2 = (0, 0, 1, \ldots)$, and it's going to circle around. So we have a sequence of vectors where the location of the 1 circles around; this does not have a limit. So the answer to Question 2.1 is no.

How about Question 2.2? We're looking at a vector and shifting by 1 over and over again; after k iterations they'll average out to $1/k$. We'll have a bit of junk, but that disappears as the number of iterations grows; so we get

$$\lim_{t\to\infty} \nu_t = \frac{1}{k}(1, 1, \ldots, 1).$$

For Question 2.3, the answer is no — we're circling round and round, so this isn't going to converge. For Question 2.4, this does converge — formally, for all π_0, it's always true that

$$\lim_{t\to infty} F_t = \frac{1}{k}(1, 1, \ldots, 1).$$

(This is *always* true — not even just with probability 1.)

§2.3.3 Gambler's ruin

For the next example:

Example 2.8 (Gambler's ruin)
We have $\mathbb{P}[X_{t+1} = X_t \pm 1] = \frac{1}{2}$ if $1 \le X_t \le n-1$, and $X_{t+1} = X_t$ if $X_t = 0$ or $X_t = n$.

Let's look at all these questions again. These are all interesting questions now.

What about Question 2.1? Last class, we saw that if $X_0 = k$, meaning that $\pi_0 = (0, \ldots, 1, 0, \ldots)$, then we know eventually X_τ will be n with probability k/n and $X_\tau = 0$ with probability $(1 - k/n)$ — eventually we're either going to be at 0 or n. So for this value of π_0, we get

$$\lim_{t\to\infty} \pi_t = \left(1 - \frac{k}{n}, 0, 0, \ldots, \frac{k}{n}\right).$$

So if we start at specifically this π_0, that's what's going to happen.

And then what will be the general formula? Let's think about what this means. This means instead of someone coming to the casino with 5 dollars, they have probability 1/2 of starting with 5 dollars, and probability 1/2 of starting with 10. Then what'll happen is with probability 1/2 they'll have the first limit, and with probability 1/2 they'll have the second. So this is a weighted average of the others, which means we get

$$\pi_t = \left(1 - \bullet, 0, \ldots, 0, \frac{\pi_0(0)\cdot 0 + \pi_1(1)\cdot 1 + \cdots + \pi_0(n)\cdot n}{n}\right),$$

where $\bullet$ is the last coordinate. (If we start at 0 we know we'll never end up at n; if we start at 1 we'll end up there with probability $1/n$; and so on. So overall we get this weighted expression.)

Student Question. *What does this mean?*

Answer. From the casino perspective, maybe half of people are careful and spend 5 dollars, and half of people are not careful and spend 10 dollars. If there's a random person going into the casino, then I have to average these two cases, because I don't know who's going to come. And the $**$ quantity is the average of all these scenarios.

How abotu Question2.2? We saw that if the first limit exists, then $\lim_{t\to\infty}\nu_t = \lim_{t\to\infty}\pi_t$. But unlike in the previous examples, note that the answers to these two questions depend on how we started. In the previous examples — the cyclic walk and coin tosses — this limit didn't depend on where we started. Here it depends *strongly* on where we started — where it started determines a lot of what's going to happen.

What about Question 2.3? We'll get that

$$\lim_{t\to\infty} X_t = n \begin{cases} n & \text{with probability } ** \\ 0 & \text{with probability } 1-** \end{cases}$$

(where $**$ is that beautiful expression from before).

And what about (4)? We have

$$\mathbb{P}[\lim F_t = (0,\dots,0,1)] = **$$

(this means you spend all your time at n, and a 0-fraction in the remaining states), and

$$\mathbb{P}[\lim F_t = (1,0,\dots,0)] = 1-**.$$

Finally, we haven't talked about the stationary distributions. For this example, what are the stationary distributions — meaning that if you start at π, at the next stage you're also at π? If I start with 0 dollars I know I end up with 0, and if I start at n I know I also end up at n. So both of those are stationary distributions. And any combination of those is also stationary — if

$$\pi = (p,0,0,\dots,1-p),$$

this is going to be stationary. (This is true for all $0 \le p \le 1$.)

§2.3.4 Binary symmetric channel

Maybe we'll do the example of the binary symmetric channel too, because that's maybe the example that's most typical to this class, and it's also an important example for engineering. But there there isn't an easy proof for everything.

Before, when we talked about independent coin tosses, we had a matrix

$$\begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix},$$

with no memory of what we did before. The binary symmetric channel is some modification of this, where we tend to be similar to what we were before — so we have transition matrix

$$P = \begin{bmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix}$$

(where $\varepsilon \le 1/2$). Note that $\varepsilon = 1/2$ gives us back independent coin tosses.

Why is this called a binary symmetric channel? This was one of the first models for what happens if you broadcast information with noise — if I send a bit over a channel in the air or a line, there is some chance ε

that the bit flips, and the remaining chance it stays the same. So this is the most basic information theory model for what is noise. And a lot of information theory has to do with this very simple channel (a lot of what Shannon and others did).

One of the most useful tricks in working with the binary symmetric channel, which we'll also use today, is that it's often more useful to write this as

$$(1-2\varepsilon)\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + 2\varepsilon\begin{bmatrix}1/2 & 1/2\\ 1/2 & 1/2\end{bmatrix}.$$

Why is this useful? We like the identity matrix a lot — it says whatever we started with, we're staying with. And we also like the second matrix a lot, because it's independent coin tosses. So we've somehow expressed the thing we care about as a convex combination of two things we like.

Let's try to answer all the questions that we were trying to answer before. For Question 1.14, does $\lim_{t\to\infty}\pi_t$ exist? We have

$$\pi_t = \left((1-2\varepsilon)\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + 2\varepsilon\begin{bmatrix}1/2 & 1/2\\ 1/2 & 1/2\end{bmatrix}\right)\pi_{t-1}.$$

We like the identity matrix a lot; and for the second, we saw that no matter what we start with, we get $(1/2, 1/2)$. So we get

$$\pi_t = (1-2\varepsilon)\pi_{t-1} + 2\varepsilon(1/2, 1/2).$$

If we recast this equation, we eventually get

$$(1-2\varepsilon)^t\pi_0 + (1-(1-2\varepsilon)^t)(1/2, 1/2).$$

(Here we do induction, inducting over this iteration.) And this converges to $(1/2, 1/2)$, since $(1-2\varepsilon)^t \to 0$. The intuition is that even though I have some memory with the last step, it decays exponentially, so that eventually I'm equally likely to be in either bit.

And we know Question 2.2 has the same answer, so $\lim_{t\to\infty}\nu_t = \lim_{t\to\infty}\pi_t = (1/2, 1/2)$.

How about Question 2.3? If I look at the actual values I get, what would this Markov chain correspond to in information theory? Maybe I submit a bit across one channel, corrupted with probability ε; and then I submit it across another channel; and so on. The first equation tells us I'll eventually get something that's completely random. The question about $\lim_{t\to\infty} X_t$ — we saw that the thing fluctuates, so this will *not* exist. (In the simulation picture on your computer, you're simulating what this Markov chain is doing; so you toss a coin that's with probability 2ε heads and $1-2\varepsilon$ tails. If it's tails you copy whatever you had before; if it's heads you randomly toss a new coin. You look at the sequence of numbers your computer spits out, which are just 0's and 1's. Even if you have a long sequence of 0's, there's some chance you get a random coin next; so it's not going to converge.)

And how about $\lim_{t\to\infty} F_t$? This one Prof. Mossel actually doesn't know how to prove without some more sophisticated stuff. What fraction of time am I going to have 0 vs. 1? It has to be $(1/2, 1/2)$ — we have

$$\mathbb{P}\left[\lim_{t\to\infty} F_t = (1/2, 1/2)\right] = 1.$$

Finally, for Question 2.5, if $\pi P = \pi$, then we can solve our equation; π times the identity is π, so it's just like the equation we solved from the coin tosses, and we get $\pi = (1/2, 1/2)$.

We didn't say it before, but at least for three of the four examples we were looking at, we had the same answer to Questions 2.1, 2.2, 2.4, and Question 2.5. Part of what we'll try to do in the next few classes, starting today, is to try to understand why.

§2.4 Existence of stationary distributions

We'll now state and prove the first theorem of the class.

Theorem 2.9

Let P be the transition matrix of a finite Markov chain. Then there exists a distribution π such that $\pi P = \pi$ (i.e., such that π is stationary).

So no matter what finite Markov chain you start with, you can always find a stationary distribution π.

What does a *finite* Markov chain mean? Formally, this means there exists a fintie set S such that the chain is always in S, meaning that $\mathbb{P}[X_t \in S] = 1$ for all t. (It's what you expect — there's a finite set, and the Markov chain is always in the finite set.)

So no matter what Markov chian you're looking at, there's at least one stationary distribution, where if you start here you'll stay there forever.

What branch of math does this theorem belong to? One is linear algebra, except there's this annoying thing that π has to be a distribution. That means its entries have to be nonnegative, so there's some inequality. So it's sort of linear algebra, but with a twist. Any other suggestions? For Prof. Mossel, it's also in topology or calculus or real analysis or something. He knows 7 proofs of this and he chose one.

Proof sketch. Basically we have to prove that there's some solution to some equation — we want a solution to $\pi P = \pi$. The first thing we'll do is talk about where the *potential* solutions live. There's a name for this, called the *probability simplex* — we define Δ as the set of *all* possible distributions on S. More explicitly,

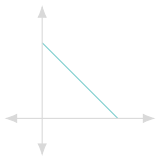
$$\Delta = \left\{ (x_i \mid i \in S) \mid x_i \geq 0, \sum x_i = 1 \right\}.$$

So Δ is the set of all possible probability distributions, where a probability distribution is a vector with nonnegative coordinates summing to 1. (This is where we're looking for a solution — we need π to belong to Δ, because we're only looking for probability distributions.)

Fact 2.10 — The set Δ is compact (meaning it is closed and bounded).

(We won't prove this.)

In two dimensions, it's just a line segment:



You can see it's bounded (it doesn't go to ∞), and it's also closed (e.g., there are no 'holes'). So it's compact.

Now we'll define a function $F: \Delta \to [0, 1]$ (meaning it takes a distribution and maps it to 0 or 1), by

$$F(\mu) = \max_i |\mu(i) - \mu P(i)|.$$

(This is called the L^∞ norm — I'm looking at μ and μP, and taking the maximum difference in coordinates.)

Fact 2.11 — This function F is continuous.

What does it mean to be continuous? It means that if we change π a bit, then $F(\pi)$ only changes a little bit. (This is because we're multiplying by a matrix, subtracting, taking absolute values, and taking maximums; and all of these are very continuous.)

Finally, we'll need one more fact (the last fact from calculus or real analysis or topology): Since Δ is compact and F is continuous, F obtains a minimum. (Every continuous function on a compact set has a minimum.) We want to show that this minimum is 0; because if π is the vector attaining this minimum and $F(\pi) = 0$, then we get $\pi = \pi P$.

So we need to show that the minimal value of F is 0. If we can show that, then we're done.

Remark 2.12. As an example if we haven't seen this kind of stuff before, if you look at $F(x) = x$ on the interval $[0, 1]$, this attains a minimum. But on $(0, 1)$, it doesn't attain a minimum — it can get arbitrarily close to 0, but it can't be 0. The point is $[0, 1]$ is closed, but $(0, 1)$ isn't. (This is a teaser for real analysis if you want to do it later.)

The logic is somewhat soft — we have a nice function that's continuous on a nice domain that's compact, and we want to show the minimum value is 0. And now we'll have a special argument due to Markov chains, showing that the minimum is smaller than any value you could think of — it's less than 1/10 or 1/1000 or 1/100000, and that'll mean it has to be 0.

Now, how can I argue that I can always find a distribution such that $\mu - \mu P$ is *small*? (This means one step of applying the Markov chain doesn't have much effect.) This is where we need the intuition about Markov chains; so far this has all been some soft math.

So I have a Markov chain on a finite state space, and I want you to find some distribution such that the difference between the chain right now and after I apply some step is going to be very small.

What we're going to do is we start with π_0 being any distribution; and then we're going to define

$$\nu_t = \frac{1}{t} \sum_{s=1}^{t} \pi_0 P^t$$

(exactly as we had before). The trick we're doing here is we're looking at a distribution *averaged over t steps*. Why is this good? Now let's look at $\nu_t - \nu_t P$; we get

$$\nu_t - \nu_t P = \frac{1}{t} \sum_{s=1}^{t} \pi_0 P^t - \sum_{s=1}^{t} \pi_0 P^{t+1}.$$

And everything cancels out, so we get

$$\frac{1}{t}(\pi_0 - \pi_0 P^{t+1}).$$

And if we look at the maximum entry of that, we get

$$\max_i |\nu_t(i) - \nu_t P(i)| = \frac{1}{2} \max_i \left|\pi_0(i) - \pi_0 P^{t+1}(i)\right| \leq \frac{1}{t}$$

(we have two probabilties, so the difference between them is at most 1).

But we could do this for *every* t, so we can get arbitrarily close to 0; and this implies the minimum is at most $1/t$ for every t, so the minimum has to be 0. And that's the end of our theorem. □

§3 February 11, 2025

§3.1 Stationary distributions

Today we'll continue talking about the long-term behavior of Markov chains. To recall what we finished with last time, we proved the following theorem:

Theorem 3.1
Every (homogeneous) finite Markov chaian has a stationary distribution π.

(We'll mostly only talk about homogeneous Markov chains, so we won't typically write 'homogeneous.')

Why can this be interpreted as a long-term behavior question about Markov chains? If you ever get to the stationary distribution π, then you always stay there — so if $X_0 \sim \pi$ (this notation means X_0 is drawn from π, or that X_0 is distributed according to π), then $X_t \sim \pi$ for all t. So that's sort of a long-term behavior — if you start from π, you're at π forever.

Here are some follow-up questions we'll be interested in today:

Question 3.2. Is there only one stationary distribution?

Question 3.3. Do we converge to a stationary distribution π in some sense (if we start with some distribution other than π)?

§3.2 Uniqueness of stationary distributions

There was at least one example we'd seen with more than one stationary distribution:

Example 3.4
In gambler's ruin, if you start from 0 you're always at 0, and if you start at n you're always at n. In fact, any combination of this works — if you start at 25% 0 and 75% n, then you'll always stay there. So there's more than one stationary distribution.

§3.2.1 Irreducibility

So let's start by trying to answer this question. For this, it'll be useful to have the following definition.

Definition 3.5. A finite Markov chain on state space Ω is called *irreducible* if for any two states $x, y \in \Omega$, there exists t such that $P^t(x, y) > 0$ (where P is the transition matrix).

Intuitively, this states that for any two states, you can get from one to the other. This condition is just a formal way of writing that — P^t is the t-step transition matrix, so $P^t(x, y) > 0$ says that you can get from x to y in t steps.

Example 3.6

The binary symmetric channel is irreducible if $0 < \varepsilon < 1$ — its matrix was

$$\begin{bmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix},$$

so we can even take $t = 1$ for all x and y (all entries of this matrix are already positive).

Example 3.7

The coupon collector is reducible, because you can't go from n coupons to 0 coupons, for example (it only goes in one direction).

It will turn out that this will play a role in understanding how many stationary distributions we have — the notion of irreducibility will come into play.

§3.2.2 Harmonic functions

But before we do that, we'll define a new notion that will also help us understand questions about uniqueness of the stationary distribution.

Definition 3.8. Consider a Markov chain with state space Ω. We say a function $F: \Omega \to \mathbb{R}$ is *harmonic* if for all $x \in \Omega$, we have

$$F(x) = \sum_y P(x, y)F(y).$$

So now we're looking at functions that map every state to a real number. And we call F harmonic if the value at where you are is the average of the values of where you can go to — where it's a *weighted* average, weighted according to the transition matrix.

First, what's the easiest example?

Example 3.9

Constant functions are always harmonic — if $F(x) = a$ for all $x \in \Omega$ (i.e., F is constant), then F is harmonic (because a is equal to the average of a bunch of values equal to a).

But we didn't define *harmonic* just as a fancy way of defining constants. So what about a harmonic function that isn't constant?

Example 3.10

In gambler's ruin with states $\{0, \dots, n\}$, the function $F(x) = \frac{x}{n}$ — the probability that you win from x — is harmonic.

To check this, we need to check that

$$F(x) = \frac{x}{n} = \frac{1}{2}\left(\frac{x+1}{n} + \frac{x-1}{n}\right)$$

for $1 \le x \le n-1$; and $F(0) = 0 = F(0)$, and $F(n) = n = F(n)$ (this is a somewhat annoying way of writing it, but the point is that 0 is the only place you can go from 0).

And you can also multiply this function by any constant:

Example 3.11

For gambler's ruin, any function of the form $F(x) = \frac{ax}{n} + b$ is harmonic.

Are there any others? The answer is no! Any harmonic function has to satisfy

$$F(x) = \frac{1}{2}(F(x-1) + F(x+1)) \quad \text{for } x \neq 0, n.$$

So we get a lot of linear constraints, and the only way to solve them is by these things.

Note that in gambler's ruin, two things happen: there is more than one stationary distribution, and there is a nonconstant harmonic function. If we believe this analogy, someone mentioned coupon collector before. How many stationary distributions are there for coupon collector? You start wherever you start, and go on and on, and eventually you have n coupons; so the only stationary distribution is just at n. So if we believe this analogy, there should be no nonconstant harmonic functions.

Example 3.12

For coupon collector, the only stationary distribution is $(0, 0, \ldots, 0, 1)$. And if F is harmonic, then

$$F(n-1) = \frac{n-1}{n}F(n-1) + \frac{1}{n}F(n),$$

which implies $F(n-1) = F(n)$. Similarly we'll get $F(n-2) = F(n-1)$, and $F(n-3) = F(n-2)$, and so on; this just says F has to be constant.

So there's something going on here — it seems like there are many harmonic functions when there's more than one stationary distribution, and when there's only one stationary distribution the only harmonic functions are constant. So far, we've seen three concepts — stationary distributions, harmonic functions, and irreducibility — and we want to connect them. The next thing we'll do is prove the following claim:

Claim 3.13 — If a finite Markov chain with transition matrix P is irreducible, then the only harmonic functions are constants.

So if the chain is irreducible — which says that you can get from every state to any other state — then there's only the boring harmonic functions (only constants are harmonic functions). This is consistent with what we've seen — the example where we did have a harmonic function was gambler's ruin, where you can't get from any state to any other (if you're at 0 you're stuck at 0, and if you're at n you're stuck at n).

Student Question. *Is coupon collector irreducible?*

Answer. It's not irreducible — you can't get from n to 0. This is not an if and only if — it says *if* it is irreducible, then you only have constant harmonic functions. The converse is false; hopefully we'll understand this better later in the class.

If you've heard of the maximum principle (e.g., from PDEs or the heat equation or complex analysis), this is probably the simplest proof that uses it.

Proof. We want to prove that under these conditions, only constant functions are harmonic. So consider a chain as in the statement (meaning that it's irreducible), and let $F: \Omega \to \mathbb{R}$ be harmonic. And we want to show that F is constant.

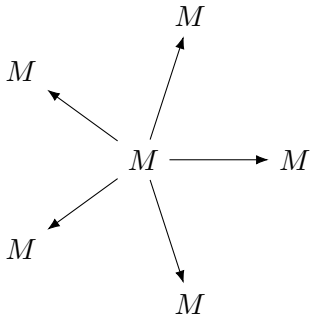
So here's what we're going to do (this is why it's called the maximum or minimum principle): Let $M = \max_{x \in \Omega} F(x)$ be the maximum value that F obtains, and let $x_0 \in \Omega$ be a point where this maximum M

is obtained (meaning that $M = F(x_0)$). So we're starting from this maximum, and we want to show that actually this maximum is the value of the function *everywhere*.

We'll draw a picture: We start from x_0, which is the maximum value. Now from x_0 we can go to a bunch of places. And we know $F(x_0) = \sum_y P(x, y)F(y)$. What do we know about the values $F(y)$? All of them are *at most* M — we know $F(y) \leq M$ for all y. So we have M on the left-hand side, and it's the average of values that are at most M. And if we have some value M that's the average of a bunch of other values less than or equal to M, what can we conclude? Everything actually in the weighted average has to genuinely be equal to M. If $P(x_0, y) = 0$ we don't get anything, but if $P(x_0, y) > 0$ then y is in the average; so we get

$$F(y) = M \quad \text{for all } y \text{ with } P(x_0, y) > 0.$$

So the value M is sort of percolating — it starts at x_0, and now everywhere we can go to with positive probability, we know the value there is also M (because M is a weighted average of all these values, where all the weights are positive).



And what are we going to do next? We proved that if we had the maximum at x_0, everyone connected to it had the same value. So now we can do it with each of these guys — we can repeat the argument. When we do this again and again, we'll eventually get

$$F(y) = M \quad \text{for all } y \text{ such that there exists } t \text{ with } P^t(x, y) > 0.$$

And since the chain is irreducible, we'll eventually get *everywhere* (if we start from x and keep going to neighbors) — we get all y's. So in other words, this means $F(y) = M$ for all $y \in \Omega$, so F is constant. □

That's the philosophy of the proof (if you've seen other uses of the maximum principle) — that the maximum *propagates* (if you start with the maximum point, then everywhere around it also has to be the maximum, and since you can get from everywhere to everywhere, this means M propagates to all points).

Student Question. *How did we know F is harmonic for P^2?*

Answer. We actually didn't need that. We started from x_0, and we know its value is M; and now all the guys it's connected to are also M. Now we use y as our new x_0; so the value at y is the average of all the points we can go to from y (using P, not P^2); so all those guys are also M. Where we used P^t is just to keep track of where we can eventually get to starting from x_0.

§3.2.3 Uniqueness for irreducible chains

Maybe this sounds a bit weird — we started talking about distributions, but now we're talking about functions. So now let's try to connect these.

Theorem 3.14

Consider a finite irreducible Markov chain with transition matrix P. Then P has a unique stationary distribution.

So if it's irreducible, then there aren't many stationary distributions; there's just one.

Usually in math, when we state something is unique, we need to prove there's at least one and at most one. Here we just need to prove there's at most one, because last class we proved that *every* finite Markov chain has at least one.

Proof. We already know there is *at least* one stationary distribution; so now we need to prove there's *at most* one. As a hint, the proof is linear algebra, and very neat; it's just two lines (or maybe four).

First, from Claim 3.13, we know that the only harmonic functions are constant. We can describe the harmonic condition in terms of a matrix times a vector — F is harmonic if $PF = F$ (where we write F as a column vector). Let's think about this for a second — we have

$$(PF)_x = \sum_y P(x,y)F(y).$$

So the equation for being harmonic is really just a linear algebra equation $PF = F$; or equivalently, as $(P - I)F = 0$.

So now can anyone translate the information that the only harmonic functions are constants to some linear algebraic information about $P - I$? We're trying to solve a linear algebraic equation $(P - I)F = 0$, and the solution is a 1-dimensional vector space; the fast way of writing this is to say that

$$\operatorname{rank}(P - I) = |\Omega| - 1$$

(it's 1 minus the maximum possible) — the dimension of the solution space is 1, and the rank of the matrix is the complement of that with respect to the size of the space. As a review of what rank is, when you solve linear equations, there's a bunch of dimensions you can look at. One is the dimension spanned by the rows of the matrix; this is actually the same as the dimension spanned by the columns. And when you solve an equation, you get the complement dimension to the size of the matrix — if the space of solutions has dimension 1, then this means the row or column span is $n - 1$.

Now why does this lead to something about stationary distributions? The tricky thing is that we can solve this equation by multiplying by column vectors, but we can also solve this equation multiplying by *row* vectors on the left — the dimension of the space of solutions to $\pi(P - I) = 0$ is *also* 1. That's the thing about the proof that's really linear algebraic — harmonic functions are about column vectors where we multiply the matrix on the left, and stationary distributions are row vectors where we multiply the matrix on the right; but the rank is the same either way.

But there's at least one stationary distribution. And if there were two, they'd have to be linearly independent (because of the condition their probabilities sum to 1 — the only way to be dependent is if one were a multiple of the other, and if e.g. one were twice the other and one had sum 1, the other would have sum 2). So we can't have two distributions π_1 and π_2 which are both stationary, i.e., such that $\pi_1(P - I) = \pi_2(P - I) = 0$. This implies there's a unique stationary distribution, which is what we wanted. □

So if the chain is irreducible, there's a unique stationary distribution.

This is a neat theorem, and maybe for some of us it's the most abstract we've seen, so we should take time to digest it.

Student Question. *Why is* $\operatorname{rank}(P - I)$ *the size of the state space minus* 1*?*

Answer. This is the rank-nullity theorem from linear algebra — there's one thing which is the dimension spanned by the rows, and there's another which is the dimension of the space of solutions, and the sum of these two things is the size of the matrix.

§3.3 Convergence to the stationary distribution

So we're making some progress in our goal for this morning — one thing we wanted to understand was when there is one stationary distribution, and we at least found a sufficient condition (irreducibility). The other thing we want to do is find when, in the long run, we converge to π. So the next thing we'll do is find a simple setting where we can do it.

Definition 3.15. A finite Markov chain is ergodic if there exists some t such that all entries of P^t are strictly positive.

So there's some power of the matrix where if I take it to this power, all the entries are positive.

What's the relationship between this and being irreducible — which is stronger or weaker? If there's some t such that all the entries are positive, then clearly in t steps you can get from any state to any other state. So ergodic immediately implies irreducible.

But we wouldn't define this if it was just equivalent to irreducibility. So are there examples which are irreducible but not ergodic?

Example 3.16

Suppose we have two states 0 and 1, where you switch states with probability 1. Then obviously you can get from any state to any other state; but from 0 to 0 you need an even number of steps, and from 0 to 1 you need an odd number. Here the matrix is

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

and every power of this matrix is either itself or the identity. So this is irreducible but not ergodic.

Theorem 3.17

If a finite Markov chain with transition matrix P is ergodic, then for any distribution ν, we have $\nu P^t \to \pi$ as $t \to \infty$, where π is the unique stationary distribution.

(We know there is a unique stationary distribution because ergodic implies irreducible.) Note that νP^t is where you are after t steps if you start with ν; so this says that if you run the chain for a long time, the distribution converges to π.

Before we prove this, let's write it in terms of probabilities — this implies that

$$\mathbb{P}[X_t = a \mid X_0 = b] \to \pi(a) \quad \text{as } t \to \infty$$

(no matter where we start, after t steps where t is large, I'm going to be very close to the stationary distribution π).

So that's sort of nice — it gives us convergence to the stationary distribution, which is something we wanted.

Today is a bunch of linear algebra proofs (later we'll see more probability proofs, but today is a bunch of linear algebra).

Proof. We're going to define a very boring matrix

$$\Pi = \begin{bmatrix} \pi \\ \pi \\ \vdots \end{bmatrix}$$

(where every row is π). What we're actually going to show is that $P^t \to \Pi$ as $t \to \infty$. Why is this good for us? Then if we look at some ν, we have $\nu P^t \to \nu\Pi$ (since we're just taking a linear combination of the rows). And Π is just the row vector π over and over again. So we have a row vector $\nu = (\nu_1, \dots, \nu_n)$ which is arbitrary; all we know is that $\sum \nu_i = 1$. And we're multiplying it by a matrix consisting of a bunch of rows π. When we multiply

$$\begin{bmatrix} \nu_1 & \nu_2 & \cdots \end{bmatrix} \begin{bmatrix} - & \pi & - \\ - & \pi & - \\ & \vdots & \end{bmatrix},$$

we're taking a linear combination of these rows, where the coefficients ν_i sum to 1; and all the rows are π, so we just get π back.

So if we can show $P^t \to \Pi$, then $\nu P^t \to \nu\Pi = \pi$.

So now we're just playing with matrices. We're going to have to use the fact that our chain is ergodic at some point. So how are we going to use it?

Claim 3.18 — Let k be such that all entries of P^k are strictly positive. Then we can write

$$P^k = \varepsilon\Pi + (1-\varepsilon)Q,$$

where Q is a transition matrix for some Markov chain and $\pi Q = \pi$ (i.e., π is stationary for Q).

(We can imagine we want to get many things done, so I assign 500 executive notes hoping things work out. I want to get Π, but I'm not going to get it after k steps; but we can push an ε-bit of it and hope it works out.)

Why can we do this? Let's think about it. All entries of P^k are strictly positive, so they're at least ε for some ε. So I can subtract ε from all of them and still get something positive. Instead of subtracting ε, I subtract $\varepsilon\Pi$; so I still get something positive. And since P^k originally had row sums 1 (and the same is true of Π), when we subtract $\varepsilon\Pi$ off and divide by $1-\varepsilon$, the row sums of Q are still 1.

In more detail:

Proof of Claim 3.18. Since all entries of P^k are strictly positive, there exists $\varepsilon > 0$ such that all are greater than ε (since we have finitely many values, so we can just take their minimum).

Now we're going to look at the matrix $B = P^k - \varepsilon\Pi$. This still has all entries nonnegative. And the sum of each row of B is $1-\varepsilon$, because the sum of each row of P^k is 1 and the sum of each row of Π is ε. And everything is positive, so I can write $B = (1-\varepsilon)Q$ where Q is a transition matrix (meaning its entries are nonnegative and each row sums to 1).

We still have to show $\pi Q = \pi$, so let's check that. We'll start from what we know — we know $\pi P^k = \pi$. And we can write this in a longer way as

$$\pi((\varepsilon\Pi) + (1-\varepsilon)Q) = \pi.$$

And we already showed that for *every* vector ν, we have $\nu\Pi = \pi$; so we get

$$\varepsilon\pi = (1-\varepsilon)\pi Q = \pi,$$

and now if we solve, we get $\pi Q = \pi$. □

But we're more ambitious; we don't just want to talk about vectors. So what can we say about ΠQ? This is just Π — because you have a bunch of rows, and for each one you multiply it by Q and get π.

And what about $Q\Pi$? Each row of Q is a probability vector, and when we look at a probability vector and multiply it by Π, we're getting a convex combination of rows that are all π, so we get π back. So $Q\Pi$ is also Π.

And this makes us happy, because we want to eventually get Π. Why? When we look at P^{2k}, we'll square this expression; and what will we get? There will be four terms — we'll get

$$P^{2k} = (1-\varepsilon)^2 Q^2 + \varepsilon(1-\varepsilon)\Pi Q + (1-\varepsilon)\varepsilon Q\Pi + \varepsilon^2\Pi^2 = (1-\varepsilon)^2 Q^2 + (1-(1-\varepsilon)^2)\Pi$$

(since $\Pi Q = Q\Pi = \Pi^2 = \Pi$). This makes us happy in terms of taking over the world — first we just got ε, but now we get something nearly 2ε, and we can keep pushing — we have

$$P^{tk} = (1-\varepsilon)^t Q^t + (1-(1-\varepsilon)^t)\Pi.$$

And as $t \to \infty$, $(1-\varepsilon)^t \to 0$ and $1-(1-\varepsilon)^t \to 1$, so this is going to converge to Π.

Are we done with the proof? Not exactly — we need to deal with powers that aren't multiples of k. We've shown that if we look at the subsequence P^k, P^{2k}, P^{3k}, …, we converge to Π. But what about things like P^{2k+5}? So now we'll deal with this. We have

$$P^{tk+j} = (1-\varepsilon)^t Q^t P^j + (1-(1-\varepsilon)^t)\Pi$$

(since $P\Pi = \Pi P = \Pi$ as well). And we just need to consider $1 \le j < k$.

And this P^j doesn't matter — we still have $(1-\varepsilon)^t \to 0$ and $1-(1-\varepsilon)^t \to 1$, so this still converges to Π as $t \to \infty$. □

§3.4 An example

Here's a question for us. Here's a chain we can talk about, which is similar to something we saw last time but not the same:

Example 3.19

Consider the chain where the state space is $\{0, \dots, k-1\}$, and the rule is

$$\mathbb{P}[X_{t+1} = X_t \pm 1 \bmod k] = \frac{1}{2}.$$

We can draw this chain on a circle; wherever we're standing, we toss a coin, and with probability $\frac{1}{2}$ we go one way and with probability $\frac{1}{2}$ we go the other.

This is irreducible — you can get from any state to any other, just by going around.

Is it ergodic? It's ergodic if k is odd, and it's not ergodic if k is even. When k is even, you're walking even–odd–even–odd–…. But if k is odd, it is actually ergodic — when you complete a cycle, you break this even-odd cycle, and if you do this enough times you can show it's ergodic.

So what this means is that if k is odd, then we do actually know $P^t \to (\frac{1}{k}, \frac{1}{k}, \dots)$ (since π is the stationary distribution).

What if k is even? This is a brainteaser for next class.

§4 February 13, 2025

§4.1 Review

We're talking about homogeneous finite Markov chains. We'll draw a big blob denoting all of them; and somewhere in between we have the irreducible ones; and somewhere inside we have the ergodic ones. So that's the picture we had so far.

What do we know for all of them? If we have *any* finite Markov chain, we know there *exists* a stationary distribution — there exists π such that $\pi P = \pi$. In the *irreducible* case, we know there exists one (because it's inside the previous circle), but in fact there exists a *unique* π such that $\pi P = \pi$. And for ergodic ones, we had the sort of signature of long-term behavior that whatever we start with, if we run for enough steps, eventually we converge to π (i.e., $\nu P^t \to \pi$ as $t \to \infty$, for any ν).

Today we'll begin the class by trying to understand what happens if we're irreducible, in the long-term behavior. And then we'll see, what happens if we're not irreducible, and there's maybe more than one stationary distribution — what can we say?

§4.2 Irreducible Markov chains

Here's what we'll prove for irreducible Markov chains that aren't necessarily ergodic. What we *really* want to have is that $P^t\nu$ converges to something; but we know we can't get this just from irreducibility. For the ergodic case, we got $P^t\nu \to \pi$. But this is too much to ask for if we're just talking about irreducible chains — for example, the chain that alternates between 0 and 1. So what can we hope for? We can hope that if we look at the *time-average*, we get this convergence.

Theorem 4.1

Let P be the transition matrix of an irreducible chain, and let ν be any distribution for X_0. Then

$$\frac{1}{t}\sum_{s=1}^{t} \nu P^s \to \pi \quad \text{as } t \to \infty,$$

where π is the unique stationary distribution.

This was one of the notions of convergence we initially talk about — we look at what happens after one step, two steps, three steps, and so on. And we average these and see, does this converge to something? And this converges to π even if the chain is irreducible but not ergodic.

Example 4.2

For the Markov chain that alternates between 0 and 1, νP^t doesn't converge — it alternates betwen $(1,0)$ and $(0,1)$. But when we average, we'll have

$$\frac{1}{t}\sum_{s=1}^{t} \nu P^s \to \left(\frac{1}{2}, \frac{1}{2}\right),$$

which is indeed the stationary distribution.

So whenever we have an irreducible chain and we look at these averages, we're going to get the stationary distribution.

Proof. The proof is pretty similar to some of the proofs we've seen before; it'll have a real analysis or calculus flavor.

Let's call this sequence $\mu_t = \frac{1}{t}\sum_{s=1}^{t} \nu P^t$, and suppose μ_t *doesn't* converge to π. In the ε-definition of convergence, this means we don't get arbitrarily close to π, i.e., there exists $\varepsilon > 0$ and infinitely many k's such that we're ε-far from π, meaning that $\|\mu_{t_k} - \pi\| > \varepsilon$. These things are vectors, so when we write $\|\bullet\|$ we actually mean in norm. It doesn't matter which norm we're using, so we'll use the L^2 norm (the length of the vector, defined by $\|\nu\| = \sqrt{\sum_a \nu(a)^2}$).

So we have π (our stationary distribution), and this sequence that does not converge to π. Maybe sometimes it goes to π and sometimes it goes far away; but because it doesn't converge to π, I can find a little ball around π and I know infinitely many of these guys are not in this ball.

Now if we remember the proof that every finite Markov chain has a stationary distribution, what we did there was that we defined a function F on distributions by

$$F(\rho) = \max_x |\rho(x) - (\rho P)(x)|.$$

(We're just measuring the difference between ρ and what happens when we apply P for one step.) And we said that F is continuous.

Now, before we looked at the set of all probability distributions Δ, and we said this was a compact set (closed and bounded). Now we're going to apply this to not Δ, but

$$\Delta \setminus \mathbb{B}(\varepsilon, \pi) = \{\rho \mid \sum_x \rho_x = 1,\ \rho_x \geq 0 \text{ for all } x,\ \|\rho - \pi\| \geq \varepsilon\}.$$

This is also closed and bounded, so it's also compact.

The picture is that we have this set of probability distributions, and π; and then we're taking a hole out (we're taking an open ball out, so that we get a closed set).

So we have a continuous function and a compact set. And we also know that

$$F(\mu_{t_k}) = \max_x \left| \left(\frac{1}{t_k} \sum_{s=1}^{t_k} \nu P^s - \frac{1}{t_k} \nu P^{s+1} \right)(x) \right|.$$

And what we said last time was that in this sum, most of the terms cancel — we have P^s and P^{s+1} — and we're only left with two terms, which means this is at most $1/t_k$.

So the μ_{t_k}'s are the guys that are going to lead to our contradiction. They live outside this ball, but on the other hand, $F(\mu_{t_k})$ is very small. So what we concludee from this is that the minimum of F on $\Delta \setminus \mathbb{B}(\pi, \varepsilon)$ (which exists by compactness) has to be 0. And this implies there exists some μ such that $\mu = \mu P$ and $\|\mu - \pi\| \geq \varepsilon$.

So what did we do here? We had a proof two classes ago saying there exists a stationary distribution. Now we're repeating the same proof, but instead of over all probability distributions, we're doing it over all probability distributions where we take out this little hole. We have a continuous function and a compact set, so by looking at the minimum we get a stationary distribution, but this time in the set with a hole.

And why does this make us happy? We found *two* stationary distributions — both π and μ — and this is a contradiction. (This is a proof by contradiction — we assumed our thing doesn't converge to π, and applied this somewhat convoluted logic of proving the same theorem again in a different setting to get a contradiction.) □

So we've amde some progress in our picture — not only is there a unique π such that $\pi P = \pi$, but there's also some long-term behavior — we know

$$\frac{1}{t}\sum_{s=1}^{t} \nu P^s \to \pi,$$

in the case where the chain is irreducible but not ergodic.

Student Question. *Did we use anything about irreducibility except the uniqueness of stationary distributions?*

Answer. No, we only used uniqueness.

Student Question. *How did we get that the minimum of F is 0?*

Answer. We had these examples μ_{t_k} in our domain, and we showed that $F(\mu_{t_k}) < 1/t_k$. But we have infinitely many t_k's, and $1/t_k \to 0$ as $k \to \infty$. And the minimum has to be nonnegative and less than all these numbers, so it has to be 0.

Student Question. *What does $\Delta \setminus \mathbb{B}(\varepsilon, \pi)$ mean?*

Answer. We used Δ for the set of *all* possible distributions — if our Markov chain has 5 states, this would be all 5-dimensional vectors where each coordinate is nonnegative and their sum is 1 (because we're doing probability). In general, if we have n states, we're looking at vectors in n dimensions with nonnegative coordinates summing to 1. So

$$\Delta = \{(\mu_1, \dots, \mu_n) \mid \mu_i \geq 0 \text{ for all } i, \sum \mu_i = 1\}.$$

And we write

$$\mathbb{B}(\pi, \varepsilon) = \{\mu \mid \|\mu - \pi\| < \varepsilon\}.$$

When we're doing $\Delta \setminus \mathbb{B}(\pi, \varepsilon)$, we get the original conditions from Δ, but another condition — the complement of the one in $\mathbb{B}(\pi, \varepsilon)$ — so we get

$$\Delta \setminus \mathbb{B}(\pi, \varepsilon) = \{\mu \mid \mu_i \geq 0 \text{ for all } i, \sum \mu_i = 1, \|\mu - \pi\| \geq \varepsilon\}.$$

§4.3 Reducing Markov chains

Now we want to understand what's happening on the interface between irreducible and reducible chains — what happens if the chain is not necessarily irreducible?

It'll be useful to recall the Gambler's Ruin:

Example 4.3 (Gambler's Ruin)

Our state space is $\{0, \dots, n\}$. We have $P(0,0) = P(n,n) = 1$, and for $0 < x < n$, we have

$$P(x, x \pm 1) = 1/2.$$

This will be our example when we talk about some definitions we'll see for the reducible case. Now we'll see some definitions and theorems for chains that might be reducible.

Goal 4.4. 'Break' any chain into irreducible ones.

Things that are irreducible, you can't reduce them any further. So if you look at reducible chains, we'd like to reduce them to irreducible ones. So the goal is to come up with some definitions and theorems that will let us break any chain into irreducible ones.

§4.3.1 Essential states

So we'll come up with some definitions, and then think about them in the context of gambler's ruin.

Notation 4.5. For two states x and y, we write $x \to y$ if there exists $t > 0$ such that $P^t(x, y) > 0$.

In words, this means you can get from x to y in a finite number of steps.

Definition 4.6. We say x and y *commute* if $x \to y$ and $y \to x$; we write this by $x \leftrightarrow y$.

Example 4.7

In gambler's ruin, the states 1 to $n-1$ all commute with each other; 0 commutes with itself, and n commutes with itself.

(It's not always the case that a state commutes with itself, so here we say that explicitly.)

This makes sense — if you have 5 dollars you can get to 3, and vice versa; but if you have n dollars, then you have to stay at n.

Claim 4.8 — The relation $\leftrightarrow$ is an equivalence relation.

As a reminder, what does it mean to be an equivalence relation? We need to show that $x \leftrightarrow y$ and $y \leftrightarrow z$ implies $x \leftrightarrow z$. (In principle, we also have to show it's symmetric, but here that's clear.)

Proof. In fact, we know more — if $x \to y$ and $y \to z$, then there exist s and t such that $P^s(x, y) > 0$ and $P^t(y, z) > 0$; and then we can conclude $P^{s+t}(x, z) > 0$ (if I can go in s steps from x to y and in t steps from y to z, then in $s + t$ steps I can get from x to z). So this means $x \to z$. And we can do the same for the arrow going in the other direction. □

Student Question. *If there's a state in the Markov chain that only emits probability but doesn't accept it, does that violate things?*

Answer. Good question, we have to be a bit careful. The standard definition of equivalence relation requires that everything is equivalent to itself; we won't require that here, so this is actually a slightly weaker thing.

Definition 4.9. We say that a state x is *essential* if whenever $x \to y$, we also have $y \to x$ (i.e., $x \leftrightarrow y$).

So states are essential if they have the very special property that if you can get from x to y, then you can also get back from y to x.

Example 4.10

In gambler's ruin, the essential states are 0 and n.

We already see that in gambler's ruin, 'essential' is a good definition because that's where we end up. That's the goal of this definition, and what we'll try to prove.

We'll need one or two more simple claims.

Claim 4.11 — If $x \to y$ and x is essential, then y is essential.

Proof. What we need to show is that if $y \to z$, then $z \to y$. so let's suppose that $y \to z$. That means $x \to y \to z$, so $x \to z$. But since x is essential, that means $z \to x$. And we know $x \to y$, so $z \to x \to y$; this means $z \to y$. So we've showed that if $y \to z$, then $z \to y$. □

Student Question. *In Gambler's Ruin, the two essential states are* 0 *and* n*; does this say we can go from* 0 *to* n*?*

Answer. No — the essential states don't necessarily need to go to each other. In gambler's ruin, you can't go from 0 to n or n to 0, so the definition doesn't apply. For example, when $x = 0$, the statement says if you can go from 0 to y, then you can go from y to 0. But the only place you can get to from 0 is 0 itself, so $y = 0$ is the only relevant case.

Here's a slightly more complicated example:

Example 4.12

As in gambler's ruin, we'll have states $0, 1, \ldots, n-1$. But then after $n-1$ we'll have two states n_{happy} and n_{sad} (since rich people also get sad), where from $n-1$ you go to each with equal probability. And maybe these two states n_{happy} and n_{sad} talk to each other (you stay at one with probability $1/2$ and move to the other with probability $1/2$). Then both of these are essential states, and they talk to each other. So you don't end up just in one state, but rather in a sort of small Markov chain thaat runs on its own.

Let's prove one more claim of this flavor. So far, what we've seen is that if $x \to y$ and x is essential, then y is essential. In particular, if $x \leftrightarrow y$, then they're both the same with respect to essentialness.

Claim 4.13 — Let S be an equivalence class of essential states. Then $P|_S$ (i.e., the matrix P restricted to S) is a Markov transition matrix.

So we have some essential states and some equivalence classes of essential states.

Definition 4.14. We write $P|_S$ to denote the matrix P restricted to S, i.e.,

$$P|_S(x, y) = P(x, y) \quad \text{only for } x, y \in S.$$

So we're looking at only the rows and columns corresponding to S, and throwing away everything else. And the claim is that if I restrict to an equivalence class of essential states, I get a smaller Markov chain on its own.

Proof. What do we need to prove in order to show that this is a transition matrix for a Markov chain? We already know the entries of $P|_S$ are nonnegative (because this is true of the original), so all we need to show is that its rows sum to 1.

Why? We know all the things you can get to from an essential state are also essential and are equivalent to you — so if we fix $x \in S$, then we know $\sum_{y \in \Omega} P(x, y) = 1$ (where the sum is over all y, not just the ones in S). But if $P(x, y) > 0$, then in fact $x \to y$, so $x \leftrightarrow y$, so $y \in S$. So we're summing the same thing as before, we've just removed some 0's — everything else is going to be the exact same. □

To draw a picture, we have a big matrix P (a big square), and $P|_S$ is a smaller square. And what this claim is saying is that if we look at a row in S, then all the numbers in the row outside our small square are 0's.

Example 4.15

In gambler's ruin, we get the very fascinating Markov matrices on one item — for $S = \{0\}$ we get $P|_S = [1]$, and likewise for $S = \{n\}$ we get $P|_S = [1]$.

Example 4.16

In our variant, we have the same situation for $S = \{0\}$. Meanwhile, for $S = \{n_{\text{happy}}, n_{\text{sad}}\}$, $P|_S$ is going to be independent coin tosses, i.e.,

$$P = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}.$$

We'll now add some words to this claim:

Claim 4.17 — Let S be an equivalence class of essential states. Then $P|_S$ (i.e., the matrix P restricted to S) is a Markov transition matrix that is irreducible.

This is just because in an equivalence class, we can go from any point to any other point.

§4.3.2 Finding essential states

So we've found some reduction already — maybe we find one class over here, and another over there, and so on; and if you can find these equivalence classes, you're sort of happy. But we haven't yet shown what happens if we run the chain, or how we find them.

So now we want to prove something interesting about this.

Claim 4.18 — Every finite Markov chain has at least one essential state.

We want to reduce every Markov chain to these irreducible chains, so it'll be good if we can find essential states; this claim says we can always do so.

Proof. Let's start with some y_1. If y_1 is essential, then we're done. Otherwise, what should we do? If y_1 is not essential, then there is some y_2 such that $y_1 \to y_2$ but $y_2 \not\to y_1$. (If you go to only yourself with probability 1 then you're essential; so you have to go to some y_1, and since you're not essential there has to be y_2 like this.)

And now we can continue this. What's the idea? We're starting from y_1; and then we have to go to some other state y_2 where there's no way to go back to y_1. Now if y_2 is essential, we can ask this question again. We already know you don't go back to y_1; if y_2 went back to itself with probability 1 it'd be essential. So there has to be some state y_3 with $y_2 \to y_3$ but $y_3 \not\to y_2$. And y_3 also can't go back to y_1 — if it did, then we'd have $y_2 \to y_3 \to y_1$, but we said we can't go from y_2 to y_1. So now y_3 can't go either to y_2 or to y_1.

And what happens if we keep doing this? If you can't go back to any of the previous states, then it's a finite Markov chain, so we'll have to stop this at some point, and we'll be at an essential state. In other words, since we only have finitely many states, we'll end at an essential state. □

So we're making progress — we have these essential equivalence classes, and when we restrict to them we have a smaller chain that's irreducible. We know they *exist*; now we want to understand a bit more about how we get into them.

So the next claim is that eventually you always hit these states.

Claim 4.19 — For every finite Markov chain and every state x, we have

$$\mathbb{P}_x[X_t \text{ is eventually essential}] = 1.$$

This means for any state x, if we start from x, we may spend some time in nonessential states, but we're eventually going to hit the essential states. For example, in gambler's ruin, we might spend some time going left and right, but eventually we're going to end up at 0 or n.

Proof. We'll write $h(x) = \mathbb{P}_x[X_t \text{ is eventually essential}]$. What can we say about this function $h(x)$? For one thing, it's between 0 and 1, because it's a probability.

(Note that $h(x)$ has no dependence on t — it means there is some t such that X_t is essential.)

Note that the maximum of h is 1 — if x is essential then $h(x) = 1$, and we have some essential states, so $h(x) = 1$.

Anything else? We claim that h is harmonic. Why? I'm sitting at some state right now, and I want to know, what's the chance I'm eventually essential? On the next step I'll be at some y, and I average the probabilities I'll eventually be harmonic over those steps.

Now we want to understand the *minimum*. So we let $a = \min_x h(x)$; and we want to show that $a = 1$.

How are we eventually going to show that $a = 1$? We'd like to show that h is constant, using the same idea from before; but that was for irreducible chains, so we have to apply the same logic but not the direct result.

So let's assume the minimum is attained at $h(y_1) = a < 1$. This means y_1 is *not* essential. And what does that mean? It means there exists y_2 such that $P(y_1, y_2) > 0$ but $y_2 \not\to y_1$. And $h(y_1)$ is an average of a bunch of things, one of which is $h(y_2)$; because $h(y_1)$ is the minimum, it has to be that $h(y_2)$ is also a. In other words, $h(y_1) = \sum_z P(y_1, z)h(z)$, and one of these terms is $h(y_2)$; and this implies $h(y_2) = a$.

And I can continue to get y_3 (with $h(y_3) = a$), then y_4, and so on. But this is a contradiction — eventually I'm going to get that y_i is essential, but then $h(y_i) = 1$, which is a contradiction. □

So this is the same logic from the proof that essential states exist. I start from the worst state — the one with the lowest probability of going to an essential state — which in particular is not essential. Then I can apply the same logic from the previous proof to say I can go from y_1 to y_2 but not back. And now the fact that h is harmonic means $h(y_1)$ is the average of a bunch of terms, one of which is also $h(y_2)$; but $h(y_1)$ is the minimum, so I also get $h(y_2) = a$. And then I can continue.

We'll finish with one last claim, which we will not prove (it's essentially a similar proof).

Claim 4.20 — Let S be an equivalence class of essential states. Then the function

$$h(x) = \mathbb{P}_x[X_t \text{ will eventually end in } S]$$

is harmonic; $h \equiv 1$ on S; and $h \equiv 0$ on any equivalence class of essential states other than S.

So now I'm asking about a *specific* equivalence class of essential states, rather than all of them. Of course, if I start from that equivalence class, I'm going to stay there. If I start somewhere else I'll stay in that somewhere else. And in general this is going to be a harmonic function. (We essentially did the proof, so we're not going to repeat it again.)

What's the high-level message? Reducible chains are more complicated. But what really happens is for a finite amount of time you're hanging around; and after this you'll end up in one of these equivalence classes of essential states. And then you're in a finite irreducible Markov chain, so whatever we proved about those we can do there. There's this initial phase; so at the beginning, maybe you go from kindergarden to elementary school to high school to college, but then you either become a banker or work for the government or become a doctor, and then you stay there for the rest of the life. So those are the essential states. At the beginning there are maybe some choices you make that affect your probability of landing in each of these states.

Next week there is a weird thing because of the holiday, so the next class is next Thursday.

§5 February 20, 2025

It's been a week since we last met, so let's recall what questions we're talking about.

§5.1 Questions

In general, we're talking about the long-term behavior of Markov chains. As an example of what we did in the last few lectures, we looked at questions like:

> **Question 5.1.** If we start with distribution ν and run for t steps, when do we converge with π — i.e., when do we have $\nu P^t \to \pi$?

This is a question about matrices and vectors, and linear algebra or calculus.

Today we're going to talk about *random* quantities and their limits. For example, in the first lecture, we defined a random vector F_t, where for every state we looked at the fraction of time we spent in that state — i.e.,

$$F_t(x) = \frac{1}{t}\#\{1 \le s \le t \mid X_s = x\}.$$

This is a random quantity — it measures, for each individual sample of my Markov chain, how much time I spent in each state.

Usually people don't use this notation, though this is how we talked about it in the first lecture. More generally, the setup we'll talk about today is:

> **Question 5.2.** Given a finite irreducible Markov chain with state space Ω, and some function $f : \Omega \to \mathbb{R}$, what can we say about
>
> $$Z_t = \frac{1}{t}\sum_{s=1}^{t} f(X_s)?$$

(We're only going to talk about finite irreducible Markov chains today.)

So we have a function $F : \Omega \to \mathbb{R}$ that measures where we are. (You could even think of a function taking vector values, e.g., $F : \Omega \to \mathbb{R}^d$.) Let's explain what's happening with Z_t — I'm running my Markov chain, and I want to summraize what's happened up to time t. So for that, I measure the Markov chain — states are some abstract thing, so I measure them by applying F. This gives me a number (or vector), and I can average this over all times. (States are not numbers, so I can't average them.)

How does f match up to F? For example, we can take $F : \Omega \to \mathbb{R}$ to be $f(y) = \mathbf{1}(y = x)$. Then when we look at Z_t, it'll be the fraction of time spent at state x (up to time t). To match the *full* vector F, we take $f : \Omega \to \mathbb{R}^{|\Omega|}$ where $f(y) = (\mathbf{1}(y = x))_{x\in\Omega}$ — so for each state in Ω, we write a 1 in the coordinate corresponding to that state, and 0's everywhere else. Then the corresponding Z_t is another way of writing our F_t from before.

§5.2 Philosophy

There are only two tools in probability we know about analyzing random vectors like this (from e.g. 18.600) — the law of large numbers and the central limit theorem. So we'd somehow like to apply one of those to say something about Z_t.

We'll first explain the philosophy of what we're trying to do, and then we'll write it down.

> **Goal 5.3.** Reduce to independent (or i.i.d.) random variables, and use the strong law of large numbers.

So that's the idea of what we want to do. It kind of sounds tricky because in Markov chains, there's some dependence — where I am next depends on where I was before. The tool for how to get i.i.d. random variables is something called *excursions*.

The picture is like the following: let's draw our finite state space Ω, and for concreteness let's imagine we start at some state z. What happens I have no idea; but because the chain is irreducible, I know eventually it's going to go back to z. And that's going to be one excursion — when I go from z all the way back to z.

So that's Excursion 1. After that, I'm going to start from z again, and then I'm going to have an *independent* excursion. That's sort of a complicated random variable — it's a path of a Markov chain, and it could be any number of steps — but I go some number of steps and generate my random variable. And then I go again and get a third excursion, and I continue doing this over and over again.

And because these are i.i.d. random variables, I can apply the strong law of large numbers to them.

So this is the tool we're going to use; now we have to formalize the definition adn see what it is.

§5.3 Excursions

So the tool is going to be excursions. To make things simpler, we're going to fix a starting state $z \in \Omega$, and we're going to define a bunch of random variables.

First, we're going to define the times when we're at z. These are defined inductively:

Definition 5.4 (Times of return to z)**.** We define $T_0 = 0$. Then we inductively define

$$T_i = \min\{t > T_{i-1} \mid X_t = z\}.$$

So we're just looking at when we're at z — the first time we're at z is time 0 by definition; then I walk and walk, and T_1 is the second time I'm at z, then T_2 is the third time, and so on.

I'm also interested in the lengths, or gaps, between these times:

Definition 5.5. We define $U_i = T_i - T_{i-1}$.

This is how long it took me to get to z from the last time I was at z.

That's one interesting parameter — how much time it takes me to get back. But we also want how much time is spent at each *state*.

Definition 5.6. We define $U_i(x) = \#\{s \mid X_s = x,\ T_{i-1} < s \le T_i\}$.

This is similar to U_i, but I'm only looking at the times when I'm at state x. So we look at times s when we're at x, but only on the ith excursion, meaning that $T_{i-1} < s \le T_i$. So this is a mouthful, but it's the number of times that x gets visited on the ith excursion.

Student Question. *What is an excursion?*

Answer. Prof. Mossel first learned the word 'excursion' in math. It means you go for a hike or something and then come back to where you were. With the Markov chain, we think about running it forever, and breaking it into paths. Think about the way Prof. Mossel walks in class as a Markov chain. His base is behind the table; in the next minute he stays here, so he goes from z back to z. Then maybe he wants to draw a picture on the right board, so he walks, walks, drops a chalk, picks it up, goes on a long excursion, and eventually comes back. That's his second excursion. In the second excursion he visited a lot of states; in the first one he just visited the one state z.

So now we have this random variable $U_i(x)$ telling me how many times I visited state x on excursion i.

Example 5.7

We have $U_i(z) = 1$ — the excursion ends the first time I go back to z; so it doesn't matter how long it is, once we return to z we stop. So we visit z exactly once on the excursion.

Example 5.8

We have $\sum_{x\in\Omega} U_i(x) = U_i$ — the total length of the excursion is the total time I spend at all states.

§5.4 Applying the strong law of large numbers

Now, I want to apply the law of large numbers, so what do we need for that?

Claim 5.9 — For each $x \in \Omega$, the random variables $U_i(x)$ (over $i = 1, 2, \dots$) are i.i.d.; similarly $\{U_i\}_{i=1}^{\infty}$ are also i.i.d.

This is clear — when we start from z, how much time we spend at x is going to be the same distribution from each excursion. Because it's a Markov chain, when I start from z I don't care what happened before; we're sampling from the same distribution each time.

The other thing we need is when we want to apply the strong law of large numbers, we need to know something about expectations.

Claim 5.10 — We have $\mathbb{P}[0 \le U_i(x) \le U_i] = 1$.

(This is clear — these are all clearly nonnegative, and clearly $U_i(x) \le U_i$ because $U_i = \sum_x U_i(x)$.)

Claim 5.11 — We have $\mathbb{E}[U_i(x)] \le \mathbb{E}[U_i] < \infty$.

We don't really care whata this number is, but we want it to be finite — for the law of large numbers to apply, we need the expectation to exist (and not be infinite).

What do we need to use for this proof — can you find a Markov chain where $\mathbb{E}[U_i]$ wouldn't be finite?

Example 5.12

As an example, we can take gambler's ruin on $\{0, 1, 2\}$ with $z = 1$. What's the expected time of going back to 1 (i.e., $\mathbb{E}[U_1]$)? This is infinite, because I *never* go back to 1 (so all this excursion story doesn't work).

What's bad about this example in terms of the conditions we talked about? This example is not irreducible.

So if we want to prove this claim, obviously we have to use the fact that it's irreducible.

Why does this help us? The idea is that we can get from any state to any other state, so we should eventually return to z. But we have to be careful — the random walk on $\mathbb{Z}$ will return to 0 infinitely many times, but the expectation here is actually infinite.

Proof. The idea is that because we're finite and irreducible, there exists some M and some $\varepsilon > 0$ such that for every $y \in \Omega$, we have

$$\mathbb{P}_y[\text{get to } z \text{ in} \le M \text{ steps}] \ge \varepsilon.$$

(In words, if we start from y, we'll get to z in at most M steps with probability at least ε.) (The definition of irreducibility just says that all these quantities are *positive*; but we have finitely many states, so we can just take the minimum, and that'll be our ε.)

And what does this tell us about

$$\mathbb{P}_y[\text{don't get to } z \text{ in} \leq nM \text{ steps}]?$$

The probability we don't get there in M steps is at most $1-\varepsilon$, so the probability we fail n times is at most $(1-\varepsilon)^n$ (because we're doing this n times), which means we get

$$\mathbb{P}_y[\text{don't get to } z \text{ in} \leq nM \text{ steps}] \leq (1-\varepsilon)^n.$$

And why is this useful? Since U_i is a positive random variable, we have

$$\mathbb{E}[U_i] = \sum_{k=1}^{\infty} \mathbb{P}[U_i \geq k].$$

And this is an exponentially decreasing sum (if we break into blocks of size M), so it's finite — more precisely, this is bounded by

$$M(1 + (1-\varepsilon) + (1-\varepsilon)^2 + \cdots) \leq \frac{M}{\varepsilon} < \infty. \qquad \square$$

Student Question. *Why does that first probability statement follow from irreducibility?*

Answer. Let's recall what irreducibility says — irreducibility says that for every state $y \in \Omega$, there exists some number M_y such that $P^{M_y}(y,z) > 0$. Let's call this probability $P^{M_y}(y,z) = \varepsilon_y$. Now what we're going to do is define

$$M = \max_y M_y \quad \text{and} \quad \varepsilon = \min_y \varepsilon_y.$$

And then we get the statement that we have — it means for every state, if I go for at most M steps, I have probability at least ε of going back to z. This is where we use finiteness — if it's not finite then we can't take a maximum or minimum.

Now we already want to apply the strong law of large numbers. So what do we get?

Claim 5.13 — For every $x \in \Omega$, we have

$$\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N} U_i(x) \to \mathbb{E}[U_i(x)] \text{ as } n \to \infty\right] = 1.$$

What does this mean? I have all these excursions — I have Excursion 1 and ask how much time I spend at x, and then I have Excursion 2 and ask how much time I spend at x, and so on. So this is the average time you spend along each excursion at x; and this says that time converges to $\mathbb{E}[U_i(x)]$. That's the strong law of large numbers.

Similarly, we have the same thing for the U_i's themselves.

Claim 5.14 — We have $\mathbb{P}[\frac{1}{N}\sum_{i=1}^{N} U_i \to \mathbb{E}[U_i] \text{ as } N \to \infty] = 1$.

So the average length of all your excursions is just the expected length of each.

Now, what can we say about the quantity from the beginning of the lecture, the fraction of time we spend at x?

Claim 5.15 — For all $x \in \Omega$, we have

$$\mathbb{P}\left[\frac{1}{t}\#\{0 \le s < t \mid X_s = x\} \to \frac{\mathbb{E}[U_i(x)]}{\mathbb{E}[U_i]}\right] = 1.$$

This makes sense — the fraction of time I spend at x (in total) should be the expected time I spend at state x in the excursion, divided by the total length of what the excursion should be.

This is sort of cool — it's already telling us that this *random quantity* on the left-hand side converges to a *number* (the expected values of $U_i(x)$ and U_i are numbers).

Proof. This is just algebra, though it's maybe a bit confusing because of all the random quantities.

We want to understand what happens at time t, so we'll break it into excursions. So we'll define $N(t)$ as the number of excursions we've completed up to time t, i.e.,

$$N(t) = \max\{i \mid T_i \le t\}$$

(the greatest i such that the ith excursion ended by time t). The annoying issue here is that we're looking at some time t, and there's no reason I just completed an excursion — I might be at a time when I'm in the middle of an excursion, and I still have to understand what's happening there.

So we're looking at

$$\frac{\#\{0 \le s < t \mid X_s = x\}}{t},$$

and I want to write this in terms of excursions. I can't do this exactly, but I can upper-bound and lower-bound it in terms of excursions. I can count the number of times I visited state x *up to* the time I finished my last excursion, for a lower bound; and I can imagine completing the excursion to get an upper bound. So for the numerator, we get

$$\sum_{i=1}^{N(t)} U_i(x) \le \#\{0 \le s < t \mid X_s = x\} \le \sum_{i=1}^{N(t)+1} U_i(x).$$

And we can do the same thing with the denominators; so we get

$$\frac{\sum_{i=1}^{N(t)} U_i(x)}{\sum_{i=1}^{N(t)+1} U_i} \le \frac{\#\{s \mid X_s = x\}}{t} \le \frac{\sum_{i=1}^{N(t)+1} U_i(x)}{\sum_{i=1}^{N(t)} U_i}$$

(this ugly algebra is because I'm not always at the end of an excursion; I might be in the middle, and then I have to compare what's happened in the beginning and what's happened in the end).

How do I make these quantities even more ugly in a way that makes it transparent what it converges to? (It's ugliness for the sake of beauty, so we can do it, maybe.) We want to use that these averages converge to the expected values, so we'll divide by $N(t)$ and $N(t)+1$; then we get

$$\frac{\frac{1}{N(t)}\sum_{i=1}^{N(t)} U_i(x)}{\frac{1}{N(t)+1}\sum_{i=1}^{N(t)+1} U_i} \cdot \frac{N(t)}{N(t)+1} \le \frac{\#\{s \mid X_s = x\}}{t} \le \frac{\frac{1}{N(t)+1}\sum_{i=1}^{N(t)+1} U_i(x)}{\frac{1}{N(t)}\sum_{i=1}^{N(t)} U_i} \cdot \frac{N(t+1)}{N(t)}.$$

Now we have this very beautiful inequality; we like all these terms. So let's try to look at this term-by-term.

We know the length of each excursion is finite; this implies

$$\mathbb{P}\left[\frac{N(t)}{N(t)+1} \to 1 \quad \text{and} \quad \frac{N(t+1)}{N(t)} \to 1\right] = 1.$$

So these annoying terms that we added for beauty are going to converge to 1, and we don't care about them.

For the rest of the terms, we can just write what the limit is — we have

$$\frac{1}{N(t)}\sum_{i=1}^{N(t)} U_i(x) \to \mathbb{E}[U_i(x)]$$

by the strong law of large numbers, and

$$\frac{1}{N(t)+1}\sum_{i=1}^{N(t)+1} U_i \to \mathbb{E}[U_i]$$

(and the same is true for the other side). (This is a bit informal; what we really mean is that the probability these convergences happen is 1.)

And if these extra factors become 1 and the two terms on the end are the same, we conclude what we want — we get that

$$\frac{\#\{0 \le s < t \mid X_s = x\}}{t} \to \frac{\mathbb{E}[U_i(x)]}{\mathbb{E}[U_i]}$$

with probability 1. And that's the end of the proof. □

§5.5 Convergence to the stationary distribution

So we're pretty happy — we showed that tis random quantity actually converges to a number. But is this the number we *wanted* it to converge to? We never talked about U_i's until this lecture — there were some other numbers we cared about. We really wanted this to converge to $\pi(x)$.

So we're sort of happy, because we got that this random quantity

$$Z_t = \frac{\#\{0 \le s < t \mid X_s = x\}}{t}$$

converges to a number. But we're sort of unhappy, because we wanted that number to be $\pi(x)$ (where π is the stationary distribution — we know there's a unique one because the chain is irreducible).

So we're somewhat happy and somewhat unhappy — we got a number, but we still haven't shown that the thing converges to π.

Let's more formally state two things we know, and then we'll try to connect them.

Question 5.16. What do we know about $\mathbb{E}_z[Z_t]$? What does it converge to as $t \to \infty$?

This is *not* a random quantity — Z_t is a random quantity, but its expected value is just a number.

This we want to connect to the linear algebra we saw before — Z_t is the fraction of time up to t that I spend in s. So

$$\mathbb{E}_z[Z_t] = \frac{1}{t}(\nu(I + P + \cdots + P^{t-1}))(x)$$

where ν is the vector that has 0's everywhere except on z, where it has a 1. So νI is where I am at the start; and then νP is where I am at the next step; and so on. And I want to know how much time I spend at x, so I look at the xth coordinate, and divide by t. (Everything we did before was about expectations.)

And last class, we showed that this thing converges to π — we showed that

$$\frac{1}{t}(\nu(I + P + \cdots + P^{t-1})) \to \pi,$$

and therefore we have

$$\mathbb{E}_z[Z_t] \to \pi(x)$$

(because everything inside converges to π, and we're taking its xth coordinate).

So what do we know? On one hand, we know $\mathbb{E}_z[Z_t] \to \pi(x)$ as $t \to \infty$. On the other hand, we know that

$$\mathbb{P}\left[Z_t \to \frac{\mathbb{E}[U_i(x)]}{\mathbb{E}[U_i]}\right] = 1.$$

So these are the two facts we know — the expected value of a sequence of random variables converge to $\pi(x)$, and on the other hand, the random variables themselves converge to a *constant*. So you would want to conclude that these two constants are teh same, i.e., that

$$\pi(x) = \frac{\mathbb{E}[U_i(x)]}{\mathbb{E}[U_i]}.$$

(The random variables themselves, with probability 1, converge to some constant; and their expectations converge to some other constant; and what we want to argue is that these constants are the same.)

We will use the following fact, which sounds trivial but we can't prove it because you need measure theory to do so.

Fact 5.17 — If $0 \le Z_t \le 1$ for all t and $\mathbb{P}[Z_t \to a] = 1$ (where a is some constant), then $\mathbb{E}[Z_t] \to a$.

If you know measure theory this is simple. But the point is that combining all these facts, we get that

$$\frac{\mathbb{E}[U_i(x)]}{\mathbb{E}[U_i]} = \pi(x).$$

Because a sequence of random variables have expectations converging to $\pi(x)$, and the random variables themselves converge to $\mathbb{E}[U_i(x)]/\mathbb{E}[U_i]$, so these two numbers have to be the same.

§5.6 The ergodic theorem for Markov chains

What this gives us is the following theorem. (It's pretty long, but we've essentially already proved it.)

Theorem 5.18 (Ergodic theorem for Markov chains)

Consider a finite irreducible Markov chain with state space Ω, and a function $f: \Omega \to \mathbb{R}$. Then for all $z \in \Omega$, we have

$$\mathbb{P}_z\left[\frac{1}{t}\sum_{s=0}^{t-1} f(X_s) \to \pi(f)\right] = 1,$$

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

Maybe this doesn't look like what we just proved, but in fact it is.

Like in the usual strong law of large numbers, there's two types of averaging. There's the averaging with $\pi(f)$ where we just average f according to π; that's the easy type of averaging. The more sophisticated kind of averaging is what's happening on the left — now we have a random process where we're starting our Markov chain at 0 and compute $f(X_0)$, then I walk and compute $f(X_1)$, and so on. And I average all these random values; and with probability 1 they're going to converge to the simpler average $\pi(f)$. (The left is a random quantity; the right is always a number.)

Why do we say we've already proved this theorem? As an easier warmup, for which kinds of functions have we already proved it? We've already proved this statement (which we'll call $(**)$) if f is an indicator of a

state, i.e., $f(x) = \mathbf{1}[x = y]$ for some fixed state y. Why did we already prove it in that case? In that case, on the left-hand side I'm counting the fraction of time where I visit y, and on the right I'm saying it's going to converge to $\pi(y)$. And that's exactly what we proved — if I take a particular state y and look at what fraction of time I spend at y, that converges to $\pi(y)$.

But then we're essentially done — for general f, we can write $f(x)$ in the fascinating way

$$f(x) = \sum_y f(y)\mathbf{1}[x = y].$$

(If we have 5 states, then f is a weighted sum of five indicator functions — that I'm in the first state, the second state, and so on, with the appropriate weights.)

And then we have

$$\frac{1}{t}\sum_{s=0}^{t-1} f(X_s) = \sum_{y\in\Omega} f(y) \cdot \frac{1}{t}\sum_{s=0}^{t-1} \mathbf{1}[X_s = y].$$

(We did some exchanging of sums — $f(X_s)$ itself we can write as a sum of indicator functions, and then we have a sum over states and a sum over times, and we can exchange them.) And we know

$$\frac{1}{t}\sum_{s=0}^{t-1} \mathbf{1}[X_s = y] \to \pi(y)$$

with probability 1. So all of these terms just converge to $\pi(y)$, which means the whole sum on the right-hand side converges to $\sum_{y\in\Omega} f(y)\pi(y)$ — i.e.,

$$\mathbb{P}\left[\sum_{y\in\Omega} f(y) \cdot \frac{1}{t}\sum_s \mathbf{1}[X_s = y] \to \sum_y \pi(y)f(y)\right] = 1.$$

What's happening is we have these random sums, and these fixed numbers $f(y)$. These random sums converge to constants $\pi(y)$; so we get a bunch of constants times $f(y)$-values.

Next class we'll do something with more examples, when we talk about reversible chains. But this is an important theorem, so Prof. Mossel thought it was good to see the proof.

Student Question. *Does it matter whether the sum starts at* 1 *vs.* 0*?*

Answer. No, it doesn't change anything. Always just one term in the sum doesn't change anything — our functions are always bounded, so they're negligible when we average and $t \to \infty$.

Student Question. *Is there an intuitive reason why the fraction of time spent at t converges to $\mathbb{E}[U_i(x)]/\mathbb{E}[U_i]$?*

Answer. Prof. Mossel thinks of it as a different way to chop time. One way to chop time is you can try recording at each step where you are. Another way is that I break my time into pieces, where each piece ends when I get a new z. And at each of these pieces of time, I see how many times I visited state x. If you think about a specific excursion, $U_i(x)$ is the number of times you visited x in that excursion, and U_i is the total amount; this is very random. But then I take a very long stretch; the total amount of time I spend at x is the total of all the $U_i(x)$, and the total time spent on the excursion is the total of all the U_i.

§6 February 25, 2025

Today will be more of an examples day, with less proofs. We're going to talk about reversibility and see a bunch of examples.

§6.1 Reversible Markov chains

Definition 6.1. We say a distribution π satisfies the *detailed balance equations* for a (homogeneous) Markov chain with transition matrix P if for all states x and y, we have

$$\pi(x)P(x,y) = \pi(y)P(y,x). \tag{6.1}$$

There are lots of words here, but one object is π (a distribution) and the other is P (the matrix describing transition probabilities). This is an equation that involves both π and P; all these quantities ($\pi(x)$, $P(x,y)$, and so on) are numbers.

Definition 6.2. We say P is *reversible* if such a π exists.

So if you have a special π satisfying these equations, then we say the Markov chain P is reversible.

Why do we care about this? First, in many interesting examples, this holds. But why does it help us?

Claim 6.3 — If π satisfies the detailed balance equations (6.1) for P, then π is stationary for P.

We definitely care about stationary distributions; and this tells us if you have this special property (satisfying the detailed balance equations), then you are in fact stationary. And because we like stationary distributions, this will make us happy. So that's a good reason to care about it.

Proof. What we need to check is that if π satisfies the detailed balance equations, then $\pi = \pi P$. Let's look at a specific coordinate $(\pi P)(y)$. (Recall that π is a row vector and P is a matrix, so πP is a row vector.) We have

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

(This is always true; we haven't used reversibility here.) Now we're going to use reversibility, because $\pi(x)P(x,y)$ also appears in (6.1) — the fact that π is reversible says that this is the same as

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

And why does this make me happy? Because you can pull out the $\pi(y)$, and the rest sums to 1 — we get

$$\pi(y)\sum_x P(y,x) = \pi(y)\cdot 1 = \pi(y)$$

(since from y, the sum of probabilities of transitioning everywhere else is going to be 1). So we got that

$$(\pi P)(y) = \pi(y);$$

this is the equation we wanted, so we're happy. □

This looks like some cute algebra — in general, checking $(\pi P)(y) = \pi(y)$ involves multiplying a matrix by a vector. But (6.1) is a simple equation — you don't have to add any numbers, you just do one multiplication on each side.

Why do we call this reversible? We'll have a claim that will maybe explain the name.

Claim 6.4 — If π satisfies the detailed balance equations for P, then

$$\mathbb{P}_\pi[X_0 = x_0, \dots, X_n = x_n] = \mathbb{P}_\pi[X_0 = x_n, X_1 = x_{n-1}, \dots, X_n = x_0].$$

When we write a subscript, it denotes where we're starting — so $\mathbb{P}_\pi$ means we start from the distribution π. So we're starting from π and asking what I see for the next n steps — what's the probability I first see x_0, then x_1, then x_2, and so on? And this claim says the probability of seeing a sequence going forwards and backwards is the same — seeing $x_0, x_1, \ldots, x_n$ has the same probability as seeing $x_n, x_{n-1}, \ldots, x_0$.

Student Question. *Could you explain again what the subscript means?*

Answer. This means we first sample x_0 according to π; then given that I'm at x_0 I go to x_1 according to the transition probabilities in P; and so on. So the left-hand side is

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

and similarly the right-hand side is

$$\pi(x_n)P(x_n,x_{n-1})\cdots P(x_1,x_0).$$

Proof. The idea is we can keep sliding $\pi(x_0)$ to the right (in the above equations) using the detailed balance equations. First, using the detailed balance equations on $\pi(x_0)P(x_0,x_1)$, we get

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

(using the detailed balance equations to replace $\pi(x_0)P(x_0,x_1)=\pi(x_1)P(x_1,x_0)$). We can write this instead as

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

And now we have the term $\pi(x_1)P(x_1,x_2)$, and we can do the same thing. So at each step, we'll look at one term and reverse the direction — now I replace this with $P(x_2,x_1)\pi(x_2)$. Then on the next step, I'll have $\pi(x_2)P(x_2,x_3)$, and I'll replace it with $P(x_3,x_2)\pi(x_3)$. So I sort of percolate the switches of the order of x_i and x_{i+1}, and the π keeps moving to the right. I keep doing this, and in the end (by induction) I'm going to get

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

And what's written here is just the right-hand side of our equation (i.e., $\pi(x_n)P(x_n,x_{n-1})\cdots P(x_1,x_0)$). □

The cool way to think about the name is, suppose you like science fiction. Then if you live in a reversible world, you don't know whether you're going forwards or backwards in time — you wouldn't be able to distinguish, because the probabilities of what you'd be seeing is the same. So that's the source of the name — you can't know the direction of time just by observing a sequence (it'll have the same probability either way).

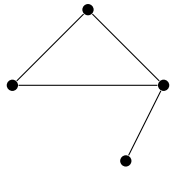
§6.2 Random walks on a graph

We promised some examples, so let's look at some.

Example 6.5 (Random walk on a graph)

We're given a graph $G=(V,E)$; we write $\deg(x)$ for the number of neighbors of x (more formally, $\deg(x)=\#\{y \mid (x,y)\in E\}$). We define the transition matrix by

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



In this graph, the vertex on the right has degree 3, and the vertex on the bottom has degree 1.

If I start the random walk at the vertex on the bottom, we'll have to go up to the vertex on the right. Here our life is more interesting — I roll a 3-sided dice and go to one of the neighbors. Maybe I go to the left; then I have two options, and maybe I go back. So whenever I'm at a vertex, I choose one of the neighbors at random and go there.

Question 6.6. Can we find a stationary distribution?

For some intuition, when I walk on this graph for a long time, am I going to spend more time on the bottom vertex or the right vertex? You'd spend more time on the right vertex — one way to think about it is that whenever I'm on the bottom, the next step I'm going to be on the right. And there's also times I visit the right vertex that have nothing to do with the bottom vertex (I might be here and then go around the triangle 5 times). So it sounds like I should spend more time at this vertex.

And that's right. In fact, we'll prove the following claim.

Claim 6.7 — The distribution given by

$$\pi(x) = \frac{\deg(x)}{\sum_y \deg(y)}$$

satisfies the detailed balance equations (and is therefore stationary).

So the way we'll show this is a stationary distribution is by showing it satisfies the detailed balance equations. This implies π is stationary; but it also means if I start from π, I can't distinguish going forwards vs. backwards in time.

Proof. First, we have to check that π is actually a distribution; but this is easy because these numbers are all nonnegative and sum to 1 (since we're dividing by the sum of degrees).

And starting with the side $\pi(x)P(x,z)$ in the detailed balance equations, if z is a neighbor of x, then we have

$$\pi(x)P(x,z) = \frac{\deg(x)}{\sum_y \deg(y)} \cdot \frac{1}{\deg(z)}$$

(the first term corresponds to the $\pi(x)$ part, and the second to the $P(x,z)$ part). And now what happens is the $\deg(x)$'s cancel, and we get

$$\pi(x)P(x,z) = \frac{1}{\sum_y \deg(y)}.$$

This quantity doesn't depend on either x or z — it's a constant — so it has to be the same no matter what x and z are (as long as they're neighbors), which means it's in particular equal to $\pi(z)P(z,x)$.

That's if z is a neighbor of x. Meanwhile, if z is *not* a neighbor of x, then $\pi(x)P(x,z) = 0$ (because $P(x,z) = 0$), and similarly $P(z,x)\pi(z)$ is also 0.

So this completes the proof that π satisfies the detailed balance equations, and therefore it's stationary. □

This is a very natural Markov chain — you do a random walk on a graph. The stationary distribution is easy to describe — it's proportional to the degree of your vertex — so this is a nice example.

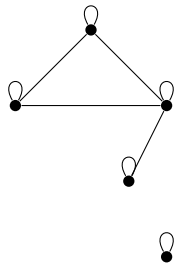
Student Question. *What happens if the graph has no edges?*

Answer. Then this Markov chain isn't defined — it's not defined if there is some vertex $x \in V$ with no neighbors (i.e., $\deg(x) = 0$).

There's various things you can do in the case where there exist isolated vertices. One way to define this Markov chain for *all* graphs is to 'make x a neighbor of itself' (for all x). More formally, this means we define

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

Sometimes the way people in graph theory think about this is that you add a *self-loop*. So maybe we have a graph similar to what we had before, but with an isolated vertex on the side; and you add a little loop from every vertex to itself.



In this case, what would be a good candidate for a stationary distribution (or something that satisfies the detailed balance equations)?

Claim 6.8 — The distribution π given by

$$\pi(x) = \frac{\deg(x)+1}{\sum_y (\deg(y)+1)}$$

satisfies the detailed balance equations.

We won't prove this, because it's the same proof.

Now we actually have two chains. We'll actually call the case with self-loops P' and π' (so that we can talk about both at once).

Question 6.9. When is P (or P') irreducible? When is it ergodic?

Let's start with irreducibility — that means you can get from any state to any other state (in some number of steps).

Claim 6.10 — Both P (if it's defined) and P' are irreducible if and only if the graph is connected.

A connected graph means you can get (using the edges of the graph) from any point to any other point. We won't prove this, but it should be clear, since the two notions are the same (we can traverse any edge with positive probability).

What about ergodicity?

Claim 6.11 — The chain P' is ergodic if and only if G is connected.

The reason is you can go from every vertex to itself, so you don't have any issues with only being able to reach some vertices at some times (you can go from a vertex to itself as long as you want to lengthen your times).

Claim 6.12 — The chain P is ergodic if and only if G is connected and not bipartite.

This is harder, and we won't prove it. But as some intuition, one case you might be worried about is something like a triangle — you might think there's a cycle of 3. But in fact, you can get from 1 to itself either by going $1 \to 2 \to 1$ (in 2 steps) or $1 \to 2 \to 3 \to 1$ (in 3 steps). And that means you can also do 7 steps, or 8 steps, or 9 steps, or 10, or 11, and so on. So the only problem is if you're bipartite (where you have even steps and odd steps). This is some intuition, but not a proof; the proof can be a harder exercise.

Maybe let's even be a little more specific and look at some *specific* graphs.

Example 6.13 (Discrete cube)

Consider the graph $G = (\{0,1\}^n, E)$ where $(x, y) \in E$ if and only if x and y differ at exactly one coordinate. (This is called the *discrete cube* or *binary cube*.)

Let's draw the first few examples of this. For $n = 2$:

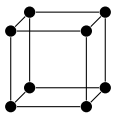
$$0 \text{ — } 1$$

Here we just go back and forth between 0 and 1; there's no randomness.

For $n = 2$, this is a more interesting chain:

$$\begin{array}{ccc} 01 & \text{—} & 11 \\ | & & | \\ 00 & \text{—} & 10 \end{array}$$

And for $n = 3$, we have eight points.



What we already know is that both P and P' are irreducible; but these graphs are bipartite, so P is not ergodic. (If I start from all-0's, then on even steps I'm always at a point where the sum of coordinates is even, and on odd steps I'm always at one where the sum of coordinates is odd.)

Also, the stationary distribution (of both P and P') is uniform, because the degrees of all vertices are the same.

Now let's define another variant chain:

Example 6.14

We define the *lazy version* of a chain P by

$$P''(x, y) = \begin{cases} \frac{1}{2}P(x, y) & \text{if } y \neq x \\ \frac{1}{2} & \text{if } y = x. \end{cases}$$

You can do this for any chain. This is called the lazy version because at each step, with probability $\frac{1}{2}$ I decide to do nothing, and with probability $\frac{1}{2}$ I do what I would have done before.

Claim 6.15 — For these graphs, P'' is ergodic, and the uniform distribution π also satisfies the detailed balance equations.

So far, this sounds somewhat boring. But there's one question about P'' you can already ask:

Question 6.16. After how many steps of P'' does one forget where they started?

We haven't defined mathematically what it means to forget where you started. But suppose I'm starting on the bottom-left. After one step I'll be at one of its neighbors (or at this vertex). So I haven't forgotten where I'm started — I might be exactly where I started, and even if I'm not, I'm at distance 1 from it. So how many steps should it take me to completely forget where I started?

We'll talk more about this later; but for this example, it's actually so simple you can talk about it without knowing anything.

A natural suggestion is n — after n steps, I've updated about n coordinates. But does this mean I've forgotten where I started? It's not clear.

But here's an equivalent description of P'' — pick a coordinate $i \in \{1, \dots, n\}$ uniformly at random, and assign a random $\{0,1\}$-value to x_i. So I pick one coordinate at random (of the n), and assign it a random value that doesn't depend on anything I've seen. And I don't change the other coordinates.

So the claim is there's an equivalent way of describing P'' that doesn't look like a random walk on a graph, but a simple procedure. I have a random pointer, I don't look at what the value before was, and I just put a random bit there. And then maybe I choose a different location (or the same location) and randomize again, and so on; and I do this over and over again until the end of time.

And the idea is that once you've used this procedure to re-randomize *all* the bits, you have no information about where you started. Randomizing doesn't mean we flip to the opposite value; it means we're tossing a $\{0,1\}$-valued coin and putting a new value here. Once we do that, it doesn't matter what bit we started with there; so once I cover all the bits, I'm done.

Claim 6.17 — Once each bit is randomized at least once, the state of the Markov chain is independent of where we started.

And how long does this take? It's the coupon collector problem (you think of the coordinates as the coupons to be collected). So if we collect all coordinates ('coupons'), then we are at this state.

So it sounds like (we're not being completely formal here) after about $n \log n$ steps, we have no idea where we started.

But it could be that even after n steps, you already don't know — this is just a proof that after $n \log n$ steps we don't know. Can we argue that after n steps we *do* maybe remember something about where we started? (This doesn't prove that the original guess of n was wrong.)

First, why is it true that if I wait $n/10$ steps, I definitely haven't forgotten where I started? Intuitively it's not possible to have touched all the coordinates; but how do we formalize that? One way is that if I start from all-0's and update $n/10$ coordinates, 90% of the coordinates are going to be 0. But a typical vertex has half-0's and half-1's. So if I started from all-0's, I'll have too many 0's, and I know I'm not at a random state.

Suppose I do $2n$ steps. How many coupons have I collected when I did coupon collector for $2n$ steps? We know you haven't collected all; but have you collected almost all, or a positive fraction, or…?

Claim 6.18 — If I start from all-0's, after $100n$ steps, a linear fraction of the coordinates are not touched.

And what does this mean about the number of 0's I have in my vector? Suppose there's 1% of coordinates I haven't touched. What's the fraction of 0's and 1's I have? The coordinates I haven't touched are all 0's, and the others I randomized, so half will be 0's and half 1's. So if I haven't touched an ε-fraction of coordinates, then with high probability the x I'm at will have at least approximately

$$\varepsilon + \frac{1}{2}(1-\varepsilon)$$

0's. And the main thing about this fraction is it's more than $\frac{1}{2}$.

A random vector of 0's and 1's should have half 0's and half 1's (with fluctuations of $o(n)$, since they're Gaussian). But with this probability, I'll have a fraction of coordinates that is greater than $\frac{1}{2}$ (with again some $o(n)$ Gaussian fluctuations). So that's why n updates or $2n$ updates is not enough.

Student Question. *Can you argue that each step doesn't have enough entropy?*

Answer. As a different procedure, we could start at coordinate 1 and randomize it, then coordinate 2 and randomize it, and so on. After n steps I'm at a random vector, and I spent less entropy than in the original procedure (since not only each time I update a bit I spend entropy, but the location of the random index was also a waste of entropy). So this is more subtle — the procedure of choosing a random bit and randomizing it is actually wasting entropy. So in this case, the random walk is just wasteful — it's not a very efficient way of using the randomness you have.

§6.3 The Ehrenfest chain

Now let's look at another example chain.

Example 6.19 (Ehrenfest chain)

We have two urns and n balls. At each step, we choose a ball uniformly at random and move it to the opposite urn.

As a brainteaser, do you see a connection between this chain and the random walk on the hypercube from before? You can think of each ball as being a coordinate — for example, let's call the two urns 0 and 1. And we'll name the balls 1, 2, 3, 4 (after the coordinates). So the picture with ball 1 in urn 0 and the others in urn 1 corresponds to the string 0111. And then if I move 3 from the right to the left urn, this corresponds to changing $0111 \to 0101$. This sounds like a reasonable story.

If that intuition is correct, what should be the stationary distribution? In the stationary distribution for the random walk on the discrete cube, how many 0's and 1's do we have? The stationary distribution is uniform (over all strings). So I'm tossing n coins and asking you how many 0's and 1's there are; and the distribution of 0's and 1's is binomial.

Here, for the state of the chain, all we need to know is how many balls are in one urn (then we automatically know the number of balls in the other). So we'll use x to denote the number of balls in urn 1.

Claim 6.20 — The distribution π given by $\pi(x) = 2^{-n}\binom{n}{x}$ satisfies the detailed balance equations with respect to P (and is therefore stationary).

We never actually wrote P down; we'll do so when proving the claim. But the intuition from the discrete cube says that this should be the case.

First, what is P? If I have x balls in urn 1, where can I go to? I can either move to $x+1$ (if I move another ball to this urn) or $x-1$ (if I move a ball from here to the other urn). And we'll have

$$P(x, x-1) = \frac{x}{n} \quad \text{and} P(x, x+1) = \frac{n-x}{n}$$

(since to go to $x-1$, we have to choose a ball in this urn).

Proof of claim. We have to check that $\pi(x)P(x, x+1) = \pi(x+1)P(x+1, x)$ (and you have to do the same with $x-1$, but it's the same equation). So we just need to check that

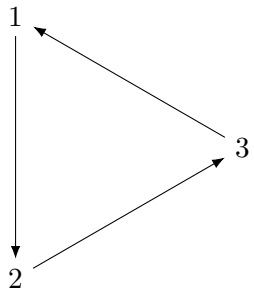
$$2^{-n}\binom{n}{x}\frac{n-x}{n} = 2^{-n}\binom{n}{x+1}\frac{x+1}{n}.$$

We won't do the algebra, but you can check and this is correct (you just look at the definitions of the binomial coefficients and you see this is true). □

§6.4 How common is reversibility?

> **Question 6.21.** Now that we love reversible chains and the detailed balance equations so much, does *every* chain have π satisfying the detailed balance equations?

The answer is no. As a simple example, we can look at the cyclic walk.



The issue here is that if $P(x, y) > 0$, then $P(y, x) = 0$; so this doesn't look like a good start to solving $\pi(x)P(x, y) = \pi(y)P(y, x)$.

As a more sophisticated question for intuition:

> **Question 6.22.** What if I look at a *random* Markov chain? Would it be reversible or not?

First, what even *is* a model for a random Markov chain on 5 states?

As one suggestion for how to generate a random Markov chain, we can take each row to be a random probability vector. So the first row we choose some numbers that are positive and normalize them to sum to 1; and we do the same for the second row, and the third row, and so on. Is this going to be reversible or not?

The answer is no; we'll give some physics intuition, in terms of degrees of freedom. To be stationary, you need to have $\pi = \pi P$; we'll say this consists of $n-1$ equations. To be reversible, you get $\pi(x)P(x, y) = \pi(y)P(y, x)$. Also here you have to be careful how many equations you have, but you have at least $\frac{n^2}{4}$. There's no reason that so many equations are going to be satisfied — being stationary is much milder (it's an n-dimensional condition, but to be reversible you have to be very lucky — you have to satisfy all these n^2 equations — and usually this won't happen).

§7 February 27, 2025

Today we'll talk a bit more about reversible chains from a different perspective.

§7.1 Generic chains are not reversible

Before that, last time we made an informal physics argument, and today we'll talk about that a bit more.

Definition 7.1. A chain P is *reversible* if there exists π such that $\pi(x)P(x,y) = \pi(y)P(y,x)$ for all x and y.

Last time, we argued:

Claim 7.2 — If n is large enough, 'most' Markov chains are not reversible.

Here's the physics proof, which we'll do a bit more accurately this time (though it's still not really a proof).

First, we'll find the dimension of the space of *all* Markov chains on n states. One way to think about this is as the dimension of the space of matrices that describe this Markov chain. And the condition is that each row is a probability vector. We have n rows, each of which is a probability vector; and there's no relationship between the rows. So this is n times the dimension of the spce of all n-dimensional probability vectors, which is $n(n-1)$ (since if we choose $n-1$ numbers, they have to sum to 1, so the last one is determined).

So the dimension of the space of *all* Markov chains is $n(n-1)$.

Now let's try to find the dimension of *reversible* Markov chains on n states. The way we'll think about it is — if we have a reversible chain, how are we going to encode it? We can encode it by two parameters. First, we'll specify π. And then how much of the probability matrix do we have to specify? We have $\pi(x)P(x,y) = \pi(y)P(y,x)$, and that means it's enough to encode what happens above the diagonal (including the diagonal itself). So we write all these numbers above the diagonal, and then using these equations we can also get the numbers below the diagonal.

So I have to choose one probability vector π, which is dimension $n-1$. And the triangle we're choosing has $\frac{n^2-n}{2} + n$ entries. So we get

$$n - 1 + \frac{n^2-n}{2} + n = \frac{n^2}{2} + \frac{3n}{2} - 1.$$

This is not actually an equality, because in fact there are more constraints. We know each row of the matrix sums to 1, so we have some extra constraints; so maybe we can make this a $\leq$.

But there are also other cheats. If $\pi(x) = 0$, then I can't solve $\pi(x)P(x,y) = \pi(y)P(y,x)$. And there's another question — what if there's more than one π corresponding to the same chain? So maybe we're overcounting.

But still this argument is sort of okay. The main lesson is that we essentially have dimensions n^2 vs. $n^2/2$, so most Markov chains are not reversible. Because this argument is sort of cheating, we don't know what 'large enough' means; but it's actually true for all $n \geq 3$.

Student Question. *Are we being rigorous about what we mean by 'most'?*

Answer. One way to make a rigorous statement of this is if I sample each row of my probability matrix uniformly at random among all probability vectors that sum to 1, the probability the Markov chain is reversible will be 0 when $n \geq 3$.

If you want to be more general, if you sample from any probability that has continuous distributions or so on — any reasonable model of sampling probability matrices will lead to a nonreversible chain.

So you can actually make this statement a lot more general.

§7.2 Birth and death chains

One lesson from where this argument fails is that if $P(x, y) = 0$ for many x and y, maybe we have a good chance for reversibility. Why? Because whenever $P(x, y)$ and $P(y, x)$ are simultaneously 0, this equation means nothing — it just says $0 = 0$.

One example of that has the very dramatic name of *birth and death chains*.

Definition 7.3. In a *birth and death chain*, we have states $\{0, 1, \dots, n\}$, and $P(x, y) = 0$ if $|x - y| > 1$.

So you can only go up or down by 1, or stay where you are. We can draw 0 through n in a line; and generically you can go one up, one back, or stay where you are. If we want to parametrize what's left, we can write $P(x, x+1) = p(x)$ and $P(x, x-1) = q(x)$ (these are the probabilities of going right and left), and then $P(x, x) = 1 - p(x) - q(x)$.

Claim 7.4 — For all p and q, the corresponding birth and death chain is reversible.

(When we write 'for all p and q' we really mean all p and q for which this is reasonably defined, e.g., p and q are positive and less than 1 and so on.)

Proof. The idea is basically that we have the right number of equations. What are all the detailed equations I have to solve? I get the equations

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

So these are the equations we have to solve; we can solve them recursively, and we get

$$\pi(x+1) = \frac{\pi(x)p(x)}{q(x+1)},$$

which eventually gets the beautiful equation

$$\pi(x) = \pi(0) \cdot \frac{p(0)}{q(1)} \cdot \frac{p(1)}{q(2)} \cdots \frac{p(x-1)}{q(x)}.$$

Why do we declare victory now? We expressed $\pi(x)$ as *something*; but do we know everything in this equation? There's one piece of information missing; I have to choose $\pi(0)$. And for that, we need to make it so that π is a probability distribution — so we normalize such that $\sum \pi(i) = 1$. This issue of normalization will come up a lot in this lecture, so it's good to think about; in this case it's simple (we expressed everything in terms of $\pi(0)$, so we can write everything explicitly and get some ugly formula). □

Student Question. *What happens if some of the q's are 0?*

Answer. What does it mean that a q is 0 — what if $q(5) = 0$? That means it doesn't go back. So when I look at stationary distributions, that means I actually have to start my chain after 5. So we maybe want to say $p \neq 0$ and $q \neq 0$ in this statement. Otherwise we have to restrict to a smaller subset of the chain.

§7.3 Sampling from distributions

Now we'll take a philosophical detour from this question for a bit, but we'll return to reversible distributions either this lecture or the next. First, here's the high-level question that we'll discuss, and then we'll specialize to Markov chains.

> **Question 7.5.** Given a description of a distribution π, how do you sample from π?

So I describe some distribution to you; how do you sample from it? People in philosophy actually are asking this question, but many people have asked it at different times, and it has different interpretations.

First we'll talk a bit about some interpretations not related to Markov chains, and then some that are.

§7.3.1 The von Neumann question

One is the von Neumann question. Von Neumann was interested in randomness, and how to generate it.

Say I want to sample i.i.d. bits using my computer. And I'm building one of the first computers in the world, and I have this physical machine with diodes and electricity gadgets, which can sample something that's 0 or 1 with *some* probability, but it's not unbiased.

> **Question 7.6** (von Neumann version)**.** Given access to i.i.d. bits X_i with $\mathbb{P}[X_i = 1] = p$ and $\mathbb{P}[X_i = 0] = 1 - p$ for *unknown* p (let's say $\frac{1}{3} < p < \frac{2}{3}$), how can we sample i.i.d. Y_i with $\mathbb{P}[Y_i = 0] = \mathbb{P}[Y_i = 1] = \frac{1}{2}$?

Here we don't really know the exact value of p — we have a physical apparatus that generates bits, and maybe we want it to be $\frac{1}{2}$–$\frac{1}{2}$, but we can't guarantee that exactly (but we *do* know they're not 0, because you did something reasonable).

For some values of p, you can sort of see how to do it. If $p = \sqrt{1/2}$, then you can look at one coin of $\sqrt{1/2}$ and another; the chance they're both heads is $\sqrt{1/2}^2$, and then I'm happy. So you can see that for various values of p, you can write various complicated functions of them that get you $1/2$. But here it's very annoying because you don't even know p.

One thing in line with the philosophy of this class is maybe you can design a Markov chain with stationary distribution $(1/2, 1/2)$, and run it for a very long time, and then you'll be very *close* to $(1/2, 1/2)$. But von Neumann wanted *exactly* $(1/2, 1/2)$.

Which Markov chain have we seen with stationary distribution $(1/2, 1/2)$? The binary symmetric channel — we have two states 0 and 1, where the probability of moving to the other state is p and the probability of staying where you are is $1 - p$. Then $\pi_t(p) \to 1/2$ as $t \to \infty$; so if I run this chain for a very long time, I'll be very *close* to $1/2$. If you actually run the calculation, you'll get

$$\mathbb{P}[Y_t = 1/2] = 1/2 \pm \eta^t.$$

But von Neumann actually got a solution which gives you the *exact* answer $1/2$, without knowing p! It's actually related to what we're going to talk about, even though it doesn't involve Markov chains. (The first time Prof. Mossel saw this, he thought it was amazing; so hopefully we'll enjoy it too.)

Proof. Sample (X_1, X_2). If $(X_1, X_2) = (0, 1)$, then we say $Y_1 = 0$. If $(X_1, X_2) = (1, 0)$, then we say $Y_1 = 1$. Otherwise, we discard (X_1, X_2) and try with (X_3, X_4).

So I'm going to sample pairs of bits, and I declare that if I see 01 that's 0 and if I see 10 that's 1; if I get 00 or 11 I don't want to touch it, so I'll throw it away and try again.

Why does this work? The events $(0,1)$ and $(1,0)$ are equally likely — both have probability $p(1-p)$. So conditioned on either of these happening, the probability of each is $\frac{1}{2}$ — in math,

$$\mathbb{P}[(X_1, X_2) = (1,0) \mid (X_1, X_2) = \{(0,1),(1,0]\}] = \frac{1}{2}.$$

So if I make a decision, it's equally likely to be 1 or 0. □

In this course, we're mostly going to talk about *approximate* sampling, as in the earlier suggestion (where we ran the binary symmetric channel a long time). Meanwhile, this is what's called *exact* sampling. We'll talk less about exact sampling, but this is a very nice example.

§7.3.2 Some more examples

Maybe that makes us happy. Now von Neumann is building a computer, and he has this physical device that generates biased bits; and this shows how to unbias them. So from now on, we're going to assume that we have access to i.i.d. coin tosses. In fact, we'll assume even more — that we have access to i.i.d. $U_1, \dots, U_n \sim \mathsf{Unif}[0,1]$. Given lots of coin tosses, how can I sample a uniform number in $[0,1]$? We can just do base-2 — I'm not really going to give you a real number (you can't even write one down), but I'll toss a coin for the first bit, then the second, and so on; and after a billion bits you don't care anymore, so I'll just give you the first billion bits and be happy. So there isn't really a big difference between uniform bits and $\mathsf{Unif}[0,1]$.

We'll see one more easy example, and then get to more Markov chain-type examples.

Example 7.7

Assume I have i.i.d. bits and uniform independent numbers in $[0,1]$. How do I generate a uniform random *permutation* on n elements $1, \dots, n$?

So I have $1, \dots, n$, and I want to generate a random permutation of these guys. When is this relevant? Is there any place where people use random permutations? When you shuffle cards, that's what you're trying to do. But now you have a computer and want a random permutation; what will you do?

Here's one way to do it — generate $U_1, \dots, U_n \sim \mathsf{Unif}[0,1]$; and then use their order to determine your permutation. So we define $\mathrm{perm}(i)$ as the position of U_i in the list $(U_1, \dots, U_n)$. The picture to have in mind is that we draw $1, \dots, n$ on the x-axis, and we write these uniform numbers as dots; and then the smallest one is going to be 1, the next-smallest is 2, and so on.

§7.4 Distributions with unknown normalization

So those are a few examples. Now we'll talk about the kinds of questions of this type that people use Markov chains a lot for.

A popular formulation of a distribution is the following: You have x living in some high-dimensional space. And the description of the distribution is that

$$\pi(x) \propto \exp(E(x)),$$

where $E(x)$ is something you can easily compute, called the *energy* of x. What we mean by 'proportional' is that

$$\pi(x) = \frac{1}{Z}\exp(E(x))$$

for some constant Z.

This is a formulation used everywhere — it started in physics, then chemistry, then biology, then machine learning, and finance, and whatever area you want — somehow, this is a very useful way to describe distributions. So you have something explicit dependin gon x (which is easy to compute), and something inexplicit (this Z) which is kind of annoying to compute.

As a comment, what is Z? Note that $Z = \sum_x \exp(E(x))$ (because the probabilities have to sum to 1). The reason this is complicated is this sum is in high dimensions, so it'll be a very big (exponentially sized) sum, which will make it hard to actually compute.

§7.4.1 Some examples

Let's see some examples, to get a feeling for this (there's lots of examples, but if you've never seen this before it might seem weird).

We'll start with physics, because this is where it came from (in this case, $E(x)$ is actually the energy of x); but then we can talk about other areas.

Example 7.8 (Ising model)

The Ising model is a model of magnets. We have $x \in \{\pm 1\}^V$ where V is the vertex set of some graph (e.g., a very large grid). By statistical physics reasoning,

$$E(x) = \beta \sum_{i \sim j} x_i x_j$$

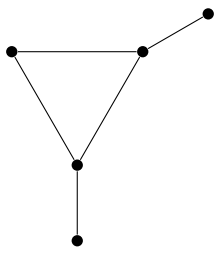
(where $i \sim j$ means i is a neighbor of j). The physics intuition is that these are the spins of atoms, and close-by atoms want to have the same spin, but they don't always have the same spin; they just like to. So this energy pushes them to have the same spin.

This distribution describes how a random magnet behaves; and then if I want to sample a random magnet from my computer, then I have to sample from this distribution.

Here's an example from theoretical computer science, about sampling uniform random colorings:

Definition 7.9. Given a graph $G = (V, E)$ and a set of q colors $[q] = \{1, \dots, q\}$, a *coloring* is a map $x : V \to [q]$ such that if $(u, v) \in E$, then $x(u) \neq x(v)$.

So we're coloring every vertex by a color, such that any two adjacent vertices have different colors.



Maybe in society, you don't really want to be unique; you just want to be different from your friends, since this makes you feel unique — the fact that there are 2 billion people like you doesn't bother you, because you don't know them.

And we want to sample a *uniform* random coloring.

Example 7.10

Suppose I want to sample a uniformly chosen q-coloring of a graph. Then we'll take

$$E(x) = \begin{cases} 0 & \text{if } x \text{ is a legal coloring} \\ -\infty & \text{otherwise.} \end{cases}$$

So the configurations that are legal all have weight 1 (when normalized), and the configurations which are illegal all have probability 0.

As another example, it used to be popular to try to sample redistricting maps — there's this thing wehre people draw districts for elections, and we have to figure out whether they're drawn for political motivations or not. And one way to deal with this is to sample random ones and see how much like them the actual ones are. Here $E(x)$ would measure how nice the curves are that define the districts, and how far they are from being completely balanced, and so on.

§7.5 The Metropolis chain

Now we want to sample from these things. And there's a very cool idea, called the Metropolis chain.

Definition 7.11. Suppose we're given a description π of a distribution on state space Ω (of the form described earlier), and an *unrelated* symmetric Markov chain ψ, the *Metropolis chain* is the new chain defined by

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

Symmetric means that $\psi(x,y) = \psi(y,x)$. So in words, to define our chain, we do the same as ψ, but then we correct it a little bit.

So I'm interested in π, I have an unrelated chain ψ, and I define a new Markov chain P. There's a couple of things we want to understand about this chain.

First, why can we compute all these quantities? That's maybe the first question. The fact that our distribution has the shape $\pi(x) = \exp(E(x))/Z$ means that $\pi(x)/\pi(y)$ is easy to compute, because we don't need Z — it's equal to

$$\frac{\pi(x)}{\pi(y)} = \exp(E(x) - E(y)).$$

So there's nothing hard about computing this — $\pi(x)$ on its own is hard to compute, because we don't know Z, but this ratio is easy to compute because Z cancels out.

The other comment is, how should I think about running this algorithm? If you want the implementation, given x, I sample y according to ψ. Then I have to decide whether to actually do it or not. So I *reject* this move (meaning I stay at x instead of moving to y) with probability $\max\{0, 1 - \pi(y)/\pi(x)\}$.

This is the same as what we defined originally — I get $\psi(x,y)$ because I sampled y according to ψ. And how do I get this minimum? Well, 1 minus this minimum is $\max\{0, 1 - \pi(y)/\pi(x)\}$. So that means with this probability I should *not* take this step, and instead stay where x is. (And I can do this by sampling a uniform random number and checking how it compares to $1 - \pi(y)/\pi(x)$.)

So we consider the step of going to y, and then we generate a uniform number to decide whether I actually like this step or not.

The main reason we're studying this is the following claim.

Claim 7.12 — The distribution π is reversible (and therefore stationary) with respect to P.

So Metropolis et. al. didn't invent this chain for no reason, but because the stationary distribution of this chain *is* the distribution we're after. So now we can take steps on this chain and converge to our distribution.

Student Question. *Why is this called the Metropolis chain?*

Answer. Metropolis was a person. The history is that this was invented in a paper by Metropolis, Rosenbluth, Rosenbluth ($\times 2$), Teller ($\times 2$). (This was an interesting paper because it involved two married couples.) This came out of the Manhattan project — they were interested in high-energy simulations. If you look at their paper, they exactly have this energy function and these pictures of configurations they want to sample, and this is where they came up with this question. It's a very interesting paper (and it's really a physics paper — no one talked about sampling then, so it was a new area, and they wrote it in physics language).

Let's try to find some examples where it's easy to explain what's going on.

Example 7.13 (Ising model)

Suppose we run the Ising model on the triangle (where we have three vertices x_1, x_2, and x_3), where

$$E(x) = x_1x_2 + x_1x_3 + x_2x_3$$

(we're taking $\beta = 1$). And suppose $\psi(x, y)$ is the random walk on the discreete hypercube $\{\pm 1\}^3$. What this means is I pick a bit and at random I flip it. This is a simple chain with nothing to do with physics (you can run it just by coin tosses); and we want to use it to get π.

This distribution is not very complicated (it only has 8 steps, so you can compute and normalize and everything). But we'll use it to illustrate.

Let's suppose my current configuration x_0 is $(-,-,-)$. Then we have

$$\psi(x_0, y) = \begin{cases} \frac{1}{3} & \text{for } y \text{ with one } + \\ 0 & \text{otherwise.} \end{cases}$$

(I'm just flipping one bit.) Suppose I choose $y = (+,-,-)$ — so this is where ψ wanted to move to.

Now, how does the chain decide whether to actually accept this move or to stay at x? Now I have a randomized decision — I move to y with probability $\min\{1, \pi(y)/\pi(x_0)\}$. What is $\pi(y)/\pi(x_0)$? Well, in x_0 everything has the same sign, so I get 3 (for the energy); if I change one of them to -1, then I get two terms of $-$'s and one $+$ in the energy, so I get -1. So this is

$$\min\left\{1, \frac{\exp(-1)}{\exp(3)}\right\}.$$

So how do I implement this? I sample $U \sim \mathsf{Unif}[0, 1]$; and if $U \le \exp(-4)$ I move to y, and otherwise I stay at x_0.

So that's one step of this chain — I try to follow the simple chain ψ that I like, but I have to take π into account. And the only way I take π into account (as with the von Neumann trick) is that it tells me whether to reject or not — do I actually take this move, or do I stay at x and try again?

The last thing we'll do in this class is to prove this is actually reversible.

Proof of Claim 7.12. For $x \neq y$, we have to look at $\pi(x)P(x,y)$. Putting the $\pi(x)$ inside the minimum, this is

$$\psi(x,y) \cdot \min\left\{\pi(x), \pi(x) \cdot \frac{\pi(y)}{\pi(x)}\right\} = \psi(x,y) \cdot \min\{\pi(x), \pi(y)\}.$$

And now we can use the fact that ψ is symmetric, so this is equal to

$$\psi(y,x)\min\{\pi(y), \pi(x)\},$$

and by the same algebra this is also equal to $\pi(y)P(y,x)$. □

At first this seems very weird. So Prof. Mossel tried to highlight some of the key features. First, π is given in some specific form that first looks very weird, but is actually very common; and in this particular form, ratios are very easy to compute, which lets you cook up this chain.

§7.6 Forecast

Next class, we'll do something slightly more general — we'll deal with the case where ψ is not necessarily symmetric. Then we'll talk about something very important — the *speed* of convergence to π. We'll talk about this for many lectures. Because the original motivation is applied — we don't just want something in the limit, we actually want to sample from π. If I know *eventually* I'll converge to π but only after the universe has ended, that's not useful. So we want to analyze how quickly the convergence happens, and that leads to interesting mathematical questions.

§8 March 4, 2025

§8.1 The general Metropolis chain

Last class, we saw the Metropolis chain when the proposal distribution was symmetric; now we'll see the general case (where it's not necessarily symmetric).

The input is a description of the desired probability distribution π, of the form

$$\pi(x) = \frac{1}{Z}\exp(E(x)).$$

And our goal is to sample from π. We also have a Markov chain that we like on the state space Ω — we'll call it $\psi(x,y)$.

Definition 8.1 (Metropolis chain)**.** The *Metropolis chain* is defined as

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

for $y \neq x$, and $P(x,x) = 1 - \sum_{y \neq x} P(x,y)$.

Claim 8.2 — The distribution π satisfies the detailed balance equations for P.

(This means π is the stationary distribution, and so on.)

Student Question. *Where did ψ come from?*

Answer. It's some Markov chain on the state space that we like — maybe there's a very simple chain that's easy to implement, and that's the one we're going to use. And the goal is to find π, which is some more complicated distribution.

Proof. There's one thing we didn't check last class — that P is actually a Markov chain. (We checked the detailed balance equations, but didn't actually check it's a Markov chain.) What do we have to check for this? In principle, it could be that $\sum_{y\neq x} P(x,y) > 1$, and then we'd have a problem. (All the other numbers are nonnegative, but it's not clear that this one is — i.e., $P(x,y) \geq 0$ for $y \neq x$ by definition, but we need to check $P(x,x) \geq 0$.)

So why is $\sum_{y\neq x} P(x,y) \leq 1$? Well, it's at most $\sum_{y\neq x} \psi(x,y)$. And ψ is a Markov chain, so this is at most 1. This implies $P(x,x) = 1 - \sum_{y\neq x} P(x,y) \geq 0$. So we're happy.

That was more of just a sanity check; all of this means P is a Markov chain. Now we want to check the detailed balance equation — so we consider $x \neq y$, and we want to look at $\pi(x)P(x,y)$. We can put the π inside to say

$$\pi(x)P(x,y) = \min\left\{\pi(x)\psi(x,y), \pi(y)\psi(y,x)\right\}.$$

And it doesn't matter what this expression is — as long as it's symmetric in x and y, we're done. Syntactically if I exchange the roles of x and y, then I'll get exactly the two terms swapped, which means we'll get exactly the same thing; so this is also equal to $\pi(y)P(y,x)$. This means we satisfy the detailed balance equations. □

So we have this somewhat complicated looking chain that lets us sample from π, in the sense that if we take enough steps, then if the chain is ergodic, we'll converge to π.

Student Question. *Last class we required ψ to be symmetric, but here we're relaxing that a little?*

Answer. Yes. Last class ψ was symmetric, and the chain was exactly the same, except that we wrote it in a different way — last time, we assumed that $\psi(x,y) = \psi(y,x)$, and the way we wrote it was as

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

In the special case where ψ is symmetric, we get the same thing — moving ψ inside, this is

$$\min\left\{\psi(x,y), \psi(x,y) \cdot \frac{\pi(y)}{\pi(x)}\right\},$$

which is the same as what we have (since $\psi(x,y) = \psi(y,x)$ in this case). So what we're doing here is a slightly more general version of what we did last class, but the same philosophy.

§8.2 Examples of Metropolis chains

Now let's consider a simple examples.

Example 8.3

Suppose π is a uniform vertex from a graph $G = (V,E)$, and $\psi(x,y)$ is a random walk on G, i.e.,

$$\psi(x,y) = \begin{cases} \frac{1}{\deg(x)} & \text{if } y \sim x \\ 0 & \text{otherwise.} \end{cases}$$

This sounds like a very silly question — if I have a graph G encoded on my computer (where the vertices are $1, \dots, n$ and the edges are written down), you can just sample a uniform random vertex directly. But sometimes the graph isn't given in this form — for example, if the graph is the web, you might not have direct access to a random webpage. Instead, all you can do is go from one webpage to another.

So what's going to be easy for us — the thing that's $\psi(x, y)$ — is not choosing a random vertex, but going from one to another.

So if the entire graph is on your computer then sampling a uniform vertex is easy. But if the graph is huge and not in the memory of your computer, then it's not; instead you have to use something to explore the graph, and we'll use a random walk.

As a sanity check, note that ψ is not symmetric — we don't necessarily have $\deg(x) = \deg(y)$.

What's the Metropolis chain here? It's going to be

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

(since $\pi(y) = \pi(x)$), when $y \sim x$.

Remark 8.4. There's another way of writing $P(x, y)$ which is useful for the implementation of the algorithm — we can write it as

$$\psi(x, y) \cdot \min\left\{1, \frac{\pi(y)}{\pi(x)} \cdot \frac{\psi(y, x)}{\psi(x, y)}\right\}.$$

Why write it in this way? This means what we're doing is we're following $\psi(x, y)$, but sometimes we reject — if this second number is less than 1, then with probability 1 minus this number, we reject.

So in this case, we can rewrite this as

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

In other words, what we're doing is we take a random walk step. Then if $\frac{\deg(x)}{\deg(y)} < 1$, we reject the step with probability $1 - \frac{\deg(x)}{\deg(y)}$ (and stay at x instead).

Let's look at the philosophy here. I want to sample a uniform random vertex. What's easy for me to do is take a (random) walk on the graph. And in a step of my algorithm, what I do is I choose a uniform random neighbor y of x, and move to y. But then I think, do I actually want to move to y or not? If it's a lower-degree vertex, I'm going to say, 'yeah I like this move' and actually take it. If it's a higher-degree vertex, then I think, 'well, maybe I don't like this' and with some probability I reject the move.

Why does this intuitively make sense? The random walk ψ prefers higher-degree vertices — we saw that the stationary distribution for ψ is proportional to $\deg(x)$. And what Metropolis does is it 'corrects' for this by preferring to stay at lower-degree vertices. So that's the philosophy — the random walk visits high-degree vertices too often, so if we're at a low-degree vertex we want to stay there longer, by rejecting some of the moves going out.

Let's make this very concrete.

Example 8.5

What is the transition matrix P for this problem if G is a star?

This means there's some vertex v_0 in the center; and then v_1, v_2, …, v_n are all adjacent to it and nothing else. What will the Metropolis chain do here?

It'll do one thing when it's at v_0, and something else when it's at all of the tips. When you're at v_0, all your neighbors have lower degree, so you'll never reject; it's $\frac{1}{n}$ to everyone else. And if I'm at a tip v_i (for $i > 0$), what'll happen? It'll go back with probability $\frac{1}{n}$, and stay with probability $1 - \frac{1}{n}$. So in this case, where there's a huge imbalance between the central guy and all the tips, what'll happen to compensate is that when we're at a tip, we almost always stay there (and only with probability $\frac{1}{n}$ we go back to the center). One way to think about this is that according to ψ, the central guy has the same mass as all the other guys together. So there's a factor of n imbalance, and to compensate for this, there's only a $\frac{1}{n}$ chance that we actually go back.

Now let's look at another example.

Example 8.6 (Metropolis for graph coloring)

Suppose our input is a graph $G = (V, E)$ and a set of colors $[q] = \{1, \dots, q\}$, and π is a uniformly chosen coloring of G.

Let ψ be the random walk on all colorings of G — more precisely, on $[q]^n$ — where

$$\psi(x, y) = \begin{cases} \frac{1}{n(q-1)} & \text{if } y \text{ differs from } x \text{ in one coordinate} \\ 0 & \text{otherwise.} \end{cases}$$

When we say *coloring*, we mean a *legal coloring*, where any two vertices which are neighbors have different colors.

So for ψ, coloring is very complicated; we don't like that. But what I like is choosing one vertex at random and choosing a color for that vertex. (For ψ, these colorings don't have to be legal — I just color however I want.)

What would the Metropolis chain look like here? First, is ψ symmetric? Yes — $\frac{1}{n(q-1)}$ doesn't involve anything about x and y, it's just a number, so whenever x and y differ in one coordinate it's the same.

We have $\psi(x, y)$, and then $\min\{1, \frac{\pi(y)}{\pi(x)}\}$. Because it's uniform, you might think $\pi(y)$ and $\pi(x)$ are always the same, so this is always 1. But we defined ψ even on illegal colorings — maybe I chose a green vertex and tried to color it blue, but one of its neighbors was already blue, in which case this would produce an illegal coloring. So we only have to look at the edge case where I'm trying to take a step which would not result in a legal coloring.

So the Metropolis chain is very simple — I choose a uniform random vertex $v \in V$ and a random color for v (which is different from its current color). Then we have to decide whether to reject or accept. If this is a legal coloring, we accept; otherwise we reject.

So this is a simple algorithm — you just choose a random vertex and try to color it with a random color different than what we have right now. If it is legal you keep the move; otherwise you do not.

§8.3 Questions about Metropolis chains

Now let's talk about a few general mathematical questions we're interested in. We want to converge to π; for this, the minimum we need is the following:

Question 8.7. When is the Metropolis chain P irreducible?

Another question that's very important:

Question 8.8. How fast is the convergence to π?

§8.4 Irreducibility for graph coloring

We won't answer these questions in generality, but let's think about these questions for graph coloring. (And then we'll also develop some general theory.)

Under what conditions is this chain irreducible?

A first guess might be when the graph is disconnectd. As an example, let's consider the graph with three vertices and no edges. Suppose I want to color this with two colors. Is the Metropolis chain going to be good? Yes — I'm going to update one color each time, and there's no constraints; I'm just taking a random walk on this discrete cube. So the fact that it's not connected is not a big deal.

Here's a graph that's more annoying — graphs with edges actually turn out to be more annoying. Suppose I have a single edge, and I want to color it in two colors. That's not hard — there's only two colorings (12 and 21). But suppose my current coloring is 12. Where can I go to from here?

So we're starting with 12, and the goal is a uniform random 2-coloring. But the problem is I'm stuck — if I update the left vertex I can only update it to 2, and that's illegal, so I can't do that (and similarly I can't update the other vertex to 1).

As a bigger example (generalizing this), if I want a 3-coloring of a triangle, we have the same issue — if I have 1, 2, and 3 then everything is rigid (when I want to update any vertex, I'm stuck).

Are there problems other than complete graphs?

The single-edge example also generalizes in a different way. If you take *any* bipartite graph (for example, a 2-path colored 121), I'm stuck.

So there's something to think about when this chain actually moves around. We don't want it to be stuck — if I'm stuck, I can't do anything.

What's a good thing for us, and what's a bad thing for us? The more colors, the better (the more freedom you have to change colors), and the lower the degree is, the better (a high-degree node means you have lots of constraints, so it's easier to get stuck; and we don't want that). Here's one claim that gives a positive instantiation of that.

Claim 8.9 — If $\max_{v \in V} \deg(v) \leq d$ and $q \geq d + 2$, then P is ergodic.

So as long as the number of colors is at least the max degree plus 2, then the chain is ergodic (you can get from any state to any other state, and if you take some power of the matrix, eventually everything will be good).

Remark 8.10. You might also need the graph to not be empty for ergodicity. If you just have one vertex and the number of colors is 2, then it's going to alternate between colors 0 and 1, which means it's not going to be ergodic.

This proof isn't really about Markov chains; it's really a question in graph theory.

We'll just prove that P is irreducible. (It's true that it's ergodic as long as the graph is nonempty.)

Proof. What do we want to show? We have some initial legal coloring, and we want to get to a target legal coloring. And we want to show that if we start from the initial coloring, there's a way for us to get to the final coloring.

As a first attempt, we can try to change the colors one by one from the initial to the final coloring. But this doesn't actually work. As a mini-example, suppose I have vertices v_1, v_2, and v_3, initially with colors 1, 2, and 3; and I want v_1 to eventually be color 2. Then we'd like to change v_1 from 1 to 2; but we can't do that, because it's illegal. (Imagine I'm allowing you to use 5 colors.)

What should we do here? We want to change v_1 from color 1 to 2. We don't like 1 and 2, because they're involved in what we're doing here. So what we'll do is we look at all the neighbors which are 1 or 2, and change them to something else. But what if later, we run into issues where we need to change this newly colored vertex to something else? The idea is this won't happen, because the final coloring is valid.

Let's see the proof in actuality. Let's give names to the coloring — so let $\sigma : V \to [q]$ and $\tau : V \to [q]$ be two legal colorings. Our goal is to show that we can get from σ to τ using moves of P (i.e., moves that, according to P, have nonzero probability).

We're going to do this by induction — we're going to show that you can go from $\sigma = \sigma_0 \to \sigma_1 \to \sigma_2 \to \cdots \to \sigma_n = \tau$. And what we want from the σ_i's is that σ_i agrees on the first i vertices with τ — so $\sigma_i(v_j) = \tau(v_j)$ for all $j \neq i$ (where $V = \{v_1, \ldots, v_n\}$). So we're going to go vertex-by-vertex. We're starting at σ, and then I'll show I can color the first vertex right, then the second vertex, and so on.

So let's think about how to get from σ_{i-1} to σ_i. This means we have vertices v_1, …, v_{i-1}, which are all already colored like they should be according to τ. Then I have the vertices v_i, v_{i+1}, …, v_n, whose colors we don't know anything about. And our goal is to move v_i to also be colored according to τ.

So we know what color we want v_i to have — its color according to τ. But maybe this is impossible because some of its neighbors in v_{i+1}, …, v_n already have that color. So we might have to change some of those first.

So what we do is first that if $j > i$ and $j \sim i$, then we update $\sigma_i(v_j)$ to a color that is not in $\{\tau(v_i), \sigma_{i-1}(v_i)\}$. The point is we have no problems with v_1, …, v_{i-1} — τ is a legal coloring, so they don't conflict with $\tau(v_i)$. But we might have conflicts with the guys on the right.

And why can we do this? Each of these vertices v_j is of degree at most d. And j is a neighbor of i, so one of its neighbors is v_i; we have at most $d-1$ other neighbors. And we have $d+2$ colors, so there has to be a color that's legal and different from these two.

So first we update all these guys v_j that might be problematic. And after you do that, you can update $\sigma_i(v_i)$ to $\tau(v_i)$.

So it's a 2-stage process — first you make sure the people you haven't updated yet and don't care about won't cause you a problem in updating this guy. Then you update this guy, and you've captured another guy; and you can keep doing this. □

Remark 8.11. What's the maximum length of the path between beginning and end, according to this? It's at most n^2 or nd or something like that — at each step, I only have to update at most d guys (the neighbors of v_i), and this happens n times.

So the length of the path is at most $n(d+1)$.

§8.5 Graph coloring using Metropolis for specific graphs

Let's look at a few quick examples; this will lead us to one example that we'll analyze in much detail. We're continuing with this example of graph coloring using Metropolis, and now we'll look at some specific graphs.

Example 8.12

Consider $G = K_n$ (the complete graph on n vertices).

What do we need for the chain to be ergodic? Everyone's connected to everyone, and we have n vertices. So the Metropolils chain is going to be ergodic if $q \geq n+1$ (by the above claim). This is necessary — if we have n colors we can't do it (if I have n people and n colors I can't move anyone — whenever I try to move anyone, I'm stuck).

So our claim is not completely useless — it's tight in at least some cases.

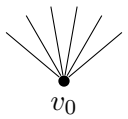
Student Question. *Is this true even for $n = 2$ with 3 colors? Is it still ergodic? Isn't there a cycle with the six different colorings?*

Answer. There's a difference between the chain we've described and the chain you're thinking about (the Glauber dynamic, where you always change). Here we have a self-loop, if you try to change a color and it's not allowed; and if you have self-loops, you're always ergodic. The point is what might happen is I choose the second vertex and try to color it 1, which means nothing happens; so there's a chance that I go back to myself.

The cycle you're thinking about is 12, 32, 31, 21, 23, 13. So this is a chain of six things, and I can go left or right. But there's also a constant probability I stay where I am — if I look at 32 and try to update the left guy to 2, I'm instead going to stay there. And that's why it's ergodic. (So it's always 1/3 I move to the left, 1/3 I move to the right, and 1/3 I stay where I am.)

Example 8.13

Let $G = S_n$ (this means the star with n tips).



What does the claim say? It says that if $q \geq n+2$, then I'm ergodic. Is this tight nor not? Do I actually need $n+2$ colors for this chain to be ergodic, or do I actually need less?

You can actually do this with 3 colors! So it's very far from being tight.

Claim 8.14 — When $G = S_n$ and $q = 3$, the chain is ergodic.

What's the idea? Once the center guy is okay the tips are fine (you can change them one by one). So the heart of the issue is, how do I change the center guy? Say the center is blue and I want it to become green. How do I do that? I have three colors, so let's first fix all the tips to red; and then I can make the center guy green.

So let's say I start with some configuration where the center guy si red, and I have a bunch of tips, but I don't care what they are. Then first I'm going to move to the same configuration where all the tips are green (I can do that because green is different from red, so there are no constraints). Once I do that, I can move the center guy to be orange (which was my goal), and all the other guys are going to be green. And once I've done that, I can move all the tips to whatever I want.

So this is a proof by picture — we went from the source to the destination (and the question marks on the tips in both means that we don't care what's written there — whatever it is, this is going to work).

So this sounds great; maybe the star is easy. Here's the claim we're goign to prove next time.

Claim 8.15 — The number of steps needed to get 'close' to the uniform distribution for S_n with 3 colors is exponential in n.

(We haven't formally defined what this means yet, but we will later.)

So we're very happy that we found a Markov chain that's ergodic with 3 colors for coloring this graph. But if you'll run it and you want to sample from this distribution — you work for this coloring graph consultancy, and you say, I have good news and bad news. The good news is the chain is ergodic; the bad news is it's going to take exponentially long (in the hundred thousand nodes I have in the graph) to converge to the uniform coloring. So it sort of sounds like everything is okay, but practically the length of the time it'll take you to get to the stationary distribution is very, very long in this example; that's one of the things we'll prove next lecture.

§8.6 Total variation distance

Before the end, we'll give one definition that measures how close we are — that measures the distance between two probability distributions — that we need in order to formalize statements like this.

Definition 8.16. Given two probability distributions μ and ν on a finite state space Ω, the *total variation distance* between μ and ν is defined as

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

In analysis, this is usually called the L^1 distance between two vectors; we take $\frac{1}{2}$ because we want the maximum total variation distance to be 1. If your probability measures live in completely different parts of Ω, then this is going to be exactly 1; if $\mu = \nu$ then of course this is going to be 0.

Maybe one thing we'll check next class is that this is actually equal to

$$1 - \sum_{x \in \Omega} \min\{\mu(x), \nu(x)\}.$$

Next class we'll formalize the above claim using the total variation distance; and we'll see that the total variation satisfies the triangle inequality and prove some properties. And then we'll show the claim — sometimes the Markov chain is ergodic and very nice, but it takes exponentially long to converge, and then you're not happy.

§9 March 6, 2025

Note: I missed this class, so these notes are just my attempt to fill in details from the slides.

§9.1 Total variation distance

Definition 9.1. Given two probability distributions μ and ν on the same space, their *total variation distance* is defined as

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

There are several equivalent ways to write the total variation distance.

Claim 9.2 — We have $\|\mu - \nu\|_{\mathsf{TV}} = 1 - \sum_{x \in \Omega} \min\{\mu(x), \nu(x)\}$.

Proof. We have $\sum_x \mu(x) = \sum_x \nu(x) = 1$, so we can replace the 1 on the right-hand side with $\sum_x \frac{1}{2}(\mu(x) + \nu(x))$ to get that

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

□

Claim 9.3 — We have $\|\mu - \nu\|_{\mathsf{TV}} = \sup_{A\subseteq\Omega}(\mu(A) - \nu(A))$.

Proof. For any $A \subseteq \Omega$, we can write

$$\mu(A) - \nu(A) = \sum_{x\in A}(\mu(x) - \nu(x)) \quad \text{and} \quad \mu(\Omega \setminus A) - \nu(\Omega \setminus A) = \sum_{x\notin A}(\mu(x) - \nu(x)).$$

Since $\mu(A) - \nu(A) = -(\mu(\Omega \setminus A) - \nu(\Omega \setminus A))$, this means we can write

$$\mu(A) - \nu(A) = \frac{1}{2}\sum_{x\in A}(\mu(x) - \nu(x)) + \frac{1}{2}\sum_{x\notin A}(\nu(x) - \mu(x)) \le \frac{1}{2}\sum_{x\in\Omega}|\mu(x) - \nu(x)|.$$

Equality holds if we take $A = \{x \in \Omega \mid \mu(x) \ge \nu(x)\}$. □

Note that $\|\bullet\|_{\mathsf{TV}}$ is a metric (in particular, it satisfies the triangle inequality — we can see this by applying the triangle inequality for $|\bullet|$ for each $x \in \Omega$), and that it's always between 0 and 1.

§9.2 Rate of convergence to equilibrium

We're often interested in not just *whether* a Markov chain eventually converges to its stationary distribution, but *how fast* it does so; and the total variation distance lets us quantify this.

Theorem 9.4

Let P be a (finite) ergodic Markov chain with stationary distribution π. Then for any distribution μ, we have $\lim_{t\to\infty} \mu P^t = \pi$. Moreover, there exist constants C and $\alpha < 1$ such that for any distribution μ, we have

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

Proof sketch. We proved earlier that $\lim_{t\to\infty} \mu P^t = \pi$, and the quantitative bound essentially falls out of the proof we gave. Specifically, in that proof we defined

$$\Pi = \begin{bmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{bmatrix},$$

and tried to show that $P^t \to \Pi$. To do so, we took k such that all entries of P^k were positive (we can find such a k because P is ergodic), and wrote

$$P^k = \varepsilon\Pi + (1-\varepsilon)Q,$$

where Q is a transition matrix for some Markov chain and π is stationary for Q. (We can simply take ε to be the minimum entry of P^k, and subtract out $\varepsilon\Pi$; the resulting Q will automatically have these properties.) Then we'll inductively have

$$P^{tk} = (1-\varepsilon)^t Q^t + (1-(1-\varepsilon)^t)\Pi$$

for any $t \geq 0$. And $\mu\Pi = \pi$, so we get

$$\mu P^{tk} = (1-\varepsilon)^t \mu Q^t + (1-(1-\varepsilon)^t)\pi,$$

and therefore

$$\|\mu P^{tk} - \pi\|_{\mathsf{TV}} = (1-\varepsilon)^t \sum_x \left|\mu Q^t(x) - \pi(x)\right|.$$

We can bound $\sum_x |\mu Q^t(x) - \pi(x)| \leq 2$ (since these are probability distributions), so we get the desired exponential decay statement (setting $\alpha = (1-\varepsilon)^{1/k}$ and $C = 2(1-\varepsilon)^{-1}$). □

In some sense, this means we get exponentially fast convergence to π as $t \to \infty$. However, α might be very close to 1, or C might be very large. So it's possible that the rate of convergence is very slow in terms of the size of our Markov chain, as we'll see in the next example.

§9.3 Slow mixing and bottlenecks

Now we'll prove the statement mentioned last lecture — that the Metropolis algorithm for coloring a star (with n tips) with $q = 3$ colors takes *exponentially* long (in n) to get close to its stationary distribution (which is the uniform distribution).

Proposition 9.5 (Slow mixing of star coloring)

Let P be the transition matrix for the Metropolis algorithm for coloring the star with $q = 3$ and $n+1$ nodes. Then there exists a state x such that for all t, we have

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

(We use δ_x to denote the distribution which places a 1 on x and 0's everywhere else.)

This means until t is exponential in n, we'll be far from the stationary distribution.

In order to prove this, we'll prove a more general result for proving 'slow mixing.'

Theorem 9.6

Suppose P is an ergodic Markov chain with stationary distribution π. Suppose that Ω can be partitioned into 3 sets L, B, and R such that

$$P(x,y) = 0 \quad \text{for all } x \in L \text{ and } y \in R.$$

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

$$\left\|\delta_x P^t - \pi\right\|_{\mathsf{TV}} \geq \pi(R \cup B) - t \cdot \frac{\pi(B)}{\pi(L)}.$$

Before we prove this, we'll see how it applies to star coloring.

Proof of Proposition 9.5. Let v_0 be the center of the star, and $v_1, \ldots, v_n$ the tips. (We'll use 0, 1, and 2 to denote the three colors.) We'll define the partition $\Omega = L \cup R \cup B$ as follows:

- Let B be the set of colorings where all the tips are colored with 1 or all are colored with 2, i.e.,

$$B = \{\sigma \mid \sigma(v_1) = \cdots = \sigma(v_n) \in \{1, 2\}\}.$$

- Let L be the set of colorings not in B where v_0 is colored with 0.
- Let R be the set of colorings not in B where v_0 is colored with 1 or 2.

Note that if we start with a coloring $\sigma \in L$, then v_0 is colored with 0, and some tips are colored with 1 and others with 2. To get to another coloring $\tau \in R$, we'd have to recolor v_0 with 1 or 2. But this is impossible, because that would conflict with one of the tips. So we have $P(\sigma, \tau) = 0$.

So we can apply Theorem 9.6 to this partition. We have $|B| = 4$ (there's 2 choices for the common color of the tips, and then 2 choices for the center), $|L| = 2^n - 2$ (there's 2 choices for each of the tips — they can be colored with 1 or 2 — and we remove the two colorings where they're all 1 or all 2), and $|R| = 2(2^n - 1)$ (for the same reason). There are $3 \cdot 2^n$ total colorings, so Theorem 9.6 gives that there exists x such that

$$\left\|\delta_x P^t - \pi\right\|_{\mathsf{TV}} \geq \pi(R \cup B) - t \cdot \frac{\pi(B)}{\pi(L)} \geq \frac{2}{3} - \frac{4t}{2^n - 2}. \qquad \square$$

Now we'll prove Theorem 9.6. The intuition is that to get from L to R, we need to pass through B; and we think of B as 'small,' so this is unlikely.

Proof of Theorem 9.6. First note that for every $x \in L$, we have

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

(since in order to get from x to R, we must pass through B at some point). Now imagine that we choose a random x according to π, conditioned on x being in L. Then we have

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

(since we can remove the conditioning at the cost of a $\pi(L)$-factor). But this is just

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

(since for each s, we have $\mathbb{P}_\pi[X_s \in B] = \pi(B)$).

In particular, this means there exists $x \in L$ with

$$\mathbb{P}_x[X_t \in R \cup B] \leq t \cdot \frac{\pi(B)}{\pi(L)},$$

which means (using the description of total variation distance from Claim 9.3) that

$$\left\|\delta_x P^t - \pi\right\|_{\mathsf{TV}} \geq \left|\pi(R \cup B) - \mathbb{P}_x[X_t \in R \cup B]\right| \geq \pi(R \cup B) - t \cdot \frac{\pi(B)}{\pi(L)}.$$

$\square$

§9.4 Strong stationary times and fast mixing

Theorem 9.6 gives a way to prove that a Markov chain converges *slowly*.

Question 9.7. Can we prove that a Markov chain converges *quickly*?

We'll see two ways to do so — *strong stationary times* (which we'll discuss this lecture) and *coupling* (which we'll discuss next lecture).

Definition 9.8. Given a sequence of random variables X_0, X_1, ..., a *stopping time* τ is a random variable taking values 0, 1, 2 ..., ∞ such that for all t, there exists a set A_t such that $\tau = t$ if and only if $(X_0, \dots, X_t) \in A_t$.

Intuitively, this means we should be able to decide whether to stop at time t based on just the information we've seen so far (namely, $X_0, \dots, X_t$).

Definition 9.9. 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 the sequence Z_t.

Example 9.10

Consider the lazy random walk on the hypercube — let $Z_t = (I_t, B_t)$ where $I_t \sim \mathsf{Unif}\{1, \dots, n\}$ and $B_t \sim \mathsf{Unif}\{\pm 1\}$; and for $x \in \{\pm 1\}^n$, let $f_t(x, (i, b))$ be the point obtained by replacing x_i with b and keeping all other coordinates as they are.

Let $\tau = \min(t \mid \{I_1, \dots, I_t\} = \{1, \dots, n\})$. This is called the *refresh time* (because when it occurs, it means we've resampled (or 'refreshed') all coordinates). Note that τ has the same distribution as coupon collector.

Definition 9.11. A *strong stationary time* for a Markov chain X_t is a randomized stopping time τ such that for all x, y, and t, we have

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

Intuitively, this means that if I know I'm stopping at some time t, then my distribution should be the stationary one.

Example 9.12

The refresh time is a strong stationary time. This is because once I've resampled every coordinate at least once, they're all independent and uniformly random (regardless of what the initial distribution was).

Claim 9.13 — If τ is a strong stationary time, then

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

(This follows just by summing the definition over all $0 \le s \le t$.)

Claim 9.14 — If τ is a strong stationary time, then for any starting distribution μ, we have

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

Proof sketch. By Claim 9.2, we can write

$$\left\|\mu P^t - \pi\right\|_{\mathsf{TV}} = 1 - \sum_x \min\{\pi(x), \mathbb{P}[X_t = x]\}.$$

And we have $\mathbb{P}[X_t = x] \geq \mathbb{P}[\tau \leq t]\pi(x)$, so we get

$$\min\{\pi(x), \mathbb{P}[X_t = x]\} \geq \pi(x)\mathbb{P}[\tau \leq t],$$

and therefore

$$\left\|\mu P^t - \pi\right\|_{\mathsf{TV}} \leq 1 - \sum_x \pi(x)\mathbb{P}[\tau \leq t] = 1 - \mathbb{P}[\tau \leq t] = \mathbb{P}[\tau > t].$$

(This probability is with respect to the starting distribution μ; but that's an average over the probabilities with starting distributions $\mathbb{P}_x$, so it's upper-bounded by $\max_x \mathbb{P}_x[\tau > t]$.) □

§10 March 11, 2025

§10.1 Strong stationary times

We started talking about how to prove upper bounds on the mixing time (the time it takes to get close to the stationary distribution). Last time, we started discussing how to do this using strong stationary times.

Definition 10.1. A *strong stationary time* τ for a Markov chain X_t is a stopping time such that

$$\mathbb{P}_x[X_\tau = y \text{ and } \tau = t] = \pi(y)\mathbb{P}_x[\tau = t].$$

Last time, we saw one example:

Example 10.2

The refresh time for the lazy random walk on the discrete cube $\{0,1\}^n$ is a strong stationary time.

One way to think about this random walk is we choose a coordinate at random and update it to a random value (independently of what's there right now). And the refresh time is the first time we've updated each coordinate at least once. We saw that was an example of a strong stationary time.

The claim we ended class with last time was the following:

Claim 10.3 — If τ is a strong stationary time and μ is any starting distribution, then

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

So we can bound the distance from the stationary distribution after t steps by the probability we haven't stopped within t steps (more precisely, the maximum of this probability over all starting states x).

§10.1.1 Example — the lazy random walk on the hypercube

Let's try to apply this to a concrete example — the refresh time for the lazy random walk on the hypercube. So we want to say we're close in total variation; and the upper bound we'll get is the maximum (over wherever we start) of the probability it takes us more than t steps to stop.

How long does it take us to stop? Eventually we need to know $\mathbb{P}_x[\tau > t]$ — the probability I have to wait more than t steps to refresh all the coordinates. For what values of t do we expect this to be small?

This is the coupon collector problem — τ is the coupon collection time (instead of collecting coupons we're collecting coordinates). Usually when you learn about coupon collector, you're not asking about $\mathbb{P}_x[\tau > t]$, but instead $\mathbb{E}[\tau]$; and this we know is about $n \log n$. And we can apply Markov and do something. But in fact, if we want $\mathbb{P}_x[\tau > t]$, we can get something more accurate.

Claim 10.4 — For all x and all c, we have $\mathbb{P}_x[\tau \geq n \log n + cn] \leq e^{-c}$.

The chain is symmetric, so it doesn't matter what x we start at. (Of course, if $c < 0$ this is not very interesting.)

So the conclusion from this is that for the lazy random walk on the hypercube, for any μ, we have

$$\left\|\mu P^t - \pi\right\|_{\mathsf{TV}} \leq e^{-c} \quad \text{if } t \geq n \log n + cn.$$

So if we wait $n \log n + 100n$, then the total variation will be e^{-100}, which is small; if we wait $2n \log n$ steps then it'll be even smaller.

Proof of Claim 10.4. We'll use the first moment method (or union bound). Let Z be the number of coordinates that haven't been updated by time t. What we want to know is $\mathbb{P}_x[\tau > t]$. For τ to be bigger than t, there has to be at least one coordinate we haven't updated; so this is equal to $\mathbb{P}[Z \geq 1]$. And because Z takes integer values, this is at most $\mathbb{E}[Z]$.

And now for each coordinate we can compute the probability we haven't updated it. There are n coordinates, and the probability we haven't updated a particular coordinate is $(1 - \frac{1}{n})^t$ (since we have to have not chosen it t times). So we get

$$\mathbb{E}[Z] = n\left(1 - \frac{1}{n}\right)^t \leq ne^{-t/n}.$$

Then plugging in $t = n \log n + cn$, we get $\mathbb{E}[Z] \leq ne^{-t/n} = e^{-c}$. □

So this is just a claim about coupon collector from which we get an upper bound. This example is a bit superficial because if I really want to generate a random string of n bits, I can do it by generating each bit one at a time. But this inefficient algorithm of choosing a random bit and updating it is not too terrible (there's an extra factor of $\log n$, but that's all).

§10.1.2 Example — top-to-random card shuffling

Let's do a more interesting example. Prof. Mossel has a deck of big cards in his office, but forgot to bring them; so he'll simulate shuffling cards by shuffling identical papers of paper which are our feedback forms. But assume they're all different (e.g., we've already written our feedback). We'll do a specific kind of card shuffling which is easy to analyze with stopping times, called *top-to-random shuffling*.

Maybe you're not so good at card shuffling (e.g., doing the riffle shuffle where you have to use two hands), so you do something simple — I take the top card and put it at a random place, then take the top card and put it at a random place again, and so on. I continue until I'm bored, and then I declare that it's shuffled.

Just to be clear, the random place can be *anywhere* — it can be the top, and it can also be under everybody.

So let's try to analyze this example (with n cards).

Whenever we analyze a Markov chain, before we talk about its mixing time, we want to check that it's irreducible and ergodic. If it's not ergodic I'm not going to get a uniform deck, so it had better be ergodic. Why is it ergodic?

Claim 10.5 — This chain is ergodic.

Proof. This is a bit tricky — suppose the top card has to move to place 5, so I put it there. And maybe the second card has to move to place 7; but when we put it there, the top card (which was in place 5) moves up. So we have to be a bit careful.

So what we say is, I start from the top card and put it at the bottom. Then I take the second card and say, where should it be with respect to the card I just put at the bottom? And I put it either directly above it or directly below it (so they're in the right order). Then I look at the third card and put it in the right place with respect to those two. So I can get from any deck to any other deck in at most n steps.

This shows it's irreducible; and it's ergodic because I can always stay where I am.

So the chain is at least ergodic, which means it's not a terrible idea to shuffle my deck this way. □

There's another thing we have to check. How do we know the stationary distribution is uniform?

Remark 10.6. The process isn't actually reversible, so we can't use reversibility — if I move the top card to position 7, that's a positive probability transition, but there's 0 probability of moving backwards. So the chain actually isn't reversible.

By symmetry, you can check that the uniform distribution is stationary (the chain is ergodic, so that's the only stationary distribution). There are many ways to see this — one is to check that the chain is doubly stochastic (if you look at the huge $n! \times n!$ matrix, it's not just that the sum of each row is 1, but the sum of each column is also 1; and that means the uniform distribution is stationary).

So we converge to the stationary distribution, and it's uniform, so we're happy. But now we want to claim that this doesn't take too long.

Goal 10.7. Find a strong stationary time τ (in order to upper bound the mixing time).

The idea is to consider the time when we've touched every card at least once.

Claim 10.8 — Given that at some step there are k cards under the original bottom card, all orderings of these k cards are equally likely.

Here's mentally what I'm doing. Let's suppose we start the shuffle again; and we fold the original bottom card to illustrate. I take the top card and put it at a random place, and then the next card, and so on. At some point I'm going to take the top card and put it under the original bottom card. When I put it under the original bottom card, the claim is that for this card, among all the 1! orderings, they're all equally likely. (This is of course true.) Then I'll perform a bunch more steps, and at some point I'll take the top card and put it under the original bottom card. It's equally likely to go above or below the previous one; so those two orderings are equally likely. Then for the next card that goes under the original bottom card, it'll be equally likely to go in the three possible places, so all 3! orderings will be equally likely.

Eventually, the original bottom card is at the top. We know all the cards under this card are equally likely to be in any of the $(n-1)!$ possible orderings (by induction). Now I have this card and I put it at a random place; so I get one of the $n!$ orderings uniformly at random.

So you can check Claim 10.8 by induction, and it implies the following:

Claim 10.9 — Let τ be one step after the first time the original bottom card makes it to the top. Then τ is a strong stationary time.

This is sort of a more interesting situation. Of course we can generate a random permutation in easier ways if we have a computer. But if you have a deck of cards, you're not going to use a computer — you're going to do something with your hands — and this is a reasonable thing to do.

But we're not done — we have to analyze how long this takes. (I have to wait until this card gets all the way to the top, and then one more step.)

So what do we know about this strong stationary time?

Question 10.10. What can we say about $\mathbb{P}_x[\tau > t]$?

(It doesn't matter what x we start at.)

This is like coupon collector in reverse — to get a card on the bottom, there's only one place we can put it, so the probability that happens is $\frac{1}{n}$. But then we have two places we can put the next card, so that probability is $\frac{2}{n}$. And so on. More precisely, we can write

$$\tau = Y_1 + Y_2 + \cdots + Y_{n-1} + 1$$

where Y_1 is the first time a card gets below the original bottom, Y_2 is (from Y_1) the second time a card gets below the original bottom; and so on.

And what do we know about the Y_i's? This is basically a coupon collector problem — we have $Y_1 \sim \mathsf{Geom}(\frac{1}{n})$, $Y_2 \sim \mathsf{Geom}(\frac{2}{n})$, and so on, and the Y_i are independent. At the beginning I'm trying to do something very difficult — there's only one spot I can hit that's going to work (so I try, try, try and mostly fail). But after that it's a bit easier — there's 2 slots I can put it. So we basically have coupon collector, just in the opposite order (the hardest step is the first step, and then it becomes easier and easier, and at the end it's very easy).

So we get that τ is equal to the coupon collector time on $n-1$ cards, plus 1. Using the bound we had before, we get a similar bound for the mixing time. So this implies for any starting distribution, if $t \geq n \log n + cn + 1$ (we ignore n vs. $n-1$), then

$$\left\|\mu P^t - \pi\right\|_{\mathsf{TV}} \leq e^{-c}.$$

So the analysis ends up being the same, even though the random walk or the process is very different (in one we're updating coordinates, and in the other we're doing this more complicated procedure, but it ends up reducing to the same problem).

Student Question. *When we look at these probability questions, it seems we're only proving this when we're past the expected mixing time, or we're exponentially close. Do we ever care about the early times, when we're really far away from it?*

Answer. That's also interesting. It's a question of what you want. Sometimes it's also interesting to understand what happens at the beginning. For example, maybe you want to know it's not too likely you're in any particular state — maybe we just want that the entropy of the walk is distributed enough. In this course, we'll consider waiting a bit more so you get close to the target. But in physical questions, people often ask, if the process runs for so long, what can we say about it? Maybe we'll be far in total variation, but what else can we say? These are interesting questions, but we won't focus too much on that in this course.

§10.2 Coupling

Now we'll introduce another method of bounding mixing times, which is coupling. We'll first talk about couplings of two random variables, and then couplings of Markov chains.

Definition 10.11. A *coupling* of two probability distributions μ and ν on the same space Ω is a pair of random variables X and Y such that (X, Y) is a random variable on Ω^2 and $X \sim \mu$ and $Y \sim \nu$.

What does the word *coupling* mean? I have two distributions, and I want to see how they behave together. One way to explore that is by looking at a pair of random variables where the first coordinate behaves like μ and the second like ν. There's usually many ways to couple two random variables.

Now we'll do the same for Markov chains:

Definition 10.12. A *coupling* for a Markov chain with transition matrix P is a process (X_t, Y_t) where X_t is a Markov chain with transitions P, and Y_t is also a Markov chain with transitions P; and if $X_s = Y_s$, then $X_t = Y_t$ for all $t \geq s$.

So what is a coupling of Markov chains? A coupling of Markov chains is again creating something that might be correlated, where X_t is run according to the rules of the Markov chain marginally, and so is Y_t. And my goal is eventualy to make them the same. So I have the rule that when they eventually become together, they stay together. (In general coupling allows me to separate them; but when I talk about couplings of Markov chains, once they're together I enforce that they stay together.)

§10.3 Coupling and mixing times

This is a very abstract and general setting, but why does it help us? First of all, it's closely related to total variation distance.

Claim 10.13 — We have $\|\mu - \nu\|_{\mathsf{TV}} = \inf\{\mathbb{P}[X \neq Y] \mid (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}$.

(For a finite space, we can replace inf with min.)

We could compute $\|\mu - \nu\|_{\mathsf{TV}}$ just by computing $|\mu(x) - \nu(x)|$ for all x and summing. This is a much more complicated way to compute the total variation distance — there are infinitely many ways of coupling μ and ν, and the total variation distance will be the probability that X and Y don't agree in the best one. One thing to remember from basic probability is if I just tell you the marginal distributions of X and Y, there's many ways of filling out the table of their joint distribution. This claim says if you look at the best way of filling in that table — if you look at each of them and consider $\mathbb{P}[X \neq Y]$ — the minimum will give you the total variation distance between μ and ν. So I'm looking at all ways of filling in the $\Omega \times \Omega$ table for the joint distribution of X and Y, and looking at all the non-diagonal entries (the ones where $X \neq Y$) and summing them; and the best gives me the total variation distance.

Let's look at a simple example.

Example 10.14

Let $\Omega = \{0, 1\}$. Let $\mu(0) = \mu(1) = \frac{1}{2}$, and let $\nu(0) = \frac{1}{4}$ and $\nu(1) = \frac{3}{4}$.

First let's compute the total variation distance — we have

$$\|\mu - \nu\|_{\mathsf{TV}} = \frac{1}{2}\left(\frac{1}{4} + \frac{1}{4}\right) = \frac{1}{4}.$$

Suppose $X \sim \mu$ and $Y \sim \nu$, and X and Y are independent. That's one coupling; what's $\mathbb{P}[X \neq Y]$ for this coupling? Well, it's

$$\mathbb{P}[X \neq Y] = \mu(0)\nu(1) + \mu(1)\nu(0) = \frac{1}{2} \cdot \frac{3}{4} + \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{2}.$$

So this is not equal to $\frac{1}{4}$; that means that's not the coupling we should take (it's not the one that makes $\mathbb{P}[X = Y]$ as large as possible).

What's a better way to do it? Let's draw a 2-dimensional table (where X is in the rows and Y is in the columns). I can fill in numbers however I want, as long as X individually is distributed according to μ and Y according to ν.

First, $\mu(0) = \frac{1}{2}$ and $\nu(0) = \frac{1}{4}$; I want to put as much as I can at $(0, 0)$, which is $\frac{1}{4}$. Similarly, at $(1, 1)$ I should put as much as I can, which is $\frac{1}{2}$. Then I need to make sure X and Y have the right marginal probability; the marginal probability for X being 0 should be $\frac{1}{2}$, so I should put a $\frac{1}{4}$ in $(0, 1)$.

	0	1
0	$\frac{1}{4}$	$\frac{1}{4}$
1	0	$\frac{1}{2}$

This gives the right marginals for X and Y, so I've found a coupling. And what's $\mathbb{P}[X \neq Y]$? The only thing that contributes is the top-right entry, which is $\frac{1}{4}$; and that's indeed the total variation distance between μ and ν.

So in general, you have to find the right dependent structure between μ and ν — they have to be correlated in the right way — to get the minimum.

We won't prove this claim — we kind of proved it by example, and you can sort of see how to generalize this example to any distributions you want. (You always try to make them agree as much as you can; then there's the constraints of the marginals. Instead, we'll now see how you apply it to Markov chains.)

Claim 10.15 — Let (X_t, Y_t) be a coupling of Markov chains, where $X_0 = x$ and $Y_0 = y$. Then

$$\left\|\delta_x P^t - \delta_y P^t\right\|_{\mathsf{TV}} \leq \mathbb{P}_{x,y}[T > t],$$

where $T = \min\{t \mid X_t = Y_t\}$.

So for any coupling of the chains started from x and y, the total variation distance is at most the probability that they haven't matched. That's how we use couplings — we start from one place x and another place y and find a smart way to run both of them so that they match up as soon as possible, and this gives a bound on the mixing time.

Student Question. *Is there always a coupling that achieves equality?*

Answer. Great question; the answer is no.

We'll see one example; for this, we need a corollary of this claim. Claim 10.15 gives a bound on what happens when I start from x vs. y. How is this related to the mixing time?

Claim 10.16 — If $\left\|\delta_x P^t - \delta_y P^t\right\|_{\mathsf{TV}} \leq \eta$ for all $x, y \in \Omega$, then for any starting distribution μ, we have

$$\left\|\mu P^t - \pi\right\|_{\mathsf{TV}} \leq \eta.$$

We won't prove any of these claims; but we'll do an example, and go over the philosophy. The philosophy is we want to show we get close to the stationary distribution. Coupling says it's enough to show that if you can start from one state and another, there's a way for you to run these two chains in smart, correlated ways such that they get together as soon as possible. And if they get together really fast, you get a bound on the mixing time.

The two chains on their own are just Markov chains. Your creativity is correlating the ways in which these two chains run so that they get together as soon as possible. So that's the game now, and we'll see one example of how to do it.

§10.3.1 Example — lazy random walk on the cycle

Example 10.17

Consider the lazy random walk on the cycle $\{0, \dots, k-1\}$ — so we have 0, 1, …, $k-1$ in a cycle, where with probability $\frac{1}{2}$ I stay where I am, with probability $\frac{1}{4}$ I move left one step, adn with probability $\frac{1}{4}$ I move right one step.

What is my goal right now? One of you is doing this random walk on the cycle, and another of you is also doing so. I want to correlate the ways you're doing the random walks so that you get together as soon as possible.

What would be a good way for us to do this? Lots of things work, but the question is what's nicest to analyze. (Going in opposite directions should work, but it's difficult to analyze.)

One idea is that to avoid the parity issue (where we miss each other), one person is going to stay still and the other is going to move. So we flip a coin and decide who moves, and then the person who moves decides if they move left or right.

So here's the coupling we do (at each step).

- Flip coin #1; this decides who moves (X or Y).
- Flip coin #2; this decides whether the mover moves left or right.

(We have the usual caveat that because this is a coupling of Markov chains, this is only if $X_t \neq Y_t$; if $X_t = Y_t$ then they move together.)

Student Question. *How come we're randomizing who moves and where they move, if the goal was to find the minimum time — can't we find that directly?*

Answer. The minimum time is the minimum over *all* ways of deciding how they move together. That's a huge way; instead of analyzing all possible ways they could move together, we're just suggesting one specific way. And that'll give us a bound on the total variation distance. (In general, finding the actual minimum is hard; we just want to find one coupling and hope it's good.)

How are we going to analyze the time it takes until our two walkers meet? We'll define D_t as the clockwise distance from X_t to Y_t. What can we say about D_t? It's a gambler's ruin — it either goes up by 1 or down by 1 (one of them moves, and they either get closer or farther). So D_t goes up by 1 or down by 1, independently at each step — which means D_t is just gambler's ruin. So letting T be the meeting time, we have

$$\mathbb{P}_{x,y}[X_t \neq Y_t] = \mathbb{P}_{x,y}[T > t] \leq \frac{\mathbb{E}[T]}{t}$$

(by Markov's inequality; you can do better, but we'll just do this). And we computed the expected time for gambler's ruin in one of the first lectures; it's $k(n-k)$, or in this case $D_0(k - D_0)$. So we get

$$\mathbb{P}_{x,y}[X_t \neq Y_t] \leq \frac{D_0(k-D_0)}{t} \leq \frac{k^2}{4t}$$

(since this thing is maximized in the middle, where $D_0 = \frac{k}{2}$).

Student Question. *How is the probability the same as in gambler's ruin — there we had $\frac{1}{2}$ probability of going back and $\frac{1}{2}$ of going forward, but here it's not the same?*

Answer. When we coupled, we said how they move together. The rule for that is we flip a coin, and decide if I move or you move; and then flip another coin to decide whether this person moves left or right. The first thing to check is that both you and I follow the rule of the Markov chain. But now we're looking at a different quantity involving both of us, the distance between us. What happens to

this distance? Suppose I stay and you move; you move either left or right, so you either move closer to me or farther from me. So at each step, this clockwise distance goes up by 1 or down by 1. When the distance gets to 0 or n, that means we've matched and we're going to stay together until the end. So now we're looking at a quantity involving both of us, this clockwise distance; and this clockwise distance actually *is* gambler's ruin.

§10.4 Logistics

Some general comments about the state of the class: there is a new version of the homework. Next week on Thursday there'll be a midterm. This will be completely closed-book. We're not asked to memorize things, but you sort of have to understand stuff. Prof. Mossel only gave this midterm once, during the first semester of COVID; so ours will be harder. His plan is to post the old midterm on Canvas, and maybe ask ChatGPT to give another one which will be similar to his; we'll see how good of a job it'll do.

It'll be during class, next Thursday.

§11 March 13, 2025

Today we'll see more examples of analyzing mixing times; and if there's time left, we'll talk about Markov chains with infinite state space.

§11.1 Example — lazy random walk on the k-level binary tree

We'll look at the lazy random walk on a specific graph, or family of graphs. In the k-level binary tree $G = (V, E)$, we start with one root, and it has two children, and each of those has two children, and so on; and we do this for k levels.

Recall that in a lazy random walk, with probability $\frac{1}{2}$ you do nothing; and if you do something, you go to one of your neighbors, chosen uniformly at random.

This isn't a case where sampling from the stationary distribution itself is difficult — we know the weight of every vertex should be proportional to its degree. But we'd still like to see how this random walk behaves.

We have $|V| = 2^k - 1$; and this chain is ergodic, and the stationary distribution is proportional to the degree sequence.

How might we prove upper or lower bounds on how long the chain takes to get to the stationary distribution?

§11.1.1 Lower bounds

For lower bounds, we'll use the bottleneck theorem. The idea is we'll take L to be the half of the vertices on the left, R to be the half of the vertices on the right, and B to be the root.

The bottleneck theorem tells us there exists x such that

$$\left\|xP^t - \pi\right\|_{\mathsf{TV}} \geq \pi(B \cup L) - t \cdot \frac{\pi(B)}{\pi(R)}.$$

Here $\pi(B \cup L)$ is clearly at least $\frac{1}{2}$ (L alone would be almost $\frac{1}{2}$, and with B it's a bit bigger). And what can we say about $\pi(B)/\pi(R)$? This is

$$\frac{\pi(B)}{\pi(R)} = \frac{\deg(\text{root})}{\sum_{v \in R} \deg(v)}$$

(because the stationary distribution is the normalized degree sequence). The degree of the root is 2; and $\sum_{v \in R} \deg(v)$ should be roughly twice the number of edges in the right subtree. In any tree, the number of edges is roughly the same as the number of vertices (minus 1); so we're going to get something like

$$\frac{1}{2} - t \cdot \frac{2}{2^{k-1} - 10}$$

(the 10 is not important or exact, but it's certainly a bound). So to get 0.1-close to π, we need

$$t \geq 0.001 \cdot 2^{k-2} \geq cn$$

(for instance). So you need at least a constant times n steps in order to get 0.1-close in total variation distance (for instance).

(Since this is just an upper bound, it doesn't matter if we compute these things directly, as long as we're undercounting or overcounting in the right direction.)

§11.1.2 Upper bounds

How about an upper bound — how would we prove that you don't need *too* long to sample from this random walk?

We could try to do a coupling. One suggestion is we start by moving one at a time, until they get to the same level; and after that, you move them together.

So here's a coupling approach to upper-bounding the mixing time: Suppose X_0 is started at some vertex x, and Y_0 at some vertex y. The basic coupling procedure is the following. Coupling means we run the two chains simultaneously such that each individually does the right thing, and we do it in a smart way so that they get together.

Case 1 (They're at the same level). Then we make the same move type for them (up, down, or stay — so if this one goes up then that one does too, and so on).

Case 2 (They're not at the same level). Then we toss a coin to decide which moves and which stays. (We know each individually should stay with probability $\frac{1}{2}$, so we toss a coin to decide which stays and moves.) This is the same philosophy as last time — we want them to eventually be at the same level, so we don't want them to swap locations; and if we just make one of them move, then that should eventually happen.

So how do we analyze the time until they meet?

> **Student Question.** *When they're not at the same level, what would be the problem with both moving down at the same time?*
>
> **Answer.** It's not really a problem. When you define a coupling, you can do many things — you could say that if they both want to move down, they can both move down. There's a slight issue when one is at the bottom and the other isn't — then it'll want to go down and the other can't.
>
> But also philosophically, if they always moved down and up together, they'd never get to the same level (if they didn't start there), so they'll never meet. So philosophically this isn't a good idea. What we want is that if they're not at the same level, we sort of make it volatile by moving one of them, so that they eventually do get to the same level.

Goal 11.1. Bound the time $\tau = \min\{t \mid X_t = Y_t\}$.

(Note that τ isn't a number — it's a random variable — and we want to say it's typically not too large.)

First, by symmetry we can assume without loss of generality that X_0 is closer or at the same distance to the root as Y_0 (because their roles are indistinguishable).

The trick here is the following claim.

Claim 11.2 — We always have $\tau \le \tau'$, where $\tau' = \min\{t \mid Y_t = \text{root}\}$.

The point is Y_0 starts below X_0. Now, we made sure they can't swap — the one that starts below always stays below (or equal). At some point when it gets to the root, it has to be above (or equal), so they have to be equal at that point. They might have been equal before that point, but they're definitely equal at that point.

So even though our coupling was somewhat complicated, to analyze it we just have to analyze *one* of the random walks. This is kind of cool — the definition of the coupling involved both random walks, but we can analyze the stopping time we're interested in just by analyzing *one* of the walks.

So now it's just a question about one random walk — I start one random walk on a tree, and I want to know how long it takes until I get to the top.

Claim 11.3 — We have $\mathbb{E}[\tau'] \le 6n$.

(The constant 6 is not super important.)

Proof sketch. The idea is it's kind of like gambler's ruin, where you have some probability of going up and some of going down. It's a bit different because we have this lazy thing, but it's gambler's ruin where we stay with probability $\frac{1}{2}$, go up with probability $\frac{1}{6}$, and go down with probability $\frac{1}{3}$. Then we can write a function — if ℓ is the level, we can write $f(\ell)$ for the expected time to get to the root from level ℓ. And then we can write some equations — $f(0) = 0$, and in general

$$f(\ell) = \frac{1}{2}f(\ell) + \frac{1}{3}f(\ell+1) + \frac{1}{6}f(\ell-1) + 1.$$

This is for $0 < \ell < k$; and at the last level you have

$$f(k) = \frac{1}{2}f(k) + \frac{1}{2}f(k-1).$$

You solve these equations (we're not going to do it), and you see that $f(k) \le 6n$. □

Why does this make us happy? This implies that for any starting distribution μ, we have

$$\left\|\mu P^t - \pi\right\|_{\mathsf{TV}} \le \mathbb{P}[\tau \ge t] \le \mathbb{P}[\tau' \ge t] \le \frac{\mathbb{E}[\tau']}{t} \le \frac{6n}{t}.$$

So the lower bound from bottleneck told us that we needed some linear number of steps; and if we want the total variation distance to be less than 0.1 we maybe need to take $t = 60n$. Maybe 60 isn't the right constant, but $60n$ is still linear in n. So the upper and lower bounds match up to the constant — the number of steps you need in order to mix is linear in n.

§11.2 Example — random transposition card shuffling

This time Prof. Mossel brought his deck of cards (which he never uses except in classes because these cards are huge). This is a slightly more complicated probabilistically, but mathematically it's nicer (it's a random walk on groups with nice generators).

We put all the cards on the table (we don't actually do this because they're too big, but pretend we did). Then we choose one card uniformly at random, choose another uniformly at random, and swap them. (My two hands don't know each other; it's possible that the two cards I choose are the same, in which case nothing happens.)

Question 11.4. How long does it take for this card shuffle to get close to the stationary distribution?

§11.2.1 Lower bounds

Maybe it feels like if we haven't touched some card, it shouldn't be close to stationary. But we have to be careful. With random walks on the hypercube, we found an upper bound saying after we've refreshed all the coordinates, you're stationary. But actually, you don't have to touch all the coordinates. In fact, by the time you've actually touched all but $\sqrt{n}$ of the coordinates, you're mixed. So you *can* do something like this, but you have to be careful.

We'll think about the deck of cards as a permutation. So let's start from the identity permutation. And the quantity we're going to look at is closely related to the above suggestion — let's look at the number of fixed points, i.e.,

$$\#\{i \mid \sigma(i) = i\}.$$

When we start with the identity permutation, there are n fixed points — everything is a fixed point.

Question 11.5. When we look at a *random* permutation, how many fixed points are there?

On average, there's 1 (you can analyze it in more detail, but this will be enough for us) — under π, we have

$$\mathbb{E}[\text{fixed points}] = \sum_i \mathbb{P}[i \text{ is a fixed point}] = \sum_i \frac{1}{n} = 1$$

(since if you look at a particular i, it's equally likely to map to any of the values).

So we start with something that has n fixed point, and we end with something with expected 1 fixed point. How can we use this to prove a lower bound?

We'll define an event $A = \{\sigma \mid \sigma \text{ has at least 2 fixed points}\}$.

Claim 11.6 — We have $\pi[A] \le \frac{1}{2}$.

This is just Markov's inequality.

Claim 11.7 — If $t \le \frac{n}{2} - 4$, then $\mathbb{P}_{\text{id}}[X_t \in A] = 1$.

At each point, we're updating at most two guys; so this means there's 4 (or maybe 8) coordinates I haven't been able to touch. At each step I can touch at most 2 coordinates, so I can destroy at most 2 fixed points. So if I wait until $\frac{n}{2}$ minus a little bit, I'll still have a bunch of fixed points (in particular, there will be at least 2).

(This means the probability starting from the identity.)

This conclusion means that for $t \leq \frac{n}{2} - 4$, we have

$$\left\|\pi - \delta_{\mathrm{id}} P^t\right\|_{\mathsf{TV}} \geq \mathbb{P}_{\mathrm{id}}[X_t \in A] - \pi(A) \geq \frac{1}{2}.$$

So we definitely need at least $\frac{n}{2}$ steps (maybe a little less).

Can anyone improve on this argument? This argument is correct, so that's a good start. But can we use the same philosophy to get a better bound?

This reminds us of coupon collector. It's kind of a different game where each day I get 2 coupons; and we said in order to collect n coupons, I need at least $\frac{n}{2}$ days. But actually, we saw that typically we need $n \log n$ days (in the ordinary coupon collector problem). So we can make this a bit stronger.

Claim 11.8 — If $t \leq \frac{n}{2} \log \frac{n}{2} - 10n$, then $\mathbb{P}_{\mathrm{id}}[X_t \in A] \geq 0.9$.

This we actually didn't analyze — what's the chance that I have a bunch of coupons left (rather than 1)? We won't do the complete proof because it's a small variant of what we did before, but the proof follows by analyzing coupon collector. It's slightly different than the usual coupon collector because we don't want to know when we got all the coupons; we want to know when we got all the coupons but two. But it's basically the same analysis.

And this means for this value of t, we have

$$\left\|\pi - \delta_{\mathrm{id}} P^t\right\|_{\mathsf{TV}} \geq 0.4.$$

§11.2.2 An upper bound by coupling

So these are some good ideas for the lower bound; now let's talk about ideas for the upper bound.

We have two proofs for upper bounds. The reason Prof. Mossel is thinking right now is which subset to give. He can give us the empty set, a somewhat sophisticated upper bound using coupling (which is suboptimal), or a more sophisticated one using strong stationary times (which is nearly optimal).

We maybe have time for both the proofs, so we'll start with the coupling proof.

The main idea of the coupling proof is to actually think about a different encoding of this Markov chain. So the first thing we have to do is think about this chain a bit differently, and understand why it's the same chain.

Here's an equivalent description of the Markov chain: At time t, I choose a value $V_t \sim \mathsf{Unif}\{1, \ldots, n\}$. Independently, I choose a position $P_t \sim \mathsf{Unif}\{1, \ldots, n\}$. So I'm choosing two uniform numbers; and what I'm going to do is I switch the card with face value V_t with the card in position P_t.

This sounds very weird — of course when I shuffle the cards, I don't know the values of them (they're face-down). But I'm telling you I'm going to choose the card that has value 3 (by looking at all of them), and then switch it with the one in position 5. Of course that's not how I implement it in real life. But it's the same process — all I'm doing is choosing one card uniformly at random and another uniformly at random, and swapping them. The fact I decided to do it in this weird way (where I'm describing the first card by value rather than position) doesn't matter.

The reason we're doing it in this weird way is that this will help us with the coupling.

So now I want to couple two Markov chains. The idea is very simple — I'm going to use the same value V_t and the same position P_t for both chains.

So now I have two decks of cards. (In reality, our two decks are disjoint, but let's imagine they're copies of the same deck of cards.)

Now I'm going to choose value 3 in the first deck and in the second, and I'm going to put it in position 5 in both decks (swapping it with whatever was originally there).

Why does this make me happy? Because when I did this, I put value 3 in the same place in both decks — I made sure that the value 3 in the two chains is now going to be at the same place.

So that sounds like a good idea. Is there any way the two permutations I have will agree *less* after a step like this, or can they only agree more? They can only agree more, so we just have to understand when they'll agree more.

Claim 11.9 — Let a_t be the number of cards at the same position under this coupling. Let τ_{i+1} be the number of shuffles between the first time with $a_t \geq i$ and $a_t \geq i+1$. Then

$$\tau_{i+1} \sim \mathsf{Geom}\left(\frac{(n-i)^2}{n^2}\right).$$

There's something a bit annoying about the notation here — I'm not writing that the number of cards agreeing *is* i or $i+1$. This is because we're moving two cards, so the number of places they agree can jump buy 1 or 2. (It could be that I have i cards, and then in the next step I have $i+2$.)

(The proof of this is also going to show why it never gets worse.)

Proof. If the cards with value V_t are in the same position (originally), then there'll be no change in a_t. (This is because if the special cards I chose are in the same location in the two deck, then we're going to move them to position 5, and the things in position 5 to this position, so there's no change in the number of agreements.)

Otherwise, if the cards in position P_t are the same, then it's a similar argument — if not this card is the same, but the card at the other position I chose, then there's also no change in a_t.

But if the location of this card is different, and what I had under that position is different, then I have to improve by at least 1 — so otherwise a_t improves by at least 1.

So we have to get a difference. To get a difference is hard; and how hard? It has a chance of $\frac{n-i}{n}$ (the more cards I have in common, the harder it is to get a difference). And I have to get a difference in both the first card and the second; the chance of that is $(\frac{n-i}{n})^2$. □

So what does that imply? We have a sum of geometric random variables, so we know what the expectation is — if τ is the coupling time, then we know

$$\mathbb{E}[\tau] \leq \sum_i \mathbb{E}[\tau_i] = n^2 \sum_{i=1}^{n-1} \frac{1}{i^2} \leq \frac{\pi^2 n^2}{6}$$

(we have $\sum \frac{1}{i^2} = \frac{\pi^2}{6}$ by some magical formula from calculus).

By Markov, this implies that $O(n^2)$ steps gets you close to th estationary distribution. If you wait a hundred billion times $\frac{\pi^2 n^2}{6}$ steps, you'll be at distance at most 1 over a hundred billion from stationary (this is the usual application of Markov that we've seen).

§11.2.3 A better upper bound by strong stationary times

Prof. Mossel sort of gave us a hint of what's going to happen — we found a lower bound of order $n \log n$, and we found an upper bound of order n^2 and said it's not optimal; so what's going to happen now is we're going to work harder to get $n \log n$.

So now the goal is to find a better upper bound using strong stationary times. The strong stationary time is going to be somewhat involved. It'll involve marking the cards — this sounds very bad when you talk about cards, but this is what we're going to do.

So we'll do the following marking procedure: I start with no marked cards. Now everything is position — I'm not looking at values — and I'll think of the two cards I choose as my left hand and right hand (they don't talk to each other, so they could be the same card). So let R_t and L_t be the cards chosen by my right and left hand at step t. And here's how I'm going to mark (a card that's marked will stay marked forever): I mark R_t if it's currently unmarked and one of two things happens: Either L_t is marked, or $L_t = R_t$.

So what should happen for me in order to mark the first card? I start with no marked cards; I choose my right card and left card and can't do anything, and so on. The first time I mark my first card is when the right equals left. So it's going to be very hard to mark at the beginning. But once cards start being marked, it becomes much easier (all I need is that the left hand touches a markec card).

Claim 11.10 — The time when all cards are marked is a strong stationary time.

So by the time I've marked all the cards, I really have a random deck.

This is quite tricky — it's a paper from 30 or 40 years ago, and it's not obvious even after the fact.

Proof. We'll show that the set of all marked cards is in a random order at all steps. So we'll prove this by induction — that at every point in time, if we restrict to the set of marked cards, they come in a uniform random order.

At the first step, there's no marked cards, so they're in a random order.

What about the next step? Nothing happens for a while. Suddenly the two hands agree; this is one card, so obviously it's in a random order.

What happens when I get the next card? A bunch of things could happen — either the right hand has to agree with the left hand on some other card, or the right hand should be there and the left hand on *this* card. In the first case they stay in the same order, and in the second they swap orders; this means the two permutations are equally likely.

So I'm marking this card if two events happen — the left hand agrees with the right hand on this card, or the left hand is on the originally marked card. Each agrees with probability $\frac{1}{2}$, so staying in the same order or swapping is equally likely.

Now what happens on the inductive step? Say I have $k-1$ marked cards, and consider what happens when the next card gets chosen. Then there are k ways for it to get marked — I choose one of the previously marked cards, or this card itself. These are all equally likely, so I'm swapping it with a uniform guy including itself, which means I get a uniform order. □

So now we have to analyze this marking procedure. How do we find this strong stationary time?

Claim 11.11 — Let τ be the the time when we've marked all cards. Then $\mathbb{E}[\tau] \le 10n \log n$ and $\mathrm{Var}[\tau] \le 10n^2$.

If you believe this claim, it implies that

$$\left\|\mu P^t - \pi\right\|_{\mathsf{TV}} \le \frac{c}{x^2} \quad \text{if } t = 10(n \log n + cx).$$

So the total variation distance shrinks fast once you're past some constant $n \log n$ (where c is some constant).

We're not going to prove this — it follows from the usual fact that the total variation distance is bounded by $\mathbb{P}[\tau > t]$, and for this we have the expected value and variance and Chebyshev's inequality. So we won't do this computation; instead we'll talk a little bit about how we bound this τ.

The idea is that last class, we had inverted coupon collector. Here we also have something like inverted coupon collector — the hardest thing is getting the first marked card (with probability $1/n$). Then we need to get a second marked card. For this, what should happen? Either it's like before, or I have the two hands choosing different cards; so this is slightly bigger probability. What'll happen is lots of the probability comes from the left card being one of the ones already marked, and the right card being one of the others. So you see these are some geometric random variables.

At least, unlike the previous case here we mark at most one card each time, so there are no jumps of 2. So we can write

$$\tau = \tau_0 + \tau_1 + \cdots + \tau_{n-1}$$

where τ_i is the time from marking the ith card to the $(i+1)$st.

Claim 11.12 — We have $\tau_i \sim \mathsf{Geom}(\frac{(i+1)(n-i)}{n^2})$, and the τ_i are independent.

It's clear that these times are independent — once I marked i cards, that tells me nothing about what'll happen. Why this probability?

Proof. We have to check the scenarios where the two agree and are unmarked, or the left is marked and the right isn't. The first case is $\frac{n-i}{n} \cdot \frac{1}{n}$, and the second is $\frac{i}{n} \cdot \frac{n-i}{n}$. So the success probability is

$$\frac{1}{n} \cdot \frac{n-i}{n} + \frac{i}{n} \cdot \frac{n-i}{n}.$$

To repeat the logic, there's one scenario where we choose one of the unmarked cards and choose it again; the other scenario is we choose one of the cards that is marked, and another that is not marked. □

And from here it's analysis of independent random variables; we know how to compute the expectation and variance, so we get this claim.

Remark 11.13. On Tuesday we'll do some kind of review, by request.

§12 April 1, 2025

§12.1 Continuous-time Markov chains

We're still going to talk about Markov chains now, but we'll spend the next few lectures talking about *continuous-time* Markov chains. The basic picture that you should have in mind for continuous-time Markov chains is that we have some time axis, and maybe we're at some state 1; and then at another random time we're going to move to state 3, then at another random time we'll move to 7, then so on. These times are no longer integers (as in the discrete-time setting) — they might be 0 and then π and then $\sqrt{2}$, or so on.

A standard way to define this kind of Markov chain uses Poisson processes. What we'll do today is review Poisson processes (and maybe also generalize them, depending on how quickly this goes). In order to do that, we'll start by reviewing exponential random variables and Poisson random variables. So that's the plan.

We're going to use a bunch of facts, so Prof. Mossel will put the facts on the left board, and the proofs on the other boards.

§12.2 Exponential random variables

Definition 12.1. We say T is a *exponential random variable* with parameter λ, written $T \sim \mathsf{Exp}(\lambda)$, if T is always positive and

$$\mathbb{P}[T > t] = e^{-\lambda t} \quad \text{for all } t \geq 0.$$

There's a bunch of properties (some of which are on the slides):

Fact 12.2 — The density of T is given by $f_T(t) = \lambda e^{-\lambda t}$ (for $t \geq 0$).

We can also compute a bunch of moments:

Fact 12.3 — We have $\mathbb{E}[T] = 1/\lambda$ and $\mathrm{Var}[T] = 1/\lambda^2$.

We're not going to rederive these; hopefully we've seen these in 18.600 (for example).

Because we're going to talk about processes, we also want to understand what happens with *sums* of exponential random variables.

Fact 12.4 — If T_n is the sum of n i.i.d. $\mathsf{Exp}(\lambda)$ random variables, then its density is given by

$$f_{T_n}(t) = \lambda e^{-\lambda t} \cdot \frac{(\lambda t)^{n-1}}{(n-1)!} \quad \text{for } t \geq 0.$$

Proof. If we have a sum of a bunch of i.i.d. random variables, how do we derive the density? The buzzword is *convolution* — the density of a sum of independent random variables can be written as the integral of one density times the other — but one is at s, and the other is at $t - s$.

We'll prove this by induction. As a sanity check, for $n = 1$ there's just one exponential; the extra factor on the right is just 1, so we get $\lambda e^{-\lambda t}$, which we said is the density.

So now we'll do the induction step. Now we want to compute $f_{T_{n+1}}(t)$. This has to be a sum of two things, which we'll call s and $t - s$, and both of these have to be positive (since all the random variables we're talking about are positive). So s goes from 0 to t, and we'll get

$$f_{T_{n+1}}(t) = \int_0^t f_{T_n}(s) f_{T_1}(t-s)\, ds.$$

(*Convolution* is what you want to do when you want to compute the density of a sum of independent random variables. In general you have some integral from $-\infty$ to ∞ of the density of one guy at s, times the density of the other at $t - s$. Here the random variables are positive, so instead of going from $-\infty$ to ∞ we just go from 0 to t, because outside this one of the two terms is 0.)

Now we can plug in our formula for f_{T_1} and the induction hypothesis, and we get

$$\int_0^t \lambda e^{-\lambda s} \frac{(\lambda s)^{n-1}}{(n-1)!} \cdot \lambda e^{-\lambda(t-s)}\, ds.$$

Then we can do the integral; the $\lambda e^{-\lambda t}$ comes out of the integral, so we get

$$\lambda e^{-\lambda t} \int_0^t \lambda \cdot \frac{(\lambda s)^{n-1}}{(n-1)!}\, ds.$$

And then we just have the integral of s^{n-1}, so we get another $1/n$ factor; so we get

$$\lambda e^{-\lambda t} \cdot \frac{(\lambda t)^n}{n!},$$

which is what we wanted. □

We'll use this fact today. Another fact used a lot about exponential random variables, one of the reasons they're so popular, is what's called the 'lack of memory.

Fact 12.5 (Lack of memory) — If $s, t > 0$, then for $T \sim \mathsf{Exp}(\lambda)$, we have

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

Why the name? Maybe in life you're waiting for some airplane to drop a suitcase full of money from the sky; this is your exponential random variable, and it has some distribution. And you're asking, given that I've waited 50 years by now, what's the chance that I'll have to wait at least 10 more years? And this is just exactly the original probability (that I'd have to wait 10 years) — it's not that I get rewarded for being patient.

(This is just Bayes' law — we have $\mathbb{P}[T > s + t] = e^{-\lambda(s+t)}$ and $\mathbb{P}[T > s] = e^{-\lambda s}$, and you divide them and get $e^{-\lambda t}$.)

Here's something you might not have seen. (You can do this more generally, but we'll just do it with one race.)

Fact 12.6 (Exponential race) — Suppose $S \sim \mathsf{Exp}(\lambda)$ and $T \sim \mathsf{Exp}(\mu)$ are independent. Then:

(a) Letting $U = \min(S, T)$, we have $U \sim \mathsf{Exp}(\lambda + \mu)$.

(b) We have $\mathbb{P}[S < T] = \frac{\lambda}{\lambda+\mu}$ (and $\mathbb{P}[T < S] = \frac{\mu}{\lambda+\mu}$).

(c) If I is the identity of the winner, then I and U are independent.

So we have two independent random variables — for example, the time until I get a suitcase falling into my hands, and the time until I sneeze. Then it turns out their minimum is also exponential, and we can compute the probability that each wins (it's proportional to their parameters).

So I have these two independent random variables and I want to understand how they play with respect to each other — who wins, how long do I have to wait until someone wins, and what can I tell about these two things? First we look at this marginally — we look at $\min(S, T)$ and ask, what kind of random variable is it? And it's exponential. Then we ask, what's the chance S vs. T is the first event to happen? The answer is proportional to λ vs. μ (S wins with probability $\frac{\lambda}{\lambda+\mu}$). And the last thing is, suppose I know it happens now. Does that change the probability I sneezed vs. got a suitcase full of money? The answer is no. If I tell you the time at which it happens, the ratio of the two probabilities remains the same — it doesn't tell you anything. You might think that if I wait very long it's more likely to be the suitcase, but it doesn't work that way — the ratio of the two probabilities is the same, no matter what time it happens.

Proof. First, for (a), we can use the definition — we have

$$\mathbb{P}[\min(S, T) > t] = \mathbb{P}[S > t \text{ and } T > t] = \mathbb{P}[S > t]\mathbb{P}[T > t] = e^{-\lambda t}e^{-\mu t} = e^{-(\lambda+\mu)t}$$

(since they're independent), which is exactly what we'd expect from an exponential random variable with parameter $\lambda + \mu$. (That was the definition — to be an exponential random variable with parameter $\lambda + \mu$, I should have $\mathbb{P}[U > t] = e^{-(\lambda+\mu)t}$.)

Now for (b), we want $\mathbb{P}[S < T]$. So I'll go over all possible values of S; and then I want to look at the density that S is there, and I want T to be bigger than that. So we have

$$\mathbb{P}[S < T] = \int_0^\infty \lambda e^{-\lambda t}\mathbb{P}[T > t]\, dt.$$

And then we can just compute that this is

$$\int_0^\infty \lambda e^{-\lambda t}e^{-\mu t}\, dt = \lambda \int_0^\infty e^{-(\lambda+\mu)t}\, dt = \frac{\lambda}{\lambda + \mu}.$$

Now let's do (c). This is actually similar. We're not used to doing proofs like this, but here's what we're going to do: Let f be the density of the random variable $U = \min(S, T)$, but restricted just to the set where $S < T$. (We have to compute the density on the set where $S < T$, and on the set where $S > T$.)

What happens for this density? There's the density of S, which is $\lambda e^{-\lambda t}$. And then the only thing I know about T is that it's bigger than this, which has probability $e^{-\mu t}$. So our joint density on this set is

$$f = \lambda e^{-\lambda t}\mathbb{P}[T > t] = \lambda e^{-\lambda t}e^{-\mu t} = \frac{\lambda}{\lambda+\mu} \cdot (\lambda+\mu)e^{-(\lambda+\mu)t}.$$

Why does this make us happy? The first term is $\mathbb{P}[S < T]$, and the second term is the density of U. So the density on the set $S < T$ is the product of the probability $S < T$ times the density of U where we don't say anything; and that shows that they're independent. □

Remark 12.7. The conceptual reason that this kind of proof is hard is there's some part here that's discrete, and some that's continuous — I is discrete (it takes values either that S wins or T wins) and U is continuous. We're maybe not used to thinking about what it means for a discrete random variable to be independent from a continuous one. But that's what we have here, and this is one of the ways to check it.

Remark 12.8. As a comment (which we're not going to prove): This generalizes to races with n exponentials. Here there were two — either I'm sneezing, or the airplane drops something nice. Maybe it's either of those, or I break my leg, or I solve the Riemann hypothesis; where there's all kinds of independent events. Then we have the same thing — the minimum time any of them happens is the sum of parameters; the probability any one of them wins is proportional to its parameter; and who wins and the time at which it happens are independent. You can do this by induction, or you can use the same proof.

§12.3 Poisson random variables

That's all we need about exponential random variables before we talk about Poisson processes; but we also have to talk about Poisson random variables. For these we'll say a little less, but we'll still say something.

Exponential random variables take real (positive) values; Poisson random variables take *integer* values.

Definition 12.9. A random variable N is *Poisson* with parameter λ, written $N \sim \mathsf{Poisson}(\lambda)$, if for every $n = 0, 1, 2, \ldots$ we have

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

You can compute a bunch of things (which we're not going to do):

Fact 12.10 — We have $\mathbb{E}[N] = \lambda$ and $\mathrm{Var}[N] = \lambda$.

You can compute other moments too; we're not going to do that right now. One claim we will use a lot:

Fact 12.11 — If $N_1 \sim \mathsf{Poisson}(\lambda_1)$ and $N_2 \sim \mathsf{Poisson}(\lambda_2)$ are independent, then

$$N_1 + N_2 \sim \mathsf{Poisson}(\lambda_1 + \lambda_2).$$

So if you have two independent Poissons, their sum is Poisson (with parameter the sum of their two parameters). This generalizes (e.g., by induction) to k independent Poisson random variables.

§12.4 Poisson processes

Now we'll give one of the possible definitions of a Poisson process. The *Poisson process* with *intensity* λ is defined using independent $\mathsf{Exp}(\lambda)$ random variables τ_1, τ_2, τ_3, ... There's two ways to define the Poisson process. One is to define

$$T_n = \sum_{i=1}^{n} \tau_i$$

(where $T_0 = 0$); we call this the *nth arrival time*. There's another way to look at this — we can consider the parameter

$$N(s) = \max\{n \mid T_n \le s\},$$

which is called the *number of arrivals* until time s.

Both of these objects (the collections T_n and $N(s)$) describe the same thing; they're two descriptions of the same object.

As a picture, imagine we have our time axis, and then a point 0; and then we draw an $\times$ (where the length of this interval is τ_1). Then we draw another $\times$, where the length of the interval between them is τ_2. Then we draw another, where the length between them is τ_3. And so on.

We have $T_0 = 0$; then T_1 is the time when the first thing happens, which is τ_1; then T_2 is the time when the next thing happens, which is $\tau_1 + \tau_2$; and so on.

If we look at $N(s)$, imagine I draw a function — so we draw a y-axis plotting $N(s)$ in red. We'll have $N(s) = 0$ from 0 to the first $\times$ (because nothing happens); then up to the next one it'll be 1; and then from then on it'll be 2; and so on. This tells me how many arrivals have happened — at the beginning it's 0 arrivals, then there's 1, and so on.

These describe the same thing. If I know the T_n, then I can define $N(s)$. And conversely, if I know the graph of $N(s)$, then I know when it jumps, so I know when I had an arrival. So these describe the same thing; they're two different ways to look at the same object.

Let's derive some basic properties of this process.

Claim 12.12 — Fix $s \ge 0$. Then:

(a) We have $N(s) \sim \mathsf{Poisson}(\lambda s)$.

(b) We have that $N(t+s) - N(s)$ is a Poisson process with parameter λ, which is independent of the Poisson process up to time s (i.e., it's independent of $(N(r) \mid r \le s)$).

Again, let's look at the graph. What this says is if I fix a time s and look at how many arrivals I expect to have up to time s, it'll be Poisson; that's the first part. The second part says, suppose there was a bunch of stuff happening, but then I just came into the room, and I want to know what happens from this point on. The second part says that what happens from now on is itself a Poisson process, and it's independent of what came before.

We'll prove the first part of the claim, and say something about how you prove the second part.

Proof of (a). How do we prove that the number of arrivals up to time s is a $\mathsf{Poisson}(\lambda s)$ random variable? I want to somehow compute $\mathbb{P}[N(s) = n]$. And we somehow have to do this in terms of the τ_n's or T_n's. What does this mean? This means the nth arrival happened before s, and the $(n+1)$st happened after — so we have

$$\mathbb{P}[N(s) = n] = \mathbb{P}[T_n \le s,\ T_n + \tau_{n+1} > s].$$

And now what? It's the same philosophy — we're going to integrate over the density of T_n, which we computed earlier (since it's the sum of n independent exponentials). So we get

$$\int_0^s \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} \cdot \mathbb{P}[t + s_{n+1} > s]\, dt$$

(the first factor is the density that T is at t; and then given that we're here, we want the probability that $T_n + \tau_{n+1} > s$, and T_n is now t).

And we can rewrite the latter factor as $\mathbb{P}[\tau_{n+1} > s - t]$, so we can rewrite this as

$$\int_0^s \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} \cdot e^{-\lambda(s-t)}\, dt$$

(since τ_{n+1} is itself just an exponential). Now everything has to miraculously cancel, so let's see that it does — we get $e^{-\lambda s}$ which comes out of the integral, the $e^{-\lambda t}$ and $e^{+\lambda t}$ cancel; and so we get

$$e^{-\lambda s} \int_0^s \lambda \cdot \frac{(\lambda t)^{n-1}}{(n-1)!}\, dt = e^{-\lambda s} \cdot \frac{(\lambda s)^n}{n!}.$$

And that's what we wanted — that's the probability that a Poisson random variable with parameter λs takes the value n. □

Proof sketch of (b). The proof of (b) — Prof. Mossel doesn't know a very nice way of writing it, so he'll just say it in words. The proof of (b) follows from the memoryless property of exponentials. The basic idea is I got to this point s; I know that some exponential started to tick (waiting to arrive), and it didn't arrive so far. Memorylessness tells me that given that it hasn't arrived so far, it doesn't matter when it started — the time until it arrives is still going to be exponential with parameter λ. So this next time is going to be exponential with parameter λ; and so is the next; and so on. □

We'll write one more (somewhat abstract) theorem about Poisson processes, though we won't prove it. Prof. Mossel doesn't think we'll use it, but it's just for our information.

Theorem 12.13

If $N(s)$ is the number of arrivals in a Poisson process with parameter λ, then:

(1) $N(0) = 0$.

(2) $N(t+s) - N(s) \sim \mathsf{Poisson}(\lambda t)$ for all t.

(3) $N(t)$ has 'independent increments' — i.e., if $t_1 \le t_2 \le t_3 \le \cdots$, then $N(t_2) - N(t_1)$, $N(t_3) - N(t_2)$, ... are independent random variables.

Conversely, if (1), (2), and (3) hold, then $(N(s))$ is a Poisson process with parameter λ.

First, (1) is obvious. In (2), we're looking at the number of arrivals between $t + s$ and s; and this number is a Poisson random variable.

So in terms of the picture that we just erased, if I look at the number of arrivals in one interval, then the next, and so on, then these numbers are all Poisson, and they're independent Poisson random variables.

The interesting part of the theorem is that the converse holds — if we have some process with $N(0) = 0$, $N(t+s) - N(s) \sim \mathsf{Poisson}(\lambda t)$, and the process has independent increments, then it's actually Poisson. So this is an if and only if — if you're a Poisson process then you satisfy (1), (2), and (3), while if you satisfy (1), (2), and (3) then you're a Poisson process.

We actually basically proved the forwards direction (that Poisson processes satisfy (1), (2), and (3)); we're not going to prove the reverse direction, but it's not too hard.

Remark 12.14. Note that the statement $N(t+s)-N(s)$ is a Poisson process is stronger than the claim that $N(t+s)-N(s) \sim \mathsf{Poisson}(\lambda t)$. The second statement is about what happens when we *fix* t (it's about a single random variable), and the first is about the entire process. So we've actually proved something stronger than (2); but for the converse it's actually good to have the weaker property (because if we just have this (and the other statements), then actually we can conclude you're a Poisson process).

§12.5 The binomial process

Prof. Mossel's dilemma is that he doesn't know if he wants to tell us about the binomial process or not. Why would you want to know about the binomial process? Prof. Mossel's experience as someone who learned about Poisson processes and taught them many times is that they're confusing, and one way to get intuition for them is the binomial process. But on the other hand, you have to upload another model to your brain. We don't need the binomial process per se; it's just to give you intuition for the Poisson process.

We vote to hear about it, so Prof. Mossel will tell us about binomial processes.

Again, why is he telling us about the binomial process? The basic philosophy is that the Poisson process is 'equal' to the limit as $n \to \infty$ of the binomial process. What's nice about the binomial process is that there's nothing continuous — there's no real numbers, and everything is discrete.

Here's the binomial process. There's two parameters, n and λ. What we're going to do is the following:

- We divide the interval $[0,\infty)$ into intervals of length $1/n$.
- We drop a ball in each interval independently with probability λ/n.

So everything is discrete. We have our interval $[0,\infty)$, and we have these tiny sub-intervals; and in each interval I either have a ball or don't (but the chance of actually having a ball is pretty small — it's λ/n, which as $n \to \infty$ becomes tiny).

What is the connection? Now I can ask, for example:

Question 12.15. What is the number of balls in the interval $[0,t]$?

What kind of random variable describes this number of balls? It's like coin tosses, so it's binomial. How many experiments do we have in $[0,t]$? We partitioned $[0,t]$ into intervals of length $1/n$, so our number of trials is nt (we're cheating a bit since this might not be an integer, but we won't worry about that), and the probability of success is λ/n. So this is $\mathsf{Binom}(nt, \lambda/n)$. And by the Poisson approximation, as $n \to \infty$ this converges to $\mathsf{Poisson}(\lambda t)$.

So we've discretized time into tiny intervals and said that the probability of success on each tiny interval is the same everywhere.

Here's another question:

Question 12.16. What's the time of the first arrival (i.e., ball)?

This is a geometric random variable — we're having experiments, and we're waiting until we have a success. So it's geometric with parameter λ/n. We also want to scale time — this is the number of trials we need until we get to success, but we want to measure real time, and each trial takes time $1/n$; so this random variable is $\frac{1}{n}\mathsf{Geom}(\frac{\lambda}{n})$. And as $n \to \infty$, this converges to $\mathsf{Exp}(\lambda)$. (We'll check this in a little bit.)

So what this is saying is that whenever you're confused, you can think about the binomial process. And that's just coin tosses, so we can understand everything — we just have questions like how long does it take

to go from one head to the next? And then we take the limit and that gives us a way of thinking about the Poisson process, which is confusing.

How do we check this fact?

Fact 12.17 — As $n \to \infty$, we have $\frac{1}{n}\mathsf{Geom}(\frac{\lambda}{n}) \to \mathsf{Exp}(\lambda)$.

Proof. We'll just try to plug in the definition of the exponential — what's $\mathbb{P}[\frac{1}{n}\mathsf{Geom}(\frac{\lambda}{n}) > t]$? We can write this as $\mathbb{P}[\mathsf{Geom}(\frac{\lambda}{n}) > nt]$. What it means for a geometric to be bigger than some number means that I've failed for that many steps; so this is

$$\left(1 - \frac{\lambda}{n}\right)^{nt}.$$

And using the calculus approximation $1 - x \approx e^{-x}$ (when x is tiny), this converges to

$$e^{-\lambda/n \cdot nt} = e^{-\lambda t},$$

which is what we wanted. □

One reason the Poisson process is so important is that it's the limit of the discrete binomial processes, which are very natural.

§12.6 Compound Poisson processes

Now we'll start a new topic. We're almost ready to define continuous-time Markov chains. We'll start with a simpler thing — something that's called a *compound* Poisson process.

Definition 12.18 (Compound Poisson process)**.** Consider a Poisson process with intensity λ, and i.i.d. random variables $Y_1, Y_2, \ldots$. The *compound process* is given by $S(t) = \sum_{i=1}^{N(t)} Y_i$ (where $N(t)$ is the number of arrivals of the Poisson process).

Let's draw a picture. So we have our Poisson process, with a $\times$ at the point of the first arrival, then the second, and so on. And at each of these points, something interesting is happening — we're getting a Y_i. And we're summing these. So we plot $S(t)$ on the y-axis in red.

At first, nothing happens, so I have 0. Then at this first $\times$ something happens, and I get one random variable Y_1; so I take that value. Then at the next $\times$ I get another random variable Y_2, so I sum that (and now I'm at $Y_1 + Y_2$). And so on.

Why would I study something like this — what could it model? Maybe it's really coming from something like telephone or shopping online — t is the time when the user buys something on Amazon, and Y_i is how much money they spent. Maybe I'm Amazon and want to know how much money the user has spent up to time t. For some time nothing happens, and then someone comes in and spends some money, and now I have money Y_1. And then nothing happens and then something more happens, and now I have money $Y_1 + Y_2$. So this describes stuff happening at random times.

Next time we'll discuss these processes again, and we'll talk also about Poisson filling and related things.

§13 April 3, 2025

Today we'll continue talking about Poisson processes, with the goal of talking about continuous-time Markov chains next week.

§13.1 Sums of random numbers of i.i.d. random variables

We'll talk about compound Poisson processes. But first, we'll state a theorem that we'll use:

Theorem 13.1

Let N be an integer-valued random variable, and let Y_1, Y_2, ... be i.i.d. random variables which are independent of N. Let $S = \sum_{i=1}^{N} Y_i = Y_1 + \cdots + Y_N$.

(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 $\mathbb{E}[N] < \infty$, then $\operatorname{Var}[S] = \mathbb{E}[N]\operatorname{Var}[Y_i] + \operatorname{Var}[N]\mathbb{E}[Y_i]^2$.

This is a sort of interesting situation. The usual situation in probability is that we're summing independent random variables. But here, the *number* of variables that we sum is random on its own. The first statement is maybe intuitive — the expected value of a sum is the sum of expectations, and here the number of summands N is random, so you'd expect to multiply by $\mathbb{E}[N]$. Maybe the second statement is less intuitive. Each guy Y_i has its own variance, so you multiply by the number of terms you get, which gets the first term. But there's also additional variance coming from the random number of terms you sum, and that's what the second term is.

Proof of (1). We want to compute $\mathbb{E}[S]$. First, if we know what N is, then we know how to compute $\mathbb{E}[S]$; so we'll first condition on the sum. And then using the tower property of conditional expectation, we can write

$$\mathbb{E}[S] = \mathbb{E}[\mathbb{E}[S \mid N]].$$

And we know how to describe $\mathbb{E}[S \mid N]$ (since now N is a constant) — this is just going to be $N\mathbb{E}[Y_i]$, so

$$\mathbb{E}[S] = \mathbb{E}[N\mathbb{E}[Y_i]].$$

And now $\mathbb{E}[Y_i]$ is a number, so it comes out of the expectation, and we get $\mathbb{E}[S] = \mathbb{E}[N]\mathbb{E}[Y_i]$. □

Proof of (2). We're going to use the law of total variance, which is like the tower property of the conditional expectation: This states that

$$\operatorname{Var}[S] = \mathbb{E}[\operatorname{Var}[S \mid N]] + \operatorname{Var}[\mathbb{E}[S \mid N]].$$

(You can prove this by just doing the algebra.) In our case, $\operatorname{Var}[S \mid N]$ is the variance of a sum where we've fixed the number of summands. Then we just sum the variances, so this is $N\operatorname{Var}[Y_i]$, which means the first term is $\mathbb{E}[N\operatorname{Var}[Y_i]]$. And we said earlier that $\mathbb{E}[S \mid N] = N\mathbb{E}[Y_i]$. So we get

$$\operatorname{Var}[S] = \mathbb{E}[N\operatorname{Var}[Y_i]] + \operatorname{Var}[N\mathbb{E}[Y_i]].$$

The terms $\operatorname{Var}[Y_i]$ and $\mathbb{E}[Y_i]$ come out, because they're just numbers; so we get

$$\operatorname{Var}[S] = \mathbb{E}[N]\operatorname{Var}[Y_i] + \operatorname{Var}[N]\mathbb{E}[Y_i]^2$$

(since when you pull out a constant from the variance, you get a square). □

§13.2 Compound Poisson processes

This is a nice theorem. Let's see why we're mentioning it — one reason is that we have these compound Poisson processes, whose definition we'll now recall.

Definition 13.2. A *compound Poisson process* is defined by

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

where $N(t)$ is a Poisson process with intensity λ and the Y_i are i.i.d.

Remember a Poisson process has times when stuff happens, and at the times when stuff happens, we're summing some size of interactions; and we want to understand how this quantity is behaving.

We can apply the above theorem because the Y_i are i.i.d., and $N(t)$ is an integer random number that's independent of the Y_i. So for each t we can apply the theorem to say that:

Claim 13.3 — We have $\mathbb{E}[S(t)] = \mathbb{E}[N(t)]\mathbb{E}[Y_i] = \lambda t \mathbb{E}[Y_i]$.

(Recall we saw last class that $N(t) \sim \mathsf{Poisson}(\lambda t)$, so its mean is λt.)

The variance is similar:

Claim 13.4 — We have

$$\operatorname{Var}[S(t)] = \mathbb{E}[N(t)] \operatorname{Var}[Y_i] + \operatorname{Var}[N(t)]\mathbb{E}[Y_i]^2 = (\lambda t) \operatorname{Var}[Y_i] + (\lambda t)\mathbb{E}[Y_i]^2 = \lambda t \mathbb{E}[Y_i^2].$$

(The variance of a Poisson is the same as its expected value, so $\operatorname{Var}[N(t)]$ is also λt; and in the last equality we're using the fact that $\operatorname{Var}[Y_i] = \mathbb{E}[Y_i^2] - \mathbb{E}[Y_i]^2$ to write this more compactly.)

§13.3 Poisson thinning and superposition

Now we're going to do something a bit different with compound Poisson processes, called *Poisson thinning*.

Definition 13.5. Consider a compound Poisson process (with intensity λ) where $Y_i \in \{1, \dots, k\}$. For $1 \le a \le k$, define $N_a(t) = \#\{i \le N(t) \mid Y_i = a\}$.

Before, we thought of the Y_i's as amounts of money. But now maybe we're thinking about exploding stars, so 1 means a small star, 2 a medium star, and 3 a huge star; so now the Y_i's take integer values.

Example 13.6

We can draw a one-dimensional picture for this: we draw our time axis starting at 0, and we draw our Poisson process with a bunch of $\times$'s on the axis. And each of these points has one of three colors (maybe with probabilities 1/2, 1/3, and 1/6); maybe the first is red, then green, then red, then yellow, then green, green, red, yellow. So for each point independently, I choose its color.

If we put t all the way on the right, then $N(t)$ is the total number of points, which is 8. But we'll also have $N_{\text{yellow}}(t) = 2$ (counting the number of yellow points), and similarly for the other colors.

The interesting claim here is:

Claim 13.7 — For each $1 \le a \le k$, we have that $N_a(t)$ is a Poisson process with intensity $\lambda \mathbb{P}[Y_i = a]$. Moreover, these processes are independent.

So the process of points colored by color a is itself; and its intensity is the original intensity λ times the probability that $Y_i = a$.

Maybe in this example we had here, maybe the probabilities of red, green, and yellow were $1/2$, $1/3$, and $1/6$. Then what this is telling us is that $N_{\text{yellow}}(t)$ is a Poisson process with new intensity $\lambda/6$.

So each of them individually is a Poisson process with a smaller parameter (λ times the probability of seeing this point).

And the second part is maybe more surprising — that these processes are *independent*! What does this mean? Maybe I have big stars exploding, and medium stars exploding, and small stars exploding. If I saw 100 small stars exploding last night, you might think maybe we had lots of stars exploding, so you might expect that there were lots of medium or big stars exploding. But that's actually false — information about the small stars exploding doesn't tell you anything about the medium ones. Similarly, here the yellow process tells you nothing about the red ones.

Student Question. *What if there's only two colors? If you know all the events are red, doesn't that mean there's no blue events?*

Answer. They're not independent conditioned on $N(t)$ — if I tell you the number of events and all the red events, then you know the number of green events. But if I just tell you what the red events are (and not the total number), that doesn't tell you anything about the green events.

We won't give the complete proof, but we'll give the gist of the proof — we'll do a proof sketch for the case $k = 2$.

Proof sketch for $k = 2$. Here we have two colors, so $Y_i \in \{1, 2\}$; and for the proof, we want to compute

$$\mathbb{P}[N_1(t) = j \text{ and } N_2(t) = k]$$

(the probability that the number of events of the first type up to t is j, and the second is k). To make the notation simpler, let $\mathbb{P}[Y_i = 1] = p$, so $\mathbb{P}[Y_i = 2] = 1 - p$.

So now I have small stars and big stars. What's the probability I have j small stars exploding and k big stars exploding? We can write this as

$$\mathbb{P}[N(t) = j + k] \cdot \binom{j+k}{j} \cdot p^j (1-p)^k$$

(since we need $j + k$ total explosions of stars; and then we know these explosions are independent, so we need j of them to be small stars and k to be big stars, and we get this). Now let's write this in a more ugly way — $N(t)$ is a Poisson random variable, so we get

$$e^{-\lambda t} \frac{(\lambda t)^{j+k}}{(j+k)!} \cdot \frac{(j+k)!}{j!\,k!} \cdot p^j (1-p)^k$$

(usually we don't like unpacking binomial coefficients, but here we'll do this because we see a $(j+k)!$, which looks a bit like part of a binomial coefficient). And now the point is that I want to write this so that it's a product of two independent things. The first thing should be $e^{-\lambda p t}$ (because I want a Poisson process not of rate λ, but of rate λp); and the second should be $e^{-\lambda(1-p)t}$. These two terms together are going to give me the $e^{-\lambda t}$. I want to get

$$e^{-\lambda p t} \frac{(\lambda p t)^j}{j!} \cdot e^{-\lambda(1-p)t} \frac{(\lambda t(1-p))^k}{k!}.$$

And we can check that this gives the above thing (we have λ^j and λ^k, p^j and $(1-p)^k$, and so on, so it all works out). And the point is that this is

$$\mathbb{P}[\mathsf{Pois}(\lambda pt) = j] \cdot \mathbb{P}[\mathsf{Pois}(\lambda(1-p)t) = k],$$

which is what we needed.

We're not going to do the full proof, but in principle, if we wanted to, what's missing? (This is the gist of the proof, but what do I have to do if I want to be super formal?) In principle, it could be that I'm telling you last night I had a lot of explosions of small stars, and two weeks ago I had a lot of explosions of small stars, and tonight I had a few; what does that tell you about big stars? So in principle, we have to consider multiple values of t, not just one. Also, here we just looked at what happens on the interval $[0, t]$; but really, we have to look at all different intervals and show that what happens there are independent, and also independent of the other process. That ends up being a lot of computations like this, but they're all of the same flavor as this one, so we won't do them. □

This process is called *thinning*. Now we'll go in the other direction. Suppose you are studying big stars, and your labmate is studying green stars, and another labmate is studying yellow stars; and the PI who is very important is studying all stars. And I say my process is a Poisson process with this parameter, and your friend says theirs is a Poisson process with some other parameter, and so on; and they're independent. And the PI thinks very hard and says, the total process is going to be a POisson process with the sum of parameters.

Here we started with one process and labelled all the points with colors, and said each individual process was an independent Poisson process. But we can also go in the other direction — we can take a bunch of independent Poisson processes and take their union — and we again get a Poisson process.

Definition 13.8. Let $N_1(t)$, $N_2(t)$, …, $N_k(t)$ be independent Poisson processes with parameters λ_1, λ_2, …, λ_k. Their *superposition* is given by $N(t) = N_1(t) + \cdots + N_k(t)$.

So we're just looking at all the points I see of all the types — I'm taking a union of these processes.

Claim 13.9 — The process $N(t)$ is a Poisson process with intensity $\lambda_1 + \cdots + \lambda_k$.

Proof. We'll use the fact that when we sum independent Poisson random variables, we get a Poisson whose parameter is the sum of their individual parameters. In principle, we have to do this for every interval. So let's consider

$$N(t+s) - N(s) = \sum_{i=1}^{k} (N_i(t+s) - N_i(s)).$$

This is a sum of k independent Poissons, so it's a Poisson with parameter $t(\sum_i \lambda_i)$. That's one part — we have to argue that in each interval of length t, we get some Poisson with parameter proportional to the length of the interval.

Second, we need to argue that if $[a_1, b_1]$, …, $[a_n, b_n]$ are disjoint intervals, then $N(b_1) - N(a_1)$, …, $N(b_n) - N(a_n)$ are independent random variables. Why? We have a picture where I'm drawing two intervals, and maybe I'll draw two colors — so I have the green process, and I have the red process. What I'm saying is that what I have on the left and right are both Poissons proportional to the interval lengths. And we want to argue that they're independent. Why? The N in the left interval $(N(b_1) - N(a_1))$ is the green $N(b_1) - N(a_1)$ plus the red one; and similarly for the right interval. And these things are all independent of each other; since we're summing independent random variables, they're going to stay independent. □

Remark 13.10. We did the easier proof (the second one) in more completeness just to show the level of annoyance you need to do to actually prove statements like this.

§13.4 An example

Let's see some examples. (This isn't a complicated example; it's just sort of to practice this notion.)

Example 13.11

Consider three independent Poisson processes with intensities 3, 2, and 1. What is the probability that in the first 6 arrivals, there are exactly 3 points from the first process, 2 point from the second, and 1 point from the third?

Let's think a little about a concrete example. In astronomy, imagine you have your astronomers, and the more advanced one is studying small stars, the next advanced is studying medium stars, and the least advanced is studying big stars (because those are rarest). You're eating lunch and you ask yourself, in the next 6 stars that explode, what's the chance that I observe half of them, you observe a third of them, and the last person observes a sixth?

This is kind of tricky because the time of these explosions is random. An easier problem to solve is the question of in the next week, what's the chance that I observe some number of explosions, you observe some number of explosions, and the third person does? These are independent, so we can multiply those numbers and be happy. But we're not looking at a fixed time period here.

One option is to ask what's the chance that this occurs up to time t, and then integrate. But can we do this without an integral?

We're going to use this dual description of thinning and superposition. So what we're going to say is we'll think about it in the following way. An equivalent description of the problem (from what we've learned) is that there is a Poisson process with rate $3+2+1=6$ (the sum of the rates), where each point is of type I, II, or III with probabilities $1/2$, $1/3$, and $1/6$ (respectively). That's an equivalent way to describe the process. Here we were originally thinking about it as three independent process; but we said you can go back and forth between looking at the union of independent processes and superposition and thinning.

So then this probability is just the question, among 6 points, what's the chance I have three of the first type, two of the second, and one of the third? This is just

$$\binom{6}{3,2,1} \cdot \left(\frac{1}{2}\right)^3 \left(\frac{1}{3}\right)^2 \left(\frac{1}{6}\right).$$

I don't have to condition on time, because each star that comes is going to be small with probability $1/2$, medium with probability $1/3$, and big with probability $1/6$. So this is a useful way to look at Poisson processes that's often used.

§13.5 Conditional independence of arrival times

Now we'll see one more fact about Poisson processes, which is again intuitive and useful, but maybe still a little surprising. We'll start from some intuition.

Now I just have one Poisson process, and I'm looking at some time t.

As a silly question, if I know there's 0 points here (in $[0,t]$), then I know there's 0 points; that's not interesting.

Now, if I tell you $N(t)=1$ (so there's only one point here), where do you expect this point to be in the interval? You'd expect it to be uniformly distributed.

What if I tell you $N(t)=2$? Then you'd still expect these points to be independent (where the first is the first point, and the second the second point).

The philosophy is in the binomial process, I have a bunch of tiny cells; $N(t) = 1$ means I just have one ball, so each cell is equally likely. If $N(t) = 2$, now I have two balls, and all the combinations should be equally likely.

That's what we're going to prove now — that if you condition on the number of balls, their locations are uniform and independent.

Claim 13.12 — Let T_1, T_2, ... be the arrival times of a Poisson process with intensity λ. Let $U_1, U_2, \dots, U_n \sim \mathsf{Unif}[0,t]$ be i.i.d., and let $V_1 < V_2 < \cdots < V_n$ be the U_i in increasing order.

Then if we condition on $N(t) = n$, then the distribution of $(T_1, \dots, T_n)$ is equal to the distribution of $(V_1, \dots, V_n)$.

So we have these random points, but we put them in increasing order — V_1 is the smallest of the U_i's, V_2 is the next-smallest, and so on.

This is still a complicated distribution (somewhat) — if I want the distribution of just T_1, I'm supposed to go to my computer, generate independent uniform numbers at random, and look at the smallest one. If I want the joint distribution of (T_1, T_n), I generate my n independent random variables and say T_1 is the smallest and T_n is the largest. So T_1 on its own is not uniformly distributed — it's going to be closer to the beginning of the interval — but it's going to be closer in the same way as if I drop uniform n points, the smallest of them is going to be closer to the beginning of the interval.

We can do two proofs. The proof in the slides is a proof with densities, but we can also do a proof with just the binomial process. (There's a little ugliness with the latter because of issues with having two balls in the same set.)

Proof. We want to show something is uniform; so we're going to compute the conditional density that $T_1 = t_1$, $T_2 = t_2$, ..., $T_n = t_n$ (where $t_1 < \cdots < t_n \le t$), where we condition on $T(t) = n$. First we're going to do the conditioning part. This means we divide by the probability of having exactly n points up to time n, which is

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

And then we have to compute the density. What's the density? For each point, the chance that the next point is at time t is given by an exponential density, so we get $\lambda e^{-\lambda(t_i - t_{i-1})}$ (given that the last point was at t_{i-1}, what's the probability the next is at t_i)? And then we also have to multiply by the probability that there's no points after the last one, which is $e^{-\lambda(t-t_n)}$. So we get

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

Why did we write this ugly expression on the board? Everything cancels out — we have all these exponentials, but they telescope, so eventually we're going to get an $e^{-\lambda t}$ from them. And then we have λ^n from the product, and we're dividing by that. So in the end, we're just left with

$$\frac{n!}{t^n}.$$

And why does this make us happy? This expression doesn't use any information about $t_1, \dots, t_n$ — it's the same for all the points — so it's uniform (among all the feasible points). So we've shown that the density is constant, so the distribution is uniform (on the set $\{(t_1, \dots, t_n) \mid t_1 \le t_2 \le \cdots \le t_n \le t\}$). □

Remark 13.13. If you think a little bit about why we know if I hvae a uniform distribution on this, the density is $n!/t^n$ — if I ignored the order and just looked at $0 \le t_1, \dots, t_n \le t$, then the volume would be t^n. And here I have a constraint that I'm looking at one particular order, so I get $1/n!$ of this volume.

So i we let $A = \{(t_1, \dots, t_n) \mid t_1 \le \cdots \le t_n \le t\}$, we can note that

$$\text{Vol}(A) = \frac{1}{n!}\text{Vol}([0,t]^n) = \frac{t^n}{n!}.$$

This means the uniform distribution on A has density $\frac{n!}{t^n}$ for all points in A.

§13.6 Non-homogeneous Poisson processes

Maybe in the last few minutes, we'll talk about something which we might not see in this class a lot (but for those of us who are going into finance, it's useful there, so we'll mention it).

Definition 13.14. We say $N(s)$ is a *Poisson process* with rate $\lambda(r)$ of:

(1) $N(0) = 0$.

(2) $N(t+s) - N(s) \sim \text{Poisson}(\int_s^{t+s} \lambda(r)\,dr)$.

(3) $N(t)$ has independent increments.

So here the intensity λ is not fixed — the rate changes with time. And to see how many events I have on some interval, I integrate the rate over that interval; and it should be Poisson with that parameter.

Let's see some basic properties of these non-homogeneous Poisson processes. (You can study these things very extensively, but we'll just talk about some basic properties.)

Our original definition of Poisson processes was that we had an exponential random variable and waited for it, and then we had an arrival. And then we had another exponential random variable we waited for, and another arrival. Now we don't even have a clear description of what is an arrival, so let's do that.

Claim 13.15 — Let τ_1 be the first arrival time. Then its density is given by

$$f_{\tau_1}(t) = \lambda(t)\exp\left(-\int_0^t \lambda(s)\,ds\right).$$

Proof. The density is the derivative of the CDF, so

$$f_{\tau_1}(t) = \frac{d}{dt}\mathbb{P}[\tau_1 \le t] = -\frac{d}{dt}\mathbb{P}[\tau_1 > t] = -\frac{d}{dt}\mathbb{P}\left[\text{Pois}\left(\int_0^t \lambda(r)\,dr\right) = 0\right].$$

And the probability that a Poisson random variable is 0 is just $e^{-\lambda}$ (where λ is its parameter), so this is

$$-\frac{d}{dt}e^{-\int_0^t \lambda(r)\,dr}.$$

And we know how to take the derivative (the exponential stays there, and you apply the fundamental theorem of calculus and get a $-\lambda(t)$ factor). □

One more question about this process:

Question 13.16. What is the joint density $f_{\tau_1,\tau_2}(t_1, t_2)$?

One reason to find this joint density is to ask, are τ_1 and τ_2 independent?

Proof. This is going to be similar — we start with the density for the first arrival, which is $\lambda(t_1)\exp(-\int_0^{t_1}\lambda(r)\,dr)$. And what's the next term going to be? Then I have a Poisson where I'm starting at t_1 and going up to t_1+t_2; so this is going to be $\lambda(t_1+t_2)\exp(-\int_{t_1}^{t_1+t_2}\lambda(r)\,dr)$. (This is the same logic — we want the first Poisson to be 0 and take the derivative, and then we want the second to also be 0 and take the derivative of that.)

From this expression, are τ_1 and τ_2 independent or not? They're not — there's no way to factorize this as a function of t_1 and a function of t_2. □

Next time we'll actually define continuous time Markov chains.

§14 April 8, 2025

Today we'll finally start talking about continuous-time Markov chains.

§14.1 Continuous-time Markov chains

Definition 14.1. Let P be the transition matrix of a finite discrete Markov chain, and let $T_0 < T_1 < T_2 < \cdots$ be the arrival times of a Poisson process with rate λ. Let $Y_0, Y_1, \ldots$ be the discrete-time Markov chain. Then the corresponding *continuous-time Markov chain* (X_t) (where $t \in \mathbb{R}_{\geq 0}$) is defined by

$$X_t = Y_k \text{ if } T_k \leq t < T_{k+1}.$$

We have two building blocks — one is a discrete-time Markov chain, and the other is this Poisson point process. This is one way to define continuous-define Markov chains; we'll see at least a couple.

We already understand discrete-time Markov chains very well. So now we'll draw a picture. We horizontally draw $\mathbb{N}$, the time axis for the discrete-time Markov chain (we draw dots horizontally labelled 0, 1, 2, 3, 4); and we draw states on top of these (Y_0, Y_1, Y_2, Y_3, Y_4). So this is the discrete-time Markov chain.

For the continuous-time one, now we'll have a real axis with real numbers. We'll look at a Poisson process, and write down T_1, T_2, T_3, T_4. What we're saying is that in the first interval (up to T_1) it takes the value Y_0, then in the second interval it takes Y_1, then in the next it takes Y_2, and so on.

So instead of times being discrete (0, 1, 2, ...), there are *random* times when stuff happens; and these random times are defined by a Poisson point process.

This definition is in some sense nice because it's minimal — once we have Poisson processes and discrete-time Markov chains, we can define this. But it doesn't seem organic in some way. So let's try to give a second definition which is maybe more organic, and then we'll try to see that they're the same.

First, we need to define what's a *rate matrix*.

Definition 14.2. We say Q is a *rate matrix* if $Q(x,y) \geq 0$ for all $x \neq y$, and $Q(x,x) = -\sum_{y\neq x} Q(x,y)$.

So rows sum to 0, and all the non-diagonal entries are nonnegative.

Definition 14.3. The *continuous-time Markov chain defined by rate matrix Q* is defined in the following way: Given that $X_t = x$ (for some real number t), we let $\tau_{x,y} \sim \mathsf{Exp}(Q(x,y))$ for all $y \neq x$. At time $t + \min_{y\neq x} \tau_{x,y}$, we transition to the state y minimizing $\tau_{x,y}$ (i.e., we transition to the state $\arg\min_{y\neq x} \tau_{x,y}$).

What does the second definition say? Here there's nothing discrete; this is more of an engineering definition. I now know I'm at a time t, and my state is X_t. And from X_t I want to move to different places, and there's a competition between all the places I might want to move to. Here we have to draw what Q is. Maybe $X_t = 0$, and Q looks like

$$\begin{bmatrix} -3 & 2 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix}.$$

So there's two things I might want to do — I might want to move to state 1 or to state 2, corresponding to exponential random variables with parameters 1 and 2 (respectively).

So now I'm going ot have exponential random variables $\tau_{0,1}$ and $\tau_{0,2}$. Because $\tau_{0,1}$ is smaller (in our picture), we ignore the other one (and completely forget about it); and at this point $t + \tau_{0,1}$, we move to $X_t = 1$.

Maybe the next row of the matrix is $(3, -5, 2)$. Then I want to go to either state 0 or 2, with exponential random variables with parameters 3 and 2; so we again compete between them, and whoever's the winner I move there.

So at each transition of the Markov chain, there's an exponential race; the winner is where I go to, and the time I transition is decided by this winner.

§14.2 Equivalence of definitions

In math, when we give two definitions, we claim they're equivalent; so that's what we're going to do next. We'll spend some time now to understand why these definitions are equivalent. Informally, the bottom line is that we'll show the two definitions 14.1 and 14.3 are equivalent. But one of them talks about Q and the other about P, and in one there's a λ-Poisson process, and so on, so we have to see how they match. We'll show that they're equivalent with $Q = \lambda(P - I)$.

We'll have two claims, for the two different directions.

Claim 14.4 — Consider Q as in Definition 14.3. Let $\lambda = \max_x -Q(x,x)$, and let

$$P(x,y) = \frac{Q(x,y)}{\lambda} \text{ if } y \neq x \quad \text{and} \quad P(x,x) = 1 - \sum_{y\neq x} P(x,y).$$

Then the chain in Definition 14.3 agrees with the chain in Definition 14.1.

We're starting from one parametrization instead of Q, and telling you how it should be parametrized in terms of λ and P.

Proof. What do I have to show in order to prove this claim? We're starting with one definition (the more engineering one, where I'm at a state x and there are all these places I want to move to, and I go to the minimum); and there's the other definition where we start with a discrete-time Markov chain and embed it in random times. How do we check that with this parametrization, we get the same thing?

We'll check that given that $X_t = x$, we transition in the same way in both chains. So let's see, what are the two chains?

According to Definition 14.3, we have independent $\tau_{x,y} \sim \mathsf{Exp}(Q(x,y))$; the time to transition is $\min_{y\neq x} \tau_{x,y}$, and the state we transition to is the minimizer y.

According to Definition 14.1, how long does it take me to transition? There's going to be an $\mathsf{Exp}(\lambda)$, which is the time to transition. And if we transition, then we're moving to state y with probability $P(x,y)$. But this could also mean that we don't transition — we'll have a probability of $1-P(x,x)$ that we transition, and otherwise we don't transition. So let's try to normalize this — we transition with probability $1-P(x,x)$, and then transition to y with probability

$$\frac{P(x,y)}{1-P(x,x)}.$$

So if we actually want to look at the time we transition, we're doing Poisson thinning — sometimes we transition and sometimes we don't. The chance we're transitioning is $1-P(x,x)$. So overall, we get that by Poisson thinning, the time to transition is

$$\mathsf{Exp}(\lambda\cdot(1-P(x,x))),$$

and the transition probabilities are given by $\frac{P(x,y)}{1-P(x,x)}$ where $x\neq y$.

Now there is some computation we have to check. What do we have to check? First of all, we have to check that the time to transition is the same — that $\min_{y\neq x}\tau_{x,y}$ is in fact exponential with parameter $\lambda(1-P(x,x))$. And then we have to check that given that we transition, the transition probabilities are the same.

First, why is $\min_{y\neq x}\tau_{x,y}$ exponential? We know that it's exponential with parameter $\sum_{y\neq x}Q(x,y)$, which is $\mathsf{Exp}(-Q(x,x))$. So now we have to check that this is equal to $\lambda(1-P(x,x))$, and that's correct.

So the transition rate is the same in the two cases; now we have to check that the probability of transitioning to each state is the same. So we need to check that for $y\neq x$, the given probability of transition is the same. The easiest way to do this is to look at the ratios. For the chain in Definition 14.1, we have

$$\frac{\mathbb{P}[\text{transition to } y']}{\mathbb{P}[\text{transition to } y]}=\frac{P(x,y')}{P(x,y)}=\frac{Q(x,y')/\lambda}{Q(x,y)/\lambda}=\frac{Q(x,y')}{Q(x,y)},$$

where the latter is the ratio according to Definition 14.3. So once I switch, the ratios of probabilities is the same; and that means I get the same chain. □

We actually sort of proved both directions here — we wrote this in terms of going from Definition 14.3 to 14.1, but the same computation lets you go from Definition 14.1 to 14.3. We'll just write the claim:

Claim 14.5 — Given P and λ as in Definition 14.1, define $Q=\lambda(P-I)$. Then the chain from Definitions 14.1 and 14.3 are equivalent.

This is essentially the same proof (this is the easier direction; it's more direct). Once I have λ and P, I have the picture I just erased, where I have points on the x-axis; the next point comes by a Poisson point process with parameter λ, and which place I move is given by P. It's essentially the same proof, so we're not going to do it again.

§14.3 The heat kernel

Now we want to understand — like in discrete-time Markov chains, we want to quickly understand what's happening with our chains in the long-run. For this, we'll define something like the t-time transition probability.

Definition 14.6. The *heat kernel* of a continuous-time Markov chain is defined by

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

Remark 14.7. Prof. Mossel doesn't like the notation H because this isn't a random variable, it's just a function — we usually use capital letters for random variables, but $H_t(x, y)$ is just a number.

So if you fix t, this is a matrix telling you, if I start from x and wait for t units of time (where t is continuous — a real number), what's the chance I get to state y? (The notation $\mathbb{P}_x$ means that we're starting at $X_0 = x$.)

We'd like to understand these matrices, so that we can talk about limits and stationarity.

Claim 14.8 — We have $H_t = e^{tQ}$.

For those of us who haven't seen exponents of matrices, there are many ways of defining it. One way, which should be good enough for us, is to just think about Taylor expansions.

Definition 14.9. For a matrix A, we define $e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!}$.

For us, the matrix is tQ.

Student Question. *For the case of an infinite state space, are these chains still equivalent, and does this claim make sense?*

Answer. Under some additional conditions, yes; without these additional conditions, no (there can be some explosions in the rate). But the answer is basically yes unless something crazy happens.

Proof. We want to prove $H_t = e^{tQ}$; how would we do this? We have to somehow get an infinite expansion; how would we do that?

There are many choices. We're going to use Definition 14.1, and we're going to condition on $N(t)$ (which is the number of arrivals, or the number of transitions, up to time t).

So we're going to look at

$$\mathbb{P}_x[X_t = y] = \sum_{k=0}^{\infty} \mathbb{P}[N(t) = k] \cdot \mathbb{P}_x[X_t = y \mid N(t) = k]$$

(so I'm breaking the probability space based on how many transitions I have).

Remark 14.10. There's a continuous time Markov chain of where are the erasers — there's a continuous state space of where they are, and somehow they always tend to accumulate on the right, which gives us more time to think.

Let's think about what these terms are. The first just comes from the Poisson point process — its parameter λt, so $N(t) \sim \mathsf{Pois}(\lambda t)$. And the second comes from the discrete-time Markov chain; so we get

$$\sum_{k=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!} P^k(x, y).$$

Now we can do some basic algebra to rearrange this to

$$e^{-\lambda t} \sum_{k=0}^{\infty} \frac{(\lambda t P)^k}{k!}(x, y)$$

(I've just pulled the P inside the parentheses). And this should be familiar — without the (x, y), that matrix is just $e^{\lambda t P}$. So we see that

$$H_t = e^{-\lambda t} e^{\lambda t P}$$

(as a matrix — because for every (x, y), its (x, y)th entry is given by the Taylor expansion of the right-hand side at (x, y)). And if you want to be fancy, you can write this as

$$e^{-\lambda tI} e^{\lambda tP} = e^{\lambda t(P-I)} = e^{tQ}$$

(since $Q = \lambda(P - I)$). □

§14.4 Stationary distributions

Now that we have this, maybe we can talk about stationary distributions.

Definition 14.11. We say π is *stationary* for Q if $\pi H_t = \pi$ for all t.

So if I start from π and go for t time, I get π — that's the stationarity property I had before, except that now it's not for a discrete number of steps, but a continuous time.

Claim 14.12 — A distribution π is stationary for Q if and only if $\pi Q = 0$, if and only if $\pi P = \pi$.

We're always using the same parametrization in our mind where $Q = \lambda(P-I)$. So this says you're stationary for the continuous chain under some algebraic condition, and that's the same algebraic condition we had for the *discrete* Markov chain — you're stationary for the discrete chain if and only if you're stationary for the other. There's really no difference (except the algebra is written a bit differently).

Proof. The second part (that $\pi Q = 0$ if and only if $\pi P = \pi$) is easy — $\pi Q = 0$ is the same as $\pi\lambda(P-I) = 0$. And $\lambda > 0$, so this is the same as $\pi(P - I) = 0$, which is the same as $\pi P = \pi$. So this is just simple algebra.

How about the first part (that π is stationary for Q if and only if $\pi Q = 0$)?

One direction is easy — suppose $\pi Q = 0$. Then if we look at

$$\pi H_t = \sum_{k=0}^{\infty} \pi \frac{(tQ)^k}{k!},$$

there are two terms here. One is the $k = 0$ term, which gives us π. And then we have

$$\sum_{k=1}^{\infty} \frac{\pi(tQ)^k}{k!}.$$

And the point is that $\pi Q = 0$, so $\pi Q^2 = \pi Q^3 = \cdots = 0$. So all these terms are equal to 0, which means we just get $\pi H_t = \pi$. So this is one direction — if we know $\pi Q = 0$, then we get that it's stationary.

What about the other direction? (Prof. Mossel gives us the challenge of doing it without taking derivatives or using eigendecompositions or anything; he doesn't know how to.)

For the other direction, suppose that $\pi H_t = \pi$ for all t. Then we can do something similar, but we have to somehow take the Taylor expansion to the next order, not just the 0th order — we get

$$\pi\left(I + tQ + \sum_{k=2}^{\infty} \frac{(tQ)^k}{k!}\right) = \pi.$$

The πI and the π cancel, and we get that

$$t(\pi Q + O(t)) = 0.$$

Now we divide by t, and get that $\pi Q + O(t) = 0$. As $t \to 0$, this gives me that $\pi Q = 0$. (So I do some calculus — I do some Taylor expansion around 0 and solve the equation. The first term $\pi I = \pi$ is trivial, the second term gives me $\pi Q = 0$, and all the other terms have at least t^2. Then I divide by t, which is like taking a derivative; and I get πQ plus something going to 0 as $t \to 0$ is 0; this tells me $\pi Q = 0$.) □

Remark 14.13. You can also say that e^{tQ} is the solution to some linear differential equation, and that gives you another perspective of looking at this.

Student Question. *How did you go from the first equation to the second?*

Answer. We're starting with the Taylor expansion

$$\pi\left(I + tQ + \sum_{k=2}^{\infty} \frac{(tQ)^k}{k!}\right) = \pi.$$

First $\pi I = \pi$, so I can cancel this term with π on the right-hand side. Then I have πtQ, and I move the t on the outside. Then I have a bunch of other terms. But in calculus terms, these are all at least quadratic in t. (These things are all matrices, but when you have a tiny number and multiply it by a fixed matrix, it's still tiny.) And the way I wrote t^2 is that it's $t \cdot O(t)$. So I got this equation

$$t(\pi Q + O(t)) = 0.$$

And I divide by t and get $\pi Q + O(t) = 0$. Then I take $t \to 0$; the right-hand side has no t and the $O(t)$ term goes to 0, so I just get $\pi Q = 0$. (This is just a way to write the calculus proof.)

§14.5 An example

We'll now see some examples, and we'll prove a theorem next class. We'll probably be done with continuous-time Markov chains this week (the idea is once you understand discrete-time Markov chains and Poisson processes, you can sort of combine them to extract a lot of information about continuous-time Markov chains).

We'll first see one definition, and then a (relatively) cute example.

Definition 14.14. We say Q is *reversible* with respect to π if

$$\pi(x)Q(x,y) = \pi(y)Q(y,x) \quad \text{for all } x, y. \tag{14.1}$$

Claim 14.15 — If Q is reversible with respect to π, then π is stationary.

Proof. The philosophy is we want to use the fact that we understand discrete-time Markov chains. So you just do the algebra — this condition (14.1) implies that

$$\pi(x)P(x,y) = \pi(y)P(y,x) \quad \text{for all } x,\ y$$

(you're removing I and multiplying by λ, but that doesn't affect this equation). This implies π is stationary for P, which implies that π is stationary for Q. □

We defined this just for the example, so let's now see the example.

Example 14.16 (Barber shop)

A barber cuts hair at a rate of 3 customers per hour. Customers arrive at a rate of 2 customers per hour. But they leave if both chairs in the waiting room are occupied.

Find the stationary distribution.

So we have a barbershop; there's one chair for the person getting a haircut, and two chairs for the people waiting. Customers arrive at a rate of 2 per hour; the barber is a bit faster and cuts at the rate of 3 per hour. But if a customer is unlucky and when they come in the two chains are taken, they say 'forget it, I'll come another day.'

First, what are the states of this Markov chain? It could be that there's nobody there, that someone's getting a haircut, someone's getting a haircut and someone's sitting in a chair, or someone's getting a haircut and two people are sitting in chairs. So the states are 0, 1, 2, and 3 (respectively).

How should I approach this question? First, what is Q? We know the states are $\Omega = \{0, 1, 2, 3\}$. You'll have $Q(x, x+1) = 2$ for $x = 0$, 1, 2 (corresponding to new customers arriving), and $Q(x, x-1) = 3$ for $x = 1$, 2, 3 (corresponding to customers being served).

Prof. Mossel gave us the hint that we're going to solve the detailed balance equations (14.1). Do we know somehow from previous things we saw in class that we *should* be expecting to solve this? We don't always have reversibility; so here why do we expect that? This is a birth-death chain where you always go up or down by 1, and we saw in discrete time that there you can always solve the detailed balance equations. So this is a continuous analog of birth and death chains, so we hope that we can solve the equations that we need to solve.

So let's write all of them: We want

$$\pi(0)Q(0,1) = \pi(1)Q(1,0), \quad \pi(1)Q(1,2) = \pi(2)Q(2,1), \quad \pi(2)Q(2,3) = \pi(3)Q(3,2).$$

If we solve these equations, we get $\pi(1) = \frac{2}{3}\pi(0)$ and similarly $\pi(2) = \frac{2}{3}\pi(1)$ and $\pi(3) = \frac{2}{3}\pi(2)$.

Like in the discrete case, we have another condition that $\pi(0) + \pi(1) + \pi(2) + \pi(3) = 1$, so we have to scale; and if you do that, you get

$$\pi = \frac{1}{65}(27, 18, 12, 8).$$

So we found the stationary distribution — 27/65 of the time there's no one there, 18/65 of the time there's one person getting a haircut and no one is waiting, and so on.

Here's the next question:

Example 14.17

What fraction of (potential) customers are serviced?

This means I also count the people who come and look and see there's two people in the waiting already, so I'm going to go.

This is proportional to the time when there's three people there, so the answer is

$$\frac{65-8}{65} = \frac{57}{65}.$$

The reason is that 8/65 of the time, the system is busy, and the arrivals are memoryless. This is important (otherwise we wouldn't know that). The point is we don't care what happened with the previous arrivals — the fact I just had 10 people come in and it's super crowded means I don't know anything about what happens next. So I know nothing about what happens next based on the fact that the system is full right now. (This requires some thinking — if you had a different system where having a lot of people now made it less likely for people to come soon, then this wouldn't be right. So it's important we have a Poisson process with this independence property.)

§15 April 10, 2025

We'll continue talking about continuous-time Markov chains.

§15.1 Fraction of time spent at states

Here's a theorem which we should have proved before the last example we saw last class, a theorem about the asymptotic fraction of time spent at each state.

Theorem 15.1 (Ergodic theorem)

Let P be an irreducible chain, and let $Q = \lambda(P - I)$ be the corresponding rate matrix. Let $f : \Omega \to \mathbb{R}$, and consider $\frac{1}{t}\int_0^t f(X_s)\,ds$. Then we have

$$\mathbb{P}_x\left[\frac{1}{t}\int_0^t f(X_s)\,ds \to \sum_{y\in\Omega}\pi(y)f(y) \text{ as } t\to\infty\right] = 1$$

(where π is the stationary distribution of the chain, and x is arbitrary).

So we want to understand how much time is spent in each state; and one way to measure it is by looking at this integral $\int_0^t f(X_s)\,ds$ — keep in mind the example where f is the indicator function of one state, i.e.,

$$f(x) = \begin{cases} 1 & \text{if } x = \omega \\ 0 & \text{otherwise.} \end{cases}$$

If we take this f, then we're measuring the amount of time spent at state ω. Of course we want to normalize, so we divide by t.

The right-hand side is a simple average — the average of time spent at each state according to the stationary distribution. The left-hand side is complicated — I'm starting the Markov chain at some x, and each time it jumps I get a new f. Then I take this integral or sum and divide by the length of the interval.

Someone wrote in the feedback form that we didn't prove this when talking about the barbershop example from last class:

Example 15.2

In the barbershop example, this theorem implies that no matter how many customers we start with, the fraction of time spent with 0, 1, 2, or 3 customers (respectively) converges to $\pi(0)$, $\pi(1)$, $\pi(2)$, and $\pi(3)$ (as we computed). We used this implicitly when we asked the question last time of 'what fraction of customers won't be served' — we said it's $\pi(3)$, but for this to hold we actually need it to be the case that the fraction of time you're at 3 is actually $\pi(3)$.

Prof. Mossel thought about proving it one way, and then about proving it another way, and then he thought he'd just sketch the two ways. So we won't actually prove it; he'll just tell us two approaches to proving it.

Proof sketch 1. The idea is to repeat the proof for discrete-time chains. What does it mean to repeat that proof? The most difficult thing is you have to remember what that proof is, so let's recall what that is.

We fix some state $x \in \Omega$, and partition time into *excursions* from x. So the picture is we have our Markov chain, we start at x, it walks and walks and comes back; that's the first part. Then in the second part, we start at x again and walk and walk and walk and return to x. And so on.

These excursions are i.i.d. — what happens in one excursion and another are independent. And for each excursion, we can look at the time spent at each state — the time in each excursion that we spend at state 1, 2, 3, and so on. Then we just sum these numbers and use the law of large numbers — we have a sum of independent random variables, and we're interested in all excursions how many times I visited 1, so that's a sum of i.i.d. random variables; and the total length of all the excursions is also a sum of i.i.d. random variables. So that's the proof for discrete-time Markov chains. And we can do the same thing here. □

Proof sketch 2. We can also *use* the result for discrete-time chains — for those, I know the fraction of time I spend in state 1 is $\pi(1)$. So what's the difference between the fraction of times when I measure in the discrete-time and continuous-time Markov chains?

In our picture, we have a discrete-time Markov chain X_0, X_1, ... at points of the lattice. And in the continuous-time Markov chain, we have a Poisson process, so the first interval is X_0, the second is X_1, and so on.

So the time I'm spending in the discrete time and continuous time Markov chain aren't the same, but we can relate them — you kind of need to weight things by this Poisson process describing how much time you spend at each state.

So if for the discrete chain we spend $N_1(k)$ time at state 1 up to time k, and $N_2(k)$ at state 2 up to time k, and so on, then for the continuous chain, this will become $\sum_{i=1}^{N_1(k)} Z_{i,1}$ for state 1, and for state 2 we're going to get $\sum_{i=1}^{N_2(k)} Z_{i,2}$, and so on; where $Z_{i,j}$ are i.i.d. $\mathsf{Exp}(\lambda)$ random variables. If for the discrete chain up to this time I was at state 0 twice, then the first interval is going to be an exponential random variable with length λ, and so is the second, and they're independent.

This sounds like it's more random — for the discrete chain I told you I spent two steps in this state and three in that, but now we have more randomness. So what's our tool to kill randomness? We're averaging, so we can use the law of large numbers to say that when we divide by t, by the law of large numbers it's concentrated around the mean — these fractions converge to $\pi(1)$, $\pi(2)$, and so on. □

Student Question. *How do you go between k and t?*

Answer. You have to be a bit careful. What happens is first you do this for k. And then you say for each k, I'm going to find the right t. So there's some additional steps with normalizing and dividing, which we're skipping; but that's one of the points we have to take care of.

§15.2 Convergence to stationary and mixing times

We'll state one more theorem, which is about mixing times — we like mixing times, so we'll talk about what happens with mixing times for continuous chains in terms of what happens with mixing times for discrete chains.

Maybe we'll start with a more basic question. Recall that for discrete-time chains, we had that

$$\delta_x P^t \to \pi \quad \text{if } P \text{ is ergodic.}$$

If P was just irreducible, that wasn't enough — we needed P to be ergodic.

For continuous chains, we claim that irreducibility is enough — you don't actually need it to be ergodic.

Let's see why by looking at an example.

Example 15.3

Suppose that

$$P = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

In the discrete chain, I'm just alternating between 0 and 1, so I'm not going to converge to $\pi = (\frac{1}{2}, \frac{1}{2})$.

It doesn't matter what λ is, so let's suppose $\lambda = 1$; what happens if I look at $Q = e^{P-I}$? It's a chain with two states 0 and 1. And what it does is it goes from 0 to 1 according to a $\mathsf{Exp}(1)$ random variable, and from

1 to 0 in the same way. It's kind of clear there's not going to be periodicity, since the time I go from one state to the other is random.

Another way to see this is that we can look at the matrix — can anyone write the matrix $P - I$ as $\lambda(P' - I)$ where P' is ergodic? We can take

$$P' = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \quad \text{and} \quad \lambda = 2.$$

Then if I look at $P' - I$, I get

$$P' - I = \begin{bmatrix} -1/2 & 1/2 \\ 1/2 & -1/2 \end{bmatrix},$$

which means that $2(P' - I) = P - I$. One way to think about this is that I have an exponential with twice the rate, but with probability $1/2$ I stay and with probability $1/2$ I go — this is Poisson thinning. So one interpretation is I wait $\mathsf{Exp}(1)$ time and definitely go; but another is that I wait $\mathsf{Exp}(2)$ time and toss a coin whether to stay or go. These two interpretations are equivalent by Poisson thinning.

So the point is for us, we only need P to be irreducible, not ergodic, in order to get convergence to stationary; and the trick is that we can always do something like this (replacing P with some average of P and I, e.g., $P' = \frac{1}{2}(P + I)$), to get something that's ergodic.

Theorem 15.4

Let P be the transition matrix of an irreducible Markov chain, and let $Q = P - I$ and consider $H_t = e^{tQ}$. Let $\widetilde{P} = \frac{1}{2}(P + I)$. Then for all $x \in \Omega$, we have

$$\|\delta_x H_t - \pi\|_{\mathsf{TV}} \downarrow 0 \quad \text{as } t \to \infty$$

(where π is the stationary distribution). Moreover, we have

$$\|\delta_x H_k - \pi\|_{\mathsf{TV}} \le \|\delta_x \widetilde{P}^k - \pi\|_{\mathsf{TV}} + \mathbb{P}\left[\mathsf{Poisson}(2k) < k\right] \le \|\delta_x \widetilde{P}^k - \pi\|_{\mathsf{TV}} + \left(\frac{2}{e}\right)^k.$$

We're assuming $\lambda = 1$ for simplicity, because we don't want to carry it around; we define $\widetilde{P}$ because we want things to be ergodic. Then the claim is, I'm starting from x and asking where I'll be in t steps; that'll be $\delta_x H_t$. And I ask how far away this is from π (where π is the stationary distribution). So we do converge to the stationary distribution (as long as P is irreducible).

And furthermore, we can find a bound on how fast we converge. First, we want to say the second part of the theorem is stronger than the second part. I want to show that the second thing goes to 0. We do have to think a bit about why $\|\delta_x H_t - \pi\|_{\mathsf{TV}}$ is decreasing in t; but what we get at k is at most what we get for $\widetilde{P}$ (which we know goes to 0, because $\widetilde{P}$ is ergodic) plus some probability of some Poisson random variable being very big, which also goes to 0 as $k \to \infty$. So overall, we'll have

$$\|\delta_x H_k - \pi\|_{\mathsf{TV}} \to 0$$

as $t \to \infty$. Again, this is the idea of trying to analyze what's happening for a continuous-time chain by analyzing what's happening for the discrete-time chain, and that's what this proof is going to do for us.

Student Question. *Isn't this just an integer subsequence, though?*

Answer. Yes. So in order for the second statement to imply the first, I need to show that $\|\pi_x H_t - \pi\|_{\mathsf{TV}}$ is actually decreasing (if I can show this, then convergence to 0 on an integer subsequence is enough).

Proof. We'll first show that $\left\|\delta_x e^{tQ} - \pi\right\|_{\mathsf{TV}}$ is decreasing. Why? I want to show that $\left\|\delta_x e^{(t+h)Q} - \pi\right\|_{\mathsf{TV}}$ is smaller than this — if I go a little more in time, I get something smaller. First, we can rewrite this as

$$\left\|\delta_x e^{(t+h)Q} - \pi e^{hQ}\right\|_{\mathsf{TV}}.$$

All we did right now was replaced π by πe^{hQ} — π is stationary, so this is just equal to π. Now I'm going to pull the e^{hQ} outside, so this is

$$\left\|(\delta_x e^{tQ} - \pi)e^{hQ}\right\|_{\mathsf{TV}}.$$

And now we're done — we have something that's sort of small in L^1, and we apply a Markov matrix to it, so that just makes it smaller. (This is left as an exercise — to check that this is less than $\left\|\delta_x e^{tQ} - \pi\right\|_{\mathsf{TV}}$.)

A maybe better way to think about it is that after I run e^{tQ}, some fraction of me is in π, and some fraction isn't. The fraction that's in π is going to stay there forever, and the fraction that's not, some of it might go to π and some won't. This is coupling. That's what's written here, in somewhat cumbersome math. More explicitly, if

$$\left\|\delta_x e^{tQ} - \pi\right\|_{\mathsf{TV}} = 1 - \eta,$$

this means I can couple $\delta_x e^{tQ}$ and π such that they agree with probability $1 - \eta$. Then if I run for h more time, they will still agree with probability at least $1 - \eta$. (We're going to show that it's actually going to improve, but certainly it'll be at least $1 - \eta$.)

Now we want to work with the matrix $\widetilde{P}$. For this, we need essentially the same algebraic claim that we had for the little example.

Claim 15.5 — If $\widetilde{H}_t = e^{(\widetilde{P}-I)}$, then $\widetilde{H}_t = H_{t/2}$.

I want to work with $\widetilde{P}$ because it's ergodic (I don't like P so much, because it's just irreducible). And for this, I actually want to work with this first heat kernel. I claim that this isn't a big difference — it's just going to change time by a factor of 2 (running the slower chain for t time is the same as running the faster chain for $t/2$ time).

Proof. This follows from the same calculation we did before — that $I - \widetilde{P} = \frac{1}{2}(I - P)$ (and then I just put this in the exponent, so whenever I get t it becomes $t/2$ for the other). □

Now we want to prove that

$$\|\delta_x H_k - \pi\|_{\mathsf{TV}} \le \|\delta_x \widetilde{P}^k - \pi\|_{\mathsf{TV}} + \mathbb{P}\left[\mathsf{Poisson}(2k) < k\right]. \tag{15.1}$$

I'm interested in the left-hand side, so I want to convert it into the slower chain; we get

$$\|\delta_x H_k - \pi\|_{\mathsf{TV}} = \left\|\delta_x \widetilde{H}_{2k} - \pi\right\|_{\mathsf{TV}}.$$

And we can write $\widetilde{H}_{2k}$ as

$$\sum_{j=0}^{\infty} \mathbb{P}[N_{2k} = j]\widetilde{P}^j$$

(where j corresponds to the number of transitions we have — so we're just writing $\widetilde{H}_{2k}$ in a complicated way by conditioning on how many steps I've taken). So this becomes

$$\left\|\delta_x \sum_{j=0}^{\infty} \mathbb{P}[N_{2k} = j]\widetilde{P}^j - \pi\right\|_{\mathsf{TV}}.$$

Then using the triangle inequality for the total variation distance, I get that this is at most

$$\sum_{j=0}^{\infty} \mathbb{P}[N_{2k} = j] \left\| \delta_x - \widetilde{P}^j - \pi \right\|_{\mathsf{TV}}$$

(the first thing is some convex combination of all $\delta_x - \widetilde{P}^j - \pi$ terms, weighted by these probabilities; and we can pull these weights out by the triangle inequality).

Now this is an infinite sum, and we'll say there are two options — either $j < k$, or $j \geq k$. If $j > k$, I'm going to get

$$\left\| \delta_x \widetilde{P}^k - \pi \right\|_{\mathsf{TV}}.$$

The reason I can replace $k+1, k+2, \ldots$ by k is because this is decreasing (the more steps I take, the closer I get). And for other terms, I can bound the probability that the summand is small by $\mathbb{P}[N(2k) < k]$; and then I don't know anything, so I just bound the total variation distance by 1. So I get

$$\left\| \delta_x \widetilde{P}^k - \pi \right\|_{\mathsf{TV}} + \mathbb{P}[N(2k) < k] \cdot 1.$$

What I'm saying is that — we expect $\mathsf{Pois}(2k)$ to take $2k$ steps. We're saying that either we take at least k steps, in which case we can bound the thing by the first term; or we take much fewer steps than the expected, in which case we get the second term.

Going from the first step to the second step has nothing to do with Markov chains, it's just about Poisson random variables — we just want to show that for a Poisson random variable with parameter $2k$, the probability it's less than k is small. This is not super important; the idea is the following. We want to say we converge to the stationary distribution fast. How fast we converge we don't really know — it really depends on $\widetilde{P}$. So we have our continuous chain; we look at what the discrete chain $\widetilde{P}$ is doing. And how bad are we going to be? We'll be far away if either $\widetilde{P}$ itself is slow, or we're very unlucky and our continuous-time chain just didn't move a lot (because we could get very unlucky and it might not move much). And now we'll evaluate the more boring term, which is just something about a Poisson random variable.

Student Question. *Did we throw out the rest of the sum, with $j \geq k+1$?*

Answer. We didn't throw it out — it's part of the first term. Really, we get a term of

$$\left\| \delta_x \widetilde{P}^k - \pi \right\|_{\mathsf{TV}} \sum_{j \geq k} \mathbb{P}[N(2k) = j].$$

But all these probabilities sum to at most 1, so we can forget about the sum and just replace it with 1.

In each term, I'm bounding by 1 something that's not actually 1 — in the first case, the probability I take at least k steps, and in the second term, the total variation distance. Neither of these is 1, but that's good enough for the proof.

We run a vote and no one wants to see the proof of the Poisson random variable claim, so we won't do it (there's some factorials and exponentials and you can use Stirling's formula and do what you have to do, and you get this). □

§15.3 Connections to differential equations

Prof. Mossel was trying to find some continuous Markov chain problem that you have to use differential equations in order to solve, and couldn't find anything simple; but he'll still give us the differential equations perspective, and then give us some problems that are actually easier to solve without differential equations.

Our goal right now is to get that $H_t = e^{tQ}$ (we won't assume we know this; our goal is to get this) using differential equations. So obviously we're not going to use $H_t = e^{tQ}$ as the definition; we're going to start with the definition

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

We're going to start from this probability expression, and we'll try to apply a differential equation to get this e^{tQ} expression. So that's the goal.

Then what we'll do is use the following. Note that

$$\sum_k H_t(i,k)H_s(i,j) = H_{t+s}(i,j).$$

Why? In probability, we like to think about this as conditioning on the intermediate state. What's the chance of getting in $t+s$ steps from i to j? Well, let's look at where we are after time t. We'll be in some state k, so we can ask what's the probability we're at state k after time t, and then what's the probability we move from there to j after additional time s.

Now we're going to do some calculus. Consider

$$H_{t+h}(i,j) - H_t(i,j)$$

(think of h as tiny — we're eventually going to divide by h to get some derivative). This is going to be

$$\sum_k H_h(i,k)H_t(k,j) - H_t(i,j).$$

And now we're going to be a little more careful and separate this based on whether $k = i$; so we can write

$$H_{t+h}(i,j) - H_t(i,j) = \sum_{k \neq i} H_h(i,k)H_t(k,j) + (H_h(i,i) - 1)H_t(i,j).$$

Why did I do it this way? I'm thinking of h as small, and I want to get derivatives, so I want everything to be small. In a small interval of time, I'm very unlikely to move; so $H_h(i,k)$ is going to be small whenever $k \neq i$. And I'm stuck with one big term $H_h(i,i)$, which is very close to 1; so $H_h(i,i) - 1$ is also going to be small. So now I've written this in a form where all terms are a small term times something.

Now, if $k \neq i$, I know that

$$\lim_{h \to 0} \frac{H_h(i,k)}{h} = Q(i,k).$$

(That's the meaning of the rate matrix.) What can we do with the other term? I want to understand what happens with

$$\lim_{h \to 0} \frac{H_h(i,i) - 1}{h}.$$

The point is that the $H_t(i, \bullet)$ are probabilities, so they sum to 1. So instead of working with this guy that I don't like, I can rewrite this as

$$\lim_{h \to 0} \frac{(1 - \sum_{k \neq i} H_h(i,k)) - 1}{h}.$$

(I'm just using the fact that H_h is a transition matrix, so the overall transition probability from i to everyone sums to 1.) The 1's cancel, and I already know that when I divide each $H_h(i,k)$ by h, I get the corresponding Q. So this gives me

$$-\sum_{k \neq i} Q(i,k).$$

And by the definition of Q, this is exactly $Q(i,i)$ (because each row of Q has to sum to 0).

So what is this telling me? It's telling me that if I look at $H_t'(i,j)$ (now I'm actually writing derivatives), this is

$$H_t'(i,j) = \sum_k Q(i,k)H(k,j).$$

For someone who likes matrices, this equation says that if we look at the matrix H', we get

$$H_t' = QH_t.$$

This implies (if we remember our differential equations) that $H_t = e^{tQ}$.

Usually when we solve differential equations, there's some degrees of freedom. How do I know this is e^{tQ}, and not one of its relatives? The point is that H_0 is the identity, so we know we can't shift by anything — it has to actually be e^{tQ}.

§15.4 The waiting time paradox

This is the end of the proof for continuous-time chains for today. Do we want to start from a question that's a paradox, or a question that's not a paradox? We vote for the paradox.

We're going to have a Poisson process with parameter λ, and we're going to fix some huge number t^*. Let S be the last arrival before t^*, and let T be the first arrival after t^*.

> **Question 15.6.** What can you say about $\mathbb{E}[T-S]$?

Why is it called the waiting time paradox (we don't see why it's a paradox yet)? The Number 1 buses don't come according to a Poisson process (they come in clumps), but suppose they do. So I come to the station; and I'm looking at the last bus that arrives before I'm at the station, and the first bus that arrives after. What's the gap between the last bus that arrived, and the bus that I'm actually going to take? (Prof. Mossel comments that as someone who sometimes takes the buses, it feels very long.)

Maybe let's start with a simpler question. We can write

$$\mathbb{E}[T-S] = \mathbb{E}[T-t^*] + \mathbb{E}[t^*-S].$$

First, what's $\mathbb{E}[T-t^*]$? This difference is $\mathsf{Exp}(\lambda)$ because of memorylessness; so we'll have

$$\mathbb{E}[T-t^*] = \lambda.$$

Can anyone argue about the other interval t^*-S? Prof. Mossel wants an intuitive argument using the fact that t^* is very large. The Poisson process is just random points I put on the line. So if I look at it in the right direction and the left direction, it shouldn't matter — there's weird stuff happening around 0, but otherwise it shouldn't matter. So we should have $\mathbb{E}[t^*-S] = \lambda(1+o(1))$, which means this should be about 2λ.

> **Student Question.** *Why is it exactly λ?*
>
> **Answer.** We didn't argue that a Poisson process backwards is also a Poisson process. This is actually true if you do it the right way — if you define a Poisson process on the infinite line, and so on. But the reason we don't know it is because there's something special happening at time 0 — for example, if I ask you what's the last arrival time before time 0.1, with high probability that's 0. This is why we want t^* to be very large — then it doesn't really matter where you started, and the thing looks the same forwards and backwards.

There's two things to take away — first, it's about 2λ, and second, it's definitely bigger than λ.

Why is this a paradox? It's a Poisson process with rate λ. So it should be the case that the expected time between two buses is λ! There's no cheating here — if I look at any bus, the expected time to the next bus is λ, not 2λ. But if I look at myself and the gaps between me and the buses that just came and are about to come, somehow the gap between buses became 2λ instead of λ.

There's no cheating in either calculation — it is true that the gap between buses has expectation λ, and also that if I look at some time immediately in the future, the expected length of the interval I'm in is actually 2λ.

So now two things can happen — either we can decide that math is inconsistent and leave the room, or we can try to explain what's going on.

One way to think about this is that with my arrival time, I'm sort of sampling a random point; when I do that, I'm more likely to fall in a big interval than a short interval. There are short intervals and long intervals; I'm sampling a random point, and the likelihood of sampling a longer interval is bigger. So when I look from my self-centered perspective, I'm more likely to choose a big interval — this biases the length of the interval to be bigger than just the way they come on their own.

Student Question. *But aren't there more smaller intervals, so it should in some sense work itself out?*

Answer. Let's draw a picture — suppose there's four small intervals (very short), and one big interval (very long) — let's for fun say they're 1, 1, 1, 1, 6. There are two questions we can ask. First, what's the expected length of an interval? That's

$$\frac{1}{5}(1+1+1+1+6) = 2.$$

Now I can ask, what's the expected length of an interval chosen by a random point? (This doesn't have to do with Poisson processes; I just come at a random point, and ask the expected length of the interval.) With probability 0.6 it's going to be 6, and otherwise it's 1. So we get

$$0.6 \cdot 6 + 0.4 \cdot 1 = 4.$$

So it went up by a factor of 2 — you're more likely to choose a bigger interval, and this is why you compute the average in a different way.

§15.5 Another example

Here's one more example. Prof. Mossel asked ChatGPT to find an example you need differential equations for; it disappointed him, in that you don't need differential equations for this, but it's still an example.

Example 15.7

Suppose you have a Markov chain with states 0, 1, 2, 3, and

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & -3 & 2 & 0 \\ 0 & 1 & -3 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Find the proability of absorption at 0 starting from 1.

What the first and last rows mean is that if I'm at state 0 or 3 I always stay there (I can't move); from states 1 and 2 I might do something.

So how do I find the probability that I'm going to end up at 0 (and once I'm at 0 I get stuck there forever) when I start at 1?

There's no difference between the discrete and continuous case, so we could solve it for the discrete case. Or we could just do it directly — we can write p_1 as the probability of ending at 0 starting from 1, and p_2 as the probability of ending at 0 starting from 2. And then we can write some equations.

From 1, what happens? I'll wait and wait, and there's going to be an $\mathsf{Exp}(1)$ and $\mathsf{Exp}(2)$. So with probability $\frac{1}{3}$ I'll just end up at 0 and get absorbed (because of exponential races), and with probability $\frac{2}{3}$ I'm going to go to state 2, and then who knows what happens. So I get

$$p_1 = \frac{1}{3} \cdot 1 + \frac{2}{3} \cdot p_2.$$

And similarly, we get

$$p_2 = \frac{1}{3} \cdot p_1 + \frac{2}{3} \cdot 0.$$

Then we can solve these equations.

What's the philosophy? One explanation of what this matrix means is at state 1, I have an exponential race between $\mathsf{Exp}(1)$ and $\mathsf{Exp}(2)$, and whoever wins, I go to that state. And when I have an exponential race, the probabilities of winning are proportional to the rates. If 0 wins then I'm absorbed; otherwise I go to 2. And from state 2, with some probability I go to state 1, and otherwise I get absorbed by the other state.

The other question ChatGPT asked:

Example 15.8

What's the expected time of absorption?

Next week we're going to start with martingales.

§16 April 15, 2025

Today we're going to start to talk about martingales.

§16.1 Conditional expectation

Before we do this, we'll talk a bit about conditional probabilities and expectations.

Let's first consider discrete random variables. For a discrete random variable, we know

$$\mathbb{P}[Y = y \mid X = x] = \frac{\mathbb{P}[Y = y \text{ and } X = x]}{\mathbb{P}[X = x]}.$$

So that's the definition of conditional probability. We know that if we somehow have more information about X than Y, then we can compute the overall probability that $Y = y$ by summing over all possibilities of X, and writing

$$\mathbb{P}[Y = y] = \sum_x \mathbb{P}[Y = y \mid X = x]\mathbb{P}[X = x].$$

We'll recall the definition of conditional expectations:

Definition 16.1. We define $\mathbb{E}[Y \mid X = x] = \sum_y y\mathbb{P}[Y = y \mid X = x]$.

So we're thinking about $\mathbb{P}[\bullet \mid X = x]$ as a probability distribution, and we're doing the standard definition of expectation.

What happens if instead of Y, I had the expected value of some *function* of Y? We'd have

$$\mathbb{E}[g(Y) \mid X = x] = \sum_y g(y)\mathbb{P}[Y = y \mid X = x].$$

And in general (we won't prove this, because you should've seen it in 18.600):

Claim 16.2 — The conditional expectation $\mathbb{E}[\bullet \mid X = x]$ satisfies all the usual properties of expectations, for each fixed $x \in \Omega$.

Let's see an example.

Example 16.3 (Dice paradox)

Consider a game where you roll a die until you get a 6.

- If the number you roll is ever odd, then you lose and get $Y = 0$ dollars.
- Otherwise, you gain $Y = \#\text{rolls}$.

Let X be 1 if you didn't lose (i.e., you made some money) and 0 otherwise. What is $\mathbb{E}[Y \mid X = 1]$?

So if I roll and get a 2 and then 1, then the game is over and I get 0 dollars. Meanwhile, if I get 2, 4, 2, 6, then I had 4 rolls, so I get 4 dollars. And given that I won (I made some money), what can you say about the expected amount of money I made?

One student answers that it should be 3, because it's a geometric random variable with probability $\frac{1}{3}$ — we just restrict to even numbers and wait until 6, which happens with probability $\frac{1}{3}$. So what we should get is $\mathbb{E}[\mathsf{Geom}(\frac{1}{3})] = 3$.

Let's do the computations and see what happens. First, let's compute $\mathbb{P}[X = 1]$. Either the first roll is a 6, or the first roll is even (not a 6) and then I get a 6, or so on; so

$$\mathbb{P}[X = 1] = \frac{1}{6} + \frac{1}{3}\cdot\frac{1}{6} + \frac{1}{3^2}\cdot\frac{1}{6} + \cdots = \frac{1}{6}\sum_{k=0}^{\infty}\frac{1}{3^k} = \frac{1}{6}\cdot\frac{1}{1-1/3} = \frac{1}{4}.$$

So that's the probability that I'm actualy going to win.

Now I have to compute, what's the probability $\mathbb{P}[Y = k \mid X = 1]$? One way to do it is that this is

$$\frac{\mathbb{P}[Y = k \text{ and } X = 1]}{\mathbb{P}[X = 1]}.$$

And this is a similar computation — it's

$$\frac{\frac{1}{6}\cdot\frac{1}{3^{k-1}}}{\frac{1}{4}}.$$

(The numerator means the first roll was 6 and the others were all 2 or 4.) And now we can just do the computations and get

$$\mathbb{E}[Y \mid X = 1] = 4\sum_{k=1}^{\infty}\frac{1}{6}\cdot\frac{1}{3^{k-1}}\cdot k.$$

This can be written as some constant times the expectation of a $\mathsf{Geom}(\frac{2}{3})$. We get that this is

$$\sum_{k=1}^{\infty} \frac{2}{3}\left(\frac{1}{3}\right)^{k-1} k.$$

This is exactly $\mathsf{Geom}(\frac{2}{3})$ (since I succeed with probability $\frac{2}{3}$ and fail with probability $\frac{1}{3}$), so this is

$$\mathbb{E}[\mathsf{Geom}(2/3)] = \frac{3}{2}.$$

So it's not 12, or even 3; it's $\frac{3}{2}$. The intuitive reason it's less than 3 is that if I condition on the fact that everything was even, then I'm actually telling you it's very unlikely I rolled many times — if they're all even, probably you didn't roll too many times, so the expectation should be lower. Why it should be lower by a factor of 2 is harder to explain.

The first time Prof. Mossel came up with this question was in the last 5 minutes of a class like this; then he computed the infinite sums correctly and got $\frac{3}{2}$, then the bell rang, then he said 'obviously the answer is 3 but I have no time to look where I lost the factor of 2 in the computation,' and then he stepped out of the class and realized this intuition that conditioning on $X = 1$ biases the measure. So this example is somewhat confusing.

Now let's actually give the formal definition of conditional expectation.

Definition 16.4. We define $\mathbb{E}[Y \mid X]$ as a random variable which takes the value $\mathbb{E}[Y \mid X = x]$ whenever $X = x$ (for all $x \in \Omega$).

The important, somewhat confusing, thing to remember is that $\mathbb{E}[Y \mid X]$ is a random variable — it's not a number anymore. So this is not a number, it's a random variable; and this random variable depends on the value of X itself (which is a random variable). The way to think about $\mathbb{E}[Y \mid X]$ is as the information we have about Y given X — X itself is random, so we get different information depending on the randomness in X.

Here are some basic properties:

Fact 16.5 (Law of total expectation) — We have $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$.

Note that $\mathbb{E}[Y \mid X]$ is a random variable, so we can take its expectation. This says that if we average this information $\mathbb{E}[Y \mid X]$ over all possible unknown values of X, we get back $\mathbb{E}[Y]$.

Fact 16.6 (Linearity) — We have $\mathbb{E}[Y + Z \mid X] = \mathbb{E}[Y \mid X] + \mathbb{E}[Z \mid X]$.

(Note that all of these things are random variables.)

Fact 16.7 — If $Z = g(X)$ is a deterministic function of X, then $\mathbb{E}[ZY \mid X] = Z\mathbb{E}[Y \mid X]$.

This is because whenever you fix X, I know Z exactly, so Z comes out of the expectation.

§16.1.1 Some examples

Let's do a couple more examples. The first one is the previous example — it's complicated, but we've already done the computations.

Example 16.8

In the above game, what is $\mathbb{E}[Y \mid X]$?

Now this is a random variable — when X takes the value 0 it takes one value, and when X takes the value 1 it takes another value. When $X = 0$, I know $Y = 0$; and when $X = 1$, we computed $\mathbb{E}[Y \mid X = 1] = \frac{3}{2}$. So

$$\mathbb{E}[Y = X] = \begin{cases} 0 & \text{if } X = 0 \\ \frac{3}{2} & \text{if } X = 1. \end{cases}$$

We can write this more compactly as

$$\mathbb{E}[Y \mid X] = \frac{3}{2}X$$

(this is the same thing, just written differently).

Now let's do another example.

Example 16.9

We have n independent coin tosses, where $\mathbb{P}[\mathsf{head}] = p$. We define S_k as the number of heads in the first k tosses, and S_n as the number of heads in all n tosses (where $n \geq k$).

- What is $\mathbb{E}[S_n \mid S_k]$?
- What is $\mathbb{E}[S_k \mid S_n]$?

Both of these are random variables — I'm asking you what's the expected number of heads I get in the larger number of tosses given the first k, and the expected number of heads in the first k given the total?

For the first one, one way to do this is to write

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

where X_i iss the ith toss. So we break it into the part we know, and the part we don't know anything about. Then by linearity of expectation, we get

$$\mathbb{E}[S_k \mid S_k] + \sum_{i=1}^{n} \mathbb{E}[X_i \mid S_k].$$

Why is this a simpler expression? When I know S_k, S_k is a function of S_k, so that goes out of the expectation, and we just have

$$\mathbb{E}[S_k \mid S_k] = S_k.$$

Meanwhile, S_k is independent of the X_i's (since S_k only tracks the tosses up to the kth), which means the conditioning doesn't do anything; and we get

$$S_k + \sum_{i=k+1}^{n} \mathbb{E}[X_i] = S_k + (n-k)p.$$

Now what about the second one? I'm telling you the total number of heads, and they should be uniformly distributed everywhere, so it should be $\frac{k}{n} \cdot S_n$. Another way to see this, using linearity of expectation, is that it's

$$\mathbb{E}\left[\sum_{i=1}^{k} X_i \;\middle|\; S_n\right] = \sum_{i=1}^{k} \mathbb{E}[X_i \mid S_n].$$

And then it's easier to think about this argument — what's the probability the ith toss is heads, given that I have a total of S_n heads? Well, the total is just $\frac{S_n}{n}$, since I have S_n heads and each of the tosses is equally likely to be heads. So I get

$$\mathbb{E}[S_k \mid S_n] = k \cdot \frac{S_n}{n} = \frac{k}{n} S_n.$$

§16.1.2 Conditional expectations for continuous random variables

We'll also quickly recall what happens with continuous random variables. It's the same story, but we replace conditional probabilities with conditional densities.

Definition 16.10. The conditional density of Y given that $X = x$ is defined by

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

Different classes use different notations, so this notation is maybe a bit nonstandard. The picture to have in mind is that (X, Y) has a joint density $f(x, y)$ — this means if I look at the vector (X, Y) and ask what's the probability it belongs to some two-dimensional set A, this is going to be

$$\mathbb{P}[(X,Y) \in A] = \iint_A f(x,y)\,dx\,dy.$$

So that's the little f on the top of the right-hand side.

The notation $f_X(x)$ is just the marginal density of X — so now I'm looking at the probability X belongs to a *one*-dimensional set A, and we'll have

$$\mathbb{P}[X \in A] = \int_A f_X(x)\,dx.$$

What's the relationship of f_X to f — how do I write the marginal in terms of the joint distribution? I can just integrate out over y — we have

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy.$$

(The two x's on the two sides are the same, and I'm integrating over y.)

The formulas are analogous to what we had in the discrete case: Just as we wrote $\mathbb{P}[Y = y]$ as a sum of conditional probabilities in the discrete case, we can do the same with densities — we have

$$f_Y(y) = \int f_Y(y \mid X = x) f_X(x)\,dx.$$

You always think of the analog with marginal probabilities — I integrate the conditional densities over all x, weighted by the marginal probability of x.

The definition of conditional expectations is also the same (with probabilities replaced by densities):

Definition 16.11. We define $\mathbb{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y f_Y(y \mid X = x)\,dy$.

So we integrate over all possibilities of y, but we have this information about X, and we take it into account in our density.

The other definitions are the same — the definition of $\mathbb{E}[Y \mid X]$ we gave before doesn't care if you're continuous or discrete, so it carries over as written. And the basic properties still all hold, and so on.

Let's do a simple example.

Example 16.12

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

Whenever you look at a conditional expectation and you're confused, it's helpful to imagine fixing some value of Y. If we fix $Y = 0$, then X is going to be uniform between 0 and 2, so its expected value is 2. If we fix $Y = 2$, then X has to be 0. In between, X is going to be uniform between 0 and $2 - Y$, so the answer is going to be

$$\mathbb{E}[X \mid Y] = 1 - \frac{Y}{2}.$$

So the quick answer is that conditioning on $Y = y$, we have $X \sim \mathsf{Unif}[0, 2 - y]$, which means

$$\mathbb{E}[X \mid Y = y] = 1 - \frac{y}{2}.$$

And this is true for every value of y, so this means

$$\mathbb{E}[X \mid Y] = 1 - \frac{Y}{2}.$$

We didn't actually follow the recipe written above — in principle we should've calculated the marginals and joint distributions and divided and so on, but if we do that we'll get the same answer.

§16.2 Jensen's inequality

One other thing we'll talk about, which is generally useful, is Jensen's inequality.

Theorem 16.13 (Jensen's inequality)

If φ is a convex function, then $\mathbb{E}[\varphi(X)] \geq \varphi(\mathbb{E}[X])$.

Example 16.14

The function $x \mapsto x^2$ is convex, so $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$.

What does this give for conditional expectations?

Corollary 16.15

We have $\mathbb{E}[\varphi(Y) \mid X] \geq \varphi(\mathbb{E}[Y \mid X])$.

On one hand this is trivial, and on the other hand it's very confusing. When I write $\mathbb{E}[\varphi(X)] \geq \varphi(\mathbb{E}[X])$, what types of objects do I have — are these numbers, variables, operators in a Banach space, ...? They're both numbers; we're saying one number is bigger than another number.

In the corollary, we've written it very innocently, but these are not numbers! We've said that $\mathbb{E}[\varphi(Y) \mid X]$ is a random variable; and the right-hand side is a function of a random variable, so it's also a random variable.

So we're saying one random variable is bigger than another random variable; what does that mean? What we mean is that this is true with probability 1 — whenever you look at these two random variables, the one on the left is going to be bigger than the one on the right.

On the other hand, the proof is going to be very simple (we're not going to prove Jensen's inequality, just the implication). If I fix the value of x, then the two sides become expectations (conditioned on $X = x$), and I apply Jensen's inequality to get that this is true for every x — so this is true because it's true conditioned on $X = x$ for all x.

§16.3 Martingales

We'll be able to define martingales today, so let's do that.

Remark 16.16. There's going to be some notation for a sequence of random variables we'll see a lot, and there isn't standard notation, so Prof. Mossel will let us choose. We'll need a lot of vectors of random variables like $(X_1, \dots, X_n)$. One option is to just write $(X_1, \dots, X_n)$. Last class, Prof. Mossel wrote $X_{[n]}$ because it's shorter. In information theory, they also don't like writing $(X_1, \dots, X_n)$ all thee time; so their notation is X_1^n.

We vote for the second option (it's close between (2) and (3)).

Now with this notation fixed, we can define martingales.

Definition 16.17. Consider two sequences of random variables $(M_i)_{i\geq 0}$ and $(X_i)_{i\geq 0}$. We say that $(M_i)_{i\geq 0}$ is a martingale with respect to $(X_i)_{i\geq 0}$ if:

(1) M_n is a function of $X_{[n]}$ (i.e., it's a function of $X_0, \dots, X_n$).

(2) We have $\mathbb{E}[M_{n+1} \mid X_{[n]}] = M_n$.

So this is the definition. Let's try to get some intuition. You can think of M_n as the value of some stock on day n, and $X_{[n]}$ might be $M_{[n]}$ (you just look at this stock up to day n), or maybe $X_{[n]}$ stores *all* of the stock market up to day n. So in other words, X_1 is the information you have at day 1, X_2 is the information you have at day 2, and so on. And M is what you care about — maybe you care only about a particular stock, not all the market. And you want that the particular stock is a function of all the information you have up to time n, and more importantly, that things are fair — the expected value of the stock tomorrow, given what we know today, should be the current value today. (If it's worth more, then people would buy it more; if it's worth less, people should sell it.) So this is sort of saying people are betting that the value in the future should be equal to the value right now.

This definition came before stock markets, but it came from games of chance, which are the same — people wanted to know how much money they would make, so they started looking at quantities like this.

We'll generalize this a bit first. Things can stay the same, or they can get better or worse. This is a martingale. There's also something called a *supermartingale*, and something that's called a *submartingale*. Condition (1) is the same, but Condition (2) is different. Annoyingly, for supermartingales instead of $=$ you have $\leq$, and for submartingales you have $\geq$.

Definition 16.18. Consider two sequences of random variables $(M_i)_{i\geq 0}$ and $(X_i)_{i\geq 0}$. We say that $(M_i)_{i\geq 0}$ is a supermartingale with respect to $(X_i)_{i\geq 0}$ if:

(1) M_n is a function of $X_{[n]}$ (i.e., it's a function of $X_0, \dots, X_n$).

(2) We have $\mathbb{E}[M_{n+1} \mid X_{[n]}] \leq M_n$.

Definition 16.19. Consider two sequences of random variables $(M_i)_{i\geq 0}$ and $(X_i)_{i\geq 0}$. We say that $(M_i)_{i\geq 0}$ is a submartingale with respect to $(X_i)_{i\geq 0}$ if:

(1) M_n is a function of $X_{[n]}$ (i.e., it's a function of $X_0, \ldots, X_n$).

(2) We have $\mathbb{E}[M_{n+1} \mid X_{[n]}] \geq M_n$.

So if it's a supermartingale you expect it to go down, and if it's a submartingale you expect it to go up.

Before looking at some examples, let's understand a bit what are supermartingales and submartingales. We'll do the martingale claim first:

Claim 16.20 — If M_n is a martingale, then $\mathbb{E}[M_{n+1}] = \mathbb{E}[M_n] = \cdots = \mathbb{E}[M_0]$.

So what would be the claim for supermartingales?

Claim 16.21 — If M_n is a supermartingale, then $\mathbb{E}[M_{n+1}] \leq \mathbb{E}[M_n] \leq \cdots \leq \mathbb{E}[M_0]$.

Claim 16.22 — If M_n is a submartingale, then $\mathbb{E}[M_{n+1}] \geq \mathbb{E}[M_n] \geq \cdots \geq \mathbb{E}[M_0]$.

Let's prove one of them; they're all the same.

Proof for supermartingales. We know that $\mathbb{E}[M_{n+1} \mid X_{[n]}] \leq M_n$. Again, this is an inequality of random variables — the random variable on the left is always less than the random variable on the right. Now we take the expected values of these random variables, and get

$$\mathbb{E}[\mathbb{E}[M_{n+1} \mid X_{[n]}]] \leq \mathbb{E}[M_n]$$

(if one random variable is always less than the other, then the expected value of the first is also less than that of the second). And by the tower property of conditional expectations, the left-hand side is just $\mathbb{E}[M_{n+1}]$. □

So I started with the data $\mathbb{E}[M_{n+1} \mid X_{[n]}] \leq \mathbb{E}[M_n]$ — no matter what the stock market is up to today, the expected value tomorrow is less than the current value right now. Then I take an expectation over everything the market could have done up to now, and we get $\mathbb{E}[\mathbb{E}[M_{n+1} \mid X_{[n]}]] \leq \mathbb{E}[M_n]$. And we know the expected value of a conditional expectation is just the original expectation, so we get what we want.

§16.4 Some examples

Let's look at some examples.

Claim 16.23 — Let P be a transition matrix for a discrete-time Markov chain. Let h be a harmonic function. Then if $(X_n)_{n\geq 0}$ is the Markov chain, then $(h(X_n))_{n\geq 0}$ is a martingale (with respect to $(X_n)_{n\geq 0}$).

So we have our Markov chain (X_n) from the first half of the class, and we have a function h which is harmonic. And we're applying this harmonic function to the Markov chain; and the claim is that this is a martingale. In this case all the information I have is where the Markov chain went, and the information I'm interested in is $h(X_n)$.

Proof. There are two things I have to check. First, I have to check that M_n is a function of $X_{[n]}$. And it's actually a very simple function, because $M_n = h(X_n)$. So Condition (1) is okay. (Usually this is just a sanity check; you don't have to do any math, you just have to check that this holds.)

Now we have to check Condition (2) — what's $\mathbb{E}[M_{n+1} \mid X_{[n]}]$, the conditional expectation of my chain at time $n+1$ given that I know all the information up to time n? Well, first it's a Markov chain, so to know what happens on the next step I just need to know where I am right now; that means this is just

$$\mathbb{E}[M_{n+1} \mid X_n].$$

And what's that? It's

$$\mathbb{E}[M_{n+1} \mid X_n] = \sum_y P(X_n, y)h(y).$$

This is because the probability I'll be at y is $P(X_n, y)$.

And because the function is harmonic, this is equal (by definition) to $h(X_n)$ — harmonic means that when I average over where I'll go in the next state, I get my current value. And then we're done. □

Next class, we'll start with this claim and see a bunch of examples of how you can apply it in practice; and then we'll also derive some more general properties of martingales and applications.

§17 April 17, 2025

§17.1 Martingales from Markov chains

At the end of last class, we showed that if h is harmonic for a Markov chain $\{X_n\}$, then if we look at the sequence $M_n = h(X_n)$, it's a martingale.

Before we give a couple of examples of this, we'll state something a little more general — we'll look at functions of *two* parameters which are 'sort of harmonic.'

Claim 17.1 — Let $f : \Omega \times \mathbb{N} \to \mathbb{R}$ be such that for all $x \in \Omega$ and $n \in \mathbb{N}$, we have

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

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

Why is this claim more general than what we saw last time? If we take $f(x, n) = h(x)$ where h is harmonic, then we recover the claim from last class — you should think of the time as n, and if f doesn't depend on n, then you get $h(x) = \sum_y P(x, y)h(y)$, which is exactly the definition of being harmonic. This is more general — now the function is allowed to depend on time — and maybe we'll see an example of this in a little bit.

It's the same proof, but let's do it.

Proof. We want to check that M_n is a martingale, so we have to check what's

$$\mathbb{E}[M_{n+1} \mid X_{[n]}]$$

(recall that this means we're given all the X_i's up to time n). This is

$$\mathbb{E}[f(X_{n+1}, n+1) \mid X_{[n]}].$$

Now, it's a Markov chain, so we know the history doesn't matter other than the last time; so it doesn't matter whether I look at all the history up to time n, or I just look at X_n. So this is just

$$\mathbb{E}[f(X_{n+1}, n+1) \mid X_n].$$

And we're looking at what X_{n+1} can be; so we sum over all y, what's the chance that X_{n+1} takes the value y, given that now we're at X_n? So this is

$$\mathbb{E}[f(X_{n+1}, n+1) \mid X_n] = \sum_y \mathbb{P}[X_{n+1} = y \mid X_n] f(y, n+1).$$

And then we can write this as

$$\sum_y P(X_n, y) f(y, n+1) = f(X_n, n) = M_n.$$

So the logic is we want to compute this expected value conditioned on X_n, so we look at the probability we take each particular value y conditional on X_n, and multiply by the corresponding value; and by the equation defining $f(x, n)$, we get exactly that this is $f(X_n, n) = M_n$. □

§17.1.1 Some applications to gambler's ruin

Now let's see some examples for gambler's ruin.

Example 17.2 (Gambler's ruin)
Consider an 'unfair' form of gambler's ruin, where $S_n = \sum_{i=1}^n X_i$ where

$$X_i = \begin{cases} 1 & \text{with probability } p \\ -1 & \text{with probability } 1-p. \end{cases}$$

Claim 17.3 — The sequence $M_n = (\frac{1-p}{p})^{S_n}$ is a martingale.

This is sort of an interesting martingale — it's something exponential in the value of S_n.

Proof. This is just a function of the state — there's no dependence on time — so we can use the claim from last class. So it suffices to show that

$$h(x) = \left(\frac{1-p}{p}\right)^x$$

is harmonic — if we show this, then we're done (because we know if I apply a harmonic function to a Markov chain, I'm going to get a martingale). So we want to check that

$$h(x) = ph(x+1) + (1-p)h(x-1).$$

You can write this out and see that it's satisfied. □

There's one value of p for which this is not an interesting claim. Well, the values $p = 0$ and $p = 1$ are not interesting, because you get ∞ or 0. But also, when $p = \frac{1}{2}$, this is not interesting — then you get $1^{S_n} = 1$. Of course the constant function is a martingale, but that's not an interesting claim.

So now let's try to find some (more interesting) martingales when $p = \frac{1}{2}$. First, what's the harmonic function for the usual gambler's ruin Markov chain with $p = \frac{1}{2}$? The function $h(x) = x$ is harmonic, because

$$x = \frac{1}{2}(x+1) + \frac{1}{2}(x-1).$$

This implies S_n is a martingale. This makes sense — in each round I either win a dollar or lose a dollar with equal probability, so the expected value of where we will be is exactly where we are right now.

Let's see a more interesting example.

Claim 17.4 — The sequence $S_n^2 - n$ is a martingale.

Proof. We want to take the function $f(x, n) = x^2 - n$ in Claim 17.1. So we need to show that

$$x^2 - n = \frac{1}{2}((x+1)^2 - (n+1)) + \frac{1}{2}((x-1)^2 - (n+1)).$$

This is true (we can check that everything cancels out); so we apply Claim 17.1 and get that $S_n^2 - n$ is a martingale. □

Student Question. *Aside from being an example of a martingale, why is this interesting?*

Answer. We'll see later. First it's kind of curious that if I tell you just about this process, you square it and subtract n and get a martingale. If you're bored in this class, you can try to find a degree-3 polynomial that's a martingale for this. And if you're really really bored, you can try to find polynomials of every degree.

We have one more example, but maybe we'll skip it and talk about some basic properties of martingales.

§17.2 Basic properties of martingales

We want to analyze how martingales behave. The general philosophy of probability classes — 18.600 and this class — is that we really want to understand what happens as $n \to \infty$. For example, the law of large numbers talks about the fraction of heads and tails if I toss a coin lots of times; when we talked about Markov chains we wanted to understand what happens if we run our chain a long time. Similarly, here we want to understand what happens in the long run. To do this, we'll need to develop some basic properties of martingales; so that's what we'll do now.

Claim 17.5 — If M_n is a martingale and φ is convex, then $\varphi(M_n)$ is a submartingale.

Remember a submartingale means that it goes up in expectation. We sort of saw the proof of this before; it follows from Jensen's inequality.

Proof. We'll look at $\mathbb{E}[\varphi(M_{n+1}) \mid X_{[n]}]$. Jensen's inequality says that if you have a convex φ, you can take it out of the expectation (even with conditioning); so by Jensen's inequality, we get

$$\mathbb{E}[\varphi(M_{n+1}) \mid X_{[n]}] \geq \varphi(\mathbb{E}[M_{n+1} \mid X_{[n]}]).$$

And because M_n is a martingale, we have $\mathbb{E}[M_{n+1} \mid X_{[n]}] = M_n$; so the right-hand side is $\varphi(M_n)$. □

Does this shed some light on some of the weird martingales we've seen before? We had a square somewhere — we said $S_n^2 - n$ is a martingale. This claim is sort of saying that if we looked at S_n^2, it'd be a submartingale. A submartingale might be a martingale, but usually in Jensen's, unless everything is constant or linear it becomes a strict inequality. So you don't expect S_n^2 itself to be a martingale; that's why we have to subtract more and more as time goes on.

Another question:

Question 17.6. Would the same proof work if (M_n) is just a submartingale (and not an actual martingale)?

The answer is no in general.

Example 17.7

Take our favorite convex function $\varphi(x) = -x$. This is a convex function, because it was linear; and if something was increasing in expectation before, now it's going to be *decreasing* in expectation. So if (M_n) is a submartingale, then $(-M_n)$ is actually a *super*martingale.

But the answer is yes under some conditions:

Claim 17.8 — If M_n is a submartingale and φ is convex and increasing, then $\varphi(M_n)$ is a submartingale.

Proof. If we have a submartingale, then when we say

$$\mathbb{E}[\varphi(M_{n+1}) \mid X_{[n]}] \geq \varphi(\mathbb{E}[M_{n+1} \mid X_n])$$

(this is still correct — it's Jensen's inequality), now we know that $\mathbb{E}[M_{n+1} \mid X_{[n]}] \geq M_n$. And because φ is monotone, this means

$$\varphi(\mathbb{E}[M_{n+1} \mid X_n]) \geq \varphi(M_n). \qquad \square$$

One very important example of this claim is the following.

Example 17.9

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

§17.2.1 Squares of martingales

Squares are important in general. Why? Because in math we like squares, and if you remember, when we proved the law of large numbers, we used Chebyshev's inequality which is about squares. And Cauchy–Schwarz also involves squares. So we like squares and understand them quite well; so let's make some more claims about squares. All of these claims look like I don't know arithmetic, but they're interesting because of that.

Claim 17.10 — We have $\mathbb{E}[M_{n+1}^2 \mid X_{[n]}] - M_n^2 = \mathbb{E}[(M_{n+1} - M_n)^2 \mid X_{[n]}]$.

Why does this look like I don't know arithmetic? In arithmetic, it's usually not the case that $(a-b)^2 = a^2 - b^2$. But I'm telling you that in this particular case, when you condition on $X_{[n]}$ and take expected values, it really is true.

Proof. We just expand the right-hand side — it's

$$\text{RHS} = \mathbb{E}[M_{n+1}^2 - 2M_{n+1}M_n + M_n^2 \mid X_{[n]}].$$

So far we haven't done anything interesting; now we'll do something slightly more interesting. We can use linearity of expectation to split this up term-by-term. And the nice thing about martingales is M_n is a function of $X_{[n]}$, so when I condition on $X_{[n]}$, M_n goes out of the expectation. So I get

$$\text{RHS} = \mathbb{E}[M_{n+1}^2 \mid X_{[n]}] - 2M_n\mathbb{E}[M_{n+1} \mid X_{[n]}] + M_n^2.$$

And it's a martingale, so $\mathbb{E}[M_{n+1} \mid X_{[n]}] = M_n$. So overall, I get

$$\mathbb{E}[M_{n+1}^2 \mid X_{[n]}] - 2M_n^2 + M_n^2,$$

which is what I wanted (it's equal to the LHS). $\square$

That seems like a nice property for martingales. Here's another one, which is actually used a lot — so often that it has a name.

Claim 17.11 (Orthogonality of increments) — If M_n is a martingale, then

$$\mathbb{E}[(M_n - M_0)^2] = \sum_{i=1}^{n} \mathbb{E}[(M_i - M_{i-1})^2].$$

This is a bit similar to the previous claim. Usually we know how to telescope $M_n - M_0 = (M_n - M_{n-1}) + (M_{n-1} - M_{n-2}) + \cdots$. But usually, a square of a sum is not the sum of the individual squares. But for martingales, if you take the expected value then it turns out to be true.

The proof is the same philosophy to the previous claim, so we won't prove it.

§17.2.2 Predictable processes

We'll use these analytic statements later. For now, we're going to collect some more probabilistic statements about martingales. Our goal now is to somehow formalize the idea that you cannot win (in expectation) in a fair game. First of all, we want to understand, what are the kinds of things you're allowed to do in a fair game? For this, we need a definition; this definition sort of tells you what kinds of things you can do in the stock market (that's not cheating).

Definition 17.12. Let (M_n) be a martingale with respect to (X_n). We say that (H_n) is a *predictable process* if for each n, H_n is a function of $X_0, \ldots, X_{n-1}$.

So (M_n) is the martingale, and (X_n) is the information. And H_n is supposed to be a function of what we've seen so far (or really, strictly before this time).

How should you think about this is that H_n tells you how much of a stock you should buy or sell on day n. It can depend on what's happened on the previous days. You can't know what's happening in the stock market today — you only know what's happened *before* today. Then X_n is going to tell me what happens on the stock market today, in particular M_n tells me the value of the stock. And H_n tells me how much I buy or sell.

Then how do you formalize the claim that you can't make money in a fair game?

Theorem 17.13

(1) Suppose (M_n) is a supermartingale and (H_n) is predictable with respect to (X_n), and $0 \le H_n \le c$ for all n, then

$$W_n = \sum_{m=1}^{n} H_m(M_m - M_{m-1})$$

is a supermartingale.

(2) Suppose (M_n) is a martingale and (H_n) is predictable, and $|H_n| \le c$. THen W_n is a martingale.

For (1), *supermartingale* means the market is not favorable to me — I go down in expectation, never up. In a stock market that I know is going to go down, how do I make money? You sell (or short) — if I can short, I can make money. The theorem has to be true, so in this setting, we require that I can only buy (if I can sell, then I can obviously make money).

And W_n is how much money I'm going to have, if I buy H_m on day m. The intuition is that on day m I have my super-smart trading strategy, which is telling me to buy H_m amount of stock on this date; and

then how much money I make or lose is H_m times the change in value of the stock. And I sum over all days, and that's the value of my portfolio.

So (1) says in an unfair martingale, if you're only allowed to buy, this goes down. Of course H_m could just be 0 — you could just stay out of the market — in which case you won't lose, but the point is there's nothing you can do to gain.

For (2), now we're assuming M_n is a martingale, but we're allowing you to both buy *and* sell. So if the market is actually fair, you can use the sophisticated strategy of doing nothing (where you start with 0 and end with 0), or you can do whatever sophisticated strategy you want involving buying and selling and shorting, and you still end with what you started (in expectation).

From the philosophy of this, what can you say if M_n is a submartingale? If you only short then you'll lose money, and if you only buy then you'll gain money. We won't state it, but it can be proved in the same way.

Student Question. *What's the significance of H_n being bounded by c?*

Answer. There's a lot of paradoxes in probability once you have infinite amounts of money. We'll see some of them later today, but some of them are really crazy. The paradox here is something like this: I'm going to choose $\mathsf{Geom}(\frac{1}{2})$ (so it's 1 or 2 or 3 with exponentially decreasing probability). And I put 10^x dollars in one envelope, and 10^{x+1} in another (where x is the geometric random variable).

Now you come for my job interview as a trader. Because I have so much money, I give you a present as part of the interview. I tell you one envelope has 10^x and the other has 10^{x+1}. I'm so generous I show you what's inside one of the envelopes, and after seeing it you have to decide whether to take that envelope or take the other.

What's your optimal strategy? It turns out that the optimal strategy is to *always* choose the other envelope. If you do the math and compute, what's the expected gain if you choose the original vs. the other, you should always choose the other.

This has to do with the fact that if you think about the amount of money I need to have in order to run to this exercise, with probability 2^{-x} I need to have 10^x dollars. So the expected money involved is infinite. When infinite money becomes involved, lots of things in probability become very dicey.

You don't need things to be *quite* this bounded (you can assume weaker things than $|H_n| \le c$), but you do need something to prevent this sort of thing.

Let's prove one of these; we'll prove (1), because (2) follows from (1).

Proof of (1). We want to understand $\mathbb{E}[W_{n+1} \mid X_{[n]}]$ — the expected amount of money I have in time $n+1$, given everything that's happened up to time n. We can write this as

$$\mathbb{E}[W_{n+1} \mid X_{[n]}] = \mathbb{E}[W_n + H_{n+1}(M_{n+1} - M_n) \mid X_{[n]}].$$

There's everything that happened until today, and then there's what I'm doing today; so I don't need to write the whole sum.

Of these three terms, what's already determined as a function of $X_{[n]}$, and what's still random. I know W_n by time n, so it can come out of the expectation. And I also know H_n (by the definition of a predictable process) and M_n. So the only thing that's still random is M_{n+1} (which we don't yet know) — everything else goes out of the expectation. So we get

$$\mathbb{E}[W_{n+1} \mid X_{[n]}] = W_n + H_{n+1}(\mathbb{E}[M_{n+1} \mid X_{[n]}] - M_n).$$

And now we just use the fact that it's a supermartingale, so it goes down in expectation; that means

$$\mathbb{E}[M_{n+1} \mid X_{[n]}] - M_n \le 0.$$

And $H_{n+1} \geq 0$ by assumption; so $\mathbb{E}[W_{n+1} \mid X_{[n]}] \leq W_n$, which is what we wanted. □

It's all very simple mathematically, but conceptually it's somewhat interesting.

§17.2.3 Stopping times

Let's see an application of this to stopping times.

Definition 17.14. We say a random variable T is a stopping time for the sequence (X_n) if the event $\{T = k\}$ is determined by $X_0, X_1, \ldots, X_k$.

So whether I stop is determined by what I've seen so far.

Claim 17.15 — If (M_n) is a martingale and T is a stopping time (both with respect to (X_n)), then the new sequence $(M_{\min(n,T)})_n$ is also a martingale.

In stories, what does $M_{\min(n,T)}$ mean? You're a trader, so M_n is the value of the stock you're following every day. Before you were following M_n, and now you have this new strategy $M_{\min(n,T)}$. What does this mean? Once we hit some state, we actually hit that value — so you have a stopping time, and that stopping time is telling you, at that point I'm getting out of the market. So e.g., I have this stopping time telling me that if it looks like $\cos(\pi x/5)$, then I'm leaving the market; and from there, you just leave it as cash. This could be a better strategy than your friend, who stays in the market and loses all their money.

This is sort of intuitive given what we said before — it sounds I'm not cheating when I use this strategy (I decide when to stop based on the information I've seen so far), so I shouldn't be able to make money from this.

Proof. We can apply Theorem 17.13 with

$$H_m = \begin{cases} 0 & \text{if } m \geq T \\ 1 & \text{if } m < T. \end{cases}$$

Is this predictable? What do I need in order for this to be predictable? I need H_m to be a function of the variables up to time $m-1$. Is it?

First of all, let's check that this actually gives the correct thing — then

$$W_m = \sum_{m=1}^{\min(m,T)} (M_m - M_{m-1}) = M_{\min(m,T)}$$

(we can assume $M_0 = 0$).

So now we just need to make sure that we can determine whether $T \geq m$ from $X_0, \ldots, X_{m-1}$. This is because it's the complement of the event $\{T \leq m-1\}$, and *this* event I can determine by $X_0, \ldots, X_{m-1}$. So its complement, I also know by $X_0, \ldots, X_{m-1}$. □

§17.3 Confusing infinities

Since this was a confusing class, let's spend some time on confusing infinities. We'll start with a martingale example, and then if we have some time we'll see some non-martingale examples.

For martingales, we know that if (M_n) is a martingale, then it doesn't change in expectation — we have

$$\mathbb{E}[M_0] = \mathbb{E}[M_1] = \mathbb{E}[M_2] = \cdots.$$

This also means if T is a stopping time, then

$$\mathbb{E}[M_0] = \mathbb{E}[M_{\min(n,T)}]$$

for all n (because $(M_{\min(n,T)})$ is also a martingale).

Question 17.16. What happens if we take $n \to \infty$ — is it true that $\mathbb{E}[M_0] = \mathbb{E}[M_T]$?

Prof. Mossel runs a vote for what we think; as usual, we're all correct. It's true sometimes, and not true some other times; we need more conditions.

Here's a bad example, which also has to do with ∞.

Example 17.17

Take $S_n = \sum_{k=1}^{n} X_k$, where $X_k \sim \mathsf{Unif}\{\pm 1\}$ are i.i.d., and take

$$T = \min\{n \mid S_n = 5\}.$$

So I start from 0 dollars, and my goal in life is to make 5 dollars. And what I do is, I just wait until I have 5 dollars. That sounds like a great strategy.

Claim 17.18 — We have $\mathbb{P}[T < \infty] = 1$, and $\mathbb{E}[S_T] = 5 \neq 0 = S_0$.

So this will eventually end (we haven't proved this, but it's true); and of course, when I have 5 dollars, I have 5 dollars, which is not 0.

So Prof. Mossel has just taught us a way to make money — start from 0 dollars, play a fair game, and when you get 5 dollars, go out.

But don't go out and start a hedge fund based on this... It turns out that the way the process goes, the expected amount of money you lose before you gain 5 dollars is infinite. So the above claim is true, but we have to notice two things. First, $\mathbb{E}[T] = \infty$ — so you have to be *very* patient. And also, $\mathbb{E}[\min_{t \le T} S_t] = -\infty$ — so you have to have the ability to borrow an infinite amount of money.

Prof. Mossel has in the slides the original example:

Example 17.19 (Cassanova's martingale)

You play a fair game starting with 1 dollar. Each time you lose, you double your wager. You exit the first time you win.

So I'm determined to make money; here's how we do it. We toss a coin, playing with 1 dollar. If I win, I'm happy; if I lose, I lost a dollar. Now I bet 2 dollars. If I win, I'm happy (I made more than I lost); if I lose, next time I bet 4 dollars. Eventually I'll win, and I've made money. (We don't have to double; we can also triple, or multiply by 10.)

The problem is that in order to do this, you need an infinite amount of money. So it's the same story — it's a martingale, and at the stopping time you have more money than you started with, but you need infinite money for this.

§18 April 22, 2025

Prof. Mossel is not here today, so Jason (one of his postdocs) is teaching.

The plan today is we'll talk about *optional stopping* — what that means and when you can use it. Then we'll see some cool applications. There's lots of quant interview-style questions; hopefully after today you'll know what to do with them. Then at the end, we'll talk about something called *martingale convergence* (we might not get to there — Jason found out he'd be doing this 12 hours ago, so he doesn't exactly know the timing of how long things will take).

§18.1 Review

Last time, we saw the following definition:

Definition 18.1. An integer-valued random variable T is a stopping time with respect to $X_0, X_1, \dots$ if the event $\{T = k\}$ is determined by $X_0, \dots, X_k$.

So you can look at the first k things and determine whether to stop at that time.

Last time, we also saw the following fact:

Corollary 18.2

If M is a martingale and T a stopping time, then $(M_{T\wedge n})$ is also a martingale.

The same is also true for supermartingales.

§18.2 Optional stopping

It turns out there's a lot of power in this fact. The first thing we can do is use it to prove a 'baby' version of optional stopping.

Theorem 18.3

Suppose M is a martingale. Then $\mathbb{E}[M_T] = \mathbb{E}[M_0]$ if T is a stopping time with $\mathbb{P}[T < \infty] = 1$, and $|M_{T\wedge n}| \le K$ for all n (where K is a constant).

If T was a fixed time, we know that $\mathbb{E}[M_T] = \mathbb{E}[M_0]$ by the martingale property. This says that we can extend this to stopping times as well, but we need some other conditions — for example, boundedness.

Why do yo need this? Here's a bad example that *doesn't* satisfy this.

Example 18.4

Suppose M is a simple random walk on the integers — so you start at 0 and take a step by ± 1 each time. Set $T = \min\{M_t = 1\}$.

Does this theorem actually hold here, and if not, what's the issue? We have $\mathbb{E}[M_T] = 1$ — I stop the martingale when it's 1. But it started at 0, so

$$\mathbb{E}[M_T] = 1 \neq \mathbb{E}[M_0] = 0.$$

(You can show that T is finite with probability 1, so M_T makes sense.)

What was the issue? The boundedness condition isn't true — if T hasn't happened yet but n is a trillion, it could be that M is way negative. So $M_{T\wedge n}$ isn't bounded.

We'll see in the proof why you need something like this assumption. (This is just one assumption you could make that lets things work; there are other assumptions you could make instead, which we'll also see.)

Proof. We'll use Corollary 18.2, which we can think of as an even more baby version of optional stopping. By this corollary, we know that

$$\mathbb{E}[M_{T\wedge n}] = \mathbb{E}[M_0]$$

for each $n \in \mathbb{N}$ (this is just a consequence of Corollary 18.2 and the martingale property). How can we try to get rid of the n? We can try taking limits as $n \to \infty$ — we can say

$$\lim_{n\to\infty} \mathbb{E}[M_{T\wedge n}] = \mathbb{E}[M_0].$$

But what we really want is to put the limit on the *inside*. Why can we do this?

This is where we use the assumption $|M_{T\wedge n}| \le k$. We know that $T \wedge n \to T$ (almost surely), since $T < \infty$ almost surely; this means $M_{T\wedge n} \to M_T$ almost surely. And by assumption we have $|M_{T\wedge n}| \le K$ for every n. So we can use the dominated convergence theorem (or bounded convergence theorem, or whatever it's called) to say

$$\lim_{n\to\infty} \mathbb{E}[M_{T\wedge n}] = \mathbb{E}\left[\lim_{n\to\infty} M_{T\wedge n}\right] = \mathbb{E}[M_T]. \qquad \square$$

Remark 18.5. The part $\lim_{n\to\infty} \mathbb{E}[M_{T\wedge n}] = \mathbb{E}[M_0]$ is true in general; to get $\mathbb{E}[M_T] = \mathbb{E}[M_0]$, you need some result that allows you to push the limit inside. Here we used a boundedness assumption to use the dominated convergence theorem; we'll see other conditions later that also allow you to do this.

§18.3 Some applications — 'fun' with optional stopping

Now let's see some nice applications of this. The key to all of them is you have to find the right martingale and the right stopping time to get the computations you want.

§18.3.1 Gambler's ruin

Example 18.6 (Gambler's ruin)

Say we have integers $a < 0 < b$. We perform a random walk starting at 0, where at every step we move right with probability p and left with probability $q = 1 - p$ (we assume $p > \frac{1}{2}$).

What is the probability of hitting a before b?

One way you can solve this is using recurrences; but we'll see that optional stopping gives you very simple and easy ways to do it — it's just a matter of picking the right martingale to do the work for you.

What's a good martingale to look at here? Let's define a martingale

$$M_t = \left(\frac{1-p}{p}\right)^{X_t},$$

where X_t is your position at time t. The first thing we should check is, is this actually a martingale? This means you're supposed to do two things. One is that you have to check that it has a mean, so it's integrable; but X_t is always between $-t$ and t, so for any given t these are all integrable.

So now we just need to make sure that it has the right conditional expectations to check that it's a martingale.

Claim 18.7 — M_t is a martingale.

Proof. If we want to compute $\mathbb{E}[M_{t+1} \mid M_t]$, we move right one step with probability p and left one step with probability $1-p$, so we get

$$\mathbb{E}[M_{t+1} \mid M_t] = p \cdot M_t \cdot \frac{1-p}{p} + (1-p) \cdot M_t \cdot \frac{p}{1-p} = M_t(1-p) + M_t p = M_t. \qquad \square$$

Now let's try to use optional stopping to answer this question. We have our martingale, so now let's define our stopping time. I want to figure out, do I hit a before b? So a nice stopping time to think about is the time I hit either a or b — we define

$$T = \min\{n \mid X_n \in \{a, b\}\}.$$

Now we've got the martingale and stopping time, so there's only one thing left to do — to try to apply the optional stopping theorem.

First, T really is a stopping time. Then we have to check, does the stopped martingale satisfy the boundedness condition? We can check that it does: we always have

$$|M_{T \wedge n}| \leq \left(\frac{1-p}{p}\right)^a$$

(we assumed $p > 1/2$, so $\frac{1-p}{p} < 1$, which means the worst case is when the exponent is as negative as possible, and we stop when we hit either a or b).

So it's a stopping time, and it's bounded. We should also check T is finite with probability 1, but you can hopefully convince yourself that's true (no matter where I am in the interval, there's a $(1-p)^{b-a}$ probability that you always go to the left and end up at a).

This means we can apply optional stopping. We have $\mathbb{E}[M_0] = 1$ (since we start at 0). Optional stopping tells us this is equal to $\mathbb{E}[M_T]$. And how can we rewrite M_T in terms of the probabilities we care about? If we let p_a be the probability we stop at a, then we get

$$\mathbb{E}[M_T] = p_a \left(\frac{1-p}{p}\right)^a + (1-p_a)\left(\frac{1-p}{p}\right)^b$$

(since if we stop at a then M_T ends up being $(\frac{1-p}{p})^a$, and if we stop at b then it's $(\frac{1-p}{p})^b$).

So we've got a very nice equation

$$1 = p_a \left(\frac{1-p}{p}\right)^a + (1-p_a)\left(\frac{1-p}{p}\right)^b$$

based purely on optional stopping. And if you do the algebra, this tells you

$$p_a = \frac{1 - (\frac{1-p}{p})^b}{(\frac{1-p}{p})^a - (\frac{1-p}{p})^b}.$$

It's a nice application of optional stopping — I know what the value of the thing is when it stops, the only things I don't know are these probabilities p_a and $1-p_a$. And you can solve for it using optional stopping.

One thing we should mention is we said $p > \frac{1}{2}$. The reason fro that is that if $p = \frac{1}{2}$, then M_t is just 1, and this would tell you nothing.

So now we'll do the case where $p = \frac{1}{2}$. For this, we'll have to use a different martingale, but it's actually an easier one.

Example 18.8

Suppose we have the same setup, but $p = \frac{1}{2}$.

Last time, we used the martingale $M_t = (\frac{1-p}{p})^{X_t}$, but that doesn't say anything when $p = \frac{1}{2}$. When $p = \frac{1}{2}$, we're doing a simple random walk left and right; so we can use a simpler martingale, just X_t itself. (It goes up or down by 1 with equal probability at each step, so it's a martingale.)

So we'll just look at X_t directly. Now we want to apply optional stopping again. For the same reasons, we have $\mathbb{P}[T < \infty] = 1$, and it's even easier to see that $|X_{T\wedge n}|$ is bounded — it can't leave the interval $[a, b]$, so we always have $|X_{T\wedge n}| \leq \max\{|a|, |b|\}$. So we've got the boundedness condition, and we can again apply optional stopping to say that

$$\mathbb{E}[X_T] = \mathbb{E}[X_0] = 0$$

(because we started at 0). And similarly to before, we can write $\mathbb{E}[X_T]$ as

$$\mathbb{E}[X_T] = p_a \cdot a + (1 - p_a) \cdot b$$

(since if we stop at a then the value is a, and if we stop at b the value is b); and if you do a bit of algebra, you get

$$p_a = \frac{b}{b - a}.$$

Just as a sanity check, when b is really really big (e.g., $a = -5$ and $b = 10^7$), I'd expect to hit -5 before 10^7. And indeed, we'd have 10^7 in the numerator and $10^7 + 5$ in the denominator, so this number would be very close to 1; so this checks out. So at least we've got the right signs and everything.

But this is kind of a one-line way to compute these probabilities, based purely on optional stopping.

We can compute other interesting quantities that aren't super obvious how to compute, again using optional stopping, but we're going to have to be a little bit more careful this time.

Suppose we want to understand a different quantity, instead of the probability we stop at a or b:

Example 18.9

What's the expected *time* it takes to get to a or b?

In other words, we want to understand $\mathbb{E}[T]$ itself.

Again, we need to define a martingale to do the work for us, and then we want to apply optional stopping and go through the same machinery. But now we have to think a bit more carefully about what martingale we want to use. This might not be super obvious, but it turns out the right martingale to use is the following: let's look at

$$M_t = X_t^2 - t.$$

Claim 18.10 — M_t is a martingale.

Proof. We can check that

$$\mathbb{E}[M_{t+1} - M_t \mid M_t] = \mathbb{E}[(X_t + Z_{t+1})^2 - (t+1) - X_t^2 + t \mid M_t],$$

where Z_{t+1} is the step at time t. If you expand the inside, the X_t^2 terms cancel, and we always have $Z_{t+1}^2 = 1$, so that cancels out the 1 from $t + 1$ vs. t. So all we're left with is the cross-term $\mathbb{E}[2X_tZ_{t+1} \mid M_t]$. And Z_{t+1} is independent of what happened before and has expectation 0; so this whole expectation is 0, which means we have a honest martingale. □

Now the interesting part is, can we use optional stopping, and what does it give us? What we'd *like* to say is that

$$0 = \mathbb{E}[M_0] \stackrel{?}{=} \mathbb{E}[M_T] = \mathbb{E}[X_T^2 - T] = \mathbb{E}[X_T^2] - \mathbb{E}[T].$$

If we have this chain of equalities, we can move $\mathbb{E}[T]$ to the other side, and what we'd get is that

$$\mathbb{E}[T] = \mathbb{E}[X_T^2].$$

Are we happy with this, or did we cheat? We did slightly cheat — there's a condition in our theorem that the (stopped) martingale has to be bounded at all times. But the issue is that's not actually the case for us. If T is very big for some reason and I haven't stopped yet, X_t could be very close to 0 but t could be very big, so the whole thing could be very negative. So this equality was a cheat, at least for now.

But let's ignore that for a second; we'll talk about how to fix that later. If this is all valid and we get $\mathbb{E}[T] = \mathbb{E}[X_T^2]$. And what do we do now? We already figured out the probabilities p_a and $p_b = 1 - p_a$, so we can again write this out as

$$\mathbb{E}[X_T^2] = p_a \cdot a^2 + (1 - p_a) \cdot b^2 = \frac{b}{b-a} \cdot a^2 + \frac{-a}{b-a} \cdot b^2 = -ab$$

(remember that a is negative).

So now the thing we have to do is figure out how we get rid of this cheat. But it'll be similar to what we did in the proof of optional stopping; we just have to be a bit more careful.

The point is the version we stated is a baby version that won't always be satisfied; but you can usually tweak the proof to make it work for the application you're interested in, as long as it's true there.

Here, instead of directly going from $\mathbb{E}[M_0]$ to $\mathbb{E}[M_T]$, we'll use the same idea of the proof: We're always allowed to say

$$\mathbb{E}[M_{T \wedge n}] = 0$$

(now we're just using optional stopping, but instead of waiting until time T, we're waiting until $T \wedge n$ for some *fixed* n); so by running the same argument, we get

$$\mathbb{E}[T \wedge n] = \mathbb{E}[X_{T \wedge n}^2].$$

(This is just because $(M_{T \wedge n})$ is a honest martingale.)

Now we need to figure out a way to get rid of n. First, we're allowed to put limits on the outside and say

$$\lim_{n \to \infty} \mathbb{E}[T \wedge n] = \lim_{n \to \infty} \mathbb{E}[X_{T \wedge n}^2].$$

And as always, the issue is, can we pass limits *inside* the expectation?

First, on the left we can use the *monotone convergence theorem* — the sequence $T \wedge n$ increases to T (as we make n bigger and bigger, this only goes up). So this is a monotone sequence, which means we can use the monotone convergence theorem to say that

$$\lim_{n \to \infty} \mathbb{E}[T \wedge n] = \mathbb{E}\left[\lim_{n \to \infty} (T \wedge n)\right] = \mathbb{E}[T].$$

(Whenever I have a positive increasing sequence, I'm always allowed to put limits inside by monotone convergence, because I don't have to worry about weird oscillations.)

Now we need to figure out the limit on the right. But we can again use one of these convergence theorems. We know that $X_{T \wedge n} \in [a, b]$, so $X_{T \wedge n}^2 \leq \max\{a^2, b^2\}$. That means this sequence is bounded by a constant, so you can use the dominated convergence theorem to get that

$$\lim_{n \to \infty} \mathbb{E}[X_{T \wedge n}^2] = \mathbb{E}\left[\lim_{n \to \infty} X_{T \wedge n}^2\right] = \mathbb{E}[X_T^2].$$

So you have to be a bit careful about what convergence theorem you use when the version of the theorem we stated doesn't apply, but 95% of the time there will be a convergence theorem that gets you the conclusion you want.

§18.3.2 Wald's identity

These were three cute applications of using optional stopping to compute quantities that it's maybe not otherwise so obvious how to compute. Now let's do a couple more applications.

Theorem 18.11 (Wald)

Suppose Y_1, Y_2, ... are i.i.d. random variables with mean μ and T is a stopping time with $\mathbb{E}[T] < \infty$. Then we have

$$\mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] = \mu\mathbb{E}[T].$$

Note that $\mathbb{E}[T] < \infty$ is stronger than just saying $\mathbb{P}[T < \infty] = 1$.

So I'm summing up a random number of these Y_t's, where I stop at a stopping time. This is a natural statement — it's saying I don't need to worry about some weird correlations. If I expect there to be 15 of the Y's, then the expected sum is just going to be $\mathbb{E}[Y_i]$ times 15.

Let's see how to prove this. Again, the moral of today is just to find a martingale to do the work for us, and then (try to) apply optional stopping.

What martingale should we use (which should ideally have something to do with $\sum_t Y_t$)? We'll define

$$M_n = \sum_{i=1}^{n} Y_i - \mu n.$$

This is a martingale because each time I add a new Y_i, I subtract off its mean, which means the expected increment is 0. So this is a martingale.

And what we're trying to compute is kind of $\mathbb{E}[M_T]$. To avoid the boundedness issue, what we can say again is that for any fixed n, we have

$$\mathbb{E}[M_{T\wedge n}] = \mathbb{E}[M_0] = 0.$$

And what is $\mathbb{E}[M_{T\wedge n}]$? Just writing this out, it's

$$\mathbb{E}\left[\sum_{i=1}^{T\wedge n} Y_i\right] - \mu\mathbb{E}[T \wedge n]$$

(this is just using Corollary 18.2 — I'm stopping the martingale early, so I don't need to worry about any of the boundedness stuff). And rewriting this, it implies that

$$\mu\mathbb{E}[X_{T\wedge n}] = \mathbb{E}\left[\sum_{i=1}^{T\wedge n} y_i\right].$$

Now ideally we want to take limits again on both sides — as usual, we have

$$\lim_{n\to\infty} \mu\mathbb{E}[X_{T\wedge n}] = \lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{T\wedge n} y_i\right].$$

But now we have to justify why we can push things inside.

On the left-hand side, we can again use monotone convergence — the sequence $T \wedge n$ is increasing and converges to T, so you can apply the same monotone convergence to say that

$$\lim_{n\to\infty} \mu\mathbb{E}[T \wedge n] = \mu\mathbb{E}[T].$$

The trickier thing is understanding why we can push the limit inside the right-hand side — is

$$\lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{T\wedge n} y_i\right] = \mathbb{E}\left[\sum_{i=1}^{T} Y_i\right] ?$$

So we need to figure out the right argument for why *this* might be the case.

We again want to use one of these convergence theorems (dominated convergence or something), so we need to understand the guy on the inside a bit more as we increase n.

For any n, we can always say that

$$\left|\sum_{i=1}^{T\wedge n} Y_i\right| = \left|\sum_{i=1}^{\infty} Y_i \mathbf{1}\{i \le T\wedge n\}\right|$$

(we write an infinite sum, but the only terms we keep are the ones between 1 and $T \wedge n$). I'm allowed to push absolute values inside sums, so we can bound this by

$$\left|\sum_{i=1}^{\infty} Y_i \mathbf{1}_{\{i\le T\wedge n\}}\right| \le \sum_{i=1}^{\infty} |Y_i|\, \mathbf{1}\{T\wedge n \ge i\}.$$

And now we can drop the n, and always upper-bound this by

$$\sum_{i=1}^{\infty} |Y_i|\, \mathbf{1}\{T \ge i\}.$$

We want to apply one of the convergence theorems. We've gotten rid of all the n's, so this is just a single random variable. And I can apply the dominated convergence theorem if it's the case that this is an integrable random variable (the DCT says that if I have a family of random variables and they're all bounded by an integrable one, then I can pass limits inside). So we need to understand, is the expected value of this guy finite?

And if you do the calculation, you have

$$\mathbb{E}\left[\sum_{i=1}^{\infty} |Y_i|\, \mathbf{1}\{T \ge i\}\right] = \sum_{i=1}^{\infty} \mathbb{E}[|Y_i|\, \mathbf{1}\{T \ge i\}]$$

(this is a positive sequence, so I can always push limits inside). I need to do something about the terms on the inside. First, $|Y_i|$ and $\mathbf{1}\{T \ge i\}$ are *independent* — this is because whether $T \ge i$ or not depends just on $Y_1, \ldots, Y_{i-1}$. (That's because $\{T \ge i\}$ is the complement of $\{T \le i-1\}$, and you can tell whether a stopping time has occurred based on just the stuff that's happened already — if I'm including Y_i in the sum, then I know that I'm supposed to include it based on just $Y_1, \ldots, Y_{i-1}$.)

So we've got an expectation of two guys that are independent from each other, which means we can pull them apart — so

$$\mathbb{E}[|Y_i|\, \mathbf{1}\{T \ge i\}] = \mathbb{E}[|Y_i|] \cdot \mathbb{E}[\mathbf{1}\{T \ge i\}]$$

by independence. And $\mathbb{E}[|Y_i|]$ is just a number (which doesn't depend on i, since they're i.i.d.), which we'll call M. And $\mathbb{E}[\mathbf{1}\{T \ge i\}]$ is just $\mathbb{P}[T \ge i]$.

So putting all these things together, the expectation we wanted becomes

$$\mathbb{E}\left[\sum_{i=1}^{\infty} |Y_i| \cdot \mathbf{1}\{T \ge i\}\right] = M \sum_{i=1}^{\infty} \mathbb{P}[T \ge i].$$

And we have $\sum_{i=1}^{\infty} \mathbb{P}[T \geq i] = \mathbb{E}[T]$ (this is a general fact about $\mathbb{N}$-valued random variables, called 'layered cake' or something). And we assumed $\mathbb{E}[T] < \infty$, so the whole expectation is finite.

So that was a lot, but the point is we had a bunch of random variables $\sum_{i=1}^{T \wedge n} Y_i$, and we showed they're uniformly bounded by a single guy $\sum_{i=1}^{\infty} |Y_i| \mathbf{1}\{T \geq i\}$ who has finite expectation (i.e., is integrable). So the punchline is that you can apply dominated convergence to push the limit inside and get that

$$\lim_{n \to \infty} \mathbb{E}\left[\sum_{i=1}^{T \wedge n} Y_i\right] = \mathbb{E}\left[\sum_{i=1}^{T} Y_i\right].$$

And finally, on the left-hand side we had $\mu \mathbb{E}[T]$; so we've proved the theorem.

Remark 18.12. The theorem looks simple, but there are these weird things to check where you have to interchange limits and expectations and check that this is allowable; and that's where the condition $\mathbb{E}[T] < \infty$ comes up. So this is a bit more subtle than you might expect from the statement.

§18.3.3 Coin flipping patterns

We'll probably get to see one more cool application (and we probably won't get to the martingale convergence stuff, but that's okay). Here's where we will learn how to dominate quant interviews.

Example 18.13

How long in a sequence of coin flips does it take to see some pattern (e.g., how long does it take to see a consecutive sequence of HTH)?

There's a stopping time associated with this, the first time we see the pattern in the sequence — for example, if the sequence is TTTHHTH, the HTH is the first time we see this, so the stopping time is $T_{\mathsf{HTH}} = 7$.

We'll use HTH for concreteness, but it's not super hard to generalize.

It turns out you can answer this using Markov chain methods and recurrences and stuff like that, but it gets kind of annoying and ugly. It turns out that with optional stopping and martingales, there's a very clever way to understand these things in a much simpler way; it's just a question of finding an even more creative martingale than we did before.

What we'll do is think about martingales in terms of gambling (martingales were maybe introduced because of gambling applications — the point is that if I have a martingale and I'm gambling on the outcomes, I can't make any money; it's a fair game). So we're going to use that kind of interpretation.

Let's focus on the HTH configuration, and let's imagine we have a stream of coins

$$\mathsf{THHTTHTH}.$$

And let's imagine we have gamblers trying to make money on this sequence of coins.

Imagine that at each time t, I get a new Gambler t. What he does is, he's going to try to predict this sequence HTH. So at time t, he's going to bet \$1 on H. In this sequence, he was wrong — it was actually T.

If he loses, what does he do? He loses his dollar and quits. But if he wins, then he'll bet double-or-nothing the next step.

So each gambler who comes at time t is first going to bet H; then if they're right they'll have \$2 and bet on T, and if they win they'll get \$4 and bet on H, and so on.

So a new guy comes at time $t+1$, bets H there, and succeeds, so he makes a dollar. He's trying to predict the sequence HTH, so he bets T on the next step. He'll be incorrect there (because it's H), so he loses all his money.

But then we have a new gambler who comes on time $t+2$, who'll be correct and then correct and then wrong; and so on.

We'll have new gamblers coming in at each step. And finally, the guy who comes in at time $t+5$ is going to succeed. There'll also be the guys who came after him (the one at $t+6$ is wrong, the one at $t+7$ is right), but we don't go any further than this because at that point, the $t+5$ gambler has won the game.

So this is kind of convoluted, but we have all these gamblers betting in this way, and we're going to figure out what's the right martingale based on this.

Each gambler, when he arrives and makes this sequence of bets — it's ± 1 flips and they're betting with no information on the rest. So for each guy who shows up, their expected payoff is 0. (It doesn't even matter how they bet; they have fair odds every time, so their expected payout is 0.)

A bit more formally, let's define a new martingale

$$M_t = P_t - t,$$

where P_t is the total money received by all the gamblers up to time t. For example, the first guy contributes 0, the guy who wins contributes 8, and so on. The t reflects the pay-in — each guy pays a dollar to start the game.

Then M_t is a martingale (since P_t is the sum of what the gamblers get out and t is the sum of what they pay in, and the game is fair).

Now let τ be the stopping time for HTH (or any sequence you want, but let's just do that one). We're going to apply optional stopping. We won't check this time that it's valid, but it turns out that it is. Because it's a martingale (and at the beginning no one has entered or won anything), we have

$$0 = \mathbb{E}[M_0] = \mathbb{E}[M_\tau] = \mathbb{E}[P_\tau] - \mathbb{E}[\tau].$$

This is promising — it tells us $\mathbb{E}[\tau] = \mathbb{E}[P_\tau]$ is the expected money these gamblers get when they bet in this way.

But how much is that? Every gambler who came before the winner has lost, so they have 0 payout (they were betting that their part of the sequence would be HTH, and that didn't happen). So everyone loses except possibly the last three people (they were all predicting the next three would be HTH, and they lost because it didn't happen). The third-last person predicted correctly — he predicted that it would be HTH, and he was right. So he first bet \$1, then bet \$2, then bet \$4 (he does double or nothing at every step), and his payout is \$8.

The person after him makes nothing, because at the beginning he predicted H, and it wasn't.

And the last person (who arrived at the end) predicted H and they were right, but then the game ended. So their payout is \$2.

So the point is that everyone's payout is 0 except possibly the last 3; the third-last won and got 8, the second-last lost and got 0; and the last got 2. And this is true no matter what the sequence is — we always have

$$P_\tau = 8 + 0 + 2 = 10.$$

So moving things over, we get $\mathbb{E}[\tau] = 10$.

This is kind of a weird argument — interpreting a sequence of gamblers and saying that in the end the game is fair, and I know how much they won at the end of the game, so somehow that magically tells you by optional stopping what the answer is.

Let's see another example of this:

Example 18.14

What if instead of HTH, we're considering HHH?

You can do the same construction, where we have a bunch of coin flips and a bunch of gamblers. All the gamblers ending before the winner have payoff 0. The winner has payoff 8 (for the same reason as before). The person after him predicted the first two right, and then the game stopped, so they got 4; the last got 2. So here the total payoff was 14; and when we do the same optional stopping thing, we get 14 (in place of 10).

So this means it takes longer to get HHH than HTH. Intuitively this makes some sense because with HHH, if I have a sequence of H's and I get a T, I'm starting from ground 0 again.

You can do this for any sequence you want; you'll get a very similar calculation, where you just need to figure out who in the end made money (by looking at, when I shift over the sequence, is the overlap correct?).

§19 April 24, 2025

§19.1 Convergence of martingales

Today we're going to talk about convergence of martingales. In probability theory, after we define stuff, we often want to understand what happens in some limit — we want to show that in some sense, sometimes martingales converge.

Before stating a theorem, what's an example of a martingale that converges or a martingale that doesn't converge?

Example 19.1

The simple random walk $S_n = \sum_{i=1}^n X_i$ (where $X_i \sim \mathsf{Unif}\{\pm 1\}$ are i.i.d.) does not converge (wherever we are, the next step we move up or down with probability $\frac{1}{2}$).

Example 19.2

If we let $S_n = \sum_{i=1}^n X_i$ where the X_i's are independent with $X_i \sim \mathsf{Unif}\{\pm 2^{-i}\}$, this does converge.

So this is sort of a random binary number — we choose a digit, then another digit, and so on. So this does converge to a number. This is obviously a martingale — the expected value of where I am next time is unbiased.

Example 19.3 (Unbiased gambler's ruin)

If I'm in a casino, and with probability $\frac{1}{2}$ I gain a dollar or lose a dollar, but I leave once I lose all my money or get n dollars, then this converges (since it becomes fixed from some point on).

Now let's state the theorem and see how it fits into these examples.

Theorem 19.4

Let $(M_n)_{n\geq 0}$ be a supermartingale such that $M_n \geq 0$ for all n. Then there exists a random variable M_∞ such that:

- $M_\infty = \lim_{n\to\infty} M_n$ with probability 1.
- $\mathbb{E}[M_\infty] \leq \mathbb{E}[M_0]$ (more generally, $\mathbb{E}[M_\infty] \leq \mathbb{E}[M_n]$ for all n).

Recall that a supermartingale means it goes down in expectation. (In particular, if we have a martingale this theorem also holds.) The condition that's interesting is we need it to always be nonnegative. Then there exists a limit. Note that M_∞ is a *random variable*, not a number (as in the law of large numbers).

§19.1.1 Examples

Let's see if this covers some of the examples we have here. For Example 19.2, we required $M_n \geq 0$ in the theorem, but the same theorem holds if we instead require $M_n \geq -10^6$ (since we can always shift). And in that example we always have $S_n \geq -2$, so the theorem guarantees there's a limit.

And unbiased gambler's ruin also satisfies these conditions.

The simple random walk does *not* satisfy this condition, because a simple random walk can get arbitrarily negative.

Another question: supermartingales go down in expectation, so of course I should need to have $\mathbb{E}[M_\infty] \leq \mathbb{E}[M_0]$.

Question 19.5. If I just have martingales, can I replace $\mathbb{E}[M_\infty] \leq \mathbb{E}[M_0]$ by equality?

First, if M_n is *bounded* — meaning that $\mathbb{P}[a \leq M_n \leq b] = 1$ for all n — then by applying the theorem to both M_n and $-M_n$ (which is also a martingale), we get that $\mathbb{E}[M_\infty] = \mathbb{E}[M_0]$.

But now let's think about some example. This will be essentially a random walk; but there's no convergence for random walks, so we'll have to change it a little bit.

Example 19.6

Let $S_n = 10 + \sum_{i=1}^n X_i$ where $X_i \sim \mathsf{Unif}\{\pm 1\}$ are i.i.d. Define a stopping time $T = \min\{n \mid S_n = 9\}$. Define $M_n = S_{\min(n,T)}$.

So I'm going left and right, winning and losing dollars. But I'm very conservative, so if I started with 10 dollars, then I stop when I have 9.

We saw in class that M_n is a martingale, and we always have $M_n \geq 9$, so the theorem applies. And $\mathbb{E}[M_\infty] = 9$ — we know that for a random walk we'll eventually visit everything, even if it takes very long; so eventually we're going to reach 9 and stop there, which means M_∞ is just the constant 9 (i.e., $\mathbb{P}[M_\infty = 9] = 1$). So in this case, $\mathbb{E}[M_\infty] < \mathbb{E}[M_0] = 10$.

So even for martingales, we cannot be too optimistic here (unless the martingale is bounded both from above and from below).

§19.1.2 Some proof ideas

We're not actually going to give the proof of this theorem, which is somewhat complicated. But we'll see some ideas from the proof.

This is the main idea, and results like this are used a lot when dealing with martingales.

Lemma 19.7 (Max inequality for (super)martingales)

Let $X_n \geq 0$ be a supermartingale. Then for all $x > 0$, we have

$$\mathbb{P}\left[\sup_t X_t \geq x\right] \leq \frac{\mathbb{E}[X_0]}{x}.$$

When we write $X_n \geq 0$, we mean that $X_n \geq 0$ with probability 1 for all n.

What does this lemma say? Imagine we just have a martingale. It's some process that goes up and down and up and down. I'm interested in not what happens at a particular time, but the supremum — the largest possible value obtained by this process at all times. And I'm saying the probability this is bigger than x is at most $\mathbb{E}[X_0]/x$.

Note that if we didn't have the supremum, this would be Markov's inequality — as a sanity check, for each *fixed* t, we have

$$\mathbb{P}[X_t \geq x] \leq \frac{\mathbb{E}[X_t]}{x} \leq \frac{\mathbb{E}[X_0]}{x}$$

(the first inequality is by Markov; the second is because it's a supermartingale). So for each fixed t, this is an easy statement. But we can't do a union bound over all t's, because there's infinitely many. So this is a much stronger statement — it's not just that each individual t satisfies this, but in fact the supremum of the *whole* process satisfies the same inequality.

Proof. Let T be the stopping time $T = \min\{t \mid X_t \geq x\}$. And then we're going to define a new supermartingale $M_t = X_{\min(t,T)}$. Because T is a stopping time, we know that M_t is going to be a supermartingale. We still need to do some limiting argument — what we want is to say that

$$\mathbb{P}\left[\sup_t X_t \geq x\right] = \lim_{t\to\infty} \mathbb{P}[M_t \geq x].$$

Here we are using some probability — the event $\{\sup_t X_t \geq x\}$ is the limit of the events $\{M_t \geq x\}$ (because there has to be some time where we go above x).

And for the right-hand side, we can just use Markov — this is

$$\mathbb{P}[M_t \geq x] \leq \frac{\mathbb{E}[M_t]}{x} \leq \frac{\mathbb{E}[M_0]}{x} = \frac{\mathbb{E}[X_0]}{x}.$$ □

Remark 19.8. The advantage of M over X is that X just tracks a specific time, but M keeps track of everything that had happened up to time t — if we ever go above x, then M will stay above x. So M kind of does the union bound for you; and we get to do this for free, because it's still a supermartingale.

Why does this help us prove Theorem 19.4? We'll give a sketchy sketch.

Proof sketch for Theorem 19.4. The idea is we'll fix some numbers $0 \leq a < b$.

Claim 19.9 — The probability that M_n crosses from a to b infinitely many times is 0.

We want to prove something converges. And the way we're going to do that is, well, let's first prove that it doesn't do the following crazy thing — I'm going to fix some number a and some number b (you should think of them as being very close, though we won't draw it that way). We have a discrete process, but we'll draw it as if it's continuous. And what we want to show is that it doesn't fluctuate between a and b infinitely many times. If it does that, obviously it doesn't converge (because it keeps going below a and

above b). What we want to show is that this behavior doesn't happen — no matter what a and b I'm looking at, either it'll eventually stay above a or below b, but it won't fluctuate infinitely many times.

This is where we'll be sketchy (to make this precise, you really need to define a bunch of stopping times). But the point is, each time you cross from a to b, how much do you pay? If $X_2 \le a$, then by this lemma we have

$$\mathbb{P}\left[\sup_{t\ge 2} X_t \ge b\right] \le \frac{a}{b}.$$

If I'm in a supermartingale, the chance that I ever get to some threshold above my expectation is at most the ratio of my expectation and that threshold (by Lemma 19.7). So each time I cross from a to b, I pay a $\frac{a}{b}$ probability factor. (It's something that may or may not happen, but the probability it happens is $\frac{a}{b}$.) And I want it to happen *infinitely* often. For this, I have to be very very fortunate — this will happen with probability $(\frac{a}{b})^\infty = 0$.

This is for *fixed* a and b. But maybe my martingale doesn't care about a and b — what if it lives above them and there's some other a and b where it goes up and down and doesn't converge. So how do I deal with all these a and b's together? I want to prove there's *no* $0 < a < b$ such that my martingale fluctuates between a and b forever. Here we showed this is true for *one* pair $0 < a < b$, but now I want to show it's true for all pairs $0 < a < b$. So how am I going to do that? (I have to show it when $a = 1$ and $b = 2$, but also when $a = 1$ and $b = 1.0001$.)

We want to union bound, but the problem is that we can't union-bound over all numbers, because there's uncountably infinitely many of them. We really need to take a union bound over something *countably* infinite. And the way you do that is by taking a union bound over all *rational* numbers (or you can use dyadic numbers, or anything like that).

So to conclude, we take a union bound over all $(a, b) \in \mathbb{Q}^2$. If you oscillate many times between a and b infinitely often, even if they're not rational, then you can find a number slightly bigger than a and slightly smaller than b that are rational, and it's also going to oscillate between those. So it's enough to just look at rationals. □

Student Question. *Can you explain the first equality in the proof of the lemma, that*

$$\mathbb{P}\left[\sup_t X_t \ge t\right] = \lim_{t\to\infty} \mathbb{P}[M_t \ge x]?$$

Answer. This goes back to a simple fact in probability theory that if you have a countably infinite collection of events (A_n), then

$$\mathbb{P}\left[\bigcup_{n=1}^{\infty} A_n\right] = \lim_{k\to\infty} \mathbb{P}\left[\bigcup_{n=1}^{k} A_n\right].$$

Now, the events we're looking at are $X_1 \ge x$, $X_2 \ge x$, $X_3 \ge x$, ...

Student Question. *But why can't it be that it approaches x from below with positive probability without ever crossing x?*

Answer. Good question. We're cheating a little bit. There's a bunch of ways of fixing it. One way to do it is instead of this, we can say

$$\mathbb{P}\left[\sup_t X_t \ge x\right] \ge \lim_{t\to\infty} \mathbb{P}[M_t \ge x - \varepsilon] \le \frac{\mathbb{E}[M_t]}{x-\varepsilon} \le \frac{\mathbb{E}[X_0]}{x-\varepsilon},$$

and then you take the limit as $\varepsilon \to 0$.

§19.2 Polya's urn

Now let's see some other examples. Here's a very classical and important example in probability (and somehow it's also very popular in Bayesian inference and machine learning, though we won't explain why).

Example 19.10 (Polya's urn)

Imagine we have a white ball and a black ball (they're actually white and red pieces of chalk). And imagine I have an urn (this is an empty chalkbox). I'm very bored, so I decide to play a game. I shut my eyes, pick a ball at random inside the urn. I look at what color I got. It's a black ball. So I put another black ball into the urn.

Then I shuffle the urn. Then I take a random ball. This one is white, so I put another white ball in.

And I continue — every time I pick a ball at random, and I put another ball of that color in (this is some sort of *reinforcing* — every time I see a color, I add another ball of the same color).

So at time 0, there's two balls — one black and one white. Then at each step, I pick a random ball. Then I return it with another ball of the same color.

Claim 19.11 — Let M_t be the fraction of black balls. Then M_t is a martingale.

So the fraction of black balls is a martingale.

We'll prove this in a second. But what does this imply, by the theorem? BY the theorem, $M_t \to M_\infty$, and $\mathbb{E}[M_\infty] = \frac{1}{2}$. So we're going to converge to some random variable, which tells me the eventual fraction of black balls; and this random variable has expected value $\frac{1}{2}$. (This equality is because M_t is bounded — it's a fraction, so it's always between 0 and 1.)

Proof of claim. First, what are the transition probabilities for M_t? This is in fact also a Markov chain — where I am on the next step only depends on where I am right now, not on previous history. Here M_{t+1} can be one of two things. If I don't draw a black ball, then I have the same number of black balls, but instead of having $t+2$ balls, then I have $t+3$, so we get

$$\frac{t+2}{t+3} \cdot M_t.$$

And this happens with probability $1 - M_t$ (the probability I draw a white ball). Otherwise, with probability M_t I do draw a black ball; then I have another black ball, so it'll be

$$\frac{(t+2)M_t + 1}{t+3}$$

($(t+2)M_t$ is the number of black balls I had before, and now I have another). So

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

So those are the transitions. Now I need to check that it's a martingale — what's $\mathbb{E}[M_{t+1} - M_t \mid M_t]$? I'll subtract M_t everywhere, because it'll make the computations simpler; this is

$$\mathbb{E}[M_{t+1} - M_t \mid M_t] = \frac{1}{t+3} M_t \cdot (1 - M_t) + \frac{M_t + 1}{t+3} \cdot M_t = 0.$$

So this proves it's a martingale. □

Remark 19.12. This computation is a bit confusing because M_t plays two roles — the $(1 - M_t)$ and M_t factors are probabilities, and the $\frac{1}{t+3}M_t$ and $\frac{M_t+1}{t+3}$ factors are the differences in the fractions.

Question 19.13. What's the limiting distribution?

This looks like a very complicated process; so this is pretty hard to guess. But sometimes you should do what ChatGPT does and hallucinate.

Claim 19.14 — If you start with one black and one white ball, then $M_\infty \sim \mathsf{Unif}(0, 1)$.

So if you want to draw a completely uniform number between 0 and 1, one way to do it is to do what we just described — the fraction of black balls you get will be uniform. Polya discovered this well before anyone started talking about martingales.

This sounds weird, so let's see the proof (the proof is actually not very hard and kind of silly).

Proof. We'll actually show by induction that it's *always* uniform — we'll show that for all t, we have

$$\mathbb{P}\left[M_t = \frac{j}{t+2}\right] = \frac{1}{t+1} \quad \text{for each } j = 1, \dots, t+1$$

(we'll have $t + 2$ balls, and we always have at least 1 white and 1 black).

And you can do this by induction; we won't do it, because it's just a computation. □

Here's something you can't guess:

Question 19.15. What happens if I start from 2 black balls and 1 white ball?

Now it's not going to be uniform — the expected value is supposed to be $\frac{2}{3}$.

This leads to something called β-distributions. One reason people in Bayesian inference and ML like them so much is because they have this Polya interpretation — there the extra ball is an extra piece of evidence, and the fact it's such a nice process to compute with makes the Bayesian people very happy.

§19.3 Branching processes

Next, we'll talk a bit about *branching processes*. These are often called Galton–Watson processes, but Prof. Mossel forgot which of Galton and Watson invented them. There's a funny historical story about why they invented them, but here's one reason people care about them right now.

Goal 19.16. Understand the dynamics of biological processes.

The idea is that you start from an individual. This individual has a random number of children (let's say 2). Then each of these 2 individuals have a random number of children (let's say the first has 1, and the second has 3). Then each of these has a random number of children (let's say 2, 0, 2, 3). So I have this process where each individual gives birth to a random number of individuals, and they keep doing the same thing. We want to understand what happens to this process.

How do we formalize this?

Definition 19.17. Let Y be a nonnegative integer random variable (taking values 0, 1, 2, ...), and let Y_i^j be i.i.d. copies of Y.

Let $Z_0 = x$. Then we define $Z_{n+1} = \sum_{i=1}^{Z_n} Y_n^i$.

So $Z_0 = x$ is the number of individuals we start with — this is fixed, and usually 1. We think of Z_{n+1} as the number of individuals at level $n+1$. This is the complicated part. What's happening here is in our picture, $Z_0 = 1$ (we just had one person). Then what's Z_1? I have i running from 1 to $Z_0 = 1$, and just taking the number of children they have; they had 2 children, so now $Z_1 = 2$.

Now each individual here gives birth to a random number of children, and I sum those to get $Z_2 = 1+3 = 4$.

And we want to understand how this process behaves in the limit.

For this, it's useful to have the following.

Claim 19.18 — Let $\mu = \mathbb{E}[Y]$. Then Z_n/μ^n is a martingale.

So if I look at the total population size divided by what I expect it to be (I expect it to be μ^n, because each generation should be a factor of μ bigger than the previous one), this is going to be a martingale.

Question 19.19. Is this also a Markov chain?

The answer is yes — in order to know how big the population is in the next generation, you have a certain number of people and they're giving births to independent number of kids, so you just need to know that number.

So this is both a Markov chain and a martingale. Let's prove that it's a martingale:

Proof. Because it's a Markov chain, it's enough to condition on just the last step; we have

$$\mathbb{E}\left[\frac{Z_{n+1}}{\mu^{n+1}} \,\middle|\, Z_n\right] = \frac{1}{\mu^{n+1}}\mathbb{E}\left[\sum_{i=1}^{Z_n} Y_n^i \,\middle|\, Z_n\right].$$

And if I condition on Z_n, these are just Z_n i.i.d. random variables, so I get

$$\frac{\mu Z_n}{\mu^{n+1}} = \frac{Z_n}{\mu^n},$$

which is exactly what we wanted. □

Question 19.20. What will happen with this martingale, as a function of μ?

It's kind of difficult because this doesn't really correspond to our world (where typically 2 parents have a children, not 1 parent has a child — so if you want to think about it in our world, you have to look at one of the genders only). If each woman has an average of 1.5 daughters, what will happen to the world? Well, you'd expect it to explode — each generation will be a factor of 1.5 larger than the previous. If each woman has an average of 0.5 daughters, then the population will die (because it shrinks every step). What about if every woman has exactly 1 daughter in expectation? This is harder to think about.

Claim 19.21 (Subcritical) **—** If $\mu < 1$, then $\mathbb{P}[Z_n \to 0] = 1$.

So eventually the population will die.

Proof. You can do this with or without martingales.

For a proof without martingales, you can consider $\mathbb{P}[\sup_{m \geq n} Z_n > 0]$. Of course, this is just $\mathbb{P}[Z_n > 0]$ (since once I'm at 0, I stay at 0 forever — once I become extinct, I stay extinct). And we can write this as $\mathbb{P}[Z_n \geq 1]$. And now we can just use Markov — we have

$$\mathbb{P}[Z_n \geq 1] \leq \mathbb{E}[Z_n] = \mu^n$$

(by the fact that Z_n/μ^n is a martingale). This goes to 0, so we're done — the probability there'll be any time after n where I'm nonzero is at most μ^n, which means I have to converge to 0. □

Claim 19.22 — If $\mu = 1$, then it's still true that $\mathbb{P}[Z_n \to 0] = 1$, assuming that $\mathbb{P}[Y = 1] \neq 1$.

There's one case where $\mu = 1$ that's not very interesting — where everyone has exactly 1 child. Then I have one person in every generation, so nothing interesting happens. But if that's not the case — if there's some randomness in the number of children people have, and the expectation is 1 — then the population is going to die.

Proof. Here we'll use the fact that Z_n is a martingale, so we know that $Z_n \to Z_\infty$ for some Z_∞. And the question is, can we converge to something nonzero? Suppose $Z_\infty \neq 0$. Now, all the Z_n's are integers, so Z_∞ also has to be an integer. So there has to be some $k \in \mathbb{N}$ such that

$$\mathbb{P}[Z_n \to k] > 0.$$

Let's think about what that means about the world population. If $k = 10$, this means from some point on, there's always 10 people — there's 10 people, who give birth to 10 people, and so on. But this is where the fact there's some variability comes in — if I have 10 people, then the probability they have exactly 10 children is less than 1, and the probability that happens again is less than 1, and so on. So the probability this happens for infinitely many generations is 0. Writing this in words, this is impossible, since for every $k \neq 0$ and n, we have

$$\mathbb{P}[Z_n = Z_{n+1} = Z_{n+2} = \cdots = k] = 0.$$ □

This leaves us with the most interesting case, which is when $\mu > 1$. This we'll do next class.

Question 19.23. Any guesses on what happens when $\mu > 1$? Does it go extinct? Can it never go extinct?

Well, one situation where you might go extinct is if you have 0 children wth huge probability, and tons of children with low probability, you might just go extinct in the first generation. And one situation where you never go extinct is if you always have 1 or 3 children — then you're certainly never going extinct.

So as some examples to keep in mind:

Example 19.24

If $\mathbb{P}[Y = 1] = \mathbb{P}[Y = 2] = \frac{1}{2}$, then $\mu = \frac{3}{2}$, and there's never extinction.

Here's an example where μ is much bigger:

Example 19.25

If $\mathbb{P}[Y = 0] = 0.9$ and $\mathbb{P}[Y = 100] = 0.1$, then $\mu = 10$. But the probability of extinction is at least 0.9 (because it could be that just on the first generation, you get extinct). (Next class we'll talk about whether it's less than 1.)

The last homework will be posted below the weekend, and will be due a week and a half later.

§20 April 29, 2025

We'll continue where we left off last time, talking about branching processes.

§20.1 Review

To recall the setup, Y is an integer-valued random variable. We have many independent copies of it — i.e., i.i.d. $Y_i^j \sim Y$. Then we have a generation process — we start with $Z_0 = x$, and in general

$$Z_{n+1} = \sum_{i=1}^{Z_n} Y_n^i.$$

Think of the Z's as different generations. At generation 0 we have x individuals. And to go from generation n to $n+1$, if there's 0 people in generation n then there'll still be 0 people; otherwise each of the Z_n people in generation n give birth to Y_n^i individuals.

We saw that a very important parameter is $\mu = \mathbb{E}[Y]$, the expected number of children. Last time, we saw that:

Fact 20.1 — If $\mu \leq 1$, then $\mathbb{P}[Z_n \to 0] = 1$.

§20.2 The supercritical case

Today we're going to talk about what happens when $\mu > 1$. We already know that

$$W_n = \frac{Z_n}{\mu^n}$$

(the number of individuals divided by the expected number) is a martingale, and therefore $W_n \to W_\infty$. (We know that a nonnegative martingale always converges.) It could be that W_∞ is 0 — this abstract math tells us that a nonnegative martingale converges, but doesn't tell us whether it's 0 or not. So that's the main question we want to understand.

Question 20.2. Could it be that $Z_n \to 0$, or that $W_\infty = 0$, if $\mu > 1$?

This question in general is pretty hard. We'll solve it in an easier case — we'll assume that Y takes only finitely many values, i.e., that $\mathbb{P}[Y \leq K] = 1$ for some K — every individual has at most K children.

Remark 20.3. If the number of children can be unbounded, it's actually pretty tricky; the people who proved it are actually pretty modern (Prof. Mossel knows some of the people who proved it). But the bounded case is classical. If you're interested, this is called the *Kesten branching process theorem.*

How are we going to analyze this case? First we'll try to understand:

Question 20.4. What is $\lim_{n\to\infty} \mathbb{P}[Z_n = 0]$?

So we want to understand what's the limiting probability that the process goes extinct. For simplicity, we'll also assume that $Z_0 = 1$.

Notation 20.5. We define $p_k = \mathbb{P}[Y = k]$ as the probability an individual has k children, and we define a generating function

$$f(x) = \sum_{k=0}^{K} p_k x^k.$$

This is a polynomial that takes in a real number x (which we'll think of as a probability) and computes this value.

Notation 20.6. We write $q_n = \mathbb{P}[Z_n = 0]$.

So this is a sequence of numbers; and we want to analyze it.

Can we say anything about this sequence? It's a probability, so it's always between 0 and 1. But can we say anything else?

Claim 20.7 — The sequence q_n is monotone increasing.

Proof. If there's nobody left in the nth generation, then there's also no one left in the next generation — $Z_n = 0$ implies $Z_{n+1} = 0$. So the event $\{Z_{n+1} = 0\}$ contains $\{Z_n = 0\}$, which means $q_{n+1} \geq q_n$. □

So we have an increasing sequence of numbers, which means it has a limit $q_n \to \rho$.

Claim 20.8 — We have $q_n = f(q_{n-1})$.

So it's not just an increasing sequence of numbers — it has some nice algebraic properties (we're just putting a number into a function and applying it over and over again).

Proof. For this, it helps to draw a picture. I have my branching process; it starts from 1 individual, and then maybe they have 3 children. And I want to understand, after time n, what's the chance there's no offspring? Then each of these 3 children needs to have no offspring after $n-1$ steps, and these are independent.

So to understand $\mathbb{P}[Z_n = 0]$, we'll condition on the first generation — we have

$$\mathbb{P}[Z_n = 0] = \sum_{k=0}^{K} \mathbb{P}[Z_1 = k] \cdot \mathbb{P}[Z_n = 0 \mid Z_1 = k].$$

(We just break the probability according to what happens in the first generation.) Now let's write what this is — the chance of having exactly k children is just p_k. And then if I have k children, in order for the nth generation to have none, each of these k children has to go extinct in $n-1$ generations. So we'll have

$$\mathbb{P}[Z_n = 0] = \sum_{k=0}^{K} p_k q_{n-1}^k$$

(since all k children have to become distinct in $n-1$ generations). And this is just equal to $f(q_{n-1})$. □

So somehow we've transferred the probabilities to a question about calculus — we have an increasing sequence where the next element is a simple function of the previous one, so we should be able to analyze what happens.

Claim 20.9 — We have $q_n \to \rho$ where ρ is the smallest root of $f(x) = x$ in the interval $[0, 1]$.

So from what we have, we can already deduce what's the limit of this sequence.

Proof. Let's start from the easiest part: Can we show that ρ has to be a root of $f(x) = x$? Since the q_n are monotone increasing and live in $[0, 1]$, it has to have a limit; so $\rho = \lim_{n\to\infty} q_n$ exists. Now we can apply f to both sides of this equation; and we get that

$$f(\rho) = f\left(\lim_{n\to\infty} q_n\right).$$

And we can move f into the limit because f is a polynomial, so it's continuous; we get that

$$f(\rho) = \lim_{n\to\infty} f(q_n) = \lim_{n\to\infty} q_{n+1} = \rho.$$

So we've shown that $f(\rho) = \rho$ — whatever this probability is, it *is* a root of the equation $f(\rho) = \rho$.

But the claim was stronger — it says that it's the *smallest* root. So let ρ^* be the smallest root in $[0, 1]$. Then we know $0 = q_0 \le \rho^*$ (the probability of extinction at time 0 is just 0). Now we'll apply f inductively — f is a monotone function (it's a polynomial whose coefficients are all positive), so we'll get that $q_n \le \rho^*$ for all n. Then taking the limit gives that $\rho \le \rho^*$. But ρ is a root, and this is the smallest root, so they have to be equal. □

Where have we used the fact that $\mu > 1$? Nowhere — everything we wrote right now is correct even if $\mu \le 1$. What this should tell us is that if $\mu \le 1$, then the only root is 1. Note that $f(1)$ is always 1 — it's a sum of probabilities. So in the case $\mu \le 1$, it turns out there's only one root, and this root is 1.

So now somehow we need to understand the relationship between $\mu > 1$ and what that says about the smallest root of $f(x) = x$. So now this is really some sort of calculus question:

Question 20.10. How is μ related to ρ (the smallest root of $f(x) = x$)?

Remark 20.11. As a sanity check, if everyone has at least one child, the smallest root of $f(x) = x$ is 0 — f won't have any constant coefficient, so $f(0) = 0$. This is a good sanity check, because if everyone has at least one child, the process will certainly go forever.

First, let's forget about the relationship between μ and ρ, and answer a simpler question:

Question 20.12. What's the relationship between μ and f?

We have $f'(1) = \mu$, and that's going to help us.

So now we're going to summarize all the calculus properties we know about f:

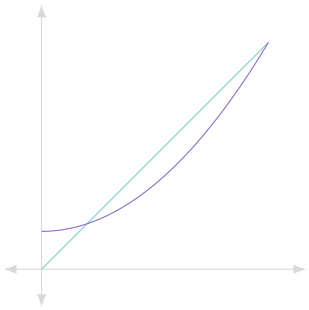
Claim 20.13 — We have that f is monotone increasing and convex, and $f'(1) = \mu$ and $f(1) = 1$.

Proof. The reason it's convex is that all its coefficients are positive, so if we take a second derivative, they're still positive. And we have

$$f'(1) = \sum_{k=1}^{K} k p_k = \mathbb{E}[Y].$$ □

Now let's try to draw this function; we want to understand where are the roots of $f(x) = x$, so we'll plot f and the line $y = x$.

First let's look at the picture when $\mu > 1$. Then the derivative at 1 is bigger than 1, so the function goes below $y = x$ at 1. And then the function is convex; and at 0, it's greater than (or equal to) 0. So it'll look like this:



So we get a single root which is less than 1.

As a sanity check, let's also draw a picture when $\mu < 1$. Then we'll have a similar picture, but it'll start out above $y = x$ (at 1) and stay there, so the only root is at 1.

What is the conclusion of all this calculus?

Proposition 20.14

If $\mu > 1$, then $\lim_{n\to\infty} \mathbb{P}[Z_n = 0] = \rho < 1$, where ρ is the smallest root of $f(x) = x$.

So we showed that in the case where the expected number of children is greater than 0, there's a positive probability this process will last forever. We know this probability is not 1 (if I have 0 or 4 children with probabilities $\frac{1}{2}$, then the probability the process dies in the first generation is already $\frac{1}{2}$); but the probability the process dies is also not 1.

This makes us somewhat happy. Here's a somewhat harder theorem that we'll also prove.

Theorem 20.15

If Y takes at most K values and $\mu > 1$, then the martingale $W_n = Z_n/\mu^n$ converges to a random variable W, and $\mathbb{P}[W = 0] = \rho$.

We know that the W_n's always converge, and they themselves become 0 with probability ρ. But the theorem isn't immediate — it could be that Z_n doesn't go to 0, but Z_n/μ^n does (since μ^n is exponentially increasing). So this is a stronger statement. It says that we know that with some probability it's going to be 0, but otherwise Z_n is going to be of order μ^n — it'll be exponentially large. Proposition 20.14 just says that with probability ρ it'll be 0 and otherwise it's nonzero; but Theorem 20.15 says that with probability ρ it'll be 0 and otherwise it'll be $\Omega(\mu^n)$.

Proof. We already know by martingale convergence that $W_n = Z_n/\mu^n$ converges to some random variable W; now we just want to understand the probability that W is 0. First note that

$$\mathbb{P}[W = 0] = \sum_{k=1}^{K} \mathbb{P}[Z_1 = k]\mathbb{P}[W = 0 \mid Z_1 = k]$$

(this is the same recursion we had before). And what does it mean to have $W = 0$? This means all the children have to have $W = 0$ as well. So the right-hand side is equal to $f(\mathbb{P}[W = 0])$.

What does this mean? It means $\mathbb{P}[W = 0]$ is either 1 or ρ — there are two roots of the equation $f(x) = x$, which are 1 and ρ. So it's enough to show that W is not always 0, i.e., that $\mathbb{P}[W = 0] \neq 1$. And there's some work in this — this is nontrivial — but that's the only thing we need to show.

How are we going to show this? We're going to take some threshold, which can be any number you want; we can write

$$\mathbb{P}[W \geq 0.1] \geq \lim_{n\to\infty} \mathbb{P}[W_n \geq 0.2]$$

(this is because W is the limit of all the W_n's, so if the W_n's are at least 0.2 then W is at least 0.1). So it's enough to show that the limit on the right is positive (then that'll tell us $W \geq 0.1$ with positive probability).

Now we don't have to deal with the limit, and we can just deal with the W_n's — it suffices to show that

$$\lim_{n\to\infty} \mathbb{P}[W_n \geq 0.2] > 0.$$

So that's our goal right now. And we're going to use something called the *second moment method*, or the *Payley–Zygmund inequality*; we'll say what that is and use it, and maybe prove it later.

Theorem 20.16 (Payley–Zygmund inequality)

If X is a random variable with $X \geq 0$, then for $0 \leq \eta \leq 1$ we have

$$\mathbb{P}[X \geq \eta \mathbb{E}[X]] \geq (1-\eta)^2 \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

This is some inequality about random variables; why is it relevant for us? We want to show $\mathbb{P}[W_n \geq 0.2]$ is something. And what we get from it — because W_n is a martingale, 0.2 is actually $0.2\mathbb{E}[W_n]$ (since $W_0 = 1$, so $\mathbb{E}[W_n] = 1$ for all n). So we get

$$\mathbb{P}[W_n \geq 0.2\mathbb{E}[W_n]] \geq 0.8^2 \cdot \frac{1}{\mathbb{E}[W_n^2]}$$

(for the numerator $\mathbb{E}[W_n]^2$, we just said that $\mathbb{E}[W_n] = 1$).

We want to show that the limit of this number is strictly greater than 0; so it's enough to show the limit of the right-hand side is strictly greater than 0. That means we need to show the number we're dividing by doesn't go to ∞.

Goal 20.17. Show that $\lim_{n\to\infty} \mathbb{E}[W_n^2] < \infty$.

And how are we going to do this? Here we're going to use the fact that martingales have orthogonal increments; recall that

$$\mathbb{E}[W_n^2] = \sum_{m=1}^{n} \mathbb{E}[(W_m - W_{m-1})^2].$$

Now we still have to compute the thing on the inside. First, we'll condition on Z_{m-1} — we have

$$\mathbb{E}[(W_n - W_{n-1})^2 \mid Z_{n-1}] = \mathbb{E}\left[\left(\frac{1}{\mu^n}\sum_{i=1}^{Z_n - 1}(Y_{n-1}^i - \mu)\right)^2 \,\middle|\, Z_n\right].$$

Why does this make us happy? We have a sum of independent random variables, so the expected square of the sum is the expected sum of squares; and we get

$$\frac{1}{\mu^{2n}} Z_{n-1}\sigma^2,$$

where $\sigma^2 = \mathrm{Var}(Y)$. And if we take the expected value of this, we get that

$$\mathbb{E}[(W_n - W_{n-1})^2] = \mathbb{E}\left[\frac{1}{\mu^{2n}} Z_{n-1}\sigma^2\right].$$

We have $\mathbb{E}[Z_{n-1}] = \mu^{n-1}$, so we get

$$\mathbb{E}[(W_n - W_{n-1})^2] = \sigma^2 \cdot \frac{1}{\mu^{n+2}}.$$

Why does this make us happy? Now we sum these guys and get

$$\mathbb{E}[W_n^2] \le \sigma^2 \sum \frac{1}{\mu^m} \le \sigma^2 \cdot \frac{1}{1-\mu} < \infty.$$

So we've got that all the variances are uniformly bounded, and that makes us happy.

To repeat the logic, the above calculation gives that $\lim_{n\to\infty} \mathbb{E}[W_n^2] < \infty$, which gives that $\lim_{n\to\infty} \mathbb{P}[W_n \ge 0.2] > 0$, which shows that $\mathbb{P}[W \ge 0.1] > 0$. □

Payley–Zygmund is not hard to prove (it's one line), but we won't do it right now, because we'll instead talk about something which is useful for the homework.

§20.3 The geometry of conditional expectations

We'll now talk about a L^2 (i.e., Euclidean distance) interpretation of conditional expectations and martingales. We'll start with a lemma about conditional expectations.

Lemma 20.18

The conditional expectation $\mathbb{E}[Y \mid X]$ is the function of X that minimizes $\mathbb{E}[(f(X) - Y)^2]$ among all functions f.

We know $\mathbb{E}[Y \mid X]$ is a function of X; and this says it's the one that minimizes this quadratic thing. Geometrically, this says conditional expectation is a *projection* — $\mathbb{E}[Y \mid X]$ is just the projection of Y onto all the functions of X. You'll see in the proof that it's all about projections — so there's some really clear geometric meaning of what it means to be $\mathbb{E}[Y \mid X]$, that you're the closest function of X to Y.

Proof. Let f be some function; we need to show that $\mathbb{E}[(Y - f(X))^2]$ is bigger than what we'd get if we replaced $f(X)$ with the conditional expectation.

What we'll do is a trick where we add and subtract the conditional expectation, and also introduce conditioning on X — so we'll rewrite the above expression as the much uglier expression

$$\mathbb{E}[(Y - f(X))^2] = \mathbb{E}[\mathbb{E}[((Y - \mathbb{E}[Y \mid X]) + (\mathbb{E}[Y \mid X] - f(X)))^2] \mid X].$$

Why would we do this? We're hoping the first thing is the minimal thing, so we're writing a minimal thing plus some error term, and hoping things work out.

When we expand things out, we'll have two terms we're happy about — the square of the first term is the thing we want, and the second term is the expectation of a square, so it's positive. And we all know that $(a+b)^2 = a^2 + b^2$, right? Well, no — we also have another term, which is the first annoying thing times the second. But why did we decide to condition on X? When we do that, which terms come out of the expectation? The second term is going to come out of the expectation, because when we condition on X, both $\mathbb{E}[Y \mid X]$ and $f(X)$ are functions of X, so we'll knwo them. So we're going to have

$$\mathbb{E}[Y - \mathbb{E}[X]](\mathbb{E}[Y \mid X] - f(X)).$$

But the first expectation is 0! So we're going to get that the cross-term disappears. And we get that this is at least

$$\mathbb{E}[(Y - \mathbb{E}[Y \mid X])^2]$$

(the cross-term is 0, and the square of the last term is nonnegative). □

We'll state another theorem, which we won't prove but is good to know (and is useful for our homework).

Theorem 20.19 (L^2 convergence of martingales)

Let $(M_t)_{t\geq 0}$ be a martingale such that $\sup_t \mathbb{E}[M_t^2] < \infty$. Then there is some M_∞ such that:

- $M_t \to M_\infty$.
- $\mathbb{E}[M_t] \to \mathbb{E}[M_\infty]$.
- $\mathbb{E}[(M_t - M_\infty)^2] \to 0$.

We want to think about these things as vectors in Euclidean space; the right way to say that these vectors are bounded is the $\sup_t \mathbb{E}[M_t^2] < \infty$ condition. And under this condition, we have stronger convergence than anything we've seen so far. So if (M_t) is a martingale — a very special sequence of vectors in $\mathbb{R}^n$ — which is bounded in L^2, then they converge to another vector, in all of these very strong senses.

One example where you can apply this is the example in the homework — note that this condition holds if the M_t's are bounded (i.e., $\mathbb{P}[a \leq M_t \leq b] = 1$ for all t).

§20.4 Martingales as successive best guesses

We'll now talk about another mini-mini-topic about martingales (we might not prove anything, but it's a useful way to think about them, and might again be useful for the homework); and that's probably the last thing we'll talk about regarding martingales. Next class we might talk about either paradoxes, or Brownian motion, or both.

Claim 20.20 — The sequence $\mathbb{E}[X]$, X, X, X, X, ... is a martingale (where X is a random variable).

Proof. From the second point on, this is clear (it's just the same information); and $\mathbb{E}[X \mid \mathbb{E}[X]] = \mathbb{E}[X]$. □

Fascinating, right? But what this says is that if you don't know anything about X, then your best guess about it is maybe $\mathbb{E}[X]$; but once you know it, that becomes your best guess.

Slightly more interesting:

Claim 20.21 — The sequence $\mathbb{E}[X]$, $\mathbb{E}[X \mid Y]$, X, X, X, ... is a martingale.

Here we don't know anything to start with, so our best guess is $\mathbb{E}[X]$; then maybe we find out Y, so we update to $\mathbb{E}[X \mid Y]$; and then we find out X itself, so that's our best guess. And you can check that $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$ and so on, so this works.

More generally:

Claim 20.22 — The sequence $\mathbb{E}[X]$, $\mathbb{E}[X \mid Y_1]$, $\mathbb{E}[X \mid Y_1, Y_2]$, ... is a martingale.

So suppose my goal is to understand if four years from January 1, the MIT math department will still exist given everything happening in the world. The first day I have my expectation; then there's a conditional expectation after what I learn after the first day of January, then the second, and so on. I collect more and more information, and update my probabilities accordingly.

Example 20.23

When Scott Sheffield teaches this, he has a beautiful example in 18.600 exemplifying this notion by a texting sequence between a couple. There's a number representing the probability that the couple stays together. Maybe the girl texts the boy 'Oh, I have to talk to you.' and maybe it goes down. And then it's 'I really love you.' and then it goes up. And then it's like 'I met this guy at work' and it goes on. And you see that the conditional probabilities of this couple eventually staying together goes up and down as we get more and more information; and that's a martingale.

§20.5 Paradoxes

In the last ten minutes, we'll talk a bit about paradoxes.

Example 20.24 (An infinite expectation paradox)

A corrupt banker dies. He suggests the following deal:

- The banker will spend n days in hell.
- At the end of the n days, we'll toss a biased coin. The banker will stay in hell with probability $\frac{1}{n}$, and otherwise will go to heaven.

So this is how the powers above decide to deal with the banker: We'll decide you suffer for some number of days (which you get to choose). Then we toss a coin; if you suffered a lot, you'll be more likely to go to heaven.

In the banker's utility, hell is $-\infty$ (staying in hell is really terrible, you really don't want that). And they like heaven; let's say it's 10^{20}. (It's not ∞, but it's pretty good.)

The banker's decision is very simple — they decide how long they're staying in hell, and then the thing happens. What will they do?

If they stay 2 days then it's $(\frac{1}{2}, \frac{1}{2})$; if they stay 3 then it's $(\frac{1}{3}, \frac{2}{3})$; and so on. So what will they do?

The difference is going to be $\frac{1}{n}$ vs. $\frac{1}{n+1}$, which is tiny; but we're multiplying that by $-\infty$. So the banker is going to stay in hell forever — they always want to increase the probability, so they'll wait for another day.

Example 20.25 (Two envelope paradox)

There are two envelopes; Prof. Mossel writes down some amounts of money on them (for example, one might be $1/\pi$ and the other might be $1/\pi + 1/10^9$).

We put one in one envelope and one in another. Devise a strategy such that no matter how much money there are in the envelopes, your expected gain is more than the average of the two envelopes.

You're allowed to look at one envelope; and you have to decide whether to accept it or take the other envelope.

There's a simple strategy — pick one at random, and then your expectation is the average. But we want to find a strategy that guarantees no matter what these numbers are (as long as they're not equal), you'll make more than the average of the two numbers.

Let's start from an easy case. Suppose that one has \$1 and the other has \$2. What will you do? You take one envelope, check that it's 1, and if it isn't you take it; if it is 1 then you take the other.

What if I promise you instead that the numbers are different, and each is an integer number in $\{1, \dots, 10\}$? Here's what you'd do — you guess a threshold randomly between 1 and 10 (e.g., 1.5, 2.5, ...), and put this

number in your head. You look at the envelope; and if the number is bigger than the threshold, you take it, otherwise you don't.

Now if your threshold is bigger than both numbers you make the average; if it's less then you still make the average; if it's in between, you make the bigger number.

In general, the strategy is to pick a random real number — maybe $x \sim \mathsf{Exp}(1)$. The point is we want the chance it lands between any two real numbers to be positive — no matter what two real numbers you have, the probability that x lies between them is positive.

And you reject if the envelope you see is less than x, and otherwise accept.

What's the logic? We have these two numbers; and my threshold can come in three different places. If my threshold is below both, I always accept, so I get the average. If it's above both, I always reject, so I also get the average. But if it's in the middle, then I always get the bigger number, and this is where I gain.

§21 May 1, 2025

Today we'll talk about Brownian motion; we'll start by talking about it more qualitatively than quantitatively.

§21.1 Discrete vs. continuous processes

Let's talk about some of the processes we've seen, and try to classify them as discrete or continuous — we'll make a table and think about both whether time is discrete or continuous, and whether space is. We started with discrete time Markov chains, and here it's sort of clear what the situation is — time was discrete (it's just 1, 2, ...), and so was space (we had finitely many states).

The next thing we looked at was continuous time Markov chains. There space was discrete. With time, it wasn't clear — time was continuous, but changes happened at discrete times. So time was continuous, but we had discrete jumps.

You can also look at random walks or certain martingales where time is still discrete, but space can be continuous or discrete — for example, we could have a random walk where each jump is a Gaussian.

But one question we can ask ourselves, as a preliminary question:

> **Question 21.1.** Is there a 'good' random process that is continuous both in time and in space?

This is a sort of philosophical question. When we do physics, everything (time and space) is continuous, so it's natural to ask if we can find such a process.

Brownian motion is going to be such a process. But can you come up with your own processes that are continuous in time and space, and random?

Some potential answers:

> **Example 21.2**
>
> You can just forget about randomness, and let the process be $f(t) = t$.

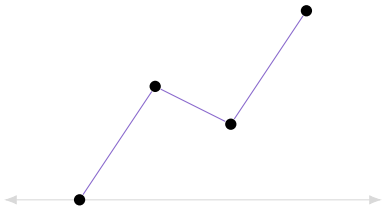
This is continuous in time and space, so you're happy; but there is no randomness here, so you're sort of cheating.

If you want to be more sophisticated...

Example 21.3

Take a random walk $S_n = \sum_{i=1}^n X_i$ where $X_i \sim \mathcal{N}(0,1)$. This is not going to be continuous in time, but we can interpolate.

So what this process looks like is I take a bunch of normal steps, and I connect these points by straight line.



Is this cheating or not? It's less cheating than before; but it's still sort of cheating, because in between the discrete times there's still no randomness. So it's deterministic at non-integer times, in the sense that if you know what happens at the integer times, you also know what happens at non-integer times. So this is unsatisfying (and it definitely doesn't look like the stock market or something like that).

§21.2 Definition of Brownian motion

Today, we're going to talk about Brownian motion. At a high level, it's a random process with continuous time and space. It's random at *all* time scales. It's a *Markov process* — we didn't talk about that, but it's the analog of Markov chains for things with continuous paths. And it's also a martingale.

So it really has lots of nice properties, and it appears in many places (maybe we'll talk a bit about applications).

Definition 21.4. We say that $(B(t) : t \geq 0)$ is a Brownian motion if the following properties hold:

- We have $B(0) = 0$.
- $B(t)$ has *independent increments* — if $0 \leq t_0 < t_1 < t_2 < \cdots$, then the random variables $B(t_1) - B(t_0)$, $B(t_2) - B(t_1)$, ... are independent.
- We have $B(t) - B(s) \sim \mathcal{N}(0, t-s)$ for all $t \geq s$.
- $\mathbb{P}[\text{the function } t \mapsto B(t) \text{ is continuous}] = 1$.

We can think about t in some interval, or $t \geq 0$, or t ranging over all of $\mathbb{R}$; we'll just use $t \geq 0$.

For the second condition, we're partitioning time into some finite number of intervals; and then we're looking at our process and looking at the difference between times 1 and 0, and times π and 1. And these differences are supposed to be independent. So what's happened so far (with respect to increments) doesn't tell me anything about what happens in the future.

The third property says these increments are not just independent, but *normal* (with mean 0, and variance the length of the interval).

The last probability is the farthest away from stuff we've seen so far. What we're defining is a random function; and we're saying this random function is continuous with probability 1.

You can imagine sampling from Brownian motion, and each time you get some continuous function (we plot a bunch of them on the board).

These all start from 0. And if I look at the continuous function, and look at various increments, we know the first increment is normal with variance proportional to the length of that interval, and so is the next increment, and they're independent.

§21.3 Some properties of Brownian motion

These are a lot of nice properties. We'll now say some other nice properties you can deduce from them.

Fact 21.5 — For fixed $s \geq 0$, the process $W(t) = B(t+s) - B(s)$ is also a Brownian motion.

So I look at the Brownian motion at time $t+s$, and subtract what I had at time s, that's also a BM.

We just have to check all the properties. We have $W(0) = B(s) - B(s) = 0$. The increments of W are the increments of B (because the $B(s)$ terms cancel); so if they were independent they're still independent, and if they were normal they're still normal. And we had a continuous function and we're subtracting a constant, so it's still continuous.

Fact 21.6 — If $s > 0$, then $B(st)/\sqrt{s}$ is also a Brownian motion.

This is sort of an interesting transformation — I scale time by s, and I scale space by $\sqrt{s}$; and then I still get a Brownian motion. So maybe if $s = 100$, time goes faster by a factor of 100 and I scale the y-axis by a factor of 10, and I get a Brownian motion.

We can check this in the same way. If I start with 0 and multiply things by 100 or divide by 10 I still get 0. The increments are just getting scaled, so if they were independent before then they're still independent. So the only thing we have to check is the normal distribution condition. And we have

$$\frac{B(st)}{\sqrt{s}} \sim \frac{1}{\sqrt{s}}\mathcal{N}(0, st).$$

But when I divide a random variable by a constant, its variance gets divided by the *square* of that constant. So this is

$$\mathcal{N}\left(0, st \cdot \frac{1}{\sqrt{s}^2}\right) = \mathcal{N}(0, t).$$

So this says shifting time still gives you a BM; scaling (accelerating time and space) also gives you a BM.

The next property is *inverting* — you can invert time.

Fact 21.7 — The process $W(t) = B(1-t) - B(1)$ for $0 \leq t \leq 1$ is also a Brownian motion.

So if I start Brownian motion at the end (shifting that to 0) and go back in time, I also get a Brownian motion.

Again, we can just check all the properties — we just checked 0, and the increments are the same increments, and if it was continuous before then it still is.

Another nice property, which is by definition:

Fact 21.8 — $B(t)$ is continuous with probability 1.

But the counterpart of this is... it's continuous, but what about higher-order derivatives?

Claim 21.9 — $B(t)$ is nowhere differentiable with probability 1.

So there's no point at which you can take a derivative of Brownian motion — this random process is continuous, but there's no point where it has a derivative.

Remark 21.10. In the early days of real analysis, people thought every continuous function has a derivative, except maybe at finitely many points. It took a while to construct counterexamples; but somehow the *canonical* example of a random function is nowhere differentiable.

Here's some intuition for why — we know $B(t+h) - B(t) \sim \mathcal{N}(0, h)$. And if you sample something from $\mathcal{N}(0, h)$, its typical magnitude will be the standard deviation — so it's typically of order $\sqrt{h}$. And then

$$\frac{B(t+h) - B(t)}{h} \approx \frac{\sqrt{h}}{h} = \frac{1}{\sqrt{h}},$$

which diverges. And this somehow is the reason it's nowhere differentiable — you look at closer and closer points, and show that enough of the times it's going to be high.

Remark 21.11. Usually when people talk about Brownian motion (we definitely won't get to this), the derivative has a name — we just said there's no derivative; but the derivative has a name, and it's called white noise. There's some weaker notions of derivatives for functions that don't have a derivative, and the derivative of this one is called white noise.

§21.4 Existence of Brownian motion

This sounds like a great process. But as mathematicians, when we define something, we want to prove it exists.

Question 21.12. We defined this beautiful process; but how do we know it exists?

We won't give the full proof, but we'll give a sketch.

The proof is actually going to use something along the ideas we saw in the earlier example, of interpolation.

What does it mean Brownian motion exists? It means we can define a function that satisfies all these properties. For laziness, we'll just do it for $t \in [0, 1]$.

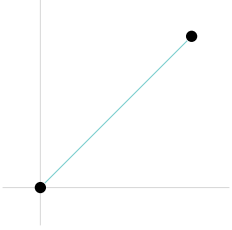
The idea is to use interpolation on smaller and smaller scales. How are we going to do this? Let

$$D_n = \left\{ \frac{m}{2^n} \mid m \text{ is an integer} \right\}.$$

We're first going to define our Brownian motion on D_n, in the following way.

First we'll define some functions. The first function we'll define is F_0, by $F_0(0) = 0$ and $F_0(1) = Z_1$, where $Z_1 \sim \mathcal{N}(0, 1)$. And then we're going to interpolate in between.

So we don't know how to construct Brownian motion, so we're starting just by making things okay at 0 and 1. At 0 we know we need to have 0, and at 1 we need $\mathcal{N}(0, 1)$, so we choose a value like that; and in between we just define it linearly.



What's the problem we had this process before? In the middle we don't have enough randomness; we want more randomness in the middle.

How are we going to generate more randomness in the middle?

First, as a side calculation, what's $\mathrm{Var}(F_0(1/2))$? This is some random variable (we're interpolating), and it's actually normal. What's its variance? Well, $F_0(1/2)$ is just $\frac{1}{2}Z_1$, and $Z_1 \sim \mathcal{N}(0,1)$, and the variance gets squared; so

$$\mathrm{Var}(F_0(1/2)) = \frac{1}{4}.$$

But what we *wanted* was that $\mathrm{Var}(B(1/2)) = \frac{1}{2}$. So how much variance are we missing to get from $\frac{1}{4}$ to $\frac{1}{2}$? We somehow need to get another $\frac{1}{4}$ of variance.

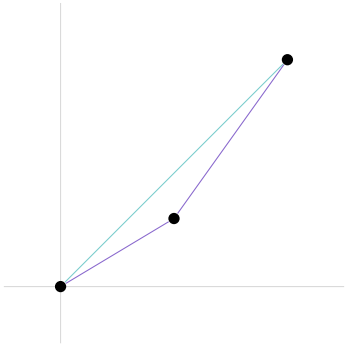
So what we're going to do is, just in the middle, we're going to add another normal with variance $\frac{1}{4}$.

So for the next step, we'll define F_1 by $F_1(0) = 0$ and $F_1(1) = 0$, and

$$F_1(1/2) = \frac{Z_{1/2}}{2}$$

where $Z_{1/2} \sim \mathcal{N}(0,1)$. So I'm correcting the function I had before. I don't touch 0 or 1 (which I was already happy about), and I'm correcting at $1/2$ by adding another random variable with the right variance; and I interpolate in between.

So now we can plot F_0 in blue, and $F_0 + F_1$ in purple:



And then the next points I'm going to look at are $1/4$ and $3/4$ — those guys will have variance too low, so I'm going to add another normal random variable at those points and interpolate; and I'm going to continue.

So more generally, we define

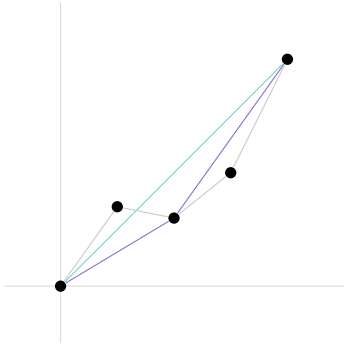
$$F_n(t) = 2^{-(n+1)/2} Z_t$$

for $t \in D_n \setminus D_{n-1}$ (this means t is an odd number divided by 2^n), where $Z_t \sim \mathcal{N}(0,1)$. And we define

$$F_n(t) = 0$$

for $t \in D_{n-1}$ (these are the ones we're already happy with, because they have the right variance already, so we don't want to touch them). And then we interpolate in between.

So maybe at the next level I fix $1/4$ (maybe I go up a bit) and $3/4$ (maybe I go down a bit).



So every time I make a small normal correction at the midpoints of my intervals; these normal corrections are getting smaller and smaller, and we can continue in this way.

We're not going to give the full proof, but here are some of the ideas.

The first thing we'll do is prove continuity. For that, we'll use the fact that:

- Each $F_n(t)$ is continuous (this is obvious — we just defined a finite number of points and interpolated between them).
- They're not too large. Specifically, with probability 1,

$$\sup |F_n(t)| \leq (1.1)^{-n}$$

for all sufficiently large n.

Student Question. *When we define these F's, is our Brownian motion the sum?*

Answer. Yes. There's a difference between the B and F — the B is the sum of the F. So we eventually define

$$B(t) = \sum_{n=0}^{\infty} F_n(t).$$

This is why we set $F_1(1) = 0$ and so on — once we fix some scale, we never touch it again.

So when we sum a bunch of continuous functions which are getting exponentially smaller, we're going to get that B_t is continuous.

Once it's continuous, that actually makes our life easier — because then we can check the claims about increments for dyadic numbers. There it's easy to check — we stop at some finite stage, so it's a bunch of normals, and we can check we get the right thing. And then we take a limit (for any two numbers, we look at dyadic numbers very close to them; and then the difference will be Gaussian corresponding to the difference of those dyadic numbers, which is going to converge to the difference of our actual numbers).

§21.5 History of Brownian motion

Why did people come up with Brownian motion? We'll talk about that, and then mention some of the math that comes from this motivation.

The original motivation was that this was a model of random fluctuations of a particle in fluids. What does this mean? In the 19th century there was a person named Brown, and he looked at a microscope and he had pollen in water, and he noticed it was moving in a crazy way — it's not moving in any particular direction, it's moving all the time; what is this thing? (This is why it's called Brownian motion.)

Similarly, Einstein sort of considered the same model for movement of molecules. He sort of derived Brownian motion — there's some differential equation describing where it is at some point. This gave some predictions. Then there was Perrin, who won the Nobel prize for confirming those predictions.

Then the actual definition we see here was done twice by Lévy (a very smart French mathematician who never wrote anything down — so he did it in the 1930s, but only his students knew about it) and then at MIT by Wiener (who defined it in similar ways to what we did right now).

There are many, many applications; we'll talk about two or three. The ones Prof. Mossel was planning to mention are to random walks, statistics, and maybe finance. And we'll see how much time we have.

§21.6 The invariance principle

We've seen the central limit theorem — if I toss a fair coin n times, I get some number S_n; and if we normalize it, we get

$$\frac{S_n - \mathbb{E}[S_n]}{\sigma[S_n]} \to \mathcal{N}(0,1).$$

But now, what if I'm more sophisticated and don't just want to look at *one* time n?

Question 21.13. What if I look at the graph of

$$\frac{S_n - \mathbb{E}[S_n]}{\sigma[S_n]}$$

for all n? What is this graph going to look like?

We don't want to write the thing where we subtract $\mathbb{E}[S_n]$ and divide by $\sigma(S_n)$ all the time, so we'll assume we have mean 0 and variance 1. This isn't really a restriction — if I'm interested in coin tosses, then I can subtract $\frac{1}{2}$ to get something with mean 0, and scale to get something with variance 1. So for coin tosses, we'd take $X_i = 2(\mathsf{Toss}_i - \frac{1}{2})$.

Theorem 21.14

Let $S_n = \sum_{i=1}^n X_i$ where X_i are i.i.d. with $\mathbb{E}[X_i] = 0$ and $\mathrm{Var}[X_i] = 1$. Define

$$W_n(t) = \frac{1}{\sqrt{n}} S_{nt}$$

(with linear interpolation). Then $W_n(t) \to B(t)$ for $t \in [0,1]$.

So S_{nt} for every $t \in [0,1]$ tells me how many wins I had at time t; and then I divide by $\sqrt{n}$. This is exactly the graph we described — how far away I am from the number of heads I expected to get (normalized by $\sqrt{n}$). And we wanted to know what this graph looked like.

And the theorem says that it's not just that if I look at this graph after n time I'll see a normal; if I look at this *whole* graph, I'll see a Brownian motion.

This is much stronger than CLT — there we're just looking at a specific n. But here I'm looking at the whole graph of how much I deviate at time 0.1 and 0.2 and 0.3 and $1/\pi$; I'm going to get a whole curve, and this curve is going to look like BM. The way you usually think about this in probability is that instead of just looking at one time, you're looking at the whole process.

§21.7 The Kolgomorov–Smirnov test

A very simple application of this is the following (it's one of the first appearances of Brownian motion in statistics).

The setup is the following: Suppose you have a distribution with CDF (cumulative distribution function) F. Now, I'm going to look at $X_1, X_2, \ldots$, which are i.i.d. samples from F. And then I'm going to define a new function, which is a random function — I'm going to define $F_n(x)$ as the fraction of $X_1, \ldots, X_n$ that are at most x.

So I have my data $X_1, \ldots, X_n$. And then I want to draw the CDF of this data. What is the CDF of this data? For every x, I'm just saying, what fraction of my samples are less than (or equal to) x? That's what summarizes my data.

You expect that F_n will eventually converge to F; in fact, the law of large numbers tells you that.

Question 21.15. How fast does it converge?

So we look at their difference — we define

$$D_n = \sup_x |F_n(x) - F(x)|.$$

Here F is the eventual distribution, and F_n is what I happened to see in my sample of the first n points; and D_n is the difference between the two.

Question 21.16. How big is this number, when n is very large?

Let's look at $x = 5$; what's the difference between the number of samples less than 5 inw hat I see vs. what I expect?

Well, if we fix a, what can we say about $F_n(a) - F(a)$? This is

$$\frac{\sum_{i=1}^n \mathbf{1}\{X_i \le a\} - F(a)}{n}.$$

These are mean-0 random variables (because the probability I'm at most a is $F(a)$). So by the law of large numbers this should go to 0; and by the CLT the fluctuations should be of order $1/\sqrt{n}$ (when we divide — before dividing, we'd expect $\sqrt{n}$).

So hopefully, the right quantity to look at is $\sqrt{n}D_n$.

Theorem 21.17

We have $\sqrt{n}D_n \to \sup_{0\le t\le 1} |B(t)|$.

It's the same logic we tried explaining for the random walk. If we looked at a specific a, this would just be CLT (it's a sum of random variables, so it should have a normal law). But now we're not just interested in one x, but all of them — is there any place we deviate a lot? And what this is telling you is the deviation behaves like a BM, so the maximum deviation should be like the maximum of BM.

How is this used in practice? One way it can be used is to *test* if the X_i's are drawn from F: You compute D_n an check if it's much larger than $\sup_{0\le t\le 1} |B(t)|$. This random variable has some distribution — maybe it's very unlikely to be bigger than 10. So if your data has it bigger than 10, then you say maybe it's coming from some other distribution.

Student Question. *Do we know the distribution of* $\sup |B(t)|$*?*

Answer. It doesn't have a very explicit form, but a lot is known about it, and it's not very crazy — for example, it won't be much bigger than a $\mathcal{N}(0,1)$. So for example, it's very very unlikely to be bigger than 20.

§21.8 The Black–Scholes model of stocks

This class is a bit eclectic with many applications, but we'll give another one.

Stocks are not Brownian motion. Black–Scholes also don't say it's BM. What happens with stocks is that it's not that the *amount* it goes up and down is random, but rather the *percent* is — it's multiplicative, rather than additive.

So the model is that the asset price $X(t)$ is given by

$$\log X(t) = \mu + \sigma^2 B(t).$$

So it's not that the price of the stock is a BM; it's the *log* of the stock price that's a BM. This is also good because stock prices are always positive (while BM can be positive or negative). But the point is that the random action is by how much it goes up by *percentage*.

So that's the assumption of the model. Prof. Mossel doesn't know if they invented it or other people considered it, but the main question they were trying to answer is:

> **Question 21.18.** How do you price an option?

About 3.5 people know what an option is, so Prof. Mossel will tell us; he'll give us the simplest example.

> **Example 21.19**
>
> A European call option with maturity date T and strike price K gives the holder the right to purchase a share for K dollars at time T.

(We have a capital T here, but this is not random; same with K.)

What does this mean? The closest Prof. Mossel got to purchasing an option was when he was our age, and his friends would start startups. What if I buy this coffee for you (that's how much I'm willing to pay). And in 5 years, you give me the option to buy 1% of your company for 10000 dollars. So I pay a little bit right now for the possibility to buy some portion of your company later. I'm not buying it right now.

All his friends refused; he lost a lot of money (some of them are billionares, so maybe he should've offered them more than one coffee).

So I'm paying money right now to have the possibility of buying something that might be worth a lot in the future.

A lot of this is happening with AI companies (secondary markets). For example, maybe I'm at Open AI and have a lot of equity but I can't really sell it, so someone says, why don't I buy your stock later for a lot of money, but I'll give you a little bit of money right now. So you reduce the risk for yourself working at Open AI — you get some money right now, and if the company does really well in the future, you've right now agreed to sell stuff at a price less than it'll be worth than.

The value of the option should be

$$\mathbb{E}[\max(0, X(T) - K)].$$

This is because $X(T) - K$ is how much money I'm going to make if I buy it for K. I'm only going to buy it if this is actually a good transaction for me, i.e., if this value is greater than 0; so that's why we have the max.

So that's the value of the option. And Black–Scholes computed the price of this, and many other options. All the calculations are based on using Brownian motion.

Next class is going to be a bit similar to what we did before the first midterm. Prof. Mossel will try to generate some practice midterms (so far ChatGPT is giving him very bad problems, but he has midterms from the last time he taught the class).

§22 May 13, 2025

Today Prof. Mossel will tell us about some other topics in probability related to stochastic processes that we didn't get to talk about.

§22.1 Random matrices

The first thing we'll talk about is random matrices. In the simplest case, we're still looking at i.i.d. random variables, but instead of drawing them in a line (writing x_1, x_2, ... horizontally and looking at the limit), we'll draw them in an *array* — so we'll have

$$M = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix}.$$

Of course, drawing a collection of random variables in a different way doesn't mean much, but we'll be asking different questions. In classical probability, you might ask:

Question 22.1. What can we say about $\sum_{i,j\in[n]} X_{ij}$?

Then you have the law of large numbers and the central limit theorem and all these things.

But questions for random matrices are very different:

Question 22.2. What can we say about the eigenvalues, or the singular values?

We write $\lambda(M)$ for the set of eigenvalues, and $\sigma(M)$ for the set of singular values.

Question 22.3. What is $\mathbb{P}[M \text{ is invertible}]$?

Like in all of probability, we want to answer these questions as $n \to \infty$.

This is a very active area in probability. Interestingly, it started from physics (like many good things in math) — the eigenvalues of M were proposed as a model for energy levels in atomic physics. And these days it's also used a lot in machine learning — it's a standard model of data in ML, and it's sort of related to problems like matrix completion, which is itself related to the Netflix challenge (try to predict which movies you will watch, given everything you've watched before), and so on.

You can spend a semester on this. This is also related to what's called *noncommutative probabilities* and von Neumann algebras and a bunch of stuff; if you want to dive deep into the field, there's a lot of interesting math.

That's the first model we wanted to mention, what happens when you mix probability and linear algebra.

§22.2 Connections to differential equations

Another topic we like as much as linear algebra is differential equations, so next we'll talk about some connections to differential equations.

There is a big area called SDEs, *stochastic differential equations*. Prof. Mossel isn't going to actually give the example, but here's a very simple example of a stochastic differential equation. In a standard differential equation, you might say

$$dX_t = U_t\, dt.$$

This differential equation we know how to solve — it means $X_t = \int U_t\,dt$.

But in stochastic differential equations, you can add another term — maybe you say

$$dX_t = U_t\,dt + V_t\,dB_t,$$

where B_t is a Brownian motion.

Let's see what intuitively this means. If we didn't have the second term, we're just saying how much X_t is moving at time t, well, it's moving by U_t. And now we're saying actually it's also moving *randomly*, according to some Brownian motion that goes up and down (and the amount it goes is given by V_t times how much the Brownian motion changes).

You can imagine that you have a molecule and there's a deterministic force U_t acting on it, but there are also atoms hitting it from all kinds of other directions, randomly. Or with a stock, maybe there's just more demand for items not from Country X; this is your U_t; but there's also some randomness coming, so maybe the price of the stock is impacted by both of these things.

Of course, for this to actually be a stochastic differential equation, maybe we take U_t to be X_t (or some function of X_t). But even in the simplest form, this is interesting.

Student Question. *Can U_t itself be random?*

Answer. Yes, U_t can be a function of X_t, and X_t is random, so U_t would be too.

Just so that if we ever hear the name we know what it means, there's a thing in the area called Itô's formula:

Theorem 22.4 (Itô's formula)

Suppose that $Y_t = g(X_t)$, where $g : \mathbb{R} \to \mathbb{R}$ is a 'nice' deterministic function. Then

$$dY_t = g'(X_t)\,dX_t + \frac{1}{2}g''(X_t)(dX_t)^2.$$

So it turns out that when X_t looks like the above thing, with some randomness, the chain rule becomes more complicated. The first term $g'(X_t)\,dX_t$ would be the usual chain rule. But because of the randomness, we also have this second term — somehow the randomness adds an extra variance term. And in general, whenever you look at SDEs, there's this second-order term that comes from the variance.

§22.3 Probability in high dimensions and geometry

The next thing we'll talk about is connections to high-dimensional geometry and convexity. We won't explain everything here, but it's sort of related to the picture we had with random matrices. As a motivating question:

Question 22.5. What can I say about a (random) vector $X_1^n = (X_1, \ldots, X_n)$ where X_i are i.i.d. from some distribution?

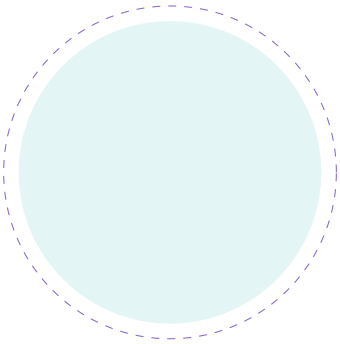
We already said we know a bunch of things — the average of the coordinates converges by LLN, and so on. But if you want to think about it as a geometric question, here's an example of a question you can ask:

Question 22.6. Suppose I have some subset $S \subseteq \mathbb{R}^n$, and I know $\mathbb{P}[X_1^n \in S] = \frac{1}{2}$. Let

$$S_\varepsilon = \{y \mid d(x, S) \le \varepsilon\}.$$

What can I say about $\mathbb{P}[X_1^n \in S_\varepsilon]$?

So S takes up half of the space according to our probability measure. And then I look at a slightly bigger set — I had my original set S, and I enlarged it by a little bit.



The answer is definitely bigger than $\frac{1}{2}$, because this is a bigger set. But how much bigger than $\frac{1}{2}$?

For example, it's known that if X_i are i.i.d. Gaussians, and d (the distance) is defined correctly, then we have a *concentration of measure* phenomenon, which says that

$$\mathbb{P}[X_1^n \in S_\varepsilon] \geq 1 - e^{-cn}.$$

(The distance d has to be normalized correctly based on the dimension; we won't go into the details of that.) The point is that if you increase a set by a little bit, and originally it took up half the space, after you enlarge by a little bit it takes up almost all of the space.

These days, something connecting SDEs and this, used a lot, is something called *stochastic localization* — which uses SDEs and Itô's formula to answer questions in high-dimensional geometry.

§22.4 Parametrized families of dependent random variables

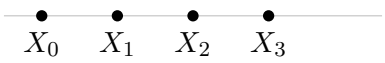
Now we'll talk a little bit about dependent random variables. We talked in this class about some dependent random variables — Markov chains and martingales. But many other dependent random variables are studied.

We've already seen in class, when we talked about metropolis, that in physics there are distributions like the Ising model, the Potts model, and a bunch of other models, which model physical phenomena — which looked like

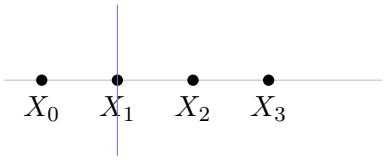
$$\mathbb{P}[(\sigma_1, \ldots, \sigma_n)] = \frac{1}{Z} \exp(H(\sigma))$$

(where $H(\sigma)$ is the 'energy' of σ). We saw some examples, and this is used a lot in physics. At the same time, like all good ideas in physics, people in machine learning took them; and this is called *graphical models* or *Markov random fields*. And a lot of people are studying these kinds of models.

We won't say too much about it, but maybe Prof. Mossel will give us one or two points to think about. One point is, why do we call it a Markov random field? In a Markov chain, if we imagine drawing the time axis, there's time 0 and X_0, then time 1 and X_1, and so on.

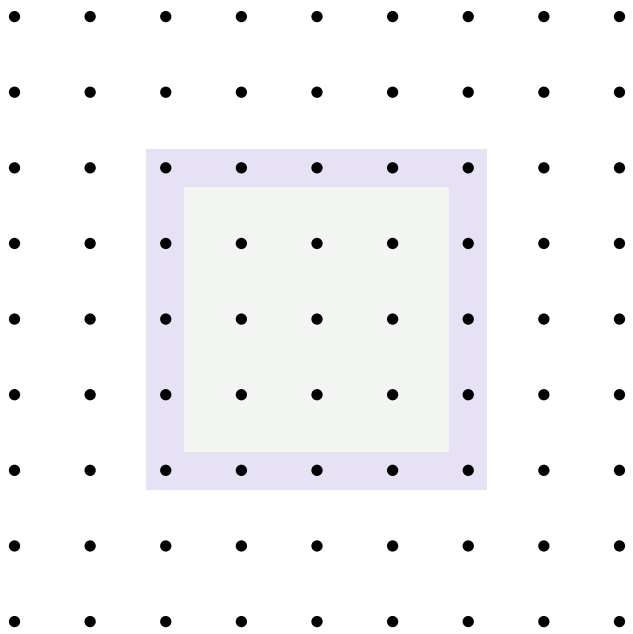


For Markov chains, the Markov property says that if I cut the line (e.g., at X_1), then when I condition on X_1, what happens on the left and right is independent. So for Markov chains, for every t, conditioned on X_t, we have that $(X_0, \ldots, X_{t-1})$ is independent of $(X_{t+1}, \ldots)$. So I condition on the middle guy and tell you the value was 7 here; and then what happens before and after are independent (because all the information is transmitted via this middle guy).



For Markov random fields, often there's a graph G. For the Ising model, maybe the graph is a 2-dimensional grid; and I have one random variable at each of the points.

And I can again condition on a bunch of these random variables. Maybe I condition on a red part which forms a box; and maybe the red part separates the yellow part (on the inside) from the white part (on the outside).



In a Markov random field, if red separates yellow from white, then conditioned on X_{red} (i.e., all X_i where i is red), then we have that X_{white} and X_{yellow} are independent (where X_{white} is the collection of all the random variables whose coordinate is white).

So we sort of generalized the notion of time — instead of time being 0, 1, 2, 3, ..., now time is a graph. And what that means is if we separate one set from another, then all the information is carried through the intermediate set — there's no additional information transmitted between yellow and white, beyond what's in red.

Also something very exciting that happened in the last 20–30 years is that in two-dimensional models, like Ising or Potts or percolation, there's some connections to complex analysis and SDEs, and this led to solutions of lots of big problems. Some names are 'quantum gravity' and 'SL evolutions' and so on. So something very interesting happens in two dimensions.

More from the engineering or machine learning side, related to Markov chains, the theory of Markov chains is fundamental for *Markov decision processes*. We're not going to define what this is, because there's a bunch of definitions and they're not the same. But one place it was developed a lot was robotics. You can think of a MDP as a process where what you do depends on the past, but you also have to make decisions, and those decisions might change what happens next. So it's a process where you don't just observe passively, but also take active decisions based on what you see. This is closely related to what's called *reinforcement learning*, which we probably heard in the context of LLMs and deep learning and so on (some kind of process where what you do changes what happens, and so on).

§22.5 Probability and algebra

Maybe two more quick connections: For people who like algebra, you can do random walks on groups — I can take a group and start walking with generators, and ask, what does the structure of the group tell me about the random walk, or vice versa?

Or you can take random generators for a group, and ask what happens.

So this is a very rich area. And there are other algebraic structures too.

§22.6 Ergodic theory

One more mathematical connection, or generalization of everything we've seen, is something called ergodic theory. We saw the ergodic theorem for Markov chains, which said that for a Markov chain that's ergodic, eventually it'll behave like the stationary distribution. But in fact, when ergodic theory was described, it's even more general.

Prof. Mossel will tell us a bit about ergodic theory. How did ergodic theory start? Well, we have a space A, and we have a measure μ — so A is a set and μ is a measure. And we have a map $T : A \to A$. People were thinking about physics — I have a fluid and I mix it and look at what happens after 1 second, and this is what T does (it mixes some stuff around). The key property we assume is that T is *measure-preserving*, which (for reasons we're not going to explain) means that:

Definition 22.7. We say T is measure-preserving if $\mu(T^{-1}(B)) = \mu(B)$.

This essentially says that for every set, after you move it, it has the same measure.

That's the basic setup for ergodic theory. And the main question they ask is:

Question 22.8. Suppose $f : A \to \mathbb{R}$. What can we say about $\frac{1}{n}\sum_{i=1}^{n} f(T^i x)$?

So f is something that measures something — e.g., it tells me, for every point in the system, what's the pressure or temperature or something. So I start with a point and look at its pressure, then I move it and ask what's its pressure, and then after I move it another step, and so on. And I look at the average of all these things.

Question 22.9. In what sense does this converge to $\mathbb{E}_\mu[f]$?

Ergodic theory finds conditions for this to hold. There's really a big theory of convergence for every x or almost all x or so on. Interestingly, even though this sounds like a big abstraction, this abstract setup is very useful and interesting, and there's lots of connections to other things (including number theory and so on).

§22.7 Mafia

Prof. Mossel will tell us a story about a stochastic process he analyzed, coming from the game Mafia.

When Prof. Mossel was young, he played Mafia, and at some point people asked, why don't you moderate a game of Mafia? And the question they wanted to ask was, usually when we play, maybe there's 9 players with 2 mafia and one policeman and so on. But he had to moderate a game with 25 people, and he didn't know how large the different groups should be. And it was also a very long game. So while moderating and being very bored, he was trying to come up with the math telling you how much mafia and policemen and so on you should have.

In the setup, we have N players and M mafia (we'll suppose there are no police). We'll make an assumption, which is maybe not completely standard — when someone dies, they are eliminated and their identity is revealed. So you see their card, and you see if they are mafia or a law-abiding citizen.

Here's the asymptotic question Prof. Mossel wanted to ask:

> **Question 22.10.** As $N \to \infty$, if everyone plays optimally, how large should M be for the game to be fair (i.e., mafia and citizen have equal chances of winning)?

So everyone has perfect poker faces, no one cheats at night and sees what happens at night, and so on. By fair, we mean the chance the mafia wins is about 50%, and the chance the citizens win is also about 50%.

Another thing is, it's a group game. If you're mafia, you don't care about yourself; you care that the last person alive is a mafia (and if you're a citizen, you care that the last person alive is a citizen).

So, how do you model this as a stochastic process?

For this, you also have to figure out what the mafia does at night, and what the citizens do during the day, and what the mafia do during the day.

First, what's the mafia's optimal play? At night they should kill a random non-mafia, and during the day they should behave just like citizens. Then the citizens don't learn anything about the mafia, so during the day they should just kill someone random. You have to prove this, but it turns out you get this stochastic process: So at night a random non-mafia is killed, and in the day a random person is killed.

Given this, as $N \to \infty$, how large should M be so that the chances are about 50%? This is a very simple stochastic process — there's two sets, and sometimes you kill a non-mafia, and sometimes you kill someone random.

The actual analysis is by martingales and so on; they wrote a paper about this. But before that, any guesses? Someone suggests that $\frac{1}{3}$ should be Mafia. The truth is actually a constant times $\sqrt{N}$. Why? If there's about $\sqrt{N}$ of them, it'll take about $\sqrt{N}$ time to kill a mafia, and in that time they kill $\sqrt{N}$ citizens.

Another way to think about it is that the square comes from the fact one group is killed in half the rounds, and the other in all.

So the result they got, which you can prove formally with martingales:

> **Theorem 22.11**
> Taking $M = \Theta(\sqrt{N})$ leads to a fair game.

Let's make it more interesting — let's add policemen. To make the rules clear, each policeman acts separately (on their own). What they do at night is, for each policeman, the mafia do their thing and then you say P1, open your eyes, point to someone. And then P1 points to someone, say Daniel. And the moderator nods if Daniel is a mafia, and shakes their head if not. And then we do the same with P2, and so on.

Suppose we have some policemen. So now the question is more complicated.

> **Question 22.12.** Suppose we have N people, M mafia, and P policemen.

This is the actual problem Prof. Mossel had to solve when moderating. What happens when you have some policemen — does this really change the picture, or not?

What's a good strategy for policemen, for those of you who've played the game? Well, what's a good strategy in a big game?

At the moment you claim to be a policeman, the mafia can also claim to be a policeman. So the right maneuver is to say I'm a policeman, X, Y, and Z are mafia, and maybe also give the identity of people who

are not. And then you say, in order to prove it, please kill me in this round, so that you know next round that what I said is true.

Actually what sophisticated players do is use it even when they're not a policeman. And people think, let's not waste a turn killing this person, and just believe them... But in a big game, this makes sense. So you query for a little bit, and then you reveal your identity by sacrificing yourself.

How much is a little bit — how much should you wait before you reveal yourself? And the policemen are not allowed to coordinate with one another, which is also kind of interesting.

You probably want to find $1/P$ of the possible mafias. Prof. Mossel doesn't know what the optimal strategy is for a fixed number of policemen, but let's try the following strategy:

Example 22.13

Suppose there are 10 policemen. Maybe we partition the circle into 4 parts. Each policeman picks one of the quarters, queries everyone in that quarter, and then comes out. So each of them tosses two coins; maybe the first decides to query Q1, 2 and 3 decide to query Q2, 4 and 6 Q3, and so on.

What do you need for this strategy to work?

First, what's the chance that Policeman 1 will be able to query everyone in this quarter before they're eliminated? The mafia eliminates people at random. The number of people I have to query is essentially a quarter of the days, so the chance I survive is essentially $\frac{3}{4}$.

So let's suppose these policemen all get lucky and manage to query everyone in their quarters. Then these guys can all come out, say kill me in the next few rounds, and then the citizens know all the mafia and the mafias know all the citizens. And then the thing is symmetric, so as long as we have more citizens than mafia as this stage, they win.

So if all the mafia get revealed, then all we need at this point is for

$$|\text{mafia}| < |\text{non-mafia}|.$$

If you do this analysis carefully (which we haven't), then the conclusion is the following:

Theorem 22.14

If $P \geq 2$, then for the game to be fair, we need $M = \Theta(N)$.

So it goes from $\sqrt{\bullet}$ scaling to linear scaling. This was very surprising. The title of the paper was at some point 'with 2 policemen you can save the world' or something.

There's actually an open problem in these papers — the case $P = 1$ is still open, and it's related to sme questions in cryptography. Why is crypto coming into the game? Because in this strategy, we assumed that people only get information about who's mafia. But if you have cryptography, maybe the policemen can secretly tell citizens about other citizens who are not mafia. If we have some cryptographic protocol where everyone is talking to everyone and no one knows who is talking to who, then the policeman can say listen, I know X, Y, Z are all good people without revealing themselves.

Prof. Mossel is going to teach this class again in the fall.