Math 154: Probability Theory, Lecture Notes

Kevin Yang

CONTENTS

1. Week 1, starting Tue. Jan. 23, 2024

1.1. Probability spaces and events.

Definition 1.1. Take a set Ω. A σ-algebra $\mathscr{F}$ is a collection of subsets of Ω such that

- $\Omega, \emptyset \in \mathscr{F}$.
- If $\{A_n\}_{n=1}^{\infty}$ is a collection of sets in $\mathscr{F}$, then $\cup_{n=1}^{\infty} A_n \in \mathscr{F}$ and $\cap_{n=1}^{\infty} A_n \in \mathscr{F}$.

Sets in $\mathscr{F}$ are called *events*. A probability measure $\mathbb{P}$ on $(\Omega, \mathscr{F})$ is a function $\mathbb{P} : \mathscr{F} \to [0,1]$ such that

- $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$
- If $\{A_n\}_{n=1}^{\infty}$ is a pairwise disjoint collection of sets in $\mathscr{F}$, then $\mathbb{P}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.
- If $\{E_n\}_{n=1}^{\infty}$ are in $\mathscr{F}$ and $E_1 \subseteq E_2 \subseteq \ldots$, then $\mathbb{P}(E_n) \to \mathbb{P}(\cup_{k=1}^{\infty} E_k)$.
- If $\{B_n\}_{n=1}^{\infty}$ are in $\mathscr{F}$ and $B_1 \supseteq B_1 \supseteq \ldots$, then $\mathbb{P}(B_n) \to \mathbb{P}(\cap_{n=1}^{\infty} B_n)$.
- The previous two bullet points are necessary parts of the definition. They must follow

The data $(\Omega, \mathscr{F}, \mathbb{P})$ is called a *probability space*.

Example 1.2. A coin is tossed. In this case, $\Omega = \{H, T\}$ (heads or tails). We can take $\mathscr{F} = 2^{\Omega}$. It contains $\{H, T\}$ (the coin lands heads or tails), $\{H\}$ (the coin lands heads), $\{T\}$ (the coin lands tails), and $\emptyset$ (the coin lands neither heads or tails). We have $\mathbb{P}(H) = 1 - \mathbb{P}(T)$, and $\mathbb{P}(\{H, T\}) = 1$ and $\mathbb{P}(\emptyset) = 0$. If it is a fair coin, then $\mathbb{P}(H), \mathbb{P}(T) = \frac{1}{2}$.

Example 1.3. A six-sided dice is thrown. $\Omega = \{1, 2, 3, 4, 5, 6\}$. We can take $\mathscr{F} = 2^{\Omega}$. In general, if Ω is finite, one should always take $\mathscr{F} = 2^{\Omega}$. If $X \in \mathscr{F}$ has size 1, then $\mathbb{P}(X) = \frac{1}{6}$. Then, use the additivity property to extend all of $\mathbb{P}$. (For example, $\mathbb{P}(\{1, 2\}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

Lemma 1.4. *(1) $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$, where $A^C = \Omega \setminus A$.*
(2) If $B \supseteq A$, then $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geqslant \mathbb{P}(A)$.
(3) If $A_1, \ldots, A_n \in \mathscr{F}$, then

$$\mathbb{P}\left(\cup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \ldots + (-1)^{n+1} \mathbb{P}(A_1 \cap \ldots \cap A_n).$$

For $n = 2$, this reduces to $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
(4) If $A_1, \ldots, A_n, \ldots \in \mathscr{F}$, then $\mathbb{P}(\cup_{n=1}^{\infty} A_n) \leqslant \sum_{n=1}^{\infty} \mathbb{P}(A_n)$. This is the union bound

Proof. Take the sequence $A_1 = A$ and $A_2 = A^C$ (and $A_n = \emptyset$ for all $n \geqslant 3$). We have $\mathbb{P}(A) + \mathbb{P}(A^C) = 1$, so point (1) follows. For point (2), write $B = A \cup (B \setminus A)$. Set $A_1 = A$, $A_2 = B \setminus A$, and $A_n = \emptyset$ for $n \geqslant 3$. Thus $\mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(A \cup (B \setminus A)) = \mathbb{P}(B)$, so point (2) follows. We will not prove point (3), since it is not really useful, but it's the same general principle as point (2). For point (4), we first define an auxiliary sequence $B_n = A_n \setminus \cup_{k=1}^{n-1} A_k$ and $B_1 = B$. Then B_n are pairwise disjoint. So $\mathbb{P}(\cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} \mathbb{P}(B_n)$. But $\cup_{n=1}^{\infty} B_n = \cup_{n=1}^{\infty} A_n$, and $B_n \subseteq A_n$, so $\mathbb{P}(B_n) \leqslant \mathbb{P}(A_n)$, and point (4) follows. □

Lemma 1.5. *Let $\{A_n\}_{n=1}^\infty$ be in $\mathcal{F}$. Then $(\cup_{n=1}^\infty A_n)^C = \cap_{n=1}^\infty A_n^C$ and $(\cap_{n=1}^\infty A_n)^C = \cup_{n=1}^\infty A_n^C$. One can take $A_n = \emptyset$ or $A_n = \Omega$ for all $n \geqslant N$ for some N to take finite unions and intersections.*

Proof. Take $x \in (\cup_{n=1}^\infty A_n)^C$. Thus, $x \notin A_n$ for any n. So $x \in A_n^C$ for all n, which means $x \in \cap_{n=1}^\infty A_n^C$. Now, take $x \in \cap_{n=1}^\infty A_n^C$, so $x \notin A_n$ for all n. This means $x \notin \cup_{n=1}^\infty A_n$, thus $x \in (\cup_{n=1}^\infty A_n)^C$. This shows $(\cup_{n=1}^\infty A_n)^C = \cap_{n=1}^\infty A_n^C$. The other claim follows by the same argument. □

Example 1.6. Let $A, B \in \mathcal{F}$. Suppose $\mathbb{P}(A) = \frac{3}{4}$ and $\mathbb{P}(B) = \frac{1}{3}$. We can bound $\mathbb{P}(A \cap B)$ as follows. First,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B).$$

We know $\mathbb{P}(A \cup B) \leqslant 1$, so $\mathbb{P}(A \cap B) \geqslant \frac{3}{4} + \frac{1}{3} - 1 = \frac{1}{12}$. Also, we know $\mathbb{P}(A \cup B) \geqslant \mathbb{P}(A)$, so $\mathbb{P}(A \cap B) \leqslant \frac{3}{4} + \frac{1}{3} - \frac{3}{4} = \frac{1}{3}$.

1.2. **Conditional probability.**

Definition 1.7. Take $B \in \mathcal{F}$ so that $\mathbb{P}(B) > 0$. The *conditional probability of A given B* is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

The idea is that one takes Ω, and restricts to a smaller probability space with set B. The σ-algebra is just given by taking $\mathcal{F}$ and intersecting with B (feel free to try to show that this is a σ-algebra). $\mathbb{P}(\cdot|B)$ is the "natural" probability measure on this probability space.

Example 1.8. Two fair dice are thrown. Condition on the first showing 3. What is the probability that the sum of the two rolls is > 6? Let A be the event where the sum of the two rolls is > 6 and B is the event where the first roll is a 3. We have

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A \cap B)}{\frac{1}{6}}.$$

Note that $A \cap B$ is the event where the second roll is $4, 5, 6$, and the first roll is a 3. In particular, there are 3 outcomes out of 36 that are okay, so the probability of $\mathbb{P}(A \cap B) = \frac{3}{36}$. This shows $\mathbb{P}(A|B) = \frac{1}{2}$.

Example 1.9. A coin is flipped twice independently. What is the probability that both are heads, given that one is a heads. It is not $\frac{1}{2}$. Indeed, let A be the event of two heads, and B is the event where one is a heads. There are four total outcomes, three of which have at least one heads. So $\mathbb{P}(B) = \frac{3}{4}$. On the other hand, $A \cap B$ is just the event of two heads, so its probability is $\frac{1}{4}$. This shows $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1}{3}$.

Lemma 1.10 (Law of total probability)**.** *We say that $B_1, \ldots, B_n \in \mathcal{F}$ form a partition of Ω if they are pairwise disjoint, positive probability, and $\cup_{i=1}^n B_i = \Omega$. For any partition $B_1, \ldots, B_n$ and any event A, we have*

$$\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

In particular, for any events A, B (where $B \neq \Omega, \emptyset$), we have $\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^C)\mathbb{P}(B^C)$.

Proof. Since $B_1, \ldots, B_n$ is a partition, the collection $A \cap B_1, \ldots, A \cap B_n$ are disjoint and $\cup_{k=1}^n A \cap B_k = A$. (To see this, note that clearly $A \cap B_k \subseteq A$, so it suffices to show that $A \subseteq \cup_{k=1}^n A \cap B_k$. Take $x \in A$. Then $x \in \Omega$, and since $B_1, \ldots, B_n$ is a partition, we know $x \in B_k$ for some k. Thus $x \in A \cap B_k$, and thus $x \in \cup_{k=1}^n A \cap B_k$.) From the first sentence, we get $\mathbb{P}(A) = \mathbb{P}(\cup_{k=1}^n A \cap B_{kj}) = \sum_{k=1}^n \mathbb{P}(A \cap B_k)$. By the definition of conditional probability, we have $\mathbb{P}(A \cap B_k) = \mathbb{P}(A|B_k)\mathbb{P}(B_k)$. Combining the previous two sentences finishes the proof. □

Theorem 1.11 (Bayes' formula)**.** *This will be helpful for the homework For any events A, B of positive probability, we have $\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$.*

Proof. It suffices to combine $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$ and $\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$. Indeed, this implies $\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$. Now, divide by $\mathbb{P}(B)$ on both sides (which one can do because B has positive probability!). □

1.3. **Independence.**

Definition 1.12. We say events A, B are *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Independent and disjoint are *totally different* notions! This is the same as $\mathbb{P}(A|B) = \mathbb{P}(A)$.

We say a family of events $\{A_i\}_{i=1}^\infty$ are *jointly independent* if $\mathbb{P}(\cap_{i=1}^n A_i) = \prod_{i=1}^n \mathbb{P}(A_i)$. We say it is *pairwise independent* if A_i, A_j are independent for all $i \neq j$.

Example 1.13. Let $\Omega = \{abc, acb, cab, cba, bca, bac, aaa, bbb, ccc\}$. Each element in Ω occurs with probability $\frac{1}{9}$. Let A_k be the event where the k-th letter (for $k = 1, 2, 3$) is a. We know that A_1, A_2, A_3 are pairwise independent. Indeed, $A_1 \cap A_2$ is the event where the first and second letter are both a. Thus, $A_1 \cap A_2 = \{aaa\}$, so $\mathbb{P}(A_1 \cap A_2) = \frac{1}{9}$. Note that $\mathbb{P}(A_1)\mathbb{P}(A_2) = \frac{1}{3}\frac{1}{3} = \frac{1}{9}$. Similar arguments apply to A_1, A_3 and A_2, A_3 (try it!).

But, A_1, A_2, A_3 are not jointly independent. Indeed, $A_1 \cap A_2 \cap A_3 = \{aaa\}$, so its probability is $\frac{1}{9}$. But $\mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3) = \frac{1}{3}\frac{1}{3}\frac{1}{3} = \frac{1}{27}$.

Example 1.14. We pick a card uniformly at random from a deck of 52. Each has probability $\frac{1}{52}$. Let A be the event where a king is picked, and B is the event where a spade is picked. Then $\mathbb{P}(A) = \frac{4}{52} = \frac{1}{13}$, and $\mathbb{P}(B) = \frac{1}{4}$. Also, $\mathbb{P}(A \cap B) = \frac{1}{52}$. So A, B are independent.

Lemma 1.15. *If A, B are independent, then A^C, B are independent and A^C, B^C are independent.*

Proof. We claim $\mathbb{P}(A^C \cap B) + \mathbb{P}(A \cap B) = \mathbb{P}(B)$. (This follows because $A^C \cap B$ and $A \cap B$ are disjoint and union to B.) Since A, B are independent, this implies $\mathbb{P}(A^C \cap B) = \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) = (1 - \mathbb{P}(A))\mathbb{P}(B)$. But $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$, so we get $\mathbb{P}(A^C \cap B) = \mathbb{P}(A^C)\mathbb{P}(B)$, which means A^C, B are independent. To show that A^C, B^C are independent, use the first result (but replace A by B and B by A^C). □

Example 1.16. Two fair dice are rolled independently. Let A be the event where the sum of the rolls is 7. Let B be the event where the first roll is 1. Then A, B are independent.

Indeed, $\mathbb{P}(A|B) = \frac{1}{6}$ (since a six is needed on the second roll). But $\mathbb{P}(A) = \frac{6}{36}$, since for any value of the first roll, there is exactly one value of the second roll to realize A. If we change 7 to 1, then A, B are no longer independent.

Definition 1.17. Fix an event B with positive probability. We say that A_1, A_2 are *conditionally independent* (given/conditioning on B) if $\mathbb{P}(A_1 \cap A_2|B) = \mathbb{P}(A_1|B)\mathbb{P}(A_2|B)$.

Lemma 1.18. *Fix B. Then A_1, A_2 are conditionally independent given B if and only if $\mathbb{P}(A_1|A_2, B) = \mathbb{P}(A_1|B)$.*

Proof. Suppose conditional independence of A_1, A_2. Then

$$\begin{aligned}\mathbb{P}(A_1|A_2, B) &= \frac{\mathbb{P}(A_1 \cap A_2 \cap B)}{\mathbb{P}(A_2 \cap B)} \\ &= \frac{\mathbb{P}(A_1 \cap A_2|B)\mathbb{P}(B)}{\mathbb{P}(A_2|B)\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(A_1|B)\mathbb{P}(A_2|B)\mathbb{P}(B)}{\mathbb{P}(A_2|B)\mathbb{P}(B)} \\ &= \mathbb{P}(A_1|B).\end{aligned}$$

Now suppose that $\mathbb{P}(A_1|A_2, B) = \mathbb{P}(A_1|B)$. Then

$$\begin{aligned}\mathbb{P}(A_1 \cap A_2|B) &= \frac{\mathbb{P}(A_1 \cap A_2 \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(A_1|A_2, B)\mathbb{P}(A_2 \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(A_1|B)\mathbb{P}(A_2|B)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A_1|B)\mathbb{P}(A_2|B).\end{aligned}$$

This finishes the proof. □

Example 1.19. Suppose I have two coins. One is fair, and the other one has probability of heads equal to $\frac{1}{3}$. I choose one of the two coins uniformly at random, and I toss it twice (independently). Let X be the value of the first flip and Y be the value of the second flip. Then X and Y are conditionally independent given that I choose the fair coin. (Same is true if I condition on choosing the non-fair coin.)

1.4. **Some examples.**

(1) (Symmetric random walk, "gambler's ruin") Let's play a game. We flip a coin repeatedly. If it lands heads, I get one dollar. If it lands tails, I lose a dollar. (Suppose this is a fair coin for now.) I want to save N dollars, at which point I stop the game, so that I can retire happily. But if I end up with zero dollars at any point, we stop the game, since I can't play anymore.

Suppose I start with $0 < k < N$ dollars. What is the probability that I win?

- Let $p_k = \mathbb{P}_k(A)$ be the event that I win if we start at k dollars. By the law of total probability, if B is the event that we toss a heads, then

$$\mathbb{P}_k(A) = \mathbb{P}_k(A|B)\mathbb{P}(B) + \mathbb{P}_k(A|B^C)\mathbb{P}(B^C).$$

We have $\mathbb{P}_k(A|B) = p_{k+1}$ and $\mathbb{P}_k(A|B^C) = p_{k-1}$ and $\mathbb{P}(B), \mathbb{P}(B^C) = \frac{1}{2}$. So $p_k = \frac{1}{2}(p_{k+1} + p_{k-1})$. But also $p_0 = 0$ and $p_N = 1$. We will talk later in this class about how to solve this equation efficiently, but one can check that $p_k = 1 - \frac{k}{N}$ solves this equation.

(2) (Testimonies) We are in court over whether or not Kevin stole the piece of chalk. We have two witnesses Alf and Bob. Alf tells the truth with probability α and Bob commits perjury with probability β. There is no collusion between these two (as in whether Kevin did it or not, their testimonies are independent). Let A be the event where Alf says Kevin stole it, and B be the event where Bob says Kevin stole it. Let T be the event where Kevin stole it. What is probability that Kevin stole it given that Alf and Bob said so, in terms of $\tau = \mathbb{P}(T)$?

- We need to compute $\mathbb{P}(T|A \cap B)$. By Bayes' rule, we have

$$\mathbb{P}(T|A \cap B) = \frac{\mathbb{P}(A \cap B|T)\mathbb{P}(T)}{\mathbb{P}(A \cap B)}.$$

We have $\mathbb{P}(A \cap B|T) = \mathbb{P}(A|T)\mathbb{P}(B|T) = \alpha\beta$, so the numerator is $\alpha\beta\tau$. For the bottom, by the law of total probability, we have

$$\begin{aligned}\mathbb{P}(A \cap B) &= \mathbb{P}(A \cap B|T)\mathbb{P}(T) + \mathbb{P}(A \cap B|T^C)\mathbb{P}(T) \\ &= \alpha\beta\tau + (1-\alpha)(1-\beta)(1-\tau).\end{aligned}$$

So, $\mathbb{P}(T|A \cap B) = \frac{\alpha\beta\tau}{\alpha\beta\tau + (1-\alpha)(1-\beta)(1-\tau)}$.

(3) (Simpson's paradox)

2. Week 2, starting Tue. Jan. 30, 2024

2.1. Random variables.

Definition 2.1. A *random variable* is a function $X : \Omega \to \mathbb{R}$ such that for any $x \in \mathbb{R}$, the event $\{X \leqslant x\}$ is in $\mathscr{F}$. The function $F(x) := \mathbb{P}(X \leqslant x)$ is the *distribution function* associated to X.

We say X is *discrete* if it only takes values in a countable set $\{x_1, \ldots, x_n, \ldots\}$ of $\mathbb{R}$. We say X is *continuous* if its distribution function can be represented as

$$F(x) = \mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f(u)du,$$

where $f : \mathbb{R} \to [0, \infty)$ is called the *probability density function* (it needs to be integrable, i.e. $\int_{\mathbb{R}} f(u)du < \infty$).

It is a fact that if X, Y are random variables and $a, b \in \mathbb{R}$, then $aX + bY$ is a random variable!

Lemma 2.2. *A distribution function F satisfies*

(1) If $x \leqslant y$, then $F(x) \leqslant F(y)$ (even if $x < y$, we can still have $F(x) = F(y)$!)
(2) $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$.
(3) $F(x + h) \to F(x)$ as $h \to 0$ from above.

Proof. (1) If $x \leqslant y$, then $\{X \leqslant x\} \subseteq \{X \leqslant y\}$.

(2) Let $A_n := \{X \leqslant -a_n\}$, where $a_n \to \infty$ is strictly increasing. Then $F(a_n) = \mathbb{P}(A_n)$. But $A_n \supseteq A_m$ for all $m \geqslant n$. So $F(a_n) = \mathbb{P}(A_n) \to \mathbb{P}(\cap_{m=1}^{\infty} A_m) = \mathbb{P}(\emptyset) = 0$.

Let $B_n := \{X \geqslant b_n\}$, where $b_n \to \infty$ is strictly increasing. Note that $B_n \subseteq B_m$ if $m \geqslant n$. Then $F(b_n) = \mathbb{P}(B_n) = \mathbb{P}(\cup_{m=1}^{\infty} B_m) = \mathbb{P}(\Omega) = 1$.

(3) Let $A_n = \{X \leqslant x + h_n\}$, where $h_n \to 0$ is strictly decreasing. Then $\cap_{n=1}^{\infty} A_n = \{X \leqslant x\}$, and $A_n \supseteq A_m$ if $m \geqslant n$. So $F(x+h_n) = \mathbb{P}(A_n) \to \mathbb{P}(\cap_{n=1}^{\infty} A_n) = \mathbb{P}(X \leqslant x) = F(x)$.

□

Definition 2.3. Suppose X is a discrete random variable. Its *probability mass function* (or *pmf*) is the function $f : \mathbb{R} \to [0, 1]$ such that $f(x) = \mathbb{P}(X = x)$. This is generally much easier to compute than the distribution function!

Example 2.4 (Bernoulli distribution)**.** Any random variable which is valued in $\{0, 1\}$. For example, the outcome of flipping a coin is Bernoulli, if we interpret heads as 1 and tails as 0. If the probability of heads is p, then its pmf if $p(1) = p$ and $p(0) = 1 - p$ and $p(x) = 0$ for $x \neq 0, 1$. The distribution function is $F(x) = 0$ for all $x < 0$, and $F(x) = 1 - p$ for all $x \in [0, 1)$, and $F(x) = 1$ for all $x \geqslant 1$.

For shorthand, we write $X \sim \text{Bern}(p)$.

Example 2.5 (Binomial distribution)**.** Let $X_1, \ldots, X_n$ be *independent* Bernoulli random variables. Set $Y = X_1 + \ldots + X_n$. This is a *binomial* random variable. It is discrete, since it takes values in $\{0, 1, \ldots, n\}$. Its probability mass function satisfies $p(x) = 0$ if $x \notin \{0, 1, \ldots, n\}$. For any $k \in \{0, 1, \ldots, n\}$, $p(k)$ is the probability of flipping exactly k heads. There are $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ ways to choose k out of n flips to be heads. The probability of flipping this particular sequence of heads and tails is $p^k(1-p)^{n-k}$. So $p(k) = \binom{n}{k} p^k (1-p)^{n-k}$.

For shorthand, we write $X \sim \text{Bin}(n, p)$.

Example 2.6 (Poisson distribution)**.** X takes values in the set $\{0, 1, 2, \ldots\}$. Its pmf is defined to be

$$\mathbb{P}(X = k) = p_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Here, $\lambda > 0$ is a fixed parameter. Note that

$$\sum_{k=0}^{\infty} p_\lambda(k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{\lambda} e^{-\lambda} = 1,$$

so $p_\lambda(\cdot)$ is indeed a probability mass function. For shorthand, we write $X \sim \text{Pois}(\lambda)$.

Example 2.7 (Geometric distribution)**.** Flip a coin repeatedly with probability of heads being p. Let X be the first time that the coin turns up heads. This takes values in $\{1, \ldots\}$. Its pmf is $\mathbb{P}(X = k) = p(k) = (1-p)^{k-1} p$. This is called the geometric distribution, since

$$\sum_{k=1}^{\infty} p(k) = p \sum_{k=0}^{\infty} (1-p)^k = p \frac{1}{1 - (1-p)} = 1$$

is a geometric series.

Definition 2.8. A *random vector* of dimension n is a vector $\mathbf{X} = (X_1, \ldots, X_n)$ such that $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ are random variables. If $X_1, \ldots, X_n$ are discrete random variables, then the pmf of $\mathbf{X}$ is defined to be the function

$$p(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n).$$

2.2. **Independence of random variables.**

Definition 2.9. A collection of random variables $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ (i.e. on the same probability space) are *jointly independent* if for open or closed subsets $A_1, \ldots, A_n \subseteq \mathbb{R}$, we have

$$\mathbb{P}\left(\cap_{i=1}^n \{X_i \in A_i\}\right) = \prod_{i=1}^n \mathbb{P}(X_i \in A_i).$$

We say they are pairwise independent if X_i, X_j are independent for all $i \neq j$.

Lemma 2.10. *Let $X_1, \ldots, X_n$ be independent discrete random variables with pmfs $p_1, \ldots, p_n$. Then $X_1, \ldots, X_n$ are jointly independent if and only if for any $x_1, \ldots, x_n \in \mathbb{R}$, we have*

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^n p_i(x_i).$$

Proof. If $X_1, \ldots, X_n$ are jointly independent, just take the formula for joint independence above and set $A_i = \{x_i\}$ for all i. For the other direction, we have

$$\begin{aligned}
\mathbb{P}\left(\cap_{i=1}^n \{X_i \in A_i\}\right) &= \sum_{x_1 \in A_1, \ldots, x_n \in A_n} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \\
&= \sum_{x_1 \in A_1} \cdots \sum_{x_n \in A_n} p_1(x_1) \ldots p_n(x_n) \\
&= \sum_{x_1 \in A_1} p_1(x_1) \ldots \sum_{x_n \in A_n} p_n(x_n) \\
&= \mathbb{P}(X_1 \in A_1) \ldots \mathbb{P}(X_n \in A_n).
\end{aligned}$$

□

Example 2.11. A coin flips heads with probability p and tails with probability $1 - p$. Let X be the number of heads and Y be the number of tails. These are *not* independent. (As for the details why, $\mathbb{P}(\{X = 1\} \cap \{Y = 1\}) = 0$ but $\mathbb{P}(X = 1)\mathbb{P}(Y = 1) = p(1 - p)$.)

Suppose that N is a Poisson random variable of parameter λ (it is independent of the coin). Then X and Y are independent! Indeed,

$$\begin{aligned}
\mathbb{P}(X = x, Y = y) &= \mathbb{P}(X = x, Y = y | N = x + y)\mathbb{P}(N = x + y) \\
&= \binom{x+y}{x} p^x (1-p)^y \frac{\lambda^{x+y}}{(x+y)!} e^{-\lambda} \\
&= \frac{(\lambda p)^x (\lambda(1-p))^y}{x!y!} e^{-\lambda}.
\end{aligned}$$

Since this has the form of $f(x)f(y)$, this means independence. To see this exactly,

$$\begin{aligned}\mathbb{P}(X=x) &= \sum_{y=0}^{\infty} \mathbb{P}(X=x, Y=y) \\ &= \frac{(\lambda p)^x}{x!} e^{-\lambda p} \sum_{y=0} \frac{(\lambda(1-p))^y}{y!} e^{-\lambda(1-p)} \\ &= \frac{(\lambda p)^x}{x!} e^{-\lambda p}.\end{aligned}$$

In particular, the number of heads and the number of tails are Poisson random variables of parameters λp and $\lambda(1-p)$, and they are independent of each other!

Lemma 2.12 (Convolution formula)**.** *Suppose X, Y are independent discrete random variables that take values in $\mathbb{Z}$. Let p_X and p_Y be their pmfs. Then $Z = X + Y$ takes values in $\mathbb{Z}$, and its pmf is*

$$p_Z(z) = \sum_{k \in \mathbb{Z}} p_X(z-k) p_Y(k).$$

Proof. For any $z \in \mathbb{Z}$, the event $\{Z = z\}$ is equal to $\cup_{k\in\mathbb{Z}} \{X = z-k\} \cap \{Y = k\}$. These events in the union are disjoint, since X, Y cannot obtain two values simultaneously. So, by independence, we have

$$\begin{aligned}\mathbb{P}(Z=z) &= \mathbb{P}(\cup_{k\in\mathbb{Z}} \{X = z-k\} \cap \{Y=k\}) = \sum_{k\in\mathbb{Z}} \mathbb{P}(X = z-k, Y=k) \\ &= \sum_{k\in\mathbb{Z}} \mathbb{P}(X=z-k)\mathbb{P}(Y=k).\end{aligned}$$

□

Example 2.13. Take X_1, X_2 independent Bernoullis of parameter p (so $\mathbb{P}(X_1 = 1), \mathbb{P}(X_2 = 1) = p$). Let $Z = X_1 + X_2$. By the convolution formula and the fact that X_1, X_2 cannot attain values other than 0 and 1, we have

$$\begin{aligned}\mathbb{P}(Z=0) &= \sum_{k\in\mathbb{Z}} \mathbb{P}(X=-k)\mathbb{P}(Y=k) = \mathbb{P}(X=0)\mathbb{P}(Y=0) = (1-p)^2, \\ \mathbb{P}(Z=1) &= \sum_{k\in\mathbb{Z}} \mathbb{P}(X=1-k)\mathbb{P}(Y=k) \\ &= \mathbb{P}(X=1)\mathbb{P}(Y=0) + \mathbb{P}(X=0)\mathbb{P}(Y=1) = 2p(1-p), \\ \mathbb{P}(Z=2) &= \sum_{k\in\mathbb{Z}} \mathbb{P}(X=2-k)\mathbb{P}(Y=k) = \mathbb{P}(X=1)\mathbb{P}(Y=1) = p^2.\end{aligned}$$

In particular, $Z \sim \mathrm{Bin}(2, p)$!

Lemma 2.14. *If X, Y are independent, then so are $f(X)$ and $g(Y)$ (for any functions f, g).*

2.3. **Expectation.**

Definition 2.15. Let X be a discrete random variable with pmf p. Its *expectation* is $\mathbb{E}(X) = \sum_{x:p(x)>0} xp(x)$.

Lemma 2.16. *(1) If $X \geqslant 0$ with probability 1, then $\mathbb{E}(X) \geqslant 0$. Thus, if X, Y satisfy $X \leqslant Y$, then $\mathbb{E}(X) \leqslant \mathbb{E}(Y)$.*
*(2) If $a, b \in \mathbb{R}$, then $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ (*linearity of expectation*; note that X, Y do not have to be independent!).*
(3) If $X = c$ with probability 1, then $\mathbb{E}(X) = c$.

Proof. (1) We have $\mathbb{E}(X) = \sum_{x:p(x)>0} xp(x)$. Since $p(x) > 0$ only if $x \geqslant 0$ by assumption, we know $xp(x) \geqslant 0$, so $\mathbb{E}(X) \geqslant 0$.
(2) We have

$$\begin{aligned}
\mathbb{E}(aX + bY) &= \sum_z z\mathbb{P}(aX + bY = z) \\
&= \sum_z z \sum_w \mathbb{P}(aX + bY = z|Y = w)\mathbb{P}(Y = w) \\
&= \sum_z z \sum_w \mathbb{P}(aX + bw = z)\mathbb{P}(Y = w) \\
&= \sum_z z \sum_w \sum_s \mathbb{P}(aX + bw = z|X = s)\mathbb{P}(X = s)\mathbb{P}(Y = w) \\
&= \sum_{w,s} (as + bw)\mathbb{P}(X = s)\mathbb{P}(Y = w) \\
&= \sum_s \left(\sum_w (as + bw)\mathbb{P}(Y = w) \right) \mathbb{P}(X = s) \\
&= \sum_s (as + b\mathbb{E}(Y))\, \mathbb{P}(Y = s) \\
&= a\mathbb{E}(X) + b\mathbb{E}(Y).
\end{aligned}$$

(3) By definition, we have $\mathbb{E}(X) = \sum_{x:p(x)>0} xp(x)$. Only $x = c$ has $p(x) > 0$, so $\mathbb{E}(X) = cp(c) = c$ since $p(c) = 1$.

□

Example 2.17. If $X \sim \mathrm{Bern}(p)$, then $\mathbb{E}(X) = p$. If $X \sim \mathrm{Bin}(n, p)$, then $X = Y_1 + \ldots + Y_n$ where $Y_i \sim \mathrm{Bern}(p)$, so $\mathbb{E}(X) = np$. If $X \sim \mathrm{Pois}(\lambda)$, then

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{\infty} \frac{k\lambda^k}{k!} e^{-\lambda} \\
&= \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} \\
&= \lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = \lambda.
\end{aligned}$$

Now, suppose X has pmf $p(k) = Ak^{-2}$ for $k \geqslant 1$ (where A is a "normalization constant", so that $\sum_{k\geqslant 1} p(k) = 1$). Then $\mathbb{E}X = \sum_{k=1}^{\infty} Ak^{-1} = \infty$.

2.4. **Variance and higher moments.**

Definition 2.18. Given a random variable X, its *variance* is $\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2$. Its *$k$-th moment* (for any $k \geqslant 0$) is $\mathbb{E}X^k$. We will often take k to be an integer.

Given any random variables X, Y, the *covariance* between X and Y is $\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$. In particular, we have $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$. We say X, Y are *uncorrelated* if $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Lemma 2.19. *(1) For any random variables X and Y, we have*

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2, \quad \mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

In particular, if X, Y are uncorrelated, then $\mathrm{Cov}(X, Y) = 0$.

(2) If X, Y are independent, then X, Y are uncorrelated.

(3) If $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are random variables, and $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are real numbers, then

$$\mathrm{Cov}(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{n} b_j Y_j) = \sum_{i,j=1}^{n} a_i b_j \mathrm{Cov}(X_i, Y_j).$$

This is often called bilinearity *of the covariance.*

(4) For any $a \in \mathbb{R}$, we have $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$. (In words, the variance is "quadratic".)

(5) There exists a constant c such that $X = c$ with probability 1 if and only if $\mathrm{Var}(X) = 0$.

Proof. (1) By definition, we have $\mathrm{Cov}(X, Y) = \mathbb{E}[XY - \mathbb{E}(X)Y - X\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$, since $\mathbb{E}[\cdot]$ is always a constant (we also use linearity of expectation here). The formula for variance follows by taking $Y = X$.

(2) If X, Y are independent, then

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_z z\mathbb{P}(XY = z) \\
&= \sum_z z \sum_w \mathbb{P}(XY = z | Y = w)\mathbb{P}(Y = w) \\
&= \sum_z z \sum_w \mathbb{P}(wX = z | Y = w)\mathbb{P}(Y = w) \\
&= \sum_z z \sum_w \mathbb{P}(Y = w)\mathbb{P}(wX = z) \\
&= \sum_z z \sum_w \mathbb{P}(Y = w) \sum_s \mathbb{P}(wX = z | X = s)\mathbb{P}(X = s) \\
&= \sum_{w,s} ws\mathbb{P}(Y = w)\mathbb{P}(X = s) \\
&= \sum_w w\mathbb{P}(Y = w) \sum_s \mathbb{P}(X = s) = \mathbb{E}(Y)\mathbb{E}(X).
\end{aligned}$$

(3) We have

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i \sum_{j=1}^{n} b_j Y_j\right] = \mathbb{E}\left[\sum_{i,j=1}^{n} a_i b_j X_i Y_j\right]$$
$$= \sum_{i,j=1}^{n} a_i b_j \mathbb{E}[X_i Y_j]$$

and

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] \mathbb{E}\left[\sum_{j=1}^{n} b_j Y_j\right] = \left\{\sum_{i=1}^{n} a_i \mathbb{E}[X_i]\right\}\left\{\sum_{j=1}^{n} b_j \mathbb{E}[Y_j]\right\}$$
$$= \sum_{i,j=1}^{n} a_i b_j \mathbb{E}[X_i]\mathbb{E}[Y_j].$$

Plug this into $\mathrm{Cov}(\sum_{i=1}^n a_i X_i, \sum_{j=1}^n b_j Y_j) = \mathbb{E}\left[\sum_{i=1}^n a_i X_i \sum_{j=1}^n b_j Y_j\right] - \mathbb{E}\left[\sum_{i=1}^n a_i X_i\right] \mathbb{E}\left[\sum_{j=1}^n b_j Y_j\right]$ to get the formula.

(4) Use part (3) with $n = 1$ and $a_1, b_1 = a$ and $X_1, Y_1 = X$.

(5) If $X = c$ with probability 1, then $\mathbb{E}(X) = c$ and $X - \mathbb{E}(X) = 0$ with probability 1. So $\mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[(c - c)^2] = 0$. If $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = 0$, then $X = \mathbb{E}(X)$ with probability 1. Indeed, if $X = d$ for $d \neq \mathbb{E}(X)$ with positive probability p, since $(X - \mathbb{E}(X))^2 \geqslant 0$ with probability 1, we would get $\mathbb{E}[(X - \mathbb{E}(X))^2] \geqslant p(d - \mathbb{E}(X))^2 > 0$, a contradiction.

□

Example 2.20. Let $X \sim \mathrm{Bern}(p)$. We saw before that $\mathbb{E}X = p$. Now, note that $X^2 = X$, since $X \in \{0, 1\}$, so that $\mathbb{E}X^2 = \mathbb{E}X = p$ as well. Thus, its variance is $\mathbb{E}X^2 - (\mathbb{E}X)^2 = p - p^2$. Now, assume that $X \sim \mathrm{Pois}(\lambda)$. We saw that $\mathbb{E}X = \lambda$. We compute

$$\mathbb{E}X^2 = \sum_{k=0}^{\infty} \frac{k^2 \lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{k^2 \lambda^k}{k!}$$
$$= e^{-\lambda} \sum_{k=1}^{\infty} \frac{k \lambda^k}{(k-1)!}$$
$$= e^{-\lambda} \lambda \sum_{k=0}^{\infty} \frac{(k+1) \lambda^k}{k!}$$
$$= e^{-\lambda} \lambda \frac{d}{d\lambda} \left(\sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{k!}\right)$$
$$= e^{-\lambda} \lambda \frac{d}{d\lambda} \left(\lambda e^{\lambda}\right)$$
$$= \lambda^2 + \lambda.$$

Hence, the variance of $X \sim \text{Pois}(\lambda)$ is $\text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$. Notice how this does not scale quadratically in λ!

2.5. **Cauchy-Schwarz and Hölder inequalities.**

Lemma 2.21. *Suppose X, Y are two random variables. Then for any $a > 0$, we have $|\mathbb{E}(XY)| \leqslant \frac{a^2\mathbb{E}(X^2)}{2} + \frac{\mathbb{E}(Y^2)}{2a^2}$. We also have $|\mathbb{E}(XY)| \leqslant (\mathbb{E}(X^2))^{1/2}(\mathbb{E}(Y^2))^{1/2}$.*

Proof. For the first inequality, we first note $(aX - \frac{1}{a}Y)^2 = a^2X^2 + \frac{Y^2}{a^2} - 2XY \geqslant 0$ (it is non-negative because it is the square of something). Thus, $XY \leqslant \frac{a^2X^2}{2} + \frac{Y^2}{2a^2}$. Now, take expectations to get $\mathbb{E}(XY) \leqslant \frac{a^2\mathbb{E}(X^2)}{2} + \frac{\mathbb{E}(Y^2)}{2a^2}$. In the case where $\mathbb{E}(XY) \geqslant 0$, this is the first claim. If $\mathbb{E}(XY) < 0$, use the claim after replacing X by $-X$. To prove the second claim, use the first claim for $a = \sqrt{2}\frac{\sqrt{\mathbb{E}(Y^2)}}{\sqrt{\mathbb{E}(X^2)}}$. □

Lemma 2.22. *Suppose $p \in [1, \infty) \cup \{\infty\}$ and suppose $\frac{1}{p} + \frac{1}{q} = 1$. Then $|\mathbb{E}(XY)| \leqslant (\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|Y|^q)^{1/q}$. (Note that if $p = q = 2$, this recovers Cauchy-Schwarz.)*

Proof. It suffices to instead use $XY \leqslant \frac{a^p|X|^p}{p} + \frac{|Y|^q}{a^q q}$ for any $a > 0$, take expectation, and choose a appropriately. □

3. Week 3, starting Tue. Feb. 6, 2024

3.1. **Law of the unconscious statistician.** Here's a quick trick that we introduced last week.

Lemma 3.1. *Take any function $f : \mathbb{R} \to \mathbb{R}$ (piecewise continuous, say). Take any random variable X with pmf p. Then*

$$\mathbb{E}[f(X)] = \sum_{x:p(x)>0} f(x)p(x).$$

Proof. By definition, we have

$$\begin{aligned}
\mathbb{E}[f(X)] &= \sum_w w\mathbb{P}[f(X) = w] \\
&= \sum_w w \sum_{s:f(s)=w} \mathbb{P}[X = s] \\
&= \sum_w \sum_{s:f(s)=w} w\mathbb{P}[X = s] \\
&= \sum_w \sum_{s:f(s)=w} f(s)\mathbb{P}[X = s] \\
&= \sum_s f(s)\mathbb{P}[X = s].
\end{aligned}$$

□

3.2. **Continuous random variables.**

Definition 3.2. A random variable X is said to be *continuous* if its distribution function can be written as $\mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} p(u)du$ for an integrable function p. This function p is the *density* or *probability density* function (or pdf for short).

Lemma 3.3. *Suppose X has pdf p. Then*

(1) $\int_{\mathbb{R}} p(x)dx = 1$
(2) $\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b p(x)dx$
(3) If p is continuous, then $p(x) \geqslant 0$ for all $x \in \mathbb{R}$
(4) $\mathbb{P}(X = x) = 0$ *for all* $x \in \mathbb{R}$

Proof. (1) $\mathbb{P}(X \leqslant A) = \int_{-\infty}^{A} p(u)du$. Now send $A \to \infty$. The LHS converges to 1.
(2) We have $\mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(X \leqslant b) - \mathbb{P}(X \leqslant a) = \int_{-\infty}^{b} p(u)du - \int_{-\infty}^{a} p(u)du = \int_a^b p(u)du$.
(3) For any $\varepsilon > 0$, we can pick $h > 0$ small enough so that $|p(y) - p(x)| \leqslant \varepsilon$ for all $y \in [x, x+h]$. In particular, for the sake of contradiction, suppose $p(x) < 0$ at x. Then $p(y) < 0$ for all $y \in [x, x+h]$ if h is small enough. But $\mathbb{P}(x \leqslant X \leqslant x+h) = \int_x^{x+h} p(y)dy < 0$ if this were to be the case, which is ridiculous.
(4) Use part (2) and the fact that the integral of any function on an interval of length 0 is 0.

□

Remark 3.4. There is the issue now of which σ-algebra to take, since $\mathbb{R}$ is *not* a finite set. This is a delicate issue of "measure theory", which is beyond the scope of this course (and, to be honest, kind of besides the point of probability theory and statistics; it's just a necessary evil to be *fully general*). For the purposes of this course (and really most situations one finds themselves in), as long as events are constructed by countable unions and intersections of events of the form $\{X \leqslant A\}$, one can integrate on them.

Example 3.5. There are three "main" examples of continuous random variables that we will be interested in. The first is the *normal* or *Gaussian* distribution. We say $X \sim N(\mu, \sigma^2)$ (where $\mu, \sigma \in \mathbb{R}$) if its pdf is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

μ is called the "mean" (for a reason we will see shortly), and σ^2 is the variance (we will prove this shortly). We also call σ the standard deviation. (Pretend $\sigma > 0$. If $\sigma = 0$, then $X \sim N(\mu, \sigma^2)$ just means $X = \mu$ with probability 1.) From this formula, it is not hard to see that if $X \sim N(0, \sigma^2)$, then $X + \mu \sim N(\mu, \sigma^2)$ and $cX \sim N(0, c^2\sigma^2)$. Proving this requires a little something, but you can take this for granted. (We will see a proof soon.)

The fact this integrates to 1 over $x \in \mathbb{R}$ is not easy to see! Let us do this really quickly. First, it suffices to assume that $\mu = 0$, since by change of variables, we have $\int_{\mathbb{R}} p(u)du = \int_{\mathbb{R}} p(u+\mu)du$ for all $\mu \in \mathbb{R}$. Moreover, by change of variables $u = x/\sigma$, it

suffices to assume that $\sigma = 1$. So, we need to show that

$$\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}dx\right)^2 = 1.$$

The LHS is equal to

$$\frac{1}{2\pi}\int_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2}}dxdy.$$

If we use polar coordinates $r^2 = x^2 + y^2$ and $dxdy = rdrd\theta$, we have

$$\begin{aligned}\frac{1}{2\pi}\int_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2}}dxdy &= \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2}}rdrd\theta \\ &= \frac{1}{2\pi}\int_0^{2\pi}\frac{d}{dr}e^{-\frac{r^2}{2}}drd\theta \\ &= \frac{1}{2\pi}\int_0^{2\pi} d\theta = 1.\end{aligned}$$

Example 3.6. We say $X \sim U([a,b])$ (or X is uniform on $[a,b]$) if its density function is $p(x) = \frac{1}{b-a}$ if $x \in [a,b]$, and $p(x) = 0$ if $x \notin [a,b]$. (If $a = b$, then this just means $X = a$ with probability 1.)

Example 3.7. We say $X \sim \mathrm{Exp}(\lambda)$ if its pdf is $p(x) = \lambda e^{-\lambda x}$ for $x \geqslant 0$ and $p(x) = 0$ for $x < 0$. (This is called an exponential random variable.)

Example 3.8. Here is another family of examples to keep in mind. We say X has a *power law* tail if its pdf satisfies $p(x) = A(1+x)^{-m}$ for some $m \geqslant 0$. Note that we must take $m > 1$ for this to even have finite integral on $\mathbb{R}$! The bigger m is, the less likely this random variable is going to be big.

Definition 3.9. Let X be a continuous random variable with pdf p. Take any function $f : \mathbb{R} \to \mathbb{R}$. Its *expectation* is $\mathbb{E}f(X) := \int_{-\infty}^{\infty} f(u)p(u)du$, provided that this integral converges absolutely. Its k-th moment is $\mathbb{E}X^k$. Its variance is $\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$. The covariance of X, Y is still $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Example 3.10. Let $X \sim N(0,1)$. Choose $f(x) = x$. Then $\mathbb{E}f(X) = \mathbb{E}X = \int_{-\infty}^{\infty} x\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}dx = 0$, since the integrand is odd. In particular, this agrees with calling μ (which in this case is 0) the mean. Next, choose $f(x) = x^2$. How do we compute its expectation? Well, first write

$$\mathbb{E}f(X) = \mathbb{E}X^2 = \int_{-\infty}^{\infty} x^2\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}dx = \int_{-\infty}^{\infty}(x^2-1)\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}dx + 1.$$

One can verify directly that $(x^2-1)e^{-\frac{x^2}{2}} = \frac{d^2}{dx^2}e^{-\frac{x^2}{2}} = -\frac{d}{dx}(xe^{-\frac{x^2}{2}})$. Thus, by the fundamental theorem of calculus, the integral on the far RHS is 0, since $xe^{-\frac{x^2}{2}}$ vanishes as $x \to \pm\infty$. In general, if $X \sim N(\mu, \sigma^2)$, then

$$\mathbb{E}X = \mu, \quad \mathbb{E}X^2 = \sigma^2 + \mu^2.$$

Of course, one can play a similar game to prove this, but we'll see a much easier way to do it. In particular, we will show that if $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim N(0,1)$ (provided $\sigma \neq 0$).

Example 3.11. As this and the previous example indicate, computing expectations often involve integration-by-parts. Let $X \sim \text{Exp}(\lambda)$. Then

$$\begin{aligned}\mathbb{E}X &= \int_0^\infty \lambda x e^{-\lambda x} dx = -\int_0^\infty x \frac{d}{dx} e^{-\lambda x} dx \\ &= \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda},\end{aligned}$$

where the last step uses u-substitution $u = \lambda x$. For the second moment $\mathbb{E}X^2$, we have

$$\begin{aligned}\mathbb{E}X^2 &= \int_0^\infty \lambda x^2 e^{-\lambda x} dx = -\int_0^\infty x^2 \frac{d}{dx} e^{-\lambda x} dx \\ &= 2\int_0^\infty x e^{-\lambda x} dx = \frac{2}{\lambda^2},\end{aligned}$$

where the last step uses our knowledge of $\mathbb{E}X = \lambda^{-1}$. Continuing in similar fashion, we can compute $\mathbb{E}X^k$ for any integer $k \geqslant 0$.

3.3. **Independence.**

Definition 3.12. Suppose that $X_1, \ldots, X_n$ are continuous random variables with pdfs $p_1, \ldots, p_n$. We say they are *jointly independent* if for any open or closed intervals $I_1, \ldots, I_n \subseteq \mathbb{R}$, we have

$$\mathbb{P}\left(\cap_{i=1}^n \{X_i \in I_i\}\right) = \prod_{i \in I} \mathbb{P}(X_i \in I) = \prod_{i=1}^n \int_{I_i} p_i(x) dx.$$

We say they are *pairwise independent* if X_i, X_j are independent for all choices of $i \neq j$. *Again, these notions are not the same!*

Lemma 3.13. *Let $X_1, \ldots, X_n$ be any random variables. Then they are jointly independent if and only if for any functions $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{E}\left[\prod_{i=1}^n f_i(X_i)\right] = \prod_{i=1}^n \mathbb{E}[f_i(X_i)].$$

Note that in the previous lemma, the random variables do not have to be continuous!

Lemma 3.14 (Convolution formula)**.** *Let X_1, X_2 be independent continuous random variables with pdfs p_1, p_2. Then $Z = X_1 + X_2$ is a continuous random variable with pdf*

$$p(z) = \int_{\mathbb{R}} p_1(z-u) p_2(u) du.$$

Proof. Same as in the discrete variable case. □

Lemma 3.15. *Lemma 2.19 is still true if the random variables therein are continuous random variables!*

3.4. **Change of variables.**

Theorem 3.16. *Let X be a continuous random variable with pdf p. Let $h : \mathbb{R} \to \mathbb{R}$ be a smooth, strictly monotone function. Then the random variable $Y = h(X)$ is continuous with pdf q given by*

$$q(y) = p(h^{-1}(y)) \left| \frac{1}{h'[h^{-1}(y)]} \right|.$$

If F is instead strictly decreasing, then $q(y) = p(-F(y))|F'(y)|$.

Proof. It suffices to show that for any $A \in \mathbb{R}$ and the proposed choice of q, we have

$$\mathbb{P}(Y \leqslant A) = \int_{-\infty}^{A} q(y)dy.$$

We have

$$\mathbb{P}(Y \leqslant A) = \mathbb{P}(F(X) \leqslant A) = \int_{\{x \in \mathbb{R} : F(x) \leqslant A\}} p(x)dx.$$

Since F is strictly increasing, we know that F is invertible, and the set $\{x \in \mathbb{R} : F(x) \leqslant A\}$ is equal to $[-\infty, F^{-1}(A)]$. Thus,

$$\mathbb{P}(Y \leqslant A) = \int_{-\infty}^{F^{-1}(A)} p(x)dx.$$

Now, make the change of variables $u = F^{-1}(x)$, i.e. $x = F(u)$. We have $dx = F'(u)du$. Moreover, this change of variables sends $[-\infty, F^{-1}(A)]$ to $[-\infty, A]$. Thus,

$$\mathbb{P}(Y \leqslant A) = \int_{-\infty}^{A} p(F(u))F'(u)du.$$

□

Example 3.17. Suppose X is uniform on $[0, 1]$, and $h(x) = -\log x$. This is smooth on $x > 0$ and strictly decreasing. Its inverse is $h^{-1}(x) = e^{-x}$. Its derivative is $h'(x) = -\frac{1}{x}$. So, the previous theorem tells us how to compute the distribution of $h(X)$; it turns out to be $\mathrm{Exp}(1)$! (This is on the HW.)

Example 3.18. This is one of the first ways we are taught how to sample from a distribution. Suppose X has pdf p. Recall $F(x) = \int_{-\infty}^{x} p(u)du$. To find its inverse, we need to know, given any $x \in [0, 1]$, for what value c is $F(c) = \int_{-\infty}^{c} p(u)du = x$. This $c(x)$ function is known as a *quantile* of x. In general, closed forms for quantiles are not available. Nevertheless, it turns out that $F(X)$ is uniform on $[0, 1]$ anyway; this is on the HW.

Example 3.19. Suppose $X \sim N(0, \sigma^2)$. We claim that $X + \mu \sim N(\mu, \sigma^2)$. To see this rigorously, note $X + \mu = h(X)$, where $h(x) = x + \mu$. Its derivative is $h'(x) = 1$, and its inverse is $h^{-1}(x) = x - \mu$. Thus, the previous formula says that the pdf of $X + \mu$ is $p(x - \mu)$, where p is the pdf for $N(0, 1)$. But $p(x - \mu)$ is the pdf for $N(\mu, 1)$. Similarly, one can use the function $h(x) = \sigma x$ to show that $\sigma X \sim N(0, \sigma^2)$.

3.5. **Random vectors.**

Definition 3.20. Let $X_1, \ldots, X_n$ be continuous random variables, so that $\mathbf{X} = (X_1, \ldots, X_n)$ is a random vector in $\mathbb{R}^n$. The pdf of $\mathbf{X}$ is the function $p(x_1, \ldots, x_n)$ such that for any

open or closed subset $E \subseteq \mathbb{R}^n$, we have

$$\mathbb{P}(\mathbf{X} \in E) = \int_E p(u_1, \ldots, u_n) du_1 \ldots du_n.$$

Now, suppose $X_1, \ldots, X_n$ are discrete random variables. The pmf of $\mathbf{X}$ is the function $p(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$.

Finally, if $X_1, \ldots, X_j$ are continuous and $X_{j+1}, \ldots, X_n$ are discrete, then the *density function* of $\mathbf{X}$ is defined as follows (in which $E \subseteq \mathbb{R}^j$ is any open or closed set):

$$\mathbb{P}((X_1, \ldots, X_j) \in E, X_{j+1} = x_{j+1}, \ldots, X_n = x_n) = \int_E p(x_1, \ldots, x_j, x_{j+1}, \ldots, x_n) dx_1 \ldots dx_j.$$

Example 3.21. Suppose $X_1, \ldots, X_n$ are independent with pdfs $p_1, \ldots, p_n$. Then $p(x_1, \ldots, x_n) = \prod_{i=1}^n p(x_i)$. Indeed, for any $E = E_1 \times \ldots \times E_n$ where $E_1 \subseteq \mathbb{R}$ are open or closed, by independence, we have

$$\mathbb{P}(\mathbf{X} \in E) = \mathbb{P}(\cap_{i=1}^n \{X_i \in E_i\}) = \prod_{i=1}^n \mathbb{P}(X_i \in E_i).$$

Now, for any open or closed $E \subseteq \mathbb{R}^n$, we can always approximate E by a disjoint union of rectangles. This requires some work, but it can be done. This example applies to continuous or discrete random variables.

Definition 3.22. Let $X_1, \ldots, X_n$ be continuous random variables, so that $\mathbf{X} = (X_1, \ldots, X_n)$ has pdf $p(x_1, \ldots, x_n)$. For any function $f : \mathbb{R}^n \to \mathbb{R}$, the expectation of f is

$$\mathbb{E} f(X_1, \ldots, X_n) = \int_{\mathbb{R}^n} f(x_1, \ldots, x_n) p(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

If $X_1, \ldots, X_n$ are instead discrete and $\mathbf{X}$ has pmf $p(x_1, \ldots, x_n)$, then

$$\mathbb{E} f(X_1, \ldots, X_n) = \sum_{(x_1, \ldots, x_n)} f(x_1, \ldots, x_n) p(x_1, \ldots, x_n).$$

Suppose $X_1, \ldots, X_j$ are continuous and $X_{j+1}, \ldots, X_n$ are discrete. Then

$$\mathbb{E} f(X_1, \ldots, X_n) = \int_{\mathbb{R}^j} \sum_{(x_{j+1}, \ldots, x_n)} f(x_1, \ldots, x_j, x_{j+1}, \ldots, x_n) p(x_1, \ldots, x_j, x_{j+1}, \ldots, x_n) dx_1 \ldots dx_j.$$

Example 3.23. Suppose X_1, X_2 are continuous pdfs such that $X_1 = X_2$, and X_1, X_2 have pdf p. Then the pdf of $\mathbf{X}$ is a little funny; it has the form $p(x_1, x_2) = p(x_1) \text{“}\delta_{x_1 = x_2}\text{”}$. This $\delta_{x=y}$ vanishes whenever $x \neq y$, and it reduces to integration only when $x = y$. In particular, for any $E \subseteq \mathbb{R}^2$, let E_1 be the set of all $x \in \mathbb{R}$ for which $(x, x) \in E$. Then we have

$$\mathbb{P}(\mathbf{X} \in E) = \int_E p(x, y) \delta_{x=y} dx dy = \int_{E_1} p(x) dx.$$

This example is not too important, since we will never use it in this class, but I want to mention it just to let you know that things can be a little weird if one is too reckless and does not throw out complete redundancies in $\mathbf{X}$.

3.6. **Multivariate Gaussians.**

Definition 3.24. Recall that a square matrix is positive definite if it is real symmetric and all its eigenvalues are strictly positive.

We say a random vector $\mathbf{X} \in \mathbb{R}^n$ is a *multivariate Gaussian*, written as $\mathbf{X} \sim N(\mathbf{m}, \Sigma)$ (where $\mathbf{m} = (m_1, \ldots, m_n) \in \mathbb{R}^n$ and Σ is a positive-definite matrix of dimension $n \times n$), if its pdf is given by (for $\mathbf{x} = (x_1, \ldots, x_n)$)

$$p(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left\{-\frac{(\mathbf{x}-\mathbf{m}) \cdot \Sigma^{-1}(\mathbf{x}-\mathbf{m})}{2}\right\}$$

Because Σ is positive definite, it is invertible.

Example 3.25. Let $X_1, \ldots, X_n$ are independent $N(m_i, \sigma_i^2)$ for $i = 1, \ldots, n$. Then $\mathbf{X} = (X_1, \ldots, X_n)$ is a multivariate Gaussian with $\mathbf{m} = (m_1, \ldots, m_n)$ and Σ diagonal with $\Sigma_{ii} = \sigma_i^2$. Indeed, by independence, the pdf of $\mathbf{X}$ is

$$\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x_i-m_i)^2}{2\sigma_i^2}} = \frac{1}{\sqrt{\prod_{i=1}^n 2\pi\sigma_i^2}} \exp\left\{-\frac{\sum_{i=1}^n (x_i - m_i)\sigma_i^{-2}(x_i - m_i)}{2}\right\}.$$

One can check that the determinant of $2\pi\Sigma$ is the product of its diagonal entries $2\pi\sigma_i^2$, and that $(\mathbf{x}-\mathbf{m}) \cdot \Sigma^{-1}(\mathbf{x}-\mathbf{m}) = \sum_i (x_i - m_i)\sigma_i^{-2}(\sigma_i - m_i)$, since the inverse of a diagonal matrix with positive entries is the diagonal matrix given by inverting the diagonal entries.

Lemma 3.26. *The pdf $p(\mathbf{x})$ for $N(\mathbf{m}, \Sigma)$ is, in fact, a pdf (so that $\int_{\mathbb{R}^n} p(\mathbf{x}) d\mathbf{x} = 1$.*

Proof. As in the $n = 1$ case, one can shift $\mathbf{u} = \mathbf{x} - \mathbf{m}$ and assume $\mathbf{m} = 0$. We must show

$$\frac{1}{\sqrt{\det(2\pi\Sigma)}} \int_{\mathbb{R}^n} \exp\left\{-\frac{\mathbf{x} \cdot \Sigma^{-1}\mathbf{x}}{2}\right\} d\mathbf{x} = 1.$$

Since Σ is real symmetric with positive eigenvalues, by the spectral theorem in linear algebra, we can write $\Sigma = O^T D O$, where O is orthogonal (so $OO^T = O^T O = I$) and D is diagonal with positive diagonal entries $D_1, \ldots, D_n$. In particular, $\Sigma = O^T D^{-1} O$ and $\det \Sigma = \det D$. So, the LHS of the previous display is equal to

$$\frac{1}{\sqrt{\det(2\pi D)}} \int_{\mathbb{R}^n} \exp\left\{-\frac{O\mathbf{x} \cdot D^{-1} O\mathbf{x}}{2}\right\} d\mathbf{x}.$$

Since O is orthogonal, the change of variables $\mathbf{u} = O\mathbf{x}$ satisfies $d\mathbf{u} = d\mathbf{x}$. So, the previous display equals

$$\begin{aligned}\frac{1}{\sqrt{\det(2\pi D)}} \int_{\mathbb{R}^n} \exp\left\{-\frac{\mathbf{x} \cdot D^{-1}\mathbf{x}}{2}\right\} d\mathbf{x} &= \frac{1}{\sqrt{\det(2\pi D)}} \int_{\mathbb{R}^n} \prod_{i=1}^n e^{-\frac{x_i^2}{2D_i}} dx_i \\ &= \int_{\mathbb{R}^n} \prod_{i=1}^n \frac{1}{\sqrt{2\pi D_i}} e^{-\frac{x_i^2}{2D_i}} dx_i.\end{aligned}$$

We used the fact that the determinant of a diagonal matrix is the product of its entries above. The last integral is the product of integrals of pdfs of one-dimensional Gaussians, which are all 1, so the proof is complete. □

Lemma 3.27. *Let $\mathbf{X} \sim N(\mathbf{m}, \Sigma)$.*
(1) If $\mathbf{X} \sim N(\mathbf{m}, \Sigma)$, then $\mathbf{X} + \mathbf{w} \sim N(\mathbf{m} + \mathbf{w}, \Sigma)$ and $M\mathbf{X} \sim N(\mathbf{m}, M^*\Sigma M)$.
(2) For any $i = 1, \ldots, n$, we have $\mathbb{E}X_i = m_i$.
(3) For any $i, j = 1, \ldots, n$, we have $\mathrm{Cov}(X_i, X_j) = \Sigma_{ij}$.

Proof. (1) Omitted.

(2) Set $\mathbf{Y} = \mathbf{X} - \mathbf{m}$. Then $\mathbf{Y} \sim N(0, \Sigma)$. But the pdf for $N(0, \Sigma)$ is symmetric about the origin, so $\mathbb{E}Y_i = -\mathbb{E}Y_i = 0$. Thus, $\mathbb{E}X_i = \mathbb{E}Y_i + m_i = m_i$.

(3) For notational convenience, let us assume $\mathbf{m} = (0, \dots, 0)$, so that $\mathbb{E}X_i, \mathbb{E}X_j = 0$ and thus $\operatorname{Cov}(X_i, X_j) = \mathbb{E}X_iX_j - \mathbb{E}X_i\mathbb{E}X_j$. We want to show that

$$\Sigma_{ij} = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \int_{\mathbb{R}^n} x_i x_j p(\mathbf{x}) d\mathbf{x}.$$

Because Σ is real symmetric and positive definite, by the spectral theorem in linear algebra, we can write $\Sigma = O^T D O$, where D is diagonal with entries $D_1, \dots, D_n > 0$ and O is an orthogonal matrix satisfying $OO^T = O^T O = I$. So, we have $\Sigma^{-1} = O^T D^{-1} O$. Moreover, we have $\det \Sigma = \det D$. Hence, we have

$$p(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi D)}} \exp\left\{-\frac{O\mathbf{x} \cdot DO\mathbf{x}}{2}\right\}.$$

Now, let A be the $n \times n$ matrix such that $A_{ij} = A_{ji} = \frac{1}{2}$. Then $x_i x_j = \mathbf{x} \cdot A\mathbf{x} = O\mathbf{x} \cdot OAO^T O\mathbf{x}$. Thus, we want to show

$$\Sigma_{ij} = \frac{1}{\sqrt{\det(2\pi D^{-1})}} \int_{\mathbb{R}^n} O\mathbf{x} \cdot OAO^T O\mathbf{x} \exp\left\{-\frac{O\mathbf{x} \cdot DO\mathbf{x}}{2}\right\} d\mathbf{x}.$$

The multivariable change-of-variables formula implies that the u-substitution $u = O\mathbf{x}$ implies $d\mathbf{u} = d\mathbf{x}$. Thus, the RHS of the previous display is

$$\frac{1}{\sqrt{\det(2\pi D^{-1})}} \int_{\mathbb{R}^n} \mathbf{x} \cdot OAO^T \mathbf{x} \exp\left\{-\frac{\mathbf{x} \cdot D\mathbf{x}}{2}\right\} d\mathbf{x} \tag{3.1}$$

$$= \int_{\mathbb{R}^n} \mathbf{x} \cdot OAO^T \mathbf{x} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi D_i^{-1}}} e^{-\frac{D_i x_i^2}{2}} dx_i. \tag{3.2}$$

If $Z = (Z_1, \dots, Z_n)$ where $Z_i \sim N(0, D_i^{-1})$ are independent, then the previous display is equal to the expectation of $Z \cdot OAO^T Z$. It requires a linear algebra, but this can be shown to equal $(O^T D O)_{ij} = \Sigma_{ij}$.

□

4. Week 4, starting Tue. Feb. 13, 2024

4.1. Triangle inequality.

Lemma 4.1. *We have $|\mathbb{E}X| \leqslant \mathbb{E}|X|$ for any random variable X.*

Proof. Suppose X is discrete. Then $|\mathbb{E}X| = |\sum_x xp(x)| \leqslant \sum_x |x|p(x) = \mathbb{E}|X|$. If X is continuous, we have $|\mathbb{E}X| = \left|\int_{\mathbb{R}} xp(x)dx\right| \leqslant \int_{\mathbb{R}} |x|p(x)dx = \mathbb{E}|X|$. □

4.2. Laplace and Fourier transforms, i.e. moment generating functions and characteristic functions.

Definition 4.2. For any random variable X, we define its Laplace transform/moment generating function (MGF) to be the function $m_X(\xi) := \mathbb{E}e^{\xi X}$. We define its Fourier transform/characteristic function to be $\chi_X(\xi) := \mathbb{E}e^{i\xi X} = m_X(i\xi)$.

Lemma 4.3. *(1) If X, Y are independent, then $m_{X+Y}(\xi) = m_X(\xi)m_Y(\xi)$ and $\chi_{X+Y}(\xi) = \chi_X(\xi)\chi_Y(\xi)$.*
(2) We have $m_X(0) = \chi_X(0) = 1$.
(3) We have $|\chi_X(\xi)| \leqslant 1$ for all $\xi \in \mathbb{R}$.

Proof. (1) We have $m_{X+Y}(\xi) = \mathbb{E}e^{\xi(X+Y)} = \mathbb{E}e^{\xi X}e^{\xi Y} = \mathbb{E}e^{\xi X}\mathbb{E}e^{\xi Y} = m_X(\xi)m_Y(\xi)$. For the other identity, use $\chi(\xi) = m(i\xi)$.
(2) We have $m_X(0), \chi_X(0) = \mathbb{E}e^{0X} = \mathbb{E}1 = 1$.
(3) Since $|e^{ix}| = 1$ for all $x \in \mathbb{R}$, we have $|\xi_X(\xi)| = |\mathbb{E}e^{i\xi X}|$. Now, by the *triangle inequality*, we have $|\mathbb{E}e^{i\xi X}| \leqslant \mathbb{E}|e^{i\xi X}| = 1$.

□

Theorem 4.4 (An inversion theorem)**.** *Suppose X, Y are random variables such that $m_X(\xi) = m_Y(\xi)$ for all ξ in a neighborhood of 0. Then X, Y have the same distribution, i.e. $\mathbb{P}[X \in A] = \mathbb{P}[Y \in A]$ for all open, closed, half-open, or half-closed subsets $A \subseteq \mathbb{R}$. The same is true for χ in place of m.*

Example 4.5. Let $X \sim \mathrm{Bern}(p)$. Then $\mathbb{E}e^{\xi X} = (1-p) + pe^{\xi}$. Now, suppose $Y \sim \mathrm{Bin}(n, p)$. We can compute

$$\mathbb{E}e^{\xi Y} = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} e^{k\xi} = \sum_{k=0}^{n} \binom{n}{k} [pe^{\xi}]^k (1-p)^{n-k} = (pe^{\xi} + (1-p))^n.$$

On the other hand, we know $Y = X_1 + \ldots + X_n$, so $\mathbb{E}e^{\xi Y} = \prod_{j=1}^{n} \mathbb{E}e^{\xi X_j} = \prod_{j=1}^{n} [pe^{\xi} + (1-p)] = [pe^{\xi} + (1-p)]^n$. This is another illustration that $Y = X_1 + \ldots + X_n$ for independent $X_j \sim \mathrm{Bern}(p)$.

Example 4.6. The sum of independent Gaussians is Gaussian. Let $X \sim N(0, \sigma_1^2)$ and $Y \sim N(0, \sigma_2^2)$. In HW3, you showed that $\mathbb{E}e^{\xi X} = e^{\frac{\xi^2\sigma_1^2}{2}}$ and $\mathbb{E}e^{\xi Y} = e^{\frac{\xi^2\sigma_2^2}{2}}$. From this, we know that $\mathbb{E}e^{\xi(X+Y)} = e^{\frac{\xi^2(\sigma_1^2+\sigma_2^2)}{2}}$. This shows that $X+Y$ has the same Laplace transform as $N(0, \sigma_1^2 + \sigma_2^2)$. So, by the inversion theorem, we know that $X + Y \sim N(0, \sigma_1^2 + \sigma_2^2)$.

Theorem 4.7 (Another inversion theorem)**.** *Let X be a discrete random variable with pmf $p(x)$. Suppose $f : \mathbb{R} \to \mathbb{C}$ is a function that satisfies $\frac{1}{2\pi}\int_{\mathbb{R}} f(\xi)e^{-ix\xi}d\xi = p(x)$ for all $x \in \mathbb{R}$. Then $\chi_X(\xi) = \mathbb{E}e^{i\xi X} = f(\xi)$. The same is true if X is a continuous random variable with pdf $p(x)$.*

Example 4.8. On HW4, you are introduced to the Cauchy distribution, which is a continuous one with pdf $p(x) = \frac{1}{\pi(1+x^2)}$. Its Fourier transform $\chi_X(\xi)$ is not so easy to compute, but it turns out to equal $e^{-|\xi|}$ (you are asked to do this computation). It is noticeably easier to show that $\frac{1}{2\pi}\int_{\mathbb{R}} e^{-|\xi|}e^{-ix\xi}d\xi = \frac{1}{\pi(1+x^2)}$. The inversion theorem now shows $\mathbb{E}e^{i\xi X} = e^{-|\xi|^2}$ if X is Cauchy.

4.3. **How to compute moments.**

Lemma 4.9. *For any random variable X and integer $k \geqslant 0$, we have*

$$\frac{d^k}{d\xi^k}\mathbb{E}e^{\xi X}|_{\xi=0} = \mathbb{E}X^k,$$
$$\frac{d^k}{d\xi^k}\mathbb{E}e^{i\xi X}|_{\xi=0} = i^k\mathbb{E}X^k.$$

Proof. By the chain rule, we have $\frac{d^k}{d\xi^k}e^{\xi X} = X^k$ and $\frac{d^k}{d\xi^k}e^{i\xi X} = i^k X^k$. Now take expectation on both sides. □

Example 4.10. If $X \sim \mathrm{Bern}(p)$. Then $\mathbb{E}X^k = \mathbb{E}X$ for all $k \geqslant 0$ because X is either 0 or 1. On the other hand, $\mathbb{E}e^{\xi X} = (1-p) + pe^{\xi}$, and e^{ξ} stays put whenever we take derivatives.

Example 4.11. If $X \sim N(0,1)$, then $\mathbb{E}e^{\xi X} = e^{\frac{\xi^2}{2}}$. We have $\frac{d}{d\xi}e^{\frac{\xi^2}{2}} = \xi e^{\frac{\xi^2}{2}}$ and $\frac{d^2}{d\xi^2}e^{\frac{\xi^2}{2}} = (\xi^2+1)e^{\frac{\xi^2}{2}}$ and $\frac{d^4}{d\xi^4}e^{\frac{\xi^2}{2}} = (\xi^4+3\xi^2+3)e^{\frac{\xi^2}{2}}$, so if we set $\xi = 0$, we get $\mathbb{E}X = 0$ and $\mathbb{E}X^2 = 1$ and $\mathbb{E}X^4 = 3$. This is what you showed on HW3, but in an easier way!

4.4. **Some inequalities.**

Lemma 4.12. *Suppose X, Y are two random variables. Then for any $a > 0$, we have $|\mathbb{E}(XY)| \leqslant \frac{a^2\mathbb{E}(X^2)}{2} + \frac{\mathbb{E}(Y^2)}{2a^2}$. We also have $|\mathbb{E}(XY)| \leqslant (\mathbb{E}(X^2))^{1/2}(\mathbb{E}(Y^2))^{1/2}$.*

Proof. For the first inequality, we first note $(aX - \frac{1}{a}Y)^2 = a^2X^2 + \frac{Y^2}{a^2} - 2XY \geqslant 0$ (it is non-negative because it is the square of something). Thus, $XY \leqslant \frac{a^2X^2}{2} + \frac{Y^2}{2a^2}$. Now, take expectations to get $\mathbb{E}(XY) \leqslant \frac{a^2\mathbb{E}(X^2)}{2} + \frac{\mathbb{E}(Y^2)}{2a^2}$. In the case where $\mathbb{E}(XY) \geqslant 0$, this is the first claim. If $\mathbb{E}(XY) < 0$, use the claim after replacing X by $-X$. To prove the second claim, use the first claim for $a = \sqrt{2}\frac{\sqrt{\mathbb{E}(Y^2)}}{\sqrt{\mathbb{E}(X^2)}}$. □

Example 4.13. Given two random variables X, Y, the correlation coefficient between them is $\sigma(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$. Note that this does not change if we replace X, Y by $\bar{X} = X - \mathbb{E}X$ and $\bar{Y} = Y - \mathbb{E}Y$, respectively. By Cauchy-Schwarz, we know that $|\sigma(X, Y)| \leqslant 1$. This means the correlation coefficient is a way to measure dependence of X, Y on each other without their size influencing anything.

Lemma 4.14. *Suppose $p \in [1, \infty) \cup \{\infty\}$ and suppose $\frac{1}{p} + \frac{1}{q} = 1$. Then $|\mathbb{E}(XY)| \leqslant (\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|Y|^q)^{1/q}$. (Note that if $p = q = 2$, this recovers Cauchy-Schwarz.)*

Proof. It suffices to instead use $XY \leqslant \frac{a^p|X|^p}{p} + \frac{|Y|^q}{a^q q}$ for any $a > 0$, take expectation, and choose a appropriately. □

Lemma 4.15 (Chebyshev inequality)**.** *Let X be a random variable. Then for any $p \geqslant 1$ and $C > 0$, we have $\mathbb{P}[X \geqslant C] \leqslant \frac{\mathbb{E}|X|^p}{C^p}$.*

More generally, if $\varphi : \mathbb{R} \to \mathbb{R}$ is an increasing function, then $\mathbb{P}[X \geqslant C] \leqslant \frac{\mathbb{E}\varphi(X)}{\varphi(C)}$.

This is sometimes called Markov's inequality if $p = 1$. Although the first claim is true if $p > 0$, it is not useful if $p < 1$.

Proof. We prove the general version; for the first statement, take $\varphi(x) = x^p$. We have

$$\mathbb{P}[X \geqslant C] \leqslant \mathbb{P}[\varphi(X) \geqslant \varphi(C)] = \mathbb{E}\mathbf{1}_{\varphi(X)\geqslant\varphi(C)} \leqslant \mathbb{E}\mathbf{1}_{\varphi(X)\geqslant\varphi(C)}\frac{\varphi(X)}{\varphi(C)}.$$

Since $\varphi(X)/\varphi(C) \geqslant 1$, we can drop the indicator for an upper bound. □

Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is convex if $f''(x) \geqslant 0$ for all x. Equivalently, for any $t \in [0, 1]$ and $x, y \in \mathbb{R}$, we have $f(tx + (1 - t)y) \leqslant tf(x) + (1 - t)f(y)$. (In words, the graph of f sits below its tangent line.) By inducting on the number of points, we can show that for any $x_1, \ldots, x_n$ and $p_1, \ldots, p_n$ such that $p_1 + \ldots + p_n = 1$, we have $f(\sum_{i=1}^n p_i x_i) \leqslant \sum_{i=1}^n p_i f(x_i)$.

Lemma 4.16 (Jensen's inequality)**.** *Take any random variable X and any convex function f. We have $f(\mathbb{E}X) \leqslant \mathbb{E}f(X)$.*

Proof. If X is a discrete random variable, then $f(\mathbb{E}X) = f(\sum_x xp(x))$. By convexity, this is $\leqslant \sum_x f(x)p(x) = \mathbb{E}f(X)$. If X is a continuous random variable, one has to use an approximation argument (which we omit). □

4.5. **Some applications of these inequalities.**

Lemma 4.17. *For any random variable X and $p \geqslant 1$, we have $|\mathbb{E}X|^p \leqslant \mathbb{E}|X|^p$.*

Proof. We give two proofs. First, note that $f(x) = |x|^p$ if convex if $p \geqslant 1$. (It suffices to prove this for $x \geqslant 0$ since $f(x) = f(-x)$. Now compute $f''(x) = p(p-1)x^{p-2}$ for $x \geqslant 0$, which is non-negative if $p \geqslant 1$.) Thus, we can now use Jensen. The second proof is based on Hölder. Let $Y = 1$ be the constant random variable, so that $|\mathbb{E}X| = |\mathbb{E}XY| \leqslant (\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|Y|^q)^{1/q} = (\mathbb{E}|X|^p)^{1/p}$. Now raise both sides of this inequality to the p-th power. □

Lemma 4.18 ("Reverse Hölder inequality)**.** *Suppose f is* concave*, i.e. $-f$ is convex. Then $f(\mathbb{E}X) \geqslant \mathbb{E}f(X)$. For example,* $\log|\mathbb{E}X| \geqslant \mathbb{E}\log|X|$.

Proof. By Jensen, we know $-f(\mathbb{E}X) \leqslant -\mathbb{E}f(X)$, so by taking negatives, we conclude the first claim. The second follows by noting that $x \mapsto \log|x|$ is concave (take $x > 0$ and take two derivatives). □

4.6. **The Law of Large Numbers.**

Theorem 4.19. *Let $X_1, \dots, X_N$ be independent random variables such that $\mathbb{E}X_j = 0$ for all $j = 1, \dots, N$. Define $Y = N^{-1}\sum_{j=1}^N X_j$. Then for any $\varepsilon > 0$, we have*

$$\mathbb{P}[|Y| \geqslant \varepsilon] \leqslant \frac{\sum_{j=1}^N \mathbb{E}|X_j|^2}{N^2\varepsilon^2} \leqslant \frac{1}{N\varepsilon^2}\sup_{j=1,\dots,N}\mathbb{E}|X_j|^2.$$

In particular, if $X_1, \dots, X_N$ have the same distribution, then $\mathbb{P}[|Y| \geqslant \varepsilon] \leqslant \frac{\mathrm{Var}(X_1)}{N\varepsilon^2}$.

Proof. By Chebyshev, we have

$$\mathbb{P}[|Y| \geqslant \varepsilon] \leqslant \frac{\mathbb{E}|Y|^2}{\varepsilon^2} = \frac{\frac{1}{N^2}\sum_{i,j=1}^N \mathbb{E}X_iX_j}{\varepsilon^2}.$$

Since X_i, X_j are independent, we know $\mathbb{E}X_iX_j = \mathbb{E}X_i\mathbb{E}X_j = 0$ if $i \neq j$. Thus, $\mathbb{P}[|Y| \geqslant \varepsilon] \leqslant \varepsilon^{-2}N^{-2}\sum_{i=1}^N \mathbb{E}|X_i|^2$. □

Example 4.20. If $X_1, \dots, X_N$ are independent $N(0,1)$, then we have already shown that $Y = N^{-1}\sum_{i=1}^N X_i \sim N(0, \frac{1}{N})$. In this case,

$$\mathbb{P}[|Y| \geqslant \varepsilon] = 2\int_\varepsilon^\infty \frac{1}{\sqrt{2\pi N^{-1}}}e^{-\frac{Nx^2}{2}}dx.$$

This vanishes as $N \to \infty$ if $\varepsilon > 0$ is fixed. Indeed, we know that $N^{1/2}\exp[-Nx^2/2] \leqslant C_\varepsilon \exp[-\sqrt{N}x]$ for all $x \geqslant \varepsilon$ if $C_\varepsilon > 0$ is sufficiently large depending only on ε. But the integral of $\exp[-\sqrt{N}x]$ from $x = \varepsilon$ to $x = \infty$ is $\leqslant \exp[-\sqrt{N}\varepsilon]$, which vanishes as $N \to \infty$.

Example 4.21. Let $X_1, \dots, X_N$ be Cauchy random variables (independent!), i.e. continuous with pdf $p(u) = \frac{1}{\pi(1+u^2)}$ for $u \in \mathbb{R}$. You will show on HW4 that $Y = N^{-1}\sum_{i=1}^N X_i$ is also Cauchy for all N. Thus, the law of large numbers does not apply! Why?

5. Week 5, starting Tue. Feb. 19, 2024

5.1. **Just a reminder.** These notes are not designed to be a substitute for lecture; they're more or less meant to help organize my thoughts for class, and in case anybody finds them helpful. In particular, these notes do not cover every detail said in class. Also, it means that typos may or may not be corrected even after lecture.

5.2. Random vectors.

Definition 5.1. A *discrete* random vector of dimension (or length) n is a vector $\mathbf{X} \in \mathbb{R}^n$ such that $\mathbf{X} = (X_1, \ldots, X_n)$ and $X_1, \ldots, X_n$ are discrete random variables. Its probability mass function $p : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$p(x_1, \ldots, x_n) = \mathbb{P}[\mathbf{X} = (x_1, \ldots, x_n)], \quad (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

A *continuous* random vector of dimension (of length) n is a vector $\mathbf{X} \in \mathbb{R}^n$ such that $\mathbf{X} = (X_1, \ldots, X_n)$ and $X_1, \ldots, X_n$ are continuous random variables. Its probability density function $p : \mathbb{R}^n \to \mathbb{R}$ is defined by the following, in which $U \subseteq \mathbb{R}^n$ is an arbitrary open set:

$$\mathbb{P}[\mathbf{X} \in U] = \int_U p(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

Lemma 5.2. *(1) If $X_1, \ldots, X_n$ are independent discrete random variables with pmfs $p_1, \ldots, p_n$, then $\mathbf{X} = (X_1, \ldots, X_n)$ is a discrete random vector with pmf $p(x_1, \ldots, x_n) = \prod_{i=1}^n p_i(x_i)$.*
(2) The same is true if we have continuous random variables, and pmf is replaced by pdf.

Proof. (1) For any $x_1, \ldots, x_n \in \mathbb{R}$, by independence, we have $\mathbb{P}[X_1 = x_1, \ldots, X_n = x_n] = \prod_{i=1}^n \mathbb{P}[X_i = x_i] = \prod_{i=1}^n p_i(x_i)$.
(2) Take any open set of the form $U = (a_1, b_1) \times \ldots \times (a_n, b_n)$. By independence, we have

$$\begin{aligned}
\mathbb{P}[\mathbf{X} \in U] &= \mathbb{P}[X_1 \in (a_1, b_1), \ldots, X_n \in (a_n, b_n)] = \prod_{i=1}^n \mathbb{P}[X_i \in (a_i, b_i)] \\
&= \prod_{i=1}^n \int_{a_i}^{b_i} p_i(x_i) dx_i = \prod_{i=1}^n \int_{\mathbb{R}} \mathbf{1}_{x_i \in (a_i, b_i)} p_i(x_i) dx_i \\
&= \int_{\mathbb{R}^n} \left(\prod_{i=1}^n \mathbf{1}_{x_i \in (a_i, b_i)} p_i(x_i) \right) dx_1 \ldots dx_n \\
&= \int_U \prod_{i=1}^n p_i(x_i) dx_1 \ldots dx_n.
\end{aligned}$$

□

Example 5.3. Let $X \sim \text{Pois}(\lambda)$ and $Y = X$. Then $\mathbf{X} = (X, Y)$ is a random vector whose pmf is $p(x, y) = \mathbf{1}_{x=y} p_{\text{Pois}(\lambda)}(x)$.

Example 5.4. Let us define the function

$$p(x,y) = \begin{cases} \frac{1}{4} & (x,y) = (0,0) \\ \frac{1}{4} & (x,y) = (0,1) \\ \frac{1}{4} & (x,y) = (1,0) \\ \frac{1}{4} & (x,y) = (1,1) \\ 0 & \text{else} \end{cases}$$

This describes two Bernoulli random variables whose distribution is a little unclear. If we write $\mathbf{X} = (X,Y)$ as a random vector with this pdf, then we have

$$\mathbb{E}X = \sum_{x,y} xp(x,y) = p(1,0) + p(1,1) = \frac{1}{2}.$$

A similar computation shows that $\mathbb{E}Y = \frac{1}{2}$. Thus, we know that $X, Y \sim \mathrm{Bern}(\frac{1}{2})$. What is their covariance? In particular,

$$\mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y = \sum_{x,y} xyp(x,y) - \frac{1}{4} = p(1,1) - \frac{1}{4} = 0.$$

One can actually show that X, Y are independent. I will leave that as an exercise. On the other hand, we can also consider the pdf

$$p(x,y) = \begin{cases} \frac{1}{2} & (x,y) = (0,0) \\ 0 & (x,y) = (0,1) \\ 0 & (x,y) = (1,0) \\ \frac{1}{2} & (x,y) = (1,1) \\ 0 & \text{else} \end{cases}$$

In this case, one can also show that $\mathbb{E}X = \mathbb{E}Y = \frac{1}{2}$, so that $X, Y \sim \mathrm{Bern}(\frac{1}{2})$. But, it is clear that $X = Y$, so they are not independent.

Example 5.5. Let $Y \sim N(X,1)$, where X is some continuous random variable. If we condition on $X = x$, then $Y \sim N(x,1)$. In particular, Y has a random mean given by X. Then $\mathbf{X} = (X,Y)$ is a continuous random vector. Its pdf is given by

$$p(x,y) = p_X(x) \times \frac{1}{[2\pi]^{1/2}} \exp\left\{ -\frac{(y-x)^2}{2} \right\}.$$

5.3. **Conditional expectation.**

Definition 5.6. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a discrete random vector. The conditional expectation of $f(\mathbf{X})$ given $X_{i_1}, \ldots, X_{i_k}$ is defined to be the function (here, $f : \mathbb{R} \to \mathbb{C}$ is any function)

$$\begin{aligned} (x_{i_1}, \ldots, x_{i_k}) &\mapsto E[f(\mathbf{X}) | X_{i_1} = x_{i_1}, \ldots, X_{i_k} = x_{i_k}] \\ &= \sum_{\substack{x_j \in \mathbb{R} \\ j \notin \{i_1, \ldots, i_k\}}} f(x_1, \ldots, x_n) \frac{p(x_1, \ldots, x_n)}{\sum_{\substack{x_j \in \mathbb{R} \\ j \notin \{i_1, \ldots, i_k\}}} p(x_1, x_2, \ldots, x_n)}. \end{aligned}$$

This is a function of the random variables $X_{i_1}, \dots, X_{i_k}$, so we will often just write $\mathbb{E}[f(\mathbf{X})|X_{i_1}, \dots, X_{i_k}]$. The idea is to take expectation with respect to the probability measure obtained by conditioning on the value of $X_{i_1}, \dots, X_{i_k}$. If $\mathbf{X}$ is instead continuous, then

$$(x_{i_1}, \dots, x_{i_k}) \mapsto E[f(\mathbf{X})|X_{i_1} = x_{i_1}, \dots, X_{i_k} = x_{i_k}]$$
$$= \int_{\substack{x_j \in \mathbb{R} \\ j \notin \{i_1, \dots, i_k\}}} f(x_1, \dots, x_n) \frac{p(x_1, \dots, x_n)}{\int_{\substack{x_j \in \mathbb{R} \\ j \notin \{i_1, \dots, i_k\}}} p(x_1, x_2, \dots, x_n) \prod_{j \notin \{i_1, \dots, i_k\}} dx_j} \prod_{j \notin \{i_1, \dots, i_k\}} dx_j$$

Lemma 5.7. *(1) Suppose $f(\mathbf{X}) = f(X_{i_1}, \dots, X_{i_k})$, i.e. f depends only on $X_{i_1}, \dots, X_{i_k}$. Then $\mathbb{E}[f(\mathbf{X})|X_{i_1}, \dots, X_{i_k}] = f(X_{i_1}, \dots, X_{i_k})$. In particular, conditional expectation does nothing to functions that depend only on what we condition on. More generally, for any other function g, we have $\mathbb{E}[f(\mathbf{X})g(\mathbf{X})|X_{i_1}, \dots, X_{i_k}] = f(\mathbf{X})\mathbb{E}[g(\mathbf{X})|X_{i_1}, \dots, X_{i_k}]$.*
(2) We have $\mathbb{E}[f(\mathbf{X})+g(\mathbf{X})|X_{i_1}, \dots, X_{i_k}] = \mathbb{E}[f(\mathbf{X})|X_{i_1}, \dots, X_{i_k}]+\mathbb{E}[g(\mathbf{X})|X_{i_1}, \dots, X_{i_k}]$ and $\mathbb{E}[cf(\mathbf{X})|X_{i_1}, \dots, X_{i_k}] = c\mathbb{E}[f(\mathbf{X})|X_{i_1}, \dots, X_{i_k}]$ for any $c \in \mathbb{R}$ deterministic.
(3) All of the lemmas (like Hölder, Cauchy-Schwarz, Jensen, etc.) hold for conditional expectation.
(4) (Law of iterated/total expectation). We have

$$\mathbb{E}\left\{\mathbb{E}[f(\mathbf{X})|X_{i_1}, \dots, X_{i_k}]\right\} = \mathbb{E}[f(\mathbf{X})].$$

(5) Suppose $f(\mathbf{X}) = f(X_{j_1}, \dots, X_{j_\ell})$, and $X_{j_1}, \dots, X_{j_\ell}$ are each jointly independent of $X_{i_1}, \dots, X_{i_k}$. Then $\mathbb{E}[f(\mathbf{X})|X_{i_1}, \dots, X_{i_k}] = \mathbb{E}[f(\mathbf{X})]$.

Example 5.8. Recall the first example with independent Bernoulli's. For any function $f : \mathbb{R} \to \mathbb{R}$, we have $\mathbb{E}[f(X)|Y] = \mathbb{E}[f(X)]$ by point (5) in the lemma. On the other hand, take the second example with identical Bernoulli's. In this case, for any function $f : \mathbb{R} \to \mathbb{R}$, we have $\mathbb{E}[f(X)|Y] = \mathbb{E}[f(Y)|Y] = f(Y)$ by point (1) in the lemma. Now, if you had a pair of Bernoulli's such that $X = Y$ with probability q and $X \neq Y$ with probability $1 - q$, then $\mathbb{E}[f(X)|Y] = qf(Y) + (1 - q)f(Z(Y))$, where $Z(Y) = 0$ if $Y = 1$ and $Z(Y) = 1$ if $Y = 0$.

Example 5.9. Recall the Gaussian example, where $Y \sim N(X, 1)$. We have $\mathbb{E}[Y|X] = X$, since the mean of Y is X (which is deterministic once we condition on it). By the law of iterated expectation, we can also compute $\mathbb{E}[Y] = \mathbb{E}\{\mathbb{E}[Y|X]\} = \mathbb{E}X$.

5.4. **Martingales.**

Definition 5.10. Suppose $(X_n)_{n \geqslant 1}$ is a sequence of random variables. We say the sequence $(M_N)_{N \geqslant 0}$ is a *martingale* with respect to the filtration generated by $(X_n)_{n \geqslant 1}$ if:

- For any $N \geqslant 0$, we have that M_N is a function of $X_1, \dots, X_N$ only.
- For any $N \geqslant 0$, we have $\mathbb{E}[M_{N+1}|X_1, \dots, X_N] = M_N$.

In the case where $N = 0$, then we identify $X_1, \dots, X_N$ with the empty set.

Lemma 5.11. *Suppose M_N is a martingale with respect to the filtration generated by $(X_n)_{n \geqslant 1}$. Then $\mathbb{E}M_N = M_0$ for any deterministic time.*

Proof. We have $\mathbb{E}[M_N] = \mathbb{E}\{\mathbb{E}[M_N|X_1, \dots, X_{N-1}]\} = \mathbb{E}M_{N-1}$. If we proceed inductively, we conclude. □

Example 5.12 (Symmetric simple random walk)**.** Suppose $X_n \overset{i.i.d.}{\sim} \mathrm{Bern}(\frac{1}{2})$, and define $Y_n = 1$ if $X_n = 1$ and $Y_n = -1$ if $X_n = 0$. In other words, we have $Y_n = (-1)^{1+X_n}$. Then the sequence $M_N = Y_1 + \ldots + Y_N$ (with $M_0 = 0$, though this initial value does not matter) is a martingale with respect to the filtration generated by $X_1, \ldots, X_N$. To check this, we first note that M_N is clearly a function of just $X_1, \ldots, X_N$. Next, we have

$$\begin{aligned}\mathbb{E}[M_{N+1}|X_1, \ldots, X_N] &= \mathbb{E}[M_N + Y_{N+1}|X_1, \ldots, X_N] \\ &= \mathbb{E}[M_N|X_1, \ldots, X_N] + \mathbb{E}[Y_{N+1}|X_1, \ldots, X_N] \\ &= M_N + \mathbb{E}[Y_{N+1}] = M_N.\end{aligned}$$

Example 5.13 (Biased simple random walk)**.** Suppose now that $X_n \overset{i.i.d.}{\sim} \mathrm{Bern}(p)$ for $p \neq 0, \frac{1}{2}, 1$. Define $W_n = X_n - p$. Then $M_N = W_1 + \ldots + W_N$ is a martingale as well.

Definition 5.14. Consider a sequence of random variables $(X_n)_{n\geqslant 1}$. A *stopping time* is a random variable τ valued in non-negative integers such that for any $n \geqslant 0$, if we condition on $X_1, \ldots, X_n$, then the indicator function $\mathbf{1}_{\tau\leqslant n}$ is deterministic.

Example 5.15. Take either simple random walk model. For any subset $A \subseteq \mathbb{R}$, the random variable $\tau = \inf\{N \geqslant 0 : M_N \in A\}$ is a stopping time. Indeed, if we condition on $X_1, \ldots, X_N$, then we know M_N, and in particular, we know if $\tau \leqslant N$ or not.

On the other hand, if we let τ_{not} be the last time that $M_N \in [-10, 10]$, for example, this is not a stopping time. Indeed, if we condition on $X_1, \ldots, X_N$, we do not know if $\tau_{\mathrm{not}} \leqslant N$; this would imply some knowledge about the future.

Theorem 5.16 (Doob's optional stopping theorem)**.** *Let M_N be a martingale with respect to a filtration generated by $(X_n)_{n\geqslant 1}$, and suppose τ is a stopping time such that at least one of the following hold:*

- *$\tau \leqslant C$ for some deterministic constant $C > 0$ with probability* 1.
- *We have $\sup_{N\leqslant\tau} |M_N| < \infty$.*
- *$\mathbb{E}\tau < \infty$ and $\sup_{N\leqslant\tau} |M_{N+1} - M_N| < \infty$.*

Then the process $M_{N\wedge\tau} := M_{\min(N,\tau)}$ is a martingale with respect to the same filtration, and $\mathbb{E}M_\tau = M_0$.

Example 5.17. Take the symmetric simple random walk (and assume $M_N = 0$). Let τ be the first time that $M_N = -a$ or $M_N = b$ (where $a, b > 0$ are deterministic integers). This is a stopping time as we explained earlier. Moreover, $|M_N| \leqslant \max(a, b) =: a \vee b$ for all $N \leqslant \tau$. Thus, by Doob's optional stopping, we know that $M_{N\wedge\tau}$ is a martingale. In particular, for any $N \geqslant 1$, we have $\mathbb{E}M_{N\wedge\tau} = \mathbb{E}M_0 = 0$.

Now, here comes a little finessing. We claim that as we send $N \to \infty$, then $\mathbb{E}M_{N\wedge\tau} \to \mathbb{E}M_\tau$. This is true even for the biased simple random walk. To see this, note that $|\mathbb{E}M_{N\wedge\tau} - \mathbb{E}M_\tau| \leqslant \sup_{k\leqslant\tau} |M_k|\mathbb{P}[\tau > N] \leqslant C\mathbb{P}[\tau > N]$ for some $C > 0$. But if $\tau > N$, then there cannot be an occurrence of $a + b$ up steps in the M process before time N (one can argue with down steps as well). The occurrence of $a + b$ up steps in a sequence of $a + b$ many total steps happens with strictly positive probability, say q. Thus, the probability that $\tau > N$ is at most $q^{N/(a+b)}$. This goes to 0 as $N \to \infty$. Thus, we

deduce $\mathbb{E}M_{N\wedge\tau} - \mathbb{E}M_\tau \to 0$ as $N \to \infty$. Ultimately, we get $\mathbb{E}M_\tau = 0$. Now, note

$$\mathbb{E}M_\tau = b\mathbb{P}[M_\tau = b] - a\mathbb{P}[M_\tau = -a] = 0.$$

Also, $\mathbb{P}[M_\tau = b] = 1 - \mathbb{P}[M_\tau = -a]$. From this, we get $\mathbb{P}[M_\tau = -a] = \frac{b}{b+a}$. Note that as $b \to \infty$, this approaches 1. Make sure this makes intuitive sense! Also, why does this argument break down for the biased simple random walk?

5.5. **A little fun fact about Gaussian tail probabilities.**

Lemma 5.18. *A random variable X satisfies $\mathbb{P}[|X| \geqslant C] \leqslant \exp\{-KC^2\}$ for all $C \geqslant 0$ (for some constant $K > 0$) if and only if $\mathbb{E}|X|^{2q} \leqslant C_1(2q-1)!!C_2^q$ for all $q \geqslant 1$, where $C_1, C_2 > 0$ are fixed constants. Moreover, we have $C_2 \leqslant K^{-1}$.*

Proof. We prove one direction; the other is on the HW (it is spelled out; actually, one can even assume $\mathbb{P}[|X| \geqslant C] \leqslant L\exp\{-KC^2\}$ for some constant $L \geqslant 0$). Suppose that $\mathbb{E}|X|^{2q} \leqslant C_1(2q-1)!!C_2^q$ for all $q \geqslant 1$, where $C_1, C_2 > 0$ are fixed constants. By Chebyshev, we have

$$\mathbb{P}[|X| \geqslant C] \leqslant e^{-\lambda C^2}\mathbb{E}e^{\lambda|X|^2},$$

where $\lambda > 0$ will be chosen shortly. By Taylor expansion, we have

$$\mathbb{E}e^{\lambda|X|^2} = \sum_{k=0}^{\infty} \frac{\lambda^k \mathbb{E}|X|^{2k}}{k!} \leqslant C_1 \sum_{k=0}^{\infty} \frac{(2k-1)!!\lambda^k C_2^k}{k!}.$$

This is $\leqslant \mathbb{E}e^{\lambda|Z|^2}$, where $Z \sim N(0, \sigma_{C_2}^2)$ is a Gaussian of variance depending on C_2. A simple integration (see me in office hours if you want to have this spelled out) shows that this is finite if λ is sufficiently small depending only on C_2. □

5.6. **Azuma's inequality and Doob's maximal inequality.**

Lemma 5.19. *Suppose that M_N is a martingale with respect to a filtration generated by $(X_n)_{n\geqslant 1}$. Suppose that $\sup_{N\geqslant 0}|M_{N+1} - M_N| \leqslant C$ for some deterministic $C < \infty$. Then there exists $K > 0$ such that for any $\varepsilon > 0$, we have*

$$\mathbb{P}\left[|M_N| \geqslant \varepsilon\right] \leqslant \exp\left\{-\frac{K\varepsilon^2}{NC^2}\right\}.$$

In particular, we have $\mathbb{E}|M_N|^{2q} \leqslant C_1(2q-1)!!N^qC_2^q$ for all $q \geqslant 1$ and for some constants $C_1, C_2 > 0$.

Lemma 5.20. *Suppose that M_N is a martingale with respect to a filtration generated by $(X_n)_{n\geqslant 1}$. Let $Z_N := \max_{0\leqslant k\leqslant N}|M_N|$. Then for any $p > 1$, we have $\mathbb{E}|Z_N|^p \leqslant (\frac{p}{p-1})^p\mathbb{E}|M_N|^p$.*

Both inequalities require the martingale structure and, in particular, the use of an appropriate stopping time! We will see these next week.

6. Week 6, starting Tue. Feb. 26, 2024

6.1. **Proof of Azuma's martingale inequality.** Assume that $N \geqslant 2$. By the Chebyshev inequality, we have

$$\mathbb{P}[M_N \geqslant \varepsilon] \leqslant e^{-\lambda\varepsilon}\mathbb{E}\exp[\lambda M_N] = e^{-\lambda\varepsilon}\mathbb{E}\{\exp(\lambda M_{N-1})\mathbb{E}[e^{\lambda(M_N - M_{N-1})}|X_1,\dots,X_{N-1}]\}.$$

By Taylor expansion, as long as λ is bounded above in absolute value independently of N, we have

$$\begin{aligned} e^{\lambda(M_N - M_{N-1})} &\leqslant 1 + \lambda(M_N - M_{N-1}) + \lambda^2 K_C |M_N - M_{N-1}|^2 \\ &\leqslant 1 + \lambda(M_N - M_{N-1}) + \lambda^2 K_C C^2. \end{aligned}$$

We get $\mathbb{E}[e^{\lambda(M_N - M_{N-1})}|X_1,\dots,X_{N-1}] \leqslant 1 + \lambda^2 K_C C^2 \leqslant \exp[\lambda^2 K_C C^2]$. Thus,

$$e^{-\lambda\varepsilon}\mathbb{E}\exp[\lambda M_N] \leqslant e^{-\lambda\varepsilon}e^{\lambda^2 K_C C^2}\mathbb{E}\exp[\lambda M_{N-1}].$$

Continuing inductively, we get

$$\mathbb{P}[M_N \geqslant \varepsilon] \leqslant e^{-\lambda\varepsilon}e^{N\lambda^2 K_C C^2}.$$

Now, choose $\lambda = \frac{\varepsilon}{LN}$, where L is a large constant depending only on C such that $L \geqslant 10K_C C^2$, for example. On the other hand, if M_N is a martingale, then $-M_N$ is a martingale, so the same argument shows

$$\mathbb{P}[M_N \leqslant -\varepsilon] \leqslant e^{-\lambda\varepsilon}e^{N\lambda^2 K_C C^2}.$$

Thus, by a union bound, we have

$$\mathbb{P}\left[|M_N| \geqslant \varepsilon\right] \leqslant \mathbb{P}[M_N \geqslant \varepsilon] + \mathbb{P}[M_N \leqslant -\varepsilon] \leqslant 2\exp\left\{-\frac{K\varepsilon^2}{NC^2}\right\}.$$

On the HW, you showed that this implies the moment bounds in Lemma 5.18. Lemma 5.18 then implies the previous estimate but without the 2 on the RHS. □

6.2. **Proof of Doob's maximal inequality.** We will assume that $M_0 = 0$; otherwise, just replace M_N by $M_N - M_0$. Fix $t > 0$, and let $\tau_t := \inf\{k \geqslant 0 : |M_k| \geqslant t\} \wedge N$ be the minimum of N and the first time k that $|M_k| \geqslant t$. Note that the event $\{Z_N \geqslant t\}$ is equal to the event $\{|M_\tau| \geqslant t\}$. Thus,

$$\mathbb{P}[Z_N \geqslant t] = \mathbb{P}[|M_\tau| \geqslant t] \leqslant \frac{\mathbb{E}[\mathbf{1}_{|M_\tau|\geqslant t}|M_\tau|^p]}{t^p}.$$

Note that if $\mathbf{M}_k$ is a martingale, then $|\mathbf{M}_k|^p = |\mathbb{E}[\mathbf{M}_\ell|X_1,\dots,X_k]|^p \leqslant \mathbb{E}[|\mathbf{M}_\ell|^p|X_1,\dots,X_k]$ for all $k \leqslant \ell$ by Jensen. Apply this to $\mathbf{M}_k = M_{\tau\wedge k}$, which is a martingale by Doob's optional stopping. We deduce

$$\mathbb{E}[\mathbf{1}_{|M_\tau|\geqslant t}|M_\tau|^p] \leqslant \mathbb{E}[\mathbf{1}_{|M_\tau|\geqslant t}|M_N|^p].$$

This proves, for any p, that

$$\mathbb{P}[Z_N \geqslant t] \leqslant t^{-p}\mathbb{E}[\mathbf{1}_{|M_\tau|\geqslant t}|M_N|^p] = t^{-p}\mathbb{E}[\mathbf{1}_{Z_N\geqslant t}|M_N|^p].$$

We now need a lemma. It is very similar to the layer cake formula from towards the beginning of the semester.

Lemma 6.1. *Let X be a random variable. Then for any $q \geqslant 1$, we have*

$$\mathbb{E}[|X|^q] = q\int_0^\infty t^{q-1}\mathbb{P}[|X| > t]\mathrm{d}t.$$

Proof. We focus on the case where X is a continuous random variable, since the discrete random variable case was on the HW. By definition, we have

$$\mathbb{E}[|X|^q] = \int_{\mathbb{R}} |x|^q p(x)\mathrm{d}x = \int_0^\infty x^q p(x)\mathrm{d}x + \int_0^\infty x^q p(-x)\mathrm{d}x.$$

Now, write $p(x) = -\frac{\mathrm{d}}{\mathrm{d}x}\int_x^\infty p(u)\mathrm{d}u$. Using this and integration-by-parts, we have

$$\int_0^\infty x^q p(x)\mathrm{d}x = \int_0^\infty qx^{q-1}\int_x^\infty p(u)\mathrm{d}u.$$

By the same token, if we write $p(-x) = -\frac{\mathrm{d}}{\mathrm{d}x}\int_{-\infty}^{-x} p(u)\mathrm{d}u$, we have

$$\int_0^\infty x^q p(-x)\mathrm{d}x = \int_0^\infty qx^{q-1}\int_{-\infty}^{-x} p(u)\mathrm{d}u.$$

But, notice that $\int_x^\infty p(u)\mathrm{d}u + \int_{-\infty}^{-x} p(u)\mathrm{d}u = \mathbb{P}[X \geqslant x] + \mathbb{P}[X \leqslant -x] = \mathbb{P}[|X| \geqslant x]$. The claim now follows. □

Now, given the layer cake formula and the probability bound from before, we have

$$\begin{aligned}\mathbb{E}[|Z_N|^p] &= p\int_0^\infty t^{p-1}\mathbb{P}[Z_N \geqslant t]\mathrm{d}t \\ &\leqslant p\int_0^\infty t^{p-2}\mathbb{E}[|M_N|\mathbf{1}_{Z_N \geqslant t}]\mathrm{d}t.\end{aligned}$$

By Hölder, we have $\mathbb{E}[|M_N|\mathbf{1}_{Z_N\geqslant t}] \leqslant (\mathbb{E}[|M_N|^p])^{\frac{1}{p}}\mathbb{P}[Z_N \geqslant t]^{\frac{p}{p-1}} \leqslant (\mathbb{E}[|M_N|^p])^{\frac{1}{p}}\mathbb{P}[Z_N \geqslant t]$, where the last bound follows because probabilities are valued in $[0,1]$, and $\frac{p}{p-1} \geqslant 1$ for all $p \geqslant 1$. Thus,

$$\begin{aligned}\mathbb{E}[|Z_N|^p] &\leqslant (\mathbb{E}[|M_N|^p])^{\frac{1}{p}} p\int_0^\infty t^{p-2}\mathbb{P}[Z_N \geqslant t]\mathrm{d}t \\ &= \frac{p}{p-1}(\mathbb{E}[|M_N|^p])^{\frac{1}{p}}\mathbb{E}|Z_N|^{p-1}.\end{aligned}$$

Again, by Hölder, we have $\mathbb{E}|Z_N|^{p-1} \leqslant (\mathbb{E}|Z_N|^p)^{\frac{p-1}{p}}$. Moving this to the LHS of the previous inequality and raising both sides to the p-th power finishes the proof. □

7. Week 7, starting Tue. Mar. 19, 2024

7.1. Convergence in distribution.

Definition 7.1. Let X be a random variable. We say $\{X_n\}_{n=1}^\infty$ *converges in distribution* to X if for any $a \leqslant b$ fixed (independent of n), we have $\mathbb{P}[a \leqslant X_n \leqslant b] \to \mathbb{P}[a \leqslant X \leqslant b]$ as $n \to \infty$.

Theorem 7.2 (Levy's continuity theorem)**.** *The following are equivalent.*

(1) $\{X_n\}_{n=1}^\infty$ converges in distribution to X
(2) For any $\xi \in \mathbb{R}$, we have $\mathbb{E}e^{i\xi X_n} \to \mathbb{E}e^{i\xi X}$.
(3) For any smooth, compactly supported function $f : \mathbb{R} \to \mathbb{R}$ (i.e. it is smooth and it vanishes outside a compact subset of $\mathbb{R}$), we have $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$.

We will not prove this theorem, since it belongs to a domain of mathematics called "Fourier analysis", but it is very useful in probability theory. Note that $f(x) = |x|^p$ is *not* compact supported for any $p > 0$, so that this is not really saying anything very strong.

Example 7.3. Suppose $X_n = X$ for all n. Then $X_n \to X$ in distribution as $n \to \infty$ clearly.

Example 7.4. Suppose $X_n = Y$ for all n, where $Y \sim N(0,1)$, and suppose $X = -Y$. Note that $X \sim N(0,1)$. Then clearly X_n, X take very different values for all n, but we claim $X_n \to X$ in distribution! Indeed, since $X_n \sim N(0,1)$ for every n, we have $\mathbb{P}[a \leqslant X_n \leqslant b] = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{u^2}{2}} \mathrm{d}u$. But this is also true for X in place of X_n, because $X \sim N(0,1)$. This example illustrates the fact that convergence in distribution, as its name suggests, depends only on the distribution of the random variable.

Example 7.5. Suppose X_n are i.i.d. with $\mathbb{E}X_n = 0$ and $\mathbb{E}X_n^2 < \infty$. Define $S_N = N^{-1/2}(X_1 + \ldots + X_N)$. Then the law of large numbers implies that $S_N \to 0$ in distribution. To check this, one can use Levy's continuity theorem; it suffices to show that for any $\xi \in \mathbb{R}$ fixed, we have $\mathbb{E}e^{i\xi S_N} \to \mathbb{E}e^{i\xi 0} = 1$. To see this, we write

$$\mathbb{E}e^{i\xi S_N} - 1 = \mathbb{E}(e^{i\xi S_N} - 1)\mathbf{1}_{|S_N| \geqslant N^{-1/3}} + \mathbb{E}(e^{i\xi S_N} - 1)\mathbf{1}_{|S_N| < N^{-1/3}}.$$

For the second term, note that if $|S_N| < N^{-1/3}$, then calculus (e.g. Taylor series) implies $|e^{i\xi S_N} - 1| \leqslant |S_N| \leqslant N^{-1/3} \to 0$. Thus, the second term on the RHS vanishes as $N \to \infty$. For the first term, note that $|e^{i\xi S_N} - 1| \leqslant 2$, since $|e^{i\xi S_N}| \leqslant 1$. Thus,

$$\mathbb{E}|e^{i\xi S_N} - 1|\mathbf{1}_{|S_N| \geqslant N^{-1/3}} \leqslant 2\mathbb{P}(|S_N| \geqslant N^{-1/3}) \leqslant 2N^{2/3}\mathbb{E}|S_N|^2 \lesssim 2N^{-1/3},$$

where the second-to-last bound follows by Chebyshev and the last bound follows by our proof of the law of large numbers. Since this vanishes as $N \to \infty$, we see that $|\mathbb{E}e^{i\xi S_N} - 1| \to 0$ as $N \to \infty$, which is what we wanted.

In principle, convergence in distribution is a very weak statement; restricting to expectations of smooth, compactly supported functions is a very restrictive thing to do. The lemma below tells us when we can relax this condition in the context of moments.

Lemma 7.6. *Let $\{X_n\}_{n=1}^\infty$ be such that $\sup_{n \geqslant 1} \mathbb{E}|X_n|^p < \infty$ for some fixed $p > 0$ and $X_n \to X$ in distribution. Then for any $0 \leqslant r < p$, we have $\mathbb{E}|X_n|^r \to \mathbb{E}|X|^r$ and $\mathbb{E}X_n^r \to \mathbb{E}X^r$.*

We will also not give the proof of this, because it requires a number of tools from measure theory (i.e. material covered in Math 114), but again, it is very useful to know.

Example 7.7. Here is a counterexample illustrating why we cannot have $r = p$ in the previous lemma; it will not be important to understand it, but it is maybe worth at least looking at. Let the probability space be $[0,1]$, and let $X_n(u) = n^{1/2}\mathbf{1}_{u \in [0, n^{-1}]}$. (Remember that random variables are just functions on probability spaces!) We claim that $X_n \to 0$ in distribution. To see this, we use Levy's continuity theorem; it suffices to show $\mathbb{E}e^{i\xi X_n} \to 1$ as $n \to \infty$. To prove this, we have

$$\mathbb{E}e^{i\xi X_n} = \int_0^1 e^{i\xi X_n(u)} \mathrm{d}u = \int_0^{1/n} e^{i\xi n^{1/2}} \mathrm{d}u + \int_{1/n}^1 \mathrm{d}u = \int_0^{1/n} e^{i\xi n^{1/2}} \mathrm{d}u + (1 - n^{-1}).$$

For the first term, again note that $|e^{i\xi n^{1/2}}| = 1$, so the first term is $\leqslant n^{-1} \to 0$ in absolute value. Next, note that $\mathbb{E}|X_n|^2 = \int_0^1 |X_n(u)|^2 \mathrm{d}u = \int_0^{1/n} n \mathrm{d}u = 1$. Thus, we clearly do not have $\mathbb{E}|X_n|^2 \to \mathbb{E}0^2$, even though $\sup_n \mathbb{E}|X_n|^2 < \infty$ and $X_n \to 0$ in distribution.

7.2. **Central limit theorem.**

Theorem 7.8. *Suppose $\{X_i\}_{i=1}^\infty$ are i.i.d. random variables with $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = 1$. Define $S_N = N^{-1/2}(X_1 + \ldots + X_N)$. Then $S_N \to G$ in distribution, where $G \sim N(0,1)$.*

Proof. There are a number of different proofs of quite different flavor; you will see glimpses of two in the HW. We give one based on the Fourier transform; it is clean, but it requires the i.i.d. assumption (one does not even need independence in full generality of the CLT). For convenience, we will also assume that $\mathbb{E}|X_i|^3 < \infty$, though this is not necessary and can be removed by being a lot of more tedious and careful.

By Levy's continuity theorem, it suffices to show that for any $\xi \in \mathbb{R}$, we have

$$\mathbb{E}e^{i\xi S_N} \to \mathbb{E}e^{i\xi G} = e^{-\frac{\xi^2}{2}};$$

the last identity was proven on an earlier HW (for $\xi = -i\rho$ with $\rho \in \mathbb{R}$). First, by independence, we have

$$\mathbb{E}e^{i\xi S_N} = \mathbb{E}\prod_{j=1}^N e^{iN^{-1/2}\xi X_j} = \prod_{j=1}^N \mathbb{E}e^{iN^{-1/2}\xi X_j} = \chi(iN^{-1/2}\xi)^N.$$

(The last identity is just setting notation $\chi(i\rho) = \mathbb{E}e^{i\rho X_j}$; since X_j are i.i.d., this does not depend on j.) Now, by Taylor expansion, we have

$$\chi(iN^{-1/2}\xi) = \chi(0) + iN^{-\frac{1}{2}}\xi\chi'(0) - \frac{1}{2}N^{-1}\xi^2\chi''(0) + \mathcal{E},$$

where

$$|\mathcal{E}| \leqslant N^{-3/2}|\xi|^3 \sup_{u\in\mathbb{R}} |\chi'''(iu)|.$$

It is not hard to see that $\chi(0) = \mathbb{E}e^0 = 1$ and $\chi'(0) = \mathbb{E}X_j = 0$ and $\chi''(0) = \mathbb{E}X_j^2 = 1$ and $\chi'''(iu) = -i\mathbb{E}X_j^3 e^{iuX_j}$. By assumption on the third moment and by using $|e^{iuX_j}| = 1$, we get $|\chi'''(iu)| < \infty$. Thus, we have

$$\mathbb{E}e^{i\xi S_N} = \left(1 - \frac{1}{2}N^{-1}\xi^2 + O(N^{-3/2}|\xi|^3)\right)^N,$$

where $O(\cdot)$ means something which is bounded above by some constant times $\cdot$. It is now standard calculus to check that (since $N^{-3/2} \ll N^{-1}$) $\mathbb{E}e^{i\xi S_N} \to e^{-\frac{1}{2}\xi^2}$ as $N \to \infty$. □

Corollary 7.9. *We checked in the law of large numbers proof that $\sup_{N\geqslant 1} \mathbb{E}|S_N|^2 < \infty$. Thus, we have $\mathbb{E}|S_N|^r \to \mathbb{E}|G|^r$ for any $0 \leqslant r < 2$ and $\mathbb{E}S_N^r \to \mathbb{E}G^r$ for any $0 \leqslant r < 2$; here $G \sim N(0,1)$.*

Example 7.10. Let $X_i \sim \mathrm{Bern}(\frac{1}{2})$ be i.i.d., and define $S_N = X_1 + \ldots + X_N$. We expect that S_N is about size N up to some constant. So, here is a reasonable question.

Approximate

$$\mathbb{P}\left[\frac{1}{2}N + aN^{1/2} \leqslant S_N \leqslant \frac{1}{2}N + bN^{1/2}\right],$$

where $0 \leqslant a \leqslant b$. Clearly, this is the same as

$$\mathbb{P}\left[2a \leqslant \frac{2S_N - N}{N^{1/2}} \leqslant 2b\right].$$

Now, note that $2S_N - N = (2X_1 - 1) + \ldots + (2X_N - 1)$. Moreover, note that $Y_j = 2X_j - 1$ satisfies $\mathbb{E}Y_j = 0$ and $\mathbb{E}Y_j^2 = 1$. Thus, the CLT says that $N^{-1/2}(2S_N - N) \to G$, where $G \sim N(0,1)$. In particular, the previous probability converges to $\mathbb{P}[2a \leqslant G \leqslant 2b] = \int_{2a}^{2b} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \mathrm{d}u$, which although is not obvious to compute in general, is much easier.

7.3. **Lindeberg exchange method.** Here is another proof of the CLT. Take any smooth, compactly supported function $f : \mathbb{R} \to \mathbb{R}$. The setting is the following.

- $\{X_i\}_{i=1}^{\infty}$ are i.i.d. random variables with $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = 1$ and $\mathbb{E}|X_i|^3 < \infty$.
- $S_N = N^{-1/2}(X_1 + \ldots X_N)$.
- $\{G_i\}_{i=1}^{\infty}$ are i.i.d. $N(0,1)$. Everything is independent from $\{X_i\}_{i=1}^{\infty}$.
- $Z_N = N^{-1/2}(G_1 + \ldots + G_N)$.
- Define $Y_N = N^{-1/2}(G_1 + X_2 + \ldots + X_N)$; it is S_N but X_1 is swapped with G_1.

The idea of this method is a little odd, as in it is a little weird that it works but it turns out to be very useful. It also says more than the CLT, though we discuss this point later. We aim to show

$$|\mathbb{E}f(S_N) - \mathbb{E}f(Y_N)| \leqslant CN^{-3/2}.$$

In words, the price to exchange X_1 for G_1 is order $N^{-3/2}$ at most. (Actually, if $\mathbb{E}X_i^3 = 0$, then it is at most order N^{-2}. If $\mathbb{E}X_i^4 = 3$, then it is at most order $N^{-5/2}$; the more moments one matches with those of $N(0,1)$, the better this price becomes, at least for the buyer.) To see why this is useful, we can then replace X_2 by G_2 for a price of $\leqslant CN^{-3/2}$, and so forth, eventually replacing S_N by Z_N for a total cost of $\leqslant N \times CN^{-3/2} = CN^{-1/2} \to 0$. But $Z_N \sim N(0,1)$, so the CLT follows.

To prove the bound in the previous display, we Taylor expand:

$$f(S_N) = f(Y_N) + N^{-1/2}(X_1 - G_1)f'(Y_N) + \frac{1}{2}N^{-1}(X_1 - G_1)^2 f''(Y_N) + O(N^{-3/2}).$$

Next, upon setting $\tilde{Y}_N = N^{-1/2}(X_2 + \ldots + X_N)$, we have

$$N^{-1/2}(X_1 - G_1)f'(Y_N) = N^{-1/2}(X_1 - G_1)f'(\tilde{Y}_N) + N^{-1}G_1(X_1 - G_1)f''(\tilde{Y}_N) + O(N^{-3/2}).$$

Taking expectation and using independence and $\mathbb{E}X_1 = 0$ and $\mathbb{E}G_1^2 = 1$, we have

$$\mathbb{E}N^{-\frac{1}{2}}(X_1 - G_1)f'(Y_N) = -N^{-1}\mathbb{E}G_1^2 f''(\tilde{Y}_N) = N^{-1}\mathbb{E}f''(\tilde{Y}_N).$$

Next, we also have

$$\frac{1}{2}N^{-1}(X_1 - G_1)^2 f''(Y_N) = \frac{1}{2}N^{-1}(X_1 - G_1)^2 f''(\tilde{Y}_N) + O(N^{-3/2})$$
$$= \frac{1}{2}N^{-1}(X_1^2 + G_1^2)f''(\tilde{Y}_N) - N^{-1}X_1 G_1 f''(\tilde{Y}_N) + O(N^{-3/2}).$$

We again take expectation and use independence and $\mathbb{E}X_1 = 0$ and $\mathbb{E}X_1^2 = 1$ and $\mathbb{E}G_1^2 = 1$ to get

$$\frac{1}{2}N^{-1}\mathbb{E}(X_1 - G_1)^2 f''(Y_N) = N^{-1}\mathbb{E}f''(\tilde{Y}_N) + O(N^{-3/2}).$$

Putting everything together, we get the desired bound. Now, note that this argument is not just giving the CLT; it is saying that one can essentially replace X_j individually with any A_j such that $\mathbb{E}A_j = 0$ and $\mathbb{E}A_j^2 = 1$ and $\mathbb{E}|A_j|^3 < \infty$. Using the same argument, one can also prove the following (the multivariable CLT).

Theorem 7.11. *Let $\{\mathbf{X}_i\}_{i=1}^{\infty}$ be i.i.d. random vectors in $\mathbb{R}^{\mathrm{d}}$ such that the entries are i.i.d. and satisfy $\mathbb{E}\mathbf{X}_i(j) = 0$ and $\mathbb{E}\mathbf{X}_i(j)^2 = 1$ (and $\mathbb{E}|\mathbf{X}_i(j)|^3 < \infty$, though this is not crucial).*

Define $\mathbf{S}_N = N^{-1/2}(\mathbf{X}_1 + \ldots + \mathbf{X}_N)$. For any smooth, compactly supported function $f : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$, we have $\mathbb{E}f(\mathbf{S}_N) \to \mathbb{E}f(\mathbf{G})$, where $\mathbf{G} \sim N(0, \mathrm{Id}_{\mathrm{d}})$ is a d-dimensional Gaussian, i.e. its components are independent $N(0,1)$.

The idea is to replace the entries of each $\mathbf{X}_i$ one by one; one accumulates now $\mathrm{d}N$ many errors of the form $O(N^{-3/2})$. The independence of the entries of $\mathbf{X}$ is also unnecessary. In general, define the $\mathrm{d} \times \mathrm{d}$ covariance matrix

$$\Sigma_{jk} = \mathbb{E}\mathbf{X}_i(j)\mathbf{X}_i(k).$$

Since $\mathbf{X}_i$ are i.i.d., this does not depend on i. Then, if we drop the independence of entries assumption, we deduce that $\mathbb{E}f(\mathbf{S}_N) \to \mathbb{E}f(\mathbf{W})$, where $\mathbf{W} \sim N(0, \Sigma)$ is a Gaussian of dimension d with the correct covariance matrix. In particular, the CLT is really a statement about linear algebra in some sense.

7.4. **A brief word on Brownian motion.**

Lemma 7.12. *Suppose $\{\mathbf{X}^{(N)}\}_N$ is a sequence of random vectors in $\mathbb{R}^d$, and assume that $\mathbf{X}^{(N)} \to N(0, \Sigma)$ for some Σ in distribution. Let $\mathbf{A}$ be any fixed, deterministic matrix. Then $\mathbf{A}\mathbf{X}^{(N)}$ converges to $N(0, \Sigma_{\mathbf{A}})$ in distribution, where $\Sigma_{\mathbf{A}}$ is some matrix depending on $\Sigma, \mathbf{A}$ (it is equal to $\mathbf{A}\Sigma\mathbf{A}^*$, but this is not important).*

Proof. For any compactly supported function $F : \mathbb{R}^d \to \mathbb{R}$, we can consider the compactly supported function $F_{\mathbf{A}} : \mathbb{R}^d \to \mathbb{R}$ given by $F_{\mathbf{A}}(x) = F(\mathbf{A}x)$. Now, we have

$$\mathbb{E}F(\mathbf{A}\mathbf{X}^{(N)}) = \mathbb{E}F_{\mathbf{A}}(\mathbf{X}^{(N)}) \to \mathbb{E}F_{\mathbf{A}}(\mathbf{X}) = \mathbb{E}F(\mathbf{A}\mathbf{X}),$$

where $\mathbf{X} \sim N(0, \Sigma)$. It suffices to show that $\mathbf{AX} \sim N(0, \Sigma_{\mathbf{A}})$. To this end, we have

$$\begin{aligned}\mathbb{E}F(\mathbf{AG}) &= \int_{\mathbb{R}^d} F(\mathbf{Ax}) \frac{1}{(2\pi \det \Sigma)^{d/2}} \exp\left\{-\frac{\mathbf{x}^*\Sigma^{-1}\mathbf{x}}{2}\right\} d\mathbf{x} \\ &= \int_{\mathbb{R}^d} F(\mathbf{y}) \frac{1}{(2\pi \det \Sigma \det \mathbf{A}^2)^{d/2}} \exp\left\{-\frac{\mathbf{y}(\mathbf{A}\Sigma\mathbf{A}^*)^{-1}\mathbf{y}}{2}\right\} d\mathbf{y},\end{aligned}$$

which is just $\mathbb{E}F(\mathbf{W})$ with $\mathbf{W} \sim N(0, \mathbf{A}\Sigma\mathbf{A}^*)$. □

Theorem 7.13. *For each $k \geqslant 1$, let $\{X_{k,i}\}_{i=1}^{\infty}$ be a sequence of independent random variables such that $\mathbb{E}X_{k,i} = 0$ and $\mathbb{E}X_{k,i}^2 = 0$. Assume that $\{X_{k,i}\}_{k,i}$ are jointly independent.*

Now, define a process $\{\mathbf{B}_k^{(N)}\}_{k\geqslant 0}$ given by $\mathbf{B}_0^{(N)} = 0$ and

$$\mathbf{B}_k^{(N)} = \mathbf{B}_{k-1}^{(N)} + N^{-\frac{1}{2}} \sum_{i=1}^{N} X_{k,i}.$$

For any fixed integer $T \geqslant 1$, the vector $(\mathbf{B}_1^{(N)}, \dots, \mathbf{B}_T^{(N)})$ converges to $(\mathbf{B}_1, \dots, \mathbf{B}_T)$ in distribution as $N \to \infty$, where $(\mathbf{B}_1, \dots, \mathbf{B}_T) \sim N(0, \Sigma(T))$ with covariance matrix

$$\Sigma(T)_{j\ell} := \min(j, \ell).$$

Proof. We make another claim. For any $k \geqslant 1$, define $\mathbf{Z}_k^{(N)} := \mathbf{B}_k^{(N)} - \mathbf{B}_{k-1}^{(N)}$. For $k = 0$, define $\mathbf{Z}_0^{(N)} = \mathbf{B}_0^{(N)} = 0$. We claim that the vector $(\mathbf{Z}_1^{(N)}, \dots, \mathbf{Z}_T^{(N)})$ converges in distribution to $N(0, \mathrm{Id})$, i.e. a vector $(\mathbf{Z}_1, \dots, \mathbf{Z}_T)$ whose entries are independent standard $N(0,1)$ Gaussians. Let us prove this first. By definition,

$$\mathbf{Z}_k^{(N)} = \mathbf{B}_k^{(N)} - \mathbf{B}_{k-1}^{(N)} = N^{-\frac{1}{2}} \sum_{i=1}^{N} X_{k,i}.$$

By linearity of expectation, we know $\mathbb{E}\mathbf{Z}_k^{(N)} = 0$. We also know

$$\mathbb{E}|\mathbf{Z}_k^{(N)}|^2 = N^{-1} \sum_{i=1}^{N} \mathbb{E}X_{k,i}^2 + N^{-1} \sum_{i\neq j} \mathbb{E}X_{k,i}X_{k,j} = 1,$$

since $\mathbb{E}X_{k,i}^2 = 1$ and $X_{k,i}, X_{k,j}$ are independent mean-zero random variables. Moreover, we know that $\mathbf{Z}_1^{(N)}, \dots, \mathbf{Z}_T^{(N)}$ are jointly independent by assumption that $X_{k,i}$ are jointly independent. Thus, by the vector-valued central limit theorem, we know $(\mathbf{Z}_1^{(N)}, \dots, \mathbf{Z}_T^{(N)})$ converges in distribution to $N(0, \mathrm{Id}_T)$.

Let us see how this helps us conclude. We know that a linear transformation of a Gaussian vector is another Gaussian. Since $(\mathbf{B}_1^{(N)}, \dots, \mathbf{B}_T^{(N)})$ is a linear transformation of $(\mathbf{Z}_1^{(N)}, \dots, \mathbf{Z}_T^{(N)})$, we know that $(\mathbf{B}_1^{(N)}, \dots, \mathbf{B}_T^{(N)})$ must converge to $(\mathbf{B}_1, \dots, \mathbf{B}_T)$, which is a Gaussian $N(0, \Sigma(T))$ for some covariance matrix $\Sigma(T)$. Here, $\mathbf{B}_k - \mathbf{B}_{k-1} = \mathbf{Z}_k$ (again, with $\mathbf{B}_0 = 0$). The entries of this are just the covariances between entries of $(\mathbf{B}_1, \dots, \mathbf{B}_T)$. In particular, we have $\mathbb{E}\mathbf{B}_k = \mathbb{E}\sum_{i=1}^{k} \mathbf{Z}_i = 0$, and we have the following

(without loss of generality, assume $j \leqslant \ell$):

$$\Sigma(T)_{j\ell} = \mathbb{E}\mathbf{B}_j\mathbf{B}_\ell = \mathbb{E}\left[\sum_{\alpha=1}^{j}\mathbf{Z}_\alpha\sum_{\beta=1}^{\ell}\mathbf{Z}_\beta\right]$$
$$= \sum_{\alpha=1}^{j}\sum_{\beta=1}^{\ell}\mathbb{E}\mathbf{Z}_\alpha\mathbf{Z}_\beta = \sum_{\alpha=1}^{j}\mathbb{E}\mathbf{Z}_\alpha^2 = j,$$

where the last line follows because $\mathbf{Z}_i$ are independent $N(0,1)$. Thus, we know $\Sigma(T)_{j\ell} = \min(j,\ell)$, and we are done. □

Definition 7.14. We say that a random continuous function $\mathbf{B} : [0,\infty) \to \mathbb{R}$ *has the law of Brownian motion* if $\mathbf{B}(0) = 0$ and for any integer $k \geqslant 1$ and $0 < t_1 < \ldots < t_k < 0$, the vector

$$(\mathbf{B}(t_1), \ldots, \mathbf{B}(t_k))$$

has distribution given by $N(0, \Sigma[t_1, \ldots, t_k])$, where $\Sigma[t_1, \ldots, t_k]_{j\ell} = \min(t_j, t_\ell)$.

In principle, *we do not know if Brownian motion as defined above even exists!* It is really asking for an "infinite-dimensional vector", which is one problem on its own. We also stipulated that the function $\mathbf{B}$ is *continuous, which is not at all necessarily compatible with the distribution that we require for finitely-many time samples*. We will see later in this class that both issues can be resolved.

8. Week 8, starting Tue. Mar. 26, 2024

8.1. Introduction to Markov chains.

Definition 8.1. Fix a countable set S (this is called the "state space"). We say a process $\{X_n\}_{n=0}^{\infty}$ is a *Markov chain* if it satisfies the following *Markov condition*:

$$\mathbb{P}\left[X_{n+1} = s | X_0 = x_0, \ldots, X_n = x_n\right] = \mathbb{P}[X_{n+1} = s | X_n = x_n]$$

for all $n \geqslant 0$ and $x_0, x_1, \ldots, x_n, s \in S$. We say it is *homogeneous* or *time-homogeneous* if $\mathbb{P}[X_{n+1} = j | X_n = i] = \mathbb{P}[X_1 = j | X_0 = i]$ for all $n \geqslant 0$ and $i, j \in S$.

Example 8.2 (Simple random walk)**.** Fix $p \in [0,1]$, and let $\{Y_i\}_{i=1}^{\infty}$ be a sequence of independent random variables such that $\mathbb{P}[Y_i = \pm 1] = \frac{1}{2}$. We claim that the sequence $X_n = \sum_{i=1}^{n} Y_i$ (with $X_0 = 0$) is a time-homogeneous with state space $\mathbb{Z}$. To see this, note that

$$\begin{aligned}\mathbb{P}\left[X_{n+1} = s | X_0 = x_0, \ldots, X_n = x_n\right] &= \mathbb{P}\left[X_n + Y_{n+1} = s | X_0 = x_0, \ldots, X_n = x_n\right] \\ &= \mathbb{P}\left[Y_{n+1} = s - x_n | X_0 = x_0, \ldots, X_n = x_n\right] \\ &= \mathbb{P}\left[Y_{n+1} = s - x_n | X_n = x_n\right] \\ &= \mathbb{P}\left[X_n + Y_{n+1} = s | X_n = x_n\right] \\ &= \mathbb{P}[X_{n+1} = s | X_n = x_n].\end{aligned}$$

Intuitively, the next location of the random walk depends only on its current position, not its past steps. It is time-homogeneous because $\mathbb{P}[X_{n+1} = i | X_n = j] = \mathbb{P}[Y_{n+1} =$

$i - j | X_n = j] = \mathbb{P}[Y_{n+1} = i - j]$, and this probability is independent of $n \geqslant 0$ since Y_i are i.i.d.

Example 8.3. Suppose that $\{S_n\}_{n=0}^{\infty}$ is a *symmetric* simple random walk, and set $X_n := |S_n|$. This is a Markov chain. Here is an intuitive explanation why. Assume $a > 0$. If we know that $|S_n| = a$, then either $S_n = a$ or $S_n = -a$. If $S_n = a$, then it will jump to either $a - 1$ or $a + 1$ with equal probability. In this case, we know that $|S_n|$ will jump to either $a - 1$ or $a + 1$ with equal probability. If $S_n = -a$, then it will jump to $-a + 1$ or $-a - 1$ with equal probability, and thus $|S_n|$ will jump to either $a - 1$ or $a + 1$, again with equal probability. On the other hand, if $|S_n| = 0$, so that $S_n = 0$, then S_{n+1} must be 1 or -1, in which case $|S_{n+1}| = 1$, so that $|S_{n+1}|$ is deterministic once we condition on $|S_n| = 0$. The point is that conditioning on previous values $|S_k|$ for $k < n$ does not affect these probabilities.

Example 8.4. Suppose we roll a fair die repeatedly. Let $\{X_n\}_{n=0}^{\infty}$ be the largest number rolled in the first n trials. This *is* a Markov chain with state space $\{1, \ldots, 6\}$. To see why, we compute

$$\mathbb{P}\left[X_{n+1} = s | X_0 = x_0, \ldots, X_n = x_n\right] = \begin{cases} 0 & s < x_n \\ \mathbb{P}[X_{n+1} = s] & s \geqslant x_n \end{cases}$$

On the other hand, the same is true for $\mathbb{P}[X_{n+1} = s | X_n = x_n]$. It is also time-homogeneous if the rolls are independent. Indeed, $\mathbb{P}[X_{n+1} = s]$ is independent of $n \geqslant 0$ in this case.

Example 8.5. Here is a process which is *not* Markov. Suppose there are restaurants A, B, and C. Each night, Kevin chooses a restaurant to go to based on restaurants he has gone to in the past. Each night, Kevin takes the restaurant he went to the previous night, looks at the other two restaurants, and picks the one he has been to more often with probability $1/3$ and the one has been to less often with probability $2/3$ (if he has been to them an equal number of times, then its 50/50).

The process of restaurants $X_n \in \{A, B, C\}$ visited by Kevin is *not* Markov. Indeed, to determine which restaurant is favorable at night n, Kevin must know all of the previous values X_k for $k < n$ in general. If Kevin just chose one of the two other restaurants with equal probability every night, then it would be Markov.

Example 8.6 (Google PageRank)**.** Here is a somewhat similar example that *is* Markov. It is a very naive example of web-surfing. Suppose there are three pages on the internet, A, B, and C. Suppose A links to both B and C, and B links to C, but C links to nothing. Kevin is a very naive internet user. With 85% chance, Kevin will choose one of the two links randomly and follow that link (if there are no links, Kevin will stay put). With 15% chance, Kevin will choose a random page to visit. The process $\{X_n\}$ is Markov (it is the process of which page Kevin is one at click n), because where he ends up next depends only on where he is now.

Of course, this is very simplistic, especially given the massive amount of webpages on the internet. But here is a question – if Kevin surfs the internet for a very long time, where do you think he will end up with the highest likelihood? It seems like C, since all roads link to C, and C links to nowhere, and Kevin does not really like to not click links. How

does one justify this more precisely? We will get to this question next week (but you will see the idea worked out in an example of it on the HW this week).

Definition 8.7. Suppose $\{X_n\}$ is a time-homogeneous Markov chain on a finite state space $S = \{1, \dots, |S|\}$. The *transition matrix* P associated to this Markov chain is the $|S| \times |S|$ matrix whose entries are given by $S_{ij} = \mathbb{P}[X_1 = j | X_0 = i]$ for $i, j \in S$.

Lemma 8.8. *Let $\{X_n\}_{n \geqslant 0}$ be a time-homogeneous Markov chain with state space $S = \{1, \dots, |S|\}$, and let P be its transition matrix. Consider the vector $\mathbf{v}$ whose entries are $\mathbf{v}_j^{(n)} = \mathbb{P}[X_n = j]$. Then we have $\mathbf{v}^{(n+1)} = \mathbf{v}^{(n)} P$. More generally, we have $\mathbf{v}^{(n+k)} = \mathbf{v}^{(n)} P^k$.*

Proof. The second claim follows by induction in k. We prove the first claim. It suffices to show that for any j, we have

$$\mathbf{v}_j^{(n+1)} = (\mathbf{v}^{(n)} P)_j = \sum_i \mathbf{v}_i^{(n)} P_{ij} = \sum_i \mathbb{P}[X_n = i] \mathbb{P}[X_{n+1} = j | X_n = i].$$

By the law of total probability, the far RHS is just $\mathbb{P}[X_{n+1} = j]$. But so is the far LHS by definition, so we are done. □

Example 8.9. Suppose I have three states, A, B, and C. At each step, I look at my current position, and I move to one of the other two positions with equal probability. In particular, the probability of going $A \to B$ is $\frac{1}{2}$ and $A \to C$ is $\frac{1}{2}$ by $A \to A$ is 0. Similarly, going $B \to A$ is $\frac{1}{2}$ and $B \to C$ is $\frac{1}{2}$ and $B \to B$ is zero. Finally, we also have $C \to A$ is $\frac{1}{2}$ and $C \to B$ is $\frac{1}{2}$ and $C \to C$ is zero. In this case, we have

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}.$$

Just as a reality check, suppose initially that $\mathbb{P}[X_0 = A] = 1$ and $\mathbb{P}[X_0 \in \{B, C\}] = 0$. What is $\mathbb{P}[X_1 = s]$ equal to for $s = B, C$? We know its $\frac{1}{2}$, but let's use the previous lemma to see this. By said lemma, we know

$$\begin{aligned} \begin{pmatrix} \mathbb{P}[X_1 = A] & \mathbb{P}[X_1 = B] & \mathbb{P}[X_1 = C] \end{pmatrix} &= \begin{pmatrix} \mathbb{P}[X_0 = A] & \mathbb{P}[X_0 = B] & \mathbb{P}[X_0 = C] \end{pmatrix} P \\ &= \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}. \end{aligned}$$

What about X_2? One can reason out that it is $\begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$, but to check this using the lemma, note that

$$\begin{aligned} \begin{pmatrix} \mathbb{P}[X_2 = A] & \mathbb{P}[X_2 = B] & \mathbb{P}[X_2 = C] \end{pmatrix} &= \begin{pmatrix} \mathbb{P}[X_1 = A] & \mathbb{P}[X_1 = B] & \mathbb{P}[X_1 = C] \end{pmatrix} P \\ &= \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{4} + \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}. \end{aligned}$$

What about X_n for big n? Of course, one does not want to keep multiplying matrices over and over, especially when the dimension of the matrix is not so small. There is a trick we will learn and use next week, but that you get a taste of on this week's HW (in a worked out example).

8.2. **Recurrence vs. transience.**

Definition 8.10. Let $\{X_n\}$ be a time-homogeneous Markov chain on a countable state space S. Let $p_{jj}(n) = \mathbb{P}[X_n = j | X_0 = j]$ for any $j \in S$. We say that state j is *recurrent* (sometimes people use the word "persistent") if

$$\sum_{n=1}^{\infty} p_{jj}(n) = \infty,$$

and we say that state j is *transient* if

$$\sum_{n=1}^{\infty} p_{jj}(n) < \infty.$$

Intuitively, recurrence means eventually returning with probability 1, and transient means eventually stop returning.

Understanding recurrence and transience has many applications; it is a fundamental question regarding the long-time behavior of a Markov chain (whose interest we already came upon when looking at the PageRank example). It turns out that in the case of finite state spaces, this question is *completely* resolved. Let us go into this more.

Definition 8.11. Fix a time-homogeneous Markov chain with state space S. We say a subset $I \subseteq S$ is *closed* if the probability of starting in and leaving I is 0. We say I is *communicating* if for any $i, j \in I$, we have $\mathbb{P}[X_n = j | X_0 = i] > 0$ and $\mathbb{P}[X_m = i | X_0 = j] > 0$ for some n, m. (In words, it is possible to get from i to j and vice versa in a finite number of steps.) Note that S is always closed.

Example 8.12. In the PageRank example, the whole state space $S = \{A, B, C\}$ is closed, since there is always at least a 7.5% chance of going to any arbitrary page. However, if we restricted Kevin to only ever follow links, then any set not containing C cannot be closed. In particular, $\{A\}$ and $\{A, B\}$ are not closed, because there is a positive probability of going to C. Also, the set $\{A, C\}$ is not closed, since $A \to B$ is possible, but $B \to A$ is impossible, so $\{B, C\}$ is closed.

On the other hand, the only subset of S which is communicating is $\{C\}$ if we restrict Kevin to only follow links. For example, if $\{A, C\}$ is not communicating, since going from C to A in any finite number of steps is impossible.

Theorem 8.13. *Let $\{X_n\}$ be a time-homogeneous Markov chain with finite state space S. Let I be closed and communicating. Then every $i \in I$ is recurrent. Now, suppose that for some $j \in S$ and closed $I \subseteq S$ which does not contain j, the probability of going from j to I in finitely many steps is positive. Then j is transient.*

We will not prove this now, because it requires quite a technical argument, but you are more than allowed to use it (since this is how you get familiar with it anyways).

Example 8.14. Back to the PageRank example with restriction to only following links, note that A, B both have a positive probability to visit $\{C\}$, which is closed and communicating. Thus, A, B are both transient. But, C is recurrent, because $\{C\}$ is closed and communicating.

Example 8.15. Let $\{X_n\}$ be a symmetric simple random walk on $\mathbb{Z}$. Note that $\mathbb{Z}$ is closed and communicating. Thus, if the previous theorem applied to infinite state spaces, then every point in $\mathbb{Z}$ would be recurrent. This happens to be true, but not because of the previous theorem. In particular, if X_n is an *asymmetric* simple random walk with increments given by $\mathbb{P}[Y_n = 1] = p$ and $\mathbb{P}[Y_n = -1] = 1 - p$ for some $p \neq \frac{1}{2}$, then the origin is transient. The symmetric simple random walk case will be dealt with in the next section. The asymmetric case is on the HW (as is the case of symmetric simple random walk in higher dimensions).

8.3. **Recurrence of the symmetric simple random walk in one dimension.**

Theorem 8.16. *Let $X_n = \sum_{i=1}^n Y_i$ for $n \geqslant 1$ and $X_0 = 0$, where Y_i are i.i.d. and $\mathbb{P}[Y_i = \pm 1] = \frac{1}{2}$. Then $\{X_n\}_n$ has $0 \in \mathbb{Z}$ as a recurrent state, i.e.*

$$\sum_{n=0}^{\infty} \mathbb{P}[X_n = 0] = \infty.$$

This subsection is dedicated to the proof of this result. First, we note that $\mathbb{P}[X_n = 0]$ if n is odd, so that

$$\sum_{n=0}^{\infty} \mathbb{P}[X_n = 0] = \sum_{n=0}^{\infty} \mathbb{P}[X_{2n} = 0] = 1 + \sum_{n=1}^{\infty} \mathbb{P}[X_{2n} = 0].$$

Now, we claim that

$$\mathbb{P}[X_{2n} = 0] = \binom{2n}{n} 2^{-2n}.$$

To prove this, note that if $X_{2n} = 0$, then it must have taken exactly n steps to the right/upwards and n steps to the left/downwards. There are exactly $\binom{2n}{n}$ many ways to do this, since one takes the $2n$ steps and just chooses which n of them are to the right/upwards. Moreover, any particular sequence of $2n$ steps has probability 2^{-2n}:

$$\mathbb{P}[Y_1 = s_1, \ldots, Y_{2n} = s_{2n}] = \prod_{i=1}^{2n} \mathbb{P}[Y_i = s_i] = \prod_{i=1}^{2n} \frac{1}{2} = 2^{-2n},$$

where $s_1, \ldots, s_{2n}$ are any fixed numbers in $\{-1, +1\}$. Ultimately, we must compute

$$\sum_{n=0}^{\infty} \mathbb{P}[X_{2n} = 0] = \sum_{n=0}^{\infty} \binom{2n}{n} 2^{-2n}.$$

Lemma 8.17 (Stirling's formula)**.** *For large N, we have*

$$N! \sim \sqrt{2\pi N} \left(\frac{N}{e}\right)^N.$$

Here, $\sim$ means that if you divide the LHS by the RHS, the limit of the ratio is 1 as $N \to \infty$.

Let us take this for granted for now and use it. (Though, first, let us at least see why this formula kind of might be true, i.e. let us get a little feeling for this formula. Note that $\mathbb{P}[X_{2n} = 0]$ is the probability that a sum of i.i.d. mean zero, variance 1 random variables is equal to 0. Now, define $Z_{2n} = (2n)^{-1/2} X_{2n}$. We believe that $Z_{2n} \to N(0,1)$ by the CLT. But the pdf of a CLT has this $\sqrt{2\pi}$ in there somewhere. This explains the $\sqrt{2\pi N}$ on the RHS. The extra factor of $\sqrt{N}$ is because the Gaussian also has the variance inside the square root. But the variance of X_N, say, is N.)

Now, we have

$$\begin{aligned}\binom{2n}{n} &= \frac{2n}{n!(2n-n)!} = \frac{2n}{(n!)^2} \\ &\sim \sqrt{4\pi n}\left(\frac{2n}{e}\right)^{2n} \times \left[\frac{1}{\sqrt{2\pi n}}\left(\frac{e}{n}\right)^n\right]^2 \\ &= \frac{\sqrt{4\pi n}}{2\pi n}\left(\frac{2n}{e}\right)^{2n}\left(\frac{e}{n}\right)^{2n} = \frac{1}{\sqrt{\pi n}} 2^{2n}.\end{aligned}$$

Now, if we multiply by 2^{-2n}, we get that

$$\binom{2n}{n} 2^{-2n} \sim \frac{1}{\sqrt{\pi n}},$$

and thus

$$\sum_{n=1}^{\infty} \binom{2n}{n} 2^{-2n} \sim \sum_{n=1}^{\infty} \frac{1}{\sqrt{\pi n}}.$$

(Note that this term is just the $N(0, n/2)$ pdf at $x = 0$!) To see this diverges, we use the following test from calculus.

Lemma 8.18. *We have*

$$\sum_{n=1}^{\infty} n^{-\alpha} = \begin{cases} \infty & \alpha \leqslant 1 \\ \text{finite} & \alpha > 1 \end{cases}$$

Proof of Stirling's formula. Let us give a heuristic for Stirling's formula. By taking logarithms, it is enough to show

$$\log N! \approx N \log N - N + \frac{1}{2}\log(2\pi N).$$

By log rules, we know that $\log N! = \sum_{k=1}^{N} \log k$. We now *roughly approximate*

$$\sum_{k=1}^{N} \log k \approx \int_0^N \log x \mathrm{d}x.$$

The antiderivative of $\log x$ is $x \log x - x$; one can just differentiate this to check it. Thus, by the fundamental theorem of calculus, we have

$$\sum_{k=1}^{N} \log k \approx \int_0^N \log x \mathrm{d}x = (x \log x - x)|_{x=0}^{N} = N \log N - N.$$

This is almost correct! Indeed, it is off by the term $\frac{1}{2}\log(2\pi N)$, which is smaller than N or $N\log N$ as $N \to \infty$, so we are on the right track. This is the heuristic. The only thing between the heuristic and the actual proof is getting this $\frac{1}{2}\log(2\pi N)$. Where could the argument have missed this term? It must have been in the approximation of the sum of logs by the integral. In particular, we must also account for the term

$$\int_0^N \log[x] - \log(x)\mathrm{d}x = \int_0^N \log\frac{[x]}{x}\mathrm{d}x,$$

where $[x]$ is the smallest integer that is greater than or equal to x. Note that as x gets large, we know that $[x]/x$ converges to 1, since $|[x] - x| \leqslant 1$ always. Since $\log 1 = 0$, we expect that the integrand gets small when x gets large. This is why when we integrate over a domain of length N, we get something which is only $\log N$ in size! The details to make this precise are quite heavy, but please come see me in office hours if you want to talk about this! □

What about the asymmetric simple random walk, where there is a preference for direction? In particular, let $S_n = W_1 + \ldots + W_n$, where W_i are i.i.d. and $W_i = 1$ with probability p and $W_i = -1$ with probability $1-p$, where $p \neq \frac{1}{2}$. Note that $\mathbb{E}W_i = p - (1-p) = 2p - 1$. By the law of large numbers and CLT, we expect that $S_n \approx (2p-1)n$ plus something which looks like $N(0, Cn^{1/2})$. Thus, if $S_0 = 0$, then for $S_n = 0$ to be true for large n, we need something like $Z = -(2p-1)n$, where $Z \sim N(0, Cn^{1/2})$. But Z really does not like to be much bigger than $n^{1/2}$, and $n^{1/2} \ll n$ for large n, so this is very unlikely, and thus the asymmetric simple random walk is transient. Of course, this is a *heuristic* and by no means a proof. A rigorous way to show this is detailed in the HW.

What about the symmetric simple random walk in higher dimension d? In particular, let $\mathbf{X}(n) = (X_1(n), \ldots, X_d(n))$ where $X_i(n)$ are independent symmetric simple random walks. It turns out that for $d \leqslant 2$, it is recurrent, but for $d \geqslant 3$, it is transient. Again, showing this is detailed in the HW, and it follows the same line of reasoning. One way of interpreting this is that all roads lead to Rome, but only if you cannot fly or dig...

8.4. **A little interlude for some linear algebra.**

Theorem 8.19 (Perron-Frobenius theorem)**.** *Let $\{X_n\}_n$ be a homogeneous Markov chain with finite state space S. Let P be the associated transition matrix. Then P always has $\lambda = 1$ as a left-eigenvalue and an associated eigenvector π*

Lemma 8.20. *Let M be a square matrix of size $n \times n$. Let $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ be some enumeration of the eigenvalues. Then $\mathrm{Tr}M = \lambda_1 + \ldots + \lambda_n$ and $\det M = \lambda_1 \times \ldots \times \lambda_n$.*

Let's see how these two results help us, with a concrete example. Take the transition matrix

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}.$$

This is the Markov chain where there are three states, and at each step, one just chooses one of the two states they are not currently at with equal probability. Note that the whole state space is closed and communicating.

Perron-Frobenius tells us that $\lambda_1 = 1$ is an eigenvalue. What about the other eigenvalues? Usually, one computes the characteristic polynomial and finds its roots, but it's actually easier in this case. Let's see why. Note that $\mathrm{Tr} P = 0$ and

$$\det P = -\frac{1}{2}\det\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} + \frac{1}{8} + \frac{1}{8} = \frac{1}{4}.$$

Thus, we have $\lambda_1+\lambda_2+\lambda_3 = 1+\lambda_2+\lambda_3 = 0$ and $\lambda_1\lambda_2\lambda_3 = \lambda_2\lambda_3 = \frac{1}{4}$. One can solve this system of two equations in two variables, and one can check that $\lambda_2, \lambda_3 = -\frac{1}{2}$. Note how much simpler this is than finding the characteristic polynomial! Also note that $\lambda_1 = 1$ and the other eigenvalues are < 1 in absolute value, in particular that 1 has multiplicity 1.

Now, let $\mathbf{v}_j$ be the left eigenvector for λ_j, so that $\mathbf{v}_j P = \lambda_j \mathbf{v}_j$. Recall from linear algebra that any vector can be written as a linear combination of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$. Also, recall $\mathbf{v}^{(n)}$ as the probability vector after n steps in the Markov chain. Write $\mathbf{v}^{(0)} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \alpha_3\mathbf{v}_3$. We know that

$$\begin{aligned} \mathbf{v}^{(n)} &= \mathbf{v}^{(0)}P^n = \alpha_1\mathbf{v}_1P^n + \alpha_2\mathbf{v}_2P^n + \alpha_3\mathbf{v}_3P^n \\ &= \alpha_1\mathbf{v}_1 + \alpha_2\lambda_2^n\mathbf{v}_2 + \alpha_3\lambda_3^n\mathbf{v}_3. \end{aligned}$$

In particular, computing $\mathbf{v}^{(n)}$ is easy, *once we know what* $\mathbf{v}_j$ *are*. Actually, problem 2 on the HW tells you that for large n, it's enough to only know $\mathbf{v}_1$, approximately!

Let's try another example. Take

$$Q = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

This is the Markov chain where one just keeps swapping between states A, B, or one just stays put at state C. Note that there are *two* closed and communicating subsets of the state space, and that the entire state space is not communicating. Again, Perron-Frobenius tells us that $\lambda_1 = 1$ is an eigenvalue. Now, note that $\mathrm{Tr} Q = 1$ and $\det Q = -1$. Thus, we know $\lambda_1 + \lambda_2 + \lambda_3 = 1 + \lambda_2 + \lambda_3 = 1$ and $\lambda_1\lambda_2\lambda_3 = \lambda_2\lambda_3 = -1$. One can check that $\lambda_2 = 1$ and $\lambda_3 = -1$ solves this system. In particular, note that 1 has multiplicity 2 now. This more or less follows because the number of closed, communicating subsets is 2. We will discuss this point next week.

9. Week 9, starting Tue. Apr. 2, 2024

9.1. **Invariant measure and stationary distribution.**

Definition 9.1. Let P be the transition matrix for a Markov chain $\{X_n\}_n$ with finite state space S of size N. We say that a row vector $\pi \in \mathbb{R}^N$ is an *invariant measure* or *stationary distribution* if

- The entries π_j are all non-negative, and $\pi_1 + \ldots + \pi_N = 1$.
- We have $\pi P = \pi$.

Moreover, we say that P or $\{X_n\}_n$ is *reversible* with respect to π if for any i, j, we have $\pi_i P_{ij} = \pi_j P_{ji}$.

Example 9.2. Consider a two-state Markov chain such that $P_{ij} = \mathbf{1}_{i \neq j}$, i.e. it jumps between states A and B at each time. The transition matrix is

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

It is easy to see that $\pi = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$ is an invariant measure of P. Is it reversible? All we have to do is check $\pi_i P_{ij} = \pi_j P_{ji}$ for all $i, j = 1, 2$. Note that in general, one can assume that $i \neq j$, since if $i = j$, this is trivially true. In this specific case, one can see that $\pi_i = \frac{1}{2}$ for all $i = 1, 2$, so we just have to check that $P_{ij} = P_{ji}$, i.e. that P is symmetric. This is clear.

In general, if π_j is constant in j, then being reversible with respect to π means P is a symmetric matrix.

Example 9.3. Consider a two-state Markov chain which will jump from A to B, but which will stay at B forever. Its transition matrix is

$$P = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

One can check that $\pi = \begin{pmatrix} 0 & 1 \end{pmatrix}$ is an invariant measure. It is also reversible, since

$$\pi_1 P_{12} = 0 \quad \text{and} \quad \pi_2 P_{21} = 0.$$

Example 9.4. What about non-reversible Markov chains and stationary distributions? You will show on the HW that one needs to consider states spaces that have size at least 3. Consider the transition matrix

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}.$$

One can check that $\pi = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$ is an invariant measure. But, we claim that P is not reversible with respect to π. To see this, note that $\pi_1 P_{12} = \frac{1}{3}\frac{1}{2} = \frac{1}{6}$, but $\pi_2 P_{21} = 0$. One can see the non-reversibility intuitively. The Markov chain describes something walking along a triangle which either stays put with probability $1/2$, or it goes to the right with probability $1/2$. If one were to play a movie of this Markov chain for very long time, if you watch the movie in reverse, you would see something that stays put with probability $1/2$, or moves left with probability $\frac{1}{2}$.

9.2. **Perron-Frobenius theorem, in more detail.**

Theorem 9.5. *Let P be the transition matrix for a Markov chain with finite state space S of size N. The following are true.*

(1) There exists an eigenvalue $\lambda = 1$ with left eigenvector π of P. Moreover, π is a stationary distribution of P.

(2) Suppose μ is another eigenvalue of P. Then $|\mu| \leqslant 1$ (note that μ can be complex!).

(3) Suppose that for some $k \geqslant 0$, the matrix P^k has entries that are all positive. Then the eigenvalue $\lambda = 1$ has multiplicity 1, and $|\mu| < 1$ for any other eigenvalue μ.

Before we prove this theorem, let us see what it is actually saying.

(1) Point (1) is saying that any Markov chain with finite state space has at least one stationary distribution.
(2) Point (2) says that any other eigenvalue cannot be bigger than 1 in absolute value.
(3) Point (3) says that under some positivity condition, the matrix P^k converges to projection onto the stationary distribution of π. Let us make this a little more precise. Let $\mu_1, \ldots, \mu_{N-1}$ be the eigenvalues of P that are not λ, and let $\mathbf{v}_1, \ldots, \mathbf{v}_{N-1}$ be the corresponding left eigenvalues. Take the vector $\mathbf{p}$ such that $\mathbf{p}_j = \mathbb{P}[X_0 = j]$. Linear algebra says that we can write

$$\mathbf{p} = \pi + \alpha_1 \mathbf{v}_1 + \ldots + \alpha_{N-1}\mathbf{v}_{N-1}.$$

By applying P^k and using $\mu_j^k \to 0$ for any j as $k \to \infty$, we deduce $\mathbf{p}P^k \to \pi$.

Our discussion of point (3) above actually implies the following.

Lemma 9.6. *Let P be the transition matrix of a Markov chain $\{X_n\}_n$ with finite state space S. Suppose P has eigenvalue $\lambda = 1$ with multiplicity 1. Then, for any $s \in S$,*

$$\mathbb{P}[X_n = s] \to \pi_s.$$

Here's a question. Is there a probabilistic interpretation of the positive condition in point (3) of the Perron-Frobenius theorem? This is answered by the following.

Proposition 9.7. *Suppose P is the transition matrix of a Markov chain with finite state space S. The following are equivalent.*

- *There exists $k \geqslant 0$ such that P^k has strictly positive entries.*
- *The Markov chain is "irreducible" and "aperiodic". In particular:*
  - *The only closed subset is S itself, and S is communicating (this is what "irreducible" means).*
  - *Take any $i, j \in S$. Consider the set $\{k \geqslant 0 : \mathbb{P}[X_k = j | X_0 = i] > 0\}$. Then the greatest common divisor of the elements in this set is 1. (This is what "aperiodic" means.)*

Example 9.8. Take the transition matrix

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

This is irreducible, since the two states are communicating. However, it is periodic (i.e. not aperiodic). Indeed, look at the set of $k \geqslant 0$ such that $\mathbb{P}[X_k = A | X_0 = A]$. Then k must be even. Thus, the greatest common divisor of such k is 2, not 1.

Let us try to make sense of this in view of the previous proposition. Since this Markov chain is not aperiodic, we want to show that P^k cannot have strictly positive entries for any $k \geqslant 0$. One can inspect this directly, since $P^2 = \mathrm{I}_2$ and thus $P^{2k} = \mathrm{I}_2$ and $P^{2k+1} = P$. One can also verify this using Perron-Frobenius. Indeed, if P^k had strictly positive entries for some k, then P could only have one eigenvalue $\lambda = 1$, and any other eigenvalue is < 1 in absolute value. We computed the invariant measure to be $\pi = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$. Thus, we would have $\mathbb{P}[X_n = A] \to \frac{1}{2}$ for n large enough, regardless of what $\mathbb{P}[X_0 = A]$ is. But we can choose $\mathbb{P}[X_0 = A] = 1$, in which case $\mathbb{P}[X_n = A] = 1$ if n is even and 0 if n is odd.

Example 9.9. Take the transition matrix

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}.$$

You will show on the HW that $\lambda = 1$ has multiplicity 1, and any other eigenvalue is < 1 in absolute value. Thus, to make sense of the above results, let us try to find k such that P^k has purely positive entries. One can compute

$$P^2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}.$$

So $k = 2$ works.

Proof of the proposition. Suppose that there exists $k \geqslant 0$ such that P^k has strictly positive entries. Fix any $i, j \in S$. We claim that

$$\mathbb{P}[X_{k+1} = j | X_0 = i], \mathbb{P}[X_k = j | X_0 = i] > 0.$$

Note that $\mathbb{P}[X_k = j | X_0 = i] = (P^k)_{ij}$. This is positive by assumption. Now, note that

$$\mathbb{P}[X_{k+1} = j | X_0 = i] = (P^{k+1})_{ij} = (PP^k)_{ij} = \sum_{\ell} P_{i\ell}(P^k)_{\ell j}.$$

We know that $P_{i\ell} > 0$ for some ℓ, since the sum of $P_{i\ell}$ over all ℓ must equal 1, and they are non-negative. But assumption, we also know $(P^k)_{\ell j} > 0$ for all ℓ. Thus, the above display is > 0. In particular, the set of m such that $\mathbb{P}[X_m = j | X_0 = i]$ contains $k, k+1$. The greatest common divisor of this set must then be 1. Thus, the chain is aperiodic. It is irreducible because it is possible to go from any state to any state in k steps by assumption.

Now, suppose that the chain is irreducible and aperiodic. I will leave this part as an exercise (come see me in office hours if you would like to see the proof). □

9.3. **Proof of Perron-Frobenius.**

9.3.1. *Proof of (1).* Take the vector $\mathbf{p}$ such that $\mathbf{p}_1 = 1$ and $\mathbf{p}_j = 0$ for all $j \geqslant 2$. (This means start the chain at state 1 with probability 1.) For integer $T \geqslant 1$, define

$$\pi^{(T)} := \frac{1}{T} \sum_{k=1}^{T} \mathbf{p} P^k.$$

Note that the entries of $\mathbf{p}P^k$ are all between 0 and 1. In particular, the vectors $\pi^{(T)}$ belong to a compact subset of $\mathbb{R}^{|S|}$. In particular, there exists a sequence $\{T_\ell\}_{\ell=1}^{\infty}$ such that $T_\ell \to \infty$ as $\ell \to \infty$, and $\pi^{(T_\ell)}$ converges to some π entrywise. We claim that π is a

stationary distribution. To see this, note that

$$\pi P = \lim_{T_\ell \to \infty} \frac{1}{T_\ell} \sum_{k=1}^{T_\ell} \mathbf{p} P^k P = \lim_{T_\ell \to \infty} \frac{1}{T_\ell} \sum_{k=2}^{T_\ell+1} \mathbf{p} P^k$$
$$= \lim_{T_\ell \to \infty} \frac{1}{T_\ell} \sum_{k=1}^{T_\ell} \mathbf{p} P^k + \lim_{T_\ell \to \infty} \frac{1}{T_\ell} \mathbf{p} P^{T_\ell+1} - \lim_{T_\ell \to \infty} \frac{1}{T_\ell} \mathbf{p} P.$$

The first term in the second line is just π. The second term and third term are vectors in $\mathbb{R}^{|S|}$ whose entries are all in $[0, 1]$ and thus have length bounded by $|S|^{1/2}$. Dividing by T_ℓ and sending $T_\ell \to \infty$ shows that the last two terms vanish, so we get $\pi P = \pi$.

Remark 9.10. This argument is known as the Krylov-Bogoliubov argument, and it is the only general way we know how to construct stationary distributions of more complicated Markov chains.

9.3.2. *Proof of (2).* Assume that part (3) is true. There exists a sequence of transition matrices P_n that converges to P and such that the entries of P_n are all strictly positive. Part (3) implies that the claim is true for P_n. But eigenvalues are continuous in matrix entries, so we can take limits.

9.3.3. *Proof of (3).* Assume that $k = 1$; the eigenvalues of P^k are just k-th powers of the eigenvalues of P, so it actually is enough to assume $k = 1$ for a complete proof, but let's just assume it to make things simpler. Suppose P has an eigenvalue μ such that $\mu \neq 1$ and $|\mu| = 1$. We can find $m \geqslant 1$ such that the real part of μ^m is negative. One can also check that P^m has strictly positive entries by induction on m. Now, take $\varepsilon > 0$ smaller than the minimal entry of P^m, and consider the matrix $P^m - \varepsilon \mathrm{Id}$. This matrix has $\mu^m - \varepsilon$ as an eigenvalue. But $|\mu^m - \varepsilon| > 1$ since $|\mu^m| = 1$ and the real part of μ^m is strictly negative. We now claim that any matrix Q whose entries are positive and whose rows sum to $\leqslant 1$ cannot have an eigenvalue that is > 1 in absolute value; applying this claim to $Q = P^m - \varepsilon \mathrm{Id}$ gives the desired contradiction.

To prove this matrix fact, suppose Q has an eigenvalue η such that $|\eta| > 1$ with eigenvector $\mathbf{w}$. Then $Q^j \mathbf{w} = \eta^j \mathbf{w}$. The maximal entry of $\eta^j \mathbf{w}$ goes to ∞ in absolute value as $j \to \infty$. However, we also know that

$$|(Q^{j+1}\mathbf{w})_i| = \sum_\ell Q_{i\ell} |(Q^j \mathbf{w})_\ell| \leqslant \max_\ell |(Q^j \mathbf{w})_\ell|$$

since the entries of Q are non-negative and its rows sum to 1. By taking a max over i, we deduce that the max of $|(Q^j \mathbf{w})_\ell|$ is non-increasing in j. This contradicts what we had before, so we are done.

It suffices to show that $\lambda = 1$ has multiplicity 1.