Math 154: Probability Theory, Lecture Notes

KEVIN YANG

CONTENTS

1. Week 1, starting Tue. Jan. 23, 2024 2
1.1. Probability spaces and events 2
1.2. Conditional probability 3
1.3. Independence 4
1.4. Some examples 5
2. Week 2, starting Tue. Jan. 30, 2024 6
2.1. Random variables 6
2.2. Independence of random variables 8
2.3. Expectation 9
2.4. Variance and higher moments 11
2.5. Cauchy-Schwarz and Hölder inequalities 13
3. Week 3, starting Tue. Feb. 6, 2024 13
3.1. Law of the unconscious statistician 13
3.2. Continuous random variables 14
3.3. Independence 16
3.4. Change of variables 17
3.5. Random vectors 17
3.6. Multivariate Gaussians 18
4. Week 4, starting Tue. Feb. 13, 2024 20
4.1. Triangle inequality 20
4.2. Laplace and Fourier transforms, i.e. moment generating functions and characteristic functions 20
4.3. How to compute moments 21
4.4. Some inequalities 23
4.5. Some applications of these inequalities 24
4.6. The Law of Large Numbers 24
5. Week 5, starting Tue. Feb. 19, 2024 25
5.1. Just a reminder 25
5.2. Random vectors 25
5.3. Conditional expectation 26
5.4. Martingales 27
5.5. A little fun fact about Gaussian tail probabilities 29
5.6. Azuma's inequality and Doob's maximal inequality 29

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 $\mathbb{E}(X) \geqslant 0$.*
*(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

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

(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 $X \sim \text{Bern}(p)$, then $\mathbb{E}(X) = p$. If $X \sim \text{Bin}(n, p)$, then $X = Y_1 + \ldots + Y_n$ where $Y_i \sim \text{Bern}(p)$, so $\mathbb{E}(X) = np$. If $X \sim \text{Pois}(\lambda)$, 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 $\mathbf{X} = (X_1, \dots, X_n)$ and $X_1, \dots, X_n$ are discrete random variables. Its probability mass function $p : \mathbb{R}^n \to \mathbb{R}$ is defined by

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

A *continuous* random vector of dimension (of length) n is a vector $\mathbf{X} \in \mathbb{R}^n$ such that $\mathbf{X} = (X_1, \dots, X_n)$ and $X_1, \dots, X_n$ 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:

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

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

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

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

□

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 $\mathbb{E}M_N = \mathbb{E}M_0$ 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 $\sup_{N \geqslant 1} |M_{N+1} - M_N| < \infty$.*

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

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). 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 $Z \sim N(0, \sigma^2_{C_2})$ 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.