

Lecture 1: Axioms of Probability

EECS 126 (UC Berkeley)

Spring 2026 — January 20, 2026

A probability space provides the mathematical framework for modeling random experiments. In this lecture we define the axioms of probability, introduce sample spaces and events, work through several foundational examples, and derive the birthday paradox bound.

1 Probability Spaces

Definition 1.1 (Probability space)

A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where:

- Ω is the *sample space* — the set of all possible outcomes of the experiment.
- $\mathcal{F} \subseteq 2^\Omega$ is a σ -algebra of subsets of Ω , called *events*.
- $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ is a *probability measure* satisfying the Kolmogorov axioms.

The elements $\omega \in \Omega$ are called *base outcomes* (or *samples*). A σ -algebra $\mathcal{F}$ is a family of subsets of Ω that is closed under countable unions, countable intersections, and complements, and that contains Ω itself. In many discrete settings we simply take $\mathcal{F} = 2^\Omega$ (the power set), so that every subset of Ω is an event.

Example 1.1 (Tossing two fair coins)

Toss two fair coins. The sample space is

$$\Omega = \{HH, HT, TH, TT\}. \quad (1.1)$$

Each base outcome is equally likely, so we assign $\mathbb{P}(\{\omega\}) = \frac{1}{4}$ for each $\omega \in \Omega$. The event space is the power set $\mathcal{F} = 2^\Omega$, which has $|\mathcal{F}| = 2^{|\Omega|} = 2^4 = 16$ elements.

Example 1.2 (Toss until heads)

Toss a coin repeatedly until the first heads. Then

$$\Omega = \{H, TH, TTH, TTTH, \dots\} \quad (1.2)$$

is a *countably infinite* sample space.

Example 1.3 (Waiting for a bus)

Wait at a bus stop for the next bus, which arrives at some time in $[0, T]$. Now

$$\Omega = [0, T] \quad (1.3)$$

is an *uncountably infinite* (continuous) sample space. In this setting the choice of $\mathcal{F}$ becomes nontrivial; one typically takes $\mathcal{F}$ to be the Borel σ -algebra on $[0, T]$.

2 Kolmogorov Axioms

Definition 1.2 (Kolmogorov axioms)

A function $\mathbb{P} : \mathcal{F} \rightarrow \mathbb{R}$ is a **probability measure** on $(\Omega, \mathcal{F})$ if it satisfies:

1. **Nonnegativity.** $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$.
2. **Normalization.** $\mathbb{P}(\Omega) = 1$.
3. **Countable additivity.** If $A_1, A_2, \dots \in \mathcal{F}$ are pairwise disjoint ($A_i \cap A_j = \emptyset$ for $i \neq j$), then

$$\mathbb{P}\left(\bigcup_{i \geq 1} A_i\right) = \sum_{i \geq 1} \mathbb{P}(A_i). \quad (1.4)$$

Why countable additivity? In a continuous sample space such as $\Omega = [0, 1]$, every individual point has probability zero: $\mathbb{P}(\{x\}) = 0$. If we only had *finite* additivity, we could not recover $\mathbb{P}([0, 1]) = 1$ from the probabilities of individual points. Countable additivity bridges this gap and is essential for the theory of continuous distributions.

Example 1.4 (Biased coin)

Flip a biased coin: heads with probability p , tails with probability $1 - p$. Set

$$\Omega = \{H, T\}, \quad \mathcal{F} = 2^\Omega = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}. \quad (1.5)$$

Define

$$\mathbb{P}(H) = p, \quad \mathbb{P}(T) = 1 - p, \quad \mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(\{H, T\}) = 1. \quad (1.6)$$

Then $(\Omega, \mathcal{F}, \mathbb{P})$ is a valid probability space.

3 Fundamental Probability Facts

The following basic identities are direct consequences of the Kolmogorov axioms.

Theorem 1.1 (Complement rule)

For any event $A \in \mathcal{F}$,

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A). \quad (1.7)$$

Proof. Since A and A^c are disjoint with $A \cup A^c = \Omega$, countable additivity gives $\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1$. $\square$

Theorem 1.2 (Inclusion–exclusion)

For any two events $A, B \in \mathcal{F}$,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \quad (1.8)$$

More generally, for events $A_1, A_2, \dots, A_n$,

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i < j} \mathbb{P}(A_i \cap A_j) + \sum_{i < j < k} \mathbb{P}(A_i \cap A_j \cap A_k) - \dots + (-1)^{n+1} \mathbb{P}(A_1 \cap \dots \cap A_n). \quad (1.9)$$

Theorem 1.3 (Union bound)

For any events $A_1, \dots, A_n$,

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathbb{P}(A_i). \quad (1.10)$$

Proof. Define $B_1 = A_1$ and $B_k = A_k \setminus \bigcup_{i < k} A_i$ for $k \geq 2$. The B_k are disjoint with $\bigcup B_k = \bigcup A_k$, and $\mathbb{P}(B_k) \leq \mathbb{P}(A_k)$, so $\mathbb{P}(\bigcup A_k) = \sum \mathbb{P}(B_k) \leq \sum \mathbb{P}(A_k)$. $\square$

4 Discrete Probability and Counting

In a discrete sample space, the probability of an event is just the sum of the probabilities of its base outcomes:

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}). \quad (1.11)$$

Definition 1.3 (Uniform sample space)

When all outcomes are equally likely, we say Ω is a **uniform sample space**, and for any event A ,

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}. \quad (1.12)$$

5 The Birthday Paradox

Example 1.5 (Birthday paradox)

Let n people be in a room. What is the probability that at least two share a birthday? Let $K = 365$ and assume birthdays are uniform. Let $A =$ “at least two people share a birthday.” Compute via the complement:

$$\mathbb{P}(A^c) = \frac{K(K-1)(K-2)\cdots(K-n+1)}{K^n} = \prod_{j=0}^{n-1} \left(1 - \frac{j}{K}\right). \quad (1.13)$$

Using $1 - x \leq e^{-x}$:

$$\mathbb{P}(A^c) \leq \prod_{j=0}^{n-1} e^{-j/K} = e^{-\sum_{j=0}^{n-1} j/K} = e^{-n(n-1)/(2K)}. \quad (1.14)$$

Theorem 1.4 (Birthday bound)

$\mathbb{P}(A) \geq 1 - e^{-n(n-1)/(2K)}$. For $K = 365$ and $n = 23$, this gives $\mathbb{P}(A) \geq 50\%$.

Balls-and-bins interpretation. The birthday problem is equivalent to throwing n balls uniformly into K bins and asking for the probability of a collision. This connection arises frequently in hashing, load balancing, and networking.

End of Lecture 1.

Lecture 2: Conditional Probability and Independence

EECS 126 (UC Berkeley)
Spring 2026 — January 22, 2026

In this lecture we introduce conditional probability, derive the multiplication (chain) rule, state the law of total probability and Bayes' rule, work through the classic false-positive medical testing example, and then develop the theory of independence and conditional independence.

1 Conditional Probability

Definition 2.1 (Conditional probability)

Let A, B be events with $\mathbb{P}(B) > 0$. The **conditional probability** of A given B is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \quad (2.1)$$

Intuitively, $\mathbb{P}(A \mid B)$ is the probability that A occurs when we know that B has occurred. Conditioning on B restricts the sample space from Ω to B ; the event B becomes the new “ Ω ” for A .

In a uniform sample space, $\mathbb{P}(A) = |A|/|\Omega|$. Conditioning on B gives $\mathbb{P}(A \mid B) = |A \cap B|/|B|$, which is simply the fraction of outcomes in B that also lie in A .

Example 2.1 (Two dice – conditioning on the sum)

Roll two standard dice. Let $A = \{\text{1st die is 6}\}$, $B = \{\text{sum of two dice is 7}\}$, and $C = \{\text{sum of two dice is 11}\}$.

Computing $\mathbb{P}(A \mid B)$: The outcomes in B (sum = 7) are $\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$, so $\mathbb{P}(B) = 6/36$. The only outcome in $A \cap B$ is $(6, 1)$, so $\mathbb{P}(A \cap B) = 1/36$. Therefore

$$\mathbb{P}(A \mid B) = \frac{1/36}{6/36} = \frac{1}{6}. \quad (2.2)$$

Computing $\mathbb{P}(A \mid C)$: The outcomes in C (sum = 11) are $\{(5, 6), (6, 5)\}$, so $\mathbb{P}(C) = 2/36$. The only outcome in $A \cap C$ is $(6, 5)$, so

$$\mathbb{P}(A \mid C) = \frac{1/36}{2/36} = \frac{1}{2}. \quad (2.3)$$

2 Multiplication Rule (Chain Rule)

Theorem 2.1 (Chain rule of probability)

For any events $A_1, A_2, \dots, A_n$ with $\mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_{n-1}) > 0$,

$$\mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_n) = \mathbb{P}(A_1) \mathbb{P}(A_2 | A_1) \mathbb{P}(A_3 | A_1 \cap A_2) \dots \mathbb{P}(A_n | A_1 \cap \dots \cap A_{n-1}). \quad (2.4)$$

Proof. By repeated application of the definition of conditional probability. Write $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) \mathbb{P}(A_2 | A_1)$, then $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1 \cap A_2) \mathbb{P}(A_3 | A_1 \cap A_2)$, and continue inductively. Each factor is well-defined because $\mathbb{P}(A_1 \cap \dots \cap A_{k-1}) > 0$ implies $\mathbb{P}(A_1 \cap \dots \cap A_k) \leq \mathbb{P}(A_1 \cap \dots \cap A_{k-1})$, and we are given the intersection has positive probability up to $n - 1$ terms. $\square$

Birthday paradox via the chain rule. The chain rule gives an alternative derivation of the birthday bound. Let $A_i = \{\text{person } i \text{ doesn't share a birthday with persons } 1, \dots, i-1\}$. Then

$$\mathbb{P}(A_1 \cap \dots \cap A_n) = 1 \cdot \frac{K-1}{K} \cdot \frac{K-2}{K} \dots \frac{K-(n-1)}{K} = \prod_{j=0}^{n-1} \left(1 - \frac{j}{K}\right) \approx e^{-n(n-1)/(2K)}, \quad (2.5)$$

recovering the same bound from Lecture 1.

3 Law of Total Probability

Theorem 2.2 (Law of total probability)

Let $A_1, A_2, \dots, A_N$ be a **partition** of Ω (i.e. the A_i are pairwise disjoint — mutually exclusive — and their union is Ω — collectively exhaustive). Then for any event B ,

$$\mathbb{P}(B) = \sum_{i=1}^N \mathbb{P}(A_i) \mathbb{P}(B | A_i). \quad (2.6)$$

Proof. Since $\{A_i\}$ partitions Ω , we can decompose B as $B = \bigcup_{i=1}^N (A_i \cap B)$, and these pieces are disjoint. By countable additivity, $\mathbb{P}(B) = \sum_{i=1}^N \mathbb{P}(A_i \cap B) = \sum_{i=1}^N \mathbb{P}(A_i) \mathbb{P}(B | A_i)$, where the last step uses the definition of conditional probability. $\square$

4 Bayes' Rule

Theorem 2.3 (Bayes' rule)

Let $A_1, \dots, A_N$ partition Ω . Given an event B with $\mathbb{P}(B) > 0$,

$$\mathbb{P}(A_i | B) = \frac{\mathbb{P}(B | A_i) \mathbb{P}(A_i)}{\sum_{j=1}^N \mathbb{P}(B | A_j) \mathbb{P}(A_j)}. \quad (2.7)$$

The numerator uses the definition of conditional probability, and the denominator is $\mathbb{P}(B)$ expanded by the law of total probability.

Example 2.2 (False positive quiz – testing for a rare disease)

A medical test for a rare disease has the following properties:

- If a person *has* the disease, the test is positive with probability 0.95 (sensitivity).
- If a person *does not* have the disease, the test is negative with probability 0.95 (specificity), i.e. false positive rate = 0.05.
- The prior prevalence is $\mathbb{P}(A) = 0.001$ (1 in 1000 people have the disease).

A person tests positive. What is $\mathbb{P}(\text{disease} | \text{positive})$?

Define $A = \{\text{has disease}\}$, $B = \{\text{tests positive}\}$. By Bayes' rule:

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(B | A) \mathbb{P}(A)}{\mathbb{P}(B | A) \mathbb{P}(A) + \mathbb{P}(B | A^c) \mathbb{P}(A^c)} = \frac{(0.95)(0.001)}{(0.95)(0.001) + (0.05)(0.999)} \approx 0.0187. \quad (2.8)$$

Despite the test being 95% accurate, a positive result only indicates a $\approx 1.87\%$ chance of actually having the disease. This is because the disease is so rare that the false positives from healthy people vastly outnumber the true positives.

Sensitivity to the prior. If the prior were $\mathbb{P}(A) = 0.01$ instead ($10\times$ more common), then $\mathbb{P}(A | B) \approx 16\%$ — still low, but dramatically higher. The base rate (prior) has an enormous effect on the posterior.

5 Independence

Definition 2.2 (Independence of two events)

Two events A and B are **independent** if the occurrence of one provides no information about the occurrence of the other. Formally:

$$\mathbb{P}(A | B) = \mathbb{P}(A), \quad \text{or equivalently,} \quad \mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B). \quad (2.9)$$

The second form says the joint probability equals the product of the marginals.

Key remarks on independence:

1. **Disjoint $\neq$ independent.** A, B disjoint means $\mathbb{P}(A \cap B) = 0$. For this to equal $\mathbb{P}(A)\mathbb{P}(B)$, we need $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$. So disjoint events with positive probability are *dependent*, not independent.
2. **Base outcomes are always dependent.** Individual outcomes $\{\omega\}$ in any finite Ω are

always dependent (knowing one occurred rules out the others).

Example 2.3 (Dice: independence depends on the conditioning event)

Roll two fair dice. Let $A = \{\text{1st die is 6}\}$, $B = \{\text{sum is 7}\}$, $C = \{\text{sum is 11}\}$.

Are A and B independent? $\mathbb{P}(A | B) = 1/6 = \mathbb{P}(A)$, so **yes**, $A \perp B$.

Are A and C independent? $\mathbb{P}(A | C) = 1/2 \neq 1/6 = \mathbb{P}(A)$, so **no**, A and C are dependent. Intuitively, knowing the sum is 11 makes it more likely that the first die was a 6 (since (6, 5) is one of only two outcomes).

6 Conditional Independence

Definition 2.3 (Conditional independence)

Events A and B are **conditionally independent** given C if

$$\mathbb{P}(A \cap B | C) = \mathbb{P}(A | C) \mathbb{P}(B | C). \quad (2.10)$$

Independence and conditional independence are distinct concepts:

1. Dependent events can become conditionally independent.
2. Independent events can become conditionally dependent.

Example 2.4 (Two indistinguishable coins)

You have two coins: one is double-headed, the other is double-tailed. You pick one at random and flip it twice. Let $H_i = \{\text{flip } i \text{ is heads}\}$, $i = 1, 2$.

Q1: Are H_1, H_2 (unconditionally) independent? No. $\mathbb{P}(H_1 | H_2) = 1 \neq 1/2 = \mathbb{P}(H_1)$. If the second flip is heads, you certainly picked the double-headed coin, so the first flip is heads with certainty.

Q2: Conditioning on the coin, are they independent? Let $A = \{\text{picked the double-headed coin}\}$. Then $\mathbb{P}(H_1 \cap H_2 | A) = \mathbb{P}(H_1 | A) \mathbb{P}(H_2 | H_1 \cap A) = 1 \cdot 1 = 1$, and $\mathbb{P}(H_1 | A) \mathbb{P}(H_2 | A) = 1 \cdot 1 = 1$. These are equal, so $H_1 \perp H_2 | A$. Once we know which coin was picked, each flip is determined and independent.

Example 2.5 (Fair coin – independent but conditionally dependent)

Toss a fair coin twice. Let $H_1 = \{\text{1st toss is H}\}$, $H_2 = \{\text{2nd toss is H}\}$, and $D = \{\text{1st toss} \neq \text{2nd toss}\}$.

Q1: $H_1 \perp H_2$? Yes, by independence of fair coin flips.

Q2: $H_1 \perp H_2 | D$? No. $\mathbb{P}(H_1 | D) \mathbb{P}(H_2 | D) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. But $\mathbb{P}(H_1 \cap H_2 | D) = 0$, since both flips being heads contradicts D (the flips must differ). Since $0 \neq 1/4$, the events are conditionally dependent given D .

7 Joint Independence vs. Pairwise Independence

Definition 2.4 (Independence of a collection)

Events $\{A_i\}_{i \in S}$ are **jointly (mutually) independent** if for every subset $T \subseteq S$,

$$\mathbb{P}\left(\bigcap_{i \in T} A_i\right) = \prod_{i \in T} \mathbb{P}(A_i). \quad (2.11)$$

Pairwise independence only requires this for all pairs $|T| = 2$.

Pairwise independence $\nRightarrow$ joint independence. Consider the fair coin example: H_1, H_2, D are pairwise independent (check all three pairs), but they are *not* jointly independent because $\mathbb{P}(H_1 \cap H_2 \cap D) = 0 \neq \mathbb{P}(H_1)\mathbb{P}(H_2)\mathbb{P}(D) = 1/8$.

8 Renewal: Rolling a 6 Before a 7

Example 2.6 (Roll a sum of 6 before a sum of 7)

Roll two fair dice independently and repeatedly. Let $E = \{\text{a sum of 6 appears before a sum of 7}\}$.

The outcomes that sum to 6 are $\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$, so $\mathbb{P}(S) = 5/36$ where $S = \{\text{1st roll sums to 6}\}$. Similarly, $\mathbb{P}(T) = 6/36$ where $T = \{\text{1st roll sums to 7}\}$.

By the law of total probability, conditioning on the first roll:

$$\mathbb{P}(E) = \mathbb{P}(E | S) \mathbb{P}(S) + \mathbb{P}(E | T) \mathbb{P}(T) + \mathbb{P}(E | (S \cup T)^c) \mathbb{P}((S \cup T)^c). \quad (2.12)$$

We have $\mathbb{P}(E | S) = 1$, $\mathbb{P}(E | T) = 0$, and $\mathbb{P}(E | (S \cup T)^c) = \mathbb{P}(E)$ by the renewal/memoryless property (each roll is independent, so if neither 6 nor 7 appears, we restart). Thus

$$\mathbb{P}(E) = 1 \cdot \frac{5}{36} + 0 \cdot \frac{6}{36} + \mathbb{P}(E) \cdot \frac{25}{36}. \quad (2.13)$$

Solving: $\mathbb{P}(E)(1 - 25/36) = 5/36$, so $\mathbb{P}(E)(11/36) = 5/36$, giving

$$\mathbb{P}(E) = \frac{5}{11}. \quad (2.14)$$

Lecture 3: Discrete Random Variables and Expectation

EECS 126 (UC Berkeley)

Spring 2026 — January 27, 2026

We transition from events to random variables — functions that assign numerical values to outcomes — and develop the key tools of expectation, linearity, LOTUS, and variance.

1 Random Variables and PMFs

Definition 3.1 (Random variable)

A **(discrete) random variable** is a function $X : \Omega \rightarrow \mathbb{R}$ that assigns a real number to each outcome $\omega \in \Omega$. We write $\mathbb{P}(X = x)$ for the probability of the event $\{\omega \in \Omega : X(\omega) = x\}$.

Definition 3.2 (Probability mass function)

The **probability mass function (PMF)** of a discrete random variable X is the function $p_X : \mathbb{R} \rightarrow [0, 1]$ defined by

$$p_X(x) = \mathbb{P}(X = x). \quad (3.1)$$

The PMF satisfies $\sum_{x \in \mathcal{X}} p_X(x) = 1$, where $\mathcal{X}$ denotes the set of all values taken by X (the *support*).

Example 3.1 (Functions of random variables)

Let X be a random variable with $p_X(x) = 1/5$ for $x \in \{-2, -1, 0, 1, 2\}$. Define $Y = |X|$. Then Y takes values in $\{0, 1, 2\}$ and

$$p_Y(y) = \sum_{\{x : |x|=y\}} p_X(x) = \begin{cases} 1/5 & \text{if } y = 0, \\ 2/5 & \text{if } y = 1, \\ 2/5 & \text{if } y = 2. \end{cases} \quad (3.2)$$

In general, if $Y = g(X)$, then $p_Y(y) = \sum_{\{x : g(x)=y\}} p_X(x)$.

2 Popular Discrete Distributions

Definition 3.3 (Bernoulli distribution)

$X \sim \text{Bernoulli}(p)$: a single coin toss.

$$\mathbb{P}(X = k) = \begin{cases} 1 - p & k = 0, \\ p & k = 1. \end{cases} \quad (3.3)$$

Definition 3.4 (Binomial distribution)

$X \sim \text{Binomial}(n, p)$: number of successes in n independent Bernoulli(p) trials.

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n. \quad (3.4)$$

Note: $\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1$ by the binomial theorem, confirming this is a valid PMF.

Definition 3.5 (Geometric distribution)

$X \sim \text{Geometric}(p)$: number of trials until the first success.

$$\mathbb{P}(X = k) = (1-p)^{k-1} p, \quad k = 1, 2, 3, \dots \quad (3.5)$$

The geometric distribution is the discrete analog of “waiting time.”

Definition 3.6 (Poisson distribution)

$X \sim \text{Poisson}(\lambda)$: models the number of arrivals in a fixed interval.

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots \quad (3.6)$$

The Poisson is the limiting distribution of $\text{Binomial}(n, \lambda/n)$ as $n \rightarrow \infty$, and arises naturally in modeling rare events and arrival processes.

Example 3.2 (Chess match – a PMF computation)

Alice and Bob play a match: the first to win a game wins the match. If there are 10 consecutive draws, the match is declared drawn. Each game independently: $\mathbb{P}(\text{A wins}) = 0.3$, $\mathbb{P}(\text{B wins}) = 0.4$, $\mathbb{P}(\text{draw}) = 0.3$.

Let L be the duration (number of games). Then for $\ell = 1, 2, \dots, 9$:

$$p_L(\ell) = (0.3)^{\ell-1} (1 - 0.3) = (0.3)^{\ell-1} (0.7), \quad (3.7)$$

since the first $\ell - 1$ games must be draws and the ℓ -th decisive. For $\ell = 10$: $p_L(10) = (0.3)^9$ (all 9 games drawn, match ends).

$\mathbb{P}(\text{Alice wins match}) = \sum_{\ell=1}^9 (0.3)^{\ell-1} (0.3) = 0.3 \sum_{k=0}^8 (0.3)^k$. Using $\sum_{k=0}^{\infty} r^k = 1/(1-r)$ for $|r| < 1$, in the limit of no cap on draws: $\mathbb{P}(\text{Alice wins}) = 0.3/(1-0.3) = 3/7$.

3 Expectation

Definition 3.7 (Expected value)

The **expected value** (or **mean**) of a discrete random variable X is

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \mathbb{P}(\{\omega\}) = \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}(X = x), \quad (3.8)$$

provided the sum converges absolutely (i.e. $\sum |x| \mathbb{P}(X = x) < \infty$).

Theorem 3.1 (Linearity of expectation)

For *any* random variables X, Y (not necessarily independent) and constant c :

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y], \quad \mathbb{E}[cX] = c\mathbb{E}[X]. \quad (3.9)$$

No independence assumption is needed. This is the single most useful property of expectation.

Definition 3.8 (Indicator random variable)

For an event A , the **indicator** $\mathbf{1}_A$ is

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \omega \in A, \\ 0 & \omega \notin A. \end{cases} \quad (3.10)$$

Observe that $\mathbf{1}_A \sim \text{Bernoulli}(\mathbb{P}(A))$, so $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)$.

Example 3.3 (Seats on a plane)

A full flight has n passengers with assigned seats, but every passenger ignores their assignment and sits in a uniformly random seat. What is $\mathbb{E}[\text{number who sit in their own seat}]$?

Let $X_i = \mathbf{1}\{\text{passenger } i \text{ sits in their own seat}\}$. Then $X = X_1 + X_2 + \cdots + X_n$. The X_i are *not* independent (if $n = 2$ and passenger 1 is in their seat, passenger 2 must be too). But linearity of expectation does not require independence:

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i] = \sum_{i=1}^n \mathbb{P}(\text{passenger } i \text{ is in their seat}) = n \cdot \frac{1}{n} = 1. \quad (3.11)$$

Regardless of the number of passengers, on average exactly 1 person sits in their own seat.

4 LOTUS (Law of the Unconscious Statistician)

Theorem 3.2 (LOTUS)

If X is a discrete random variable and $g : \mathbb{R} \rightarrow \mathbb{R}$ is any function, then

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) \mathbb{P}(X = x). \quad (3.12)$$

More generally, for a function of several random variables:

$$\mathbb{E}[f(X_1, \dots, X_n)] = \sum_{x_1, \dots, x_n} f(x_1, \dots, x_n) p_{X_1, \dots, X_n}(x_1, \dots, x_n). \quad (3.13)$$

The name reflects that one can compute $\mathbb{E}[g(X)]$ directly from the distribution of X , without first finding the distribution of $Y = g(X)$.

5 Independence of Random Variables

Definition 3.9 (Independent random variables)

Random variables X and Y are **independent** (written $X \perp Y$) if their joint distribution factors into the product of the marginals:

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \mathbb{P}(Y = y) \quad \text{for all } x, y. \quad (3.14)$$

Theorem 3.3 (Expectation of a product of independent RVs)

If $X \perp Y$, then

$$\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]. \quad (3.15)$$

Proof. By LOTUS and independence:

$$\begin{aligned} \mathbb{E}[XY] &= \sum_{x,y} xy p_{X,Y}(x,y) = \sum_{x,y} xy p_X(x) p_Y(y) \\ &= \left(\sum_x x p_X(x) \right) \left(\sum_y y p_Y(y) \right) = \mathbb{E}[X] \mathbb{E}[Y]. \end{aligned} \quad (3.16)$$

The key step is factoring the joint PMF into marginals (by independence), which allows separating the double sum into a product of single sums. $\square$

The converse is false. $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ does *not* imply $X \perp Y$. The condition $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ is called *uncorrelation*, which is strictly weaker than independence. As an exercise, construct dependent random variables that are uncorrelated.

6 Tail Sum Formula

Theorem 3.4 (Tail sum formula)

Let X be a random variable taking values in $\mathbb{N} = \{0, 1, 2, \dots\}$. Then

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k). \quad (3.17)$$

Proof. Starting from the definition of expectation:

$$\begin{aligned} \mathbb{E}[X] &= \sum_{k=1}^{\infty} k \mathbb{P}(X = k) = \sum_{k=1}^{\infty} \sum_{j=1}^k 1 \cdot \mathbb{P}(X = k) \\ &= \sum_{j=1}^{\infty} \sum_{k=j}^{\infty} \mathbb{P}(X = k) = \sum_{j=1}^{\infty} \mathbb{P}(X \geq j). \end{aligned} \quad (3.18)$$

The key step is writing $k = \sum_{j=1}^k 1$ to introduce a double sum, then swapping the order of summation (Tonelli's theorem ensures this is valid for nonnegative terms). Geometrically, this corresponds to “counting vertically then horizontally”: instead of weighting each column k by its height k , we sum across rows $j = 1, 2, \dots$, each contributing $\mathbb{P}(X \geq j)$. $\square$

Intuition for the tail sum. Think of $\mathbb{E}[X]$ as the total area under the “staircase” of the survival function $\mathbb{P}(X \geq k)$. Each unit step at height k contributes $\mathbb{P}(X \geq k)$ to the total. This is the same idea as computing the area under a curve by integrating horizontally rather than vertically.

7 Variance

Definition 3.10 (Variance and standard deviation)

The **variance** of X measures the spread of X around its mean:

$$\text{Var}(X) \triangleq \sigma_X^2 \triangleq \mathbb{E}[(X - \mathbb{E}[X])^2]. \quad (3.19)$$

The **standard deviation** is $\sigma_X = \sqrt{\text{Var}(X)}$.

An equivalent computational formula (often easier to use):

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2. \quad (3.20)$$

Derivation of the computational formula. Expand the square using linearity:

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2. \end{aligned} \quad (3.21)$$

Note that $\mathbb{E}[X]$ is a *constant*, so $\mathbb{E}[2X \cdot \mathbb{E}[X]] = 2\mathbb{E}[X] \cdot \mathbb{E}[X]$ and $\mathbb{E}[(\mathbb{E}[X])^2] = (\mathbb{E}[X])^2$. $\square$

Theorem 3.5 (Scaling property of variance)

For constants $a, c \in \mathbb{R}$:

$$\text{Var}(aX + c) = a^2 \text{Var}(X). \quad (3.22)$$

Adding a constant shifts the mean but does not change the spread. Scaling by a scales the variance by a^2 (and the standard deviation by $|a|$).

Proof. Let $\mu = \mathbb{E}[X]$. Then $\mathbb{E}[aX + c] = a\mu + c$, so

$$\text{Var}(aX + c) = \mathbb{E}[(aX + c - a\mu - c)^2] = \mathbb{E}[(a(X - \mu))^2] = a^2 \mathbb{E}[(X - \mu)^2] = a^2 \text{Var}(X). \quad (3.23)$$

□

Coming up next lecture: covariance $\text{Cov}(X, Y)$, variance of sums, and the beginning of continuous random variables.

End of Lecture 3.

Lecture 4: Variance, Covariance, and Common Discrete Distributions

EECS 126 (UC Berkeley)

Spring 2026 — January 29, 2026

Reading: Bertsekas & Tsitsiklis, Ch. 2; §4.2–4.3. Ch. 3 next week.

These are my personal notes from Professor Kannan Ramchandran's EECS 126 (Probability and Random Processes) course at UC Berkeley.

Last lecture we defined expectation and variance for a single random variable. In this lecture we study how variance behaves for *sums* of random variables, which leads to covariance and the correlation coefficient. We then compute the mean and variance of the four workhorse discrete distributions (Bernoulli, binomial, geometric, Poisson), apply these to the coupon collector problem, prove the memoryless property of the geometric distribution, study how Poisson random variables merge and split, and finally introduce conditional expectation.

4.1 Variance of Sums and Covariance

Recall from Lecture 3 the two forms of the variance,

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2, \quad (4.1)$$

and the scaling property

$$\text{Var}(aX + c) = a^2 \text{Var}(X). \quad (4.2)$$

Shifts by a constant c are not recorded in variance calculations: shifting every outcome by c moves the mean by c as well, so the deviations $X - \mathbb{E}[X]$ are unchanged.

The natural next question is: what is the variance of a *sum* $X + Y$? Unlike the mean, the answer depends on how X and Y interact.

Theorem 4.1 (Variance of a sum)

For any random variables X, Y with finite variances,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]). \quad (4.3)$$

Proof.

$$\begin{aligned} \text{Var}(X + Y) &= \mathbb{E}[(X + Y)^2] - (\mathbb{E}[X + Y])^2 \\ &= \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\ &= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - (\mathbb{E}[X])^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - (\mathbb{E}[Y])^2 \\ &= \underbrace{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}_{\text{Var}(X)} + \underbrace{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}_{\text{Var}(Y)} + 2\mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y] \\ &= \text{Var}(X) + \text{Var}(Y) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]). \quad \square \end{aligned}$$

No independence was used. The derivation above holds for *any* X and Y . Independence only enters at the very end, where it makes the cross term vanish (Corollary 4.1 below).

The cross term in (4.3) is important enough to get its own name.

Definition 4.1 (Covariance)

The **covariance** of random variables X and Y is

$$\text{Cov}(X, Y) \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \quad (4.4)$$

In particular, $\text{Cov}(X, X) = \text{Var}(X)$. With this notation,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y). \quad (4.5)$$

Equivalence of the two forms. Write $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$, which are constants. Expanding the product and using linearity,

$$\begin{aligned} \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] &= \mathbb{E}[XY - \mu_Y X - \mu_X Y + \mu_X \mu_Y] \\ &= \mathbb{E}[XY] - \mu_Y \mathbb{E}[X] - \mu_X \mathbb{E}[Y] + \mu_X \mu_Y \\ &= \mathbb{E}[XY] - \mu_Y \mu_X - \mu_X \mu_Y + \mu_X \mu_Y \\ &= \mathbb{E}[XY] - \mu_X \mu_Y. \end{aligned} \quad \square$$

Interpreting the sign. $\text{Cov}(X, Y) > 0$ means X and Y tend to be above their means together (and below their means together). $\text{Cov}(X, Y) < 0$ means when one is above its mean, the other tends to be below. For a sum, positive covariance *inflates* the spread and negative covariance *dampens* it.

Corollary 4.1 (Independent summands)

If $X \perp Y$, then $\text{Cov}(X, Y) = 0$, and therefore

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y). \quad (4.6)$$

Proof. In Lecture 3 we proved that $X \perp Y$ implies $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Hence $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$, and the cross term in (4.3) vanishes. $\square$

Corollary 4.2 (Variance of a sum of many random variables)

For random variables $X_1, \dots, X_n$ with finite variances,

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i < j} \text{Cov}(X_i, X_j). \quad (4.7)$$

In particular, if the X_i are independent (in fact, *pairwise uncorrelated* is enough), then

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i). \quad (4.8)$$

Proof. Let $\mu_i = \mathbb{E}[X_i]$. Since $\mathbb{E}[\sum_i X_i] = \sum_i \mu_i$, the centered sum is $\sum_i (X_i - \mu_i)$. Squaring a sum gives a double sum over all pairs (i, j) :

$$\begin{aligned} \text{Var}\left(\sum_i X_i\right) &= \mathbb{E}\left[\left(\sum_i (X_i - \mu_i)\right)^2\right] = \mathbb{E}\left[\sum_i \sum_j (X_i - \mu_i)(X_j - \mu_j)\right] \\ &= \sum_i \sum_j \text{Cov}(X_i, X_j). \end{aligned} \quad (4.9)$$

The diagonal terms $i = j$ give $\text{Cov}(X_i, X_i) = \text{Var}(X_i)$. The off-diagonal terms come in symmetric pairs, $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$, so they contribute $2 \sum_{i < j} \text{Cov}(X_i, X_j)$. If every pair is uncorrelated, all off-diagonal terms are zero. $\square$

Uncorrelated does not imply independent. Covariance only detects *linear* dependence. For example, let X be uniform on $\{-1, 0, 1\}$ and $Y = X^2$. Then $\mathbb{E}[X] = 0$ and $\mathbb{E}[XY] = \mathbb{E}[X^3] = \frac{1}{3}(-1) + \frac{1}{3}(0) + \frac{1}{3}(1) = 0$, so $\text{Cov}(X, Y) = 0 - 0 \cdot \mathbb{E}[Y] = 0$. But Y is completely determined by X : for instance $\mathbb{P}(Y = 0 \mid X = 0) = 1 \neq \frac{1}{3} = \mathbb{P}(Y = 0)$. This was a question Ramchandran asked us at end of previous lecture.

4.2 The Correlation Coefficient

Covariance has units (units of X times units of Y), so its size alone does not tell us how *strongly* two variables are related. Normalizing by the standard deviations removes the units.

Definition 4.2 (Correlation coefficient)

For random variables X, Y with $\text{Var}(X) > 0$ and $\text{Var}(Y) > 0$, the **correlation coefficient** is

$$\rho(X, Y) \triangleq \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}}. \quad (4.10)$$

Ramchandran's vector analogy. Think of random variables as vectors, with “dot product” $\langle U, V \rangle = \mathbb{E}[UV]$ and “length” $\sqrt{\mathbb{E}[U^2]}$. For the centered variables $U = X - \mathbb{E}[X]$ and $V = Y - \mathbb{E}[Y]$:

- the dot product is $\mathbb{E}[UV] = \text{Cov}(X, Y)$;
- the lengths are $\sqrt{\mathbb{E}[U^2]} = \sqrt{\text{Var}(X)}$ and $\sqrt{\mathbb{E}[V^2]} = \sqrt{\text{Var}(Y)}$.

So ρ plays the role of $\cos \theta = \frac{u \cdot v}{|u| |v|}$, the cosine of the angle between two vectors. For ordinary vectors, the Cauchy–Schwarz inequality $|u \cdot v| \leq |u| |v|$ forces $-1 \leq \cos \theta \leq 1$. The same inequality, proved for random variables below, forces $-1 \leq \rho \leq 1$.

Lemma 4.1 (Cauchy–Schwarz inequality for random variables)

For any random variables U, V with $\mathbb{E}[U^2] < \infty$ and $\mathbb{E}[V^2] < \infty$,

$$(\mathbb{E}[UV])^2 \leq \mathbb{E}[U^2] \mathbb{E}[V^2]. \quad (4.11)$$

Proof. **Case 1:** $\mathbb{E}[V^2] = 0$. Since $V^2 \geq 0$ and its mean is 0, we must have $V = 0$ with probability 1. Then $UV = 0$ with probability 1, so $\mathbb{E}[UV] = 0$ and both sides equal 0.

Case 2: $\mathbb{E}[V^2] > 0$. For every real number t , the random variable $(U - tV)^2$ is nonnegative, so its expectation is nonnegative. Expanding with linearity:

$$0 \leq \mathbb{E}[(U - tV)^2] = \mathbb{E}[U^2] - 2t\mathbb{E}[UV] + t^2\mathbb{E}[V^2]. \quad (4.12)$$

This holds for every t , so choose the t that minimizes the right-hand side, namely $t^* = \mathbb{E}[UV]/\mathbb{E}[V^2]$. Substituting,

$$\begin{aligned} 0 &\leq \mathbb{E}[U^2] - 2 \frac{\mathbb{E}[UV]}{\mathbb{E}[V^2]} \mathbb{E}[UV] + \frac{(\mathbb{E}[UV])^2}{(\mathbb{E}[V^2])^2} \mathbb{E}[V^2] \\ &= \mathbb{E}[U^2] - 2 \frac{(\mathbb{E}[UV])^2}{\mathbb{E}[V^2]} + \frac{(\mathbb{E}[UV])^2}{\mathbb{E}[V^2]} \\ &= \mathbb{E}[U^2] - \frac{(\mathbb{E}[UV])^2}{\mathbb{E}[V^2]}. \end{aligned} \quad (4.13)$$

Multiplying both sides by $\mathbb{E}[V^2] > 0$ gives $(\mathbb{E}[UV])^2 \leq \mathbb{E}[U^2] \mathbb{E}[V^2]$. $\square$

Theorem 4.2 (Range of the correlation coefficient)

Whenever $\rho(X, Y)$ is defined,

$$-1 \leq \rho(X, Y) \leq 1. \quad (4.14)$$

Proof. Apply Cauchy–Schwarz to the centered variables $U = X - \mathbb{E}[X]$ and $V = Y - \mathbb{E}[Y]$. By definition, $\mathbb{E}[UV] = \text{Cov}(X, Y)$, $\mathbb{E}[U^2] = \text{Var}(X)$, and $\mathbb{E}[V^2] = \text{Var}(Y)$, so

$$(\text{Cov}(X, Y))^2 \leq \text{Var}(X) \text{Var}(Y). \quad (4.15)$$

Taking square roots, $|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}$. Dividing by the (positive) right-hand side gives $|\rho(X, Y)| \leq 1$. $\square$

Extreme and zero values. $|\rho| = 1$ exactly when equality holds in Cauchy–Schwarz, which happens when $U - t^*V = 0$ with probability 1, i.e. when Y is an exact linear function $Y = aX + b$ of X (with $\rho = +1$ if $a > 0$ and $\rho = -1$ if $a < 0$). At the other end, $\rho = 0$ means *uncorrelated*, which, as we just saw, is weaker than independent.

4.3 Mean and Variance of Common Distributions

We now compute $\mathbb{E}[X]$ and $\text{Var}(X)$ for the four distributions introduced in Lecture 3. A recurring theme: the direct sums are often messy, and a better *representation* of the random variable makes the computation easy.

Bernoulli

Theorem 4.3 (Bernoulli mean and variance)

If $X \sim \text{Bernoulli}(p)$, then

$$\mathbb{E}[X] = p, \quad \text{Var}(X) = p(1 - p). \quad (4.16)$$

Proof. X takes only the values 0 and 1, with $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$.

$$\mathbb{E}[X] = 1 \cdot \mathbb{P}(X = 1) + 0 \cdot \mathbb{P}(X = 0) = 1 \cdot p + 0 = p. \quad (4.17)$$

For the second moment, by LOTUS,

$$\mathbb{E}[X^2] = \sum_x x^2 \mathbb{P}(X = x) = 1^2 \cdot \mathbb{P}(X = 1) + 0^2 \cdot \mathbb{P}(X = 0) = p. \quad (4.18)$$

Therefore

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1 - p). \quad (4.19)$$

□

Binomial

Theorem 4.4 (Binomial mean and variance)

If $X \sim \text{Binomial}(n, p)$, then

$$\mathbb{E}[X] = np, \quad \text{Var}(X) = np(1 - p). \quad (4.20)$$

The direct approach is

$$\mathbb{E}[X] = \sum_{x=0}^n x \cdot \mathbb{P}(X = x) = \sum_{x=0}^n x \binom{n}{x} p^x (1 - p)^{n-x}, \quad (4.21)$$

which is doable but unpleasant. Instead, think of the binomial as a *sum of n Bernoulli trials*.

Proof. Let $X_i = \mathbf{1}\{\text{trial } i \text{ is a success}\}$ for $i = 1, \dots, n$. Each $X_i \sim \text{Bernoulli}(p)$, the X_i are independent (the trials are independent), and the total number of successes is

$$X = X_1 + X_2 + \dots + X_n. \quad (4.22)$$

Mean. By linearity of expectation (which needs no independence),

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i] = \sum_{i=1}^n p = np. \quad (4.23)$$

Variance. Because the X_i are independent, Corollary 4.1 lets us add the variances:

$$\text{Var}(X) = \sum_{i=1}^n \text{Var}(X_i) = \sum_{i=1}^n p(1 - p) = np(1 - p). \quad (4.24)$$

□

Applying direct sum gives us the same result: Using the identity $x \binom{n}{x} = n \binom{n-1}{x-1}$ for

$x \geq 1$ (the $x = 0$ term is zero),

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=1}^n n \binom{n-1}{x-1} p^x (1-p)^{n-x} = np \sum_{x=1}^n \binom{n-1}{x-1} p^{x-1} (1-p)^{(n-1)-(x-1)} \\ &= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{(n-1)-j} = np (p + (1-p))^{n-1} = np,\end{aligned}\quad (4.25)$$

where we substituted $j = x - 1$ and used the binomial theorem. The indicator method reaches the same answer with far less algebra.

Geometric

Recall that $X \sim \text{Geometric}(p)$ counts the number of independent trials up to and including the first success:

$$p_X(k) = \mathbb{P}(X = k) = (1-p)^{k-1} p, \quad k = 1, 2, 3, \dots \quad (4.26)$$

Reading the PMF. Here k marks the trial on which the first Success occurs ($k = S$). So $p_X(k)$ is the probability that the first success happens on trial k : the first $k - 1$ trials are failures (probability $(1-p)^{k-1}$) and trial k is a success (probability p). Note that $p_X(k)$ is *not* the probability of k successes; a geometric random variable always records exactly one success, and k counts how long we waited for it.

Definition 4.3 (Survival function)

For an integer-valued random variable X , the **survival function** (or *tail*) is $k \mapsto \mathbb{P}(X > k)$. For $X \sim \text{Geometric}(p)$,

$$\mathbb{P}(X > k) = (1-p)^k, \quad k = 0, 1, 2, \dots \quad (4.27)$$

Proof. The event $\{X > k\}$ means “no success in the first k trials,” i.e. we tried k times and failed every time. The trials are independent and each fails with probability $1 - p$, so $\mathbb{P}(X > k) = \underbrace{(1-p) \cdots (1-p)}_{k \text{ times}} = (1-p)^k$. $\square$

Theorem 4.5 (Geometric mean)

If $X \sim \text{Geometric}(p)$ with $0 < p \leq 1$, then

$$\mathbb{E}[X] = \frac{1}{p}. \quad (4.28)$$

The direct sum $\mathbb{E}[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p$ requires a trick. A better method is the tail sum formula from Lecture 3, which applies because X takes values in the nonnegative integers.

Proof. By the tail sum formula,

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k). \quad (4.29)$$

For integer-valued X , $\{X \geq k\} = \{X > k - 1\}$. Substituting $j = k - 1$ (so j runs from 0 to ∞),

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X > k - 1) = \sum_{j=0}^{\infty} \mathbb{P}(X > j). \quad (4.30)$$

Now plug in the survival function (4.27):

$$\mathbb{E}[X] = \sum_{j=0}^{\infty} (1 - p)^j. \quad (4.31)$$

This is a geometric series $\sum_{j \geq 0} x^j$ with ratio $x = 1 - p$. Since $0 < p \leq 1$ we have $|x| < 1$, so the series converges to $1/(1 - x)$:

$$\mathbb{E}[X] = \frac{1}{1 - (1 - p)} = \frac{1}{p}. \quad (4.32)$$

□

Exam note: tail sum formula will be on exam. For a random variable X taking values in $\{0, 1, 2, \dots\}$,

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k) = \sum_{k=0}^{\infty} \mathbb{P}(X > k). \quad (4.33)$$

It expresses the mean as the sum of the tail probabilities. It follows from writing $X = \sum_{k=1}^{\infty} \mathbf{1}\{X \geq k\}$ and taking expectations of both sides.

The direct sum, for comparison. Differentiating the geometric series $\sum_{k \geq 0} x^k = \frac{1}{1-x}$ term by term (valid for $|x| < 1$) gives $\sum_{k \geq 1} kx^{k-1} = \frac{1}{(1-x)^2}$. With $x = 1 - p$,

$$\mathbb{E}[X] = p \sum_{k=1}^{\infty} k(1 - p)^{k-1} = p \cdot \frac{1}{(1 - (1 - p))^2} = p \cdot \frac{1}{p^2} = \frac{1}{p}. \quad (4.34)$$

The variance of the geometric is derived in Section 4.5, using the memoryless property.

Poisson

Theorem 4.6 (Poisson mean and variance)

If $X \sim \text{Poisson}(\lambda)$, so that $p_X(k) = \frac{e^{-\lambda} \lambda^k}{k!}$ for $k = 0, 1, 2, \dots$, then

$$\mathbb{E}[X] = \lambda, \quad \text{Var}(X) = \lambda. \quad (4.35)$$

Both computations rely on the Taylor series of the exponential,

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots, \quad (4.36)$$

and on the cancellation $\frac{k}{k!} = \frac{1}{(k-1)!}$ for $k \geq 1$.

Proof of the mean.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{\infty} k \cdot \frac{e^{-\lambda} \lambda^k}{k!} \\
&= \sum_{k=1}^{\infty} k \cdot \frac{e^{-\lambda} \lambda^k}{k!} && \text{(the } k=0 \text{ term is 0)} \\
&= \sum_{k=1}^{\infty} \frac{e^{-\lambda} \lambda^k}{(k-1)!} && \text{(cancel } k \text{ against } k!) \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} && \text{(factor out } \lambda e^{-\lambda}) \\
&= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} && \text{(substitute } j = k-1) \\
&= \lambda e^{-\lambda} e^{\lambda} = \lambda. && \text{(Taylor series (4.36))} \quad \square
\end{aligned}$$

For the variance, computing $\mathbb{E}[X^2] = \sum k^2 \frac{e^{-\lambda} \lambda^k}{k!}$ directly is awkward because k^2 does not cancel cleanly against $k!$. The fix is the **factorial moment trick**: since $X^2 = X(X-1) + X$, linearity gives

$$\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X], \quad (4.37)$$

and $k(k-1)$ cancels against $k!$ just as k did.

Proof of the variance. Step 1: the factorial moment.

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{k=0}^{\infty} k(k-1) \frac{e^{-\lambda} \lambda^k}{k!} \\
&= \sum_{k=2}^{\infty} k(k-1) \frac{e^{-\lambda} \lambda^k}{k!} && \text{(the } k=0, 1 \text{ terms are 0)} \\
&= \sum_{k=2}^{\infty} \frac{e^{-\lambda} \lambda^k}{(k-2)!} && \text{(cancel } k(k-1) \text{ against } k!) \\
&= e^{-\lambda} \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} && \text{(factor out } e^{-\lambda} \lambda^2) \\
&= e^{-\lambda} \lambda^2 e^{\lambda} = \lambda^2. && \text{(same trick: substitute } j = k-2) \quad (4.38)
\end{aligned}$$

Step 2: the second moment. By (4.37),

$$\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] = \lambda^2 + \lambda. \quad (4.39)$$

Step 3: the variance.

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = (\lambda^2 + \lambda) - (\lambda)^2 = \lambda. \quad (4.40)$$

□

Mean equals variance (Poisson distribution). The Poisson is the only one of our four distributions whose mean and variance are equal. This is consistent with the Poisson being the limit of $\text{Binomial}(n, \lambda/n)$: the binomial variance is $n \cdot \frac{\lambda}{n} \left(1 - \frac{\lambda}{n}\right) = \lambda \left(1 - \frac{\lambda}{n}\right) \rightarrow \lambda$ as $n \rightarrow \infty$.

4.4 The Coupon Collector Problem

Example 4.1 (Coupon collector)

There are N different coupons you want to collect, found inside cereal boxes. Each box contains exactly one coupon, whose type is **uniformly random** among the N types and **independent** of all other boxes. You buy boxes until you have received all N distinct coupons. What is the expected number of boxes you must buy?

Let T be the total time (number of boxes) to collect all N coupons. Break the process into N stages. For $i = 1, \dots, N$, let

$$T_i = \# \text{ of boxes bought to obtain the } i\text{-th new coupon}, \quad (4.41)$$

counting from right after the $(i-1)$ -th new coupon was obtained. Then the total time is the sum of the stage times,

$$T = \sum_{i=1}^N T_i. \quad (4.42)$$

Each stage is geometric. During stage i we already hold $i-1$ distinct coupons, so $N - (i-1) = N - i + 1$ types would be new. Each box is independently a “success” (a new type) with probability

$$p_i = \frac{N - i + 1}{N}, \quad \text{so} \quad T_i \sim \text{Geometric}\left(\frac{N - i + 1}{N}\right). \quad (4.43)$$

- $i = 1$: $T_1 \sim \text{Geometric}\left(\frac{N-1+1}{N}\right) = \text{Geometric}(1)$, so $T_1 = 1$. The first box always gives a new coupon.
- $i = 2$: $T_2 \sim \text{Geometric}\left(\frac{N-2+1}{N}\right) = \text{Geometric}\left(\frac{N-1}{N}\right)$.
- $i = N$: $T_N \sim \text{Geometric}\left(\frac{1}{N}\right)$. The last coupon is the hardest to find.

Expected total. Since $\mathbb{E}[T_i] = 1/p_i = \frac{N}{N-i+1}$, linearity of expectation gives

$$\mathbb{E}[T] = \mathbb{E}\left[\sum_{i=1}^N T_i\right] = \sum_{i=1}^N \mathbb{E}[T_i] = \sum_{i=1}^N \frac{N}{N-i+1}. \quad (4.44)$$

Write out the terms, from $i = 1$ (denominator N) to $i = N$ (denominator 1):

$$\begin{aligned} \mathbb{E}[T] &= \frac{N}{N} + \frac{N}{N-1} + \frac{N}{N-2} + \cdots + \frac{N}{2} + \frac{N}{1} \\ &= N \left(\frac{1}{N} + \frac{1}{N-1} + \cdots + \frac{1}{2} + \frac{1}{1} \right) = N \sum_{j=1}^N \frac{1}{j}, \end{aligned} \quad (4.45)$$

where we substituted $j = N - i + 1$ (as i runs from 1 to N , j runs from N down to 1). The sum is the N -th **harmonic number** $H_N = \sum_{j=1}^N \frac{1}{j}$, so

$$\mathbb{E}[T] = N H_N. \quad (4.46)$$

Lemma 4.2 (Size of the harmonic numbers)

For every $N \geq 1$,

$$\ln(N+1) \leq H_N \leq 1 + \ln N, \quad (4.47)$$

and more precisely $H_N = \ln N + \gamma + o(1)$, where $\gamma \approx 0.5772$ is the **Euler–Mascheroni constant**.

Proof of the bounds. Compare each term $1/j$ with the area under the curve $1/x$. Since $1/x$ is decreasing, on the interval $[j, j+1]$ we have $1/x \leq 1/j$, and on $[j-1, j]$ we have $1/x \geq 1/j$. Hence

$$\int_j^{j+1} \frac{dx}{x} \leq \frac{1}{j} \leq \int_{j-1}^j \frac{dx}{x} \quad (j \geq 2 \text{ for the right inequality}). \quad (4.48)$$

Summing the left inequality over $j = 1, \dots, N$ gives $H_N \geq \int_1^{N+1} \frac{dx}{x} = \ln(N+1)$. Summing the right inequality over $j = 2, \dots, N$ and adding the $j = 1$ term gives $H_N \leq 1 + \int_1^N \frac{dx}{x} = 1 + \ln N$. (The sharper statement with γ is a classical result we take as given.) $\square$

Corollary 4.3 (Coupon collector asymptotics)

$$\mathbb{E}[T] = N H_N \approx N \ln N + \gamma N \approx N \ln N. \quad (4.49)$$

So we need to buy about a factor of $\ln N$ more boxes than the N coupons we are trying to collect.

Numbers and intuition. For $N = 50$, $H_{50} \approx 4.499$, so $\mathbb{E}[T] \approx 225$ boxes, about $4.5\times$ the number of coupon types. Most of the cost comes at the end: the final coupon alone takes $\mathbb{E}[T_N] = N$ boxes on average, while the entire first half of the collection takes fewer than N boxes in total ($\sum_{i \leq N/2} \frac{N}{N-i+1} \leq \frac{N}{2} \cdot 2 = N$). This $N \ln N$ behavior shows up whenever we need to “hit every bin” by random sampling: hashing, load balancing, and covering every node in a network.

4.5 The Memoryless Property

Theorem 4.7 (Memoryless property of the geometric)

Let $X \sim \text{Geometric}(p)$ with $0 < p < 1$. For all integers $s, t \geq 0$,

$$\mathbb{P}(X > s+t \mid X > s) = \mathbb{P}(X > t). \quad (4.50)$$

In words: given that we have already failed s times, the probability of failing t more times is the same as if we were starting over from scratch.

Proof. First note that $\mathbb{P}(X > s) = (1-p)^s > 0$, so the conditional probability is well defined. By the definition of conditional probability,

$$\mathbb{P}(X > s+t \mid X > s) = \frac{\mathbb{P}(\{X > s+t\} \cap \{X > s\})}{\mathbb{P}(X > s)}. \quad (4.51)$$

Since $s+t \geq s$, the event $\{X > s+t\}$ is contained in $\{X > s\}$ (if we have $s+t$ failures, we certainly have s failures), so the intersection is just $\{X > s+t\}$:

$$\mathbb{P}(X > s+t \mid X > s) = \frac{\mathbb{P}(X > s+t)}{\mathbb{P}(X > s)}. \quad (4.52)$$

Both numerator and denominator are values of the geometric survival function (4.27):

$$\frac{\mathbb{P}(X > s+t)}{\mathbb{P}(X > s)} = \frac{(1-p)^{s+t}}{(1-p)^s} = (1-p)^t, \quad (4.53)$$

which is exactly the survival function of the geometric evaluated at t , i.e. $\mathbb{P}(X > t)$. $\square$

Two remarks on memorylessness.

1. **The remaining wait is again geometric.** Given $X > s$, the number of additional trials $X - s$ has the same distribution as X : for $k \geq 1$,

$$\mathbb{P}(X - s = k \mid X > s) = \frac{\mathbb{P}(X = s+k)}{\mathbb{P}(X > s)} = \frac{(1-p)^{s+k-1}p}{(1-p)^s} = (1-p)^{k-1}p. \quad (4.54)$$

This is the form we use in the derivations below.

2. **Uniqueness.** The geometric is the *only* memoryless distribution on $\{1, 2, 3, \dots\}$. Its continuous analog, the exponential distribution, will be the only memoryless continuous distribution.

To use memorylessness to recompute $\mathbb{E}[X]$, we need to split an expectation according to which of several events occurred. This is the expectation version of the law of total probability from Lecture 2.

Theorem 4.8 (Total expectation theorem)

Let $A_1, \dots, A_n$ be a partition of Ω with $\mathbb{P}(A_i) > 0$ for each i , and define the conditional expectation given an event, $\mathbb{E}[X \mid A_i] = \sum_x x \mathbb{P}(X = x \mid A_i)$. Then

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{P}(A_i) \mathbb{E}[X \mid A_i]. \quad (4.55)$$

Proof. For each value x , the law of total probability gives $\mathbb{P}(X = x) = \sum_i \mathbb{P}(A_i) \mathbb{P}(X = x \mid A_i)$. Substituting into the definition of expectation and swapping the order of summation (justified by absolute convergence),

$$\begin{aligned} \mathbb{E}[X] &= \sum_x x \mathbb{P}(X = x) = \sum_x x \sum_i \mathbb{P}(A_i) \mathbb{P}(X = x \mid A_i) \\ &= \sum_i \mathbb{P}(A_i) \sum_x x \mathbb{P}(X = x \mid A_i) = \sum_i \mathbb{P}(A_i) \mathbb{E}[X \mid A_i]. \end{aligned} \quad \square$$

Example 4.2 (Geometric mean via memorylessness)

Let $X \sim \text{Geometric}(p)$. Condition on the outcome of the first trial, using the partition $\{X = 1\}$ (first trial succeeds) and $\{X > 1\}$ (first trial fails):

$$\mathbb{E}[X] = \mathbb{P}(X = 1) \mathbb{E}[X \mid X = 1] + \mathbb{P}(X > 1) \mathbb{E}[X \mid X > 1]. \quad (4.56)$$

- $\mathbb{P}(X = 1) = p$ (success on the first try), and then $X = 1$ exactly, so $\mathbb{E}[X \mid X = 1] = 1$.
- $\mathbb{P}(X > 1) = 1 - p$. If the first trial failed, the geometric “tells us we start again”: by memorylessness the remaining wait $X - 1$ is again $\text{Geometric}(p)$, and we have already

spent one trial. So

$$\mathbb{E}[X \mid X > 1] = \mathbb{E}[(X - 1) + 1 \mid X > 1] = \mathbb{E}[X] + 1. \quad (4.57)$$

Substituting,

$$\begin{aligned} \mathbb{E}[X] &= p \cdot 1 + (1 - p)(\mathbb{E}[X] + 1) \\ &= p + \mathbb{E}[X] + 1 - p\mathbb{E}[X] - p \\ &= \mathbb{E}[X] + 1 - p\mathbb{E}[X]. \end{aligned} \quad (4.58)$$

Subtracting $\mathbb{E}[X]$ from both sides gives $0 = 1 - p\mathbb{E}[X]$, i.e. $p\mathbb{E}[X] = 1$, so

$$\mathbb{E}[X] = \frac{1}{p}, \quad (4.59)$$

matching the tail sum computation.

Two checks on this argument.

1. **Finiteness.** Subtracting $\mathbb{E}[X]$ from both sides is only legitimate because $\mathbb{E}[X] < \infty$, which we already know from the tail sum. (If $\mathbb{E}[X] = \infty$, the equation $\infty = \infty + 1 - \infty$ says nothing.)
2. **Sanity check the answer.** Any correct formula must satisfy $\mathbb{E}[X] \geq 1$ (you need at least one trial) and $\mathbb{E}[X] \rightarrow \infty$ as $p \rightarrow 0$. A sign slip in the algebra can produce $\mathbb{E}[X] = \frac{1}{1+p}$, which fails both tests, since it is less than 1 for every $p > 0$.

Theorem 4.9 (Geometric variance (supplementary))

If $X \sim \text{Geometric}(p)$, then

$$\text{Var}(X) = \frac{1 - p}{p^2}. \quad (4.60)$$

Proof. We use the same first-trial conditioning, now on $\mathbb{E}[X^2]$ (which is finite, since the tail of X decays geometrically).

$$\mathbb{E}[X^2] = \mathbb{P}(X = 1)\mathbb{E}[X^2 \mid X = 1] + \mathbb{P}(X > 1)\mathbb{E}[X^2 \mid X > 1]. \quad (4.61)$$

- Given $X = 1$: $X^2 = 1$, so $\mathbb{E}[X^2 \mid X = 1] = 1$.
- Given $X > 1$: by memorylessness, $X = 1 + X'$ where $X' \sim \text{Geometric}(p)$ has the same distribution as X . Expanding the square,

$$\mathbb{E}[X^2 \mid X > 1] = \mathbb{E}[(1 + X')^2] = 1 + 2\mathbb{E}[X'] + \mathbb{E}[X'^2] = 1 + 2\mathbb{E}[X] + \mathbb{E}[X^2]. \quad (4.62)$$

Substituting, with $\mathbb{E}[X] = 1/p$:

$$\begin{aligned} \mathbb{E}[X^2] &= p \cdot 1 + (1 - p)(1 + 2\mathbb{E}[X] + \mathbb{E}[X^2]) \\ &= p + (1 - p) + 2(1 - p)\mathbb{E}[X] + (1 - p)\mathbb{E}[X^2] \\ &= 1 + \frac{2(1 - p)}{p} + (1 - p)\mathbb{E}[X^2]. \end{aligned} \quad (4.63)$$

Move the $\mathbb{E}[X^2]$ terms to the left:

$$\mathbb{E}[X^2] - (1-p)\mathbb{E}[X^2] = p\mathbb{E}[X^2] = 1 + \frac{2(1-p)}{p} = \frac{p+2-2p}{p} = \frac{2-p}{p}. \quad (4.64)$$

Dividing by p , $\mathbb{E}[X^2] = \frac{2-p}{p^2}$. Finally,

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}. \quad (4.65)$$

□

Cross-check with the factorial moment trick. Differentiating $\sum_{k \geq 0} x^k = \frac{1}{1-x}$ twice gives $\sum_{k \geq 2} k(k-1)x^{k-2} = \frac{2}{(1-x)^3}$. With $x = 1-p$,

$$\mathbb{E}[X(X-1)] = \sum_{k=2}^{\infty} k(k-1)(1-p)^{k-1}p = p(1-p) \sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2} = p(1-p) \cdot \frac{2}{p^3} = \frac{2(1-p)}{p^2}, \quad (4.66)$$

so $\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] = \frac{2(1-p)}{p^2} + \frac{p}{p^2} = \frac{2-p}{p^2}$, in agreement.

4.6 Poisson Merging and Splitting

Poisson random variables model counts of arrivals (packets at a router, calls at a switchboard, customers at a queue). Two structural facts make them especially convenient: independent Poisson counts *merge* into a Poisson count, and randomly *splitting* a Poisson count produces independent Poisson counts.

Theorem 4.10 (Poisson merging (sum of independent Poissons))

Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ be independent. Then

$$X + Y \sim \text{Poisson}(\lambda + \mu), \quad (4.67)$$

with mean $\mathbb{E}[X + Y] = \lambda + \mu$.

Proof (supplementary). Fix $n \in \{0, 1, 2, \dots\}$. Since $X, Y \geq 0$, the event $\{X + Y = n\}$ splits into

the disjoint events $\{X = k, Y = n - k\}$ for $k = 0, \dots, n$:

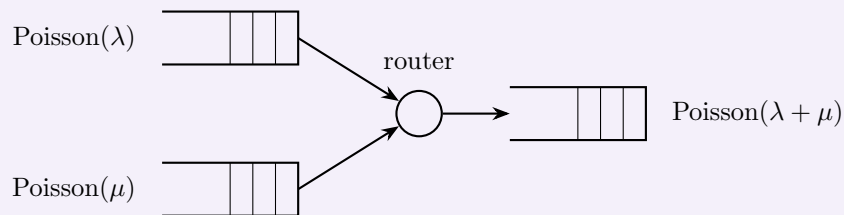
$$\begin{aligned}
 \mathbb{P}(X + Y = n) &= \sum_{k=0}^n \mathbb{P}(X = k, Y = n - k) \\
 &= \sum_{k=0}^n \mathbb{P}(X = k) \mathbb{P}(Y = n - k) && \text{(independence)} \\
 &= \sum_{k=0}^n \frac{e^{-\lambda} \lambda^k}{k!} \cdot \frac{e^{-\mu} \mu^{n-k}}{(n-k)!} && \text{(Poisson PMFs)} \\
 &= e^{-(\lambda+\mu)} \sum_{k=0}^n \frac{\lambda^k \mu^{n-k}}{k! (n-k)!} && \text{(collect exponentials)} \\
 &= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{k=0}^n \frac{n!}{k! (n-k)!} \lambda^k \mu^{n-k} && \text{(multiply and divide by } n!) \\
 &= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{k=0}^n \binom{n}{k} \lambda^k \mu^{n-k} \\
 &= \frac{e^{-(\lambda+\mu)} (\lambda + \mu)^n}{n!}. && \text{(binomial theorem)} \tag{4.68}
 \end{aligned}$$

This is exactly the $\text{Poisson}(\lambda + \mu)$ PMF evaluated at n . The mean $\lambda + \mu$ also follows directly from linearity of expectation. $\square$

Corollary 4.4 (Many independent Poissons)

If $X_1, \dots, X_m$ are independent with $X_i \sim \text{Poisson}(\lambda_i)$, then $\sum_{i=1}^m X_i \sim \text{Poisson}(\sum_{i=1}^m \lambda_i)$. (Apply the theorem repeatedly; $X_1 + \dots + X_k$ is independent of X_{k+1} .)

Networking picture: merging. Two independent packet streams feed a router through separate buffers. If the numbers of packets arriving in each buffer during an interval are $\text{Poisson}(\lambda)$ and $\text{Poisson}(\mu)$, the number of packets in the merged output buffer is $\text{Poisson}(\lambda + \mu)$.



Splitting is the reverse operation: each arriving packet is routed to one of two outputs by an independent coin flip. Given the total count $X = n$, the number Y sent to the first output is the number of successes in n independent Bernoulli(p) trials. This is a “binomial with a random number of trials.”

Theorem 4.11 (Poisson splitting)

Let $X \sim \text{Poisson}(\lambda)$, and suppose that, conditioned on $X = n$, the random variable Y satisfies

$$Y \mid X = n \sim \text{Binomial}(n, p). \quad (4.69)$$

Let $Z = X - Y$ be the remainder. Then

$$Y \sim \text{Poisson}(\lambda p), \quad Z \sim \text{Poisson}(\lambda(1 - p)), \quad \text{and} \quad Y \perp Z. \quad (4.70)$$

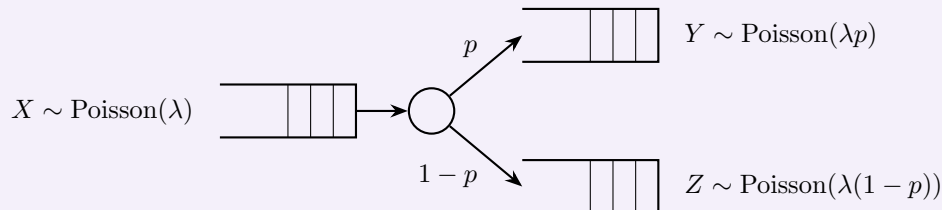
Proof (supplementary). We compute the joint PMF of (Y, Z) and show that it factors. Fix integers $a, b \geq 0$. The event $\{Y = a, Z = b\}$ is the same as $\{X = a + b, Y = a\}$ (the total is $a + b$, of which a went to the first output). By the chain rule,

$$\begin{aligned} \mathbb{P}(Y = a, Z = b) &= \mathbb{P}(X = a + b) \mathbb{P}(Y = a \mid X = a + b) \\ &= \frac{e^{-\lambda} \lambda^{a+b}}{(a+b)!} \cdot \binom{a+b}{a} p^a (1-p)^b && \text{(Poisson, then binomial)} \\ &= \frac{e^{-\lambda} \lambda^{a+b}}{(a+b)!} \cdot \frac{(a+b)!}{a! b!} p^a (1-p)^b && \text{(expand the binomial coefficient)} \\ &= e^{-\lambda} \frac{\lambda^a p^a}{a!} \cdot \frac{\lambda^b (1-p)^b}{b!} && \text{(cancel } (a+b)!, \text{ split } \lambda^{a+b}) \\ &= \underbrace{\frac{e^{-\lambda p} (\lambda p)^a}{a!}}_{\text{Poisson}(\lambda p) \text{ PMF at } a} \cdot \underbrace{\frac{e^{-\lambda(1-p)} (\lambda(1-p))^b}{b!}}_{\text{Poisson}(\lambda(1-p)) \text{ PMF at } b}, && \text{(split } e^{-\lambda} = e^{-\lambda p} e^{-\lambda(1-p)}) \end{aligned} \quad (4.71)$$

Marginal of Y . Sum over $b \geq 0$. The second factor is a PMF in b , so it sums to 1, leaving $\mathbb{P}(Y = a) = \frac{e^{-\lambda p} (\lambda p)^a}{a!}$. Thus $Y \sim \text{Poisson}(\lambda p)$. Summing over a instead shows $Z \sim \text{Poisson}(\lambda(1 - p))$.

Independence. The joint PMF equals the product of the two marginals for every (a, b) , which is the definition of $Y \perp Z$. $\square$

Why the independence is surprising. If the total X were a fixed number n , then Y and $Z = n - Y$ would be perfectly (negatively) dependent: learning Y would tell you Z exactly. It is the *randomness of the total count* that makes the two outputs independent. A large value of Y suggests X was large, which exactly offsets what it would otherwise say about Z . This is special to the Poisson.



4.7 Conditional PMFs and Conditional Expectation

Definition 4.4 (Conditional PMF)

Let X, Y be discrete random variables and let y satisfy $p_Y(y) > 0$. The **conditional PMF of X given $Y = y$** is

$$p_{X|Y}(x | y) \triangleq \mathbb{P}(X = x | Y = y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}. \quad (4.72)$$

It is a genuine PMF in x . For each fixed y with $p_Y(y) > 0$:

- $p_{X|Y}(x | y) \geq 0$, since it is a ratio of a probability and a positive number;
- it sums to one over x , by marginalizing the joint PMF:

$$\sum_x p_{X|Y}(x | y) = \frac{1}{p_Y(y)} \sum_x p_{X,Y}(x, y) = \frac{p_Y(y)}{p_Y(y)} = 1. \quad (4.73)$$

So for each fixed value y , “ X given $Y = y$ ” is just another random variable with its own distribution, and we can take its expectation. (Note that the fraction is $p_{X,Y}(x, y)/p_Y(y)$, a ratio of PMF values, rather than $\mathbb{P}(X \cap Y)/\mathbb{P}(Y)$, since X and Y are random variables rather than events.)

Definition 4.5 (Conditional expectation given $Y = y$)

For y with $p_Y(y) > 0$, the **conditional expectation of X given $Y = y$** is

$$\mathbb{E}[X | Y = y] \triangleq \sum_x x \cdot p_{X|Y}(x | y) = \sum_x x \cdot \mathbb{P}(X = x | Y = y). \quad (4.74)$$

Like any expectation, $\mathbb{E}[X | Y = y]$ is just a **real number**.

By “sweeping” through the values of y in (4.74), we get a (possibly different) number $\mathbb{E}[X | Y = y]$ for every value y . That is, $y \mapsto \mathbb{E}[X | Y = y]$ is a *function* of y .

Definition 4.6 (Conditional expectation as a random variable)

Define $g(y) = \mathbb{E}[X | Y = y]$. The **conditional expectation of X given Y** is the random variable

$$\mathbb{E}[X | Y] \triangleq g(Y). \quad (4.75)$$

It is a function of Y , and therefore $\mathbb{E}[X | Y]$ is itself a **random variable**: it takes the value $g(y)$ with probability $p_Y(y)$.

Notation: $\mathbb{E}[X | Y = y]$ **versus** $\mathbb{E}[X | Y]$. These are different objects. $\mathbb{E}[X | Y = y]$ (lowercase y , a specific value) is a number. $\mathbb{E}[X | Y]$ (capital Y) is a random variable.

Example 4.3 (Die roll, conditioned on parity)

Roll a fair six-sided die. Let X be the result and $Y = \mathbf{1}\{X \text{ is even}\}$.

Given $Y = 1$, X is uniform on $\{2, 4, 6\}$, so $\mathbb{E}[X | Y = 1] = \frac{2+4+6}{3} = 4$. Given $Y = 0$, X is uniform on $\{1, 3, 5\}$, so $\mathbb{E}[X | Y = 0] = \frac{1+3+5}{3} = 3$. Hence

$$\mathbb{E}[X | Y] = g(Y) = \begin{cases} 4 & \text{if } Y = 1, \\ 3 & \text{if } Y = 0, \end{cases} \quad \text{i.e.} \quad \mathbb{E}[X | Y] = 3 + Y. \quad (4.76)$$

This is a random variable taking the values 3 and 4, each with probability $\frac{1}{2}$. Notice that its mean is $\frac{1}{2}(3) + \frac{1}{2}(4) = 3.5 = \mathbb{E}[X]$. That is no coincidence, as the next lecture will show.

Law of Iterated Expectation (Tower Rule): TBD next lecture. This was on today's agenda but was not reached in lecture. It states that $\mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[X]$; the total expectation theorem of Section 4.5 is a special case.

4.8 Summary: Common Discrete Distributions

Distribution	PMF $\mathbb{P}(X = k)$	Support	$\mathbb{E}[X]$	$\text{Var}(X)$
Bernoulli(p)	$p^k(1-p)^{1-k}$	$\{0, 1\}$	p	$p(1-p)$
Binomial(n, p)	$\binom{n}{k}p^k(1-p)^{n-k}$	$\{0, \dots, n\}$	np	$np(1-p)$
Geometric(p)	$(1-p)^{k-1}p$	$\{1, 2, \dots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$\frac{e^{-\lambda}\lambda^k}{k!}$	$\{0, 1, 2, \dots\}$	λ	λ

End of Lecture 4.

Lecture 5: Iterated Expectation, Conditional Variance, and Continuous Random Variables

EECS 126 (UC Berkeley)

Spring 2026 — February 3, 2026

Reading: Bertsekas & Tsitsiklis, §4.2–4.3; Ch. 3.

These are my personal notes from Professor Kannan Ramchandran's EECS 126 (Probability and Random Processes) course at UC Berkeley.

We open with a famous game whose expected payout is infinite. We then finish the story of conditional expectation from Lecture 4: the law of iterated expectation (the tower rule), which we use to analyze a random walk, and its companion for variances, the law of total variance. The second half of the lecture moves from discrete to *continuous* random variables, introducing probability density functions (PDFs) and cumulative distribution functions (CDFs).

5.1 The St. Petersburg Paradox

Example 5.1 (St. Petersburg game)

I toss a fair coin until I get a heads. If I toss n times, I give you $\$2^n$. For example,

$$H \mapsto \$2, \quad TH \mapsto \$4, \quad TTH \mapsto \$8, \quad \dots \quad (5.1)$$

Question: how much would you pay to play this game?

Let N be the number of tosses in the game and $W = 2^N$ the amount of money you win. The game ends on the first heads, so $N \sim \text{Geometric}(\frac{1}{2})$:

$$\mathbb{P}(N = n) = \underbrace{\frac{1}{2} \cdot \frac{1}{2} \cdots \frac{1}{2}}_{n \text{ flips}} = 2^{-n}, \quad n = 1, 2, 3, \dots \quad (5.2)$$

(The first $n - 1$ flips are tails and the n -th is heads; each has probability $\frac{1}{2}$.) By LOTUS,

$$\mathbb{E}[W] = \sum_{n=1}^{\infty} 2^n \cdot \mathbb{P}(N = n) = \sum_{n=1}^{\infty} 2^n \cdot \frac{1}{2^n} = \sum_{n=1}^{\infty} 1 = \infty. \quad (5.3)$$

The expected payout is infinite. This is the **St. Petersburg paradox**.

Why is it a paradox? If expected value were the whole story, you should be willing to pay *any* finite price to play. Yet almost nobody would pay more than a few dollars: half the time the game pays \$2, and $\mathbb{P}(W \leq \$8) = \frac{7}{8}$. The infinite mean is driven entirely by astronomically rare, astronomically large payouts (each outcome n contributes exactly \$1 to the mean). Two standard resolutions:

1. **Diminishing utility.** Daniel Bernoulli argued that the *value* of money grows more slowly than the amount, e.g. logarithmically. The expected log-payout is finite: $\mathbb{E}[\log_2 W] = \mathbb{E}[N] = 2$ (the geometric mean from Lecture 4), which corresponds to a “fair price” of about $2^2 = \$4$.

2. **Finite bankroll.** No casino can pay more than some $\$2^M$. If payouts are capped at 2^M , then

$$\mathbb{E}[W] = \sum_{n=1}^M 2^n 2^{-n} + \sum_{n=M+1}^{\infty} 2^M 2^{-n} = M + 2^M \cdot 2^{-M} = M + 1. \quad (5.4)$$

Even a billion-dollar bankroll ($2^{30} \approx 1.07 \times 10^9$) gives an expected payout of only \$31.

5.2 Conditional Expectation as a Random Variable

Recall from Lecture 4 that for each value y with $\mathbb{P}(Y = y) > 0$,

$$\mathbb{E}[X \mid Y = y] = \sum_x x \cdot \mathbb{P}(X = x \mid Y = y) \quad (5.5)$$

is a *number*. Letting $g(y) = \mathbb{E}[X \mid Y = y]$, the conditional expectation $\mathbb{E}[X \mid Y] \triangleq g(Y)$ is a function of Y , and therefore $\mathbb{E}[X \mid Y]$ is a *random variable*.

Example 5.2 (Tossing a fair coin a random number of times)

We toss a fair coin N times, where N is itself a random variable taking values in $\{1, 2, 3, \dots\}$. Let X be the total number of heads. Given $N = n$, the n tosses are independent fair Bernoulli trials, so $X \mid N = n \sim \text{Binomial}(n, \frac{1}{2})$.

$N = 1$: one toss gives 0 or 1 heads, each with probability $\frac{1}{2}$:

$$\mathbb{E}[X \mid N = 1] = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2} \quad (\text{a number}). \quad (5.6)$$

$N = 2$: two tosses give 0, 1, or 2 heads, with probabilities $\frac{1}{4}$ (TT), $\frac{1}{2}$ (HT or TH), and $\frac{1}{4}$ (HH):

$$\mathbb{E}[X \mid N = 2] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1 \quad (\text{a number}). \quad (5.7)$$

General $N = n$: by the binomial mean from Lecture 4,

$$\mathbb{E}[X \mid N = n] = n \cdot \frac{1}{2} = \frac{n}{2} \quad (\text{a number}). \quad (5.8)$$

Sweeping over all values n and replacing n by the random variable N ,

$$\underbrace{\mathbb{E}[X \mid N]}_{\text{r.v.}} = \underbrace{\frac{N}{2}}_{\text{r.v.}}. \quad (5.9)$$

This is an **equality of random variables**: for every outcome of the experiment, the two sides take the same value.

The outcomes are not equally likely. For $N = 2$, averaging the possible head counts as $\frac{0+1+2}{3} = 1$ happens to give the right answer, but only by coincidence: the three outcomes have probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$, not $\frac{1}{3}$ each. For $N = 3$ the same shortcut would still give $\frac{3}{2}$ by symmetry, but for a *biased* coin it fails immediately. Always weight by the conditional PMF.

5.3 The Law of Iterated Expectation (Tower Property)

Since $\mathbb{E}[X | Y]$ is a random variable, we can take *its* expectation. The answer is remarkably simple.

Theorem 5.1 (Law of iterated expectation)

For random variables X, Y with $\mathbb{E}[|X|] < \infty$,

$$\mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[X]. \quad (5.10)$$

The inner expectation is over X (with Y held fixed); the outer expectation is over Y .

Proof. Write $\mathbb{E}[X | Y] = g(Y)$ with $g(y) = \mathbb{E}[X | Y = y]$. The outer expectation is the expectation of a function of Y , so by LOTUS (summing over the values y with $\mathbb{P}(Y = y) > 0$):

$$\begin{aligned} \mathbb{E}[\mathbb{E}[X | Y]] &= \sum_y \mathbb{E}[X | Y = y] \cdot \mathbb{P}(Y = y) && \text{(outer expectation, LOTUS)} \\ &= \sum_y \sum_x x \cdot \mathbb{P}(X = x | Y = y) \cdot \mathbb{P}(Y = y) && \text{(definition of } \mathbb{E}[X | Y = y]) \\ &= \sum_y \sum_x x \cdot \mathbb{P}(X = x, Y = y) && \text{(definition of conditional probability)} \\ &= \sum_x x \sum_y \mathbb{P}(X = x, Y = y) && \text{(swap the order of summation)} \\ &= \sum_x x \cdot \mathbb{P}(X = x) && \text{(marginalize out } Y) \\ &= \mathbb{E}[X]. && (5.11) \end{aligned}$$

The third line uses $\mathbb{P}(X = x | Y = y) \mathbb{P}(Y = y) = \mathbb{P}(X = x, Y = y)$, i.e. $p_{X|Y}(x | y) p_Y(y) = p_{X,Y}(x, y)$. Swapping the sums is justified by $\mathbb{E}[|X|] < \infty$ (absolute convergence). $\square$

Intuition: an average of averages. $\mathbb{E}[X | Y = y]$ is the average of X within the “group” where $Y = y$. The tower rule says: average these group averages, weighting each group by how likely it is, and you recover the overall average. The total expectation theorem from Lecture 4, $\mathbb{E}[X] = \sum_i \mathbb{P}(A_i) \mathbb{E}[X | A_i]$, is the special case where Y records which event A_i of a partition occurred.

A useful companion result lets us pull functions of Y through the inner expectation: once we condition on Y , any function of Y is a known constant.

Theorem 5.2 (Taking out what is known)

For any function f with $\mathbb{E}[|f(Y) X|] < \infty$,

$$\mathbb{E}[f(Y) X | Y] = f(Y) \mathbb{E}[X | Y], \quad (5.12)$$

and consequently

$$\mathbb{E}[f(Y) X] = \mathbb{E}[f(Y) \mathbb{E}[X | Y]]. \quad (5.13)$$

Proof. Fix y with $\mathbb{P}(Y = y) > 0$. Given $Y = y$, the random variable $f(Y)$ equals the constant $f(y)$, so it factors out of the conditional expectation:

$$\begin{aligned}\mathbb{E}[f(Y)X \mid Y = y] &= \sum_x f(y)x \mathbb{P}(X = x \mid Y = y) \\ &= f(y) \sum_x x \mathbb{P}(X = x \mid Y = y) = f(y) \mathbb{E}[X \mid Y = y].\end{aligned}\quad (5.14)$$

Replacing y by Y gives $\mathbb{E}[f(Y)X \mid Y] = f(Y) \mathbb{E}[X \mid Y]$. Now apply the tower rule to the random variable $f(Y)X$:

$$\mathbb{E}[f(Y)X] = \mathbb{E}[\mathbb{E}[f(Y)X \mid Y]] = \mathbb{E}[f(Y) \mathbb{E}[X \mid Y]].\quad (5.15)$$

□

Where the brackets go. In (5.13), $f(Y)$ stays *inside* the outer expectation, multiplied by the random variable $\mathbb{E}[X \mid Y]$. It is *not* true in general that $\mathbb{E}[f(Y)X] = \mathbb{E}[f(Y)] \cdot \mathbb{E}[X \mid Y]$; the right-hand side of that would not even be a number. For instance, with $f(Y) = Y^2$: $\mathbb{E}[Y^2X] = \mathbb{E}[Y^2 \mathbb{E}[X \mid Y]]$.

Example 5.3 (Random number of tosses, with $N \sim \text{Geometric}(p)$)

Suppose in Example 5.2 that N is generated by a *first coin* with heads probability p , tossed until its first heads, so that $N \sim \text{Geometric}(p)$. We then toss a *second*, fair coin N times and let X be the number of heads. From (5.9), $\mathbb{E}[X \mid N] = \frac{N}{2} = \frac{1}{2} \cdot N$.

Mean. By the tower rule, linearity, and $\mathbb{E}[N] = \frac{1}{p}$ for $N \sim \text{Geometric}(p)$ (Lecture 4),

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid N]] = \mathbb{E}\left[\frac{N}{2}\right] = \frac{1}{2} \mathbb{E}[N] = \frac{1}{2} \cdot \frac{1}{p} = \frac{1}{2p}.\quad (5.16)$$

Sanity check: if $p = 1$, then $N = 1$ always, and one fair toss gives $\mathbb{E}[X] = \frac{1}{2}$. ✓

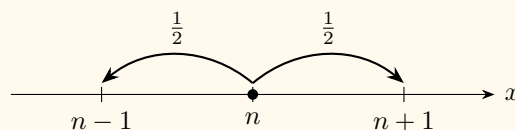
A mixed moment. Taking out what is known with $f(N) = N$, and using $\mathbb{E}[N^2] = \frac{2-p}{p^2}$ from Lecture 4,

$$\mathbb{E}[NX] = \mathbb{E}[N \mathbb{E}[X \mid N]] = \mathbb{E}\left[N \cdot \frac{N}{2}\right] = \frac{1}{2} \mathbb{E}[N^2] = \frac{2-p}{2p^2}.\quad (5.17)$$

5.4 The Symmetric Random Walk

Example 5.4 (Random walk on the integers)

A drunk man is walking along the integer line. He starts at the origin, $X_0 = 0$. At each step, independently of everything before, he moves one unit right with probability $\frac{1}{2}$ or one unit left with probability $\frac{1}{2}$. Let X_n be his location after n steps.



Questions: What is $\mathbb{E}[X_n]$? What is $\mathbb{E}[X_n^2]$, and how far from the origin is he typically?

Write the $(n+1)$ -th step using two indicator random variables,

$$\mathbf{1}_+^{(n+1)} = \begin{cases} 1 & \text{if step } n+1 \text{ is } +1, \\ 0 & \text{otherwise,} \end{cases} \quad \mathbf{1}_-^{(n+1)} = \begin{cases} 1 & \text{if step } n+1 \text{ is } -1, \\ 0 & \text{otherwise.} \end{cases} \quad (5.18)$$

Exactly one of them equals 1, so the step is $\mathbf{1}_+^{(n+1)} - \mathbf{1}_-^{(n+1)} \in \{+1, -1\}$, and

$$X_{n+1} = X_n + \mathbf{1}_+^{(n+1)} - \mathbf{1}_-^{(n+1)}. \quad (5.19)$$

Notation. An indicator only takes the values 0 and 1, so a step of -1 is written by *subtracting* the indicator $\mathbf{1}_-^{(n+1)}$. The superscript $(n+1)$ records that each step is a fresh, independent coin flip. Similarly, the location X_n and the step taken at time n are different random variables, so we keep separate names for them.

The mean: $\mathbb{E}[X_n] = 0$ for all n

Take expectations in (5.19). By linearity, and since $\mathbb{E}[\mathbf{1}_\pm^{(n+1)}] = \mathbb{P}(\text{step is } \pm 1) = \frac{1}{2}$,

$$\mathbb{E}[X_{n+1}] = \mathbb{E}[X_n] + \mathbb{E}[\mathbf{1}_+^{(n+1)}] - \mathbb{E}[\mathbf{1}_-^{(n+1)}] = \mathbb{E}[X_n] + \frac{1}{2} - \frac{1}{2} = \mathbb{E}[X_n]. \quad (5.20)$$

So one step never changes the expected location.

Theorem 5.3 (Mean of the symmetric random walk)

For every $n \geq 0$, $\mathbb{E}[X_n] = 0$.

Proof. By induction on n . **Base case:** $X_0 = 0$ is a constant, so $\mathbb{E}[X_0] = 0$. **Inductive step:** suppose $\mathbb{E}[X_n] = 0$. By (5.20), $\mathbb{E}[X_{n+1}] = \mathbb{E}[X_n] = 0$. Unrolled, this is just a chain of equalities:

$$\mathbb{E}[X_n] = \mathbb{E}[X_{n-1}] = \mathbb{E}[X_{n-2}] = \cdots = \mathbb{E}[X_1] = \mathbb{E}[X_0] = 0. \quad (5.21)$$

□

Two more ways to see it.

- Sum of steps.** Let $\xi_i \in \{+1, -1\}$ be the i -th step, so $X_n = \xi_1 + \xi_2 + \cdots + \xi_n$. Each step has mean $\mathbb{E}[\xi_i] = (+1) \cdot \frac{1}{2} + (-1) \cdot \frac{1}{2} = 0$, and by linearity $\mathbb{E}[X_n] = \sum_{i=1}^n \mathbb{E}[\xi_i] = 0$.
- Symmetry.** Flipping the direction of every step turns the walk X_n into $-X_n$, and since left and right are equally likely, the flipped walk has exactly the same distribution. So $\mathbb{E}[X_n] = \mathbb{E}[-X_n] = -\mathbb{E}[X_n]$, which forces $\mathbb{E}[X_n] = 0$.

A zero mean does *not* mean the walker stays near the origin. He drifts away, but is equally likely to drift left or right, so the drifts cancel in the average. To measure how far he wanders, we need the second moment.

The second moment: $\mathbb{E}[X_n^2] = n$

Suppose $X_n = k$. The next step is ± 1 with probability $\frac{1}{2}$ each, independent of the past, so

$$\mathbb{P}(X_{n+1}^2 = (k+1)^2 \mid X_n = k) = \frac{1}{2}, \quad \mathbb{P}(X_{n+1}^2 = (k-1)^2 \mid X_n = k) = \frac{1}{2}. \quad (5.22)$$

These are the same two equally likely moves ($k \rightarrow k+1$ and $k \rightarrow k-1$), just with the resulting location squared. Therefore

$$\begin{aligned} \mathbb{E}[X_{n+1}^2 \mid X_n = k] &= \frac{1}{2}(k+1)^2 + \frac{1}{2}(k-1)^2 \\ &= \frac{1}{2}(k^2 + 2k + 1) + \frac{1}{2}(k^2 - 2k + 1) \\ &= k^2 + 1. \end{aligned} \quad (5.23)$$

The edge case $k = 0$. When $k = 0$, $(k+1)^2 = (k-1)^2 = 1$, so the two events above are the *same* event, which has probability 1 (not $\frac{1}{2}$). The computation of $\mathbb{E}[X_{n+1}^2 \mid X_n = 0]$ is still correct, because we are really averaging over the two *moves*: $\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1 = 0^2 + 1$.

We found $g(k) = \mathbb{E}[X_{n+1}^2 \mid X_n = k] = k^2 + 1$. Replacing the value k by the random variable X_n (the “pattern matching” step, exactly as in $\mathbb{E}[X \mid N] = N/2$):

$$\mathbb{E}[X_{n+1}^2 \mid X_n] = X_n^2 + 1. \quad (5.24)$$

By the law of iterated expectation,

$$\begin{aligned} \mathbb{E}[X_{n+1}^2] &= \mathbb{E}[\mathbb{E}[X_{n+1}^2 \mid X_n]] && \text{(law of iterated expectation)} \\ &= \mathbb{E}[X_n^2 + 1] \\ &= 1 + \mathbb{E}[X_n^2]. && \text{(linearity)} \end{aligned} \quad (5.25)$$

Each step adds exactly 1 to the second moment, starting from $\mathbb{E}[X_0^2] = 0$:

$$\mathbb{E}[X_0^2] = 0, \quad \mathbb{E}[X_1^2] = 1, \quad \mathbb{E}[X_2^2] = 2, \quad \dots \implies \mathbb{E}[X_n^2] = n, \quad (5.26)$$

(formally, by induction exactly as for the mean). In particular $\mathbb{E}[X_{n+1}^2] = 1 + n$.

Theorem 5.4 (Spread of the symmetric random walk)

For every $n \geq 0$,

$$\text{Var}(X_n) = \mathbb{E}[X_n^2] - (\mathbb{E}[X_n])^2 = n - 0^2 = n, \quad \text{so} \quad \sigma_{X_n} = \sqrt{n}. \quad (5.27)$$

Cross-check with Lecture 4. With $X_n = \xi_1 + \dots + \xi_n$ and independent steps, each step has $\text{Var}(\xi_i) = \mathbb{E}[\xi_i^2] - 0^2 = 1$ (since $\xi_i^2 = 1$ always). The variance of a sum of independent random variables is the sum of the variances, so $\text{Var}(X_n) = n$. Same answer, no conditioning needed.

Key takeaway. The walker’s expected position is 0 forever, but his typical distance from the origin is $\sqrt{n}$: after 10,000 steps he is typically about 100 units away. This $\sqrt{n}$ growth is the signature of *diffusion*, and it reappears throughout the course.

5.5 Conditional Variance and the Law of Total Variance

Definition 5.1 (Conditional variance)

For y with $\mathbb{P}(Y = y) > 0$, the **conditional variance of X given $Y = y$** is the variance of the conditional distribution of X :

$$\text{Var}(X | Y = y) \triangleq \mathbb{E}[(X - \mathbb{E}[X | Y = y])^2 | Y = y] = \mathbb{E}[X^2 | Y = y] - (\mathbb{E}[X | Y = y])^2. \quad (5.28)$$

This is a number for each y , so $\text{Var}(X | Y)$, obtained by replacing y with Y , is a function of Y , and hence a **random variable**.

(The second form follows exactly as in Lecture 3, applied to the conditional distribution of X given $Y = y$.)

Theorem 5.5 (Law of total variance)

For random variables X, Y with $\mathbb{E}[X^2] < \infty$,

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X | Y)] + \text{Var}(\mathbb{E}[X | Y]). \quad (5.29)$$

Proof. Abbreviate $\mu(Y) = \mathbb{E}[X | Y]$, so that $\text{Var}(X | Y) = \mathbb{E}[X^2 | Y] - \mu(Y)^2$.

Take expectations and use the tower rule on $\mathbb{E}[X^2 | Y]$:

$$\mathbb{E}[\text{Var}(X | Y)] = \mathbb{E}[\mathbb{E}[X^2 | Y]] - \mathbb{E}[\mu(Y)^2] = \mathbb{E}[X^2] - \mathbb{E}[\mu(Y)^2]. \quad (5.30)$$

Next, $\mu(Y)$ is a random variable; apply the variance formula to it, and use the tower rule on its mean, $\mathbb{E}[\mu(Y)] = \mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[X]$:

$$\text{Var}(\mu(Y)) = \mathbb{E}[\mu(Y)^2] - (\mathbb{E}[\mu(Y)])^2 = \mathbb{E}[\mu(Y)^2] - (\mathbb{E}[X])^2. \quad (5.31)$$

Adding the two, the $\mathbb{E}[\mu(Y)^2]$ terms cancel:

$$\mathbb{E}[\text{Var}(X | Y)] + \text{Var}(\mathbb{E}[X | Y]) = \mathbb{E}[X^2] - \mathbb{E}[\mu(Y)^2] + \mathbb{E}[\mu(Y)^2] - (\mathbb{E}[X])^2 = \text{Var}(X). \quad (5.32)$$

□

Interpretation: within-group plus between-group. Group the outcomes by the value of Y . Then $\mathbb{E}[\text{Var}(X | Y)]$ is the average spread *within* each group, and $\text{Var}(\mathbb{E}[X | Y])$ is the spread *between* the group means. Total spread is the sum of the two. Since both terms are nonnegative, $\text{Var}(X) \geq \mathbb{E}[\text{Var}(X | Y)]$: on average, learning Y can only reduce our uncertainty about X .

Example 5.5 (Variance of heads in a random number of tosses)

Continue the example with $N \sim \text{Geometric}(p)$ and $X | N \sim \text{Binomial}(N, \frac{1}{2})$. From the binomial formulas of Lecture 4,

$$\mathbb{E}[X | N] = \frac{N}{2}, \quad \text{Var}(X | N) = N \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{N}{4}. \quad (5.33)$$

Using $\mathbb{E}[N] = \frac{1}{p}$ and $\text{Var}(N) = \frac{1-p}{p^2}$:

$$\mathbb{E}[\text{Var}(X | N)] = \mathbb{E}\left[\frac{N}{4}\right] = \frac{1}{4} \cdot \frac{1}{p} = \frac{1}{4p}, \quad (5.34)$$

$$\text{Var}(\mathbb{E}[X | N]) = \text{Var}\left(\frac{N}{2}\right) = \frac{1}{4} \text{Var}(N) = \frac{1-p}{4p^2}. \quad (5.35)$$

By the law of total variance,

$$\text{Var}(X) = \frac{1}{4p} + \frac{1-p}{4p^2} = \frac{p}{4p^2} + \frac{1-p}{4p^2} = \frac{1}{4p^2}. \quad (5.36)$$

Sanity check: if $p = 1$, then $N = 1$ and $X \sim \text{Bernoulli}(\frac{1}{2})$, whose variance is $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. ✓

5.6 Continuous Random Variables

Continuous random variables (CRVs) let us measure uncertainty in a principled way for quantities that vary continuously, such as velocity, time, and distance.

Coming up. Important continuous distributions include the *uniform*, *exponential*, and *gamma* distributions, which we develop in upcoming lectures. The exponential is the continuous analog of the geometric (Lecture 4).

Why discrete tools break. Let $X \sim \text{Uniform}[0, 1]$, a number chosen “uniformly at random” from $[0, 1]$. Then:

- $\mathbb{P}(X = x) = 0$ for any specific number x (there are uncountably many equally likely candidates);
- yet $\mathbb{P}(X < \frac{1}{2}) = \frac{1}{2}$.

So we cannot build probabilities by adding up point masses as we did with PMFs. Instead, we assign probabilities directly to *sets* that are “allowable events.”

Definition 5.2 (Allowable events (Borel sets))

The allowable events for a real-valued random variable include all intervals,

$$(-\infty, a), \quad [a, b], \quad (a, b], \quad \dots \quad (5.37)$$

together with everything obtainable from intervals by **countable unions**, **countable intersections**, and **complements**, e.g. $(a, b) \cup (c, d)$. These are the **Borel sets**; they form the Borel σ -algebra from Lecture 1. In practice, any “reasonable” set is allowable.

Recall that for a discrete random variable (DRV) and a set of values B ,

$$\mathbb{P}(X \in B) = \sum_{x \in B} p_X(x). \quad (5.38)$$

For a CRV, the sum over points becomes an integral of a density.

Definition 5.3 (Continuous random variable and PDF)

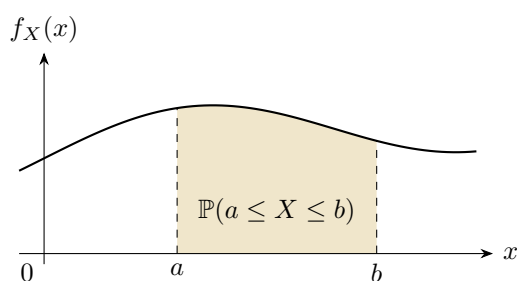
A random variable X is **continuous** if there is a function $f_X : \mathbb{R} \rightarrow [0, \infty)$, called its **probability density function (PDF)**, such that for every allowable set B ,

$$\mathbb{P}(X \in B) = \int_B f_X(x) dx. \quad (5.39)$$

A density must satisfy

$$f_X(x) \geq 0 \text{ for all } x, \quad \text{and} \quad \int_{-\infty}^{\infty} f_X(x) dx = 1. \quad (5.40)$$

In particular, $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx$ is the area under the density between a and b .

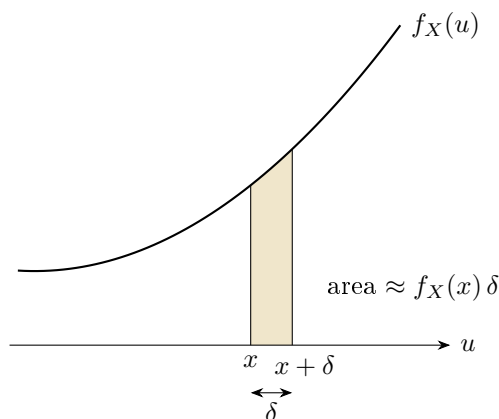
**Theorem 5.6 (Single points have probability zero)**

If X is continuous, then $\mathbb{P}(X = x) = 0$ for every $x \in \mathbb{R}$.

Proof. Fix x and any $\delta > 0$. The event $\{X = x\}$ is contained in $\{x \leq X < x + \delta\}$, so

$$\mathbb{P}(X = x) \leq \mathbb{P}(x \leq X < x + \delta) = \int_x^{x+\delta} f_X(u) du. \quad (5.41)$$

(We use the dummy variable u inside the integral so as not to confuse it with x , which now appears in the limits.) As $\delta \rightarrow 0$, this is the integral of f_X over a shrinking interval, which tends to 0. If f_X is continuous at x , we can see this explicitly: the sliver is approximately a rectangle of height $f_X(x)$ and width δ , so the integral is $\approx f_X(x) \delta \rightarrow 0$. Since $\mathbb{P}(X = x)$ is smaller than a quantity that can be made arbitrarily small, $\mathbb{P}(X = x) = 0$. $\square$



A density is not a probability. $f_X(x)$ has units of *probability per unit length*. Wherever f_X is continuous,

$$f_X(x) = \lim_{\delta \rightarrow 0} \frac{\mathbb{P}(x \leq X < x + \delta)}{\delta}, \quad \text{i.e.} \quad \mathbb{P}(x \leq X < x + \delta) \approx f_X(x) \delta \text{ for small } \delta. \quad (5.42)$$

Don't confuse $f_X(x)$ with a probability: it is a *density*. Only areas under f_X are probabilities, so $f_X(x)$ itself may exceed 1, as the next example shows.

Example 5.6 (An unbounded density)

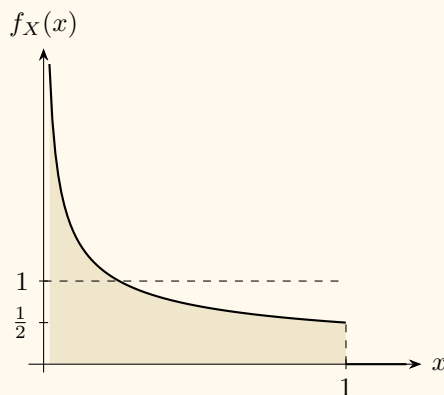
Let

$$f_X(x) = \begin{cases} \frac{1}{2\sqrt{x}} & 0 < x \leq 1, \\ 0 & \text{otherwise.} \end{cases} \quad (5.43)$$

It is a valid PDF. Clearly $f_X \geq 0$. For the total area, write $\frac{1}{2\sqrt{x}} = \frac{1}{2}x^{-1/2}$ and integrate:

$$\int_{-\infty}^{\infty} f_X(x) dx = \int_0^1 \frac{1}{2} x^{-1/2} dx = \frac{1}{2} \left[2x^{1/2} \right]_0^1 = \left[\sqrt{x} \right]_0^1 = 1 - 0 = 1. \quad (5.44)$$

It exceeds 1, and is even unbounded. $f_X(1) = \frac{1}{2}$, but $f_X(0.01) = \frac{1}{2(0.1)} = 5$, and $f_X(x) \rightarrow \infty$ as $x \rightarrow 0^+$. Nevertheless every probability is finite, e.g. $\mathbb{P}(X \leq 0.01) = \int_0^{0.01} \frac{1}{2} x^{-1/2} dx = \sqrt{0.01} = 0.1$.



5.7 The Cumulative Distribution Function

Definition 5.4 (Cumulative distribution function (CDF))

The **cumulative distribution function** of any random variable X (discrete or continuous) is

$$F_X(x) \triangleq \mathbb{P}(X \leq x), \quad x \in \mathbb{R}. \quad (5.45)$$

For a continuous random variable,

$$F_X(x) = \int_{-\infty}^x f_X(u) du. \quad (5.46)$$

Theorem 5.7 (Properties of the CDF)

Every CDF satisfies:

1. **Nondecreasing:** if $x \leq y$ then $F_X(x) \leq F_X(y)$, since $\{X \leq x\} \subseteq \{X \leq y\}$.
2. **Limits:** $\lim_{x \rightarrow -\infty} F_X(x) = 0$ and $\lim_{x \rightarrow +\infty} F_X(x) = 1$.
3. **Intervals:** $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$, because $\{X \leq b\}$ is the disjoint union of $\{X \leq a\}$ and $\{a < X \leq b\}$.

Theorem 5.8 (Recovering the PMF or PDF from the CDF)

- **CRV:** wherever f_X is continuous,

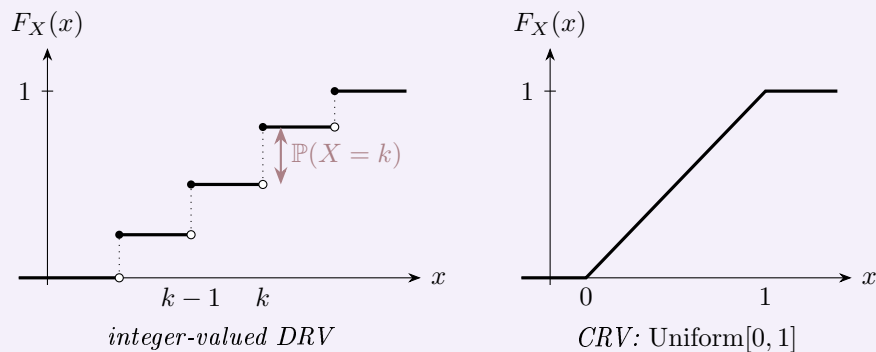
$$f_X(x) = \frac{dF_X(x)}{dx}. \quad (5.47)$$

- **Integer-valued DRV:** for every integer k ,

$$\mathbb{P}(X = k) = F_X(k) - F_X(k - 1). \quad (5.48)$$

Proof. **CRV:** $F_X(x) = \int_{-\infty}^x f_X(u) du$, so this is the fundamental theorem of calculus. **DRV:** because X only takes integer values, there are no possible values strictly between $k - 1$ and k . So $\{X \leq k\}$ is the disjoint union of $\{X \leq k - 1\}$ and $\{X = k\}$, giving $F_X(k) = F_X(k - 1) + \mathbb{P}(X = k)$. (For a discrete random variable taking non-integer values, use the jump of F_X at the value instead.) $\square$

Picturing the two kinds of CDF. For an integer-valued DRV, F_X is a *staircase*. On each open interval $(k - 1, k)$ there are no possible values, so no new probability accumulates and F_X stays flat, holding the value $F_X(k - 1)$. It does not drop back to zero: all the probability collected up through $k - 1$ is still counted. At k itself, F_X jumps up by exactly $\mathbb{P}(X = k)$. For a CRV, single points carry no probability, so F_X rises continuously with no jumps.



Corollary 5.1 (Consequences for continuous random variables)

If X is continuous, then for all $a \leq b$:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx = F_X(b) - F_X(a), \quad (5.49)$$

$$\mathbb{P}(X = a) = 0, \quad \mathbb{P}(X < a) = \mathbb{P}(X \leq a). \quad (5.50)$$

Because points do not contribute, it does not matter whether the endpoints of an interval are included: $\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b)$. This is *false* for discrete random variables, where $\mathbb{P}(X < k)$ and $\mathbb{P}(X \leq k)$ differ by $\mathbb{P}(X = k)$.

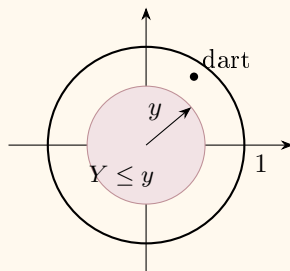
5.8 Example: A Random Dart on a Disk

Example 5.7 (Random dart on the unit disk)

A dart lands uniformly at random on a disk of radius 1 (the unit circle and its interior). “Uniformly” means the probability of landing in any region is proportional to its area:

$$\mathbb{P}(\text{dart lands in region } R) = \frac{\text{Area}(R)}{\text{Area of full disk}} = \frac{\text{Area}(R)}{\pi(1)^2}. \quad (5.51)$$

Let Y be the dart’s distance from the center. Find the PDF $f_Y(y)$.



It is easiest to compute $F_Y(y) = \mathbb{P}(Y \leq y)$ first. For $0 \leq y \leq 1$, the event $\{Y \leq y\}$ means the dart lands in the shaded disk of radius y :

$$F_Y(y) = \mathbb{P}(Y \leq y) = \frac{\text{Area shaded}}{\text{Area full disk}} = \frac{\pi y^2}{\pi(1)^2} = y^2. \quad (5.52)$$

Outside $[0, 1]$: $F_Y(y) = 0$ for $y < 0$ (distance is never negative) and $F_Y(y) = 1$ for $y > 1$ (the dart is always on the disk).

Differentiating,

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{d}{dy}(y^2) = 2y, \quad 0 \leq y \leq 1, \quad (5.53)$$

and $f_Y(y) = 0$ otherwise. Check: $\int_0^1 2y \, dy = [y^2]_0^1 = 1$. ✓

By the CDF,

$$\mathbb{P}(Y \in (0.5, 0.6)) = F_Y(0.6) - F_Y(0.5) = 0.6^2 - 0.5^2 = 0.36 - 0.25 = 0.11. \quad (5.54)$$

Or, equivalently, by the PDF: $\int_{0.5}^{0.6} 2y \, dy = [y^2]_{0.5}^{0.6} = 0.11$.

Three traps in this example.

1. **Y is not uniform on $[0, 1]$.** The dart’s *position* is uniform on the disk, but its *distance* from the center is not: the density $2y$ grows with y , because there is more area at larger radii. The ring $0.5 < Y < 0.6$ has probability 0.11, while the same-width band $0 < Y < 0.1$ near the center has only $0.1^2 = 0.01$.
2. **Don’t stop at the CDF.** y^2 is F_Y , not f_Y ; the PDF requires the extra differentiation step.
3. **$\frac{1}{\pi r^2}$ is a 2-D density.** The constant $\frac{1}{\pi r^2}$ (here $\frac{1}{\pi}$) is the density of the dart’s *position* per unit *area*. It is not f_Y , which is a density per unit *length* along the radius.

The CDF method. This “CDF first, then differentiate” strategy is the standard way to find the density of a function of a random variable.

5.9 Example: Meeting at Random Times

Not finished in lecture. We ran out of time after posing this problem. The full solution is given below.

Example 5.8 (Will A and B meet?)

A and B each arrive at a meeting spot at a uniformly random time between 4:00 pm and 5:00 pm, independently of each other. Whoever arrives first waits 10 minutes and then leaves. What is $\mathbb{P}(A \text{ and } B \text{ meet})$?

Setup. Measure time in minutes after 4:00 pm, so A and B are independent $\text{Uniform}[0, 60]$ random variables. The pair (A, B) is a point in the 60×60 square $[0, 60]^2$. Because A and B are independent and each uniform, the point (A, B) is uniform on the square:

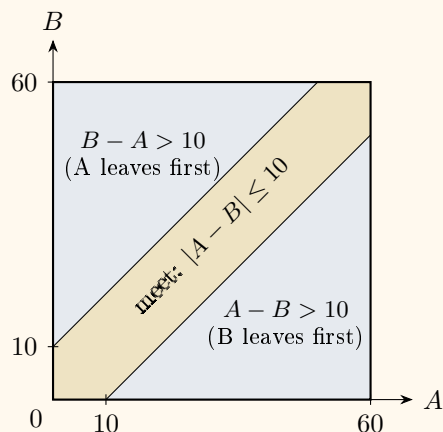
$$\mathbb{P}((A, B) \in R) = \frac{\text{Area}(R)}{60 \times 60} = \frac{\text{Area}(R)}{3600}. \quad (5.55)$$

(Joint distributions of several continuous random variables are developed formally in upcoming lectures; this “probability = area fraction” rule is the two-dimensional version of the uniform distribution, exactly as in the dart example.)

When do they meet? Whoever arrives first waits 10 minutes, so they meet exactly when their arrival times differ by at most 10 minutes:

$$\{A \text{ and } B \text{ meet}\} = \{|A - B| \leq 10\}. \quad (5.56)$$

This is the shaded diagonal band in the figure below.



Compute the complement. The two blue corner regions, where they miss each other, are right triangles:

- $\{A - B > 10\}$: A arrives more than 10 minutes after B , so B has already left. The triangle has vertices $(10, 0)$, $(60, 0)$, $(60, 50)$ and legs of length 50.
- $\{B - A > 10\}$: the mirror image, with vertices $(0, 10)$, $(0, 60)$, $(50, 60)$, also with legs of length 50.

Each triangle has area $\frac{1}{2} \cdot 50 \cdot 50 = 1250$, so together they cover 2500 out of 3600:

$$\mathbb{P}(\text{miss}) = \frac{2 \cdot 1250}{3600} = \frac{2500}{3600} = \frac{25}{36}. \quad (5.57)$$

Therefore

$$\mathbb{P}(A \text{ and } B \text{ meet}) = 1 - \frac{25}{36} = \frac{11}{36} \approx 0.306. \quad (5.58)$$

Equivalently, the band has area $3600 - 2500 = 1100$, and $\frac{1100}{3600} = \frac{11}{36}$.

Checks and generalization.

1. **Boundaries don't matter.** The lines $|A - B| = 10$ have zero area, so it makes no difference whether “waits 10 minutes” includes the endpoint. This is the two-dimensional version of $\mathbb{P}(X < a) = \mathbb{P}(X \leq a)$.
2. **General formula.** With a window of length T and waiting time $w \leq T$, the same picture gives

$$\mathbb{P}(\text{meet}) = 1 - \left(1 - \frac{w}{T}\right)^2. \quad (5.59)$$

With $w = 10$, $T = 60$: $1 - (5/6)^2 = 11/36$. Sanity checks: $w = 0$ gives 0 (they must arrive at the exact same instant), and $w = T$ gives 1.

End of Lecture 5.

Lecture 7: The Gaussian Distribution, Derived Distributions, and Order Statistics

EECS 126 (UC Berkeley)

Spring 2026 — February 10, 2026

Reading: Bertsekas & Tsitsiklis, §3.3, §4.1.

These are my personal notes from Professor Kannan Ramchandran's EECS 126 (Probability and Random Processes) course at UC Berkeley.

Topics: the Gaussian distribution, change of variables (derived distributions), and order statistics.

7.1 The Gaussian Distribution

Definition 7.1 (Gaussian (normal) distribution)

$X \sim \mathcal{N}(\mu, \sigma^2)$, with $\mu \in \mathbb{R}$ and $\sigma > 0$, if

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R}. \quad (7.1)$$

The **standard normal** is $Z \sim \mathcal{N}(0, 1)$:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad z \in \mathbb{R}. \quad (7.2)$$

The constant $\frac{1}{\sqrt{2\pi}}$ is the normalization factor that makes $\int_{-\infty}^{\infty} f_Z(z) dz = 1$.

Definition 7.2 (Standard normal CDF)

$$\Phi(z) \triangleq \mathbb{P}(Z \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du. \quad (7.3)$$

u is a dummy variable of integration; z appears only in the upper limit. Φ has no closed form, so its values are read from a table.

Theorem 7.1 (Key Gaussian integral)

For every $\alpha > 0$,

$$\int_{-\infty}^{\infty} e^{-\alpha x^2/2} dx = \sqrt{\frac{2\pi}{\alpha}}, \quad (7.4)$$

or equivalently

$$I(\alpha) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\alpha x^2/2} dx = \frac{1}{\sqrt{\alpha}} = \alpha^{-1/2}. \quad (7.5)$$

Proof. Substitute $u = \sqrt{\alpha} x$, so $du = \sqrt{\alpha} dx$ and $\alpha x^2 = u^2$:

$$\int_{-\infty}^{\infty} e^{-\alpha x^2/2} dx = \int_{-\infty}^{\infty} e^{-u^2/2} \frac{du}{\sqrt{\alpha}} = \frac{1}{\sqrt{\alpha}} \int_{-\infty}^{\infty} e^{-u^2/2} du = \frac{1}{\sqrt{\alpha}} \cdot \sqrt{2\pi}, \quad (7.6)$$

where the last step uses $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = 1$ from (7.2). Dividing by $\sqrt{2\pi}$ gives (7.5). $\square$

7.2 Mean and Variance of the Standard Normal

Theorem 7.2 (Standard normal moments)

If $Z \sim \mathcal{N}(0, 1)$, then $\mathbb{E}[Z] = 0$ and $\text{Var}(Z) = 1$.

Proof of $\mathbb{E}[Z] = 0$.

$$\mathbb{E}[Z] = \int_{-\infty}^{\infty} z f_Z(z) dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} z e^{-z^2/2} dz. \quad (7.7)$$

The integrand $g(z) = z e^{-z^2/2}$ is odd, $g(-z) = -g(z)$, so the two halves cancel:

$$\begin{aligned} \int_0^{\infty} z e^{-z^2/2} dz &= \left[-e^{-z^2/2} \right]_0^{\infty} = 0 - (-1) = 1, \\ \int_{-\infty}^0 z e^{-z^2/2} dz &= \left[-e^{-z^2/2} \right]_{-\infty}^0 = -1 - 0 = -1, \end{aligned} \quad (7.8)$$

and therefore $\mathbb{E}[Z] = \frac{1}{\sqrt{2\pi}}(1 - 1) = 0$. $\square$

Proof of $\text{Var}(Z) = 1$. Differentiate $e^{-\alpha x^2/2}$ with respect to α :

$$\frac{d}{d\alpha} \left(e^{-\alpha x^2/2} \right) = -\frac{x^2}{2} e^{-\alpha x^2/2} \implies x^2 e^{-\alpha x^2/2} = -2 \frac{d}{d\alpha} \left(e^{-\alpha x^2/2} \right). \quad (7.9)$$

Then

$$\begin{aligned} \mathbb{E}[Z^2] &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-\alpha x^2/2} dx \right]_{\alpha=1} \\ &= \left[\int_{-\infty}^{\infty} \frac{-2}{\sqrt{2\pi}} \frac{d}{d\alpha} \left(e^{-\alpha x^2/2} \right) dx \right]_{\alpha=1} \\ &= \left[-2 \frac{d}{d\alpha} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\alpha x^2/2} dx \right) \right]_{\alpha=1} && \text{(Leibniz rule)} \\ &= \left[-2 \frac{d}{d\alpha} I(\alpha) \right]_{\alpha=1} = \left[-2 \frac{d}{d\alpha} \alpha^{-1/2} \right]_{\alpha=1} && \text{by (7.5)} \\ &= \left[-2 \left(-\frac{1}{2} \alpha^{-3/2} \right) \right]_{\alpha=1} = \left[\alpha^{-3/2} \right]_{\alpha=1} = \left[\frac{1}{\sqrt{\alpha^3}} \right]_{\alpha=1} = 1. \end{aligned} \quad (7.10)$$

The Leibniz rule (differentiation under the integral sign) justifies moving $\frac{d}{d\alpha}$ outside the integral. Finally,

$$\text{Var}(Z) = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 = 1 - 0^2 = 1. \quad (7.11)$$

$\square$

Why only $\alpha = 1$? I was confused why we set $\alpha = 1$, and whether it could be something else. The identity $I(\alpha) = \alpha^{-1/2}$ holds for *every* $\alpha > 0$, so I is a function of α that we can

differentiate, and differentiating in α is what produces the factor x^2 . We set $\alpha = 1$ only *after* differentiating, because the standard normal has $e^{-x^2/2}$, i.e. $\alpha = 1$. Setting $\alpha = 1$ first would leave the constant $I(1) = 1$, whose derivative is 0.

Cross-check by integration by parts. With $u = x$ and $dv = \frac{1}{\sqrt{2\pi}} x e^{-x^2/2} dx$, so $v = -\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$:

$$\mathbb{E}[Z^2] = \left[-\frac{x}{\sqrt{2\pi}} e^{-x^2/2} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0 + 1 = 1. \quad (7.12)$$

7.3 Linear Transformations of Gaussians

Theorem 7.3 (Shifting and scaling a standard normal)

If $Z \sim \mathcal{N}(0, 1)$, $\sigma > 0$, and $X = \sigma Z + \mu$, then $X \sim \mathcal{N}(\mu, \sigma^2)$.

Proof. Start with the CDF of X . Since $\sigma > 0$, dividing by σ preserves the inequality:

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\sigma Z + \mu \leq x) = \mathbb{P}\left(Z \leq \frac{x - \mu}{\sigma}\right) = F_Z(z), \quad z = \frac{x - \mu}{\sigma}. \quad (7.13)$$

Differentiate with the chain rule, using $\frac{dz}{dx} = \frac{d}{dx} \left(\frac{x - \mu}{\sigma} \right) = \frac{1}{\sigma}$:

$$\begin{aligned} f_X(x) &= \frac{d}{dx} F_X(x) = \frac{d}{dx} F_Z(z) = f_Z(z) \frac{dz}{dx} = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \cdot \frac{1}{\sigma} \\ &= \frac{1}{\sqrt{2\pi} \sigma} e^{-(x-\mu)^2/2\sigma^2}. \quad \square \end{aligned}$$

Is $f_Z(z) \cdot \frac{1}{\sigma}$ already the answer? I was unsure whether the line $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \cdot \frac{1}{\sigma}$ was correct. It is: substituting $z = \frac{x-\mu}{\sigma}$ gives exactly the $\mathcal{N}(\mu, \sigma^2)$ density. The factor $\frac{1}{\sigma}$ in front comes from $\frac{dz}{dx}$.

Should μ and σ have a subscript Z ? I was also confused about this. No: Z always has mean 0 and variance 1. Here μ and σ are the shift and scale applied to Z , and they become the mean and standard deviation of X : $\mathbb{E}[X] = \sigma \mathbb{E}[Z] + \mu = \mu$ and $\text{Var}(X) = \sigma^2 \text{Var}(Z) = \sigma^2$. If anything, they are μ_X and σ_X .

Corollary 7.1 (Standardization)

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$, and

$$F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right). \quad (7.14)$$

7.4 Symmetry of Φ and Standardization

Theorem 7.4 (Symmetry of the standard normal CDF)

For every $y \in \mathbb{R}$,

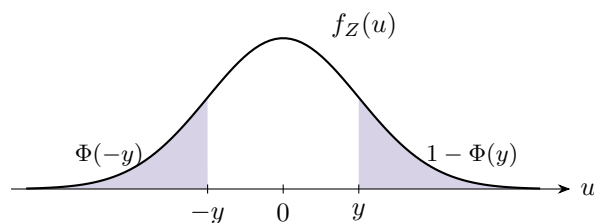
$$\Phi(-y) = 1 - \Phi(y). \quad (7.15)$$

So a table only needs $\Phi(y)$ for $y \geq 0$.

Proof. f_Z is even: $f_Z(-v) = f_Z(v)$. Substitute $u = -v$, $du = -dv$:

$$\begin{aligned} \Phi(-y) &= \int_{-\infty}^{-y} f_Z(u) du = \int_{\infty}^y f_Z(-v) (-dv) = \int_y^{\infty} f_Z(v) dv \\ &= \mathbb{P}(Z \geq y) = 1 - \mathbb{P}(Z < y) = 1 - \Phi(y), \end{aligned} \quad (7.16)$$

using $\mathbb{P}(Z < y) = \mathbb{P}(Z \leq y)$ for a continuous random variable. $\square$

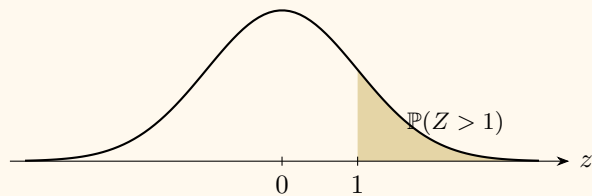


Example 7.1 (Annual rainfall)

The annual rainfall in a city is $X \sim \mathcal{N}(60, 20^2)$, in inches. Find $\mathbb{P}(X > 80)$.

Standardize with $Z = \frac{X-60}{20} \sim \mathcal{N}(0, 1)$:

$$\begin{aligned} \mathbb{P}(X > 80) &= \mathbb{P}\left(\frac{X - 60}{20} > \frac{80 - 60}{20}\right) \\ &= \mathbb{P}\left(\frac{X - 60}{20} > 1\right) = \mathbb{P}(Z > 1) \\ &= 1 - \mathbb{P}(Z \leq 1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587. \end{aligned} \quad (7.17)$$



Why $1 - \Phi(1)$? I was confused why the answer is $1 - \Phi(1)$ rather than $\Phi(1)$. $\Phi(1)$ is the area to the *left* of 1; $\mathbb{P}(Z > 1)$ is the area to the *right*, which is its complement.

7.5 Sums of Independent Gaussians

Theorem 7.5 (Sum of independent Gaussians)

If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are independent, then

$$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2). \quad (7.18)$$

The parameters follow from Lecture 4: $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] = \mu_X + \mu_Y$, and by independence $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) = \sigma_X^2 + \sigma_Y^2$. The content of the theorem is that $X + Y$ is again *Gaussian*. This can be shown with a convolution, or in one line with moment generating functions.

Proof: MGF proof TBD in a later lecture.

Theorem 7.6 (Converse)

If X and Y are independent and $X + Y$ is Gaussian, then X and Y are each Gaussian (allowing a constant, i.e. variance 0).

This direction is much harder; its proof is beyond the scope of this course.

7.6 Derived Distributions

Given X with known density f_X and $Y = g(X)$, find f_Y .

Theorem 7.7 (CDF method)

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \in \{x : g(x) \leq y\}) = \int_{\{x : g(x) \leq y\}} f_X(x) dx, \quad (7.19)$$

and then

$$f_Y(y) = \frac{dF_Y(y)}{dy}. \quad (7.20)$$

If g is strictly increasing, $\{x : g(x) \leq y\} = \{x \leq g^{-1}(y)\}$, so $F_Y(y) = F_X(g^{-1}(y))$. If g is strictly decreasing, the inequality flips: $\{x : g(x) \leq y\} = \{x \geq g^{-1}(y)\}$, so $F_Y(y) = 1 - F_X(g^{-1}(y))$.

Corollary 7.2 (Monotone change of variables)

If g is strictly monotone and differentiable, then

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|. \quad (7.21)$$

Proof. Increasing g : $\frac{d}{dy} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}$, with $\frac{dg^{-1}(y)}{dy} > 0$. Decreasing g : $\frac{d}{dy} [1 - F_X(g^{-1}(y))] = -f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}$, with $\frac{dg^{-1}(y)}{dy} < 0$. In both cases the result is (7.21). $\square$

For $X = \sigma Z + \mu$: $g^{-1}(x) = \frac{x - \mu}{\sigma}$ and $\frac{dg^{-1}(x)}{dx} = \frac{1}{\sigma}$, which is the factor $\frac{1}{\sigma}$ found in Section 7.3.

Example 7.2 (Square root of a uniform)

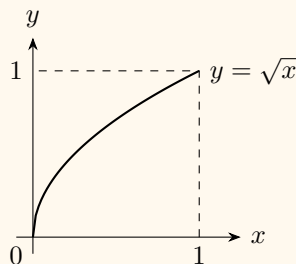
Let $X \sim \text{Uniform}[0, 1]$ and $Y = \sqrt{X}$. For $0 \leq y \leq 1$, since $\sqrt{x}$ is increasing on $[0, 1]$ and both sides are nonnegative,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(\sqrt{X} \leq y) = \mathbb{P}(X \leq y^2) = F_X(y^2) = \int_0^{y^2} 1 \, dx = \left[x \right]_0^{y^2} = y^2 - 0 = y^2, \quad (7.22)$$

with $F_Y(y) = 0$ for $y < 0$ and $F_Y(y) = 1$ for $y > 1$. Differentiating,

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} (y^2) = 2y, \quad 0 \leq y \leq 1. \quad (7.23)$$

Check with (7.21): $g^{-1}(y) = y^2$, $\frac{dg^{-1}(y)}{dy} = 2y$, $f_X(y^2) = 1$, so $f_Y(y) = 1 \cdot 2y = 2y$.



$Y = \sqrt{X}$ has the same density $2y$ as the distance of a random dart from the center of the unit disk (Lecture 5).

7.7 Order Statistics

Let $X_1, X_2, \dots, X_n$ be i.i.d. samples with common density f_X and CDF F_X .

Definition 7.3 (Order statistics)

$$X^{(k)} \triangleq \text{the } k\text{-th smallest of } X_1, X_2, \dots, X_n, \quad k = 1, \dots, n. \quad (7.24)$$

In particular $X^{(1)} = \min_i X_i$ and $X^{(n)} = \max_i X_i$.

Example 7.3 (Ordering three samples)

$n = 3$, $X_i \sim \text{Uniform}(0, 1)$, observed $(X_1, X_2, X_3) = (0.3, 0.8, 0.4)$. Then

$$X^{(1)} = 0.3, \quad X^{(2)} = 0.4, \quad X^{(3)} = 0.8. \quad (7.25)$$

Question: what is $f_{X^{(k)}}(x)$, the density of $X^{(k)}$?

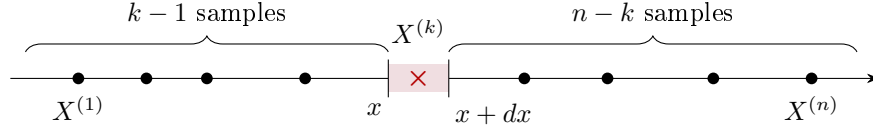
By the definition of a density,

$$\mathbb{P}(X^{(k)} \in (x, x + dx)) \approx f_{X^{(k)}}(x) \, dx. \quad (7.26)$$

For the k -th smallest sample to lie in $(x, x + dx)$, we need

- $k - 1$ samples in $(-\infty, x)$: probability $[F_X(x)]^{k-1}$;
- 1 sample in $(x, x + dx)$: probability $f_X(x) dx$;
- $n - k$ samples in $(x + dx, \infty)$: probability $[1 - F_X(x)]^{n-k}$,

where the probabilities multiply by independence.



Count the ways to assign the n samples to the three regions:

- which sample lies in $(x, x + dx)$: n ways;
- which $k - 1$ of the remaining $n - 1$ lie to the left: $\binom{n-1}{k-1}$ ways;
- the remaining $n - k$ lie to the right: 1 way.

$$n \binom{n-1}{k-1} = n \cdot \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} = \frac{n!}{(k-1)!(n-k)!}. \quad (7.27)$$

Why only 1 way for the third group? I was confused whether the right-hand group contributes $\binom{n-k}{n-k}$ more choices. It does not: once the middle sample and the $k - 1$ left samples are chosen, the other $n - k$ samples are forced to the right. (Two or more samples landing in $(x, x + dx)$ has probability of order $(dx)^2$, which is negligible.)

Theorem 7.8 (Density of the k -th order statistic)

$$f_{X^{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} [F_X(x)]^{k-1} [1 - F_X(x)]^{n-k} f_X(x). \quad (7.28)$$

Proof. Multiply the number of assignments by the probability of each:

$$f_{X^{(k)}}(x) dx = \frac{n!}{(k-1)!(n-k)!} [F_X(x)]^{k-1} [1 - F_X(x)]^{n-k} f_X(x) dx, \quad (7.29)$$

and divide both sides by dx . □

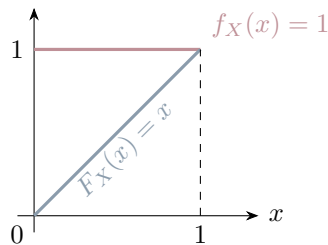
Why y ? I was confused why the formula was written in y rather than x . The argument is a dummy variable: $f_{X^{(k)}}(y)$ and $f_{X^{(k)}}(x)$ are the same function. What matters is that one variable is used throughout, and that F_X and f_X are the CDF and PDF of a *single* sample, not of $X^{(k)}$.

Check, $k = n$ (the maximum). $\mathbb{P}(X^{(n)} \leq x) = \mathbb{P}(\text{all } X_i \leq x) = [F_X(x)]^n$, and $\frac{d}{dx} [F_X(x)]^n = n [F_X(x)]^{n-1} f_X(x)$, matching (7.28) with $k = n$.

Corollary 7.3 (Uniform samples)

If $X_i \sim \text{Uniform}[0, 1]$, then $f_X(x) = 1$ and $F_X(x) = x$ on $[0, 1]$, so

$$f_{X^{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1} (1-x)^{n-k}, \quad 0 \leq x \leq 1. \quad (7.30)$$


Example 7.4 (Third smallest of four uniforms)

$X_i \sim \text{Uniform}[0, 1]$, $n = 4$. Find $\mathbb{P}(X^{(3)} < \frac{1}{2})$.

With $n = 4$, $k = 3$: $\frac{n!}{(k-1)!(n-k)!} = \frac{4!}{2!1!} = \frac{24}{2} = 12$. By (7.30),

$$\begin{aligned} \mathbb{P}(X^{(3)} < \tfrac{1}{2}) &= F_{X^{(3)}}(\tfrac{1}{2}) = \int_0^{1/2} f_{X^{(3)}}(x) dx = \int_0^{1/2} 12x^2(1-x) dx \\ &= 12 \int_0^{1/2} (x^2 - x^3) dx = 12 \left[\frac{x^3}{3} - \frac{x^4}{4} \right]_0^{1/2} \\ &= 12 \left(\frac{1}{24} - \frac{1}{64} \right) = 12 \left(\frac{8}{192} - \frac{3}{192} \right) = 12 \cdot \frac{5}{192} = \frac{60}{192} = \frac{5}{16}. \end{aligned} \quad (7.31)$$

Check: $X^{(3)} < \frac{1}{2}$ exactly when at least 3 of the 4 samples are below $\frac{1}{2}$. The number below $\frac{1}{2}$ is Binomial($4, \frac{1}{2}$), so

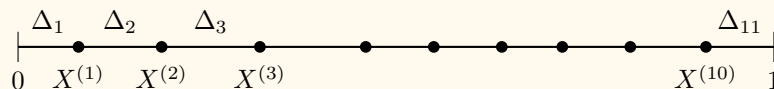
$$\mathbb{P}(X^{(3)} < \tfrac{1}{2}) = \binom{4}{3} \left(\tfrac{1}{2}\right)^4 + \binom{4}{4} \left(\tfrac{1}{2}\right)^4 = \frac{4}{16} + \frac{1}{16} = \frac{5}{16}. \quad (7.32)$$

Example 7.5 (Expected minimum by symmetry)

$X_i \sim \text{Uniform}[0, 1]$, $n = 10$. Find $\mathbb{E}[X^{(1)}]$, the expected smallest sample.

The 10 samples cut $[0, 1]$ into 11 gaps:

$$\Delta_1 = X^{(1)} - 0, \quad \Delta_i = X^{(i)} - X^{(i-1)} \quad (2 \leq i \leq 10), \quad \Delta_{11} = 1 - X^{(10)}. \quad (7.33)$$



The sum telescopes:

$$\sum_{i=1}^{11} \Delta_i = (X^{(1)} - 0) + (X^{(2)} - X^{(1)}) + \cdots + (X^{(10)} - X^{(9)}) + (1 - X^{(10)}) = 1. \quad (7.34)$$

By symmetry, all 11 gaps have the same distribution, so $\mathbb{E}[\Delta_i] = \mathbb{E}[\Delta_1]$ for every i . Taking expectations,

$$\mathbb{E}\left[\sum_{i=1}^{11} \Delta_i\right] = \sum_{i=1}^{11} \mathbb{E}[\Delta_i] = 11 \mathbb{E}[\Delta_1] = \mathbb{E}[1] = 1 \implies \mathbb{E}[X^{(1)}] = \mathbb{E}[\Delta_1] = \frac{1}{11}. \quad (7.35)$$

Do the gaps really sum to 1? I was unsure about this. They do: the gaps tile $[0, 1]$ exactly, which is the telescoping sum above.

Why are the gaps symmetric? Place 11 independent uniform points on a circle of circumference 1 and cut the circle at one of them. The other 10 points are i.i.d. uniform on the resulting interval $[0, 1]$, and the 11 arcs between consecutive points are exactly the gaps $\Delta_1, \dots, \Delta_{11}$. On the circle no arc is special, so all gaps have the same distribution.

Cross-check with (7.30). With $k = 1$: $f_{X^{(1)}}(x) = \frac{10!}{0!9!}(1-x)^9 = 10(1-x)^9$. Substituting $u = 1 - x$,

$$\mathbb{E}[X^{(1)}] = \int_0^1 x \cdot 10(1-x)^9 dx = \int_0^1 (1-u) 10 u^9 du = 10 \left(\frac{1}{10} - \frac{1}{11} \right) = 10 \cdot \frac{1}{110} = \frac{1}{11}. \quad (7.36)$$

General case. $X^{(k)} = \Delta_1 + \dots + \Delta_k$, so for n uniform samples,

$$\mathbb{E}[X^{(k)}] = k \mathbb{E}[\Delta_1] = \frac{k}{n+1}. \quad (7.37)$$

Next lecture: moment generating functions (MGFs).

End of Lecture 7.

Lecture 8: Sums of Independent Random Variables and Moment Generating Functions

EECS 126 (UC Berkeley)

Spring 2026 — February 12, 2026

Reading: Bertsekas & Tsitsiklis, §4.4–4.6.

These are my personal notes from Professor Kannan Ramchandran's EECS 126 (Probability and Random Processes) course at UC Berkeley.

Topics: sums of independent random variables (convolution), tail bounds, and moment generating functions (MGFs).

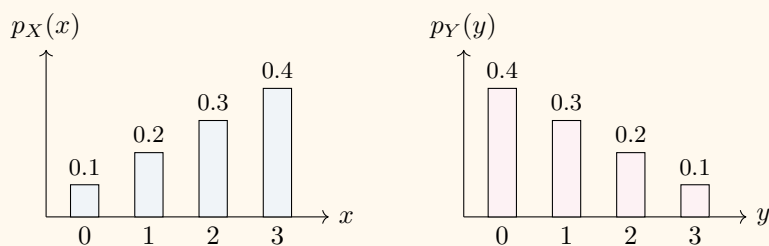
8.1 Discrete Convolution

Example 8.1 (Sum of two independent discrete random variables)

Let X and Y be independent, each taking values in $\{0, 1, 2, 3\}$, with

$$p_X = (0.1, 0.2, 0.3, 0.4), \quad p_Y = (0.4, 0.3, 0.2, 0.1) \quad (8.1)$$

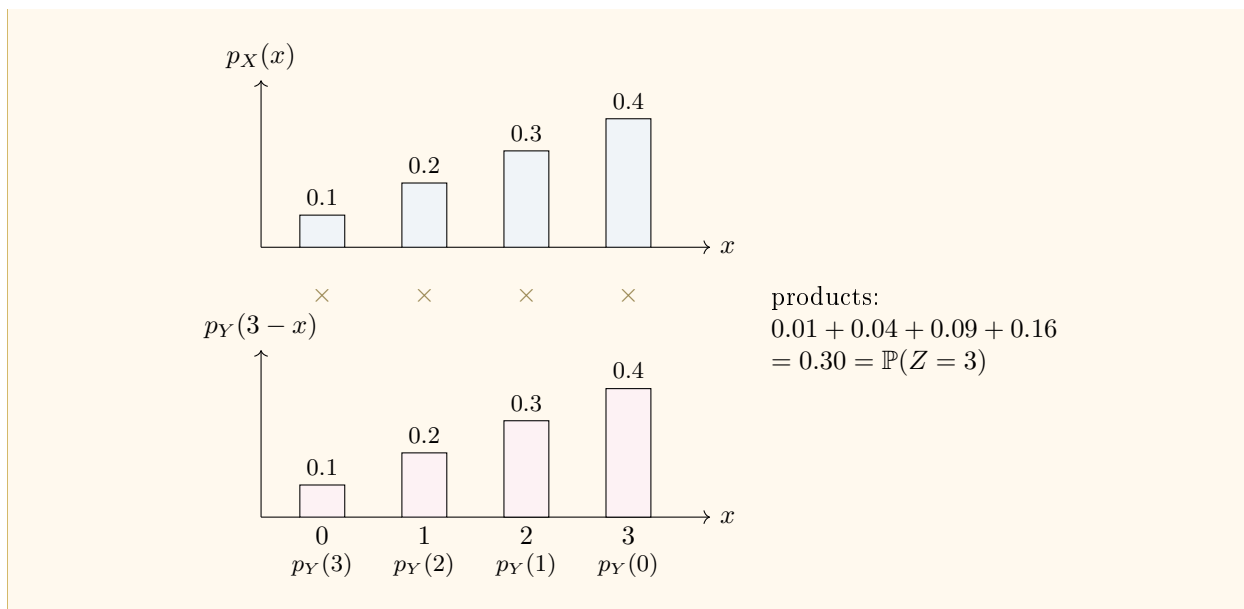
at the values 0, 1, 2, 3 respectively. Let $Z = X + Y$. Find $\mathbb{P}(Z = 3)$.



$Z = 3$ occurs exactly for the pairs $(x, y) \in \{(0, 3), (1, 2), (2, 1), (3, 0)\}$, i.e. $y = 3 - x$:

$$\begin{aligned} \mathbb{P}(Z = 3) &= \mathbb{P}(X + Y = 3) = \sum_{(x,y): x+y=3} \mathbb{P}(X = x, Y = y) \\ &= \sum_{(x,y): x+y=3} \mathbb{P}(X = x) \mathbb{P}(Y = y) \quad (\text{independence}) \\ &= \sum_x \mathbb{P}(X = x) \mathbb{P}(Y = 3 - x) \\ &= p_X(0) p_Y(3) + p_X(1) p_Y(2) + p_X(2) p_Y(1) + p_X(3) p_Y(0) \\ &= (0.1)(0.1) + (0.2)(0.2) + (0.3)(0.3) + (0.4)(0.4) \\ &= 0.01 + 0.04 + 0.09 + 0.16 = 0.30. \end{aligned} \quad (8.2)$$

To see the products line up, flip p_Y and shift it by 3: plot $p_Y(3 - x)$ against x , directly under $p_X(x)$, then multiply bar by bar and add.



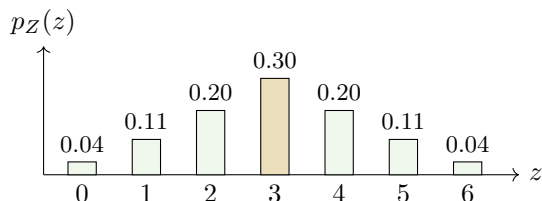
Definition 8.1 (Discrete convolution)

If X and Y are independent and integer-valued, then $Z = X + Y$ has PMF

$$p_Z(z) = (p_X * p_Y)(z) \triangleq \sum_x p_X(x) p_Y(z - x). \quad (8.3)$$

Here x is a dummy summation variable.

Applying (8.3) at every z in the example gives the full PMF of Z :



Check: $0.04 + 0.11 + 0.20 + 0.30 + 0.20 + 0.11 + 0.04 = 1$.

8.2 Continuous Convolution

Theorem 8.1 (Density of a sum of independent random variables)

If X and Y are independent continuous random variables and $Z = X + Y$, then

$$f_Z(z) = (f_X * f_Y)(z) \triangleq \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx. \quad (8.4)$$

Here x is a dummy variable of integration.

Proof by conditioning. Find the conditional CDF of Z given $X = x$, then differentiate:

$$\begin{aligned}
 F_{Z|X}(z | x) &= \mathbb{P}(Z \leq z | X = x) = \mathbb{P}(X + Y \leq z | X = x) \\
 &= \mathbb{P}(x + Y \leq z | X = x) \\
 &= \mathbb{P}(Y \leq z - x | X = x) \\
 &= \mathbb{P}(Y \leq z - x) && \text{(independence)} \\
 &= F_Y(z - x).
 \end{aligned} \tag{8.5}$$

Differentiating in z ,

$$f_{Z|X}(z | x) = \frac{d}{dz} F_Y(z - x) = f_Y(z - x). \tag{8.6}$$

Multiply both sides by the marginal $f_X(x)$; the left side becomes the joint PDF of X and Z :

$$f_{X,Z}(x, z) = f_{Z|X}(z | x) f_X(x) = f_Y(z - x) f_X(x). \tag{8.7}$$

To get the marginal f_Z , integrate over all possible x :

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Z}(x, z) dx = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx. \tag{8.8}$$

□

Why does the conditioning drop? I was unsure why we can remove “ $| X = x$ ”. Once $X = x$ is fixed, the only random variable left in $\{Y \leq z - x\}$ is Y , and Y is independent of X , so conditioning on $X = x$ does not change its distribution.

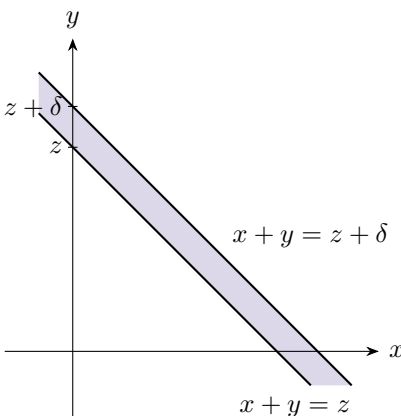
Proof by a thin strip. $\mathbb{P}(z \leq Z \leq z + \delta)$ is the probability that (X, Y) lands in the strip between the lines $y = z - x$ and $y = z + \delta - x$:

$$\begin{aligned}
 \mathbb{P}(z \leq Z \leq z + \delta) &= \int_{-\infty}^{\infty} \int_{z-x}^{z-x+\delta} f_{X,Y}(x, y) dy dx \\
 &= \int_{-\infty}^{\infty} \int_{z-x}^{z-x+\delta} f_X(x) f_Y(y) dy dx && \text{(independence)} \\
 &= \int_{-\infty}^{\infty} f_X(x) \left[\int_{z-x}^{z-x+\delta} f_Y(y) dy \right] dx \\
 &\approx \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) \delta dx,
 \end{aligned} \tag{8.9}$$

where the inner integral over an interval of width δ is approximated by a rectangle of height $f_Y(z - x)$. Since also $\mathbb{P}(z \leq Z \leq z + \delta) \approx f_Z(z) \delta$,

$$f_Z(z) \delta \approx \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) \delta dx, \tag{8.10}$$

and dividing by δ and letting $\delta \rightarrow 0$ gives (8.4). □



The strip has negative slope because $Z = X + Y$ means $Y = Z - X$. What we integrate over the strip is the joint *density*, so $\mathbb{P}(z \leq Z \leq z + \delta)$ is the probability mass of the strip, not its area.

8.3 Example: Sum of Two Uniforms

Example 8.2 (Sum of two independent uniforms)

Let $X, Y \sim \text{Uniform}[0, 1]$ be independent, so $f_X(x) = \frac{1}{b-a} = \frac{1}{1-0} = 1$ on $[0, 1]$, and the same for f_Y . Find f_Z for $Z = X + Y$.

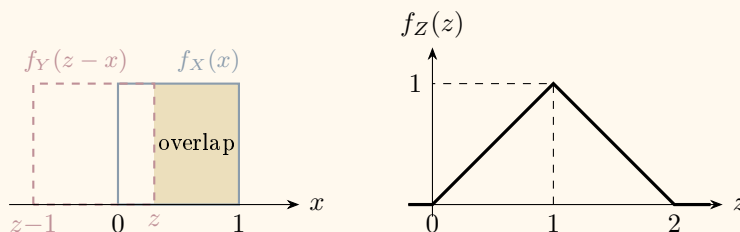
By (8.4), writing each density as an indicator,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx = \int_{-\infty}^{\infty} \mathbf{1}\{0 \leq x \leq 1\} \cdot \mathbf{1}\{0 \leq z-x \leq 1\} dx. \quad (8.11)$$

The second indicator is 1 when $z-1 \leq x \leq z$. So the integrand is 1 exactly on $[\max(0, z-1), \min(1, z)]$, and $f_Z(z)$ is the length of that interval:

$$\begin{aligned} 0 \leq z \leq 1: \quad f_Z(z) &= \int_0^z 1 dx = z, \\ 1 \leq z \leq 2: \quad f_Z(z) &= \int_{z-1}^1 1 dx = 1 - (z-1) = 2-z, \end{aligned} \quad (8.12)$$

and $f_Z(z) = 0$ otherwise. Check: the triangle has area $\frac{1}{2} \cdot 2 \cdot 1 = 1$.



Sliding overlap. As a function of x , $f_Y(z - x)$ is a box on $[z - 1, z]$ that slides to the right as z increases. $f_Z(z)$ is the length of its overlap with the box $[0, 1]$: it grows from 0 at $z = 0$, is largest at $z = 1$ (exact overlap), and shrinks back to 0 at $z = 2$. The same flip-and-slide picture is how convolution is used in signal processing.

8.4 Bounding Tail Probabilities

Example 8.3 (How big must a buffer be?)

Over n time slots, P_i packets arrive in slot i , where $P_1, \dots, P_n \sim \text{Poisson}(1)$ are i.i.d. A buffer holds $5n$ packets. How likely is it to overflow? We want to bound

$$\mathbb{P}\left(\sum_{i=1}^n P_i > 5n\right). \quad (8.13)$$

Theorem 8.2 (Markov's inequality)

If $X \geq 0$ and $a > 0$, then

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \quad (8.14)$$

Proof. Pointwise, $X \geq a \mathbf{1}\{X \geq a\}$: if $X \geq a$ the right side is $a \leq X$, and otherwise it is $0 \leq X$. Taking expectations, $\mathbb{E}[X] \geq a \mathbb{E}[\mathbf{1}\{X \geq a\}] = a \mathbb{P}(X \geq a)$. $\square$

Let $S_n = \sum_{i=1}^n P_i \geq 0$. By linearity of expectation and $\mathbb{E}[P_i] = 1$ for $\text{Poisson}(1)$,

$$\mathbb{P}(S_n > 5n) \leq \mathbb{P}(S_n \geq 5n) \leq \frac{\mathbb{E}[S_n]}{5n} = \frac{\sum_{i=1}^n \mathbb{E}[P_i]}{5n} = \frac{\sum_{i=1}^n 1}{5n} = \frac{n}{5n} = \frac{1}{5}. \quad (8.15)$$

Too loose. The bound $\frac{1}{5}$ does not shrink as n grows, even though S_n/n stays close to its mean 1, far below 5, for large n . Markov only uses $\mathbb{E}[S_n]$.

The exponential trick. For $t > 0$, the function $x \mapsto e^{tx}$ is strictly increasing, so the event is unchanged by applying it to both sides. Apply Markov to the nonnegative random variable e^{tS_n} :

$$\mathbb{P}(S_n > 5n) = \mathbb{P}(e^{tS_n} > e^{5nt}) \leq \frac{\mathbb{E}[e^{tS_n}]}{e^{5nt}}. \quad (8.16)$$

Why e^{tx} ? I was confused why this particular function. We want to apply a function T so that $\mathbb{E}[T(\sum_i P_i)]$ splits into a product $\prod_i \mathbb{E}[T(P_i)]$ by independence. That requires $T(x+y) = T(x)T(y)$, which the exponential satisfies: $T(x+y) = e^{t(x+y)} = e^{tx}e^{ty} = T(x)T(y)$. The parameter $t > 0$ is a scaling factor we are free to choose.

Since the P_i are independent and identically distributed,

$$\mathbb{E}[e^{tS_n}] = \mathbb{E}[e^{tP_1+tP_2+\dots+tP_n}] = \mathbb{E}[e^{tP_1}] \mathbb{E}[e^{tP_2}] \dots \mathbb{E}[e^{tP_n}] = \left(\mathbb{E}[e^{tP_1}]\right)^n. \quad (8.17)$$

For $P \sim \text{Poisson}(\lambda)$, using the Taylor series $e^x = \sum_{k \geq 0} \frac{x^k}{k!}$,

$$\mathbb{E}[e^{tP}] = \sum_{k=0}^{\infty} e^{tk} \frac{e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^t \lambda)^k}{k!} = e^{-\lambda} e^{e^t \lambda} = e^{\lambda(e^t - 1)}. \quad (8.18)$$

Substituting into (8.16),

$$\mathbb{P}(S_n > 5n) \leq \frac{(e^{\lambda(e^t - 1)})^n}{e^{5nt}} = e^{\lambda n(e^t - 1) - 5nt} \stackrel{\lambda=1}{=} e^{n(e^t - 1 - 5t)}. \quad (8.19)$$

Choosing t . The bound holds for every $t > 0$, so minimize the exponent $g(t) = e^t - 1 - 5t$:

$$g'(t) = e^t - 5 = 0 \implies t^* = \ln 5, \quad g''(t) = e^t > 0, \quad (8.20)$$

$$g(t^*) = e^{\ln 5} - 1 - 5 \ln 5 = 4 - 5 \ln 5 \approx -4.05. \quad (8.21)$$

Therefore

$$\mathbb{P}\left(\sum_{i=1}^n P_i > 5n\right) \leq e^{-(5 \ln 5 - 4)n} \approx e^{-4.05n}. \quad (8.22)$$

For $n = 10$: Markov gives 0.2, while this bound gives $e^{-40.5} \approx 2.7 \times 10^{-18}$.

Is the bound $\propto e^{-n}$? I was unsure about the rate. Any $t > 0$ with $g(t) < 0$ gives exponential decay $e^{ng(t)}$; the best rate is $5 \ln 5 - 4 \approx 4.05$, at $t^* = \ln 5$. This method (Markov applied to e^{tX} , then optimizing over t) is called the **Chernoff bound**. The quantity $\mathbb{E}[e^{tP}]$ is the MGF of P , defined next.

8.5 Moment Generating Functions

Definition 8.2 (Moment generating function)

The **moment generating function** of a random variable X is

$$M_X(s) \triangleq \mathbb{E}[e^{sX}] = \begin{cases} \sum e^{sx} p_X(x) & (\text{discrete}), \\ \int_{-\infty}^{\infty} e^{sx} f_X(x) dx & (\text{continuous}), \end{cases} \quad (8.23)$$

for $s \in \mathbb{R}$ (it may equal ∞ for some s).

MGFs are useful for:

1. tight tail bounds (Section 8.4);
2. computing all the moments $\mathbb{E}[X^n]$ of a distribution;
3. sums of independent random variables: convolution becomes multiplication.

Theorem 8.3 (MGFs generate moments)

If $M_X(s) < \infty$ for all s in an open interval around 0, then for every $n \geq 1$,

$$M_X^{(n)}(0) = \mathbb{E}[X^n]. \quad (8.24)$$

Proof. Here s is a constant inside the expectation and X is the random part, so $\frac{d}{ds}e^{sX} = Xe^{sX}$. Differentiating n times and moving the derivative inside the expectation (allowed because M_X is finite near 0),

$$\frac{d^n}{ds^n}M_X(s) = \frac{d^n}{ds^n}\mathbb{E}[e^{sX}] = \mathbb{E}\left[\frac{d^n}{ds^n}e^{sX}\right] = \mathbb{E}[X^n e^{sX}]. \quad (8.25)$$

At $s = 0$, $e^{0 \cdot X} = e^0 = 1$, so $M_X^{(n)}(0) = \mathbb{E}[X^n e^{sX}]|_{s=0} = \mathbb{E}[X^n]$. □

Example 8.4 (MGF of the exponential)

Let $X \sim \text{Exponential}(\lambda)$: $f_X(x) = \lambda e^{-\lambda x}$ for $x \in [0, \infty)$. Then

$$M_X(s) = \int_0^\infty e^{sx} \lambda e^{-\lambda x} dx = \lambda \int_0^\infty e^{(s-\lambda)x} dx. \quad (8.26)$$

If $s < \lambda$, the exponent $s - \lambda$ is negative and

$$M_X(s) = \lambda \left[\frac{1}{s-\lambda} e^{(s-\lambda)x} \right]_0^\infty = \lambda \left(0 - \frac{1}{s-\lambda} \right) = \frac{\lambda}{\lambda-s}. \quad (8.27)$$

If $s \geq \lambda$, the integrand $e^{(s-\lambda)x} \geq 1$ for all $x \geq 0$, so the integral diverges:

$$M_X(s) = \begin{cases} \frac{\lambda}{\lambda-s} & s < \lambda, \\ \infty & s \geq \lambda. \end{cases} \quad (8.28)$$

Moments. $M'_X(s) = \frac{\lambda}{(\lambda-s)^2}$ and $M''_X(s) = \frac{2\lambda}{(\lambda-s)^3}$, so

$$\mathbb{E}[X] = M'_X(0) = \frac{1}{\lambda}, \quad \mathbb{E}[X^2] = M''_X(0) = \frac{2}{\lambda^2}, \quad \text{Var}(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}. \quad (8.29)$$

Theorem 8.4 (MGFs determine the distribution)

If $M_X(s) = M_Y(s) < \infty$ for all s in some open interval $(-\varepsilon, \varepsilon)$ around 0, then X and Y have the same distribution.

So if we recognize an MGF, we know the distribution. (The proof is beyond the scope of this course.)

Not finished in lecture. We ran out of time after starting the list of MGF properties. The main properties are filled in below.

Theorem 8.5 (Properties of MGFs)

1. $M_X(0) = 1$.
2. If $Y = aX + b$, then $M_Y(s) = e^{sb} M_X(as)$.
3. If X and Y are independent, then $M_{X+Y}(s) = M_X(s) M_Y(s)$. More generally, for independent $X_1, \dots, X_n$, $M_{X_1+\dots+X_n}(s) = \prod_{i=1}^n M_{X_i}(s)$, and $(M_{X_1}(s))^n$ if they are also identically distributed.

Proof. 1. $M_X(0) = \mathbb{E}[e^{0 \cdot X}] = \mathbb{E}[1] = 1$.

2. $M_Y(s) = \mathbb{E}[e^{s(aX+b)}] = \mathbb{E}[e^{sb} e^{(as)X}] = e^{sb} \mathbb{E}[e^{(as)X}] = e^{sb} M_X(as)$, since e^{sb} is a constant.
3. $M_{X+Y}(s) = \mathbb{E}[e^{s(X+Y)}] = \mathbb{E}[e^{sX} e^{sY}] = \mathbb{E}[e^{sX}] \mathbb{E}[e^{sY}] = M_X(s) M_Y(s)$, using independence of e^{sX} and e^{sY} . The n -term version follows by induction. $\square$

Property 3 is the payoff: the density of $X + Y$ requires a convolution integral, but its MGF is just a product.

8.6 MGF Proofs for Sums of Independent Random Variables

Lemma 8.1 (MGF of the Gaussian)

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then for all $s \in \mathbb{R}$,

$$M_X(s) = e^{\mu s + \sigma^2 s^2 / 2}. \quad (8.30)$$

Proof. First let $Z \sim \mathcal{N}(0, 1)$. Complete the square in the exponent, $z^2 - 2sz = (z - s)^2 - s^2$:

$$\begin{aligned} M_Z(s) &= \int_{-\infty}^{\infty} e^{sz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z^2 - 2sz)/2} dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-s)^2/2 + s^2/2} dz \\ &= e^{s^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-s)^2/2} dz = e^{s^2/2} \cdot 1, \end{aligned} \quad (8.31)$$

since the last integrand is the $\mathcal{N}(s, 1)$ density. Now $X = \sigma Z + \mu$ (Lecture 7), so by Property 2 with $a = \sigma$, $b = \mu$,

$$M_X(s) = e^{\mu s} M_Z(\sigma s) = e^{\mu s} e^{\sigma^2 s^2 / 2} = e^{\mu s + \sigma^2 s^2 / 2}. \quad (8.32)$$

$\square$

Theorem 8.6 (Sum of independent Gaussians)

If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are independent, then

$$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2). \quad (8.33)$$

Proof. By Property 3 and the Gaussian MGF,

$$\begin{aligned} M_{X+Y}(s) &= M_X(s) M_Y(s) = e^{\mu_X s + \sigma_X^2 s^2 / 2} \cdot e^{\mu_Y s + \sigma_Y^2 s^2 / 2} \\ &= e^{(\mu_X + \mu_Y)s + (\sigma_X^2 + \sigma_Y^2)s^2 / 2}, \end{aligned} \quad (8.34)$$

which is the MGF of $\mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$, finite for all s . Since MGFs determine the distribution, $X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$. $\square$

This completes the proof left as TBD in Lecture 7.

Corollary 8.1 (Sum of independent Poissons)

If $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ are independent, then $X + Y \sim \text{Poisson}(\lambda + \mu)$.

Proof. By (8.18), $M_X(s) = e^{\lambda(e^s-1)}$ and $M_Y(s) = e^{\mu(e^s-1)}$, so

$$M_{X+Y}(s) = e^{\lambda(e^s-1)} e^{\mu(e^s-1)} = e^{(\lambda+\mu)(e^s-1)}, \quad (8.35)$$

the MGF of $\text{Poisson}(\lambda + \mu)$. (Compare the convolution proof of Poisson merging in Lecture 4.) $\square$

End of Lecture 8.

Lecture 9: Concentration Inequalities

EECS 126 (UC Berkeley)

Spring 2026 — February 17, 2026

Reading: Bertsekas & Tsitsiklis, §5.1; supplementary notes on concentration inequalities.

These are my personal notes from Professor Kannan Ramchandran's EECS 126 (Probability and Random Processes) course at UC Berkeley.

Topics: Markov's inequality, Chebyshev's inequality, the Chernoff bound, Jensen's inequality, Hoeffding's inequality, and a first look at the law of large numbers and the central limit theorem. Each bound trades more information about X for a tighter estimate of a tail probability.

9.1 Markov's Inequality

Theorem 9.1 (Markov's inequality)

If $X \geq 0$ and $a > 0$, then

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \quad (9.1)$$

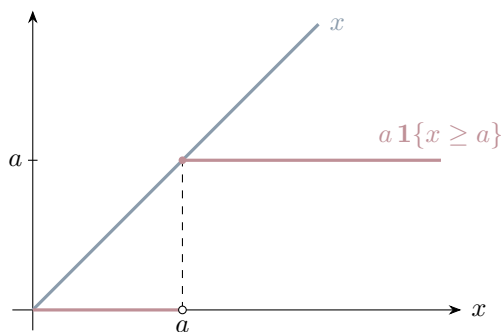
Proof. Compare X with the random variable $a \mathbf{1}\{X \geq a\}$ pointwise:

- if $X \geq a$: $a \mathbf{1}\{X \geq a\} = a \leq X$;
- if $X < a$: $a \mathbf{1}\{X \geq a\} = 0 \leq X$, since $X \geq 0$.

So $X \geq a \mathbf{1}\{X \geq a\}$ always. Taking expectations, and using $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)$,

$$\mathbb{E}[X] \geq a \mathbb{E}[\mathbf{1}\{X \geq a\}] = a \mathbb{P}(X \geq a). \quad (9.2)$$

Dividing by $a > 0$ gives (9.1). □



For $x \geq 0$, the line x lies above the step $a \mathbf{1}\{x \geq a\}$; averaging both sides gives Markov.

Example 9.1 (Markov for an exponential)

Let $X \sim \text{Exponential}(1)$, so $\mathbb{E}[X] = 1$ (Lecture 8) and $\mathbb{P}(X \geq a) = \int_a^\infty e^{-x} dx = \left[-e^{-x}\right]_a^\infty = e^{-a}$. Markov gives

$$\mathbb{P}(X \geq a) \leq \frac{1}{a}. \quad (9.3)$$

At $a = 5$: Markov gives 0.2, while the exact value is $e^{-5} \approx 0.0067$.

Markov cannot be improved using only the mean. Fix $m = \mathbb{E}[X]$ and $a \geq m$. The random variable

$$X = \begin{cases} a & \text{w.p. } \frac{m}{a}, \\ 0 & \text{w.p. } 1 - \frac{m}{a} \end{cases} \quad (9.4)$$

has $\mathbb{E}[X] = a \cdot \frac{m}{a} = m$ and $\mathbb{P}(X \geq a) = \frac{m}{a} = \frac{\mathbb{E}[X]}{a}$, so (9.1) holds with equality. To do better, we need more information about X .

9.2 Chebyshev's Inequality

Theorem 9.2 (Chebyshev's inequality)

If X has mean μ and variance σ^2 , then for every $c > 0$,

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}. \quad (9.5)$$

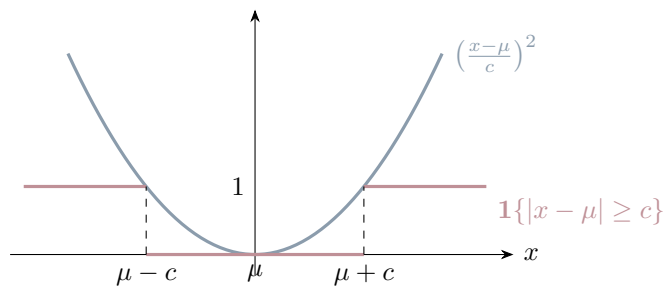
With $c = k\sigma$:

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \quad (9.6)$$

Proof. Both sides of $|X - \mu| \geq c$ are nonnegative, so squaring does not change the event: $\{|X - \mu| \geq c\} = \{(X - \mu)^2 \geq c^2\}$. Apply Markov to the nonnegative random variable $(X - \mu)^2$ with $a = c^2$:

$$\mathbb{P}(|X - \mu| \geq c) = \mathbb{P}((X - \mu)^2 \geq c^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{c^2} = \frac{\sigma^2}{c^2}. \quad (9.7)$$

□



The parabola $\left(\frac{x-\mu}{c}\right)^2$ lies above the indicator; averaging both sides gives Chebyshev. It uses the variance, not just the mean.

Example 9.2 (Chebyshev for a uniform)

Let $X \sim \text{Uniform}[0, 1]$. Then $\mu = \frac{1}{2}$, $\mathbb{E}[X^2] = \int_0^1 x^2 dx = \frac{1}{3}$, and

$$\sigma^2 = \mathbb{E}[X^2] - \mu^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}. \quad (9.8)$$

With $c = 0.4$, the exact value is $\mathbb{P}(|X - \frac{1}{2}| \geq 0.4) = \mathbb{P}(X \leq 0.1) + \mathbb{P}(X \geq 0.9) = 0.1 + 0.1 = 0.2$,

while

$$\mathbb{P}(|X - \tfrac{1}{2}| \geq 0.4) \leq \frac{1/12}{(0.4)^2} = \frac{1}{12 \cdot 0.16} = \frac{1}{1.92} \approx 0.52. \quad (9.9)$$

Example 9.3 (Polling with Chebyshev)

To estimate the fraction p of voters who support a candidate, poll n people independently. Let $X_i \sim \text{Bernoulli}(p)$ indicate support and $M_n = \frac{1}{n} \sum_{i=1}^n X_i$. How large must n be so that $\mathbb{P}(|M_n - p| \geq 0.01) \leq 0.05$?

$\mathbb{E}[M_n] = p$ and $\text{Var}(M_n) = \frac{p(1-p)}{n}$ (Lecture 4 and Lecture 10). Since p is unknown, bound the variance:

$$p(1-p) = \frac{1}{4} - \left(p - \frac{1}{2}\right)^2 \leq \frac{1}{4}. \quad (9.10)$$

By Chebyshev with $c = \varepsilon$,

$$\mathbb{P}(|M_n - p| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}. \quad (9.11)$$

Setting $\frac{1}{4n\varepsilon^2} \leq \delta$ with $\varepsilon = 0.01$, $\delta = 0.05$:

$$n \geq \frac{1}{4\varepsilon^2\delta} = \frac{1}{4(0.01)^2(0.05)} = \frac{1}{0.00002} = 50,000. \quad (9.12)$$

9.3 The Chernoff Bound

Theorem 9.3 (Chernoff bound)

For every b and every $s > 0$,

$$\mathbb{P}(X \geq b) \leq e^{-sb} M_X(s), \quad M_X(s) = \mathbb{E}[e^{sX}], \quad (9.13)$$

and therefore

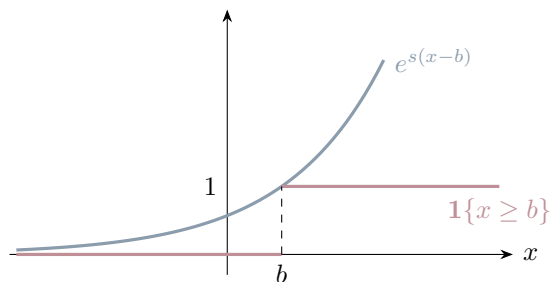
$$\mathbb{P}(X \geq b) \leq \min_{s>0} e^{-sb} M_X(s). \quad (9.14)$$

For the lower tail, $\mathbb{P}(X \leq b) \leq \min_{s<0} e^{-sb} M_X(s)$.

Proof. For $s > 0$, $x \mapsto e^{sx}$ is strictly increasing, so $\{X \geq b\} = \{e^{sX} \geq e^{sb}\}$. Apply Markov to the nonnegative random variable e^{sX} :

$$\mathbb{P}(X \geq b) = \mathbb{P}(e^{sX} \geq e^{sb}) \leq \frac{\mathbb{E}[e^{sX}]}{e^{sb}} = e^{-sb} M_X(s). \quad (9.15)$$

This holds for every $s > 0$, so it holds for the minimizing s . For $s < 0$, $x \mapsto e^{sx}$ is strictly decreasing, so $\{X \leq b\} = \{e^{sX} \geq e^{sb}\}$, and the same argument gives the lower-tail bound. $\square$



For any $s > 0$, the exponential $e^{s(x-b)}$ lies above the indicator $\mathbf{1}\{x \geq b\}$; averaging gives $\mathbb{P}(X \geq b) \leq e^{-sb} M_X(s)$, and we choose s to make the curve hug the indicator as tightly as possible.

Example 9.4 (Gaussian tail)

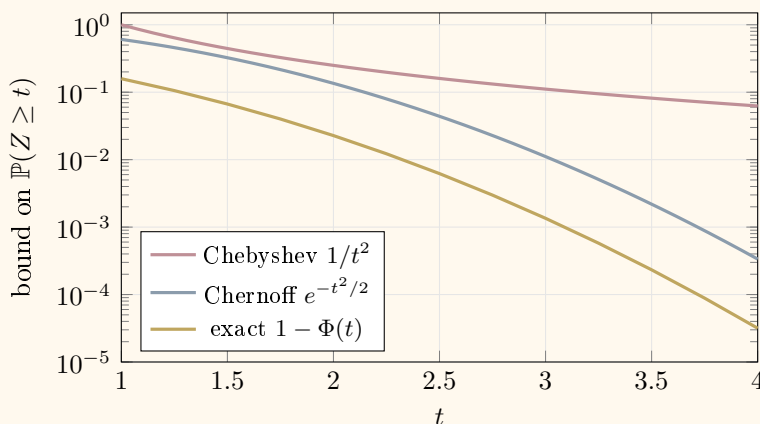
Let $Z \sim \mathcal{N}(0, 1)$, so $M_Z(s) = e^{s^2/2}$ (Lecture 8). For $t > 0$,

$$\mathbb{P}(Z \geq t) \leq \min_{s>0} e^{-st+s^2/2}. \quad (9.16)$$

The exponent $h(s) = \frac{s^2}{2} - st$ has $h'(s) = s - t = 0$ at $s^* = t$, and $h(t) = \frac{t^2}{2} - t^2 = -\frac{t^2}{2}$. Therefore

$$\mathbb{P}(Z \geq t) \leq e^{-t^2/2}. \quad (9.17)$$

Compare with Chebyshev, which gives $\mathbb{P}(Z \geq t) \leq \mathbb{P}(|Z| \geq t) \leq \frac{1}{t^2}$:



Chebyshev decays polynomially; Chernoff decays like the true tail, up to a smaller factor.

Example 9.5 (Chernoff for coin flips)

Let $S_n \sim \text{Binomial}(n, p)$, a sum of n independent Bernoulli(p) flips, so $M_{S_n}(s) = (1-p+pe^s)^n$: each flip has $\mathbb{E}[e^{sX_i}] = (1-p)e^0 + pe^s$, and the MGF of an independent sum is the product (Lecture 8). For $p < q < 1$,

$$\mathbb{P}(S_n \geq qn) \leq \min_{s>0} e^{-sqn} (1-p+pe^s)^n = \min_{s>0} \exp\left(n[-sq + \ln(1-p+pe^s)]\right). \quad (9.18)$$

Set the derivative of the bracket to zero:

$$-q + \frac{pe^s}{1-p+pe^s} = 0 \implies pe^s = q(1-p+pe^s) \implies pe^s(1-q) = q(1-p) \implies e^{s^*} = \frac{q(1-p)}{p(1-q)}, \quad (9.19)$$

and $s^* > 0$ since $q > p$. At s^* , $1-p+pe^{s^*} = 1-p + \frac{q(1-p)}{1-q} = \frac{(1-p)(1-q+q)}{1-q} = \frac{1-p}{1-q}$, so the bracket equals

$$\begin{aligned} -q \ln \frac{q(1-p)}{p(1-q)} + \ln \frac{1-p}{1-q} &= -q \ln \frac{q}{p} - q \ln \frac{1-p}{1-q} + \ln \frac{1-p}{1-q} \\ &= -q \ln \frac{q}{p} + (1-q) \ln \frac{1-p}{1-q} \\ &= -\left[q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p} \right] \triangleq -D(q \| p). \end{aligned} \quad (9.20)$$

Therefore

$$\mathbb{P}(S_n \geq qn) \leq e^{-n D(q \| p)}. \quad (9.21)$$

Numbers. $n = 100$ fair flips ($p = \frac{1}{2}$), at least 75 heads ($q = \frac{3}{4}$):

$$D\left(\frac{3}{4} \parallel \frac{1}{2}\right) = \frac{3}{4} \ln \frac{3/4}{1/2} + \frac{1}{4} \ln \frac{1/4}{1/2} = \frac{3}{4} \ln 1.5 + \frac{1}{4} \ln 0.5 \approx 0.304 - 0.173 = 0.131. \quad (9.22)$$

Markov	$\frac{\mathbb{E}[S_n]}{75} = \frac{50}{75}$	≈ 0.67
Chebyshev	$\mathbb{P}(S_n - 50 \geq 25) \leq \frac{\text{Var}(S_n)}{25^2} = \frac{25}{625}$	$= 0.04$
Chernoff	$e^{-100 D(3/4 \ 1/2)} \approx e^{-13.1}$	$\approx 2.1 \times 10^{-6}$
Exact	$\sum_{k=75}^{100} \binom{100}{k} 2^{-100}$	$\approx 2.8 \times 10^{-7}$

9.4 Jensen's Inequality

Definition 9.1 (Convex function)

A function g is **convex** if for all x, y and $\lambda \in [0, 1]$,

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y) : \quad (9.23)$$

every chord lies above the graph. If g is differentiable, convexity is equivalent to every tangent line lying below the graph:

$$g(x) \geq g(c) + g'(c)(x-c) \quad \text{for all } x, c. \quad (9.24)$$

g is **concave** if $-g$ is convex. Examples: x^2 , e^{sx} , $|x|$ are convex; $\ln x$ and $\sqrt{x}$ are concave on $(0, \infty)$.

Theorem 9.4 (Jensen's inequality)

If g is convex, then

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]. \quad (9.25)$$

If g is concave, the inequality reverses: $g(\mathbb{E}[X]) \geq \mathbb{E}[g(X)]$.

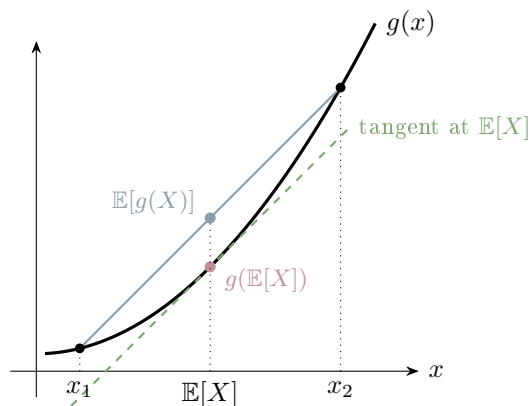
Proof. Take the tangent line (9.24) at $c = \mathbb{E}[X]$ and substitute $x = X$:

$$g(X) \geq g(\mathbb{E}[X]) + g'(\mathbb{E}[X])(X - \mathbb{E}[X]). \quad (9.26)$$

Take expectations. $g(\mathbb{E}[X])$ and $g'(\mathbb{E}[X])$ are constants, and $\mathbb{E}[X - \mathbb{E}[X]] = 0$:

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) + g'(\mathbb{E}[X]) \cdot 0 = g(\mathbb{E}[X]). \quad (9.27)$$

(For a convex g that is not differentiable, the same argument works with any supporting line through $(\mathbb{E}[X], g(\mathbb{E}[X]))$.) For concave g , apply the result to $-g$. $\square$



For X equal to x_1 or x_2 with probability $\frac{1}{2}$ each, $\mathbb{E}[g(X)]$ is the midpoint of the chord, which lies above $g(\mathbb{E}[X])$ on the curve.

Example 9.6 (Consequences of Jensen)

1. $g(x) = x^2$ (convex): $(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$, i.e. $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \geq 0$.
2. $g(x) = e^{sx}$ (convex): $e^{s\mathbb{E}[X]} \leq M_X(s)$, so an MGF is never smaller than $e^{s\mathbb{E}[X]}$.
3. $g(x) = \ln x$ (concave): $\mathbb{E}[\ln X] \leq \ln \mathbb{E}[X]$. For X uniform on positive numbers $\{x_1, \dots, x_n\}$,

$$\frac{1}{n} \sum_{i=1}^n \ln x_i \leq \ln \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \implies \left(\prod_{i=1}^n x_i \right)^{1/n} \leq \frac{1}{n} \sum_{i=1}^n x_i, \quad (9.28)$$

the arithmetic–geometric mean inequality.

4. In the St. Petersburg game (Lecture 5), $\mathbb{E}[\log_2 W] = 2$ is finite even though $\mathbb{E}[W] = \infty$, consistent with $\mathbb{E}[\log_2 W] \leq \log_2 \mathbb{E}[W]$.

9.5 Hoeffding's Inequality

Chernoff requires the MGF. For *bounded* random variables, we can bound the MGF without knowing the distribution.

Lemma 9.1 (Hoeffding's lemma)

If $a \leq X \leq b$ and $\mathbb{E}[X] = 0$, then for every $s \in \mathbb{R}$,

$$\mathbb{E}[e^{sX}] \leq e^{s^2(b-a)^2/8}. \quad (9.29)$$

Proof. Since $\mathbb{E}[X] = 0$, we have $a \leq 0 \leq b$. Write $x \in [a, b]$ as the convex combination $x = \lambda a + (1 - \lambda)b$ with $\lambda = \frac{b-x}{b-a}$. By convexity of e^{sx} (9.23),

$$e^{sx} \leq \frac{b-x}{b-a} e^{sa} + \frac{x-a}{b-a} e^{sb}. \quad (9.30)$$

Substitute $x = X$ and take expectations, using $\mathbb{E}[X] = 0$:

$$\mathbb{E}[e^{sX}] \leq \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb}. \quad (9.31)$$

Let $p = \frac{-a}{b-a} \in [0, 1]$ and $h = s(b-a)$, so $sa = -ph$ and $sb = h - ph$. The right side becomes

$$(1-p)e^{-ph} + pe^{h-ph} = e^{-ph}(1-p+pe^h) = e^{L(h)}, \quad L(h) = -ph + \ln(1-p+pe^h). \quad (9.32)$$

Differentiate:

$$\begin{aligned} L'(h) &= -p + \frac{pe^h}{1-p+pe^h}, & L(0) &= 0, & L'(0) &= -p + p = 0, \\ L''(h) &= \frac{pe^h(1-p+pe^h) - pe^h \cdot pe^h}{(1-p+pe^h)^2} = \frac{pe^h}{1-p+pe^h} \cdot \frac{1-p}{1-p+pe^h} = u(1-u) \leq \frac{1}{4}, \end{aligned} \quad (9.33)$$

where $u = \frac{pe^h}{1-p+pe^h} \in [0, 1]$ and $u(1-u) \leq \frac{1}{4}$ (as for $p(1-p)$ above). By Taylor's theorem with remainder, for some ξ between 0 and h ,

$$L(h) = L(0) + L'(0)h + \frac{L''(\xi)}{2} h^2 \leq 0 + 0 + \frac{1}{8} h^2 = \frac{s^2(b-a)^2}{8}. \quad (9.34)$$

□

Theorem 9.5 (Hoeffding's inequality)

Let $X_1, \dots, X_n$ be independent with $a_i \leq X_i \leq b_i$, and $S_n = \sum_{i=1}^n X_i$. For every $t > 0$,

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \quad (9.35)$$

In particular, if the X_i are i.i.d. in $[a, b]$ with mean μ and $M_n = \frac{S_n}{n}$,

$$\mathbb{P}(M_n - \mu \geq \varepsilon) \leq e^{-2n\varepsilon^2/(b-a)^2}, \quad \mathbb{P}(|M_n - \mu| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2/(b-a)^2}. \quad (9.36)$$

Proof. Let $C = \sum_i (b_i - a_i)^2$. Each centered term $X_i - \mathbb{E}[X_i]$ has mean 0 and lies in an interval of length $b_i - a_i$. By Chernoff, independence (Lecture 8), and Hoeffding's lemma, for $s > 0$,

$$\begin{aligned} \mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) &\leq e^{-st} \mathbb{E} \left[e^{s \sum_i (X_i - \mathbb{E}[X_i])} \right] = e^{-st} \prod_{i=1}^n \mathbb{E} \left[e^{s(X_i - \mathbb{E}[X_i])} \right] \\ &\leq e^{-st} \prod_{i=1}^n e^{s^2(b_i - a_i)^2/8} = \exp \left(-st + \frac{s^2 C}{8} \right). \end{aligned} \quad (9.37)$$

Minimize the exponent: $-t + \frac{sC}{4} = 0$ at $s^* = \frac{4t}{C}$, giving

$$-\frac{4t}{C}t + \frac{16t^2}{C^2} \cdot \frac{C}{8} = -\frac{4t^2}{C} + \frac{2t^2}{C} = -\frac{2t^2}{C}. \quad (9.38)$$

For the sample mean, $\{M_n - \mu \geq \varepsilon\} = \{S_n - n\mu \geq n\varepsilon\}$, so take $t = n\varepsilon$ and $C = n(b-a)^2$: $\frac{2t^2}{C} = \frac{2n^2\varepsilon^2}{n(b-a)^2} = \frac{2n\varepsilon^2}{(b-a)^2}$. The lower tail is the same bound applied to $-X_i$, and adding the two tails gives the factor 2. $\square$

Example 9.7 (Polling with Hoeffding)

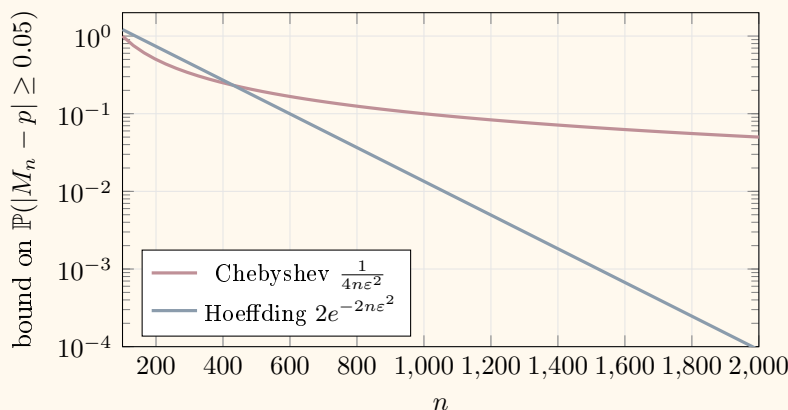
Return to polling: $X_i \in [0, 1]$, so $b - a = 1$. By (9.36),

$$\mathbb{P}(|M_n - p| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2} \leq \delta \iff n \geq \frac{\ln(2/\delta)}{2\varepsilon^2}. \quad (9.39)$$

With $\varepsilon = 0.01$, $\delta = 0.05$:

$$n \geq \frac{\ln 40}{2(0.01)^2} = \frac{3.689}{0.0002} \approx 18,445, \quad (9.40)$$

compared with 50,000 from Chebyshev. The two bounds as functions of n (with $\varepsilon = 0.05$):



Chebyshev decays like $\frac{1}{n}$; Hoeffding decays exponentially in n .

9.6 A First Look at the LLN and the CLT

Law of large numbers. For i.i.d. X_i with mean μ and variance σ^2 , Chebyshev gives

$$\mathbb{P}(|M_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \longrightarrow 0, \quad (9.41)$$

which is the weak law of large numbers (proved in full in Lecture 10).

Central limit theorem. $Z_n = \frac{\sqrt{n}(M_n - \mu)}{\sigma}$ is approximately $\mathcal{N}(0, 1)$ for large n (Lecture 10), so

$$\mathbb{P}(|M_n - \mu| \geq \varepsilon) = \mathbb{P}\left(|Z_n| \geq \frac{\varepsilon\sqrt{n}}{\sigma}\right) \approx 2\left(1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)\right). \quad (9.42)$$

For polling, $\sigma^2 \leq \frac{1}{4}$, and $2(1 - \Phi(1.96)) = 0.05$, so we need $\frac{\varepsilon\sqrt{n}}{\sigma} \geq 1.96$:

$$n \geq \frac{(1.96)^2 \sigma^2}{\varepsilon^2}, \quad \text{which holds for every } p \text{ once} \quad n \geq \frac{(1.96)^2}{4(0.01)^2} = 9,604. \quad (9.43)$$

This is an *approximation*, not a guaranteed bound like Chebyshev or Hoeffding.

Method	Requires	Decay	Polling n
Markov	$X \geq 0, \mathbb{E}[X]$	$1/a$	—
Chebyshev	$\text{Var}(X)$	$1/c^2, 1/n$	50,000
Chernoff	$M_X(s)$	exponential	—
Hoeffding	independent, bounded X_i	exponential in n	18,445
CLT	finite variance, large n	Gaussian (approximate)	9,604

End of Lecture 9.

Lecture 10: Laws of Large Numbers and the Central Limit Theorem

EECS 126 (UC Berkeley)

Spring 2026 — February 19, 2026

Reading: Bertsekas & Tsitsiklis, Ch. 5; supplementary notes on convergence.

These are my personal notes from Professor Kannan Ramchandran's EECS 126 (Probability and Random Processes) course at UC Berkeley.

Topics: the weak law of large numbers (WLLN), convergence in probability, almost sure convergence and the strong law of large numbers (SLLN), and the central limit theorem (CLT).

10.1 Recap: Concentration Inequalities

The following bounds were derived in Lecture 9; we restate the ones used below.

Theorem 10.1 (Markov's inequality)

If $X \geq 0$ and $a > 0$, then

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \quad (10.1)$$

Theorem 10.2 (Chebyshev's inequality)

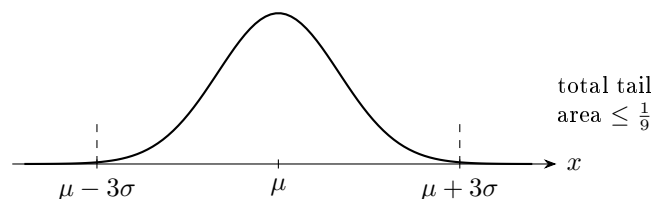
If X has mean μ and variance σ^2 , then for every $c > 0$,

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\text{Var}(X)}{c^2} = \frac{\sigma^2}{c^2}. \quad (10.2)$$

With $c = k\sigma$:

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \quad (10.3)$$

The left side is the probability that X deviates from its mean by at least c . For $k = 3$, the two tails beyond $\mu \pm 3\sigma$ have total probability at most $\frac{1}{9}$, for *any* distribution with finite variance.



Theorem 10.3 (Chernoff bound)

For every b ,

$$\mathbb{P}(X \geq b) \leq \min_{s>0} e^{-sb} M_X(s), \quad M_X(s) = \mathbb{E}[e^{sX}]. \quad (10.4)$$

Example 10.1 (Chernoff bound for a Gaussian tail)

Let $X \sim \mathcal{N}(0, \sigma^2)$, so $M_X(s) = e^{\sigma^2 s^2/2}$ (Lecture 8). With $b = k\sigma$,

$$\mathbb{P}(X \geq k\sigma) \leq \min_{s>0} e^{-sk\sigma} e^{\sigma^2 s^2/2} = \min_{s>0} e^{h(s)}, \quad h(s) = \frac{\sigma^2 s^2}{2} - k\sigma s. \quad (10.5)$$

Minimize the exponent:

$$h'(s) = \sigma^2 s - k\sigma = 0 \implies s^* = \frac{k}{\sigma}, \quad h(s^*) = \frac{\sigma^2}{2} \cdot \frac{k^2}{\sigma^2} - k\sigma \cdot \frac{k}{\sigma} = \frac{k^2}{2} - k^2 = -\frac{k^2}{2}. \quad (10.6)$$

Therefore

$$\mathbb{P}(X \geq k\sigma) \leq e^{-k^2/2}. \quad (10.7)$$

Comparison at $k = 3$ (two-sided, $\mathbb{P}(|X| \geq 3\sigma)$):

Chebyshev	$\frac{1}{k^2} = \frac{1}{9}$	≈ 0.111
Chernoff	$2e^{-k^2/2} = 2e^{-4.5}$	≈ 0.0222
Exact	$2(1 - \Phi(3))$	≈ 0.0027

Chebyshev uses only the variance, so it holds for every distribution and is loose for the Gaussian. Chernoff uses the whole MGF and is much tighter.

10.2 The Weak Law of Large Numbers

Let $X_1, X_2, \dots, X_n$ be i.i.d. with mean μ and variance $\sigma^2 < \infty$. The **sample mean** (empirical mean estimate) is

$$M_n = \frac{1}{n} \sum_{i=1}^n X_i. \quad (10.8)$$

Theorem 10.4 (Weak law of large numbers)

For every $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| \geq \varepsilon\right) \longrightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (10.9)$$

The probability that the sample mean deviates from the population mean by at least ε vanishes as $n \rightarrow \infty$.

Proof. Compute the mean and variance of M_n , then plug into Chebyshev. By linearity,

$$\mathbb{E}[M_n] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n} \cdot n\mu = \mu. \quad (10.10)$$

By $\text{Var}(aX) = a^2 \text{Var}(X)$ and independence (Lecture 4),

$$\text{Var}(M_n) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \left(\frac{1}{n}\right)^2 \text{Var}\left(\sum_{i=1}^n X_i\right) = \left(\frac{1}{n}\right)^2 \cdot n\sigma^2 = \frac{\sigma^2}{n}. \quad (10.11)$$

By Chebyshev (10.2) with $X = M_n$ and $c = \varepsilon$,

$$\mathbb{P}(|M_n - \mu| \geq \varepsilon) \leq \frac{\text{Var}(M_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (10.12)$$

□

Formally, for all $\varepsilon, \delta > 0$ there exists $n(\varepsilon, \delta)$ such that

$$\mathbb{P}(|M_n - \mu| \geq \varepsilon) \leq \delta \quad \text{for all } n \geq n(\varepsilon, \delta). \quad (10.13)$$

From (10.12), $\frac{\sigma^2}{n\varepsilon^2} \leq \delta \iff n \geq \frac{\sigma^2}{\varepsilon^2\delta}$, so $n(\varepsilon, \delta) = \frac{\sigma^2}{\varepsilon^2\delta}$ samples suffice. For example, with $\sigma^2 = 1$, $\varepsilon = 0.1$, $\delta = 0.05$: $n \geq \frac{1}{(0.01)(0.05)} = 2000$.

What are ε and δ capturing? I had a question of what ε captures in this statement.

- ε is the **accuracy**: how close M_n must be to μ .
- δ is the **confidence**: the probability of a deviation of at least ε is at most δ . (It bounds a probability, not the deviation itself.)
- $n(\varepsilon, \delta)$ is the number of samples needed for a given accuracy and confidence.

In words: for $n \geq n(\varepsilon, \delta)$, $|M_n - \mu| < \varepsilon$ with probability at least $1 - \delta$.

10.3 Convergence in Probability

Definition 10.1 (Convergence in probability)

A sequence $Y_1, Y_2, \dots$ **converges in probability** to a constant a , written $Y_n \xrightarrow{p} a$, if for every $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|Y_n - a| \geq \varepsilon) = 0. \quad (10.14)$$

The WLLN says exactly that $M_n \xrightarrow{p} \mu$.

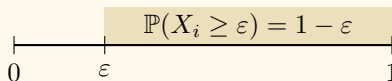
Example 10.2 (Minimum of uniforms)

Let $X_1, X_2, \dots$ be i.i.d. Uniform $[0, 1]$ and $Y_n = \min(X_1, \dots, X_n)$. Does Y_n converge in probability, and to what?

As we draw more samples, the smallest one should approach 0. Fix $0 < \varepsilon < 1$. Since $Y_n \geq 0$,

$$\begin{aligned} \mathbb{P}(|Y_n - 0| \geq \varepsilon) &= \mathbb{P}(Y_n \geq \varepsilon) = \mathbb{P}(\min(X_1, \dots, X_n) \geq \varepsilon) \\ &= \mathbb{P}(X_1 \geq \varepsilon, X_2 \geq \varepsilon, \dots, X_n \geq \varepsilon) \quad (\text{all must be } \geq \varepsilon) \\ &= \prod_{i=1}^n \mathbb{P}(X_i \geq \varepsilon) \quad (\text{independence}) \\ &= (1 - \varepsilon)^n \rightarrow 0 \quad \text{as } n \rightarrow \infty, \end{aligned} \quad (10.15)$$

since $0 < 1 - \varepsilon < 1$. (For $\varepsilon \geq 1$ the probability is 0.) Therefore $Y_n \xrightarrow{p} 0$.



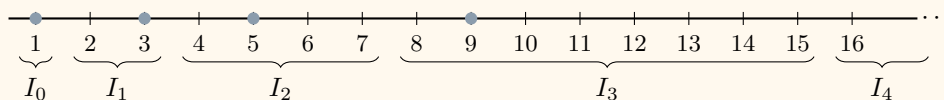
Example 10.3 (Discrete-time arrival process)

Time runs in slots $n = 1, 2, 3, \dots$, grouped into intervals

$$I_k = \{2^k, 2^k + 1, \dots, 2^{k+1} - 1\}, \quad k = 0, 1, 2, \dots, \quad (10.16)$$

so $I_0 = \{1\}$, $I_1 = \{2, 3\}$, $I_2 = \{4, \dots, 7\}$, $I_3 = \{8, \dots, 15\}$, and $|I_k| = 2^k$. There is exactly 1 arrival in each interval I_k , at a time chosen uniformly at random within I_k . Let

$$Y_n = \begin{cases} 1 & \text{if there is an arrival at time } n, \\ 0 & \text{otherwise.} \end{cases} \quad (10.17)$$



Each time in I_k is equally likely to hold the arrival, so

$$\begin{aligned} \mathbb{P}(Y_1 = 1) &= 1, \\ \mathbb{P}(Y_2 = 1) &= \mathbb{P}(Y_3 = 1) = \frac{1}{2}, \\ \mathbb{P}(Y_4 = 1) &= \dots = \mathbb{P}(Y_7 = 1) = \frac{1}{4}, \end{aligned} \quad (10.18)$$

and in general

$$\mathbb{P}(Y_n = 1) = \frac{1}{2^k} \quad \text{if } n \in I_k. \quad (10.19)$$

As $n \rightarrow \infty$, the interval index k containing n also goes to ∞ , so

$$\lim_{n \rightarrow \infty} \mathbb{P}(Y_n = 1) = \lim_{k \rightarrow \infty} \frac{1}{2^k} = 0. \quad (10.20)$$

For $0 < \varepsilon \leq 1$, $\mathbb{P}(|Y_n - 0| \geq \varepsilon) = \mathbb{P}(Y_n = 1) \rightarrow 0$ (and it is 0 for $\varepsilon > 1$), so $Y_n \xrightarrow{p} 0$.

Converging, yet arrivals keep coming. Initially I was wondering how Y_n can converge to 0 when, no matter which n I pick, infinitely many arrivals still come after it. This is why convergence in probability is a *weak* notion: it only says that at each fixed large n , an arrival at that particular time is unlikely. It says nothing about a single realization of the whole sequence, which still has $Y_n = 1$ once in every interval. This weakness belongs to convergence in probability, not to the WLLN itself.

10.4 Almost Sure Convergence and the Strong Law

Each outcome ω of the experiment produces a whole *realization*: an ordinary sequence of numbers $Y_1(\omega), Y_2(\omega), \dots$

Definition 10.2 (Almost sure convergence)

Y_n converges almost surely (a.s.) to Y , written $Y_n \xrightarrow{\text{a.s.}} Y$, if

$$\mathbb{P}\left(\left\{\omega : \lim_{n \rightarrow \infty} Y_n(\omega) = Y(\omega)\right\}\right) = 1. \quad (10.21)$$

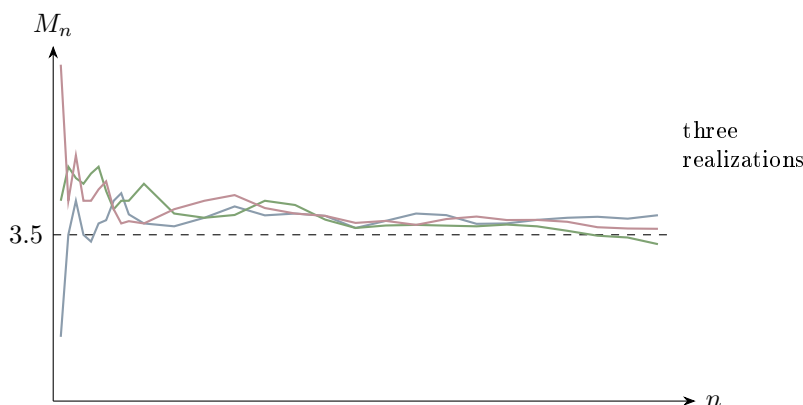
That is, the realization converges as an ordinary sequence for every outcome outside an event of probability 0.

Theorem 10.5 (Strong law of large numbers)

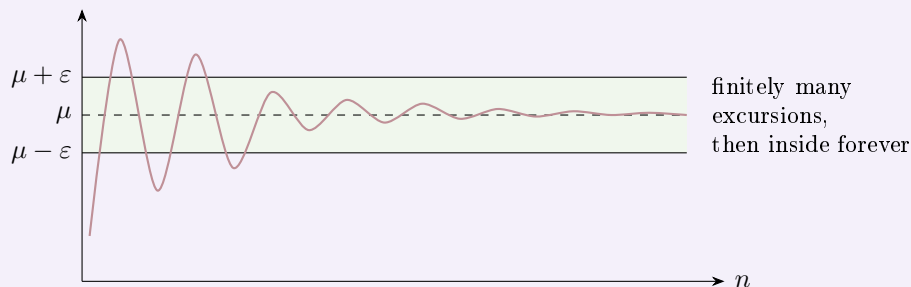
Let $\{X_n\}_{n \geq 1}$ be i.i.d. with mean μ . Then

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i = \mu\right) = 1, \quad \text{i.e.} \quad M_n \xrightarrow{\text{a.s.}} \mu. \quad (10.22)$$

(The proof is beyond the scope of this course.) For fair die rolls, $\mu = \frac{1+2+\dots+6}{6} = \frac{21}{6} = 3.5$, and every realization of M_n (except on an event of probability 0) converges to 3.5:



Excursions: SLLN versus WLLN. Initially I was wondering whether the SLLN allows any excursions at all outside a band $\mu \pm \varepsilon$. It does: almost every realization leaves the band only *finitely many* times, and after some point stays inside forever. Convergence in probability (the WLLN) only requires that, at each fixed large n , few realizations are outside the band; a single realization may still leave it infinitely often.



Theorem 10.6 (Almost sure convergence implies convergence in probability)

If $Y_n \xrightarrow{\text{a.s.}} Y$, then $Y_n \xrightarrow{p} Y$. The converse is not necessarily true.

Proof. Fix $\varepsilon > 0$ and let

$$A_n = \bigcup_{m \geq n} \{|Y_m - Y| \geq \varepsilon\}, \quad (10.23)$$

the event that the realization leaves the ε -band at some time $m \geq n$. Then $A_1 \supseteq A_2 \supseteq \dots$. An outcome in every A_n leaves the band infinitely often, so its realization does not converge to Y ; hence $\mathbb{P}(\bigcap_n A_n) \leq \mathbb{P}(Y_n \not\rightarrow Y) = 0$. By continuity of probability for decreasing events (a consequence of countable additivity, Lecture 1), $\mathbb{P}(A_n) \rightarrow \mathbb{P}(\bigcap_n A_n) = 0$. Since $\{|Y_n - Y| \geq \varepsilon\} \subseteq A_n$,

$$\mathbb{P}(|Y_n - Y| \geq \varepsilon) \leq \mathbb{P}(A_n) \rightarrow 0. \quad (10.24)$$

□

The converse fails: the arrival process. Every realization has exactly one arrival in each interval I_k , so $Y_n(\omega) = 1$ for infinitely many n and $Y_n(\omega) = 0$ for infinitely many n . The sequence $Y_n(\omega)$ never converges, for *every* outcome ω , so $\mathbb{P}(Y_n \rightarrow 0) = 0$. Thus $Y_n \xrightarrow{p} 0$ but Y_n does not converge almost surely.

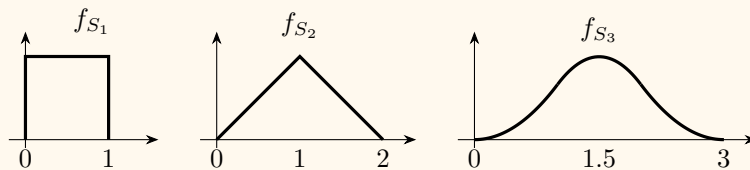
10.5 The Central Limit Theorem

Let $X_1, X_2, \dots$ be i.i.d. and $S_n = \sum_{i=1}^n X_i$. What happens to the distribution of S_n as n grows?

Example 10.4 (Sums of uniforms)

$X_i \sim \text{Uniform}[0, 1]$. f_{S_1} is a box and f_{S_2} is the triangle from Lecture 8. Convolving once more, $f_{S_3}(z) = \int_{z-1}^z f_{S_2}(x) dx$:

$$\begin{aligned} 0 \leq z \leq 1: & \int_0^z x dx = \frac{z^2}{2}, \\ 1 \leq z \leq 2: & \int_{z-1}^1 x dx + \int_1^z (2-x) dx = \frac{1-(z-1)^2}{2} + \left(2z - \frac{z^2}{2} - \frac{3}{2}\right) = \frac{-2z^2 + 6z - 3}{2}, \\ 2 \leq z \leq 3: & \int_{z-1}^2 (2-x) dx = \frac{(3-z)^2}{2}. \end{aligned} \quad (10.25)$$



f_{S_3} is symmetric about 1.5 and already looks bell-shaped.

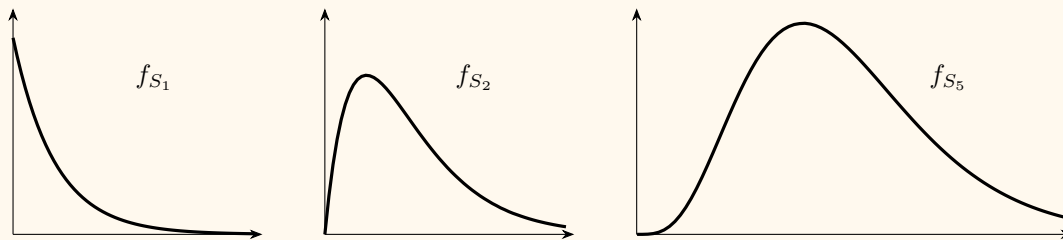
Example 10.5 (Sums of exponentials)

$X_i \sim \text{Exponential}(\lambda)$, so $f_{S_1}(x) = f_{X_1}(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. By convolution,

$$f_{S_2}(x) = \int_0^x \lambda e^{-\lambda u} \lambda e^{-\lambda(x-u)} du = \lambda^2 e^{-\lambda x} \int_0^x du = \lambda^2 x e^{-\lambda x}, \quad x \geq 0. \quad (10.26)$$

By induction, $f_{S_n}(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}$:

$$f_{S_{n+1}}(x) = \int_0^x \frac{\lambda^n u^{n-1} e^{-\lambda u}}{(n-1)!} \lambda e^{-\lambda(x-u)} du = \frac{\lambda^{n+1} e^{-\lambda x}}{(n-1)!} \int_0^x u^{n-1} du = \frac{\lambda^{n+1} x^n e^{-\lambda x}}{n!}. \quad (10.27)$$



Even starting from a very skewed density, f_{S_n} becomes bell-shaped.

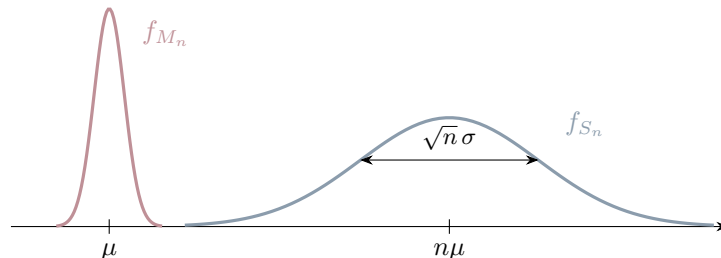
What happens to the mean and variance? By linearity and independence,

$$\mathbb{E}[S_n] = n\mu, \quad \text{Var}(S_n) = n\sigma^2. \quad (10.28)$$

The variance always grows, and the mean grows too unless $\mu = 0$, so f_{S_n} drifts and spreads out. What about the sample mean $M_n = \frac{S_n}{n}$?

$$\mathbb{E}[M_n] = \mu, \quad \text{Var}(M_n) = \frac{\sigma^2}{n} \longrightarrow 0, \quad (10.29)$$

so f_{M_n} collapses to a spike at μ . Neither shows the shape of the limit:



So M_n and S_n are both the wrong objects. The right normalization subtracts the mean and divides by the standard deviation of S_n :

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}. \quad (10.30)$$

Then

$$\mathbb{E}[Z_n] = \frac{\mathbb{E}[S_n] - n\mu}{\sqrt{n}\sigma} = \frac{n\mu - n\mu}{\sqrt{n}\sigma} = 0, \quad \text{Var}(Z_n) = \frac{\text{Var}(S_n)}{(\sqrt{n}\sigma)^2} = \frac{n\sigma^2}{n\sigma^2} = 1. \quad (10.31)$$

Standardizing does not make Z_n Gaussian. I had a question of whether the CLT just says that Z_n becomes Gaussian once it is normalized. $\mathbb{E}[Z_n] = 0$ and $\text{Var}(Z_n) = 1$ for every n , but for finite n , Z_n is generally *not* Gaussian (for uniform X_i , Z_n even has bounded support). The CLT is a statement about the *limit*: the CDF of Z_n approaches Φ as $n \rightarrow \infty$.

Definition 10.3 (Convergence in distribution)

Z_n **converges in distribution** to Z , written $Z_n \xrightarrow{d} Z$, if

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq x) = \mathbb{P}(Z \leq x) \quad (10.32)$$

at every x where the CDF of Z is continuous.

Theorem 10.7 (Central limit theorem)

Let $X_1, X_2, \dots$ be i.i.d. with mean μ and variance $0 < \sigma^2 < \infty$, and $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$. Then for every $x \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq x) = \Phi(x), \quad \Phi(x) = \mathbb{P}(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-u^2/2} du, \quad (10.33)$$

where $Z \sim \mathcal{N}(0, 1)$ and u is a dummy variable. That is, $Z_n \xrightarrow{d} \mathcal{N}(0, 1)$.

Implications.

1. In the limit, the distribution of Z_n “wipes out” all information about the distribution of the X_i except the mean and the variance.
2. If there are a large number of independent small disturbances, their aggregate effect is approximately normally distributed. This is why the Gaussian is a good model for noise.

The finite-variance assumption is needed: without it, Z_n need not approach a Gaussian.

Proof via MGFs. Assume $M_{X_1}(s) < \infty$ for s in an interval around 0. Standardize each term: $Y_i = \frac{X_i - \mu}{\sigma}$, i.i.d. with $\mathbb{E}[Y_i] = 0$ and $\mathbb{E}[Y_i^2] = \text{Var}(Y_i) = 1$. Then

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\sum_{i=1}^n (X_i - \mu)}{\sqrt{n}\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i. \quad (10.34)$$

By the MGF properties of Lecture 8 (scaling, then independent sums),

$$M_{Z_n}(s) = \mathbb{E}\left[e^{\frac{s}{\sqrt{n}} \sum_{i=1}^n Y_i}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\frac{s}{\sqrt{n}} Y_i}\right] = \left(M_Y\left(\frac{s}{\sqrt{n}}\right)\right)^n. \quad (10.35)$$

Taylor-expand M_Y around 0, using $M_Y^{(k)}(0) = \mathbb{E}[Y^k]$ (Lecture 8):

$$M_Y(t) = M_Y(0) + M_Y'(0)t + \frac{M_Y''(0)}{2}t^2 + o(t^2) = 1 + \mathbb{E}[Y]t + \frac{\mathbb{E}[Y^2]}{2}t^2 + o(t^2) = 1 + \frac{t^2}{2} + o(t^2). \quad (10.36)$$

With $t = \frac{s}{\sqrt{n}}$,

$$M_Y\left(\frac{s}{\sqrt{n}}\right) = 1 + \frac{s^2}{2n} + o\left(\frac{1}{n}\right). \quad (10.37)$$

Take logarithms and use $\ln(1+x) = x + O(x^2)$ as $x \rightarrow 0$:

$$\ln M_{Z_n}(s) = n \ln \left(1 + \frac{s^2}{2n} + o\left(\frac{1}{n}\right) \right) = n \left(\frac{s^2}{2n} + o\left(\frac{1}{n}\right) \right) = \frac{s^2}{2} + o(1) \rightarrow \frac{s^2}{2}. \quad (10.38)$$

So $M_{Z_n}(s) \rightarrow e^{s^2/2}$, which is the MGF of $\mathcal{N}(0, 1)$ (Lecture 8). Finally, we use without proof the *continuity theorem*: if $M_{Z_n}(s) \rightarrow M_Z(s)$ for all s in an interval around 0, then $Z_n \xrightarrow{d} Z$. Hence $Z_n \xrightarrow{d} \mathcal{N}(0, 1)$. $\square$

This proof assumes the MGF exists near 0. The CLT holds with only a finite variance; that more general proof uses characteristic functions and is beyond the scope of this course.

End of Lecture 10.