CMM 2026 Power Round

Caltech Math Meet

January 2026

Instructions

- You will have one week to work on the Power Round. Your solutions should be turned in by **12:00 PM (noon) on Sunday, January 18th** via Gradescope. No late submissions will be accepted. Submissions must be in the form of a PDF, and you may typeset solutions in LaTeX, or write them by hand. You can submit as many times as you want before the due date, but only the latest Power Round submission from your team will be graded. When you submit your solutions on Gradescope, you must assign pages to the correct problems.
- We strongly encourage you to LaTeX your solutions. If you do not know how to create a math symbol or do something in LaTeX, you may use Overleaf online guides, Detexify, or the TeX Stack Exchange.
- You may ask questions about the test on the Piazza forum, on which you may ask *public* or *private* questions. Public questions can be seen by all teams, so please do not reveal any parts of your solutions in public questions. However, you can ask a public question if you think that a question would be useful for everyone to see. We will not provide hints on any of the problems, but we will clarify ambiguities and fix any errors in the problems.
- Instructions for how to join Piazza and Gradescope will be emailed to coaches as well as posted in the CMM Discord server.
- On any problem, **you may cite without proof previous problems, theorems, and facts** in the Power Round, **as well as anything in the Appendix**. However, you may not cite future problems. The problems are graded separately, so it will be very confusing for us if you cite parts of your solutions to previous problems. If you want to use the same idea in multiple problems, please write it out each time.
- You must show your work and justify your reasoning, regardless of whether the problem says "show", "prove", "find", "compute", etc.
- **You may not use references such as books, calculators, online resources, computer programs, or generative AI, other than LaTeX help and dictionaries for teams whose members for whom English is a foreign language.** Except for asking questions on Piazza, you may not discuss the content of the Power Round with people outside your team.

Contents

Notation

- $\mathbb{Z}$: The set of integers
- $\mathbb{Z}_{>0}$: The positive integers
- $\mathbb{R}$: The real numbers
- $\mathbb{R}_{>0}$: The positive real numbers
- $[a, b]$: Interval of all real numbers between a and b, inclusive
- $\{A, B, C\}$: Set containing A, B, C
- $X \in Y$: X is in the set Y
- $X \subset Y$: X is a subset of Y not equaling Y
- $X \subseteq Y$: X is a subset of Y, possibly equal to Y
- $X \cap Y$: Intersection of X and Y
- $X \cup Y$: Union of X and Y
- $\exists$: There exists
- $\forall$: For all

1 Introduction to Probability and Statistics

1.1 Discrete Probability

Suppose that each event A has finitely many different outcomes $O_1, O_2, \ldots, O_k$. Then for an event O_1, we denote by $P(O_1)$ the chance of it happening. (Then since some event must be the outcome, $P(O_1) + P(O_2) + \cdots + P(O_k) = 1$.) We denote by $P(A \cap B)$ the chance of events A and B both happening, $P(A \cap B \cap C)$ the chance of events A, B and C all happening, etc.

For example, if one coin is flipped, $P(\text{head}) = \frac{1}{2}$ and $P(\text{tail}) = \frac{1}{2}$. If a random integer is chosen from the set $\{1 \ldots n\}$, then $P(1) = P(2) = \cdots = P(n) = \frac{1}{n}$.

1.1.1 Conditional Probability

We denote by $P(A \mid B)$ the probability that A happens given that B happens. For example, if x is the outcome of a die roll, then $P(x > 5 \mid x > 3) = \frac{1}{3}$. $P(x \text{ is even} \mid x^2 \text{ is even}) = 1$. If a and b are outcomes of two different coin flips, $P(a \text{ is a head} \mid b \text{ is a head}) = \frac{1}{2}$. Conditional probability is a notion of how *related* different random variables are, and how much they affect each other.

Remark 1.1 (Compound Events). $P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B)$.

Problem 1.1 (2 points). A family has two children, at least one of which was born on a Tuesday. What is the probability that the first one was born on a Tuesday?

Theorem 1.2 (Bayes' Theorem)
For events A and B,

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

Problem 1.2 (3 points). Suppose you are being tested for a disease which everyone has a 1% chance of having. If you have it, then the test will come up positive. If you don't, then the test will come up negative with 97% chance and positive with 3% chance. Given that the test came up positive, what is the probability that you actually have the disease?

Independent Events
Events A and B are **independent** if and only if $P(A) = P(A \mid B)$, or equivalently $P(B) = P(B \mid A)$.

Remark 1.3. Otherwise, we say that events A and B are *non-independent*, or *dependent*.

Remark 1.4 (Independent Probabilities Multiply). Events A and B are independent if and only if $P(A \cap B) = P(A)P(B)$.

For example, the outcome of coin flip A and the outcome of coin flip B are independent, because $P(A \text{ is a head}) = P(A \text{ is a head} \mid B \text{ is a head}) = \frac{1}{2}$, but the parity of die roll X and the parity of its square are not independent, since $P(X \text{ is even}) = \frac{1}{2}$ but $P(X \text{ is even} \mid X^2 \text{ is even}) = 1$.

Problem 1.3 (13 points). Determine whether or not each of the following pairs of events are independent:

a) (2 points) If x is the outcome of one die roll, (x is even) and (x is a multiple of 3)

b) (2 points) If x is the outcome of one die roll, (x is a multiple of 3) and (x is a multiple of 4)

c) (2 points) If x is the outcome of one die roll, (x is odd) and (x is prime)

d) (2 points) For a randomly chosen day, (it is a Friday) and (it is the 13th of a month). *(For the sake of simplicity, assume every month has* 31 *days, so every* 31 *days it becomes a new month.)*

e) (2 points) In a random permutation of $\{1, 2, 3\}$, (the 1st number is greater than the 2nd number) and (the 2nd number is greater than the 3rd number)

f) (3 points) In a sequence of coin flips, (the substring "HH" first appears before the substring "TT" does) and (the substring "HT" first appears before the substring "TH" does)

1.2 Continuous Probability

If a random variable takes on values in a continuous range, then instead of discrete probabilities, we will use a probability density function that yields relative likelihoods of outcomes.

Since there will be uncountably many outcomes, we will analyze the infinitesimally small increments with calculus. **If you are not already familiar with calculus, please read through Section 4** on derivatives and integrals, useful inequalities to know, and series and convergence.

1.2.1 Continuous Distributions

Probability Density Function

For a continuous random variable x sampled from a distribution X (so it can take on uncountably many values on a whole interval of the real line), the **probability density function** (PDF) is the function $f_X : \mathbb{R} \to \mathbb{R}$ whose value at any given point is the *relative likelihood* that the random variable will be equal to that value. We have $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$.

The *absolute likelihoods* of each outcome are all zero, but *relative likelihoods* show how probable different outcomes are compared to each other.

For example, for a random variable taking on real values in the interval $[0, 1]$ we can define the probability density function $f_X : [0, 1] \to \mathbb{R}$ such that $f(x)$ is the relative likelihood of x being the outcome, and all the probability densities sum to 1: $\int_0^1 f_X(x)\, dx = 1$.

Problem 1.4 (2 points). Suppose X is a random variable taking real values between 0 and 1 such that the likelihood of each outcome is proportional to its value. What is the appropriate probability density function f from $[0, 1]$ to $\mathbb{R}$?

Joint Probability Density Function

Given two random variables x and y, chosen from probability spaces X and Y, which may or may not be independent, their **joint probability density function** is the function $f_{X,Y} : X \times Y \to \mathbb{R}$ representing the relative likelihood of each ordered pair (x, y) occurring, which satisfies $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy\, dx = 1$.

Independent Events (Continuous Case)
Random variables x and y are taken from probability distributions X and Y such that f_X and f_Y are the probability density functions of x and y, and $f_{X,Y}$ is the joint probability density function for x and y. Then x and y are independent if and only if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

for all $(x,y) \in X \times Y$.

1.3 Expected Value

Discrete Expected Value
The **expected value** $E[x]$ of a *discrete* random variable x that has countably many outcomes is the average of its outcomes, weighted by how likely they are to occur. If $O_1, O_2, O_3, \ldots$ are all the possible outcomes, then $E[x] = \sum_{i=1}^{\infty} xP(O_i)$.

You can think of this as multiplying each outcome by a certain "weight", which is its chance of occurring.

For example, the expected value of a die roll is $\sum_{i=1}^{6} i \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = 3.5$. The expected value of the remainder when a die roll is divided by 4 is $\frac{1+2+3+0+1+2}{6} = \frac{3}{2}$.

Problem 1.5 (2 points). An unfair coin, with $P(\text{head}) = \frac{1}{n}$, is flipped repeatedly. What is the expected number of times it must be flipped until a head appears?

We can also extend this definition to the continuous case in which there are uncountably many outcomes, by replacing the summation with an integral:

Continuous Expected Value
If $f(x)$ is the probability density function of a *continuous* random variable x, then the expected value of x, denoted $E[x]$, is equal to $\int_{-\infty}^{\infty} xf(x)\,dx$.

For example, if a random variable takes on values in the interval $[0,1]$ with linear probability density function $f(x) = 2x$, then $E[x] = \int_{-\infty}^{\infty} x \cdot 2x\,dx = \int_{-\infty}^{\infty} 2x^2\,dx = \frac{2}{3}x^3|_0^1 = \frac{2}{3}$, which can also be found using centroid properties.

Undefined Expected Value
When the sum $\sum xP(x)$ (for discrete) or integral $\int xf_x(x)\,dx$ (for continuous) do not converge, expected value is said to be *undefined* and cannot be worked with.

Theorem 1.5 (Linearity of Expectation)
Expected value is linear, so for random variables X and Y which may be independent or non-independent, we have $E[X+Y] = E[X] + E[Y]$, and $E[aX+bY] = aE[X] + bE[Y]$ for any constants a and b.

Proof. Let random variables x and y be chosen from probability spaces X and Y respectively. For the discrete case,

$$E[x+y] = \sum_{(x,y)\in X\times Y} (x+y)P(x,y) = \sum_{(x,y)\in X\times Y} xP(x,y) + \sum_{(x,y)\in X\times Y} yP(x,y)$$

$$= \sum_{x\in X} x\left[\sum_{y\in Y} P(x,y)\right] + \sum_{y\in Y} y\left[\sum_{x\in X} P(x,y)\right]$$

$$= \sum_{x\in X} xP(x) + \sum_{y\in Y} yP(y) = E[x] + E[y].$$

For the continuous case, if $f_{X\times Y}(x,y) : X \times Y \to [0,1]$ is the probability density function which outputs the likelihood of each ordered pair outcome (x,y), $f_X : X \to [0,1]$ is the probability density function for x alone, and $f_Y : Y \to [0,1]$ is the probability density function for y alone, then we have

$$E[x+y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{X\times Y}(x,y)\,dy\,dx$$
$$= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}xf_{X\times Y}(x,y)\,dy\right]dx + \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}yf_{X\times Y}(x,y)\,dx\right]dy$$
$$= \int_{-\infty}^{\infty}x\left[\int_{-\infty}^{\infty}f_{X\times Y}(x,y)\,dy\right]dx + \int_{-\infty}^{\infty}y\left[\int_{-\infty}^{\infty}f_{X\times Y}(x,y)\,dx\right]dy$$
$$= \int_{-\infty}^{\infty}xf_X(x)\,dx + \int_{-\infty}^{\infty}yf_Y(y)\,dy = E[x]+E[y].$$

□

Theorem 1.6 (Expectation With Independent Events)
If random variables X and Y are **independent**, then $E[X] = E[X \mid Y = y]$ for every possible outcome y of Y.

Here, $E[X \mid Y = y]$ denotes the expected value of the variable X given that the random variable Y is equal to y.

Problem 1.6 (4 points). Prove Theorem 1.6 in the discrete case and the continuous case separately. (Hint: Use summations for the discrete case, and integral with probability density functions for the continuous case.)

Fact 1.7. If the random variable X is always less than or equal to the random variable Y, then $E[X] \le E[Y]$.

Problem 1.7 (4 points). Prove that $E[X] \le \sqrt{E[X^2]}$ for (1) a discrete random variable X, and (2) a continuous random variable X. (Hint: Use Holder's Inequality.)

Theorem 1.8 (Tower Rule)
For random variables x and y on the same probability space (not necessarily independent), $E[X] = E[E[X \mid Y]]$.

(Here, the random variable $E[X \mid Y]$ is defined on the probability space of possible outcomes of Y, and $E[E[X \mid Y]]$ is the expectation of this random variable over all possible outcomes of Y.)

Problem 1.8 (4 points). Prove Theorem 1.8 in the discrete case and the continuous case. (Hint: Use Fubini's Theorem.)

Theorem 1.9 (Independent Expectations Multiply)
For independent random variables x and y which have expected values $E[x]$ and $E[y]$, $E[xy] = E[x]E[y]$.

Problem 1.9 (4 points). Prove Theorem 1.9 in the discrete case and the continuous case.

Remark 1.10. Note that the expectation of the square of a die roll, $E[x^2] = \frac{91}{6}$, is not equal to the product of the expected values of the die rolls: $E[x]E[x] = E[x]^2 = \frac{49}{4}$, because die roll x is the same as die roll x and thus they are not independent. However, the expectation of the product of two different die rolls, $E[x_1x_2] = \frac{49}{4}$, *is* equal to the product of their individual die rolls $E[x_1]E[x_2] = \frac{49}{4}$, since they are completely distinct and independent events.

For the rest of this Power Round, we will assume that random variables are continuous (as opposed to discrete), unless otherwise specified.

1.4 Martingales

Martingale

A **discrete-time martingale** is a sequence of random variables $X_1, X_2, \ldots$, such that the expected value of the absolute value of each variable is bounded, and also the expected value of each variable given the values of all previous variables is the previous variable:

- $E[|X_n|] < \infty$
- $E[X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1] = X_{n-1}$.

An example of a discrete-time martingale is a sequence $X_1, X_2, \ldots$ in which a coin is flipped to determine whether each X_{k+1} is equal to $X_k + 1$ or $X_k - 1$. The first condition is satisfied because we end up at most n units away from the starting point after n steps, and the second condition is satisfied because X_n is equal to $X_{n-1} + 1$ with $\frac{1}{2}$ chance and $X_{n-1} - 1$ with $\frac{1}{2}$ chance, averaging to X_{n-1}.

Problem 1.10 (3 points). Prove that for a martingale, $E[X_n] = E[X_1]$ for all positive integers n.

Theorem 1.11 (Optional Stopping Theorem)

Let $X = (X_1, X_2, X_3, \ldots)$ be a discrete-time martingale. Define a *stopping function* f which takes in sequences of numbers of any finite length and outputs either True or False.

For each sample $(x_1, x_2, x_3, \ldots)$ from the martingale X, we define k to be the smallest positive integer for which $f(x_1, x_2, \ldots, x_k) =$ True. Assume that the value k exists with probability 1. Suppose that **at least one** of the following conditions holds:

- There exists a constant C_1 for which $k \leq C_1$ with probability 1. **(Bounded Time)**
- $E[k]$ is finite, and there exists a constant C_2 for which, if $k > t$, then $E[|x_{t+1} - x_t|] \leq C_2$ with probability 1. **(Bounded Increments and Bounded Expected Time)**
- There exists a constant C_3 such that $|x_{\min(t,k)}| \leq C_3$ with probability 1 for all positive integers t. **(Bounded Values)**

Then $E[x_k] = E[x_1]$.

Proof. We will prove that each of the conditions is sufficient. Suppose we have a stopping function f such that each sample from the martingale eventually stops with probability 1. From this stopping function we can define a new discrete-time martingale $X' = (X_1', X_2', X_3', \ldots)$ which is a variation on X, defined by:

$$x_n' = \begin{cases} x_{n-1}' & \text{if } f(x_1', x_2', \ldots, x_{k-1}') = \text{True for some } k \leq n-1 \\ X_n \mid x_{n-1}', x_{n-2}', \ldots, x_1' & \text{if } f(x_1', x_2', \ldots, x_{k-1}') = \text{False for all } 1 \leq k \leq n-1 \end{cases}$$

In other words, as soon as the sequence $x_1', x_2', x_3', \ldots$ reaches a stopping state (an output of *True*)), every term thereafter is the same as the previous, so for example if x_1' and x_1', x_2' are not stopping states but x_1', x_2', x_3' is, then the sample would look like: $x_1', x_2', x_3', x_3', x_3', \ldots$.

Problem 1.11 (2 points). Prove that X' is a martingale.

Let us also define another random variable S depending on the martingale X: $S = |x_1| + |x_2 - x_1| + |x_3 - x_2| + \cdots + |x_k - x_{k-1}|$, where k is defined from before as the "stopping time" of the sample $(x_1, x_2, x_3, \dots)$ from X.

Problem 1.12 (2 points). Show that for every n, $|X'_n| \leq S$.

Problem 1.13 (4 points). Show that if at least one of the three conditions specified in the Optional Stopping Theorem is met, then $E[S]$ is finite.

Theorem 1.12 (Lebesgue's Dominated Convergence Theorem)
If a sequence $X_1, X_2, X_3, \dots$ of variables depending on a random event always converges to another variable Y also depending on the event regardless of the event's outcome, and $\{X_1, X_2, \dots\}$ are all bounded by a random variable with finite expected value, then $\lim_{n\to\infty} E[X_n] = E[Y]$. In other words, the limit of the expected values of the X_i's is equal to the expected value of their limit.

In this case, we can take the "event" to be the sample from the martingale X. Each sample defines with probability 1 the outcome of the variable x_k, and also those of $X'_1, X'_2, \dots, S$. By definition, the sequence $X'_1, X'_2, \dots$ converges to x_k regardless of which sample from X was picked, and by Problem 1.12, all $|X'_n|$ are bounded by S. So the Dominated Convergence Theorem gives that $\lim_{n\to\infty} E[X'_n] = E[x_k]$. Since X' is a martingale, by Problem 1.10 $\lim_{n\to\infty} E[X'_n] = \lim_{n\to\infty} E[x_1] = E[x_1] = E[x_k]$. □

Problem 1.14 (3 points). Find a counterexample of a martingale, along with a stopping function for which the value of k exists with probability 1, which does not satisfy $E[x_k] = E[x_1]$.

Problem 1.15 (3 points). Suppose you are walking on a number line with endpoints 0 and 100, and you are currently at 67. Every second, you flip a coin and walk one step forward if it is heads, and one step backward if it is tails, until you reach either 100 or 0. What is the probability that you reach 100 before 0? (Hint: Find a martingale.)

Problem 1.16 (3 points). Consider the same situation as in the previous problem. What is the expected number of seconds it will take you to reach one of the endpoints, either 100 or 0? Use a martingale. (Hint: Use the Optional Stopping Theorem.)

Problem 1.17 (3 points). Consider a slight modification to the situation: you are walking on a number line between 0 and 100 and are currently at 67, except that this time the coin is unfair; the chance of heads is h and the chance of tails is $1 - h$, for some value of h between 0 and 1. What is the probability that you reach 100 before 0, in terms of h?

Problem 1.18 (4 points). This time, you are walking on a number line between 0 and 1 and are currently at 0.6, but the fairness of the coin depends on where you are; every second, if you are located at coordinate x then the probability that you will get a head and move 0.05 units forward is equal to x, while you get a tail and move 0.05 units backward with probability $1 - x$. What is the probability that you reach 1 before 0?

1.5 Variance

Variance

The **variance** $\text{Var}(X)$ of a random variable X, often denoted by σ^2, is the expected value of the squared distance from the mean of the distribution. We can say $\text{Var}(X) = E[(X - E[X])^2]$.

Remark 1.13 (Alternate definition of Variance). By Linearity of Expectation, $\text{Var}(X) = E[X^2] - E[X]^2$.

Problem 1.19 (2 points). Compute the variance of one die roll.

Problem 1.20 (2 points). Compute the variance of a random real number sampled uniformly from the interval $[a, b] \subset \mathbb{R}$.

Theorem 1.14 (Independent Variances Add)

The variance of the sum of two independent random variables is the sum of their variances: $\text{Var}(x + y) = \text{Var}(x) + \text{Var}(y)$.

Problem 1.21 (2 points). Prove Theorem 1.14.

Problem 1.22 (2 points). Compute the variance of the sum of n die rolls.

Theorem 1.15 (Variance Scales Quadratically)

When a probability distribution is scaled, i.e. every outcome is multiplied by a scale factor λ, the variance scales by a factor of λ^2.

Problem 1.23 (2 points). Prove Theorem 1.15.

Standard Deviation

The **standard deviation** σ of a random variable is the square root of its variance.

Each probability distribution has a standard deviation σ. We say that a sample X from the distribution is "within n standard deviations from the mean" if $|X - E[X]| \leq n\sigma$. We say that X is "$\frac{|X-E[X]|}{\sigma}$ standard deviations away from the mean".

Problem 1.24 (2 points). How many standard deviations away from the mean are each of the five elements in the set $\{0, 3, 4, 6, 12\}$?

Problem 1.25 (3 points). (Chebyshev's Inequality) Show that for any data set of real numbers with standard deviation > 0, and any positive integer n, at most $\frac{1}{n^2}$ of the data lies at least n standard deviations away from the mean.

Problem 1.26 (2 points). Use the previous problem to obtain a 75% confidence interval for the sum of 1000 die rolls. That is, an interval which contains 75% of the sums obtained when sets of 1000 dice are rolled and summed.

1.6 Recursion

When we have an infinite iterated processes such as a sequence of random variables, calculating a property of the process can sometimes be reduced down to a later iteration of itself, which is still the same process. From this, an equation can be written which describes the infinite iterated process in a succinct, finite way, and which can be solved.

For example, to calculate the value of $\sqrt{2+\sqrt{2+\sqrt{2+\sqrt{\dots}}}}$, we can set it equal to a variable S, and then recognize that the part of the expression ignoring the leftmost $2+$ is identical to S (this is the recursive step), which then yields a finite equation in S: $S = \sqrt{2+S}$, which we can then solve to obtain $S = 2$.

The following is an example of using recursion in an infinite process of sampling random variables:

Example 1.16
You are playing a game in which you start with \$1, and flip a coin repeatedly. Every head you get increases your money by 50%, but as soon as you get a tail, you must stop. What is your expected final amount of money?

We can solve this with recursion. If E is the expected value of the amount of money that you leave with in the end, then if you start by flipping a head, it is as if you have started playing the game all over again, but with an initial \$1.50 instead of \$1. So, the expected value of the amount of money you leave with assuming this initial head is $\frac{3}{2}H$. There is a 50% chance of the first flip being a head, so your total expected final amount of money is

$$E = \frac{1}{2}\left(\frac{3}{2}E\right) + \frac{1}{2}(1).$$

Solving this equation, we get $E = \boxed{\$2}$.

Problem 1.27 (2 points). If instead of 50% every head increases your money by 25%, what is your new expected final amount of money? (You still start the game with \$1.)

Problem 1.28 (2 points). If instead of 50% every head increases your money by 100%, creating an equation like in the previous example yields $E = E + \frac{1}{2}$, or $0 = \frac{1}{2}$, which is clearly false. What happened?

Problem 1.29 (2 points). If instead of 50% every head increases your money by 200%, creating an equation like in the previous example yields $E = 2E + \frac{1}{2}$, which implies that E is negative. But never in the game do you lose money, so the expected value of your final amount of money must be positive. What happened?

Problem 1.30 (5 points). A real number is randomly chosen between 0 and 1, and written in base 3 as a string of 0's, 1's, and 2's: $0.\overline{xyz\ldots}_3 = x \cdot \frac{1}{3} + y \cdot \frac{1}{9} + z \cdot \frac{1}{27} + \ldots$. It is then interpreted as a decimal in base 9: $0.\overline{xyz\ldots}_9 = x \cdot \frac{1}{9} + y \cdot \frac{1}{81} + z \cdot \frac{1}{729} + \ldots$. What is the expected value of the square of this base-9 decimal? (Hint: Use variance.)

2 Convergence and the Central Limit Theorem

When many independent random variables are added together, the probability distribution of their sum often looks quite nice – it concentrates around its mean value, and sometimes even approaches a Normal Distribution.

2.1 A Concentration Inequality

We can find an exponentially decreasing upper bound for the tail of a distribution that is the sum of many independent variables. Intuitively, the sum must be sufficiently concentrated around its mean value.

Let X_1, X_2, X_3, ... be independent random variables, not necessarily from the same probability distribution, such that $X_i \in [0,1]$ for each i. Let S_n be the raw sum $X_1 + X_2 + \cdots + X_n$ of the first n variables, $\mu_n = E[S_n]$, and $t > 0$ be a real number. We wish to bound $P(S_n \geq \mu_n + t)$ from above.

Let $\lambda > 0$ be a real number. Then $P(S_n \geq \mu_n + t) = P(e^{\lambda S_n} \geq e^{\lambda(\mu_n + t)})$.

Problem 2.1 (2 points). Show that $P(e^{\lambda S_n} \geq e^{\lambda(\mu_n+t)}) \leq e^{-\lambda(\mu_n+t)} \left(E[e^{\lambda X_1}]E[e^{\lambda X_2}] \cdots E[e^{\lambda X_n}]\right)$.

We wish to bound each of the individual $E[e^{\lambda X_i}]$ terms. Here we will use the fact that each X_i is contained in the interval $[0,1]$.

Problem 2.2 (3 points). Prove that $\frac{e^{\lambda x}-1}{\lambda x} \leq \frac{e^{\lambda}-1}{\lambda}$ for all $x \in [0,1]$.

Thus, $e^{\lambda X_i} \leq 1 + X_i(e^{\lambda} - 1)$ for each variable X_i, so $E[e^{\lambda X_i}] \leq 1 + E[X_i](e^{\lambda} - 1)$, and

$$e^{-\lambda(\mu_n+t)} \left(E[e^{\lambda X_1}]E[e^{\lambda X_2}] \cdots E[e^{\lambda X_n}]\right) \leq e^{-\lambda(\mu_n+t)} \prod_{i=1}^{n} (1 + E[X_i](e^{\lambda} - 1)).$$

Problem 2.3 (3 points). Show that $\prod_{i=1}^{n} (1 + E[X_i](e^{\lambda} - 1)) \leq \left(1 + \frac{e^{\lambda}-1}{n}\mu_n\right)^n$.

Thus, we obtain the bound $P(S_n \geq \mu_n + t) \leq e^{-\lambda(\mu_n+t)} \left(1 + \frac{e^{\lambda}-1}{n}\mu_n\right)^n$, for any choices of $t, \lambda > 0$. Since we can choose λ to be whatever we want, we can let $e^{\lambda} = \frac{\mu_n + t}{\mu_n}$ to make things cancel nicely. Then we get

$$P(S_n \geq \mu_n + t) \leq \left(\frac{\mu_n}{\mu_n + t}\right)^{\mu_n + t} \left(1 + \frac{t}{n}\right)^n.$$

Because of the Taylor Expansion of e^t, $\left(1 + \frac{t}{n}\right)^n$ approaches e^t from the left as $n \to \infty$, so

$$\boxed{P(S_n \geq \mu_n + t) \leq \left(\frac{\mu_n}{\mu_n + t}\right)^{\mu_n + t} e^t}.$$

This is called **Chernoff's Inequality**.

> *Remark 2.1.* Using the fact that $x - (1+x)\ln(1+x) \leq \frac{x^2}{2(1+x/3)}$ for all $x > 0$, this bound can be revised to $P(S_n \geq \mu_n + t) \leq e^{-\frac{t^2}{2(\mu_n + t/3)}}$.

2.2 Moments

The **moments** $\{\mu_0, \mu_1, \mu_2, \dots\}$ of a probability distribution are quantitative values used to describe its shape. If x is the random variable on the probability distribution, then the nth moment is related to the expected value of x^n. The first moment is *expected value*, the second is *variance*, the third is *skew*, and the fourth is *kurtosis*.

Moments

The nth *raw* moment of a random variable x which takes on values in the set X is:

$$\mu_n' = E[x^n] = \begin{cases} \sum_{x \in X} x^n P(x) & \text{if } x \text{ is discrete with probability function } P \\ \int_{-\infty}^{\infty} x^n f(x)\, dx & \text{if } x \text{ is continuous with probability density function } f \end{cases}$$

The nth moment *about* c for a real number c is:

$$\mu_n = E[(x-c)^n] = \begin{cases} \sum_{x \in X} (x-c)^n P(x) & \text{if } x \text{ is discrete with probability function } P \\ \int_{-\infty}^{\infty} (x-c)^n f(x)\, dx & \text{if } x \text{ is continuous with probability density function } f \end{cases}$$

Note that the raw moments are the moments *about* 0. The second and higher moments will by default be about the mean c of the distribution, which are called the *central moments*.

We will by default be working with the continuous case.

Remark 2.2 (Nonexistent Moments). If $\sum_{x \in X} |x^n| P(x) = \infty$ or $\int_{-\infty}^{\infty} |x^n| f(x)\, dx = \infty$, then the moment μ_n is said to *not exist*.

Problem 2.4 (4 points). Prove that if the nth moment exists, then so do all lower moments (0th through $(n-1)$th).

Problem 2.5 (3 points). Prove that the moments about a real number b can be calculated from the moments about a real number a in the following way:

$$E[(x-b)^n] = \sum_{i=0}^{n} \binom{n}{i} E[(x-a)^i](a-b)^{n-i}$$

Normalized Moments

The **normalized** nth central moment, $\tilde{\mu}_n$, is the nth central moment divided by the nth power of the standard deviation σ:

$$\tilde{\mu}_n = \frac{\mu_n}{\sigma^n} = \frac{E[(x-\mu_1)^n]}{\sigma^n}.$$

Each normalized moment is a dimensionless ratio that is scale invariant; if the probability distribution is scaled by a factor then the moments will change, but the *normalized* moments will remain the same.

2.2.1 Uniqueness

When do the moments uniquely determine the distribution? (Unique up to equivalence class, that is.)

Equivalent Distributions

We define two probability distributions X_1 and X_2 to be **equivalent** if and only if their *cumulative distribution functions*, $P(X_1 \leq x)$ and $P(X_2 \leq x)$ for all $x \in \mathbb{R}$, are equal.

Theorem 2.3 (Bounded Interval Uniqueness)
Every probability distribution on a **bounded** interval is uniquely determined by its moments.

Theorem 2.4 (Carleman's Condition for Uniqueness)
If

$$\sum_{i=1}^{\infty} \frac{1}{\mu_{2i}^{1/2i}} = \infty,$$

then the set of all moments $\{\mu_i\}$ uniquely determines the distribution.

2.3 Normal Distributions

Normal Distribution
The **normal distribution**, or **Gaussian**, with mean μ and variance σ^2 is defined by the probability density function

$$f(x) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Normal distributions are also sometimes casually referred to as *bell curves*.

We also have a notion of summing two distributions, however it is more complicated than just summing their probability density functions.

Convolution of Distributions
The sum, or **convolution**, of two **independent random variables** is the probability density function of the sum of those two variables. If independent random variables x and y have probability density functions f_x and f_y, then the probability density function f_z of $z = x + y$ is the convolution of f_x and f_y:

$$f_z(z) = \begin{cases} \sum_{x \in X} P(x = x) P(y = z - x) & \text{for the discrete case} \\ \int_{-\infty}^{\infty} f_x(x) f_y(z - x)\, dx & \text{for the continuous case} \end{cases}$$

Theorem 2.5 (Convolution of Normal Distributions)
The convolution of any two normal distributions is normal.

Problem 2.6 (4 points). Prove Theorem 2.5: Let X be a random sample from $\mathcal{N}(\mu_1, \sigma_1^2)$ and Y a random sample from $\mathcal{N}(\mu_2, \sigma_2^2)$. Show that the probability distribution of $X + Y$ is $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

2.4 Central Limit Theorem

The **Central Limit Theorem (CLT)** explains why the normal distribution appears so frequently in probability and statistics. Roughly speaking, when many independent random variables are added together, the resulting sum, after being appropriately *centered* and scaled, has a distribution that is close to normal. This phenomenon occurs even when the original random variables are not normally distributed.

Independent and Identically Distributed
We say that a set of random variables is **independent and identically distributed** (sometimes abbreviated i.i.d.) if each variable's value does not affect any of the other variables' values (independent) and all the variables are samples from the same probability distribution (identically distributed).

Standardized Sample Mean
The **standardized sample mean** of independent and identically distributed random variables X_1, X_2, $\ldots$, X_n from a distribution with mean μ and variance σ^2, is $\frac{\sqrt{n}}{\sigma}\left(\frac{X_1+X_2+\cdots+X_n}{n} - \mu\right)$.

Theorem 2.6 (Central Limit Theorem (CLT))
Suppose that $X_1, X_2, X_3, \ldots$ is an infinite sequence of independent and identically distributed variables from a probability distribution with mean μ and variance σ^2. Then the probability distribution of the sum of the first n variables approaches a normal distribution as $n \to \infty$. More specifically, the probability distribution of the nth standardized sample mean S_n converges to $\mathcal{N}(0,1)$ as $n \to \infty$.

2.4.1 Proving CLT Using Moments

We will first prove CLT using Moment Generating Functions (MGF)'s. Moment Generating Functions are a way of condensing all the moments of a distribution into one formula, and succinctly describe the distribution.

Theorem 2.7 (Taylor Series)
For an infinitely differentiable real-valued function f, its value at $x \in \mathbb{R}$ can be approximated by its Taylor series evaluated at the point a for any $a \in \mathbb{R}$:

$$f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n.$$

Remark 2.8. For many functions, the Taylor series evaluated at a only converges to $f(x)$ when $|a-x|$ is sufficiently small. However, some functions such as e^x equal their Taylor series evaluated at any point.

Moment Generating Function (MGF)
Let X be a random variable. Then X defines a function M_X called its **moment generating function (MGF)**, which is defined on all real t by $M_X(t) = E[e^{tX}]$. The Taylor Series expansion of e^{tX} about 0 is

$$1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \ldots$$

so by Linearity of Expectation,

$$M_X(t) = 1 + tE[X] + \frac{t^2E[X^2]}{2!} + \frac{t^3E[X^3]}{3!} + \cdots = 1 + t\mu_1 + \frac{t^2\mu_2}{2!} + \frac{t^3\mu_3}{3!} + \ldots.$$

Thus, all the moments of X can be obtained from the function $M_X(t)$. In fact, any moment μ_k is equal to the value of the k'th derivative of $M_X(t)$ at $t = 0$.

Problem 2.7 (3 points). Compute the moment generating function of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ to be

$$M_{\mathcal{N}(\mu,\sigma^2)}(t) = e^{t\mu + \frac{\sigma^2 t^2}{2}}.$$

Moment generating functions help us access the moments of a probability distribution, which is useful because two probability distributions can sometimes be proved identical if they have the same moments.

Let $X_1, X_2, X_3, \ldots$ be a sequence of independent samples from a probability distribution with mean μ and variance σ^2. First, we will normalize them by defining the sequence of variables $Y_1, Y_2, Y_3, \ldots$ such that $Y_i = \frac{X_i - \mu}{\sigma}$ for all positive integers i.

Problem 2.8 (2 points). Show that each Y_i has mean 0 and variance 1.

We will also create a normalized version of the sum, by defining the n'th **standardized sample mean** as $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$. Since variance scales quadratically, the variable S_n also has mean 0 and variance 1. Our

goal is to show that the probability distribution of S_n is exactly the same as $\mathcal{N}(0,1)$. We will explicitly find the limit of each moment of S_n as $n \to \infty$.

For each positive integer k, the kth moment is $E[S_n^k] = E\left[\left(\frac{1}{\sqrt{n}}\right)^k (Y_1 + Y_2 + \cdots + Y_n)^k\right]$. By Linearity of Expectation, we are able to expand the product to obtain the sum of many terms of the form $E\left[\left(\frac{1}{\sqrt{n}}\right)^k Y_1^{k_1} Y_2^{k_2} \cdots Y_n^{k_n}\right]$ with $k_1 + k_2 + \cdots + k_n = k$. **For each such n-tuple $(k_1, k_2, \ldots, k_n)$, let $m_{(k_1,k_2,\ldots,k_n)}$ be the multiplicity of $E\left[\left(\frac{1}{\sqrt{n}}\right)^k Y_1^{k_1} Y_2^{k_2} \cdots Y_n^{k_n}\right]$ in this sum. So all the multiplicities should add to n^k in total.**

Problem 2.9 (9 points). We will do casework on the possible n-tuples of exponents $(k_1, k_2, \ldots, k_n)$. Show the following:

a) (2 points) The sum of $\left(m_{(k_1,k_2,\ldots,k_n)} E\left[\left(\frac{1}{\sqrt{n}}\right)^k Y_1^{k_1} Y_2^{k_2} \cdots Y_n^{k_n}\right]\right)$ over **all** n-tuples $(k_1, k_2, \ldots, k_n)$ of nonnegative integers summing to k **such that at least one of the k_i's is equal to** 1, is 0.

b) (4 points) The sum of $\left(m_{(k_1,k_2,\ldots,k_n)} E\left[Y_1^{k_1} Y_2^{k_2} \cdots Y_n^{k_n}\right]\right)$ over **all** n-tuples $(k_1, k_2, \ldots, k_n)$ of nonnegative integers summing to k **with no k_i equal to** 1**, but with at least one** $k_i \geq 3$, is less than or equal to $n^{\frac{k-1}{2}}$ times a constant depending on k, and thus the sum of $E\left[\left(\frac{1}{\sqrt{n}}\right)^k Y_1^{k_1} Y_2^{k_2} \cdots Y_n^{k_n}\right]$ over those n-tuples approaches 0 as $n \to \infty$.

c) (3 points) The sum of $\left(m_{(k_1,k_2,\ldots,k_n)} E\left[\left(\frac{1}{\sqrt{n}}\right)^k Y_1^{k_1} Y_2^{k_2} \cdots Y_n^{k_n}\right]\right)$ over **all** n-tuples $(k_1, k_2, \ldots, k_n)$ of nonnegative integers summing to k **such that each k_i is either** 0 **or** 2 (if any such n-tuples exist) is 0 if k is odd, and approaches $(k-1)!! = (k-1)(k-3)\cdots(3)(1)$ as $n \to \infty$ if k is even.

Summing all the expected values, we get that $E[S_n^k] \to (k-1)!!$ as $n \to \infty$ if k is even, and $E[S_n^k] \to 0$ if k is odd. Thus, the moments of S_n are: $\mu_1 = 0$, $\mu_2 = 1$, $\mu_3 = 0$, $\mu_4 = 3$, $\mu_5 = 0$, $\mu_6 = 15$, ...

Problem 2.10 (3 points). Prove that the moments of S_n as $n \to \infty$ are identical to the moments of $\mathcal{N}(0,1)$.

Problem 2.11 (3 points). Prove that if a probability distribution has the same moments as $\mathcal{N}(0,1)$, then it is equivalent to $\mathcal{N}(0,1)$.

Thus, the normalized sum S_n of $X_1, X_2, \ldots, X_n$ approaches the standard normal distribution as n grows large, which means that the sum of many samples from a given probability distribution approaches a normal distribution as the number of samples grows to infinity.

Another way to prove CLT is by using the **Lagrange Error Bound**.

2.4.2 Proving CLT Using the Lagrange Error Bound

Lagrange Error Bound

A $(k+1)$-times differentiable function $f(x)$ can be approximated by its Taylor expansion about a as

$$P_k(x) = f(a) + f'(a)(x-a) + \frac{f''(a)(x-a)^2}{2!} + \cdots + \frac{f^{(k)}(a)(x-a)^k}{k!}.$$

The "error" of this approximation is defined to be $R_k(x) = f(x) - P_k(x)$, and if $f^{(k)}$ is continuous on the interval between a and x, then we have that

$$R_k(x) = \frac{f^{(k+1)}(z)(x-a)^{k+1}}{(k+1)!}$$

for some real number z between a and x.

As before, let $X_1, X_2, X_3, \ldots$ be independent and identically distributed variables from a probability distribution X with $E[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Define $Y_i = \frac{X_i - \mu}{\sigma}$ to be the normalized versions of X_i, so that $E[Y_i] = 0$ and $\mathrm{Var}(Y_i) = 1$, and define the standardized sample mean $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$.

Problem 2.12 (2 points). Show that the moment generating function $M_{S_n}(t)$ of S_n is equal to $\left(E[e^{tY_1/\sqrt{n}}]\right)^n$.

Problem 2.13 (2 points). Let $e^x = 1 + x + \frac{x^2}{2} + R(x)$. Use the Lagrange Error Bound to find a bound on $R(x)$.

Problem 2.14 (2 points). Apply the previous problem with $x = \frac{tY_1}{\sqrt{n}}$ to show that $E[e^{tY_1/\sqrt{n}}] = 1 + \frac{t^2}{2n} + R(t, Y_1, n)$, where $|R(t, Y_1, n)|$ is bounded by $\frac{1}{n^{3/2}}$ times a constant not depending on n.

Problem 2.15 (5 points). Show that

$$\lim_{n\to\infty} \left(1 + \frac{t^2}{2n} + R(t, Y_1, n)\right)^n = e^{t^2/2}.$$

The conclusion can then be drawn in the same way as in the previous proof: the moments of the probability distribution of S_n approach those of $\mathcal{N}(0,1)$, so S_n must converge to $\mathcal{N}(0,1)$.

2.4.3 Finding a Rate of Convergence

Finally, we can prove CLT with a completely different method that does not use moments, but rather the fact that if we have many independent and identically distributed random variables, then replacing one of them with a normal distribution doesn't change the overall mean by very much. The effect of this small increment can then be bounded using calculus. This proof also allows us to explicitly calculate the rate of convergence to the normal distribution **assuming that we can control** $E[|X|^3]$.

Consider independent and identically distributed random variables $X_1, X_2, X_3, \ldots$ from a distribution X which is already normalized (so $E[X] = 0$ and $\mathrm{Var}[X] = 1$), **and such that** $E[|X|^3] \le Q$ **for some finite real number** Q. The nth standardized sample mean is $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$; we wish to show that S_n converges to $\mathcal{N}(0,1)$.

Indicator Function

Given two real numbers x and r, we will define the **indicator function** $\text{Ind}_r(x)$ to equal 1 if $x \geq r$, and 0 otherwise. Given a probability distribution X and any real number r, $E[\text{Ind}_r(X)]$ represents **the probability that a random sample from X is $\geq r$.**

Remark 2.9. The functions $E[\text{Ind}_p(X)] = P(X \geq p)$ correspond to the Cumulative Distribution Function (CDF) at each real number p, another function that describes the distribution as an alternative to the PDF.

So, another way to show that S_n converges to $\mathcal{N}(0,1)$ is by showing that their indicator functions converge to each other. However, as the indicator function is not differentiable, in order to do calculus we will need to approximate it by differentiable functions.

Let f be a thrice-differentiable function (1st, 2nd, and 3rd derivatives exist) with a bounded third derivative: $|f^{(3)}(x)| \leq C$. As going from S_n to $\mathcal{N}(0,1)$ is possibly a big leap, we will make a smaller jump by comparing the expected value $E[f(S_n)]$ to $E\left[f\left(S_n - \frac{X_n}{\sqrt{n}} + \frac{G_n}{\sqrt{n}}\right)\right]$, where G_n ("G" for Gaussian) is a random sample taken from $\mathcal{N}(0,1)$. Intuitively, we are taking a baby step toward the normal distribution by changing one of the n random variables in the mean to a Gaussian (normally distributed variable). Through many baby steps, we can eventually reach a full normal distribution.

Problem 2.16 (4 points). Use the Lagrange Error Bound to show that

$$E[f(S_n)] = E\left[f\left(T_{n-1} + \frac{X_n}{\sqrt{n}}\right)\right] = E[f(T_{n-1})] + \frac{E[f''(T_{n-1})]}{2n} + Error, \quad |Error| \leq \frac{E[|X_n|^3]}{6n^{3/2}}C$$

where we define T_{n-1} as $\frac{1}{\sqrt{n}}\sum_{i=1}^{n-1} X_i$.

By the same logic,

$$E\left[f\left(T_{n-1} + \frac{G_n}{\sqrt{n}}\right)\right] = E[f(T_{n-1})] + \frac{E[f''(T_{n-1})]}{2n} + Error, \quad |Error| \leq \frac{E[|G_n|^3]}{6n^{3/2}}C.$$

Problem 2.17 (3 points). Compute $E[|G_n|^3]$ to be $\sqrt{\frac{8}{\pi}}$.

Since the first parts of the expressions for $E\left[f(S_n)\right]$ and $E\left[f\left(T_{n-1} + \frac{G_n}{\sqrt{n}}\right)\right]$ are identical, their difference is bounded by the sum of their errors, which is $\frac{E[|X_n|^3]+\sqrt{8/\pi}}{6n^{3/2}}C$. Now that we have quantified the effect of replacing one of the random variables with a Gaussian, we can iterate the process on each of the random variables.

Problem 2.18 (3 points). Iterate on previous results to show that

$$\left|E\left[f\left(S_n\right)\right] - E\left[f\left(\frac{G_1}{\sqrt{n}} + \frac{G_2}{\sqrt{n}} + \cdots + \frac{G_n}{\sqrt{n}}\right)\right]\right| \leq n \cdot \frac{E[|X_n|^3] + \sqrt{8/\pi}}{6n^{3/2}}C,$$

where $G_1, G_2, \ldots, G_n$ are all random samples taken from $\mathcal{N}(0,1)$.

Problem 2.19 (2 points). Show that the probability distribution of $\frac{G_1}{\sqrt{n}} + \frac{G_2}{\sqrt{n}} + \cdots + \frac{G_n}{\sqrt{n}}$ is $\mathcal{N}(0,1)$.

Hence, we can replace $\frac{G_1}{\sqrt{n}} + \frac{G_2}{\sqrt{n}} + \cdots + \frac{G_n}{\sqrt{n}}$ with a single sample from $\mathcal{N}(0,1)$, which we will call G. Then $|E[f(S_n)] - E[f(G)]| \leq \frac{E[|X_n|^3]+\sqrt{8/\pi}}{6n^{1/2}}C$. We can now relate S_n to G, with the caveat that we have to apply to both of them a function which is thrice differentiable and has bounded third derivative. We would ideally like to apply the indicator function to both of them, so instead we may formulate a function which behaves similarly to the indicator but meets the desired criteria.

We can approximate the indicator function $\text{Ind}_r(x)$ with a "smoother" function $f_{r,\epsilon}(x)$, where $\epsilon > 0$ describes how close of an approximation it is to the indicator (the smaller the value of ϵ, the closer the approximation). We define $f_{r,\epsilon}$ for real numbers r and $\epsilon > 0$ as follows:

$$f_{r,\epsilon}(x) = \begin{cases} 0 & x \leq r-\epsilon \\ \frac{\int_{-1}^{(x-r)/\epsilon}(t+1)^3(t-1)^3\,dt}{\int_{-1}^{1}(t+1)^3(t-1)^3\,dt} & r-\epsilon \leq x \leq r+\epsilon \\ 1 & x \geq r+\epsilon. \end{cases}$$

Problem 2.20 (6 points). Verify the conditions for the function $f_{r,\epsilon}(x)$:

a) (3 points) It is thrice differentiable (1st, 2nd, and 3rd derivatives exist).

b) (3 points) The absolute value of its third derivative is bounded by a fixed value, $\frac{105}{16\epsilon^3}$.

Thus, $f_{r,\epsilon}(x)$ meets the desired criteria. We wish to compare the indicator functions of S_n and G, and $f_{r,\epsilon}$ is now our way of going between those two random variables. So, we bound the indicator function with $S_{r,\epsilon}$'s.

Problem 2.21 (3 points). Show that

$$E[\text{Ind}_p(S_n)] \in \Big[E[f_{p+\epsilon,\epsilon}(S_n)], E[f_{p-\epsilon,\epsilon}(S_n)]\Big]$$

and

$$E[\text{Ind}_p(G)] \in \Big[E[f_{p+\epsilon,\epsilon}(G)], E[f_{p-\epsilon,\epsilon}(G)]\Big].$$

We already showed that these two intervals are quite similar; corresponding endpoints differ by at most $\frac{E[|X_n|^3]+\sqrt{8/\pi}}{6n^{1/2}} \cdot \frac{105}{16\epsilon^3}$. However, we do not yet know how wide the intervals are; the wider they are, the more room for $E[\text{Ind}_p(S_n)]$ and $E[\text{Ind}_p(G)]$ to differ. However, we can take advantage of the fact that we know everything about the probability distribution of G.

Problem 2.22 (5 points). The width of the interval $\Big[E[f_{p+\epsilon,\epsilon}(G)], E[f_{p-\epsilon,\epsilon}(G)]\Big]$ can be bounded in the following way:

a) (2 points) First, prove that $\Big[E[f_{p+\epsilon,\epsilon}(G)], E[f_{p-\epsilon,\epsilon}(G)]\Big] \subseteq \Big[E[\text{Ind}_{p+2\epsilon}(G)], E[\text{Ind}_{p-2\epsilon}(G)]\Big]$.

b) (3 points) Use knowledge about the normal distribution to prove that

$$E[\text{Ind}_{p-2\epsilon}(G)] - E[\text{Ind}_{p+2\epsilon}(G)] \leq \frac{4\epsilon}{\sqrt{2\pi}}.$$

From this, we know that $E[\text{Ind}_p(G)]$ is contained in an interval at most $\frac{4}{n^{1/8}\sqrt{2\pi}}$ wide and with endpoints differing by at most $\frac{E[|X_n|^3]+\sqrt{8/\pi}}{6n^{1/2}} \cdot \frac{105}{16\epsilon^3}$ from the endpoints of another interval that contains $E[\text{Ind}_p(S_n)]$. Thus,

$$\Big|E[\text{Ind}_p(S_n)] - E[\text{Ind}_p(G)]\Big| \leq \frac{4\epsilon}{\sqrt{2\pi}} + \frac{E[|X_n|^3]+\sqrt{8/\pi}}{6n^{1/2}} \cdot \frac{105}{16\epsilon^3}.$$

Problem 2.23 (3 points). Find a constant depending only on the probability distribution X (not depending on n) such that the absolute value of the right-hand side of the inequality is always at least this constant times $n^{-1/8}$.

We can obtain a convergence rate of $n^{-1/8}$ by setting ϵ equal to $n^{-1/8}$ for each n; then, the right-hand side evaluates to $n^{-1/8}$ times a constant in terms of $E[|X|^3]$. As $n \to \infty$, $n^{-1/8} \to 0$, so the indicator functions, at every real number p, of S_n and G approach each other at a fixed rate. Thus, the standardized sample means converge to $\mathcal{N}(0,1)$. Note that by the previous problem, $n^{-1/8}$ is the fastest rate of convergence that this method gives us. The difference

$$\left|E[\text{Ind}_p(S_n)] - E[\text{Ind}_p(G)]\right| = \left|P(S_n \geq p) - P(G \geq p)\right|$$

is bounded by a constant times $n^{-1/8}$.

Remark 2.10. The Berry–Esseen theorem provides a slightly stronger result: $|P(S_n \geq p) - P(G \geq p)| \leq Cn^{-1/2}$, for some constant C depending on the probability distribution X.

2.4.4 Application of CLT

Theorem 2.11 (Law of Large Numbers (LLN))
In general, the Law of Large Numbers (LLN) states that the average of a large number of independent and identically distributed samples from a probability distribution converges to the mean of the probability distribution.

Given a sequence of independent and identically distributed variables $X_1, X_2, X_3, \ldots$ from a distribution with mean μ, let $\overline{X_n} = \frac{X_1+X_2+\cdots+X_n}{n}$ be the raw mean of the first n samples.

Weak LLN: For any $\epsilon > 0$, $\lim_{n\to\infty} P(|\overline{X_n} - \mu| < \epsilon) = 1$.

Strong LLN: $P(\lim_{n\to\infty} \overline{X_n} = \mu) = 1$.

Problem 2.24 (3 points). Prove the Weak LLN using CLT.

3 Borel-Cantelli Lemmas

In this section, we will be dealing with scenarios where we need to know whether or not infinitely many events happen **almost surely** (meaning with probability 1). The main tool to do so that we'll explore is the following:

Theorem 3.1 (Borel-Cantelli Lemmas)
Let $(A_1, A_2, \ldots)$ be a sequence of events.

1. If $\sum_{n=1}^{\infty} P(A_n) < \infty$ then
$$P(\text{infinitely many } A_n \text{ happen}) = 0.$$
2. If $\sum_{n=1}^{\infty} P(A_n) = \infty$ **and** $(A_1, A_2, \ldots)$ **are independent**, then
$$P(\text{infinitely many } A_n \text{ happen}) = 1.$$

Problem 3.1 (3 points). Prove the infinite monkey theorem: let S be a set and $X_0, X_1, \ldots$ be independent identically distributed variables taking values in S such that for all $s \in S$, $P(X_0 = s) > 0$. Prove that for all finite sequences $s_0, s_1, \ldots, s_k$ of elements of S,
$$P(\exists n : X_n = s_0, X_{n+1} = s_1, \ldots, X_{n+k} = s_k) = 1$$

Problem 3.2 (5 points). Prove that there exists a *simply normal number*: a real number $x \in [0,1]$ such that for any $d > 1$ and $a \in \{0, \ldots, d-1\}$ the digit a occurs in the base d representation of x with asymptotic frequency $1/d$:
$$\lim_{n\to\infty} \frac{\text{number of times } a \text{ occurs in the first } n \text{ digits of } x}{n} = \frac{1}{d}.$$

Problem 3.3 (5 points). Find a sequence of independent random variables $(X_1, X_2, \ldots)$ with $P(X_n \in \{-n, n, 0\}) = 1$ and $E[X_n] = 0$ for each n, and such that the weak Law of Large Numbers (LLN) holds but not the strong LLN: for $Y_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ it holds for any $\varepsilon > 0$ that $P(|Y_n| \geq \varepsilon) \to 0$, but $P(\lim_{n\to\infty} Y_n = 0) \neq 1$.

3.1 Markov Chains

Consider a set S, which is either finite or countably infinite (countably infinite meaning there is a sequence whose elements are exactly the set S). An infinite sequence of random variables $X_0, X_1, X_2 \ldots$ is said to be a **Markov Chain on S** if for any sequence of $x_0, x_1, \ldots, x_n \in S$, it holds that
$$P(X_n = x_n \mid X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) = P(X_n = x_n \mid X_{n-1} = x_{n-1}).$$

S is called the **state space** of the Markov Chain. Intuitively, a Markov Chain is a process which only remembers the immediate last state. The previous states are still potentially relevant: it is not true in general that $P(X_3 = x \mid X_0 = y) = P(X_3 = x)$, for example, but all the influence the value of X_0 has in X_3 is described by its influence on X_2.

For most cases we are interested in, the Markov chain will be **time homogeneous**, meaning that
$$P(X_n = y \mid X_{n-1} = x)$$

depends only on x and y, but not on n. For time homogeneous Markov Chains, we can describe them entirely by their **initial conditions** (the distribution of X_0) and the **transition matrix** P, the matrix with rows and columns corresponding to the elements of S whose entry in row x and column y is given by

$$P(x,y) = P(X_n = y \mid X_{n-1} = x).$$

Associated Graph
For a time homogeneous Markov Chain, we can associate a (directed, possibly with self-loops) graph on the state space S, where we connect $y \to x$ if $P(x,y) > 0$.

From now on, unless stated otherwise, assume Markov Chains are time homogeneous.

Problem 3.4 (2 points). Prove that for all $m, n \in \mathbb{Z}_{>0}$,

$$P(X_{m+n} = y \mid X_n = x) = P^m(x,y)$$

where P^m represents matrix multiplication. (For more on matrices, see Appendix 4.4.)

One of the most classical examples of Markov Chains is the **simple random walk**. The simple random walk on $\mathbb{Z}$ is given by $X_0 = 0$, with transition matrix

$$P(x,y) = \begin{cases} \frac{1}{2} & \text{if } x = y+1 \\ \frac{1}{2} & \text{if } x = y-1 \\ 0 & \text{otherwise} \end{cases}$$

More generally, for a graph G, the simple random walk on G starts at some vertex and has transition matrix

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

where $\deg(y)$ is the number of neighbors of y in the graph.

Irreducible Markov Chain
A Markov chain $X_0, X_1, \ldots$ on a state space S is said to be **irreducible** if for all ordered pairs $(x,y) \in S$ there is some $m \in \mathbb{Z}_{>0}$ such that $P^m(x,y) > 0$.

Problem 3.5 (2 points). Prove that a Markov Chain being irreducible is equivalent to its associated graph being **(strongly) connected**, meaning that there is an directed path (through arrows) from every vertex to any other vertex.

So if our state space is finite and our Markov Chain starts at a single point (meaning that there is some x_0 for which $P(X_0 = x_0) = 1$), we can often (but not always) reduce to the case of an irreducible one by deleting unreachable vertices of the associated graph. In fact, we can always delete vertices to obtain some irreducible Markov Chain (but possibly not with the same starting point) where entries of the transition matrix for the remaining vertices are the same. However, for Markov Chains on infinite state spaces, we have an interesting case where that's not true:

Example 3.2
Consider a Markov chain on $\mathbb{Z}$ where $X_0 = 0$ and

$$P(x,y) = \begin{cases} \frac{1}{2}, & \text{if } y = x+1 \\ \frac{1}{2}, & \text{if } y = x+2 \\ 0, & \text{otherwise} \end{cases}$$

It is reducible, but we can't get an irreducible Markov Chain by deleting vertices, since for any $x \in \mathbb{Z}$,

$P^m(x,x) = 0$ for all m.

Recurrent and Transient

Given a Markov Chain, for every $x \in S$, define its *hitting time* to be a variable defined by

$$T_x = \min\{n > 0 : X_n = x\}.$$

We say a Markov Chain with $P(X_0 = x_0) = 1$ is **recurrent** if $P(T_{x_0} \text{ is finite}) = 1$, and **transient** otherwise. This definition captures the idea of a Markov Chain which keeps returning to its starting point versus one which escapes off to infinity.

This intuition is formalized by the following:

Theorem 3.3 (Equivalent Conditions for Recurrence)

Consider an irreducible Markov Chain with $P(X_0 = x_0) = 1$. Then, the following are equivalent:

1. The Markov chain is recurrent.
2. For some $x \in X$ it holds that
$$P(X_n = x \text{ infinitely often}) = 1.$$
3. For all $x \in X$ it holds that
$$P(X_n = x \text{ infinitely often}) = 1.$$
4. For some $x \in X$ it holds that $\sum_m P^m(x,x) = \infty$.
5. For all $x \in X$ it holds that $\sum_m P^m(x,x) = \infty$.

Problem 3.6 (4 points). Prove that items 2, 3, 4, 5 from the above theorem are equivalent.

Problem 3.7 (2 points). Consider a Markov chain on $\mathbb{Z}_{>0}$ where $X_0 = 2026$ and

$$P(x,y) = \begin{cases} \frac{1}{2}, \text{ if } y = x+1 \\ \frac{1}{2}, \text{ if } y = 1 \\ 0, \text{ otherwise} \end{cases}$$

Prove that it is recurrent.

Problem 3.8 (7 points). The *independent random walk* on $\mathbb{Z}^d$ (the d-dimensional space of integers) is given by

$$P(x,y) = \begin{cases} 2^{-d} & \text{if } |x_i - y_i| = 1 \text{ for all } i \in \{1,\dots,d\} \\ 0 & \text{otherwise.} \end{cases}$$

Equivalently, the independent random walk on $\mathbb{Z}^d$ is a Markov process taking values on $\mathbb{Z}^d$ where in each coordinate, the process is independent of the other coordinates, and equal to a simple random walk. For which integers d is the independent random walk transient, and for which is it recurrent?

4 Appendix (Introduction to Calculus)

This is an introduction to calculus, containing many useful definitions and theorems that will be used for analyzing continuous probability distributions. **There are no problems to solve here – this is just useful information that will help you solve the other problems.**

4.1 Derivatives and Integrals

Limits

The limit of a function is a value that it approaches over time. Given a function $f(x)$, its left limit $\lim_{x\to c^-} f(x)$ equals to a real number $r \in \mathbb{R}$ if and only if for every $\epsilon > 0$, there exists a $\delta > 0$ such that for all $x \in [c-\delta, c]$, $f(x) \in (r-\epsilon, r+\epsilon)$. $f(x)$ *approaches* r as x approaches c from the left. Alternatively, if $f(x)$ approaches $\pm\infty$ as $x \to c^-$, then we say that $\lim_{x\to c^-} f(x) = \pm\infty$. Similarly, we can define the right limit $\lim_{x\to c^+} f(x)$. If the left and right limits both exist and are equal, then the **limit** $\lim_{x\to c} f(x)$ is their common value.

Continuous Function

A continuous function is one which you can draw without lifting your pen (there are no breaks, jumps, or holes).

The derivative of a continuous function measures its instantaneous rate of change.

Differentiation

For any continuous function $f(x)$, we can define its *left derivative* at the point x to be

$$\lim_{h\to 0^-} \frac{f(x+h)-f(x)}{h},$$

where $h \to 0^-$ means that h is a negative number approaching 0. Similarly, the *right derivative* is

$$\lim_{h\to 0^+} \frac{f(x+h)-f(x)}{h},$$

where $h \to 0^+$ means that h is a positive number approaching 0. If the left and right derivatives are equal and finite, then we say that $f(x)$ is **differentiable** at x, with **derivative** $\frac{d}{dx}(f(x)) = f'(x)$ equal to both derivatives. Here $\frac{d}{dx}$ means "derivative with respect to x". When it is clear what variable we are differentiating with respect to, the derivative may be abbreviated f'.

Remark 4.1. $f^{(n)}(x)$ is the n'th derivative of $f(x)$, if it is defined. $f(x)$ is n-times differentiable if one can take the derivative n times: $f(x), f'(x), \ldots, f^{(n-1)}(x)$ are all differentiable.

Some derivatives that will be useful to know for this power round are $\boxed{\frac{d}{dx}(x^n) = nx^{n-1}}$, and $\boxed{\frac{d}{dx}(e^x) = e^x}$.

Theorem 4.2 (Product Rule)
For functions $u(t)$ and $v(t)$, $\frac{d}{dt}(u(t)v(t)) = u(t)\frac{d}{dt}(v(t)) + v(t)\frac{d}{dt}(u(t))$.

Theorem 4.3 (Quotient Rule)
For a function $h(t) = \frac{f(t)}{g(t)}$, if $g(t) \neq 0$ then $h'(t) = \frac{f(t)g'(t) - f'(t)g(t)}{[g(t)]^2}$.

Theorem 4.4 (Chain Rule)
For functions $u(t)$ and $v(t)$, $\frac{d}{dt}(u(v(t))) = v'(t)u'(v(t))$.

Integration is the opposite of differentiation – instead of finding the function that resembles the rate of change, it is finding a new function that has a rate of change equal to the original function.

Integration

An integral is geometrically the area under a curve. Given a function $f(x)$, its **indefinite integral** $\int f(x)\,dx$ is a function which has derivative with respect to x equal to $f(x)$. For an interval $[a,b] \subseteq \mathbb{R}$, the **definite integral** $\int_a^b f(x)\,dx$ represents the sum of $f(x)$ over all x-values on $[a,b]$, or the signed area under the curve of $f(x)$ on this interval.

Theorem 4.5 (Fundamental Theorem of Calculus)

For a differentiable function $f(x)$,

$$\int_a^b f'(x)\,dx = f(x)\Big|_a^b = f(b) - f(a).$$

Some integrals that will be useful to know for this Power Round are $\boxed{\int x^n\,dx = \frac{x^{n+1}}{n+1} + C}$ for a constant C, and $\boxed{\int e^x\,dx = e^x + C}$ for a constant C. Integration by Parts is a way to compute integrals of complicated expressions that are the product of two things.

Theorem 4.6 (Integration by Parts)

The Product Rule states that $(uv)' = uv' + vu'$ (for functions u and v in terms of another variable t), so taking the integral of both sides we get $uv = \int uv' + \int vu'$, often written as $\boxed{\int u\,dv = uv - \int v\,du}$.

Theorem 4.7 (Fubini's Theorem)

Let $X, Y \subseteq \mathbb{R}$, and $f : X \times Y \to \mathbb{R}$ be an integrable function. Then the order of the integrals can be interchanged, so

$$\int_X \int_Y f(x,y)\,dy\,dx = \int_Y \int_X f(x,y)\,dx\,dy.$$

Theorem 4.8 (Factorization of Products)

Let $X, Y \subseteq \mathbb{R}$, and $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$ be integrable functions. Then

$$\int_X \int_Y f(x)g(y)\,dy\,dx = \left(\int_X f(x)\,dx\right)\left(\int_Y g(y)\,dy\right).$$

4.2 Inequalities

These are a few more tools which will be useful for bounding probabilities and probability distributions.

Theorem 4.9 (AM-GM Inequality)

For n nonnegative real numbers $x_1, x_2, \ldots, x_n$,

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \geq \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

with equality holding if and only if $x_1 = x_2 = \cdots = x_n$. The left side is the arithmetic mean (AM) and the right side geometric mean (GM).

Remark 4.10. In probability and statistics, the term "mean" will by default refer to arithmetic mean, not geometric mean.

Theorem 4.11 (Hölder's Inequality)
For sequences $\{x_k\}$ and $\{y_k\}$ of real variables, and real numbers p and q such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$\sum_{k=1}^{\infty} |x_k y_k| \leq \left(\sum_{k=1}^{\infty} |x_k|^p \right)^{\frac{1}{p}} \left(\sum_{k=1}^{\infty} |y_k|^q \right)^{\frac{1}{q}}.$$

Also, for real-valued functions f and g on an interval $S \subset \mathbb{R}$,

$$\int_S |f(x)g(x)|\, dx \leq \left(\int_S |f(x)|^p\, dx \right)^{\frac{1}{p}} \left(\int_S |g(x)|^q\, dx \right)^{\frac{1}{q}}.$$

We can bound $n!$ above and below by noting that

$$\ln n! = \sum_{k=1}^{n} 1 \cdot \ln k$$

(where ln is *natural log*, the inverse function of e^x) can be thought of as the area of several rectangles under the graph of $\ln x$. This yields the approximation

$$\ln n! \approx \int_1^n \ln x\, dx = n \ln n - n + 1,$$

i.e.

$$n! \approx e \left(\frac{n}{e} \right)^n$$

With other methods, we can refine this to Stirling's approximation:

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

where we say that $f(n) \sim g(n)$ if $\lim_{n\to\infty} f(n)/g(n) = 1$.

4.3 Convergence and Divergence

For a sequence of real numbers $a_1, a_2, \ldots$, we say that the series $\sum_{n=1}^{\infty} a_n$ converges to $S \in \mathbb{R}$ if

$$\lim_{N\to\infty} \sum_{n=1}^{N} a_n = S$$

and otherwise we say it diverges. There are several criteria for telling whether a series converges or diverges, and we'll list some here:

Theorem 4.12 (Test for divergence)
If the series $\sum_{n=1}^{\infty} a_n$ converges, then

$$\lim_{n\to\infty} a_n = 0$$

Theorem 4.13 (Comparison Test)
Let $(a_n)_{n\in\mathbb{Z}_{>0}}, (b_n)_{n\in\mathbb{Z}_{>0}}$ be sequences of nonnegative real numbers with $a_n \leq b_n$ for all n. If $\sum_{n=1}^{\infty} b_n$ converges, then $\sum_{n=1}^{\infty} a_n$ converges.

Theorem 4.14 (Limit Comparison Test)
Let $(a_n)_{n\in\mathbb{Z}_{>0}}, (b_n)_{n\in\mathbb{Z}_{>0}}$ be sequences of nonnegative real numbers with $b_n \neq 0$ for all n. If

$$\lim_{n\to\infty} \frac{a_n}{b_n}$$

exists, is finite and nonzero, then $\sum_{n=1}^{\infty} b_n$ converges if and only if $\sum_{n=1}^{\infty} a_n$ converges.

Theorem 4.15 (p-series)
The series

$$\sum_{n=1}^{\infty} \frac{1}{n^p}$$

converges if and only if $p > 1$.

We also say that the integral $\int_a^b f(x)\,dx$ *converges* if it equals a finite real number, and *diverges* otherwise.

Theorem 4.16 (Integral Comparison Test)
If $|f(x)| \leq |g(x)|$ on $[a, b]$, and $\int_a^b g(x)\,dx$ converges (it is equal to a finite real number), then $\int_a^b f(x)\,dx$ also converges.

If $f(x) \leq g(x)$ on $[a, b]$, and $\int_a^b f(x)\,dx = +\infty$ (so it diverges), then $\int_a^b g(x)\,dx$ also diverges to $+\infty$. Likewise, if $\int_a^b g(x)\,dx$ diverges to $-\infty$, then $\int_a^b f(x)\,dx$ also diverges to $-\infty$.

4.4 Matrices

An $m \times n$ **matrix** (here m and n may be either positive integers or ∞) is a table — more formally, a function – $a(x, y)$ defined for $x, y \in \mathbb{Z}_{<0}$ and $x \leq m$, $y \leq n$ (where by convention we say every positive integer is at most ∞). We say that this matrix has m rows indexed by x, and n columns indexed by y.

For an $m \times n$ matrix $A = (a(x, y))$ and an $n \times k$ matrix $B = (b(z, w))$, we define $C = AB$ to be the $m \times k$ matrix with entries

$$c(x, w) = \sum_{j=1}^{n} a(x, j)b(j, w).$$

In the case where n is infinite, this is only defined if all of the required sums are convergent, which will always be the case in the contexts we're considering.