Probability Theory Review

- probability space setup
- conditional probability
- random variables and their distributions
- joint distributions, conditional distributions, independence of random variables
- expectation, variance, covariance
- inequalities, confidence intervals, Weak Law of Large Numbers, Central Limit Theorem
- Markov chains

Too much material! Some of these slides will be skipped in lecture, but are included so you can look them over later.

Basic Probability Facts

We can derive facts from the probability axioms:

- $\mathbb{P}(\varnothing) = 0$.
- If $A^c := \Omega \setminus A$ is the complement of A, then $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
- Inclusion-Exclusion: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
- Union Bound: $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$. Then, by induction, $\mathbb{P}(\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^n \mathbb{P}(A_i)$.
- Generalized Inclusion-Exclusion: $\mathbb{P}(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n (-1)^{n+1} \sum_{S \subseteq \{1,\ldots,n\}, |S|=i} \mathbb{P}(\bigcap_{j \in S} A_j)$.
- Note: In probability, $\mathbb{P}(A \cap B)$ is abbreviated as $\mathbb{P}(A, B)$.

Probability Space Framework

A probability space consists of $(\Omega, \mathscr{F}, \mathbb{P})$ where:

- Ω is the set of all outcomes;
- $\mathscr{F}$ is the family of events, where an event is a subset of Ω;
- a probability law $\mathbb{P} : \mathscr{F} \to [0,1]$ satisfying $\mathbb{P}(\Omega) = 1$ and for all countable disjoint $A_1, A_2, A_3, \ldots$, $\mathbb{P}(\bigcup_{i=1}^\infty A_i) = \sum_{i=1}^\infty \mathbb{P}(A_i)$.

Important: Probabilities are assigned to *events*. If X is a random variable, then $\mathbb{P}(X)$ makes no sense!

Discrete probability:

- The sample space Ω is countable.
- The probability law is fully specified by giving the probability of each individual outcome $\mathbb{P}(\{\omega\})$ for $\omega \in \Omega$.
- For any event A, $\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\})$.
- We abbreviate $\mathbb{P}(\{\omega\})$ as $\mathbb{P}(\omega)$.

Conditional Probability

Given an event B with $\mathbb{P}(B) > 0$, define

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

Interpretation: After observing B, we move from the probability space Ω with law $\mathbb{P}$ to the smaller space B with law $\mathbb{P}(\cdot \mid B)$.

Why use conditional probability?

- Product Rule: $\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A \mid B)$. Calculate probabilities step-by-step!
- More generally, $\mathbb{P}(\bigcap_{i=1}^n A_i) = \mathbb{P}(A_1) \prod_{i=2}^n \mathbb{P}(A_i \mid \bigcap_{j=1}^{i-1} A_j)$.
- $B_1, \ldots, B_n$ partition Ω if they are disjoint and $\bigcup_{i=1}^n B_i = \Omega$.
- Law of Total Probability: If $B_1, \ldots, B_n$ partition Ω, then $\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A \cap B_i) = \sum_{i=1}^n \mathbb{P}(B_i)\mathbb{P}(A \mid B_i)$.
- "Divide and conquer" strategy for calculating $\mathbb{P}(A)$.

Continuous Probability

Continuous probability:

- Consider the uniform distribution on $\Omega = [0,1]$.
- The sample space is *uncountable*.
- It is no longer enough to specify the probability of each outcome. In fact, $\mathbb{P}(\omega) = 0$ for each $\omega \in [0,1]$.
- Instead, to fully specify the probability law, we must give the probability of *intervals* $[a,b]$, where $a < b$.
- For the uniform distribution, $\mathbb{P}([a,b]) = b - a$.

Role of the probability space:

- Usually, we care more about random variables; the probability space sits in the background, forgotten.
- Why did Kolmogorov need to set up a probability space? Answer: to unify probability with the rest of mathematics.

Bayes Rule

If $B_1, \ldots, B_n$ are possible causes, and A is the effect, then for each $i = 1, \ldots, n$,

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(B_i)\mathbb{P}(A \mid B_i)}{\sum_{j=1}^n \mathbb{P}(B_j)\mathbb{P}(A \mid B_j)}.$$

Why is this so important?

- $\mathbb{P}(B_i)$ is the *prior probability*. How likely is B_i, before you observe anything?
- $\mathbb{P}(A \mid B_i)$ is the *conditional probability*. The likelihood of the effect given the cause B_i.
- Bayes Rule tells you to multiply the two effects, and then renormalize.
- Then, you can *infer* what the probability of the cause is, given the effect. "Update your beliefs after observing A."

Independence

If $\mathbb{P}(A \mid B) = \mathbb{P}(A)$, we say that A and B are independent.

- Interpretation: Even after observing B, A is just as likely to occur as before.
- This definition is not so symmetric.

Alternative definition: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

- This definition works even when $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$.

What is the difference between disjoint and independent?

- Disjoint: $A \cap B = \varnothing$. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
- Independent: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
- Not the same! If A and B are disjoint, then after observing B, you know A cannot happen: $\mathbb{P}(A \mid B) = 0$.
- So if A and B are disjoint, then they are *not* independent (except for edge cases).

Distribution of a Random Variable

For a discrete RV X, the probability mass function (PMF), denoted p_X, is the function $p_X(x) := \mathbb{P}(X = x)$.

- For discrete probability spaces, it is enough to specify the probability of individual outcomes.
- Similarly, for discrete RVs, just specify the PMF.

For a continuous RV X, the PMF no longer makes sense.

- For continuous probability spaces, we needed to specify the probability of *intervals*.
- Similarly, here we need some way of specifying $\mathbb{P}(X \in [a,b])$ for intervals $[a,b]$.
- We use a probability density function (PDF): a function $f_X : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.
- Then, $\mathbb{P}(X \in [a,b]) = \int_a^b f_X(x)\,dx$.

Mutual Independence

What does it mean for events $A_1, \ldots, A_n$ to be independent?

Pairwise independence:

- Each pair of events is independent.
- Not as useful!

Mutual independence:

- For all subsets $S \subseteq \{1, \ldots, n\}$, $\mathbb{P}(\bigcap_{i \in S} A_i) = \prod_{i \in S} \mathbb{P}(A_i)$.
- For three events, this means pairwise independence plus another condition: $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$.
- Pairwise independence does *not* imply mutual independence.
- When we mention multiple objects being "independent", we mean mutually independent, unless otherwise stated.

Distribution of a Random Variable

Definition that works for both discrete and continuous RVs:

- The cumulative distribution function (CDF) is the function $F_X : \mathbb{R} \to [0,1]$ given by $F_X(x) := \mathbb{P}(X \leq x)$.
- Conversion table to/from CDF:

	X takes values in $\mathbb{Z}$	X continuous
to	$F_X(x) = \sum_{s=-\infty}^{x} p_X(s)$	$F_X(x) = \int_{-\infty}^{x} f_X(s)\,ds$
from	$p_X(x) = F_X(x) - F_X(x-1)$	$f_X(x) = \frac{d}{dx} F_X(x)$

What do we mean by "give the distribution of X"?

- Giving the CDF always works. Or, give the PMF (discrete) or PDF (continuous).
- Give a named distribution with parameters.

Random Variables

A random variable (RV) X is a function $X : \Omega \to \mathbb{R}$.

- What can you do with functions? Add them, multiply them.
- Can you take unions of functions? No! So, for random variables, $X_1 \cup X_2$ is nonsense!

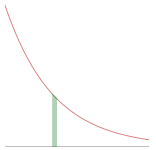
Notation:

- For $x \in \mathbb{R}$, $\{X = x\}$ is an *event*. So, $\{X = x\} \subseteq \Omega$. What does it mean?
- $\{X = x\} := \{\omega \in \Omega : X(\omega) = x\}$. The set of outcomes for which X takes on the value x.
- Another way to write this is $\{X = x\} = X^{-1}(\{x\})$ (the inverse image of $\{x\}$ under the function X).
- Similarly, for $A \subseteq \mathbb{R}$, $\{X \in A\}$ is the event $X^{-1}(A)$.
- Abbreviate $\mathbb{P}(\{X = x\})$ as $\mathbb{P}(X = x)$.

Interpretation of the PDF

For a continuous RV X, what is the interpretation of f_X?

Let $\delta > 0$ be very small. Then, $\mathbb{P}(X \in [x, x+\delta]) = \int_x^{x+\delta} f_X(s)\,ds$. Plot of $f_X(x)$ vs. x:



This is approximately $f_X(x) \cdot \delta$.

So,

$$f_X(x) \approx \frac{\mathbb{P}(X \in [x, x+\delta])}{\delta}.$$

$f_X(x)$ is like the "probability per unit length" near x.

Multiple Random Variables

When we have two random variables, X and Y, we describe their *joint distribution* via:

- (both discrete) the joint PMF $p_{X,Y}(x,y) := \mathbb{P}(X = x, Y = y)$;
- (both continuous) the joint PDF $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ such that $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1$.

The definitions extend for more than two RVs.

The notation $\mathbb{P}(X = x, Y = y)$ is shorthand for $\mathbb{P}(\{X = x\} \cap \{Y = y\}) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\})$.

How to calculate probabilities: for $A \subseteq \mathbb{R}^2$,

- (discrete) $\mathbb{P}((X,Y) \in A) = \sum_{(x,y)\in A} p_{X,Y}(x,y)$;
- (continuous) $\mathbb{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y$.

Conditional Distributions

The Law of Total Probability has analogues for RVs too.

- (X is $\mathbb{Z}$-valued, Y is $\mathbb{Z}$-valued) $p_X(x) = \sum_{y=-\infty}^{\infty} p_{X,Y}(x,y) = \sum_{y=-\infty}^{\infty} p_Y(y) p_{X|Y}(x \mid y)$.
- (X is $\mathbb{Z}$-valued, Y is continuous) $p_X(x) = \int_{-\infty}^{\infty} f_Y(y) p_{X|Y}(x \mid y)\,\mathrm{d}y$
- (X is continuous, Y is $\mathbb{Z}$-valued) $f_X(x) = \sum_{y=-\infty}^{\infty} p_Y(y) f_{X|Y}(x \mid y)$
- (X is continuous, Y is continuous) $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x \mid y)\,\mathrm{d}y$

Finding the probability of an event by conditioning on X:

- (X is $\mathbb{Z}$-valued) $\mathbb{P}(A) = \sum_{x=-\infty}^{\infty} p_X(x)\mathbb{P}(A \mid X = x)$
- (X is continuous) $\mathbb{P}(A) = \int_{-\infty}^{\infty} f_X(x)\mathbb{P}(A \mid X = x)\,\mathrm{d}x$

Marginalization

Given the joint distribution of (X,Y), we can recover the *marginal distribution* of X via:

- (discrete, Y takes values in $\mathbb{Z}$) $p_X(x) = \sum_{y=-\infty}^{\infty} p_{X,Y}(x,y)$;
- (continuous) $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(y)\,\mathrm{d}y$.

Example where X and Y take on two values each:

	$y=0$	$y=1$	p_X
$x=0$	$p_{X,Y}(0,0)=0.1$	$p_{X,Y}(0,1)=0.3$	$p_X(0)=0.4$
$x=1$	$p_{X,Y}(1,0)=0.2$	$p_{X,Y}(1,1)=0.4$	$p_X(1)=0.6$
p_Y	$p_Y(0)=0.3$	$p_Y(1)=0.7$	

- From the joint, we can recover the marginals.
- From the marginals of X and Y, we *do not have enough information to recover the joint*.

Bayes Rule for RVs

X is $\mathbb{Z}$-valued, Y is $\mathbb{Z}$-valued:

$$p_{X|Y}(x \mid y) = \frac{p_X(x)p_{Y|X}(y \mid x)}{p_Y(y)} = \frac{p_X(x)p_{Y|X}(y \mid x)}{\sum_{x'=-\infty}^{\infty} p_X(x')p_{Y|X}(y \mid x')}$$

X is $\mathbb{Z}$-valued, Y is continuous:

$$p_{X|Y}(x \mid y) = \frac{p_X(x)f_{Y|X}(y \mid x)}{f_Y(y)} = \frac{p_X(x)f_{Y|X}(y \mid x)}{\sum_{x'=-\infty}^{\infty} p_X(x')f_{Y|X}(y \mid x')}$$

X is continuous, Y is $\mathbb{Z}$-valued:

$$f_{X|Y}(x \mid y) = \frac{f_X(x)p_{Y|X}(y \mid x)}{p_Y(y)} = \frac{f_X(x)p_{Y|X}(y \mid x)}{\int_{-\infty}^{\infty} f_X(x')p_{Y|X}(y \mid x')\,\mathrm{d}x'}$$

X is continuous, Y is continuous:

$$f_{X|Y}(x \mid y) = \frac{f_X(x)f_{Y|X}(y \mid x)}{f_Y(y)} = \frac{f_X(x)f_{Y|X}(y \mid x)}{\int_{-\infty}^{\infty} f_X(x')f_{Y|X}(y \mid x')\,\mathrm{d}x'}$$

Conditional Distributions

Independence of RVs: X and Y are independent if for all A, B, $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$.

- (both discrete) the joint PMF factorizes: $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ for all x, y;
- (both continuous) the joint PDF factorizes: $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all x, y.
- If X and Y are independent, then so are $f(X)$ and $g(Y)$ (any functions of them).
- "i.i.d." means "independent and identically distributed".

Conditional distributions:

- (conditional PMF) $p_{X|Y}(x \mid y) := \mathbb{P}(X = x \mid Y = y) = p_{X,Y}(x,y)/p_Y(y)$;
- (conditional PDF) $f_{X|Y}(x \mid y) := f_{X,Y}(x,y)/f_Y(y)$.

Expectation

For a random variable X, the *expectation* of X, $\mathbb{E}[X]$, is:

- (discrete, X is $\mathbb{Z}$-valued) $\mathbb{E}[X] = \sum_{x=-\infty}^{\infty} x p_X(x)$;
- (continuous) $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x$.
- Interpretation: What is the center of mass of the PMF/PDF?

What about a function of X, such as $\mathbb{E}[X^2]$?

- Say X is continuous. By the definition, $\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x f_{X^2}(x)\,\mathrm{d}x$. What is f_{X^2}?
- Easier way: $\mathbb{E}[g(X)]$ is $\sum_{x=-\infty}^{\infty} g(x)p_X(x)$, or $\int_{-\infty}^{\infty} g(x)f_X(x)\,\mathrm{d}x$.

Linearity of expectation: $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, always. Independence is not required!

Expectation

Tail Sum Formula: If $X \geq 0$ (shorthand for $\mathbb{P}(X \geq 0) = 1$), then:

- $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq x)\,dx = \int_0^\infty \mathbb{P}(X > x)\,dx$ (always holds).
- If X is $\mathbb{N}$-valued, then $\mathbb{E}[X] = \sum_{x=1}^\infty \mathbb{P}(X \geq x) = \sum_{x=0}^\infty \mathbb{P}(X > x)$.
- The second version can be derived from the first version (think about it).

Conditional expectation: For an event A,

- (discrete, X is $\mathbb{Z}$-valued) $\mathbb{E}[X \mid A] = \sum_{x=-\infty}^\infty x p_{X|A}(x) = \sum_{x=-\infty}^\infty x \mathbb{P}(X = x \mid A)$;
- (continuous) $\mathbb{E}[X \mid A] = \int_{-\infty}^\infty x f_{X|A}(x)\,dx$.

Covariance, Variance

For random variables X and Y, the *covariance* is $\operatorname{cov}(X,Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

- Fact: If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, i.e., $\operatorname{cov}(X,Y) = 0$.
- Warning: If $\operatorname{cov}(X,Y) = 0$, then X and Y are not necessarily independent.

Then, the variance is $\operatorname{var} X := \operatorname{cov}(X,X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

- Interpretation: How spread out is the distribution of X?
- Variance of sum: $\operatorname{var}(X+Y) = \operatorname{var} X + \operatorname{var} Y + 2\operatorname{cov}(X,Y)$. Similar to $(a+b)^2 = a^2 + b^2 + 2ab$.
- Generally, $\operatorname{var}(cX) = c^2 \operatorname{var} X$ and $\operatorname{var} \sum_{i=1}^n X_i = \sum_{i=1}^n \operatorname{var} X_i + \sum_{i,j \in \{1,\ldots,n\},\, i \neq j} \operatorname{cov}(X_i, X_j)$.
- So if $X_1, \ldots, X_n$ are independent, $\operatorname{var} \sum_{i=1}^n X_i = \sum_{i=1}^n \operatorname{var} X_i$.

Indicators

For an event A, the indicator random variable $\mathbb{1}_A$ is the random variable such that $\mathbb{1}_A(\omega) = 1$ if $\omega \in A$, and $\mathbb{1}_A(\omega) = 0$ if $\omega \notin A$.

- Fact: A and B are independent if and only if $\mathbb{1}_A$ and $\mathbb{1}_B$ are independent as RVs.
- Since $\mathbb{1}_A$ takes values in $\{0,1\}$, then raising $\mathbb{1}_A$ to any power does not change it: $\mathbb{1}_A^n = \mathbb{1}_A$.
- $\mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A)$ and $\operatorname{var} \mathbb{1}_A = \mathbb{P}(A)[1 - \mathbb{P}(A)]$.
- Product of indicators: $\mathbb{1}_A \mathbb{1}_B = \mathbb{1}_{A \cap B}$.

Indicator method for calculating expectation/variance:

- If the random variable X counts the number of occurrences, write X as the sum of indicator RVs.
- Example: Pick a permutation of $\{1, \ldots, n\}$ uniformly at random; X is the number of fixed points. For $i = 1, \ldots, n$, X_i indicates if the ith position is fixed, so $X = \sum_{i=1}^n X_i$.

Discrete Distributions Reference

Bernoulli(p) ($p \in [0,1]$):

- Same as Binomial($1,p$).

Binomial(n,p) ($n \in \mathbb{Z}^+$, $p \in [0,1]$):

- PMF: $p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x \in \{0,1,\ldots,n\}$.
- $\mathbb{E}[X] = np$, $\operatorname{var} X = np(1-p)$.

Geometric(p) ($p \in (0,1]$):

- PMF: $p_X(x) = p(1-p)^{x-1}$ for $x \in \mathbb{Z}^+$.
- $\mathbb{E}[X] = 1/p$, $\operatorname{var} X = (1-p)/p^2$.

Poisson(λ) ($\lambda \in (0,\infty)$):

- PMF: $p_X(x) = \exp(-\lambda)\lambda^x / x!$ for $x \in \mathbb{N}$.
- $\mathbb{E}[X] = \lambda$, $\operatorname{var} X = \lambda$.

Continuous Distributions Reference

Uniform($[a,b]$) ($-\infty < a < b < \infty$):

- PDF: $f_X(x) = 1/(b-a)$ for $x \in [a,b]$.
- $\mathbb{E}[X] = (a+b)/2$, $\operatorname{var} X = (b-a)^2/12$.

Exponential(λ) ($\lambda \in (0,\infty)$):

- PDF: $f_X(x) = \lambda \exp(-\lambda x)$ for $x \geq 0$.
- $\mathbb{E}[X] = 1/\lambda$, $\operatorname{var} X = 1/\lambda^2$.

Normal(μ, σ^2) ($\mu \in \mathbb{R}$, $\sigma^2 \geq 0$):

- PDF: $f_X(x) = (2\pi\sigma^2)^{-1/2} \exp[-(x-\mu)^2/(2\sigma^2)]$.
- $\mathbb{E}[X] = \mu$, $\operatorname{var} X = \sigma^2$.

Notes on Distributions

Binomial(n,p): Models the number of successes in n independent trials, if each success has probability p.

- X is the sum of n i.i.d. Bernoulli(p).

Geometric(p): Models the number of flips needed to get Heads, with a coin of bias p.

- Memoryless: For $k, n \in \mathbb{Z}^+$, $p_{X|\{X>n\}}(n+k) = p_X(k)$.

Poisson(λ): Models rare events.

- Independent Poisson(λ) + Poisson(μ) = Poisson($\lambda + \mu$).
- For $i = 1, \ldots, n$, if $X_i \sim$ Bernoulli(λ/n), then $\sum_{i=1}^n X_i$ is approximately Poisson(λ) for large n.

Exponential(λ): Models radioactive decay.

- Memoryless: For $s, t \in \mathbb{R}$, $\mathbb{P}(X > s+t \mid X > s) = \mathbb{P}(X > t)$.

Normal(μ, σ^2): Models noise/sum of i.i.d. effects.

- Sum of independent Normal(μ_1, σ_1^2) and Normal(μ_2, σ_2^2) is Normal($\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2$).

Some Important Problems (Brief Solutions)

Birthday problem: Throw k balls into n bins, independently and uniformly at random. What is the probability of no collisions?

- Answer: $(1-1/n)(1-2/n)\cdots(1-(k-1)/n)$.
- Approximation: $\approx \exp(-1/n)\exp(-2/n)\cdots\exp(-(k-1)/n) \approx \exp(-k^2/n)$.

Indicators: Pick a permutation of $\{1,\ldots,n\}$ uniformly at random. Let X be the number of fixed points. What are $\mathbb{E}[X]$ and $\operatorname{var} X$?

- Use indicators. $\mathbb{E}[X] = \operatorname{var} X = 1$.

Coupon collector: Each time you buy a box, you get one of n coupons independently and uniformly at random. Expected number of boxes needed to collect all coupons?

- Let X_i be the time to get the ith new coupon; $X_i \sim \text{Geometric}((n-(i-1))/n)$.
- By linearity of expectation, $\mathbb{E}[X] = n\sum_{i=1}^{n} i^{-1} \approx n \ln n$.

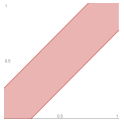
Some Important Problems (Brief Solutions)

Sum of independent Poisson: $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$ are independent; show $X+Y \sim \text{Poisson}(\lambda+\mu)$.

- Key idea: Use Law of Total Probability.
- For $k \in \mathbb{N}$, $\mathbb{P}(X+Y=k) = \sum_{x=0}^{k} \mathbb{P}(X=x, Y=k-x) = \sum_{x=0}^{k} \mathbb{P}(X=x)\mathbb{P}(Y=k-x)$. Work through the algebra.

Lunch meeting: A and B independently arrive uniformly in the interval $[0,1]$; they will eat together if they arrive within 0.25 of each other. What is the probability they eat together?

- Key idea: Draw a diagram!



(X,Y) is uniformly distributed on the unit square; the probability is the shaded area $(9/16)$.

Inequalities, Weak Law of Large Numbers

Markov's Inequality:

- If $X \geq 0$, then for $t > 0$, $\mathbb{P}(X \geq t) \leq \mathbb{E}[X]/t$.

Chebyshev's Inequality:

- For $\varepsilon > 0$, $\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq (\operatorname{var} X)/\varepsilon^2$.
- $|X - \mathbb{E}[X]| \geq \varepsilon$ means X is at least ε-far from its mean; X does not belong to $(\mathbb{E}[X]-\varepsilon, \mathbb{E}[X]+\varepsilon)$.

Weak Law of Large Numbers (WLLN):

- Suppose $X_1,\ldots,X_n$ are i.i.d., mean μ, variance σ^2. Define the *sample mean* $\bar{X}_n := (\sum_{i=1}^{n} X_i)/n$.
- Quick check: is $\bar{X}_n$ the same as μ? No, $\bar{X}_n$ is a random variable! It depends on $X_1,\ldots,X_n$!
- Calculate: $\mathbb{E}[\bar{X}_n] = \mu$ and $\operatorname{var} \bar{X}_n = \sigma^2/n$.
- Variance shrinks with n. By Chebyshev, $\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \to 0$ as $n \to \infty$.

Weak Law of Large Numbers

What is the interpretation of the WLLN?

- Take an ε-wide interval around μ. As $n \to \infty$, all of the probability mass enters this interval.
- The distribution of $\bar{X}_n$ becomes more and more sharply peaked around μ.
- When n is very large, $\bar{X}_n$ is almost a constant (almost exactly μ).

What is the use of the WLLN?

- Suppose we do not know μ.
- If we collect enough samples (n is large), then $\bar{X}_n$ is basically the same thing as μ.

Confidence Intervals

Given n samples and $\delta > 0$, a $1-\delta$ confidence interval for μ is a random interval $(\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon)$ so that with probability $\geq 1-\delta$, μ lies in the interval.

Say what?

- n: The number of samples taken.
- δ: The probability that the confidence interval fails.
- ε: The tolerance, or width, of the confidence interval.
- Need $\mathbb{P}(\mu \notin (\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon)) = \mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \leq \delta$.
- You will be given two out of the three: n, δ, ε. Solve for the quantity you are not given.

Central Limit Theorem

From the WLLN, we know that if we add $X_1,\ldots,X_n$ and divide by n, then we lose almost all information about the distribution: for large n, $(\sum_{i=1}^{n} X_i)/n$ is basically a constant, μ.

Can we choose a different scaling to retain more information about the distribution?

- Sum the centered RVs: $\sum_{i=1}^{n}(X_i - \mu)$.
- Scale by $n^{-1/2}$: $Z_n := n^{-1/2}\sum_{i=1}^{n}(X_i - \mu)$.
- Central Limit Theorem (CLT): As $n \to \infty$, Z_n converges in distribution to Normal$(0,\sigma^2)$, in the sense that

$$\mathbb{P}(Z_n \leq z) \to \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx.$$

Finite-State Markov Chains

What is a Markov chain?

- Is it the transition probability matrix P? No!

A Markov chain is a sequence of RVs $(X_n)_{n\in\mathbb{N}}$ taking values in a finite *state space* S, satisfying the *Markov property*.

Markov Property: For all $n \in \mathbb{N}$ and feasible sequences of states $i_0, i_1, \ldots, i_{n-1}, i, j$,

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0)$$

does *not* depend on $i_0, i_1, \ldots, i_{n-1}$ and n.

- "Not depending on n" is *time-homogeneity*.
- We call this quantity $P(i,j)$ and we put it in the (i,j) entry of an $|S| \times |S|$ transition probability matrix P.

Distribution of a Markov Chain

The distribution of X_0 is called the *initial distribution*.

- To discuss the distributions of $(X_n)_{n\in\mathbb{N}}$, we use different notation. For all $n \in \mathbb{N}$ and $i \in S$, $\pi_n(i) := \mathbb{P}(X_n = i)$.
- We think of π_n as a *row vector* (of length $|S|$).
- Quick quiz: is π_n a random variable? No!

Transition of distribution:

- In matrix notation, $\pi_1 = \pi_0 P$.
- Hitting the distribution vector by the transition matrix advances the dynamics by one step!
- Then, $\pi_n = \pi_0 P^n$.

First Step Equations

Hitting time: What is the expected time to hit state $j \in S$?

- For $i \in S$, define $\beta(i) := \mathbb{E}[\text{steps to hit } j \mid X_0 = i]$.
- Boundary condition: $\beta(j) = 0$.
- For $i \neq j$, $\beta(i) = 1 + \sum_{k\in S} P(i,k)\beta(k) = \sum_{k\in S} P(i,k)[\beta(k)+1]$.
- This is a system of linear equations which we can solve.
- If we start from initial distribution π_0, the answer is $\sum_{i\in S} \pi_0(i)\beta(i)$.

Probability of hitting A before B: Let $A, B \subseteq S$. What is the probability of hitting A before B?

- For $i \in S$, define $\alpha(i) := \mathbb{P}(\text{hit } A \text{ before } B \mid X_0 = i)$.
- Boundary conditions: $\alpha(i) = 0$ for $i \in B$; $\alpha(i) = 1$ for $i \in A$.
- For $i \in S \setminus (A \cup B)$, then $\alpha(i) = \sum_{k\in S} P(i,k)\alpha(k)$.
- Again, solve the system of linear equations.

Classification of States

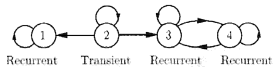


Figure: Figure taken from *Introduction to Probability* by Bertsekas and Tsitsiklis, 2nd edition.

A *class* is a set of states in which every state can talk to any other state.

- A class is recurrent if it only has arrows pointing inwards.
- A class is transient if it has arrows pointing to other classes.

The probability mass "leaks out of" transient classes and remains in recurrent classes. Therefore, only recurrent classes matter for long-term Markov chain behavior.

Long-Run Behavior: Stationarity

A Markov chain with exactly one recurrent class (no transient classes) is called *irreducible*.

- These are the simplest building blocks for analyzing long-term behavior of a Markov chain.

Stationary distribution: The probability distribution π (a row vector) is called a *stationary (or invariant) distribution* if $\pi = \pi P$.

- If the initial distribution is π, then the distribution remains π for all time.
- The equation $\pi = \pi P$ is a system of linear equations, the *balance equations*, which can be solved for π.
- Every Markov chain has at least one stationary distribution.
- If the Markov chain is irreducible, the stationary distribution is unique.

Long-Run Behavior: Convergence

MC Law of Large Numbers: If $(X_n)_{n\in\mathbb{N}}$ is an irreducible MC, then for any state $i \in S$ and any π_0, $n^{-1}\sum_{m=0}^{n-1} \mathbb{1}_{\{X_m = i\}}$ converges, as $n \to \infty$, to $\pi(i)$, where π is the stationary distribution.

- Interpretation: The *fraction of time spent in state i* converges to $\pi(i)$.
- Quick quiz: is $n^{-1}\sum_{m=0}^{n-1} \mathbb{1}_{\{X_m = i\}}$ a RV? Yes, it depends on $X_0, X_1, \ldots, X_{n-1}$.
- Convergence occurs in the sense that for all $\varepsilon > 0$, $\mathbb{P}(|n^{-1}\sum_{m=0}^{n-1} \mathbb{1}_{\{X_m = i\}} - \pi(i)| \geq \varepsilon) \to 0$ as $n \to \infty$.
- One consequence: If we take expectations, $n^{-1}\sum_{m=0}^{n-1} \mathbb{P}(X_m = i) = n^{-1}\sum_{m=0}^{n-1} \pi_m(i)$ converges to $\pi(i)$.

In other words, the *average of the distributions* over the first n time steps, $n^{-1}\sum_{m=0}^{n-1} \pi_m$, converges to π.

Do the distributions themselves converge?

Long-Run Behavior: Convergence

Aperiodicity:

- An MC is *periodic* it can be grouped into $d > 1$ groups of states such that each group only flows into the next group.
- Otherwise, it is *aperiodic*.
- If the MC has a self-loop, it is aperiodic.

MC Convergence Theorem: If the MC $(X_n)_{n \in \mathbb{N}}$ is irreducible and aperiodic, then $\pi_n \to \pi$ as $n \to \infty$. In other words, $\mathbb{P}(X_n = i) \to \pi(i)$ as $n \to \infty$ for every $i \in S$.

- The theorem holds, *regardless of the initial distribution* π_0.
- If the chain is periodic, then convergence can still happen for *some* initial distributions.

Next Lectures

The next two lectures will cover applications of discrete mathematics and probability theory.

These lectures will not be covered on the final, so they are optional.

Nevertheless, they may still help you practice the concepts of discrete math/probability.