EECS126 Notes

Lecturers: Kannan Ramchandran and Abhay Parekh
Scribe: Michael Whitmeyer

Jan 2020 - May 2020 (with some examples from Fall 2018)

Lecture 1: Overview and beginning of CS70 Review

0.1 Motivation

1. Uncertainty is all around us!
2. This course is about formalizing how to predict things.
3. Actually has origins in gambling
4. First need to develop **model** (requires understanding of the problem as an experiment), and then need to **solve** (using combinatorics, calculus, common sense, etc). As engineers, we need to do both, but often it is (perhaps unexpectedly) the modelling that is more difficult than the solving.
5. Last but certainly not least (for many of you), foundational for ML/AI.

0.2 Content

Definition 0.1. A **Sample Space** Ω of an experiment is the set of all outcomes of the experiment. The outcomes must be *mutually exclusive* (ME) and *collectively exhaustive* (CE)

Example 0.2. Toss two fair coins. Then we have $\Omega = \{HH, HT, TH, TT\}$. Can check that these outcomes are ME and CE.

Definition 0.3. An **Event** is simply an allowable subset of Ω.

Example 0.4. In Ex 0.2 an event would be getting at least 1 Head

Definition 0.5. A **Probability Space** $(\Omega, \mathcal{F}, P)$ is a mathematical construct that allows us to model these "experiments". Here $\mathcal{F}$ denotes the set of all possible events, where each event is a set containing 0 or more base outcomes (for discrete Ω this is simply the power set of Ω). And $P : \mathcal{F} \mapsto [0, 1]$ is a function assigning probabilities to each event.

All of Probability Theory rests on just 3 (2.5?) axioms (Kolmogorov):

1. $\Pr(\emptyset) = 0$
2. $\Pr(\Omega) = 1$
3. $Pr(A_1 \cup A_2 \cup ...) = Pr(A_1) + Pr(A_2) + ...$ for all disjoint $A_1, A_2, ...$ This can be finite or we can take $n \to \infty$ and this becomes *countable additivity*.

We immediately have the following fundamental facts:

1. $Pr(A^c) = 1 - Pr(A)$
2. $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$
3. Union bound: $Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} Pr(A_i)$
4. Inclusion-Exclusion:

$$Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum Pr(A_i) - \sum_{i<j} Pr(A_i \cap A_j) + \sum_{i<j<k} Pr(A_i \cap A_j \cap A_k) - ...$$

In the discrete setting, just from the axioms, we have that $Pr(A) = \sum_{\omega \in A} Pr(\omega)$. If our sample space is *uniform*, then we have that $Pr(A) = \frac{|A|}{|\Omega|}$

0.3 Conditional Probability

Definition 0.6. In general, we use the notation $Pr(A|B)$ = the probability that event A has occurred *given* that we know that B has occurred.

Proposition 0.7 (Bayes Rule)**.**

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$

"B is the new Ω"

Example 0.8. We roll two six sided die, and observe that the sum of the two die is 11. What is is the probability that the first die was a 6? Here let A = event of a 6 on the first die and B = event of sum being 11.

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{Pr(\{6,5\})}{Pr(\{6,5\}) + Pr(\{5,6\})} = \frac{1}{2}$$

Bayes Rule directly extends to the **Product Rule**, which says that

$$Pr(A_1 \cap A_2 \cap ... \cap A_n) = Pr(A_1)Pr(A_2|A_1)Pr(A_3|A_1 \cap A_2) \cdots Pr(A_n|A_1 \cap ... \cap A_{n-1})$$

We also can develop the **Law of Total Probability**, which says that for ME and CE events $A_1, ..., A_n$, we have that

$$Pr(B) = Pr(A_1 \cap B) + ... + Pr(A_n \cap B) = \sum_{i=1}^{n} Pr(A_i)Pr(B|A_i)$$

This can be easily visualized via the following picture:

TODO 1. include picture

Lecture 2: Independence, Bayes Rule, Discrete Random Variables

We will begin with a cool example.

Example 0.9 (Birthday Paradox)**.** Want to estimate Pr(at least two people in a group of size n share the same birthday). First we note that $|\Omega| = k^n = 365^n$. The problem is that this event is a bit complicated. So we will consider the complement: $A^c =$ "no two people share a birthday" Then, since the distributions are uniform, we have

$$Pr(A^c) = \frac{|A^c|}{|\Omega|} = \frac{365 * 364 * ... * (365 - n + 1)}{365^n}$$

$$= 1(1 - \frac{1}{k})(1 - \frac{2}{k}) \cdots (1 - \frac{n-1}{k})$$

$$\approx e^{-1/k} e^{-2/k} \cdots e^{-(n-1)/k} = e^{-\frac{1}{k}(1+...+n-1)}$$

$$\approx e^{-n^2/k}$$

So then we have that

$$Pr(A) = 1 - Pr(A^c) \approx 1 - e^{-n^2/k}$$

It turns out, for $k = 365$ and $n = 23$, we get a roughly 50% chance of two people having the same birthday!

0.4 Bayes Theorem

Bayes Theorem was motivated by disease testing.

Example 0.10 (False Positive Quiz)**.** We are testing for a rare disease, and our test has the following properties:

- If person has disease, we detect with 0.95 probability.
- If person doesn't have the disease, test is negative wp 0.95
- Random person has disease wp 0.001

Let A be the event that the person has the disease, and B be the event that the person tests positive. We would like to calculate $Pr(A|B)$. We have via Bayes Thm that

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B|A)Pr(A) + Pr(B|A^c)Pr(A^c)}$$

$$= \frac{(0.95)(0.001)}{(0.95)(0.001) + (0.999)(0.05)} = 0.0187$$

Most doctors, when asked, said this probability was 95%. The main contributing factor here is the fact that the prior $Pr(A) = 0.001$ is so small. If we change the scenario and have $Pr(A) = 0.01$, then our new probability $Pr(A|B) = 0.16$, so we should be more worried. Note that the doctor would actually be correct if the disease were present in 1/2 of the population.

Definition 0.11. Two events are **Independent** if the occurrence of one provides no information about the occurrence of the other. i.e.

$$Pr(A|B) = Pr(A)$$

which is equivalent to saying

$$Pr(A \cap B) = Pr(A)Pr(B)$$

Extending this, a collection of events S are independent if

$$Pr(\bigcap_{i \in S} A_i) = \prod_{i \in S} Pr(A_i)$$

Remark 0.12. Pairwise Indpendence does not imply joint independence.

Remark 0.13. Being disjoint does not imply independence, nor does the implication hold in the other direction. If A, B are disjoint, then $Pr(A \cap B) = 0$, and independence tells us that $Pr(A \cap B) = Pr(A)Pr(B)$, which would tell us for two events to be both disjoint and independent, at least one of the two events must have zero probability of occurring. Note that this tells us that base outcomes of our probability space, which are all disjoint by definition, and all have nonzero probability by definition, must not be independent.

Definition 0.14. Conditional Independence is when $Pr(A \cap B|C) = Pr(A|C)Pr(B|C)$. Then we say that A and B are "conditionally independent given C".

TODO 2. some picture showing how we can represent independence in the venn diagram? Like the ratios have to be the same

Example 0.15. say we have two coins, one with tails on both sides, one with heads on both sides. We pick one up at random, and we flip it twice. We also let H_i be the event that that i'th flip is a heads. Note immediately that H_1 and H_2 are decidedly not independent. Now we denote A as the event of us picking the two-headed coin. Then we have that $Pr(H_1 \cap H_2|A) = Pr(H_1|A)Pr(H_2|A \cap H_1) = Pr(H_1|A)Pr(H_2|A)$. So this is an example of events that are conditionally independent but not themselves independent.

Exercise 0.16. Construct an example of RVs that are indepedent, but not conditionally independent.

Example 0.17. I roll two fair die. What is the probability I see a 6 before I see a 7? Let's use independence to attack this problem.

Lets condition on the first roll. Let S be the event that the first roll of the two die is a 6, and T be the event that the first roll is a 7. Let E be the event we are looking for, that we see a 6 before we see a 7. Then we have

$$Pr(E) = Pr(E|S)Pr(S) + Pr(E|T)Pr(T) + Pr(E|(S \cup T)^c)Pr((S \cup T)^c)$$

$$= 1 * 5/36 + 0 * 6/36 + Pr(E) * 25/36$$

$$\rightarrow Pr(E) = 5/11$$

0.5 (Discrete) Random Variables

Definition 0.18. Random Variables associate a real number with each possible event. They are inherently a *function* $f : \Omega \to \mathbb{R}$.

Why is this useful? If we are stuck with only events, we have no numbers to work with, we can't calculate means and variances and we cannot do statistics. Heads and tails only gets us so far, but if we assign the value 0 or 1 now we can do *math*.

Example 0.19 (Some random variables). 1. The RV X has value i if the throw of a die is i.

2. X^2 is a perfectly valid random variable.

Consider rolling two four-sided die. Then M_k is the *event* that the min is k, whereas we can say M is the *random variable* that is equal to the value of the minimum of the two die. By enumerating all the possible values of the two die roll (which are all equal probability), we can see that $M = 1$ wp $7/16$, $M = 2$ wp $5/16$, $M = 3$ wp $3/16$, and $M = 4$ wp $1/16$. This mapping from values of a RV to probabilities for discrete random variables is known as a **probability mass function** or **PMF**, and in a way it defines the random variable.

Just to go over some notation, we have the PMF of a RV X is $P_X(x)$, and we say $P_X(x) = Pr(\{X = x\})$, often simply denoted $Pr(X = x)$. We also have to have (in order for our PMf to be valid), that

$$\sum_x P_X(x) = 1, \qquad P(X \in S) = \sum_{x \in S} P_X(x)$$

Example 0.20 (chess). Imagine Vishy Anand is playing Kasparov in chess (when they are at the height of their power). They play 10 games, and for each individual game, the probability Anand wins is 0.3, the probability that Kasparov wins is 0.4, and then probability that they draw is 0.3. The first to win a game wins the match, and if there are ten consecutive draws then the match is drawn. **Question:** what is the PMF of the deration of the match L? We have that

$$P_L(l) = \begin{cases} 0.3^9 & l = 10 \\ 0.3^{l-1} \cdot 0.7 & 1 \leq l \leq 9 \end{cases}$$

Question: What is the probability that Anand wins the match?

$$\Pr(\text{A wins the match}) = \sum_{l=0}^{9} (0.3)^l (0.3)$$

which can be simplified using the formula for a geometric series.

Lecture 3: Expectation, Uniform, Geometric, Binomial and Poisson Distributions

Agenda:

1. Recap of Discrete RVs and Probability Mass Functions (PMF)
2. Expectation
3. Some popular Disrete RVs
4. Variance

As a reminder, **discrete random variables** (DRVs) associate a real number with each possible outcome, so they are really just functions from $\Omega \to \mathbb{R}$. The distribution or **PMF** is the collection of values $\{a, P_X(a) : a \in \mathcal{A}\}$ where $\mathcal{A}$ is the set of all possible values taken by the RV X

Remark 0.21 (Functions of RVs are still RVs)**.** Let $Y = g(X)$. Then we have

$$P_Y(y) = \sum_{\{x|g(x)=y\}} P_X(x)$$

An RV itself is a function, and the function of a function is still a function!

Example 0.22. Let $Y = |X|$, where X is uniformly distributed between -2 and 2. So then $P_X(x) = 1/5, \quad \forall x \in \{-2, -1, 0, 1, 2\}$. Then $P_Y(y) = 2/5$ for $y = 1, 2$ and $P_Y(y) = 1/5$ for $y = 0$.

0.6 Expectation

Definition 0.23. We have the **expectation** of a discrete RV X that takes on values in a set $\mathcal{X}$ is

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x P_X(x)$$

Alternatively, we also have $\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) P(\omega)$

Theorem 0.24. *Let $Y = g(X)$. Then we have that*

$$\mathbb{E}[Y] = \sum_y y Pr(Y = y) = \sum_x g(x) Pr(X = x)$$

Note that there are no restrictions on the function g, so it holds for any function.

Proof. We start by noting

$$Pr(Y = y) = \sum_{x:g(x)=y} Pr(X = x)$$

Then, we have that

$$\mathbb{E}[Y] = \sum_y y Pr(Y = y) = \sum_y y \sum_{x:g(x)=y} Pr(X = x)$$

$$= \sum_y \sum_{x:g(x)=y} g(x) Pr(X = x)$$

$$= \sum_x g(x) Pr(X = x)$$

□

We can then use the above theorem to prove linearity of expectations!

Theorem 0.25. *(Linearity of Expectation)*
We have

$$E[X + Y] = E[X] + E[Y]$$

for arbitrary X and Y that are defined on the same probability space. This of course generalizes (via induction) to more than just two RVs.

Proof. We let $g(X, Y) = X + Y$. Then, according to the above theorem, we have that

$$\mathbb{E}[X + Y] = \sum_{x,y} (x + y) Pr(X = x, Y = y)$$

$$= \sum_{x,y} x Pr(X = x, Y = y) + \sum_{x,y} y Pr(X = x, Y = y)$$

$$= \sum_x \sum_y x Pr(X = x, Y = y) + \sum_y \sum_x y Pr(X = x, Y = y)$$

$$= \sum_x x \sum_y Pr(X = x, Y = y) + \sum_y y \sum_x Pr(X = x, Y = y)$$

$$= \sum_x x Pr(X = x) + \sum_y y Pr(Y = y) = \mathbb{E}[X] + \mathbb{E}[Y]$$

Where the last step follows from the law of total probability. □

Example 0.26. The average of the sum of two rolls of the dice X_1, X_2 is

$$\mathbb{E}[X] = \mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2] = 7$$

Example 0.27. Suppose Prof Ramchandran collects homeworks from n students, shuffles them randomly, and then hands them back (at random). What is the expected number of students who get their homework back? More formally, what is the expected number of fixed points in a random permutation of n points?

Solution: Let X_i be the indicator RV that equals 1 if student i gets his/hw back, and equals 0 otherwise. Then we can note that the number of students who get their homework back X, is exactly equal to $X_1 + ... + X_n$. So then

$$\mathbb{E}[X] = \mathbb{E}[X_1 + ... + X_n] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum i = 1^n \Pr(X_i = 1) = n \cdot \frac{1}{n} = 1$$

Remarkably, we see that the expected number of fixed points is always 1, regardless of how large or small n is. This technique of defining indicator RVs and applying linearity of expectation is extremely powerful and will come up over and over again in this course.

Remark 0.28. The X_i's in the previous example are **not** independent (exercise: why?), yet we can still apply linearity of expectation!

Definition 0.29. We define the **Variance** of a RV X, sometimes denoted σ_X^2 is

$$\mathsf{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

And furthermore the **standard deviation** is

$$\sigma_X = \sqrt{\mathsf{Var}(X)}$$

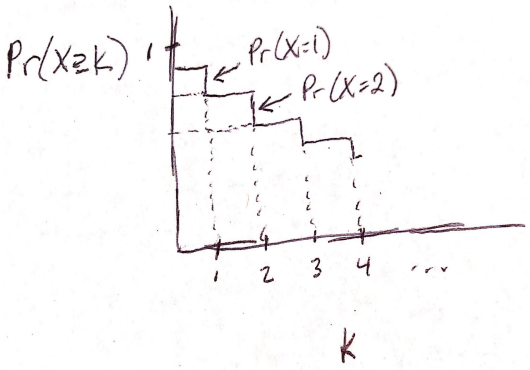
Exercise 0.30. From the definition of variance derive that

$$\mathsf{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Theorem 0.31 (Discrete Tail Sum Formula)**.** *The (discrete) tail sum formula that we know and love for* ***positive valued*** *random variables is*

$$E[X] = \sum_x xP(X = x) = \sum_{k=1}^{\infty} P(X \geq k)$$

There is a derivation that just does some tricks with algebra inside summations, but first I will give a hopefully more intuitive picture of what is going on here.



Consider the above picture. The regular formula for expectation, $\mathbb{E}[X] = \sum_{k=1}^{\infty} kPr(X = k)$, is equivalent to calculating the area of the above graph horizontally, while the tail sum formula $\sum_{k=1}^{\infty} Pr(X \geq k)$, is equivalent to calculating the area of the above graph vertically.

Proof. Now we give the (less intuitive) algebraic proof:

$$E[X] := \sum_{x=1}^{\infty} xPr(X = x)$$

notice here that $xPr(X = x) = \sum_{k=1}^{x} Pr(X = x)$, so then we have:

$$E[X] = \sum_{x=1}^{\infty} \sum_{k=1}^{x} Pr(X = x)$$

$$= \sum_{k=1}^{\infty} \sum_{x=k}^{\infty} Pr(X = x)$$

$$= \sum_{k=1}^{\infty} Pr(X \geq k)$$

□

Exercise 0.32. Convince yourself that $\sum_{x=1}^{\infty}\sum_{k=1}^{x} Pr(X = x) = \sum_{k=1}^{\infty}\sum_{x=k}^{\infty} Pr(X = x)$. It may help to draw a graph of an arbitrary distribution, with $Pr(X = x)$ as the y-axis and x as the x-axis.

0.7 Some Popular Discrete Random Variables

Definition 0.33 (Discrete Uniform RV)**.** The discrete uniform distribution over $[n] = \{1, ..., n\}$ has PMF:

$$P_X(k) = \frac{1}{n}, \forall k \in [n]$$

We can easily see that for uniform X, we have $\mathbb{E}[X] = \frac{n+1}{2}$.

Definition 0.34 (Bernoulli ("coin flip") RV)**.** The Bernoulli(p) RV takes on the value 1 with probability p, and 0 with probability $1 - p$. Explicitly:

$$P_X(k) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \end{cases}$$

And we can easily calculate that $\mathbb{E}[X] = p$. We also have

$$\mathsf{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1 - p)$$

Definition 0.35 (Indicator RV)**.** An **indicator RV** of an event A takes on the value of 1 if A happens/is true and 0 otherwise:

$$X = \{1\}_A = \mathbb{1}_A = \begin{cases} 1 & \text{A is true} \\ 0 & \text{else} \end{cases}$$

We can note then that

$$\mathbb{E}[\mathbb{1}_A] = \sum_x x \Pr(X = x) = \Pr(A)$$

Definition 0.36 (Binomial Random Variable)**.** If $X \sim Bin(n, p)$ then we define the PMF (probability mass function) as:

$$P_X(k) = Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

The Binomial distribution is by definition also just the sum of n iid Bernoulli variables with parameter p. $X = \sum_{i=1}^{n} B_i$ where $B_i \sim Ber(p)$

It is not too difficult to calculate the expecation and variance of a binomial random variable, precisely because it can be represented as the sum of n iid Bernoullis. We have that we can use linearity of expectations to calculate the expectation of $X \sim Bin(n, p)$. We have

$$\mathbb{E}[X] = \mathbb{E}[\sum_{i=1}^{n} B_i] = \sum_{i=1}^{n} \mathbb{E}[B_i] = \sum_{i=1}^{n} p = np$$

Example 0.37. Let $Y = aX + b$. Then we have that

$$\begin{aligned}\mathsf{Var}(Y) = \mathsf{Var}(aX+b) &= \mathbb{E}[(aX+b) - \mathbb{E}[aX+b]] \\ &= \mathbb{E}[((aX+b) - (a\mathbb{E}[X]+b))^2] \\ &= \mathbb{E}[(aX - a\mathbb{E}[X])^2] \\ &= a^2\mathbb{E}[(X-\mathbb{E}[X])^2] \\ &= a^2\mathsf{Var}(X)\end{aligned}$$

Note that adding a constant does not affect the variance, nor should it intuitively, as we are simply shifting where the variable occurs and not affecting the spread of the variable at all. Multiplying by a constant, however, should and does affect the spread and therefore the variance of an RV.

Definition 0.38 (Geometric Random Variable)**.** A geometric random variable counts the time until the first success. We have that the PMF of a geomtric random variable with parameter p (the probability of success is p) is as follows:

$$Pr(X = k) = (1-p)^{k-1}p$$

The above formula makes sense because we need the first $k-1$ events to be failures, which happen with probability $1-p$, and then we need the k^{th} event to be a success, which happens with probability p. Also intuitively, if we want to calculate the probability $Pr(X > k)$, then we need the first k events to all be failures, and it does not matter at all what happens after that. Therefore, $Pr(X > k) = (1-p)^k$, which tells us that the CDF of a geometric RV is $Pr(X \leq k) = 1 - Pr(X > k) = 1 - (1-p)^k$. As a sanity check, we can differentiate the CDF and find that it does indeed equal the PDF.

We can also calculate more easily the expectation of geometric random variable using the tail sum formula. We have that

$$E[X] = \sum_{i=1}^{\infty} P(X \geq k) = \sum_{i=1}^{\infty}(1-p)^{k-1} = \frac{1}{p}$$

Where the last step follows from the formula for infinite geometric series.

Lecture 4: (Co)variance, Correlation, Conditional / Iterated Expectation, Law of Total Variance

Agenda

1. Recap of expectation, their properties, and popular RVs
2. Memoryless property of Geometric(p) RVs
3. Conditional RV and Iterated Expectation
4. Covariance

0.8 Geometric RV and Properties, Poisson RV

Example 0.39 (Coupon Collector Problem)**.** Imagine we have N balls of different colors, and we sample with replacement. What is the expected number of trials before we see all of the colors? To address this problem, we start by defining a few variables. Let C_r be the number of samplings required until we see at least r distinct colors. Then we know that $C_1 = 1$ and is in fact not at all random. We further define X_i as the number of samplings required to see i distinct colors given that we have already seen $i-1$ colors. We note here also that each X_i is a geometric random variable with parameter (probability of success) $p = \frac{N-i+1}{N}$ We also note that we have

$$C_N = \sum_{i=1}^{N} X_i$$

Then,

$$\mathbb{E}[C_N] = \sum_{i=1}^{N} \mathbb{E}[X_i] = 1 + \frac{N}{N-1} + \frac{N}{N-2} + ... + N \approx N \log N$$

Definition 0.40 (Poisson Random Variable)**.** We define $X \sim Pois(\lambda)$ with the follwing PMF:

$$Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

In general, the parameter λ describes a *rate*, i.e. the number of customers entering the store in a hour. We can calculate for $X \sim Pois(\lambda)$ the expectation:

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!}$$

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

Exercise 0.41. Prove that for $X \sim Pois(\lambda)$:

$$\mathsf{Var}(X) = \lambda$$

Exercise 0.42 (Poisson Merging)**.** Prove that for $X \sim Pois(\lambda)$ and $Y \sim Pois(\mu)$ we have $X + Y \sim Pois(\lambda + \mu)$ (this is done in discussion)

Exercise 0.43 (Poisson splitting)**.** Prove that if we "split" $X \sim Pois(\lambda)$ into two paths, by having an arrival take one path with probability p and the other with probability $1-p$. Prove that the number of arrivals to the first path Y is $Pois(p\lambda)$ and is moreover independent of the number of arrivals to the second path Z, which is distributed according to $Pois((1-p)\lambda)$. (this is a homework problem)

We now explore the relationship between the binomial distribution and poisson distribution. The poisson distribution actually turns out to be a limit of the binomial distribution, as we let n get large and p go to zero. Specifically, we must have that $\lim_{n\to\infty} np_n = \lambda$. Consider letting $p = \frac{\lambda}{n}$. Then we get the PMF for the binomial becomes:

$$Pr(X=k) = \binom{n}{k}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k}$$

$$= \frac{n(n-1)...(n-k+1)}{n^k}\frac{\lambda^k}{k!}(1-\frac{\lambda}{n})^{n-k}$$

$$= \frac{n}{n}\frac{n-1}{n}...\frac{n-k+1}{n}\frac{\lambda^k}{k!}(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^k$$

Now, as we let $n \to \infty$, we can see that the first k left terms go to 1, as well as the rightmost term. We also know that the second to rightmost term approaches $e^{-\lambda}$. This leaves us with the pdf for the poisson distribution: $Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$.

Example 0.44 (St. Petersburg Paradox)**.** I keep tossing a fair coin until I get heads. If this takes n tosses, then I get 2^n dollars. How much should I pay to play this game? Well, if W is the amount I win, we can calculate:

$$\mathbb{E}[W] = \sum_{k=0}^{\infty} 2^k \frac{1}{2^k} = 2(\frac{1}{2} + 4\frac{1}{4} + 8\frac{1}{8} + ...$$

$$= 1+1+1+.... = \infty$$

So I should pay an unbounded amount to play this game? Bernoulli said we should actually calculate $\log U$ where U is our utility/payout, in which case we would only pay \$4 to play this game.

Lemma 0.45. *If X and Y are independent,*

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$$

Proof.

$$\mathbb{E}[XY] = \sum_x \sum_y xyP_{XY}(x,y)$$

$$= \sum_x \sum_y xyP_X(x)P_Y(y) = \sum_x xP_X(x)\sum_y yP_Y(y) = \mathbb{E}[X]\,\mathbb{E}[Y]$$

□

Remark 0.46. The reverse is generally **not** true:

$$\mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[XY] \not\Rightarrow \text{X is independent of Y}$$

Lemma 0.47. *If X and Y are independent,*

$$\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$$

Proof. WLOG, we can say $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, since variance is not affected by the shifting of a constant and therefore subtracting out the means does not alter the variance. Then if we let $Z = X + Y$, we have

$$\mathsf{Var}(Z) = \mathsf{Var}(X+Y) = \mathbb{E}[(X+Y)^2] = \mathbb{E}[X^2 + Y^2 + 2XY]$$
$$= \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\,\mathbb{E}[X]\,\mathbb{E}[Y] = \mathsf{Var}(X) + \mathsf{Var}(Y)$$

□

The above two lemmas of course generalize to more than just two independent variables via induction. We can also use the above lemma to calculate the variance of a binomial very easily, since a binomial $X \sim Bin(n,p)$ is equal to $B_1 + ... + B_n$, where each $B_i \sim Bern(p)$. Then, we have that (since the B_i are iid)

$$\mathsf{Var}(X) = \mathsf{Var}(\sum_{i=1}^{n} B_i) = \sum_{i=1}^{n} \mathsf{Var}(B_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

The above lemma raises the question, what if X_1 and X_2 are not independent, but we would like to calculate $\mathsf{Var}(X_1 + X_2)$?

Definition 0.48 (Covariance)**.** Consider $\mathsf{Var}(X+Y) = \mathbb{E}[(X+Y-\mathbb{E}[X]-\mathbb{E}[Y])^2]$. Now let $\hat{X} = X - \mathbb{E}[X]$ and $\hat{Y} = Y - \mathbb{E}[Y]$. Then

$$\mathsf{Var}(X+Y) = \mathbb{E}[(\hat{X}+\hat{Y})^2]$$
$$= \mathbb{E}[\hat{X}^2] + \mathbb{E}[\hat{Y}^2] + 2\mathbb{E}[\hat{X}\hat{Y}]$$

This last term, $\mathbb{E}[\hat{X}\hat{Y}]$, is called the **covariance** of X and Y, and it tells us how they change with each other. We have that

$$\mathsf{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$
$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Intuitively, the covariance between two variables is related to how they affect each other. If X_1 increasing causes X_2 to generally increase, then the covariance will be positive. If X_1 increasing causes X_2 to generally decrease, then the covariance will be negative.

Definition 0.49 (Correlation Coefficient)**.**

$$\rho(X,Y) = \frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}(X)\mathsf{Var}(Y)}}$$

The above is known as the correlation coefficient of two variables, and is always between -1 and 1. This can be proved using the Cauchy-Schwarz Inequality (try it!)

0.9 Conditioning of RVs

When we consider $X|Y$, the first thing to note is that this is just another random variable, with its own PMF $P_{X|Y}(x|y) = \Pr(X = x|Y + y)$. Furthermore, we must still have that

$$\sum_x P_{X|Y}(x|y) = 1$$

Lemma 0.50 (Memorylessness of Geometric RVs). *We have that for geometric RV X*

$$Pr(X = k + m|X > k) = Pr(X = m)$$

Proof.

$$Pr(X = k + m|X > k) = Pr(X = k + m \cap X > k)/Pr(X > k)$$
$$= \frac{Pr(X = k + m)}{Pr(X > k)} = \frac{(1-p)^{k+m-1}p}{(1-p)^k}$$
$$= (1-p)^{m-1}p = Pr(X = m)$$

□

We can use a clever conditioning trick, along with the memorylessness property, to calculate the variance of a geometric random variable. First, we need $\mathbb{E}[X^2]$. We have by total probability:

$$\mathbb{E}[X^2] = \mathbb{E}[X^2|X = 1]Pr(X = 1) + \mathbb{E}[X^2|X > 1]Pr(X > 1)$$
$$= p + (1-p)\mathbb{E}[(1 + X)^2]$$

Where $\mathbb{E}[X^2|X > 1] = \mathbb{E}[(1+X)^2]$ follows from the memorylessness property (convince yourself this is true). Then,

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

Then we have that

$$\mathsf{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

Example 0.51 (Romance is dead). $2m$ people form couples. 50 years from now, the probability that any person is alive is p. Now suppose that there are A people alive after 50 years. Let S be the number of couples for which both people are still alive. We would like to find $\mathbb{E}[S|A = a]$. In order to do this, we further define X_i as the indicator that the first person of couple i survives, and Y_i as the indicator that the second person of couple i survives. Then $S = \sum_i X_i Y_i$. Then, we have

$$\mathbb{E}[S|A = a] = \mathbb{E}[\sum_i X_i Y_i|A = a] = \sum_i \mathbb{E}[X_i Y_i|A = a]$$
$$= m\mathbb{E}[X_i Y_i|A = a] = mPr(X_i Y_i = 1|A = a)$$
$$= m\frac{a}{2m}\frac{a-1}{2m-1}$$
$$= m\frac{\binom{2m-2}{a-2}}{\binom{2m}{a}}$$

Why have we included the last equality, rather than simplifying further? Because it lends itself to an alternate interpretation of the solution. Consider couple i. What is the probability that they survive, given that $A = a$? Well $\binom{2m-2}{a-2}$ is the number of ways for a people to survive including this specific couple, and $\binom{2m}{a}$ is the number of ways for a people to survive in general. More formally, we have:

$$Pr(X_i Y_i = 1|A = a) = \frac{Pr(A = a|X_i Y_i = 1)Pr(X_i Y_i = 1)}{Pr(A = a)}$$

Now, A is a $Bin(2m, p)$ and $A|X_iY_i = 1$ is a $Bin(2m-2, p)$. So we have:

$$= \frac{\binom{2m-2}{a-2}p^{a-2}(1-p)^{2m-a}p^2}{\binom{2m}{a}p^a(1-p)^{2m-a}}$$

$$= \frac{\binom{2m-2}{a-2}}{\binom{2m}{a}}$$

And all we have to do is use linearity of expectations (and multiply by m) to get the same answer as above.

Lecture 5: Iterated Expectation, Continuous Probability, Uniform, Exponential Distributions

Agenda

1. Law of Iterated Expectations
2. Continuous probability (CDF, Uniform, Exp)

0.10 Iterated Expectation

Recall how conditional expectation works:

$$\mathbb{E}[X|Y=y] = \sum_x x \Pr(X=x|Y=y)$$

We say that $\mathbb{E}[X|Y=y]$ is the "expectation of X w.r.t. the distributions of X *conditioned on* $Y=y$, and it is really just a number.

Definition 0.52. Let X and Y be RVs. Then $\mathbb{E}[X|Y]$ is also a RV, the conditional expectation of X given Y, which has the value $\mathbb{E}[X|Y=y]$ with probability $\Pr(Y=y)$. It is important but subtle to note that $\mathbb{E}[X|Y]$ is a RV itself.

Example 0.53. Suppose we roll a die N times. Let X be the sum of the die rolls. Then we have that

$$\mathbb{E}[X|N=1] = \frac{7}{2}$$
$$\mathbb{E}[X|N=2] = 7$$

and in general, $\mathbb{E}[X|N=n] = \frac{7n}{2}$, and in general:

$$\mathbb{E}[X|N] = \frac{7N}{2}$$

The difference here is subtle, but the last equality is actually a much stronger statement, as it equates random variables rather than just numbers.

Out of this comes a natural question: since $\mathbb{E}[X|Y]$ is a RV, what is its expectation?

Theorem 0.54 (Iterated Expectations/Tower Rule)**.**

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$$

Proof.

$$\mathbb{E}[\mathbb{E}[X|Y]] = \sum_y \mathbb{E}[X|Y=y] \Pr(Y=y)$$
$$= \sum_y \sum_x x \Pr(X=x|Y=y) \Pr(Y=y)$$
$$= \sum_x x \sum_y \Pr(X=x, Y=y)$$
$$= \sum_x x \Pr(X=x) = \mathbb{E}[X]$$

□

Example 0.55. We roll a die N times where $N \sim Geom(p)$. As before, X represents the sum of the N die rolls. Then we have:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|N]] = \mathbb{E}[\frac{7N}{2}] = \frac{7}{2}\mathbb{E}[X] = \frac{7}{2p}$$

Example 0.56 (Drunken walk on a line)**.** Suppose we take a random walk, starting at the origin, on a discretized line. Then if X_{n+1} is our location at time $n+1$, then we have the recurrence:

$$X_{n+1} = X_n + \mathbb{1}_+ - \mathbb{1}_-$$

where naturally $\mathbb{1}_+$ is an indicator for drunk taking a $+1$ step, and likewise $\mathbb{1}_-$ is an indicator for drunk taking a -1 step. Then we have:

$$\mathbb{E}[X_{n+1}] = \mathbb{E}[X_n] + 1/2 - 1/2 = \mathbb{E}[X_n] = 0$$

But what about the variance of the walk? $\mathbb{E}[X_n^2] = ?$. We have

$$\Pr(X_{n+1}^2 = (k+1)^2 | X_n = k) = \Pr(X_{n+1}^2 = (k-1)^2 | X_n = k) = 1/2$$

So then we have:

$$\mathbb{E}[X_{n+1}^2 | X_n = k] = \frac{(k+1)^2 + (k-1)^2}{2} = k^2 + 1$$

$$\Rightarrow \mathbb{E}[X_{n+1}^2 | X_n] = X_n^2 + 1$$

Then we can calculate

$$\begin{aligned}\mathbb{E}[X_{n+1}^2] &= \mathbb{E}[\mathbb{E}[X_{n+1}^2 | X_n]] = \mathbb{E}[X_n^2] + 1 \\ &= \mathbb{E}[X_{n-1}^2 + 1 + 1\end{aligned}$$

and then, after noting $\mathbb{E}[X_0^2] = 0$, we can see that

$$\mathsf{Var}(X_n) = \mathbb{E}[X_n^2] = n$$

0.11 Continuous Probability

Continuous RVs is a concept you should be relatively familiar with from CS70, but we will go over it quickly again and there are some subtleties to make sure are clear.

In most settings, a **continuous sample space** is more natural than a discrete one (such as distance, time, temperature, etc). For a continuous RV, there is no such thing as $\Pr(X = x)$. Well there is, but it's just equal to zero and generally pretty meaningless. We need to instead define probability over sets that have "length" and quantify "allowable events". What "allowable events" refers to here gets more into measure theory, which we are not going to get into in this course, as in virtually all engineering applications the distinction is unimportant. So rather than talking about $\Pr(X = x)$, we instead talk about f_X, which is the **probability density function (PDF)** of a continuous random variable.

Definition 0.57. X is a **continuous RV** if

1. $\exists$ a non-negative function f_X s.t.

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

is well-defined.

2. for every interval $B \in \mathbb{R}$,

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

Remark 0.58. $\Pr(X = a) = 0$, which means that $\Pr(X < a) = \Pr(X \leq a)$, which means for the rest of these notes I will be lazy and interchange $<$ and $\leq$ for continuous RVs at will.

This density has the property that if we wish to calculate the probability that our random variable falls in a small δ sized interval, we have

$$Pr(X \in [x, x + \delta]) = \int_{x}^{x+\delta} f_X(t)dt \approx f_X(x)\delta$$

for small enough delta, of course. Then we have

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

Hence the name "density function". Note that it is perfectly fine for the density to be greater than 1 at any particular point, as it is not a probability. We have only the requirement that the **integral of f_X over its domain must be equal to 1** (think about why this must be), and that the **density must be nonnegative**. Another useful interpretation may be to think of PDF values at certain points as relative likelihoods; that is, if $f_X(s) = 2f_X(t)$, then we are twice as likely to see values in a small δ neighborhood around s than values in a small δ neighborhood around t (if the density is continuous).

Example 0.59. let $f_X(x) = \frac{1}{2\sqrt{x}}$ for $0 < x < 1$ and take on the value 0 otherwise. Then we have that it is nonnegative, and that

$$\int_{0}^{1} f_X(x)dx = 1$$

So this is a valid PDF.

Now we mention the **cumulative distribution function**, which completely analogously to the discrete case is simply $\Pr(X < x)$. Since the CDF would be

$$F(x) = \int_{-\infty}^{x} f_X(t)dt$$

The CDF has the following properties:

1. $F_X(\infty) = 1$
2. $F_X(-\infty) = 0$
3. if X is discrete, then

$$\Pr(X = k) = F_X(k) - F_X(k - 1)$$

and in the continuous case:

$$f_X(x) = \frac{d}{dx}F_X(x)$$

Where the last fact follows from the fundamental theorem of calculus (if F is differentiable). This can be a very useful fact, as often the CDF is easier to calculate than the PDF.

Example 0.60. Imagine throwing darts at a unit circle. We model this by saying that the location of where the dart lands in the circle is completely random (i.e. uniform over the circle). We would

like to find the CDF and PDF of Y, which is the distance from the origin of where the dart lands. We have that

$$\Pr(Y \leq y) = \frac{\text{area of circle of radius y}}{\text{area of whole circle}}$$
$$= \frac{\pi y^2}{\pi} = y^2$$

. Then we have that simply

$$f_Y(y) = \frac{d}{dy} F_Y(y) = 2y$$

We can calculate that:

$$P(0.5 < Y < 0.6) = F_Y(0.6) - F_Y(0.5) = 0.36 - 0.25 = 0.11$$

.

We have some analogous definitions and lemmas that pretty much follow from the discrete case:

Definition 0.61. The **expectation** of a continuous RV X is

$$\int_{-\infty}^{\infty} x f_X(x) dx$$

Lemma 0.62.

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Lemma 0.63. *if X, Y are independent, then*

$$F_{X,Y}(x, y) = F_X(x) F_Y(y)$$

Now we go over some popular continuous RVs.

Definition 0.64 (Uniform RV)**.** If $X \sim Unif[a, b]$, then it must have constant probability density between a and b and zero density everywhere else, which tells us that

$$f_X(x) = \frac{1}{b-a}$$

for $x \in [a, b]$.

We can calculate for $X \sim U[a, b]$ that:

$$\mathbb{E}[X] = \int_a^b x \frac{1}{b-a} dx = \frac{a+b}{2}$$

and also

$$\mathsf{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{(b-a)^2}{12}$$

Exercise 0.65. verify the Variance of a uniform RV between a and b is actually what we claimed above.

Definition 0.66 (Exponential RV)**.** Lets say we wanted to find a continuous RV that had the same "memoryless" property as the discrete Geometric RV, and analogously measured "time to success" (or failure, however you want to look at it). But now this time is a continuous thing, say the amount of time before a lightbulb burns out. Specifically, for the memoryless property, we want $Pr(X > t + s|X > s) = Pr(X > t)$. That is, we want

$$\frac{Pr(X > t + s \cap X > s)}{Pr(X > s)} = \frac{Pr(X > t + s)}{Pr(X > s)} = Pr(X > t)$$

The question then becomes, what function $g(t) = \Pr(X > t)$ satisfies $\frac{g(s+t)}{g(s)} = g(t)$? Well, eventually, we might notice that $g(x) = e^x$ works! The problem is, this increases $g(x)$ as x increases, which is not the behavior we want if we are to keep the analogy. Well, $g(x) = e^{-x}$ also works, and it is monotonically decreasing, so that is better! In fact, we can even throw in a constant $g(x) = e^{-\lambda x}$, for increased versatility, and it still is monotonically decreasing and memoryless. Then we have

$$F_X(x) = 1 - Pr(X > x) = 1 - e^{-\lambda x}$$

$$\Rightarrow \frac{d}{dx} F_X(x) = f_X(x) = \lambda e^{-\lambda x}$$

for any $\lambda > 0$. We can further check that this integrates to 1 over its domain (since it is measuring time to success, this is a positive random variable):

$$\int_0^\infty f_X(x) = \lambda \int_0^\infty e^{-\lambda x} = 1$$

as desired. And with that I conclude the most long winded introduction to the exponential random variable that has ever been.

Exercise 0.67. Show that if $X \sim Exp(\lambda)$, then

$$\mathbb{E}[X] = \frac{1}{\lambda}$$

$$\mathsf{Var}(X) = \frac{1}{\lambda^2}$$

Definition 0.68 (Laplace Distribution)**.** Let $Z = X - Y$, where $X, Y \sim exp(\lambda)$, and X and Y are independent. Then how is Z distributed? Well, if $X > Y$, then by the memoryless property we have that Z is simply an exponential RV. This happens with probability $1/2$, so we have $f_Z(z) = \frac{1}{2}\lambda e^{-\lambda z}$. What if then $Y > X$? Then once again by the memoryless property, we get that Z is simply a negated exponential RV: $f_Z(z) = \frac{1}{2}\lambda e^{+\lambda z}$. So putting this together we have what is known as the **Laplace Distribution:**

$$f_Z(z) = \frac{1}{2}\lambda e^{-\lambda|z|}$$

Lecture 6: Normal Distribution, Continuous Analogs, Derived Distributions

Agenda: see above

0.12 Review

Exercise 0.69. Let R be the distance from the origin of a point randomly sampled on a unit ball (in $\mathbb{R}^3$).

1. what is the CDF of R?
2. PDF?
3. Expectation?

0.13 Normal Distribution

Definition 0.70 (Normal Distribution)**.** Let $X \sim \mathcal{N}(\mu, \sigma^2)$, where μ is the mean of the distribution and σ is the standard deviation. Here is the PDF:

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

We call the pdf of $X \sim \mathcal{N}(0,1)$ is $F_X(x) = \Phi(x)$, which cannot be expressed in elementary functions

The PDF of the normal is clearly positive. We would like to also show that it integrates to 1:

Proof. We will show this when $\mu = 0$ and $\sigma^2 = 1$. The idea is to show that

$$\left(\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x\right)^2 = 1$$

We have that:

$$\begin{aligned}\left(\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x\right)^2 &= \left(\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x\right)\left(\int_{-\infty}^{\infty} f_Y(y)\mathrm{d}y\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y \\ &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-(x^2+y^2)/2}\mathrm{d}x\mathrm{d}y\end{aligned}$$

Now the trick here is to use polar integration, recalled that $\mathrm{d}y\mathrm{d}x = r\mathrm{d}r\mathrm{d}\theta$. Then we have

$$\int_0^{2\pi}\int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-(r^2)/2} r dr d\theta$$

$$= \int_{-\infty}^{\infty} e^{-(r^2)/2} r dr$$

We can use u substitution to solve this integral, which will evaluate to 1 (think about the pdf of the exponential RV! Or just do it manually).

□

Some properties of Normal distributions:

1. if X, Y are independent normals, then $Z = X + Y$ is also normal $Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.
2. The sum of two dependent normals isn't always Normal. Consider $X \sim \mathcal{N}(0, 1)$, and $Y = X$ w.p. $1/2$ and $-X$ wp $1/2$. Then both X and Y are normal but $X + Y$ is not normal.
3. We have that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then
$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Example 0.71. Let $X \sim \mathcal{N}(2, 16)$. We wish to find $Pr(-2 < X < 6)$. We have

$$Pr(-2 < X < 6) = Pr(-4 < X - 2 < 4) = Pr(-1 < \frac{X - 2}{4} < 1)$$

$$= Pr(-1 < \mathcal{N}(0, 1) < 1) = \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 \approx 0.68$$

Where Φ is the CDF of the standard normal distribution.

Exercise 0.72. convince yourself that $\Phi(1) - \Phi(-1) = 2\Phi(1) - 1$ if you haven't already.

Example 0.73. Suppose male height is distributed as $\mathcal{N}(70, 5)$ and female height is $\mathcal{N}(64, 4)$. What's the probability that a random chosen male is taller than a randomly chosen female? Express your answer in terms of Φ.

Ans: Let X be the boys height and Y the girls height. We want to calculate $P(X - Y > 0) = P(Y - X < 0)$. Note that $Y - X \sim \mathcal{N}(-6, 9)$, and so

$$\frac{Y - X + 6}{\sqrt{9}} \sim \mathcal{N}(0, 1)$$

So we have

$$P(Y - X < 0) = P(\frac{Y - X + 6}{3} < 2) = \Phi(2)$$

0.14 Continuous Analogs of Discrete RVs

For **joint distributions**, we can generalize from the discrete case:

$$P(A) = \sum_{(x,y) \in A} P_{X,Y}(x, y)$$

and analogously:

$$P(A) = \int_A f_{X,Y}(x, y) \mathrm{d}x\mathrm{d}y$$

The definitions of marginal probabilities, conditional probabilities, multiplication rule, and Bayes Rule all carry over naturally into the domain of continuous probability, all you need to do is replace summations with integrals and p_X's with f_X's.

Example 0.74. We can have discrete and continuous RVs defined jointly. For example. For example,

let X be the outcome of a die roll, and $Y \sim Exp(X)$. Then we have

$$p_X(x) = \frac{1}{6}$$

and

$$f_{Y|X}(y|x) = xe^{-xy}$$

Example 0.75. Let $X \sim Bern(1/2)$ and $Y = 2X$. We have that the distribution of Y is

$$\Pr(Y = y) = \Pr(2X = y) = \Pr(X = \frac{y}{2})$$

more generally, if X is discrete RV, and $Y = f(X)$, then

$$\Pr(Y = y) = \Pr(f(X) = y) = \Pr(X \in f^{-1}(y))$$

Be careful! Is it then true that in the continuous case if $X \sim U[0, 1]$ and $Y = 2X$. Is it then true that

$$f_Y(y) = \Pr(Y = y) = \Pr(2X = y) = \Pr(X = y/2) = f_X(\frac{y}{2})$$

NO. There are many things wrong here, first of all the quantity $\Pr(Y = y) = 0, \forall y$. Second, this does not integrate to 1:

$$\int_0^2 f_Y(y)\mathrm{d}y = \int_0^2 f_X(y/2)\mathrm{d}y = 2$$

Instead, we have to derive the CDF of Y properly using the CDF. It IS true that:

$$F_Y(y) = \Pr(Y \leq y) = \Pr(2X \leq y) = \Pr(X \leq \frac{y}{2}) = F_X(\frac{y}{2}$$

and then

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X(y/2) = \frac{1}{2} f_X(\frac{y}{2})$$

Which we can check does integrate to 1.

Example 0.76 (More on the relationship between Exponential and Geometric RVs)**.** Toss a coin every δ seconds, and let the probability of heads $p = 1 - e^{-\lambda\delta}$, with $\delta << 1$. Let $N \sim Geom(p)$ and $X \sim exp(\lambda)$. Then we have that $F_N(n) = Pr(N < n) = 1 - e^{-\lambda n\delta} = F_X(n\delta)$. If you graph $F_N(n)$ and $F_X(n\delta)$, then you can see how the exponential is the limit of the geometric as $\delta \to 0$.

TODO 3. There's a lot of graphs the rest of this lecture that would take a lot of time to latex.

It is useful to know that the Covariance is a multilinear function, meaning

$$\mathsf{Cov}(X + Y, W + Z) = \mathsf{Cov}(X, W) + \mathsf{Cov}(X, Z) + \mathsf{Cov}(Y, W) + \mathsf{Cov}(Y, Z)$$

And it is also useful to note that the variance $\mathsf{Var}(X) = \mathsf{Cov}(X, X)$. We also have that $\mathsf{Cov}(aX + b, Y) = a\mathsf{Cov}(X, Y)$. This yields the following useful identity:

$$\mathsf{Var}\left(\sum X_i\right) = \sum_i \mathsf{Var}(X_i) + \sum_i \sum_{j \neq i} \mathsf{Cov}(X_i, X_j)$$

The **Tower Rule** or **Iterated Expectation** or the **Law of Total Expectation** also holds in the continuous case. We have:

$$\mathbb{E}_Y[\mathbb{E}_X[X|Y]] = \int_Y f_Y(y) \int_X x f_{X|Y}(x|y)\mathrm{d}x\mathrm{d}y$$

$$= \int_Y \int_X x f_{X|Y}(x|y) f_Y(y) \mathrm{d}x \mathrm{d}y = \int_X \int_Y x f_{X,Y}(x,y) \mathrm{d}y \mathrm{d}x$$
$$= \int_X x f_X(x) \mathrm{d}x = \mathbb{E}[X]$$

The above result should make intuitive sense when you think about it, and the intuition is quite similar to the intuition behind discrete total probability. If we want to find $\mathbb{E}[X]$, and the instances of Y subdivide our probability space, it may be easier to calculate $\mathbb{E}[X|Y]$ for every Y. But then we have to weight each expectation the probability that particular instance of Y happens, hence the outside expectation over the Y variable.

Example 0.77. Consider trying to estimate X given some information Y with the estimate $\mathbb{E}[X|Y]$. Well we have that the error E is $E = X - \mathbb{E}[X|Y]$, and we further have that

$$\mathbb{E}[E] = \mathbb{E}[X] - \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X] - \mathbb{E}[X] = 0$$

and therefore $\mathbb{E}[X|Y]$ is called an *unbiased estimator*. We will learn more about this later in the semester though when we talk about MMSE.

Lecture 7: Order Statistics, Convolution, Moment Generating Functions

Agenda:

1. Law of Total Variance
2. Order Statistics
3. Convolution
4. Moment Generating Functions

0.15 Conditional Variance and Law of Total Variance

Definition 0.78 (Conditional Variance)**.** Let X, Y be RVs. We can define the conditional variance $\mathsf{Var}(X|Y=y)$ as the variance of the conditional distribution $P(X=x|Y=y)$.

Remark 0.79. $\mathsf{Var}(X|Y)$ is a RV that assumes the value $\mathsf{Var}(X|Y=y)$ with probability $\Pr(Y=y)$.

Lemma 0.80 (Total Variance)**.** *We have that*

$$\mathsf{Var}(X) = \mathbb{E}[\mathsf{Var}(X|Y)] + \mathsf{Var}(\mathbb{E}[X|Y])$$

I will now try to offer some sort of intuition before the formal proof. We want to answer the question: how much does X vary? Well, if we fix Y, we could take the expectation over all the $y \in Y$ of $\mathsf{Var}(X|Y)$. But even if we are fixing Y, there is still some variance in X, and therefore some variance in $\mathbb{E}[X|Y]$, which is where the second term comes into play. The first term is the expected variance from the mean of $X|Y$; the second is the variance of that mean.

Proof. We have that

$$\begin{aligned}
\mathsf{Var}(X) &= \mathbb{E}[X^2] + \mathbb{E}[X]^2 \\
&= \mathbb{E}[\mathbb{E}[X^2|Y]] - (\mathbb{E}[\mathbb{E}[X|Y]])^2 \\
&= \mathbb{E}[\mathsf{Var}(X|Y) + \mathbb{E}[X|Y]^2] - (\mathbb{E}[\mathbb{E}[X|Y]])^2 \\
&= \mathbb{E}[\mathsf{Var}(X|Y)] + \left(\mathbb{E}[\mathbb{E}[X|Y]^2] - (\mathbb{E}[\mathbb{E}[X|Y]])^2\right) \\
&= \mathbb{E}[\mathsf{Var}(X|Y)] + \mathsf{Var}(\mathbb{E}[X|Y])
\end{aligned}$$

□

Remark 0.81 (Geometric interpretation of the Law of Total Variance)**.** *Read only if interested. If it doesn't make a lot of sense yet, don't worry too much about it, but maybe come back to it after you've learned about Kalman Filters!*
First, we perform some manipulation that will be useful later:

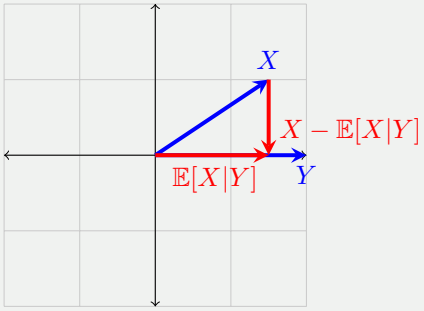
$$\begin{aligned}
\mathbb{E}[\mathsf{Var}(X|Y)] &= \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[[X|Y])^2|Y]] \\
&= \mathbb{E}[(X - \mathbb{E}[X|Y])^2] \\
&= \mathsf{Var}(X - \mathbb{E}[X|Y])
\end{aligned}$$

Later in the semester, you will learn about the geometric representation of RVs. For now, just assume we represent RVs geometrically with their length equal to their standard deviation. We further assume WLOG all the RVs are zero mean (mean doesn't matter when looking at variance).

Then, interestingly, the law of total variance is simply an expression of the pythagorean theorem! Consider:

$$\mathsf{Var}(X) = \mathbb{E}[\mathsf{Var}(X|Y)] + \mathsf{Var}(\mathbb{E}[X|Y]) = \mathsf{Var}(X - \mathbb{E}[X|Y]) + \mathsf{Var}(\mathbb{E}[X|Y])$$

In the geometric representation of RVs, $\mathbb{E}[X|Y]$ is actually a projection of X onto the subspace spanned by Y. This is getting way into stuff that you haven't learned yet, but intuitively this should make sense. In linear algebra, if we want the best estimate of X given Y, then we project X onto the subspace Y. Analogously (again, more details at the end of the semester), $\mathbb{E}[X|Y]$ is the best estimate of X given Y. Then, we have that $\mathsf{Var}(X)$ is the square of the length of X (which is the standard deviation), and $\mathsf{Var}(X - \mathbb{E}[X|Y]) + \mathsf{Var}(\mathbb{E}[X|Y])$ is the sum of the squares of the lengths of the two vectors that add to form X. The below diagram should help:



Example 0.82. We have a biased coin, we toss is n times, and we let X be the number of heads, and $Y \sim U[0,1]$ be the probability of heads (the bias of the coin). First, we have that

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[nY] = n\,\mathbb{E}[Y] = \frac{n}{2}$$

Now, we can calculate the variance:

$$\begin{aligned}\mathsf{Var}(X) &= \mathsf{Var}(\mathbb{E}[X|Y]) + \mathbb{E}[\mathsf{Var}(X|Y)] \\ &= \mathsf{Var}(nY) + \mathbb{E}[nY(1-Y)] \\ &= n^2\mathsf{Var}(Y) + n\,\mathbb{E}[Y] - n\,\mathbb{E}[Y^2] \\ &= \frac{n^2}{12} + \frac{n}{2} - \frac{n}{3} = \frac{n^2}{12} + \frac{n}{6}\end{aligned}$$

Compare this value to tossing a fair coin n times, which has variance $\frac{n}{4}$.

Example 0.83 (Random number of Random Variables)**.** Say we have $Y = X_1 + ... + X_N$, where the X_i are all independent and N is also random. What is $\mathsf{Var}(Y)$? First, we have:

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|N]] = \mathbb{E}[N\mathbb{E}[X_i]] = \mathbb{E}[N]\mathbb{E}[X_i]$$

Where the second to last equality follows from linearity of expectations. Also, since N is given in the inner expectation, we can treat it as a constant (until the outer expectation). We then have:

$$\begin{aligned}\mathsf{Var}(Y) &= \mathbb{E}[\mathsf{Var}(Y|N)] + \mathsf{Var}(\mathbb{E}[Y|N]) \\ &= \mathbb{E}[N\mathsf{Var}(X_i)] + \mathsf{Var}(N\mathbb{E}[X_i]) \\ &= \mathbb{E}[N]\mathsf{Var}(X_i) + \mathbb{E}[X_i]^2\mathsf{Var}(N)\end{aligned}$$

0.16 Order Statistics

Let X be a continuous RV for which $x_1, x_2, ..., x_n$ are values of a random sample of size n. We can then reorder the x_i's from smallest to largest (we don't need to worry about ties, as we are in a continuous sample space here!).

Example 0.84. Suppose $X \sim U[0,1]$, and $n = 4$, and we observe $x_1 = 0.5, x_2 = 0.7, x_3 = 0.2, x_4 = 0.1$. Then we can order them as

$$x^{(1)} = 0.1, \quad x^{(2)} = 0.2, \quad x^{(3)} = 0.5, \quad x^{(4)} = 0.7$$

where $x^{(i)}$ i the i^{th} smallest observation

We call $X^{(i)} = \left(x^{(i)}\right)$ the i^{th} **order statistic**.

Theorem 0.85. *If X has pdf $f_X(x)$, the marginal pdf of the i^{th} order statistic is*

$$f_{X^{(i)}}(y) = \frac{n!}{(i-1)!(n-i)!}(F_X(y))^{i-1}(1 - F_X(y))^{n-i} f_X(y)$$

Proof. We present a sketch of the proof. We would like to calculate

$$\Pr(X^{(i)} \in \{y, y + dy\}) \approx f_{X^{(i)}}(y)dy$$

We need $i - 1$ of the samples to be less than y, which is the $(F_X(y))^{i-1}$ term. We also need exactly one to be right around y, which is approximately $f_X(y)dy$. Finally, we need $(n - i)$ of the samples to be greater than y, which is the $(1 - F_X(y))^{n-i}$ term. Lastly, we have to count how many ways we can pick with ones come first and which one is the the i^{th} largest (which exactly determines which ones come after y), which is $n * \binom{n}{i-1} = \frac{n!}{(i-1)!(n-i)!}$. Combining all of these together yields the exact expression we were looking for! □

Example 0.86 (Special case when X is uniform)**.** Supppose $X \sim U[0,1]$. Recall that $f_X(x) = 1$, and $F_X(x) = x$ (convince yourself if you've forgotten why this is true!). Then we can plug in and see that

$$f_{X^{(i)}}(y) = \frac{n!}{(i-1)!(n-i)!} y^{i-1}(1 - y)^{n-i}$$

for $0 < y < 1$. This is a special case of a *Beta Distribution*.

Exercise 0.87. What is the probability that the 9th smallest of of ten draws from $X \sim U[0,1]$ is greater than 0.8?

0.17 Convolution

Let $Z = X + Y$, where X and Y are both continuous and independent. We would like to calculate the PDF of Z. We can relate Z to X using total probability:

$$f_Z(z) = \int_x f_{X,Z}(x, z)$$

Furthermore, we have

$$F_{Z|X} = \Pr(X + Y \leq z | X = x) = \Pr(Y \leq z - x | X = x)$$

$$= \Pr(Y \leq z - x) = F_Y(z - x)$$
$$\Rightarrow f_{Z|X}(z|x) = f_Y(z - x)$$

Now, incorporating this into our original expression for $f_Z(z)$, we have

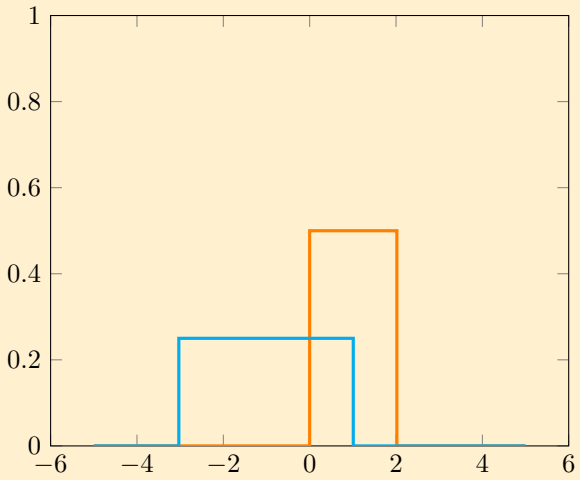
$$f_Z(z) = \int_x f_X(x) f_Y(z - x) dx = (f_X * f_Y)(z)$$

which is called a **convolution.** Intuitively, the expression should make sense, as we are just integrating over all possible combinations of X and Y that could sum to z. The discrete case is entirely analogous:

$$\Pr(Z = z) = \sum_k \Pr(X = k) \Pr(Y = n - k)$$

Example 0.88. Suppose $X, Y \sim U[0, 1]$ are independent. What is $f_Z(z)$, where $Z = X + Y$? We could do an integral and get the right answer via the definition of the convolution, but we can also visually see that it becomes a triangle:

TODO 4. tikz :(



As a general remark, convolution always creates more uncertainty than we started out with. In your homework you will show that if $X, Y \sim \mathcal{N}(0, 1)$ are independent, then $Z = X + Y \sim \mathcal{N}(0, 2)$.

0.18 Moment Generating Functions (MGFs)

Definition 0.89 (Moment Generating Functions)**.** We define the **Moment Generating Function** of an RV X as

$$M_X(s) = \mathbb{E}[e^{sX}]$$

Whats the point of MGFs? It seems like a fairly arbitrary definition. Well, first recall the Taylor series for e:

$$e^{sX} = 1 + sX + \frac{(sX)^2}{2!} + \frac{(sX)^3}{3!} + ...$$
$$\Rightarrow \mathbb{E}[e^{sX}] = 1 + s\mathbb{E}[X] + \frac{s^2}{2!}\mathbb{E}[X^2] + \frac{s^3}{3!}\mathbb{E}[X^3] + ...$$

Then, we can observe that

$$\frac{d}{ds}\mathbb{E}[e^{sX}]\Big|_{s=0} = \mathbb{E}[X]$$

and

$$\frac{d^2}{ds^2}\mathbb{E}[e^{sX}]\Big|_{s=0} = \mathbb{E}[X^2]$$

continuing in this manner, we can see that

$$\frac{d^n}{ds^n}M_X(s)\Big|_{s=0} = \mathbb{E}[X^n]$$

Which is an extremely useful property of the MGF and can help with many computations. We also note that $M_X(0) = 1$ must be true.

Lecture 8: MGFs, Bounds/Concentration Inequalities (Markov, Chebyshev, Chernoff)

Agenda:

1. MGF's (examples and properties)
2. Limit theorems (Markov, Chebyshev, Chernoff)

0.19 Properties of MGFs

Recall that the Moment Generating Function (MGF) of an RV X is the transform:

$$M_X(s) = \mathbb{E}[e^{sX}] = \sum_{k=0}^{\infty} \frac{s^k \, \mathbb{E}[X^k]}{k!} = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx$$

and that

$$\frac{d^n}{ds^n} M_X(s)\Big|_{s=0} = \mathbb{E}[X^n]$$

Some utilities of the MGf:

1. Finding higher moments often becomes easier (derivatives are usually easier than integrals!)
2. Convolution becomes multiplication in the MGF domain, which is often much easier (again, avoiding integrals)
3. Great analytical tool to prove things (such as the CLT!)

And here are some properties to keep in mind:

1. $M_X(0) = 1$
2. if $X > 0$, then $M_X(-\infty) = 0$
3. if $X < 0$, then $M_X(\infty) = 0$
4. if $Y = aX + b$, we have

$$M_Y(s) = \mathbb{E}[e^{s(aX+b)}] = e^{sb} \, \mathbb{E}[e^{asX}] = e^{sb} M_X(as)$$

Example 0.90 (MGF of exponential RV)**.** Let $X \sim exp(\lambda)$. Then we have that:

$$M_X(s) = \mathbb{E}[e^{sX}] = \int_0^{\infty} e^{sX} \lambda e^{-\lambda x} \mathrm{d}x$$

$$= \lambda \int_0^{\infty} e^{x(s-\lambda)} \mathrm{d}x$$

$$= \lambda \frac{e^{x(s-\lambda)}}{s-\lambda}\Big|_0^{\infty}$$

Which, if $\lambda > s$, equals

$$= \frac{\lambda}{\lambda - s}$$

It is fine here that the MGF is not defined for all s, as we only need for it to be defined around $s = 0$ so that we can take derivatives evaluated at $s = 0$.

We can use the MGF of an exponential to easily calculate moments:

$$\mathbb{E}[X] = M_X'(0) = \left.\frac{y}{(y-s)^2}\right|_{s=0} = \frac{1}{\lambda}$$

and we note that

$$\mathbb{E}[X^k] = \left.\frac{d^k}{ds^k} M_X(s)\right|_{s=0} = \left.\frac{\lambda k!}{(\lambda - s)^{k+1}}\right|_{s=0} = \frac{k!}{\lambda^k}$$

Example 0.91 (MGF of a Poisson)**.** We have that for $X \sim Pois(\lambda)$ that

$$M_X(s) = \sum_{k=0}^{\infty} e^{sk} \Pr(X = k) = \sum_{k=0}^{\infty} e^{sk} \frac{e^{-\lambda}\lambda^k}{k!}$$

$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^s\lambda)k}{k!} = e^{-\lambda} e^{e^s\lambda} = e^{-\lambda+\lambda e^s}$$

Which is valid for all values of s.

Example 0.92 (MGF of Normal RV)**.** Let $X \sim \mathcal{N}(0, 1)$. Then we have

$$\mathbb{E}[e^{sX}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx} e^{-x^2/2} \mathrm{d}x$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx - x^2/2} \mathrm{d}x$$

$$= \frac{e^{s^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x^2/2 - sx + s^2/2)}$$

$$= e^{s^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-s)^2/2}$$

$$= e^{s^2/2}$$

Where in the third line we have multiplied by a clever choice of 1 in order to complete the square in the exponent, and in the last line I have used that fact that $\frac{1}{\sqrt{2\pi}} e^{-(x-s)^2/2}$ is the PDF of a standard normal that has been shifted by s, and so must integrate to 1.

Now, if $Y \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = \sigma X + \mu$ and we have

$$\mathbb{E}[e^{sY}] = \mathbb{E}[e^{s(\sigma X + \mu)}]$$

$$= e^{\mu s} \mathbb{E}[e^{\sigma Y s}] = e^{\mu s + \sigma^2 s^2/2}$$

Remark 0.93. An interesting and useful fact (that we will not prove in this course) is that a given MGF corresponds to a unique CDF. This is related to the fact that $M_X(s)$ is just a Laplace transform of $f_X(x)$. Inversions are usually performed by just pattern matching.

Remark 0.94 (Convolving densities corresponds to multiplying their transforms)**.** Take $Z = X + Y$, and assume X and Y are independent. Then we have that

$$M_Z(s) = \mathbb{E}[e^{sZ}] = \mathbb{E}[e^{s(X+Y)}] = \mathbb{E}[e^{sX}]\mathbb{E}[e^{sY}] = M_X(s)M_Y(s)$$

Example 0.95 (MGF of binomial)**.** We can use the above remark very nicely in computing the MGF of a binomial RV, because we can use the fact that a binomial is simply the sum of bernoullis. We have $X \sim Bin(n,p) = Y_1 + ... + Y_n$, where $Y_i \sim Ber(p)$. We have then

$$M_{Y_i}(s) = \mathbb{E}[e^{Y_i s}] = (1-p)e^{s\cdot 0} + pe^s = 1 - p + pe^s$$

$$\Rightarrow M_X(s) = (1 - p + pe^s)^n$$

Example 0.96 (Summing of a random number of random variables)**.** Let $Y = X_1 + ... + X_N$, where $X_1, ..., X_N$ are i.i.d. and N is a RV. We then have that

$$M_Y(s) = \mathbb{E}[e^{Ys}] = \mathbb{E}[\mathbb{E}[e^{Ys}|N]]$$

$$= \mathbb{E}[\mathbb{E}[e^{s(X_1+...+X_N)}|N]] = \mathbb{E}[M_X(s)^N]$$

$$= \mathbb{E}[e^{N\ln(M_X(s))}]$$

$$= M_N(\ln M_X(s))$$

Example 0.97 (Sum of Geometric number of exponential RVs)**.** We will begin with the fact that if $N \sim Geom(p)$, then

$$M_N(s) = \frac{pe^s}{1-(1-p)e^s}$$

Then, if $Y = X_1 + ... + X_N$, where each X_i is an iid exponential RV. Then, from the previous example, we have that $M_Y(s) = M_N(\ln(M_X(s)))$ We also have from before that $M_{X_i}(s) = \frac{\lambda}{\lambda - s}$. Then, we have

$$M_Y(s) = \frac{pM_{X_i}(s)}{1-(1-p)M_{X_i}(s)}$$

$$= \frac{p\frac{\lambda}{\lambda-s}}{1-(1-p)\frac{\lambda}{\lambda-s}}$$

0.20 Limiting Behavior of RV's

Suppose we observe a sequence $X_1, X_2, ..., X_n$ i.i.d. samples. We let

$$M_n = \frac{\sum X_i}{n}$$

be the **sample mean** (which makes sense, as it is just an average). We have:

1. $\mathbb{E}[M_n] = \frac{n\,\mathbb{E}[X_i]}{n} = \mu$
2. Assuming $\mathsf{Var}(X_i) < \infty$, we have

$$\mathsf{Var}(M_n) = \frac{1}{n^2}\sum \mathsf{Var}(X_i) = \frac{\mathsf{Var}(X_i)}{n} \to 0$$

as $n \to \infty$

A natural question is then: What happens to the "deviation" $|M_n - \mathbb{E}[M_n]| = |M_n - \mu|$

Definition 0.98 (Markov Bound)**.** For a *nonnegative random variable*, we have that

$$aP(X \geq a) \leq \mathbb{E}[X]$$

Proof. We begin by defining the indicator variable Z where $Z = 1$ if $X \geq a$ and is zero otherwise. Then $aZ \leq X$ by definition of Z. Taking expectations of both sides yields the desired result

□

Example 0.99. Let $X \sim U[0, 1]$. Then we have

$$\Pr(X > 3/4) \leq \frac{1/2}{3/4} = 2/3$$

and

$$\Pr(X > 1) \leq \frac{1}{2}$$

Which seems pretty stupid/not very powerful. But it is this way because it makes very little assumptions on the RV. We don't hate on Markov too much because it is actually the building block for many other bounds, and is often very useful when we don't know much or anything about the higher moments of our RV.

Remark 0.100 (Markov Inequality intuition)**.** Say my distribution has a mean of μ, and I want to maximize the probability that $Pr(X \geq k\mu)$. How would I do this? I would do this by letting X take on a value of $k\mu$ with probability $\frac{1}{k}$, and $X = 0$ otherwise. This achieves the correct expectation while still maximizing $Pr(X \geq k\mu)$.

Definition 0.101 (Chebyshev's Inequality)**.** Chebyshev's Inequality states

$$Pr(|X - \mu| \geq a) \leq \frac{\mathsf{Var}(X)}{a^2}$$

where $\mu = \mathbb{E}[X]$

Proof. We know that $\mathsf{Var}(X) = \mathbb{E}[(X - \mu)^2]$. We then have that

$$Pr((X - \mu)^2 \geq a) \leq \frac{\mathsf{Var}(X)}{a}$$

by Markov's inequality. This implies

$$\Rightarrow Pr(|X - \mu| \geq \sqrt{a}) \leq \frac{\mathsf{Var}(X)}{a}$$

$$\Rightarrow Pr(|X - \mu| \geq a) \leq \frac{\mathsf{Var}(X)}{a^2}$$

□

Note that Chebyshev's inequality holds for any random variable, not just positive ones (in contrast to Markov's inequality. We also note the special case:

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

when X has mean μ and variance σ^2.

Remark 0.102 (Weak Law of Large Numbers)**.** We can use Chebyshev's inequality to show a result known as the **Weak Law of Large Numbers**. Suppose we have an average of a bunch of i.i.d. RVs $M_n = \frac{X_1+\ldots+X_n}{n}$. Then we have that $\mathsf{Var}(M_n) = \frac{n\mathsf{Var}(X_i)}{n^2} = \frac{\mathsf{Var}(X_i)}{n}$. This implies via Chebyshev's inequality that:

$$Pr(|M_n - \mathbb{E}[X_i]| \geq a) \leq \frac{\sigma}{na^2}$$

Definition 0.103 (Chernoff Bound)**.** Suppose we know the MGF of our random variable $M_X(s) = \mathbb{E}[e^{sX}]$. Note that this is a positive RV, so we can apply markov's inequality:

$$Pr(e^{sX} \geq a) \leq \frac{\mathbb{E}[e^{sX}]}{a} = \frac{M_X(s)}{a}$$

$$\Rightarrow Pr(e^{sX} \geq e^{as}) \leq \frac{\mathbb{E}[e^{sX}]}{e^{as}}$$

$$\Rightarrow Pr(sX \geq as) \leq \frac{\mathbb{E}[e^{sX}]}{e^{as}}$$

where the last step follows since $f(x) = e^x$ is monotonic. Then, if $s > 0$ we have

$$\Rightarrow Pr(X \geq a) \leq \frac{\mathbb{E}[e^{sX}]}{e^{as}}$$

alternatively, if $s < 0$ we have

$$\Rightarrow Pr(X \leq a) \leq \frac{\mathbb{E}[e^{sX}]}{e^{as}}$$

Note that the chernoff bound is a function of s. We often have to choose the optimal choice for s to get a good bound (take derivative and set to zero!). Also if we recall the taylor series for e^x, the idea behind a chernoff bound is that it can use all the moments of a RV to bound said RV. Compare this to Markov's, which only uses the first moment, and Chebyshev's, which only uses the second moment. This might lead one to think that Chernoff is *always* better than applying markov/chebyshev bounds, or even applying markov's bound to higher moments of the random variable. This leads to the following remark:

Remark 0.104 (Is Chernoff always better than Markov/Chebyshev?)**.** In short, no. Consider using Markov's inequality to bound a higher moment of our RV X. This yields (provided the higher moment is positive of course) $Pr(X \geq a) \leq \frac{\mathbb{E}[X^k]}{a^k}$. Here I claim:

$$\inf_{k>0} \frac{\mathbb{E}[X^k]}{a^k} \leq \inf_{s>0} \frac{\mathbb{E}[e^{sX}]}{e^{as}}$$

Why is this true? Lets examine the RHS:

$$\frac{\mathbb{E}[e^{sX}]}{e^{as}} = \frac{1}{e^{as}} \sum_k \frac{s^k \mathbb{E}[X^k]}{k!}$$

$$= \sum_k \left(\frac{(as)^k e^{-as}}{k!} \right) \frac{\mathbb{E}[X^k]}{a^k}$$

Now, the above expression is simply averaging over the moment bounds where you let the moment be distributed as a Poisson random variable with parameter as, and in general, the minimum over the moment bounds will be smaller than the average (no matter how the averaging is done, and so minimizing over s doesn't change anything), and thus we get the result.

Lecture 9: Convergence, Weak and Strong Law of Large Numbers, Central Limit Theorem

Agenda:

1. Recap of Limit Theorems (Chernoff)
2. Laws of Large Numbers (WLLN, convergence in probability)
3. Central Limit Theorem

0.21 Recap of Bounds

Example 0.105. Let $X \sim \mathcal{N}(0,1)$. We can bound the tail probabilities of X using the chernoff bound:

$$\Pr(X \geq k) \leq \frac{\mathbb{E}[e^{sX}]}{e^{sk}}$$

Recalling the MGF of a normal distribution, we have:

$$= \frac{e^{s^2/2}}{e^{sk}} = e^{s^2/2 - sk}$$

Minimizing this expression over $s > 0$ corresponds to minimizing the exponent. Taking the derivative, we see

$$-k + s = 0 \Rightarrow s^* = k$$

Plugging this optimal value s^* in, we get:

$$\Pr(X \geq k) \leq e^{-k^2/2}$$

Which is actually exponential decreasing, which is much closer to the true behavior of the normal distribution.

Exercise 0.106. Extend the above exercise to show that for $X \sim \mathcal{N}(0,1)$, we have

$$\Pr(|X| \geq k) \leq 2e^{-k^2/2}$$

Definition 0.107 ((Weak) Law of Large Numbers (WLLN))**.** If we perform an experiment n times independently and

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- If X_i has mean μ and variance σ^2
- $\mathbb{E}[M_n] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\mathbb{E}[X_i]n = \mu$
- $$\mathsf{Var}(M_n) = \frac{1}{n^2}\sum \mathsf{Var}(X_i) = \frac{\sigma^2}{n}$$

This tells us that if $X_1, ..., X_n$ are i.i.d. RV's with mean μ and finite variance, then for every $\epsilon > 0$, we have

$$\Pr(|M_n - \mu| \geq \epsilon) \to 0$$

as $n \to \infty$

Proof. The proof of the last claim is quite simple and only uses Chebyshev's inequality. It tells us at what "rate" this probability goes to zero as $n \to \infty$. We have

$$\Pr(|M_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2 n} \to 0 \quad \text{as } n \to \infty$$

□

What does the WLLN tell us? It tells us that

$$\lim_{n\to\infty} \Pr(|M_n - \mu| \geq \epsilon) = 0$$

Remark 0.108. Similar to the definition of the limit. For any $\epsilon > 0$, $\delta > 0$, there exists some $n_0(\epsilon, \delta)$ such that

$$\Pr(|M_n - \mu| \geq \epsilon) \leq \delta)$$

for all $n > n_0(\epsilon, \delta)$. We say then that M_n **converges in probability** to μ.

Example 0.109. Let $Y_n = \min(X_1, ..., X_n)$ for $X_i \sim U[0, 1]$. We have that

$$\Pr(|Y_n - 0| \geq \epsilon) = \Pr(|X_1| > \epsilon, |X_2| \geq \epsilon, ..., |X_n > \epsilon) = (1 - \epsilon)^n$$

Which goes to zero for all $\epsilon > 0$ as $n \to \infty$. This tells us that Y_n converges in probability to 0.

Example 0.110. Suppose we have an arrival process where we divide the number line into exponentially increasing sized intervals:

$$I_k = \{2^k, 2^k + 1, ..., 2^{k+1} - 1\}$$

And suppose we have exactly one arrival in each interval. So we let $Y_n = 1$ if there is an arrival at time n, and $Y_n = 0$ if there is no arrivals. We then have that

$$\Pr(Y_1 = 1) = 1$$

$$\Pr(Y_2 = 1) = \Pr(Y_3 = 1) = 1/2$$

$$\Pr(Y_n = 1) = \frac{1}{2^k} \quad \text{if } n \in I_k$$

This implies that

$$\lim_{n\to\infty} \Pr(|Y_n - 0| \geq \epsilon) = \lim_{n\to\infty} \Pr(Y_n = 1) = \lim_{k\to\infty} \frac{1}{2^k} = 0$$

Which tells us that Y_n converges in probability to 0.

The above highlights the weakness of convergence of probability. We can see of course that for any finite n, there are certainly an infinite number of 1's (arrivals) after n, yet it still converges in probability. This is fixed by something known as **almost sure** convergence, which we will not get deep into in this course.

Question: What happens to $S_n = \sum_{i=1}^n X_i$. This is just a bunch of convolutions! In particular, if each $X_i \sim U[0, 1]$, we know that convolving two uniform pdfs looks like a triangle pdf. Convolving yet again gives us a quadratic polynomial. Each time we convolve the width gets higher (the variance blows up) and the order of the polynomial becomes larger. This general phenomenon happens for non-uniform iid RVs (amazingly) as well!

Our problem is that the mean and variance of S_n both blow up as $n \to \infty$. To fix this, we define

$$\widehat{S_n} = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

which we can verify has zero mean and unit variance.

Theorem 0.111 (Central Limit Theorem). *The CLT says that*

$$\lim \Pr(\widehat{S_n} \leq x) = \Phi(x)$$

where $\Phi(x)$ is the CDF of the standard normal distribution! This type of convergence is known as convergence in distribution.

Proof. We present a sketch of the proof. Note that $S_n \to \mathcal{N}(0,1)$ implies that $S_n \to \mathcal{N}(n\mu, n\sigma^2)$. □

Exercise 0.112. See **Sinho's notes** on modes of convergence (I will hopefully type up my own sometime soon, but these are very good)

Remark 0.113 (SLLN vs WLLN). The WLLN, as we already discussed, says that

$$Pr(|M_n - \mathbb{E}[X_i]| \geq a) \leq \frac{\sigma}{na^2}$$

Which tells us that

$$\lim_{n\to\infty} Pr(|M_n - \mathbb{E}[X_i]| \geq a) = 0$$

The Strong Law of Large Numbers, on the other hand, says something stronger. It says that:

$$Pr(\lim_{n\to\infty} M_n = \mu) = 1$$

On the surface, these look similar. But the key difference is that for some $\epsilon > 0$, the SLLN says that $|M_n - \mu| > \epsilon$ will only happen a *finite* number of times (in other words, there exists some N such that $n > N \Rightarrow |M_n - \mu| < \epsilon$). On the other hand, the WLLN makes no such guarantee. More specifically, the WLLN says that M_n converges *in probability*, while the SLLN says M_n converges *almost surely* or *with probability one*. For more details on the difference between these two things, you should refer to Sinho's notes or the course notes.

Lecture 10: Information Theory

Agenda:

1. Recap of WLLN
2. Proof of CLT
3. Introduction to Information Theory (Entropy, Compression)

0.22 Proof of CLT

Recall the CLT:

Theorem 0.114 (Central Limit Theorem). *The CLT says that*

$$\lim \Pr(\widehat{S_n} \leq x) = \Phi(x)$$

where $\Phi(x)$ is the CDF of the standard normal distribution! This type of convergence is known as convergence in distribution.

where we had defined $\widehat{S_n}$ as

$$\widehat{S_n} = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

Proof. Let

$$Z_n = \frac{\sum_{i=1}^{n} X_i}{\sqrt{n}}$$

where each X_i is iid and $\mathbb{E}[X_i] = 0$ and $\mathsf{Var}(X_i) = 0$. We also note that if $Y \sim \mathcal{N}(0,1)$, then $M_Y(s) = e^{s^2/2}$ and furthermore $\log M_Y(s) = s^2/2$. So it suffices to show that the log of the MGF of Z_n is $s^2/2$. We have

$$M_{Z_n}(s) = \mathbb{E}[e^{sZ_n}] = \mathbb{E}[\exp(\frac{s}{\sqrt{n}} \sum_{i=1}^{n} X_i)]$$

$$= \mathbb{E}[\exp(\frac{s}{\sqrt{n}} X_1) \cdots \exp(\frac{s}{\sqrt{n}} X_n)]$$

$$= \mathbb{E}[\exp(\frac{s}{\sqrt{n}} X_1)] \cdots \mathbb{E}[\exp(\frac{s}{\sqrt{n}} X_n)]$$

$$= \left[M_X(\frac{s}{\sqrt{n}}) \right]^n$$

Now, recall that $M_X(0) = 1$, and $M_X'(0) = 0$, and $M_X''(0) = 1$, by our assumptions and the properties of the MGF. Now we consider:

$$\lim_{n\to\infty} \log M_{Z_n}(s) = \lim_{n\to\infty} \left[n \log M_X(\frac{s}{\sqrt{n}}) \right] = \lim_{n\to\infty} \left[\frac{\log M_X(\frac{s}{\sqrt{n}})}{\frac{1}{n}} \right]$$

Now, letting $y = \frac{1}{\sqrt{n}}$

$$= \lim_{y\to 0} \left[\frac{\log M_X(sy)}{y^2} \right]$$

Now notice that the limit of both the numerator and the denominator is zero, so we can use L'Hopital's rule!

$$= \lim_{y\to 0} \left[\frac{sM_X'(sy)}{2yM_X(sy)} \right]$$

The numerator and denominator once again both go to zero. L'Hopital again!

$$== \lim_{y \to 0} \left[\frac{s^2 M_X''(sy)}{2M_X(sy) + 2ysM_X'(sy)} \right] = \frac{s^2}{2}$$

□

Example 0.115 (Polling Example)**.** Suppose we ask n randomly sampled voters if they support candidate X. So $X_i = 1$ if yes, and zero otherwise. Suppose we want a 95% confidence interval that $|M_n - p| < \epsilon$, where p is the true probability that each voter supports our candidate, and $M_n = \frac{1}{n}\sum X_i$ is the empirical mean. Well, Chebyshev tells us that

$$\Pr(|M_n - p| \geq a) \leq \frac{\mathsf{Var}(M_n)}{a^2}$$

But now we note that $\mathsf{Var}(X_i) = p(1-p) \leq 1/4$, which tells us that $\mathsf{Var}(M_n) = \frac{1}{n}\mathsf{Var}(X_i) \leq \frac{1}{4n}$. Now, suppose we want to know our p value to within 0.1 with probability at least 95%. Mathematically, we want:

$$\Pr(|M_n - p| \geq 0.1) \leq 0.05$$

and we know

$$\Pr(|M_n - p| \geq 0.1) \leq \frac{\mathsf{Var}(M_n)}{0.1^2} \leq \frac{1}{4n(0.01)}$$

Which implies that in order for us to obtain a 95% confidence interval, we need to set $n \geq 500$. If $a = 0.01$, then we would need $n \geq 50000$ for a 95% confidence interval!

Now, let's compare this with the CLT method. The CLT tells us that

$$\frac{M_n - \mathbb{E}[M_n]}{\sqrt{\mathsf{Var}(M_n)}} \to \mathcal{N}(0,1)$$

and we want

$$\Pr(|M_n - p| \geq 0.1) \leq 0.05$$
$$\Leftrightarrow \Pr(\frac{|M_n - p|}{\frac{1}{2\sqrt{n}}} \geq \frac{0.1}{2\sqrt{n}}) \leq 0.05$$

But notice that the left hand side is roughly a standard normal. To get a 95% confidence interval for a normal distribution, we use the fact that we know 95% of the probability mass lies within 2 standard deviations, and in this case a standard deviation is 1. So we have

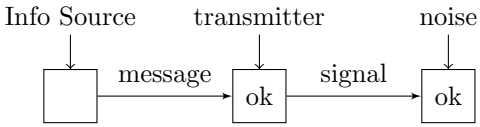
$$0.2\sqrt{n} \geq 2 \Rightarrow n \geq 100$$

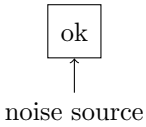
Which we can see is much better than the result Chebyshev gives us.

0.23 Intro to Info Theory

The field of information theory was pioneered by Claude Shannon in his seminal 1948 paper "A Mathematical Theory of Communication". There is a great textbook on the topic "Elements of Information" by Cover and Thomas, which is a highly recommended resource. Also, if you are interested in this topic further, you should take EECS 229A!

Shannon was concerned with the question, how much information can I reliably send over a noisy channel?





There are two things we can concern ourselves with.

1. How much can we compress our information in the presence of no noise? This is known as the **Source Coding** problem.
2. How much information can we send in the presence of noise? This is known as the **Channel Coding Problem**

Shannon was able to answer both of these questions, and he was also even able to say that we can separately optimize for both of these criterion and arrive at a globally optimal solution!

Lecture 11: Info Theory, Binary Erasure Channel

Agenda:

1. Information theory overview (Entropy, AEP, Capacity of BEC)

Recall that there were two fundamental questions Shannon was exploring:

1. How much can we compress our information in the presence of no noise? This is known as the **Source Coding** problem.
2. How much information can we send in the presence of noise? This is known as the **Channel Coding** problem.

Theorem 0.116 (Source Coding Theorem). *Given N i.i.d. RV's $X_1, ..., X_n$, each having entropy $H(X)$, then these can be compressed with a source coding channel into no more than $N(H(X) + \epsilon)$ bits, $\forall \epsilon > 0$ as $N \to \infty$.*
Conversely, we also have that compression to fewer than $NH(X)$ bits is impossible without loss of information.

Definition 0.117 (Entropy). The entropy of a discrete RV X is defined as

$$H(X) := \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}$$

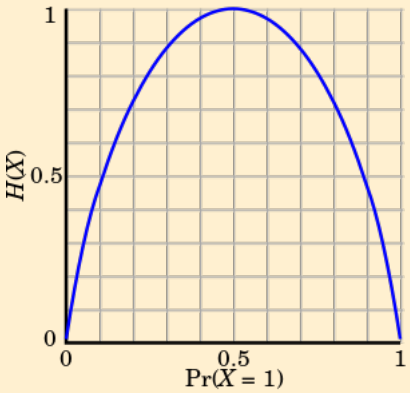
$$= \mathbb{E}[\log \frac{1}{P_X(x)}]$$

We can interpret this definition very roughly by noting that the quantity $\log \frac{1}{P_X(x)}$ roughly corresponds to the "surprise" of seeing the outcome x. Then the entropy corresponds to the "average surprise" of our distribution. Another interpretation of entropy is that is correlated to the uncertainty of the random variable.

Example 0.118. When $X \sim Bern(p)$, then

$$H(X) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} =: H(p)$$

We can graph this quantity as p varies from 0 to 1. Note that at 0 and 1, the quantity is 0, and at $p = 1/2$, the quantity is 1. We also can calculate that $H(0.11) = 1/2$. This tells us that if we have a really long sequence of $Bern(0.11)$ RV's, then roughly half of the bits are "redundant", i.e. they can be compressed.



We further can naturally define

1.

$$H(X,Y) = \sum_{x,y} P_{X,Y}(x,y) \log \frac{1}{P_{X,Y}(x,y)}$$

Exercise 0.119. Show that

$$H(X,Y) = H(X) + H(Y|X)$$

where

$$H(Y|X) =$$

Example 0.120. We now consider a motivating example for the AEP. Suppose we flip a coin n times independently. What is a "typical" sequence? Well, there are 2^n total sequences, but a "typical" sequence has np heads and $n(1-p)$ tails. The probability of a particular "typical sequence" S is:

$$\begin{aligned} P(S) &= p^{np}(1-p)^{n(1-p)} \\ &= 2^{np\log p} 2^{n(1-p)\log(1-p)} \\ &= 2^{n(p\log p + (1-p)\log(1-p)} = 2^{-nH(p)} \end{aligned}$$

Our next question is then, how many such typical sequences are there? Well, there are exactly $\binom{n}{np}$, which it turns out is approximately $2^{nH(p)}$ for large n! How do we know this? Well it uses Stirling's approximation, and we won't go into detail here, but the first steps look something like this:

$$\binom{n}{np} = \frac{n!}{(np)!(n(1-p))!}$$

and we use the fact that $n! \approx \left(\frac{n}{e}\right)^n$.

What does this example tells us? Well, we have $2^{nH(p)}$ sequences, and all of these sequences occur with probability $2^{-nH(p)}$. This means virtually all of the probability must be used up by these "typical sequences"! This is known as the **Asymptotic Equipartition Property**, and is really quite a mind-boggling phenomenon, which is hopfully illustrated by this following example:

Example 0.121. Suppose our sequence of RVs are iid $Bern(0.11)$, and we are sending sequences of $n = 1000$ of these bits. We know that $H(p) = 0.5$. This tells us that our "typical set" is composed of the set of approximately 2^{500} sequences containing roughly $1000 * 0.11$ 1's and $1000 * 0.89$ 0's, and each of the sequences in this typical set have roughly equal probability. Then the source coding theorem tells us we can transmit these sequences of 1000 bits with on average only around 500 bits!

How could we achieve this in practice? This is a very difficult question, and one that information theorists do not typically concern themselves with. We can, however, consider the following computationally infeasible scheme:

1. put each of the 2^{500} "typical" sequences into a lookup table with 2^{500} entries
2. if the input sequence is in the typical set, simply send a "0" following by the bit string that is the index of the typical sequence in the lookup table. The decoder can just look up the typical sequence in his copy of the lookup table when he receives the compressed message.
3. if the sequence is not "typical", just send a "1" followed by the whole sequence. This happens with probability that goes to zero as $n \to \infty$.

The bit at the beginning is simply to let the receiver know whether to look in the lookup table or to just look at the next 1000 bits. This scheme is entirely infeasible because we cannot store 2^{500} size lookup table

in our computer, and much research in the last 50 years has been devoted to achieving the source coding theorem in practice.

Theorem 0.122 (AEP). *We now formalize the Asymptotic Equipartition Property. If $X_1, ..., X_n$ are i.i.d. $\sim P_X(x)$, then*

$$-\frac{1}{n}\log P(X_1, ..., X_n) \longrightarrow H(X)$$

in probability as $n \to \infty$

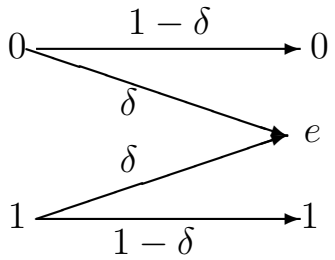
Proof. Recall that a function of RV's is still an RV. Then, if the X_i's are iid, so are $\log P(X_i)$, so we have

$$-\frac{1}{n}\log P(X_1, ..., X_n) = -\frac{1}{n}\sum \log P(X_1) \to -\mathbb{E}[\log P(X_i)] = H(X)$$

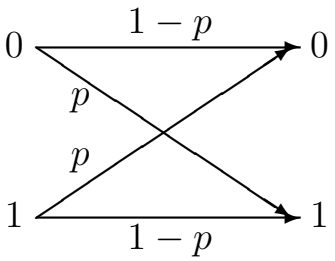
Where we have use the WLLN. □

0.24 Capacity of BEC

We have a **Binary Erasure Channel** looks like this:



Where this models the "noise" of a channel which takes a bit and erases it (maps it to e) with some probability δ. We also have a **Binary Symmetric Channel**, which looks like this:



As you can see, it flips each bit independently with probability p.

For the rest of this lecture, we will focus on the Binary Erasure Channel. Intuitively, we note that a binary erasure channel should have a higher capacity, which we will define shortly, than a BSC, because a BEC tells you exactly which bits have been corrupted.

Definition 0.123 (Capacity). We say that the **Capacity** of a channel is the maximium rate of reliable communication for that channel. Mathematically,

$$Rate = R = \frac{L_n}{n}$$

where L_n is the length of your message, and n is the length of your encoding.

Say m is the message your encoder receives (so $m \in \{0,1\}^{L_n}$), and at the end your decoder outputs a guess $\hat{m}$. We would like to minimize the probability of error:

$$P_e^{(n)} = \max_m \Pr[m \neq \hat{m}]$$

We say that rate R is **achievable** for the channel if for every positive number n that is "long enough", there exists and encoder and decoder functions f_n and g_n respectively such that

$$P_e^{(n)} \to 0$$

as $n \to \infty$. The largest achievable rate R is called the **capacity** of our channel.

Theorem 0.124. *We have that the capacity of a BEC channel is*

$$C_{BEC(p)} = 1 - p$$

bits per channel use.

This is really a remarkable result (think about why!). We have to show two things:

1. The **converse**: We need to be able to show that it is not possible to achieve a rate of $1 - p + \epsilon$ for any $\epsilon > 0$.
2. **achievability**: We would like to show that there is actually a scheme (even if it is computationally infeasible) that achieves this $1 - p$ rate

The proof of the converse goes as follows: Suppose there is a genie which is actually helping you encode and decode your message by *telling you in advance* exactly which bits will be erased. We can show that even with this help, we cannot achieve a capcity better than $1 - p$ as $n \to \infty$.

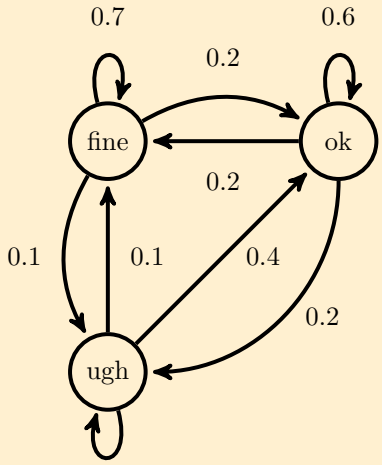
Lecture 12

Definition 0.125 (Markov Chain)**.** A Markov chain is a sequence of random variables $X_0, X_1, X_2, ...$ satisfying the following 3 conditions:

1. The assumption that
$$Pr(X_{n+1} = c_{n+1} | X_0 = c_0, ..., X_n = c_n) = Pr(X_{n+1} = c_{n+1} | X_n = c_n)$$
2. $X_0, X_1, ...$ take on values from some set S
3. X_0 is an arbitrary pmf on S.

For a homogeneous discrete time markov chain, we say $Pr(X_{n+1} = i | X_n = j) = P_{ji}$

Example 0.126. Here is an example of a markov chain, represented with a diagram. It represents



the three fundamental states of any Berkeley student.

We can specify what our initial state X_0 is, and answer questions such as what is the $Pr(X_5 = ugh)$? Over 16 weeks, what fraction of time are you ok?

Example 0.127 (PageRank)**.** Pagerank is google's algorithm for returning search results. It is now much more complicated, but at it's core it uses markov chains to determine how popular each website on the internet is. There are a few ways to formulate this notion:

1. Score each page i with π_i, such that
$$\pi_i = \sum_j \pi_j P_{ji}$$
$$\sum_i \pi_i = 1$$
The first equation is called a balance equation. More on that later
2. bot randomly picks link on each page it visits. π_i is the equal to the probability that the bot is on page i at some point in time $t >> 0$.
3. π_i = fraction of time bot spends on page i.

All three of these formulations are equivalent.

Let's start working our way more towards these balance equations. First, we are interested in what happens to finite states as $n \to \infty$. We define $r_{ij}(n)$ as the probability of going from state i to state j in n time steps. Well, $r_{ij}(1) = P_{ij}$, since there is only one way to get from i to j in one time step. $r_{ij}(n)$ is more complicated, but luckily we can actually write it in terms of $r_{ij}(n-1)$ as follows:

$$r_{ij}(n) = \sum_{k \in \mathcal{S}} r_{ik}(n-1) P_{kj}$$

The above are a form of the *Chapman Kolmogorov Equations*. They should intuitively make sense: the only way to get from i to j in n steps is if you first get to somewhere else in $n-1$ steps, and then make the last step to state j. We are just summing over all the possible places you could be at time step $n-1$. Lets examine $r_{ij}(2)$. We have that

$$r_{ij}(2) = \sum_{k \in \mathcal{S}} r_{ik}(1) P_{kj} = \sum_{k \in \mathcal{S}} P_{ik} P_{kj}$$

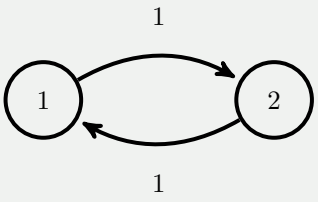
$$= \begin{bmatrix} P_{i1} & P_{i2} & \cdots & P_{im} \end{bmatrix} \begin{bmatrix} P_{1j} \\ P_{2j} \\ \vdots \\ P_{mj} \end{bmatrix}$$

Now further recall from CS70 our transition probability matrix:

$$\begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1m} \\ P_{21} & P_{22} & \cdots & P_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ P_{m1} & P_{m2} & \cdots & P_{mm} \end{bmatrix}$$

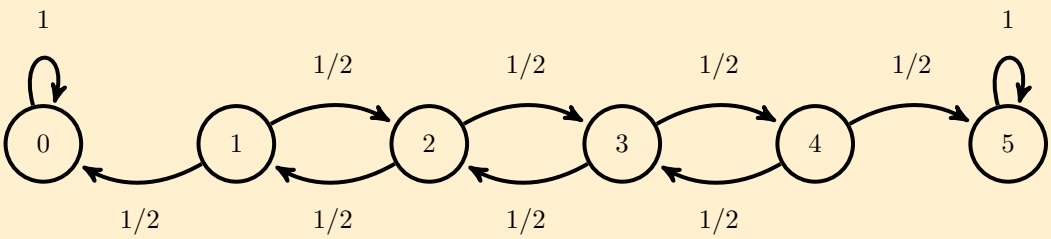
We can then see that $r_{ij}(2) = (P^2)_{ij}$. This is quite convenient! It is also very easy to then see that $r_{ij}(n) = (P^n)_{ij}$, or the $(i,j)^{th}$ entry of P^n. If the values of each column of P^n converge to the same value, then this tells us that no matter where we start out, you have an equal probability of ending up in a given state.

Remark 0.128. Consider the below markov chain:



It is easy to see that we will "ping pong" infinitely back and forth, and it is entirely deterministic which state we are in at any given time (given that we know where we started).

Example 0.129 (2 spiders, 1 fly)**.** Consider that below markov chain:



We will intuitively always end up either at state 0 or state 5, and it is much more likely that we get stuck at 5 if we start in state 4 than if we start in state 1. So once again, $Pr(X_n = i)$ is **not** always independent of where we start. In which situations is it? Stay tuned...

We first note that a more concise way to write (1) in the PageRank example would be $\pi P = \pi$ where $\sum_i \pi_i = 1$. Such a π satisfying these two equations is called the **stationary distribution** of a markov chain. We further note the following definitions:

Definition 0.130. We say a state i is **recurrent** if for all other states j

$$j \text{ accessible from } i \Rightarrow i \text{ accessible from } j$$

Furthermore, we say that a state is **positive recurrent** if $\mathbb{E}[T_i] < \infty$, where T_i is the time to return to state i after leaving it. Otherwise if $\mathbb{E}[T_i] = \infty$ then the state is called **null recurrent**. If a state is not recurrent at all then it is called **transient**.

Definition 0.131. The **class** of a state i is $\{j : j \text{ accessible from } i \text{ and } i \text{ accessible from } j\}$

Proposition 0.132. *The states in a class are either all recurrent or all transient.*

Proof. Let i and j be in the same class and suppose towards a contradiction that i is recurrent while j is transient. Since i and j are in the same class we know that they are accessible to each other, i.e. there exists some path from i to j and likewise some path from j to i. Since j is transient, $\exists k$ such that k is accessible from j but j is not accessible from k. This implies that there exists a path from i to k (going through j), and furthermore since i is recurrent that means there must be some path from k to i. But now we have a contradiction, as there is a path from k to i, and a path from i to j, so there must be a path from k to j. □

If the above proof was confusing, it is very helpful to draw it out!

Definition 0.133. Consider $s_i = \{n : r_{ii}(n) > 0\}$. Then we define the **periodicity** of a state as $GCD(s_i)$. In english, the periodicity of a state is the GCD of the all the possible times we could return to that state. In the "ping pong" example, both states have a periodicity of 2. If a state has a self loop, then its periodicity is trivially one.

Proposition 0.134. *All the states in a class have the same period.*

Proof. We start by denoting $d(s)$ as the period of state s. Once again, consider i and j in a communicating class together. We know i and j are accessible from each other, so WLOG consider a path of length n from i to j and a path of length m from j to i. Then there is a path of $n + m$ from i to i, so it follows from the definition of the period that $n + m$ is divisible by $d(i)$. Consider any path from j to j. Say it has length t. This creates yet another path from i to i of length $n + t + m$ (first go from i to j, then j to j, then back to i). B the same logic, we have that $n + m + t$ is divisible by $d(i)$. This implies that t is divisible by $d(i)$, for all t such that there is a path of length t from j back to j. Since this holds for all t, this means that $d(i)$ is a factor of $\{n : r_{jj}(n) > 0\}$, and by definition it is less than or equal to the greatest common factor, $d(j)$ (we can even claim that $d(j)$ is divisible by $d(i)$, by why overcomplicate things?). Reversing the roles of i and j in the above argument implies that $d(j) \leq d(i)$, which implies that $d(j) = d(i)$, as desired. □

In general, any MC with a single aperiodic recurrent class (and some transients) must converge in the following sense:

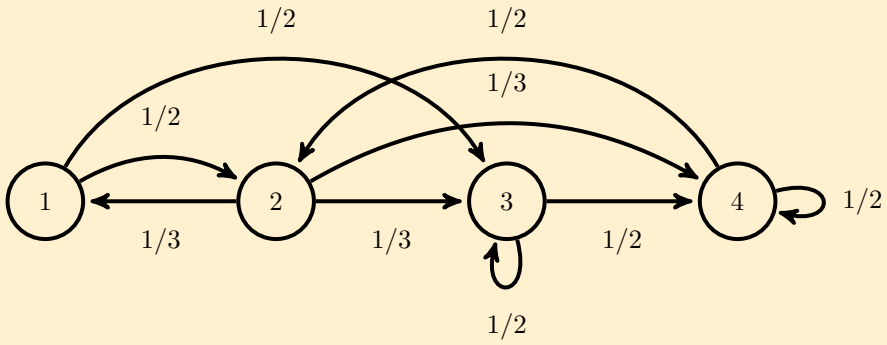
1. for each state j,
$$\lim_{n\to\infty} P_{ij}^{(n)} = \pi_j, \forall i$$

2. The π_j are given by a system of equations:
$$\pi_j = \sum_{k=1}^{m} \pi_k P_{kj} \qquad \sum_i \pi_i = 1$$

3. $\pi_i = 0$ if state i is transient, and $\pi_i > 0$ if i is recurrent.

Now, what if we wanted to find the expected amount of time to get from one state to another, given that we are in stationarity? We will now develop a tool known as **first-step equations** to deal specifically with this omnipresent problem. It is actually easiest to see with an example.

Example 0.135. Consider the MC below.



we further define x_i as the expected amount of steps we must take to reach a certain special state, say 1 in this case. Then trivially we can observe that $x_1 = 0$. What about x_2? Well, with 1/3 probability, we are done, but we could also could go to states 3 and 4. By splitting up into cases, we can see

$$\begin{aligned} x_2 = 1 &+ Pr(\text{we go to 1})\mathbb{E}[\text{time to 1 from 1}] \\ &+ Pr(\text{we go to 2})\mathbb{E}[\text{time to 1 from 2}] \\ &+ Pr(\text{we go to 3})\mathbb{E}[\text{time to 1 from 3}] \\ &+ Pr(\text{we go to 4})\mathbb{E}[\text{time to 1 from 4}] \\ &= 1 + 1/3 * 0 + 0x_2 + 1/3x_3 + 1/3x_4 \end{aligned}$$

Note that we include the 1+ since we need to take at least one step no matter what happens. Using similar logic, we can come up with the following system of equations for the other nodes:

$$x_3 = 1 + 1/2x_3 + 1/2x_4$$

$$x_4 = 1 + 1/2x_2 + 1/2x_4$$

Which leaves us with three equations and three unknowns, which means we can solve for each x_i.

In general, we can use the same idea to define a **mean recurrence time** t_s^* = the average number of steps the MC takes to return to state s. Then we have

$$t_s^* = 1 + \sum_{i=1}^{m} P_{si} t_i$$

where t_i is of course the expected amount of time to get from state i to state s.

Example 0.136. We can consider the same markov chain from the previous example, but a more general hitting time problem. Given sets A and B such that $A \cap B = \emptyset$, we want to find the probability that we reach a node in set A before we reach a node in set B. Using the MC from the previous example, we can let $A = \{1\}$ and $B = \{4\}$. Then now we can define x_i = probability that we reach A before B given we start in state i. Then trivially, $x_1 = 1$ and $x_4 = 0$. We also have by splitting into cases.

$$\begin{aligned} x_2 = &Pr(\text{we go to 1}) * Pr(\text{we get to } A \text{ first given go to 1}) \\ &+Pr(\text{we go to 2}) * Pr(\text{we get to } A \text{ first given go to 2}) \\ &+Pr(\text{we go to 3}) * Pr(\text{we get to } A \text{ first given go to 3}) \\ &+Pr(\text{we go to 4}) * Pr(\text{we get to } A \text{ first given go to 4}) \\ &= 1/3 * 1 + 0 * x_2 + 1/3 * x_3 + 1/3 * 0 \end{aligned}$$

Similarly, we can formulate another equation for x_3, which would allow us to solve for our two unknowns.

Remark 0.137. There is an inherent connection to the material we have (or will) learned about in CS188. We can think about collecting a reward R_i every time to MC is in state i, and then we can further define r_i as the expected reward we get starting from state i until we reach some set A. Then we have

$$r_i = R_i \quad \forall i \in A$$

$$r_i = R_i + \sum_j P_{ij} r_j \quad \forall i \notin A$$

We can also think about adding in a "discount factor" so if $X(n) = i$ then we receive reward $\beta^n R_i$, where β is the discount factor. Then similarly we have:

$$r_i = R_i \quad \forall i \in A$$

$$r_i = R_i + \beta \sum_j P_{ij} r_j \quad \forall i \notin A$$

Exercise 0.138. Suppose Alice commutes between 2 houses every week. If the weather is great (this happens with probability p), she grabs her fishing rod and fishes on her way to the other house. There are N fishing rods. Assuming this has been going on for a very long time already, what is the probability that she has no rods when the weather is good? *Hint:* Set up a MC with $N + 1$ states.