EE126 Notes

Michael Whitmeyer

August 2018 - December 2018

1 Lecture 1: CS70 Review

[kind of boring]

2 Lecture 2: CS70 Review

Example 2.1 (example of two random variables that are conditionally independent but not themselves independent)**.** Let there be two coins, one that shows up with heads with probability 0.9, and the other that shows up head with probability 0.1.

$$P(\text{heads}|C = C_1) = 0.1$$

$$P(\text{heads}|C = C_2) = 0.9$$

Where C is the coin we pick. Let's pick one coin at random, and toss it two times, where H_i is an indicator that the i^{th} toss is heads. We can see that, conditioned on C, the H_i are independent, but that if we do not condition on C, then they are not conditionally independent.

$$P(H_2 = 1|H_1 = 1) \neq P(H_2 = 1)$$

but

$$P(H_2 = 1|H_1 = 1, C = C_i) = P(H_2 = 1|C = C_i)$$

or equivalently

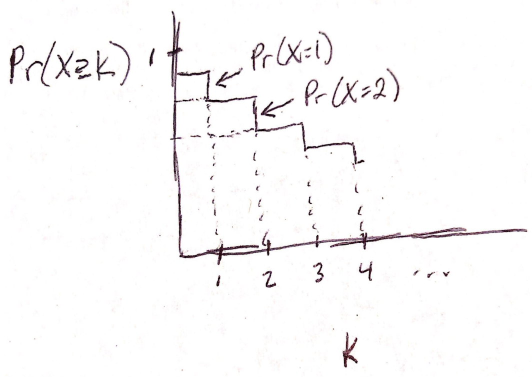
$$P(H_2 = 1, H_1 = 1|C = C_i) = P(H_2 = 1|C = C_i)P(H_1 = 1|C = C_i)$$

which shows conditional independence but not independence.

Theorem 2.2 (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, $E[X] = \sum_{k=1}^{\infty} kPr(X = k)$, 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 2.3. 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.

Theorem 2.4. *Let $Y = g(X)$. Then we have that*

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

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

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

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

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

□

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

Theorem 2.5. *(Linearity of Expectation)*
We have

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

for arbitrary X and Y.

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

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

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

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

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

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

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

Definition 2.6 (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

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

Similarly, we have that (since the B_i are iid)

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

Example 2.7. Let $Y = aX + b$. Then we have that

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

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

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 2.8 (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

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

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

Example 2.9 (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 2.10 (Poisson Random Variable)**.** We define $X \sim Pois(\lambda)$ with the follwing PMF:

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

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

Now we talk about **Conditioning with RVs**. We start with simply Bayes' Rule:

$$Pr(X = k|Y = m) = \frac{Pr(X = k \cap Y = m)}{Pr(Y = m)}$$
$$\Rightarrow Pr(X = x, Y = y) = Pr(X = x|Y = y)Pr(Y = y) = Pr(Y = y|X = x)Pr(X = x)$$

And **Total Probability** (which we have probably (ha) already used but define here explicitly) for $A_1, ..., A_n$ events that form a disjoint cover of the entire probability space:

$$Pr(X = x) = \sum_{i=1}^{n} P(A_i)P(X = x|A_i)$$

Example 2.11 (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

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

3 Lecture 3

[CS70 Review, I might've missed this lecture]

4 Lecture 4

Example 4.1 (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 $A|X_i Y_i = 1$ 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.

Let $Z = X + Y$. The most basic way to calculate the PMF of Z is through what is known as a **convolution**. We have

$$Pr(Z=z) = \sum_x Pr(X=x, Y=z-x)$$

Intuitively this should make sense, as we are summing over all the possible ways X and Y could add up to equal z.

4.1 Continuous RVs

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. 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. Instead, we talk about f_X, which is the **density** or the **probability density function (PDF)** of a continuous random variable. 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).

Since the CDF would be

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

It follows from the fundamental theorem of calculus (if F is differentiable), that $f_X(x) = \frac{d}{dx}F_X(x)$. This can be a very useful fact, as often the CDF is easier to calculate than the PDF.

Definition 4.2 (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 satisfies $\frac{f(s+t)}{f(s)} = f(t)$? Well, eventually, we might notice that $f(x) = e^x$ works! The problem is, this increases $f(x)$ as x increases, which is not the behavior we want if we are to keep the analogy. Well, $f(x) = e^{-x}$ also works, and it is monotonically decreasing, so that is better! In fact, we can even throw in a constant $f(x) = e^{-\lambda x}$, 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.

There are several analogous properties that we mentioned for discrete RVs that also still work in the continuous setting. For example, for **independent RVs** X and Y, we have $Pr(X \leq x, Y \leq y) = F_X(x)F_Y(y)$. Furthermore, for any arbitrary function g, we have that $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx$.

Definition 4.3 (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:**

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

Definition 4.4 (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:

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

Note that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then we have that $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$. (Check this yourself if it is not clear).

5 Lecture 5

Example 5.1 (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 $\delta \to 0$.

Example 5.2. 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 5.3. convince yourself that $\Phi(1) - \Phi(-1) = 2\Phi(1) - 1$ if you haven't already.

TODO 1. There's a lot of graphs the rest of this lecture that would take a lot of time to latex.

Definition 5.4 (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

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

6 Lecture 6

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:

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

Definition 6.1 (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!)

Now we examine something know as the **Tower Rule** or **Iterated Expectation** or the **Law of Total Expectation** (there's even a couple of more names but these are the most common). Consider:

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

$$= \int_Y \int_X x f_{X|Y}(x|y) f_Y(y) dx dy = \int_X \int_Y x f_{X,Y}(x, y) dy dx$$

$$= \int_X x f_X(x) dx = \mathbb{E}[X]$$

The above result should make intuitive sense when you think about it, and the intuition is quite similar to the intuition behind 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 6.2. 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.

7 Lecture 7

Definition 7.1 (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

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

□

Remark 7.2 (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.*
First, we do perform some manipulation that will be useful later:

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

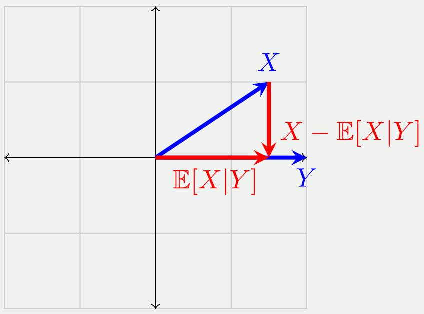
$$= \mathbb{E}[(X - \mathbb{E}[X|Y])^2]$$

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

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 7.3 (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:

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

$$= \mathbb{E}[N\mathsf{Var}(X_i)] + \mathsf{Var}(N\mathbb{E}[X_i])$$

$$= \mathbb{E}[N]\mathsf{Var}(X_i) + \mathbb{E}[X_i]^2\mathsf{Var}(N)$$

Definition 7.4 (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.

Example 7.5. If $Y = aX + b$, then $M_Y(s) = \mathbb{E}[e^{s(aX+b)}] = e^{sb}\mathbb{E}[e^{asX}] = e^{sb}M_X(as)$

Example 7.6 (MGF of exponential RV). Let $X \sim exp(\lambda)$. Then we have that:

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

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

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

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 $s = 0$.

Example 7.7 (MGF of Normal RV)**.** Let $X \sim \mathcal{N}(0, 1)$. Then we have

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

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

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

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

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

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 7.8 (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 7.9 (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 7.10 (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 7.11 (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

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

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

7.1 Bounds

Definition 7.12 (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

□

Remark 7.13 (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 7.14 (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. □

Remark 7.15 (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 $M_n = \frac{X_1+...+X_n}{n}$. 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 7.16 (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.

Remark 7.17 (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.

Example 7.18 (more MGF practice)**.** Let $X \sim exp(1)$, and $Y = aX + b$.

8 Lecture 9

TODO 2. Talk about Convergence here?

Remark 8.1 (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

Definition 8.2 (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

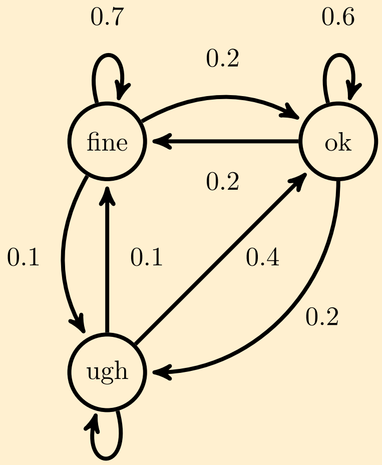
$$Pr(X_{n+1} = c_{n+1}|X_0 = c_0, ..., X_n = c_n) = Pr(X_{n+1} = c_{n+1}|X_n = c_n)$$

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 8.3. 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 8.4 (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

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

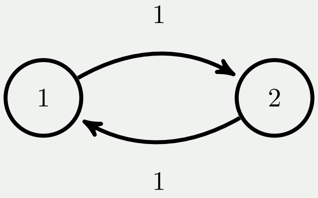
$$= \begin{bmatrix} P_{i1} & P_{i2} & \cdots & P_{im} \end{bmatrix} \begin{bmatrix} P_{1j} \\ P_{2j} \\ \vdots \\ P_{mj} \end{bmatrix}$$

Now further recall from CS70 our transition probability matrix:

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

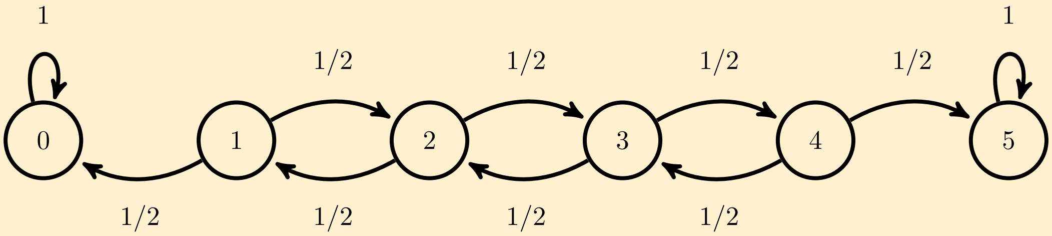
We can then see that $r_{ij}(2) = (P^2)_{ij}$. This is quite convenient! It is also very easy to then see that $r_{ij}(n) = (P^n)_{ij}$, or the $(i,j)^{th}$ entry of P^n. If the values of each column of P^n converge to the same value, then this tells us that no matter where we start out, you have an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

Remark 8.5. 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 8.6 (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 8.7. 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 8.8. The **class** of a state i is $\{j : j$ accessible from i and i accessible from $j\}$

Proposition 8.9. *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 8.10. 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 8.11. *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. □