MATH 2B RECITATION 1/19/12

ALDEN WALKER

1. Random Thoughts

I like this example, because it is one of those things that pretty much everybody can understand, but it is extremely counter-intuitive. Suppose you pick a sequence A of heads and tails (e.g. $A = HTHHT$). You start flipping a coin and look at the last 5 flips. How many flips do you expect to make before the sequence A appears? This time is called the (expected) *waiting time* for the sequence A.

Now let's try something different. I pick a sequence A (e.g. $A = THTH$), and you pick a sequence B (e.g. $HTHH$). We flip a coin, and whichever sequence shows up first wins. What are the odds that A beats B? John Conway made up one of the most clever algorithms I have seen. For sequences X, Y (say of length 4), define the XY as follows: write X above Y. If all the letters match, put a 1 over the first letter of X; if not, put a zero. Next, compare the last three letters in X to the first three in Y (shift Y to the right). If they match, put a 1 over the second letter of X, and so on. After this procedure, you have a string of 1's and 0's. Interpreted as a binary number, this is XY. Here is AB:

$$\begin{array}{ccccc}
0 & 1 & 0 & 1 & = 5 \\
\hline
\mathrm{T} & \mathrm{H} & \mathrm{T} & \mathrm{H} & \\
\mathrm{H} & \mathrm{T} & \mathrm{H} & \mathrm{H} &
\end{array}$$

It turns out that the odds that Y beats X are $XX - XY : YY - YX$. In our case, the odds that B beats A are $AA - AB : BB - BA = 10 - 5 : 9 - 0 = 5 : 9$, so B beats A with probability $5/14 \approx 0.357$.

Here's the weird part: the expected waiting time for A is 20, and the waiting time for B is 18, so B has a shorter waiting time (you expected it to show up sooner), but when we pit A and B against each other, A wins the majority of the time. To see why this might be true, note that when we play two sequences against each other, we are interested in conditional probabilities, but the waiting time doesn't have to do with that. Intuitively, it's very confusing.

2. Random variables

A *random variable* is a formal model of a random event in real life. Technically, a random variable isn't actually a "random" thing — it assigns probability to various values, based on the likelihood that, when "evaluated", or "sampled", it would be that value. This is vague, but accurate. The real definition is hard.

3. Distributions

You're currently learning about distributions. The word *distribution* basically means the collection of functions and pictures which help you understand and compute probabilities involving a random variable. To say that a random variable "has" a distribution means that the probability that that random variable is a certain value is given by the density function associated to a distribution.

Terms:

Probability density function (PDF): The key function that describes how a random variable is distributed is the probability density function. If ϕ is the probability density function for a distribution, then for a random variable X with that distribution, $P(a \leq X \leq b) = \int_a^b \phi(x)dx$.

Cumulative distribution function (CDF): If ϕ is the PDF for a distribution, then the CDF Φ is defined $\Phi(a) = \int_{-\infty}^a \phi(x)dx$. Note this is $\Phi(a) = P(X \leq a)$ for a random variable X following this distribution.

A distribution can be *discrete* or *continuous*. The definitions above work for either, if you allow integration with point masses, so for example the Binomial(n, p) distribution has a PDF with its mass concentrated at the integers. In the case of a discrete distribution, we often define the PDF by giving its point mass at various values (usually the integers).

Note that we must have $\Phi(\infty) = \int_{-\infty}^{\infty} \phi(x)dx = 1$. This corresponds to the fact that a random variable must take *some* value with probability 1.

If you have *any* function ϕ whose total integral is 1 (and ϕ isn't stupid or silly in some way), then ϕ *is* the PDF for a distribution, and it defines the behavior of a random variable. This distribution might not have a name, though. The only thing which is special about the distributions with names is that they are nice and tend to describe most things that we need to describe. If a new distribution comes along which somebody discovers a nice use for, we give it a name.

3.1. **Parameters.** Sometimes, many distributions are related because they look and behave similarly. For example, a normal distribution with mean 0 and a normal distribution with mean 1. For this reason, we give a single name to a whole family of distributions. Consider a normal distribution with mean μ and standard deviation σ. Here μ and σ are the parameters, and they give a whole family of individual distributions.

3.2. **Example.** What's the distribution of a random variable X modelling a single coin flip? We identify H with 0 and T with 1. Then the PDF is two delta functions at 0 and 1, each with value 1/2. More commonly, we will state this has $P(X = 0) = 1/2$ and $P(X = 1) = 1/2$.

3.3. **Example.** I tell you that I've got a random variable X with the distribution given by the PDF defined $\phi(x) = 1/x^2$ if $x > 1$ and 0 otherwise. What's $P(X > 5)$?

This is simply $\int_5^{\infty} 1/x^2 dx = 1 - \int_1^5 1/x^2 dx = 4/5$.

3.4. **Example.** A random variable X has a Poisson distribution with mean λ. What's $P(X = 10)$.

Giving the distribution for X by name means that we know the PDF. We consult it to see that $P(X = 10) = e^{-\lambda}\lambda^{10}/10!$.

4. Joint Distributions and Functions

It is very common to create a joint distribution, which is just the distribution of the pair (X, Y) given two random variables X and Y. You can of course recover the original distribution of X or Y by summing over all possibilities for the other variable. You can do the same thing with more than two variables.

You can also define functions of random variables, as in: if X is the number rolled on a die, what is the distribution of X^2. Page 154 contains facts that will be helpful to you:

- Functions of independent random variables are independent.
- Disjoint blocks of independent random variables are independent, e.g. if all X_i are independent, then (X_1, X_2) and (X_3, X_4) are independent.
- Functions of disjoint blocks of independent random variables are independent: with the same setup as above, X_1X_2 and X_3X_4 are independent

4.1. **Example (Indicator Functions).** Let X_i be the outcome of flipping a coin on the ith trial, where heads is indicated by 1 and tails by 0. Then note that the number of heads in 5 flips is indicated by $Y = X_1 + \cdots + X_5$. Note that the distribution of Y is the binomial $(5, 0.5)$ distribution (basically by definition).

4.2. **Example (Joint Distributions).** What is the probability that if two integers X, Y less than or equal to 100 are picked at random we have $X \times Y \leq 50$? Here we are asking about a probability dealing with the ordered pair (X, Y), so we are really interested in the joint distribution of (X, Y). Let B be the set of all such ordered pairs. We can write the answer as $P((X, Y) \in B) = \sum_{k=1}^{100} P((X, Y) \in B \,|\, Y = k)P(Y = k)$. Now,

$$P((X, Y) \in B \,|\, Y = k) = P(X \leq 50/k) = \frac{\lfloor 50/k \rfloor}{100}$$

So we've got

$$P((X, Y) \in B) = \sum_{k=1}^{100} \frac{1}{100} \frac{\lfloor 50/k \rfloor}{100} \approx 0.0207$$

5. Expectation

This is exactly what you think it should be. Expectation is defined $E(X) = \sum_x xP(X = x)$. The mean of a distribution is defined $\mu = \sum_x xP(x)$. In the case that X does not take discrete values (for a continuous distribution, like the normal), then we understand this to mean $\mu = \int_{-\infty}^{\infty} x\phi(x)dx$, where ϕ is the PDF. Note that if you "pick a variable from a distribution," the expected value of the variable is the mean of the distribution. Expectation has some nice properties:

- $E(X_1 + \cdots + X_n) = \sum_i E(X_i)$, even if they are not independent!
- If X_1 is an indicator function for an event A, $E(X_1) = P(A)$.
- Tail sum formula (for X which takes values in $\{0, \dots, n\}$): $E(X) = \sum_{j=1}^n P(X \geq j)$.
- $E(g(X)) = \sum_x g(x)P(X = x)$
- $E(XY) = E(X)E(Y)$ if X and Y are independent.

Have a look at the tables on pages 180 and 181 in your book for more exciting facts.

It is always possible just to go back to the definition to figure out an expected value, but very often, the addition law makes things **much** easier!

5.1. **Example.** If I roll two dice, what is the total expected number shown? This is just asking for $E(X+Y)$, where X and Y are the values of each die. By the addition formula above, we know that the expected value is $E(X + Y) = E(X) + E(Y) = 3.5 + 3.5 = 7$. Note that $E(X)$ is easy to calculate as $(1/6)\sum_{i=1}^6 i$.

5.2. **Example (E is not probability).** Suppose I flip a coin. Then the expected number of heads is 1/2 and the probability of getting heads is 1/2, the same. This does not continue, obviously, but it can lead to some confusing situations. If you flip a coin twice, what is the probability of seeing at least one head? Since expectation adds, the expected number of heads is 1. However, the probability of seeing a head is NOT one of course—it's 3/4.

The fact that the expectation can be a real number between 0 and 1 can be used to make fallacious arguments about it being a probability—don't be fooled!

5.3. **Example.** What is your expected payoff if you play a fair game of three card monte in which a \$1 bet gets you \$3 if you win? Your probability of winning is 1/3, so letting X be the money you have won after a single hand, we have $E(X) = (1/3)(2) + (2/3)(-1) = 0$. In other words, you actually don't expect to lose money! However, no fair game of three card monte exists, so don't get too excited.

5.4. **Example.** Suppose 5 people draw a ball out of bag of 10 balls, and they each replace it before the next person draws. How many distinct balls do we expect to be drawn (you would think it would be slightly less than 5 because of the chance of overlap).

Let B_i be the indicator function for ball i, i.e. it is 1 if ball i is drawn (by anybody), and 0 otherwise. Then $E(B_i) = 1 \times P(B_i = 1) + 0 \times P(B_i = 0) = 1 - \left(\frac{9}{10}\right)^5 = 0.4095$.

Now the total number of balls drawn is $\sum_{i=1}^{10} B_i$, and by the summation formula (which works even though the fact that ball i was drawn is not independent of the fact that B_j was drawn),

$$E(\sum_i B_i) = \sum_i E(B_i) = 10 \cdot 0.4095 \approx 4.1$$

Which coincides with our intuition. Notice that if the number of balls is n, then the expected number of distict balls picked is

$$n\left(1 - \left(\frac{n-1}{n}\right)^5\right) = \frac{n^5 - (n-1)^5}{n^4} = 5 - O(\frac{1}{n})$$

So the number of balls picked goes to 5, as we would expect.

5.5. **Example (Just checking).** Let's verify the expected value of a Poisson random variable X with mean λ is, in fact, λ. We have

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} \\ &= e^{-\lambda}\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} \\ &= e^{-\lambda}\lambda e^{\lambda} \\ &= \lambda \end{aligned}$$

5.6. **Example (E and Functions).** It is not true in general that $E(g(X)) = g(E(X))$, even though that would be nice. Let's do an example with this.

Suppose that someone is going to choose n points in the interval $[0, 1]$ uniformly at random. You want to record which points these are (approximately) using as little space as possible. In fact, you only have a single bit (0 or 1) to record these points. Here are the rules: before the points are chosen, you get to pick a subset A of $[0, 1]$ (WLOG we may assume they are intervals) for the 0 state of your bit to correspond to. The 1 state must correspond to the full interval $[0, 1]$. If all the points fall into A, you get to record those points with a 0. If any point falls outside of A, you must return 1. Call the outcome of determining whether to return A or $[0, 1]$ given n points by $f(X_1, \ldots, X_n)$. Here is the main point: you want the returned state to have as small a measure as possible, specifically, you want to minimize $E = E(m(f(X_1, \ldots, X_n)))$.

You can see the tradeoff: if you make A large, it is more likely that you will return A and the measure will be smaller than 1. However, the larger A gets, the worse the gain for returning A is.

Let's try to find an expression for E in terms of n and the measure of A and see if we can minimize it.

$$m(f(X_1, \ldots, X_n)) = \begin{cases} 1 & \text{if not all } X_i \text{ are in } A \\ m(A) & \text{if all } X_i \in A \end{cases}$$

Then we can write out the formula for expected value:

$$E(m(f(\{X_i\})) = 1P(f(\{X_i\}) = [0, 1]) + m(A)P(f(\{X_i\}) = A)$$

Now, the X_i are all independent, so

$$P(f(\{X_i\}) = A) = P(X_1 \in A, \ldots, X_n \in A) = P(X_1 \in A) \cdots P(X_n \in A) = m(A)^n$$

Then $P(f(\{X_i\}) = [0, 1] = 1 - m(A)^n$. Therefore, if we let $x = m(A)$, we see that

$$E(m(f(\{X_i\})) = (1 - x^n) + x(x^n) = x^{n+1} - x^n + 1$$

This is a polynomial, so to minimize it we must only look for zeros of the derivative, which is $(n+1)x^n - nx^{n-1}$. Solving, we get $x = \frac{n}{n+1}$. The second derivative at this point is $\frac{n^{n-1}}{(n+1)^{n-2}}$, which is positive, so we really have found a minimum.

In recitation, this is all that I will talk about, so I'm putting a page break here; however, there is a super-cool generalization of this which starts on the next page (available online):

Let's take it one step further and think about what happens if the number of points that we are given is now random under a Poisson distribution of known mean λ. In that case, the setup is the same, except now the number of points, K, is random:

$$P(f(X_1,\ldots,X_K)=A)=\sum_{k=0}^{\infty}P(K=k)P(f(X_1,\ldots,X_k)=A)=\sum_{k=0}^{\infty}e^{-\lambda}\frac{\lambda^k}{k!}m(A)^k=e^{\lambda(m(A)-1)}$$

and clearly then $P(f(X_1,\ldots,X_K)=[0,1])=1-e^{\lambda(m(A)-1)}$, so again letting $x=m(A)$,

$$E=1\left(1-e^{\lambda(x-1)}\right)+xe^{\lambda(x-1)}$$

Again, we differentiate and get $\lambda x e^{\lambda(x-1)}+e^{\lambda(x-1)}-\lambda e^{\lambda(x-1)}$, which is easy to solve for $x=\frac{\lambda-1}{\lambda}$, which is a minimum. Here is a plot with a comparison between the optimal measure with a given number of points and a random number of points:

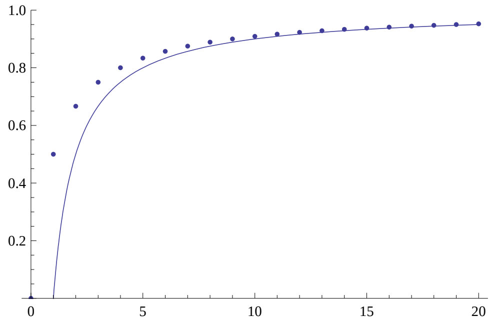

FIGURE 1. Optimal measure of A as a function of n (dots) and as a function of λ

The fact that the solid line lies below the dots makes sense because the Poisson distribution tends to bunch up to the left of the mean, i.e. getting fewer points than the mean is more likely that getting more points. Therefore, it is better to make A smaller.

This problem comes from an actual real-life problem involving how best to store data efficiently in a small amount of space.