Massachusetts Institute of Technology | Handout 58
6.042J/18.062J: Mathematics for Computer Science | November 30, 1995
Professors David Karger and Tom Leighton

Notes 23

1 Expectations of Products

First let's recall some basic facts about expectations from last time.

1. $E[R_1 + R_2] = ER_1 + ER_2$. (*Linearity of expectation.*)
2. $E[aR] = a \cdot ER$, where a is a scalar.
3. $E[R + a] = ER + a$, where again a is a scalar.

Today, we'll look at the expectation of the *product* of two random variables. Unlike the case of sums, the identity $E[R_1R_2] = E[R_1] \cdot E[R_2]$ for products does not necessarily hold. However, it does hold in the important special case when R_1 and R_2 are independent.

Theorem 1.1 *Let R_1 and R_2 be independent random variables. Then $E[R_1R_2] = E[R_1] \cdot E[R_2]$.*

Proof. We have

$$\begin{aligned} E[R_1R_2] &= \sum_x x \cdot Pr(R_1R_2 = x) \\ &= \sum_{x_1}\sum_{x_2} x_1x_2 Pr(R_1 = x_1 \text{ and } R_2 = x_2) \\ &= \sum_{x_1}\sum_{x_2} x_1x_2 Pr(R_1 = x_1) \cdot Pr(R_2 = x_2). \end{aligned}$$

The last equality follows from the independence of R_1 and R_2 (this is where we needed it). Now, we can factor this expression as

$$\left(\sum_{x_1} x_1 \cdot Pr(R_1 = x_1)\right)\left(\sum_{x_2} x_2 \cdot Pr(R_2 = x_2)\right).$$

Now the first of these terms is, by definition, ER_1, and the second of these terms is ER_2. Thus we have

$$E[R_1R_2] = E[R_1] \cdot E[R_2].$$

■

So for example, let's suppose we roll two fair, 6-sided dice whose outcomes are independent. Then if we let R_1 denote the value of the first die, and R_2 denote the value of the second die, then

$$E[R_1R_2] = E[R_1] \cdot E[R_2] = \frac{7}{2} \cdot \frac{7}{2} = \frac{49}{4}.$$

Now, what if the two dice are rigged so that they always give the same value as one another? In this case, R_1 and R_2 always assume the same values, and we have

$$E[R_1R_2] = \sum_{i=1}^{6} i^2 \cdot Pr(R_1 = i) = \frac{91}{6} \neq E[R_1]^2.$$

So this is a case of two dependent random variables for which the expectation of their product is not the product of their expectations.

Finally, by induction on n, it is easy to prove the following corollary of Theorem 1.1

Theorem 1.2 *If $R_1, \ldots, R_n$ are mutually independent, then*

$$E[R_1R_2 \cdots R_n] = (ER_1)(ER_2) \cdots (ER_n).$$

2 Conditional Expectation

Conditional expectation is often useful in doing a "case analysis" to compute an expectation.

Definition 2.1 *The conditional expectation of X given an event A is defined as*

$$E[X \mid A] = \sum_j j \cdot Pr[(X = j) \mid A].$$

Example 1 Let S be the number of spots we see in a typical die roll.

$$\begin{aligned} E[S \mid S = 4] &= 4 \\ E[S \mid S \leq 4] &= 2.5 \end{aligned}$$

Suppose there are two assembly plants F and S for computers. Plant F is fast but sloppy: they make 70% of the computers, but a computer from their plant has a 2% probability of breaking down on any given day of use. Plant S is slow and careful: they make 30% of the computers, but the breakdown probability is only 1%. These computers get randomly mixed, so I buy one at random from the store. What is the expected number of days my computer works?

Let D be the number of days my computer works. Let I_S be the event that I bought the computer from S, and let I_F be that I bought it from F.

What is the expected working time from the slow factory? It's determined by MTF calculation on one of the computers from S, that is, a negative binomial distribution with failure probability $p = 1\%$.

$$\begin{aligned}
E[D \mid I_S] &= \sum_k k \cdot \Pr[D = k \mid I_S] \\
&= \sum_k k \cdot (1-p)^k p \\
&= \frac{1-p}{p} \\
&= 99
\end{aligned}$$

One the other hand, the expected working time given I_F is 49 days.

But what we really want to know is the expected working time of my computer when I don't know which factory it came from. Fortunately, there is a useful analogue to the theorem of total probability:

Theorem 2.2 *Let $\{A_i\}$ be a set of disjoint events that exhaustively partition the sample space. Then for any random variable R,*

$$E[R] = \sum_i E[R \mid A_i] \Pr[A_i]$$

Note that the theorem of total probability follows as a special case if we take R to be the indicator variable for an event whose probability we want to determine.

Proof.

$$
\begin{aligned}
E[R] &= \sum_k k \cdot \Pr[R = k] \\
&= \sum_k k \cdot \left(\sum_i \Pr[R = k \mid A_i] \Pr[A_i] \right) \qquad \text{(theorem of total probability)} \\
&= \sum_i \left(\sum_k k \cdot \Pr[R = k \mid A_i] \right) \Pr[A_i] \\
&= \sum_i E[R \mid A_i] \Pr[A_i]
\end{aligned}
$$

■

It follows that in for our computer,

$$
\begin{aligned}
E[D] &= .3 \cdot 99 + .7 \cdot 49 \\
&= 64
\end{aligned}
$$

Sometimes people talk about the conditional expectation of X given a random variable Y, written $E[X \mid Y]$. This quantity is a *random variable.* To compute its value at as point s in the sample space, we first compute the value of Y there, say $Y(s) = k$, and then compute the expectation $E[X \mid Y = k]$.

Theorem 2.3 $E[E[X \mid Y]] = E[X]$

3 Wald's Theorem

Suppose I flip 3 standard coins and then roll a die for each coin that comes up heads. How many spots do I expect to see? Well, if H is a random variable denoting the number of heads I see, and $D_1, D_2, \ldots$ are random variables for die rolls, then the random variable describing the number of spots can be written as

$$
S = \sum_{i=1}^{H} D_i.
$$

This sum is a bit strange, because the upper summation index is a random variable? How can I compute it's expectation? One way is by conditional expectations:

$$\begin{aligned}
E[S] &= \sum_{k=0}^{3} E[\sum_{i=1}^{H} D_i \mid H = k] \Pr[H = k] \\
&= \sum_{k=0}^{3} E[\sum_{i=1}^{k} D_i] \Pr[H = k] \\
&= \frac{3}{8}(E[D_1]) + \frac{3}{8}(E[D_1] + E[D_2]) \\
&\quad + \frac{1}{8}(E[D_1] + E[D_2] + E[D_3]) \\
&= \frac{3}{8}3.5 + \frac{3}{8}7 + \frac{1}{8}10.5 \\
&= \frac{21}{4}
\end{aligned}$$

Is there an easier way to figure this out? Well, $E[H] = \frac{7}{2}$, and $E[D] = \frac{3}{2}$, so the expectation turns out to be $E[H] \cdot E[D]$. But why?

What about the rule for expectation sums? I have a problem: I don't know how many copies of $E[D]$ to add up!

Could it be the product rule for expectations? No, this would apply if I multiplied the number of heads by the outcome of rolling one die. Here, I roll multiple dice rather than multiplying the outcome of one die.

To deduce my result, I need to prove a general result called *Wald's Theorem.*

Theorem 3.1 *Let R be a random variable, Q a random variable independent of R, and let T be the sum of Q samples of R—that is,*

$$T = \sum_{i=1}^{Q} R_i.$$

Then $ET = E[Q] \cdot E[R]$.

We have already observed this in the die-rolling example.

Proof. Picture the random experiment as follows. First we generate an infinite number of samples of R, call them $R_1, R_2, \ldots$; then we use Q to decide how many of these copies to add up in computing T. This experiment works on a different sample space than the original, but gives the same distribution. Then:

$$
\begin{aligned}
E[T] &= E[\sum_{i=1}^{Q} R_i] \\
&= \sum_k E[\sum_{i=1}^{Q} R_i \mid Q = k] \cdot \Pr[Q = k] \\
&= \sum_k E[\sum_{i=1}^{k} R_i] \cdot \Pr[Q = k] \qquad (*) \\
&= \sum_k \left(\sum_{i=1}^{k} E[R_i] \right) \cdot \Pr[Q = k] \\
&= \sum_k \left(\sum_{i=1}^{k} E[R] \right) \cdot \Pr[Q = k] \\
&= \sum_k (kE[R]) \cdot \Pr[Q = k] \\
&= E[R] \sum_k k \cdot \Pr[Q = k] \\
&= E[R]E[Q]
\end{aligned}
$$

■

Where in this proof did I use the hypothesis that Q and the R_i are independent? When I deduced at the line labeled (*) that

$$E[\sum_{i=1}^{Q} R_i \mid Q = k] = E[\sum_{i=1}^{k} R_i].$$

If the R_i are dependent on Q, this might not work.

Example 2 Consider random variables R and Q such that R is always equal to Q, and Q takes values 1 and 2 with equal probability. Then

$$E(T) = E(\sum_{i=1}^{Q} R_i) = (\frac{1}{2})(1 + (2 + 2)) = 2.5 \neq 2.25 = (1.5)(1.5) = E(Q)E(R)$$

4 An Application to Branching Processes

Branching processes are used to model population growth, nuclear chain reactions, mutations, and contagious diseases.

Suppose someone is infected with a new disease. Every day, he encounters $n = 10$ healthy people and infects each of them with probability $p = 5\%$. What is the expected number of infected people after k days?

Let S_k be the number infected after k days. We can write a recurrence for S_{k+1} in terms of S_k. Let the sick people at day k be numbered $1, \ldots, S_k$. Let H_i denote a variable distributed as the number of healthy people infected by sick person i. H_i has a binomial distribution with parameters n and p. Then:

$$S_{k+1} = S_k + \sum_{i=1}^{S_k} H_i$$

It follows from linearity of expectation that

$$E[S_{+1}] = E[S_k] + E[\sum_{i=1}^{S_k} H_i].$$

Then by Wald's theorem,

$$E[S_{k+1}] = E[S_k] + E[S_k] \cdot E[H]$$

where H is a variable with the binomial distribution with parameters n and p, so that $E[H] = np$. Therefore, if we let $e_{k+1} = E[s_{k+1}]$, we have $e_0 = 1$, and

$$\begin{aligned} e_{k+1} &= e_k + e_k np \\ &= (1 + np)e_k \end{aligned}$$

It follows that $e_k = (1 + np)^k$.

Eventually, we will have more sick people than people. What erroneous assumption is in our model? That each sick person meets a different collection of 10 healthy people. Once there are enough sick people, it will become less likely for a sick person to meet so many healthy people. Also, many sick people will encounter the same healthy person. But our "approximation" will be valid as long as the sick population is not too big.

Now suppose any sick person recovers from the disease each day with probability r. Let B_i be the indicator variable for whether the i^{th} sick person gets better. Then $E[B_i] = r$. Then

$$S_{k+1} = S_k + \sum_{i=1}^{S_k} (H_i - B_i)$$

It follows that

$$E[S_{k+1}] = E[S_k] + E[S_k] \cdot E[H] - E[S_k]E[B_i]$$

and so

$$\begin{aligned} e_{k+1} &= e_k(1 + np - r) \\ &= (1 + np - r)^k \end{aligned}$$

Therefore, the progress of the disease is determined entirely by the parameter $(1+np-r)$. If this exceeds 1, it looks like we have an epidemic on our hands (but expectation isn't telling the whole story: there's a chance the first person gets better without making anyone sick). If less than 1, it looks pretty innocuous. In fact, next time we will see that if less than 1, the disease almost certainly dies out after a certain number of days.

5 Coupon Collecting

Suppose there are n different baseball cards. I buy them one at a time and stop when I get one of each. How long do I expect to continue?

Let W_k be the time I wait until I get the k^{th} card, and let $X_k = W_{k+1} - W_k$. Then clearly,

$$W_n = \sum X_k$$

and therefore the waiting time is

$$\sum E[X_k].$$

To determine $E[X_k]$, consider the "experiment" being performed during this k^{th} time interval. I have k cards, and I want to find the $(k+1)^{st}$. Each time I buy a card, my probability of getting a new card is $p = (n-k)/n$. Each attempt is independent, so I have a sequence of Bernoulli Trials. By the previous example on negative binomial distributions ("mean time to failure," only now it is success) my expected waiting time is

$$\begin{aligned}\frac{1}{1-p} &= \frac{1}{1-k/n} \\ &= \frac{n}{n-k}\end{aligned}$$

It follows that the expected waiting time to get all the cards is

$$\begin{aligned}\sum_{k=1}^{n-1} \frac{n}{n-k} &= n\sum_{k=1}^{n-1}\frac{1}{k} \\ &\approx n \ln n\end{aligned}$$

where the last approximation comes from our recognition of the *Harmonic series.*

(So to get all 1000 distinct baseball cards, we expect to buy about 7000).

We were actually a bit sloppy in our analysis of the coupon collecting problem. When we talked about the experiment in the k^{th} time interval, we referred to the trials being done in this interval. But which trials in our sequence of purchases make up the k^{th} experiment? The ones after we got k cards. It follows that we are actually conditioning on a particular value of W_k. In other words, we are computing $E[X_k]$ using conditional expectations! What saves us is that $E[X_k]$ turns out to be independent of W_k:

$$\begin{aligned} E[X_k] &= \sum_z E[X_k \mid W_k = z] \cdot \Pr[W_k = z] \\ &= \sum_z \frac{n}{n-k} \Pr[W_k = z] \\ &= \frac{n}{n-k} \sum_z \Pr[W_k = z] \\ &= \frac{n}{n-k} \end{aligned}$$

6 Some More Examples for Tutorial

6.1 A two-player game

Here's a game you don't want to play unless you know what you're doing. Two players, A and B, agree on a probability distribution $f(x)$ for a random variable X over $\mathbb{N}$. That is, $f(x) = Pr(X = x)$.

Now player A picks a number a at random according to f. Then player B independently picks a number b according to f; if $b \leq a$, then B repeats this until she gets a number $b > a$. Each time through the loop, B pays A \$1.

What is B's expected loss? The answer is ∞.

To get some intuition for this, consider first the finite case: $f(x)$ is non-zero only for a finite set of numbers. Let t be the largest such number. Then with positive probability, A will pick t, in which case B can never pick a bigger number and so her loss will be infinite.

More precisely, if R is a random variable denoting the number of times B pays \$1, then we have

$$ER = \sum_{x \in \mathbb{N}} Pr(R > x).$$

But $Pr(R > x) \geq f(t)$ for all $x \geq t$, so this sum is infinite.

Now let's consider the case of an arbitrary distribution function f. Here, there may be no "biggest number" t, so we have to use a different argument.

Consider the game as follows: A draws the first number, and then B draws infinitely many more numbers, paying \$1 until a number comes up that's bigger than the first number. So B will pay at least \$m if among the first m numbers, the first is biggest.

For a fixed m, let A denote this event: that the first of m numbers drawn according to f is the biggest. We claim that $Pr(A) = \frac{1}{m}$, regardless of f. Intuitively, this is because only the order in which the numbers arrive matters. We can prove this as follows:

The sample space is the set S of all m-vectors of numbers. Let's say that two such vectors are equivalent if one is a permutation of the other. This gives a partition of S into equivalence classes: $S = \cup_i M_i$ and $\sum_i Pr(M_i) = 1$.

Now for each i, we have $Pr(A \mid M_i) \geq \frac{1}{m}$, since one we've fixed the m numbers, all orders are equally likely. (We don't necessarily have equality since numbers can be repeated.) Now we have

$$\begin{aligned} Pr(A) &= \sum_i Pr(M_i) \cdot Pr(A \mid M_i) \\ &\geq \frac{1}{m} \sum_i Pr(M_i) \\ &= \frac{1}{m}. \end{aligned}$$

Finally, we can compute the expected loss:

$$ER = \sum_{m \in \mathbb{N}} Pr(R > m) \geq \sum_{m \in \mathbb{N}} \frac{1}{m} = \infty.$$

6.2 Another two-player game

You remember Monty Hall? Well, not satisfied with the money *Let's Make a Deal* is making in syndication, he's turned to other sorts of games. Here's one:

There are two envelopes, one containing twice as much money as the other. He gives you one, but he does not allow you to open it. He then asks whether you want to keep the envelopes, or switch. Should you?

Here's some incorrect reasoning: lets say you have $\$x$ in your current envelope; so with probability $\frac{1}{2}$ the other envelope has $\$(2x)$, and with probability $\frac{1}{2}$ the other envelope has $\$x/2$. So your expected gain on swapping is $x/2 - x/4 = x/4$ — that is, you should always swap!

But this argument is incorrect, and the reason is that the probability is not 50-50 that the other envelope has either twice or half as much. For example, suppose I randomly put \$1 and \$2 in two envelopes. You open one and it contains \$2. What is the probability that the other envelope contains \$4? It's zero.

A correct argument, which gives the intuitively reasonable answer that it makes no difference whether you switch, is the following. One envelope has $\$y$ and the other has $\$2y$. So if you open an envelope and see $\$x$, then either $x = y$ with probability $\frac{1}{2}$, or $x = 2y$ with probability

$\frac{1}{2}$. So in the first case your gain on switching is x, and in the second your gain is $-x$. Thus your expected gain is zero.

Now Monty allows you to open the envelope in hand, revealing $\$x$. Again he asks whether you want to keep the money, or switch. Given this extra information - the actual number x - should you switch or not. Intuitively, it would seem that knowing the number x does not give you any edge and so it shouldn't matter whether you switch or not. But, in actuality there is a strategy which allows you to pick the larger of the two sums with probability better than $\frac{1}{2}$. The strategy is as follows - choose any distribution which has continuous support over the reals, like the normal distribution, and pick a number r from it, if $r > x$ then switch, otherwise stay. Note that whenever r is smaller than the lesser of the two sums or bigger than the greater of the two then we end up with the larger sum with probability $\frac{1}{2}$, but whenever r lies in between the two sums then we always end up with the larger sum. Thus we do better than $\frac{1}{2}$, since there is a non-zero probability of r landing exactly between the two sums. Note, however, that if we use the normal distribution to pick r then if the sums in question grow very large our probability of winning the larger amount goes to $\frac{1}{2}$. In fact it can be shown that there is no distribution that can guarantee us a winning probability greater than $\frac{1}{2}$ by any fixed constant.

Of course, all this assumes that Monty offers the switch regardless of how much money he sees you holding. Where do we use this fact?