15-251 Recitation notes 4/07/03 **Guru: Avrim Blum**

Plan for today: some interesting things about random walks. Chebyshev's inequality. In the process, getting more practice with random variables.

"A drunk man will find his way home, but a drunk bird may get lost forever"
– Shizuo Kakutani

---

1 Random walks

Consider a random walk on a line. You start at the origin, and at each time step you flip a coin, going one step to the right if you get a heads and one step to the left if you get a tails. What is the chance you will eventually get back to the origin? It turns out that, in fact, with probability 1 you will eventually return. Same in 2-dimensions. But in 3-d, there is a positive chance you will never come back. (See quotation above.)

We're going to prove all this by first showing the following claim.

Claim 1 *Suppose the probability of* *not* *returning to the origin is some $p > 0$. Then the* *expected* *number of visits to the origin will be $1/p$.*

Proof: Each time we leave the origin, with probability p we never return, and with probability $1 - p$ we do. If we do return, then at that point we are again in exactly the same situation. So we can think of this like flipping a coin of bias p, where if it comes up tails we get to flip again, but if it comes up heads, the game is over. We are asking for the expected number of flips we make. But we have already solved this (in fact, you did it on your homework!) and we know the answer is $1/p$. So, that is the expected number of total visits to the origin (counting our starting there as a visit too). ■

1.1 The 1-dimensional case

Let's calculate the expected number of visits to the origin in the first $2n$ steps. (Using "$2n$" since we can only be at the origin at even time steps...). Let's use Y_n for the random variable "number of visits to the origin in the first $2n$ steps". Can you see any way to break Y_n down into a sum of simpler indicator variables?

Here is a good way: let A_t be the event that we are at the origin at time $2t$. Let X_t be the indicator RV for A_t. So, $Y_n = \sum_{t=0}^{n} X_t$. Also, $\mathbf{E}[X_t] = \Pr(A_t)$ is something we know how to calculate. What is it? $[\binom{2t}{t}/2^{2t}.]$

Also, on Homework 5, you showed using Stirling's formula that this is $\Theta(1/\sqrt{t})$. This means that:

$$\begin{aligned}\mathbf{E}[Y_n] &= \Theta(1/\sqrt{1} + 1/\sqrt{2} + \ldots + 1/\sqrt{n}) \\ &= \Theta(\sqrt{n}).\end{aligned}$$

Now, I claim we have proven the result that the we return to the origin with probability 1. If that were false, there would be some probability $p > 0$ we never return, so the expected number of returns would be some finite number $1/p$. But, we've just shown that the expected number of returns *by time* $2n$ is $\Theta(\sqrt{n})$ which goes to infinity as $n \to \infty$. So, that can't be the case.

1.2 The 2-dimensional case

To make things simpler, let's model the 2-dimensional walk by saying that at each time step, you move both in the x-direction and in the y-direction. In other words, you can think of your x-coordinate as a random walk, and you can think of your y-coordinate as an independent random walk, and both are going on simultaneously. You will be at the origin only if both random walks happen to hit the origin at the same time.

Let's again define A_t as the event that we are at the origin at time $2t$. What is $\Pr(A_t)$? [It is $\Theta((1/\sqrt{t})(1/\sqrt{t})) = \Theta(1/t)$.]

So, the expected *number* of visits to the origin by time $2n$ is

$$\begin{aligned} \mathbf{E}[Y_n] &= \Theta(1/1 + 1/2 + 1/3 + \ldots + 1/n) \\ &= \Theta(\log n). \end{aligned}$$

(I don't remember if we did this summation in class, but if we didn't, you can think of $\sum_{i=1}^{n} 1/i$ just like the integral of $1/x$ from 1 to n.)

In any case, we again would have a contradiction if there was a positive probability $p > 0$ of never returning, because $\log(n)$ goes to infinity as $n \to \infty$.

1.3 The 3-dimensional case

Now we have $\Pr(A_t) = \Theta(1/t^{3/2})$. But in this case, $\sum_{t=1}^{\infty} 1/t^{3/2}$ is finite. So, there is some finite number K such that $\lim_{n\to\infty} \mathbf{E}[Y_n] < K$. But, by Claim 1, this means that there has to be at least a $1/K$ chance of never returning: if the chance of never returning was less than $1/K$ then the expected *number* of returns would be greater than K.

2 Chebyshev's inequality, standard deviations, and the shape of the binomial

Because it is so important, let's try to develop a feel for what the (unbiased) binomial distribution looks like. In other words, let's say we have walked for n steps. We know our expected position is the origin, but what can we say about how close we will typically be? The answer is that we will typically be within $\pm\sqrt{n}$. To talk about this formally, let's introduce the notions of *variance* and *standard deviation*.

The variance of a random variable X is defined as $\mathbf{var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2]$. In other words, on average, what is the square of your distance to the expectation. We can multiply this definition out to get:

$$\mathbf{var}[X] \;=\; \mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2] \;=\; \mathbf{E}[X^2] - (\mathbf{E}[X])^2,$$

which was the alternative definition given in class. The standard deviation $\sigma(X)$ is just defined to be the square root of the variance. A useful probabilistic fact is:

Claim 2 (Chebyshev's inequality) *Let X be a random variable with expectation μ and standard deviation σ. Then for any $t > 0$ we have:*

$$\Pr(|X - \mu| > t\sigma) \leq 1/t^2.$$

In other words, for any random variable, you will typically be within a couple standard deviations of the expectation. Let's prove this fact, and then use it to argue about what we will see in a random walk.

Proof: First of all, if Y is a non-negative random variable and k is some positive number, we know that $\Pr(Y > k\mathbf{E}[Y]) < 1/k$. (Because otherwise, this fact by itself would bring the expectation up too high).

Now, let's just define the random variable $Y = (X - \mu)^2$. By definition, $\mathbf{E}[Y] = \mathbf{var}[X]$, right? So, $\Pr(Y > t^2\mathbf{E}[Y]) < 1/t^2$. But, the event that $Y > t^2\mathbf{E}[Y]$ is the *same* as the event that $|X - \mu| > t\sigma$. So, $\Pr(|X - \mu| > t\sigma)$ is also less than $1/t^2$. ■

Now, what is the variance for our n-step random walk? Let X be the position of our walk after n steps. Let's define $X_i = 1$ if we took a step to the right in our ith step, and $X_i = -1$ if we took a step to the left on our ith step. So $X = X_1 + \ldots + X_n$, and $\mathbf{E}[X_i] = 0$.

There are two parts to variance. The easy part is $(\mathbf{E}[X])^2$, which is 0 since $\mathbf{E}[X] = 0$.

What about $\mathbf{E}[X^2]$? We can multiply it out and get $\mathbf{E}[X^2] = \mathbf{E}[\sum_i \sum_j X_i X_j] = \sum_i \sum_j \mathbf{E}[X_i X_j]$. So, what is $\mathbf{E}[X_i X_j]$? [If $i = j$ then $X_i X_j$ is always 1, so $\mathbf{E}[X_i X_j] = 1$. If $i \neq j$ then all four possibilities $(1,1), (1,-1), (-1,1), (-1,-1)$ are equally likely, so in this case $\mathbf{E}[X_i X_j] = 0$.]

This means that $\mathbf{E}[X^2] = n$, so $\mathbf{var}[X] = n$. [In fact, more generally, if $X_1, \ldots, X_n$ are pairwise-independent random variables, then the variance of the sum is the sum of the variances.] Since standard-deviation is the square-root of variance, we have that for our random walk, the standard deviation $\sigma(X) = \sqrt{n}$.

So, combining with Chebyshev's inequality, we have that we will typically be within a couple $\sqrt{n}$'s of the origin.