CS174 Lecture 8 John Canny

More on Coupon Collecting

Recall that coupon collecting is equivalent to placing m balls in n bins so that no bin is empty. Last time we derived an upper bound for the probability that some bin is empty which is

$$\Pr[\text{some bin empty}] < n \exp(-m/n)$$

Then we showed that if you fix this probability and rearrange the equation, the number of balls you need is:

$$m > n \ln(n) + \Omega(n)$$

That also implies the result that we were interested in about stable marriages: After about $n \ln n$ proposals under the random algorithm, every female will have been proposed to, and a stable marriage will result. That means if you started with random permutations as preferences, the proposal algorithm would run for at most $n \ln n$ steps. It doesnt tell us however, what the lower bound on the running time is, or the expected number of rounds. We tackle that next:

Expected number of rounds to collect all coupons

Now lets turn to an analysis of the expected value of m to hit all of the bins (or collect all coupons). Let X be the number of balls placed when the last bin is hit, and:

Let X_0 be the number of trials til the first bin is hit (=1).
Let X_1 be the number of trials after the 1st bin is hit until the 2nd bin is hit.
Let X_2 be the number of trials after the 2nd bin is hit until the 3rd bin is hit.
$\vdots$
Let X_{n-1} be the number of trials after the (n-1)st bin is hit until the nth bin is hit.

Then the total number of trials is:

$$X = \sum_{i=0}^{n-1} X_i$$

The trials counted in each X_i are called epochs. In epoch i (corresponding to X_{i-1}), the probability that any trial hits the next bin is

$$p_i = \frac{n-i}{n}$$

That's because there are i bins that have already been hit, and $n - i$ that havent. X_i is called a geometrically distributed random variable. We will see why in a second.

Geometrically distributed random variables

The value of X_i is the first k such that a certain event happens for the first time. The probability of that event is always the same, p_i. The distribution of X_i looks like this:

$$\Pr[X_i = k] = (1 - p_i)^{k-1} p_i$$

Which explains why its called geometric. The sequence of probabilities forms a geometric series with ratio $(1 - p_i)$.

Let's compute the mean and variance for a geometric r.v. First the expected value:

$$\mathrm{E}[X_i] = \sum_{k=1}^{\infty} k(1 - p_i)^{k-1} p_i$$

and we notice that the $k(1 - p_i)^{k-1}$ term looks like a derivative,

$$\mathrm{E}[X_i] = \sum_{k=1}^{\infty} -p_i \frac{d}{dp_i}(1 - p_i)^k$$

The derivative commutes with the sum, because both are linear operations:

$$\mathrm{E}[X_i] = -p_i \frac{d}{dp_i} \sum_{k=1}^{\infty}(1 - p_i)^k$$

and we can sum the geometric series:

$$-p_i \frac{d}{dp_i} \sum_{k=1}^{\infty}(1 - p_i)^k = -p_i \frac{d}{dp_i}\left(\frac{1 - p_i}{p_i}\right) = \frac{1}{p_i}$$

So the expected value of a geometric r.v. is just $1/p_i$. For the variance, we use the formula

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$$

We just computed $\mathrm{E}[X_i]$ so all we need now is $\mathrm{E}[X_i^2]$.

$$\mathrm{E}[X_i^2] = \sum_{k=1}^{\infty} k^2(1 - p_i)^{k-1} p_i$$

And we do some rearranging to make it look like a second and a first derivative:

$$\mathrm{E}[X_i^2] = p_i \sum_{k=1}^{\infty}(k+1)k(1 - p_i)^{k-1} - p_i \sum_{k=1}^{\infty} k(1 - p_i)^{k-1}$$

Replacing the terms with the derivative expressions gives:

$$\mathrm{E}[X_i] = p_i \sum_{k=1}^{\infty} \frac{d^2}{dp_i^2}(1 - p_i)^{k+1} - p_i \sum_{k=1}^{\infty} \frac{d}{dp_i}(1 - p_i)^k$$

Then we can swap the derivatives and sums:

$$\mathrm{E}[X_i] = p_i \frac{d^2}{dp_i^2} \sum_{k=1}^{\infty} (1-p_i)^{k+1} - p_i \frac{d}{dp_i} \sum_{k=1}^{\infty} (1-p_i)^k$$

And solving for the sums of the geometric series gives:

$$\mathrm{E}[X_i] = p_i \frac{d^2}{dp_i^2} \frac{(1-p_i)^2}{p_i} - p_i \frac{d}{dp_i} \frac{(1-p_i)}{p_i}$$

evaluating derivatives gives:

$$\mathrm{E}[X_i] = p_i \frac{2}{p_i^3} - \frac{1}{p_i} = \frac{2-p_i}{p_i^2}$$

The variance is the difference between this value and the square of $\mathrm{E}[X_i]$:

$$\mathrm{Var}[X_i] = \frac{2-p_i}{p_i^2} - \frac{1}{p_i^2} = \frac{1-p_i}{p_i^2}$$

Expected Time for Coupon Collecting

We have divided the set of choices into epochs, with X_i being the number of placements during the i^{th} epoch. The expected value of X is

$$\mathrm{E}[X] = \sum_{i=0}^{n-1} \mathrm{E}[X_i] = \sum_{i=0}^{n-1} \frac{1}{p_i} = \sum_{i=0}^{n-1} \frac{n}{n-i}$$

And we change variables and use $j = n - i$, which gives

$$\sum_{i=0}^{n-1} \frac{n}{n-i} = n \sum_{j=1}^{n} \frac{1}{j} = nH_n \approx n \ln(n)$$

for the expected number of rounds to hit all of the bins.

Another Upper Bound on Coupon Collecting

The expected value calculation we just did gives us another way to bound the probability that coupon collecting takes longer than a specified value close to $n \ln n$. Once we have the expected value and variance of X, we can use Chebyshev to bound the probability that the running time is larger than expected. The variance is

$$\mathrm{Var}[X] = \sum_{i=0}^{n-1} \mathrm{Var}[X_i] = \sum_{i=0}^{n-1} \frac{1-p_i}{p_i^2} = \sum_{i=0}^{n-1} \frac{ni}{(n-i)^2}$$

Applying the substitution $j = n - i$ gives:

$$\sum_{i=0}^{n-1} \frac{ni}{(n-i)^2} = \sum_{j=1}^{n} \frac{n(n-j)}{j^2} = n^2 \sum_{j=1}^{n} \frac{1}{j^2} - n \sum_{j=1}^{n} \frac{1}{j}$$

The first sum approaches $\pi^2/6$ as $n \to \infty$, so we have

$$\mathrm{Var}[X] \approx \frac{n^2\pi^2}{6} - nH_n$$

That is, the standard deviation is approximately (ignoring lower order terms)

$$\sigma_X \approx \frac{n\pi}{\sqrt{6}}$$

Now we can use Chebyshev. Suppose we want to ensure that the probability of an empty bin is less than 0.01. Pick $t = 10$ in the Chebyshev formula:

$$\Pr[|X - \overline{X}| \geq t\sigma_X] \leq \frac{1}{t^2}$$

Which requires that $X - \mathrm{E}[X] \geq t\sigma_X$ or

$$X \geq n \ln n + 10n\pi/\sqrt{6}$$

That's roughly the same kind of bound we obtained earlier using a direct probability analysis. It's worth noting however, that if we want to make the probability very small, the Chebyshev bounds don't help very much. The actual probability falls off exponentially with distance from the mean, whereas the Chebyshev bound falls off only as $1/t^2$.