CS410 Final Exam

Russell Miller

December 9, 2011

Professor Shrimpton,
This final was very similar to our last homework assignment for me. In the last few weeks I've been understanding less and less of what you've showed us in class. I would have asked more questions, but I felt like we'd never get through your examples if I asked as many questions as it would take to really understand what you were saying. I didn't want to hinder the rest of the class. I tried to make up for it by going to your office, but I really didn't know what questions to ask. Doing Chernoff bounds has been feeling like sort of a magic show to me – as if there is something behind a curtain that I don't have the privilege to see. Some of these more complicated Computer Science questions involving graph coloring and bit strings really stump me because, frankly, I don't know what the question is asking. I think there is some priori knowledge of these advanced algorithms or concepts that I'm just missing. I've always prided myself on being a really hard worker, and the reason I'm writing all of this to you is that no matter how ***hard*** *I have worked on some of these problems, or how* ***badly*** *I've wanted to be able to solve them, I keep working myself into a rut. Under each answer in the pages that follow, I tried to express that I spent hours trying to splice these problems and break them down to a simpler level, to no avail. Also note that I was working on number 5 when the clock struck 6:00 and didn't have it completed nor did I get nearly enough time to check my work on it, by the time I had skipped ahead to it. Please do not misinterpret this write-up as lazily thrown together or apathetic about the topic. I did thoroughly enjoy the challenge of your class this term. That is all.*

1. Consider building a network in the following way. Given n machines, insert an edge between any pair of machines with probability $p = c/n$ for some $c \leq n$. For what value of c is the expected number of edges $(n-1)$? For this c, give the best upper bound you can on the probability that there are at least $2(n-1)$ edges in the network.

Let us define a random variable X to represent the total number of edges. Let us also consider an individual edge between two machines. Let X_i be an indicator random variable that is 1 if and only if there is an edge between those two machines. Then X will be a binomial random variable that is equal to the sum of these indicators.

$$X = \sum_{i=1}^{m} X_i$$

where m is the total number of combinations of 2 machines. We can see that $m = \binom{n}{2}$.
Now we will find the expected number of edges $E[X]$.

$$E[X] = E[\sum_{i=1}^{\binom{n}{2}} X_i] = \sum_{i=1}^{\binom{n}{2}} E[X_i] = \binom{n}{2} p$$

We're told that $p = c/n$.

$$E[X] = \binom{n}{2} \left(\frac{c}{n}\right) = \frac{c(n-1)}{2}$$

In order to find c, let us assume that $E[X] = n - 1$ and solve for c.

$$\begin{aligned} \frac{c(n-1)}{2} &= n-1 \\ c(n-1) &= 2(n-1) \\ c &= 2 \end{aligned}$$

Let's get the variance of X in attempt to find a good bound. The variance of a Binomial random variable with parameters n and p is $np(1-p)$, but $X \sim B(\binom{n}{2}, 2/n)$.

$$Var[X] = \binom{n}{2} \left(\frac{2}{n}\right) (1 - \left(\frac{2}{n}\right)) = (n-1)(1 - \left(\frac{2}{n}\right)) = \frac{(n-1)(n-2)}{n}$$

We now have a Chebyshev bound.

$$Pr(|X - E[X]| \geq a) \leq \frac{Var[X]}{a^2}$$

We need to find a, and we're bounding the condition of $X \geq 2E[X]$. If $X = 2E[X]$, then $X - E[X] = E[X] = a$.

$$Pr(|X - E[X]| \geq E[X]) \leq \frac{Var[X]}{(E[X])^2} = \frac{\frac{(n-1)(n-2)}{n}}{(n-1)^2} = \frac{n-2}{n(n-1)}$$

2. Consider a graph $G = (V, E)$ with the following properties: there exists a partitioning of the vertex set V into subsets V_1 and V_2 of sizes r and s (respectively); for all $x \in V_1$ and $y \in V_2$ there is an edge between x and y; there are no edges between vertices in V_1, and no edges between vertices in V_2. Such a graph is denoted $K_{r,s}$ and is called a complete bipartite graph (with vertex partitioning into sets of size r and s.) Clearly the number of edges in $K_{r,s}$ is rs.
Prove that there is a two-coloring of edges of $K_{r,s}$ with at most

$$\binom{r}{a}\binom{s}{b}2^{1-ab}$$

monochromatic $K_{a,b}$ as subgraphs.

First we should determine how many $K_{a,b}$ there are in $K_{r,s}$. This is $n = \binom{r}{a}\binom{s}{b}$. Next we will let S_i be bernoulli random variables that are 1 if and only if K_i is monochromatic. We have the following.

$$Pr(S_1 \neq 1 \cap S_1 \neq 1 \cap ... \cap S_n \neq 1)$$

Which can be rewritten as

$$1 - Pr(S_1 = 1 \cup S_1 = 1 \cup ... \cup S_n = 1)$$

Or

$$1 - \sum_{i=1}^{n} Pr(\mu = 1)$$

Alright, I was trying to follow along with what seemed like similar exercises in my notes and I'm going to be completely honest, I have no idea how to prove anything about coloring graphs. I drew a $K_{2,3}$ graph and couldn't 2-color it by hand and I have no idea how to show that it's possible. Maybe it's not. I'm sure there's some trick here that I'm just not seeing. I don't know what else to try here.

3. Let $X_1, X_2, ..., X_m$ be independent and identically distributed indicator random variables, and let $\mu = E[X_i]$ for all i. If we want

$$\Pr\left(\left|\left(\frac{1}{m}\sum_{i=1}^{m} X_i\right) - \mu\right| \geq \varepsilon\mu\right) \leq \delta$$

for some $0 < \varepsilon < 1$ and $0 \leq \delta \leq 1$, how many indicators m do we require? (Hint: think Chernoff, and derive a bound on m that is a function of ε, δ, μ.)

Using Corollary 4.6 from Mitzenmacher and Upfal, we can rearrange this to be

$$Pr\left(\left|\left(\frac{1}{m}\sum_{i=1}^{m}\right) - \varepsilon\mu\right| \geq \delta\right) \leq 2e^{-\mu\varepsilon^2/3}$$

The portion related to m here, $\frac{1}{m}\sum\limits_{i=1}^{m} X_i$, can be thought of as $\frac{\mu}{m}$ because as m approaches the size of the population μ is based on, this value approaches μ. So ε is directly related to m in the following way.

$$m \leq \frac{\delta}{\frac{1}{m} - \varepsilon}$$

I am aware this is probably the worst answer you'll see for this question. At this point I'm pretty sure it's safe to say that I don't know how to "derive a Chernoff bound." I tried and tried to figure it out, and I don't see a connection between the derivations we do in class and the ones you're asking us to do. Sorry.

4. The following approach is often called *reservoir sampling*. Suppose we have a sequence of items passing by one at a time. We want to maintain a sample of one item with the property this it is uniformly distributed over all the items that we have seen at each step. Moreover, we want to accomplish this without knowing the total number of items in advance or storing all of the items that we see.
Consider the following algorithm, which stores just one item in memory at all times. When the first item appears, it is stored in the memory. When the kth item appears, it replaces the item in memory with the probability $1/k$. Explain why this algorithm solves the problem.

First let's look at a simple example with only 4 items. We want all 4 items to have a $1/4$ probability of being selected as the sample.
Let's look at the probability that each one is selected (and not replaced).
Item 1: $(1)(1/2)(2/3)(3/4) = 6/24 = 1/4$
(1st item stored, then 2nd,3rd,4th not stored)
Item 2: $(1)(1/2)(2/3)(3/4) = 6/24 = 1/4$
(Doesn't matter if 1st item stored, 2nd one is, 3rd and 4th not stored)
Item 3: $(1)(1)(1/3)(3/4) = 3/12 = 1/4$
(Doesn't matter if 1st or 2nd are stored, 3rd is, 4th is not)
Item 4: $(1)(1)(1)(1/4) - 1/4$
(Doesn't matter if 1st,2nd,3rd are stored, 4th is)
To show that it doesn't matter if previous items are stored, think of each value that is 1 as being a sum of all possible choices for what is stored in the previous item viewings.

This illustrates the idea that the uniform probability permeates through the sequence of items. Since no matter how the $n - 1$ storages were selected, the nth choice is always going to be $1/n$ probability, then this algorithm is definitely another implementation of a uniform random item sampler.

5. Consider throwing m balls uniformly and independently into bins labeled $0, 1, ..., n-1$. We say there is a k-*gap* starting at bin i if bins $i, i+1, ..., i+k-1$ are empty.

a. Determine the expected number (μ) of k-gaps.

Let G be a random variable representing the number of k-gaps total and

$$G = \sum_{i=0}^{n-k} G_i$$

where each G_i is a random variable that is 1 when there is a k-gap at bin i , and 0 otherwise. Then to find μ

$$E[G] = \sum_{i=0}^{n-k} E[G_i] = \sum_{i=0}^{n-k} Pr(G_i = 1)$$

So now we need to know what is the probability that there is a k-gap starting at bin i.

The probability of just one bin being empty is

$$\left(1 - \frac{1}{n}\right)^m \leq e^{-\frac{m}{n}}$$

The probability of k bins in a row being empty would then be bounded by

$$(e^{-\frac{m}{n}})^k = e^{-\frac{mk}{n}}$$

So the expected value $\mu = E[G]$ would be

$$E[G] = \sum_{i=0}^{n-k} e^{-\frac{mk}{n}} = (n-k)(e^{-\frac{mk}{n}})$$

b. Consider the case $n = 10$ and $k = 3$.
What is the probability that a 3-gap starts at bin 2, given that a 3-gap starts at bin 1?
What is the probability that a 3-gap starts at bin 5, given that a 3-gap starts at bin 1?

Pr($B_4 = 0 | B_1 = 0 \cap B_2 = 0 \cap B_3 = 0$):
Each bin i has a probability of 1/10 of getting landed in, for each ball that was thrown. In order to throw m balls and that bin still be empty, the probability would be $(9/10)m$.

$$\begin{aligned} Pr(B_4 = 0 | B_1 = 0 \cap B_2 = 0 \cap B_3 = 0) &= \frac{Pr(B_1 = 0 \cap B_2 = 0 \cap B_3 = 0 \cap B_4 = 0)}{Pr(B_1 = 0 \cap B_2 = 0 \cap B_3 = 0)} \\ &= \frac{((9/10)m)^4}{((9/10)m^3} \\ &= (9/10)m \end{aligned}$$

Ran out of time here.

c. Use Chernoff bounds to give an upper bound on the number of k-gaps. If it helps, assume that k divides n. (Hint, use what you learned in part (b).)

Again, sorry about the incomplete and poorly written final.