CS 174: Combinatorics and Discrete Probability Fall 2012

Homework 5

Due: Thursday, October 4, 2012 by **9:30am**

Instructions: *You should upload your homework solutions on bspace. You are strongly encouraged to type out your solutions using LaTeX. You may also want to consider using mathematical mode typing in some office suite if you are not familiar with LaTeX. If you must handwrite your homeworks, please write clearly and legibly. We will not grade homeworks that are unreadable. You are encouraged to work in groups of 2-4, but you* **must** *write solutions on your own. Please review the homework policy carefully on the class homepage.*

Note: You *must* justify all your answers. In particular, you will get no credit if you simply write the final answer without any explanation.

Problem 1. *(Exercise 5.9 from MU – 5 points)* Consider the probability that every bin receives exactly one ball when n balls are thrown randomly into n bins.

(a) Give an upper bound on this probability using the Poisson approximation.

Solution: Let Y_i be a Poisson random variable for bin i, and recall that $m = n$

$$\Pr[Y_i = j] = \frac{e^{-n/n}(\frac{n}{n})^j}{j!}$$
$$\Pr[Y_i = 1] = \frac{1}{e}$$

Applying Theorem 5.7, we can bound the probability of X using $Y = \cup_{i=1}^{n}\{Y_i = 1\}$.

$$\begin{aligned}\Pr[X] &\leq e\sqrt{n}\Pr[Y] \\ &= e\sqrt{n}\left(\frac{1}{e}\right)^n\end{aligned}$$

(b) Determine the *exact* probability of this event.

Solution: There are n^n ways to throw n balls into n bins. There is only one way, given a specific ordering of the balls, to put a single ball into each bin. And there are $n!$ possible orderings of the balls. Thus the exact probability of n balls in n bins is:

$$\Pr[X] = \frac{n!}{n^n}$$

Problem 2. *(Exercise 5.13 from MU – 5 points)* Let Z be a Poisson random variable with mean μ, where $\mu \geq 1$ is an integer. First, show that $\Pr[Z = \mu + h] \geq \Pr[Z = \mu - h - 1]$ for $0 \leq h \leq \mu - 1$, and use this to conclude that $\Pr[Z \geq \mu] \geq 1/2$.

Solution:

1. Show that $\Pr[Z = \mu + h] \geq \Pr[Z = \mu - h - 1]$ for $0 \leq h \leq \mu - 1$.

$$\begin{aligned}
\Pr[Z = \mu + h] &= \frac{e^{-\mu}\mu^{\mu+h}}{(\mu+h)!} \\
&= \frac{e^{-\mu}\mu^{\mu-h-1}}{(\mu-h-1)!} \cdot \frac{\mu^{2h+1}}{(\mu+h)\ldots\mu\ldots(\mu-h)} \\
&= \Pr[Z = \mu - h - 1] \cdot \prod_{i=1}^{h} \frac{\mu^2}{(\mu+i)(\mu-i)} \\
&\geq \Pr[Z = \mu - h - 1]
\end{aligned}$$

2. Using part (1), we can see that $\Pr[Z \geq \mu] \geq 1/2$.

$$\begin{aligned}
\Pr[Z \geq \mu] &= \sum_{h=\mu}^{\infty} \Pr[Z = h] \\
&= \sum_{h=0}^{\infty} \Pr[Z = \mu + h] \\
&\geq \sum_{h=0}^{\mu-1} \Pr[Z = \mu + h] \\
&\geq \sum_{h=0}^{\mu-1} \Pr[Z = \mu - h - 1] \qquad \text{by part (a)} \\
&= \Pr[Z < \mu]
\end{aligned}$$

Since $\Pr[Z < \mu] + \Pr[Z \geq \mu] = 1$ and $\Pr[Z \geq \mu] \geq \Pr[Z < \mu]$ from above, we have $\Pr[Z \geq \mu] \geq 1/2$.

Problem 3 *(Exercise 5.14 from MU – 5 points)* Let $Y_1, \ldots, Y_n$ be Poisson random variables with mean $\mu(= m/n)$. Let $X_1, X_2, \ldots, X_n$ be the random variables denoting the number of balls in each bin when m balls are thrown in n bins. In class, we showed that for any non-negative function, f,

$$\mathbb{E}[f(Y_1, \ldots, Y_n)] \geq \mathbb{E}[f(X_1, \ldots, X_n)] \Pr[\sum_{i=1}^{n} Y_i = m]$$

When f is monotonically increasing, show that

$$\mathbb{E}[f(Y_1, \ldots, Y_n)] \geq \mathbb{E}[f(X_1, \ldots, X_n)] \Pr[\sum_{i=1}^{n} Y_i \geq m]$$

Use this and problem 2 to conclude that $\mathbb{E}[f(X_1, \ldots, X_n)] \leq 2\mathbb{E}[f(Y_1, \ldots, Y_n)]$ (see Theorem 5.10).

Solution: We have

$$
\begin{aligned}
\mathbb{E}\left[f(Y_1,\ldots,Y_n)\right] &= \sum_{k=0}^{\infty} E\left[f(Y_1,\ldots,Y_n)|\sum_{i=1}^{n} Y_i = k\right] \Pr\left[\sum_{i=1}^{n} Y_i = k\right] \\
&\geq \sum_{k=m}^{\infty} \mathbb{E}\left[f(Y_1,\ldots,Y_n)|\sum_{i=1}^{n} Y_i = k\right] \Pr\left[\sum_{i=1}^{n} Y_i = k\right] \quad \text{(f is non-negative)} \\
&= \sum_{k=m}^{\infty} \mathbb{E}\left[f(X_1^{(k)},\ldots,X_n^{(k)})\right] \Pr\left[\sum_{i=1}^{n} Y_i = k\right] \quad (X_i^k) - k \text{ balls in } n \text{ bins}) \\
&\geq \sum_{k=m}^{\infty} \mathbb{E}\left[f(X_1,\ldots,X_n)\right] \Pr\left[\sum_{i=1}^{n} Y_i = k\right] \quad (f \text{ is monotonically increasing}) \\
&= \mathbb{E}\left[f(X_1,\ldots,X_n)\right] \Pr\left[\sum_{i=1}^{n} Y_i \geq m\right]
\end{aligned}
$$

Then $\sum_{i=1}^{n} Y_i$ has the Poisson distribution with parameter m. Since by problem (2), $\Pr[\sum_{i=1}^{n} Y_i \geq m] \geq 1/2$, by substitution into the previous result we get the desired bound.

Problem 4 *(Exercise 4.12 – 5 points)* Let $X_1,\ldots,X_n$ be geometric random variables with mean 2. Let $X = \sum_{i=1}^{n} X_i$ and $\delta > 0$,

(a) Derive a bound on $\Pr[X \geq (1+\delta)2n]$ by applying a Chernoff bound to a squence of $(1+\delta)(2n)$ independent coin tosses.

Solution: Note that $X - i$ is distributed according to the number of fair coins flipped before obtaining a head. Additionally, X is distributed according to the number of fair coins flipped before getting n heads. Consider the variables Y_i that are $+1$ if the i^{th} flip is tails and -1 if the i^{th} flip is heads. Let $Y = \sum_{i=1}^{(1+\delta)2n} Y_i$. Then

$$\Pr[X \geq (1+\delta)2n] = \Pr[Y \geq (1+\delta)2n - 2n] = \Pr[Y \geq 2\delta n] \leq e^{-\frac{(2\delta n)^2}{2n}} = e^{-\delta^2 n/2}$$

(b) Consider the quantity $\mathbb{E}[e^{tX}]$ and derive a Chernoff bound for $\Pr[X \geq (1+\delta)(2n)]$ using Markov's inequality for the random variable e^{tX}.

Solution: we compute $\mathbb{E}[e^{tX}]$ as follows

$$\begin{aligned}
\mathbb{E}[e^{tX}] &= \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}] \\
&= \prod \sum_{k=1}^{\infty} p(1-p)^{k-1} e^{tk} \\
&= \prod \sum_{k=1}^{\infty} (1/2)^k e^{tk} \\
&= \prod \left(\sum_{k=0}^{\infty} \left(\frac{e^t}{2} \right)^k \right) - 1 \\
&= \prod \left(\frac{1}{1 - e^t/2} \right) - 1 \\
&= \left(\frac{e^t}{2 - e^t} \right)^n
\end{aligned}$$

Now we can use Markov's inequality with the moment generating function to get a Chernoff bound.

$$\begin{aligned}
\Pr[X \geq (1+\delta)(2n)] &= \Pr\left[e^{tX} \geq e^{2n(1+\delta)t} \right] \\
&\leq \frac{\mathbb{E}[e^{tX}]}{e^{2n(1-\delta)t}} \\
&= \left[\left(\frac{e^t}{2 - e^t} \right) \cdot e^{-2(1+\delta)t} \right]^n \\
&= \left[\frac{e^{-(1+2\delta)t}}{2 - e^t} \right]^n
\end{aligned}$$

Setting the derivative to zero, we find that

$$t = ln \left(\frac{1+2\delta}{1+\delta} \right)$$

and

$$\Pr[X \geq (1+\delta)(2n)] \leq \left(\frac{(1+\delta)^{2+2\delta}}{(1+2\delta)^{1+2\delta}} \right)^n .$$

(c) Which bound is better?

Solution: If δ is large, then (a) is better, if δ is small, then (b) is better.

We ignore the n, on the outside, and take logs. The lower the result, the better the bound. We see that the log of probability given by (a) is $\ln(e^{-\delta^2/2}) = -\delta^2 n/2$. For (b) we must work a little harder.

$$
\begin{aligned}
&\ln\left(\frac{(1+\delta)^{2+2\delta}}{(1+2\delta)^{1+2\delta}}\right) \\
&= (1+2\delta)\ln\left(1-\frac{\delta}{1+2\delta}\right)+\ln(1+\delta) \\
&= (1+2\delta)\left(-\frac{\delta}{1+2\delta}-\frac{\delta^2}{2(1+2\delta)^2}-\frac{\delta^3}{3(1+2\delta)^3}-\cdots\right)+\delta-\frac{\delta^2}{2}+\frac{\delta^3}{3}-\cdots \\
&\le \left(-\delta-\frac{\delta^2}{2(1+2\delta)}-\frac{\delta^3}{3(1+2\delta)^2}\right)+\left(\delta-\frac{\delta^2}{2}+\frac{\delta^3}{3}\right) \\
&= -\delta^2\left(\frac{1}{2(1+2\delta)}+\frac{\delta}{3(1+2\delta)^2}+\frac{1}{2}-\frac{\delta}{3}\right) \\
&\le \delta^2\left(\frac{3+5\delta+2\delta^2}{3(1+2\delta)^2}\right)
\end{aligned}
$$

For small δ this is clearly $< -\delta^2/2$. However, if you make δ large (like 10) then

$$
\begin{aligned}
&\ln\left(\frac{(1+\delta)^{2+2\delta}}{(1+2\delta)^{1+2\delta}}\right) \\
&= (1+2\delta)\ln\left(1-\frac{\delta}{1+2\delta}\right)+\ln(1+\delta) \\
&\approx -2\delta\ln(1/2)+\ln(\delta) \\
&\ge -\frac{\delta^2}{2}
\end{aligned}
$$

Problem 5. *(Exercise 4.25 from MU - 10 points)* In this exercise, we design a randomized algorithm for the following packet routing problem. We are given a network that is an undirected connected graph, G, where nodes represent processors and the edges between the nodes represent wires. We are also given a set of N packets to route. For each packet we are given a source node, a destination node, and the exact route (path in the graph) that the packet should take from the source to the destination. (We may assume that there are no loops in the path.) In each time step, at most one packet can traverse an edge. A packet can wait at any node during any time step, and we assume unbounded queue sizes at each node.

A *schedule* for a set of packets specifies the timing for the movement of packets along their respective routes. That is, it specifies which packet should move and which should wait at each time step. Our goal is to produce a schedule for the packets that tries to minimize the total time and the maximum queue size needed to route all the packets to their destinations.

(a) The dilation, d, is the maximum distance travelled by any packet. The congestion, c, is the maximum number of packets that must traverse a single edge during the entire course of the routing. Argue that the time required for any schedule should be at least $\Omega(c+d)$. (Hint: Show that the time should be at least $\max\{c,d\}$ which is $\Omega(c+d)$.)

Solution: Fix any schedule, and suppose the schedule has length T. By definition of dilation, there exists a packet that travels a distance d, and it takes at least d time steps to travel a distance d, so $T \geq d$. Let e be the edge with congestion c. Since at each time step at most one packet can pass through e, it must take c time steps for all c packets passing through e to go through, so $T \geq c$. Therefore, $T \geq max\{c, d\} = \Omega(c+d)$ and this holds for every schedule.

(b) Consider the following *unconstrained* schedule, where many packets may traverse an edge during a single time step. Assign each packet an integral delay chosen randomly, independently, and uniformly from the interval $[1, \lceil \alpha c / \log(Nd) \rceil]$, where α is a *sufficiently large* constant. A packet that is assigned a delay of x waits in its source node for x time steps; then it moves on to its final destination through its specified route without ever stopping. Give an upper bound on the probability that more than $O(\log(Nd))$ packets use a particular edge e at a particular time step t.

Solution: Fix a time step t and an edge e. At most c packets use the edge e at some time, and we may assume WLOG that exactly c packages use the edge e at some time (since this is the worst case). Let X be the r.v. for the number of packets traversing e at time t. We write $X = \sum_i X_i$, where the indicator r.v. X_i is 1 if packet i passes through e at time t, and 0 otherwise. Clearly, $\mathbb{E}[X_i] = \frac{\log(Nd)}{\alpha c}$, so $\mathbb{E}[X] = \frac{\log(Nd)}{\alpha}$. Also, the X_i are independent because the packet delays are independent. So we may apply the Chernoff bound in the form $\Pr[X \geq (1+\delta)\mu] \leq \exp\left(-\frac{\delta^2}{2+\delta}\mu\right)$, with $\mu = \frac{\log(Nd)}{\alpha}$ and $\delta = b\alpha - 1$ to get

$$\Pr\left[X \geq b\log(Nd)\right] \leq \exp\left(-\frac{(b\alpha-1)^2}{\alpha(b\alpha+1)}\log(Nd)\right)$$

(Here b is a constant that we can choose.) Now if we set (for example) $b = 5$ and $\alpha = 2$ the exponent in the above bound becomes $\frac{81}{22}\log(Nd) \geq 3\log(Nd)$. Thus we have

$$\Pr\left[X \geq 5\log(Nd)\right] \leq \exp\left(-3\log(Nd)\right) = \frac{1}{(Nd)^3}$$

Hence, the probability that more than $O(\log(Nd))$ packages use e at time step t is at most $\frac{1}{(Nd)^3}$. [Note: We chose $\frac{1}{(Nd)^3}$ here for use in the next part. More generally, we can achieve an upper bound of any poly $\left(\frac{1}{Nd}\right)$ by replacing b and α with correspondingly larger constants.]

(c) Again using the unconstrained schedule of part (b), show that the probability that more than $O(\log(Nd))$ packets pass through any edge at any time step is at most $1/(Nd)$ for a sufficiently large α.

Solution: We need to take a union bound over all edges e that are used and over all time steps t. To do this, we need upper bounds on both the number of edges and the number of time steps:

- Each packet uses at most d distinct edges, so the total number of edges used is at most Nd.
- The total number of time steps is at most $d + \frac{\alpha c}{\log(Nd)}$ The congestion c is bounded by N, so this total number of time steps is at most $d + \frac{\alpha N}{\log(Nd)} \leq d + N \leq 2Nd$, for sufficiently large N. (Recall that $\alpha = 2$ is a constant.)

Now, we may apply a union bound to deduce that the probability that there exists some e, t such that more than $5\log(Nd)$ packets use the edge e at time step t is at most $Nd \cdot 2Nd \cdot \frac{1}{(Nd)^3} \leq \frac{2}{N}$. Therefore, with probability$1 - O(1/N)$, we obtain a schedule in the unconstrained model with low congestion, namely one wherein at every time step, at most $5\log(Nd)$ packets traverse any particular edge.

(d) Use the unconstrained schedule to devise a simple randomized algorithm that, with high probability, produces a schedule of length $O(c + d\log(Nd))$ using queues of size $O(\log(Nd))$ and following the constraint that at most one packet crosses an edge per time step. (By high probability, we mean $1 - O(1/N)$.)

Solution: Note that it suffices to handle the case where the schedule in the unconstrained model has low congestion (i.e., at every time step, at most $5\log(Nd)$ packets traverse any edge), since by part (c) this occurs with probability $1 - O(1/N)$. (With probability $O(1/N)$, our schedule will do arbitrarily poorly, which is OK.) We turn such an unconstrained schedule into a real schedule by replacing every time step in the unconstrained schedule by $s = 5\log(Nd)$ time steps in the real schedule; we want it to be the case that for each $i = 1, 2, \ldots$, the locations of the packets in the real schedule after the $(is)^{th}$ time step will be the same as that in the unconstrained schedule after the i^{th} time step. (This ensures that there is no interference between steps in the unconstrained schedule, so the analysis of parts (b) and (c) still holds.)

Since in the unconstrained schedule at most s packets traverse any particular edge in a single time step, all of these packets can traverse this edge in s time steps in the real schedule without violating the constraint that at most one packet crosses an edge per time step. Once a packet crosses an edge, it waits at the other end of the edge until the next time step on the unconstrained schedule. Clearly, we only need queues of size $s = O(\log(Nd)$ to implement this scheme. The length of the unconstrained schedule is $d + \frac{\alpha c}{\log(Nd)}$, so the length of the real schedule is s times that, which is $O(c + d\log(Nd))$ (recall that $\alpha = 2$ is a constant).