CSCI 6220/4030: Homework 2

Assigned Thursday September 28 2017. Due at beginning of class Thursday Oct 12 2017.

Remember to typeset your submission, and label it with your name. Please start early so you have ample time to see me during office hours. Provide mathematically convincing arguments for the following problems. Ask me if you are unclear whether your arguments are acceptable.

1. A discrete Poisson random variable X with parameter μ follows the following probability distribution on $j = 0, 1, 2, \ldots$:
$$\mathbb{P}(X = j) = \frac{\mathrm{e}^{-\mu}\mu^j}{j!}$$
   (i) Find the moment generating function of a Poisson random variable with parameter μ.
   (ii) Using the moment generating function, compute the mean and variance of a Poisson random variable.
   (iii) If X and Y are independent Poisson r.v.s with parameters μ_X and μ_Y, what is the moment generating function of $X + Y$, and what distribution does this suggest $X + Y$ follows?
   (iv) Prove your conjecture from the previous part by computing the probability mass function $\mathbb{P}(X + Y = k)$ for $k = 0, 1, \ldots$

2. Consider a collection $X_1, \ldots, X_n$ of n independent integers chosen uniformly from the set $\{0, 1, 2\}$. Let $X = \sum_{i=1}^n X_i$ and $0 < \delta < 1$. Derive a Chernoff bound for $\mathbb{P}(X \geq (1 + \delta)n)$ and $\mathbb{P}(X \leq (1 - \delta)n)$.

3. Consider a collection $X_1, \ldots, X_n$ of n independent geometrically distributed random variables with mean 2. Let $X = \sum_{i=1}^n X_i$ and $\delta > 0$.
   (i) Derive a bound on $\mathbb{P}(X \geq (1 + \delta)(2n))$ by applying a Chernoff bound to a sequence of $(1 + \delta)(2n)$ fair coin tosses.
   (ii) Directly derive a Chernoff bound on $\mathbb{P}(X \geq (1 + \delta)(2n))$ using the moment generating function for geometric random variables.
   (iii) Which bound is better?

4. Consider the following view of Randomized Quicksort. Every point in the algorithm where it decides on a pivot element is called a node. Suppose the size of the set to be sorted at a particular node is s. The node is called good if the pivot element divides the set into two parts, each of size not exceeding size $2s/3$. Otherwise the node is called bad. The nodes can be thought of as forming a tree in which the root node has the whole set to be sorted and its children have the two sets formed after the first pivot step and so on.
   (i) Show that the number of good nodes in any path from the root to a leaf in this tree is not greater than $c \log_2 n$, where c is some positive constant.
   (ii) Show that, with high probability (greater than $1 - 1/n^2$), the number of nodes in a given root to leaf path of the tree is not greater than $c' \log_2 n$, where c' is another constant.
   (iii) Show that, with high probability (greater than $1 - 1/n$), the number of nodes in the longest root to leaf path is not greater than $c' \log_2 n$.
   (iv) Use your answers to show that the running time of Quicksort is $O(n \ln n)$ with probability at least $1 - 1/n$.

5. Suppose that we can obtain independent samples $X_1, X_2, \ldots$ of a random variable X and that we want to use these samples to estimate $\mathbb{E}[X]$. Given t samples, we use the sample average $\left(\sum_{i=1}^t X_i\right)/t$ for our estimate of $\mathbb{E}[X]$. We want the estimate to be within $\varepsilon\mathbb{E}[X]$. of the true

value $\mathbb{E}[X]$ with probability at least $1-\delta$. We may not be able to use Chernoff's bound directly to bound how good our estimate of X is if we do not know its moment generating function. We develop an alternative approach that requires only having a bound on the variance of X. Let $r = \sqrt{\mathrm{Var}(X)}/\mathbb{E}[X]$.

(i) Show using Chebyshev's inequality that $O(r^2/(\varepsilon^2\delta))$ samples are sufficient to solve the problem.

(ii) Suppose that we need only a weak estimate that is within $\varepsilon\mathbb{E}[X]$ of $\mathbb{E}[X]$ with probability at least $3/4$. Argue that $O(r^2/\epsilon)$ samples are enough for this weak estimate.

(iii) Show that, by taking the median of $O(\ln(1/\delta))$ weak estimates, we can obtain an estimate within $\varepsilon\mathbb{E}[X]$ of $\mathbb{E}[X]$ with probability at least $1-\delta$. Conclude that we need only $O(r^2\ln(1/\delta)/\varepsilon^2)$ samples.

6. Consider n balls thrown randomly into n bins. Let $X_i = 1$ if the ith bin is empty and 0 otherwise. Let $X = \sum_i X_i$. Let Y_i for $i = 1, \ldots, n$ be independent Bernoulli random variables that are 1 with probability $p = (1-1/n)^n$. Let $Y = \sum_{i=1}^n Y_i$.

   (i) Show that $\mathbb{E}[X_1 X_2 \cdots X_k] \leq \mathbb{E}[Y_1 Y_2 \cdots Y_k]$ for any $k \geq 1$.

   (ii) Show that $\mathbb{E}[\mathrm{e}^{tX}] \leq \mathbb{E}[\mathrm{e}^{tY}]$ for all $t \geq 0$. (Hint: use the expansion for e^x and compare $\mathbb{E}[X^k]$ to $\mathbb{E}[Y^k]$.)

   (iii) Derive a Chernoff bound for $\mathbb{P}(X \geq (1+\delta)\mathbb{E}[X])$.