CS5330: Randomized Algorithms

Problem Set 2—Solutions

Due: January 29, 6:30pm

Instructions. The *exercises* at the beginning of the problem set do not have to be submitted—though you may. The first three exercises should be easy review. The remaining two exercises are more interesting and it is recommended that you do them. There are three problems to submit.

- Please submit the problem set on IVLE in the appropriate folder. (Typing the solution using latex is recommended.) If you want to do the problem set by hand, please submit it at the beginning of class.
- Start each problem on a separate page.
- If you submit the problem set on paper, make sure your name is on each sheet of paper (and legible).
- If you submit the problem set on paper, staple the pages together.

Remember, that when a question asks for an algorithm, you should:

- First, give an overview of your answer. Think of this as the executive summary.
- Second, describe your algorithm in English, giving pseudocode if helpful.
- Third, give an example showing how your algorithm works. Draw a picture.

You may then give a proof of correctness, or explanation, of why your algorithm is correct, an analysis of the running time, and/or an analysis of the approximation ratio, depending on what the question is asking for.

Advice. Start the problem set early—questions may take time to think about. Come talk to me about the questions. Talk to other students about the problems.

Collaboration Policy. The submitted solution must be your own unique work. You may discuss your high-level approach and strategy with others, but you must then: (i) destroy any notes; (ii) spend 30 minutes on facebook or some other non-technical activity; (iii) write up the solution on your own; (iv) list all your collaborators. Similarly, you may use the internet to learn basic material, but do not search for answers to the problem set questions. You may not use any solutions that you find elsewhere, e.g. on the internet. Any similarity to other students' submissions will be treated as cheating.

Review (***Not to be submitted.***)

Exercise 1. [Variance]

Recall that for a random variable X, the variance of X ($\mathrm{Var}[X]$) is defined as $\mathrm{E}\left[(X-\mu)^2\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$, where $\mu = \mathrm{E}[X]$.

a. Let X be an indicator random variable that is 1 with probability p and 0 with probability $1-p$. Prove that $\mathrm{Var}[X] = p(1-p)$.

b. Let X be the number of heads when you flip a coin n times where each flip is heads with probability $1/2$. Calculate $\mathrm{Var}[X]$.

Exercise 2. [Chebychev]

Given a random variable X, prove that:

$$\Pr\left[|X-\mu| \geq t\right] \leq \frac{\mathrm{Var}[X]}{t^2}$$

This is known as Chebychev's Inequality. (Hint: use Markov's Inequality to bound $\Pr\left[(X-\mu)^2 \geq t^2\right]$.)

Exercises (***To be done, but may not be submitted.***)

Exercise 3. [Balls in Bins]

Assume you have m balls and n bins. Each ball is placed in a bin chosen uniformly at random. Let X_i be the number of balls in bin i.

a. What is $\mathrm{E}[X_i]$?

b. What is $\Pr[X_i = 0]$?

c. For $m = n$, show that the expected number of empty bins (after the process is complete) is approximately n/e.

d. For $m = 2n\log(n)$, show that the expected number of empty bins (after the process is complete) is < 1.

Exercise 4. [QuickSelect]

Consider the QuickSelect algorithm which returns the kth item in an (unsorted) array:

Algorithm 1: QuickSelect$(A, k, begin, end)$

```
p ← Random(begin, end)
i ← Partition(A, p)
if k < i then return QuickSelect(A, k, begin, i − 1).
if k > i then return QuickSelect(A, k − i, i + 1, end).
if k = 1 then return A[i].
```

Prove that QuickSelect takes $O(n)$ expected time.

Standard Problems (to be submitted)

Problem 1. [Flipping Coins]

Imagine you flip a fair coin n times. Each time you flip the coin, it comes up heads with probability 1/2 and tails with probability 1/2. A *streak of length k* occurs when the coin comes up heads k times in a row at some point during the sequence. For example, consider the following sequence:

T T H H H T H T T T H H H H

In this sequence there is one streak of length 4, three streaks of length 3, and five streaks of length 2.

Problem 1.a. What is the expected number of streaks of length k?

Solution: Let X_i be the indicator random variable that is equal to 1 if there is a streak of length k that begins at position i. The probability $\Pr[X_i] = 1/2^k$. There are $n - k + 1$ possible positions in which a streak of length k can begin. Hence by linearity of expectation, the expected number of streaks is equal to $\text{Exp}\left[\sum_{i=1}^{n-k+1} X_i\right] = \sum_{i=1}^{n-k+1} \text{Exp}[X_i] = \frac{n-k+1}{2^k}$.

Problem 1.b. Show that the probability of a streak of length $\log n + 2$ is small (e.g., $< 1/2$).

Solution: Define indicator random variables X_i, as above. In this case, $\Pr[X_i] = 1/2^{\log n+2} = \frac{1}{4n}$. We then take a union bound over all possible X_i: $\Pr[X_1 \cup X_2 \cup \ldots \cup X_{n-k+1}] \leq \sum_{j=1}^{n-k+1} \Pr[X_i] \leq \frac{n}{4n}$ (where $k = \log n + 2$). Thus we conclude that the probability of a streak of length $\log n + 2$ is $\leq 1/4$.

Problem 1.c. Show that, for sufficiently large n, the probability that there is no streak of length at least $\lfloor \log n - 2 \log \log n \rfloor$ is less than $1/n$.

Hint: break the sequence of flips up into disjoint blocks of $\lfloor \log n - 2 \log \log n \rfloor$ consecutive flips, and use the fact that streaks in different blocks are independent.

Solution: For each chunk of length $k = \lfloor \log n - 2 \log \log n \rfloor$, the probability that all the flips in the chunk are heads is $\geq \log^2 n/n$. Notice that each chunk is independent, and there are at least $n/\log n$ chunks. Hence the probability that all the chunks contain at least one tails is at most:

$$\left(1 - \frac{\log^2 n}{n}\right)^{n/\log n} \leq \left(\frac{1}{e}\right)^{\log n} \leq 1/n$$

Thus with high probability, there is at least one streak of length $k = \lfloor \log n - 2 \log \log n \rfloor$. (Be sure to get the direction of the inequalities correct.)

Problem 2. [Estimators]

Imagine you have a large array A of size n containing all 0's and 1's. (The array might represent a large population, and $A[i]$ might indicate whether the person likes laksa.) We want to estimate the number of $1's$ in the array. That is, let $T = \sum_{i=1}^{n} A[i]$ be the number of $1's$ in array A.

We solve this problem via sampling. We repeat the following procedure s times: choose a random index in the array and check whether it is a 0 or a 1. Let X_i be the value of the ith sample. Let $Y = (n/s) \sum_{i=1}^{s} X_i$. The value Y is an estimator for the sum T.

a. What is $\mathrm{E}[X_i]$?

b. What is $\mathrm{Var}[X_i]$?

c. What is $\mathrm{E}[Y]$?

d. What is $\mathrm{Var}[Y]$? (Hint: Show that $\mathrm{Var}[Y] = (n^2/s)\mathrm{Var}[X_1]$.)

From these results, do you expect that Y is a good estimator of T?

Solution: For each X_i, $\Pr[X_i = 1] = T/n$, so $\mathrm{E}[X_i] = T/n$.

For an indicator random variable Z with probability p, the variance is equal to $\mathrm{E}\left[Z^2\right] - \mathrm{E}[Z]^2 = p - p^2 = p(1-p)$. (Recall $\mathrm{E}\left[Z^2\right] = p \cdot 1^2 + (1-p) \cdot 0^2$.) Since X_i is an indicator random variable with probability T/n, the variance of X_i is equal to $(T/n)(1 - T/n)$.

We can determine $\mathrm{E}[Y] = (n/s)\sum X_i$ by linearity of expectation, i.e., it is equal to $(n/s) \cdot (s) \cdot (T/n) = T$. That is, $\mathrm{E}[Y] = T$.

In general, for two independent random variables W and Z, $\mathrm{Var}[W + Z] = \mathrm{Var}[W] + \mathrm{Var}[Z]$. For a constant a, $\mathrm{Var}[aZ] = a^2\mathrm{Var}[Z]$. Both of these facts can be proved directly from the definition of variance. (Notice that $\mathrm{Var}[X + X] \neq \mathrm{Var}[X] + \mathrm{Var}[X]$ because X is surely not independent from X!)

Thus, we can compute:

$$
\begin{aligned}
\mathrm{Var}[Y] &= \mathrm{Var}\left[(n/s)\sum_{i=1}^{s} X_i\right] \\
&= (n^2/s^2)\sum_{i=1}^{s} \mathrm{Var}[X_i] \\
&= (n^2/s)\mathrm{Var}[X_1] \\
&= (n^2/s)(T/n)(1 - T/n) \\
&= Tn/s - T^2/s
\end{aligned}
$$

Using this, we conclude that the variance of Y is reasonably high. On the one hand, Y is a good estimate because in expectation it is equal to T. On the other hand, the variance is sufficiently high that it will not return a very good estimate unless s is reasonably large. (Imagine trying to use Chebychev's Inequality to show that the estimate is good; it will not provide a very good estimate!) However, for large enough s, it will work!

Problem 3. [Permutations]

For Chinese New Year, we all decide to play Secret Santa[1]. For a class of size n, we each put one present into a big pile (leading to a pile of n presents), and we each get a random present from the pile.

Problem 3.a. What is the expected number of people who get their own present (i.e., who get from the pile the same present they put into the pile)?

Solution: Let $X_j = 1$ if person j receives their own present and 0 otherwise. The $\mathrm{E}[X_j] = \Pr[X_j = 1] = 1/n$. By linearity of expectation, the expected number of people who receive their own present are $\sum \mathrm{E}[X_j] = n \cdot (1/n) = 1$.

Problem 3.b. In fact, the assignment of presents to people is a random permutation. Let π be a random permutation of $[1, \ldots, n]$, uniformly chosen from all possible permutations of n. We say that $v_1, v_2, \ldots, v_k$ is a cycle of length k if $\pi(v_i) = v_{i+1}$ for all $i < k$, and $\pi(v_k) = v_1$. For example, consider the following permutation:

$$\sigma = \mathbf{3\ 5\ 1\ 2\ 4}$$

(That is, $\sigma(1) = 3$, $\sigma(2) = 5$, etc.) This permutation consists of two cycles: $(1, 3)$ and $(2, 5, 4)$. The first cycle has length 2 and the second cycle has length 3.

Fix a person j. How many cycles of length 1 contain person j? How many cycles of length 2 contain person j? How many cycles of length x contain person j?

Solution: The answer for all of the above is $(n-1)!$. If you exclude person j, there are $(n-1)!$ possible permutations, so there are $(n-1)!$ permutations where person j is in a cycle of length 2. There are $n-1$ possible cycles of length 2 containing person j, and there are $(n-2)!$ possible permutations of the remaining $n-2$ people, leading to $(n-1)(n-2)! = (n-1!)$ possible permutations in which j is in a cycle of length 2.

In general, there are $\binom{(n-1)}{x-1}((x-1)!)$ cycles of length x containing person j. (We multiple by $(x-1)!$ since those $x-1$ people can be arranged in the cycle in any order.) There are $(n-x)!$ permutations of the remaining people. Thus the total number of permutations is:

$$\frac{(n-1)!}{(n-x)!(x-1)!}(x-1)!(n-x)! = (n-1)!$$

Thus for any length cycle, the number of cycles containing person j is $(n-1)!$.

Problem 3.c. Every permutation can be decomposed into cycles. What is the expected number of cycles in a randomly chosen permutation π?

[1] Why can't Santa deliver presents for Chinese New Year too? He's had plenty of rest since Christmas.

Hint: Define the random variable S_i to be the size of the cycle containing i. For example, in the permutation σ above, $S_1 = 2$, $S_2 = 3$, $S_3 = 2$, $S_4 = 3$, $S_5 = 3$. Define the random variable Y_i to be $1/S_i$. Again, in the example above for permutation σ, $Y_1 = 1/2$, $Y_2 = 1/3$, $Y_3 = 1/2$, $Y_4 = 1/3$, $Y_5 = 1/3$.

Solution: Define the random variables as specified, and notice that $\sum_{i=1}^{n} Y_i$ is equal to the number of permutations. We first calculate $\mathrm{E}[Y_i]$ and $\Pr[S_i = k]$.

The key observation here is that i is equally likely to be part of a cycle of each length! Recall that there are $n!$ possible permutations. We already observed that there are $(n-1)!$ possible permutations in which i is part of a cycle of length x for every x.

From this we conclude that $\Pr[S_i = k] = (n-1)!/n! = 1/n$. We can then calculate $\mathrm{Exp}[Y_i] = \sum_{i=1}^{n} (1/i) \Pr[S_i = i]$ (by the definition of expectation), and hence $\mathrm{Exp}[Y_i] = \sum_{i=1}^{n} (1/i)(1/n) \approx \log n/n$.

We conclude, by linearity of expectation, that $\mathrm{Exp}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \mathrm{Exp}[Y_i] \approx \log n$. That is, in expectation, there are $\log n$ cycles in a random permutation.

Problem 3.d. (Optional.) For some $k > n/2$, what is the probability that the permutation contains a cycle of length k? What is the probability that the permutation contains no cycle of length $> n/2$? (Hint: notice that if a permutation has a cycle of length $k > n/2$, then it does *not* have a cycle of length $k' \neq k$ where $k' > n/2$.)

Solution: The total number of permutations containing a cycle of length $k > n/2$ is $\binom{n}{k}(k-1)!(n-k)! = (n!/((n-k)!k!))(k-1)!(n-k)! = n!/k$. Since there are $n!$ permutations in total, that means that the probability that a random permutation has a cycle of length k is $1/k$. (Notice that we avoid double counting cycles because there can only be one cycle of length $k > n/2$ in the permutation.)

Since the events of having a cycle of length $k > n/2$ and $k' > n/2$ are disjoint (when $k \neq k'$), we conclude that the probability of having any cycle of length $> n/2$ is $\sum_{j=n/2+1}^{n} (1/j) \leq \int_{n/2}^{n} (1/j) \leq \ln(n/(n/2)) = \ln(2)$. Thus the probability that the prisoners win is at least $1 - \ln(2)$.

Challenge Problem (Optional)

Problem 4. Imagine you have n bins and n balls. Consider the following game:

Repeat the following until there are no bins left:

- Place each ball into a bin chosen uniformly at random from the remaining set of bins.
- Remove every bin that has at least one ball.
- Collect all n balls and continue.

For example, after the first iteration, in expectation, there will be about n/e bins left. The process terminates when there are no bins left. Prove that, in expectation, the process terminates in $O(\log^* n)$ rounds, where $\log^* n$ is the iterated logarithm[2].

[2] Define $\log^* 1 = 0$ and $\log^* n = 1 + \log^*(\log n)$.