University of Virginia **cs4102: Algorithms** 1 October 2020

Randomness and Sorting Exercise 1

Response by: **Charles Wang (cw5mj)**

Collaborators and Resources: George Cao (glc6qrx)

A Matching Pair

Problem 1: Linear Time Expected

Inner Loop: Consider these cases separately:

Case 1: Assuming $s_1 \in U$, show that the expected number of times the inner loop will iterate is $\Theta(n)$.

Since $s_1 \in U$, The number of possible socks s_2 that are the same as s_1 is 0. The loop iterates until $s_1 = s_2$ and since that will never happen, the loop will iterate until it reaches sock $\frac{n}{4}+1$. Thus, the number of iterations is $\frac{n}{4}+1 \in \Theta(n)$

Case 2: Assuming $s_1 \in M$, show that the expected number of times the inner loop will iterate is $\Theta(1)$.

The probability for 2 or more iterations would be $\frac{n/4-1}{n-1} \leq \frac{1}{4}^1$
The probability for 3 or more iterations would be $\frac{n/4-1}{n-1} * \frac{n/4-2}{n-2} \leq \frac{1}{4}^2$
...
Let X be the number of iterations that the inner loop iterates. The probability that the algorithm continues after iteration i can be written as

$$\sum_{i=1}^{\frac{n}{4}+1} Pr[X \geq i] \leq \sum_{i=1}^{\frac{n}{4}+1} (\frac{1}{4})^{i-1} \leq \sum_{i=1}^{\infty} (\frac{1}{4})^{i-1} = \frac{4}{3}$$

Therefore, the expected number of iterations is $\frac{4}{3} \in \Theta(1)$.

Outer Loop: Show that the expected number of times the outer loop will iterate is $\Theta(1)$. **Hint:** Bound the probability that the algorithm continues after iteration i (i.e., the first time that $s_1 \in M$ occurs is after iteration i).

Similarly to Case 2, the probability that the outer loop takes more than i iterations to find sock $s \in M$ is $\frac{1}{4}^{i-1}$. This is a geometric series that converges to $\frac{1}{1-1/4} = \frac{4}{3}$.
Therefore, the expected number of iterations is $\frac{4}{3} \in \Theta(1)$.

Putting it together: Use all three of your answers above to find the overall expected running time of this algorithm in terms of the number of comparisons.

Since the outer loop is expected to run in constant time, and the inner loop may iterate $\Theta(n)$ times, the overall comparisons/running time is $\Theta(n)$.

Problem 2: A Better Algorithm

Algorithm: Select 2 random socks. Do not put the socks back. If they do not match, continue selecting pairs of socks.

Correctness: Since there are n socks and $\frac{3n}{4}+1$ are unique, in the worst case, where every pair contains one sock in U and one sock in M, after the algorithm loops $\frac{n}{4}-1$ times, the remaining socks are all guaranteed to be identical.

Running Time: The probability of both socks being identical is $\frac{3n/4+1}{n} * \frac{3n/4}{n-1} \geq \frac{9}{16}$.
Therefore, the probability that the algorithm continues, or the probability of picking either two unique socks or one unique sock and one from the identical group, is $\leq \frac{7}{16}$.

Let X be the number of iterations that the algorithm iterates. The probability that the algorithm continues after iteration i can be written as

$$\sum_{i=1}^{\frac{n}{4}} Pr[X \geq i] \leq \sum_{i=1}^{\frac{n}{4}} (\frac{7}{16})^i \leq \sum_{i=1}^{\infty} (\frac{7}{16})^i = \frac{16}{9}$$

Therefore, the expected number of iterations is $\frac{16}{9} \in \Theta(1)$.

Problem 3: The Count-Min Sketch and Frequency Estimation

Expected value of estimate. Show that $f_i \leq \mathbb{E}[X_i] \leq f_i + N/M$. **Hint:** Try defining an indicator random variable $X_{ij} = 1$ if $h(x_i) = h(x_j)$ and 0 otherwise.

$f_i \leq \mathbb{E}[X_i]$ is trivial, as the expected value for the X_i will be at least the number of queries f_i that have actually been made.
For proving the other part of the inequality, let indicator variable X_{ij} indicate whether $h(x_i) = h(x_j)$. $X_{ij} = 1$ if $h(x_i) = h(x_j)$ and 0 otherwise.
With this definition, we have $Pr[X_{ij} = 1] = \frac{1}{M}$, as there are M possible buckets.
Since we have N total queries, the expected value for the total number of collisions is

$$\sum_{n=1}^{N} Pr[X_{ij} = 1] = \sum_{n=1}^{N} \frac{1}{M} = \frac{N}{M}$$

The expected value should be less than the minimum number of queries plus the expected collisions, namely, $\mathbb{E}[X_i] \leq f_i + \frac{N}{M}$.
Therefore, $f_i \leq \mathbb{E}[X_i] \leq f_i + \frac{N}{M}$

Quality of estimates. Show that $\Pr[X_i \geq f_i + 2N/M] \leq 1/2$. **Hint:** You may use without proof that for a random variable X taking on non-negative values, $\Pr[X \geq t] \leq \mathbb{E}[X]/t$.

Since we are given $\Pr[X \geq t] \leq \mathbb{E}[X]/t$, we can say $\Pr[X_i \geq f_i + \frac{2N}{M}] \leq \frac{\mathbb{E}[X]}{f_i + 2N/M}$
$\Pr[X_i - f_i \geq \frac{2N}{M}] \leq \frac{\mathbb{E}[X_i]}{f_i + 2N/M}$

Let $Y_i = X_i - f_i$. Then, $\mathbb{E}[Y_i] = \mathbb{E}[X_i] - f_i$
Therefore, we have

$$\Pr[X_i \geq f_i + \frac{2N}{M}] = \Pr[Y_i \geq \frac{2N}{M}] \leq \frac{\mathbb{E}[Y_i]}{2N/M} = \frac{\mathbb{E}[X_i] - f_i}{2N/M} = \frac{f_i + N/M - f_i}{2N/M} = \frac{N/M}{2N/M} = \frac{1}{2}$$

$$\therefore \Pr[X_i \geq f_i + \frac{2N}{M}] \leq 1/2$$

Reducing the error. We can reduce the error by using multiple independent copies of the above data structure. Namely, we initialize multiple arrays of counters $C_1 = (c_{1,1}, \ldots, c_{1,M}), \ldots, C_k = (c_{k,1}, \ldots, c_{k,M})$ and multiple independent hash functions $h_1, \ldots, h_k$ (each sampled independently from the family of universal hash functions). When we obtain a query q, we increment *all* of the counters $c_{1,h_1(q)}, \ldots, c_{k,h_k(q)}$. To obtain an estimate for a query q, we take the *minimum* count across all of the arrays of counters: $X_i = \min_{i=1,\ldots,k} c_{i,h_i(q)}$. Show that $\Pr[X_i \geq f_i + 2N/M] \leq 1/2^k$.

If you have a minimum $X_i \geq f_i + \frac{2N}{M}$, that means it is less than all the other $X_{i=1,\ldots,k}$.
The probability of the minimum being greater than or equal to $f_i + \frac{2N}{M}$ is equal to the probability of all of the other $X_{j=1,\ldots,k}$ being greater than or equal to $f_j + \frac{2N}{M}$.
Therefore, we can do

$$Pr[X_i \geq f_i + \frac{2N}{M}] = \prod_{j=1}^{k} Pr[X_j \geq f_j + \frac{2N}{M}] \leq \prod_{i=1}^{k} \frac{1}{2} = (\frac{1}{2})^k$$