Abhimanyu M A **(1111002)** , Sumesh T A **(1111065)**
Department of Computer Science and Engineering
IIT Kanpur

August 29, 2011

ASSIGNMENT 1

Problem 1. **Analysis of number of comparisons in Random Quick Select** We shall first prove that the expected number of comparisons is less than $\mathbf{4}\boldsymbol{n}$.

Lemma 1. *Given a* $\boldsymbol{x}$ *randomly selected from* $\boldsymbol{S}$*, the expected size of the larger set among* $\boldsymbol{S_{<x}}$ *and* $\boldsymbol{S_{>x}}$ *is* $\frac{\mathbf{3}|\boldsymbol{S}|}{\mathbf{4}}$

Proof. Now let us look at the smaller set, the size of the smaller set can vary anywhere from $\mathbf{0}$ to $\frac{|\boldsymbol{S}|}{\mathbf{2}}$, since $\boldsymbol{x}$ is taken uniformly at random

$$\boldsymbol{E}[\text{—Smaller Set—}] = \frac{|\boldsymbol{S}|}{\mathbf{4}} \tag{1-1}$$

This gives us that the expected size of the larger set is

$$\boldsymbol{E}[\text{—Larger Set—}] = \frac{\mathbf{3}|\boldsymbol{S}|}{\mathbf{4}} \tag{1-2}$$

□

Now consider the first $\boldsymbol{x}$ chosen it will have to make $|\boldsymbol{S}|$ comparisons, and finally it will be splitting the set $\boldsymbol{S}$ into two and we can see that we will be discarding at least one of them. To ensure that we get the worst case expected probability we assume that the smaller set is discarded and hence the size of the instance now reduces to $\frac{\mathbf{3}|\boldsymbol{S}|}{\mathbf{4}}$. Now taking this as the set the same step is repeated hence.

$$\boldsymbol{E}[\text{Number of comparisons}] \leq |\boldsymbol{S}| + \frac{\mathbf{3}|\boldsymbol{S}|}{\mathbf{4}} + \left(\frac{\mathbf{3}|\boldsymbol{S}|}{\mathbf{4}}\right)^{\mathbf{2}} + \ldots \tag{1-3}$$

This is an infinite geometric series with $\boldsymbol{r} = \frac{\mathbf{3}}{\mathbf{4}}$, and hence taking $|\boldsymbol{S}| = \boldsymbol{n}$ we get.

$$\boldsymbol{Sum} = \frac{\boldsymbol{n}}{\mathbf{1} - \frac{\mathbf{3}}{\mathbf{4}}} = \mathbf{4}\boldsymbol{n} \tag{1-4}$$

Hence

$$\boldsymbol{E}[\text{Number of comparisons}] \leq \mathbf{4}\boldsymbol{n} \tag{1-5}$$

⟶ Answer

Problem 2. **Candidate Selection Problem**

We shall do this problem by careful application of partition theorem. Now we need to find the probability of Professor Dixon hiring the best qualified candidate.

$$\boldsymbol{P}[\varepsilon] \tag{2-1}$$

Where $\boldsymbol{\varepsilon}$ is the event that the most eligible candidate is hired

Since computing this value will be very hard we shall apply partitioning on it. We consider the event $\boldsymbol{X_i}$ where the best candidate is hired provided that he is at position $\boldsymbol{i}$.
It is easy to see that since the best candidate can occur in any position from $\mathbf{1}$ to $\boldsymbol{n}$ with equal probability(since we have assumed that the applicants appear in a uniformly random order)
If we denote

$$\boldsymbol{p_i} = \boldsymbol{Pr}[\boldsymbol{X_i}] \tag{2-2}$$

then we can see that

$$\boldsymbol{P}[\varepsilon] = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{p_i} \tag{2-3}$$

We shall now construct a qualification array $\boldsymbol{A}$ where $\boldsymbol{A_i}$ = number of people behind him + 1. It is easy to see that for the best qualified candidate $\boldsymbol{A}[\boldsymbol{i}] = \boldsymbol{n}$.
Since Professor Dixon rejects all candidates appearing in positions from $\mathbf{1}$ to $\boldsymbol{k}$. We can see that

$$\boldsymbol{p_i} = \mathbf{0} \forall \boldsymbol{i} = [\mathbf{1} \ldots \boldsymbol{k}] \tag{2-4}$$

Now we shall assume the case of $\boldsymbol{i} > \boldsymbol{k}$.

Lemma 1. *The best candidate at* $\boldsymbol{i^{th}}$ *position is hired iff the maximum element of* $\boldsymbol{A}[\mathbf{1} \ldots \boldsymbol{i}-\mathbf{1}]$ *is at a position* $\leq \boldsymbol{k}$.

Proof. If the maximum element of $\boldsymbol{A}[\mathbf{1} \ldots \boldsymbol{i} - \mathbf{1}]$ is at a position $\leq \boldsymbol{k}$, then no person from $\boldsymbol{k} + \mathbf{1}$ to $\boldsymbol{i} - \mathbf{1}$ will be better qualified than the most qualified person from $\mathbf{1}$ to $\boldsymbol{k}$.
Now if the maximum element of $\boldsymbol{A}[\mathbf{1} \ldots \boldsymbol{i} - \mathbf{1}]$ is at a position $> \boldsymbol{k}$, then either that person will be hired, or some other person from the $\boldsymbol{(k+1)^{th}}$ to this person will be hired, and hence the person at $\boldsymbol{i^{th}}$ position won't be hired. □

We can now define $\boldsymbol{p_i}$ as the probability that the best qualified person in $\boldsymbol{A}[\mathbf{1} \ldots (\boldsymbol{i} - \mathbf{1})]$ appears from $\mathbf{1}$ to $\boldsymbol{k}$. Since all permutations of $\boldsymbol{n} - \mathbf{1}$ persons are possible in the $\boldsymbol{i} - \mathbf{1}$ positions.

$$\boldsymbol{Pr}[\text{Maximum of A[1...(i-1)] is from i to k]} = \frac{\boldsymbol{k}}{\boldsymbol{i} - \mathbf{1}} \tag{2-5}$$

Hence we can see that

$$\boldsymbol{p_i} = \frac{\boldsymbol{k}}{\boldsymbol{i} - \mathbf{1}} \tag{2-6}$$

Now we can express $\boldsymbol{P}[\varepsilon]$

$$\boldsymbol{P}[\varepsilon] = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{p_i} = \frac{\mathbf{1}}{\boldsymbol{n}} \left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \mathbf{0} + \sum_{\boldsymbol{i=k+1}}^{\boldsymbol{n}} \frac{\boldsymbol{k}}{\boldsymbol{i} - \mathbf{1}} \right) \tag{2-7}$$

Now we can take $\boldsymbol{k}$ outside, and we get

$$\boldsymbol{P[\varepsilon] = \frac{k}{n} \sum_{i=k+1}^{n} \frac{1}{i-1}} \tag{2-8}$$

Expressing this in terms of the Harmonic Number $\boldsymbol{H_n}$.

$$\boldsymbol{P[\varepsilon] = \frac{k}{n} \left(\sum_{i=1}^{n-1} \frac{1}{i} - \sum_{i=1}^{k-1} \frac{1}{i} \right)} \tag{2-9}$$

$$\boldsymbol{P[\varepsilon] = \frac{k}{n} \left(H_{n-1} - H_{k-1} \right)} \tag{2-10}$$

We know that $\boldsymbol{H_x \approx lnx + \gamma}$ where $\boldsymbol{\gamma}$ is the Euler- Mascheroni Constant.

$$\boldsymbol{P[\varepsilon] = \frac{k}{n} \left(ln(n-1) - ln(k-1) \right)} \tag{2-11}$$

$$\boldsymbol{P[\varepsilon] = \frac{k}{n} \left(ln(\frac{n-1}{k-1}) \right)} \tag{2-12}$$

Which is the required answer for part (i).

Now to find the value for which the probability is maximum we differentiate, we first express

$$\boldsymbol{P[\varepsilon] \leq \frac{k}{n} ln \left(\frac{n}{k} \right)} \tag{2-13}$$

Differentiating this ***w.r.tk*** we get

$$\boldsymbol{\frac{\partial}{\partial k} \left(\frac{k}{n} ln \left(\frac{n}{k} \right) \right) = \frac{1}{n} \left(ln \left(\frac{n}{k} \right) - 1 \right) = 0} \tag{2-14}$$

Since we know that $\boldsymbol{n}$ is finite then maximizing this

$$\boldsymbol{ln \left(\frac{n}{k} \right) = 1} \tag{2-15}$$

$$\boldsymbol{\frac{n}{k} = e} \tag{2-16}$$

$$\boldsymbol{k = \frac{n}{e}} \tag{2-17}$$

Now substituting for this value of $\boldsymbol{k}$ we get

$$\boldsymbol{P[\varepsilon] \leq \frac{1}{e}} \tag{2-18}$$

$\longrightarrow$ Answer

Problem 3.

Randomization in Error Rectification

The algorithm we define is as follows,

(a) Pick a number uniformly and randomly from $\mathbf{0}$ to $\boldsymbol{n-1}$ let this be $\boldsymbol{x}$.

(b) Now we compute $\boldsymbol{y = (z - x)\ mod\ n}$. This gives us two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ such that $\boldsymbol{(x + y) mod n = z}$.

(c) Substituting this value for $\boldsymbol{z}$, we get $\boldsymbol{F((x + y)\ mod\ n) = (F(x) + F(y)) mod m}$. So now we do not have to find $\boldsymbol{F(z)}$,but instead we find $\boldsymbol{F(z) = (F(x) + F(y))\ mod\ m}$. This completes our algorithm.

We shall now analyze the probability of $\boldsymbol{F(z)}$ being corrupt

Since $\boldsymbol{x}$ is a uniformly and randomly picked number

$$\boldsymbol{P}[\boldsymbol{F(x)} \text{ has been corrupted }] = \frac{\mathbf{1}}{\mathbf{5}} \tag{3-1}$$

$$\boldsymbol{P}[\boldsymbol{F(x)} \text{ has not been corrupted }] = \frac{\mathbf{4}}{\mathbf{5}} \tag{3-2}$$

We see that since $\boldsymbol{x}$ is a uniformly random variable, $\boldsymbol{y}$ can also be any of the numbers uniformly (but not randomly). We can see that here we do not require randomness, but the possibility that all values of $\boldsymbol{y}$ are equally possible, this means that adversary will NOT be able to change certain numbers alone, and give a failure probability greater than the fraction of table entries corrupted, This enables us to say that the distribution of $\boldsymbol{F(y)}$ will not change with the change according to the value $\boldsymbol{z}$ chosen and hence

$$\boldsymbol{P}[\boldsymbol{F(y)} \text{ has been corrupted }] = \frac{\mathbf{1}}{\mathbf{5}} \tag{3-3}$$

$$\boldsymbol{P}[\boldsymbol{F(y)} \text{ has not been corrupted }] = \frac{\mathbf{4}}{\mathbf{5}} \tag{3-4}$$

Since we find that we cannot argue the independence of the random variables. We shall use Union Theorem. $\boldsymbol{F(z)}$ is corrupted if either $\boldsymbol{F(x)}$ or $\boldsymbol{F(y)}$ is corrupted.

$$\boldsymbol{P}[\boldsymbol{F(z)} \text{ has been corrupted }] \leq \boldsymbol{P}[\boldsymbol{F(x)} \text{ has been corrupted }] + \boldsymbol{P}[\boldsymbol{F(y)} \text{ has been corrupted }] \tag{3-5}$$

(3-6)

$$\boldsymbol{P}[\boldsymbol{F(z)} \text{ corrupted}] \leq \frac{\mathbf{2}}{\mathbf{5}} \tag{3-7}$$

$$\boldsymbol{P}[\boldsymbol{F(z)} \text{ is correct}] \geq \frac{\mathbf{3}}{\mathbf{5}} \tag{3-8}$$

This completes our answer to part (i) of the problem.

For the second part, we run the same algorithm three times, and take the value than has been repeated most number of times, if no value has been repeated, then we output the first value.

We check the following probabilities.

Probability that all three values of $\boldsymbol{F(z)}$ are correct =

$$\boldsymbol{P}[\text{All 3 correct}] \geq \left(\frac{\boldsymbol{3}}{\boldsymbol{5}}\right)^{\boldsymbol{3}} = \frac{\boldsymbol{27}}{\boldsymbol{125}} \tag{3-9}$$

Since all three runs of the algorithm are independent of each other we have been able to apply the product rule.

Probability that any two values of $\boldsymbol{F(z)}$ are correct =

$$\boldsymbol{P}[\text{Some 2 correct}] \geq \binom{\boldsymbol{3}}{\boldsymbol{2}} \left(\frac{\boldsymbol{3}}{\boldsymbol{5}}\right)^{\boldsymbol{2}} \left(\frac{\boldsymbol{2}}{\boldsymbol{5}}\right) = \frac{\boldsymbol{54}}{\boldsymbol{125}} \tag{3-10}$$

Probability that the first value is correct and the rest two are wrong.

$$\boldsymbol{P}[\text{Some 2 correct}] \geq \left(\frac{\boldsymbol{3}}{\boldsymbol{5}}\right) \left(\frac{\boldsymbol{2}}{\boldsymbol{5}}\right)^{\boldsymbol{2}} = \frac{\boldsymbol{12}}{\boldsymbol{125}} \tag{3-11}$$

We can see that the following is a partition of the probable cases and hence if we are allowed to repeat our algorithm 3 times with independent values of $\boldsymbol{x}$ each time, then the total success probability.

$$\begin{aligned} \boldsymbol{P}[\boldsymbol{F^{(3)}(z)} \text{ is correct }] &\geq \frac{\boldsymbol{27}}{\boldsymbol{125}} + \frac{\boldsymbol{54}}{\boldsymbol{125}} + \frac{\boldsymbol{12}}{\boldsymbol{125}} = \frac{\boldsymbol{93}}{\boldsymbol{125}} && (3\text{-}12) \\ &\geq \boldsymbol{0.744} && (3\text{-}13) \end{aligned}$$

$\longrightarrow$ Answer

Problem 4.

Set Balancing

We shall first look at a single entry. Let $\boldsymbol{A_{ij}}$ be the element at the $\boldsymbol{i^{th}}$ row and $\boldsymbol{j^{th}}$ column of the matrix $\boldsymbol{A}$. We also define $\boldsymbol{Ab_i}$ as the $\boldsymbol{i^{th}}$ value of the column vector $\boldsymbol{Ab}$.

Proving that $\|\boldsymbol{Ab}\|_{\boldsymbol{\infty}}$ is of order $\boldsymbol{O(\sqrt{nlogn})}$ is equivalent to proving that

$$\|\boldsymbol{Ab}\|_{\boldsymbol{\infty}} \leq \boldsymbol{C}\sqrt{\boldsymbol{nlogn}} \tag{4-1}$$

for some finite $\boldsymbol{C}$ with very high probability.

Lemma 1. $|Ab_i| \leq c\sqrt{nlogn}$ *with very high probability where* Ab_i *is the* i^{th} *value of the Column Vector* Ab.

Proof. We can write Ab_i as

$$Ab_i = \sum_{j=1}^{n} A_{ij} b_j \tag{4-2}$$

Where A_{ij} is from the set $\{0, 1\}$ and b_j is randomly picked from $\{-1, +1\}$
It is very easy now to see the connection with the random walk problem.
In short we can see that if $A_{ij} = 0$ then it has no effect to the sum, this can be seen as equivalent to "not moving" in a step in the random walk. We shall now reorder the sum in such a way that all the 0's occur at the end, since b_i is uniformly random this will not change the problem. Let us assume,

$$m = \sum_{j=1}^{n} A_{ij} \tag{4-3}$$

This will count the total number of non-zero entries in the row, now we can see that this problem is equivalent to the random walk problem.
We shall now define a new row $A`_i$ of size m with all zero entries removed, and B with all the entries in the same index as the zero entries of A_i removed.

Now let us define a random variable

$$Y_i = \{1 \text{ if } B_i = +1 \quad , 0 \text{ if } B_i = -1\} \tag{4-4}$$

Expected value of $Y = \sum Y_i$

$$E[Y] = \frac{m}{2} \tag{4-5}$$

We can see that Y is the result of a series of bernouli trials and hence we can apply Chernoff bound on this.

We need to prove that

$$P[Y > (1 + \delta)\frac{m}{2}] \leq \frac{1}{m^c} \tag{4-6}$$

Applying Chernoff Bound, we get

$$e^{\frac{\mu\delta^2}{2}} \leq \frac{1}{m^c} \tag{4-7}$$

$$e^{\frac{m\delta^2}{4}} \leq e^{clnm} \tag{4-8}$$

$$\frac{m\delta^2}{4} \leq c\, lnm \tag{4-9}$$

$$\tag{4-10}$$

Now we find the value of $\boldsymbol{\delta}$ that will make this valid.

$$\boldsymbol{\delta = \sqrt{\frac{4c\ ln\ m}{m}}} \tag{4-11}$$

Applying this value for $\boldsymbol{\delta}$ we also set $\boldsymbol{c = 2}$ get

$$\boldsymbol{P[Y > (1 + \sqrt{\frac{8\ ln\ m}{m}})\frac{m}{2}] \leq \frac{1}{m^2}} \tag{4-12}$$

Rewriting

$$\boldsymbol{P[(Y - \frac{m}{2}) > \sqrt{8m\ ln\ m}) \leq \frac{1}{m^2}} \tag{4-13}$$

From this we can get

$$\boldsymbol{P[Ab_i > \sqrt{8m\ ln\ m}] \leq \frac{1}{m^2}} \tag{4-14}$$

Now $\boldsymbol{m}$ was the number of non-zero entries in the row and setting $\boldsymbol{E[m] = \frac{n}{2}}$

$$\boldsymbol{P[Ab_i > 2\sqrt{n\ ln\ n}] \leq \frac{4}{n^2}} \tag{4-15}$$

And hence we say $\boldsymbol{|Ab_i| \leq cnlogn}$ with very high probability. □

Now we apply union theorem on the result, and hence

$$\boldsymbol{P[Max(Ab_i) > 2\sqrt{n\ ln\ n}] \leq n\frac{4}{n^2}} \tag{4-16}$$

$$\boldsymbol{P[Max(Ab_i) > 2\sqrt{n\ ln\ n}] \leq \frac{4}{n}} \tag{4-17}$$

Hence we can say that $\boldsymbol{\|Ab\|_\infty = O(\sqrt{nlogn})}$ with very high probability

$\longrightarrow$ Answer

Submitted by Abhimanyu M A ***(1111002)*** *, Sumesh T A* ***(1111065)*** *on August 29, 2011.*