CS6170: Randomized Algorithms
Problem Set #2

NAME: Swathi Shree Narashiman
ROLL NO: EE22B149

MARKS: 35
DUE: March 2, 23:59

Problem 1 **5 marks**

Consider a universe $\mathscr{U}$ of size m, and hash table T of size n. We will say that a family $\mathscr{H}$ of hash functions from $[m] \to [n]$ is k-perfect, if for every set $S \subseteq \mathscr{U}$ where $|S| = k$, there exists an $h \in \mathscr{H}$ such that for every $x \neq y \in S$, $h(x) \neq h(y)$.

Show that if $m = 2^{\Omega(n)}$, then there exists an n-perfect family $\mathscr{H}$ of hash functions from $[m] \to [n]$ such that $|\mathscr{H}| = m^{O(1)}$.

Note that the question is only asking you to prove the existence of such a family, not to give an efficient algorithm to construct it. Start as follows: Fix a set $S \subseteq \mathscr{U}$ such that $|S| = n$, and compute the probability that a random function from $[m] \to [n]$ is perfect for S. Now, try to find the value of k so that the probability that there is some set $S \subseteq [n]$ for which k random functions are not perfect is less than 1.

Solution:

Let $h \in \mathscr{H}$, be a perfect hashing function and let $|S| = n$,

The number of possible subsets is given by $\binom{m}{n}$

For the given h, let the number of elements from [m] mapped to the i^{th} position be m_i.

$$\sum_i m_i = m$$

For h to be perfect hashing, let S' be the set of all possible subsets of size n, these subsets can have only 1 element from the m_i elements in the i^{th} position

$$|S'| = \prod_i m_i$$

By AM-GM inequality,

$$|S'| = \prod_i m_i \leq (\frac{m}{n})^n$$

For the entire hash family, the number of subsets is given by,

$$|\mathscr{H}| \left(\frac{m}{n}\right)^n$$

$\mathscr{H}$ is a perfect hash family if:

$$\binom{m}{n} \leq |\mathscr{H}| \left(\frac{m}{n}\right)^n$$

This gives us,

$$|\mathcal{H}| \geq \frac{\binom{m}{n}}{\left(\frac{m}{n}\right)^n} = \binom{m}{n}\left(\frac{n}{m}\right)^n$$

$$|\mathcal{H}| \geq \frac{(1-\frac{1}{m})(1-\frac{2}{m})(1-\frac{3}{m})...(1-\frac{n-1}{m})}{(1-\frac{1}{n})(1-\frac{2}{n})(1-\frac{3}{n})...(1-\frac{n-1}{n})}$$

Using, for x<=1

$$\frac{1}{1-x} \geq 1+x$$

$$|\mathcal{H}| \geq (1-\frac{1}{m})(1-\frac{2}{m})(1-\frac{3}{m})...(1-\frac{n-1}{m})(1+\frac{1}{n})(1+\frac{2}{n})(1+\frac{3}{n})...(1+\frac{n-1}{n})$$

Let's asymptotically approximate $1+x \approx e^x$; $1-x \approx e^{-x}$,

$$|\mathcal{H}| \approx e^{-\frac{n^2}{2m}} e^{\frac{n-1}{2}} \approx e^{\frac{n}{2}(1-\frac{n}{m})}$$

If, $m = 2^{\Omega(n)}$

$$n = O(\log m)$$

since m grows exponentially with n, $1-\frac{n}{m} \approx 1$,

$$|\mathcal{H}| \approx e^{\frac{n}{2}} = e^{\frac{1}{2}O(\log m)} = m^{O(1)}$$

Hence proved.

Note: The proof I have done above is without the use of Sterling approximation

Problem 2 — 5 marks

Suppose that we have a hash table of size n to which we hash using the following strategy: Choose two hash functions h_1 and h_2 uniformly at random. For an element, we compute $h_1(x)$ and $h_2(x)$ and try to place x in $h_1(x)$ followed by $h_2(x)$. If both the positions are occupied, then we rehash the table by choosing new set of hash functions. What is the expected number of elements x that you can hash before the first rehashing?

Solution: Let X_i denote an indicator random variable given by,

$$X_i = \begin{cases} 1 & \text{, if } i^{th} \text{ element is successfully hashed after i-1 successful hashes} \\ 0 & \text{, Unsuccessful hash after i-1 successful hashes} \end{cases}$$

. Clearly,

$$Pr[X_1 = 1] = 1$$

for $i \geq 1$, the i-1 elements are successfully hashed and they occupy positions given by a set S, where |S| = i-1

$$Pr[X_i = 1 | X_1 = 1, X_2 = 1, ..., X_{i-1} = 1] = 1 - Pr[h_1(i) \in S, h_2(i) \in S] = 1 - \frac{(i-1)^2}{n^2}$$

$$Pr[X_i = 0] =$$

$$Pr[X_1 = 1].Pr[X_2 = 1|X_1 = 1].Pr[X_3 = 1|X_1 = X_2 = 1].....Pr[X_i = 0|X_1 = X_2 = .. = X_{i-1} = 1]$$

$$Pr[X_i = 0] = (1 - \frac{1}{n^2})(1 - \frac{4}{n^2})....(\frac{(i-1)^2}{n^2})$$

Using $1 - x \leq e^{-x}$,

$$Pr[X_i = 0] \leq \frac{(i-1)^2}{n^2} e^{-\frac{1}{n^2}\sum_{j=1}^{i-2} j^2} = \frac{(i-1)^2}{n^2} e^{-\frac{1}{6n^2}(i-1)(i)(2i-1)}$$

$$Pr[X_i = 0] \leq \frac{(i-1)^2}{n^2} e^{-\frac{1}{6n^2}(i-1)^3}$$

Let X denote the number of successful insertions,

$$E[X] \approx \sum_i (i-1) Pr[X_i = 0] \leq \sum_i \frac{(i-1)^3}{n^2} e^{-\frac{1}{6n^2}(i-1)^3}$$

Let's bound E[X] by integral,

$$E[X] \leq \int_0^n \frac{x^3}{n^2} e^{-\frac{x^3}{6n^2}} dx$$

Let,

$$u = \frac{x^3}{6n^2}$$

$$du = \frac{x^2}{2n^2} dx$$

$$E[X] \approx \int_0^n 2xe^{-u} du = 2(6n^2)^{\frac{1}{3}} \int_0^n u^{\frac{1}{3}} e^{-u} du \leq 2(6n^2)^{\frac{1}{3}} \Gamma\left(\frac{4}{3}\right)$$

The integral is a gamma function which converges to a constant,

$$E[X] \approx \theta(n^{\frac{2}{3}})$$

Therefore, the expected number of elements hashed before rehashing is $\theta(n^{\frac{2}{3}})$

Problem 3 — 5 marks

Consider the scenario where we are using a Bloom filter to store a set of size m. Unlike what we did in class, let us assume that we choose a hash function from a pairwise-independent hash family $\mathcal{H}$. How large should the Bloom filter be so that the probability of false positives is at most δ.

Solution:

Let us choose a pairwise independent hash function $h \in \mathcal{H}$, this gives us

$$Pr[h(x_1) = y_1 \cap h(x_2) = y_2] = Pr[h(x_1) = y_1] * Pr[h(x_2) = y_2]$$

A Bloom filter is a probabilistic data structure used to store a set S with m elements. It consists of:

- A bit array B of length n.
- k hash functions $h_1, h_2, \ldots, h_k : U \to \{1, 2, \ldots, n\}$.

The set S is stored as follows: For each $x \in S$, we set

$$B[h_i(x)] = 1, \quad \text{for every } 1 \leq i \leq k.$$

If multiple hash functions map to the same position, the bit is set to 1 only once.

A membership query is answered as follows: Given $x \in U$, answer "yes" if

$$B[h_i(x)] = 1, \quad \text{for every } 1 \leq i \leq k.$$

If $x \in S$, then the membership query answers correctly with probability 1. Hence, let us calculate the error when $x \notin S$, that is, the case of false positives.

Let us assume that all m elements have been hashed , let r be a position in the bit string B,

$$Pr[B[r] = 0] = Pr(\textit{ None of the m elements got hashed into position } r)$$

$$Pr[B[r] = 0] = Pr(\bigcap_j \bigcap_i h_j(i) \neq r)$$

Since, $h_i's$ are sampled uniformly at random,

$$Pr[B[r] = 0] = \prod_j Pr(\bigcap_i h_j(i) \neq r) = (Pr(\bigcap_i h_j(i) \neq r))^k$$

$$Pr[B[r] = 0] = (1 - Pr(\bigcup_i h_j(i) = r))^k$$

We know that,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} P\left(\bigcap_{i=1}^{n} A_i\right)$$

By **Bonferroni inequality**:

$$P\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j)$$

Therefore,

$$Pr(\bigcup_i h_j(i) = r) \geq \sum_{i=1}^{m} P(h_j(i) = r) - \sum_{i<k} P(h_j(i) = r \cap h_j(k) = r)$$

By using pairwise independence,

$$Pr(\bigcup_i h_j(i) = r) \geq \frac{m}{n} - \binom{m}{2}\frac{1}{n^2} \geq \frac{m}{n} - \frac{m^2}{2n^2} \geq log(1+\frac{m}{n})$$

Hence,

$$Pr[B[r]=0] \leq (1-log(1+\frac{m}{n}))^k \leq e^{-klog(1+\frac{m}{n})} = (\frac{n}{m+n})^k$$

$$Pr[B[r]=1] \geq 1-(\frac{n}{m+n})^k$$

Now to give a false positive result for a given $y \notin S$, we need to have the k indices corresponding to y to be 1

$$Pr[Error] = \prod_j Pr(B[h_j(y)]=1) \approx \left(1-(\frac{n}{m+n})^k\right)^k$$

$$Pr[Error] \approx e^{-k(\frac{n}{m+n})^k} \approx \delta$$

Let us minimize this with respect to k, so we maximize the power of e, maximum occurs at k given by,

$$k = \frac{1}{\ln(1+\frac{m}{n})}$$

This gives us,

$$k = \ln(1/\delta)$$

Therefore,

$$n \approx m\ln(\frac{1}{\delta})$$

Collaborator: Nitin G

Problem 4 — 11 marks

Suppose that we throw n balls into n bins uniformly at random. Let X_i denote the indicator random variable that is 1 if the bin i is empty. Let $X = \sum_{i=1^n} X_i$ denote the number of empty bins.

(a) (3 marks) Use the Poisson approximation to prove a concentration bound for $\Pr[X \geq (1+\delta)\mathbb{E}[X]]$.

Solution: Since $X_i's$ are Benoulli random variable,

$$Pr(X_i=1) = p = Pr(i^{th} \textit{ bin is empty}) = Pr(n \textit{ balls occupy n-1 bins})$$

$$Pr(X_i=1) = p = \left(\frac{n-1}{n}\right)^n$$

By using the inequality, $1-x \leq e^{-x}$, we can make an asymptotic approximation

$$Pr(X_i=1) = p \approx e^{-\frac{n}{n}} \approx e^{-1}$$

Hence, we can calculate the E[X] as follows,

$$E[X] = E[\sum_i X_i] = \sum_i E[X_i]$$

Due to bernoulli distribution,

$$E[X] = \sum_i Pr[X_i = 1] = np \approx ne^{-1}$$

Consider $Y \sim Poisson(np)$ to be the poisson approximate of X,
By chernoff bounds, we have

$$Pr(Y \geq (1+\delta)\lambda) \leq (\frac{e^\delta}{(1+\delta)^{1+\delta}})^\lambda$$

Therefore,

$$Pr(X \geq (1+\delta)E[X]) \leq (\frac{e^\delta}{(1+\delta)^{1+\delta}})^{np}$$

Now let us try to prove that without using Poisson approximation as follows. Let $Y_i \sim \text{Ber}((1-1/n)^n)$, and let $Y = \sum_{i=1}^n Y_i$.

(b) (2 marks) Show that for any $k \geq 1$, $\mathbb{E}[X_1 X_2 \cdots X_k] \leq \mathbb{E}[Y_1 Y_2 \cdots Y_k]$.

Solution: Since X_i is a bernoulli random variable with,

$$E[X_1X_2.....X_k] = 1 * Pr[All\ k\ bins\ are\ empty] + 0 * Pr[At\ least\ 1\ bin\ is\ non\ empty]$$

This gives us,

$$E[X_1X_2.....X_k] = \frac{(n-k)^n}{n^n} = (1-\frac{k}{n})^n$$

Since, $Y_i's$ are independent bernoulli random variables

$$E[Y_1Y_2......Y_k] = Pr(\bigcap_{i=1}^n Y_i) = \prod_{i=1}^n Pr(Y_i = 1)$$

This gives us,

$$E[Y_1Y_2......Y_k] = ((1-1/n)^n)^k = (1-1/n)^{nk}$$

To prove:

$$(1-\frac{k}{n})^n \leq (1-1/n)^{nk}$$

Taking log on both sides,

$$nlog(1-\frac{k}{n}) \leq nklog(1-\frac{1}{n})$$

Let us use taylor approximation of $log(1-x) = -x - \frac{x^2}{2}$,

$$n(-\frac{k}{n}-\frac{k^2}{2n^2}) \leq nk(-\frac{1}{n}-\frac{1}{2n^2})$$

On simplification we get,

$$\frac{k^2}{n} \geq \frac{k}{n^2}$$

Since, for $k \geq 1$,

$$k \leq k^2$$

$$(1-\frac{k}{n})^n \leq (1-1/n)^{nk}$$

Therefore,

$$E[X_1X_2\cdots X_k] \leq \mathbb{E}[Y_1Y_2\cdots Y_k]$$

Hence proved.

(c) (2 marks) Show that $\mathbb{E}[e^{tX}] \leq \mathbb{E}[e^{tY}]$ for all $t > 0$.

Solution: By multinomial theorem we have,

$$E[X^k] = E[(\sum_i X_i)^k] = E[\sum_{i_1,i_2,\ldots,i_k} X_{i_1}X_{i_2}\ldots.X_{i_k}]$$

By linearity of expectations,

$$E[X^k] = \sum_{i_1,i_2,\ldots,i_k} E[X_{i_1}X_{i_2}\ldots.X_{i_k}]$$

Where $1 \leq i_1, i_2, \ldots.. \leq n$

By using the inequality we established in part (b),

$$E[X^k] \leq E[Y^k]$$

By expansion of e^x we have,

$$\mathbb{E}[e^{tX}] = \sum_{i=0}^{\infty} \frac{t^i\mathbb{E}[X^i]}{i!} \qquad \mathbb{E}[e^{tY}] = \sum_{i=0}^{\infty} \frac{t^i\mathbb{E}[Y^i]}{i!}$$

Therefore,

$$E[e^{tX}] \leq E[e^{tY}]$$

Hence proved.

(d) (4 marks) Use the parts above to bound $\Pr[X \geq (1+\delta)\mathbb{E}[X]]$. You may want to read the book on how the Chernoff bound is proved.

Solution:

$$Pr(X \geq (1+\delta)E[X]) = Pr(e^{tX} \geq e^{(1+\delta)E[X]})$$

By Markov inequality, we can write,

$$Pr(e^{tX} \geq e^{(1+\delta)E[X]}) \leq \frac{E[e^{tX}]}{e^{t(1+\delta)E[X]}}$$

By the result from part (c),

$$Pr(e^{tX} \geq e^{t(1+\delta)E[X]}) \leq \frac{E[e^{tY}]}{e^{t(1+\delta)E[X]}} ------1$$

Now, let's compute the RHS using the independence of Y_i

$$E[e^{tY}] = E[e^{t\sum_i Y_i}] = E[\prod_i e^{tY_i}] = \prod_i E[e^{tY_i}]$$

For Y$_i$, Let p $= \left(1 - \frac{1}{n}\right)^n \approx e^{-1}$ (asymptotic approximation)

$$E[e^{tY_i}] = (pe^t - p + 1)$$

By using $1 + x \leq e^x$,

$$E[e^{tY_i}] \leq e^{p(e^t-1)}$$
$$E[e^{tY}] = E[e^{tY_i}]^n \leq e^{np(e^t-1)}$$

As calculated in part (a)

$$E[X] = np$$

Therefore, from (1)

$$\Pr\left(e^{tX} \geq e^{t(1+\delta)E[X]}\right) \leq \frac{e^{np(e^t-1)}}{e^{tnp(1+\delta)}} = \left(\frac{e^{e^t-1}}{e^{t(1+\delta)}}\right)^{np}$$

Let us find the minimum value for the quantity on RHS, this happens at $t = ln(1+\delta)$

$$Pr(e^{tX} \geq e^{t(1+\delta)E[X]}) \leq \left(\frac{e^{\delta}}{e^{(1+\delta)ln(1+\delta)}}\right)^{np}$$

Therefore,

$$\Pr\left(e^{tX} \geq e^{t(1+\delta)E[X]}\right) \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{np}$$

Hence,

$$Pr(X \geq (1+\delta)E[X]) \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{np}$$

This bound is the same as the one derived in part(a) by approximating X_i to poisson distribution Y_i

Problem 5 **9 marks**

Consider the scenario of n autonomous agents in a distributed setting vying for a resource, say a printer on a network. Assume that there are n copies of the resource available, but an agent will be served by a copy of the resource if it is the only agent that has chosen that instance of the resource. If there are multiple agents that choose the same copy, then that copy gets blocked and the agents will have to wait for the next round. Our goal is to understand the number of rounds before all n agents get served.

Let us model this as a balls into bins process. Here the agents are the balls and the copies of the resource are the bins. In the first round, n balls are thrown independently and uniformly at random into n bins. After round i, we discard all balls that fell into a bin by itself in round i. We continue with the remaining balls in a similar fashion for round $i+1$, where they are thrown independently and uniformly at random into n bins.

(a) (3 marks) If there are b agents waiting to be served at the start of a round, what is the expected number of agents remaining at the start of the next round?

Solution: Let E_j be the indicator random variable that the agent j is retained in the next round. This happens if the agent chooses one of the bins that is already chosen by some other agent.

We can derive the following:

$$Pr[E_1 = 1] = 0$$

Since, the first ball definitely goes into a bin on its own.

$$Pr[E_2 = 1] = \frac{1}{n}$$

By law of total probability,

$$Pr[E_3 = 1] = Pr[E_1^c E_2 E_3] + Pr[E_1^c E_2^c E_3] = \frac{2}{n} - \frac{1}{n^2}$$

$$Pr[E_4 = 1] = Pr[E_1^c E_2 E_3 E_4] + Pr[E_1^c E_2^c E_3 E_4] + Pr[E_1^c E_2 E_3^c E_4] + Pr[E_1^c E_2^c E_3^c E_4]$$

$$Pr[E_4 = 1] = \frac{3}{n} - \frac{3}{n^2} + \frac{1}{n^3}$$

Similarly, we can get,

$$Pr[E_5 = 1] = \frac{4}{n} - \frac{6}{n^2} + \frac{4}{n^3} - \frac{1}{n^4}$$

By observing the pattern we can say that,

$$Pr[E_j = 1] = \sum_{k=1}^{j-1} (-1)^{k+1} \binom{j-1}{k} \frac{1}{n^k}$$

Let X_{i+1} be the random variable denoting the number of agents remaining at the start of the $i+1^{th}$ round,

$$X_{i+1} = \sum_{j=1}^{b} E_j$$

This implies that,

$$E[X_{i+1}] = E[\sum_j E_j] = \sum_j E[E_j] = \sum_j 1.Pr[E_j = 1]$$

$$E[X_{i+1}] = \sum_j \sum_{k=1}^{j-1} (-1)^{k+1} \binom{j-1}{k} \frac{1}{n^k}$$

$$E[X_{i+1}] = \frac{1}{n}(1+2+..+b-1) - \frac{1}{n^2}(\binom{2}{2}+\binom{3}{2}+....+\binom{b-1}{2}) + \frac{1}{n^3}(\binom{3}{3}+\binom{4}{3}+....+\binom{b-1}{3}) + ...$$

Further simplification gives us,

$$E[X_{i+1}] = \frac{b(b-1)}{2n} - \frac{(b-2)(b-1)(b)}{6n^2} + \ldots$$

Hence, we can bound the expected number of agents at the start of the next round as,

$$\frac{b(b-1)}{2n} - \frac{(b-2)(b-1)(b)}{6n^2} \leq E[X_{i+1}] \leq \frac{b(b-1)}{2n}$$

Collaborator: Nitin G

(b) (6 marks) Show that there is a constant c such that all the agents will be served within $c \log\log n$ rounds with probability at least $1 - o(1)$.

Solution: Let X_i denote the number of the agents at the start of the i^{th} round, by the relation found in part (a), we have

$$E[X_i + 1 | X_i] \leq \frac{X_i(X_i - 1)}{2n} \leq \frac{X_i^2}{2n}$$

We know that, $X_1 = n$,

$$E[X_2] \approx \frac{n}{2} = \frac{n}{2^1}$$

Hence we can derive,

$$E[X_3] \approx \frac{n}{8} = \frac{n}{2^3}$$

$$E[X_4] \approx \frac{n}{128} = \frac{n}{2^7}$$

$$E[X_5] \approx \frac{n}{2^{15}}$$

Therefore, we can derive the following,

$$E[X_i] \approx \frac{n}{2^{2^{i-1}-1}}$$

The process stops when at the start of a given round the number of agents is 1,

$$E[X_i] \approx 1$$

On simplification we get,

$$n \approx 2^{2^{i-1}-1}$$

$$2^{i-1} - 1 \approx \log n$$

$$i - 1 \approx \log(1 + \log n)$$

$$i \approx O(\log\log n)$$

Probability Analysis: By Chebysev Bounds

Let us note that X_i is the sum of Dependent bernoulli trials E_j for j agents

$$Pr[|X_i - E[X_i]| \geq \epsilon] \leq \frac{Var[X_i]}{\epsilon^2}$$

Since, $X_i = \sum_j E_j$,

$$Var[X_i] = \sum_j Var[E_j] + 2\sum_{j<k} Cov(E_j, E_k)$$

But in this setting, if $E_j = 1$, then it falls into one of the already occupied bins, and it reduces the probability that $E_k = 1$. This means that,

$$Cov(E_j, E_k) < 0$$

$$Var[E_j] = p_j(1-p_j)$$

$$Var[X_i] \leq \sum_j Var[E_j] = \sum_j p_j(1-p_j) \leq E[X_i]$$

On simplification,

$$Pr[|X_i - E[X_i]| \geq \epsilon] \leq \frac{E[X_i]}{\epsilon^2} \approx \frac{n}{\epsilon^2 2^{2^{i-1}-1}}$$

When $i = c \log\log n$,

$$Pr[|X_i - E[X_i]| \geq \epsilon] \leq \frac{1}{\epsilon^2} = o(1)$$

Therefore,

$$Pr[|X_i - E[X_i]| < 1] \geq 1 - o(1)$$

Therefore, we can approximate $X_i \approx E[X_i]$.
This means that all agents will be served within $c \log\log n$ rounds with a probability of at least 1 - o(1).

Collaborator: Nitin G