CS 498ABD: Algorithms for Big Data, Spring 2019
Midterm: February 28, 2019

Name:	Ray Ying
NetID:	xinruiy2

- This is a closed-book, closed-notes, closed-electronics exam. If you brought anything except your writing implements, put it away for the duration of the exam. In particular, you may not use *any* electronic devices other than those that are medically necessary.
- We will scan the exam into Gradescope. Please do not write outside the black boxes on each page; these indicate the area of the page that the scanner can actually see.
- This answer booklet is **double-sided**!
- If you run out of space for an answer, feel free to use the scratch pages at the back of the answer booklet, but **please clearly indicate where we should look.**
- **Please read the entire exam before writing anything.** There are five numbered problems.
- **You have 120 minutes (2 hours).**
- **Proofs are required only if we specifically ask for them.**

Do not write on this page

Probabilistic Inequalities

- Markov's inequality: For a non-negative random variable X and $t > 0$, $\Pr[X > t] \leq \mathrm{E}[X]/t$.
- Chebyshev's inequality. For a random variable X $\Pr[|X - \mathrm{E}[X]| \geq a] \leq \mathrm{Var}[X]/a^2$
- Chernoff bound for sum of non-negative bounded random variables. Let $X_1, \ldots, X_k$ be k independent binary random variables such that, for each $i \in [1,k]$, $\mathrm{E}[X_i] = \Pr[X_i = 1] = p_i$. Let $X = \sum_{i=1}^{k} X_i$. Then $\mathrm{E}[X] = \sum_i p_i$.
  - Upper tail bound: For any $\mu \geq \mathrm{E}[X]$ and any $\delta > 0$,
  $$\Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$$
  - Lower tail bound: For any $0 < \mu < \mathrm{E}[X]$ and any $0 < \delta < 1$,
  $$\Pr[X \leq (1-\delta)\mu] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}$$

  The above bounds can be simplified when $0 \leq \delta < 1$, as follows:
  $$\Pr[X \geq (1+\delta)\mu] \leq e^{\frac{-\delta^2\mu}{3}} \text{ and } \Pr[X \leq (1-\delta)\mu] \leq e^{\frac{-\delta^2\mu}{2}}$$
- Chernoff bound for sum of bounded random variables. Let $X_1, \ldots, X_k$ be k independent random variables such that, for each $i \in [1,k]$, $X_i \in [-1,1]$. Let $X = \sum_{i=1}^{k} X_i$. For any $a > 0$,
  $$\Pr[|X - \mathrm{E}[X]| \geq a] \leq 2\exp\left(\frac{-a^2}{2n}\right).$$

CS 498ABD, Spring 2019 Midterm: Problem 1

Let $h : [n] \to [m]$ be a random hash function chosen from a 3-wise independent family of hash functions. For a fixed item i let Y be the number of items $i' \neq i$ that collide with i under h.

- What is $E[Y]$?
- What is $Var[Y]$ as a function of m, n? *Hint:* Use 3-wise independence here.
- Using Chebyshev, what is $Pr[Y \geq a]$ where $a \geq 1$ is some integer. Express this as a function of a, m, n.

$$E[Y] = \sum_{i' \neq i} E[Y_{i'}] = (n-1) \times \frac{1}{m} = \frac{n-1}{m}$$

n: the total number of item hashed (contains i)

$$Var[Y] = \sum_{i'} Var[h(i') = a, i \neq i' \mid h(i) = a]$$

$$= (n-1) \cdot \frac{1}{m} \cdot \frac{m-1}{m}$$

$$Var[Y_i] = \frac{1}{m} \cdot \frac{m-1}{m} \quad \text{(binomial)}$$

(We fix an item, by 3-wise independent, we still have 2-wise independent)

$$Pr[Y \geq a] \Rightarrow Pr[Y - E[Y] \geq a - E[Y]]$$

$$Pr[Y - E[Y] \geq a - E[Y]] \leq Pr[|Y - E[Y]| \geq |a - E[Y]|]$$

$$Pr[|Y - E[Y]| \geq |a - E[Y]|] \leq \frac{Var(Y)}{(a - E[Y])^2} = \frac{(n-1) \cdot \frac{1}{m} \cdot \frac{m-1}{m}}{(a - \frac{n-1}{m})^2}.$$

this is a function of a, m, n.

CS 498ABD, Spring 2019 Midterm: Problem 2

We have seen the use of the median trick for improving the probability of success. Suppose we have an estimator X for a quantity α of interest such that $E[X] = \alpha$ and $\Pr[|X - \alpha| \geq \epsilon\alpha] < \rho$ for some $\rho < 1/2$. We wish to improve the error probability to δ for some desired δ. We have seen the use of the median trick for this. We compute independent estimators $X_1, X_2, \ldots, X_h$ in parallel and output the median Z of the computed estimators. How large should h be to guarantee that $\Pr[|Z - \alpha| \geq \epsilon\alpha] \leq \delta$ (as a function of ρ and δ)? Use one of the Chernoff inequalities and briefly justify your bound.

$$h = \frac{8}{(2\rho-1)^2} \ln\left(\frac{2}{\delta}\right)$$

By chernoff bound:

The expectation bad estimators in h copies is less than ρh. In order to have a bad median, we need half of the estimators to be bad, let Y to count the number of bad estimator. If $Y = \frac{h}{2}$, the $Y - E[Y] > \frac{h}{2} - \rho h$.

we want to show

$$\Pr[|Y - E[Y]| \geq \frac{h}{2} - \rho h] \leq 2e^{-\frac{(\frac{h}{2}-\rho h)^2}{2h}} \leq \delta$$

We want to have

$$2e^{-\frac{(\frac{h}{2}-\rho h)^2}{2h}} \leq \delta$$

$$e^{-\frac{(\frac{h}{2}-\rho h)^2}{2h}} \leq \frac{\delta}{2}$$

$$\frac{2}{\delta} \leq e^{\frac{(\frac{h}{2}-\rho h)^2}{2h}}$$

$$2\ln\left(\frac{2}{\delta}\right) \leq \frac{(\frac{h}{2}-\rho h)^2}{h}$$

$$2\ln\left(\frac{2}{\delta}\right) \leq \frac{h}{4} - \rho h + \rho^2 h$$

$$8\ln\left(\frac{2}{\delta}\right) \leq h - 4\rho h + 4\rho^2 h$$

$$8\ln\left(\frac{2}{\delta}\right) \leq h(4\rho^2 - 4\rho h + 1)$$

go to page 9

CS 498ABD, Spring 2019 Midterm: Problem 3

Let $A[1..n]$ be a sorted array of n integers. Given an integer x, one way to decide if $x \in A$ is to use binary search. In this problem, we analyze a randomized version of binary search to find x.

Consider a randomized variant of binary search where one picks a *random* index $i \in [n]$ and compares $A[i]$ with x. If $A[i] = x$, then it terminates with the answer "yes"; if $A[i] \neq x$, then it recurses appropriately.

- Write down a formal description of randomized binary search including taking care of base cases.
- Prove that the expected running time for searching any given item x is $O(\log n)$.
- **Extra credit:** Prove that the running time of the algorithm is $O(\log n)$ with high probability.

The function binaryR(A, x):

```
        i ← 1 to n
    If(A[i] == x):
        return "Yes"
    else:
        If ( |A[i]| == 1 ):
            return "false"
        If ( A[i] > x ):
            return binaryR(A[x+1, |A|], x)
        else:
            return binaryR(A[1, x-1], x)
```

We want to analysis using recursive tree,

$T(n) = 1 + T(i-1)$ $A[i] < x$ | $T(n) = 1 + T(n-i+1)$ $A[i] > x$

$E(T(n)) = 1 + E\left(\sum_{i:A[i]<x} T(i-1) p(i) + \sum_{i:A[i]>x} T(n-i+1) \cdot p(i)\right)$

Since we are randomly pick i from 1 to n, $p(i)$ is same for all i. go to page 11.

CS 498ABD, Spring 2019 Midterm: Problem 4

Recall the algorithm to estimate the number of distinct elements in a stream using an ideal hash function $h : [n] \to [0,1]$. The algorithm maintains the minimum of the hash value seen in the stream, say z, and outputs $\frac{1}{z} - 1$ as the estimator for the number of distinct elements. Suppose there was a mistake in the implementation and instead of storing the minimum hash value seen, z stored the *maximum* hash value. How would you use z now to estimate the number of distinct elements? Briefly justify your answer.

Since it's a ideal hash function, then for d distinct value, we can divide the range of $[0, 1]$ to $d+1$ piece. The minimum is $\frac{1}{d+1}$ (expected) and the maximum would by $\frac{d}{d+1}$. (expected)

$$z = \frac{d}{d+1} \Rightarrow z(d+1) = d$$

$$zd + z = d$$

$$z = d(1-z)$$

$$\frac{z}{1-z} = d.$$

CS 498ABD, Spring 2019 Midterm: Problem 5

Consider F_2 estimation via the AMS algorithm using 4-wise independent hash functions. In this problem, the high-level goal is to process two different streams coming in at two different locations and use this estimator to estimate the F_2 distance between the streams.

Let σ_1 and σ_2 be two streams. Let $\{f_{1,i} : i \in [n]\}$ and $\{f_{2,i} : i \in [n]\}$ denote the frequencies of σ_1 and σ_2, respectively. The F_2 distance between the streams is the sum

$$\sum_{i=1}^{n}(f_{1,i} - f_{2,i})^2.$$

Recall that the AMS estimator computes a value Z where the expected value of Z^2 is the F_2 of the stream. (One then takes averages and then medians of many copies to improve the accuracy.) The basic framework to estimate the F_2 distance of σ_1 and σ_2 is as follows. We first produce an estimate Z_1 for σ_1 and an estimate Z_2 for σ_2. We then somehow combine Z_1 and Z_2 to estimate the distance. Here we have two design decisions.

- When producing Z_1 and Z_2, should we use the same hash function, or two independent ones?
- How do we combine Z_1 and Z_2 so that the expected value is $\sum_{i=1}^{n}(f_{1,i} - f_{2,i})^2$.

Answer the above with some brief justification.

We should use same hash function so that we are comparing the matched $f_{1,i}$ with $f_{2,i}$.

Each time, we are adding $h(f_{1,i})$ to Z_1, $h(f_{2,i})$ to Z_2, output $(Z_1 - Z_2)^2$

$$E[(Z_1-Z_2)^2] = \left(\sum_i^n f_{1,i} \cdot h(f_{1,i}) - \sum_i f_{2,i} \cdot h(f_{2,i})\right)^2$$

as we are using same hash table,

$$= \sum_i^n (f_{1,i} h(f_{1,i}) - f_{2,i} h(f_{2,i}))^2$$

$$= \sum_i^n (f_{1,i} - f_{2,i})^2$$

$E[h(f_{1,i})] = E[h(f_{2,i})] = 0$, $E[h^2(f_{1,i})] = E[h^2(f_{2,i})] = 1$

This page is for extra work.

This page is for extra work.

$$8 \ln\left(\frac{2}{\delta}\right) \leq h\,(4p^2 - 4p + 1)$$

$$8 \ln\left(\frac{2}{\delta}\right) \leq h\,(2p-1)^2$$

$$\frac{8}{(2p-1)^2} \ln\left(\frac{2}{\delta}\right) \leq h$$

This page is for extra work.

This page is for extra work.

$$E(T(n)) = 1 + E\left(\sum_{i:A[i]<x} T(i-1) + \sum_{i:A[i]>x} T(n-i+1)\right)$$

Since we know that the expectation for i is $\frac{n}{2}$ by linerality of expectation, the by expectation, the worst case is each time we have to look into half of the array next iteration. The the expected level of recursive tree will be $\log n$. Each level takes 1 then the total running time is $\log n$.

We proved that expected size for next iteration is $\frac{n}{2}$. by markov, let X be the size for next iteration, $\Pr[X > \frac{3}{4}n] < \frac{E[X]}{\frac{3}{4}n} = \frac{2}{3}$.

The expected level if each iteration we can have $\frac{3}{4}n$ to recurse next time is $\log_{\frac{3}{4}} n$.

The probability that it ends with $O(\log_{\frac{3}{4}} n)$ is less than $1 - (\frac{2}{3})^{\log_{\frac{3}{4}} n} = 1 - (\frac{1}{2} \cdot \frac{4}{3})^{\log_{\frac{3}{4}} n}$

$$= 1 - (\frac{1}{2})^{\log_{\frac{3}{4}} n} \frac{1}{n} \geq 1 - \frac{1}{n}.$$

This page is for extra work.