New York University Tandon School of Engineering

CS-GY 6763: Midterm Exam.

Friday Mar. 14th, 2023.

47 points total (+ 4 bonus)

You have 1 hour, 15 minutes to take the exam.
Show your work to receive full (and partial) credit.

You may find the following statement of the Chernoff bound from class useful.

Theorem 1 (Chernoff Bound)**.** *Let $X_1, X_2, \ldots, X_k$ be independent $\{0,1\}$-valued random variables and let $p_i = \mathbb{E}[X_i]$, where $0 < p_i < 1$. Let $S = \sum_{i=1}^{k} X_i$ and $\mathbb{E}[S] = \mu$. For $\epsilon \in (0,1)$,*

$$\Pr[|S - \mu| \geq \epsilon\mu] \leq 2e^{-\epsilon^2\mu/3}.$$

1. Always, sometimes, never. (**15pts – 3pts each**)

Indicate whether each of the following statements is always true, sometimes true, or never true. To receive full credit, **PROVIDE A SHORT JUSTIFICATION OR EXAMPLE** to explain your choice.

(a) For random variables X and Y, $\mathrm{Var}[6X + 4Y] = 36\,\mathrm{Var}[X] + 16\,\mathrm{Var}[Y]$.

ALWAYS SOMETIMES NEVER

(b) Let X, Y be discrete random variables and s, t be values. $\Pr[X = s \text{ or } Y = t] \leq \Pr[X = s] + \Pr[Y = t]$.

ALWAYS SOMETIMES NEVER

(c) A pairwise independent hash function is universal.

ALWAYS SOMETIMES NEVER

(d) Increasing the number of tables in a locality sensitive hashing scheme increases the false negative rate.

ALWAYS SOMETIMES NEVER

(e) When used to estimate the frequency $f(v)$ of an item v in a stream, the CountMin sketch provides an overestimate, $\tilde{f}(v) > f(v)$.

ALWAYS SOMETIMES NEVER

2. Short answer. (**10pts – 5pts each**)

(a) Consider $\mathbf{x}$ drawn uniformly from a d dimensional ball.

Let $S(d) = \Pr[\|\mathbf{x}\|_2 \geq .9]$. How does $S(d)$ change as d increases? **No justification needed.**

INCREASES   DECREASES   STAYS THE SAME

Let $T(d) = \Pr[|x_1| \geq .1]$, where x_1 denotes the first entry of $\mathbf{x}$. How does $T(d)$ change as d increases? **No justification needed.**

INCREASES   DECREASES   STAYS THE SAME

(b) State and prove Chebyshev's inequality. You may use Markov's inequality without proof.

3. A Random Walk Won't Go Very Far (14pts)

Consider a random walk on the number line. Let x_t denote the postion of the walk and assume we start at the origin, so $x_0 = 0$. At all subsequent time steps, the walk either moves up or down one step with equal probability, so for $t = 1, 2, \ldots$ we have:

$$x_t = \begin{cases} x_{t-1} + 1 & \text{with probability } 1/2 \\ x_{t-1} - 1 & \text{with probability } 1/2. \end{cases}$$

(a) (8pts) Suppose we walks for n steps. Prove that, with probability $9/10$, the walk terminates no further than $O(\sqrt{n})$ away from the origin. I.e., prove that with probability $9/10$, $|x_n| \leq c\sqrt{n}$ for a constant c.

(b) (6pts) Prove that, in fact, the walk *never* strays too far away from the origin over the course of the walk. Specifically, prove that with probability $9/10$, $\max_{i \in 1,\ldots,n} |x_i| \leq c\sqrt{n \log n}$.

4. Collision Free Hashing Revisited (8pts + 4pts bonus)

(a) (8pts) Prove that if we insert m unique keys into a hash table of size $O(m^{1.5})$ using a uniformly random hash function, then there **will be a collision** with probability $\geq 9/10$ for any m above a fixed constant. Note that this is the opposite of what we prove on Homework 1, which was that there would not be a collision if the table size was set a bit larger (to $O(m^2)$.)

(b) **Optional Bonus:** Consider an alternative data structure for storing items: build two tables, each of size $O(m^{1.5})$ and choose a separate random hash function (independently at random) for each table. To insert an item, hash it to a bucket in each table and place it in the emptier bucket (you can break ties arbitrarily).

Prove that, if we insert m items into the data structure described above, then with probability $\geq 9/10$, every item will end up in a bucket by itself, i.e., there will be no collisions.