New York University Tandon School of Engineering
Computer Science and Engineering

# CS-GY 6763: Midterm Exam.

Friday Oct. 20th, 2023.

**47 points total**

You have 1 hour, 10 minutes to take the exam. 17.5 minutes per question.
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. Short answer. (12pts – 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) Consider a random variable $Y$ that takes positive integer values. For $z > 0$, $\Pr[Y \geq z] \leq \frac{\mathbb{E}[Y]}{z}$.

ALWAYS SOMETIMES NEVER

(b) For random variables $X$ and $Y$, $\mathrm{Var}[10X + 3Y] = 10^2\,\mathrm{Var}[X] + 3^2\,\mathrm{Var}[Y]$.

ALWAYS SOMETIMES NEVER

(c) Increasing the number of tables in a locality sensitive hashing scheme decreases the expected number of false positives.

ALWAYS SOMETIMES NEVER

(d) Let $X_1, \ldots, X_n$ be random variables (not necessarily independent). $\Pr\left[\max_j X_j \geq z\right] \leq \sum_{j=1}^n \Pr\left[X_j \geq z\right]$.

ALWAYS SOMETIMES NEVER

## 2. Collision Free Hashing Revisited (10pts)

You proved on Problem Set 1 that if we insert $m$ unique keys into a hash table of size $O(m^2)$ using a uniformly random hash function, then there will be no collisions with high probability.

In this problem we 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 $\geq 9/10$, there will be no collisions. I.e., after inserting all $m$ items, every item inserted will be by itself in a bucket in one of the tables.

## 3. LSH for Hamming Similarity (10pts)

For two length $d$ binary vectors $\mathbf{q}, \mathbf{y} \in \{0,1\}^d$, consider the hamming similarity:

$$s(\mathbf{q}, \mathbf{y}) = 1 - \frac{\|\mathbf{q} - \mathbf{y}\|_0}{d}.$$

Above $\|\mathbf{q} - \mathbf{y}\|_0$ is the hamming distance, $\|\mathbf{q} - \mathbf{y}\|_0 = \sum_{i=1}^d |q_i - y_i|$, where $q_i$ and $y_i$ denote the $i^{\text{th}}$ entries of $\mathbf{q}$ and $\mathbf{y}$, respectively. For example, the hamming similarity between the following two vectors is $1 - \frac{2}{6} = \frac{2}{3}$.

$$\mathbf{q} = \left[1, 0, 0, 1, 1, 0\right]$$
$$\mathbf{y} = \left[1, 0, 1, 1, 0, 0\right]$$

(a) (5pts) Construct a function $h$ as follows: define the random function $c : \{0,1\}^d \to \{0,1\}$ as $c(\mathbf{x}) = \mathbf{x}[j]$, where $j$ is a uniform random integer in $\{1, \ldots, d\}$. Then, let $g$ be a uniform random hash function from $\{0,1\} \to \{1, \ldots, m\}$. Finally, let:

$$h(\mathbf{x}) = g(c(\mathbf{x})).$$

Prove that $h$ is a locality sensitive hash function for hamming similarity.

(b) (5pts) Describe (in equations or pseudocode) a version of $h$ with $r > 1$ bands, and write down an expression for your new LSH function's collision probability as a function of the hamming similarity and $r$.

## 4. A Random Walk Won't Go Very Far (15pts)

Suppose a tourist is randomly walking up and down 5th Avenue, which can be modeled as a number line. At time $t$, their position is denoted by $x_t$. At time $t = 0$ they start at the origin, so $x_0 = 0$. At all subsequent time steps, the tourist 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) (7pts) Suppose the tourist walks for a total of $n$ steps. Show that, with probability $9/10$, they end up no more than $O(\sqrt{n})$ away from the origin. I.e., show that with probability $9/10$, $|x_n| \leq c\sqrt{n}$ for some constant $c$. **Hint:** Write $x_n$ as a sum of random variables.

(b) (8pts) Prove that, in fact, the tourist *never* strays too far away from the origin over the course of the walk. Specifically, show that with probability $9/10$, $\max_{i \in 1,\ldots,n} |x_i| \leq c\sqrt{n \log n}$.