New York University Tandon School of Engineering
Computer Science and Engineering

CS-GY 6763: Midterm Exam.
Monday Mar. 20th, 2022.
55 points total

Time limit: 1 hour and 15 minutes.

Directions

- Show your work to receive full (and partial) credit.
- Don't get stuck – if you can't see how to solve a problem skip it and move on.
- Feel free to ask me any clarifying questions.

1. Always, sometimes, never. (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) Let X be a $\{0,1\}$ indicator random variable for a random event. $\text{Var}[X] > \mathbb{E}[X]$.
ALWAYS SOMETIMES NEVER

(b) For random variables X, Y , we have $\mathbb{E}[XY] \geq \mathbb{E}[X]\mathbb{E}[Y]$.
ALWAYS SOMETIMES NEVER

(c) If random variables X, Y have the same variance, the average $Z = \frac{1}{2}(X+Y)$ has $\text{Var}[Z] \leq \frac{1}{2}\text{Var}[X]$.
ALWAYS SOMETIMES NEVER

(d) For a positive random variable Y, $\Pr[Y \geq z] \leq \frac{\mathbb{E}[Y]}{z}$.
ALWAYS SOMETIMES NEVER

2. Sum of Convex Functions is Convex (5 pts)

Let $f(x)$ and $g(x)$ be two convex functions. Show that $f(x) + g(x)$ is convex.

3. Safety First (5pts)

An airplane has 1000 critical parts, including engine components, navigation equipment, etc. Each part has been thoroughly tested, and during a given flight, each part is guaranteed not to fail with probability 9999/10000. What is the probability that no part fails during a given flight? Give the highest bound you can based on the problem information and explain how you obtained in.

4. Concentration Walks (13pts)

Consider an ant A walking along the real number line $\mathbb{R}$. At time $t = 0$, it is placed at the origin 0. At each time step, the ant randomly moves either one integer to the left or right, with equal probability. In other words, if $X^{(t)} \in \mathbb{Z}$ is the position of the ant on time step t, then

$$X^{(t+1)} = \begin{cases} X^{(t)} + 1 & \text{with probability } 1/2 \\ X^{(t)} - 1 & \text{with probability } 1/2 \end{cases}$$

In the following, let $C = (1000 \log n)\sqrt{n}$.

1. (5pts) Suppose the ant randomly walks for a total of n time steps. Give the best upper bound you can on the probability $\Pr[|X^{(n)}| > C]$.

2. (8pts) Show that the ant *never* goes farther than C steps from the origin with probability at least 2/3. In other words, prove

$$\Pr\left[\max_{i=1,2,\ldots,n} |X^{(i)}| > C\right] < \frac{1}{3}$$

5. Popflix learns to LSH (20pts)

The giant internet company *Popflix* operates a movie streaming service, and has a total of N subscribers worldwide. Popflix offers a total of d movies, and keeps track of the movies viewed by each subscriber. They store this data via a set of vectors $x^1, x^2, \ldots, x^N \in \{0,1\}^d$, where x^i is a binary vector corresponding to the i-th viewer, where for each film $j \in [d]$:

$$x^i_j = \begin{cases} 1 & \text{if person } i \text{ has seen film } j \\ 0 & \text{otherwise} \end{cases}$$

For a specific user i, Popflix would like to find other users j whose viewing habits are similar to user i. They model viewing (dis)similarity by the *Hamming Distance* $\|x^i - x^j\|_0$ between the users' corresponding vectors. Recall that the *Hamming distance* between two length d binary vectors $x, y \in \{0,1\}^d$ is given by $\|x - y\|_0 = \sum_i |x_i - y_i|$, i.e. the number of bits in which x, y differ.

Popflix recently hired an expert in sketching algorithms, who wants to use Locally Sensitive Hashing (LSH) to find similar users efficiently. To do this, they design a LSH for the Hamming Distance as follows. Let i be a uniform random integer in $\{1, \ldots, d\}$, and define the hash function $h : \{0,1\}^d \to \{0,1\}$ as $h(x) = x_i$, where x_i is the i^{th} entry in the vector x.

(a) (4pts) Given $x, y \in \{0,1\}^d$, compute $\Pr[h(x) = h(y)]$ as a function of $\|x - y\|_0$ and d, where the probability is taken over the random draw of $i \in [d]$.

(b) (8 pts) A new user just subscribed to Popflix, and so far has not watched any films. Since the start of the service, Popflix has had a total of M distinct movie views across all its N lifetime viewers. Let $q = (0, 0, \ldots, 0) \in \{0,1\}^d$ be the vector corresponding to the brand new user.

The sketching expert first decides to hash all of the subscribers vectors $X = \{x^1, \ldots, x^N\}$ into a single hash table $T = \{0,1\}$. Compute the total expected number of vectors $x_i \in X$ such that $h(x_i) = h(q)$, in terms of M, N, d.

(c) (8pts) The sketching expert now wants to build a series of locally sensitive hash tables $T_1, \ldots, T_t$ with t "tables" and r "bands" (i.e., r is the number of independent hash functions h used for each table) as in Lecture 5, using the LSH for Hamming distance they designed above. They want to hash all the subscriber vectors $X = \{x^1, \ldots, x^N\}$ into each of the tables using the LSH scheme, such that given a query point $q \in \{0,1\}^d$, they can find points x^i that are close to q. However, Popflix has strict requirements on the probability of false negatives and false positives that the LSH scheme must satisfy.

- **False Negative Rate**: For any one specific close point $x \in X$ such that $\|q - x\|_1 \leq \frac{d}{2}$, the probability we do not find x (meaning x never collides with q in any hash table) is at most $\frac{1}{10}$.
- **False Positive Rate:** For any one specific far point $y \in X$ such that $\|q - y\|_1 \geq \frac{9d}{10}$, the probability that we have to search through y in the LSH scheme (meaning that y collides with x in at least one of the tables) is at most $\frac{1}{100}$.

State integer values t, r for the number of tables and bands such that the false positive and false negative rates that Popflix desires are satisfied, and prove that this is the case. You may want to recall the useful inequalities, that hold for any $x \in (0, 1]$ and $n \geq 1$:

$$(1-x)^{n/x} \leq \left(\frac{1}{2}\right)^n \qquad \text{and} \qquad (1-x)^n \geq 1 - xn$$

The Chernoff-Hoeffding Bound

You may use the following form of the Chernoff-Hoeffding bound in your solutions. Note that you are welcome to use either this bound, or the very similar Chernoff bound which was stated in class.

Theorem 1. *Let $x_1, x_2, \ldots, x_n$ be independent random variables such that $a \leq x_i \leq b$ for all $i = 1, 2, \ldots, n$. Let $\mu = \mathbb{E}\left[\sum_{i=1}^{n} x_i\right]$. Then for any $t > 0$ we have*

$$\Pr\left[\left|\sum_i X_i - \mu\right| > t\right] \leq 2e^{-\frac{2t^2}{n(b-a)^2}}$$