COMPSCI 690: Midterm

Instructions:

- You have 2.5 hours to complete this exam – from 11:15am-1:45pm.
- If you have a question, **raise you hand and I'll come over to you.**
- You must **show your work/derive any answers** as part of the solutions to receive full credit (and partial credit if you make a mistake).
- If you need extra space to show your work you can include additional pages. Clearly mark the top of the any additional page with your **name and problem number**. On the exam **indicate that the work is finished on an extra page.**
- If you need to use the restroom, **leave you cellphone on your desk**.

1. Always, Sometimes, Never (8 points)

Indicate whether each statement is always true, sometimes true, or never true. **Give a short sentence/phrase/example explaining why.**

1. (2 points) Let $\mathbf{X}_1, \mathbf{X}_2$ both have variance σ^2. The average $\mathbf{X} = \frac{1}{2}(\mathbf{X}_1 + \mathbf{X}_2)$ has $\text{Var}[\mathbf{X}] = \frac{\sigma^2}{2}$.

   ALWAYS SOMETIMES NEVER

2. (2 points) Given random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$, $\mathbb{E}\left[\max_{i \in \{1,\ldots,n\}} \mathbf{X}_i\right] = \max_{i \in \{1,\ldots,n\}} \mathbb{E}[\mathbf{X}_i]$.

   ALWAYS SOMETIMES NEVER

3. (2 points) Assume there is a Las Vegas algorithm for solving a decision problem with expected running time T. There is a Monte-Carlo algorithm with worst-case running time $O(T)$ and $\leq 1/10$ probability of failure.

   ALWAYS SOMETIMES NEVER

4. (2 points) For $A, B \in \mathbb{R}^{n\times n}$ whose product AB has a non-zero entry *in every row*, the failure probability of Freivald's algorithm in testing if $AB = 0$ is lower than $C, D \in \mathbb{R}^{n\times n}$, whose product CD has just a single non-zero entry in one row.

   ALWAYS   SOMETIMES   NEVER

2. Short Answers (10 points)

1. (2 points) You draw coupons *with replacement* from a bag containing n unique coupons. For any $j \in [n]$, let $\mathbf{X}_j$ be the draw at which you see the j^{th} unique coupon. For any $j \in [n-1]$, give a formula for $\mathbb{E}[\mathbf{X}_{j+1} - \mathbf{X}_j]$ in terms of n and j.

2. (2 points) Consider $A \in \mathbb{R}^{n\times n}$. Let $\mathbf{x} \in \{-1, 1\}^n$ have each entry set independently to -1 with probability 1/2 and 1 with probability 1/2. Prove that $\mathbb{E}[\|A\mathbf{x}\|_2^2] = \|A\|_F^2$.

3. (2 points) Consider placing n balls into n bins, independently and uniformly at random. Let $\mathbf{X}$ be the fraction of bins that have *at most* 1 *ball* in them. Give an exact formula for $\mathbb{E}[\mathbf{X}]$ in terms of n. What is the limiting value as $n \to \infty$? **Note:** Your final value may use the constant e in it.

4. (2 points) Suppose you have a hash table with n buckets into which you have inserted $.95 \cdot n$ items. Which do you expect to perform better in this setting: linear probing or chaining? Explain in a sentence or two. You do not necessarily need to do any computations.

5. (2 points) You are standing on a number line, starting at position 0. You take n independent random steps. In each, you move right $+2$ with probability $1/2$ or left -2 with probability $1/2$. Let $\mathbf{X}$ be your final position. Use Chebyshev's inequality to give an upper bound on $\Pr[|\mathbf{X}| \geq 4\sqrt{n}]$.

3. Randomized Identifiers (10 points)

Consider a Peer-2-Peer file-sharing network. There are m files $f_1, \ldots, f_m$, each represented by a string of n bits. On any given day, a file f_i is requested for download with probability p_i. We have $\sum_{i=1}^{m} p_i = z$ and assume that all download requests are independent of each other.

1. (2 points) Which concentration inequalities could you use to bound the probability that the number of requested files on any given day exceeds $2z$? Which inequality do you expect would give the tightest bound? **Note:** You do not need to do any computation here. Just write a few sentences.

2. (2 points) Each file that is requested on a given day must be assigned a unique ID. Assume there is no centralized mechanism to pick the unique IDs, so a file f_i is hashed to $h(f_i) = F_i \bmod p$, where F_i is the value of file interpreted as a binary number, and p is a randomly chosen prime in $\{1, \ldots, O(nt \log(nt))\}$. p is selected once by the system at the beginning of operation, and then fixed.

   What is the probability that files f_i and f_j are both requested on a given day *and* assigned the same ID? Give an upper bound in terms of t, p_i, and p_j. You may assume that the choice of random prime p is independent of the random download requests.

3. (2 points) How large must we set t such that, with probability at least 99/100, all requested file IDs on a given day are unique? **Hint:** Start by giving an upper bound on the expected number of pairwise ID collisions.

4. (2 points) With the above setting of t, upper bound the probability that in a span of 10 days, there is at least one instance where two requested files on a given day have the same ID.

5. (2 points) Could you apply a Chernoff bound to give a tighter bound in part (3)? Why or why not? What if each $h(f_i)$ is selected independently and uniformly from $\{1, \ldots, t\}$, rather than via a Rabin fingerprint?

4. Approximate Geometric Median (4 points)

Consider a set of n points $x_1, \ldots, x_n \in \mathbb{R}^d$. For any $y \in \mathbb{R}^d$, let $\phi(y) = \sum_{i=1}^n \|x_i - y\|_2$. The *geometric median* is defined as $x^\star = \underset{x_i \in \{x_1, \ldots, x_n\}}{\arg\min} \phi(x_i)$. I.e., it is the point in the set whose total Euclidean distance to all other points is minimal.

Let $\mathbf{z}$ be picked uniformly at random from $\{x_1, \ldots, x_n\}$. Show that

$$\mathbb{E}[\phi(\mathbf{z})] \leq 2\phi(x^\star).$$

I.e., a randomly selected point from the set is within a two factor of minimizing the total distance to all other points in expectation. **Hint:** Use triangle inequality.

5. Randomized Low-Rank Approximation (4 points)

Consider a matrix $A \in \mathbb{R}^{n \times d}$, whose columns are all copies of the standard basis vectors $e_1, \ldots, e_n \in \mathbb{R}^n$. Assume that e_1 appears m times, and this is the most frequently appearing column. Let $\mathbf{C} \in \mathbb{R}^{n \times t}$ consist of t columns of A, sampled independently and uniformly at random. Let $\mathbf{Z} \in \mathbb{R}^{n \times t}$ be an orthonormal basis for $\mathbf{C}$. Give an upper bound (in terms of n, m and t) on

$$\Pr\left[\|A - \mathbf{Z}\mathbf{Z}^T A\|_F^2 > \min_{B:\text{rank}(B)=1} \|A - B\|_F^2\right].$$

6. Communication Complexity of Majority (BONUS: 6 points)

Alice and Bob are given n-bit strings $a, b \in \{0,1\}^n$. They would like to compute the majority bit amongst the strings – i.e., out of the $2n$ total bits, they would like to output the bit 0 or 1 that appears more times. They may output anything if both 0 and 1 appear exactly n times each.

1. (2 points) Describe a deterministic protocol that solves this problem using $O(\log n)$ bits of communication.

2. (2 points) Argue, via an indistinguishability argument, that any deterministic protocol requires $\Omega(\log n)$ bits of communication to solve the problem.

3. (2 points) Consider a variant on this problem: Alice and Bob want to compute the majority bit of $a \wedge b$, the entrywise AND of the two input vectors. Argue that any deterministic protocol for solving this problem requires $\Omega(n)$ bits of communication.