COMPSCI 514: Midterm

Date: 10/17/2019, 10:00am-11:15am

Instructions:

- Please 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. Please mark clearly with your name and problem number.
- If you have a question, raise your hand and we will come to you.

Probability, Expectation, and Variance (10 points)

1. (3 points) For two random variables $\mathbf{X}$ and $\mathbf{Y}$, indicate whether each statement is **always true**, **sometimes true**, or **never true**. Give a short sentence/phrase explaining why.

   (a) $\mathbb{E}[\mathbf{X} - \mathbf{Y}] = \mathbb{E}[\mathbf{X}] - \mathbb{E}[\mathbf{Y}]$. ALWAYS SOMETIMES NEVER

   (b) $Var[\mathbf{X} + \mathbf{Y}] = Var[\mathbf{X}] + Var[\mathbf{Y}]$. ALWAYS SOMETIMES NEVER

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

2. (3 points) Prove that for any random variable $\mathbf{X}$, $Var[\mathbf{X}] = \mathbb{E}[\mathbf{X}^2] - \mathbb{E}[\mathbf{X}]^2$.

3. (2 points) Given a random variable $\mathbf{X}$, can we conclude that $\mathbb{E}[1/\mathbf{X}] = 1/\mathbb{E}[\mathbf{X}]$? If so, prove this. If not, give an example where the equality does not hold.

4. (2 points) An airplane has 1000 critical parts – including engine components, navigation equipment, etc. Each part has been throughly tested, and during a given flight, each part is guaranteed to fail with probability at most .0001. Give as tight an upper bound as you can on the probability that *at least one part fails* during a flight.

Hash Functions (10 points)

Consider a pairwise independent hash function $\mathbf{h} : [n] \to [m]$.

1. (2 points) For any $x, y \in [n]$ with $x \neq y$ and $z \in [m]$, what is $\Pr(\mathbf{h}(x) = \mathbf{h}(y) = z)$.

2. (2 points) Is $\mathbf{h}$ a 2-universal hash function? Circle one and give a sentence explaining your answer: YES NO MAYBE.

3. (2 points) Is $\mathbf{h}$ a locality sensitive hash function? Circle one and give a sentence explaining your answer: YES NO MAYBE.

4. (2 points) Why are pairwise independent hash functions used instead of fully independent hash functions in practice?

5. (2 points) You use a pairwise independent hash function to insert items into hash table. You would like to show that the maximum number of items in any bucket is small with good probability. Circle any of the following bounds that you can apply to this problem. Explain in a couple of sentences.

   Markov's Chebyshev's Chernoff.

Finding Duplicates in Limited Space (10 points)

You have a list of n usernames that are registered on your website. When a new user signs up, you want to check whether the username they pick is already assigned.

1. (3 points) What data structure might you use to rapidly check if a username has been previously assigned? You must be sure to never let a user pick an already assigned name, but are willing to allow a small false positive rate: with small probability, a user may not be allowed to pick a name even though it is not actually assigned. (No explanation of answer needed.)

2. (3 points) What data structure/technique might you use if you also want to prevent a new user from choosing a username that is *close to any registered username*, and therefore a potential point of confusion? (No explanation of answer needed.)

3. (4 points) Generally, how do the space complexities of these two methods compare to if you just stored the registered user names in a hash table with $O(n)$ buckets? Are they higher or lower? Why?

Estimating Similarity with Min Hash (10 points)

In class we used MinHash as a locality sensitive hashing function for Jaccard similarity. We will now see how it can be used directly as an estimator of Jaccard similarity. Consider two sets A and B with Jaccard similarity $J(A, B) = s$. Let $\mathbf{MH}_1(\cdot), \mathbf{MH}_2(\cdot), \ldots, \mathbf{MH}_t(\cdot)$ be t independent instantiations of MinHash. Let $\mathbf{X}$ be the number of times that the MinHash values for A and B collide in these t instantiations. That is, $\mathbf{X} = |\{i \in [t] \text{ such that } \mathbf{MH}_i(A) = \mathbf{MH}_i(B)\}|$. Let $\tilde{\mathbf{s}} = \mathbf{X}/t$ be an estimate of s derived from $\mathbf{X}$.

1. (2 points) Show that $\mathbb{E}[\tilde{\mathbf{s}}] = s$.

2. (3 points) Show that $Var[\tilde{\mathbf{s}}] \leq \frac{1}{t}$. **Hint**: Break $\mathbf{X}$ into indicator random variables, and show that each has variance ≤ 1.

3. (3 points) Use the results from (1) and (2) to show that if we set $t \geq \frac{1}{\delta\epsilon^2}$, then

$$\Pr[|\tilde{\mathbf{s}} - \mathbf{s}| \geq \epsilon] \leq \delta.$$

That is, with good probability $\tilde{\mathbf{s}}$ is within ϵ of the true Jaccard similarity.

4. (2 points) In part (3), to give failure probability δ, t depends on $1/\delta$. What technique could you use to improve this dependence, so that the failure probability is lower?

Randomized Dimensionality Reduction (EXTRA CREDIT: 8 points)

In class we studied randomized dimensionality reduction (the Johnson-Lindenstrauss lemma) using a matrix whose entries are random Gaussian variables. In this problem we will see how one might analyze a random projection whose entries are just random signs.

Consider a vector $\vec{x} \in \mathbb{R}^d$ and let $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ be a random matrix, with each entry set independently to $1/\sqrt{m}$ with probability $1/2$ and $-1/\sqrt{m}$ with probability $1/2$.

1. (2 points) Letting $\mathbf{\Pi}(j)$ denote the j^{th} row of $\mathbf{\Pi}$, what is $\mathbb{E}[\langle \vec{x}, \mathbf{\Pi}(j) \rangle]$?

2. (2 points) What is $\mathbb{E}[\langle \vec{x}, \mathbf{\Pi}(j) \rangle^2]$?

3. (2 points) What is $\mathbb{E}[\|\mathbf{\Pi}\vec{x}\|_2^2]$? Does this value make sense if we want to use $\mathbf{\Pi}$ to give a low-distortion embedding?

4. (2 points) In practice, why might one prefer a random projection matrix whose entries are random signs rather than random Gaussian variables?