COMPSCI 614: Midterm Review

General Info: The midterm will be held **in class on March 28th** from 10:00am-11:15pm. The test will be **closed book**, with no cheatsheets or calculators allowed. 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).

Format: The test will contain 4-5 questions. The first will be a mix of True/False or Always/Sometimes/Never style questions (see examples later in this document). The rest will be short answer style questions, like homework questions, but significantly less involved.

Studying Tips:

- Do as many practice problems as you can – from this review sheet, the books, the quizzes, the homeworks, and the 'Exercise' or 'Think-Pair-Share' questions given on the slides. For quizzes/homeworks/in class questions – try to re-solve without looking at the answer key or a solution given in the next slide. Then check to see how you did.
- For all practice questions, try to solve (and write down) a solution first without resources and somewhat quickly, as you would on the exam. Then go back and more slowly work through the problem, see if you solution is correct, etc.
- We encourage you to post on Piazza to check answers/discuss approaches.

1 Concepts to Study

Foundational Probability + Concentration Bounds

- Basic probability, conditional probability, independence, k-wise independence.
- Linearity of expectation and variance.
- Markov's inequality, Chebyshev's inequality. Should know from memory and understand how they were derived.
- Union bound. Should know from memory.
- General idea of higher moment inequalities.
- Chernoff and Bernstein bounds. Don't need to memorize the formulas, but should be able to apply if given. Know when they can be applied (i.e., to sum of bounded independent random variables.)
- Basic complexity classes related to randomized algorithms ($P \subseteq ZPP \subseteq RP \subseteq BPP \subset PP$), and how to do reductions between them.
- Monte-Carlo vs. Las Vegas algorithms.

Basic Algorithmic Applications

- Polynomial identity testing and the Schwartz-Zippel lemma. Should know the statement and understand the idea behind the proof but don't need to fully memorize the proof.
- Coupon collector problem. Analysis in expectation.
- Balls-into-bins analysis and bounds. Don't need to memorize details of the analysis.
- Statistical estimation type error bounds for sampling using concentration inequalities.
- High level understanding of Quicksort analysis and bounds, but don't need to memorize.
- High level understanding of linear probing analysis and bounds, but don't need to memorize.
- Understanding of how linear probing relates to chaining and how both can be analyzed with concentration bounds.

Random Sketching and Communication Complexity

- Rabin fingerprint and it's analysis.
- Rabin-Karp pattern matching algorithm.
- Application to equality testing problem in communication complexity.
- Equality testing $\Omega(n)$ bit lower bound for deterministic algorithms.
- ℓ_0 sampling – high level idea of how the random non-zero index is recovered.
- Application of ℓ_0 sampling to low-communication graph connectivity, and low-space graph connectivity in the streaming model
- Count Sketch and its analysis. The median trick as a general approach to increasing the success probability of randomized algorithms.

Randomized Numerical Linear Algebra

- Approximate matrix multiplication. Don't need to memorize the analysis but should understand the tools used and the application of non-uniform sampling.
- The idea of importance sampling in general.
- Implicit trace estimation and Hutchinson's method. Should understand basic analysis and application to triangle counting.
- Sampling based low-rank approximation algorithm and analysis.

2 Practice Questions

Work in progress. Check back to see if more questions have been added.

1. Probability, Expectation, Variance, Concentration Bounds:

1. Upfal and Mitzenmacher, Chapers 1-4 exercises (there are a lot of great problems here, most of which are relevant to what we have covered)
2. Show that for any $\mathbf{X}$, $\mathbb{E}[\mathbf{X}^2] \geq \mathbb{E}[\mathbf{X}]^2$.
3. Show that for independent $\mathbf{X}$ and $\mathbf{Y}$, $\mathbb{E}[\mathbf{X} \cdot \mathbf{Y}] = \mathbb{E}[\mathbf{X}] \cdot \mathbb{E}[\mathbf{Y}]$.
4. Show that for independent $\mathbf{X}$ and $\mathbf{Y}$ with $\mathbb{E}[\mathbf{X}] = \mathbb{E}[\mathbf{Y}] = 0$, $Var[\mathbf{X} \cdot \mathbf{Y}] = Var[\mathbf{X}] \cdot Var[\mathbf{Y}]$. **Hint:** use part (3).
5. For the statements below, indicate if they are **always true**, **sometimes true**, or **never true**. Give a sentence explaining why. (Note, we will definitely have this style of question on the exam.)
   (a) $\Pr[\mathbf{X} = s \cap \mathbf{Y} = t] > \Pr[\mathbf{X} = s]$. ALWAYS SOMETIMES NEVER
   (b) $\Pr[\mathbf{X} = s \cup \mathbf{Y} = t] \leq \Pr[\mathbf{X} = s] + \Pr[\mathbf{Y} = t]$. ALWAYS SOMETIMES NEVER
   (c) $\Pr[\mathbf{X} = s \cap \mathbf{Y} = t] = \Pr[\mathbf{X} = s] \cdot \Pr[\mathbf{Y} = t]$. ALWAYS SOMETIMES NEVER
6. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be the number of visitors to a website on n consecutive days. These are independent and identically distributed random variables. For every i, we have $\mathbb{E}[\mathbf{X}_i] = 20,000$ and $Var[\mathbf{X}_i] = 100,000,000$.
   (a) Give an upper bound on the probability that on day i, more than $40,000$ visitors hit the website.
   (b) Let $\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$ be the average number of visitors over n days. What are $\mathbb{E}[\bar{\mathbf{X}}]$ and $Var[\bar{\mathbf{X}}]$?
   (c) Give an upper bound on the probability that $\bar{\mathbf{X}} \geq 25,000$, for $n = 100$.
7. Assume there are 1000 registered users on your site $u_1, \ldots, u_{1000}$, and in a given day, each user visits the site with some probability p_i. The event that any user visits the site is independent of what the other users do. Assume that $\sum_{i=1}^{1000} p_i = 500$.
   (a) Let $\mathbf{X}$ be the number of users that visit the site on the given day. What is $\mathbb{E}[\mathbf{X}]$.
   (b) Apply a Chernoff bound to show that $\Pr[\mathbf{X} \geq 600] \leq .01$.
   (c) Apply Markov's inequality and Chebyshev's inequality to bound the same probability. How do they compare?
8. Give an example of a random variable and a deviation t where Markov's inequality gives a tighter bound than Chebyshev's inequality.

2. Basic Algorithmic Applications:

1. Upfal and Mitzenmacher exercises 5.3 (ignoring the part about getting tight constants), 5.7, 5.11.
2. Motwani Raghavan, exercises 3.1, 3.3, 3.12, 4.8.

3. You sample n items with replacement from a set of n items. What is the expected number of unique items that you see?

4. You store n items in a hash table with $2n$ buckets using a fully random hash function. Give an upper bound on the maximum load in any bucket, which holds with probability at least $1 - 1/n$. Does the answer change significantly if you have n instead of $2n$ buckets?

5. Consider the scenario above, where you use linear probing instead of chaining. Does the expected look up time change significantly if you use n rather than $2n$ buckets?

3. Random Sketching and Communication Complexity:

1. How many bits must a Rabin fingerprint use to achieve $Pr(\mathbf{h}(x) = \mathbf{h}(y)) \leq 1/n^{10}$ for $x \neq y$. Give an answer in Big-Oh notation.

2. You have a database containing m documents, each represented as an n bit string. How large should you set t in a Rabin fingerprint so that with probability at least 99/100 no two documents in the database have the same fingerprint.

3. Count-Sketch estimates each entry of a vector $x(i)$ to error $\epsilon\|x\|_2$. For how many entries at most does it give a non-trivial approximation? I.e, where the error $\epsilon\|x\|_2$ is smaller in magnitude than the entry $x(i)$. Does this make sense in light of the space complexity used?

4. If we want to simultaneously estimate all entries of a vector x to error $\epsilon\|x\|_2$ with probability $1 - \delta$, how many repetitions of Count Sketch do we need? **Hint:** Apply a union bound.

4. Randomized Numerical Linear Algebra:

1. Consider implementing approximate matrix multiplication where you sample $\mathbf{i}_1, \ldots \mathbf{i}_t$ uniformly at random with replacement and approximate AB by $\overline{\mathbf{C}} = \frac{n}{t}\sum_{j=1}^{t} A_{:,\mathbf{i}_j} B_{\mathbf{i}_j,:}$. Assume that all entries in A, B are bounded by 1 in magnitude. Show that for $t = \frac{1}{\epsilon^2\delta}$, with probability $\geq 1 - \delta$,
$$\|AB - \overline{\mathbf{C}}\|_F \leq \epsilon n^2.$$

2. Assume that $A \in \mathbb{R}^{n\times n}$ has entries with magnitude bounded by 1. Consider sampling $\overline{\mathbf{C}} \in \mathbb{R}^{n\times t}$ where for each $j \in [t]$, $\overline{\mathbf{C}}_{:,j}$ is set to $\sqrt{\frac{n}{t}} \cdot A_{:,i}$ with probability $1/n$ for any i. Show that if we take $t = \frac{k}{\epsilon^2\delta}$ and let $\mathbf{V} \in \mathbb{R}^{n\times k}$ consist of the top k left singular vectors of $\overline{\mathbf{C}}$, then with probability $\geq 1 - \delta$,
$$\|A - \mathbf{V}\mathbf{V}^T A\|_F^2 \leq \min_{Z: Z^T Z = I} \|A - ZZ^T A\|_F^2 + \epsilon n^2.$$

3. Consider applying Hutchinson's estimator to a diagonal matrix D. Let $\tilde{T} = \frac{1}{m}\sum_{i=1}^{m} x_i^T D x_i$ be the Hutchinson's estimator of $\mathrm{tr}(T)$ using m random vectors. What is $\mathrm{Var}(\tilde{T})$? What if we implement Hutchinson's estimator using vectors with iid standard normal random variables instead of ± 1s?

4. In Freivald's algorithm (i.e., checking if $AB = C$ by checking if $ABx = Cx$ for a random vector x), one can use either random vectors drawn from $\{0, 1\}^n$ or from $\{-1, 1\}^n$. Is the same true for Hutchinson's method? Why or why not?

5. Prove that for $A, B \in \mathbb{R}^{n \times d}$ where $\text{rank}(B) = k$, that if $V \in \mathbb{R}^{n \times k}$ is an orthonormal basis for B's columns, $\|A - VV^T A\|_F^2 \leq \|A - B\|_F^2$. **Hint:** Prove that $\|A - B\|_F^2 = \|A - VV^T A\|_F^2 + \|B - VV^T A\|_F^2$.

6. Verify the cyclic property of trace – that $\text{tr}(AB) = \text{tr}(BA)$.

7. Give an example of a symmetric matrix $B \in \mathbb{R}^{n \times n}$ and an orthonormal matrix $V \in \mathbb{R}^{n \times k}$ such that $\text{tr}(V^T BV) = \sqrt{k} \cdot \|B\|_F$.