Data Science Through a Geometric Lens Fall 2019

Homework 1

Due: Monday 10/21/19, 5pm

- **Solving 3 of the following 5 problems** will lead to full credit. You may attempt more than 3 problems, but the grading will be based on the 3 problems with the highest scores.
- Email the solutions to both the instructor and TA (emails listed on the course website).
- You may work in groups of size 1-3. If you do, please hand-in a single assignment with everyone's names on it. It is strongly encouraged to type up the solutions in Latex.
- If the question asks to prove something, you must write out a formal mathematical proof.
- If the question involves analyzing an algorithm, you must formally explain the time and/or space usage, along with the approximation guarantees (when applicable).
- When you are asked to prove a bound, it suffices to prove it up to multiplicative constants, i.e., using $O(\cdot)$, $\Theta(\cdot)$, or $\Omega(\cdot)$ notation. No need to optimize (multiplicative) constants!
- You may use other resources, but you must cite them. If you use any external sources, you still must provide a complete and self-contained proof/result for the homework solution.

1 Problem 1: Probability Fundamentals

(a) Consider n bins, where several balls are thrown. We throw each ball independently in a uniformly random bin. Let X be a random variable equal to the number of balls we need to throw until every bin contains at least one ball. Show that

$$\mathbb{E}[X] = n \cdot \sum_{i=1}^{n} \frac{1}{i}.$$

Use that $\sum_{i=1}^{n} \frac{1}{i} = \Theta(\log n)$ to conclude that $X = O(n \log n)$ with probability at least 0.9.

(b) Show that Markov's inequality is tight. That is, exhibit a **non-negative** random variable Y and a value $\lambda > 0$ such that

$$\Pr(Y \geq \lambda) = \frac{\mathbb{E}[Y]}{\lambda}.$$

Provide the random variable, the expectation, the value of λ, and the details for why the inequality is tight.

(c) Show that Chebyshev's inequality is tight. That is, exhibit a random variable Z and a value $\lambda > 0$ such that

$$\Pr(|Z - \mathbb{E}[Z]| \geq \lambda) = \frac{\mathrm{Var}[Z]}{\lambda^2}.$$

Provide the random variable, the expectation, the variance, the value of λ, and the details for why the inequality is tight.

2 Problem 2: Pairwise Independence

(a) Let q be a prime number. For integers $c, d \in \{0, 1, \ldots, q-1\}$, define the hash function $h_{c,d}$ as

$$h_{c,d}(x) = cx + d \mod q.$$

Let $\mathcal{H}$ be the set of all such hash functions, defined as

$$\mathcal{H} = \big\{ h_{c,d} \mid c, d \in \{0, 1, \ldots, q-1\} \big\}.$$

Prove that $\mathcal{H}$ is a pairwise independent hash family. That is, prove that for any distinct $i \neq i'$ and any j, j', we have that

$$\Pr_{h_{c,d} \in \mathcal{H}}[\ h_{c,d}(i) = j \text{ and } h_{c,d}(i') = j'\ ] = \frac{1}{q^2},$$

where $h_{c,d} \in \mathcal{H}$ is chosen uniformly by choosing c, d at random in $\{0, 1, \ldots, q-1\}$.

Hint: It may be good to start with $q = 2$ and $\{0, 1\}$ values; then, generalize to all prime $q \geq 2$.

(b) Let $Y_1, \ldots, Y_n$ be pairwise independent random variables (see Lecture 3 notes for a definition). Prove that $\mathrm{Var}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \mathrm{Var}[Y_i]$.

(c) **Extra Credit.** Let q be a prime, and let k be an integer with $q \geq k$. Consider the set $\mathcal{H}$ of degree $k-1$ polynomials over $\mathbb{F}_q$. More precisely, let $\mathcal{H}$ be the set of polynomials $h_{\vec{c}}$ defined by a vector $\vec{c}$ of k coefficients $c_0, c_1, \ldots, c_{k-1} \in \{0, 1, \ldots, q-1\}$ such that

$$h_{\vec{c}}(x) = c_{k-1}x^{k-1} + c_{k-2}x^{k-2} + c_1 x + c_0 \mod q.$$

Prove that $\mathcal{H}$ is a k-wise independent hash family. That is, prove that for all distinct $i_1, i_2, \ldots, i_k$ and all $j_1, j_2, \ldots, j_k$, we have

$$\Pr_{\vec{c}}[\ h_{\vec{c}}(i_1) = j_1 \text{ and } h_{\vec{c}}(i_2) = j_2 \text{ and } \cdots \text{ and } h_{\vec{c}}(i_k) = j_k\ ] = \frac{1}{q^k},$$

where the probability is over uniformly random $c_0, c_1, \ldots, c_{k-1} \in \{0, 1, \ldots, q-1\}$.

Hint: Consider the $k \times k$ Vandermonde matrix, which is invertible.

3 Problem 3: Heavy Hitters

Develop a **deterministic** algorithm to estimate counts of the k heaviest elements in a stream. The stream consists of integers (not necessarily distinct) arriving one-by-one from the range $\{1, 2, \ldots, n\}$. The algorithm may compute something and update the storage, but then the stream element may not be accessed again (unless explicitly stored). The space is the maximum amount used throughout.

After seeing n elements, let f_j denote the number of times that j appeared in the stream. For each integer $j \in [n]$, the algorithm should output estimates $\tilde{f}_j$ such that $f_j - \frac{n}{k} \leq \tilde{f}_j \leq f_j$.

(a) Design a deterministic algorithm using $O(k \log n)$ space and provide the pseudo-code of the algorithm. *Hint:* Generalize the majority algorithm from class (Lecture 4).

(b) Provide theoretical guarantees for the algorithm and prove that it works as desired. That is, prove that the estimates satisfy $f_j - \frac{n}{k} \leq \tilde{f}_j \leq f_j$.

(c) Demonstrate an example showing that your analysis regarding the estimates is tight. That is, come up with a worst-case input stream, and explain the approximation of the algorithm.

4 Problem 4: Streaming Sampling

Let $a_1, a_2, \ldots, a_n$ be a stream of n integers (not necessarily distinct) in the range $\{1, 2, \ldots, n\}$. The algorithm knows n up front. Each a_i will arrive one-by-one. The algorithm may compute something and update the storage, but then the value may not be accessed again (unless it is explicitly stored). The space is the maximum amount of memory used throughout. For each of these, you must prove that the algorithm works correctly, and provide a bound on the space.

(a) Provide an algorithm using $O(\log n)$ space that samples a uniformly random element a_i from the stream (that is, the probability that a_i is output is equal to $1/n$ for each $i \in [n]$).

(b) Assume that you know $A = \|\vec{a}\|_2^2 = \sum_{i=1}^{m} a_i^2$, the sum-of-squares of the values in the stream. Provide an algorithm using $O(\log n)$ space to sample an element a_i from the stream with probability exactly $p_i = \frac{a_i^2}{A}$.

(c) Now, assume that you **do not know** A ahead of time. Provide an algorithm using $O\left(\log^2 n\right)$ space that samples a_i from the stream with probability approximately $p_i = \frac{a_i^2}{A}$. More precisely, you should sample a_i with probability $\widetilde{p_i}$ satisfying $\frac{p_i}{4} \leq \widetilde{p_i} \leq 4p_i$ for all $i \in [n]$.

Hint: Use many different samples like the ones from (b) depending on the true value of A, and in parallel, compute A exactly so that you know which sample to use for the output.

(d) **Extra Credit:** Improve your algorithm from (c). Now, given ε in the range $0 < \varepsilon < 1$, your sampling probabilities should satisfy $(1-\varepsilon)p_i \leq \widetilde{p_i} \leq (1+\varepsilon)p_i$.

5 Problem 5: Implementing a Sketching Algorithm

Implement and test one of the algorithms from the class, that is, choose one of the following options: (i) Morris for approximate counting, (ii) FM for distinct elements, or (iii) AMS for ℓ_2 estimation. Implement the algorithm and the + and ++ variants for the one you choose.

(a) Demonstrate/compare the performance of the three variants of the algorithm (normal, +, ++). Set the input size(s) to be large enough to see some difference in their performance.

(b) Provide results (in a table or plot, clearly labeled) for at least 2 different parameter settings (and list the parameters). Briefly discuss the results and any interesting observations.

(c) Provide the results of 10 repetitions for each of the two parameter settings (in a table or plot, clearly labeled), to demonstrate the probability of failure (and list the parameters). Briefly discuss the results and any interesting observations.

(d) Discuss how theory relates to practice, with quantitative results to back up your claims. For example, if the theory is pessimistic, then show that the results in practice are better with the same/improved parameters.