Scalable Algorithms for Data Analysis

Assignment 2

Aug-Nov 2022

1 Instructions

1. Deadline: 5pm on 11-Oct-2022.
2. Maximum mark : 15
3. Solve it by yourself and be honest.
4. For each question write answers on separate sheets.
5. Submit the answers through gradescope (see the course description page in piazza).

2 Definitions: Pairwise Independence

Definition 1. Let $X_1, \ldots, X_n$ be n random variables. We say that $X_1, \ldots, X_n$ are pairwise independent if the following holds. For any pair of distinct indices i and j, and values $a, b \in \mathbb{R}$,

$$\Pr[X_i = a \text{ and } X_j = b] = \Pr[X_i = a] \cdot \Pr[X_j = b]$$

Definition 2. A family of hash functions $\mathcal{H} \subseteq \{f \colon X \to Y\}$, is a *pairwise independent hash family* if the following two conditions hold.

- Uniformly distributed: for a any $x \in X$ and $y \in Y$,

$$\Pr_{h \sim \mathcal{H}}[h(x) = y] = \frac{1}{|Y|}.$$

- For any $x, x' \in X$ and $y, y' \in Y$ s.t $x \neq x'$,

$$\Pr_{h \sim \mathcal{H}}[h(x) = y \wedge h(x') = y'] = \frac{1}{|Y|^2}.$$

Here, $h \sim \mathcal{H}$ means that h is chosen uniformly at random from $\mathcal{H}$.

Definition 3. For a matrix $A \in \{0,1\}^{k \times n}$ and vector $b \in \{0,1\}^k$, define functions $h_A, h_{A,b} \colon X \to Y$ as follows:

$$\begin{aligned} h_{A,b}(x) &= (Ax + b) \mod 2 \\ h_A(x) &= Ax \mod 2 \end{aligned}$$

3 Questions

1. Consider the randomized algorithm for verifying matrix multiplication mentioned in the class. Here, the input is three $n \times n$ matrices A, B, C and we want to test whether $AB = C$? Recall the steps of the algorithm

   (a) Choose a random $n \times 1$ vector $r \in \{0,1\}^n$

   (b) Return **Yes** if $ABr = Cr$ and **No** otherwise.

   Prove that if $AB \neq C$, the the probability that the algorithm outputs **Yes** is at most 1/2. [3]

2. Let $X_1, \ldots, X_n$ be n pairwise independent random variables. Prove that $Var[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} Var[X_i]$. [2]

3. Consider the family of hash functions $\mathcal{H}_1 = \{h_{A,b} : A \in \{0,1\}^{k \times n}, b \in \{0,1\}^k\}$. Is $\mathcal{H}_1$ a pairwise independent hash family. Prove your answer. [2]

4. Consider the family of hash functions $\mathcal{H}_2 = \{h_A : A \in \{0,1\}^{k \times n}\}$. Is $\mathcal{H}_2$ a pairwise independent hash family. Prove your answer. [2]

5. Rank of a number a_i in a sequence of n distinct numbers $a_1, \ldots, a_n$ is the position of a_i in the increasing order of the numbers. Design a streaming algorithm that given n distinct numbers $a_1, \ldots, a_n$ in a streaming fashion, $0 < \epsilon, \delta < 1$, and outputs x such that $(1-\epsilon)n/2 \leq rank(x) \leq (1+\epsilon)n/2$ with probability at least $1 - \delta$. Write an algorithm, correctness proof, and analyse its space complexity. Your objective is to minimize the space complexity as much as possible. [6]