CSCE 689: Special Topics in Modern Algorithms for Data Science Fall 2023

Lecture 16 — October 4, 2023

Prof. Samson Zhou *Scribe: Ryan King*

1 Sparse Recovery

If we have an insertion-deletion stream of length $m = \Theta(n)$, where n is the size of the universe, with at most k non-zero coordinates, our goal is to recover the k non-zero elements along with their frequencies. To accomplish this, we must maintain $2k$ running sums of different linear combinations of all the coordinates (equivalent to solving $2k$ equations with $2k$ unknown variables). This algorithm results in $O(k)$ words of space.

Recall Chebyshev's Inequality. Let X be a random variable with expected value $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \text{Var}[X]$. Chebyshev's Inequality states:

$$\mathbf{Pr}\Big[|X - \mathbb{E}[[]X]| \geq t\Big] \leq \frac{Var[X]}{t^2} = \frac{\sigma^2}{t^2}$$

If we choose $t = k\sigma$ we get:

$$\mathbf{Pr}\Big[|X - \mathbb{E}[[]X]| \geq k\sigma\Big] \leq \frac{1}{k^2}$$

Intuitively, this provides a bound on the deviation of the random variable in terms of its variance.

2 Distinct Elements

Given a set S of m elements from $[n]$, let f_i be the frequency of element i, and let ε be an accuracy parameter. Define F_0 as the frequency moment of the vector:

$$F_0 = |i : f_i \neq 0|$$

The goal of the Distinct Elements problem is to output a $(1+\varepsilon)$-approximation of F_0.

2.1 Applications

While this problem can be solved exactly with enough memory, many applications require sublinear memory. Examples include:

1. **Ad allocation**: If we want to determine if an ad on a webpage is increasing traffic to our product, we cannot simply count the number of times the ad was clicked since a single

user could click the ad numerous times, resulting in misleading information about the ad's effectiveness. Instead, we want to know the number of distinct users or IP addresses that clicked on the ad. However, keeping track of all those IP addresses may not be feasible.

2. **Traffic Monitoring**: If we want to understand the number of users visiting a site or the number of unique search engine queries, we are not interested in the number of times a single user logs in or the number of times the same search was executed. However, storing the number of unique logins or searches may be costly.

3. **Computational Biology**: DNA contains short sequences called motifs. These motifs are recurring patterns that presumably have a biological function.

2.2 Estimation of F_0

Let S be a set of N numbers. Suppose we form a set S' by sampling each item of S with probability $1/2$. In expectation, the number of elements in S' is:

$$\mathbb{E}[|S'|] = \frac{N}{2}$$

and the variance is:

$$Var[|S'|] \leq \frac{N}{2}$$

We can use Chebyshev's inequality to get the following:

$$\mathbf{Pr}\Big[||S'| - \frac{N}{2}| \geq 100\sqrt{N}\Big] \leq \frac{1}{10}$$

This allows use to bound $|S'|$ with probability at least $9/10$:

$$\frac{N}{2} - 100\sqrt{N} \leq |S'| \leq \frac{N}{2} + 100$$

However, we need a bound on the $|S|$. Since $\mathbb{E}[|S'|] = \frac{N}{2}$, we can simply multiply our bounds by 2 to get the following:

$$N - 200\sqrt{N} \leq 2|S'| \leq N + 200$$