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

Lecture 17 — October 11, 2023

Prof. Samson Zhou *Scribe: Jung Hoon Seo*

1 Distinct Elements

1.1 Estimation of F_0

Consider a set S comprising N numbers, and let us assume that we create a new set, denoted as S', by sampling each element from S with a probability of $\frac{1}{2}$. Then the expectation for the S' is:

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

and the variance is:

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

Let $X_1, ...X_N$ be indicator random variables, such that $X_i = 1$ when the i-th element of S is sampled into S', and $X_i = 0$ otherwise. Let X be the sum of these random variables, $X_1, ...X_N$ such that $X = |S'|$. Then, the variance of X is:

$$\mathrm{Var}[X] = \mathrm{Var}[X_1] + ... + \mathrm{Var}[X_N] = N \cdot \mathrm{Var}[X_i]$$

and, the variance of X_i is:

$$\mathrm{Var}[X_i] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

Since the variance of each X_i is $\frac{1}{4}$, and the variance of the size of S' is equivalent to X, the variance of S' is as follows:

$$\mathrm{Var}[|S'|] = \frac{N}{4}$$

We can apply Chebyshev's inequality now since we have established a variance bound. This will allow us to analyze the probability that the number of distinct elements in S' deviates from $\mathbb{E}$, which is $\frac{N}{2}$.

$$Pr[|S'| - \frac{N}{2} \ge t]$$

By Chebyshev's inequality, we have

$$Pr[|S'| - \frac{N}{2} \ge 100\sqrt{N}] \le \frac{1}{10}$$

This indicates that the probability of a deviation in the number of distinct elements in S' from its expected value is relatively low. If we establish a bound on $|S'|$ with a probability of at least $\frac{9}{10}$, then the number of distinct elements in the sub-stream S' is:

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

By multiplying the count of distinct elements in S' by 2, we can obtain a reliable estimate of the total number of distinct elements in the original dataset, as follows:

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

If we assume $200\sqrt{N} \leq \frac{N}{200}$ then, we get

$$0.99N \leq 2|S'| \leq 1.01N$$

This confirms that by sampling each item from the universe with a probability of $\frac{1}{2}$, we can create a new universe, U'. If we define S' as the set of items in the data stream that belong to U', then the resulting output is $2|S'|$. However, while this algorithm may provide a good estimate of the actual count of distinct elements, the space required remains unreasonably large. To enhance this, we can form set S' by sampling each item from S with a probability p, instead of the probability of $\frac{1}{2}$. As a result, the expectation for the S' becomes:

$$\mathbb{E}[|S'|] = pN$$

and the variance is:

$$\text{Var}[|S'|] \leq pN$$

Using Chebyshev's inequality with a probability of at least $\frac{9}{10}$, we can establish a bound on the likelihood of the number of elements in S' deviating from its expected value, as follows:

$$pN - 100\sqrt{pN} \leq |S'| \leq pN + 100\sqrt{pN}$$

We can re-scale it by $\frac{1}{p}$, then

$$N - \frac{100}{\sqrt{p}}\sqrt{N} \leq \frac{1}{p}|S'| \leq N + \frac{100}{\sqrt{p}}\sqrt{N}$$

If we assume $\frac{100}{\sqrt{p}}\sqrt{N} \leq \varepsilon N$ then, we get

$$(1-\varepsilon)N \leq \frac{1}{p}|S'| \leq (1+\varepsilon)N$$

Therefore, with probability at least $\frac{9}{10}$, we can conclude that $\frac{1}{p}|S'|$ is a $(1+\varepsilon)$ approximation of N where the value of p is bounded by $p \geq \frac{1000}{\varepsilon^2 N}$.

1.2 Finding p value

While we know that p is bounded by $p \geq \frac{1000}{\varepsilon^2 N}$, determining N becomes a prerequisite for setting p, which poses a challenge since the primary objective is to estimate N. To determine the optimal value for p, we aim for p to be small enough to avoid an excessive number of samples while ensuring it is not too small to result in a low additive error. Suppose we know N, then for $p = \Theta(\frac{1}{\varepsilon^2 N})$, we have

$$\mathbb{E}[|S'|] = pN = \Theta(\frac{1}{\varepsilon^2})$$

To determine p such that $\mathbb{E}[|S'|] = pN = \Theta(\frac{1}{\varepsilon^2})$, we can experiment with values of p, starting with $p = 1, \frac{1}{2}, \frac{1}{4}...$, and selecting the one for which

$$\frac{1000}{\varepsilon} \leq |S'| \leq \frac{2000}{\varepsilon^2}$$

However, incorrect guesses for p may lead to an excessive number of samples. Therefore, we can employ a dynamic approach by adjusting the p values and sub-sampling accordingly to mitigate this issue.

Algorithm

1. Set $U_0 = [n]$ and for each i, sample each element of $U_i - 1$ with probability $\frac{1}{2}$
2. Start index $i = 0$ and track the number $|S \cap U_i|$ of elements S in U_i
3. If $|S \cap U_i| < \frac{2000}{\varepsilon^2} log(n)$, then increment $i = i + 1$

Explanation: Initially, we retain all distinct elements. When the sample count surpasses $\frac{2000}{\varepsilon^2} log(n)$ we dynamically increase our sampling probability by a factor of $\frac{1}{2}$, resulting in a corresponding halving of the maintained set. Simultaneously, we keep track of the values of p, which follow the sequence $\{\frac{1}{2}, \frac{1}{4}, \ldots, \frac{1}{2^n}\}$. Ultimately, this process leads us to determine the optimal value of p, resulting in an output of $2^i \cdot |S \cap U_i|$.

Summary: The algorithm stores a maximum of $\frac{2000}{\varepsilon^2} log(n)$ elements from the stream and requires $\Theta(\frac{1}{\varepsilon^2} log(n))$ words of space.