CS 395T: Sublinear Algorithms, Fall 2020 September 3, 2020

Lecture 3: Distinct Elements; Turnstile Model; l_2 Norm Estimation

Prof. Eric Price *Scribes: Aaron Lamoreaux, Stanley Wei*

NOTE: THESE NOTES HAVE NOT BEEN EDITED OR CHECKED FOR CORRECTNESS

1 Overview

In the last lecture, we discussed how to solve the streaming distinct elements problem in $O\left(\frac{1}{\epsilon^2}\log n \log\left(\frac{\log n}{\epsilon}\right)\right)$ by solving the decision version of the problem. We then showed that by tracking the minimum hashed value, we can solve this in $O\left(\frac{1}{\epsilon^2}\log\left(\frac{\log n}{\epsilon}\right)\right)$.

In this lecture we will continue discussing the distinct elements problem including how we handle the fact that we needed fully independent hash functions. We will then discuss how to solve distinct elements in the strict turnstile model (i.e. allowing deletions of elements in the stream). Finally, we will discuss how to approximate the l_2 norm in the turnstile model.

2 Distinct element bounds with only pairwise independence.

For both of the previous approaches, we needed fully independent random hash functions for the analysis.

Question 1. *How can we get h such that it only takes $o(n)$ space to store and has needed independence properties?*

One option is to use $ax + y \mod n$ which gives limited independence but takes approximately $\log n$ bits to store h. Another option is to hope that the input is sufficiently random.

We can also select h to be a cryptographic hash function such that any non uniformity in the input would require solving a hard problem to find. Thus we are justified in assuming that h is fully random. An example of a cryptographic hash function would be SHA-256.

Question 2. *How can we get good bounds with only pairwise independence?*

Recall how we solved the decision version of the problem of counting distinct elements. We first introduce a hash function $h\colon U \to [B]$, where U is the universal set. Now, define $V := \{x \in U : h(x) = 1\}$. Furthermore, denote S as the set of elements in the stream.

Assuming full independence on the choice $h(x)$ between different x, we have that the value of $\mathbb{P}(|S \cap V| = 0)$ decreases in $|S|$ for $S \in U$.

Recall how we distinguished whether $n < T$ and $n > 2T$. Given full independence, we can compute that

$$\Pr(|S \cap V| = 0) = \left(1 - \frac{1}{B}\right)^n \approx e^{-n/B}.$$

By setting $B = T$, we can now empirically approximate the probability $\Pr(|S \cap V| = 0)$. However this does not hold when we only have pairwise independence.

Recall the statement of Principle of Inclusion-Exclusion. For three sets, A, B, C, we can write by PIE that

$$Pr(A \cup B \cup C) = Pr(A) + Pr(B) + Pr(C) - Pr(A \cap B) - Pr(B \cap C) - Pr(A \cap C) + Pr(A \cap B \cap C).$$

However we can also use PIE, to derive the following inequality on partial values.

Proposition 3. *Given n events $X_1, ...X_n$, we have that*

$$\sum_i Pr(X_i) - \sum_{i,j:i \neq j} Pr(X_i \cap X_j) \leqslant Pr[\bigcup_i X_i] \leqslant \sum_i Pr[X_i].$$

If our hash function h is pairwise independent, then Proposition 2 tells us

$$\frac{n}{B} - \frac{\binom{n}{2}}{B^2} = \frac{n}{B}(1 - \frac{n-1}{B}) \leqslant Pr[\exists x \in S : h(x) = 1] \leqslant \sum_{x \in S} Pr[h(x) = 1] = \frac{n}{B}$$

In particular, if we set $B = 100n$, we obtain that the probability we see any element of the stream being hashed to 1 is between 0.995% and 1%.

Question 4. *In our decision problem, is $n > 2T$, or is $n < T$, for some threshold T? Specifically, how many samples do we need in order to distinguish between the two possibilities?*

If we choose $T = B/100$, we get that if $n > 2T$, then $Pr[\exists x \in S : h(x) = 1] \geqslant 1.98\%$, and if $n < T$, then $Pr[\exists x \in S : h(x) = 1] \leqslant 1\%$. In fact, because these are constants, we can actually distinguish between the two possibilities in the order of constant time.

Since we want to distinguish a probability to within $\pm.5\%$, we can use the bound from the weighted coin puzzle in last lecture, to say that approximately $4 \cdot 100^2$ samples are enough to distinguish between the two probabilities with a failure probability of $1/4$.

3 Distinct elements with deletions (strict turnstile model)

3.1 Turnstile steaming model

Now, we'll include deletions in our streams. We will define the *turnstile* model of data streams. Essentially, we track a vector $x \in \mathbb{R}^n$, where $n = |U|$ (the size of the universal set). Moreover, this model incorporates a series of update queries (i, α) that updates x_i to $x_i + \alpha$. The strict turnstile model is one in which $x_i \geqslant 0$ at all times. The problem of finding the number of distinct elements in here is just the number of nonzero components in x, or $\|x\|_0$ which is called the l_0 norm of x. However we may also desire to compute other functions on the vector x.

3.2 Using previous insertion only algorithm

Question 5. *How can we adopt the algorithm for insertion only to the strict turnstile model.*

For $T = 1, 2, 4, 8, ..., n$, repeat the following sequence $O(\log(\log(\mathrm{n})))$ times in parallel: first, pick a random $h : [n] \to [100T]$; next, define $V \subseteq [n]$ such that $h(i) = 1$ for all $i \in V$; finally, check if $x_V = 0$.

Notice with $1 - \frac{1}{log(n)}$ probability if it's $\geqslant$ 1.98% or $\leqslant$ 1%. In particular, if $\|x\|_0 \leqslant T$, then $Pr[x_V] = 0 \leqslant 1\%$, and if $||x||_0 \geqslant 2T$, then $Pr[x_V] = 0 \geqslant 2\%$. After $O(log(\frac{1}{\delta}))$ independent repetitions, we can correctly determine with probability $1 - \delta$. After $O(log(log(n)))$ repetitions, we can get the right answer for all T by a union bound with $\delta = \frac{1}{\log n}$.

We can also choose the maximum T which outputs "bigger" which will give us a 4-approximation of $\|x\|_0$. We can also do similar analysis with an error of approximation ϵ.

To adapt this approach to the model with deletions, we can instead keep a counter for each given h, the number of elements, x, for which $h(x) = 1$. In other words

$$\sum x_i \mathbf{I}_{h(i)=1} = \sum_{i \in V} x_i.$$

In the strict turnstile model, this is zero if and only if $x_V = 0$. We can maintain this simply under each of the (i, α) updates.

4 AMS-sketch: l_2 Norm Estimation

Lemma 6. *Recall the turnstile model from above. We would like to estimate*

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}}.$$

Let h be a 4-wise independent hash function (we will need the 4-wise independence in our analysis) such that

$$h \colon [n] \to \{-1, +1\}.$$

We will store the estimation

$$y = \sum_{i=1}^{n} x_i h(i) = \sum_{i=1}^{n} x_i v_i$$

where $v_i = h(i)$. We claim that y^2 is a good estimator of $\|x\|_2^2$.

Proof. Let's begin with analyzing the expected value of y^2.

$$\begin{aligned} E[y^2] &= E\left[\left(\sum_{i=1}^{n} x_i v_i \right)^2 \right] \\ &= E\left[\sum_{i=1}^{n} x_i^2 v_i^2 + 2 \sum_{1 \leq i < j \leq n} x_i v_i x_j v_j \right] \end{aligned}$$

Given the linearity of expectation, we can write that

$$= \sum_{i=1}^{n} x_i^2 E[v_i^2] + 2 \sum_{1 \le i < j \le n} x_i x_j E[v_i v_j].$$

Using the pairwise independence from h, we know that $E[v_i v_j] = E[v_i] \cdot E[v_j] = 0$. Additionally $E[v_i^2] = 1$. Thus we get

$$\begin{aligned} &= \sum_{i=1}^{n} x_i^2 \\ &= \|x\|_2^2. \end{aligned}$$

Thus we know that y^2 is a good estimation of $\|x\|_2^2$, however we need to know how many times to sample. Thus let's analyze the variance. Begin with the definition of variance

$$\begin{aligned} \mathrm{Var}(y^2) &= E\left[y^4\right] - E[y^2]^2 \\ &= E\left[\left|\sum_{i=1}^{n} x_i v_i\right|^4\right] - \|x\|_2^4. \end{aligned}$$

Expanding the summation and distributing expectation will give us quite a few terms. It may be clear that multiple of these terms will become 0 because of 4-wise independence. However we can explicitly write it out to get

$$\begin{aligned} &= \sum_{i=1}^{n} x_i^4 E[v_i^4] + 4 \sum_{\substack{1 \le i,j \le n \\ i \ne j}} x_i^3 x_j E[v_i^3 v_j] + 3 \sum_{\substack{1 \le i,j \le n \\ i \ne j}} x_i^2 x_j^2 E[v_i^2 v_j^2] + 6 \sum_{\substack{1 \le i,j,k \le n \\ i \ne j \ne k}} x_i^2 x_j x_k E[v_i^2 v_j v_k] \\ &\quad + \sum_{\substack{1 \le i,j,k,l \le n \\ i \ne j \ne k \ne l}} x_i x_j x_k x_l E[v_i v_j v_k v_l] - \|x\|_2^4 \end{aligned}$$

Notice that because of 4-wise independence, all the terms with odd exponents will cause that term to be 0. Thus we can write that

$$= \sum_{i=1}^{n} x_i^4 + 3 \sum_{\substack{1 \le i,j \le n \\ i \ne j}} x_i^2 x_j^2 - \|x\|_2^4.$$

We can use the fact that $\|x\|_2^4 = \sum_{1 \le i \le n} x_i^4 + \sum_{\substack{1 \le i,j \le n \\ i \ne j}} x_i^2 x_j^2$ to conclude that

$$= 2 \sum_{\substack{1 \le i,j \le n \\ i \ne j}} x_i^2 x_j^2.$$

Notice that $\mathrm{Var}(y^2) \le 2\|x\|_2^4 = 2E[y^2]^2$. Now if we repeat this a total of $\frac{4}{\epsilon^2}$ times, we will be within $\epsilon\sigma$ with a 3/4 probability. Thus we will get $\mu \pm \epsilon\sqrt{2}\mu$ with at least a 3/4 probability. We can change the number of trials to remove the $\sqrt{2}$ as well. □