Advanced Algorithms November 27th, 2024

Lecture 19: Streaming Algorithms (AMS F_2 estimator)

Notes by Ola Svensson[1]

(These notes are based on [**MW**])

In this lecture we start by stating the streaming algorithm for distinct elements and then provide the analysis (last lecture notes). We then see a classic streaming algorithm by Alon, Matias, and Szegedy [**AMS**] for estimating the ℓ_2 norm of the frequency vector. Recall the streaming setting that we consider:

- The input is a long stream $\sigma = \langle a_1, a_2, \ldots, a_m \rangle$ consisting of m elements where each element takes a value from the universe $[n] = \{1, \ldots, n\}$.
- Our central goal is to process the input stream (going from left to right) using a small amount of *space* s, i.e., to use s bits of random-access memory while calculating (approximately) some interesting function/statistics $\phi(\sigma)$.

In this lecture, we are again interested in calculating statistics based on the frequency vector vector $\mathbf{f} = (f_1, \ldots, f_n)$ of the stream, where $f_i = |\{j : a_j = i\}|$ is the number of elements of value i. Note that $f_1 + f_2 + \cdots + f_m = m$. In particular, we want to estimate the second moment

$$F_2 = \sum_{i=1}^{n} f_i^2 \,.$$

1 Naive attempt: downsampling

Since we are dealing with a "big data" problem, we may first downsample the input into a smaller length, then we calculate the second moment of the down sample and we use it to estimate the second moment of the original input. Consider the following set of two inputs.

$$1, 2, 3, 4, \ldots, n$$
$$\underbrace{1, 1, \ldots, 1}_{m \text{ times}}, \underbrace{2, 2, \ldots, 2}_{m \text{ times}}, \ldots, \underbrace{n/m, n/m, \ldots, n/m}_{m \text{ times}},$$

where $m = \Omega(\sqrt{n})$. Observe that any downsample of the first sequence gives completely distinct numbers, and any downsample of the second sequence of size $O(\sqrt{n})$ also gives almost distinct numbers with a high probability. So, any streaming algorithm that is based on downsampling sees almost the same thing, i.e., completely distinct elements, in both cases. However, the second moment of the first sequence is n and the second moment of the second one is $O(n^{3/2})$, so we don't expect a streaming algorithm based on downsampling to size at most $O(\sqrt{n})$ obtain an estimate better than $\sqrt{n}$ of the true second moment.

2 AMS Sketch

Let us first define k-wise independent family of hash functions.

Definition 1 *A family of hash functions $H = \{h : [n] \to U\}$ is a k-wise independent if for any k distinct elements $(x_1, \cdots, x_k) \in U^k$ and any numbers $(u_1, \cdots, u_k)$, we have:*

$$Pr_{h \in H}[h(x_1) = u_1 \wedge \cdots \wedge h(x_k) = u_k] = (\frac{1}{|U|})^k.$$

[1] **Disclaimer:** These notes were written as notes for the lecturer. They have not been peer-reviewed and may contain inconsistent notation, typos, and omit citations of relevant works.

Initialization:

(1) Pick a 4-wise independent hash function: $h : [n] \to \{-1, +1\}$

(2) Let $\sigma_i = h(i)$ so $\sigma \in \{-1, 1\}^n$

(3) Let $Z = 0$.

Process element of value i: $Z = Z + \sigma_i$

Output: Z^2

Note that at the end of the algorithm we have that $Z = \sum_{i=1}^{n} f_i \sigma_i$. As for distinct elements, it is crucial to bound the expected value and variance of Z^2 in the above algorithm. We start by proving that it is an unbiased estimator.

Claim 2

$$\mathbb{E}[Z^2] = ||\boldsymbol{f}||_2^2$$

Proof

$$\begin{aligned}
\mathbb{E}[Z^2] &= \mathbb{E}[(\sum_{i\in[n]} \sigma_i f_i)^2] \\
&= \mathbb{E}[\sum_{i,j\in[n]} \sigma_i \sigma_j f_i f_j] \\
&= \mathbb{E}[\sum_{i\in[n]} \sigma_i^2 f_i^2] + \mathbb{E}[\sum_{i\neq j\in[n]} \sigma_i \sigma_j f_i f_j] \\
&= \mathbb{E}[\sum_{i\in[n]} f_i^2] + \sum_{i\neq j\in[n]} \mathbb{E}[\sigma_i]\mathbb{E}[\sigma_j] f_i f_j \\
&= ||\mathbf{f}||_2^2
\end{aligned}$$

■

Next we bound the variance of the output (to later be apply to apply Chebychev's Inequality).

Claim 3

$$Var[Z^2] \leq 2||\boldsymbol{f}||_2^4$$

Proof Recall that, one can compute the variance of a random variable using the following formula:

$$\text{Var}[Z^2] = \mathbb{E}[Z^4] - (\mathbb{E}[Z^2])^2.$$

Therefore, let us first compute $\mathbb{E}[Z^4]$.

$$Z^4 = (\sum_{i\in[n]} \sigma_i f_i)(\sum_{j\in[n]} \sigma_j f_j)(\sum_{k\in[n]} \sigma_k f_k)(\sum_{l\in[n]} \sigma_l f_l)$$

Let us consider several types of terms:

- all the indexes are equal $i = j = k = l$:
  $\sum_{i\in[n]} \sigma_i^4 f_i^4 = \sum_{i\in[n]} f_i^4$
- the indexes are matched 2 by 2:
  $\binom{4}{2} \sum_{i<j} (\sigma_i \sigma_j f_i f_j)^2 = 6 \sum_{i<j} f_i^2 f_j^2$

- terms with a single (unmatched) multiplier: in this case, since the value $\mathbb{E}[\sigma_i] = 0$ for any $1 \leq i \leq n$, then the coefficient of such terms are zero. (Here we use that our hash function is 4-wise independent.)

Therefore, $\mathbb{E}[Z^4] = \sum_{i\in[n]} f_i^4 + 6\sum_{i<j} f_i^2 f_j^2$. So the variance of Z^2 is :

$$\begin{aligned}
\text{Var}[Z^2] &= \mathbb{E}[Z^4] - (\mathbb{E}[Z^2])^2 \\
&= \sum_{i\in[n]} f_i^4 + 6\sum_{i<j} f_i^2 f_j^2 - (\sum_i f_i^2)^2 \\
&= \sum_i f_i^4 + 6\sum_{i<j} f_i^2 f_j^2 - \sum_i f_i^4 - 2\sum_{i<j} f_i^2 f_j^2 \\
&= 4\sum_{i<j} f_i^2 f_j^2 \\
&\leq 2(\sum f_i^2)^2 \\
&= 2||\mathbf{f}||_2^4
\end{aligned}$$

■

We have $\mathbb{E}[Z^2] = ||\mathbf{f}||_2^2$ and $\text{Var}[Z^2] \leq 2||\mathbf{f}||_2^4$. Now we improve the precision of the estimate by repeating the algorithm for a sufficient number of times (independently) and using the average as an estimate.

(1) For $t = \frac{6}{\epsilon^2}$, maintain t i.i.d copies of the above algorithm. Let $Z_1^2, Z_2^2 \cdots Z_t^2$ be the output of these copies.

(2) Let $\tilde{Z}^2 = \frac{1}{t}\sum_{i=1}^t Z_i^2$.

(3) Output $\tilde{Z}^2$.

By linearity of expectation, we have $\mathbb{E}[\tilde{Z}^2] = ||\mathbf{f}||_2^2$. However, the variance becomes smaller. In particular, the variance of the estimate is now $\frac{\text{Var}(Z^2)}{t} \leq \frac{2}{t}||\mathbf{f}||_2^4$. By Chebyshev's inequality we get

$$\begin{aligned}
Pr[|\tilde{Z}^2 - ||\mathbf{f}||_2^2| > \epsilon||\mathbf{f}||_2^2] &\leq \frac{(\frac{2}{t})\cdot||\mathbf{f}||_2^4}{\epsilon^2||\mathbf{f}||_2^4} \\
&\leq \frac{2}{t\epsilon^2} \\
&\leq \frac{1}{3}
\end{aligned}$$

3 AMS sketch and the Johnson-Lindenstrauss lemma

We just constructed an algorithm for approximating the L_2 norm of a vector $x \in \mathbf{R}^n$ using the AMS sketch, given a stream of updates to entries of x. We generated a random matrix $A \in \mathbf{R}^{m\times n}$ by independently and uniformly sampling each entry A_{ij} from $\{1, -1\}$. We then proved that $\forall\epsilon > 0$, if the dimension $m = O(\frac{1}{\epsilon^2})$, then $\forall x \in \mathbf{R}^n$:

$$\Pr\left[\,\left|\,\|Ax\|_2^2 - m\,\|x\|_2^2\,\right| > \epsilon m\,\|x\|_2^2\right] < \frac{1}{3}$$

This implies that with probability at least $\frac{2}{3}$, $(1-\epsilon)\|x\|_2 \leq \left\|\frac{1}{\sqrt{m}}Ax\right\|_2 \leq (1+\epsilon)\|x\|_2$ – this follows because for any $0 < \epsilon < 1$, $\sqrt{1+\epsilon} < 1+\epsilon$ and $\sqrt{1-\epsilon} > 1-\epsilon$.

3.1 Sketch using a Gaussian distribution

We consider another sketch $A \in \mathbf{R}^{m\times n}$, where $\forall 1 \leq i,j \leq n$, $A_{ij} \sim \mathcal{N}(0,1)$. Examining the conditions imposed on the A's coefficients in the last lecture, we notice that both still hold:

1. $E[A_{ki}A_{kj}] = 0$, $\forall i \neq j$, $1 \leq k \leq m$ because the variables are independent.
2. $E[A_{ik}^2] = 1$, $\forall 1 \leq i \leq m, 1 \leq k \leq n$ by definition.

The new sketch has an additional property: $1 \leq i \leq n$, $(Ax)_i = \sum_{j=1}^n A_{ij}x_j \sim \mathcal{N}(0, \|x\|_2^2)$. This is known as the 2-stability of the Gaussian distribution.

Most importantly, it is possible to prove a strong Chernoff-type concentration inequality for $\|Ax\|_2$. Specifically, $\|Ax\|_2^2 = \sum_{i=1}^m (Ax)_i^2 = \sum_{i=1}^m y_i^2$, $y_i \sim \mathcal{N}(0, \|x\|_2^2)$ follows a χ^2 with m degrees of freedom. The following bound holds for this distribution:

$$\Pr\left[\left|\, \|Ax\|_2^2 - m\|x\|_2^2 \,\right| > \epsilon m \|x\|_2^2\right] < e^{-C\epsilon^2 m} \text{ for a constant } C > 0$$

3.2 Johnson - Lindenstrauss lemma

Lemma 4 *For any $\epsilon \in (0, \frac{1}{2})$, $\forall x_1, \cdots, x_n \in \mathbf{R}^d$, there exists $M \in \mathbf{R}^{m\times d}$ with $m = O(\frac{1}{\epsilon^2}\log n)$ such that for all $1 \leq i,j \leq n$:*

$$(1-\epsilon)\|x_i - x_j\|_2 \leq \|Mx_i - Mx_j\|_2 \leq (1+\epsilon)\|x_i - x_j\|_2$$

Remark This is a statement about dimensionality reduction. The dimension to which the $\mathbf{R}^d$ vectors are reduced, m, does not depend on d, only on the number of vectors.

Proof Fix two indices $i \neq j$ and let $y^{ij} = x_i - x_j$ and $M = \frac{1}{\sqrt{m}}A$, where $A \in \mathbf{R}^{m\times n}$ has i.i.d. elements sampled from $\mathcal{N}(0,1)$. By the previous result and setting $m = \frac{4}{C\epsilon^2}\log n$:

$$\Pr\left[\left|\, \left\|My^{ij}\right\|_2^2 - \left\|y^{ij}\right\|_2^2 \,\right| > \epsilon \left\|y^{ij}\right\|_2^2\right] < e^{-C\epsilon^2 m} = e^{-C\epsilon^2 \frac{4}{C\epsilon^2}\log n} = \frac{1}{n^4}$$

Next, by taking the union bound:

$$\begin{aligned}
\Pr\left[\exists i \neq j, \left|\, \left\|My^{ij}\right\|_2^2 - \left\|y^{ij}\right\|_2^2 \,\right| > \epsilon \left\|y^{ij}\right\|_2^2\right] &\leq \sum_{i\neq j} \Pr\left[\left|\, \left\|My^{ij}\right\|_2^2 - \left\|y^{ij}\right\|_2^2 \,\right| > \epsilon \left\|y^{ij}\right\|_2^2\right] \\
&< \binom{n}{2}\frac{1}{n^4} \\
&< \frac{1}{n^2}
\end{aligned}$$

Therefore $\Pr \forall i \neq j, \left|\, \|Mx_i - Mx_j\|_2^2 - \|x_i - x_j\|_2^2 \,\right| \leq \epsilon \|x_i - x_j\|_2^2 > 1 - \frac{1}{n^2}$. The probability is taken w.r.t the law of M, which allows us to conclude the existence of at least one matrix M satisfying the desired inequality. ■

Remark The bound for m is optimal, as proved by a very recent result Larsen-Nelson'16.

References

[1] Aida Mousavifar and Junxiong Wang: *Scribes of Lecture 2 in Topics in TCS 2017*

[2] N. Alon, Y. Matias, and M. Szegedy: *The space complexity of approximating the frequency moments.* In: *STOC* 1996.