Advanced Algorithms May 10, 2022

Lecture 20: Streaming Algorithms (CountSketch)

Notes by Michael Kapralov[1]

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 $\phi(\sigma)$.

The stream σ implicitly defined the *frequency vector* $f \in \mathbb{R}^n$: for every $i \in [n]$ we let f_i denote the number of occurrences of i in the stream σ.

1 AMS sketch and the Johnson-Lindenstrauss lemma

1.1 Reminder from the last lecture

Last time, we constructed an algorithm for approximating the L_2 norm of a vector $x \in \mathbb{R}^n$ using the AMS sketch, given a stream of updates to entries of x. We generated a random matrix $A \in \mathbb{R}^{m \times n}$ 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 $\forall x \in \mathbb{R}^n$:

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

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 $\sqrt{1-\epsilon} > 1 - \epsilon$.

1.2 Sketch using a Gaussian distribution

We consider another sketch $A \in \mathbb{R}^{m \times n}$, 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:

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

[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.

1.3 Johnson - Lindenstrauss lemma

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

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

Remark This is a statement about dimensionality reduction. The dimension to which the $\mathbb{R}^d$ 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 $A \in \mathbb{R}^{m \times n}$ 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 $\Pr\left[\forall i \neq j, \left| \left\|Mx_i - Mx_j\right\|_2^2 - \left\|x_i - x_j\right\|_2^2 \right| \leq \epsilon \left\|x_i - x_j\right\|_2^2 \right] > 1 - \frac{1}{n^2}$. 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.

2 CountSketch (ℓ_2-heavy hitters)

We now use ideas from the AMS sketch to design a small space algorithm for outputting (approximately) the top few elements of a data stream using a small amount of space. Specifically, we will be able to recover elements with 'large' frequencies – such elements are known as heavy-hitters. The formal definition is

Definition 2 (ℓ_2 Heavy Hitter) *We call $i \in [n]$ a ϕ - heavy hitter if $|x_i| \geq \phi \|x\|_2$ for $\phi \in (0, 1)$.*

Goal: Design a small space algorithm that,

1. Approximates x_i up to $(\phi/4)||x||_2$ error, for every $i \in [n]$.
2. Outputs a list $L \subseteq [n]$ that contains all ϕ - heavy hitters and does not contain any element that is not a $\phi/2$-heavy hitter.

Remark Note that ℓ_2-heavy hitters for constant ϕ is a more powerful primitive than ℓ_1-heavy hitters (the elements found by the Misra-Gries estimator that we discussed two lectures ago). For example, consider a stream of length n where one item occurs $\sqrt{n}$ times, and the remaining $n - \sqrt{n}$ items are distinct. Note that the 'heavy' item is heavy in ℓ_2 sense, but not in ℓ_1 sense. In particular, if one samples a random location in the stream, the probability of hitting the 'heavy' item is only $1/\sqrt{n}$. In particular,

the Misra-Gries estimator will not be able to find the 'heavy' item using $O(\log n)$ space. Nevertheless, the CountSketch algorithm recovers this 'heavy' item using $O(\log n)$ space.

The CountSketch algorithm of Charikar, Chen and Farach-Colton proceeds as follows. Choose R pairwise independent hash functions

$$h_r : [n] \to [B], \quad r = 1, 2, \ldots, R$$

mapping the universe $[n]$ to B buckets. Also choose a sequence of R sign functions

$$s_r : [n] \to \{\pm 1\}, \quad r = 1, 2, \ldots, R$$

from a pairwise independent family. The CountSketch algorithm maintains, for each bucket $b \in [B]$ and repetition $r \in R$,

$$y_{r,b} = \sum_{j \in [n] \text{ s.t. } h_r(j) = b} s_r(j) x_j,$$

where $x \in \mathbb{R}^n$ is the frequency vector of the data stream. Frequency estimation is performed as follows. First, for $i \in [n]$ and $r \in [R]$ define

$$\widehat{x}_i^r = s_r(i) y_{r, h_r(i)}.$$

Our estimate for $i \in [n]$ is then given by

$$\text{median}_{r \in [R]} \{\widehat{x}_i^r\}.$$

We now give the analysis of CountSketch. For convenience relabel elements of the universe $[n] = \{0, 1, 2, \ldots, n-1\}$ so that $x_0 \geq x_1 \geq x_2 ... \geq x_n$. We then define the head and tail of x as follows. Define the head of the signal as $H = \{0, 1, 2, ...k-1\}$ (top k frequencies) and the tail as $T = \{k, ...n-1\}$.

2.1 Bounding estimation error for fixed $r \in [R]$

We now analyze the estimation error. It is convenient to consider the contribution of the head and the tail of the signal separately to the estimation error:

$$\begin{aligned}
\widehat{x}_i^r - x_i &= s_r(i) y_{r, h_r(i)} - x_i \\
&= \sum_{j \in [n] \setminus \{i\} \text{ s.t. } h_r(j) = h_r(i)} s_r(i) s_r(j) x_j \\
&= \underbrace{\sum_{\substack{j \in H \setminus \{i\} \\ h_r(i) = h_r(j)}} s_r(i) s_r(j) x_j}_{\substack{\| \\ \text{0 if i is alone in its' bucket}}} + \underbrace{\sum_{\substack{j \in T \setminus \{i\} \\ h_r(i) = h_r(j)}} s_r(i) s_r(j) x_j}_{\substack{\| \\ \Delta(i,r)}} \qquad (1)
\end{aligned}$$

Note that if i does not collide with any of the head elements of the signal in its bucket by $\mathcal{E}_{no-collisions}(i, r)$. Note that this event is quite likely if the number of buckets is much larger than k. Formally, we have for every $r \in [R]$ and $i, j \in [n], i \neq i$

$$\Pr\left[h_r(j) = h_r(i)\right] = \frac{1}{B}.$$

We thus have that the probability that i does not collide with any of the head elements in its bucket is upper bounded by

$$\Pr\left[\exists\, j \in H\backslash\{i\} : h_r(i) = h_r(j)\right] \leq \frac{k}{B},$$

where we used the union bound over at most k elements of H. Choosing $B \geq 10k$, we get that this probability is at most $\frac{1}{10}$, giving that $\Pr\left[\mathcal{E}_{no-collisions}(i,r)\right] \geq 9/10$, and thus the first term in (1) is zero with probability at least $9/10$.

We next show that the second term, namely $\Delta(i,r)$, is small in absolute value with high probability. We first show that the expectation of the second term is zero, and then bound the variance. For the expectation we have

$$\begin{aligned}
\mathbf{E}[\Delta(i,r)] &= \mathbf{E}\left[\sum_{\substack{j\in T\backslash\{i\}\\ h_r(i)=h_r(j)}} s_r(i)s_r(j)x_j\right] \\
&= \sum_{\substack{j\in T\backslash\{i\}\\ h_r(i)=h_r(j)}} \mathbf{E}[s_r(i)s_r(j)]x_j \\
&= 0
\end{aligned}$$

We now bound the variance, conditioned on a specific choice of h_r:

$$\begin{aligned}
\mathbf{E}_s[\Delta(i,r)^2] &= \mathbf{E}[\Delta(i,r)^2] - (\mathbf{E}[\Delta(i,r)])^2 \\
&= \mathbf{E}_s\left[\left(\sum_{\substack{j\in T\backslash\{i\}\\ h_r(j)=h_r(i)}} s_r(i)\; s_r(j)\; x_j\right)^2\right] \qquad (\text{since } \mathbf{E}[\Delta(i,r)] = 0) \\
&= \mathbf{E}_s\left[\sum_{\substack{j\in T\backslash\{i\}\\ h_r(i)=h_r(j)}}\sum_{\substack{j'\in T\backslash\{i\}\\ h_r(i)=h_r(j')}} s_r(i)^2 s_r(j)s_r(j')x_jx_{j'}\right] \\
&= \sum_{j,j'} \mathbf{E}_s[s_r(j)s_r(j')]x_jx_{j'} \qquad (\text{since } s_r(i)^2 \text{ is always } 1) \\
&= \sum_{j\in T\backslash\{i\}} x_j^2 \quad (\text{ since } \mathbf{E}_s[s_r(j)s_r(j')] = 0 \text{ if } j \neq j' \text{ and } 1 \text{ if } j = j')
\end{aligned}$$

We thus have

$$\Pr\left[\Delta(i,r)^2 \geq 10 \sum_{\substack{j\in T\backslash\{i\}\\ h_r(i)=h_r(j)}} x_j^2\right] \leq \frac{1}{10},$$

where the probability is over s, conditioned on a fixed choice of h_r. Define the event $\mathcal{E}_{small-noise}(i,r)$ by

$$\mathcal{E}_{small-noise}(i,r) := \left\{\Delta(i,r)^2 \leq 10 \sum_{\substack{j\in T\backslash\{i\}\\ h_r(i)=h_r(j)}} x_j^2\right\}.$$

How small is $\sum_{\substack{j\in T\setminus\{i\}\\ h_r(i)=h_r(j)}} x_j^2$ typically? Taking the expectation over the hash function h, we get

$$\begin{aligned}\mathbf{E}_h\left[\sum_{\substack{j\in T\setminus\{i\}\\ h_r(i)=h_r(j)}} x_j^2\right] &= \mathbf{E}\left[\sum_{j\in T\setminus\{i\}} x_j^2\cdot \mathbf{1}_{[h_r(j)=h_r(i)]}\right]\\ &= \sum_{j\in T\setminus\{i\}} x_j^2\cdot \Pr\left[h_r(j)=h_r(i)\right]\\ &\leq \frac{1}{B}\sum_{j\in T} x_j^2,\end{aligned}$$

where the probability is over the choice of h_r. Now using Markov's Inequality, we get

$$\Pr\left[\sum_{\substack{j\in T\setminus\{i\}\\ h_r(i)=h_r(j)}} x_j^2 > \frac{10}{B}\sum_{j\in T} x_j^2\right] \leq \frac{1}{10}$$

We define $\mathcal{E}_{\text{small-var}}(i,r)$ by

$$\mathcal{E}_{\text{small-var}}(i) := \left\{\sum_{\substack{j\in T\setminus\{i\}\\ h_r(i)=h_r(j)}} x_j^2 < \frac{10}{B}\sum_{j\in T} x_j^2\right\}.$$

Letting x_T denote the restriction of x onto coordinates in T, we get $\sum_{j\in T} x_j^2 = \|x_T\|_2^2$. Using this notation, we get by a union bound over $\mathcal{E}_{no-collisions}(i,r), \mathcal{E}_{small-noise}(i,r)$ and $\mathcal{E}_{small-var}(i,r)$ that

$$\Pr\left[|\widehat{x}_i^r - x_i|^2 \leq \frac{100\,\|x_T\|_2^2}{B}\right] \geq 1 - \frac{1}{10} - \frac{1}{10} - \frac{1}{10} \geq 7/10.$$

2.2 Putting it together

We repeat this process $R = C_1 \log n$ times for a sufficiently large constant $C > 0$ to get $\widehat{x}_i^1, \widehat{x}_i^2, \ldots, \widehat{x}_i^R$. Recall that our final estimate is

$$\widehat{x}_i = \text{median}_{r\in[R]}\{\widehat{x}_i^r\}.$$

By standard median trick analysis we have $|\widehat{x}_i - x_i| \leq \frac{100\|x_T\|_2}{\sqrt{B}}$ with probability at least $1 - 1/n^2$ for every fixed $i \in [n]$. By a union bound over all $i \in [n]$ we thus have

$$\|\widehat{x} - x\|_\infty \leq \frac{10\,\|x_T\|_2}{\sqrt{B}}$$

with probability at least $1 - 1/n$.

To solve the original problem, just let $B = C_2 k/\phi^2$ for a sufficiently large C_2 to ensure that $\frac{10\|x_T\|_2}{\sqrt{B}} < (\phi/4)||x_T||_2 \leq (\phi/4)||x||_2$, and let the output list be defined as

$$L = \{i \in [n] : |\widehat{x}_i| > (3\phi/4)||x||_2\}.$$

Remark Note that we proved stronger upper bounds on the quality of estimation provided by CountSketch than are needed for the application to heavy hitters. Specifically, we showed that our estimate errs by at most $\frac{10\|x_T\|_2}{\sqrt{B}}$, i.e. the error depends on the ℓ_2 mass in the tail of the signal only. This in particular shows that vectors x with at most k nonzero entries can recovered **exactly** from our sketch.