CS-448 Sublinear Algorithms for Big Data Analysis March 12, 2025

Lecture 4

Lecturer: Michael Kapralov

1 The heavy-hitters problem

In this section we show how one can approximate individual entries of x presented as a stream of updates using a small sketch. Specifically, can we approximate top k elements of x? What about at least $k = 1$ case?

The answer is, in general, 'NO'. But 'YES' if max coordinate contributes a nontrivial fraction of the mass of x. We assume that entries in x are nonnegative at the end. We start the discussion by formally defining 'a coordinate contributing a nontrivial fraction of the mass of x'.

Definition 1 (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 linear sketch that,

1. Approximates x_i up to $(\phi/4)||x||$ 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 the definition above makes sense with the ℓ_2 norm replaced with the ℓ_1 norm. The fact that we are able to solve the problem under the above definition with ℓ_2 norm is surprising. 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 $1/\sqrt{n}$. Nevertheless, the CountSketch algorithm recovers this 'heavy' item using $O(\log n)$ space.

$\|x\|_2 \leq \|x\|_1$ for all x, and $\|x\|_2 \ll \|x\|_1$ for some x

The CountSketch algorithm of Charikar, Chen and Farach-Colton [1] 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

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

This is basically a hash table lookup

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

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

We now given the analysis of CountSketch. Consider an arrangement of x_i s such that $|x_1| \geq |x_2| ... \geq |x_n|$. We then define the head and tail of x as follows. Define the head of the signal as $H = \{1, 2, ...k\}$ (top k frequencies) and the tail as $T = \{k+1, ...n\}$.

1.1 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:

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

Denote the event that i does not collide with any of the head elements of the signal in its bucket in the r-th hashing 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

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

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

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

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

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

We thus have

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

where the probability is over s, conditioned on a fixed choice of h_r.

Define the event $\mathcal{E}_{small-noise}(i,r)$ by

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

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

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

1.2 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. We will use these stronger bounds when we talk about sparse recovery.

2 Sparse recovery

Definition 2 *Let x_j denote the j-th largest element (in absolute value) of x, then given a fixed k, we define $x_{(1,2\ldots,k)} = x_H$ as the head of x and $x_{(k+1,\ldots,n)} = x - x_H = x_T$ as the tail of x.*

Claim 3 *When $x_{(i)} \sim\ i^{-\alpha}$ (distribution satisfying power law) for $\alpha \in (\frac{1}{2}, 1)$,*

$$\frac{||x_{(k+1,\ldots,n)}||_2^2}{||x||_2^2} \sim\ k^{-2\alpha+1}$$

Proof We have $\sum_{i=k}^{\infty}(i^{-\alpha})^2 = \sum_{i=k}^{\infty}(i^{-2\alpha}) \approx k^{-2\alpha+1} \ll 1$, which can be seen by evaluating the integral $\int_k^{\infty} x^{-2\alpha}dx = \frac{1}{2\alpha+1}k^{-2\alpha+1}$. ■

Checking the above claim for ℓ_1 norm, we obtain $\frac{||x_{(k+1,\ldots,n)}||_1}{||x||_1} \sim\ k^{\alpha+1}$ (since $\sum_{i=k}^{\infty}(i^{-\alpha}) \approx k^{-\alpha+1}$), thus the tail is not as sparse for the ℓ_2 norm as it is for ℓ_1 norm for distributions obeying the power law.

Definition 4 (Sparse Recovery) *Given Ax for some $x \in \mathbb{R}^n$ and an integer $k \geq 1$ and a precision parameter $C > 1$, reconstruct $y \in \mathbb{R}^n$ s.t*

$$\begin{aligned} ||x - y||_p &\leq C \min_{k-sparse\ z}(||x - z||_q) \\ &= C||x_T||_q. \end{aligned}$$

Note that C can be close to 1, in which case we talk about $C = 1 + \epsilon$-approximate sparse recovery. Specific instantiations of p and q include the following.

ℓ_1/ℓ_1 guarantee : Given Ax reconstruct y such that

$$||x - y||_1 \leq C \min_{k-sparse\ z} ||x - z||_1$$

ℓ_2/ℓ_2 guarantee:

$$||x - y||_2 \leq C \min_{k-sparse\ z} ||x - z||_2$$

y satisfies the ℓ_∞/ℓ_1 guarantee if:

$$||x - y||_\infty \leq \frac{\epsilon}{k}||x_T||_1$$

y satisfies the ℓ_∞/ℓ_2 guarantee if:

$$||x - y||_\infty \leq \frac{\epsilon}{\sqrt{k}}||x_T||_2$$

Claim 5 *For every integer $k \geq 1$, every $\epsilon \in (0,1)$ if y be the vector of estimates given by CountMin with $B = \Theta(\frac{k}{\epsilon})$ and $R = C \log n$ for a sufficiently large constant $C > 0$, then with probability at least $1 - 1/n$*

$$||x - y||_\infty \leq \frac{\epsilon}{k}||x_T||_1$$

Claim 6 *For every integer $k \geq 1$, every $\epsilon \in (0,1)$ if y be the vector of estimates given by* CountSketch *with $B = \Theta(\frac{k}{\epsilon^2})$ and $R = C \log n$ for a sufficiently large constant $C > 0$, then with probability at least $1 - 1/n$*

$$||x - y||_\infty \leq \frac{\epsilon}{\sqrt{k}}||x_T||_2$$

References

[1] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.