1 Overview

In the last lecture, we discussed how to estimate the mean of a random variable $X$ using an estimator that does not depend on the choice of $\varepsilon$ when $X$ is symmetric. The main insight was that, although the median does not in general work, the median of pairwise means does. To analyze this estimator, we used Fourier analysis.

In this lecture, we will continue analyzing Heavy Hitter algorithms. We will provide an improved analysis of the Count-Min sketch, and discuss how to improve these guarantees further using the Count-Sketch.

2 Improved analysis for Count-Min

Goal: given a stream $u_1, \ldots, u_N \in [n]$, construct an estimate $\hat{x}$ of the histogram of the stream $x \in \mathbb{R}^n$, where $x_u = |\{u_i : u_i = u, i \in [N]\}|$.

2.1 Count-Min algorithm

Note: The Count-Min algorithm assumes a strict turnstile model, since the analysis assumes the final histogram $x \geq 0$.

In order to obtain an estimate for $x$, pick independently $R$ pairwise independent hash functions $h : [n] \rightarrow [B]$, for parameters $R$ and $B$ to be chosen. The algorithm will then store $R$ linear sketches $y \in \mathbb{R}^B$ of $x$, where the $j$th entry in the $i$th sketch is given by:

$$y_j^{(i)} = \sum_{u : h(u) = j} x_u$$  \hspace{1cm} (1)

Storing these sketches will require $O(RB)$ space. To obtain our final estimator $\hat{x}$, we compute, for each $u \in [n]$,

$$\hat{x}_u = \min_{i \in [R]} y^{(i)}_{h_i(u)}$$

2.2 Sketch of analysis from last time

Last lecture, we proved the following guarantee for Count-Min:
Lemma 1. Take $\hat{x} \in \mathbb{R}^n$ to be the estimate of $x$ output by Count-Min, using $R$ sketches each of length $B$. If we choose $R = 2 \log n$, then with probability $1 - \frac{1}{n}$,

$$\|\hat{x} - x\|_\infty \leq \frac{2\|x\|_1}{B}.$$  \hfill (2)

Proof sketch. Observe that, in a strict turnstile model, $y^{(i)}_{h_i(u)}$ is an estimate of $x_u$ plus some additional terms $x_{u'}$. In particular, denoting $\hat{x}_u^{(i)} = y^{(i)}_{h_i(u)}$, we have that:

$$x_u \leq \hat{x}_u^{(i)} = x_u + \sum_{v \neq u, h_i(v) = h_i(u)} x_v.$$  \hfill (3)

Therefore, by pairwise independence of the hash functions,

$$E[\hat{x}_u^{(i)} - x_u] = \sum_v E[x_v 1\{v \neq u, h_i(v) = h_i(u)\}]$$

$$\leq \sum_v x_v P(h_i(v) = h_i(u))$$

$$= \frac{\|x\|_1}{B}.$$.  \hfill (4)

As a consequence, by Markov’s inequality, $\hat{x}_u^{(i)} \leq x_u + 2\frac{\|x\|_1}{B}$ with probability at least $\frac{1}{2}$. Therefore, the by taking the minimum of all of our estimators, we may boost the success probability to $1 - \frac{1}{2^R}$. So, if we take $R = 2 \log n$, then $\|\hat{x} - x\|_\infty \leq \frac{2\|x\|_1}{B}$ with probability $1 - \frac{1}{n}$. \hfill $\square$

2.3 Sparsity-aware bounds

Suppose that the histogram $x \in \mathbb{R}^n$ only has a few non-zero coordinates. As an example, suppose that we wish to estimate the number of votes in an election, where the vast majority of voters select only one of two candidates. Is it possible to obtain error bounds that are independent of the large coordinates?

In this lecture, we will show the following:

1. $\|\hat{x} - x\|_\infty \leq O\left(\frac{\|x-H_k(x)\|_1}{k}\right)$ with $O(k \log n)$ space (where $H_k(x)$ denotes the largest $k$ coordinates of $x$). Recall that we saw how to do this previously in the insertion-only model using the FrequentElements algorithm.

2. $\|\hat{x} - x\|_\infty \leq O\left(\frac{\|x-H_k(x)\|_2}{\sqrt{k}}\right)$. We will discuss in class and in the homework why this $\ell_2$ norm bound is much better than the $\ell_1$ norm bound.

3. Fast recovery, but $O(k \log^2 n)$ space.
2.3.1 $\ell_1$-error guarantee

Lemma 2. Let $\hat{x} \in \mathbb{R}^n$ be the estimate of $x$ output by Count-Min, and let $H_k(x)$ denote the largest $k$ entries of $x$. If we choose $R = O(\log n)$ and $B = O(k)$, then with probability $1 - \frac{1}{n}$,

$$\|\hat{x} - x\|_{\infty} \leq \frac{\|x - H_k(x)\|_1}{k}$$

Proof. In order to obtain an error bound that scales with $\|x - H_k(x)\|_1$, there must be zero error when there are less than $k$ nonzero values. Now, recalling the decomposition from our previous analysis in Equation 3, and denoting $H \subset [n]$ as the indices of the top $k$ values of $x$, we may write:

$$\hat{x}_u[i] - x_u = \sum_{v \neq u} \begin{cases} x_v & \text{if } h_i(v) = h_i(u) \\ x_v & \text{if } h_i(v) = h_i(u) \\ \sum_{v \in H} E_H & \text{if } h_i(v) = h_i(u) \\ \sum_{v \notin H} E_T & \text{if } v \notin H \end{cases}$$

Now, $E_H$ is the HeavyHitter contribution, and $E_T$ is the tail contribution. If there are $\leq k$ nonzero entries in $x$, then $E_T = 0$. Thus, we want to show that the HeavyHitter contribution, $E_H$, is zero with constant probability. This follows by observing that, by a union bound and pairwise-independence of the hash functions,

$$\mathbb{P}(E_H \neq 0) \leq \mathbb{P}(\exists v \in H \setminus \{u\} : h_i(v) = h_i(u)) \leq \frac{k}{B}$$

Thus, by taking $B = 4k$, $\mathbb{P}(E_H = 0) \geq \frac{3}{4}$. We may bound the tail error $E_T$ in the same way as the original analysis of Equation 4, we may bound the expectation of the tail error as:

$$\mathbb{E}[E_T] \leq \sum_{v \notin H} x_v \mathbb{P}(h_i(v) = h_i(u)) = \frac{\|x - H_k(x)\|_1}{B}$$

Thus, by Markov’s inequality:

$$\mathbb{P}\left(E_T \geq \frac{4\|x - H_k(x)\|_1}{B}\right) \leq \frac{1}{4}$$

Combining our results, we have that, since $B = 4k$,

$$\mathbb{P}\left(x_u^{(i)} - x_u \geq \frac{4\|x - H_k(x)\|_1}{B}\right) = \mathbb{P}\left(x_u^{(i)} - x_u \geq \frac{\|x - H_k(x)\|_1}{k}\right) = \mathbb{P}(E_H + E_T \geq \frac{\|x - H_k(x)\|_1}{k}) \leq \mathbb{P}(E_H \neq 0) + \mathbb{P}(E_T \geq \frac{\|x - H_k(x)\|_1}{k}) \leq \frac{1}{2}$$
Thus, taking the minimum over our $R$ estimators as before, we have that

$$\mathbb{P}\left(\hat{x}_u - x \geq \frac{\|x - H_k(x)\|_1}{k}\right) \leq \frac{1}{2R}$$

with $RB = 4Rk$ words. In particular, $\|\hat{x} - x\|_\infty \leq \frac{\|x - H_k(x)\|_1}{k}$ with probability $1 - \frac{1}{n}$ using $O(k \log n)$ words.

2.3.2 Comparing $\ell_1$ and $\ell_2$-error guarantees

There is a problem, however, with the $\ell_1$-error guarantee that we’ve just obtained. In many real-world data streams, the coefficients decay at a rate of $i^{-\alpha}$, for $\alpha \in (0, 1)$. Several examples are shown in Figure 1. This rate of decay is problematic because, while $\|x - H_k(x)\|_1$ is not summable in this regime, $\|x - H_k(x)\|_2$ is. In particular, if $x_i = x_1i^{-\alpha}$ for $\alpha < 1$, then

$$\|x - H_k(x)\|_1 = \sum_{i=k+1}^{n} x_i$$

$$= \sum_{i=k+1}^{n} x_1i^{-\alpha}$$

$$\approx x_1 \int_{k}^{n} i^{-\alpha} di$$

$$= \frac{x_1}{1-\alpha} \left(n^{1-\alpha} - k^{1-\alpha}\right)$$

$$= \Theta(x_1 n^{1-\alpha})$$

Therefore, our error bounds on $\|\hat{x} - x\|_\infty$ are only $\frac{\|x - H_k(x)\|_1}{k} \approx \frac{n^{1-\alpha} x_1}{k}$. So we need $k = n^{1-\alpha} c^\alpha$ to have error less than $x(c)$. However, if instead, we could achieve an error bound scaling with the $\ell_2$
norm as $\|x - H_k(x)\|_2/\sqrt{k}$, then on this range of $\alpha \in (0.5, 1)$, we have that

$$\|x - H_k(x)\|_2^2 = \sum_{i=k+1}^{n} x_i^2$$

$$= x_1^2 \sum_{i=k+1}^{n} i^{-2\alpha}$$

$$\approx \frac{x_1^2}{2\alpha - 1} \left( \frac{1}{k^{2\alpha - 1}} - \frac{1}{n^{2\alpha - 1}} \right)$$

$$= \Theta \left( \frac{x_1^2}{k^{2\alpha - 1}} \right)$$

Therefore,

$$\|\hat{x} - x\|_\infty^2 \leq O \left( \frac{x_1^2}{k^{2\alpha}} \right)$$

$$= O \left( x_1^2 \right)$$

so we need only that $k = c$ in order to estimate $x(c)$ within constant factors.

Several additional notes:

- Cannot obtain better results that aren't implied by 2-norm bounds using logarithmic space
- $\ell_1$ bound is attainable deterministically (for example, the \texttt{FrequentElements}), whereas the $\ell_2$ bound requires randomization.

## 3 Count-Sketch: from $\ell_1$ to $\ell_2$ bounds

In this section, our goal remains the same: estimate the histogram $x \in \mathbb{R}^n$ of a data stream. The algorithm will be quite similar to the Count-Sketch. We hope to improve the dependence on the $\ell_1$ norm to a dependence on $\ell_2$ norm.

### 3.1 Count-Sketch algorithm

\textbf{Note:} Unlike in the Count-Min sketch, this algorithm works in the non-strict turnstile model, where the final histogram may have positive or negative entries.

As before, we choose independently $h_1, \ldots, h_R : [n] \rightarrow [B]$ pairwise independent hash functions, and additionally choose independently $s_1, \ldots, s_R : [n] \rightarrow \{\pm 1\}$ pairwise independent random signs. The algorithm will then store $R$ linear sketches $y \in \mathbb{R}^B$ of $x$, where the $j$th entry of the $i$th sketch is now given by

$$y^{(i)}_j = \sum_{u : h_i(u) = j} x_u s_i(u).$$
The algorithm will then output the estimate
\[ \hat{x}_u = \text{median}_{i \in [R]} y_{h_i(u)}^{(i)} s_i(u) \]
Compared with Equation 1, the only difference in the stored values is the random signs that are now being used. As before, the algorithm will need \( O(RB) \) space. Observe that, because each estimate \( \hat{x}_u^{(i)} = y_{h_i(u)}^{(i)} \) is no longer an upper estimate of \( x_u \), we cannot choose the minimum estimate as our final estimator.

In contrast to the Count-Min sketch, the Count-Sketch does not require the strict turnstile model.

### 3.2 Count-Sketch analysis

**Lemma 3.** Let \( \hat{x} \in \mathbb{R}^n \) be the estimate of \( x \) output by Count-Sketch, with \( R = O(\log n) \) and \( B = O(k) \). Then, with probability \( 1 - \frac{1}{n} \),
\[ \| \hat{x} - x \|_\infty \leq \frac{\| x - H_k(x) \|_2}{\sqrt{k}} \]

**Observation 1.** As we show in Homework 4 problem 2, the guarantee in Lemma 3 is a strictly better guarantee than Lemma 2.

**Proof of Lemma 3.** Let us denote \( \hat{x}_u^{(i)} = y_{h_i(u)}^{(i)} s_i(u) \). Unlike in our estimate for Count-Min, our estimate of \( x_u \) is an unbiased estimate (recall that the estimate used in Count-Min was an overestimate of \( x_u \)). In particular, we have that
\[ \hat{x}_u^{(i)} = x_u + \sum_{v \not= u \atop h_i(v) = h_i(u)} x_v s_i(v) s_i(u) \]
Hence, by pairwise independence, and by an application of the tower rule of expectation,
\[ \mathbb{E}[\hat{x}_u^{(i)}] = x_u \]
As before, we will split the error into two terms: error due to the heavy hitters, and the error due to the tail. Indeed, letting \( H \subset [n] \) be the set of \( k \) heavy hitters, we may write
\[ \hat{x}_u^{(i)} - x_u = \sum_{v \not= u \atop h_i(v) = h_i(u) \atop v \in H} x_v + \sum_{v \not= u \atop h_i(v) = h_i(u) \atop v \not\in H} x_v = E_H + E_T \]
Now, as before,
\[ \Pr(E_H \neq 0) \leq \Pr(\exists v \in H \setminus u : h_i(v) = h_i(u)) \leq \frac{k}{B} = \frac{1}{20} \]
if \( B = 20k \). Thus, it suffices to bound \( E_T \). We have that
\[ \Pr(|E_T| > \tau) = \Pr(E_T^2 > \tau^2) \leq \frac{\mathbb{E}[T^2]}{\tau^2} \]
Now,

\[
\mathbb{E}[E_T^2] = \mathbb{E}_{h,s}\left[\left(\sum_{v \neq u, \ h(v) = h(u), \ v \notin H} x_v s_i(v) s_i(u) + \sum_{v \neq u, \ v \notin H} x_v x'_v \sum_{s_i(v) s_i(v') \text{ pairwise independent}} s_i(v) s_i(v') s_i(u)^2 \right)^2\right]
\]

\[
= \mathbb{E}_{h,s}\left[\sum_{v \neq u, \ h(v) = h(u), \ v \notin H} x_v^2 + \sum_{v \neq u, \ v \notin H, \ v \neq v', \ h(v) = h(u), \ h(v') = h(v')} x_v x'_v \sum_{s_i(v) s_i(v') \text{ pairwise independent}} s_i(v) s_i(v') s_i(u)^2 \right)
\]

\[
\leq \frac{\|x - H_k(x)\|_2^2}{B}
\]

Therefore,

\[
\mathbb{P}\left(|E_T| > \frac{\|x - H_k(x)\|_2}{\sqrt{k}}\right) \leq \frac{k}{B} = \frac{1}{20}
\]

Combining these two results establishes that

\[
\mathbb{P}\left(|\hat{x}^{(i)}_u - x_u| \geq \frac{\|x - H_k(x)\|_2}{\sqrt{k}}\right) \leq \frac{1}{10}
\]

We may now apply a Chernoff bound to obtain

\[
\mathbb{P}\left(|\hat{x}^{(i)}_u - x_u| \geq \frac{\|x - H_k(x)\|_2}{\sqrt{k}}\right) \leq \mathbb{P}\left(\sum_{i \in [R]} 1 \left\{ |\hat{x}^{(i)}_u - x_u| \geq \frac{\|x - H_k(x)\|_2}{\sqrt{k}}\right\} \geq \frac{R}{2}\right)
\]

\[
\leq \exp(-\Omega(R))
\]

Hence, setting \(R = O(\log n)\), \(|\hat{x} - x|_\infty \leq \frac{\|x - H_k(x)\|_2}{\sqrt{k}}\) with probability \(1 - \frac{1}{n}\).

**Question:** Suppose that we do not care about the worst case error \(|\hat{x} - x|_\infty\), but only the error for some fixed \(u\). Can we give a better bound than \(|\hat{x}^{(i)}_u - x_u| \leq \frac{\|x - H_k(x)\|_2}{\sqrt{k}}\)? How big is it usually?

**Idea:** Exploit symmetry of \(\hat{x}^{(i)}_u\), as discussed in the last lecture.

Indeed, if \(h, s\) are fully independent, then \(\hat{x}^{(i)}_u\) is symmetric about \(x_u\), and additionally, the Fourier transform \(\mathcal{F}(\hat{x}^{(i)}_u - x_u)\) is nonnegative. Thus, we can show that the median value \(\hat{x}_u = \text{median}_{i \in [R]} \hat{x}^{(i)}_u\) converges to \(x_u\) as \(R \to \infty\).

In particular, this implies that

\[
\mathbb{P}\left(|\hat{x}_u - x_u| \geq \frac{\|x - H_k(x)\|_2}{\sqrt{k}}\sqrt{\frac{t}{R}}\right) \leq \exp(-\Omega(t))
\]

When \(t = R\), then this is the same bound as before, but gives better concentration with lower probabilities. Hence, if we don’t care about the worst-case error \(|\hat{x} - x|_\infty\), but only the error in a few fixed coordinates of \(x\), then the error will usually be a factor of \(\sqrt{\log n}\) smaller than the worst-case error. The downside of this argument, however, is that the proof requires full independence.