**CS 15-859: Algorithms for Big Data** Fall 2017

Lecture 8-1 — 10/26/2017

*Prof. David Woodruff* *Scribe: Sahil Singla*

# 1 Announcements

- Course projects suggestions have been posted online. Students should work in groups of size at most 2 (special permission needed otherwise).
- Last one-two lectures of the course will be project presentations.
- HW-4 will be just a project update.
- Send a 1 page project summary (project idea and group members) along with HW-3.

# 2 $\ell_p$-Norm Estimation for $p > 2$ (continued)

Consider a stream of addition/subtraction updates to an $n$ dimensional vector $y$, while always ensuring that $y \in [-M, M]^n$ (the turnstile model). The $\ell_p$-norm estimation problem is to estimate $\|y\|_p$ using $o(n)$ space (assume $p = O(1)$). In this section we present a result due to Andoni [1] that finds an $O(1)$-approx to $\|y\|_p$ using $\tilde{O}\left(n^{1-\frac{2}{p}}\right)$ space. This bound is tight up to $poly \log n$ factors.

## 2.1 Recap

Our sketch is of the form $P \cdot D$ where $P$ is a $s \times n$ $\mathsf{CountSketch}$ matrix and $D$ is a $n \times n$ diagonal matrix $\mathsf{diag}\left(\frac{1}{E_1^{1/p}}, \ldots, \frac{1}{E_n^{1/p}}\right)$ with $E_i$ being an exponential random variable of rate 1. Using the fact that minimum of independent exponential random variables has an exponential distribution, in the last lecture we prove that w.p. at least 4/5, we have

$$\frac{\|y\|_p^p}{10} \leq \|Dy\|_\infty^p \leq 10\|y\|_p^p. \tag{1}$$

Hence, $\|Dy\|_\infty$ is a good estimator of $\ell_p$-norm. However, the difficulty is that we cannot store $Dy$ as its size is $n$. Using a $\mathsf{CountSketch}$ matrix $P$ lets us reduce the dimension of $Dy$, but we need to ensure $\|PDy\|_\infty \approx \|Dy\|_\infty$.

Condition on $D$ satisfying Eq. (1). To prove $\|PDy\|_\infty \approx \|Dy\|_\infty$, the plan is to apply Bernstein's concentration bound. In the last lecture we show that the random variable $(PDY)_i$ satisfies $\mathbb{E}[(PDY)_i] = 0$ and $\mathbb{E}[(PDY)_i^2] = O(1/s) \cdot (n^{1-\frac{2}{p}}\|y\|_p^2)$. Now if we naively apply Bernstein's bound then to bound the noise we need a small bound on $K = \|DY\|_\infty$, which we know does not hold as this is what we are trying to estimate. In today's lecture we will show that not many coordinates $j$ have a large $|(DY)_j| = \frac{|y_j|}{E_j^{1/p}}$ ( $\geq \Omega(\frac{\|y\|_p}{\log n})$. Hence, if we condition on these coordinates going to distinct buckets then the remaining entries are small and we can get good concentration bounds.

### 2.2 Proving $\|PDy\|_\infty \approx \|Dy\|_\infty$

Condition on $D$ satisfying Eq. (1). Let $L$ denote the set of *large* coordinates $j$ s.t. $R_j := |(DY)_j| = \frac{|y_j|}{E_j^{1/p}} \geq \frac{\alpha \cdot \|y\|_p}{\log n}$, where $\alpha$ is a small constant. We call the remaining coordinates *small*.

**Claim 1.** With probability at least 9/10, the number of large coordinates $|L|$ is at most $O(\log^p n)$.

*Proof.* The probability that a particular coordinate $j$ belongs to $L$ is

$$\begin{aligned}\mathbb{P}\left[\frac{|y_j|}{E_j^{1/p}} \geq \frac{\alpha \cdot \|y\|_p}{\log n}\right] &= \mathbb{P}\left[\frac{|y_j|^p}{\alpha^p \cdot \|y\|_p^p} \cdot \log^p n \geq E_j\right] \\ &= \quad 1 - \exp\left(\frac{-|y_j|^p}{\alpha^p \cdot \|y\|_p^p} \cdot \log^p n\right) \\ &\leq \frac{|y_j|^p}{\alpha^p \cdot \|y\|_p^p} \cdot \log^p n.\end{aligned}$$

Since $\sum_j |y_j|^p = |y|_p^p$, the expected number of large coordinates is $O(\log^p n)$ (assuming $p$ and $\alpha$ are $O(1)$). Now by Markov's inequality, we get $|L| = O(\log^p n)$ with probability $\geq 9/10$. ■

Since the number of buckets $s = \tilde{O}(n^{1-\frac{2}{p}})$ is large, w.h.p. all $O(\log^p n)$ large coordinates belong to different buckets (recollect, $p = O(1)$). Conditioning on this event, we can assume that all other coordinates in a bucket are small, i.e., $K \leq \frac{\alpha \cdot \|y\|_p}{\log n}$. Applying Bernstein's bound for each bucket,

$$\mathbb{P}\left[(PDy)_i^{small} \geq \frac{\|y\|_p}{100}\right] \quad \leq \quad C\left(\exp(-\Theta(\log n)) + \exp\left(-c\frac{\log n}{100\alpha}\right)\right) \quad \leq \quad \frac{1}{n^2}.$$

Hence by union bound over all the buckets, we get that no bucket has the signed sum of small entries more than $\|y\|_p/100$.

Finally, to complete the proof, we condition on Eq. (1), on $|L| \leq O(\log^p n)$, on every bucket containing at most one coordinate from $L$, and on every bucket having signed sum of small $DY$ entries at most $\|y\|_p/100$. By union bound, all of these events simultaneously happen w.p. at least 2/3. In this case, we have

$$\|PDy\|_\infty \leq \|Dy\|_\infty + \frac{\|y\|_p}{100} \leq 10^{1/p}\|y\|_p + \frac{\|y\|_p}{100} \quad \& \quad \|PDy\|_\infty \geq \|Dy\|_\infty - \frac{\|y\|_p}{100} \geq \frac{\|y\|_p}{10^{1/p}} - \frac{\|y\|_p}{100}$$

and we can output $\|PDy\|_\infty$ to obtain an $O(1)$ approximation to $\|y\|_p$. The total space consumed is $\tilde{O}(s) = \tilde{O}(n^{1-\frac{2}{p}})$.

## 3 Heavy Hitter

Consider again a sequence of addition/deletion updates to a vector $y$ in the turnstile model. We wish to find all the "large" entries in $y$ in $o(n)$ space. We need to be careful in defining "large" because if we define it to be the largest entry, i.e. the $\ell_\infty$ norm, then from previous section we know that this requires $\Omega(n)$ space. In this section we output all *heavy hitter* coordinates $j$ that satisfy $|x_j| \geq \phi \cdot \|x\|_p$, and no coordinate $j$ with $|x_j| \leq (\phi - \epsilon) \cdot \|x\|_p$, where $p \in \{1, 2\}$ and $\phi, \epsilon$ are some constants. Most of these results appeared in the work of Charikar et al. [2].

### 3.1 $\ell_2$ vs $\ell_1$ Error

In this subsection we argue why finding all $\ell_2$-heavy hitters is more difficult than finding all $\ell_1$-heavy hitters. As an example, consider $x = \{\sqrt{n}, 1, 1, \ldots, 1\}$. For $\phi$ being a small constant, observe that there are no $\ell_1$-heavy hitters but the first coordinate is an $\ell_2$-heavy hitter.

**Claim 2.** If $j$ is a $\phi$ $\ell_1$-heavy hitter then it's also a $\phi^2$ $\ell_2$-heavy hitter.

*Proof.* By definition we know $|x_j| \geq \phi \cdot \|x\|_1$. Then, $x_j^2 \geq \phi^2 \cdot \|x\|_1^2 \geq \phi^2 \cdot \|x\|_2^2$. ■

Hence, we will mostly focus in designing randomized algorithms to achieve $\ell_2$ guarantees. Although this implies randomized algorithm for $\ell_1$ guarantees, we remark that $\ell_1$ heavy hitters can be also found deterministically, which we will discuss later [3]. Such deterministic guarantees are not possible for $\ell_2$-heavy hitters.

### 3.2 Intuitive Examples

Consider an example where you are promised that at the end of the stream only one coordinate $x_i = n$ and every other coordinate is in $\{0, 1\}$. How would you find $i$? Consider a simpler question, how would you find if $i$ is odd or even?

To answer the above latter question, we can maintain separately the sum of all the odd and all the even coordinates. Since $x_i = n$, the sum of coordinates containing $x_i$ will be larger (as the other sum is at most $n/2$) and we can find the parity of $i$. Note that this took only $O(\log n)$ space and is equivalent to finding the last bit of $i$ in its binary representation. One can now extend this idea to find each of the $\log n$ bits for $i$. To find the $j$'th bit, we keep track of the sum of all coordinates with $j$'th bit 0 and all coordinates with $j$'th bit 1. The set with the larger sum of coordinates gives the $j$'th bit. The total space is $O(\log^2 n)$.

Next, consider an example where you are promised that at the end of the stream only one coordinate $x_i = 100\sqrt{n \log\log n}$ and every other coordinate is in $\{0, 1\}$. How would you find $i$? The previous proof does not work as the now $x_i$ is not large enough to make the sum of other coordinates insignificant. We answer the simpler question of finding the parity of $i$ as the arguments similar to the previous example extend this to finding every bit of $i$. The crucial idea is that instead of maintaining the sum of all the odd and all the even coordinates, we multiply each coordinate with a random $\pm 1$, w.p. half each, and then maintain the sum of all the odd and all the even coordinates. We output the parity of $i$ corresponding to the sum of coordinates with the higher magnitude. The reason that this works is that the noise due to random plus-minus small coordinates will be $O(\sqrt{n})$ w.h.p., which is insignificant when added to $x_i$. The additional $\sqrt{\log\log n}$ factor is given so that we can take a union bound over all the $\log n$ coordinates.

### 3.3 Using CountSketch

The idea of using a random plus-minus sign in the above example indicates using the CountSketch matrix. Hence we randomly partition each coordinate into one of $B$ buckets, where bucket $j$ is $\{i : h(i) = j\}$, and maintain $c_j = \sum_{i:h(i)=j} x_i \cdot \sigma_i$ for every bucket $j$. Observe that $\sigma_i \cdot c_{h(i)}$ is an unbiased estimator for $x_i$. This is because $\mathbb{E}[\sigma_i \cdot c_{h(i)}] = \sigma_i \cdot \sum_{i':h(i)=h(i')} \sigma_{i'} x_{i'} = x_i$. Our algorithm

is to independently repeat this hashing scheme $O(\log n)$ times and for every $i$ output the median of the $\log n$ estimators for $x_i$. We argue that w.h.p. this method finds $x_i \pm \frac{\|x\|_2}{B}$, which suffices to output all heavy-hitters (we set $B$ as a function of $\phi, \epsilon$). The remaining proof bounds the noise in our estimators.

For bucket $i$, the noise is given by $\sigma_i \cdot \sum_{i' \neq i : h(i') = h(i)} \sigma_{i'} x_{i'}$. Its mean is zero and its variance is

$$\mathbb{E}\left[\left(\sigma_i \cdot \sum_{i' \neq i : h(i') = h(i)} \sigma_{i'} x_{i'}\right)^2\right] = \mathbb{E}\left[\left(\sum_{i' \neq i : h(i') = h(i)} \sigma_{i'} x_{i'}\right)^2\right] \leq \frac{\|x\|_2^2}{B}. \tag{2}$$

Hence, by Chebyshev's inequality, with constant probability the noise in the bucket is $O\left(\frac{\|x\|_2}{\sqrt{B}}\right)$ in magnitude. Since we estimate $x_i$ using $O(\log n)$ independent estimates, w.h.p., the median is $x_i \pm O\left(\frac{\|x\|_2}{\sqrt{B}}\right)$. Taking union bound over all the $n$ coordinates, we can simultaneously estimate all of them up to $\pm O\left(\frac{\|x\|_2}{\sqrt{B}}\right)$, which gives us all the heavy-hitter coordinates.

We next improve the above result slightly to argue that we can simultaneously estimate every coordinate $x_i$ to $\pm O\left(\frac{\|x_{-B/4}\|_2}{\sqrt{B}}\right)$, where $x_{-B/4}$ is vector $x$ with its largest (in magnitude) $B/4$ coordinates zeroed. This is useful for examples such as $x = (n^2, n, 1, 1, \ldots, 1)$. Although the first coordinate is primarily the only one that contributes in the $\ell_2$ norm, we would also like to output the second coordinate as a heavy hitter because it's much larger than all the remaining ones.

To prove this stronger result of $\pm O\left(\frac{\|x_{-B/4}\|_2}{\sqrt{B}}\right)$ error in the estimate of $x_i$, consider any coordinate $i$. We observe that with probability at least $3/4$ none of the top $B/4$ coordinates of $x$ land in the same bucket as $x_i$. In these cases only coordinates not in the top $B/4$ contribute in the noise variance for that bucket (see Eq. (2)). Hence the previous analysis gives an error of $\pm O\left(\frac{\|x_{-B/4}\|_2}{\sqrt{B}}\right)$.

## References


[1] Andoni, Alexandr. "High frequency moments via max-stability." Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017. APA.

[2] Charikar, Moses, Kevin Chen, and Martin Farach-Colton. "Finding frequent items in data streams." Automata, languages and programming (2002): 784-784.

[3] Nelson, Jelani, Huy L. Nguyen, and David P. Woodruff. "On deterministic sketching and streaming for sparse recovery and norm estimation." Linear Algebra and its Applications 441 (2014): 152-167.