CS 15-859: Algorithms for Big Data	Fall 2017

Lecture 7 - 2 — 10/19/2017

Prof. David Woodruff — *Scribe: Aditya Krishnan*

For context, these notes continue after the discussion on 1-Norm estimators, 2-Norm estimators and other canonical problems in the Turnstile streaming model.

1 p-Norm Estimators

The setting under consideration is the Turnstile streaming model where we have some vector $\vec{x}$ to which updates arrive as a stream. Our goal is to estimate $\|\vec{x}\|_p$ where the vector $\vec{x} \in \{-M, \ldots 0, \ldots, M\}^n$ for some $M \in \mathbb{N}$.

1.1 Estimating for small p where $0 < p < 2$

Using the techniques seen to estimate the 1-Norm in the Turnstile model, we can use similar techniques to estimate the p-Norm for $0 < p < 2$. Recall that 1-Norm estimation and 2-Norm estimation crucially used distributions (Cauchy and Gaussian respectively) in their sketches that were norm preserving. p-Norm preserving distributions are known as p-stable distributions. Specifically,

Definition. A real valued distribution $\mathcal{D}$ is called p**-stable** if for random variables $x, x_1, \ldots, x_n$ that are i.i.d from $\mathcal{D}$ and real values $a_1, \ldots, a_n \in \mathbb{R}$ we have the following property:

$$a_1 x_1 + \cdots + a_n x_n \sim \|a\|_p x$$

p-stable distributions exists for all $p \in (0, 2]$ but not for $p > 2$. Using p-stable distributions, one can perform p-Norm estimation in a similar fashion to 1-Norm estimation for $p \in (0, 2]$. It is also known that one can sample from p-stable distributions efficiently and that one can discretize them and construct sketching matrices consisting of p-stably distributed random variables with limited independence.

1.2 Estimating for $p > 2$

We will use a sketch and sample technique to estimate larger norms but since we don't have p-stable distributions, our sketch matrices need to rely on different distributions that capture the norm. Additionally, for $p > 2$, there is a $\Omega(n^{1-\frac{2}{p}})$ space complexity lower bound in the Turnstile streaming model [1]. We will now discuss a $\tilde{O}(n^{1-\frac{2}{p}})$ space algorithm for p-Norm estimation when $p > 2$.

First, we must perform a seemingly unmotivated digression into exponential random variables.

[1] Notice that for $p = \infty$ this automatically implies a linear lower bound on the space. We can't do anything smarter than store the entire n-dimensional vector $\vec{x}$

2 Exponential random variables and their stability

Definition. An exponential random variable $\boldsymbol{X}(\lambda)$ with $\lambda \in \mathbb{R}^+$ has the following PDF given by $f_{\boldsymbol{X}}$ and CDF given by $F_{\boldsymbol{X}}$

$$f_{\boldsymbol{X}(\lambda)}(x) = \lambda e^{-\lambda x} \qquad\qquad F_{\boldsymbol{X}(\lambda)}(x) = 1 - e^{-\lambda x}$$

Let us refer to $\boldsymbol{X}(1)$ as a standard exponential distribution, and from now on when we refer to an exponential without its parametrization (such as $\boldsymbol{X}$) we shall assume this is referring to a standard exponential.

Fact 1. For a scalar $t \geq 0$, the distribution corresponding to $t\boldsymbol{X}(\lambda)$ is the same as $\boldsymbol{X}(\frac{\lambda}{t})$

2.1 How stable are they?

Let $E_1, \ldots, E_n$ be independent exponential random variables, let $y \in \mathbb{R}^n$ be a vector and let $p \in \mathbb{R}$. We then investigate the distribution $q = \min\left(\frac{E_1}{|y_1|^p}, \ldots, \frac{E_n}{|y_n|^p}\right)$. To do so, let us calculate the CDF $F_q(x)$ of q.

$$F_q(x) = 1 - \mathbf{Pr}\left[q > x\right] = 1 - \mathbf{Pr}\left[\forall i,\ \frac{E_i}{|y_i|^p} \geq x\right]$$

By independence of $E_1, \ldots, E_n$

$$= 1 - \prod_{i=1}^{n} \mathbf{Pr}\left[\frac{E_i}{|y_i|^p} \geq x\right]$$

By Fact 1 we have

$$\begin{aligned} &= 1 - \prod_{i=1}^{n} \mathbf{Pr}\left[E_i\left(|y_i|^p\right) \geq x\right] = \prod_{i=1}^{n} e^{-x|y_i|^p} \\ &= 1 - e^{-x\|y\|_p^p} \end{aligned}$$

Hence q is distributed $E \cdot \frac{1}{\|y\|_p^p}$ where E is a standard exponential. We shall refer to the above property of the exponential as the *stability property.*

Let us now get back to estimating the p-Norm.

3 Sketch and Estimate

3.1 What sketch do we use and why?

Our sketch is of the form $P \cdot D$ where P is a $s \times n$ CountSketch matrix and D is a $n \times n$ diagonal matrix given by $\mathsf{diag}\left(\frac{1}{E_1^{1/p}}, \ldots, \frac{1}{E_n^{1/p}}\right)$

What does $\|Dy\|_\infty^p$ look like for any vector y?

$$\|Dy\|_\infty^p = \max_i \frac{|y_i|^p}{E_i} = \frac{1}{\min_i \frac{E_i}{|y_i|^p}}$$

By the stability property of the exponential

$$= \frac{1}{\frac{E}{\|y\|_p^p}} = \frac{\|y\|_p^p}{E}$$

What is the probability that $E \sim \mathsf{Exp}(1)$ lies within some fixed range $[0.1, 10]$?

$$\mathbf{Pr}\left[E \in [0.1, 10]\right] = 1 - e^{-10} - (1 - e^{-0.1}) = e^{-0.1} - e^{-10} > \frac{4}{5}$$

Now we know that with probability at least $\frac{4}{5}$, we have that $\frac{\|y\|_p^p}{10} \leq \|Dy\|_\infty^p \leq 10\,\|y\|_p^p$. Hence we know that $\|Dy\|_\infty^p$ is a good estimator for the p-Norm. But Dy is an n-dimensional vector! Which is why we sketch.

3.2 What is sketching doing?

Recall that the CountSketch matrix P is a $s \times n$ matrix where each column has a ± 1 (each with equal probability) in exactly one coordinate chosen uniformly at random amongst the s coordinates and all other entries are 0. The rows of P can interpreted as hash buckets where each row takes a signed sum of the entries corresponding to the non-zero values in the row. [2]

P can thus be described as a pair of functions $h : [n] \to [s]$ and $\sigma : [n] \to \{+1, -1\}$. Here $h(i)$ describes the coordinate of the non-zero value in the i^{th} column and $\sigma(i)$ describes the sign of the non-zero value in the i^{th} column. For the sake of convenience we will assume that h, σ are truly random.

3.3 Achieving $\|PDy\|_\infty \approx \|Dy\|_\infty$

To achieve this with good probability we want two things:

1. In each bucket i not containing the coordinate j for which $\left|(Dy)_j\right| = \|Dy\|_\infty$, we want the signed sum to be small. I.e we want $\left|(PDy)_i\right| \leq \frac{\|y\|_p}{100}$
2. In the buckets that do contain the coordinate j for which $\left|(Dy)_j\right| = \|Dy\|_\infty$, we want the noise to be small. I.e we want $\left|(PDy)_i - \|Dy\|_\infty\right| \leq \frac{\|y\|_p}{100}$

[2] Alternatively one can view it as Dy taking a linear combination of the columns and each entry in y getting mapped uniformly at random to one of the s coordinates. We take a signed sum of the entries that get mapped to the same coordinate.

Let us set up some notation before we start analyzing $\|PDy\|_\infty$. Let $\delta(X) = 1$ if some event X occurs and 0 otherwise. What does the value in the i^{th} hash bucket, i.e $|(PDy)_i|$, look like?

$$(PDy)_i = \sum_{j=1}^{n} \delta(h(j) = i) \cdot \sigma(j) \cdot |(Dy)_j|$$

Notice that $\mathop{\mathbb{E}}_P [(PDy)_i] = 0$ since for every $j \in [n]$ for which $h(j) = i$, we will have that with equal probability the value is $+1$ or -1 and hence in expectation the value is 0. Notice that this expectation is taken over the randomness of P. Ok, now we know that $(PDy)_i$ is mean 0 but how concentrated is it? I.e what is its variance?

$$\begin{aligned}
\mathop{\mathbb{E}}_P \left[(PDy)_i^2\right] &= \sum_{j,k} \mathop{\mathbb{E}}_P [\delta(h(j) = i) \cdot \delta(h(k) = i) \cdot \sigma(j)\sigma(k)] \cdot |(Dy)_j| \cdot |(Dy)_k| \\
&= \sum_{j \neq k} \mathbb{E}[\delta(h(j) = i) \cdot \sigma(j)] \cdot \mathbb{E}[\delta(h(k) = i) \cdot \sigma(k)] \cdot |(Dy)_j| \cdot |(Dy)_k| \\
&+ \sum_{j=1}^{s} \mathbb{E}[\delta(h(j) = i)^2 \cdot \sigma(j)^2](Dy)_j^2 \\
&= \sum_{j=1}^{s} \mathbb{E}[\delta(h(j) = i)^2 \cdot \sigma(j)^2](Dy)_j^2 = \frac{1}{s}\|Dy\|_2^2
\end{aligned}$$

$$\mathop{\mathbb{E}}_D \left[\|Dy\|_2^2\right] = \sum_{i=1}^{n} y_i^2 \cdot \mathop{\mathbb{E}}_D \left[D_{i,i}^2\right]$$

$$\mathop{\mathbb{E}}_D \left[D_{i,i}^2\right] = \int_{t \geq 0} t^{\frac{-2}{p}} e^{-t} dt = \int_0^1 t^{\frac{-2}{p}} e^{-t} dt + \int_{t \geq 1} t^{\frac{-2}{p}} e^{-t} dt$$

For $t \in [0, 1]$, $e^{-t} \leq 1$ and since $p > 2$, for $t \geq 1$, $t^{-2/p} \leq 1$. Hence we have

$$\begin{aligned}
&\leq \int_0^1 t^{\frac{-2}{p}} dt + \int_{t \geq 1} e^{-t} dt \\
&= \frac{1}{1 - \frac{2}{p}} \cdot t^{1 - \frac{2}{p}} \Big|_0^1 - e^{-t} \Big|_1^\infty = O(1)
\end{aligned}$$

Hence we have that $\mathop{\mathbb{E}}_{P,D} \left[(PDy)_i^2\right] = \frac{1}{s}\|y\|_2^2$. But, we want to relate the variance of $(PDy)_i$ to the p-Norm of y not the 2-Norm. To do this, we can use the generalized version of the Cauchy-Schwarz inequality, known as Hölder's inequality.

Fact 2. A corollary of Hölder's inequality is that if x, y are vectors and $p, q \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$ we have that

$$\langle x, y \rangle \leq \|x\|_p \cdot \|y\|_p$$

We then write $\|y\|_2^2$ in terms of the p-Norm using Hölder's inequality.

$$\|y\|_2^2 = \sum_{j=1}^{n} y_j^2 \cdot 1 \le \left(\sum_{j=1}^{n} (y_j^2)^{p/2}\right)^{2/p} \cdot \left(\sum_{j=1}^{n} 1^q\right)^{1/q}$$

The last inequality is applied using Fact 2. For the inequality to hold, it must be that $q = \frac{1}{1-\frac{2}{p}}$. Substituting the value of q into the above expression gives us that $\|y\|_2^2 \le n^{1-2/p}\,\|y\|_p^2$. Putting all this together, we have finally found the variance of $(PDy)_i$ in terms of the p-Norm of y. We finally have that

$$\mathop{\mathbb{E}}_{P,D}\left[(PDy)_i^2\right] \le \frac{1}{s} \cdot n^{1-2/p}\,\|y\|_p^2$$

Recall we wanted to show two properties of $\|PDy\|_\infty$. To show these, we will need to show concentration of $\|PDy\|_\infty$ for which we have mean and variance. To show this we look at Bernstein's bound.

Bernstein's Bound: Suppose $R_1, \ldots, R_n$ are independent random variables and for all $j, |R_j| \le K$ and $\mathbf{Var}\left[\sum_j R_j\right] = \sigma^2$ then there are constants c, C so that for all $t > 0$

$$\mathbf{Pr}\left[\left|\sum_{j=1}^{n} R_j - \mathbb{E}[\sum_{j=1}^{n} R_j]\right|\right] \le C\left(e^{\frac{-ct^2}{\sigma^2}} + e^{\frac{-ct}{K}}\right)$$

We will apply Bernstein's bound to $(PDy)_i$ by setting each $R_j := \delta(h(i) = j) \cdot \sigma(j) \cdot |(Dy)_j|$. Setting $t = \frac{\|y\|_p}{100}$ and $s = \theta(n^{1-2/p}\log(n))$ gives us that

$$e^{\frac{-ct^2}{\sigma^2}} = \theta\left(\frac{1}{n^2}\right)$$

But what about the term $K = \max_j |R_j|$, notice that can be as high as $\|Dy\|_\infty$. In fact, the bucket i which contains the $\|Dy\|_\infty$ will have such a large K. Hence applying Bernstein's bound naively will give us a poor bound. We must condition on which bucket the large value ($\|Dy\|_\infty$) sits in and then apply Bernstein's bound to show that these entries are small. Additionally we need to argue that for the buckets in which the large entry ($\|Dy\|_\infty$) sits, the rest of the noise is small. These arguments will be made in the upcoming lecture.