CS 15-859: Algorithms for Big Data Fall 2020

Lecture 8 Part 1 — Oct 29th

Prof. David Woodruff *Scribe: Lichen Zhang*

1 Recap: Turnstile Streaming Model

Recall the Turnstile Streaming Model we started to study last time:

- There is an underlying n-dimensional vector x initialized to 0^n.
- There is a long stream of updates $x_i \leftarrow x_i + \Delta_j$, for $\Delta_j \in \{-M, -M+1, \ldots, 0, 1, \ldots, M-1, M\}^n$, where $M \in \text{poly}(n)$, here x_i is the i^{th} coordinate of vector x, and Δ_j indicates the j^{th} update.
- Throughout the stream, x is promised to be in $\{-M, -M+1, \ldots, M-1, M\}^n$, so entries of x can be stored with only $O(\log n)$ bits.
- Output an approximation to a function $f(x)$ with high probability over coin tosses, which means we can use randomized algorithm.
- The goal is to use as little space in bits as possible.

There are large number of real-life applications of data streaming model. Today, we will focus on one specific problem.

2 Estimating Norms

Suppose we want to estimate the p-norm of x: $\|x\|_p = \sqrt[p]{\sum_{i=1}^n |x_i|^p}$, or $\|x\|_p^p = \sum_{i=1}^n |x_i|^p$. Our goal is to find an estimation z such that

$$(1-\epsilon)\|x\|_p^p \leq z \leq (1+\epsilon)\|x\|_p^p$$

with probability at least $9/10$. We make some remarks for specific p and their applications:

- $p = 1$ corresponds to total variation distance between distributions. For the sake of simplicity, consider two discrete probability distribution $p = (p_1, \ldots, p_n), q = (q_1, \ldots, q_n)$, then total variation distance between p and q is defined as

$$d_{\text{tv}} = \frac{1}{2}|p - q|$$

- $p = 2$ is the familiar Euclidean norm.
- $p = \infty$ is useful for anomaly detection.

2.1 Estimating Euclidean Norm

The goal is to find z such that

$$(1-\epsilon)\|x\|_2^2 \leq z \leq (1+\epsilon)\|x\|_2^2$$

Notice we can simply use a CountSketch matrix S with $O\left(\frac{1}{\epsilon^2}\right)$ rows, it satisfies JL-moment property so we will have the above property with constant probability, if we set $z = \|Sx\|_2^2$. The algorithm is as follows:

- Initially, set $x = 0^n$.
- For each update, update the vector $Sx \leftarrow Sx + \Delta_j S_{*,j}$, where $S_{*,j}$ is the j^{th} column of S.
- At the end of stream, output $\|Sx\|_2^2$.

How to store S? Recall S can be viewed as two hash functions $h : [n] \to \left[\frac{1}{\epsilon^2}\right]$ and $\sigma : [n] \to \{-1, 1\}$, and h requires pairwise independence and σ requires 4-wise independence, so only $O(\log n)$ bits suffice to store S. Thus, the total space complexity is $O\left(\frac{1}{\epsilon^2}\log n\right)$.

2.2 Estimating 1-Norm

Once again, we shall find z such that

$$(1-\epsilon)\|x\|_1 \leq z \leq (1+\epsilon)\|x\|_1$$

A natural thing to try is to use a random Cauchy matrix S, in [1], it is shown that one can store S using $O\left(\frac{1}{\epsilon}\log n\right)$ bits, and the update rule is exactly the same as in ℓ_2 case. Using $\frac{1}{\epsilon^2}$ rows, we can finally arrive at a space complexity of $O\left(\frac{1}{\epsilon^2}\log n\right)$ bits. However, recall that Cauchy random variable has undefined expectation and infinite variance, which means it does not have any concentration, so $\|Sx\|_1$ might not give rise to a good approximation of $\|x\|_1$, in fact, as we have shown in the ℓ_1 sketching case, it gives an $O(d \log d)$ approximation, where here we have $d = 1$, so an $O(1)$ approximation, not the $(1 \pm \epsilon)$ we want. We shall use a different strategy to achieve this task.

We remind readers the probability density function for a Cauchy random variable C is

$$f(x) = \frac{1}{\pi(1+x^2)}$$

If we only consider $|C|$, the half-Cauchy, then its pdf becomes

$$f(x) = \frac{2}{\pi(1+x^2)}$$

The cumulative density function is

$$F(z) = \int_0^z f(x)\ dx = \frac{2}{\pi}\arctan(z)$$

The idea is to consider the *median* of random variable $|C|$, recall median of $|C|$ is the value where its cdf outputs $1/2$. If $F(z) = \frac{1}{2}$, then $\arctan(z) = \frac{\pi}{4}$, apply tan to both sides we get $z = \tan(\frac{\pi}{4}) = 1$, so $F(1) = 1/2$, and median of $|C|$ is 1.

We will prove the following key lemma:

Lemma 1. *Let $r = \frac{\log \frac{1}{\delta}}{\epsilon^2}$, if we take r independent samples $X_1, X_2, \ldots, X_r$ from a distribution $\mathcal{F}$ and let $X = median\{X_1, X_2, \ldots, X_r\}$, then with probability $1 - \delta$, $F(X) \in [1/2 - \epsilon, 1/2 + \epsilon]$.*

Proof. Consider the indicator random variable Z_i, where $Z_i = 1$ if the sample X_i has the property that $F(X_i) \leq 1/2 - \epsilon$, so $\mathbf{Pr}[Z_i = 1] = \frac{1}{2} - \epsilon$, let $Z = \sum_{i=1}^{r} Z_i$, then

$$\begin{aligned}\mathbb{E}[Z] &= \mathbb{E}[\sum_{i=1}^{r} Z_i] \\ &= \sum_{i=1}^{r} \mathbf{Pr}[Z_i = 1] \\ &= \frac{r}{2} - r\epsilon\end{aligned}$$

One thing to observe is if the median $X \in [0, 1/2 - \epsilon]$, then at least half of Z_i's are 1, so $Z \geq r/2$. Using Chernoff bound, we get

$$\mathbf{Pr}[Z - \mathbb{E}[Z] \geq r\epsilon] \leq \exp\left(\frac{-\epsilon^2 \mathbb{E}[Z]}{3}\right)$$

Recall $r = \frac{\log \frac{1}{\delta}}{\epsilon^2}$, so

$$\begin{aligned}\mathbb{E}[Z] &= \frac{r}{2} - r\epsilon \\ &= \Theta(r) \\ &= \frac{c \cdot \log \frac{1}{\delta}}{\epsilon^2}\end{aligned}$$

So Chernoff will give a $1 - \delta$ probability, $X \notin [0, 1/2 - \epsilon] \cup [1/2 + \epsilon, 1]$, which is equivalent to $X \in [1/2 - \epsilon, 1/2 + \epsilon]$. ■

Now we know that using median of independent samples, we have its cdf being close to the cdf of true median. The question is whether the empirical median itself close to the true median? To answer this question, we shall consider the inverse cdf of $|C|$:

$$F^{-1}(X) = \tan\left(\frac{\pi}{2}\right)$$

So let's examine two endpoints for $F(X)$:

$$\begin{aligned} F^{-1}\left(\frac{1}{2}-\epsilon\right) &= \tan\left(\frac{\pi}{4}-\frac{\epsilon\pi}{2}\right) \\ &\geq 1-\pi\epsilon \\ &\geq 1-4\epsilon \\ F^{-1}\left(\frac{1}{2}+\epsilon\right) &= \tan\left(\frac{\pi}{4}+\frac{\epsilon\pi}{2}\right) \\ &\leq 1+\pi\epsilon \\ &\leq 1+4\epsilon \end{aligned}$$

The bounds are obtained through Taylor expansion of $g(\epsilon) = \tan\left(\frac{\pi}{4}-\frac{\epsilon\pi}{2}\right)$ around $\epsilon = 0$. The last bound is obtained similarly.

Thus, the empirical median $X \in [1-4\epsilon, 1+4\epsilon]$. We remark that if the distribution is very flat, then we cannot obtain such a good approximation to median by sampling. The algorithm builds upon the property of median of Cauchy:

- Draw a matrix $S \in \mathbb{R}^{r\times n}$ with each entry independent Cauchy random variables and $r = O(1/\epsilon^2)$.
- Initially, set $x = 0^n$.
- For each update, update $Sx \leftarrow Sx + \Delta_j S_{*,j}$.
- At the end of stream, output median of entries of Sx.

Why does this work? Notice $S = \begin{bmatrix} — & S_1 & — \\ — & S_2 & — \\ & \vdots & \\ — & S_r & — \end{bmatrix}$ and $Sx = \begin{bmatrix} \langle S_1, x\rangle \\ \langle S_2, x\rangle \\ \vdots \\ \langle S_r, x\rangle \end{bmatrix}$, by 1-stability of Cauchy, we can further write Sx as

$$Sx = \begin{bmatrix} \|x\|_1|C_1| \\ \|x\|_1|C_2| \\ \vdots \\ \|x\|_1|C_r| \end{bmatrix}$$

where $|C_1|, |C_2|, \ldots, |C_r|$ are half-Cauchys. The median of these r Cauchys is a $(1\pm 4\epsilon)$-approximation to the true median, which is 1, so we know median z of Sx has

$$(1-4\epsilon)\|x\|_1 \leq z \leq (1+4\epsilon)\|x\|_1$$

2.3 Estimating p-Norm: $0 < p < 2$

We can even achieve $O\left(\frac{1}{\epsilon^2}\log n\right)$ bits of space for p-norm estimation for any $0 < p < 2$, it uses the fact that p-stable distribution exists for $0 < p < 2$, i.e.,

$$a_1 \cdot W_1 + a_2 \cdot W_2 + \ldots + a_n \cdots W_n = \|a\|_p \cdot W$$

However, there's no closed-form expression for such probability density function; instead, we can generate a sample relatively easy:

Theorem 1. *If we pick* $\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ *and* $r \in [0,1]$ *uniformly random, then*

$$\frac{\sin(p\theta)}{\cos^{1/p}(\theta)} \left(\frac{\cos(\theta(1-p))}{\ln \frac{1}{r}}\right)^{\frac{1-p}{p}}$$

is a sample from a p-stable distribution.

We remark that if $p < 1$, $(\sum_{i=1}^{n} |x_i|^p)^{1/p}$ is even not a norm, since it does not satisfy triangle inequality. Also, though in Theorem 1 we require uniform randomness, it is possible to discretize them and store a sketching matrix of samples from the p-stable distribution using limited independence.

2.4 Estimating p-Norm: $p > 2$

Unfortunately, if $p > 2$, there does not exist p-stable distribution. Even more, we will later show that $\Omega\left(n^{1-2/p}\right)$ bits of space is required to approximate such a p-norm, up to a constant factor with constant probability. It is possible to achieve an $\tilde{O}\left(n^{1-2/p}\right)(O\left(n^{1-2/p}\log^2 n\right))$ bits of space with exponential random variables, and we will focus on constant approximation on parameter ϵ.

The sketch matrix will be PD, where P is a CountSketch matrix, and D is a diagonal matrix in the form of

$$D = \begin{bmatrix} 1/E_1^{1/p} & & & \\ & 1/E_2^{1/p} & & \\ & & \ddots & \\ & & & 1/E_n^{1/p} \end{bmatrix}$$

i.e., each entry is the reciprocal of an independent standard exponential variable, raise to $1/p$ power.

We first introduce or refresh memories of readers, to the exponential random variable.

Definition. An *exponential random variable* is parametrized by a parameter λ, denoted as $X \sim Exp(\lambda)$. It has probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

It has cumulative density function

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Some nice property of exponential random variable: let $E \sim Exp(\lambda)$ and $t \geq 0$, then $t \cdot E \sim Exp\left(\frac{\lambda}{t}\right)$, since

$$\begin{aligned} \mathbf{Pr}[t \cdot E \leq x] &\leq \mathbf{Pr}[E \leq \frac{x}{t}] \\ &= 1 - e^{-\frac{\lambda}{t}x} \end{aligned}$$

How does exponential random variables preserve some form of "stability"? Consider independent standard ($\lambda = 1$) exponential random variables $E_1, \ldots, E_n$m and scalars $|y_1|, \ldots, |y_n|$, let $q = \min\left(\frac{E_1}{|y_1|^p}, \ldots, \frac{E_n}{|y_n|^p}\right)$, we claim q is again an exponential random variable, consider its cdf:

$$\begin{aligned}
\mathbf{Pr}[q > x] &= \mathbf{Pr}[\forall i, \frac{E_i}{|y_i|^p} > x] \\
&= \prod_{i=1}^{n} \mathbf{Pr}[\frac{E_i}{|y_i|^p} > x] && \text{since all } E_i\text{'s are independent} \\
&= \prod_{i=1}^{n} e^{-x|y_i|^p} && \frac{E_i}{|y_i|^p} \sim Exp\left(|y_i|^p\right) \\
&= e^{-x\sum_{i=1}^{n}|y_i|^p} \\
&= e^{-x\|y\|_p^p}
\end{aligned}$$

Thus, $q \sim Exp\left(\|y\|_p^p\right)$ and $q \sim \frac{E}{\|y\|_p^p}$, for $E \sim Exp\left(1\right)$. Next, we shall relate $\|Dy\|_\infty$ to $\|y\|_p^p$:

$$\begin{aligned}
\|Dy\|_\infty &= \max_i \frac{|y_i|^p}{E_i} \\
&= \frac{1}{\min_i \frac{E_i}{|y_i|^p}} \\
&\sim \frac{1}{E \cdot \frac{1}{\|y\|_p^p}} \\
&= \frac{\|y\|_p^p}{E}
\end{aligned}$$

As a standard exponential random variable, it has high probability to fall in a fairly small range:

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

Therefore, we know $\|Dy\|_\infty \in \left[\frac{\|y\|_p}{10^{1/p}}, 10^{1/p}\|y\|_p\right]$ with probability at least $\frac{4}{5}$, so $\|Dy\|_\infty$ is a good approximation for $\|y\|_p$, the problem is $D \in \mathbb{R}^{n\times n}$, so there's no dimensionality reduction being performed.

So we shall apply a CountSketch matrix P for $\tilde{O}\left(n^{1-2/p}\right)$ number of rows, the intuition for P is as follows: P hashes coordinates of Dy into buckets and takes a signed sum of entries, we expect everything to cancel out, so that $\|PDy\|_\infty \approx \|Dy\|_\infty$. In next part of lecture, we will use Bernstein's bound and a novel technique to divide entries into large and small, to achieve this guarantee.

References

[1] Daniel M. Kane, Jelani Nelson, and David P. Woodruff. *On the Exact Space Complexity of Sketching and Streaming Small Norms*, pages 1161–1178. URL: `https://epubs.siam.`

org/doi/abs/10.1137/1.9781611973075.93, arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611973075.93, doi:10.1137/1.9781611973075.93.