CS 15-859: Algorithms for Big Data Fall 2020

Lecture 8 Part 1 — November 4

Prof. David Woodruff *Scribe: Kunming Jiang*

1 Recap: Turnstile Streaming Model

Given a size-n vector x, we want to approximate the result of the following procedure using only $O(\log n)$ bits of space:

- Initialize x to 0^n.
- Sequentially update x using the formula

$$x_i \leftarrow x_i + \Delta_j, \text{ where } \Delta_j \in \{-M, \ldots, M\}, M \in \text{poly}(n)$$

In the previous class, we have shown:

1. To test if $x = 0^n$, we can use CountSketch, which can be hashed using $O(\log n)$ bits.
2. Any deterministic algorithm would require $\Omega(n \log n)$ bits of space to test if $x = 0^n$.
3. We can deterministically extract a k-sparse vector using $O(k \log n)$ bits of space.

2 Estimating Norm in Streaming Model

In this section, we want to find z such that

$$(1-\epsilon)|x|_p^p \le z \le (1+\epsilon)|x|_p^p \text{ with prob.} > 9/10$$

where $|x|_p^p = \sum_{i=1}^{n} |x_i|^p$.

We do so by dividing p into different cases.

2.1 Euclidean Norm: $p = 2$

To find z such that $(1-\epsilon)|x|_2^2 \le z \le (1+\epsilon)|x|_2^2$, we do the following:

1. Sample a CountSketch matrix S with $1/\epsilon^2$ rows.
2. For each update $x_i \leftarrow x_i + \Delta_j$, do $Sx \leftarrow Sx + \Delta_j S_{*,i}$.
3. Output $|Sx|_2^2$ at the end.

With probability $\geq 9/10$, we have $|Sx|_2^2 = (1+\epsilon)|x|_2^2$.

Complexity: $1/\epsilon^2$ words. each with $O(\log n)$ bits. Therefore the overall data structure can be stored using $O((\log n)/\epsilon^2)$ bits.

2.2 1-Norm: $p = 1$

To find z such that $(1-\epsilon)|x|_1 \leq z \leq (1+\epsilon)|x|_1$, we do the following:

1. Sample a Cauchy matrix S with $1/\epsilon^2$ rows.
2. For each update $x_i \leftarrow x_i + \Delta_j$, do $Sx \leftarrow Sx + \Delta_j S_{*,i}$.
3. Output *median* of $|(Sx)_1|, |(Sx)_2|, \ldots, |(Sx)_{1/\epsilon^2}|$.

Note: we cannot output $|Sx|_1$, since $|Sx|_1 \leq O(d \log d)|x|_1 = c \cdot |x|_1$, where $d = 1$ and c is some constant. This bound is not tight enough for us.

To prove that median works, first recall the definition of Cauchy Distribution, $\mathcal{C}$:

- PDF of $|\mathcal{C}|$ is $f(x) = \dfrac{2}{\pi(1+x^2)}$.
- CDF of $|\mathcal{C}|$ is $F(x) = \displaystyle\int_0^z f(x)dx = \frac{2}{\pi}\arctan(z)$.

  $\tan(\pi/4) = 1 \Longrightarrow F(1) = 1/2$, so $\text{median}(|C|) = 1$.

Now we claim that, given any $r = \log\frac{1}{\delta}/\epsilon^2$ independent samples $X_1, \ldots X_r$ from F, if $X = \text{median}_i X_i$, then $F(X) \in [\frac{1}{2}-\epsilon, \frac{1}{2}+\epsilon]$ has probability $1-\delta$.

Proof. First we ask: how many X_i would satisfy $F(X_i) \leq \frac{1}{2} - \epsilon$?

We call this quantity C_l, and note that $C_l = \sum_i Y_i$, where $Y_i \in \{0,1\}$.

Since $\mathbb{E}[C_l] = r \cdot (\frac{1}{2}-\epsilon)$, by Chernoff bound, $\mathbf{Pr}[C_l \geq \frac{r}{2}] \leq \exp(-\theta(\epsilon^2 r)) \leq \frac{\delta}{2}$.

Similarly, let C_h denote the amount of X_i such that $F(X_i) \geq \frac{1}{2}+\epsilon$, $\mathbf{Pr}[C_h \geq \frac{r}{2}] \leq \frac{\delta}{2}$.

By union bound $\mathbf{Pr}[C_l \geq \frac{r}{2} \vee C_h \geq \frac{r}{2}] \leq \delta$, so we have $\leq \delta$ probability that the median lies outside of $[\frac{1}{2}-\epsilon, \frac{1}{2}+\epsilon]$, so $F(X) \in [\frac{1}{2}-\epsilon, \frac{1}{2}+\epsilon]$ has prob. $1-\delta$. ■

However, we still have to prove that variables around the median is close to the actual median.

This is done by showing $F^{-1}(X) = \tan(\frac{X\pi}{2}) \in [1-4\epsilon, 1+4\epsilon]$, so small change in median $\Longrightarrow$ small change in variable. Thus,

$$\begin{aligned}
\text{median}(|S_1 \cdot x|, |S_2 \cdot x|, \ldots) &= \text{median}(|x|_1 \cdot |\mathcal{C}_1|, \cdots, |x|_1 \cdot |\mathcal{C}_r|) \\
&= |x|_1 \cdot \text{median}(|\mathcal{C}_1|, \ldots, |\mathcal{C}_r|) \\
&= |x|_1 \cdot (1 + \Theta(\epsilon))
\end{aligned}$$

2.3 p-Norm esttimation, $0 < p < 2$

This can be achieved in $1/\epsilon^2$ words of space by using a p-stable distribution.

Definition. A distribution Π is *p-stable* if given any fixed vector $V \in \mathbb{R}^n$ and independent samples $X_1, \ldots, X_n \sim \Pi$,

$$\sum_i X_i V_i \equiv X \cdot |V|_p, \text{ where } X \sim \Pi$$

Note that we need to discretize and store a sketching matrix of samples from the p-stable distribution using limited independence. One way is to round every entry to the nearest integer multiple of $1/n^{100}$.

2.4 p-Norm estimation, $p > 2$

For $p > 2$, p-stable distribution doesn't exist, and we need $\Omega(n^{1-2/p})$ bits to approximate.

We can use exponential random variables (R.V.), $E(\lambda)$, with properties listed below:

- PDF: $f(x) = \lambda e^{-\lambda x}$ if $x \geq 0$, 0 otherwise.
- CDF: $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$.
- For any $t \cdot E$ where $t \geq 0$ is a scalar, the CDF is $F(x) = 1 - e^{-\frac{\lambda}{t}x}$.
- *Min Stable*: for independent $E_1, \ldots, E_n$ and scalar $|y_1|, \ldots, |y_n|$,
  let $q = \min(\frac{E_1}{|y_1|^p}, \ldots, \frac{E_n}{|y_n|^p})$, then $\mathbf{Pr}[q > x] = \mathbf{Pr}[\forall\ i, \frac{E_i}{|y_i|^p} \geq x] = \Pi_i e^{-x|y_i|^p} = e^{-x|y|_p^p}$.
  So q is an exponential random variable with $\lambda = |y|_p^p$.

Our sketch is $P \cdot D$, where P is an $O(n^{1-2/p}) \times n$ CountSketch matrix and

$$D = \begin{bmatrix} 1/E_1^{1/p} & & \\ & 1/E_2^{1/p} & \\ & & \ddots \end{bmatrix} \text{ is diagonal and } E_i \text{ are standard exponential R.V.s.}$$

Then, for arbitrary y.

$$|Dy|_\infty^p = \max_i(\frac{|y_i|^p}{E_i}) = \frac{1}{\min_i \frac{E_i}{|y_i|^p}} \equiv \frac{1}{E \cdot \frac{1}{|y|_p^p}} = \frac{|y|_p^p}{E}$$

Now E is a standard exponential R.V., so $\mathbf{Pr}[E \in (\frac{1}{10}, 10)] \geq \frac{4}{5}$.

$\Longrightarrow |Dy|_\infty \in [\frac{1}{10^{1/p}}|y|_p, 10^{1/p}|y|_p]$ with prob. $\geq 4/5$.

However, Dy is n-dimensional, which is too large.

In the next part of the lecture, we will argue that if does not distort $|Dy|_\infty$ too much, so $|PDy|_\infty$ is close to $|Dy|_\infty$.