CS 15-851: Algorithms for Big Data Spring 2025

Lecture 8 Part 1 — March 13

Prof. David Woodruff *Scribe: Ron Sarma*

1 Recap

1.1 Turnstile Streaming Model

Consider a high-dimensional vector $\mathbf{x} \in \mathbb{R}^n$ that

- initializes as $\mathbf{x} = 0^n$
- follows the update rule $x_i \leftarrow x_i + \Delta_j$ where $\Delta_j \in \{-M, \cdots, M\}$ is the change in x and $M \leq \text{poly}(n)$

x would naively require poly(n) memory (or poly($\log n$) bits) to store, we want to figure out the minimum space required to store and generate the output of x with high probability (9/10).

We can test if $x = 0^n$ by sketching x using a CountSketch matrix (S) with ($1/\epsilon^2$) rows.

$$|Sx|_2^2 = (1 \pm \epsilon)|x|_2^2$$

Since, S preserves the norm of x with high probability, we can update $Sx \leftarrow Sx + S_i\Delta$ and store its hash and sign functions using $O(\log n)$ bits instead of updating $x \leftarrow x + \Delta$ and storing it which requires poly(n) bits.

Any deterministic algorithm would require $\Omega(n \log n)$ bits to store x.

2 Recovering a k-sparse vector

Common practices involve sparse vectors, such as vectors monitoring differences in network traffic.

Let $\mathbf{x} \in \text{R}^n$ be a k-sparse vector with k non-zero entries and let k be small. We can identify the indices and values of the k non-zero entries of $\mathbf{x}$ using $k$polylog(n) bits of space deterministically.

Proof. Consider an $s \times n$ matrix A, such that it has $2k$ linearly independent columns and $s \geq 2k$. We want to show that A will be able to recover the k non-zero entries of x.

Consider two vectors x and y each with k non-zero entries. Suppose $Ax = Ay$, then $A(x = -y) = 0$. Now, since A has $2k$ linearly independent columns and $(x - y)$ has at most $2k$ non-zero entries, $x = y$ must hold. This also implies that x can be recovered deterministically given a fixed A.

We have only discussed two properties of A so far - the size $(s \times n)$ and it having $2k$ linearly independent columns such that $s \geq 2k$. But how does A look like? Turns out there are different kinds of matrices that qualify as A, even random matrices. But random matrices occupy large space. The Vandermonde matrix is a common option for this.

The Vandermonde matrix is a fixed matrix $A \in \mathrm{R}^{2k\times n}$ where $s = 2k$ and $A_{i,j} = j^{i-1}$

$$\begin{bmatrix} 1 & 1 & 1 & \cdots \\ 1 & 2 & 3 & \cdots \\ 1 & 4 & 9 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

We can show A has $2k$ linearly independent columns by calculating the determinant of a $2k \times 2k$ sub-matrix of A and show that it is non-zero.

Since the entries of A grows exponentially with n, it would require $O(n)$ bits to store each entry to be maintained in the stream. That is a problem.

An efficient way would be to store $(A \cdot x \mod p)$ for a large enough prime $p \in$ poly (n). $(A \cdot x \mod p)$ would have $2k$ linearly independent columns because any of its $2k \times 2k$ sub-matrix will have a non-zero determinant. And, it can be stored with $O(k \log n)$ bits of space.

3 Estimating Norms in a Streaming Model

A p-Norm of a vector $\mathbf{x} \in \mathrm{R}^n$ by definition is $|\mathbf{x}|_p^p = \sum_{i=1}^n |x_i|^p$

We want an estimator of the p norm z of x in a streaming model, such that, with high probability 9/10,

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

For the rest of the lecture, $p = 1, p = 2$ and $0 < p < 2$ are discussed in detail, wrapping up with $p > 2$.

3.1 2-Norm or Euclidean Norm Estimation: $p = 2$

Find z such that $(1-\epsilon)|\mathbf{x}|_2^2 \leq z \leq (1+\epsilon)|\mathbf{x}|_2^2$ with high probability 9/10.

Instead of updating and storing $x_i \leftarrow x_i + \Delta_j$, we can sketch $\mathbf{x}$ and store its embedded norm, which will have more efficient space complexity. For this,

- Sample using a CountSketch matrix S having $1/\epsilon^2$ rows.
- Update the sketched vector $Sx \leftarrow +\Delta S_{*i}$ for every i^{th} update in x_i.
- Output $|S\mathbf{x}|_2^2$. Since S is a subspace embedding, $|Sx|_2^2 = (1 \pm \epsilon)|x|_2^2$, and it can be stored with space complexity $1/\epsilon^2$ words, each word having $O(\log n)$ bits. Overall space complexity is $O(\frac{\log n}{\epsilon^2})$ bits.

3.2 1-Norm Estimation: $p = 1$

Find z such that $(1-\epsilon)|\mathbf{x}|_1^1 \leq z \leq (1+\epsilon)|\mathbf{x}|_1^1$ with high probability 9/10. This is a harder problem than estimating $p = 2$ Norm.

We have seen in the previous lecture that Cauchy matrices can be used to estimate $p = 1$ Norms. We can apply a similar approach here:

- Sample a Cauchy matrix S consisting of Cauchy random variables with $1/\epsilon^2$ rows. Notice that S can be stored pseudorandomly with space complexity $1/\epsilon$ words, with each word having $O(\log n)$ bits.
- For each update $\mathbf{x}_i \leftarrow \mathbf{x}_i + \Delta_j$, do $S\mathbf{x}_i \leftarrow S\mathbf{x}_i + \Delta_j S_{*i}$.
- Output median of $|(S\mathbf{x})_1|, |(S\mathbf{x})_2|, \cdots, |(S\mathbf{x})_{1/\epsilon^2}|$ instead of just $|(S\mathbf{x})|_1$ because $|(S\mathbf{x})|_1 \leq O(d \log d)|x|_1$. We saw this in the last lecture and this bound is not tight enough.

To prove that this median works, we need to look into the PDF and CDF of the Cauchy distribution, $\mathbf{C}$.

PDF of $\mathbf{C}$: $f(x) = \dfrac{2}{\pi(1+x^2)}$

CDF of $\mathbf{C}$: $F(z) = \int_0^z f(x)dx = \dfrac{2}{\pi}\arctan(z)$
For $z = 1$, $F(1) = 1/2$, because $\tan(\pi/4) = 1$. This means median$(|\mathbf{C}|) = 1$.

Next we want to show for independent samples $X_1, \cdots, X_r$ from F where $r = \dfrac{\log(1/\delta)}{\epsilon^2}$, if $X = \text{median}_i X_i$, then $F(X) \in [(1/2-\epsilon), (1+\epsilon)]$ with probability $1-\delta$.

Proof. Consider a quantity Z_i such that, $Z = \Sigma_i^r Z_i$, and

$$\begin{aligned} Z_i &= 1, \quad \text{if} F(X_i) < \frac{1}{2} - \epsilon \\ &= 0, \quad \text{otherwise} \end{aligned} \tag{1}$$

Lets look at the situation when $F(X_i) \leq \frac{1}{2} - \epsilon$

The expectation of Z: $\mathbb{E}(Z) = (\frac{1}{2} - \epsilon)r$
Applying Chernoff bound, $\mathbf{Pr}\Big[Z_i \geq \frac{r}{2}\Big] \leq \mathbf{Pr}\Big[|Z_i - E[Z]| \geq \epsilon r\Big] \leq e^{-\Theta(\epsilon^2 r)}$

Plugging in $r = \Theta\left(\frac{1}{\epsilon^2 \log(1/\delta)}\right)$, we have $\mathbf{Pr}\Big[Z_i \geq \frac{r}{2}\Big] \leq \frac{\delta}{2}$

Similarly, looking at $F(X_i) \geq \frac{1}{2} + \epsilon$, we have, $\mathbf{Pr}\Big[Z_j \geq \frac{r}{2}\Big] \leq \frac{\delta}{2}$

By union bound, $\mathbf{Pr}\Big[Z_i \geq \frac{r}{2} \vee Z_j \geq \frac{r}{2}\Big] \leq \delta$. Finally, it means the median $F(X) \in \Big[(\frac{1}{2-\epsilon}) + (\frac{1}{2} + \epsilon)\Big]$ has probability $1-\delta$.

We can also show that $F^{-1}(X) = \tan\left(\frac{X\pi}{2}\right) \in [1-4\epsilon, 1+4\epsilon]$. This shows that the medium is highly concentrated around 1.

$$|Sx|_1 = (1 \pm O(\epsilon))|x|_1$$

3.3 *p*-norm Estimation: $0 < p < 2$

We can estimate the p-norm of a vector by looking at a p-stable distribution.

Definition (p-stable)**.** For a fixed vector $\mathbf{V} \in \mathrm{R}^n$, a distribution Π is p-stable if

$$\sum_{i}^{n} Z_i v_i = Z \cdot |V|_p$$

where X_i and X are p-stable random variables $\sim \Pi$

Quite interestingly, p-stable distributions do not have a closed-form expression of their PDF unlike 1-stable and 2-stable distributions do. But p-stable distributions can be efficiently used for sampling because the memory needed for this is the same as the $p = 1$ situation - $O(\frac{1}{\epsilon^2} \log n)$ bits or $\frac{1}{\epsilon^2}$ words.

Recap of 2-stability:

$$\sum_{i}^{n} G_i v_i = G \cdot |V|_2$$

Where G_i and G are 2-stable Gaussian random variables with a well-defined closed-form PDF expression.

Recap of 1-stability:

$$\sum_{i}^{n} C_i v_i = C \cdot |V|_1$$

Where C_i and C are 2-stable Cauchy random variables with a well-defined closed-form PDF expression.

3.4 *p*-norm Estimation: $p > 2$

p-stable distributions for $p > 2$ do not exist. It is almost as if there is a phase transition at $p = 2$. However, the p-norm can be estimated using exponential random variables that have min-stability. This method requires $O(n^{1-2/p})$ bits of space.

Exponential random variables $E(\lambda)$ have the following properties:

- PDF: $f(x) = \lambda e^{-\lambda x}$, for $x \geq 0$; 0, otherwise
- CDF: $F(x) = 1 - e^{-\lambda x}$, for $x \geq 0$; 0, otherwise
- For any scalar $t \geq 0$, consider $t \cdot E$. The CDF is: $F(x) = 1 - e^{-\left(\frac{\lambda}{t} x\right)}$

- Consider independent exponential rvs $E_i \in E_1, \cdots, E_n$ and scalars $|y|_1, \cdots, |y|_n$. The *min-stable* is defined as follows:

$$q = \min\left(\frac{E_1}{|y|_1^p}, \cdots, \frac{E_n}{|y|_p}\right)$$

In the next part of the lecture we will see how we can take advantage of *min-stable* and sketch using a matrix $S = P.D$ where $P \in \mathbb{R}^{1/\epsilon^2 \times n}$ is a CountSketch matrix and $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix with $D_{i,i} = \frac{1}{E_i^{(1/p)}}$ to estimate $p > 2$ norms efficiently.