**Special Topics in Theoretical Computer Science** Spring 2019

Lecture 3 — January 14, 2019

*Prof. Huy Nguyen* *Scribe: Peter Bernstein*

# 1 Overview

Last time we were working in a setting where initially we had a vector $x \in \mathbb{R}^n$ that was initially 0s and we received updates at particular indices of the form:

$$\text{Update } (i, \delta) : x_i \leftarrow x_i + \delta$$

and when the updates were finished we'd compute $f(x)$.

In this lecture we look at examples of *linear sketching*. In this setting we can't store $x$, so instead we store a *sketch* of $x$, in particular a linear sketch. For example, we might store the product $Ax$ for some matrix $A \in \mathbb{R}^{k \times n}$. Then initially we would have $Ax^{(0)} = 0$ since $x^{(0)} = 0$ and updates would be of the form

$$\text{Update: } A(x + \delta) \leftarrow Ax + A\delta$$

This way we can maintain the sketch through the stream of updates.

Sketching is also useful beyond the streaming model. For example, if we have $n$ distributed machines each storing some data $x_1, ..., x_n$, we can maintain a sketch of the data with $A(x_1 + x_2 + ... + x_n)$.

# 2 Frequency Moments

To think about estimating the frequency moments of $x$, we should first mention the $p$-norm of $x$,

$$\|x\|_p := (x - 1^p + x_2^p + ... + x_r^p)^{1/p}$$

Note that the $p = 2$ norm is just the euclidean norm. As you increase $p$, the norms tell you more about how concentrated the data is around 1 specific coordinate. In the limit, the $p = \infty$ norm tells you exactly the largest coordinate of $x$.

Now we go through a linear sketching algorithm by Alon-Matias-Szegedy '99 to solve the 2nd moment estimation, which is just the 2-norm squared, i.e. $\|x\|_2^2$.

We're going to use linear sketching to solve this problem, and we consider $\delta \in \{-1, 1\}$, so the updates can be either positive or negative. For the basic algorithm, we have $k = 1$ and we need to choose the matrix $A$. We can choose the elements $r_1, r_2, ..., r_n$ uniformly randomly from $\{-1, 1\}$ and all we need to store is $Z = \sum_i r_i x_i$. Then we can compute our estimator for the second moment, or the 2-norm squared ($\|x\|_2^2\|$), as $Z^2$.

The distribution of $Z$ may be complicated so we start by computing our favorite statistics, the expectation and variance. The expectation of $Z^2$ is

$$\begin{aligned}\mathbb{E}[Z^2] &= \mathbb{E}[(\sum_i r_i x_i)^2] \\ &= \mathbb{E}[\sum_{i=1} \sum_{j=1} r_i r_j x_i x_j]\end{aligned}$$

Let's examine $i = 1, j = 2$. The only things that are random are the $r_i$, and they are independent, so the expectation can be split and thus $\mathbb{E}[r_1 r_2 x_1 x_2] = 0$.

In general, if $i \neq j$, then $\mathbb{E}[r_i r_j x_i x_j] = 0$ and $\mathbb{E}[r_i^2 x_i^2] = x_i^2$ because $r_i$ is either -1 or 1 so the square is always equal to 1. Thus $\mathbb{E}[Z^2] = \|x\|_2^2$.

This is good, we have an unbiased estimator of the 2-norm now. Next let's understand the variance:

We need to compute $\mathbb{E}[Z^4]$ to understand the variance.

$$\mathbb{E}[Z^4] = \sum_{i,j,k,l} \mathbb{E}[r_i, r_j, r_k, r_l, x_i, x_j, x_k, x_l]$$

It would be nice to get rid of some of these terms, so let's use the trick that if some index occurs an odd number of times, then the expectation is 0. Then

$$\mathbb{E}[Z^4] = 3 \sum_i x_i^2 \sum_{j \neq i} x_j^2 + \sum_i x_i^4$$

The 3 comes from the fact that the second index matching $i$ could be any of the other 3 indices $j, k, l$. And now we can compute the variance,

$$\begin{aligned}\text{var}[Z^2] &= \mathbb{E}[Z^4] - E[Z^2]^2 \\ &= 3 \sum_i x_i^2 \sum_{j \neq i} x_j^2 + \sum_i x_i^4 - \left( \sum_i x_i^2 \right)^2 \\ &\leq 3 \left( \sum_i x_i^2 \right)^2 - \left( \sum_i x_i^2 \right)^2 \\ &= 2 \left( \mathbb{E}[Z^2] \right)^2\end{aligned}$$

As usual, we have that the expectation is equal to what we want but a variance that's too large. We can now use our usual trick of repeating the basic algorithm many times and taking and average to reduce the variance when compared to one iteration of the algorithm. Ideally we want a variance of $O\left(\varepsilon^2 \, \mathbb{E}[Z^2]^2\right)$ so we should repeat $O\left(\frac{1}{\varepsilon^2}\right)$ times.

Let's maintain $Z_1, Z_2, ..., Z_k$ where $k = 10/\varepsilon$. Then our estimator is

$$\hat{Z}^2 = \frac{Z_1^2 + Z_2^2 + ... + Z_k^2}{k}$$

which has $\mathbb{E}[\hat{Z}^2] = \mathbb{E}[Z^2]$ and $\text{var}[\hat{Z}^2] = \frac{\text{var}(Z^2)}{k} \leq \frac{\varepsilon^2 \left(\mathbb{E}[Z^2]\right)^2}{5}$. Using Chebyshev,

$$\mathbb{P}\left[|\hat{Z}^2 - \mathbb{E}[\hat{Z}^2]\right] \leq \frac{\varepsilon^2 \left(\mathbb{E}[\hat{Z}^2]\right)^2/5}{\varepsilon^2(\mathbb{E}[\hat{Z}^2])^2} = \frac{1}{5}$$

And just like in the previous lectures we can take $\log(1/\delta)$ copies and use the median of means.

Now if we want to reduce the storage of random bits in the matrix $A$ what can we do? Do we need independently random bit? Can we use $k-$wise independence? The answer is that we only need 4-wise independence because of the terms in $\mathbb{E}[Z^4]$. The total space is then $\frac{1}{\varepsilon^2} \log \frac{1}{\delta} \log n$.

# 3 Distributional Johnson-Lindenstrauss Lemma

In JL sketching, we use a normal distribution for updates instead of $\pm 1$. A normal distribution with mean $\mu$ and variance $\sigma^2$ is denoted by $\mathcal{N}(\mu, \sigma^2)$, the density function for $\mathcal{N}(\mu, \sigma)$ is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(x-\mu)^2/2\sigma^2)$$

One important feature for Normal distribution is that it's closed under linear transformations. For example, if $X$ and $Y$ are independent random variable with normal distribution. Then $aX + bY$ has a normal distribution.

## 3.1 Description of the sketch

We have a sketching matrix $R \in \mathbb{R}^{k \times n}$ with $R_{ij} \sim \mathcal{N}(0, 1)$. Note that this is similar to the $\pm 1$ sketch in that the mean and variance are still 0 and 1 respectively. We maintain the vector $Rx$ and output $\|Rx\|_2^2/k$ as our estimator for $\|x\|_2^2$.

## 3.2 Analysis of the algorithm

Let's dive into our favorite statistics again.

$$\begin{aligned}
\mathbb{E}[\|Rx\|_2^2/k] &= \frac{1}{k}\,\mathbb{E}(x^T R^T Rx) \\
&= \frac{1}{k} x^T\,\mathbb{E}(R^T R)x \\
&= \frac{1}{k}\,\mathbb{E}[\sum_{u,v,w} x_u R_{v,u} R_{v,w} x_w]
\end{aligned}$$

If $u \neq w$ then $\mathbb{E}[x_u R_{v,u} R_{v,w} x_w] = 0$ so

$$\mathbb{E}[\|Rx\|_2^2/k] = \frac{1}{k}\,\mathbb{E}[\sum_{u=1}^{n} x_u^2 \sum_{v=1}^{k} R_{v,u}^2] = \mathbb{E}[\sum_u x_u^2] = \|x\|^2$$

Next we will show that this concentrates around its mean with a normal tail. Without loss of generality, assume that $\|x\| = 1$, since we could always scale the matrix coefficients to compensate. We want to prove the following for one side of the tail (the other side is similar):

**Lemma 1** (Distributional Johnson-Lindenstrauss Lemma)**.**

$$\mathbb{P}(1 - \epsilon \leq \|Rx\|/\sqrt{k} \leq 1 + \epsilon) \geq 1 - 2\delta$$

*Proof.* We rewrite this with $Z = Rx$, square both sides, and since the exponential function is monotonic we can prove the following to prove the lemma:

$$\mathbb{P}(\|Z\|^2 \geq (1+\varepsilon)^2 k) \leq \exp(-\varepsilon^2 k + O(k\varepsilon^3))$$

Let $Y = \|Z\|^2$ and let $\alpha = k(1+\epsilon)^2$, we have for $s > 0$ by the Markov bound

$$\mathbb{P}(Y > \alpha) = \mathbb{P}(\exp(sY) > \exp(s\alpha)) \leq \exp(-s\alpha)\,\mathbb{E}[\exp(sY)]$$

We can split $\mathbb{E}[\exp(sY)]$ by independence:

$$\mathbb{E}[\exp(sY)] = \prod_i E[\exp(sZ_i^2)] \tag{1}$$

By the closure property, $Z_i$ also follows a normal distribution. Since for every $i$

$$\mathbb{E}[Z_i] = \sum_j \mathbb{E}[r_{ij}x_j] = 0$$

$$\mathrm{var}(Z_i) = \mathbb{E}[Z_i^2] = \sum_j \mathbb{E}[r_{ij}^2 x_j^2] = \|x\|^2 = 1$$

we obtain that $Z_i \sim \mathcal{N}(0,1)$, so we can calculate $E[\exp(sZ_i^2)]$ analytically

$$\mathbb{E}[\exp(sZ_i^2)] = \frac{1}{\sqrt{2\pi}} \int \exp(st^2)\exp(-t^2/2)dt = \frac{1}{\sqrt{1-2s}}$$

So we get

$$\mathbb{P}(Y \geq \alpha) = \exp(-s\alpha)(1-2s)^{-k/2}.$$

The last step is to plug in an appropriate choice of $s$. We set $s = (1 - k/\alpha)/2$, giving

$$\mathbb{P}[Y > \alpha] \leq e^{-s\alpha}(1-2t)^{-k/2} = e^{(k-\alpha)/2}(k/\alpha)^{-k/2}$$

choosing $\alpha = k(1+\varepsilon)^2$ and plugging it into the above equation we get

$$\mathbb{P}(Y \geq \alpha) = \exp(-\varepsilon k - \varepsilon^2 k/2 + k\ln(1+\varepsilon)) = \exp(-k\varepsilon^2 + kO(\varepsilon^3))$$

Here we use the Taylor expansion $\log(1+x) = x - x^2/2 + O(x^3)$. This tail bound proves the correctness of the estimation and we obtain a better parameter for $k$. Let $\exp(-C\varepsilon^2 k) = \delta$ we have $k = O(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$ to get $1 \pm \varepsilon$ approximation with probability $1 - \delta$.

Unlike before, we cannot get away with 4-wise independence. When we split the expectation in (1) we need at least $k-$wise independence, but that still means we don't need full independence. In terms of time complexity, we need $O(k)$ to update the sketch. □