Special Topics in Theoretical Computer Science Spring 2019

Lecture 5 — 24Jan2019

Prof. Huy L. Nguyen Scribe: Biswaroop Maiti

1 Overview

In the last lecture we studied how to estimate the L_p norm of a frequency vector x for $p : 0 \leq p < 2$. In this lecture we will study how to estimate L_p norm for $p > 2$. Last time, we were able to sketch L_p norm using space polylogarithmic in n, the number of coordinates of the given vector. This time, we can sketch it using sublinear space.

In the previous lecture a key ingredient was the existence of a p- stable distribution. Unfortunately, there are no p stable distribution for $p = 2$.

1.1 Exponential Distribution

This distribution has some interesting properties, with range in $t \in [0, \infty)$. The standard exponential pdf is e^{-t}. The general exponential distribution has a parameter λ. The pdf ogf the general exponential distribution is $\lambda \cdot e^{-\lambda t}$ cdf is $1 - e^{-\lambda t}$

Expectation is λ^{-1}.

Claim 1. Let $X_1, X_2, \cdots X_n$ be independent exponential random variables with parameters $\lambda_1, \lambda_2, \cdots \lambda_n$ respectively. Then, the random variable $X := \min\{X_1, X_2, \cdots X_n\}$ is distributed according to the exponential distribution with parameter $\lambda = \sum_{i=1}^{n} \lambda_i$. $X \sim \mathsf{exp}(\lambda)$.

Proof.

$$\begin{aligned}\mathbb{P}[X \geq t] &= \Pi_{i=1}^{n} P[X_i \geq t] \\ &= \Pi_{i=1}^{n} e^{\lambda_i t} \\ &= e^{\sum_{i=1}^{n} \lambda_i t}\end{aligned}$$

□

This property is sometimes referred to as the "max stability" property of Exponential distribution. Let $X \in \mathbb{R}^n$ be the frequency vector. Let $U_1, U_2, \cdots U_n$ be iid-s that is, independent identical exponential random variables. We combine them with X_i-s in a manner that they become exponential random variables with some parameter, such that the scaling factor turns out to be the L_p norm of X.

We define:

$$Y_i = \frac{X_i}{U_i^{1/p}} \qquad Y_i^p = \frac{X_i^p}{U_i}$$

Therefore, $1/Y \sim \mathsf{exp}(\lambda t)/\|X\|_p$.

However, Y still has n dimensions. We can compress it by using the linear sketch idea from the AMS paper [1].

The total mass difference in $\frac{1}{Y_i}$ will not be much.

Step 2: We will use random hash functions.

$$h : [n] \to [m]$$

We will distribute n items into m buckets. We analyze the random variable $Z_j := \sum_{h(i)=j} Y_i \cdot \sigma_i$, where σ_i are random signs i.e. $\sigma_i \overset{\$}{\leftarrow} \{\pm 1\}$.

Therefore, every coordinate has a random sign.

Lemma 2. $\mathbb{P}\left[\|Y\|_\infty \in \left[\frac{1}{2}\|X\|_p, 2\|X\|_p\right]\right] \geq 3/4$

Proof. Let $q = \min\left(\frac{U_i}{|X_i|^p}\right)$. Note $1/q = \|Y\|_\infty^p$

$$\begin{aligned}\mathbb{P}[q \geq t] &= \mathbb{P}[\forall i : \frac{U_i}{|X_i|^p} \geq t] \\ &= \Pi_i \mathsf{exp}\{-|X_i|^p t\} \\ &= \mathsf{exp}\{-\|X\|_p^p t\}\end{aligned}$$

Therefore, q is distributed exponentially with the parameter $\|X\|_p^p$.

□

Since we know that $1/q = \|Y\|_\infty^p$, we have:

$$\|Y\|_\infty \in \left[\frac{1}{2}\|X\|_p, 2\|X\|_p\right] \iff q \in \left[\frac{1}{2^p}\frac{1}{\|X\|_p}, 2^p\frac{1}{\|X\|_p}\right]$$

What is the CDF?

$$\mathbb{P}[\cdots] = e^{-1/2^p} - e^{-2^p} \geq 3/4$$

Now we need to bound the variance in the compression from Y to Z. We need to show that the maximum is preserved.

We need to make sure that the entries of opposing signs in any bucket don't cancel each other and decrease the weight. We have two issues to deal with here:

1. No two "big" coordinates (elephants) collide i.e. fall in the same bucket (because then they might cancel each other with some probability 1/2)

2. Noise from "small" coordinates is insignificant (if a large number of small coordinates are of same sign they might add up and increase the total noise)

Given a coordinate Y_i, we call it "Big" if $|Y_i| \geq \frac{\|X\|_p}{c \log n} = \frac{M}{c \log n}$ and call it small otherwise, i.e. if $|Y_i| < \frac{M}{c \log n}$.

Let A_i be the indicator random variables for the event when the coordinate Y_i is small.

$$A_i = \begin{cases} 1, & \text{if } Y_i \geq M/l \\ 0, & \text{otherwise} \end{cases}$$

$$\begin{aligned} \mathbb{E}[A_i] &= \mathbb{P}\left[\frac{|X_i|}{U_i^{1/p}} \geq M/l\right] \\ &= \mathbb{P}\left[U_i \leq \frac{|X_i|^p \cdot l^p}{M^p}\right] \\ &= 1 - \mathsf{exp}\left(\frac{|X_i|^p \cdot l^p}{M^p}\right) \\ &\leq \frac{|X_i|^p \cdot l^p}{M^p} \end{aligned}$$

For the last step, we used the fact $\mathsf{exp}(-t) \geq 1 - t$.

Using linearity of expectation, we find the expected number of indices such that the coordinates are big as follows.

$$\begin{aligned} \mathbb{E}\left[|\text{index } i \text{ such that } Y_i \geq M/l|\right] &= \mathbb{E}\left[\sum_i A_i\right] \\ &\leq \frac{\|X\|_p^p \cdot l^p}{M^p} \\ &= l^p \end{aligned}$$

Therefore, in expectation the number of "big" coordinates is $(c \log n)^p$. With probability $\frac{99}{100}$, the number of big coordinates is at most $100(c \cdot \log n)^p$.

We have $m = \Theta(n^{1-2/p} \log n)$ buckets. We want to show that

Exercise 3. Birthday Paradox. Show that with probability 99/100, no two "big" elements collide for $m = \Theta(n^{1-2/p} \log n)$.

The small coordinates are $\{i : |y_i| < \frac{M}{c \log n}\}$. Noise for Z_j is $\sum_{i:\text{ small}} Y_i \sigma_i$

This is a complicated distribution and we shall try to understand it through its expectation and variance.

We define: $Z_j^{'} = \sum_{i:\text{ small}} Y_i \sigma_i$

We get the following.

$$\mathbb{E}[Z_j^{'}] = 0$$

$$\begin{aligned}\mathbb{E}[(Z_j^{'})^2] &= \mathbb{E}[\left(\sum_{i:\text{ small} h(i=j)} Y_i \cdot \sigma_j\right)^2] \\ &= \mathbb{E}[\sum_{i:\text{ small} h(i)=j}] \\ &= \frac{1}{m} \cdot \|Y\|_2^2\end{aligned}$$

Claim 4. $\mathbb{E}[\|Y\|_2^2] \leq n^{1-2/p}\|X\|_p^2$

Proof.

$$\begin{aligned}\mathbb{E}[Y_i^2] &= \mathbb{E}\left[\frac{X_i}{U_i^{2/p}}\right] \\ &= O(X_i^2)\end{aligned}$$

$\implies$

$$\mathbb{E}[\|Y\|_2^2] = O(\|X\|_2^2)$$

□

Now, we recall an important inequality which will let us compare 2-norm and p-norms of a vector.

Lemma 5. Holder's Inequality $\langle f, g\rangle \leq \|f\|_a \|g\|_b$ for $\frac{1}{a} + \frac{1}{b} = 1$.

Some interesting special cases of this inequality are $a = 1, b = \infty$ and $a = 2, b = 2$.

Apply Holder's inequality with $f_i = X_i^2$ and $g_i = 1$ with $a = p/2, b = \frac{1}{1-2/p}$.

$$\begin{aligned}\|X\|_2^2 &\leq \left(\sum_i X_i^{2\cdot p/2}\right)\left(\sum_i 1\right)^{1-2/p} \\ &\leq \|X\|_p^2 \cdot n^{1-2/p}\end{aligned}$$

Thus,

$$\begin{aligned}Var\left(Z_j^{'}\right) &\leq \frac{n^{1-2/p}M^2}{m} \\ &\leq M^2/(c\log n)\cdot\end{aligned}$$

Lemma 6. Bernstein's Inequality Suppose $X_1, X_2, \cdots X_n$ are independent random variables with expectation $\mathbb{E}[X_i] = 0$ and their absolute value $|X_i| \leq Q$. Then, we have

$$\mathbb{P}[\sum_{i=1}^{n} X_i \geq t] \leq \exp\left(-\frac{t^2/2}{\sum \mathbb{E}[X_i^2] + \frac{1}{3}Qt}\right)$$

When we apply this inequality with $Q = \frac{M}{c \log n}, t = \alpha M$ with a constant value of say, $\alpha = \frac{1}{2}$.

Then we have:

$$\mathbb{P}[Z_j^{'} \geq \alpha M] \leq \exp\left(-\frac{\alpha^2 M^2/2}{\frac{M^2}{c \log n} + \frac{1}{3} \cdot \frac{M}{c \log n} \cdot \alpha M}\right)$$

$$\begin{aligned}\mathbb{P}[Z_j^{'} \geq \alpha M] &\leq \exp\left(-\frac{\alpha^2 M^2/2}{\frac{M^2}{c \log n} + \frac{1}{3} \cdot \frac{M}{c \log n} \cdot \alpha M}\right) \\ &\leq \frac{1}{n^2}.\end{aligned}$$

By Union Bound, we get $\forall j : Z_j^{'} < \alpha \cdot M$.

We have left out the analysis of the U_i-s. We can view this as a ROBP and use Nisan's generator to use PRG and bound the amount of randomness used as a resource. Nisan's generator takes polylog space and therefore does not increase the space requirements being sublinear.

This is still a linear sketch but the analysis is far more involved than the case for $p < 2$.

References

[1] Noga Alon, Yossi Matias, Mario Szegedy. The Space Complexity of Approximating the Frequency Moments. J. Comput. Syst. Sci., 58(1):137–147, 1999.