CS 15-859: Algorithms for Big Data Fall 2019

Lecture 6 (Part 2) — 10/10/2019

Prof. David Woodruff *Scribe: Neil Xu*

1 Proof of Sketching Theorem (cont.)

1.1 Upper bounding the Cauchy variables

Motivation At the end of the previous part, we observed that the Cauchy random variables in our sketching matrix are at least 1 with high probability - but we did not provide an upper bound. We used the fact that the Cauchy distribution is heavy tailed in the lower bound, but this fact could also make proving an upper bound tricky. In relation to the overall proof plan, we need to upper bound the Cauchy variables to prove the upper bound portion of the theorem. We will later see that the upper bound portion of the theorem is used in proving the lower bound side.

Proposition 1. For each $|Z_j|$, its c.d.f. $F_{|Z|}(|Z_j)$ is asymptotically $1 - \Theta\left(\frac{1}{z}\right)$.

Proof. The pdf of $|Z_j|$ is

$$f_{|Z|}(z) = \frac{2}{\pi(z^2+1)}$$

for $|Z_j| \geq 0$.

Thus, the c.d.f. is

$$F_{|Z|}(z) = \int_0^z \frac{2}{\pi(x^2+1)}dx = -\frac{2}{\pi(x+1)}\Big|_0^z = -\frac{2}{\pi(z+1)} + \frac{2}{\pi} = 1 - \frac{\pi-2}{\pi} - \frac{2}{\pi(z+1)} = 1 - \Theta\left(\frac{1}{z}\right)$$

■

Let's assume there exists a well-conditioned basis of A (such as the Lowner-Johnson one we constructed before). We can write the $l1$ norm of the ith column of A (which we now assume is well-conditioned) as:

$$|RA_{*i}|_1 = |A_{*i}|_1 \sum_{j=1}^{d\log(d)} |Z_{i,j}| / (d\log(d))$$

Define $E_{i,j}$ to be the event that $|Z_{i,j}| \leq d^3$.

Proposition 2. $\mathbb{E}[Z_{ij} \mid E_{ij}] = O(\log(d))$

Proof.

$$
\begin{aligned}
\mathbb{E}[|Z_{ij}| \mid E_{ij}] &= \int_0^\infty x \cdot \Pr[|Z_{ij}| = x \mid E_{ij}]dx \\
&= \int_0^\infty x \cdot \frac{\Pr[E_{ij} \cap |Z_{ij}| = x]}{\Pr[E_{ij}]}dx \\
&= \int_0^{d^3} x \cdot \frac{\Pr[E_{ij} \cap |Z_{ij}| = x]}{\Pr[E_{ij}]}dx && \text{Probability of } E_{ij} \text{ is 0 when } x > d^3 \\
&= \int_0^{d^3} x \cdot \frac{\Pr[|Z_{ij}| = x]}{\Pr[E_{ij}]}dx \\
&= \int_0^{d^3} \frac{2x}{\pi(x+1)^2 \Pr[E_{ij}]}dx && \text{Substitute in half Cauchy p.d.f.} \\
&= \frac{1}{1 - \Theta\left(\frac{1}{d^3}\right)} \int_0^{d^3} \frac{2x}{\pi(x+1)^2}dx && \text{Proposition 1} \\
&= \frac{1}{1 - \Theta\left(\frac{1}{d^3}\right)} \left(O(1) + \int_1^{d^3} \frac{2x}{\pi(x+1)^2}dx \right) && \text{Integrand had u.b. for bounded range [0, 1]} \\
&= \frac{1}{1 - \Theta\left(\frac{1}{d^3}\right)} \left(O(1) + \int_1^{d^3} \Theta\left(\frac{1}{x}\right)dx \right) \\
&= \frac{1}{1 - \Theta\left(\frac{1}{d^3}\right)} \left(O(1) + \Theta\left(\log(x)\right)\Big|_1^{d^3} \right) \\
&= O(\log(d))
\end{aligned}
$$

■

In addition, let E the event that E_{ij} occurs for all i, j.

Proposition 3. $\Pr[E] \geq 1 - \Theta(\frac{\log(d)}{d})$.

Proof. We know $\Pr[\neg E_{ij}] = \frac{1}{\Theta(d^3)}$ from proposition 1.

By union bound $\Pr[\exists i, j : \neg E_{ij}] \leq \Theta\left(\frac{d^2 \log(d)}{d^3}\right) = \Theta\left(\frac{\log(d)}{d}\right)$ since the dimensions of RA are $d\log(d) \times d$. Thus, $Pr[E] = 1 - \Theta\left(\frac{\log(d)}{d}\right)$. ■

To actually upper bound $|Z_{ij}|$, we need to consider $\mathbb{E}[|Z_{ij}| \mid E]$.

Proposition 4. $\mathbb{E}[|Z_{ij}| \mid E] = O(\log(d))$

Proof.

$$\begin{aligned}
\mathbb{E}[|Z_{ij}| \mid E_{ij}] &= \mathbb{E}[|Z_{ij}| \mid E_{ij}, E]\Pr[E \mid E_{ij}] + \mathbb{E}[|Z_{ij}|E_{ij}, \neg E]\Pr[\neg E \mid E_{ij}] \\
&\geq \mathbb{E}[|Z_{ij}| \mid E_{ij}, E]\Pr[E \mid E_{ij}] \\
&= \mathbb{E}[|Z_{ij}| \mid E]\left(\frac{\Pr[E_{ij} \mid E]\Pr[E]}{\Pr[E_{ij}]}\right) \\
&= \mathbb{E}[|Z_{ij}| \mid E]\Pr[E] && \text{Prob. is upper bounded by 1} \\
& && E \text{ is superset of } E_{ij} \\
&\geq \mathbb{E}[|Z_{ij}| \mid E]\left(1 - \Theta\left(\frac{\log(d)}{d}\right)\right) && \text{Proposition 3}
\end{aligned}$$

Given Proposition 2, we get that:

$$O(\log(d)) \geq E[|Z_{ij}| \mid E]\left(1 - \Theta\left(\frac{\log(d)}{d}\right)\right)$$

$$\frac{O(\log(d))}{\left(1 - \Theta\left(\frac{\log(d)}{d}\right)\right)} = O(\log(d)) \geq E[|Z_{ij}| \mid E]$$

■

Given large enough d, we can lower bound the Proposition 3 lower bound for $\Pr[E]$ with a constant. Thus, with constant probability, we can invoke Proposition 4 rewrite our previous L1 norm of a column of RA as

$$\begin{aligned}
|RA_{*i}|_1 &= |A_{*i}|_1 \sum_{j=1}^{d\log(d)} O(\log(d))/(d\log(d)) \\
&= O(\log(d))\,|A_{*i}|_1
\end{aligned}$$

Thus, we can say that the sum of L1 norms of columns (still w/ constant probability):

$$\sum_{i=1}^{d} |RA_{*i}|_1 = O(\log(d)) \sum_{i=1}^{d} |A_{*i}|_1$$

We now introduce a new type of basis:

Definition. An **Auerbach basis** A is a basis for a d-dimensional subspace such that:

1. For all x: $|x|_\infty \leq |Ax|_1$
2. $\sum_{i=1}^{d} |A_{*i}|_1 = d$

and always exists for every subspace.

Proposition 5. An Auerbach basis A is well-conditioned.

Proof. plainurl

$$\begin{aligned}
\frac{|x|_1}{d} &\leq |x|_\infty && \text{Elem. mean} \leq \text{Max elem.} \\
&\leq |Ax|_1 && \text{Auerbach prop. 1} \\
&= \left| \sum_{i=1}^{n} \sum_{j=1}^{d} A_{ji} \cdot x_i \right|_1 \\
&\leq \sum_{i=1} \sum_{j=1} |A_{ji}|_1 \, |x_i|_1 && \text{triangle inequality} \\
&\leq \sum_{i=1} \sum_{j=1} |A_{ji}|_1 \, |x|_\infty \\
&= d \, |x|_\infty && \text{Auerbach prop. 1} \\
&\leq d \, |x|_1
\end{aligned}$$

So A is well-conditioned up to factor d. ■

By the 2nd property of Auerbach basis, we now can rewrite our L1 norm sum as:

$$\sum_{i=1}^{d} |RA_{*i}|_1 = O(\log(d)) \sum_{i=1}^{d} |A_{*i}|_1 = O(d \log(d))$$

To conclude the proof, we can show that, for all x:

$$\begin{aligned}
|RAx|_1 &\leq \sum_{i=1}^{d} |RA_{*i} x_i|_1 && \text{by triangle inequality} \\
&\leq |x|_\infty \sum_{i=1}^{d} |RA_{*i}|_1 \\
&= |x|_\infty \, O(d \log(d)) && \text{L1 norm sum result above} \\
&\leq O(d \log(d)) \, |Ax|_1 && \text{prop. 1 of Auerbach}
\end{aligned}$$

Thus, we've shown with constant probability, for all x, an upper bound on the L1 norm of a vector projected in the sketched subspace by $O(d \log(d))$ factor on the L1 norm of the vector projected in the original subspace.

1.2 Sketch lower bounds

Currently, we have proved the following statements:

1. There is a γ-net M s.t. $|M| \leq \left(\frac{d}{\gamma}\right)^{O(d)}$ of the set $\{Ax : |x|_1 = 1\}$.

2. For any fixed x, $|RAx|_1 \geq |Ax|_1$ with probability $1 - \exp(-d\log(d))$
3. For all x, $|RAx|_1 = O(d\log(d))\,|Ax|_1$ (what we just proved in the previous part).

To complete the sketching theorem, we now have to show the lower bound: $|Ax|_1 \leq |RAx|_1$, since we have already proven an upper bound of $|RAx|_1 = O(d\log(d))\,|Ax|_1$.

First, we can choose $\gamma = \frac{1}{d^3\log(d)}$. Thus, our γ net M will have size $|M| \leq \left(\frac{d}{\frac{1}{d^3\log(d)}}\right)^{O(d)} = d^{O(d)}$.

Using the statement 2, we get a union bound on the probability that all vectors Ax that form the net in M will satisfy $|RAx|_1 \geq |Ax|_1$ with

$$1 - d^{O(d)}\exp(-d\log(d)) = 1 - d^{O(d)}(\exp(-\log(d)))^d = 1 - d^{O(d)}d^{-d} = 1 - O(1)$$

Thus, by union bound for all net vectors y in M: $|Ry|_1 \geq |y|_1$.

If we look at an arbitrary x s.t. $|x|_1 = 1$, we know it satisfies the net property i.e.

$$|Ax - y|_1 \leq \gamma = \frac{1}{d^3\log(d)}$$

Thus, we can know decompose the sketch $|RAx|_1$:

$$\begin{aligned}
|RAx|_1 &\geq |Ry|_1 - |R(Ax-y)|_1 && \text{by triangle inequality} \\
&\geq |y|_1 - O(d\log(d))\,|Ax-y|_1 && \text{above and statement 3} \\
&\geq |y|_1 - O(d\log(d))\gamma && \text{prop. of } \gamma\text{-net} \\
&\geq |y|_1 - O\left(\frac{1}{d^2}\right) && \gamma = \frac{1}{d^3\log(d)}
\end{aligned}$$

Note that, by definition, y is of the form Ax' for some x' where $|x'|_1 = 1$, which allows us to use the Auerbach property 1 to get:

$$|y|_1 = |Ax'|_1 \geq |x|_\infty \geq |x|_1/d = 1/d$$

Consequently, for large enough d, $\frac{|y|_1}{2} \geq \frac{1}{d^2}$.

Thus $|y|_1 - O\left(\frac{1}{d^2}\right) \geq |y|_1 - \frac{|y|_1}{2} = \frac{|y|_1}{2}$, and consequently $|RAx|_1 \geq \frac{|y|_1}{2}$.

Now to lower bound $|y|_1$:

$$\begin{aligned}
|y|_1 &\geq |Ax|_1 - |Ax-y|_1 && \text{by triangle inequality} \\
&\geq |Ax|_1 - \gamma && \text{def. of } \gamma\text{-net} \\
&\geq |Ax|_1 - \frac{1}{d^3\log(d)}
\end{aligned}$$

By prop. 1 of Auerbach basis again, we can note $|Ax|_1 \geq 1/d$. Thus, $\frac{|Ax|_1}{2} \geq \frac{1}{d^3 \log(d)}$ and $|y|_1 \geq \frac{|Ax|_1}{2}$. Putting the inequalities together, we get:

$$|RAx|_1 \geq \frac{|num|_1}{2} \geq \frac{|Ax|_1}{4}$$

and consequently, we can scale R by 4 to achieve the lower bound of the sketching theorem and conclude the proof.

1.3 Sketch matrix variants

One of the challenges of the L1 regression algorithm is generating and using the random matrix of i.i.d. Cauchy variables. Thus, some work has been done on producing faster sketching matrices.

CountSketch w/ Cauchy variables One faster sketching matrix is the CountSketch matrix, but instead of the values being simply ± 1, they're replaced with i.i.d. Cauchy variables for each column [CW][MM]. This leverages sparsity of the input matrix and gets a $l1$-regression runtime of $\text{nnz}(A) + \text{poly}(d/\epsilon)$. The sketch matrix however can only satisfy a sketching theorem of $\frac{1}{d^2 \log^2(d)} |Ax|_1 \leq |RAx|_1 \leq O(d \log(d)) |Ax|_1$.

CountSketch w/ inverse exponential variables [WZ] also proposes CountSketch, but instead of using i.i.d. Cauchy variables per column, to use i.i.d. inverse exponential variables. Inverse exponential variables also have a heavy tail, similar to Cauchy variables, and can require a fewer number of rows being sampled in the $l1$ regression algorithm. The sketching matrix w/ inverse exponential variables only has a guarantee of $\frac{1}{\sqrt{d}\text{poly}(\log(nd))} |Ax|_1 \leq |RAx|_1 \leq O(d \log(d)) |Ax|_1$.

2 Fun Cauchy Distribution Fact

Normally, we expect by central limit theorem that for i.i.d. X_i with mean 0 and variance σ^2 :

$$\sum_{i=1}^{n} X_i \xrightarrow{d} \mathcal{N}(0, \sigma^2/n)$$

The variance decreasing seems intuitive, since by averaging more random variables, they mean becomes more concentrated around the true mean.

For i.i.d. X_i that are standard Cauchy variables, however, by 1-stability of the Cauchy distribution, we know that $\frac{1}{n} \sum X_i$ is just the standard Cauchy variable again, and not a normal distribution.

References

[CW] Kenneth L. Clarkson and David P. Woodruff. Low Rank Approximation and Regression in Input Sparsity Time.

[MM] Xiangrui Meng and Michael W. Mahoney. Low-distortion Subspace Embeddings in Input-sparsity Time and Applications to Robust Linear Regression. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 91–100. ACM. ISBN 978-1-4503-2029-0.

[WZ] David P Woodruff and Qin Zhang. Subspace Embeddings and p-Regression Using Exponential Random Variables. page 22.