CS 15-859: Algorithms for Big Data Fall 2020

Lecture 7 (part 1) — 10/24/20

Prof. David Woodruff *Scribe: Charlie Hou*

1 Recap

Recall that in last lecture we talked about ℓ_1 regression

$$\min_x \|Ax - b\|_1$$

This can be solved via linear programming. We want to sketch this to make the LP smaller! Recall there were two components that we needed for the ℓ_1 sketching algorithm. First, we need a sketching matrix that gets a poly(d) approximation. Second, we need a well-conditioned basis U that preserves the ℓ_1 norm (up to poly(d)).

These two components are easy to get, given the following theorem:

Theorem 1. *There is a probability space over* $(d \log d) \times n$ *matrices* R *such that for any* $n \times d$ *matrix* A*, with probability at least* $99/100$ *we have for all* x*:*

$$\|Ax\|_1 \leq \|RAx\|_1 \leq O(d \log d)\|Ax\|_1$$

The way to get the aforementioned two components via this theorem was shown in the previous lecture, and we won't show it here.

2 Proving the theorem

2.1 Proof overview

Let $R_{i,j} \sim (\frac{1}{d \log d})C$, where C is a standard Cauchy rv. Recall that a Cauchy rv has the following pdf:

$$\text{pdf}(z) = \frac{1}{\pi(1 + z^2)}$$

Remark 1. The Cauchy rv has an undefined expectation and infinite variance. It is also heavy-tailed. However, it does have one very nice property: it is 1-stable, i.e. given $a \in \mathbb{R}^n$, $\{z_i\}_{i=1}^n$ all iid standard Cauchy,

$$\sum_{i=1}^n a_i z_i \sim \|a\|_1 z$$

where z is a standard Cauchy rv.

First, let us do our usual tricks. A can be assumed to be a well-conditioned basis, because we can just write $Ax = UW^{-1}x$ for some invertible W, and the conclusions have to be true for all x. Second, $\|x\|_1$ can be assumed to be 1, since we can divide all the inequalities by $\|x\|_1$. In particular, we can assume A to be the Auerbach basis (which always exists), which has the following properties:

- For all x, $\|x\|_\infty \leq \|Ax\|_1$
- $\sum_i \|A_{*,i}\|_1 = d$

Now assume the following three facts:

- for any fixed x, $\|RAx\|_1 \geq \|Ax\|_1$ w.p. $1 - \exp(-d \log d)$
- there is a γ-net M with $|M| \leq (\frac{d}{\gamma})^{O(d)}$, of the set $\{Ax | \|x\|_1 = 1\}$
- for all x, $\|RAx\|_1 = O(d \log d)\|Ax\|_1$

The third bullet point is exactly the second in the series of inequalities we have to prove, so it remains to prove the first in the series of inequalities. Setting $\gamma = \frac{1}{d^3 \log d}$, which gives us $|M| \leq d^{O(d)}$, we get via union bound that the probability some $y \in M$ does not satisfy $\|Ry\|_1 \geq \|y\|_1$ is at most $d^{O(d)} \exp(-d \log d) = O(1)$. Then if we let $\|x\|_1$ be arbitrary, $\|Ax - y\|_1 \leq \gamma$ for some y, it gives us

$$\begin{aligned}
\|RAx\|_1 &\geq \|Ry\|_1 - \|R(Ax - y)\|_1 \\
&\geq \|y\|_1 - O(d \log d)\|Ax - y\|_1 \\
&\geq \|y\|_1 - O(d \log d)\gamma \\
&\geq \|y\|_1 - O(\frac{1}{d^2}) \\
&\geq \frac{\|y\|_1}{2}
\end{aligned}$$

Where the last inequality follows because $y = Az$ for some unit z, and $\|Az\|_1 \geq \|z\|_\infty$ by property of the Auerbach basis, and $\|z\|_\infty \geq \frac{1}{d}$ because otherwise $\|z\|_1 < 1$. Continuing the series of inequalities, we have (by the property of the net + triangle inequality, and $\|Ax\|_1 \geq \frac{1}{d}$)

$$\begin{aligned}
&\geq \frac{\|Ax\|_1 - \gamma}{2} \\
&\geq \frac{\|Ax\|_1}{3}
\end{aligned}$$

Altogether, we have

$$\frac{\|Ax\|_1}{3} \leq \|RAx\|_1 \leq O(d \log d)\|Ax\|_1$$

And so adjusting R by constant factors will give us the theorem statement.

So what remains is to prove the three facts assumed earlier.

2.2 First fact

Recall we want to show that for any fixed x, $\|RAx\|_1 \geq \|Ax\|_1$ w.p. $1 - \exp(-d \log d)$. Observe that for any row of $r \in R$,

$$\langle r, Ax \rangle = \frac{\|Ax\|_1 Z}{d \log d}$$

where Z is Cauchy (this follows from the 1-stability of Cauchy random variables). So then $RAx = (\|Ax\|_1 Z_1, \ldots, \|Ax\|_1 Z_{d \log d})/(d \log d)$, where each of the Z's are iid Cauchy. It follows that $\|RAx\|_1 = \|Ax\|_1 \sum_j \frac{|Z_j|}{d \log d}$, where the $|Z_j$ are half-Cauchy.

Notice that $P(|Z_i|_1 \geq \frac{1}{10}) \geq \frac{1}{2}$ (from integration). Then let $Y_i = 1$ if $|Z_i| \geq \frac{1}{10}$ and 0 otherwise. We then get

$$P(\sum_{i=1}^{d \log d} |Z_i| \leq \frac{d \log d}{20}) \leq P(\sum_{i=1}^{d \log d} Y_i \leq \frac{\mathbb{E}[\sum Y_i]}{2}) \leq e^{-\theta(E[\sum Y_i])} = e^{-\theta(d \log d)}$$

Where the first inequality comes from knowing that $\sum_{i=1}^{d \log d} |Z_i| \leq \frac{d \log d}{20} \implies \sum_{i=1}^{d \log d} Y_i \leq \frac{\mathbb{E}[\sum Y_i]}{2}$. So altogether, we get that

$$\begin{aligned}
\|RAx\|_1 &= \|Ax\|_1 \sum_j |Z_j|/(d \log d) \\
&\geq \|Ax\|_1 \Omega(d \log d)/(d \log d) \\
&= \Omega(1) \|Ax\|_1
\end{aligned}$$

Adjusting R by some constant factor will give you the conclusion you want.

2.3 Second fact

This was actually proven in the previous lecture. Check those notes for that proof.

2.4 Third fact

First notice that (by triangle inequality, bounding x_i, first property of Auerbach basis)

$$\begin{aligned}
\|RAx\|_1 &\leq \sum_i \|RA_{*,i} x_i\|_1 \\
&\leq \|x\|_\infty \sum_i \|RA_{*,i}\|_1 \\
&= \sum_i \|RA_{*,i}\|_1 \|Ax\|_1
\end{aligned}$$

So it suffices to prove $\sum_i \|RA_{*,i}\|_1 = O(\log d) \sum_i \|A_{*,i}\|_1$. Notice this implies $\sum_i \|RA_{*,i}\|_1 = O(d \log d)$ because for an Auerbach basis $\sum_i \|A_{*,i}\|_1 = d$.

We have

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

Let $E_{i,j}$ be the event that $|Z_{i,j}| \le d^3$, and define $Z'_{i,j} = |Z_{i,j}|$ if $E_{i,j}$ and $Z'_{i,j} = d^3$ otherwise. We know

$$\begin{aligned} E[|Z_{i,j}||E_{i,j}] &= E[Z'_{i,j}|E_{i,j}] \\ &= O(\log d) \end{aligned}$$

Where we get the last equality by just writing out the integral. Let E be the event that for all i, j, $E_{i,j}$ occurs. Then we know that

$$P(E) \ge 1 - \sum_{i,j} P(\text{not } E_{i,j}) \ge 1 - d^2 \log(d)/d^3 = 1 - \frac{\log(d)}{d}$$

Observe that

$$\begin{aligned} O(\log d) = E[Z'_{i,j}|E_{i,j}] &= E[Z'_{i,j}|E_{i,j}, E]P(E|E_{i,j}) + E[Z'_{i,j}|E_{i,j}, \text{not } E]P(\text{not } E|E_{i,j}) \\ &\ge E[Z'_{i,j}|E_{i,j}, E]P(E|E_{i,j}) \\ &= E[Z'_{i,j}|E_{i,j}, E]\frac{P(E_{i,j}|E)P(E)}{P(E_{i,j})} \\ &\ge E[Z'_{i,j}|E]P(E) \\ &\ge E[Z'_{i,j}|E](1 - \frac{\log(d)}{d}) \end{aligned}$$

Which implies that

$$E[Z'_{i,j}|E] = O(\log d)$$

So with this, we can see that

$$E[\sum_i \|RA_{*,i}\|_1|E] = \sum_i \|A_{*,i}\|_1 O(\log d)$$

So we get

$$\begin{aligned} P(\sum_i \|RA_{*,i}\|_1 \le \frac{\sum_i \|A_{*,i}\|_1 O(\log d)}{t}) &\ge P(\sum_i \|RA_{*,i}\|_1 \le \frac{\sum_i \|A_{*,i}\|_1 O(\log d)}{t}|E)P(E) \\ &\ge (1 - \frac{\log d}{d})\frac{1}{t} \\ &= O(1) \end{aligned}$$

Where the second inequality comes from Markov's inequality. So with constant probability, we get $\sum_i \|RA_{*,i}\|_1 = O(\log d) \sum_i \|A_{*,i}\|_1$, which is exactly what we wanted to show.