For context, this continues the quest that the previous scribe notes started, of performing $\ell_1$ regression fast, that is, performing the task of minimizing $\|Ax - b\|_1$ by suitable choice of $x$.

1 The Cauchy Distribution

We survey a bunch of the properties of the Cauchy distribution in this section.

The Cauchy distribution, denoted $\psi$, has probability density function

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

for $z \in \mathbb{R}$. Note that this distribution has undefined expectation and infinite variance.

A notable property of the Cauchy distribution is that it is 1-stable: if $z_1, z_2, \ldots, z_n$ are i.i.d. Cauchy, then for all $a \in \mathbb{R}^n$, we have $a_1z_1 + \ldots + a_n z_n \sim \|a\|_1 z$ where $z$ is Cauchy. We don’t prove this here but drop the words ‘Fourier transform’ and ‘convolution’ as hints.

A natural question to ask is how one can generate a Cauchy random variables and it turns out this can be done by taking the ratio of two standard normal random variables.

Recall that $R$ was the dense sketching matrix made of i.i.d. Cauchy random variables from the previous scribe notes. By 1-stability for all rows $r$ of $R$,

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

where $Z$ is Cauchy. In particular, we have

$$\frac{1}{d \log d} \left[ \|Ax\|_1 Z_1 \cdots \|Ax\|_1 Z_{d \log d} \right]$$

where $Z_1, Z_2, \ldots, Z_{d \log d}$ are i.i.d. Cauchy.

Now that we have established some basic properties of the Cauchy distribution, for the purposes of regression, we study the term $\|RAx\|_1$ and prove the sketching theorem.

Proof of Sketching Theorem. We point out that $\|RAx\|_1 = \|Ax\|_1 \sum_j |Z_j|$ where $|Z_j|$ are half-Cauchy. The value of $\sum_j |Z_j|$ is $\Omega(d \log d)$ with probability $1 - e^{-cd \log d}$. Indeed, the value of $|Z_j|$ is at least 1 with some constant probability $p$, and by Chernoff bounds, the the fraction of $|Z_j|$ with value less than 1 is at least $\frac{p}{2}$ with probability at least $1 - e^{-cd \log d}$ by Chernoff bounds. Note that this establishes that for a fixed $x$, we have $\|RAx\|_1 \geq \|Ax\|_1$ with probability $1 - e^{-cd \log d}$. We shall use a $\gamma$-net argument to generalize this to all $x$. 

1
We would be very happy if we could say a statement like \( \sum_j |Z_j| \) is \( O(d \log d) \) with high probability, but such a thing isn’t true since the Cauchy distribution is heavy tailed, so an alternate approach of proof is called for.

Note that there exists a well conditioned basis of \( A \) and without loss of generality say the basis vectors are \( A_{*,1}, A_{*,2}, \ldots, A_{*,d} \), then we have

\[
\|RA_{*,i}\|_1 = \sum_j |r_j \cdot A_{*,i}| = \|A_{*,i}\|_2 \frac{\sum_j |Z_{i,j}|}{d \log d}
\]

where \( Z_{i,j} \) is entry \((j,i)\) of the matrix \( RA \).

Let \( E_{i,j} \) be the event that \( |Z_{i,j}| \leq d^3 \). Define \( Z'_{i,j} = |Z_{i,j}| \) if \( |Z_{i,j}| \leq d^3 \) and \( Z'_{i,j} = d^3 \) otherwise.

We now analyze \( \mathbb{E}[Z_{i,j}|E_{i,j}] = \mathbb{E}[Z'_{i,j}|E_{i,j}] \), which is given by

\[
\mathbb{E}[Z'_{i,j}|E_{i,j}] = \int_0^{d^3} \frac{2z}{\pi(1+z^2) \mathbb{P}[E_{i,j}]} \, dz = \frac{2}{\pi \mathbb{P}[E_{i,j}]} \int_0^{d^3} \frac{z}{1+z^2} \, dz = \frac{2}{\pi \mathbb{P}[E_{i,j}]} \log |d^3|_1 + \Theta(1) = \Theta(\log d)
\]

Let \( E \) be the event that for all \( i, j \) the event \( E_{i,j} \) occurs. We know that

\[
\mathbb{P}[\overline{E}] \leq \frac{d^2 \log d}{d^3} = \frac{\log d}{d}
\]

And this means that

\[
\mathbb{P}[E] \geq 1 - \frac{\log d}{d}
\]

The goal is to show the sketching theorem by assuming \( E \) occurs: then the chance of \( E \) not occurring would get absorbed in the \( \frac{1}{100} \) failure probability that the sketching theorem permits.

Thus, towards this goal, we analyze \( \mathbb{E}[Z'_{i,j}|E_{i,j}] \).

\[
\mathbb{E}[Z'_{i,j}|E_{i,j}] = \mathbb{E}[Z'_{i,j}|E_{i,j}, E] \mathbb{P}[E|E_{i,j}] + \mathbb{E}[Z'_{i,j}|E_{i,j}, \overline{E}] \mathbb{P}[\overline{E}|E_{i,j}]
\]

\[
\geq \mathbb{E}[Z'_{i,j}|E_{i,j}, E] \mathbb{P}[E|E_{i,j}]
\]

\[
= \mathbb{E}[Z'_{i,j}|E] \left( \frac{\mathbb{P}[E_{i,j}|E] \mathbb{P}[E]}{\mathbb{P}[E_{i,j}]} \right)
\]

\[
\geq \mathbb{E}[Z'_{i,j}|E] \left( 1 - \frac{\log d}{d} \right)
\]

So, after all the dust settles, we get \( \mathbb{E}[Z'_{i,j}|E] = O(\log d) \).

Note that we have

\[
\|RA_{*,i}\|_1 = \|A_{*,i}\|_1 \frac{\sum_j |Z_{i,j}|}{d \log d}
\]
and since the expected value of the above expression is \( \|A_i\|_1 \log d \), we know that the expected value of \( \sum_i \|RA_i\|_1 \) is \( \log d \sum_i \|A_i\|_1 \), which means with constant probability \( \sum_i \|RA_i\|_1 \) is at most \( O(\log d) \sum_i \|A_i\|_1 \) via Markov’s inequality.

We reiterate that \( A_1, \ldots, A_d \) is a well-conditioned basis, and in the earlier scribe notes, we showed the existence of such a basis. We will use the Auerbach basis, which always exists and satisfies two properties: for all \( x \), \( \|x\|_\infty \leq \|Ax\|_1 \) and \( \sum_i \|A_i\|_1 = d \).

To see why this basis is well-conditioned, consider the following.

\[
\frac{\|x\|_1}{d} \leq \|x\|_\infty
\leq \|Ax\|_1
\leq \left\| \sum_{i=1}^d A_ix_i \right\|_1
\leq \sum_{i=1}^d \|A_i\|_1 |x_i|
\leq \|x\|_\infty \cdot \sum_{i=1}^d \|A_i\|_1
= d\|x\|_\infty \leq d\|x\|_1
\]

We know that \( \sum_i \|RA_i\|_1 \) is, with constant probability, at most \( O(\log d) \sum_i \|A_i\|_1 \), which by the Auerbach basis is \( O(d \log d) \). Thus, for all \( x \), we have

\[
\|RAx\|_1 \leq \sum_i \|RA_ix_i\|_1 \leq \|x\|_\infty \sum_i \|RA_i\|_1 = \|x\|_\infty O(d \log d) \leq O(d \log d) \|Ax\|_1
\]

Now, we look back at what we have done so that we don’t lose sight of the high level picture. First of all, we showed a constant probability upper bound of \( O(d \log d) \|Ax\|_1 \) on \( \|RAx\|_1 \) for all \( x \). And we showed that for any fixed \( x \), \( \|RAx\|_1 \geq \|Ax\|_1 \) with probability at least \( 1 - e^{-\alpha d \log d} \), claiming we would generalize this to all \( x \) with a \( \gamma \)-net argument.

Set \( \gamma = \frac{1}{d \log^3 d} \) so we get a \( \gamma \)-net with \( |M| \leq d^{O(d)} \). By union bound on all \( y \in M \), we have \( \|Ry\|_1 \geq \|y\|_1 \). Suppose we take \( x \) with unit \( \ell_1 \) norm. There is \( y \in M \) satisfying \( \|Ax - y\|_1 \leq \gamma = \frac{1}{d \log^3 d} \).

Now, from a chain of inequalities we get

\[
\|RAx\|_1 \geq \|Ry\|_1 - \|R(Ax - y)\|_1
\geq \|y\|_1 - O(d \log d) \|Ax - y\|_1
\geq \|y\|_1 - O(d \log d) \gamma
\geq \|y\|_1 - O\left( \frac{1}{d^2} \right)
\geq \frac{|y|_1}{2}
\]

The last inequality is justified by \( \|Ax\|_1 \geq \|x\|_\infty \geq \frac{|x|_d}{d} = \frac{1}{d} \).
Note that since $\|Ax - y\|_1 \leq \gamma$, which means $\|Ax - y\|_1 + \|y\|_1 - \gamma \leq 2\|RAx\|_1$. It follows that $\frac{1}{2}\|Ax - y\|_1 + \frac{1}{2}\|y\|_1 \leq 2\|RAx\|_1$. By triangle inequality, we get that
\[
\frac{1}{2}\|Ax\|_1 \leq 2\|RAx\|_1
\]
which means $\|RAx\|_1 \geq \frac{1}{2}\|Ax\|_1$.

\[\blacksquare\]