1 \( \ell_1 \) Regression

The Rough Algorithm

As we have seen in leverage score sampling for \( \ell_2 \) regression, we can also solve \( \ell_1 \) regression from sampling rows. Uniform sampling is a bad idea, since we might miss rows with large values, causing the error to be large. Sampling based on the norms of rows is insufficient too, as the rows can have the same values but could have different directions. In \( \ell_2 \) regression, leverage score sampling solves the problem, where we sample the rows based on the \( \ell_2 \) norms of an orthonormal basis. Similarly, in \( \ell_1 \) regression, we would sample the rows based on the \( \ell_1 \) norms of a “well-conditioned” basis.

We break down the algorithm into the following steps:

1. Compute a poly\((d)\) approximation. We find \( \hat{x} \) such that:

\[
\| A\hat{x} - b \|_1 \leq \text{poly}(d) \min_x \| Ax - b \|_1
\]

Let \( \hat{b} = A\hat{x} - b \) be the residual. Suppose

\[
x^* = \text{argmin}_x \| Ax - b' \|_1.
\]

Then

\[
x^* + \hat{x} = \text{argmin}_x \| Ax - b \|_1.
\]

Therefore, we can rewrite the original problem \( \min_x \| Ax - b \|_1 \) as \( \min_x \| Ax - \hat{b} \|_1 \).

2. In parallel to step 1, we also compute a well-conditioned basis \( U \) such that \( A = UW \).

\[
\frac{\|x\|_1}{\text{poly}(d)} \leq \|Ux\|_1 \leq \text{poly}(d) \|x\|_1
\]

We can further rewrite the problem as

\[
\min_x \|Ux - b'\|_1.
\]

3. Sample rows from \([U \ b']\) according to the \( \ell_1 \) norm of the rows. This step can be done in \( \text{nnz}(A) \) time.

4. Solve \( \ell_1 \) regression on the sample, where we only have \( \text{poly}(d/\epsilon) \) rows. We can solve this with linear programming in \( \text{poly}(d/\epsilon) \) time.

We will focus how to efficiently compute the first 2 steps.
Sketching Theorem

**Theorem 1.1.** There is a probability space over $(d \log d) \times n$ matrices $R$ such that for any $n \times d$ matrix $A$, with probability $\geq \frac{99}{100}$, for all $x$,

$$\|Ax\|_1 \leq \|RAx\|_1 \leq d \log d \|Ax\|_1.$$

Here, $R$ can be seen as an $\ell_1$ norm subspace embedding for $A$, as $R$ is linear, independent of $A$, and preserves the length of an infinite number of vectors.

Applying the Sketching Theorem

Once we have the matrix $R$ from the Sketching Theorem, we can compute a $d \log d$-approximation and a well-conditioned basis efficiently.

1. Computing a $d \log d$-approximation:
   - (a) Compute $RA$ and $Rb$.
   - (b) Solve $\hat{x} = \text{argmin}_x \|RAx - Rb\|_1$. $\hat{x}$ is a $d \log d$-approximation.

\[
\|A\hat{x} - b\|_1 \leq \|RA\hat{x} - Rb\|_1 \leq \|RAx^* - Rb\|_1 \leq O(d \log d) \|Ax^* - b\|_1
\]

From the Sketching Theorem, where $R$ sketches for $[A, b]$.

Here, $x^* = \text{argmin}_x \|Ax - b\|_1$. Therefore, $\hat{x}$ is a $d \log d$-approximation.

Since $R$ has $O(d \log d)$ rows, we can solve $\text{min}_x \|RAx - Rb\|_1$ using linear programming efficiently.

2. Computing a well-conditioned basis:
   - (a) Compute $RA$.
   - (b) Compute the SVD of $RA = QW^{-1}$, where $Q$ is orthonormal.
   - (c) $U = AW$ is a well-conditioned basis.

\[
\|AWx\|_1 \leq \|RAWx\|_1 = \|Qx\|_1 \\
\leq \sqrt{d \log d} \|Qx\|_2 \\
\leq \sqrt{d \log d} \|x\|_2 \\
\leq \sqrt{d \log d} \|x\|_1
\]

From the Sketching Theorem

$\therefore Q = RAW$

$\therefore \|y\|_1 \leq \sqrt{m} \|y\|_2, \forall y \in \mathbb{R}^m$

$\therefore Q$ is orthonormal

$\therefore \|y\|_2 \leq \|y\|_1, \forall y \in \mathbb{R}^m$
\[ \|AWx\|_1 \geq \frac{1}{d \log d} \|RAWx\|_1 \quad \text{From the Sketching Theorem} \]
\[ = \frac{1}{d \log d} \|Qx\|_1 \quad \therefore Q = RAW \]
\[ \geq \frac{1}{d \log d} \|Qx\|_2 \quad \therefore \|y\|_2 \leq \|y\|_1, \forall y \in \mathbb{R}^m \]
\[ \geq \frac{1}{d \log d} \|x\|_2 \quad \therefore Q \text{ is orthonormal} \]
\[ \geq \frac{1}{d^2 \log d} \|x\|_1 \quad \therefore \|y\|_1 \leq \sqrt{m} \|y\|_2, \forall y \in \mathbb{R}^m \]

Note that this method is similar to how we compute leverage scores efficiently. Here, we take advantage of the fact that when the dimensions are small, we can approximate \( \ell_1 \) norm to \( \ell_2 \) norm. Therefore, computing SVD in the \( \ell_2 \) sense works.

**Proof of the Sketching Theorem**

One dense \( R \) that works for the Sketching Theorem is a \((d \log d) \times n\) matrix, where each entry is an i.i.d. Cauchy random variables, scaled by \( \frac{1}{d \log d} \).

**Cauchy Random Variables**

Cauchy Random Variables have distribution of
\[ \text{pdf}(z) = \frac{1}{\pi (1 + z^2)}, \forall z \in \mathbb{R}. \]

This is a “heavy tail” distribution, as the tail goes down in \( \Theta \left( \frac{1}{z^2} \right) \) instead of \( \Theta \left( e^{-z^2} \right) \) like the Gaussian random variables.

Sadly, due to its heavy tail property, the expectation and variance of the Cauchy random variables are undefined. However, a good property of it is the 1-stability, i.e., suppose \( z_1, z_2, \ldots z_n \) are i.i.d Cauchy random variables, then for any \( a \in \mathbb{R}^n \),
\[ a_1 z_1 + a_2 z_2 + \cdots + a_n z_n = \|a\|_1 z, \text{ where } z \text{ is also a Cauchy random variable.} \]

We can generate Cauchy random variables efficiently, as they are the ratio of two standard normal random variables.

We now prove that \( R \) with Cauchy random variables works for the Sketching Theorem. We would prove the lower bound first. That is,
\[ \|RAx\|_1 \geq \|Ax\|_1. \]

\(^1\)Note that this is analog to the 2-stability of Gaussian random variables. Suppose we have \( g_i \sim N(0, 1) \). Then for all \( a \in \mathbb{R}^n \), we have \( \sum a_i g_i \sim N(0, \sum a_i^2) \), which is a Gaussian scaled by the \( \ell_2 \) norm of \( a \).
Suppose $r_i$ is the $i$th row of $R$.

\[
RAx = \begin{bmatrix}
\langle r_1, Ax \rangle \\
\langle r_2, Ax \rangle \\
\vdots \\
\langle r_{d \log d}, Ax \rangle \\
\end{bmatrix} = \\
\begin{bmatrix}
\|Ax\|_1 Z_1/(d \log d) \\
\|Ax\|_1 Z_2/(d \log d) \\
\vdots \\
\|Ax\|_1 Z_{d \log d}/(d \log d) \\
\end{bmatrix}
\]

$Z_i$s are i.i.d Cauchy

Therefore,

\[
\|RAx\|_1 = \frac{1}{d \log d} \|Ax\|_1 \sum_{j=1}^{d \log d} |Z_j|, \quad |Z_j| \text{ are half-Cauchy}
\]

To lower bound $\sum_{j=1}^{d \log d} |Z_j|$, we apply the Chernoff Bound. The Chernoff bound can be stated as the following:

**Theorem 1.2.** Suppose $Y_1, Y_2, \ldots, Y_r$ are i.i.d random variables where

\[
Y_i = \begin{cases}
1 & \text{w.p. } p \\
0 & \text{w.p. } 1 - p.
\end{cases}
\]

The expectation of $\sum_{i=1}^{r} Y_i$ is $E[\sum_{i=1}^{r} Y_i] = pr$. The Chernoff bound states that:

\[
\Pr \left[ \sum_{i=1}^{r} Y_i \leq \frac{1}{2} pr \right] \leq e^{-\Theta(pr)}
\]

Therefore, to lower bound $\sum_{j=1}^{d \log d} |Z_j|$, we define

\[
Y_i = \begin{cases}
1 & \text{if } |Z_i| \geq 1 \quad (\text{This happens with probability } p = \Omega(1)) \\
0 & \text{otherwise.}
\end{cases}
\]

Applying the Chernoff Bound, we get:

\[
\Pr \left[ \sum_{j=1}^{d \log d} |Z_j| \leq \frac{1}{2} p \cdot d \log d \right] \leq \Pr \left[ \sum_{j=1}^{d \log d} Y_j \leq \frac{1}{2} p \cdot d \log d \right]
\]

\[
\leq e^{-\Theta(p \cdot d \log d)} \\
\leq e^{-\Theta(d \log d)} \\
\therefore p = \Omega(1)
\]

From Chernoff Bound
This means with probability $\geq 1 - \Theta(d \log d)$, $\sum_{j=1}^{d \log d} |Z_j| = \Omega(d \log d)$. Since

$$
\|RAx\|_1 = \frac{1}{d \log d} \|Ax\|_1 \sum_{j=1}^{d \log d} |Z_j|,
$$

with appropriate chosen constant $C$ for $\sum_{j=1}^{d \log d} |Z_j| \geq C(d \log d)$, we can show the lower bound $\|RAx\|_1 \geq \|Ax\|_1$. Note that this only holds for a fixed $x$. To prove that this works for all $x$, we would have to apply this to a net. Therefore, the small failure probability $e^{-\Theta(d \log d)}$ is important here, as we would have to take a union bound on all the points in the net later.

We now try to show the upper bound:

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

Since $|Z_j|$ are heavy tailed, $\|RAx\|_1$ may be large. The c.d.f of $|Z_j|$ are asymptotic to $1 - \Theta(\frac{1}{z})$.

Therefore, the failure probability is too large to union bound all the points in a net. We have to do something different from the net argument.

To do so, we will use the existence of a well-conditioned basis $A$, which we proved in the previous lecture. Since we are proving $\|Ax\|_1 \leq \|RAx\|_1 \leq O(d \log d) \|Ax\|_1$, we can choose any basis for $A$. Therefore, we will take $A$ as a well-conditioned basis, where the columns are $A_1, A_2, \ldots, A_{nd}$.

Observe that $\|RA_{\ast i}\|_1 = \|A_{\ast i}\|_1 \sum_j |Z_{i,j}| / (d \log d)$. To bound this, we want to truncate the heavy tail of $|Z_j|$. Consider the event $E_{i,j}$ where $|Z_{i,j}| \leq d^3$. We define

$$
|Z'_{i,j}| = \begin{cases} |Z_{i,j}| & \text{if } |Z_{i,j}| \leq d^3 \\ d^3 & \text{otherwise} \end{cases}
$$

Then

$$
E[Z_{i,j}|E_{i,j}] = E[Z'_{i,j}|E_{i,j}]
$$

$$
= \frac{\int_0^d \frac{2z}{\pi(1+z^2)} dz}{1 - \Theta\left(\frac{1}{d}\right)}
$$

$$
= \frac{1}{1 - \Theta\left(\frac{1}{d}\right)} \left( \Theta\left(\frac{1}{d}\right) \left( \int_0^1 \frac{z}{1+z^2} dz + \int_1^{d^3} \Theta\left(\frac{1}{z} dz\right) \right) \right)
$$

$$
= O(\log d)
$$

Let $E$ be the event that all $E_{i,j}$ occurs. By union bound, we have $\Pr[E] \geq 1 - \frac{\log d}{d}$. We want to compute $E[Z'_{i,j}|E]$. Since $Z_{i,j}$ are not independent to each other (the rows are independent, but the columns are not), we cannot simply take the value of $E[Z_{i,j}|E_{i,j}]$.

---

1Since $\int_{\frac{1}{d}}^{\infty} \frac{1}{\pi(1+y^2)} dy = \int_{\frac{1}{d}}^{\infty} \Theta\left(\frac{1}{y^2}\right) dy = \Theta\left(\frac{1}{d}\right)$, the c.d.f is $1 - \Theta\left(\frac{1}{d}\right)$. 

5