Lecture 6 Part 1— February 22

Prof. David Woodruff *Scribe: Mahbod Majid*

1 ℓ_1 Regression

The ℓ_1 norm is defined as follows:

Definition (ℓ_1 Regression)**.** Given an $n \times d$ matrix A and a vector $b \in \mathbb{R}^n$, find $x \in \mathbb{R}^d$ such that

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_1.$$

The ℓ_1 regression problem can be solved optimally in time $\mathcal{O}(nd)$, using linear programming. However, this would be computationally expensive for large n. In order to solve this problem more efficiently, we can use sketching.

1.1 Well-Conditioned Bases

In the ℓ_2 case we knew that ℓ_2 norm of vectors are preserved under orthonormal transformations. However, in the ℓ_1 case, we need to find a well-conditioned basis. More specifically, for an $n \times d$ matrix A, we can choose an $n \times d$ matrix U with orthonormal columns such that $A = UW$, and $\|Ux\|_2 = \|x\|_2$ for all x. Drawing inspiration from the ℓ_2 case we can ask the following question for the ℓ_1 case:

> Given a matrix A, can we find a matrix U for which $A = UW$ and $\|Ux\|_1 \approx \|x\|_1$ for all x?

For simplicity, we can define the following norm for a vector x and a full rank matrix Q:

Definition (($Q, 1$)-norm)**.** Assume Q is a matrix with full rank, then for a vector $z \in \mathbb{R}^d$, we define its $(Q, 1)$-norm as follows:

$$\|z\|_{Q,1} := \|Qz\|_1.$$

It can be shown that $\|\cdot\|_{Q,1}$ is a norm.

We can consider the unit ball of $\|\cdot\|_{Q,1}$, which is defined as follows: let $C := \left\{ z \in \mathbb{R}^d \;\middle|\; \|x\|_{Q,1} \leqslant 1 \right\}$ be the unit ball of $\|\cdot\|_{Q,1}$. It can be observed that C is a convex set which is symmetric about the origin. The following theorem shows that we can find an ellipsoid E such that $E \subseteq C \subseteq \sqrt{d}E$.

Theorem 1 (Lowner-John Ellipsoid)**.** *Let K be a convex body (a compact convex set with non-empty interior) in $\mathbb{R}^d$. Moreover, assume K is symmetric about the origin. Then there exists an ellipsoid E such that*

$$E \subseteq C \subseteq \sqrt{d}E,$$

where

$$E = \left\{ z \in \mathbb{R}^d \;\middle|\; z^\top F z \leqslant 1 \right\},$$

and $F = G^\top G$ for some $G \in \mathbb{R}^{d \times d}$.

As an application of the above theorem we can show the following lemma:

Lemma 1 (Löwner-John Ellipsoid for $(Q,1)$-norm)**.** *Let Q be a full rank $d \times d$ matrix. Then there exists a full rank $d \times d$ matrix G such that*

$$\forall z \in \mathbb{R}^d : \quad (z^\top F z)^{0.5} \leqslant \|z\|_{Q,1} \leqslant \sqrt{d}(z^\top F z)^{0.5},$$

where $F = G^\top G$.

Recall that our goal is to find a matrix U such that $\|Ux\|_1 \approx \|x\|_1$ for all x. We can use the lemma above to find such a matrix U.

Theorem 2 (Existence of Well Conditioned Basis)**.** *Given a full rank $d \times d$ matrix Q, there exists full rank matrices U and G such that $Q = UG$, and*

$$\forall x \in \mathbb{R}^d : \quad \frac{1}{\sqrt{d}}\|x\|_1 \leqslant \|Ux\|_1 \leqslant \sqrt{d}\|x\|_1.$$

Moreover, we call U with the above properties a well-conditioned basis for Q.

Proof. Let G be as in Lemma 1 for Q. Let $F = G^\top G$. Then for all $z \in \mathbb{R}^d$, we have

$$(z^\top F z)^{0.5} \leqslant \|z\|_{Q,1} \leqslant \sqrt{d}(z^\top F z)^{0.5}.$$

Let $U = QG^{-1}$, and take $z = G^{-1}x$. Then for all $x \in \mathbb{R}^d$, we have

$$(x^\top x)^{0.5} \leqslant \|Ux\|_1 \leqslant \sqrt{d}(x^\top x)^{0.5}.$$

Therefore,

$$\|x\|_2 \leqslant \|Ux\|_1 \leqslant \sqrt{d}\|x\|_2.$$

Note that $\|x\|_2 \leqslant \|x\|_1 \leqslant \sqrt{d}\|x\|_2$, so we have

$$\frac{1}{\sqrt{d}}\|x\|_1 \leqslant \|Ux\|_1 \leqslant \sqrt{d}\|x\|_1,$$

as desired. ■

1.2 Net for the Unit ℓ_1 Ball

Similar to the ℓ_2 case,another ingredient we need is a net for the unit ℓ_1 ball. Consider the unit ℓ_1 ball $B_1^d = \left\{ x \in \mathbb{R}^d \;\middle|\; \|x\|_1 \leqslant 1 \right\}$. We want to construct N such that it is a γ-net for B_1^d: for all $x \in B_1^d$, there exists $y \in N$ such that $\|x - y\|_1 \leqslant \gamma$.

Lemma 2. *There exists a γ-net N for B_1^d of size at most $(\frac{2+\gamma}{\gamma})^d$.*

Proof. We construct N greedily as follows: while there exists a point $x \in B$ of distance larger than γ from every point in N, include x in N. Now we use a volume argument to show that the size of N is small. The ℓ_1-ball of radius $\gamma/2$ around every point in N contained in the ℓ_1 all of radius $1 + \gamma/2$ around 0^d, and all such balls are disjoint.

Consider the volume ratio of the ℓ_1 ball of radius $1 + \gamma/2$ to the ℓ_1 ball of radius $\gamma/2$. We have

$$|N| \leqslant \frac{\mathrm{Vol}(B_1^d(1+\gamma/2))}{\mathrm{Vol}(B_1^d \gamma/2)} = (\frac{1+\gamma/2}{\gamma/2})^d = (\frac{2+\gamma}{\gamma})^d.$$

■

Our goal is to construct a cover for the unit ℓ_1 ball, using members of the image of U. Let N be a (γ/d)-net for the unit ℓ_1-ball B, as above. Let M be the transformation of N under U, i.e. $M = \{Ux \mid x \in N\}$. Note that $|M| \leqslant (1 + \gamma/(2d)^d)/(\gamma/(2d)^d)$. We claim that M is a γ-cover for unit ℓ_1 ball B.

Claim 1. Let $A = UW$ for a well conditioned basis U, and M the transformation of a (γ/d)-net N for the unit ℓ_1-ball B under U. Then M is a γ-cover for B.

Proof. Let $x \in B$. Then there exists $z \in N$ such that $\|x - z\|_1 \leqslant \gamma/d$. Then

$$\|Ux - Uz\|_1 \leqslant \sqrt{d}\|x - z\|_2 \leqslant \sqrt{d}\|x - z\|_1 \leqslant \sqrt{d}(\gamma/d) = \gamma.$$

Therefore, $Uz \in M$ is a γ-approximation to Ux, and hence M is a γ-cover for B. ■

Therefore, for a well-conditioned basis U, there exists a γ-cover M for the unit ℓ_1 ball B, with members of the image of U. Note that here $|M| \leqslant (d/\gamma)^{\mathcal{O}(d)}$, and this would lead to an additional $\log d$ factor compared to the ℓ_2 result.

1.3 Overview of the Algorithm

A naive method to solve the ℓ_1 regression problem is to solve the problem over a sampled subset of the rows of A. Uniform sampling of the rows of A is not a good idea, as we may miss a row of A that is very different from others. Recall that sampling proportional to the squared ℓ_2 norm of U in the ℓ_2 case, led to a good sampling strategy. We can use a similar strategy in the ℓ_1 case, by sampling proportional to the ℓ_1 norm of U, where $A = UW$, and U is a well-conditioned basis for A.

The steps to solve the ℓ_1 regression problem is as follows:

1. Compute poly(d)-approximation: Find x' such that $\|Ax' - b\|_1 \leqslant \mathrm{poly}(d) \min_{x \in \mathbb{R}^d} \|Ax - b\|_1$. Let $b' = b - Ax'$ be the residual. Then we have $\|A(x + x') - b\|_1 = \|Ax - b'\|_1$ for any $x \in \mathbb{R}^d$. This can be viewed as the original problem with a change of variables.

2. Compute well-conditioned basis: Compute U such that $A = UW$, and U is a well-conditioned basis for A: $\frac{1}{\mathrm{poly}(d)}\|x\|_1 \leqslant \|Ux\|_1 \leqslant \mathrm{poly}(d)\|x\|_1$ for all $x \in \mathbb{R}^d$. We can then consider the problem of $\min_{y \in \mathbb{R}^d} \|Uy - b'\|_1$. If y is a minimizer of this problem, then $x = W^{-1}y$ is a minimizer of $\min_{x \in \mathbb{R}^d} \|Ax - b'\|_1$.

3. Sample poly(d/ε) rows from U the well-conditioned basis and the residual b' proportional to their ℓ_1 norm. According to the two above steps minimizing $\|Ux - b'\|_1$ is equivalent to minimizing the original problem.

After taking these steps applying generic linear programming to the sampled rows of U and b' will be sufficient.

Now let's focus on showing how to perform the first two steps quickly.

1. Compute a poly(d)-approximation.
2. Compute a well-conditioned basis.

1.4 Sketching Theorem

The following theorem shows that sketching matrix distributions exist that embed a subspace up to a $d \log d$ factor in ℓ_1 norm.

Theorem 3 (ℓ_1 Embedding)**.** *There is a probability space over $(d \log d) \times n$ matrices R such that for and $n \times d$ matrix A, with probability at least 0.99, for all $x \in \mathbb{R}^d$,*

$$\|Ax\|_1 \leqslant \|RAx\|_1 \leqslant (d \log d)\|Ax\|_1.$$

Note that here R is linear, and is independent of A and it preservers the ℓ_1 norm of an infinite number of vectors.

Before proving the theorem, let's see how we may apply it to the ℓ_1 regression problem.

1.5 Application of Sketching Theorem

Suppose a sketching matrix R is given such that for all $x \in \mathbb{R}^d$, we have

$$\|Ax\|_1 \leqslant \|RAx\|_1 \leqslant (d \log d)\|Ax\|_1.$$

Then we can use RA and Rb to solve the ℓ_1 regression problem. We use this to compute a poly(d)-approximation to the ℓ_1 regression problem, and then compute a well-conditioned basis, efficiently

1.5.1 Computing a $d \log d$-approximation

The algorithm is as follows:

1. Compute RA, and Rb.
2. Solve the ℓ_1 regression problem for RA and Rb. Let x' be the solution. This can be done efficiently because R reduces the size and RA and Rb have $d \log d$ rows. Then we have

   $$\|Ax' - b\|_1 \leqslant \|RAx' - Rb\|_1 \leqslant \|RAx^* - Rb\|_1 \leqslant d \log d\|Ax^* - b\|_1,$$

   where x^* is the optimal solution to the original problem.

This gives us a poly(d)-approximation to the ℓ_1 regression problem.

1.5.2 Computing a well-conditioned basis

The algorithm is as follows:

1. RA.
2. Compute W such that RAW is orthonormal (in the ℓ_2 sense)
3. Output $U = AW$.

Then $U = AW$ will be a well-conditioned basis. To see this note that

$$\begin{aligned} \|AWx\|_1 &\leqslant \|RAW\|_1 \\ &\leqslant (d \log d)^{0.5} \|RAWx\|_2 \\ &\leqslant (d \log d)^{0.5} \|x\|_2 \\ &\leqslant (d \log d)^{0.5} \|x\|_1, \end{aligned}$$

and

$$\begin{aligned} \|AWx\|_1 &\geqslant \frac{1}{d \log d} \|RAW\|_1 \\ &\geqslant \frac{1}{d \log d} \|RAWx\|_2 \\ &\geqslant \frac{1}{d \log d} \|x\|_2 \\ &\geqslant \frac{1}{d^{3/2} \log d} \|x\|_1. \end{aligned}$$

1.6 Proof of Sketching Theorem

What is a good sketching matrix? Subgaussian random variables are *not* good for sketching in ℓ_1. We should look for a family of heavy tailed distributions. One such distribution is the Cauchy distribution. One choice of R that can be shown to work is as follows: the entries of R are i.i.d. Cauchy random variables scaled by $(d \log d)^{-1}$.

Definition (Cauchy Random Variable)**.** A random variable X is Cauchy distributed if it has the density function

$$f(x) = \frac{1}{\pi(1 + x^2)}.$$

This distribution has heavy tails, and is symmetric about the origin. Furthermore its expectation is undefined and the variance is infinite.

Recall that Gaussians are 2-stable. For Cauchy random variables it can be shown that they're 1-stable.

Fact 1 (Cauchy is 1-stable)**.** X_i's are independent Cauchy random variables, then for any $a \in \mathbb{R}^n$,

$$\sum_{i=1}^{n} a_i X_i \sim \|\sum_{i=1}^{n} a_i\|_1 \cdot Z,$$

where Z is a Cauchy random variable.

Now since Cauchy is 1-stable, we know that for every row r in R

$$\langle r, Ax \rangle = \|Ax\|_1 \cdot Z/(d \log d),$$

where Z is a Cauchy random variable. Then

$$RAx = (\|Ax\|_1 \cdot Z_1, \dots, \|Ax\|_1 \cdot Z_{d \log d}) /(d \log d),$$

where $Z_1, \dots, Z_{d \log d}$ are i.i.d. Cauchy random variables. Now we can write

$$\|RAx\|_1 = \|Ax\|_1 \sum_j |Z_j|/(d \log d),$$

where $|Z_j|$'s are i.i.d. half-Cauchy random variables. We are interested in proving upper and lower bounds on this quantity.

In order to prove lower bounds, let $X_j = \mathbb{1}[|Z_j| > 0.2]$, then $X_j's$ are i.i.d. Bernoulli random variables with $\mathbb{P}[X_j = 1] \geqslant 0.01$. Then we can apply a Chernoff's bound

$$\mathbb{P}\left[\sum_j X_j \leqslant 0.01 d \log d\right] \leqslant \exp(-\Theta(d \log d)).$$

Therefore,

$$\sum_j |Z_j| = \Omega(d \log d)$$

with probability $1 - \exp(-d \log d)$.

The other direction is more difficult, since $\sum_j |Z_j|$ is heavy-tailed.

Please refer to the next lecture for the rest of the proof.