CS 15-859: Algorithms for Big Data Fall 2020

Lecture 6.2 — October 15

Prof. David Woodruff *Scribe: Jeffrey Chen*

1 L1 Regression

We consider the problem of L1 regression, where we want to find an x^* that minimizes $\|Ax - b\|_1 = \sum_i |b_i - \langle A_{i,*}, x\rangle|$. L1 cost is less sensitive to outliers, since the differences between b and Ax are not squared. If $A \in \mathbb{R}^{n\times d}$, then L1 regression can be solved exactly using linear programming (LP) in time poly(nd).

We can formulate L1 regression as an LP problem. We introduce variables α^+ and α^- and solve for

$$\min[1,\ldots,1]\cdot(\alpha^+ + \alpha^-) \text{ subject to constraints } \begin{cases} Ax + \alpha^+ - \alpha^- = b \\ \alpha^+, \alpha^- \geq 0 \end{cases}$$

α^+ and α^- are $n \times 1$ vectors, and the objective is the same as saying we want to minimize the sum of the entries in $\alpha^+ + \alpha^-$. Note that for each index i, at most one of α_i^+ and α_i^- can be nonzero, since if both were greater than 0, then we can subtract the smaller value from both α_i^+ and α_i^- to still satisfy the constraints but have a lower cost. By solving this LP, the cost will be $\min_x \|Ax - b\|_1$, and the x that minimizes the cost will be x^*. There are d variables in x, $2n$ variables in α^+ and α^-, and n constraints, so the LP can be solved in poly(nd) time. n can be much larger than d, so we want to improve this time by sketching.

2 Well-Conditioned Bases

For an $n \times d$ matrix A, we can always find an $n \times d$ matrix U with orthonormal columns such that $A = UW$ and $\|Ux\|_2 = \|x\|_2$ for all x. This can be done by writing the SVD $A = U(\Sigma V^T)$. We want to find a different matrix U such that $A = UW$ and $\|Ux\|_1 \approx \|x\|_1$ for all x. Multiplying an $n \times d$ matrix with orthonormal columns by a change of basis matrix R^{-1} gives the same subspace, so this question is equivalent to finding some change of basis matrix R such that $AR = U$ and U preserves L1 norm.

One inequality we have relating L1 and L2 norms is

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2 \tag{1}$$

for all $x \in \mathbb{R}^n$. To see the first inequality,

$$\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2} \leq \sum_{i=1}^n \sqrt{x_i^2} = \sum_{i=1}^n |x_i| = \|x\|_1$$

and to see the second inequality,

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| = \sum_{i=1}^{n} |x_i| \cdot 1 \leq \sqrt{\sum_{i=1}^{n} |x_i|^2} \sqrt{\sum_{i=1}^{n} 1^2} = \sqrt{n}\|x\|_2$$

where we use Cauchy Schwarz in the second inequality.

Consider what happens if we just let U be an orthonormal basis for the columns of A. Then

$$\|Ux\|_1 \leq \sqrt{n}\|Ux\|_2 = \sqrt{n}\|x\|_2 \leq \sqrt{n}\|x\|_1$$

and

$$\|Ux\|_1 \geq \|Ux\|_2 = \|x\|_2 \geq \frac{1}{\sqrt{d}}\|x\|_1$$

Putting these together, we get $\frac{1}{\sqrt{d}}\|x\|_1 \leq \|Ux\|_1 \leq \sqrt{n}\|x\|_1$ or $\|x\|_1 \leq \|\sqrt{d}Ux\|_1 \leq \sqrt{nd}\|x\|_1$. This is undesirable because the approximation is off by a factor that depends on $\sqrt{n}$, which can be large.

2.1 Q,1 Norm

Let $A = QW$ for any matrix Q with full column rank, and define $\|z\|_{Q,1} = \|Qz\|_1$. We show that $\|\cdot\|_{Q,1}$ is a norm:

- $\|x\| = 0$ iff $x = 0$: True because $\|x\|_{Q,1} = \|Qx\|_1 = 0$ iff $x = 0$ since Q has full column rank.
- $\|cx\| = c\|x\|$ for all c: True because $\|cx\|_{Q,1} = \|Q \cdot cx\|_1 = c\|Qx\|_1 = c\|x\|_{Q,1}$.
- $\|x + y\| \leq \|x\| + \|y\|$ for all x, y: True because $\|x + y\|_{Q,1} = \|Q(x + y)\|_1 \leq \|Qx\|_1 + \|Qy\|_1 = \|x\|_{Q,1} + \|y\|_{Q,1}$.

Let $C = \{z \in \mathbb{R}^d : \|z\|_{Q,1} \leq 1\}$ be the unit ball of $\|\cdot\|_{Q,1}$. We claim C is a convex set that is symmetric about the origin. To show convexity, let $x, y \in C$. Any point in the line can be written as $\lambda x + (1 - \lambda)y$ for some $\lambda \in [0, 1]$. By properties of norms,

$$\|\lambda x + (1 - \lambda)y\|_{Q,1} \leq \|\lambda x\|_{Q,1} + \|(1 - \lambda)y\|_{Q,1} = \lambda\|x\|_{Q,1} + (1 - \lambda)\|y\|_{Q,1} \leq \lambda \cdot 1 + (1 - \lambda) \cdot 1 = 1$$

This shows any point on the line between x and y also lies in C.

To show that C is symmetric about the origin, note that if $z \in C$, then $\|z\|_{Q,1} \leq 1$ and

$$\|-z\|_{Q,1} = \|Q(-z)\|_1 = \|Qz\|_1 = \|z\|_{Q,1} \leq 1$$

so $-z$ is also in C.

Theorem 1 (Lowner-John Theorem)**.** *Let C be a convex set symmetric about the origin. Then we can find an ellipsoid $E = \{z \in \mathbb{R}^d : z^T F z \leq 1\}$ where $F = G^T G$ for some matrix G such that $E \subseteq C \subseteq \sqrt{d}E$.*

By this theorem, we have for all z that

$$(z^T F z)^{0.5} \leq \|z\|_{Q,1} \leq \sqrt{d}(z^T F z)^{0.5}$$

Define $U = QG^{-1}$, where $A = QW$ for some full column rank matrix Q and $F = G^T G$ from the Lowner-John Theorem. Note that $A = QW = QG^{-1}GW = UGW$.

Let $z = G^{-1}x$. Then $\|Ux\|_1 = \|QG^{-1}x\|_1 = \|Qz\|_1 = \|z\|_{Q,1}$. Plugging z into $(z^T F z)^{0.5}$, we get

$$(z^T F z)^{0.5} = (x^T (G^{-1})^T G^T G G^{-1} x)^{0.5} = (x^T x)^{0.5} = \|x\|_2$$

This shows $\|x\|_2 \leq \|Ux\|_1 \leq \sqrt{d}\|x\|_2$. Applying inequality (1), we get

$$\frac{\|x\|_1}{\sqrt{d}} \leq \|x\|_2 \leq \|Ux\|_1 \leq \sqrt{d}\|x\|_2 \leq \sqrt{d}\|x\|_1$$

or equivalently

$$\|x\|_1 \leq \|\sqrt{d}Ux\|_1 \leq d\|x\|_1$$

This shows that our chosen U (or the scaled $\sqrt{d}U$) preserves L1 norm up to a factor of d, which does not depend on n.

3 Net for ℓ_1 Ball

We revisit the idea of a net from week 1, this time with the L1 norm. Consider the ℓ_1-ball $B = \{x \in \mathbb{R}^d : \|x\|_1 = 1\}$. Subset N is a γ-net if for all $x \in B$, there is a $y \in N$ such that $\|x - y\|_1 \leq \gamma$. We can find a net N by the greedy algorithm: while there is a point x of distance greater than γ from every point in N, include x in N. The ℓ_1-ball of radius $\gamma/2$ around every point in N is contained in the ℓ_1-ball of radius $1 + \gamma/2$ around 0 by the triangle inequality. All of the balls are disjoint because if two balls were not disjoint, then their centers would be at a distance less than $\gamma/2 + \gamma/2 = \gamma$ apart, so we would not add one of the balls to the net by the greedy algorithm. This means

$$|N| \leq \frac{\text{Vol}(\ell_1\text{-ball with radius } 1 + \gamma/2)}{\text{Vol}(\ell_1\text{-ball with radius } \gamma/2)} \leq \left(\frac{1 + \gamma/2}{\gamma/2}\right)^d$$

Let $A = UW$ for a well-conditioned basis U, so $\|x\|_1 \leq \|Ux\|_1 \leq d\|x\|_1$ for all x. Let N be a γ/d-net for the unit ℓ_1-ball B. Let $M = \{Ux \mid x \in N\}$, so $|M| = |N| \leq \frac{(1+(\gamma/2d))^d}{(\gamma/2d)^d}$. We claim that for every $x \in B$, there is a $y \in M$ such that $\|Ux - y\|_1 \leq \gamma$. This is because we can find $x' \in N$ such that $\|x - x'\|_1 \leq \gamma/d$, so $\|Ux - Ux'\|_1 \leq d\|x - x'\|_1 \leq \gamma$ and we can set $y = Ux'$.

4 Rough Algorithm

Remember that to solve L1 regression, we want to minimize $Ax - b$. One way to reduce the size of the problem is to sample (using leverage score sampling, for example) some of the rows of $[A\, b]$. Uniform sampling is a bad idea for regression, since there may be an important row that is very different from the others that we end up not sampling, which would lead to large error. Sampling proportional to the squared norms of the rows is also insufficient, since the important row may

have the same norm as the other rows but be in a different direction. For L2 regression, sampling proportional to the squared L2 norms of U in the SVD of $[A\, b]$ gives more weight to important rows and leads to a good approximation. For L1 regression, we will show that the analogous sampling proportional to the L1 norms of a well-conditioned basis U will give a good approximation.

Our rough algorithm proceeds as follows:

1. Compute a poly(d)-approximation to the solution. This means finding an x' such that $\|Ax' - b\|_1 \leq \text{poly}(d) \min_x \|Ax - b\|_1$. Then compute the residual $b' = b - Ax'$. We will consider the optimization problem $\min_x \|Ax - b'\|_1$. Note that $\|Ax - b'\|_1 = \|Ax - (b - Ax')\|_1 = \|A(x + x') - b\|_1$ This problem is equivalent to the original problem, since if $x^* = \arg\min_x \|Ax - b'\|_1 = \arg\min_x \|A(x + x') - b\|_1$, then $x^* - x'$ minimizes the original problem $\|Ax - b\|_1$.

2. Compute a well-conditioned basis U for A such that $A = UW$ and for all $x \in \mathbb{R}^d$,

$$\|x\|_1 / \text{poly}(d) \leq \|Ux\|_1 \leq \text{poly}(d)\|x\|_1$$

   We can then consider the problem $\min_x \|Ux - b'\|_1 = \min_x \|AW^{-1}x - b'\|_1$. This problem is equivalent to minimizing $\|Ax - b'\|_1$, since if $\hat{x} = \arg\min_x \|Ux - b'\|_1 = \arg\min \|AW^{-1}x - b'\|_1$, then $W^{-1}\hat{x}$ minimizes $\|Ax - b'\|_1$.

3. Approximate the new problem $\min_x \|Ux - b'\|_1$ by sampling poly(d/ϵ) rows from $[U\, b']$ proportional to their L1 norm. We have shown that this new problem is equivalent to the original problem.

4. Solve L1 regression on the sample, which can be done in poly(d/ϵ) time with linear programming, and output x.

It remains to show that steps 1 and 2 above can be computed efficiently.

Theorem 2 (Sketching Theorem)**.** *There is a probability distribution 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 simultaneously for all x that*

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

This theorem shows that R is a subspace embedding for the L1 norm: it is linear, independent of A, and preserves the lengths of an infinite number of vectors. Before proving the theorem, we show that we can use this theorem to get both steps 1 and 2 of the rough algorithm.

1. Let $\hat{x} = \arg\min_x \|RAx - Rb\|_1$, which we can solve efficiently because R reduces the size of the problem to poly(d). Then

$$\|A\hat{x} - b\|_1 \leq \|R(A\hat{x} - b)\|_1 \leq \|R(Ax^* - b)\|_1 \leq d \log d \|Ax^* - b\|_1$$

   where $x^* = \arg\min_x \|Ax - b\|_1$ is the optimal solution. This $\hat{x}$ is a poly(d)-approximation to the problem, which is step 1.

2. Compute RA, and compute W such that RAW is orthonormal in the L2 sense. This can be done using SVD to get $RA = U\Sigma V^T$ and letting $W = (\Sigma V^T)^{-1}$. Then output $U = AW$ as the well-conditioned basis. This is similar to our method of preconditoning for L2 norm: we let $SA = QR^{-1}$ and showed that AR is close to orthonormal, meaning the condition number is close to 1.

   To show $U = AW$ is well-conditioned, note that

$$\begin{aligned}
\|Ux\|_1 = \|AWx\|_1 &\leq \|RAWx\|_1 && \text{By Theorem 2} \\
&\leq (d\log d)^{1/2}\|RAWx\|_2 && \text{By inequality 1, } R \text{ has } d\log d \text{ rows} \\
&= (d\log d)^{1/2}\|x\|_2 && RAW \text{ is orthonormal} \\
&\leq (d\log d)^{1/2}\|x\|_1 && \text{By inequality 1}
\end{aligned}$$

   and

$$\begin{aligned}
\|Ux\|_1 = \|AWx\|_1 &\geq \|RAWx\|_1/(d\log d) && \text{By Theorem 2} \\
&\geq \|RAWx\|_2/(d\log d) && \text{By inequality 1} \\
&= \|x\|_2/(d\log d) && RAW \text{ is orthonormal} \\
&\geq \|x\|_1/(d^{3/2}\log d) && \text{By inequality 1}
\end{aligned}$$

   so

$$\|x\|_1/(d^{3/2}\log d) \leq \|Ux\|_1 \leq (d\log d)^{1/2}\|x\|_1$$

We will prove Theorem 2 in the next lecture.