CS 15-859: Algorithms for Big Data	Fall 2017
Lecture 6-1 — 10/12/2017	
Prof. David Woodruff	*Scribe: Sidhanth Mohanty*

1 ℓ_1-regression preliminaries

Given a $n \times d$ matrix A, and a vector b, the goal of ℓ_1-regression is to find x^* minimizing the value of $\|Ax - b\|_1$. We note that in this problem, the cost is less sensitive to outliers than least squares regression, and depending on the application, this might be a desirable property making ℓ_1-regression more suitable than least squares.

We note that if our goal is simply to solve this problem in polynomial time, then we can do so by writing a linear program. We introduce two new vectors of variables $\boldsymbol{\alpha}^+$ and $\boldsymbol{\alpha}^-$ and solve the following LP.

$$\begin{aligned} \text{Minimize} \quad & \vec{\mathbf{1}} \cdot (\boldsymbol{\alpha}^+ + \boldsymbol{\alpha}^-) \\ \text{Subject to} \quad & Ax + \boldsymbol{\alpha}^+ + \boldsymbol{\alpha}^- = b \\ & \boldsymbol{\alpha}^+, \boldsymbol{\alpha}^- \geq 0 \end{aligned}$$

Note that this LP takes $\mathsf{poly}(n, d)$ time to solve.

However, as always, we are curious about whether we can solve this problem better or not. And a natural tool to turn to is sketching. And the answer is, yes, we can indeed solve this problem faster in approximation via sketching. As a first step we explore the geometry of ℓ_1 space a bit more.

2 Geometry of ℓ_1 space

2.1 Löwner-John Theorem and Well Conditioned Bases

All our fast algorithms ℓ_2 regression hinge on the following fact. For an $n \times d$ matrix A, we can always choose a $n \times d$ matrix U with orthonormal columns where A can be written as UW and $\|Ux\|_2 = \|Ax\|_2$ for all $x \in \mathbb{R}^d$.

And thus naturally, we ask if can we find a U for which $A = UW$ and $\|Ux\|_1 = \|x\|_1$. Unfortunately, this isn't possible. What is true, though, is that there is U with orthogonal columns and $A = UW$ where the following inequality holds:

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

and this is easy to prove via the Cauchy-Schwartz inequality. [1]

[1] $\|y\|_1$ can be written as the dot product of y and a vector where all entries are ± 1. Cauchy-Schwartz then tells us this is at most $\sqrt{n}\|y\|_2$.

This massive amount of distortion is unsuitable for our purposes, and we do some work to strengthen this inequality to

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

which ends up being good enough for our purposes.

Let's write A as QW where Q has full column rank, and define $\|z\|_{Q,1} = \|Qz\|_1$: $\|cdot\|_{Q,1}$ ends up being a norm.

Let $C = \{z \in \mathbb{R}^d : \|z\|_{Q,1} \leq 1\|$ be the unit ball of norm $\| \cdot \|_{Q,1}$. C is convex, which can be seen by taking $x, y \in C$ and $0 \leq \alpha \leq 1$ to get

$$\begin{aligned}\|\alpha x + (1-\alpha)y\|_{Q,1} &\leq \|\alpha x\|_{Q,1} + \|(1-\alpha)y\|_{Q,1} \\ &= \alpha\|x\|_{Q,1} + (1-\alpha)\|y\|_{Q,1} \\ &\leq 1\end{aligned}$$

which means $\alpha x + (1-\alpha)y$ is in C.

Another property of C is that it is symmetric around the origin, because for any $x \in C$, we have $\|-x\|_{Q,1} = \|x\|_{Q,1} \leq 1$.

Working towards our inequality, we state the Löwner-John Theorem.

Theorem 1 (Löwner-John)**.** *For all convex bodies C, we can find ellipsoid E such that $E \subseteq C \subseteq \sqrt{d}E$ where $E = \{z \in \mathbb{R}^d : z^T G^T G z \leq 1\}$.*

We take the unit ball under the $\| \cdot \|_{Q,1}$ norm, and use the Löwner-John theorem to find an ellipsoid such that

$$\sqrt{z^T G^T G z} \leq \|z\|_{Q,1} \leq \sqrt{d} z^T G^T G z$$

Next, we define U as QG^{-1}, where U satisfies what is called the 'well-conditioned basis' property. [2]

Take some x and define $z = G^{-1}x$. Then we get

$$\|Ux\|_1 = \|QG^{-1}x\|_1 = \|Qz\|_1 = \|z\|_{Q,1}$$

Now, note that

$$z^T G^T G z = x^T (G^{-1})^T G^T G (G^{-1}) x = x^T x = \|x\|_2^2$$

This gives us $\|x\|_2 \leq \|Ux\|_1 \leq \sqrt{d}\|x\|_2$, which from norm inequalities gives

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

2.2 Net for ℓ_1-ball

Consider the unit ℓ_1 ball B. We call $N \subseteq B$ a γ-net if for all $x \in B$, there is $y \in N$ such that $\|x - y\|_1 \leq \gamma$. A suitable N can be constructed greedily by picking any $y \in B$ where y is more than

[2] Note that G is invertible since the ellipsoid must be full dimensional as Q is full rank, which makes the unit ball of the norm we defined full dimensional.

γ away from all members of N, and adding it to N while there such a point y. We note that the ℓ_1 ball of radius $\frac{\gamma}{2}$ around every point is contained in the ℓ_1 ball of radius $1+\frac{\gamma}{2}$ around $\mathbf{0}$. We note that by our construction every pair of points in N are distance at least γ apart, which means that all the balls we drew around each point are disjoint.

Since the ratio of the volumes of d-dimensional ℓ_1 balls of radius $1+\frac{\gamma}{2}$ and $\frac{\gamma}{2}$ is upper bounded by

$$\frac{\left(1+\frac{\gamma}{2}\right)^d}{\left(\frac{\gamma}{2}\right)^d}$$

so the above quantity also serves as an upper bound to the size of N.

2.3 Net for ℓ_1 subspace

Let $A = UW$ for a well-conditioned basis U where vectors are scaled so that for all x

$$\|x\|_1 \leq \|Ux\|_1 \leq d\|x\|_1$$

Now, let N be a $\left(\frac{\gamma}{d}\right)$-net for the unit ℓ_1-ball B and let $M = \{Ux | x \in N\}$ so $|M| \leq \frac{\left(1+\frac{\gamma}{2d}\right)^d}{\left(\frac{\gamma}{2d}\right)^d}$. We claim that for every $x \in B$, there is $y \in M$ for which $\|Ux - y\| \leq \gamma$. To see why this is true, pick $x' \in N$ for which $\|x' - x\|_2 \leq \frac{\gamma}{d}$, and by the well-conditioned basis property of U, we know $\|U(x'-x)\|_1 \leq \gamma$ which means that $\|Ux - y\|_1 \leq \gamma$.

3 Towards an Algorithm

We start out with a very high level algorithm overview and delve into details afterwards.

- Compute $\mathsf{poly}(d)$-approximate solution to the problem: that is, we find x' such that

$$\|Ax' - b\|_1 \leq \mathsf{poly}(d) \min_{x \in \mathbb{R}^d} \|Ax - b\|_1$$

- Compute a well-conditioned basis, that is write A as UW with the condition that for all $x \in \mathbb{R}^d$ we have

$$\frac{\|x\|_1}{\mathsf{poly}(d)} \leq \|Ux\|_1 \leq \mathsf{poly}(d)\|x\|_1$$

- Compute $b' = b - Ax'$.
- Consider the optimization problem $\min_x \|Ux - b'\|_1$ and sample $\mathsf{poly}\left(\frac{d}{\varepsilon}\right)$ rows of this problem with probabilities proportional to the ℓ_1 norm of $U|b'$ (b' is appended as a column to U).
- Now run the generic LP on the subsampled rows: this runs efficiently.

As motivation for considering the optimization problem $\|Ux - b\|_1$, we show that

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

whose proof is via the following chain of equalities.

$$\begin{aligned}
\min_{x \in \mathbb{R}^d} \|Ax - b'\| &= \min_{x+x' \in \mathbb{R}^d} \|A(x + x') - b\|_1 \\
&= \min_{x \in \mathbb{R}^d} \|Ax + Ax' - b\|_1 \\
&= \min_{x \in \mathbb{R}^d} \|Ax - b'\|_1 \\
&= \min_{W^{-1}x \in \mathbb{R}^d} \|AW^{-1}x - b'\|_1 \\
&= \min_{x \in \mathbb{R}^d} \|Ux - b'\|_1
\end{aligned}$$

Once the first two steps and the sampling step are made more efficient, we are happy since the rest can be done in $\text{nnz}(A) + \mathsf{poly}\left(\frac{d}{\varepsilon}\right)$ time.

3.1 Finding an approximate solution fast

Now, towards the goal of finding an approximate solution quickly we consider the following theorem.

Theorem 2 (Sketching Theorem)**.** *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 $\frac{99}{100}$, we have for all x:*

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

We note that this embedding simultaneously enjoys linearity and preserving the lengths of an infinite number of vectors and is also independent of A.

We defer the proof to a later section, in the next set of notes.

The computation of a $d \log d$ approximation can be done in two simple steps:

- Compute RA and Rb.
- Solve $\arg\min_{x \in \mathbb{R}^d} \|RAx - Rb\|_1$.

The sketching theorem applied to $A \circ b$ implies that x' is a $d \log d$ approximation. Since RA and Rb have $d \log d$ rows, so we can solve the ℓ_1 regression problem efficiently.

3.2 Computing a well-conditioned basis

The following algorithm accomplishes it.

- Let R be a ℓ_1-sketching matrix. Compute RA.
- Compute W so that RAW is orthonormal.

- $U = AW$ is well-conditioned because

$$\|AWx\|_1 \leq \|RAWx\|_1 \leq \sqrt{d \log d}\|RAW\|_2 = \sqrt{d \log d}\|x\|_2 \leq \sqrt{d \log d}\|x\|_1$$

  and

$$\|AWx\|_1 \geq \frac{\|RAWx\|_1}{d \log d} \geq \frac{\|RAWx\|_1}{d \log d} = \frac{\|x\|_2}{d \log d} \geq \frac{\|x\|_1}{d^{3/2} \log d}$$

A dense R that works is when all the entries are i.i.d. Cauchy random variables, scaled by $\frac{1}{d \log d}$.