ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

L1 Regression

with Lewis Weights Subsampling

Aditya Parulekar
adityaup@cs.utexas.edu
UT Austin

Advait Parulekar
advaitp@utexas.edu
UT Austin

Eric Price
ecprice@cs.utexas.edu
UT Austin

APPROX/RANDOM 2021

Problem setting

- **Linear ℓ_1 regression**

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^d} \|X\beta - y\|_1$$

- **Active Linear ℓ_1 regression**

Sampling matrix

$$\widehat{\beta} := \arg\min \|SX\beta - Sy\|_1.$$

Such that the following holds with probability at least 1 – δ

$$\|X\widehat{\beta} - y\|_1 \leq (1+\varepsilon)\min_{\beta} \|X\beta - y\|_1.$$

The method

Definition 2.1 (Subspace Embeddings). *A subspace embedding for the column space of the matrix $X \in \mathbb{R}^{n\times d}$ is a matrix S such that for all $\beta \in \mathbb{R}^d$,*

$$\|SX\beta\| = (1 \pm \varepsilon)\|X\beta\|$$

- **If we had access to all of y, we can find a subspace Embedding S for the combined matrix [X y] to solve the problem**

$$\begin{aligned}
\|X\widehat{\beta} - y\|_1 &\leq \frac{1}{1-\varepsilon}\|SX\widehat{\beta} - Sy\|_1 \\
&\leq \frac{1}{1-\varepsilon}\|SX\beta^* - Sy\|_1 \\
&\leq \frac{1+\varepsilon}{1-\varepsilon}\|X\beta^* - y\|_1 \\
&\leq (1+4\varepsilon)\|X\beta^* - y\|_1.
\end{aligned}$$

ParNeC

Definition 2.5 $\boxed{\text{(Lewis Weights)}}$. *The ℓ_1 Lewis weights of a matrix X are the unique weights $\{w_i\}_{i=1}^n$ that satisfy $w_i^2 = x_i^\top(\sum_{j=1}^n \frac{1}{w_j} x_j x_j^\top)^{-1} x_i$ for all i.*

Definition 2.3 (Sampling and Reweighting with $\{p_i\}_{i=1}^n$). *For any sequence $\{p_i\}_{i=1}^n$, let $N = \sum_i p_i$. Then, the sampling-and-reweighting distribution $\mathcal{S}(\{p_i\}_{i=1}^n)$ over the set of matrices $S \in \mathbb{R}^{N \times n}$ is such that each row of S is independently the ith standard basis vector with probability $\frac{p_i}{N}$, scaled by $\frac{1}{p_i}$. For any $k \in [N]$, let i_k denote the index such that $S_{k,i_k} = \frac{1}{p_{i_k}}$.*

Theorem 2.6 ([CP15] Theorem 2.3). *Sampling at least $O(\frac{d \log d}{\varepsilon^2})$ rows according to the ℓ_1 Lewis weights $\{w_i\}_{i=1}^n$ of a matrix $X \in \mathbb{R}^{n \times d}$ results in a subspace embedding for X with at least some constant probability. If at least $O(\frac{d \log \frac{d}{\varepsilon\delta}}{\varepsilon^2})$ rows are sampled, then we have a subspace embedding with probability at least $1 - \delta$.*

The method

- The problem is: **we do not have access to all of y**

  the Lewis weight sampling-and-embedding matrix S preserves $\|X\beta\|_1$ for all β, **but it doesn't preserve $\|X\boldsymbol{\beta} - \boldsymbol{y}\|_1$**

- **Solution**

$$\text{estimate} \quad \|X\widehat{\beta} - y\|_1 - \|X\beta^* - y\|_1 \quad \text{with} \quad \|SX\widehat{\beta} - Sy\|_1 - \|SX\beta^* - Sy\|_1$$

Lemma 4.1. *Let $X \in \mathbb{R}^{n\times d}$ have ℓ_1 Lewis weights $\{w_i\}_{i\in[n]}$. Then, for any N that is at least $O\left(\frac{d}{\varepsilon^2}\log\frac{d}{\varepsilon\delta}\right)$, there is a sampling-and-reweighting distribution $\mathcal{S}(\{p_i\}_{i=1}^n)$ satisfying $\sum_i p_i = N$ such that for all y, if $S \sim \mathcal{S}(\{p_i\}_{i=1}^n)$ and $\beta^* = \arg\min \|X\beta - y\|_1$, we have for all β*

$$(\|SX\beta^* - Sy\|_1 - \|SX\beta - Sy\|_1) - (\|X\beta^* - y\|_1 - \|X\beta - y\|_1) \le \varepsilon \cdot \|X\beta^* - X\beta\|_1 \qquad (9)$$

with probability at least $1-\delta$. Further, for constant δ, $m = O(d\log d/\varepsilon^2)$ rows suffice.

Lemma 4.1. *Let $X \in \mathbb{R}^{n\times d}$ have ℓ_1 Lewis weights $\{w_i\}_{i\in[n]}$. Then, for any N that is at least $O\left(\frac{d}{\varepsilon^2}\log\frac{d}{\varepsilon\delta}\right)$, there is a sampling-and-reweighting distribution $\mathcal{S}(\{p_i\}_{i=1}^n)$ satisfying $\sum_i p_i = N$ such that for all y, if $S \sim \mathcal{S}(\{p_i\}_{i=1}^n)$ and $\beta^* = \arg\min \|X\beta - y\|_1$, we have for all β*

$$(\|SX\beta^* - Sy\|_1 - \|SX\beta - Sy\|_1) - (\|X\beta^* - y\|_1 - \|X\beta - y\|_1) \le \varepsilon \cdot \|X\beta^* - X\beta\|_1 \tag{9}$$

with probability at least $1-\delta$. Further, for constant δ, $m = O(d\log d/\varepsilon^2)$ rows suffice.

Proof of Theorem 3.1. Applying Lemma 4.1 to $\widehat{\beta} := \arg\min \|SX\beta - Sy\|_1$, we get

$$\left(\|SX\beta^* - Sy\|_1 - \|SX\widehat{\beta} - Sy\|_1\right) \le \left(\|X\beta^* - y\|_1 - \|X\widehat{\beta} - y\|_1\right) + \varepsilon \cdot \|X\beta^* - X\widehat{\beta}\|_1$$

Since $\widehat{\beta}$ is the minimizer of $\|SX\beta - Sy\|_1$, the left side is non-negative. So,

$$\begin{aligned}\|X\widehat{\beta} - y\|_1 &\le \|X\beta^* - y\|_1 + \varepsilon \cdot \|X\beta^* - X\widehat{\beta}\|_1\\ &\le \|X\beta^* - y\|_1 + \varepsilon \cdot (\|X\beta^* - y\|_1 + \|X\widehat{\beta} - y\|_1)\end{aligned}$$

Rearranging, and assuming $\varepsilon < 1/2$,

$$\begin{aligned}\|X\widehat{\beta} - y\|_1 &\le \frac{1+\varepsilon}{1-\varepsilon}\|X\beta^* - y\|_1\\ &\le (1+4\varepsilon)\|X\beta^* - y\|_1\end{aligned}$$

Using $\varepsilon' = \varepsilon/4$ proves the theorem.

Results

Theorem 3.1. *Let $X \in \mathbb{R}^{n\times d}$ have ℓ_1 Lewis weights $\{w_i\}_{i\in[n]}$, and let $0 < \varepsilon, \delta < 1$. Then, for any N that is* $\boxed{\text{at least } O\left(\frac{d}{\varepsilon^2}\log\frac{d}{\varepsilon\delta}\right),}$ *there is a sampling-and-reweighting distribution $\mathcal{S}(\{p_i\}_{i=1}^n)$ satisfying $\sum_i p_i = N$ such that for all y, if $S \sim \mathcal{S}(\{p_i\}_{i=1}^n)$ and $\widehat{\beta} = \arg\min \|SX\beta - Sy\|_1$, we have*

$$\|X\widehat{\beta} - y\|_1 \leq (1+\varepsilon)\min_{\beta}\|X\beta - y\|_1$$

with probability $1-\delta$. If $\delta = O(1)$ is some constant, then N at least $O\left(\frac{1}{\varepsilon^2}d\log d\right)$ rows suffice.

Theorem 3.5. *For any $d \geq 2$, $\epsilon < \frac{1}{10}$, $\delta < \frac{1}{4}$, there exist sets $\mathcal{X} \in \mathbb{R}^d, \mathcal{Y} \in \mathbb{R}$ of inputs and labels, and a distribution P on $\mathcal{X}\times\mathcal{Y}$ such that any algorithm which solves Problem 2, with $\varepsilon = 1$, requires at least $m = \Omega(\frac{d}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta} + d\log\frac{1}{\delta})$ samples.*

- **which indicates that the proposed method is near optimal**

THANKS