ℓ_1 Regression via Lewis Weight Subsampling

Aditya Parulekar, Advait Parulekar, Eric Price

UT Austin

Abstract

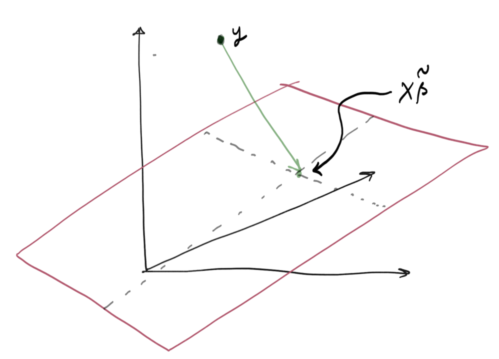
We consider the problem of finding an approximate solution to ℓ_1 regression while only observing a small number of labels. Given an $n \times d$ unlabeled data matrix X, we must choose a small set of $m \ll n$ rows to observe the labels of, then output an estimate $\widehat{\beta}$ whose error on the original problem is within a $1+\varepsilon$ factor of optimal. We show that sampling from X according to its Lewis weights and outputting the empirical minimizer succeeds with probability $1-\delta$ for $m > O(\frac{1}{\varepsilon^2} d \log \frac{d}{\varepsilon\delta})$. This is analogous to the performance of sampling according to leverage scores for ℓ_2 regression, but with exponentially better dependence on δ. We also give a corresponding lower bound of $\Omega(\frac{d}{\varepsilon^2} + (d + \frac{1}{\varepsilon^2}) \log \frac{1}{\delta})$.

Active LAD Regression

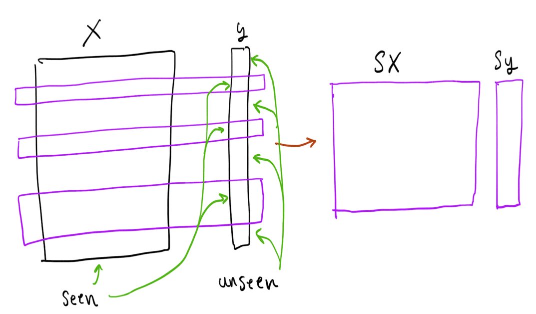
There is a full training set $\{X_i\}_{1\le i\le n}$, but **no observed** $\{y_i\}_{1\le i\le n}$. **Pick index set** I of size m to query, so you see $\{y_i\}_{i\in I}$ with $m \ll n$. Return $\tilde{\beta}$ such that with high probability

$$\|X\tilde{\beta} - y\|_1 \le (1+\epsilon)\|X\beta^* - y\|_1$$

where $\beta^* =_{\beta\in\mathbb{R}^d} \|X\beta - y\|_1$.



Strategy



- Pick a distribution over $[n]$ and sample $m \ll n$ elements, sample and re-weight according to that distribution to preserve expectations.
- Sampling represented by "sampling-and-reweighting" matrix S, so

$$X \xrightarrow{\text{Sampled}} SX$$

- return $\|SX\beta - Sy\|_1$.

Subspace Embeddings and Regression

- Can use **Lewis weight subsampling** to get **subspace embedding:**

$$(1-\varepsilon)\|X\beta\|_1 \le \|SX\beta\|_1 \le (1+\varepsilon)\|X\beta\|_1$$

  for all β.
  - Need $m = O\left(\frac{d\log(d/\varepsilon\delta)}{\varepsilon^2}\right)$ sampled rows [1]
- Show that we can use this to sample well for regression as well! (without even seeing y).

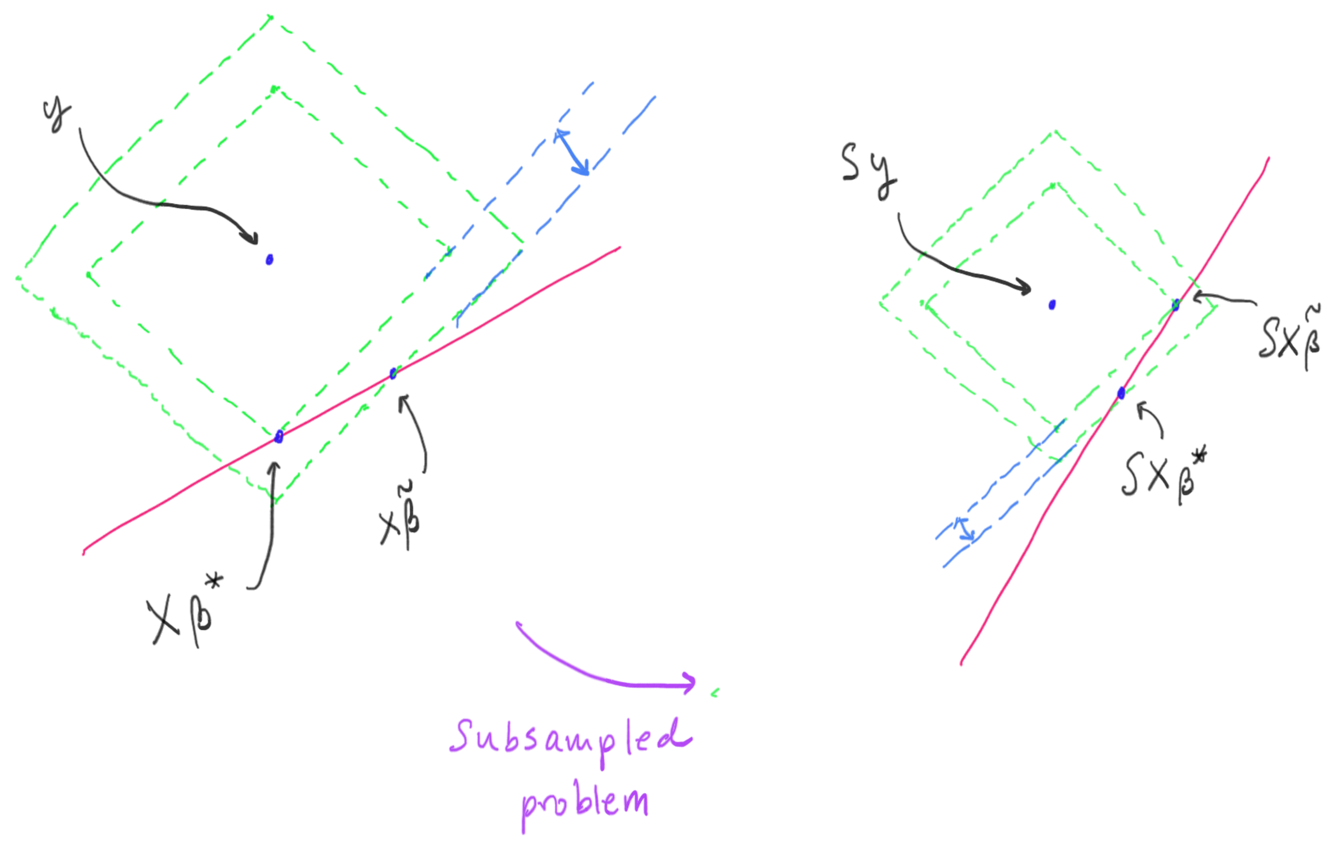


Fig. 3: Representation of the problem: on the left, $\mathbb{R}^n$, with green ℓ_1 balls projecting y onto the red column space of X. On the right, the same but for the sampled space in $\mathbb{R}^m$

Proof Sketch

Three main steps:

- **Symmetrize:**

$$\mathop{\mathbb{E}}_{S}\left[\left(\max_{\|X\beta^*-X\beta\|=1} |(\|SX\beta^* - Sy\|_1 - \|SX\beta - Sy\|_1) - (\|X\beta^* - y\|_1 - \|X\beta - y\|_1)|\right)^l\right] \le 2^l \mathop{\mathbb{E}}_{S,\sigma}\left[\left(\max_{\|X\beta^*-X\beta\|=1}\left|\sum_k \sigma_k\left(\frac{|x_{i_k}^\top\beta^* - y_{i_k}|}{p_{i_k}} - \frac{|x_{i_k}^\top\beta - y_{i_k}|}{p_{i_k}}\right)\right|\right)^l\right]$$

  where i_k are the rows sampled by S, and σ_k are independent Rademacher random variables (± 1 w.p. $1/2$).
- **Contraction Lemma:** effectively a triangle inequality that removes the effect of y

$$2^l \mathop{\mathbb{E}}_{S,\sigma}\left[\left(\max_{\|X\beta^*-X\beta\|=1}\left|\sum_k \sigma_k\left(\frac{|x_{i_k}^\top\beta^* - y_{i_k}|}{p_{i_k}} - \frac{|x_{i_k}^\top\beta - y_{i_k}|}{p_{i_k}}\right)\right|\right)^l\right] \le 2^{2l+1} \mathop{\mathbb{E}}_{S,\sigma}\left[\left(\max_{\|X(\beta^*-\beta)\|_1=1}\left|\sum_k \sigma_{i_k}\frac{x_{i_k}^\top}{p_{i_k}}(\beta^*-\beta)\right|\right)^l\right]$$

- **Subspace embedding argument:** we are in the column space of SX now, so we can modify and apply subspace embedding results

$$2^{2l+1} \mathop{\mathbb{E}}_{S,\sigma}\left[\left(\max_{\|X(\beta^*-\beta)\|_1=1}\left|\sum_k \sigma_{i_k}\frac{x_{i_k}^\top}{p_{i_k}}(\beta^*-\beta)\right|\right)^l\right] \le \varepsilon^l\delta$$

Proof Approach

- Cannot show $\|\beta^* - \tilde{\beta}\|_1$ is small
  - Because ℓ_1 minimizer is not unique
- Instead:

  $(\|SX\beta^* - Sy\|_1 - \|SX\beta - Sy\|_1) - (\|X\beta^* - y\|_1 - \|X\beta - y\|_1) \le \varepsilon \cdot \|X\beta^* - X\beta\|_1$
- So in figure 3, we would be showing that the two blue distances are close compared to $\|X\beta^* - X\tilde{\beta}\|_1$.
- The effects of the difficult y cancel in each term, and now we can show this using the subspace embedding property for Lewis weight sampling.

Results

- **Upper bound**: Need $O(\frac{d\log(d/\epsilon\delta)}{\epsilon^2})$ rows!
- **Lower bound**: No algorithm can do fewer than $\Omega(\frac{d}{\epsilon^2} + \frac{\log(1/\delta)}{\epsilon^2} + d\log\frac{1}{\delta})$

References

[1] Michael B. Cohen and Richard Peng. Lp row sampling by lewis weights. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 183–192, New York, NY, USA, 2015. Association for Computing Machinery.