L1 Regression with Lewis Weights Subsampling

Aditya Parulekar ✉

Department of Computer Science, University of Texas at Austin, TX, USA

Advait Parulekar ✉

Department of Electrical and Computer Engineering, University of Texas at Austin, TX, USA

Eric Price ✉

Department of Computer Science, University of Texas at Austin, TX, USA

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 $\Omega(\frac{d}{\varepsilon^2} + (d + \frac{1}{\varepsilon^2}) \log \frac{1}{\delta})$.

2012 ACM Subject Classification Computing methodologies → Active learning settings

Keywords and phrases Active regression, Lewis weights

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2021.49

Category RANDOM

Related Version *Full Version*: `https://arxiv.org/abs/2105.09433`

Funding *Advait Parulekar*: This author was supported by NSF Grant 2019844.
Eric Price: This author was supported in part by NSF Award CCF-1751040 (CAREER).

Acknowledgements The authors wish to thank the anonymous reviewers for several comments that improved the paper.

1 Introduction

The standard linear regression problem is, given a data matrix $X \in \mathbb{R}^{n\times d}$ and corresponding values $y \in \mathbb{R}^n$, to find a vector $\beta \in \mathbb{R}^d$ minimizing $\|X\beta - y\|_p$. Least squares regression ($p = 2$) is the most common, but least absolute deviation regression ($p = 1$) is sometimes preferred for its robustness to outliers and heavy-tailed noise. In this paper we focus on ℓ_1 regression:

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^d} \|X\beta - y\|_1 \tag{1}$$

But what happens if the unlabeled data X is cheap but the labels y are expensive? Can we choose a small subset of indices, only observe the corresponding labels, and still recover a good estimate $\widehat{\beta}$ of the true solution? We would like an algorithm that works with probability $1-\delta$ for any input (X, y); this necessitates that our choice of indices be randomized, so the adversary cannot concentrate the noise on them. Formally we define the problem as follows:

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2021).
Editors: Mary Wootters and Laura Sanità; Article No. 49; pp. 49:1–49:21
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

▶ **Problem 1** (Active L1 regression). *There is a known matrix $X \in \mathbb{R}^{n\times d}$ and a fixed unknown vector y. A learner interacts with the instance by querying rows indexed $\{i_k\}_{k\in[m]}$ adaptively, and is shown labels $\{y_{i_k}\}_{k\in[m]}$ corresponding to the rows queried. The learner must return $\widehat{\beta}$ such that with probability $1-\delta$ over the learner's randomness,*

$$\|X\widehat{\beta} - y\|_1 \leq (1+\varepsilon)\min_\beta \|X\beta - y\|_1. \tag{2}$$

Some rows of X may be more important than others. For example, if one row is orthogonal to all the others, we need to query it to have any knowledge of the corresponding y; but if many rows are in the same direction it should suffice to label a few of them to predict the rest.

A natural approach to this problem is to attach some notion of "importance" $p_1, \ldots, p_n$ to each row of X, then sample rows proportional to p_i. We can represent this as a "sampling-and-reweighting" sketch $S \in \mathbb{R}^{m\times n}$, where each row is $\frac{1}{p_i}e_i$ with probability proportional to p_i. This reweighting is such that $\mathbb{E}_S[\|Sv\|_1] \propto \|v\|_1$ for any vector v. By querying m rows we can observe Sy, and so can output the empirical risk minimizer (ERM)

$$\widehat{\beta} := \arg\min \|SX\beta - Sy\|_1. \tag{3}$$

For fixed β, $\mathbb{E}_S \|SX\beta - Sy\|_1 \propto \|X\beta - y\|_1$. The hope is that, if the p_i are chosen carefully, the ERM $\widehat{\beta}$ will satisfy (2) with relatively few samples. Our main result is that this is true if the p_i are drawn according to the ℓ_1 Lewis weights:

▶ **Theorem 1** (Informal). *Problem 1 can be solved with $m = O(\frac{1}{\varepsilon^2} d \log \frac{d}{\varepsilon\delta})$ queries. For constant $\delta = \Theta(1)$, $m = O\left(\frac{1}{\varepsilon^2} d \log d\right)$ suffices.*

Note that, while the model allows for adaptive queries, this algorithm is nonadaptive.

We next show that our sample complexity is near-optimal by demonstrating the following lower bound on the number of queries needed by any algorithm to obtain an accurate estimate. Whether the multiplicative factor of $\log d$ is necessary is an open question.

▶ **Theorem 2** (Informal). *Any algorithm satisfying Problem 1 must query $\Omega(d \log \frac{1}{\delta} + \frac{d}{\varepsilon^2} + \frac{1}{\varepsilon^2} \log \frac{1}{\delta})$ rows on some instances (X, y).*

For small δ, the upper bound is the product of d, $\frac{1}{\varepsilon^2}$, and $\log(1/\delta)$ while the lower bound is the product of each pair.

1.1 Related Work

If all the labels are known

LAD regression cannot be solved in closed form. It can be written as a linear program, but this is relatively slow to solve. One approach to speeding up LAD regression is "sketch-and-solve," which replaces (1) with (3), which has fewer constraints and so can be solved faster. The key idea here is to acquire regression guarantees by ensuring that S is a *subspace embedding* for the column space of $[X\ y]$.

For a survey on techniques to do this, we direct the reader to [17],[13], [3]. In [17], the emphasis is on *oblivious* sketches – distributions which do not require knowledge of $[X\ y]$. On the other hand, [13], [3] discuss sketches that depend on $[X\ y]$. Most relevant to us [9], which shows that sampling-and-reweighting matrices S using Lewis weights of $[X\ y]$ suffice; we give a simple proof of this in Remark 4. The problem is that *figuring out which labels are important* involves looking at all the labels.

Active ℓ_2 regression

Here we return to our setting, where only a subset of the labels is available to us. A number of works have studied this problem, including [6, 7, 8]. The ℓ_2 version of the problem was solved optimally in [2], where an algorithm was given using $O(\frac{d}{\varepsilon})$ queries to find $\widehat{\beta}$ satisfying $\mathbb{E}\left[\|X\widehat{\beta} - y\|_2^2\right] \leq (1+\varepsilon)\|X\beta^* - y\|_2^2$. Independent, identical sampling using leverage scores achieves the same guarantee using $O(d\log d + \frac{d}{\varepsilon})$ queries. Note that these results for ℓ_2 ERM only work in expectation, while our results hold with high probability.

Subspace embedding for ℓ_1 norms

Subspace embeddings for the ℓ_1 norm have been studied in a long line of work including [15], [16], [12], [5], and [4], the most recent of which describes an iterative algorithm to approximate *Lewis weights*, which are the analogue of leverage scores for importance sampling preserving ℓ_1 norms. The [4] result shows that, for the same $m = O(\frac{1}{\varepsilon^2} d \log \frac{d}{\varepsilon\delta})$ sample complexity as given in Theorem 11, a sampler sketch S based on the Lewis weights of X will have $\|SX\beta\|_1 \approx_\varepsilon \|X\beta\|_1$ for all $\beta \in \mathbb{R}^d$.

Our approach

At a very high level the goal of this paper is to replace the ℓ_2 leverage score analysis of [2] (which looks at the sample complexity of ℓ_2 regression in setting of this paper) with the ℓ_1 Lewis weight analysis in [4] (which, among other results, demonstrates that i.i.d. importance sampling with Lewis weights results in a subspace embedding). However, the differences between ℓ_1 and ℓ_2 are significant enough that very little of the [2] proof approach remains.

Per [4], the Lewis weight sampling-and-embedding matrix S preserves $\|X\beta\|_1$ for all β. The problem is that it *doesn't* preserve $\|X\beta - y\|_1$: if y has outliers, we have no idea where they are to sample them. In the ℓ_2 setting, this difficulty is addressed using the closed-form solution $\beta^* = X^\dagger y$, and the fact that the residual vector $X\beta^* - y$ is orthogonal to the column space of X. If S is a subspace embedding it will preserve $\|X\beta - X\beta^*\| \approx \|S(X\beta - X\beta^*)\|$, while orthogonality of $X\beta^* - y$ and $X\beta^*$ ensures that $\|S(X\tilde{\beta} - X\beta^*)\| \ll \|S(X\tilde{\beta} - y)\| \approx \|X\tilde{\beta} - y\|$ (here the last approximation is not because of embeddings but rather Markov's inequality). In the ℓ_1 setting, not only is β^* not expressible in closed form, but there can be many equally valid minimizers β^* that are far from each other. In Appendix A we show how this approach extends to the ℓ_1 setting to give a simple proof of Theorem 1 for a constant factor approximation (i.e., $\varepsilon = O(1)$); but the existence of multiple β^* makes $\varepsilon < 1$ seem unobtainable by this approach.

Instead, we massage the [2] subspace embedding proof into the appropriate form, as we discuss in Section 3. While S doesn't preserve the total error $\|X\beta - y\|_1$, it does preserve *relative* error $\|X\beta - y\|_1 - \|X\beta^* - y\|_1$; the effect of outliers is canceled out, so that this concentrates similarly well to $\|X\beta - X\beta^*\|_1$.

Concurrent work

A very similar set of results appears concurrently and independently in [1]. Their main result is identical to ours, with a similar proof. They also extend the result to $1 < p < 2$, but with a significantly weaker $m = \widetilde{O}(d^2/\varepsilon^2)$ bound. They do not have the $\Omega(d \log \frac{1}{\delta})$ lower bound.

2 Preliminaries: Subspace Embeddings and Importance Sampling

A key idea used in our analysis is that of a ℓ_1 subspace embedding, which is a linear sketch of a matrix that preserves ℓ_1 norms within the column space of a matrix:

▶ **Definition 3** (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 $\beta \in \mathbb{R}^d$,*

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

▶ Remark 4. Consider the simpler setting in which we had access to all of y, but we still want to subsample rows to improve computational complexity. We can view the regression loss $\|X\beta - y\|_1$ as the ℓ_1 norm of the point $[X\ y]\begin{bmatrix}\beta\\-1\end{bmatrix}$ in the column space of $[X\ y]$. Indeed, suppose $\beta^* = \arg\min \|X\beta - y\|_1$ as before and let $\widehat{\beta} = \arg\min \|SX\beta - Sy\|_1$. Then, $\widehat{\beta}$ solves problem 1 because, for $\varepsilon < \frac{1}{3}$,

$$\|X\widehat{\beta}-y\|_1 \le \frac{1}{1-\varepsilon}\|SX\widehat{\beta}-Sy\|_1 \le \frac{1}{1-\varepsilon}\|SX\beta^*-Sy\|_1 \le \frac{1+\varepsilon}{1-\varepsilon}\|X\beta^*-y\|_1 \le (1+4\varepsilon)\|X\beta^*-y\|_1.$$

One way to construct a subspace embedding is by sampling rows and rescaling them appropriately:

▶ **Definition 5** (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 $S_{k,i_k} = \frac{1}{p_{i_k}}$.*

When working in ℓ_2, there is a natural choice for re-weighting: the leverage scores of the rows [17].

▶ **Definition 6** (Leverage Scores). *The leverage score of the ith row of a matrix X, $l_i(X)$ is defined as $x_i^\top(X^\top X)^{-1}x_i$.*

For ℓ_1 subspace embeddings, the analogous weights are the ℓ_1 Lewis weights, defined implicitly as the unique weights $\{w_i(X)\}_{i=1}^n$ that satisfy $w_i(X) = l_i(WX)$ where W is a diagonal matrix with ith diagonal entry $\frac{1}{\sqrt{w_i(X)}}$. We will drop the explicit dependence on X whenever it is clear from context.

▶ **Definition 7** (Lewis Weights). *The ℓ_1 Lewis weights of a matrix X are the unique weights $\{w_i\}_{i=1}^n$ that satisfy $w_i^2 = x_i^\top(\sum_{j=1}^n \frac{1}{w_j}x_jx_j^\top)^{-1}x_i$ for all i.*

Lewis weights are defined in general for general ℓ_p norms, but we will only need the ℓ_1 Lewis weights. For basic properties of Lewis weights, we direct the reader to [4]. Using these definitions, we now state the main consequence of using Lewis weights. This result comes from a line of work on embeddings from subspaces of $L_1[0,1]$ to ℓ_1^m such as [15], but is reproduced here similar to how it is presented in [4].

▶ **Theorem 8** ([4] 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 $1-\delta$.*

2.1 Properties of Lewis Weights

We will need some properties of Lewis weights, particularly of how they change when the matrix X is modified.

▶ **Lemma 9** ([4] Lemma 5.5). *The ℓ_1 Lewis weights of a matrix do not increase when rows are added.*

▶ **Lemma 10.** *Let $X \in \mathbb{R}^{n\times d}$, and let $X' \in \mathbb{R}^{kn\times d}$ be X stacked on itself k times, with each row scaled down by k. Then, each of the Lewis weights is reduced by a factor of k.*

3 Proof Overview

▶ **Theorem 11.** *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 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.

Regression guarantees from column-space embeddings

As noted in Remark 4, it would suffice to show that $\|SX\beta - Sy\|_1 \approx \|X\beta - y\|_1$ for all β. The problem is that this is impossible without knowing y: if one random entry of y is very large, we would need to sample it to estimate $\|X\beta - y\|_1$ accurately. However, we don't actually need to estimate $\|X\beta - y\|_1$; we just need to be able to distinguish values of β for which $\|X\beta - y\|_1$ is far from $\|X\beta^* - y\|_1$ from values for which it is close. That is, it would suffice to accurately

$$\text{estimate}\quad \|X\widehat{\beta} - y\|_1 - \|X\beta^* - y\|_1 \quad\text{with}\quad \|SX\widehat{\beta} - Sy\|_1 - \|SX\beta^* - Sy\|_1 \tag{4}$$

for every possible β. In the above example where y has a single large outlier coordinate, sampling this coordinate or not will dramatically affect *both* terms, but will not affect the difference very much. As such, our key lemma, Lemma 28, states that ℓ_1 Lewis weight sampling achieves (4) with high probability. In particular, using at least $m \geq O(\frac{d}{\varepsilon^2}\log\frac{d}{\varepsilon\delta})$ rows we have

$$(\|SX\beta^* - Sy\|_1 - \|SX\beta - Sy\|_1) - (\|X\beta^* - y\|_1 - \|X\beta - y\|_1) < \varepsilon\|X(\beta^* - \beta)\|_1 \tag{5}$$

for all β with probability at least $1-\delta$. We do this by adapting the argument of [4] which shows that S is a column-space embedding with high probability. We have summarized this argument below.

Column-space embedding using Lewis weights ([4])

An important result in [4], which directly implies the high probability subspace embedding, and which will be useful to us later is the following moment bound on deviations of $\|SX\beta\|_1$.

▶ **Lemma 12** ([4] Lemma 7.4). *If N is at least $O\left(\frac{d}{\varepsilon^2}\log\frac{d}{\varepsilon\delta}\right)$, and $S \in \mathbb{R}^{N\times n}$ is drawn from the sampling-and-reweighting distribution $\mathcal{S}(\{p_i\}_{i=1}^N)$ with $\sum_i p_i = N$ and $\{p_i\}_{i=1}^n$ proportional to Lewis weights $\{w_i\}_{i=1}^n$, then*

$$\mathop{\mathbb{E}}_{S}\left[\left(\max_{\|X\beta\|_1=1} |\|SX\beta\|_1 - \|X\beta\|_1|\right)^l\right] \le \varepsilon^l\delta$$

The proof follows from this chain of inequalities:

$$\begin{aligned}\mathop{\mathbb{E}}_{S}\left[\left(\max_{\|X\beta\|_1=1} \|SX\beta\|_1 - \|X\beta\|_1\right)^l\right] &\overset{(A)}{\le} 2^l \mathop{\mathbb{E}}_{\sigma,S}\left[\left(\max_{\|X\beta\|_1=1}\left|\sum_k \sigma_k \frac{|x_{i_k}^T\beta|}{p_{i_k}}\right|\right)^l\right]\\ &\overset{(B)}{\le} 2^l \mathop{\mathbb{E}}_{\sigma,S}\left[\left(\max_{\|X\beta\|_1=1}\sum_k \sigma_k \frac{x_{i_k}^T\beta}{p_{i_k}}\right)^l\right]\\ &\overset{(C)}{\le} \varepsilon^l\delta\end{aligned}$$

where the σ_k are independent Rademacher variables, which are ± 1 with probability 1/2 each, and p_{i_k} is proportional to the ℓ_1 Lewis weight of row i_k. (A) follows by symmetrizing the objective $F \coloneqq \max_{\|X\beta\|_1=1}\|SX\beta\|_1 - \|X\beta\|_1$. (B) follows from a contraction lemma. (C) is shown by constructing a related matrix with bounded Lewis weights and applying Lemma 32 from [15] reproduced below.

▶ **Lemma 13.** *There exists constant C such that for any $X \in \mathbb{R}^{n\times d}$ with all ℓ_1 Lewis weights less than $C\frac{\varepsilon^2}{\log\left(\frac{n}{\delta}\right)}$ and $l = \log(2n/\delta)$, we have*

$$\mathbb{E}_\sigma\left[\left(\max_{\|X\beta\|_1=1}\left|\sum_{i=1}^n \sigma_i x_i^\top\beta\right|\right)^l\right] \le \frac{\varepsilon^l\delta}{2} \tag{6}$$

Regression guarantees using Lewis weight sampling

In this work, we show the following chain of inequalities.

$$\begin{aligned}&\mathop{\mathbb{E}}_{S}\left[\left(\max_{\|X\beta^*-X\beta\|=1}|(\|SX\beta^*-Sy\|_1 - \|SX\beta-Sy\|_1) - (\|X\beta^*-y\|_1 - \|X\beta-y\|_1)|\right)^l\right]\\ &\overset{(A)}{\le} 2^l \mathop{\mathbb{E}}_{S,\sigma}\left[\left(\max_{\|X\beta^*-X\beta\|=1}\left|\sum_k \sigma_k\left(\frac{|x_{i_k}^\top\beta^*-y_{i_k}|}{p_{i_k}} - \frac{|x_{i_k}^\top\beta-y_{i_k}|}{p_{i_k}}\right)\right|\right)^l\right]\\ &\overset{(B)}{\le} 2^{2l+1} \mathop{\mathbb{E}}_{S,\sigma}\left[\left(\max_{\|X(\beta^*-\beta)\|_1=1}\left|\sum_k \sigma_{i_k}\frac{x_{i_k}^\top}{p_{i_k}}(\beta^*-\beta)\right|\right)^l\right]\\ &\overset{(C)}{\le} \varepsilon^l\delta\end{aligned} \tag{7}$$

Here, for (A), we symmetrize the left hand side of (5) in Lemma 29. For (B), we apply a different contraction lemma, Lemma 30, that allows us to remove y from our expression, and then end up with the same moment bound for (C). Step (C) is essentially an application of Lemma 32 to SX, however, because we cannot immediately bound the Lewis weights of SX to confirm the constraints of the Lemma, we instead construct another matrix X'' which does not significantly alter the right hand side of inequality (7) while having bounded Lewis weights. This is done in Lemmas 33 and 34.

3.1 Lower Bounds

We will show that any algorithm must see $\Omega(d \log \frac{1}{\delta} + \frac{1}{\varepsilon^2} \log \frac{1}{\delta} + \frac{d}{\varepsilon^2})$ labels to return $\widehat{\beta}$ satisfying $\|X\widehat{\beta} - y\|_1 \leq (1+\varepsilon)\|X\beta^* - y\|_1$ with probability greater than $1-\delta$.

For the lower bound proof it is convenient to consider a *distributional* version of the problem:

▶ **Problem 2** (Distributional active L1 regression)**.** *There is an unknown joint distribution P over a finite set $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$, with $|\mathcal{Y}| = 2$. The learner is allowed to adaptively observe N i.i.d. samples from $P(\cdot|X = x)$ for the learner's choice of N values $x \in \mathcal{X}$. The learner must return $\widehat{\beta}$ satisfying*

$$\mathbb{E}_{(X,Y)\sim P}\left[|X^\top \widehat{\beta} - Y|\right] \leq (1+\varepsilon) \inf_{\beta} \mathbb{E}_{(X,Y)\sim P}\left[|X^\top \beta - Y|\right]. \tag{8}$$

with probability at least $1-\delta$.

We begin with a lemma that shows that solving the original, Problem 1, for some n polynomial in the parameters d, ε, δ is harder than solving the distributional version, Problem 2.

▶ **Lemma 14.** *A randomized algorithm that solves Problem 1 for $n = \frac{2}{\varepsilon^2}\left(\log \frac{2}{\delta} + d \log \frac{3d}{\varepsilon}\right)$ with accuracy ε and failure probability δ can be used to solve any instance of Problem 2, where $\mathcal{X}, \mathcal{Y}$, in the unit ℓ_∞ ball, with accuracy 6ε and failure probability 2δ, for small ε.*

Proof. Let $n = \frac{8}{\varepsilon^2}\left(\log \frac{2}{\delta} + d \log \frac{4d}{\varepsilon}\right)$. Construct an instance of Problem 1 in which the rows of feature matrix $\mathbf{X}$ and the corresponding label vector y are drawn i.i.d. from P. Let H be the unit ℓ_∞ ball. We have the following:

▷ Claim 15. For all $\beta \in H$, with probability at least $1-\delta$,

$$(1-\varepsilon)\, \mathbb{E}_{(X,Y)\sim P}\left[|X^\top \beta - Y|\right] \leq \frac{1}{n}\|\mathbf{X}\beta - y\|_1 \leq (1+\varepsilon)\, \mathbb{E}_{(X,Y)\sim P}\left[|X^\top \beta - Y|\right]$$

Let β° denote the minimizer $\inf_\beta \mathbb{E}_{(X,Y)\sim P}\left[|X^\top \beta - Y|\right]$. Let β^* denote the minimizer of the matrix instance $\inf_\beta \|\mathbf{X}\beta - y\|_1$, and let $\widehat{\beta}$ denote the output of the algorithm on the instance generated. Then we have

$$\begin{aligned}
(1-\varepsilon)\, \mathbb{E}_{(X,Y)\sim P}\left[|X^\top \widehat{\beta} - Y|\right] &\leq \frac{1}{n}\|\mathbf{X}\widehat{\beta} - y\|_1 \\
&\leq (1+\varepsilon)\frac{1}{n}\|\mathbf{X}\beta^* - y\|_1 && \text{with probability } 1-\delta \\
&\leq (1+\varepsilon)\frac{1}{n}\|\mathbf{X}\beta^\circ - y\|_1 \\
&\leq (1+\varepsilon)^2\, \mathbb{E}_{(X,Y)\sim P}\left[|X^\top \beta^\circ - Y|\right]
\end{aligned}$$

So with probability $1-2\delta$,

$$\mathbb{E}_{(X,Y)\sim P}\left[|X^\top \widehat{\beta} - Y|\right] \leq (1+6\varepsilon)\, \mathbb{E}_{(X,Y)\sim P}\left[|X^\top \beta^\circ - Y|\right]. \quad ◀$$

We then prove lower bounds on the accuracy for any algorithm on Problem 2. We prove three theorems that allow us to show Theorem 26: Theorems 16, 20, and 23. To do this, we make several Claims, which are proved in section 7.1.

In all our lower bounds, x is a uniform e_i, and $y \in \{0,1\}$ while y is a Bernoulli random variable. For $\Omega(\frac{d}{\varepsilon^2})$, we set $P(y = 1 \mid x = e_i)$ to $\frac{1}{2} + \varepsilon$ uniformly at random independently for each i; getting an ε-approximate solution requires getting most of the biases correct, which requires $\frac{1}{\varepsilon^2}$ samples from most of the coordinates e_i. The $\Omega(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$ instance sets $P(y=1 \mid x=e_i)$ to $\frac{1}{2}+\varepsilon$ with the same bias for each i; solving this is essentially distinguishing a ε biased coin from a $-\varepsilon$-biased coin. Finally, for $\Omega(d\log\frac{1}{\delta})$ we set $P(y=1 \mid x=e_i)=0$ except for a random hidden i^* with $P(y=1 \mid x=e_{i^*})=\frac{3}{4}$. Solving this instance requires finding i^*, but there's a δ chance the first $d\log\frac{1}{\delta}$ queries are all zero.

▶ **Theorem 16.** *For any $d \geq 2$ and $\varepsilon < \frac{1}{10}$, there exist families $\mathcal{X} \in \mathbb{R}^d, \mathcal{Y} \in \mathbb{R}$ of inputs and labels respectively such that any algorithm which solves Problem 2 with $\delta < \frac{1}{4}$ requires at least $m = \frac{3d}{2000\varepsilon^2}$ samples.*

We take $\mathcal{X}$ to be the set of standard basis vectors, and the distribution over $\mathcal{X}$ to be uniform. We will define a set $\mathcal{B}$ as being a subset of the unit hypercube $\{-1,1\}^d$ such that every element is sufficiently far from every other.

▷ Claim 17. There is a set $\mathcal{B} \subset \mathcal{H}$ with $|\mathcal{B}| \geq 2^{0.2d}$ such that for any two $\beta_1, \beta_2 \in \mathcal{B}$, we have $|\beta_1 - \beta_2| > 0.2d$

Proof. Here we just need an error correcting code with constant rate and constant relative (Hamming) distance. The existence of such a code follows from the Gilbert-Varshamov bound [10]. ◁

Fix some unknown β^*. We will have $Y = ZX^\top\beta^*$ where Z is an independent random variable with probability $\frac{1}{2}+\varepsilon$ of being 1, and $\frac{1}{2}-\varepsilon$ of being -1. This completes our description of P. We define $l(\beta)$ to be the ℓ_1 norm of the residuals for β, that is, $l(\beta) = \mathbb{E}_{(X,Y)\sim P}[\left|X^\top\beta - Y\right|]$. We have the following properties of $l(\beta)$.

▷ Claim 18. For $D, \mathcal{B}$ as chosen above, $l(\beta^*) = 1 - 2\varepsilon$.

▷ Claim 19. For $D, \mathcal{B}$ as chosen above, we have for all $\beta \in \mathcal{B}$, $l(\beta) - l(\beta^*) = \frac{2\varepsilon}{d}||\beta - \beta^*||_1$.

Proof of Theorem 16. Suppose some algorithm returns $\widehat{\beta}$ with $l(\widehat{\beta}) < (1+\frac{\varepsilon}{5})l(\beta^*) \implies ||\beta^* - \widehat{\beta}||_1 < 0.1d$ with probability $\frac{3}{4}$. By Fano's inequality,

$$H(\beta^*|\widehat{\beta}) < H\left(\frac{1}{4}\right) + \frac{\log|\mathcal{B}| - 1}{4} < 0.05d,$$

and we have a lower bound on the mutual information between the output of our algorithm and the true parameter: $I(\widehat{\beta};\beta^*) = H(\beta^*) - H(\beta^*|\widehat{\beta}) \geq 0.15d$. For an upper bound on the mutual information after seeing m samples, we use the data processing inequality.

$$\begin{aligned}
I(\beta^*;\widehat{\beta}) &\leq I(\beta^*;(Y_i)_{i\in[m]}) \leq \sum_{i=1}^m I(\beta^*;Y_i|(Y_j)_{j\in[i-1]}) \\
&= \sum_{i=1}^m H(Y_i|(Y_j)_{j\in[i-1]}) - H(Y_i|\beta^*,(Y_j)_{j\in[i-1]}) \\
&\leq \sum_{i=1}^m 1 - H(Y_i|\beta^*, I_i) \\
&\leq 4\varepsilon^2 m
\end{aligned}$$

Here we have used that

$$\begin{aligned}
H(Y_i|\beta^*,(Y_j)_{j\in[i-1]}) &\geq H(Y_i|\beta^*, I_i, (Y_j)_{j\in[i-1]}) \\
&= H(Y_i|\beta^*, I_i)
\end{aligned}$$

and that the distribution of Y_i conditioned on β^*, I_i is just an independent Bernoulli with parameter $\frac{1}{2}+\varepsilon$ and so

$$\sum_{i=1}^{m} 1 - H(Y_i|\beta^*, I_i) \leq \sum_{i=1}^{m} \left[1 + \left(\frac{1}{2}+\varepsilon\right)\log\left(\frac{1}{2}+\varepsilon\right) + \left(\frac{1}{2}-\varepsilon\right)\log\left(\frac{1}{2}-\varepsilon\right)\right]$$
$$\leq 4\varepsilon^2 m$$

So $0.15d \leq I(\beta^*; \widehat{\beta}) \leq 4\varepsilon^2 m$, and so we need $m \geq \frac{3d}{80\varepsilon^2}$. The result follows by replacing ε with 5ε. ◀

We can use the same instance to give a high probability lower bound of $\Omega(\log \frac{1}{\delta}/\varepsilon^2)$.

▶ **Theorem 20.** *For any d and $\varepsilon < \frac{1}{10}$, there exist sets $\mathcal{X} \in \mathbb{R}, \mathcal{Y} \in \mathbb{R}$ of inputs and labels respectively, and a distribution P on $\mathcal{X} \times \mathcal{Y}$ such that any algorithm which solves problem 2 requires at least $m = \frac{1}{4\varepsilon^2}\log\frac{1}{\delta}$ samples.*

Proof. Consider two instances, denoted by subscripts (1) and (2) with $\beta^*_{(1)} = -\mathbb{1}_d$ and $\beta^*_{(2)} = \mathbb{1}_d$, where $\mathbb{1}_d \in \mathbb{R}^d$ is the all-ones vector. Denote by $P_{(i)}$ the distribution over $\mathcal{X}, \mathcal{Y}$ for instance (i), and let $l_{\beta^*_{(i)}}(\beta) = \mathbb{E}_{(X,Y)\sim P_{(i)}}[|X^\top\beta - Y|]$ for $i \in \{1, 2\}$.

▷ Claim 21. For any β, $\max\{\ell_{\beta^*_{(1)}}(\beta) - \ell_{\beta^*_{(1)}}(\beta^*_{(1)}), \ell_{\beta^*_{(2)}}(\beta) - \ell_{\beta^*_{(2)}}(\beta^*_{(2)})\} > 2\varepsilon$

From this claim together with Claim 18, we have for some $i \in \{1, 2\}$, $l_{\beta^*_{(i)}}(\beta) \geq (1 + 2\varepsilon)l_{\beta^*_{(i)}}(\beta^*_{(i)})$, for all β.

Denote by $\widehat{\beta}$ the output of the algorithm. Denote by $\mathbb{P}_{(1)}$ the distribution over outputs by a algorithm interacting instance (1), and by $\mathbb{P}_{(2)}$ the distribution over outputs by a algorithm interacting instance (2). Denote by A the event that $\ell_{\beta^*_{(1)}}(\widehat{\beta}) - \ell_{\beta^*_{(1)}}(\beta^*_{(1)}) \geq 2\varepsilon$. Note that under A^c, we have $\ell_{\beta^*_{(2)}}(\widehat{\beta}) - \ell_{\beta^*_{(2)}}(\beta^*_{(2)}) \geq 2\varepsilon$. Because the algorithm fails with probability at most δ on any instance, we have $2\delta \geq \mathbb{P}_{(1)}(A) + \mathbb{P}_{(2)}(A^c)$. On the other hand, $\mathbb{P}_{(1)}(A) + \mathbb{P}_{(2)}(A^c) \geq e^{-D(\mathbb{P}_{(1)}||\mathbb{P}_{(2)})}$. We can bound the KL-divergence of the two distributions as an aggregate KL-divergence over the course of acquiring the samples.

▶ **Theorem 22** (Lemma 15.1, [11]). *If a learner interacts with two environments (1) and (2) through a policy $\pi(\cdot|I_1, Y_1, I_2, Y_2, \cdots, Y_{i-1})$ which dictates a distribution over actions I_i conditioned on the past $(I_1, Y_1, \cdots, Y_{i-1})$, and sees label Y_i distributed according to some label distribution $P_{(1),I_i}$ and $P_{(2),I_i}$, then the KL-divergence between the output of the learner on instance (1) and (2), $\mathbb{P}_{(1)}$ and $\mathbb{P}_{(2)}$ is given by*

$$D(\mathbb{P}_{(1)}||\mathbb{P}_{(2)}) = \sum_{k=1}^{d} \mathbb{E}_{(1)}\left[\sum_{i=1}^{N} \mathbb{1}\{I_i = k\} \cdot D(P_{(1),I_i}||P_{(2),I_i})\right]$$

Now, $P_{(1),k}$ is a Bernoulli with parameter $\frac{1}{2}+\varepsilon$, and $P_{(1),k}$ is a Bernoulli with parameter $\frac{1}{2}-\varepsilon$, so $D(P_{(1),k}\|P_{(1),k}) \leq 16\varepsilon^2$, and so we have

$$\sum_{k=1}^{d} \mathbb{E}_{(1)}\left[\sum_{i=1}^{N} \mathbb{1}\{I_i = k\} \cdot D(P_{(1),I_i}||P_{(2),I_i})\right] \leq \sum_{k=1}^{d} \mathbb{E}_{(1)}\left[\sum_{i=1}^{N} \mathbb{1}\{I_i = k\} \cdot 16\varepsilon^2\right]$$
$$= 16\varepsilon^2 \cdot \mathbb{E}_{(1)}\left[\sum_{k=1}^{d}\sum_{i=1}^{N} \mathbb{1}\{I_i = k\}\right] = 16\varepsilon^2 m$$

Putting this together, we have $\delta \geq e^{-16\varepsilon^2 m} \implies m \geq \frac{1}{16\varepsilon^2}\log\frac{1}{\delta}$, and the result follows by replacing ε with $\frac{1}{2}\varepsilon$. ◀

▶ **Theorem 23.** *For any $d \geq 2$, 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 $m = \frac{d}{3} \log \frac{1}{8\delta}$ samples.*

Proof. All logarithms are base 4. Consider instances in which $\mathcal{X} = \{e_1, e_2, \cdots, e_d\}$ where e_i denotes the ith standard basis vector and the distribution over $\mathcal{X}$ is uniform. We take $Y = ZX^\top \beta^*$ for some β^*, where Z is an independent Bernoulli random variable which is 1 with probability $\frac{3}{4}$, and 0 otherwise. Consider d instances labelled with subscripts $(1), (2), \cdots, (d)$, one in which each of the d standard basis is β^*, that is, $\beta^*_{(i)} = e_i$. Denote by β_j the jth coordinate of β. For each instance, we have

▷ Claim 24. For all $i \in [d], \beta \in \mathbb{R}^d$, we have $\ell_{\beta^*_{(i)}}(\beta) \geq \frac{1}{4d}$ with equality when $\beta = \beta^*_{(i)}$

We would like our algorithm to return an estimate $\widehat{\beta}$ which satisfies $\ell_{\beta^*}(\widehat{\beta}) < \frac{1}{2d}$. We first note that any choice of β only succeeds to be this close to the optimal on a single instance.

▷ Claim 25. Any $\beta \in \mathbb{R}^d$ can only satisfy $\ell_{\beta^*_{(i)}}(\widehat{\beta}) < \frac{1}{2d}$ for one $i \in [d]$.

So, we may as well enforce that the algorithm return one of $e_1, e_2, \cdots, e_d$, since any other output can be mapped to one of these to improve the performance of the algorithm.

We will allow our algorithm to sample $N = \frac{d}{3} \log \frac{1}{\delta}$ rows total. Let $\mathcal{E}$ be the event that $Y_1, Y_2, \ldots Y_N$ are all zero. Given any algorithm $\mathcal{A}$, let $F_\mathcal{A}$ denote the set of rows it samples fewer than $\log \frac{1}{\delta}$ times with probability at least $\frac{1}{2}$, in event $\mathcal{E}$. Because the total number of rows sampled is $\frac{d}{3} \log \frac{1}{\delta}$, there must be at least $\frac{2d}{3}$ rows which are sampled fewer than $\frac{1}{2} \log \frac{1}{\delta}$ times in expectation.

By Markov's inequality, these rows are sampled fewer than $\log \frac{1}{\delta}$ times with probability at least $\frac{1}{2}$, and are thus all in $F_\mathcal{A}$. Let $B_\mathcal{A}$ denote the distribution over outputs $\widehat{\beta}$ of $\mathcal{A}$ in event $\mathcal{E}$. Let $i_\mathcal{A} = \arg\min_{j \in F_\mathcal{A}} B_\mathcal{A}(j)$. Denote by $G_\mathcal{A}$ the event that row $i_\mathcal{A}$ is sampled fewer than $\log \frac{1}{\delta}$ times; by construction we have $\mathbb{P}(G_\mathcal{A}) > \frac{1}{2}$.

The subscripts are explicit because $F_\mathcal{A}, B_\mathcal{A}, i_\mathcal{A}, \mathbb{P}[G_\mathcal{A}]$ are properties of the algorithm and are independent of the instance with which it interacts. Consider the performance of this algorithm against the instance $\beta^*_{(i_\mathcal{A})}$.

Let $Y_{(i_\mathcal{A}),j,k}$ denote the label returned to the algorithm when it queries e_j for the kth time. Let $T_{(i_\mathcal{A})} = \min\{t | Y_{(i_\mathcal{A}),i_\mathcal{A},t} = 1\}$. Denote by $E_{(i_\mathcal{A})}$ the event that $T_{(i_\mathcal{A})} > \log \frac{1}{\delta}$. Because $T_{(i_\mathcal{A})}$ is a geometric random variable, we have $\mathbb{P}[E_{(i_\mathcal{A})}] > \delta$.

Now condition on the event $G_\mathcal{A} \cap E_{i_\mathcal{A}}$, which is an event with probability $\frac{1}{2}\delta$. Here our algorithm samples $i_\mathcal{A}$ fewer than $T_{i_\mathcal{A}}$ times, so it never sees a 1 and its output distribution is $B_\mathcal{A}$. It returns $i \in F_\mathcal{A} \setminus \{i_\mathcal{A}\}$ with probability at least $1 - B_\mathcal{A}(i_\mathcal{A}) \geq 1 - \frac{1}{|F_\mathcal{A}|} \geq 1 - \frac{3}{2d} \geq \frac{1}{4}$. In summary, even after $\frac{d}{3} \log \frac{1}{\delta}$ queries, no algorithm can return $\widehat{\beta}$ with $\|X\widehat{\beta} - y\| < (1+\varepsilon)\|X\beta^* - y\|$ with probability greater than $\frac{1}{8}\delta$. The result follows by replacing δ by 8δ. ◀

Putting these together we have:

▶ **Theorem 26.** *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.*

▶ **Corollary 27.** *Any algorithm that solves Problem 1 takes at least $\Omega(d \log \frac{1}{\delta} + \frac{d}{\varepsilon^2} + \frac{1}{\varepsilon^2} \log \frac{1}{\delta})$ samples for some $n = O(\frac{d \log \frac{d}{\delta}}{\varepsilon})$.*

Proof. Each of the instances that demonstrate the lower bounds above, in Lemmas 16, 20, and 23, take $|\mathcal{X}| = d$, the results follows from Lemma 14. ◀

4 Proof of Theorem 11

▶ **Lemma 28.** *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*

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

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

This lemma is proved for high probability in Section 4.1, and for constant probability in the full version of this paper, [14]. Given this, we can prove the main theorem.

Proof of Theorem 11. Applying Lemma 28 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. ◀

4.1 Proof of Lemma 28

This argument is similar to that in Appendix B of [4]. In order to prove Lemma 28, by Markov's inequality, it is sufficient to show that for some l,

$$M := \mathop{\mathbb{E}}_{S}\left[\left(\max_{\|X\beta^*-X\beta\|=1} |(\|SX\beta^* - Sy\|_1 - \|SX\beta - Sy\|_1) - (\|X\beta^* - y\|_1 - \|X\beta - y\|_1)|\right)^l\right] \le \varepsilon^l\delta$$

To show this, we will symmetrize, then use a contraction lemma to cancel the y terms. Then, with all the terms being within the column space of SX, we use the fact that S is a subspace embedding with high probability. We present two different bounds, one used for the constant probability and one for the high probability cases, but the following intermediate bound is the same for the two:

▶ **Lemma 29.** *Given a matrix $X \in \mathbb{R}^{n\times d}$, let $\mathcal{S}(\{p_i\}_{i\in[n]})$ be any sampling-and-reweighting disribution, and let i_k be the row-indices chosen by this sampling matrix such that $S_{k,i_k} = \frac{1}{p_{i_k}}$. Let σ_k be independent Rademacher variables that are ± 1 each with probability 0.5. Then,*

$$M \le 2^l \mathop{\mathbb{E}}_{S,\sigma}\left[\left(\max_{\|X\beta^*-X\beta\|=1}\left|\sum_k \sigma_k\left(\frac{|x_{i_k}^\top\beta^* - y_{i_k}|}{p_{i_k}} - \frac{|x_{i_k}^\top\beta - y_{i_k}|}{p_{i_k}}\right)\right|\right)^l\right] \quad (10)$$

This is essentially standard symmetrization; the proof is in Appendix B. To simplify the expression and eliminate the terms involving the labels, we then use a theorem from [12]:

▶ **Lemma 30** ([12] Theorem 5). *Let $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$ be convex and increasing, and let $\phi_k : \mathbb{R} \to \mathbb{R}$ be contractions such that $\phi_k(0) = 0$ for all k. Let $\mathcal{F}$ be a class of functions on $\{1, 2, 3 \ldots, n\}$, and $\|g(f)\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |g(f)|$. Then,*

$$\mathbb{E}_\sigma \left[\Phi\left(\frac{1}{2} \left\| \sum_k \sigma_k \phi_k(f(k)) \right\|_{\mathcal{F}} \right) \right] \leq \frac{3}{2} \mathbb{E}_\sigma \left[\Phi\left(\left\| \sum_k \sigma_k f(k) \right\|_{\mathcal{F}} \right) \right]$$

▶ **Lemma 31.** *For any $y \in \mathbb{R}^n$, we have*

$$\begin{aligned} &\mathop{\mathbb{E}}_{S,\sigma} \left[\left(\max_{\|X\beta^* - X\beta\| = 1} \left| \sum_k \sigma_k \left(\frac{|x_{i_k}^\top \beta^* - y_{i_k}|}{p_{i_k}} - \frac{|x_{i_k}^\top \beta - y_{i_k}|}{p_{i_k}} \right) \right| \right)^l \right] \\ &\quad \leq 2^{l+1} \mathop{\mathbb{E}}_{S,\sigma} \left[\left(\max_{\|X\beta^* - X\beta\|_1 = 1} \left| \sum_k \sigma_k \left(\frac{x_{i_k}^\top \beta^* - x_{i_k}^\top \beta}{p_{i_k}} \right) \right| \right)^l \right] \end{aligned} \tag{11}$$

Proof. We take $\Phi(x) = x^l$, which is convex and increasing for $l > 1$, let $\mathcal{F}$ be the set of functions f_β where $f_\beta(k) = \frac{x_{i_k}^\top \beta^* - x_{i_k}^\top \beta}{p_{i_k}}$ and β satisfies $\|X\beta^* - X\beta\|_1 = 1$, and let ϕ_k be defined as

$$\phi_k(z) = \frac{|x_{i_k}^\top \beta^* - y_{i_k}|}{p_{i_k}} - \frac{|x_{i_k}^\top \beta^* - z p_{i_k} - y_{i_k}|}{p_{i_k}}.$$

This satisfies

$$\phi_k(f_\beta(k)) = \phi_k\left(\frac{x_{i_k}^\top \beta^* - x_{i_k}^\top \beta}{p_{i_k}} \right) = \frac{|x_{i_k}^\top \beta^* - y_{i_k}|}{p_{i_k}} - \frac{|x_{i_k}^\top \beta - y_{i_k}|}{p_{i_k}}.$$

This is a contraction, since

$$\begin{aligned} |\phi_k(z_1) - \phi_k(z_2)| &= \left| \frac{|x_{i_k}^\top \beta^* - z_2 p_{i_k} - y_{i_k}|}{p_{i_k}} - \frac{|x_{i_k}^\top \beta^* - z_1 p_{i_k} - y_{i_k}|}{p_{i_k}} \right| \\ &\leq \frac{|z_1 p_{i_k} - z_2 p_{i_k}|}{p_{i_k}} \leq |z_1 - z_2| \end{aligned}$$

Applying Lemma 30 with these parameters, we have

$$\begin{aligned} &\mathop{\mathbb{E}}_{\sigma} \left[\left(\frac{1}{2} \max_{\|X\beta^* - X\beta\| = 1} \left| \sum_k \sigma_k \left(\frac{|x_{i_k}^\top \beta^* - y_{i_k}|}{p_{i_k}} - \frac{|x_{i_k}^\top \beta - truey_{i_k}|}{p_{i_k}} \right) \right| \right)^l \right] \\ &\quad \leq \frac{3}{2} \mathop{\mathbb{E}}_{\sigma} \left[\left(\max_{\|X\beta^* - X\beta\|_1 = 1} \left| \sum_k \sigma_k \left(\frac{x_{i_k}^\top \beta^* - x_{i_k}^\top \beta}{p_{i_k}} \right) \right| \right)^l \right] \end{aligned}$$

After taking the expectation with respect to S and multiplying both sides by 2^l, this gives the statement of the lemma. ◀

From here, we use two separate results to show the appropriate row counts for the constant and high probability cases. The constant probability case is left for the full version of this paper, [14].

For high probability row-counts, we use a lemma from [4]:

▶ **Lemma 32** (8.2, 8.3, 8.4 in [4]). *There exists constant C such that for any $X \in \mathbb{R}^{n\times d}$ with all ℓ_1 Lewis weights less than $C\frac{\varepsilon^2}{\log(\frac{n}{\delta})}$ and $l = \log(2n/\delta)$, then*

$$\mathbb{E}_\sigma\left[\left(\max_{\|X\beta\|_1=1}\left|\sum_{i=1}^n \sigma_i x_i^\top \beta\right|\right)^l\right] \le \frac{\varepsilon^l\delta}{2} \tag{12}$$

We want a similar statement, but for arbitrary matrices, with no bounds placed on the Lewis weights. To do this, we construct a new, related matrix using the following lemma, which is proved in Appendix B:

▶ **Lemma 33** (Similar to [4] Lemma B.1). *Let X be any matrix, and let W be the matrix that has the Lewis weights of X in the diagonal entries. Let $N \ge \frac{d}{\varepsilon^2}\log\frac{d}{\varepsilon\delta}$. There exist constants C_1, C_2, C_3 such that we can construct a matrix X' such that*

- *X' has C_1dN rows,*
- *$X'^\top W'^{-1}X' \succeq X^\top W^{-1}X$, (where W' is the matrix that has the Lewis weights of X' in the diagonal entries),*
- *$\|X'\beta\|_1 \le C_2\|X\beta\|_1$ for all β,*
- *the Lewis weights of X' are bounded by $\frac{C_3}{N}$.*

▶ **Lemma 34.** *Consider $X \in \mathbb{R}^{n\times d}$ with ℓ_1 Lewis weights w_i. Let p_i be some set of sampling values such that $N = \sum_i p_i$ and, for some constants C, C_1, C_4,*

$$p_i \ge \frac{\log\left(\frac{N+C_1Nd}{\delta}\right)}{C\varepsilon^2}w_i$$

Then, if $N \ge C_4\frac{d}{\varepsilon^2}\log\frac{d}{\varepsilon\delta}$ and if $S \sim \mathcal{S}(\{p_i\}_{i\in[n]})$, then

$$\mathbb{E}_{S,\sigma}\left[\left(\max_{\|X\beta\|_1=1}\left|\sum_{k=1}^N \sigma_k\frac{x_{i_k}^\top\beta}{p_{i_k}}\right|\right)^l\right] \le \frac{\varepsilon^l\delta}{2} \tag{13}$$

Proof of Lemma 34. Ideally the Lewis weights of SX would be bounded by $C\frac{\varepsilon^2}{\log\frac{N}{\delta}}$ and we could directly apply Lemma 32 to SX to obtain a bound on the moment. However, we do not know this. Instead, we first construct X' using X as described in Lemma 33. We then construct a new matrix X'' by stacking X' on top of SX. Define W'' to be the diagonal matrix consisting of the ℓ_1 Lewis weights of X''. Define, for convenience, $R = N + C_1Nd$, which is the number of rows X'' has.

We can bound the term on the left side of (13) by the same term, summing over the rows of X'' instead. That is,

$$\mathbb{E}_{S,\sigma}\left[\left(\max_{\|X\beta\|=1}\left|\sum_{k=1}^N \sigma_k\frac{x_{i_k}^\top\beta}{p_{i_k}}\right|\right)^l\right] \le \mathbb{E}_{S,\sigma}\left[\left(\max_{\|X\beta\|=1}\left|\sum_{i=1}^R \sigma_i x_i''^\top\beta\right|\right)^l\right]$$

Our goal is to apply Lemma 32 to the right side. To do this, we need to show the correct bound on its Lewis weights, and then have the term be a maximum over $\|X''\beta\|_1 = 1$, rather than $\|X\beta\|_1 = 1$.

Bounding the Lewis weights of X''. By Lemma 9, the ℓ_1 Lewis weights of a matrix do not increase when more rows are added. So, the rows in X'' that are from X' have Lewis weights that are bounded above by $C_3 \frac{\varepsilon^2}{\log\left(\frac{d}{\varepsilon\delta}\right)}$. Further,

$$\begin{aligned} X''^\top W''^{-1} X'' &= \sum_{i=1}^{R} \frac{1}{w_i''} x_i''(x_i'')^\top \\ &\succeq \sum_{i=1}^{R-N} \frac{1}{w_k''} x_k''(x_k'')^\top \qquad\qquad \text{since } \sum_{i=kC_1d^2+1}^{N} \frac{1}{w_i''} x_i''(x_i'')^\top \succeq 0 \\ &= X'^\top W'^{-1} X' \succeq X^\top W^{-1} X. \end{aligned}$$

So, any row $y_i = x_i/p_i$ in X'' that is from SX satisfies

$$\begin{aligned} w_i''^2 = y_i^\top (X''^\top W''^{-1} X'')^{-1} y_i \le y_i^\top (X^\top W^{-1} X)^{-1} y_i &= \frac{1}{p_i^2} x_i^\top (X^\top W^{-1} X)^{-1} x_i \\ &\le \left(\frac{C\varepsilon^2}{\log\left(\frac{R}{\delta}\right)} \frac{1}{w_i} \right)^2 \cdot w_i^2 = \left(\frac{C\varepsilon^2}{\log\left(\frac{R}{\delta}\right)} \right)^2 \end{aligned}$$

which means that all of the Lewis weights of X'' are less than the larger of $C\frac{\varepsilon^2}{\log\left(\frac{R}{\delta}\right)}$ and $C_3 \frac{\varepsilon^2}{\log\left(\frac{d}{\varepsilon\delta}\right)}$. Now, for small enough ε, δ, $\log \frac{R}{\delta} \le \frac{C}{C_3} \log \frac{d}{\varepsilon\delta}$, we have the Lewis weight upper bound for all rows of X'' is $C\frac{\varepsilon^2}{\log\left(\frac{R}{\delta}\right)}$

Renormalizing to maximize over $\|X''\beta\|_1 = 1$. If we define the following

$$F := \max_{\|X\beta\|_1 = 1} |\|SX\beta\|_1 - \|X\beta\|_1|$$

then,

$$\|X''\beta\|_1 = \|SX\beta\|_1 + \|X'\beta\|_1 \le (1 + C_2 + F)\|X\beta\|_1$$

So, we get

$$\begin{aligned} \left(\max_{\|X\beta\|=1} \left| \sum_{k=1}^{R} \sigma_k x_k''^\top \beta \right| \right)^l &\le (1 + C_2 + F)^l \left(\max_{\|X''\beta\|=1} \left| \sum_{k=1}^{R} \sigma_k x_k''^\top \beta \right| \right)^l \\ &\le 2^{l-1}((1+C_2)^l + F^l) \left(\max_{\|X''\beta\|=1} \left| \sum_{k=1}^{R} \sigma_k x_k''^\top \beta \right| \right)^l \end{aligned}$$

Taking expectations of either side over just the Rademacher variables,

$$\mathop{\mathbb{E}}_{\sigma} \left[\left(\max_{\|X\beta\|=1} \left| \sum_{k=1}^{R} \sigma_k x_k''^\top \beta \right| \right)^l \right] \le 2^{l-1}((1+C_2)^l + F^l) \mathop{\mathbb{E}}_{\sigma} \left[\left(\max_{\|X''\beta\|=1} \left| \sum_{k=1}^{R} \sigma_k x_k''^\top \beta \right| \right)^l \right]$$

Applying Lemma 32 to X''. Since X'' has R rows, and the correct Lewis weight bound, we can simply apply Lemma 32 to the right side above

$$\mathop{\mathbb{E}}_{\sigma} \left[\left(\max_{\|X\beta\|=1} \left| \sum_{k=1}^{R} \sigma_k x_k''^\top \beta \right| \right)^l \right] \le 2^{l-1}((1+C_2)^l + F^l)) \frac{\varepsilon^l \delta}{2}$$

Now, by Lemma 12, we know that $\mathbb{E}_S[F^l] \leq \varepsilon^l \delta$. So, taking the expectation with respect to the sampling matrices of either side of the above, we get, for small enough ε, δ,

$$\mathop{\mathbb{E}}_{S,\sigma}\left[\left(\max_{\|X\beta\|=1}\left|\sum_{k=1}^{kC_1d^2+N} \sigma_k x_k''^\top \beta\right|\right)^l\right] \leq 2^{l-1}((1+C_2)^l + \varepsilon^l \delta)\frac{\varepsilon^l \delta}{2} \leq 2^l(1+C_2)^l \frac{\varepsilon^l \delta}{2}$$

So, solving the problem for $\varepsilon' = \frac{\varepsilon}{2+2C_2}$ gives the correct bound. ◀

Finally, we can show Lemma 28

Proof of Lemma 28. Take $l = \log(2n/\delta)$, $N = 5\frac{(1+C_1)C_3}{C}\frac{d}{\varepsilon^2}\log\frac{d}{\varepsilon\delta}$. Then, we apply Lemma 29, Lemma 31, and Lemma 34 to get

$$M \leq 2^{2l}\varepsilon^l \delta$$

which, solving the problem for $\varepsilon/4$, gives the correct bound. Then, applying Markov's inequality, we get that with probability δ,

$$\max_{\|X\beta^* - X\beta\|=1} |(\|SX\beta^* - Sy\|_1 - \|SX\beta - Sy\|_1) - (\|X\beta^* - y\|_1 - \|X\beta - y\|_1)| \leq \varepsilon$$

Finally, scaling up appropriately gives, in generality,

$$|(\|SX\beta^* - Sy\|_1 - \|SX\beta - Sy\|_1) - (\|X\beta^* - y\|_1 - \|X\beta - y\|_1)| \leq \varepsilon\|X\beta^* - X\beta\|_1$$ ◀

References

1 Xue Chen and Michał Dereziński. Query complexity of least absolute deviation regression via robust uniform convergence. *arXiv preprint arXiv:2102.02322*, 2021.

2 Xue Chen and Eric Price. Active regression via linear-sample sparsification. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 663–695, Phoenix, USA, June 25–28 2019. PMLR. URL: `http://proceedings.mlr.press/v99/chen19a.html`.

3 Kenneth L. Clarkson. Subgradient and sampling algorithms for l1 regression. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, page 257–266, USA, 2005. Society for Industrial and Applied Mathematics.

4 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. `doi:10.1145/2746539.2746567`.

5 Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W Mahoney. Sampling algorithms and coresets for \ell_p regression. *SIAM Journal on Computing*, 38(5):2060–2078, 2009.

6 Michał Derezinski and Michael W Mahoney. Determinantal point processes in randomized numerical linear algebra. *Notices of the American Mathematical Society*, 68(1), 2021.

7 Michał Dereziński and Manfred K Warmuth. Unbiased estimates for linear regression via volume sampling. *arXiv preprint arXiv:1705.06908*, 2017.

8 Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Sampling algorithms for l 2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136, 2006.

9 David Durfee, Kevin A. Lai, and Saurabh Sawlani. ℓ_1 regression using lewis weights preconditioning and stochastic gradient descent. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1626–1656. PMLR, July 06–09 2018. URL: `http://proceedings.mlr.press/v75/durfee18a.html`.

10 E. N. Gilbert. A comparison of signalling alphabets. *The Bell System Technical Journal*, 31(3):504–522, 1952. doi:10.1002/j.1538-7305.1952.tb01393.x.

11 T. Lattimore and C. Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2020.

12 M. Ledoux and M. Talagrand. Comparison theorems, random geometry and some limit theorems for empirical processes. *Ann. Probab.*, 17(2):596–631, April 1989. doi:10.1214/aop/1176991418.

13 Michael W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(2):123–224, 2011. doi:10.1561/2200000035.

14 Aditya Parulekar, Advait Parulekar, and Eric Price. L1 regression with lewis weights subsampling. *CoRR*, abs/2105.09433, 2021. arXiv:2105.09433.

15 Michel Talagrand. Embedding subspaces of l1 into ln 1. *Proceedings of the American Mathematical Society*, 108(2):363–369, 1990. URL: http://www.jstor.org/stable/2048283.

16 Michel Talagrand. Embedding subspaces of lp in lpn. In J. Lindenstrauss and V. Milman, editors, *Geometric Aspects of Functional Analysis*, pages 311–326, Basel, 1995. Birkhäuser Basel.

17 David P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1–2):1–157, 2014. doi:10.1561/0400000060.

A Constant-factor approximation

▶ **Theorem 35.** *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 $O(d\log d)$, 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*

$$\|X\widehat{\beta} - y\|_1 \le 41 \min_\beta \|X\beta - y\|_1$$

with probability 0.9.

Proof. Since we just want a constant factor approximation, we can take $\mathcal{S}$ to be the distribution over constant probability Lewis weight ℓ_1-subspace embeddings, so that $\|X\beta\|_1 \le 2\|SX\beta\|_1$ with probability at least 0.9. We have

$$\begin{aligned}\|X\widehat{\beta} - y\|_1 &\le \|X\widehat{\beta} - X\beta^*\|_1 + \|X\beta^* - y\|_1\\ &\le 2\|SX\widehat{\beta} - SX\beta^*\|_1 + \|X\beta^* - y\|_1\\ &\le 2(\|SX\widehat{\beta} - Sy\|_1 + \|SX\beta^* - Sy\|_1) + \|X\beta^* - y\|_1\\ &\le 4(\|SX\beta^* - Sy\|_1) + \|X\beta^* - y\|_1\end{aligned}$$

where in the last inequality, we have used the fact that $\widehat{\beta}$ is the minimizer of $\|SX\beta - Sy\|_1$. Now, by Markov's inequality, with probability 0.9, $\|SX\beta^* - Sy\|_1 \le 10\|X\beta^* - y\|_1$. So, we have with probability 0.81,

$$\|X\widehat{\beta} - y\|_1 \le 41\|X\beta^* - y\|_1$$

Since we only used a constant-factor subspace embedding, the row count would be $O(d\log d)$. ◀

B Proofs of Lemmas

▶ **Lemma 10.** *Let $X \in \mathbb{R}^{n\times d}$, and let $X' \in \mathbb{R}^{kn\times d}$ be X stacked on itself k times, with each row scaled down by k. Then, each of the Lewis weights is reduced by a factor of k.*

Proof. Let $\{w_i\}_{i=1}^n$ be the Lewis weights of X, and let $\{w'_i\}_{i=1}^{kn}$ be the Lewis weights of X'. Let x_i be the ith row of X, and similarly let x'_i be the ith row of X'. Let the ordering of the rows be such that $x'_{jn+i} = \frac{1}{k}x_i$ for $0 \le j < k$. Let W be the diagonal matrix where $W_{ii} = w_i$. Since Lewis weights are defined circularly, we just need to check that the suggested weights work, and by uniqueness, they will be correct.

We know that $w_i^2 = x_i^\top (X^\top W^{-1} X)^{-1} x_i$. Therefore, if we take W' to be the diagonal matrix of size $kn \times kn$, and set the diagonal entries to be the Lewis weights of X divided by k, repeated k times, then we have

$$X'^\top W'^{-1} X' = \sum_{i=1}^{kn} \frac{1}{w'_i} x'_i x_i'^\top = \sum_{i=1}^{kn} \frac{k}{w_i} x'_i x_i'^\top = k \sum_{i=1}^{n} \frac{k}{w_i} \cdot \frac{1}{k^2} x_i x_i^\top$$

In the last expression above, we are only summing over the first set of rows in X', which are the scaled rows of X, and then multiplying by k since they are repeated k times. Now,

$$k \sum_{i=1}^{n} \frac{k}{w_i} \cdot \frac{1}{k^2} x_i x_i^\top = \sum_{i=1}^{n} \frac{1}{w_i} x_i x_i^\top = X^\top W^{-1} X$$

So, finally, for an arbitrary row x'_{jn+i}, which corresponds to row x_i in the original matrix, we get its Lewis weight:

$$w'^2_{jn+i} = x'^\top_{jn+i} (X'^\top W'^{-1} X')^{-1} x'_{jn+i} = \frac{1}{k^2} x_i^\top (X^\top W^{-1} X)^{-1} x_i = \frac{w_i^2}{k^2}$$

which proves that our suggested Lewis weights are consistent. ◀

▶ **Lemma 29.** *Given a matrix $X \in \mathbb{R}^{n\times d}$, let $\mathcal{S}(\{p_i\}_{i\in[n]})$ be any sampling-and-reweighting disribution, and let i_k be the row-indices chosen by this sampling matrix such that $S_{k,i_k} = \frac{1}{p_{i_k}}$. Let σ_k be independent Rademacher variables that are ± 1 each with probability* 0.5*. Then,*

$$M \le 2^l \mathop{\mathbb{E}}_{S,\sigma}\left[\left(\max_{\|X\beta^*-X\beta\|=1}\left|\sum_k \sigma_k \left(\frac{|x_{i_k}^\top \beta^* - y_{i_k}|}{p_{i_k}} - \frac{|x_{i_k}^\top \beta - y_{i_k}|}{p_{i_k}}\right)\right|\right)^l\right] \tag{10}$$

Proof. We proceed by symmetrization. Since the matrix S scales the rows by the probability they are picked with, the expectation of $\|SM\beta\|_1$ is just $\|M\beta\|_1$, for any matrix M and vector β. So, adding or subtracting the same term with a different sampling matrix S', $(\|S'X\beta^* - S'y\|_1 - \|S'X\beta - S'y\|_1) - (\|X\beta^* - y\|_1 - \|X\beta - y\|_1)$, is just adding a mean zero term, and since taking the lth power of a maximum is convex, this can only increase the expectation. That is,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S,S'}\left[\left(\max_{\|X\beta^*-X\beta\|=1} |\,(\|SX\beta^* - Sy\|_1 - \|SX\beta - Sy\|_1) - (\|X\beta^* - y\|_1 - \|X\beta - y\|_1)\,|\right)^l\right] \\
&\le \mathop{\mathbb{E}}_{S,S'}\Bigg[\Bigg(\max_{\|X\beta^*-X\beta\|=1} |\,((\|SX\beta^* - Sy\|_1 - \|SX\beta - Sy\|_1) - (\|X\beta^* - y\|_1 - \|X\beta - y\|_1)) \\
&\qquad - \left((\|S'X\beta^* - S'y\|_1 - \|S'X\beta - S'y\|_1\right) - (\|X\beta^* - y\|_1 - \|X\beta - y\|_1))\,|\Bigg)^l\Bigg]
\end{aligned}$$

So, we can bound M as

$$\begin{aligned}
M \le \mathop{\mathbb{E}}_{S,S'}\Bigg[\Bigg(&\max_{\|X\beta^*-X\beta\|=1} |\,(\|SX\beta^* - Sy\|_1 - \|SX\beta - Sy\|_1) - \\
&(\|S'X\beta^* - S'y\|_1 - \|S'X\beta - S'y\|_1)\,|\Bigg)^l\Bigg]
\end{aligned}$$

Let i_k be the indices chosen by S, and i'_k the indices chosen by S'. Rewriting this as a sum,

$$M \leq \mathop{\mathbb{E}}_{S,S'}\left[\left(\max_{\|X\beta^*-X\beta\|=1}\left|\sum_k\left(\frac{|x_{i_k}^\top\beta^*-y_{i_k}|}{p_{i_k}}-\frac{|x_{i_k}^\top\beta-y_{i_k}|}{p_{i_k}}\right)-\right.\right.\right.$$
$$\left.\left.\left.\sum_k\left(\frac{|x_{i'_k}^\top\beta^*-y_{i'_k}|}{p_{i'_k}}-\frac{|x_{i'_k}^\top\beta-y_{i'_k}|}{p_{i'_k}}\right)\right|\right)^l\right]$$

Now, since i_k and i'_k are independent and identically distributed, randomly swapping elements from either sum does not change the distribution. This amounts to adding a random sign σ_k to the terms, where $\sigma_k = \pm 1$ independently with probability $1/2$. So,

$$\begin{aligned}
M &\leq \mathop{\mathbb{E}}_{S,S',\sigma}\left[\left(\max_{\|X\beta^*-X\beta\|=1}\left|\sum_k\sigma_k\left(\frac{|x_{i_k}^\top\beta^*-y_{i_k}|}{p_{i_k}}-\frac{|x_{i_k}^\top\beta-y_{i_k}|}{p_{i_k}}\right)-\right.\right.\right.\\
&\qquad\left.\left.\left.\sum_k\sigma_k\left(\frac{|x_{i'_k}^\top\beta^*-y_{i'_k}|}{p_{i'_k}}-\frac{|x_{i'_k}^\top\beta-y_{i'_k}|}{p_{i'_k}}\right)\right|\right)^l\right]\\
&\leq \mathop{\mathbb{E}}_{S,S',\sigma}\left[\left(\max_{\|X\beta^*-X\beta\|=1}\left|\sum_k\sigma_k\left(\frac{|x_{i_k}^\top\beta^*-y_{i_k}|}{p_{i_k}}-\frac{|x_{i_k}^\top\beta-y_{i_k}|}{p_{i_k}}\right)\right|+\right.\right.\\
&\qquad\left.\left.\max_{\|X\beta^*-X\beta\|=1}\left|\sum_k\sigma_k\left(\frac{|x_{i'_k}^\top\beta^*-y_{i'_k}|}{p_{i'_k}}-\frac{|x_{i'_k}^\top\beta-y_{i'_k}|}{p_{i'_k}}\right)\right|\right)^l\right]\\
&\leq 2^l\mathop{\mathbb{E}}_{S,\sigma}\left[\left(\max_{\|X\beta^*-X\beta\|=1}\left|\sum_k\sigma_k\left(\frac{|x_{i_k}^\top\beta^*-y_{i_k}|}{p_{i_k}}-\frac{|x_{i_k}^\top\beta-y_{i_k}|}{p_{i_k}}\right)\right|\right)^l\right]
\end{aligned}$$

Where the final inequality follows from $(a+b)^l \leq 2^{l-1}(a^l+b^l)$. Putting these together,

$$M \leq 2^l\mathop{\mathbb{E}}_{S,\sigma}\left[\left(\max_{\|X\beta^*-X\beta\|=1}\left|\sum_k\sigma_k\left(\frac{|x_{i_k}^\top\beta^*-y_{i_k}|}{p_{i_k}}-\frac{|x_{i_k}^\top\beta-y_{i_k}|}{p_{i_k}}\right)\right|\right)^l\right] \tag{14}$$

◀

▶ **Lemma 33** (Similar to [4] Lemma B.1). *Let X be any matrix, and let W be the matrix that has the Lewis weights of X in the diagonal entries. Let $N \geq \frac{d}{\varepsilon^2}\log\frac{d}{\varepsilon\delta}$. There exist constants C_1, C_2, C_3 such that we can construct a matrix X' such that*

- *X' has $C_1 dN$ rows,*
- *$X'^\top W'^{-1}X' \succeq X^\top W^{-1}X$, (where W' is the matrix that has the Lewis weights of X' in the diagonal entries),*
- *$\|X'\beta\|_1 \leq C_2\|X\beta\|_1$ for all β,*
- *the Lewis weights of X' are bounded by $\frac{C_3}{N}$.*

Proof. Given matrix X, we can use Lemma B.1 from [4] to construct a new matrix X_1 that satisfies

- X_1 has $C_1 d^2$ rows,
- $X_1^\top W_1^{-1}X_1 \succeq X^\top W^{-1}X$, (where W_1 is the matrix that has the Lewis weights of X_1 in the diagonal entries),
- $\|X_1\beta\|_1 \leq C_2\|X_1\beta\|_1$ for all β,
- the Lewis weights of X_1 are bounded by $\frac{C_3}{d}$.

So, we can take this matrix and stack it on itself $k = \frac{N}{d}$ times, while scaling each row down by the same k. This will be our matrix X'. X' will then have $k = C_1 N d$ rows, which satisfies the first bullet. Also, by Lemma 10, this shrinks the Lewis weights by a factor of k, which changes the Lewis weight upper bound to $\frac{C_3}{kd} = \frac{C_3}{N}$ which is what we need. Now, since we are repeating rows k times, but each row is scaled down by k, we have $\|X_1\beta\|_1 = \|X'\beta\|_1$ for all β. Therefore, $\|X'\beta\|_1 \leq C_2\|X\beta\|_1$ for all β. Finally, as in the proof of Lemma 10, we know that since we have duplicated the rows of X_1 k times but scaled them down by k, $X_1^\top W_1^{-1} X_1 = X'^\top W'^{-1} X'$, and so we are done. ◀

B.1 Proof of Claims 15, 18, 19, 24, and 25

▷ Claim 15. For all $\beta \in H$, with probability at least $1 - \delta$,

$$(1-\varepsilon)\,\mathbb{E}_{(X,Y)\sim P}\left[|X^\top\beta - Y|\right] \leq \frac{1}{n}\|\mathbf{X}\beta - y\|_1 \leq (1+\varepsilon)\,\mathbb{E}_{(X,Y)\sim P}\left[|X^\top\beta - Y|\right]$$

Proof of Claim 15. By assumption, we know that $X^\top\beta, Y \in [-1,1]$, so, $|X^\top\beta - Y| \in [0,2]$. So, for fixed β, by Hoeffding's on the rows of $\mathbf{X}\beta - y$, we have that if $n \geq \frac{8}{\varepsilon^2}\log\frac{2}{\delta'}$, then with probability at least $1 - \delta'$,

$$\left(1 - \frac{\varepsilon}{2}\right)\mathbb{E}_{(X,Y)\sim P}\left[|X^\top\beta - Y|\right] \leq \frac{1}{n}\|\mathbf{X}\beta - y\|_1 \leq \left(1 + \frac{\varepsilon}{2}\right)\mathbb{E}_{(X,Y)\sim P}\left[|X^\top\beta - Y|\right] \quad (15)$$

Now, we construct a $\frac{\varepsilon}{2d}$-covering S of the unit ℓ_∞ ball H, with fewer than $\left(\frac{4d}{\varepsilon}\right)^d$ elements, so that for any β, there is some $\beta_c \in S$ such that $\|\beta - \beta_c\|_\infty \leq \frac{\varepsilon}{2d}$. To do this, simply take $S = \{\beta : \beta_i = k\frac{\varepsilon}{2d}, k \in \mathbb{Z} \cap [-2d/\varepsilon, 2d/\varepsilon]\}$.

Note that $\mathbf{X}$ has rows on the hypercube. So, if we denote $x_{i,j}$ to be the entry of $\mathbf{X}$ in the ith row and jth column, then $x_{i,j} \in \{-1,1\}$. Therefore, for any β,

$$\|\mathbf{X}\beta\|_1 = \sum_{i=1}^{n}|x_i^\top\beta| \leq \sum_{i=1}^{n}\sum_{j=1}^{d}|x_{i,j}\beta_j| \leq \sum_{i=1}^{n}\sum_{j=1}^{d}|\beta_j| \leq nd\|\beta\|_\infty$$

Therefore, we can apply Hoeffding's, as in (15), with $\delta' = \delta\left(\frac{\varepsilon}{4d}\right)^d$, and union bound over the set S, to get that for any $\beta \in S$, with probability at least $1 - \delta$, (15) holds.

Then, for any $\beta \in H$, by the covering property, we can find some $\beta_c \in S$ such that

$$\|\beta - \beta_c\|_\infty \leq \frac{\varepsilon}{d} \implies \|\mathbf{X}\beta - \mathbf{X}\beta_c\|_1 \leq n\varepsilon. \quad (16)$$

We have

$$\|\mathbf{X}\beta_c - y\|_1 - \|\mathbf{X}\beta_c - \mathbf{X}\beta\|_1 \leq \|\mathbf{X}\beta - y\|_1 \leq \|\mathbf{X}\beta - \mathbf{X}\beta_c\|_1 + \|\mathbf{X}\beta_c - y\|_1$$

So, combining (15) and (16), and dividing by n, we finally have that if $n \geq \frac{8}{\varepsilon^2}\left(\log\frac{2}{\delta} + d\log\frac{4d}{\varepsilon}\right)$, then for all $\beta \in H$,

$$(1-\varepsilon)\,\mathbb{E}_{(X,Y)\sim P}\left[|X^\top\beta - Y|\right] \leq \frac{1}{n}\|\mathbf{X}\beta - y\|_1 \leq (1+\varepsilon)\,\mathbb{E}_{(X,Y)\sim P}\left[|X^\top\beta - Y|\right] \quad ◁$$

▷ Claim 18. For $D, \mathcal{B}$ as chosen above, $l(\beta^*) = 1 - 2\varepsilon$.

Proof of Claim 18. The ℓ_1 error for the correct β is given by

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim P}\left|X^\top\beta^* - Y\right| \\
&\quad = \mathbb{E}_X[E_{Y\sim P(\cdot|X)}||X^\top\beta^* - Y|] && \text{by independence} \\
&\quad = \mathbb{E}_X[(\frac{1}{2}+\varepsilon)|X^\top\beta^* - X^\top\beta^*| + (\frac{1}{2}-\varepsilon)|X^\top\beta^* + X^\top\beta^*|] \\
&\quad = \mathbb{E}_X[(1-2\varepsilon)|X^\top\beta^*|] && \beta^*\in\mathcal{H} \\
&\quad = 1-2\varepsilon
\end{aligned}$$

◁

▷ Claim 19. For $D, \mathcal{B}$ as chosen above, we have for all $\beta \in \mathcal{B}$, $l(\beta) - l(\beta^*) = \frac{2\varepsilon}{d}||\beta - \beta^*||_1$.

Proof of Claim 19.

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim P}\left|X^\top\beta - Y|\right| \\
&\quad = \mathbb{E}_X\left[E_{Y\sim P(\cdot|X)}\left|X^\top\beta - Y|\right|\right] \\
&\quad = \mathbb{E}_X\left[\left(\frac{1}{2}+\varepsilon\right)\left|X^\top\beta - X^\top\beta^*\right| + \left(\frac{1}{2}-\varepsilon\right)\left|X^\top\beta + X^\top\beta^*\right|\right] \\
&\quad = (1-2\varepsilon) + 2\varepsilon\,\mathbb{E}_X[X^\top\beta - X^\top\beta^*] \\
&\quad = (1-2\varepsilon) + 2\varepsilon\frac{1}{d}||\beta-\beta^*||_1
\end{aligned}$$

◁

▷ Claim 21. For any β, $\max\{\ell_{\beta^*_{(1)}}(\beta) - \ell_{\beta^*_{(1)}}(\beta^*_{(1)}), \ell_{\beta^*_{(2)}}(\beta) - \ell_{\beta^*_{(2)}}(\beta^*_{(2)})\} > 2\varepsilon$

Proof of Claim 21.

$$\begin{aligned}
l(\beta) + l(\beta) &= 2 - 4\varepsilon + \frac{2\varepsilon}{d}\|\beta^*_{(1)} - \beta\|_1 + \frac{2\varepsilon}{d}\|\beta^*_{(2)} - \beta\|_1 \\
&\geq 2 - 4\varepsilon + \frac{2\varepsilon}{d}\|\beta^*_{(2)} - \beta^*_{(1)}\|_1 \\
&= 2
\end{aligned}$$

$$\implies \max\{\ell_{\beta^*_{(1)}}(\beta) - \ell_{\beta^*_{(1)}}(\beta^*_{(1)}), \ell_{\beta^*_{(2)}}(\beta) - \ell_{\beta^*_{(2)}}(\beta^*_{(2)})\} > 2\varepsilon, \qquad \forall\beta\in\mathbb{R}^d$$

◁

▷ Claim 24. For all $i \in [d], \beta \in \mathbb{R}^d$, we have $\ell_{\beta^*_{(i)}}(\beta) \geq \frac{1}{4d}$ with equality when $\beta = \beta^*_{(i)}$

Proof of Claim 24.

$$\begin{aligned}
\ell_{\beta^*_{(i)}}(\beta) &= \frac{1}{d}\sum_{j\neq i}|\beta_j| + \frac{\frac{1}{2}+\varepsilon}{d}|1-\beta_i| + \frac{\frac{1}{2}-\varepsilon}{d}|\beta_i| \\
&\geq \frac{\frac{1}{2}-\varepsilon}{d}\left(|\beta_i| + |1-\beta_i|\right) + \frac{2\varepsilon}{d}|1-\beta_i| \geq \frac{\frac{1}{2}-\varepsilon}{d}
\end{aligned}$$

◁

▷ Claim 25. Any $\beta \in \mathbb{R}^d$ can only satisfy $\ell_{\beta^*_{(i)}}(\widehat{\beta}) < \frac{1}{2d}$ for one $i \in [d]$.

Proof of Claim 25. Indeed, suppose β was such that $\ell_{\beta^*_{(I)}}(\beta), \ell_{\beta^*_{(J)}}(\beta) < \frac{1}{2d}$. Then we must have

$$\begin{aligned}
\frac{1}{2d} &\geq \ell_{\beta^*_{(I)}}(\beta) \\
&= \frac{1}{d}\sum_{j \neq I} |\beta_j| + \frac{\frac{1}{2} - \varepsilon}{d}\left(|\beta_I| + |1 - \beta_i|\right) + \frac{2\varepsilon}{d}|1 - \beta_I| \\
&\geq \frac{1}{d}\sum_{j \neq I} |\beta_j| + \frac{\frac{1}{2} - \varepsilon}{d} + \frac{2\varepsilon}{d}|1 - \beta_I| \\
\iff \varepsilon &\geq \sum_{j \neq I} |\beta_j| + 2\varepsilon|1 - \beta_I| \\
&\geq \sum_{j \neq I} |\beta_j| + 2\varepsilon - 2\varepsilon|\beta_I| \\
\iff 2|\beta_I| &\geq \|\beta\|_1 + 2\varepsilon
\end{aligned}$$

Similarly for J, so we would have $\|\beta\| \geq |\beta_I| + |\beta_J| \geq \|\beta\|_1 + 2\varepsilon$. ◁