# $\ell_p$-Regression in the Arbitrary Partition Model of Communication

**Yi Li** YILI@NTU.EDU.SG
*Nanyang Technological University*

**Honghao Lin** HONGHAOL@ANDREW.CMU.EDU
*Carnegie Mellon University*

**David Woodruff** DWOODRUF@ANDREW.CMU.EDU
*Carnegie Mellon University*

**Editors:** Gergely Neu and Lorenzo Rosasco

## Abstract

We consider the randomized communication complexity of the distributed $\ell_p$-regression problem in the coordinator model, for $p \in (0, 2]$. In this problem, there is a coordinator and $s$ servers. The $i$-th server receives $A^i \in \{-M, -M+1, \ldots, M\}^{n\times d}$ and $b^i \in \{-M, -M+1, \ldots, M\}^n$ and the coordinator would like to find a $(1+\varepsilon)$-approximate solution to $\min_{x\in\mathbb{R}^n} \|(\sum_i A^i)x - (\sum_i b^i)\|_p$. Here $M \leq \mathrm{poly}(nd)$ for convenience. This model, where the data is additively shared across servers, is commonly referred to as the arbitrary partition model.

We obtain significantly improved bounds for this problem. For $p = 2$, i.e., least squares regression, we give the first optimal bound of $\widetilde{\Theta}(sd^2 + sd/\epsilon)$ bits.

For $p \in (1, 2)$, we obtain an $\widetilde{O}(sd^2/\varepsilon + sd/\,\mathrm{poly}(\varepsilon))$ upper bound. Notably, for $d$ sufficiently large, our leading order term only depends linearly on $1/\epsilon$ rather than quadratically. We also show communication lower bounds of $\Omega(sd^2 + sd/\varepsilon^2)$ for $p \in (0, 1]$ and $\Omega(sd^2 + sd/\varepsilon)$ for $p \in (1, 2]$. Our bounds considerably improve previous bounds due to (Woodruff et al. COLT, 2013) and (Vempala et al., SODA, 2020).

## 1. Introduction

Regression is a lightweight machine learning model used to capture linear dependencies between variables in the presence of noise. In this problem there is a (sometimes implicit) matrix $A \in \mathbb{R}^{n\times d}$ and a vector $b \in \mathbb{R}^n$ and the goal is to find a hyperplane $x \in \mathbb{R}^d$ for which $\|Ax - b\|$ is small for some loss function $\|\cdot\|$, which throughout this paper will be a norm. Here $A$ is known as the design matrix, $b$ the response vector, and $x$ the model parameters. We focus on the over-constrained case, when $n \gg d$, which corresponds to having many more examples than features. Although more sophisticated models can often achieve lower error, regression is often the most computationally efficient and the first model of choice.

One of the most popular loss functions is the $\ell_p$-norm, or equivalently its $p$-th power $\|y\|_p^p = \sum_{i=1}^n |y_i|^p$. When $p = 2$ this is least squares regression, which corresponds to the maximum likelihood estimator (MLE) in the presence of Gaussian noise. When the noise is more heavy-tailed, often $p < 2$ is chosen as the loss function since it is more robust to outliers. Indeed, since one is not squaring the differences, the optimal solution pays less attention to large errors. For example, $p = 1$ gives the MLE for Laplacian noise. While $p < 1$ results in non-convex loss functions, heuristics are still used given its robustness properties. When $p > 2$, the loss function is even more sensitive to outliers; it turns out that such $p$ cannot be solved without incurring a polynomial dependence on $n$ in the communication model we study, see below, and so our focus will be on $p \leq 2$.

It is often the case that data is either collected or distributed across multiple servers and then a key bottleneck is the *communication complexity*, i.e., the number of bits transmitted between the servers for solving a problem. We consider the standard coordinator model of communication, also known as the message-passing model, in which there is a site designated as the coordinator who has no input, together with $s$ additional sites, each receiving an input. There is a communication channel between the coordinator and each other server, and all communication goes through the coordinator. This model is convenient since it captures arbitrary point-to-point communication up to small factors, i.e., if server $i$ wants to send a message to server $j$, server $i$ can first send the message to the coordinator and then have it forwarded to server $j$. We note that in addition to the total communication, it is often desirable to minimize the time complexity on each server, and the protocols in this paper will all be time-efficient.

A natural question in any communication model is how the input is distributed. We study the *arbitrary partition model* of (Kannan et al., 2014; Boutsidis et al., 2016), which was studied for the related task of low rank approximation. In this model, the $i$-th server receives $A^i \in \{-M, -M+1, \ldots, M\}^{n \times d}$ and $b^i \in \{-M, -M+1, \ldots, M\}^n$ and the coordinator would like to find a $(1+\varepsilon)$-approximate solution to $\min_{x \in \mathbb{R}^n} \|(\sum_i A^i)x - (\sum_i b^i)\|_p$. Here $M \leq \mathrm{poly}(nd)$ for convenience. Note that this model gives more flexibility than the so-called *row partition model* in which each example and corresponding response variable is held on exactly one server, and which is a special case of the arbitrary partition model. For example, if each row $i$ of $A$ corresponds to an item and each column $j$ to a user and an entry $A_{i,j}$ corresponds to the number of times user $i$ purchased item $j$, then it might be that each server $t$ is a different shop where the user could purchase the item, giving a value $A^t_{i,j}$, and we are interested in $\sum_{t=1}^s A^t_{i,j}$, i.e., the matrix which aggregates the purchases across the shops. This communication model is also important for *turnstile streaming* where arbitrary additive updates are allowed to an underlying vector (Muthukrishnan, 2005), as low-communication protocols often translate to low memory streaming algorithms, while communication lower bounds often give memory lower bounds in the streaming model. The number of communication rounds often translates to the number of passes in a streaming algorithm. See, e.g., (Boutsidis et al., 2016), as an example of this connection for low rank approximation. We note that for $p > 2$, there is an $\Omega(n^{1-2/p})$ lower bound in the arbitrary partition model even for just estimating the norm of a vector (Bar-Yossef et al., 2004; Gronemeier, 2009; Jayram, 2009), and so we focus on the $p < 2$ setting.

The communication complexity of approximate regression was first studied in the coordinator model in the row partition model in (Woodruff and Zhang, 2013), though their protocols for $1 \leq p < 2$ use $\widetilde{O}(sd^{2+\gamma} + d^5 + d^{3+p}/\varepsilon^2)$ communication, where $\widetilde{O}(f)$ suppresses a $\mathrm{poly}(\log(sdn/\varepsilon))$ factor. These bounds were later improved in the coordinator model and in the row partition model in (Vempala et al., 2020), though the bounds are still not optimal, i.e., their lower bounds do not depend on $\varepsilon$, are suboptimal in terms of $s$, or hold only for deterministic algorithms. Their upper bounds also crucially exploit the row partition model, and it is unclear how to extend them to the arbitrary partition model. We will substantially improve upon these bounds.

Despite the previous work on understanding the communication complexity of a number of machine learning models (see, e.g., (Vempala et al., 2020) and the references therein), perhaps surprisingly for arguably the most basic task of regression, the optimal amount of communication required was previously unknown.

**Our Results** We obtain a lower bound of $\Omega(sd^2 + sd/\varepsilon^2)$ for $p \in (0, 1]$ and a lower bound of $\Omega(sd^2 + sd/\varepsilon)$ for $p \in (1, 2]$, both of which improve the only known lower bound of $\widetilde{\Omega}(d^2 + sd)$

| | | Communication | |
|---|---|---|---|
| $0 < p < 2$ | Upper Bound | $\widetilde{O}(sd^2/\varepsilon^2)$ | Folklore |
| $p = 2$ | Upper Bound | $\widetilde{O}(sd^2/\varepsilon)$ | (Clarkson and Woodruff, 2009) |
| $0 < p \leq 2$ | Lower Bound | $\Omega(d^2 + sd)$ | (Vempala et al., 2020) |
| $p = 1$ | Upper Bound* | $\widetilde{O}(\min(sd^2 + \frac{d^2}{\varepsilon^2}, \frac{sd^3}{\varepsilon}))$ | (Vempala et al., 2020) |
| $p = 2$ | Upper Bound* | $\widetilde{O}(sd^2)$ | (Vempala et al., 2020) |
| $1 \leq p < 2$ | Upper Bound* | $\widetilde{O}(sd^{2+\gamma} + d^5 + d^{3+p}/\varepsilon^2)$ | (Woodruff and Zhang, 2013) |
| $0 < p \leq 1$ | Lower Bound | $\Omega(sd^2 + sd/\varepsilon^2)$ | Theorem 13, 16 |
| $1 < p \leq 2$ | Lower Bound | $\Omega(sd^2 + sd/\varepsilon)$ | Theorem 13, 16 |
| $1 < p < 2$ | Upper Bound | $\widetilde{O}(sd^2/\varepsilon + sd/\operatorname{poly}(\varepsilon))$ | Theorem 20 |
| $p = 2$ | Upper Bound | $\widetilde{O}(sd^2 + sd/\varepsilon)$ | Theorem 17 |

Table 1: Summary of the results for the distributed $\ell_p$ regression problem. * denotes row partition model. The upper bound in the first row uses a median sketch of the $p$-stable distribution, which is time-inefficient (see, e.g. Backurs et al., 2015, Section F.1).

by (Vempala et al., 2020). We strengthen their $d^2$ lower bound by a multiplicative factor of $s$ and incorporate the dependence on $\varepsilon$ into their $sd$ lower bound.

When $p = 2$, we obtain an upper bound of $\widetilde{O}(sd^2 + sd/\varepsilon)$ bits, which matches our lower bound up to logarithmic factors. The total runtime of the protocol is $O(\sum_i \operatorname{nnz}(A^i) + s\operatorname{poly}(d/\varepsilon))$, which is optimal in terms of $\operatorname{nnz}(A^i)$. Here for a matrix $A$, $\operatorname{nnz}(A)$ denotes the number of non-zero entries of $A$. Our results thus largely settle the problem in the case of $p = 2$.

When $p \in (1, 2)$, we obtain an upper bound of $\widetilde{O}(sd^2/\varepsilon + sd/\operatorname{poly}(\varepsilon))$ bits with a runtime of $O(\sum_i \operatorname{nnz}(A^i)(d/\varepsilon^{O(1)}) + s\operatorname{poly}(d/\varepsilon))$. Note that if the $\widetilde{O}(sd^2/\varepsilon)$ term dominates, then our upper bound is optimal up to a $1/\varepsilon$ factor due to our lower bound. Interestingly, this beats a folklore sketching algorithm for which each server sketches their input using a shared matrix of $p$-stable random variables with $\widetilde{O}(d/\varepsilon^2)$ rows, sends their sketch to the coordinator with $\widetilde{O}(sd^2/\varepsilon^2)$ total communication, and has the coordinator add up the sketches and enumerate over all $x$ to find the best solution (see, e.g., Appendix F.1 of (Backurs et al., 2015) for a proof of this for $p = 1$). Moreover, our algorithm is time-efficient, while the sketching algorithm is not. In fact, any sketch that solves the harder problem of computing an $\ell_p$-subspace embedding requires $\operatorname{poly}(d)$ distortion (Wang and Woodruff, 2019) or has an exponential dependence on $1/\varepsilon$ (Li et al., 2021). We further show that if the leverage scores of $[A\ b]$ are uniformly small, namely, at most $\operatorname{poly}(\varepsilon)/d^{4/p}$, then our runtime can be improved to $O(\sum_i \operatorname{nnz}(A^i) + s\operatorname{poly}(d/\varepsilon))$, which is now optimal in terms of $\operatorname{nnz}(A)$, with the same amount of communication. Along the way we prove a result on embedding $d$-dimensional subspaces in $\ell_p^n$ to $\ell_r$ for $1 < r < p$, which may be of independent interest.

**Open Problems** We leave several intriguing questions for future work.

First, it would be good to close the gap in our upper and lower bounds as a function of $\varepsilon$ for $p < 2$. For $1 < p < 2$, if $\operatorname{poly}(1/\varepsilon) < d$ then our bounds are off by a $1/\varepsilon$ factor, namely, our upper bound is $\widetilde{O}(sd^2/\varepsilon)$, but our lower bound is $\Omega(sd^2)$.

Second, the $\operatorname{nnz}$ term in our runtime in general has a multiplicative factor of $d/\operatorname{poly}(\varepsilon)$. This is mainly due to the use of a dense matrix for the lopsided subspace embedding of $\ell_p^n$ into $\ell_r$, and it is interesting to see whether there are sparse lopsided subspace embeddings of $\ell_p^n$ into $\ell_r$.

### 1.1. Our Techniques

**Lower Bounds** We first demonstrate how to show an $\Omega(sd/\varepsilon^2)$ lower bound for $p \in (0, 1]$ and an $\Omega(sd/\varepsilon)$ lower bound for $p \in (1, 2]$.

Let us first consider the special case of $d = 1$. Consider the $\ell_p$ regression problem $\min_{x\in\mathbb{R}} \|a \cdot x - b\|_p$, where $a$ and $b$ are uniformly drawn from $\{-1, 1\}^n$. The crucial observation is that the solution $x$ reveals the Hamming distance $\Delta(a, b)$. Specifically, when $n = \Theta(1/\varepsilon^2)$, a $(1 \pm \varepsilon)$-solution when $0 < p \leq 1$ and $(1 \pm \varepsilon^2)$-solution when $1 < p \leq 2$ suffice for us to solve the Gap-Hamming communication problem (GHD) of $a$ and $b$ (determining $\Delta(a, b) \geq c\sqrt{n}$ or $\Delta(a, b) \leq -c\sqrt{n}$). The GHD problem has an $\Omega(n)$ information cost lower bound (Braverman et al., 2016), which implies, by our choice of $n$, an $\Omega(1/\varepsilon^2)$ lower bound for $p \in (0, 1]$ and an $\Omega(1/\varepsilon)$ lower bound for $p \in (1, 2]$.

To gain the factor of $s$, we design a distributed version of GHD, the $s$-GAP problem, as follows. There are $2s$ players. Each of the first $s$ players holds a vector $a^i \in \{-1, 1\}^n$ and each of the remaining players holds a $b^i \in \{-1, 1\}^n$, with the guarantee that $\sum_i a^i = a$ and $\sum_i b^i = b$. The $2s$ players and the coordinator will collectively determine the two cases of $\Delta(a, b)$. Our goal is to show an $\Omega(sn)$ lower bound for this communication problem. To this end, we employ the symmetrization technique that was used in (Phillips et al., 2016). Specifically, Alice simulates a random player and Bob the remaining $s - 1$ players. As such, Bob will immediately know the whole vector $b$ and part of the vector $a$ (denote the set of these indices by $I$). As we will show in the proof, to determine the distance $\Delta(a, b)$, Alice and Bob still need to approximately determine $\Delta(a_{I^c}, b_{I^c})$, which requires $\Omega(|I^c|) = \Omega(n)$ communication. Note that the input distribution of each player is the same and Alice is choosing a random player. Hence, Alice's expected communication to Bob is at most $O(\chi/s)$ bits if $s$-GAP can be solved using $\chi$ bits of communication, which yields a lower bound of $\Omega(sn)$ bits for the $s$-GAP problem.

So far we have finished the proof for $d = 1$. To obtain a lower bound for general $d$, we use a padding trick. Consider $A = \mathrm{diag}(a_1, \ldots, a_d)$ and let $b$ be the vertical concatenation of $b_1, \ldots, b_d$, where each pair $(a_i, b_i)$ is drawn independently from the hard distribution for $d = 1$. One can immediately observe that $\min_x \|Ax - b\|_p^p = \sum_i \min_{x_i} \|a_i x_i - b\|_p^p$ and show that approximately solving $\min_x \|Ax-b\|_p^p$ can approximately solve a constant fraction of the $d$ subproblems $\min_{x_i} \|a_i x_i-b\|_p^p$. This further adds an $O(d)$ factor to the lower bound.

Next we discuss the $\Omega(sd^2)$ lower bound. We shall follow the idea of (Vempala et al., 2020) and construct a set of matrices $\mathcal{H} \subseteq \{-1, 1\}^{d\times d}$ with a vector $b \in \mathbb{R}^d$ such that (i) $A$ is non-singular for all $A \in \mathcal{H}$, (ii) $A^{-1}b \neq B^{-1}b$ for all $A, B \in \mathcal{H}$ and $A \neq B$ and (iii) $|\mathcal{H}| = 2^{\Omega(d^2)}$. The conditions (i) and (ii) mean that a constant-factor approximation to $\min_x \|Ax-b\|_p^p$ is exact, from which the index of $A$ in the set $\mathcal{H}$ can be inferred. Condition (iii) then implies an $\Omega(d^2)$ lower bound for solving the regression problem up to a constant factor. To gain a factor of $s$, we consider the communication game where the $i$-th player receives a matrix $A^i \subseteq \{-1, 1\}^{d\times d}$ with the guarantee that $A = \sum_i A^i$ is distributed in $\mathcal{H}$ uniformly. Then the $s$ players with the coordinator want to recover the index of $A$ in $\mathcal{H}$. We consider a similar symmetrization technique. However, the issue here is if Bob simulates $s - 1$ players, he will immediately know roughly a $\frac{1}{2}$ fraction of coordinates of $A$, which can help him to get the index of $A$ in $\mathcal{H}$. To overcome this, we choose a different strategy where Alice simulates two (randomly chosen) players and Bob simulates the remaining $s - 2$ players. In this case Bob can only know a $\frac{1}{4}$-fraction of the coordinates without communication. However, one new issue here is Bob will know partial information about the remaining coordinates. But, as we shall

show in the proof, even when conditioned on Bob's input on $s-2$ players, with high probability the entropy of the remaining coordinates is still $\Omega(d^2)$. This implies that Alice still needs to send $\Omega(d^2)$ bits to Bob, which yields an $\Omega(sd^2)$ lower bound for the original problem.

**Upper Bounds** For the $\ell_p$-regression $\min_x \|Ax-b\|_p$, a classical "sketch-and-solve" approach is to use a $(1+\varepsilon)$-subspace embedding $S$ for $B = [A\ b] \in \mathbb{R}^{n\times(d+1)}$ and reduce the problem to solving $\min_x \|SAx - Sb\|_p$, which is of much smaller size. The subspace embedding is non-oblivious and obtained by subsampling $\widetilde{O}(d/\varepsilon^2)$ rows of $B$ with respect to the Lewis weights of $B$ (Cohen and Peng, 2015). More recently, it was shown that sampling $\widetilde{O}(d/\varepsilon)$ rows according to the Lewis weights is sufficient for solving $\ell_p$-regression (Musco et al., 2022; Chen et al., 2022), instead of $\widetilde{O}(d/\varepsilon^2)$ rows needed for an $\ell_p$-subspace embedding. However, computing the Lewis weights is expensive and would incur a communication cost as well as a runtime at least linear in $n$, which is prohibitive in our setting.

Instead of embedding an $\ell_p$-subspace into $\ell_p$, we $(1+\varepsilon)$-embed an $\ell_p$-subspace into $\ell_r$ for some $1 < r < p$. Furthermore, since we are solving a regression problem, we do not need a conventional subspace embedding but only a *lopsided* one; that is, the map $S$ must not contract $\|Ax-b\|_p$ for all $x$ simultaneously but it is required not to dilate $\|Ax^*-b\|_p$ for only the optimal solution $x^*$. We show that an $S$ of i.i.d. $p$-stable variables and $O(d\log d/\operatorname{poly}(\varepsilon))$ rows suffices (see Lemma 18 for the formal statement). Such a lopsided subspace embedding for embedding a subspace of $\ell_p^n$ into $\ell_r$, to the best of our knowledge, has not appeared in the literature[1] and may be of independent interest. This lopsided subspace embedding reduces the $\ell_p$ regression problem to an $\ell_r$-regression problem of $\widetilde{O}(d/\operatorname{poly}(\varepsilon))$ rows. Importantly though, we do not need to ever explicitly communicate these rows in their entirety. Namely, we can leave the regression problem in an implicit form and now run a Lewis weight approximation algorithm, and since our effective $n$ has been replaced with $d/\operatorname{poly}(\varepsilon)$, we just need $d/\operatorname{poly}(\varepsilon)$ communication to iteratively update each of the weights in the Lewis weight algorithm, rather than $n$ communication.

For the $\ell_2$-regression problem, it is known that a $(1+\sqrt{\varepsilon})$-subspace embedding can yield a $(1+\varepsilon)$-approximate solution (see, (Bourgain and Nelson, 2013), also the [Woo14] reference therein) and so the subspace embedding $S$ needs only to have $O(d(\log d)/\varepsilon)$ rows. The servers then run gradient descent on the sketched version $\min_x \|SAx-Sb\|_2$. To ensure fast convergence in $O(\log(1/\varepsilon))$ iterations, the servers will instead solve $\min_x \|SARx-Sb\|_2$, where $R$ is a pre-conditioner to make $SAR$ have a constant condition number. Putting these pieces together leads to our near-optimal communication and runtime.

## 2. Preliminaries

**$\ell_2$ Subspace Embeddings.** For a matrix $A \in \mathbb{R}^{n\times d}$, we say a matrix $S \in \mathbb{R}^{m\times n}$ is a $(1\pm\varepsilon)$-$\ell_2$ subspace embedding for the column span of $A$ if $(1-\varepsilon)\|Ax\|_2 \le \|SAx\|_2 \le (1+\varepsilon)\|Ax\|_2$ for all $x \in \mathbb{R}^d$ with probability at least $1-\delta$. We summarize the subspace embeddings we use in this paper below:

- Count-Sketch: $m = O(d^2/(\delta\varepsilon^2))$ with $s=1$ non-zero entry per column, with each non-zero entry in $\{-1,1\}$ (Clarkson and Woodruff, 2017). Computing $SA$ takes only $O(\operatorname{nnz}(A))$ time.

1. We note that the works of (Pisier, 1983; Friedland and Guédon, 2011) consider embedding the entire space $\ell_p^n$ into $\ell_r$ instead of embedding a low-dimensional subspace of $\ell_p^n$ into $\ell_r$.

- OSNAP: $m = O((d\log(d/\delta))/\varepsilon^2)$ and has $s = O((\log(d/\delta))/\varepsilon)$ non-zeros per column, with each non-zero entry in $\{-1, 1\}$ (Nelson and Nguyên, 2013; Cohen, 2016). Computing $SA$ takes $O(s \cdot \text{nnz}(A)) = O(\text{nnz}(A)(\log(d/\delta)/\varepsilon))$ time.

**$p$-stable Distributions.** Our protocol for distributed $\ell_p$ regression will use $p$-stable distributions, which are defined below.

**Definition 1 (Zolotarev (1986))** *For $0 < p < 2$, there exists a probability distribution $\mathcal{D}_p$ called the $p$-stable distribution, which satisfies the following property. For any positive integer $n$ and vector $x \in \mathbb{R}_n$, if $Z_1, \ldots, , Z_n \sim \mathcal{D}_p$ are independent, then $\sum_{j=1}^n Z_j x_j \sim \|x\|_p Z$ for $Z \sim \mathcal{D}_p$.*

**Lewis Weights.** Below we recall some facts about Lewis weights. For more details, we refer the readers to, e.g., (Clarkson et al., 2019, Section 3.3).

**Definition 2** *Given a matrix $A \in \mathbb{R}^{n\times d}$. The* leverage score *of a row $A_{i,*}$ is defined to be $\tau_i(A) = A_{i,*}(A^TA)^\dagger(A_{i,*})^T$.*

**Definition 3 (Cohen and Peng (2015))** *For a matrix $A \in \mathbb{R}^{n\times d}$, its $\ell_p$-Lewis weights $\{w_i\}_{i=1}^n$ are the unique weights such that $w_i = \tau_i(W^{1/2-1/p}A)$ for each $i \in [n]$. Here $\tau_i$ is the leverage score of the $i$-th row of a matrix and $W$ is the diagonal matrix whose diagonal entries are $w_1, \ldots, w_n$.*

The Lewis weights are used in the construction of randomized $\ell_p$-subspace embeddings. In particular, the rescaled sampling matrix w.r.t. Lewis weights gives an $\ell_p$-subspace embedding.

**Definition 4** *Given $p_1, \ldots, p_n \in [0, 1]$ and $p \geq 1$, the* rescaled sampling matrix *$S$ with respect to $p_1, \ldots, p_n$ is a random matrix formed by deleting all zero rows from a random $n \times n$ diagonal matrix $D$ in which $D_{i,i} = p_i^{-1/p}$ with probability $p_i$ and $D_{i,i} = 0$ with probability $1 - p_i$.*

**Lemma 5 (Lewis weight sampling, Cohen and Peng (2015))** *Let $A \in \mathbb{R}^{n\times d}$ and $p \geq 1$. Choose an oversampling parameter $\beta = \Theta(\log(d/\delta)/\varepsilon^2)$ and sampling probabilities $p_1, \ldots, p_n$ such that $\min\{\beta w_i(A), 1\} \leq p_i \leq 1$ and let $S$ be the rescaled sampling matrix with respect to $p_1, \ldots, p_n$. Then it holds with probability at least $1 - \delta$ that $(1 - \varepsilon)\|Ax\|_p \leq \|SAx\|_p \leq (1 + \varepsilon)\|Ax\|_p$ (i.e., $S$ is an $\varepsilon$-subspace embedding for $A$ in the $\ell_p$-norm) and $S$ has $O(\beta \sum_i w_i(A)) = O(\beta d)$ rows.*

Cohen and Peng (2015) give an iterative algorithm (Algorithm 1) which computes the Lewis weights time-efficiently for $p < 4$.

**Lemma 6 (Cohen and Peng (2015))** *Suppose that $p < 4$ and $\beta = \Theta(1)$. After $T = \log\log(n)$ iterations in Algorithm 1, $w$ is a constant approximation to the $\ell_p$ Lewis weights.*

## 3. Distributed $\ell_p$-Regression Lower Bound

We consider the following variant of the Gap-Hamming problem (GHD).

1. Initialize $w = \mathbf{1} \in \mathbb{R}^n$.
2. For $t = 1, 2, \ldots, T$
   (a) Let $\tau \in \mathbb{R}^n$ be a $\beta$-approximation of the leverage scores of $W^{1/2-1/p}A$.
   (b) Set $w_i \leftarrow (w_i^{2/p-1}\tau_i)^{p/2}$.
3. Return $w$.

**Algorithm 1:** Iterative Algorithm to Compute the $\ell_p$ Lewis Weights

**Gap-Hamming Problem.** In the Gap-Hamming problem ($\mathsf{GHD}_{n,c}$), Alice and Bob receive binary strings $x$ and $y$, respectively, which are uniformly sampled from $\{-1,1\}^n$. They wish to decide which of the following two cases $\Delta(x,y) = \sum_{i=1}^n x_i y_i$ falls in: $\Delta(x,y) \geq c\sqrt{n}$ or $\Delta(x,y) \leq -c\sqrt{n}$, where $c$ is a constant. (If $\Delta(x,y)$ is between $-c\sqrt{n}$ and $c\sqrt{n}$, an arbitrary output is allowed.)

**Lemma 7 (Braverman et al. (2016))** *If there is a protocol $\Pi$ which solves $\mathsf{GHD}_{n,c}$ with large constant probability, then we have $I(x,y;\Pi) = \Omega(n)$, where $I$ denotes mutual information and the constant hidden in the $\Omega$-notation depends on $c$.*

### 3.1. $s$-GAP problem

In this section, we will define the $s$-GAP problem and then prove an $\Omega(sn)$ lower bound.

**Definition 8** *In the $s$-GAP problem, there are $2s$ players, where for the first $s$ players, the $i$-th player receives an $n$-bit string $a^i \in \{-1,1\}^n$, and for the remaining $s$ players, the $i$-th player receives an $n$-bits string $b^i \in \{-1,1\}^n$, with the guarantee that $a = \sum_i a^i \in \{-1,1\}^n$, $b = \sum_i b^i \in \{-1,1\}^n$ and $\Delta(a,b) \in [-c_2\sqrt{n}, c_2\sqrt{n}]$. The $2s$ players want to determine if $\Delta(a,b) \geq c_1\sqrt{n}$ or $\Delta(a,b) \leq -c_1\sqrt{n}$. Here $c_1 < c_2$ are both constants. (Similarly, if $\Delta(a,b)$ is between $-c_1\sqrt{n}$ and $c_1\sqrt{n}$, an arbitrary output is allowed).*

To prove the $\Omega(sn)$ lower bound, we use a similar symmetrization augment as in (Phillips et al., 2016) and reduce to the $\mathsf{GHD}$ problem. For the reduction, we consider $s = 4t + 2$ for simplicity, and without loss of generality by padding, and consider the following distribution $\mu$ for the inputs $a_i^j$ for players $j = 1, 2, \ldots, 2t+1$. Choose a uniformly random vector $a \in \{-1,1\}^n$. For each $i$, if $a_i = 1$, we place $(t+1)$ bits of 1 and $t$ bits of $-1$ randomly among the $2t+1$ players in this coordinate; if $a_i = -1$, we place $t$ bits of 1 and $(t+1)$ bits of $-1$ randomly among the $2t+1$ players. We remark that under this distribution, each player's inputs are drawn from the same distribution, and each coordinate of each player is 1 with probability $1/2$ and $-1$ with probability $1/2$. The distribution of $b_i^j$ is the same as that of $a_i^j$ for players $j = 2t+2, \ldots, 4t+2$.

**Theorem 9** *Any protocol that solves the $s$-GAP problem with large constant probability requires $\Omega(sn)$ bits of communication.*

**Proof** We reduce the $s$-GAP problem to the $\mathsf{GHD}$ problem using a similar symmetrization argument to that in (Phillips et al., 2016). Alice picks a random number $i \in [2t+1]$ uniformly and simulates the $i$-th player. Bob simulates the remaining $s-1$ players. We shall show that if there is an $s$-player protocol solving the $s$-GAP problem, then the coordinator will be able to solve the $\mathsf{GHD}$ problem on a constant fraction of the input vectors $a$ and $b$, which requires $\Omega(n)$ bits of communication. Note

that the input distribution of each player is the same and Alice is choosing a random player. Hence, Alice's expected communication to Bob is at most $O(\chi/s)$ bits if the $s$-GAP problem can be solved using $\chi$ bits of communication, which yields a lower bound of $\Omega(sn)$ bits for the $s$-GAP problem.

We first consider Bob's information when he simulates $s-1$ players. He knows each coordinate of $b$ directly. Consider a coordinate of $a$. If the sum of Bob's $s-1$ bits on this coordinate is 2 or $-2$, then he knows Alice's bit on this coordinate immediately, as their sum should be 1 or $-1$; while if Bob's sum is 0, he has zero information about Alice's bit on this coordinate. By a simple computation, we obtain that Bob's sum is 2 or $-2$ with probability $\frac{t}{2t+1}$ and is 0 with probability $\frac{t+1}{2t+1}$. From a Chernoff bound, we see that with probability at least $1-e^{-\Omega(n)}$, Bob learns at most $\frac{3}{5}n$ coordinates of $a$. Let $I$ denote the set of remaining indices. Then $|I| \geq \frac{2n}{5}$. We will show that Alice and Bob can solve GHD on $a_I$ and $b_I$ by simulating the protocol for the $s$-GAP problem.

Consider $\Delta(a_J, b_J)$ for $J = [n] \setminus I$. With probability at least $99/100$, it will be contained in $[-c_1\sqrt{|J|}, c_1\sqrt{|J|}]$, where $c_1$ is a sufficiently large absolute constant. Conditioned on this event, we have that whether the distance $\Delta(a_I, b_I) \geq c_2\sqrt{|I|}$ or $\Delta(a_I, b_I) \leq -c_2\sqrt{|I|}$ will decide whether $\Delta(a,b) \geq c_3\sqrt{n}$ or $\Delta(a,b) \leq -c_3\sqrt{n}$, where $c_2, c_3 > 0$ are appropriate constants (recall that we have $|I| \geq \frac{2}{5}n$ and $|J| \leq \frac{3}{5}n$). This means that, by simulating a $2s$-player protocol for the $s$-GAP problem, Alice and Bob can solve the $\mathsf{GHD}_{|I|,c_2}$ problem on $a_I$ and $b_I$, which requires $\Omega(|I|) = \Omega(n)$ bits of communication. ■

**Corollary 10** *Any protocol that solves $m$ independent copies of the $s$-GAP problem with high constant probability requires $\Omega(snm)$ bits of communication.*

**Proof** Similar to the proof of Theorem 9, Alice and Bob in this case need to solve $m$ independent copies of GHD. The direct sum theorem (Chakrabarti et al., 2001; Bar-Yossef et al., 2004) states that if the information cost of solving a communication problem with probability $2/3$ is $f$, then the information cost of solving $m$ independent copies of the same communication problem simultaneously with probability at least $2/3$ is $\Omega(mf)$. Since the information cost implies a communication lower bound, it follows from Lemma 7 and the direct sum theorem that $\Omega(knm)$ bits of communication are required. ■

### 3.2. $\Omega(sd/\varepsilon^2)$ and $\Omega(sd/\varepsilon)$ Lower Bounds

In this section, we will show an $\Omega(sd/\varepsilon^2)$ lower bound for the $\ell_p$-regression problem when $0 < p \leq 1$ and an $\Omega(sd/\varepsilon)$ lower bound when $1 < p \leq 2$.

For simplicity, we first consider the case of $d = 1$ and will later extend the result to general $d$. Consider the same input distribution as in Definition 8 with $n = 1/\varepsilon^2$, and for which the $2s$ players want to compute a $(1+\varepsilon)$-approximate solution to the $\ell_p$ regression problem

$$\arg\min_{x \in \mathbb{R}} \|ax - b\|_p^p . \tag{1}$$

In the lemma below, we shall show that using a $(1+\varepsilon)$-approximate solution for the $\ell_p$-regression problem (1), the players can distinguish the two cases to the $s$-GAP problem for the vectors $a$ and $b$, which implies an $\Omega(s/\varepsilon^2)$ lower bound. The proof, analogous to that of (Musco et al., 2022, Theorem 12.2), analyzes an objective of the form $r|1-x|^p + (n-r)|1+x|^p$ for $r = (n + \Delta(a,b))/2$ and is postponed to Appendix A.

**Lemma 11** *Suppose that $p \in (0,2]$, $n = \Theta(1/\varepsilon^2)$, and $a$ and $b$ are the vectors drawn from the distribution in Definition 8. Let $\eta = \varepsilon$ when $p \in (0,1]$ and $\eta = \varepsilon^2$ when $p \in (1,2]$. Then, any $\widetilde{x}$ such that $\|a\widetilde{x} - b\|_p^p \leq (1+\eta)\min_{x\in\mathbb{R}} \|ax-b\|_p^p$ can be used to distinguish whether $\Delta(a,b) \geq c\sqrt{n}$ or $\Delta(a,b) \leq -c\sqrt{n}$, where $c$ is an absolute constant.*

Combining this lemma with Theorem 9 yields the desired lower bound for the distributional regression problem with $d = 1$.

**Lemma 12** *Suppose that $d = 1$ and $\varepsilon > 0$. Then any protocol that computes a $(1+\varepsilon)$-approximate solution to the $s$-server distributional $\ell_p$-regression problem in the message passing model with high constant probability requires $\Omega(s/\varepsilon^2)$ bits of communication for $p \in (0,1]$ and $\Omega(s/\varepsilon)$ bits of communication for $p \in (1,2]$.*

We now extend the lower bound to general $d$ via a padding argument. Suppose that $a_1, a_2, \ldots, a_d$ and $b_1, b_2, \ldots, b_d$ are $d$ independent samples drawn from the same distribution as defined in Definition 8 with $n = \Theta(1/\varepsilon^2)$. We form a matrix $A \in \mathbb{R}^{O(d/\varepsilon^2)\times d}$ and a vector $b \in \mathbb{R}^{O(d/\varepsilon^2)}$ as

$$A = \begin{bmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_d \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{bmatrix}.$$

It then follows that

$$\min_{x\in\mathbb{R}^d} \|Ax - b\|_p^p = \sum_{i=1}^{d} \min_{x_i\in\mathbb{R}} \|a_i x_i - b_i\|_p^p.$$

We then make the following observation. If $x \in \mathbb{R}^d$ is a $(1+\varepsilon)$-approximate solution of $\min_x \|Ax - b\|_p^p$, then there must exist a constant fraction of the indices $i \in [d]$ such that $x_i$ is a $(1 + O(\varepsilon))$-approximate solution to the regression problem $\min_{x_i\in\mathbb{R}} \|a_i x_i - b_i\|_p^p$ (recall that we have the guarantee that $\Delta(a_i, b_i) \in [-c_2\sqrt{n}, c_2\sqrt{n}]$ for all $i$, and hence the objective values for each regression problem are within a constant factor). This means that from the signs of these $x_i$, we can solve a constant fraction of the $d$ independent copies of the $s$-GAP problem, which implies the following theorem immediately.

**Theorem 13** *Suppose that $\varepsilon > \frac{1}{\sqrt{n}}$ for $p \in (0,1]$ and $\varepsilon > \frac{1}{n}$ for $p \in (1,2]$. Then any protocol that computes a $(1+\varepsilon)$-approximate solution to the $s$-server distributional $\ell_p$-regression problem with $d$ columns in the message passing model with large constant probability requires $\Omega(sd/\varepsilon^2)$ bits of communication for $p \in (0,1]$ and $\Omega(sd/\varepsilon)$ bits of communication for $p \in (1,2]$.*

### 3.3. $\Omega(sd^2)$ Lower Bound for $p \in (0,2]$

In this section, we present an $\Omega(sd^2)$ lower bound for $0 < p \leq 2$. We first describe the intuition behind our lower bound. Following (Vempala et al., 2020), we construct a set of matrices $\mathcal{H} \subseteq \mathbb{R}^{d\times d}$ with a vector $b \in \mathbb{R}^d$ such that (i) $T$ is non-singular for all $T \in \mathcal{H}$, and (ii) $S^{-1}b \neq T^{-1}b$ for all $S, T \in \mathcal{H}$ and $S \neq T$. Then we uniformly sample a matrix $A \in \mathcal{H}$ and show that we can obtain the index of $A$ in the set $\mathcal{H}$ from a constant-factor approximate solution to the regression problem $\min \|Ax - b\|_p^p$. This will imply an $\Omega(d^2)$ lower bound even for $s = 2$. The construction of $\mathcal{H}$ is given in the following lemma.

**Lemma 14** *For every sufficiently large $d$, there exists a set of matrices $\mathcal{H} \subseteq \{-1,1\}^{d\times d}$ with $|\mathcal{H}| = \Omega(2^{0.49d^2})$ such that (i) $T$ is non-singular for all $T \in \mathcal{H}$, and (ii) for all distinct $S, T \in \mathcal{H}$, $S^{-1}e_d \neq T^{-1}e_d$, where $e_d$ is the $d$-th standard basis vector.*

We remark that in (Vempala et al., 2020), Lemma 14 was only shown for the case where $t > 1$, $|\mathcal{H}| = \Omega(t^{1/6d^2})$ and the matrix entries are integers in $[-t, t]$. However, using the singularity probability of random matrices in $\{-1,+1\}^{d\times d}$ and following a similar argument to (Vempala et al., 2020), we can obtain the desired bounds in Lemma 14. The detailed proof can be found in Appendix B. Note that the construction procedure of the set is close to random sampling – uniformly sample $\Omega(2^{0.49d^2})$ matrices and remove a small fraction. This property will be crucial to our proof.

To achieve an $\Omega(sd^2)$ lower bound for $s$ players, we consider the same input distribution for the $s$ players in Lemma 7 and employ a similar symmetrization argument. After sampling matrices in $\mathcal{H}$, we construct the inputs of the $s$ players to be matrices in $\{-1,+1\}^{d\times d}$ with the sum being $A$. However, if we follow the same argument and let Bob simulate $s-1 = 2t$ players, in expectation he will know a $\frac{t}{2t+1} \approx \frac{1}{2}$ fraction of the entries of $A$, and from the construction of the set $|\mathcal{H}|$ we know that there will be only $O(1)$ matrices in $\mathcal{H}$ satisfying the conditions on such entries. Hence, Alice only needs to send $O(1)$ bits of information to Bob. To solve this issue, we make the following modification. Instead, we let Alice simulate 2 players, and Bob simulates the remaining $s-2 = 2t-1$ players. In this case, Bob will know roughly a $1/4$-fraction of the entries directly; however, for the remaining entries, he will know side information. Roughly speaking, for $A_{ij}$, if Bob's sum over the $s-2$ players is 1, with probability roughly $2/3$, $A_{ij}$ is 1; if his sum over the $k-2$ players is $-1$, with probability roughly $2/3$, $A_{ij}$ is $-1$. We shall show that even having such side information, with high probability the conditional entropy of the remaining entries of $A$ is still $\Omega(d^2)$, which implies that Alice still needs to send Bob $\Omega(d^2)$ bits.

**Lemma 15** *Consider the following game of $s = 2t+1$ players, where the $i$-th player receives a $d \times d$-matrix $A^i$ such that $A^i \subseteq \{-1,1\}^{d\times d}$ with the guarantee that $A = \sum_i A^i$ is distributed in $\mathcal{H}$ uniformly. The $s$ players want to determine collectively the index of the matrix $A$ in $\mathcal{H}$. Any protocol which solves this problem with large constant probability requires $\Omega(sd^2)$ bits of communication.*

**Proof** We first describe the input distribution of each player. Suppose that matrix $A$ has been sampled from $\mathcal{H}$. For each coordinate $(i,j)$, if $A_{ij} = 1$, we place $(t+1)$ bits of 1 and $t$ bits of $-1$ randomly among the $2t+1$ players' inputs for coordinate $j$; if $A_{ij} = -1$, we place $t$ bits of 1 and $t+1$ bits of $-1$. Similarly, under this distribution, each player's inputs are drawn from the same distribution.

We then use symmetry and let Alice simulate two random players, and Bob simulates the remaining $s-2 = 2t-1$ players. Consider first Bob's information when he simulates $2t-1$ players. Via a simple computation we can get that for each coordinate, with probability $\frac{t-1}{4t+2}$ Bob's sum will be 3 or $-3$, in which case he will know $A_{ij}$ immediately. If Bob's sum is 1, he will get that $A_{ij} = 1$ with probability $\frac{2}{3}$ and $A_{ij} = -1$ with probability $\frac{1}{3}$; if Bob's sum is $-1$, he will get that $A_{ij} = -1$ with probability $\frac{2}{3}$ and $A_{ij} = 1$ with probability $\frac{1}{3}$. It follows from a Chernoff bound that with probability $1 - \exp(-d^2)$, Bob obtains the exact information of at most $0.26d^2$ coordinates and has partial information about the remaining coordinates. For the remainder of the proof we assume this event happens.

Let $\mathcal{S}$ denote the subset of $\mathcal{H}$ which agrees on the above $0.26d^2$ coordinates. From the construction of $\mathcal{H}$ we get that with at least constant probability $|\mathcal{S}| = \Omega(2^{0.2d^2})$. Condition on this

event. For simplicity, next we only consider the matrix in $\mathcal{S}$ and treat it as an $\ell$-dimensional vector after removing the known $0.26d^2$ coordinates, where $\ell = 0.74d^2$. Let $Y$ denote Bob's sum vector. We shall show that the conditional entropy $H(A \mid Y)$ remains $\Omega(d^2)$, and hence by a standard information-theoretic argument, Alice must still send $\Omega(d^2)$ bits to Bob to identify the index of the matrix in $\mathcal{S}$. From this, we get an $\Omega(sd^2)$ lower bound on the protocol for the original problem.

By a Chernoff bound, with probability $1 - \exp(-d^2)$, the Hamming distance between $A$ and $Y$ is within $\frac{1}{3}\ell \pm 0.01d^2$. We condition on this in the remainder of the proof. We now turn to bound the number of matrices in $S$ which have a Hamming distance of $\frac{1}{3}\ell$ from $Y$. For each matrix $B$, from the construction of $\mathcal{H}$ we know that each coordinate of $B$ is the same as the corresponding coordinate of $A$ with probability $1/2$. Hence, the probability that $B$ has Hamming distance $\frac{2}{3}\ell$ from $A$ is (using Stirling's formula)

$$\binom{\ell}{\frac{2}{3}\ell} \cdot 2^{-\ell} \simeq \frac{1}{\ell} \cdot \frac{3^\ell}{2^{\frac{2}{3}\ell}} \cdot 2^{-\ell} = \frac{3^\ell}{\ell\, 2^{\frac{5}{3}\ell}}.$$

Hence, the expected number of such $B$ is

$$|\mathcal{S}| \cdot \frac{3^\ell}{\ell 2^{\frac{5}{3}\ell}} > 2^{0.2d^2} \cdot \frac{3^\ell}{\ell 2^{\frac{5}{3}\ell}} \geq (1.101)^{d^2} \ .$$

From a Chernoff bound we know that with probability at least $1 - \exp(-d^2)$, the number of $B \in \mathcal{S}$ for which $B$ has a Hamming distance $\frac{1}{3}\ell$ from $Y$ is at least $(1.10)^{d^2}$.

We next turn to show that when conditioned on the event above, it is enough to show that the conditional entropy $H(A \mid Y)$ satisfies $H(A \mid Y) = \Omega(d^2)$ given Bob's vector $Y$. Let $\mathcal{T}$ be the subset of $\mathcal{H}$ which agrees on the above $0.26d^2$ coordinates and having Hamming distance within $\frac{1}{3}\ell \pm 0.01d^2$. For each matrix $T \in \mathcal{T}$, define a weight of the matrix $T$ to be $w_T = \left(\frac{2}{3}\right)^{\ell-u}\left(\frac{1}{3}\right)^u = \left(\frac{1}{3}\right)^\ell 2^{l-u}$, where $u$ is the Hamming distance between $T$ and $Y$. It follows from Bayes' Theorem that $T$ is the correct matrix with probability

$$p_T = \frac{w_T}{\sum_{i \in \mathcal{T}} w_i} \ .$$

For the denominator, we have from the conditioned events that

$$S = \sum_{i \in \mathcal{T}} w_i \geq (1.10)^{d^2} \cdot \left(\frac{1}{3}\right)^\ell 2^{\frac{2}{3}\ell - 0.01d^2} \geq (0.682)^{d^2} \ .$$

For the numerator, note that it holds for every $i \in \mathcal{T}$ that

$$w_i \leq \left(\frac{1}{3}\right)^\ell 2^{\frac{2}{3}\ell + 0.01d^2} \leq (0.629)^{d^2}.$$

It follows from the definition of the entropy that

$$H(A \mid Y) = \sum_{i \in \mathcal{T}} p_i \log \frac{1}{p_i} = \sum_{i \in \mathcal{T}} \frac{w_i}{S} \log \frac{S}{w_i} \geq \sum_{i \in \mathcal{T}} \frac{w_i}{S} \log \frac{S}{(0.629)^{d^2}} = \log \frac{S}{(0.629)^{d^2}} = \Omega\left(d^2\right) \ ,$$

which is exactly we need. The proof is complete. ■

The following theorem follows immediately from the preceding lemma.

**Theorem 16** *Suppose that $0 < p \leq 2$. Any protocol that computes a constant-factor approximate solution to the $s$-server distributional $\ell_p$-regression problem with $d$ columns in the message passing model with large constant probability requires $\Omega(sd^2)$ bits of communication.*

## 4. $\ell_2$-Regression Upper Bound

In this section, we give an $\widetilde{O}(sd^2 + sd/\varepsilon)$ communication protocol for the distributed $\ell_2$-regression problem. We first describe the high-level intuition of our protocol, which is based on the sketching algorithm in (Clarkson and Woodruff, 2009) and the sketching-based pre-conditioning algorithm in (Clarkson and Woodruff, 2017).

- Let $S_1 \in \mathbb{R}^{O(d\log(d)/\varepsilon)\times n}$ be a $(1 \pm \sqrt{\varepsilon})$-subspace embedding. We compute $\hat{A} = SA$ and $\hat{b} = Sb$ and then the problem is reduced to solving $\min_{x\in\mathbb{R}^d} \|\hat{A}x - \hat{b}\|_2^2$.
- Let $S_2 \in \mathbb{R}^{O(d\log d)\times O(d\log(d)/\varepsilon)}$ be a $(1 \pm 1/2)$ subspace embedding of $SA$. We compute a QR-decomposition of $S\hat{A} = QR^{-1}$. Then the regression problem is equivalent to solving $\min_{x\in\mathbb{R}^d} \|\hat{A}Rx - \hat{b}\|_2^2$.
- Run a gradient descent algorithm for $T = O(\log(1/\varepsilon))$ iterations. In the $t$-th iteration, compute the gradient of the objective function at the current solution $x_t$ and perform the update $x_{t+1} = x_t - (\hat{A}R)^T(\hat{A}Rx_t - \hat{b})$.
- Output $Rx_T$ as the solution.

The protocol is presented in Algorithm 2. Initially, each server computes $\hat{A}^i = \Pi_2\Pi_1 A^i$, then computes $\Pi_3\hat{A}^i$ and sends it to the coordinator. Note that $\Pi_1$ is a Count-Sketch matrix and hence we can compute $\Pi_1 A^i$ in $\mathrm{nnz}(A^i)$ time and then compute $\Pi_2\Pi_1 A^i$ in $\mathrm{nnz}(A^i) + \mathrm{poly}(d/\varepsilon)$ time. The coordinator then computes a QR-decomposition of $\Pi_3\hat{A} = \sum_i \Pi_3\hat{A}^i$. The point is that $\hat{A}R$ will be well-conditioned, which will greatly improve the convergence rate of gradient descent. Then each server will help compute the gradient at the current solution $x_t$ and the coordinator will perform the corresponding update. The following is our theorem. We defer the proof to Appendix C.

**Theorem 17** *The protocol in Algorithm 2 returns a $(1\pm\varepsilon)$-approximate solution to the $\ell_2$-regression problem with large constant probability, and the communication complexity is $\widetilde{O}(sd^2+sd/\varepsilon)$. Moreover, the total runtime of all servers of the protocol is $O(\sum_i \mathrm{nnz}(A^i) + s\cdot\mathrm{poly}(d/\varepsilon))$.*

## 5. $\ell_p$-Regression Upper Bound

In this section, we give an $\widetilde{O}(sd^2/\varepsilon + sd/\varepsilon^{O(1)})$ communication protocol for the distributed $\ell_p$-regression problem when $1 < p < 2$. We first describe the high-level intuition of our protocol.

- Let $T \in \mathbb{R}^{O(d(\log d)/\varepsilon^{O(1)})\times n}$ be a sketch matrix whose entries are scaled i.i.d. $p$-stable random variables. We compute $\hat{A} = TA$ and $\hat{b} = Tb$ and then the problem is reduced to solving $\min_{x\in\mathbb{R}^d} \|\hat{A}x - \hat{b}\|_r$.
- Run Algorithm 1 to obtain a constant approximation of the $\ell_r$ Lewis weights $w$ of $[\hat{A}\ \hat{b}]$.
- Sample $O(d/\varepsilon)$ rows of $\hat{A}$ and $\hat{b}$ proportional to $w$, and form the new matrix $A'$ and $b'$.
- Solve $x = \arg\min_{x\in\mathbb{R}^d} \|A'x - b'\|_r$ and output $x$.

The protocol is shown in Algorithm 3. To show its correctness, we first analyze $\ell_p$-to-$\ell_r$ embeddings and the algorithm for solving the $\ell_p$-regression problem using Lewis weight sampling.

1. Each Server $P_i$ initializes the same Count-Sketch matrix $\Pi_1 \in \mathbb{R}^{O(d^2/\varepsilon)\times n}$ and OSNAP matrices $\Pi_2 \in \mathbb{R}^{O(d(\log d)/\varepsilon)\times O(d^2/\varepsilon)}$, and $\Pi_3 \in \mathbb{R}^{O(d\log d)\times O(d(\log d)/\varepsilon)}$.
2. Each Server $P_i$ computes $\hat{A}^i = \Pi_2\Pi_1 A^i$ and $\hat{b}^i = \Pi_2\Pi_1 \hat{b}^i$.
3. Each Server $P_i$ computes $\Pi_3\hat{A}^i$ and sends it to the coordinator.
4. The coordinator computes a $QR$ decomposition of $\Pi_3\hat{A} = \sum_i \Pi_3\hat{A}^i = QR^{-1}$ and sends $\widetilde{R}$ to each server $P_i$, where $\widetilde{R}$ satisfies that (i) every entry of $\widetilde{R}$ is an integer multiple of $1/\operatorname{poly}(nd)$, (ii) every entry of $\widetilde{R} - R$ is in $[-1/\operatorname{poly}(nd), 1/\operatorname{poly}(nd)]$ and (iii) $\widetilde{R}$ is invertible.
5. Each server initializes $x_1 = \mathbf{0}^d$. For $t = 1, 2, \ldots, T = O(\log(1/\varepsilon))$
   (a) Each server computes $\hat{A}^i\widetilde{R}x_t - \hat{b}^i$ and sends it to the coordinator.
   (b) The coordinator computes $y_t = \hat{A}\widetilde{R}x_t - \hat{b} = \sum_i(\hat{A}^i\widetilde{R}x_t - \hat{b}^i)$ and sends it to each server.
   (c) Each server computes $(\hat{A}^i)^T y$, and sends it to the coordinator. The coordinator computes $g_t = B^T y = \sum_i(\hat{A}^i)^T y$, and makes the update $x_{t+1} = x_t - g_t$, then sends $x_{t+1}$ to each server.
   (d) The coordinator computes $\widetilde{R}x_T$ as the solution.

**Algorithm 2:** Protocol for $\ell_2$ regression in the message passing model

**$p$-stable distribution.** The best known $(1 \pm \varepsilon)$ $\ell_p$ subspace embeddings require an exponential number of rows for a $p$-stable sketch. However, as we will show in the following lemma, for $1 < r < p$, $\widetilde{O}(d/\varepsilon^{O(1)})$ rows are enough to give a $(1 \pm \varepsilon)$ (lopsided) embedding from $\ell_p$ to $\ell_r$, which is sufficient for the regression problem. The proof of the lemma is postponed to Appendix D.

**Lemma 18** *Suppose that $p > r > 1$ are constants, and $T \in \mathbb{R}^{m\times n}$ is a matrix whose entries are i.i.d. $p$-stable random variables scaled by $1/(m^{1/r}\cdot\alpha_{p,r})$, where $\alpha_{p,r}$ is a constant depending on $p$ and $r$ only. For $m = d\log d/\varepsilon^{C(\varepsilon,r)}$, where $C(\varepsilon, r)$ is a constant depending on $p$ and $r$ only, it holds for any given matrix $A \in \mathbb{R}^{n\times d}$ that*

1. *(dilation) for each $x \in \mathbb{R}^d$, $\|TAx\|_r \le (1+\varepsilon)\|Ax\|_p$ with large constant probability.*
2. *(contraction) $\|TAx\|_r \ge (1-\varepsilon)\|Ax\|_p$ for all $x \in \mathbb{R}^d$ simultaneously with high probability.*

*Furthermore, the entries of $T$ can be rounded to the nearest integer multiples of $1/\operatorname{poly}(nd)$ and the same guarantees still hold.*

**Lewis Weight Sampling.** It is known that sampling $\widetilde{O}(d/\varepsilon^2)$ rows with respect to the $\ell_p$ Lewis weights gives an $\ell_p$ subspace embedding with large constant probability when $p \in [1, 2]$ (Cohen and Peng, 2015). In the following lemma, we shall show that for $\ell_p$-regression, sampling $\widetilde{O}(d/\varepsilon)$ rows is enough.

**Lemma 19** *Let $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$ and $p \in (1, 2)$. Suppose that $S$ is a rescaled sampling matrix according to $w_i([A\ b])$ with oversampling factor $\beta = \Theta(\varepsilon^{-1}\log^2 d\log n\log(1/\delta))$ and $\widetilde{x} = \arg\min_{x\in\mathbb{R}^d}\|SAx - Sb\|_p$. With probability at least $1-\delta$, it holds that $\|A\widetilde{x} - z\|_p \le (1+\varepsilon)\min_{x\in\mathbb{R}^d}\|Ax - z\|_p$ and the number of rows that $S$ samples is $O\left(\varepsilon^{-1}d\log^2 d\log n\log(1/\delta)\right)$.*

The proof of the lemma closely follows the proof in (Chen et al., 2022) and is postponed to Appendix E. We are now ready to prove our theorem for distributed $\ell_p$-regression. Its proof is postponed to Appendix F.

1. Each server initializes the same $p$-stable variable matrix $T \in \mathbb{R}^{\log(d)/\varepsilon^{O(1)} \times n}$, OSNAP matrix $S \in \mathbb{R}^{d\log(d) \times d\log(d)/\varepsilon^{O(1)}}$ and Gaussian matrix $G \in \mathbb{R}^{(d+1)\times \log(d/\varepsilon)}$. The entries of $T$ and $G$ are rounded to the nearest integer multiples of $1/\operatorname{poly}(nd)$.
2. Each server $P_i$ computes $\hat{A}^i = TA^i$ and $\hat{b}^i = Tb^i$, and forms $B^i = [\hat{A}^i\ \hat{b}^i]$.
3. Each server initializes $w = \mathbf{1}^{d/\varepsilon^{O(1)}}$. For $j = 1, 2, \ldots, t = O(\log\log(d/\varepsilon))$
   (a) Each server $P_i$ computes $S_t\widetilde{W}^{1/2-1/r}B^i$ (where $W = \operatorname{diag}(w)$ and $\widetilde{W}^{1/2-1/r}$ is a rounded version of $W^{1/2-1/r}$) and then sends it to the coordinator.
   (b) The coordinator computes the QR-decomposition of $S\widetilde{W}^{1/2-1/r}B = QR^{-1}$. It then sends $\widetilde{R}$ to each server, where $\widetilde{R}$ satisfies that (i) every entry of $\widetilde{R}$ is an integer multiple of $1/\operatorname{poly}(nd)$, (ii) every entry of $\widetilde{R} - R$ is in $[-1/\operatorname{poly}(nd), 1/\operatorname{poly}(nd)]$ and (iii) $\widetilde{R}$ is invertible.
   (c) Each server computes $B^i\widetilde{R}G$ and sends it to the coordinator.
   (d) The coordinator computes the square of the $\ell_2$ norm of the rows in $B\widetilde{R}G$ as a vector $\tau \in \mathbb{R}^{d/\varepsilon^{O(1)}}$.
   (e) The coordinator performs $w_i \leftarrow (w^{2/r-1}\tau_i)^{r/2}$ and sends the new $w$ to all servers, after rounding each coordinate of $w$ to the nearest integer multiple of $1/\operatorname{poly}(nd)$.
4. The coordinator samples the $i$-th row of $\hat{A}$ and $\hat{b}$ with probability $q_i \geq \beta \cdot w_i \cdot \log^3(d/\varepsilon)/\varepsilon$, where $\beta$ is a sufficiently large constant. Suppose that $\mathcal{S}$ is the set of indices of the sampled rows. Each server sends the rows in $\mathcal{S}$ to the coordinator.
5. The coordinator forms the matrix $A'$ and $b'$ using the rows in $\mathcal{S}$ and each sampled row with a re-scaling factor of $1/q_i^{1/r}$.
6. The coordinator solves $x = \arg\min_{x\in\mathbb{R}^d} \|A'x - b'\|_r$ and returns the solution $x$.

**Algorithm 3:** Protocol for $\ell_p$ regression in the message passing model

**Theorem 20** *The protocol described in Figure 3 returns a $(1 \pm \varepsilon)$-approximate solution to the $\ell_p$-regression problem with large constant probability. The communication complexity is $\widetilde{O}(sd^2/\varepsilon + sd/\varepsilon^{O(1)})$ and the total runtime of all servers is $O((\sum_i \operatorname{nnz}(A^i)) \cdot (d/\varepsilon^{O(1)}) + s \cdot \operatorname{poly}(d/\varepsilon))$.*

We remark that when all leverage scores of $[A\ b]$ are $\operatorname{poly}(\varepsilon)/d^{4/p}$, the servers can first uniformly sample $O(\operatorname{poly}(\varepsilon)/d \cdot n)$ rows of $A$ using the public random bits, rescale the sampled rows and obtain an $A'$. The servers can then run the protocol on $A'$. This modified protocol will still produce a $(1+\varepsilon)$-approximate solution to the $\ell_p$-regression problem and has the same communication complexity because uniform sampling does not require communication. The runtime is now reduced to $O(\sum_i \operatorname{nnz}(A^i) + s \cdot \operatorname{poly}(d/\varepsilon))$, which is optimal in terms of $\operatorname{nnz}(A^i)$. The details, including the formal statement, can be found in Appendix G.

## Acknowledgements

Y. Li is supported in part by Singapore Ministry of Education (AcRF) Tier 1 grant RG75/21 and Tier 2 grant MOE-T2EP20122-0001. H. Lin and D. Woodruff would like to thank support from the National Institute of Health (NIH) grant 5R01 HG 10798-2 and the Office of Naval Research (ONR) grant N00014-18-1-2562.

## References

Arturs Backurs, Piotr Indyk, Eric Price, Ilya P. Razenshteyn, and David P. Woodruff. Nearly-optimal bounds for sparse recovery in generic norms, with applications to $k$-median sketching. *CoRR*, abs/1504.01076, 2015.

Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68(4):702–732, 2004.

Jean Bourgain and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. *CoRR*, abs/1311.2542, 2013.

Christos Boutsidis, David P. Woodruff, and Peilin Zhong. Optimal principal component analysis in distributed and streaming models. In Daniel Wichs and Yishay Mansour, editors, *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 236–249. ACM, 2016.

Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein. Information lower bounds via self-reducibility. *Theory Comput. Syst.*, 59(2):377–396, 2016.

Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew Chi-Chih Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, 14-17 October 2001, Las Vegas, Nevada, USA*, pages 270–278. IEEE Computer Society, 2001.

Cheng Chen, Yi Li, and Yiming Sun. Online active regression. arXiv:2207.05945 [cs.LG], 2022. URL https://doi.org/10.48550/.

Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In Michael Mitzenmacher, editor, *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pages 205–214. ACM, 2009.

Kenneth L. Clarkson and David P. Woodruff. Low-rank approximation and regression in input sparsity time. *J. ACM*, 63(6), January 2017. ISSN 0004-5411. doi: 10.1145/3019134. URL https://doi.org/10.1145/3019134.

Kenneth L. Clarkson, Ruosong Wang, and David P. Woodruff. Dimensionality reduction for tukey regression. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1262–1271. PMLR, 2019.

Michael B. Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '16, page 278–287, USA, 2016. Society for Industrial and Applied Mathematics. ISBN 9781611974331.

Michael B. Cohen and Richard Peng. $l_p$ row sampling by lewis weights. In Rocco A. Servedio and Ronitt Rubinfeld, editors, *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 183–192. ACM, 2015. doi: 10.1145/2746539.2746567. URL https://doi.org/10.1145/2746539.2746567.

Omer Friedland and Olivier Guédon. Random embedding of $\ell_p^n$ into $\ell_r^n$. *Mathematische Annalen*, 350(4):953–972, 2011.

André Gronemeier. Asymptotically optimal lower bounds on the nih-multi-party information complexity of the and-function and disjointness. In Susanne Albers and Jean-Yves Marion, editors, *26th International Symposium on Theoretical Aspects of Computer Science, STACS 2009, February 26-28, 2009, Freiburg, Germany, Proceedings*, volume 3 of *LIPIcs*, pages 505–516. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Germany, 2009.

T. S. Jayram. Hellinger strikes back: A note on the multi-party information complexity of AND. In Irit Dinur, Klaus Jansen, Joseph Naor, and José D. P. Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 12th International Workshop, APPROX 2009, and 13th International Workshop, RANDOM 2009, Berkeley, CA, USA, August 21-23, 2009. Proceedings*, volume 5687 of *Lecture Notes in Computer Science*, pages 562–573. Springer, 2009.

Ravi Kannan, Santosh S. Vempala, and David P. Woodruff. Principal component analysis and higher correlations for distributed data. In Maria-Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 1040–1057. JMLR.org, 2014.

Yi Li, David P. Woodruff, and Taisuke Yasuda. Exponentially improved dimensionality reduction for l1: Subspace embeddings and independence testing. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 3111–3195. PMLR, 2021.

Yi Li, Honghao Lin, and David P. Woodruff. *The $\ell_p$-Subspace Sketch Problem in Small Dimensions with Applications to Support Vector Machines*, pages 850–877. SIAM, 2023. doi: 10.1137/1.9781611977554.ch34.

Arvind V. Mahankali, David P. Woodruff, and Ziyu Zhang. Near-linear time and fixed-parameter tractable algorithms for tensor decompositions. arXiv:2207.07417 [cs.DS], 2022.

Cameron Musco, Christopher Musco, David P Woodruff, and Taisuke Yasuda. Active sampling for linear regression beyond the $l_2$ norm. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022, Denver, CO, USA, Oct 31–Nov 3, 2022*, pages 744–753. IEEE, 2022.

S. Muthukrishnan. Data streams: Algorithms and applications. *Found. Trends Theor. Comput. Sci.*, 1(2), 2005. doi: 10.1561/0400000002. URL https://doi.org/10.1561/0400000002.

J. Nelson and H. L. Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 117–126, 2013.

Jeff M. Phillips, Elad Verbin, and Qin Zhang. Lower bounds for number-in-hand multiparty communication complexity, made easy. *SIAM J. Comput.*, 45(1):174–196, 2016.

Gilles Pisier. On the dimension of the $\ell_p^n$-subspaces of banach spaces, for $1 \leq p < 2$. *Trans. of AMS*, 276:201–211, 1983.

Christian Sohler and David P. Woodruff. Subspace embeddings for the $l_1$-norm with applications. In Lance Fortnow and Salil P. Vadhan, editors, *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 755–764. ACM, 2011. doi: 10.1145/1993636.1993736. URL https://doi.org/10.1145/1993636.1993736.

Konstantin Tikhomirov. Singularity of random bernoulli matrices. *Annals of Mathematics*, 191(2): 593–634, 2020.

Santosh S. Vempala, Ruosong Wang, and David P. Woodruff. The communication complexity of optimization. In Shuchi Chawla, editor, *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pages 1733–1752. SIAM, 2020.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, pages 299–303, 1965.

Ruosong Wang and David P. Woodruff. Tight bounds for $\ell_p$ oblivious subspace embeddings. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 1825–1843. SIAM, 2019.

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

David P. Woodruff and Qin Zhang. Subspace embeddings and lp regression using exponential random variables. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 546–567. JMLR.org, 2013.

Vladimir M Zolotarev. *One-dimensional stable distributions*, volume 65. American Mathematical Soc., 1986.

## Appendix A. Proof of Lemma 11

We follow the analysis of a similar objective function in (Musco et al., 2022, Theorem 12.2). Suppose that $a_i = b_i$ for $r$ coordinates $i$ and $a_i \neq b_i$ for $n - r$ coordinates $i$. The objective function $\|ax - b\|_p^p$ can be rewritten as

$$r \cdot |1 - x|^p + (n - r) \cdot |1 + x|^p \, .$$

**Case $p \in (0,1)$.** The first observation is that the optimal solution $x^*$ should lie in $[-1,1]$, otherwise $x = 1$ or $x = -1$ will give a lower cost. Next, without loss of generality, we can assume that $\Delta(a,b) \geq c\sqrt{n}$, which means that $r \geq \frac{n}{2} + \frac{c}{2}\sqrt{n}$. Following a similar analysis to that in (Musco et al., 2022, Theorem 12.2), we can now obtain that the optimal solution $x^*$ satisfies $x^* > 0$ and every $x < 0$ will lead to $\|ax - b\|_p^p \geq (1+\varepsilon)\|ax^* - b\|_p^p$. The case where $\Delta(a,b) \leq -c\sqrt{n}$ is similar, where the optimal solution $x^*$ satisfies $x^* < 0$ and every $x > 0$ will lead to $\|ax - b\|_p^p \geq (1+\varepsilon)\|ax^* - b\|_p^p$. Hence, using the sign of $x$ and the fact that $x$ is a $(1+\varepsilon)$-approximate solution, we can distinguish the two cases of $\Delta(a,b)$.

**Case $p = 1$.** The objective can now be rewritten as

$$r \cdot |1 - x| + (n - r) \cdot |1 + x| \ .$$

Without loss of generality, we assume that $\Delta(a,b) \geq c\sqrt{n}$ which means that $r \geq \frac{n}{2} + \frac{c}{2}\sqrt{n}$. The only thing we have to show is that $\|ax - b\|_p^p \geq (1+\varepsilon)\|ax^* - b\|_p^p$ for all $x < 0$. On the one hand, we have that $\|ax^* - b\|_p^p \leq \|a \cdot 1 - b\|_p^p \leq n - c\sqrt{n}$. On the other hand, when $x < 0$, noting that $r > n - r$, we have that $\|ax - b\|_p^p \geq \|a \cdot 0 - b\|_p^p = n \geq (1+\varepsilon)(n - c\sqrt{n})$. The last inequality follows from our choice of $n = \Theta(1/\varepsilon^2)$. To conclude, when $p = 1$, we can also distinguish the two cases from the sign of $x$.

**Case $p \in (1,2)$.** The case of $1 < p < 2$ was shown in (Musco et al., 2022, Theorem 12.4). Similar to their analysis, we can get that (i) when $\Delta(a,b) \geq c\sqrt{n}$, the optimal solution $x^*$ satisfies $x^* > 0$ and any $x < 0$ will yield $\|ax - b\|_p^p \geq (1 + 2\varepsilon^2)\|ax^* - b\|_p^p$; (ii) when $\Delta(a,b) \leq -c\sqrt{n}$, the optimal solution $x^*$ satisfies $x^* < 0$ and any $x > 0$ will yield $\|ax - b\|_p^p \geq (1 + 2\varepsilon^2)\|ax^* - b\|_p^p$. Hence, we can deduce the sign of $x$ in the two cases, and can distinguish the two cases when $x$ is a $(1+\varepsilon^2)$-approximate solution.

**Case $p = 2$.** The optimal solution is $x^* = \frac{\sum_i a_i b_i}{\sum_i a_i^2} = \frac{\sum_i a_i b_i}{n}$ and the corresponding objective value is $n - \frac{(\sum_i a_i b_i)^2}{n}$. When $\Delta(a,b) \geq c\sqrt{n}$, the optimal solution $x^* > 0$ and $\|ax^* - b\|_2^2 \leq n - c^2$, while for all $x < 0$, from the property of the quadratic function, we get that $\|ax^* - b\|_2^2 \geq \|a \cdot (0) - b\|_2^2 = n \geq (1 + 2\varepsilon^2)(n - c^2)$ (recall that $n \leq c/(2\varepsilon^2)$). A similar analysis works when $\Delta(a,b) \leq -c\sqrt{n}$ and the proof is complete.

## Appendix B. Proof of Lemma 14

We need the following theorem on the singularity probability of random sign matrices.

**Theorem 21 (Tikhomirov (2020))** *Let $M_n \in \mathbb{R}^{n \times n}$ be a random matrix whose entries are i.i.d. Rademacher random variables. It holds that*

$$\Pr\left[M_n \text{ is singular}\right] \leq \left(\frac{1}{2} + o_n(1)\right)^n \ .$$

The proof of the following lemma follows directly from the proof in Vempala et al. (2020) with only minor modifications.

**Lemma 22** *For sufficiently large $d$, there exists a set of matrices $\mathcal{T} \subseteq \{-1,1\}^{d \times d}$ with $|\mathcal{T}| = \Omega(2^{0.49d^2})$ such that*

1. *For any* $T \in \mathcal{T}$, $\mathrm{rank}(T) = d$*;*
2. *For any* $S, T \in \mathcal{T}$ *such that* $S \neq T$, $\mathrm{span}([S_{d-1}\ T_{d-1}]) = \mathbb{R}^d$*, where* $S_{d-1}$ *denotes the first* $d-1$ *column of* $S$*.*

**Proof** We use the probabilistic method to prove the existence. Let $t = 2-\varepsilon$, where $\varepsilon$ is a sufficiently small constant. We use $\mathsf{Bad} \subset \mathbb{R}^{d\times(d-1)}$ to denote the set

$$\mathsf{Bad} = \{B \in \mathbb{R}^{d\times(d-1)} \mid \Pr[X \in \mathrm{span}(B)] \geq c \cdot t^{-d} \text{ or } \mathrm{rank}(B) < d-1\},$$

where $X \in \mathbb{R}^d$ is a vector whose entries are i.i.d. Rademacher variables and $c$ is an absolute constant.

Consider a random matrix $A \in \mathbb{R}^{d\times(d-1)}$ with i.i.d. Rademacher entries. Then

$$\Pr[A \in \mathsf{Bad}] \leq \frac{1}{c}, \tag{2}$$

since otherwise, if we use $X \in \mathbb{R}^d$ to denote a vector with i.i.d. Rademacher coordinates, we have

$$\begin{aligned}
&\Pr[\mathrm{rank}([A\ X]) < d] \\
\geq &\Pr[\mathrm{rank}([A\ X]) < d \mid A \in \mathsf{Bad}] \cdot \Pr[A \in \mathsf{Bad}] \\
> &t^{-d},
\end{aligned}$$

which violates Theorem 21.

For any fixed $A \in \mathbb{R}^{d\times(d-1)} \setminus \mathsf{Bad}$, consider a random matrix $B \in \mathbb{R}^{d\times(d-1)}$ whose entries are i.i.d. Rademacher variables,

$$\Pr[\mathrm{span}([A\ B]) = \mathbb{R}^d] \geq 1 - \Pr\left[\bigcap_{i=1}^{d-1} B_i \in \mathrm{span}(A)\right] \geq 1 - c^d t^{-d(d-1)}, \tag{3}$$

which follows from the definition of $\mathsf{Bad}$ and the independence of the columns of $B$.

Now we construct a multiset $\mathcal{S}$ of $c^{-d}t^{d(d-1)/2}$ matrices, chosen uniformly with replacement from $\{-1, 1\}^{d\times d}$. By (2) and linearity of expectation, we have

$$\mathbb{E}[|\mathcal{S} \cap S_{\mathsf{Bad}}|] \leq c^{-d}t^{d(d-1)/2} \cdot \frac{1}{c},$$

where $S_{\mathsf{Bad}}$ denotes the set of the matrices $M$ such that the first $d-1$ columns of $M$ is in $\mathsf{Bad}$. Let $\mathcal{E}_1$ denote the event that

$$|\mathcal{S} \cap S_{\mathsf{Bad}}| \leq 4\,\mathbb{E}[|\mathcal{S} \cap S_{\mathsf{Bad}}|] \leq 4c^{-(d+1)}t^{d(d-1)/2},$$

which holds with probability at least $3/4$ by Markov's inequality.

Let $S_{\mathsf{rank}}$ denote the set of $d \times d$ matrices that are not of full rank. By (2) and linearity of expectation, we have

$$\mathbb{E}[|\mathcal{S} \cap S_{\mathsf{rank}}|] \leq c^{-d}t^{d(d-1)/2} \cdot t^{-d},$$

Let $\mathcal{E}_2$ denote the event that

$$|\mathcal{S} \cap S_{\mathsf{rank}}| \leq 4\,\mathbb{E}[|\mathcal{S} \cap S_{\mathsf{rank}}|] \leq 4c^{-d}t^{d(d-1)/2} \cdot t^{-d},$$

which holds with probability at least $3/4$ by Markov's inequality.

Let $\mathcal{E}_3$ denote the event that

$$\forall S \in \mathcal{S} \setminus S_{\mathsf{Bad}}, \forall T \in \mathcal{S} \setminus \{S\}, \mathrm{span}([S_{d-1}\ T_{d-1}]) = \mathbb{R}^d.$$

Using a union bound and (3),

$$\Pr(\mathcal{E}_3) \geq 1 - |\mathcal{S}|^2 c^d t^{-d(d-1)} = 1 - o_d(1).$$

Thus by a union bound, the probability that all $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$ hold is strictly larger than zero, which implies there exists a set $\mathcal{S}$ such that $\mathcal{E}_1$ $\mathcal{E}_2$, and $\mathcal{E}_3$ hold simultaneously. Now we consider $\mathcal{T} = \mathcal{S} \setminus (S_{\mathsf{Bad}} \cup S_{\mathsf{rank}})$. Since $\mathcal{E}_1$ and $\mathcal{E}_2$ hold, we have $|\mathcal{T}| \geq \Omega(c^{-2d} t^{d(d-1)/2}) = \Omega(2^{0.49d^2})$, provided that $d$ is sufficiently large and $\varepsilon$ is sufficiently small. The event $\mathcal{E}_3$ implies that all elements in $\mathcal{T}$ are distinct, and furthermore, it holds for any $S, T \in \mathcal{T}$ with $S \neq T$ that $\mathrm{span}([S_{d-1}\ T_{d-1}]) = \mathbb{R}^d$. ■

Suppose that $\mathcal{T}$ satisfies the conditions in Lemma 22. For each $T \in \mathcal{T}$, we add $T^T$ into $\mathcal{H}$. Now suppose there exist $S, T \in \mathcal{H}$ such that $S \neq T$ and $S^{-1}e_d = T^{-1}e_d$, which means there exists $x \in \mathbb{R}^d$ such that $Sx = e_d$ and $Tx = e_d$. This implies that $x^T(S^T)_{d-1} = x^T(T^T)_{d-1} = 0$. The construction of $\mathcal{T}$ guarantees that $\mathrm{span}([(S^T)_{d-1}\ (T^T)_{d-1}]) = \mathbb{R}^d$ and it must thus hold that $x = 0$, which would result in $Sx = Tx = 0 \neq e_d$. Therefore, for any $S, T \in \mathcal{H}$ with $S \neq T$, $S^{-1}e_d \neq T^{-1}e_d$.

## Appendix C. Proof of Theorem 17

To prove the correctness of Algorithm 2, we need the following lemmas. The reader can find more detail in (Woodruff, 2014).

**Lemma 23** *Suppose that $S$ is a $(1 \pm \sqrt{\varepsilon})$-subspace embedding and $x' = \arg\min_{x \in \mathbb{R}^d} \|S(Ax - b)\|_2$. Then it holds with large constant probability that*

$$\|Ax' - b\|_2 \leq (1 + \varepsilon)\|Ax - b\|_2 .$$

*Further suppose that $x_c$ is a $(1 + \varepsilon)$-approximate solution to $\min_{x \in \mathbb{R}^d} \|S(Ax - b)\|_2$, it then holds that*

$$\|S(Ax_c - b)\|_2 \leq (1 + \varepsilon)\|Ax - b\|_2 .$$

We remark that the case where $x_c$ is the minimizer was shown by Clarkson and Woodruff (2009) and the case where $x_c$ is a $(1 + \varepsilon)$-approximate solution was recently shown by Mahankali et al. (2022).

**Lemma 24** *Suppose that $S$ is a $(1 \pm \varepsilon_0)$-subspace embedding and consider the iterative algorithm above, then*

$$\|\hat{A}Rx_{t+1} - x^*\|_2 = \varepsilon_0^m \cdot \|\hat{A}Rx_t - x^*\|_2 .$$

*As a corollary, when $t = \Omega(\log(1/\varepsilon))$, it holds that $\|\hat{A}Rx_t - \hat{b}\|_2^2 \leq (1 + \varepsilon)\|\hat{A}Rx^* - \hat{b}\|_2^2$.*

Now we are ready to prove Theorem 17.

**Proof** [of Theorem 17] Since $\Pi_1$ has $O(d^2/\varepsilon)$ rows and $\Pi_2$ has $O(d\log(d)/\varepsilon)$ columns, from Section 2 we get that with probability at least $99/100$, both $\Pi_1$ and $\Pi_2$ are $(1 \pm O(\sqrt{\varepsilon}))$ subspace embeddings, which means $\Pi_2\Pi_1$ is a $(1+\sqrt{\varepsilon})$-subspace embedding.

Let $\hat{A} = \Pi_2\Pi_1 A$ and $\hat{b} = \Pi_2\Pi_1 b$. From Lemma 23, we see that it suffices to solve $\min_{x\in\mathbb{R}^d} \|\hat{A}x - \hat{b}\|_2$. Conditioned on these events, it follows immediately from Lemma 24 that $x_T$ is a $(1\pm\varepsilon)$-approximate solution to $\min_{x\in\mathbb{R}^d} \|\hat{A}x - \hat{b}\|_2$, provided that each server uses $R$ instead of $\widetilde{R}$. To show that $\widetilde{R}$ works here, note that an initial step in the proof of Lemma 24 is that $\|S\hat{A}Rx\|_2 = 1$ for all unit vectors $x$, which implies that $\|\hat{A}Rx\|_2 \in [1-\varepsilon_0, 1+\varepsilon_0]$. For $\widetilde{R}$, we have that

$$\|S\hat{A}Rx\|_2 - \|S\hat{A}\widetilde{R}x\|_2 \le \|S\hat{A}(R-\widetilde{R})x\|_2 \le 2\|\hat{A}\|_2\|(R-\widetilde{R})x\|_2 \le 1/\operatorname{poly}(nd)\ .$$

The last inequality is due to the fact that each entry of $R-\widetilde{R}$ is $O(1/\operatorname{poly}(nd))$ and each entry of $\hat{A}$ is $O(\operatorname{poly}(nd))$. Hence, $\|ARx\| \in [1-1.1\varepsilon_0, 1+1.1\varepsilon_0]$ will still hold and a similar argument will go through, yielding that $x_T$ is a $(1\pm\varepsilon)$-approximate solution.

We next analyze the communication complexity of the protocol. For Step 3, since $\Pi_3\hat{A}^i$ is an $O(d\log d)\times d$ matrix, each server $P_i$ sends $\widetilde{O}(d^2)$ entries. Each entry of $A^i$ has magnitude $[1/n^c, n^c]$, and thus each entry of $\Pi_1 A^i$ is contained in $[1/n^c, n^{c+1}]$, each entry of $\hat{A}^i = \Pi_2\Pi_1 A^i$ is contained in $[\varepsilon/n^{c+2}, n^{c+3}/\varepsilon]$ and each entry of $\Pi_3\hat{A}^i$ is contained in $[\varepsilon^2/n^{c+4}, n^{c+5}/\varepsilon^2]$, which implies that each entry of $\Pi_3\hat{A}^i$ can be described using $O(\log(n/\varepsilon))$ bits and thus a total communication of $O(sd^2)$ bits for Step 3. In Step 4, since $\widetilde{R}$ is a $d\times d$ matrix and each entry is an integer multiple of $1/\operatorname{poly}(nd)$, the coordinator sends $\widetilde{R}$ to each server using $\widetilde{O}(sd^2)$ bits in total. In each iteration of Step 5, we note that $y_t$ is an $O(d/\varepsilon)$-dimensional vector and $g_t$ is a $d$-dimensional vector, and each of their entries has $O(\log(nd))$ precision. Hence, the total communication of each iteration is $\widetilde{O}(sd/\varepsilon)$. Putting everything together, we conclude that the total amount of the communication is $\widetilde{O}(sd^2 + \log(1/\varepsilon)\cdot(sd/\varepsilon)) = \widetilde{O}(sd^2 + sd/\varepsilon)$ bits.

We now consider the runtime of the protocol. To compute $\Pi_2\Pi_1 A^i$, notice that $\Pi_1$ is a CountSketch matrix, and hence each server takes $\operatorname{nnz}(A^i)$ time to compute $\Pi_1 A^i$ and then use $\operatorname{poly}(d/\varepsilon)$ time to compute $\Pi_2(\Pi_1 A^i)$. Hence, Step 2 takes $O(\sum_i \operatorname{nnz}(A^i))$ time. For the remaining steps, one can verify that each step takes $\operatorname{poly}(d/\varepsilon)$ time on a single server or on the coordinator. The total runtime is therefore $O(\sum_i \operatorname{nnz}(A^i) + s\cdot\operatorname{poly}(d/\varepsilon))$. ■

## Appendix D. Proof of Lemma 18

To prove the lemma, we need the following results.

**Lemma 25 (see, e.g., Friedland and Guédon (2011))** *Suppose that $\alpha \in \mathbb{R}^d$ and $\theta \in \mathbb{R}^d$ is a vector whose entries are i.i.d. $p$-stable variables. Then it holds that*

$$\left(\mathbb{E}\left|\sum_i \alpha_i\theta_i\right|^r\right)^{1/r} = \alpha_{p,r}\left(\sum_i |\alpha_i|^p\right)^{1/p}$$

*where $\alpha_{p,r}$ is a constant that only depends on $p$ and $r$.*

**Proposition 26** *Suppose that $r, s \ge 1$ and $X$ is a random variable with $\mathbb{E}\,|X|^{rs} < \infty$. It holds that*

$$\mathbb{E}\,\big||X|^r - \mathbb{E}\,|X|^r\big|^s \le 2^s\,\mathbb{E}\,|X|^{rs}\ .$$

**Proof** We have that

$$\begin{aligned}\mathbb{E}\left||X|^r - \mathbb{E}\,|X|^r\right|^s &\leq 2^{s-1}\,\mathbb{E}\left(\left||X|^r\right|^s + (\mathbb{E}\,|X|^r)^s\right)\\ &\leq 2^{s-1}(\mathbb{E}\,|X|^{rs} + (\mathbb{E}\,|X|^r)^s)\\ &\leq 2^{s-1}(\mathbb{E}\,|X|^{rs} + \mathbb{E}\,|X|^{rs})\\ &= 2^s\,\mathbb{E}\,|X|^{rs}\,.\end{aligned}$$

■

**Lemma 27 (von Bahr and Esseen (1965, Theorem 2))** *Suppose that $1 \leq r \leq 2$. Let $X_1, \ldots, X_n$ be independent zero mean random variables with $\mathbb{E}[|X_i|^r] < \infty$. Then we have that*

$$\mathbb{E}\left[\left(\sum_{i=1}^n |X_i|\right)^r\right] \leq 2\sum_{i=1}^n \mathbb{E}\left[|X_i|^r\right]\,.$$

**Lemma 28** *Suppose that $p \in (1,2)$ is a constant and $T \in \mathbb{R}^{m\times n}$ is a matrix whose entries are i.i.d. $p$-stable entries scaled by $1/(\alpha_p \cdot m^{1/p})$. For $m = d\log d/\varepsilon^{O(1)}$, given any $A \in \mathbb{R}^{n\times d}$, it holds with large constant probability that for all $x \in \mathbb{R}^d$*

$$\|TAx\|_p \leq \mathrm{poly}(d)\|Ax\|_p\,.$$

We note that Lemma 28 was shown in Sohler and Woodruff (2011) for $p = 1$. For $1 < p < 2$, a similar argument still goes through after replacing the $\ell_1$ well-conditioned basis with an $\ell_p$ well-conditioned basis.

**Proof** [of Lemma 18] First we consider the original $T$ without rounding the entries.

Now we show (1). Let $y = Ax$. From properties of $p$-stable random variables, we get that each $(Ty)_i$ follows the same distribution. From Lemma 25 we have that for every $i$, $\mathbb{E}\,|(Ty)_i|^r = \frac{\alpha_{p,r}^r}{\alpha_{p,r}^r \cdot m}\|y\|_p^r = \frac{1}{m}\|y\|_p^r$. To get concentration, we pick an $r' \in (r,p)$ and consider the $r'/r$-moment of $(Ty)_i^r$.

Similar to Lemma 25, we have that $\mathbb{E}[|(Ty)|_i^{r'}] = \frac{\beta_{p,r,r'}}{m^{r'/r}}\|y\|_p^{r'}$ is bounded, where $\beta_{p,r,r'}$ is a constant depending on $p, r, r'$ only. Let $S = \sum_i |(Ty)_i|^r$ and we have that $\mathbb{E}[S] = \|y\|_p^r$. Consider the $(r/r')$-th moment of $S$. We then have

$$\begin{aligned}\mathbb{E}\left[(S - \mathbb{E}[S])^{r'/r}\right] &= \mathbb{E}\left[\left(\sum_i \left(|(Ty)_i|^r - \frac{1}{m}\|y\|_p^r\right)\right)^{r'/r}\right]\\ &\leq 2\left(\sum_i \mathbb{E}\left[\left|(Ty)_i|^r - \frac{1}{m}\|y\|_p^r\right|\right]^{r'/r}\right) && \text{(Lemma 27)}\\ &\leq 2^{r'/r+1}\left(\sum_i \mathbb{E}\,|(Ty)_i|^{r'}\right) && \text{(Proposition 26)}\\ &\leq C\left(\sum_i \frac{1}{m^{r'/r}}\|y\|_p^{r'}\right)\\ &= C\|y\|_p^{r'}/m^{r'/r-1}\,,\end{aligned}$$

where $C$ is a constant that depends only on $r, r'$, and $p$. By Markov's inequality, we have that

$$\begin{aligned}\mathbf{Pr}\left[|S-\mathbb{E}[S]| \geq \varepsilon\,\mathbb{E}[S]\right] &\leq \mathbf{Pr}\left[|S-\mathbb{E}[S]|^{r'/r} \geq (\varepsilon\,\mathbb{E}[S])^{r'/r}\right] \\ &\leq \frac{\mathbb{E}\left[(S-\mathbb{E}[S])^{r'/r}\right]}{\varepsilon^{r'/r}\|y\|_p^{r'}} \\ &\leq \frac{C_{r'/r}}{\varepsilon^{r'/r}m^{r'/r-1}}\,.\end{aligned}$$

Hence, we can see that when $m = \Omega(1/\varepsilon^{\frac{r'}{r'-r}}) = 1/\varepsilon^{\Omega(1)}$, $|\|Ty\|_r - \|y\|_p| \leq \varepsilon\|y\|_p$ holds with large constant probability.

We next prove (2). We first show that for every $x \in \mathbb{R}^d$, $\|Ty\|_r^r \geq (1-\varepsilon)\|y\|_p^r$ holds with probability at least $1-\exp(-d\log(d)/\varepsilon^{O(1)})$. Recall that we have that we have that $\mathbb{E}\,|(Ty)_i|^r = \frac{1}{m}\|y\|_p^r$ for every $i$. Fix $k = 1/\varepsilon^{O(1)}$. Let

$$s_i = |(Ty)_{(i-1)k+1}|^r + |(Ty)_{(i-1)k+2}|^r + \cdots + |(Ty)_{ik}|^r \;\; (1 \leq i \leq m/k)\,.$$

We then have $\|Ty\|_r^r = \sum_i s_i$. Similar to (1), one can show that for each $i$, with large constant probability

$$\left|s_i - \frac{k}{m}\|y\|_p^r\right| \leq \varepsilon\frac{k}{m}\|y\|_p^r \tag{4}$$

By a Chernoff bound, with probability at least $1-\exp(-d/\varepsilon^{\Omega(1)})$, at least a $(1-\varepsilon)$-fraction of the $s_i$ satisfy (4). Conditioned on this event, it holds that

$$\|Ty\|_r^r = \sum_i s_i \geq \frac{m}{k}(1-\varepsilon)\frac{k}{m}\|y\|_p^r = (1-\varepsilon)\|y\|_p^r\,,$$

which is what we need.

The next is a standard net-argument. Let $\mathcal{S} = \{Ax : x \in \mathbb{R}^d, \|Ax\|_p = 1\}$ be the unit $\ell_p$-ball and $\mathcal{N}$ be a $\gamma$-net with $\gamma = \mathrm{poly}(\varepsilon/d)$ under the $\ell_p$ distance. It is a standard fact that the size of $\mathcal{N}$ can be $(\mathrm{poly}(d/\varepsilon))^d$. By a union bound, we have that $\|TAx\|_r \geq (1-\varepsilon)\|Ax\|_p = (1-\varepsilon)$ for all $Ax \in \mathcal{N}$ simultaneously with probability at least $9/10$. From Lemma 28, we have that with probability at least $9/10$, $\|TAx\|_p \leq \mathrm{poly}(d)\|Ax\|_p$ for all $x \in \mathbb{R}^d$. Conditioned on these events, we then have for all $x \in \mathbb{R}^d$,

$$\|TAx\|_r \leq m^{1/r-1/p}\|TAx\|_p \leq \mathrm{poly}(d/\varepsilon)\|Ax\|_p\,.$$

Then, for each $y = Ax \in \mathcal{S}$, we choose a sequence of points $y_0, y_1, \cdots \in \mathcal{S}$ as follows.

- Choose $y_0 \in \mathcal{S}$ such that $\|y - y_0\|_p \leq \gamma$ and let $\alpha_0 = 1$;
- After choosing $y_0, y_1, \ldots, y_i$, we choose $y_{i+1}$ such that

$$\left\|\frac{y - \alpha_0 y_0 - \alpha_1 y_1 - \cdots - \alpha_i y_i}{\alpha_{i+1}} - y_{i+1}\right\|_p \leq \gamma,$$

  where $\alpha_{i+1} = \|y - \alpha_0 y_0 - \alpha_1 y_1 - \cdots - \alpha_i y_i\|_p$.

The choice of $y_{i+1}$ means that

$$\alpha_{i+2} = \|y - \alpha_0 y_0 - \alpha_1 y_1 - \cdots - \alpha_i y_i - \alpha_{i+1} y_{i+1}\|_p \leq \alpha_{i+1}\gamma.$$

A simple induction yields that $\alpha_i \leq \gamma^i$. Hence

$$y = y_0 + \sum_{i \geq 1} \alpha_i y_i, \quad |\alpha_i| \leq \gamma^i \ .$$

Suppose that $y_i = Ax_i$. We have

$$\|TAx\|_r \geq \|TAx_0\|_p - \sum_{i \geq 1} \gamma^i \|TAx_i\|_p \geq (1-\varepsilon) - \sum_{i \geq 1} \gamma^i \cdot (\mathrm{poly}(d/\varepsilon)) = 1 - O(\varepsilon).$$

Rescaling $\varepsilon$, we obtain that $\|TAx\|_r^r \geq (1-\varepsilon)\|Ax\|_p^r$ for all $x \in \mathbb{R}^d$ simultaneously.

This completes the proof of the two guarantees for the original $T$, without rounding the entries. To show that the guarantees continue to hold after rounding the entries, We only need to notice that

$$\begin{aligned}\left| \|\widetilde{T}Ax\|_r - \|TAx\|_r \right| \leq \|(\widetilde{T} - T)Ax\|_r &\leq m^{\frac{1}{r}-\frac{1}{2}} \|(\widetilde{T} - T)Ax\|_2 \\ &\leq m^{\frac{1}{r}-\frac{1}{2}} \|\widetilde{T} - T\|_2 \|Ax\|_2 \\ &\leq \frac{1}{\mathrm{poly}(nd)} \|Ax\|_p \ .\end{aligned}$$

■

## Appendix E. Proof of Lemma 19

The proof of the lemma closely follows that in Chen et al. (2022). The proof is a bootstrapping argument based on the following two lemmas. For simplicity of notation, we define $R(A, b) = \min_x \|Ax - b\|_p$.

**Lemma 29 ((Musco et al., 2022, Theorem 3.18))** *There exists an absolute constant $c \in (0, 1]$ such that the following holds for all $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$ and $\gamma \in (0, 1)$. Let $x^* = \arg\min_{x \in \mathbb{R}^d} \|Ax - b\|_p$. Whenever $x \in \mathbb{R}^d$ satisfies $\|Ax - b\|_p \leq (1 + c\gamma)R(A, b)$, we have that $\|Ax^* - Ax\|_p \leq \sqrt{\gamma} R(A, b)$.*

**Lemma 30** *Let $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$ and $0 < \gamma < 1$. Let $S$ be the rescaled sampling matrix with respect to $\{p_i\}_{(i)}$ such that $p_i = \min\{\beta w_i([A\ b]), 1\}$ and $\beta = \Theta(\frac{\gamma}{\varepsilon^2} \log^2 d \log n \log \frac{1}{\delta})$. Suppose that $\widetilde{x} = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_p$ and $\|A\widetilde{x} - Ax^*\|_p \leq \sqrt{\gamma} R(A, b)$. It holds that*

$$\|A\widetilde{x} - b\|_p \leq (1 + C\varepsilon)R(A, b)$$

*with probability at least $0.99 - \delta$, where $C$ is an absolute constant.*

Assuming these two lemmas, the proof of Lemma 19 is nearly identical to that in the proof of (Chen et al., 2022, Theorem 5.6) and is thus omitted. The proof is simpler because we do not need an argument to first show that sketching by $S$ gives a $(1+O(\sqrt{\varepsilon}))$-approximate solution, which follows immediately from the fact that $S$ is a $(1+\sqrt{\varepsilon})$-subspace-embedding with large constant probability.

In the remainder of this section, we discuss the proof of Lemma 30. The proof is similar to that of (Chen et al., 2022, Lemma 5.4), which converts the bound on the target dimension obtained from an iterative argument in Musco et al. (2022) to a moment bound using the framework in Cohen and Peng (2015).

The difference is that here we can choose the weights to be the Lewis weights of $[Ab]$, while in (Chen et al., 2022, Lemma 5.4), it considers $\min_x \|Ax-z\|_p$ with $\|z\|_p \le R(A,b)$ and it samples the rows according to the Lewis weights of $A$. Specifically, let $R = R(A,b)$, $x' = x - x^*$ and $b' = b - Ax^*$. We have, as in the proof of (Chen et al., 2022, Lemma 5.4), that

$$\begin{aligned}
\|A\widetilde{x}-b\|_p^p - \|Ax^*-b\|_p^p &= \|A\widetilde{x}-b\|_p^p - \|SA\widetilde{x}-Sb\|_p^p + \|SA\widetilde{x}-Sb\|_p^p - \|SAx^*-Sz\|_p^p \\
&\quad + \|SAx^*-Sb\|_p^p - \|Ax^*-b\|_p^p \\
&\le \|A\widetilde{x}-b\|_p^p - \|SA\widetilde{x}-Sb\|_p^p + \|SAx^*-Sb\|_p^p - \|Ax^*-b\|_p^p \\
&\le \left\|Ax'-b'\right\|_p^p - \left\|SAx'-Sb'\right\|_p^p + \left\|Sb'\right\|_p^p - \left\|b'\right\|_p^p \\
&= \left\|Ax'-b'\right\|_p^p - \left\|Ax'-\overline{b}'\right\|_p^p - \left\|b'-\overline{b}'\right\|_p^p \\
&\quad - \left(\left\|SAx'-Sb'\right\|_p^p - \left\|SAx'-S\overline{b}'\right\|_p^p - \left\|Sz'-S\overline{b}'\right\|_p^p\right) \\
&\quad - \left(\left\|SAx'-S\overline{b}'\right\|_p^p - \left\|Ax'-\overline{b}'\right\|_p^p + \left\|\overline{b}'\right\|_p^p - \left\|S\overline{b}'\right\|_p^p\right) \\
&=: E_1 - E_2 - E_3,
\end{aligned}$$

where $\overline{b}$ is the vector obtained from $b$ by removing all coordinates $b_i$ such that $|b_i| \ge \frac{w_i}{\varepsilon}R$. Note that $\|\overline{b}'\|_p \le \|b'\|_p = R$ and $\|Ax'\|_p \le \sqrt{\gamma}R$. The first term can be controlled using (Musco et al., 2022, Lemma 3.5), except that the sampling probabilities are Lewis weights of $[A\ b]$ instead of $A$, but the proof still goes through because it also holds that $|(Ax)_i|^p \le \|Ax\|_p s_i^p([A\ b])$, where $s_i([A\ b])$ is the $\ell_p$-sensitivity of $[A\ b]$. The second term can be controlled by (Musco et al., 2022, Lemma 3.6), yielding that $|E_2| \le \varepsilon R^p$ with probability at least 0.99. The last term can be controlled as in (Chen et al., 2022, Lemma 5.3), where the Lewis weights of $[A\ b]$ do not affect the proof.

## Appendix F. Proof of Theorem 20

By Lemma 18(1), it holds with high constant probability that

$$\min_{x\in\mathbb{R}^d} \|T(Ax-b)\|_r \le (1+\varepsilon)\min_{x\in\mathbb{R}^d}\|Ax-b\|_p\ .$$

Suppose that $x'\in\mathbb{R}^d$ is a $(1+\varepsilon)$-approximate solution to $\min_{x\in\mathbb{R}^d}\|T(Ax-b)\|_r$, i.e.,

$$\|T(Ax'-b)\|_r \le (1+\varepsilon)\min_{x\in\mathbb{R}^d}\|T(Ax-b)\|_r\ .$$

It follows from Lemma 18(2) that

$$\|Ax' - b\|_p \leq \frac{1}{1-\varepsilon}\|T(Ax' - b)\|_r \leq (1 + O(\varepsilon)) \min_{x\in\mathbb{R}^d} \|Ax - b\|_p \, .$$

Hence, the problem is reduced to obtaining a $(1+\varepsilon)$-approximate solution to $\min_{x\in\mathbb{R}^d} \|T(Ax - b)\|_r = \min_{x\in\mathbb{R}^d} \|\hat{A}x - \hat{b}\|_r$.

Consider the iteration in Step 3. A standard analysis (see, e.g. Woodruff, 2014, Section 2.4) yields that in each iteration, with probability at least $1 - 1/\operatorname{poly}(d)$, $\tau$ is a constant approximation to the leverage score of $W^{1/p-1/2}B$. Taking a union bound, we get that with high constant probability, for all iterations it holds. Conditioned on this event happening, from Lemma 6 we get that after $t$ iterations, $w$ is a constant approximation to the $\ell_r$ Lewis weights of $B$ (in each iteration we round $w$; however, notice that if the Lewis weight $w_i$ is not 0, it should be larger than $1/\operatorname{poly}(nd)$ as the non-zero entries of the matrix $B$ are at least $1/\operatorname{poly}(nd)$[2] , and hence the rounding will not affect the approximation ratio guarantee in each iteration). From Lemma 19, the solution to $\min_{x\in\mathbb{R}^d} \|A'x - b'\|_r$ is a $(1+\varepsilon)$-approximate solution to $\min_{x\in\mathbb{R}^d} \|T(Ax - b)\|_r$, and is thus a $(1 \pm O(\varepsilon))$-approximate solution to the original problem $\min_{x\in\mathbb{R}^d} \|Ax - b\|_p$.

We next analyze the communication complexity of the protocol. For Step 3(a), $S_t\widetilde{W}^{1/2-1/p}B_i$ is a $d\log(d) \times (d+1)$ matrix and the entries of $S_t\widetilde{W}^{1/2-1/p}B_i$ are in $\operatorname{poly}(nd)$-precision as the entries of $S_t, \widetilde{W}^{1/2-1/p}$, and $B_i$ are both in $\operatorname{poly}(nd)$-precision. Hence, the total communication of all servers is $\widetilde{O}(sd^2)$. For Step 3(b), $\widetilde{R}$ is a $(d+1)\times(d+1)$ matrix and hence the total communication cost is $\widetilde{O}(sd^2)$. For 3(c), $B^i\widetilde{R}G$ is a $d/\varepsilon^{O(1)} \times O(\log d)$ matrix, and hence similarly we get that the total communication cost is $O(sd/\varepsilon^{O(1)})$. For 3(e), since $w$ is a $d/\varepsilon^{O(1)}$ vector, the total communication cost of this step is $O(sd/\varepsilon^{O(1)})$. In Step 5, since the sum of Lewis weights is $O(d)$, with high constant probability the server samples at most $\widetilde{O}(d/\varepsilon)$ rows, and hence the communication cost of this step is $O(sd^2/\varepsilon)$. Putting everything together, we get that the total communication cost is

$$\widetilde{O}\left(\log\log(d/\varepsilon)\cdot(sd^2 + sd/\varepsilon^{O(1)}) + sd^2/\varepsilon\right) = \widetilde{O}(sd^2/\varepsilon + sd/\varepsilon^{O(1)}) \, .$$

We now consider the runtime of the protocol. To compute $TA^i$, notice that $T$ has $d/\varepsilon^{O(1)}$ rows, which means it takes $O(\operatorname{nnz}(A^i)\cdot(d/\varepsilon^{O(1)}))$ times to compute $TA^i$. Hence Step 2 takes time $O((\sum_i \operatorname{nnz}(A^i))\cdot(d/\varepsilon^{O(1)}))$. For the remaining steps, one can verify that each step takes $\operatorname{poly}(d/\varepsilon)$ time on a single server or on the coordinator. The total runtime is therefore $O(\sum_i \operatorname{nnz}(A^i)\cdot(d/\varepsilon^{O(1)}) + s\cdot\operatorname{poly}(d/\varepsilon))$.

## Appendix G. Faster Runtime for Distributed $\ell_p$-Regression

We need the following auxiliary results.

**Proposition 31** *Suppose that $1 \leq p < 2$ and $A \in \mathbb{R}^{n\times d}$. The $\ell_p$-sensitivity scores of $A$ are defined as*

$$\ell_i^{(p)}(A) = \sup_{x:Ax\neq 0} \frac{|\langle a_i, x\rangle|^p}{\|Ax\|_p^p},$$

2. It is easy to see that the $\ell_r$ sensitivities, defined in Proposition 31 in Section G, are at least $\Omega(1/\operatorname{poly}(nd))$ in our setting if the corresponding rows are nonzero as if we take $x = a_i$, we can get that the $\ell_i^{(r)}(A) \geq \|a_i\|^{2p}/\|Aa_i\|_p^p$ where the denominator is at most $\operatorname{poly}(nd)$ as each entry of $A$ is in $\operatorname{poly}(nd)$. From Lemma 2.5 in Musco et al. (2022), we know that the $\ell_r$ Lewis weights are larger than the $\ell_r$ sensitivities when $r < 2$.

*where $a_i$ is the $i$-th row of $A$. It holds that $\ell_i^{(p)}(A) \leq (\tau_i(A))^{p/2}$ for all $i$.*

**Proof** Suppose that $A$ has full column rank, otherwise we can find an invertible matrix $T$ such that $AT = [A' \ 0]$, where $A'$ has full column rank, and consider $\ell_i^{(p)}(A')$ and $\tau_i(A')$ instead. It is not difficult to verify that $\ell_i^{(p)}(A') = \ell_i^{(p)}(A)$ and $\tau_i(A') = \tau_i(A)$.

Write $A = UR$, where $U \in \mathbb{R}^{n\times d}$ has orthonormal columns and $R \in \mathbb{R}^{d\times d}$ is invertible. Then

$$\ell_i^{(p)}(A) = \sup_{y\neq 0} \frac{|\langle U_i, y\rangle|^p}{\|Uy\|_p^p} \leq \sup_{y\neq 0} \frac{\|U_i\|_2^p \|y\|_2^p}{\|Uy\|_2^p} = \|U_i\|_2^p = (\tau_i(A))^{p/2},$$

as advertised. ■

**Lemma 32 ((Li et al., 2023, Lemma 5.5))** *Let $A \in \mathbb{R}^{n\times d}$ and $1 \leq p < \infty$. The matrix $A'$ is a submatrix of $A$ such that the rescaled $i$-th row $p_i^{-1/p} a_i$ is included in $A'$ with probability $p_i \geq \min(\beta s_i(A), 1)$. Then, there is a constant $c$ such that when $\beta \geq c\varepsilon^{-2} d\log(1/\varepsilon)$, the matrix $A'$ is a $(1\pm\varepsilon)$-subspace embedding of $A$ with probability at least $9/10$.*

As an immediate corollary of the auxiliary results above, we have that when $A \in \mathbb{R}^{n\times d}$ has uniformly small leverage scores, uniformly sampling its rows can give an $\ell_p$-subspace-embedding (after rescaling).

**Corollary 33** *Suppose that $1 \leq p < 2$ and the matrix $A \in \mathbb{R}^{n\times d}$ satisfies that $\tau_i(A) \leq (c\varepsilon^2\gamma/(d\log(1/\varepsilon)))^{2/p}$ for all $i$, where $\gamma \leq \varepsilon^2/(Cd\log(1/\varepsilon))$. Let $A'$ be a matrix formed from $A$ by retaining each row with probability $\gamma$ independently and then rescaling by $1/\gamma^{1/p}$. It holds with large constant probability that*

$$(1-\varepsilon)\|Ax\|_p^p \leq \|A'x\|_p^p \leq (1+\varepsilon)\|Ax\|_p^p$$

*for all $x \in \mathbb{R}^d$ simultaneously, and that $A'$ has $O(\gamma n)$ rows.*

**Proof**

By Proposition 31, $\ell_i^{(p)}(A) \leq (\tau_i(A))^{p/2} = c\varepsilon^2\gamma/(d\log(1/\varepsilon))$, so the sampling probability

$$\gamma \geq \frac{Cd\log(1/\varepsilon)}{\varepsilon^2} \cdot \ell_i^{(p)}(A)$$

satisfies the condition in Lemma 32. The conclusion follows immediately. ■

Hence, if $A$ has uniformly small leverage scores, all sites can agree on the $O(\gamma n)$ uniformly sampled rows using the public random bits and run the protocol in Algorithm 3 on the induced $A'$. By Markov's inequality, $\mathrm{nnz}(A') = O(\gamma \mathrm{nnz}(A))$ with large constant probability and we finally conclude with the following theorem.

**Theorem 34** *Suppose that $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^d$ satisfies that the leverage scores of $[A\ b]$ are all bounded by $\mathrm{poly}(\varepsilon)/d^{4/p}$. There is a protocol which outputs a $(1\pm\varepsilon)$-approximate solution to the $\ell_p$-regression problem with large constant probability, using $\widetilde{O}(sd^2/\varepsilon + sd/\varepsilon^{O(1)})$ bits of communication and running in total time (over all servers) $O(\sum_i \mathrm{nnz}(A^i) + s\cdot\mathrm{poly}(d/\varepsilon))$.*