

Low-distortion Subspace Embeddings in Input-sparsity Time and Applications to Robust Linear Regression

Xiangrui Meng*
LinkedIn Corporation
2029 Stierlin Ct, Mountain View, CA 94043
ximeng@linkedin.com

Michael W. Mahoney
Dept. of Mathematics, Stanford University
Stanford, CA 94305
mmahoney@cs.stanford.edu

ABSTRACT

Low-distortion embeddings are critical building blocks for developing random sampling and random projection algorithms for common linear algebra problems. We show that, given a matrix $A \in \mathbb{R}^{n \times d}$ with $n \gg d$ and a $p \in [1, 2]$, with a constant probability, we can construct a low-distortion embedding matrix $\Pi \in \mathbb{R}^{\mathcal{O}(\text{poly}(d)) \times n}$ that embeds \mathcal{A}_p , the ℓ_p subspace spanned by A 's columns, into $(\mathbb{R}^{\mathcal{O}(\text{poly}(d))}, \|\cdot\|_p)$; the distortion of our embeddings is only $\mathcal{O}(\text{poly}(d))$, and we can compute ΠA in $\mathcal{O}(\text{nnz}(A))$ time, i.e., input-sparsity time. Our result generalizes the input-sparsity time ℓ_2 subspace embedding by Clarkson and Woodruff [STOC'13]; and for completeness, we present a simpler and improved analysis of their construction for ℓ_2 . These input-sparsity time ℓ_p embeddings are optimal, up to constants, in terms of their running time; and the improved running time propagates to applications such as $(1 \pm \epsilon)$ -distortion ℓ_p subspace embedding and relative-error ℓ_p regression. For ℓ_2 , we show that a $(1 + \epsilon)$ -approximate solution to the ℓ_2 regression problem specified by the matrix A and a vector $b \in \mathbb{R}^n$ can be computed in $\mathcal{O}(\text{nnz}(A) + d^3 \log(d/\epsilon)/\epsilon^2)$ time; and for ℓ_p , via a subspace-preserving sampling procedure, we show that a $(1 \pm \epsilon)$ -distortion embedding of \mathcal{A}_p into $\mathbb{R}^{\mathcal{O}(\text{poly}(d))}$ can be computed in $\mathcal{O}(\text{nnz}(A) \cdot \log n)$ time, and we also show that a $(1 + \epsilon)$ -approximate solution to the ℓ_p regression problem $\min_{x \in \mathbb{R}^d} \|Ax - b\|_p$ can be computed in $\mathcal{O}(\text{nnz}(A) \cdot \log n + \text{poly}(d) \log(1/\epsilon)/\epsilon^2)$ time. Moreover, we can also improve the embedding dimension or equivalently the sample size to $\mathcal{O}(d^{3+p/2} \log(1/\epsilon)/\epsilon^2)$ without increasing the complexity.

Categories and Subject Descriptors

F.2 [ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY]: Numerical Algorithms and Problems

Keywords

subspace embedding; input-sparsity time; low-distortion embedding; linear regression; robust regression; ℓ_p regression

*Most of this work was done while the author was at ICME, Stanford University supported by NSF DMS-1009005.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

STOC'13, June 1-4, 2013, Palo Alto, California, USA.

Copyright 2013 ACM 978-1-4503-2029-0/13/06 ...\$15.00.

1. INTRODUCTION

Regression problems are ubiquitous, and the fast computation of their solutions is of interest in many large-scale data applications. A parameterized family of regression problems that is of particular interest is the *overconstrained ℓ_p regression problem*: given a matrix $A \in \mathbb{R}^{n \times d}$, with $n > d$, a vector $b \in \mathbb{R}^n$, a norm $\|\cdot\|_p$ parameterized by $p \in [1, \infty]$, and an error parameter $\epsilon > 0$, find a $(1 + \epsilon)$ -approximate solution $\hat{x} \in \mathbb{R}^d$ to:

$$f^* = \min_{x \in \mathbb{R}^d} \|Ax - b\|_p,$$

i.e., find a vector \hat{x} such that $\|A\hat{x} - b\|_p \leq (1 + \epsilon)f^*$, where the ℓ_p norm of a vector x is $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$, defined to be $\max_i |x_i|$ for $p = \infty$. Special cases include the ℓ_2 regression problem, also known as the Least Squares problem, and the ℓ_1 regression problem, also known as the Least Absolute Deviations or Least Absolute Errors problem. The latter is of particular interest as a *robust estimation* or *robust regression* technique, in that it is less sensitive to the presence of outliers than the former. We are most interested in this paper in the ℓ_1 regression problem due to its robustness properties, but our methods hold for general $p \in [1, 2]$, and thus we formulate our results in ℓ_p .

It is well-known that for $p \geq 1$, the overconstrained ℓ_p regression problem is a convex optimization problem; for $p = 1$ and $p = \infty$, it is an instance of linear programming; and for $p = 2$, it can be solved with eigenvector-based methods such as with the QR decomposition or the Singular Value Decomposition of A . In spite of their low-degree polynomial-time solvability, ℓ_p regression problems have been the focus in recent years of a wide range of random sampling and random projection algorithms, largely due to a desire to develop improved algorithms for large-scale data applications [3, 24, 10]. For example, Clarkson [9] uses subgradient and sampling methods to compute an approximate solution to the overconstrained ℓ_1 regression problem in roughly $\mathcal{O}(nd^3 \log n)$ time; and Dasgupta et al. [12] use well-conditioned bases and subspace-preserving sampling algorithms to solve general ℓ_p regression problems, for $p \in [1, \infty)$, in roughly $\mathcal{O}(nd^5 \log n)$ time. A similar subspace-preserving sampling algorithm was developed by Drineas, Mahoney, and Muthukrishnan [16] to compute an approximate solution to the ℓ_2 regression problem. The algorithm of [16] relies on the estimation of the ℓ_2 leverage scores¹ of A to be used as an importance sampling distribu-

¹Recall that for an $n \times d$ matrix A , with $n \gg d$, the ℓ_2 leverage scores of the rows of A are equal to the diagonal el-

tion, but when combined with the results of Sarlós [28] and Drineas et al. [17] (that quickly preprocess A to uniformize those scores) or Drineas et al. [15] (that quickly computes approximations to those scores), this leads to a random projection or random sampling (respectively) algorithm for the ℓ_2 regression problem that runs in roughly $\mathcal{O}(nd \log d)$ time [17, 20]. More recently, Sohler and Woodruff [29] introduced the Cauchy Transform to obtain improved ℓ_1 embeddings, thereby leading to an algorithm for the ℓ_1 regression problem that runs in $\mathcal{O}(nd^{1.376+})$ time; and Clarkson et al. [10] use the Fast Cauchy Transform and ellipsoidal rounding methods to compute an approximation to the solution of general ℓ_p regression problems in roughly $\mathcal{O}(nd \log n)$ time.

These algorithms, and in particular the algorithms for $p = 2$, form the basis for much of the large body of recent work in randomized algorithms for low-rank matrix approximation, and thus optimizing their properties can have immediate practical benefits. See, e.g., the recent monograph of Mahoney [20] and references therein for details. Although some of these algorithms are near-optimal for dense inputs, they all require $\Omega(nd \log d)$ time, which can be large if the input matrix is very sparse. Thus, it was a significant result when Clarkson and Woodruff [11] developed an algorithm for the ℓ_2 regression problem (as well as the related problems of low-rank matrix approximation and ℓ_2 leverage score approximation) that runs in *input-sparsity time*, i.e., in $\mathcal{O}(\text{nnz}(A) + \text{poly}(d/\epsilon))$ time, where $\text{nnz}(A)$ is the number of non-zero elements in A and ϵ is an error parameter. This result depends on the construction of a *sparse embedding matrix* Π for ℓ_2 . By this, we mean the following: for an $n \times d$ matrix A , an $s \times n$ matrix Π such that,

$$(1 - \epsilon)\|Ax\|_2 \leq \|\Pi Ax\|_2 \leq (1 + \epsilon)\|Ax\|_2,$$

for all $x \in \mathbb{R}^d$. That is, Π embeds the column space of A into \mathbb{R}^s , while approximately preserving the ℓ_2 norms of all vectors in that subspace. Clarkson and Woodruff achieve their improved results for ℓ_2 -based problems by showing how to construct such a Π with $s = \text{poly}(d/\epsilon)$ and showing that it can be applied to an arbitrary A in $\mathcal{O}(\text{nnz}(A))$ time [11]. (In particular, this embedding result improves the result of Meng, Saunders, and Mahoney [24], who in their development of the parallel least-squares solver LSRN use a result from Davidson and Szarek [14] to construct a constant-distortion embedding for ℓ_2 that runs in $\mathcal{O}(\text{nnz}(A) \cdot d)$ time.) Interestingly, the analysis of Clarkson and Woodruff coupled ideas from the data streaming literature with the structural fact that there cannot be too many high-leverage constraints/rows in A . In particular, they showed that the high-leverage parts of the subspace may be viewed as heavy-hitters that are “perfectly hashed,” and thus contribute no distortion, and that the distortion of the rest of the subspace as well as the “cross terms” may be bounded with a result of Dasgupta, Kumar, and Sarlós [13].

In this paper, we provide improved low-distortion subspace embeddings for ℓ_p , for all $p \in [1, 2]$, in input-sparsity time. We also show that, by coupling with recent work on fast subspace-preserving sampling from [10], these embeddings can be used to provide $(1 + \epsilon)$ -approximate solutions to

elements of the projection matrix onto the span of A . See [20, 15] for details; and note that they can be generalized to ℓ_1 and other ℓ_p norms [10] as well as to arbitrary $n \times d$ matrices, with both n and d large [21, 15].

ℓ_p regression problems, for $p \in [1, 2]$, in nearly input-sparsity time. In more detail, our main results are the following.

First, for ℓ_2 , we obtain an improved result for the input-sparsity time $(1 \pm \epsilon)$ -distortion embedding of [11]. In particular, for the same embedding procedure, we obtain improved bounds for the embedding dimension with a much simpler analysis than [11]. See Theorem 1 of Section 3 for a precise statement of this result. Our analysis is direct and does *not* rely on splitting the high-dimensional space into a set of heavy-hitters consisting of the high-leverage components and the complement of that heavy-hitting set. Since our result directly improves the ℓ_2 embedding result of Clarkson and Woodruff [11], it immediately leads to improvements for the ℓ_2 regression, low-rank matrix approximation, and ℓ_2 leverage score estimation problems that they consider.

Second, for ℓ_1 , we obtain a low-distortion sparse embedding matrix Π such that ΠA can be computed in input-sparsity time. That is, we construct an embedding matrix $\Pi \in \mathbb{R}^{\mathcal{O}(\text{poly}(d)) \times n}$ such that, for all $x \in \mathbb{R}^d$,

$$1/\mathcal{O}(\text{poly}(d)) \cdot \|Ax\|_1 \leq \|\Pi Ax\|_1 \leq \mathcal{O}(\text{poly}(d)) \cdot \|Ax\|_1,$$

with a constant probability, and ΠA can be computed in $\mathcal{O}(\text{nnz}(A))$ time. See Theorem 2 of Section 4 for a precise statement of this result. Here, our proof involves splitting the set $Y = \{Ux \mid \|x\|_\infty = 1, x \in \mathbb{R}^d\}$, where U is an ℓ_1 well-conditioned basis for the span of A , into two parts, informally a subset where coordinates of high ℓ_1 leverage dominate $\|y\|_1$ and the complement of that subset. This ℓ_1 result leads to immediate improvements in ℓ_1 -based problems. For example, by taking advantage of the fast version of subspace-preserving sampling from [10], we can construct and apply a $(1 \pm \epsilon)$ -distortion sparse embedding matrix for ℓ_1 in $\mathcal{O}(\text{nnz}(A) \cdot \log n + \text{poly}(d/\epsilon))$ time. In addition, we can use it to compute a $(1 + \epsilon)$ -approximation to the ℓ_1 regression problem in $\mathcal{O}(\text{nnz}(A) \cdot \log n + \text{poly}(d/\epsilon))$ time, which in turn leads to immediate improvements in ℓ_1 -based matrix approximation objectives, e.g., for the ℓ_1 subspace approximation problem [6, 29, 10].

Third, for ℓ_p , for all $p \in (1, 2)$, we obtain a low-distortion sparse embedding matrix Π such that ΠA can be computed in input-sparsity time. That is, we construct an embedding matrix $\Pi \in \mathbb{R}^{\mathcal{O}(\text{poly}(d)) \times n}$ such that, for all $x \in \mathbb{R}^d$,

$$1/\mathcal{O}(\text{poly}(d)) \cdot \|Ax\|_p \leq \|\Pi Ax\|_p \leq \mathcal{O}(\text{poly}(d)) \cdot \|Ax\|_p,$$

with a constant probability, and ΠA can be computed in $\mathcal{O}(\text{nnz}(A))$ time. See Theorem 4 of Section 5 for a precise statement of this result. Here, our proof generalizes the ℓ_1 result, but we need to prove upper and lower tail bound inequalities for sampling from general p -stable distributions that are of independent interest. Although these distributions don’t have closed forms for $p \in (1, 2)$ in general, we prove that there exists an order among the Cauchy distribution, a p -stable distribution with $p \in (1, 2)$, and the Gaussian distribution such that for all $p \in (1, 2)$ we can use the upper bound from the Cauchy distribution and the lower bound from the Gaussian distribution. As with our ℓ_1 result, this ℓ_p result has several extensions: in $\mathcal{O}(\text{nnz}(A) \cdot \log n + \text{poly}(d/\epsilon))$ time, we can construct and apply a $(1 \pm \epsilon)$ -distortion sparse embedding matrix for ℓ_p ; in $\mathcal{O}(\text{nnz}(A) \cdot \log n + \text{poly}(d/\epsilon))$ time, we can compute a $(1 + \epsilon)$ -approximation to the ℓ_p regression problem; and in $\mathcal{O}(\text{nnz}(A) \cdot d \log d)$ time, we can construct and apply a near-optimal (in terms of embedding dimension and distortion factor) embedding matrix.

The $(1 \pm \epsilon)$ -distortion subspace embedding (for ℓ_p , $p \in [1, 2]$), that we construct from the input-sparsity time embedding and the fast subspace-preserving sampling) has embedding dimension $s = \mathcal{O}(\text{poly}(d) \log(1/\epsilon)/\epsilon^2)$, where the somewhat large $\text{poly}(d)$ term directly multiplies the $\log(1/\epsilon)/\epsilon^2$ term. We can also improve this, showing that it is possible, without increasing the overall complexity, to decouple the large $\text{poly}(d)$ and $\log(1/\epsilon)/\epsilon^2$ via another round of sampling and conditioning, thereby obtaining an embedding dimension that is a small $\text{poly}(d)$ times $\log(1/\epsilon)/\epsilon^2$. See Theorem 7 of Section 6 for a precise statement of this result.

Remark. Subsequent to our posting the first version of this paper on arXiv [23], Clarkson and Woodruff let us know that, independently of us, they used a result from [10] to extend their ℓ_2 subspace embedding from [11] to provide a nearly input-sparsity time algorithm for ℓ_p regression, for all $p \in [1, \infty)$. This is now posted as Version 2 of [11]. Their approach requires solving a rounding problem of size $\mathcal{O}(n/\text{poly}(d)) \times d$, which depends on n (possibly very large). Our approach via input-sparsity time oblivious low-distortion ℓ_p subspace embeddings does not contain this intermediate step and it only needs $\mathcal{O}(\text{poly}(d))$ storage.

Remark. In the first version of this paper, the embedding dimension for ℓ_2 in Theorem 1 was $\mathcal{O}(d^4/\epsilon^2)$. Subsequent to the dissemination of this version, Drineas pointed out to us that our result could very easily be improved to $\mathcal{O}(d^2/\epsilon^2)$. Nelson and Nguyen also let us know that, at about the same time and using the same technique, but independent of us, they first published the $\mathcal{O}(d^2/\epsilon^2)$ embedding result [26].

2. BACKGROUND

We use $\|\cdot\|_p$ to denote the ℓ_p norm of a vector, $\|\cdot\|_2$ the spectral norm of a matrix, $\|\cdot\|_F$ the Frobenius norm of a matrix, and $|\cdot|_p$ the element-wise ℓ_p norm of a matrix. Given $A \in \mathbb{R}^{n \times d}$ with full column rank and $p \in [1, 2]$, we use \mathcal{A}_p to denote the ℓ_p subspace spanned by A 's columns. We are interested in fast embedding of \mathcal{A}_p into a d -dimensional subspace of $(\mathbb{R}^{\text{poly}(d)}, \|\cdot\|_p)$, with distortion either $\text{poly}(d)$ or $(1 \pm \epsilon)$, for some $\epsilon > 0$, as well as applications of this embedding to problems such as ℓ_p regression. We assume that $n \gg \text{poly}(d) \geq d \gg \log n$. To state our results, we assume that we are capable of computing a $(1 + \epsilon)$ -approximate solution to an ℓ_p regression problem of size $n' \times d$ for some $\epsilon > 0$, as long as n' is independent of n . Denote the running time needed to solve this smaller problem by $\mathcal{T}_p(\epsilon; n', d)$. In theory, we have $\mathcal{T}_2(\epsilon; n', d) = \mathcal{O}(n'd \log(d/\epsilon) + d^3)$ (see Drineas et al. [17]), and $\mathcal{T}_p(\epsilon; n', d) = \mathcal{O}((n'd^2 + \text{poly}(d)) \log(n'/\epsilon))$, for general p (see Mitchell [25]).

Conditioning.

The ℓ_p subspace embedding and ℓ_p regression problems are closely related to the concept of conditioning. We state here two related notions of ℓ_p -norm conditioning and then a lemma that characterizes the relationship between them.

DEFINITION 1 ([10]). *Given an $n \times d$ matrix A and $p \in [1, \infty]$, let $\sigma_p^{\max}(A) = \max_{\|x\|_2 \leq 1} \|Ax\|_p$ and let $\sigma_p^{\min}(A) = \min_{\|x\|_2 \geq 1} \|Ax\|_p$. Then, we denote by $\kappa_p(A)$ the ℓ_p -norm condition number of A : $\kappa_p(A) = \sigma_p^{\max}(A)/\sigma_p^{\min}(A)$. For simplicity, we will use κ_p , σ_p^{\min} , and σ_p^{\max} when the underlying matrix is clear.*

DEFINITION 2 ([12]). *Given an $n \times d$ matrix A and $p \in [1, \infty]$, let q be the dual norm of p . Then A is (α, β, p) -conditioned if (1) $|A|_p \leq \alpha$, and (2) for all $z \in \mathbb{R}^d$, $\|z\|_q \leq$*

$\beta \|Az\|_p$. Define $\bar{\kappa}_p(A)$ as the minimum value of $\alpha\beta$ such that A is (α, β, p) -conditioned.

LEMMA 1 ([10]). *Given an $n \times d$ matrix A and $p \in [1, \infty]$: $d^{-|1/2-1/p|} \kappa_p(A) \leq \bar{\kappa}_p(A) \leq d^{\max\{1/2, 1/p\}} \kappa_p(A)$.*

Remark. Given the equivalence established by Lemma 1, we will say that A is *well-conditioned in the ℓ_p norm* if $\kappa_p(A)$ or $\bar{\kappa}_p(A) = \mathcal{O}(\text{poly}(d))$, independent of n .

Although for an arbitrary matrix $A \in \mathbb{R}^{n \times d}$, the condition numbers $\kappa_p(A)$ and $\bar{\kappa}_p(A)$ can be arbitrarily large, we can find a matrix $R \in \mathbb{R}^{d \times d}$ such that AR^{-1} is well-conditioned. This procedure is called *conditioning*, and there exist two approaches for conditioning: via low-distortion ℓ_p subspace embedding and via ellipsoidal rounding.

DEFINITION 3. *Given an $n \times d$ matrix A and a number $p \in [1, \infty]$, $\Pi \in \mathbb{R}^{s \times n}$ is a low-distortion embedding of \mathcal{A}_p if $s = \mathcal{O}(\text{poly}(d))$ and $\forall x \in \mathbb{R}^d$:*

$$1/\mathcal{O}(\text{poly}(d)) \cdot \|Ax\|_p \leq \|\Pi Ax\|_p \leq \mathcal{O}(\text{poly}(d)) \cdot \|Ax\|_p.$$

Remark. Given a low-distortion embedding matrix Π of \mathcal{A}_p , let R be the ‘‘R’’ matrix from the QR decomposition of ΠA . Then, AR^{-1} is well-conditioned in the ℓ_p norm.

For a discussion of ellipsoidal rounding, we refer readers to Clarkson et al. [10]. In this paper, we simply cite the following lemma, which is based on ellipsoidal rounding.

LEMMA 2 ([10]). *Given an $n \times d$ matrix A and $p \in [1, \infty]$, it takes at most $\mathcal{O}(nd^3 \log n)$ time to find a matrix $R \in \mathbb{R}^{d \times d}$ such that $\kappa_p(AR^{-1}) \leq 2d$.*

Subspace-preserving sampling and ℓ_p regression.

Given $R \in \mathbb{R}^{d \times d}$ such that AR^{-1} is well-conditioned in the ℓ_p norm, we can construct a $(1 \pm \epsilon)$ -distortion embedding, specifically a subspace-preserving sampling, of \mathcal{A}_p in $\mathcal{O}(\text{nnz}(A) \cdot \log n)$ additional time and with a constant probability. This result from Clarkson et al. [10, Theorem 5.4] improves the subspace-preserving sampling algorithm proposed by Dasgupta et al. [12] by estimating the row norms of AR^{-1} (instead of computing them exactly) to define importance sampling probabilities.

LEMMA 3 ([10]). *Given a matrix $A \in \mathbb{R}^{n \times d}$, $p \in [1, \infty)$, $\epsilon > 0$, and a matrix $R \in \mathbb{R}^{d \times d}$ such that AR^{-1} is well-conditioned, it takes $\mathcal{O}(\text{nnz}(A) \cdot \log n)$ time to compute a sampling matrix $S \in \mathbb{R}^{s \times n}$ (with only one nonzero element per row) with $s = \mathcal{O}(\bar{\kappa}_p^p(AR^{-1}) d^{|p/2-1|+1} \log(1/\epsilon)/\epsilon^2)$ such that with a constant probability,*

$$(1 - \epsilon) \|Ax\|_p \leq \|SAx\|_p \leq (1 + \epsilon) \|Ax\|_p, \quad \forall x \in \mathbb{R}^d.$$

Given a subspace-preserving sampling algorithm, Clarkson et al. [10, Theorem 5.4] show it is straightforward to compute a $\frac{1+\epsilon}{1-\epsilon}$ -approximate solution to an ℓ_p regression problem.

LEMMA 4 ([10]). *Given an ℓ_p regression problem specified by $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$, and $p \in [1, \infty)$, let S be a $(1 \pm \epsilon)$ -distortion embedding matrix of the subspace spanned by A 's columns and b from Lemma 3, and let \hat{x} be an optimal solution to the subsampled problem $\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_p$. Then \hat{x} is a $\frac{1+\epsilon}{1-\epsilon}$ -approximate solution to the original problem.*

Remark. Collecting these results, we see that low-distortion ℓ_p subspace embedding is a fundamental building block (and very likely a bottleneck) for $(1 \pm \epsilon)$ -distortion ℓ_p subspace embeddings, as well as for a $(1 + \epsilon)$ -approximation to an ℓ_p regression problem. This motivates our work and its emphasis on finding low-distortion subspace embeddings.

Stable distributions.

We use properties of p -stable distributions for analyzing input-sparsity time low-distortion ℓ_p subspace embeddings.

DEFINITION 4. A distribution \mathcal{D} over \mathbb{R} is called p -stable, if for any m real numbers a_1, \dots, a_m , we have

$$\sum_{i=1}^m a_i X_i \simeq \left(\sum_{i=1}^m |a_i|^p \right)^{1/p} X,$$

where $X_i \stackrel{iid}{\sim} \mathcal{D}$ and $X \sim \mathcal{D}$. By “ $X \simeq Y$ ”, we mean X and Y have the same distribution.

By Lévy [19], it is known that p -stable distributions exist for $p \in (0, 2]$; and from Chambers et al. [7], it is known that p -stable random variables can be generated efficiently, thus allowing their practical use. Let us use \mathcal{D}_p to denote the “standard” p -stable distribution, for $p \in [1, 2]$, specified by its characteristic function $\psi(t) = e^{-|t|^p}$. It is known that \mathcal{D}_1 is the standard Cauchy distribution, and that \mathcal{D}_2 is the Gaussian distribution with mean 0 and variance 2.

Tail inequalities.

We note two inequalities from Clarkson et al. [10] regarding the tails of the Cauchy distribution.

LEMMA 5. For $i = 1, \dots, m$, let C_i be m (not necessarily independent) standard Cauchy variables, and $\gamma_i > 0$ with $\gamma = \sum_i \gamma_i$. Let $X = \sum_i \gamma_i |C_i|$. For any $t > 1$,

$$\Pr[X > t\gamma] \leq \frac{1}{\pi t} \left(\frac{\log(1 + (2mt)^2)}{1 - 1/(\pi t)} + 1 \right).$$

For simplicity, we assume that $m \geq 3$ and $t \geq 1$, and then we have $\Pr[X > t\gamma] \leq 2 \log(mt)/t$.

LEMMA 6. For $i = 1, \dots, m$, let C_i be independent standard Cauchy random variables, and $\gamma_i \geq 0$ with $\gamma = \sum_i \gamma_i$. Let $X = \sum_i \gamma_i |C_i|$. Then, for any $t > 0$,

$$\log \Pr[X \leq (1-t)\gamma] \leq \frac{-\gamma t^2}{3 \max_i \gamma_i}.$$

The following result about Gaussian variables is a direct consequence of Maurer’s inequality ([22]), and we will use it to derive lower tail inequalities for p -stable distributions.

LEMMA 7. For $i = 1, \dots, m$, let G_i be independent standard Gaussian random variables, and $\gamma_i \geq 0$ with $\gamma = \sum_i \gamma_i$. Let $X = \sum_i \gamma_i |G_i|^2$. Then, for any $t > 0$,

$$\log \Pr[X \leq (1-t)\gamma] \leq \frac{-\gamma t^2}{6 \max_i \gamma_i}.$$

3. MAIN RESULTS FOR ℓ_2 EMBEDDING

Here is our result for input-sparsity time low-distortion subspace embeddings for ℓ_2 . See also Nelson and Nguyen [26] for a similar result with a slightly better constant.

THEOREM 1. Given a matrix $A \in \mathbb{R}^{n \times d}$ and $\epsilon \in (0, 1)$, let $\Pi = SD$ where $S \in \mathbb{R}^{s \times n}$ has each column chosen independently and uniformly from the s standard basis vectors of \mathbb{R}^s and $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonal entries chosen independently and uniformly from ± 1 . Let $s = (d^2 + d)/(\epsilon^2 \delta)$. Then with probability at least $1 - \delta$,

$$(1 - \epsilon) \|Ax\|_2 \leq \|\Pi Ax\|_2 \leq (1 + \epsilon) \|Ax\|_2, \quad \forall x \in \mathbb{R}^d.$$

In addition, ΠA can be computed in $\mathcal{O}(\text{nnz}(A))$ time.

The construction of Π in this theorem is the same as in Clarkson and Woodruff [11]. There, $s = \mathcal{O}((d/\epsilon)^4 \log^2(d/\epsilon))$ in order to achieve $(1 \pm \epsilon)$ distortion with a constant probability. Theorem 1 shows that it actually suffices to set $s = \mathcal{O}((d^2 + d)/\epsilon^2)$. Surprisingly, the proof is rather simple. Let $X = U^T \Pi^T \Pi U$, where U is an orthonormal basis for \mathcal{A}_2 . Compute $\mathbf{E}[\|X - I\|_F^2]$ and apply Markov’s inequality to $\|X - I\|_F^2 \leq \epsilon^2$, which implies $\|X - I\|_2 \leq \epsilon$ and hence the embedding result. See Appendix A.1 for a complete proof.

Remark. The $\mathcal{O}(\text{nnz}(A))$ running time is indeed optimal, up to constant factors, for general inputs. Consider the case when A has an important row a_j such that A becomes rank-deficient without it. Thus, we have to observe a_j in order to compute a low-distortion embedding. However, without any prior knowledge, we have to scan at least a constant portion of the input to guarantee that a_j is observed with a constant probability, which takes $\mathcal{O}(\text{nnz}(A))$ time. Note that this optimality result applies to general p .

The results of Theorem 1 propagate to related applications, e.g., to the ℓ_2 regression problem, the low-rank matrix approximation problem and the problem of computing approximations to the ℓ_2 leverage scores. Since it underlies the other applications, only the ℓ_2 regression improvement is stated here explicitly; its proof is basically combining our Theorem 1 with Theorem 19 of [11].

COROLLARY 1. With a constant probability, a $(1 + \epsilon)$ -approximate solution to an ℓ_2 regression problem can be computed in $\mathcal{O}(\text{nnz}(A) + \mathcal{T}_2(\epsilon; d^2/\epsilon^2, d))$ time.

Remark. Although our simpler direct proof leads to a better result for ℓ_2 subspace embedding, the technique used in the proof of Clarkson and Woodruff [11], which splits coordinates into “heavy” and “light” sets based on the leverage scores, highlights an important structural property of ℓ_2 subspace: that only a small subset of coordinates can have large ℓ_2 leverage scores. (We note that the technique of splitting coordinates is also used by Ailon and Liberty [1] to get an unrestricted fast Johnson-Lindenstrauss transform; and that the difficulty in finding and approximating the large-leverage directions was—until recently [20, 15]—responsible for difficulties in obtaining fast relative-error random sampling algorithms for ℓ_2 regression and low-rank matrix approximation.) An analogous structural fact holds for ℓ_1 and other ℓ_p spaces. Using this property, we can construct novel input-sparsity time ℓ_p subspace embeddings for general $p \in [1, 2)$, as we discuss in the next two sections.

4. MAIN RESULTS FOR ℓ_1 EMBEDDING

Here is our result for input-sparsity time low-distortion subspace embeddings for ℓ_1 .

THEOREM 2. Given $A \in \mathbb{R}^{n \times d}$, let $\Pi = SC \in \mathbb{R}^{s \times n}$, where $S \in \mathbb{R}^{s \times n}$ has each column chosen independently and uniformly from the s standard basis vectors of \mathbb{R}^s , and where $C \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonals chosen independently from the standard Cauchy distribution. Set $s = \omega d^5 \log^5 d$ with ω sufficiently large. Then with a constant probability, we have $\forall x \in \mathbb{R}^d$:

$$1/\mathcal{O}(d^2 \log^2 d) \cdot \|Ax\|_1 \leq \|\Pi Ax\|_1 \leq \mathcal{O}(d \log d) \cdot \|Ax\|_1.$$

In addition, ΠA can be computed in $\mathcal{O}(\text{nnz}(A))$ time.

The construction of the ℓ_1 subspace embedding matrix is different than its ℓ_2 norm counterpart only by the diagonal elements of D (or C): whereas we use ± 1 for the ℓ_2

norm, we use Cauchy variables for the ℓ_1 norm. The proof of Theorem 2 uses the technique of splitting coordinates, the fact that the Cauchy distribution is 1-stable, and the upper and lower tail inequalities from Lemmas 5 and 6. See Appendix A.2 for a complete proof.

Remark. As mentioned above, the $\mathcal{O}(\text{nnz}(A))$ running time is optimal. Whether the distortion $\mathcal{O}(d^3 \log^3 d)$ is optimal is still an open question. However, for the same construction of Π , we can provide a “bad” case that provides a lower bound. Choose $A = (I_d \ \mathbf{0})^T$. Suppose that s is sufficiently large such that with an overwhelming probability, the top d rows of A are perfectly hashed, i.e., $\|\Pi A x\|_1 = \sum_{k=1}^d |c_k| |x_k|$, $\forall x \in \mathbb{R}^d$, where c_k is the k -th diagonal of C . Then, the distortion of Π is $\max_{k \leq d} |c_k| / \min_{k \leq d} |c_k| \approx \mathcal{O}(d^2)$. Therefore, at most an $\mathcal{O}(d \log^3 d)$ factor of the distortion is due to artifacts in our analysis.

Our input-sparsity time ℓ_1 subspace embedding of Theorem 2 improves the $\mathcal{O}(\text{nnz}(A) \cdot d \log d)$ -time embedding by Sohler and Woodruff [29] and the $\mathcal{O}(nd \log n)$ -time embedding of Clarkson et al. [10]. In addition, by combining Theorem 2 and Lemma 3, we can compute a $(1 \pm \epsilon)$ -distortion embedding in nearly input-sparsity time.

THEOREM 3. *Given $A \in \mathbb{R}^{n \times d}$, it takes $\mathcal{O}(\text{nnz}(A) \cdot \log n)$ time to compute a sampling matrix $S \in \mathbb{R}^{s \times n}$ with $s = \mathcal{O}(\text{poly}(d) \log(1/\epsilon)/\epsilon^2)$ such that with a constant probability, S embeds \mathcal{A}_1 into $(\mathbb{R}^s, \|\cdot\|_1)$ with distortion $1 \pm \epsilon$.*

Our improvements in Theorems 2 and 3 also propagate to related ℓ_1 -based applications, including the ℓ_1 regression and the ℓ_1 subspace approximation problem considered in [29, 10]. As before, only the regression improvement is stated here explicitly. For completeness, we present in Algorithm 1 our algorithm for solving ℓ_1 regression problems in nearly input-sparsity time. The brief proof of Corollary 2, our main quality-of-approximation result for Algorithm 1, may be found in Appendix A.3.

Algorithm 1 Fast ℓ_1 Regression Approximation in $\mathcal{O}(\text{nnz}(A) \cdot \log n + \text{poly}(d) \log(1/\epsilon)/\epsilon^2)$ Time

Input: $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$, and $\epsilon \in (0, 1/2)$.

Output: A $(1 + \epsilon)$ -approximation solution \hat{x} to $\min_{x \in \mathbb{R}^d} \|Ax - b\|_1$, with a constant probability.

- 1: Let $\bar{A} = (A \ b)$ and denote $\bar{\mathcal{A}}_1$ the ℓ_1 subspace spanned by A 's columns and b .
 - 2: Compute a low-distortion embedding $\Pi \in \mathbb{R}^{\mathcal{O}(\text{poly}(d)) \times n}$ of $\bar{\mathcal{A}}_1$ (Theorem 2).
 - 3: Compute $\bar{R} \in \mathbb{R}^{(d+1) \times (d+1)}$ from $\Pi \bar{A}$ such that $\bar{A} \bar{R}^{-1}$ is well-conditioned (QR or Lemma 2).
 - 4: Compute a $(1 \pm \epsilon/4)$ -distortion embedding matrix $S \in \mathbb{R}^{\mathcal{O}(\text{poly}(d) \log(1/\epsilon)/\epsilon^2) \times n}$ of $\bar{\mathcal{A}}_1$ (Lemma 3).
 - 5: Compute a $(1 + \epsilon/4)$ -approximate solution \hat{x} to $\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_1$.
-

COROLLARY 2. *With a constant probability, Algorithm 1 computes a $(1 + \epsilon)$ -approximation to an ℓ_1 regression problem in $\mathcal{O}(\text{nnz}(A) \cdot \log n + \mathcal{T}_1(\epsilon; \text{poly}(d) \log(1/\epsilon)/\epsilon^2, d))$ time.*

Remark. For readers familiar with the impossibility results for dimension reduction in ℓ_1 [8, 18, 5], note that those results apply to arbitrary point sets of size n and are interested in embeddings that are “oblivious,” in that they do not

depend on the input data. In this paper, we only consider points in a subspace, and the subspace-preserving sampling procedure of [12] that we use is data-dependent.

5. MAIN RESULTS FOR ℓ_p EMBEDDING

In this section, we use the properties of p -stable distributions to generalize the input-sparsity time ℓ_1 subspace embedding to ℓ_p norms, for $p \in (1, 2)$. Generally, \mathcal{D}_p does not have explicit PDF/CDF, which increases the difficulty for theoretical analysis. Indeed, the main technical difficulty here is that we are not aware of ℓ_p analogues of Lemmas 5 and 6 that would provide upper and lower tail inequality for p -stable distributions. (Indeed, even Lemmas 5 and 6 were established only recently [10].)

Instead of analyzing \mathcal{D}_p directly, for any $p \in (1, 2)$, we establish an order among the Cauchy distribution, the p -stable distribution, and the Gaussian distribution, and then we derive upper and lower tail inequalities for the p -stable distribution similar to the ones we used to prove Theorem 2. We state these technical results here since they are of independent interest. We start with the following lemma, which is proved in Appendix A.4 and which establishes this order.

LEMMA 8. *For any $p \in (1, 2)$, there exist constants $\alpha_p > 0$ and $\beta_p > 0$ such that*

$$\alpha_p |C| \succeq |X_p|^p \succeq \beta_p |G|^2,$$

where C is a standard Cauchy variable, $X_p \sim \mathcal{D}_p$, G is a standard Gaussian variable. By “ $X \succeq Y$ ” we mean $\Pr[X \geq t] \geq \Pr[Y \geq t]$, $\forall t \in \mathbb{R}$.

Our numerical results suggest that the constants α_p and β_p are not too far away from 1. See [23] for more details.

Lemma 8 suggests that we can use Lemma 5 (regarding Cauchy random variables) to derive upper tail inequalities for general p -stable distributions and that we can use Lemma 7 (regarding Gaussian variables) to derive lower tail inequalities for general p -stable distributions. The following two lemmas establish these results; the proofs of these lemmas are provided in Appendixes A.5 and A.6, respectively.

LEMMA 9. *Given $p \in (1, 2)$, for $i = 1, \dots, m$, let X_i be m (not necessarily independent) random variables sampled from \mathcal{D}_p , and $\gamma_i > 0$ with $\gamma = \sum_i \gamma_i$. Let $X = \sum_i \gamma_i |X_i|^p$. Assume that $m \geq 3$. Then for any $t \geq 1$,*

$$\Pr[X \geq t \alpha_p \gamma] \leq \frac{2 \log(mt)}{t}.$$

LEMMA 10. *For $i = 1, \dots, m$, let X_i be independent random variables sampled from \mathcal{D}_p , and $\gamma_i \geq 0$ with $\gamma = \sum_i \gamma_i$. Let $X = \sum_i \gamma_i |c_i|$. Then,*

$$\log \Pr[X \leq (1 - t) \beta_p \gamma] \leq \frac{-\gamma t^2}{6 \max_i \gamma_i}.$$

Given these results, here is our main result for input-sparsity time low-distortion subspace embeddings for ℓ_p . The proof of this theorem is similar to the proof of Theorem 2, except that we replace the ℓ_1 norm $\|\cdot\|_1$ by $\|\cdot\|_p^p$ and use Lemmas 9 and 10 (rather than Lemmas 5 and 6).

THEOREM 4. *Given $A \in \mathbb{R}^{n \times d}$ and $p \in (1, 2)$, let $\Pi = SD \in \mathbb{R}^{s \times n}$ where $S \in \mathbb{R}^{s \times n}$ has each column chosen independently and uniformly from the s standard basis vectors of \mathbb{R}^s , and where $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonals chosen independently from \mathcal{D}_p . Set $s = \omega d^5 \log^5 d$*

with ω sufficiently large. Then with a constant probability, we have $\forall x \in \mathbb{R}^d$:

$$1/\mathcal{O}((d \log d)^{2/p}) \|Ax\|_p \leq \|\Pi Ax\|_p \leq \mathcal{O}((d \log d)^{1/p}) \|Ax\|_p.$$

In addition, ΠA can be computed in $\mathcal{O}(\text{nnz}(A))$ time.

Then, by combining Theorem 4 and Lemma 3, we can compute a $(1 \pm \epsilon)$ -distortion embedding in $\mathcal{O}(\text{nnz}(A) \cdot \log n)$ time.

THEOREM 5. *Given $A \in \mathbb{R}^{n \times d}$ and $p \in (1, 2)$, it takes $\mathcal{O}(\text{nnz}(A) \cdot \log n)$ time to find a sampling matrix $S \in \mathbb{R}^{s \times n}$ with $s = \mathcal{O}(\text{poly}(d) \log(1/\epsilon)/\epsilon^2)$ such that with a constant probability, S embeds \mathcal{A}_p into $(\mathbb{R}^s, \|\cdot\|_p)$ with distortion $1 \pm \epsilon$.*

Similar to the ℓ_1 case, Theorem 5 improves the $\mathcal{O}(nd \log n)$ -time embedding of Clarkson et al. [10]. As we mentioned in Section 1, their construction (and hence the construction of [11]) works for all $p \in [1, \infty)$, but it requires solving a rounding problem of size $\mathcal{O}(n/\text{poly}(d)) \times d$ as an intermediate step, which may become intractable when n is very large or in a streaming environment, while our construction only needs $\mathcal{O}(\text{poly}(d))$ storage. Our improvements for ℓ_p subspace embedding also propagate to related ℓ_p -based applications. In particular, we can establish an improved algorithm for solving the ℓ_p regression problem in nearly input-sparsity time.

COROLLARY 3. *Given $p \in (1, 2)$, with a constant probability, it takes $\mathcal{O}(\text{nnz}(A) \cdot \log n + \mathcal{T}_p(\epsilon; \text{poly}(d) \log(1/\epsilon)/\epsilon^2, d))$ time to compute a $(1 + \epsilon)$ -approximate solution to an ℓ_p regression problem.*

For completeness, we also present a result for low-distortion dense embeddings for ℓ_p that the tail inequalities from Lemmas 9 and 10 enable us to construct. See Appendix A.7 for a proof of the following theorem.

THEOREM 6. *Given $A \in \mathbb{R}^{n \times d}$ with full column rank and $p \in (1, 2)$, let $\Pi \in \mathbb{R}^{s \times n}$ whose entries are i.i.d. samples from \mathcal{D}_p . If $s = \omega d \log d$ for ω sufficiently large, with a constant probability, we have $\forall x \in \mathbb{R}^d$:*

$$1/\mathcal{O}(1) \cdot \|Ax\|_p \leq \|\Pi Ax\|_p \leq \mathcal{O}((d \log d)^{1/p}) \cdot \|Ax\|_p.$$

In addition, ΠA can be computed in $\mathcal{O}(\text{nnz}(A) \cdot d \log d)$ time.

Remark. The result in Theorem 6 is based on a dense ℓ_p subspace embeddings that is analogous to the dense Gaussian embedding for ℓ_2 and the dense Cauchy embedding of [29] for ℓ_1 . Although the running time (if one is simply interested in FLOP counts in RAM) of Theorem 6 is somewhat worse than that of Theorem 4, the embedding dimension and condition number quality (the ratio of the upper bound on the distortion and the lower bound on the distortion) are much better. Our numerical implementations, both with the ℓ_1 norm [10] and with the ℓ_2 norm [24], strongly suggest that the latter quantities are more important to control when implementing randomized regression algorithms in large-scale parallel and distributed settings.

6. IMPROVED EMBEDDING DIMENSION

In Theorem 2 and Theorem 4, the embedding dimension is $s = \mathcal{O}(\text{poly}(d) \log(1/\epsilon)/\epsilon^2)$, where the $\text{poly}(d)$ term is a somewhat large polynomial of d that directly multiplies the $\log(1/\epsilon)/\epsilon^2$ term. (See the remark below for comments on the precise value of the $\text{poly}(d)$ term.) This is not ideal for the subspace embedding and the ℓ_p regression, because

we want to have a small embedding dimension and a small subsampled problem, respectively. Here, we show that it is possible to decouple the large polynomial of d and the $\log(1/\epsilon)/\epsilon^2$ term via another round of sampling and conditioning without increasing the complexity. See Algorithm 2 for details on this procedure. Theorem 7 provides our main quality-of-approximation result for Algorithm 2; its proof can be found in Appendix A.8.

Algorithm 2 Improving the Embedding Dimension

Input: $A \in \mathbb{R}^{n \times d}$, $p \in [1, 2)$, and $\epsilon \in (0, 1)$.

Output: A $(1 \pm \epsilon)$ -distortion embedding matrix $S \in \mathbb{R}^{\mathcal{O}(d^{3+p/2} \log(1/\epsilon)/\epsilon^2) \times n}$ of \mathcal{A}_p .

- 1: Compute a low-distortion embedding $\tilde{\Pi} \in \mathbb{R}^{\mathcal{O}(\text{poly}(d)) \times n}$ of \mathcal{A}_p (Theorems 2 and 4).
 - 2: Compute $\tilde{R} \in \mathbb{R}^{d \times d}$ from $\tilde{\Pi} A$ such that $A \tilde{R}^{-1}$ is well-conditioned (QR or Lemma 2).
 - 3: Compute a $(1 \pm 1/2)$ -distortion embedding matrix $\tilde{S} \in \mathbb{R}^{\mathcal{O}(\text{poly}(d) \times n)}$ of \mathcal{A}_p (Lemma 3).
 - 4: Compute $R \in \mathbb{R}^{d \times d}$ such that $\kappa_p(\tilde{S} A R^{-1}) \leq 2d$ (Theorem 2).
 - 5: Compute a $(1 \pm \epsilon)$ -distortion embedding matrix $S \in \mathbb{R}^{\mathcal{O}(d^{3+p/2} \log(1/\epsilon)/\epsilon^2) \times n}$ of \mathcal{A}_p (Lemma 3).
-

THEOREM 7. *Algorithm 2 computes a $(1 \pm \epsilon)$ -distortion embedding of \mathcal{A}_p into $(\mathbb{R}^{\mathcal{O}(d^{3+p/2} \log(1/\epsilon)/\epsilon^2)}, \|\cdot\|_p)$ in time $\mathcal{O}(\text{nnz}(A) \cdot \log n)$ with a constant probability.*

Then, by applying Theorem 7 to the ℓ_p regression problem, we can improve the size of the subsampled problem and hence the overall running time.

COROLLARY 4. *Given $p \in [1, 2)$, with a constant probability, a $(1 + \epsilon)$ -approximate solution to an ℓ_p regression problem can be computed in*

$$\mathcal{O}(\text{nnz}(A) \cdot \log n + \mathcal{T}_p(\epsilon; d^{3+p/2} \log(1/\epsilon)/\epsilon^2, d))$$

time. The second term comes from solving a subsampled problem of size $\mathcal{O}(d^{3+p/2} \log(1/\epsilon)/\epsilon^2) \times d$.

Remark. We have stated our results in the previous sections as $\text{poly}(d)$ without stating the value of the polynomial because there are numerous trade-offs between the conditioning quality and the running time. For example, let $p = 1$. We can use a rounding algorithm instead of QR to compute the R matrix. If we use the input-sparsity time embedding with the $\mathcal{O}(d)$ -rounding algorithm of [10], then the running time to compute the $(1 \pm \epsilon)$ -distortion embedding is $\mathcal{O}(\text{nnz}(A) \cdot \log n + d^8/\epsilon^2)$ and the embedding dimension is $\mathcal{O}(d^{6.5}/\epsilon^2)$ (ignoring log factors). If, on the other hand, we use QR to compute R , then the running time is $\mathcal{O}(\text{nnz}(A) \cdot \log n + d^7/\epsilon^2)$ and the embedding dimension is $\mathcal{O}(d^8/\epsilon^2)$. However, with the result from this section, the running time is simply $\mathcal{O}(\text{nnz}(A) \cdot \log n + \text{poly}(d) + \mathcal{T}_p(\epsilon; d^{3+p/2}/\epsilon^2, d))$ and the $\text{poly}(d)$ term can be absorbed by the $\text{nnz}(A)$ term.

7. ACKNOWLEDGMENTS

The authors want to thank P. Drineas for reading the first version of this paper and pointing out that the embedding dimension in Theorem 1 can be easily improved from $\mathcal{O}(d^4/\epsilon^2)$ to $\mathcal{O}(d^2/\epsilon^2)$ using the same technique. The authors also want to thank J. Nelson and H. Nguyen for letting us know about their independent work on ℓ_2 embedding.

8. REFERENCES

- [1] N. Ailon and E. Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. In *Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 185–191, 2011.
- [2] H. Auerbach. *On the area of convex curves with conjugate diameters*. PhD thesis, University of Lwów, 1930.
- [3] H. Avron, P. Maymounkov, and S. Toledo. Blendepik: Supercharging LAPACK’s least-squares solver. *SIAM Journal on Scientific Computing*, 32:1217–1236, 2010.
- [4] J. Bourgain, J. Lindenstrauss, and V. Milman. Approximation of zonoids by zonotopes. *Acta Mathematica*, 162:73–141, 1989.
- [5] B. Brinkman and M. Charikar. On the impossibility of dimension reduction in ℓ_1 . *Journal of the ACM*, 52(5):766–788, 2005.
- [6] J. P. Brooks and J. H. Dulá. The L1-norm best-fit hyperplane problem. *Applied Mathematics Letters*, 26(1):51–55, 2013.
- [7] J. M. Chambers, C. L. Mallows, and B. W. Stuck. A method for simulating stable random variables. *Journal of the American Statistical Association*, 71(354):340–344, 1976.
- [8] M. Charikar and A. Sahai. Dimension reduction in the ℓ_1 norm. In *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, pages 551–560, 2002.
- [9] K. Clarkson. Subgradient and sampling algorithms for ℓ_1 regression. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 257–266, 2005.
- [10] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Meng, and D. P. Woodruff. The Fast Cauchy Transform and faster robust linear regression. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 466–477, 2013.
- [11] K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. Technical report. Preprint: arXiv:1207.6365 (2012). To appear in STOC’13.
- [12] A. Dasgupta, P. Drineas, B. Harb, R. Kumar, and M. W. Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM Journal on Computing*, (38):2060–2078, 2009.
- [13] A. Dasgupta, R. Kumar, and T. Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing*, pages 341–350, 2010.
- [14] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces*, volume 1, pages 317–366. North Holland, 2001.
- [15] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.
- [16] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1127–1136, 2006.
- [17] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2010.
- [18] J. R. Lee and A. Naor. Embedding the diamond graph in L_p and dimension reduction in L_1 . *Geometric And Functional Analysis*, 14(4):745–747, 2004.
- [19] P. Lévy. *Calcul des Probabilités*. Gauthier-Villars, Paris, 1925.
- [20] M. W. Mahoney. *Randomized Algorithms for Matrices and Data*. Foundations and Trends in Machine Learning. NOW Publishers, Boston, 2011.
- [21] M. W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci. USA*, 106:697–702, 2009.
- [22] A. Maurer. A bound on the deviation probability for sums of non-negative random variables. *J. Inequalities in Pure and Applied Mathematics*, 4(1), 2003.
- [23] X. Meng and M. W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. Technical report. Preprint: arXiv:1210.3135 (2012).
- [24] X. Meng, M. A. Saunders, and M. W. Mahoney. LSRN: A parallel iterative solver for strongly over- or under-determined systems. Technical report. Preprint: arXiv:1109.5981 (2011).
- [25] J. E. Mitchell. Polynomial interior point cutting plane methods. *Optimization Methods and Software*, 18(5):507–534, 2003.
- [26] J. Nelson and H. Nguyen. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. *arXiv preprint arXiv:1211.1002*, 2012.
- [27] J. P. Nolan. *Stable Distributions - Models for Heavy Tailed Data*. Birkhauser, Boston, 2013. In progress, Chapter 1 online at academic2.american.edu/~jpnolan.
- [28] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 143–152, 2006.
- [29] C. Sohler and D. P. Woodruff. Subspace embeddings for the ℓ_1 -norm with applications. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, pages 755–764, 2011.

APPENDIX

A.1 Proof of Theorem 1

Let the $n \times d$ matrix U be an orthonormal basis for the range of the $n \times d$ matrix A . Rather than proving the theorem by establishing that

$$(1 - \epsilon)\|Uz\|_2 \leq \|\Pi Uz\|_2 \leq (1 + \epsilon)\|Uz\|_2$$

holds for all $z \in \mathbb{R}^d$, as is essentially done in, e.g., [16] and [11], we note that $U^T U = I_d$, and we directly bound the extent to which the embedding process perturbs this product. To do so, define $X = (\Pi U)^T (\Pi U) = U^T D^T S^T S D U$, i.e.,

$$x_{kl} = \sum_{i=1}^s \left(\sum_{j=1}^n s_{ij} d_j u_{jk} \right) \left(\sum_{j=1}^n s_{ij} d_j u_{jl} \right), \quad k, l \in \{1, \dots, d\},$$

where s_{ij} is the (i, j) -th element of S , d_j is the j -th diagonal element of D , and u_{jk} is the (j, k) -th element of U . We will use the following facts in the proof:

$$\mathbf{E}[d_{j_1} d_{j_2}] = \delta_{j_1 j_2},$$

$$\mathbf{E}[s_{i_1 j_1} s_{i_2 j_2}] = \begin{cases} \frac{1}{s^2} & \text{if } j_1 \neq j_2, \\ \frac{1}{s} & \text{if } i_1 = i_2, j_1 = j_2, \\ 0 & \text{if } i_1 \neq i_2, j_1 = j_2. \end{cases}$$

Given these, it is easy to show that $\mathbf{E}[x_{kl}] = \delta_{kl}$ and that

$$\mathbf{E}[x_{kl}^2] = \begin{cases} 1 + \frac{2}{s}(1 - \|U_{*k}\|_4^4) & \text{if } k = l, \\ \frac{1}{s}(1 - 2\langle U_{*k}^2, U_{*l}^2 \rangle) & \text{if } k \neq l. \end{cases}$$

Given these results, it is easy to obtain that

$$\begin{aligned} \mathbf{E}[\|X - I\|_F^2] &= \sum_{k,l} \mathbf{E}[(x_{kl} - \delta_{kl})^2] \\ &= \frac{2}{s} \left(\sum_k (1 - \|U_{*k}\|_4^4) + \sum_{k < l} (1 - 2\langle U_{*k}^2, U_{*l}^2 \rangle) \right) \leq \frac{d^2 + d}{s}. \end{aligned}$$

For any $\delta \in (0, 1)$, set $s = (d^2 + d)/(\epsilon^2 \delta)$. Then, by Markov's inequality,

$$\Pr[\|X - I\|_F \geq \epsilon] = \Pr[\|X - I\|_F^2 \geq \epsilon^2] \leq \frac{d^2 + d}{\epsilon^2 s} = \delta.$$

Therefore, with probability at least $1 - \delta$, we have $\|X - I\|_2 \leq \|X - I\|_F \leq \epsilon$, which implies

$$(1 - \epsilon)\|Uz\|_2 \leq \|\Pi Uz\|_2 \leq (1 + \epsilon)\|Uz\|_2.$$

A.2 Proof of Theorem 2

LEMMA 11. (Auerbach [2]) Let $(\mathcal{A}, \|\cdot\|)$ be a d -dimensional normed vector space. There exists a basis $\{e_1, \dots, e_d\}$ of \mathcal{A} , called Auerbach basis, such that $\|e_k\| = 1$ and $\|e^k\|^* = 1$ for $k = 1, \dots, d$, where $\{e^1, \dots, e^n\}$ is a basis of \mathcal{A}^* dual to $\{e_1, \dots, e_n\}$.

This Auerbach's lemma implies that a $(d, 1, 1)$ -conditioned basis matrix of \mathcal{A}_1 exists, which will be denoted by U throughout the proof. By definition, U 's columns are unit vectors in the ℓ_1 norm (thus $\|U\|_1 = d$, where recall that $\|\cdot\|_1$ denotes the element-wise ℓ_1 norm of a matrix) and $\|x\|_\infty \leq \|Ux\|_1, \forall x \in \mathbb{R}^d$. Denote by u_j the j -th row of U , $j = 1, \dots, n$. Define $v_j = \|u_j\|_1$ the ℓ_1 leverage scores of A . We have $\sum_j v_j = \|U\|_1 = d$. Let $\tau > 0$ to be determined later, and define two index sets $H = \{j \mid v_j \geq \tau\}$ and $L = \{j \mid v_j < \tau\}$. It is easy to see that $|H| \leq \frac{d}{\tau}$ where $|\cdot|$ is used to denote the size of a finite set, and $\|v^L\|_\infty \leq \tau$ where

$$v_j^L = \begin{cases} v_j, & \text{if } j \in L \\ 0, & \text{otherwise} \end{cases}, \quad j = 1, \dots, n.$$

Similarly, when an index set appears as a superscript, we mean zeroing out elements or rows that do not belong to this index set, e.g., v^L and U^L . Define

$$Y = \{y \in \mathbb{R}^n \mid y = Ux, \|x\|_\infty = 1, x \in \mathbb{R}^d\}.$$

For any $y = Ux \in Y$, we have $\|y\|_1 = \|Ux\|_1 \geq \|x\|_\infty = 1$,

$$|y_j| = |u_j^T x| \leq \|u_j\|_1 \|x\|_\infty = v_j, \quad j = 1, \dots, n,$$

and thus $\|y\|_1 \leq \|v\|_1 = d$. Define $Y^L = \{y \in Y \mid \|y^L\|_1 \geq \frac{1}{2}\|y\|_1\}$ and $Y^H = Y \setminus Y^L$. Given S , define a mapping ϕ :

$\{1, \dots, n\} \rightarrow \{1, \dots, s\}$ such that $s_{\phi(j), j} = 1, j = 1, \dots, n$, and split L into two subsets: $\hat{L} = \{j \in L \mid \phi(j) \in \phi(H)\}$ and $\bar{L} = L \setminus \hat{L}$. Consider these events:

- \mathcal{E}_U : $|\Pi U|_1 \leq \omega_1 d \log d$ for some $\omega_1 > 0$.
- \mathcal{E}_L : $\|Sv^L\|_\infty \leq \omega_2/(d \log d)$ for some $\omega_2 > 0$.
- \mathcal{E}_H : $\phi(j_1) \neq \phi(j_2), \forall j_1 \neq j_2, j_1, j_2 \in H$.
- \mathcal{E}_C : $\min_{j \in |H|} |c_j| \geq \omega_3/(d^2 \log^2 d)$ for some $\omega_3 > 0$.
- $\mathcal{E}_{\bar{L}}$: $|\Pi U^{\bar{L}}|_1 \leq \omega_4/(d^2 \log^2 d)$ for some $\omega_4 > 0$.

Recall that we set $s = \omega d^5 \log^5 d$ in Theorem 2. We will show that, with ω sufficiently large and proper choices of $\omega_1, \omega_2, \omega_3$, and ω_4 , the event \mathcal{E}_U leads to an upper bound of $\|\Pi y\|_1$ for all $y \in \text{range}(A)$, \mathcal{E}_U and \mathcal{E}_L lead to a lower bound of $\|\Pi y\|_1$ for all $y \in Y^L$ with probability at least 0.9, and $\mathcal{E}_H, \mathcal{E}_{\bar{L}}$, and \mathcal{E}_C together imply an lower bound of $\|\Pi y\|_1$ for all $y \in Y^H$.

LEMMA 12. Provided \mathcal{E}_U , we have

$$\|\Pi y\|_1 \leq \omega_1 d \log d \cdot \|y\|_1, \quad \forall y \in \text{range}(A).$$

PROOF. For any $y \in \text{range}(A)$, we can find an x such that $y = Ux$. Then, $\|\Pi y\|_1 = \|\Pi Ux\|_1 \leq |\Pi U|_1 \|x\|_\infty \leq |\Pi U|_1 \|Ux\|_1 \leq \omega_1 d \log d \cdot \|y\|_1$. \square

LEMMA 13. Provided \mathcal{E}_L , for any fixed $y \in Y^L$, we have

$$\log \Pr \left[\|\Pi y\|_1 \leq \frac{1}{4} \|y\|_1 \right] \leq -\frac{d \log d}{24\omega_2}.$$

PROOF. Let $z = \Pi y$. We have,

$$\begin{aligned} |z_i| &= \left| \sum_j s_{ij} c_j y_j \right| \simeq \left(\sum_j s_{ij} |y_j| \right) |\tilde{c}_i| \\ &\succeq \left(\sum_j s_{ij} |y_j^L| \right) |\tilde{c}_i| := \tilde{\gamma}_i |\tilde{c}_i|, \end{aligned}$$

where $\{\tilde{c}_i\}$ are independent Cauchy variables. Let $\tilde{\gamma} = \sum_i \tilde{\gamma}_i = \|y^L\|_1$. Since $|y| \leq v$, we have $\tilde{\gamma}_i \leq \|Sv^L\|_\infty$. By Lemma 6,

$$\log \Pr \left[X \leq \frac{\|y^L\|_1}{2} \right] \leq \frac{-\|y^L\|_1}{12\|Sv^L\|_\infty}.$$

By assumption \mathcal{E}_L and $\|y^L\|_1 \geq \frac{1}{2}\|y\|_1 \geq \frac{1}{2}$, we obtain the result. \square

LEMMA 14. Assume both \mathcal{E}_U and \mathcal{E}_L . If ω_1 and ω_2 satisfy

$$d \log(6d(1 + 4\omega_1 d \log d)) - \frac{d \log d}{24\omega_2} \leq \log \delta$$

for some $\delta \in (0, 1)$ regardless of d , then, with probability at least $1 - \delta$, we have

$$\|\Pi y\|_1 \geq \frac{1}{8} \|y\|_1, \quad \forall y \in Y^L.$$

PROOF. Set $\epsilon = 1/(2 + 8\omega_1 d \log d)$ and create an ϵ -net $Y_\epsilon^L \subseteq Y^L$ such that for any $y \in Y^L$, we can find a $y_\epsilon \in Y_\epsilon^L$ such that $\|y - y_\epsilon\|_1 \leq \epsilon$. Since $\|y\|_1 \leq d$ for all $y \in Y^L$, there exist such an ϵ -net with at most $(3d/\epsilon)^d$ elements (Bourgain et al. [4]). By Lemma 13, we can apply a union bound for all the elements in Y_ϵ^L :

$$\begin{aligned} \Pr[\|\Pi y_\epsilon\|_1 \geq \frac{1}{4} \|y_\epsilon\|_1, \forall y_\epsilon \in Y_\epsilon^L] &\geq 1 - \left(\frac{3d}{\epsilon}\right)^d e^{-\frac{d \log d}{24\omega_2}} \\ &= 1 - e^{d \log \frac{3d}{\epsilon} - \frac{d \log d}{24\omega_2}} \geq 1 - \delta. \end{aligned}$$

For any $y \in Y^L$, we have, noting that $y - y_\epsilon \in \text{range}(A)$,

$$\begin{aligned} \|\Pi y\|_1 &\geq \|\Pi y_\epsilon\|_1 - \|\Pi(y - y_\epsilon)\|_1 \geq \frac{1}{4} \|y_\epsilon\|_1 - \omega_1 d \log d \|y - y_\epsilon\|_1 \\ &\geq \frac{1}{4} \|y\|_1 - \left(\frac{1}{4} + \omega_1 d \log d\right) \epsilon \geq \frac{1}{8} \|y\|_1. \end{aligned}$$

So we establish a lower bound for all $y \in Y^L$. \square

LEMMA 15. *Provided \mathcal{E}_H and $\mathcal{E}_{\hat{L}}$, if $\omega_3 > 4\omega_4$, we have*

$$\|\Pi y\|_1 \geq \frac{\omega_4}{d^2 \log^2 d} \|y\|_1, \quad \forall y \in Y^H.$$

PROOF. For any $y = Ux \in Y^H$, we have,

$$\begin{aligned} \|\Pi y\|_1 &\geq \|\Pi(y^H + y^{\hat{L}})\|_1 \geq \|\Pi y^H\|_1 - \|\Pi U^{\hat{L}} x\|_1, \\ &\geq \sum_{j \in H} |c_j| |y_j| - \|\Pi U^{\hat{L}}\|_1 \|x\|_\infty \geq \min_{j \in H} |c_j| \|y^H\|_1 - \|\Pi U^{\hat{L}}\|_1 \\ &\geq \left(\frac{\omega_3}{2d^2 \log^2 d} - \frac{\omega_4}{d^2 \log^2 d}\right) \|y\|_1 \geq \frac{\omega_4}{d^2 \log^2 d} \cdot \|y\|_1, \end{aligned}$$

which creates a lower bound for all $y \in Y^H$. \square

We continue to show that, with ω sufficiently large, by setting $\tau = \omega^{1/4}/(d \log^2 d)$ and choosing $\omega_1, \omega_2, \omega_3$, and ω_4 properly, we have each event with probability at least $1 - 0.08 = 0.92$ and thus

$$\Pr[\mathcal{E}_U \cap \mathcal{E}_L \cap \mathcal{E}_H \cap \mathcal{E}_{\hat{L}} \cap \mathcal{E}_C] \geq 0.6.$$

Moreover, the condition in Lemma 14 holds with $\delta = 0.1$, and the condition in Lemma 15 holds. Therefore, $\Pi = SC$ has the desired property with probability at least 0.5, which would conclude the proof of Theorem 2.

LEMMA 16. *With probability at least 0.92, \mathcal{E}_U holds with $\omega_1 = 500(1 + \log \omega)$.*

PROOF. With S fixed, we have,

$$\begin{aligned} |\Pi U|_1 &= |SCU|_1 = \sum_{k=1}^d \sum_{i=1}^s \left| \sum_{j=1}^n s_{ij} c_j u_{jk} \right| \\ &\simeq \sum_{k=1}^d \sum_{i=1}^s \sum_{j=1}^n (|s_{ij} u_{jk}|) |\tilde{c}_{ik}|, \end{aligned}$$

where $\{\tilde{c}_{ik}\}$ are *dependent* Cauchy random variables. Then

$$\sum_{k=1}^d \sum_{i=1}^s \sum_{j=1}^n |s_{ij} u_{jk}| = \sum_{k=1}^d \sum_{j=1}^n |u_{jk}| = |U|_1 = d.$$

Apply Lemma 5,

$$\Pr[|\Pi U|_1 \geq td | S] \leq \frac{2 \log(sdt)}{t}.$$

Setting $\omega_1 = 500(1 + \log \omega)$ and $t = \omega_1 \log d$, we have

$$\frac{2 \log(sdt)}{t} = \frac{2 \log(\omega \omega_1 d^6 \log^5 d)}{\omega_1 \log d} \leq 0.08.$$

We assume that $\log d \geq 1$ and $\log \omega \geq 1$. \square

LEMMA 17. *For any $\delta \in (0, 0.1)$, if $s \geq d/\tau$, we have,*

$$\Pr \left[\|Sv^L\|_\infty \geq \left(1 + 2 \log \frac{d}{\delta\tau}\right) \cdot \tau \right] \leq \delta.$$

PROOF. Let $X_{ij} = s_{ij} v_j^L$. Then $\mathbf{E}[X_{ij}] = v_j^L/s$, $\mathbf{E}[X_{ij}^2] = (v_j^L)^2/s$, and $0 \leq X_{ij} \leq v_j^L \leq \tau$. Fixed i , X_{ij} are independent, $j = 1, \dots, n$. By Bernstein's inequality,

$$\begin{aligned} \log \Pr \left[\sum_j X_{ij} \geq \frac{\|v^L\|_1}{s} + t \right] &\leq \frac{-t^2/2}{\|v^L\|_2^2/s + \tau t/3} \\ &\leq \frac{-t^2/2}{\tau(\|v^L\|_1/s + t/3)} \leq \frac{-t^2/(2\tau)}{d/s + t/3}. \end{aligned}$$

where we use Holder's inequality: $\|v^L\|_2^2 \leq \|v^L\|_1 \|v^L\|_\infty \leq d\tau$. To obtain a union bound for all i with probability $1 - \delta$, we need

$$\frac{-t^2/(2\tau)}{d/s + t/3} + \log s \leq \log \delta.$$

Given $\delta < 0.1$, it suffices if $s = d/\tau$ and $t = 2 \log(d/(\delta\tau))\tau$. Note that $\|v^L\|_1/s \leq \|v\|_1/s = \tau$. We have

$$\Pr \left[\|Sv^L\|_\infty \geq \left(1 + 2 \log \frac{d}{\delta\tau}\right) \cdot \tau \right] \leq \delta.$$

Increasing s will decrease the failure rate, so it holds for all $s \geq d/\tau$. \square

LEMMA 18. *With probability at least 0.92, \mathcal{E}_L holds with $\omega_2 = (15 + \log \omega)/\omega^{1/4}$.*

PROOF. By Lemma 17, with probability at least 0.92, \mathcal{E}_L holds with

$$\omega_2 = \frac{1 + 2 \log \frac{\omega^{1/4} d^2 \log^2 d}{0.08}}{\omega^{1/4} \log d} \leq \frac{15 + \log \omega}{\omega^{1/4}}.$$

\square

LEMMA 19. *With above choices of ω_1 and ω_2 , the condition in Lemma 13 holds with $\delta = 0.1$ for sufficiently large ω .*

PROOF. With $\omega_1 = 500(1 + \log \omega)$, $\omega_2 = (15 + \log \omega)/\omega^{1/4}$, the first term in

$$d \log(6d(1 + 4\omega_1 d \log d)) - \frac{d \log d}{24\omega_2}$$

increases much slower than the second term as ω increases, while both are at the order of $d \log d$. Therefore, if ω is sufficiently large, the condition hold with $\delta = 0.1$. \square

LEMMA 20. *If $\omega \geq 160$, event \mathcal{E}_H holds with probability at least 0.92.*

PROOF. Given $j_1, j_2 \in H$ and $j_1 \neq j_2$, let $X_{j_1 j_2} = 1$ if $\phi(j_1) = \phi(j_2)$ and $X_{j_1 j_2} = 0$ otherwise. It is easy to see that $\Pr[X_{j_1 j_2} = 1] = \frac{1}{s}$. Therefore,

$$\Pr[\mathcal{E}_H] \geq 1 - \sum_{j_1 < j_2} \Pr[X_{j_1 j_2} = 1] \geq 1 - \frac{|H|^2}{s} \geq 1 - \frac{d^2}{s\tau^2} \geq 1 - \frac{1}{\omega^{1/2}}.$$

It suffices if $\omega \geq 160$. \square

LEMMA 21. *With probability at least 0.92, event \mathcal{E}_C holds with $\omega_3 = 1/(8\omega^{1/4})$.*

PROOF. Let c be a Cauchy variable. We have

$$\Pr[|c| \leq t] = \frac{2}{\pi} \tan^{-1} t \leq \frac{2t}{\pi}.$$

$|H|$ is at most $d/\tau = \omega^{1/4}d^2 \log^2 d$. Then

$$\begin{aligned} \Pr[\mathcal{E}_C] &\geq 1 - |H| \cdot \Pr\left[|c| < \frac{\omega_3}{d^2 \log^2 d}\right] \\ &\geq 1 - \omega^{1/4}d^2 \log^2 d \cdot \frac{2\omega_3}{\pi d^2 \log^2 d}. \end{aligned}$$

Therefore, $\omega_3 = 1/(8\omega^{1/4})$ would suffice. \square

LEMMA 22. *With probability at least 0.92, event $\mathcal{E}_{\hat{L}}$ holds with $\omega_4 = 25000(1 + \log \omega)/\omega^{3/4}$. Thus with ω sufficiently large and the above choice of ω_3 , the condition in Lemma 15 $\omega_3 > 4\omega_4$ holds.*

PROOF. We have,

$$\mathbf{E}[|U^{\hat{L}}|_1] = \frac{|H|}{s} |U^L|_1 \leq \frac{\omega^{1/4}d^2 \log^2 d}{\omega d^5 \log^5 d} \cdot d = \frac{1}{\omega^{3/4}d^2 \log^3 d}.$$

By Markov's inequality,

$$\Pr\left[|U^{\hat{L}}|_1 \geq \frac{25}{\omega^{3/4}d^2 \log^3 d}\right] \leq 0.04.$$

Assume that $|U^{\hat{L}}|_1 \leq \frac{25}{\omega^{3/4}d^2 \log^3 d}$. Similar to the proof of Lemma 16, we have

$$\begin{aligned} |\Pi U^{\hat{L}}|_1 &= \sum_{k=1}^d \sum_{i \in \phi(H)} |\sum_j s_{ij} c_j u_{jk}^{\hat{L}}| \\ &\simeq \sum_{k=1}^d \sum_{i \in \phi(H)} \left(\sum_j s_{ij} |u_{jk}^{\hat{L}}| \right) |\tilde{c}_{ik}|, \end{aligned}$$

where $\{\tilde{c}_{ik}\}$ are dependent Cauchy variables. Apply Lemma 5,

$$\Pr[|\Pi U^{\hat{L}}| \geq |U^{\hat{L}}|t] \leq \frac{2 \log(|H|dt)}{t}$$

It suffices to choose $t = 1000(1 + \log \omega) \log d$ to make the RHS less than 0.04. So with probability at least 0.92, we have $\mathcal{E}_{\hat{L}}$ holds with $\omega_4 = 25000(1 + \log \omega)/\omega^{3/4}$. \square

A.3 Proof of Corollary 2

By Theorem 2 and Lemma 3, we know that Steps 2 and 4 of Algorithm 1 succeed with a constant probability. Conditioning on this event, we have

$$\begin{aligned} \|A\hat{x} - b\|_1 &\leq \frac{1}{1 - \epsilon/4} \|SA\hat{x} - Sb\|_1 \leq \frac{1 + \epsilon/4}{1 - \epsilon/4} \|SAx^* - Sb\|_1 \\ &\leq \frac{(1 + \epsilon/4)^2}{1 - \epsilon/4} \|Ax^* - b\|_1 \leq (1 + \epsilon) \|Ax^* - b\|_1, \end{aligned}$$

where the last inequality is due to $\epsilon < 1/2$. By Theorem 2, Step 2 takes $\mathcal{O}(\text{nnz}(A))$ time, and Step 3 takes $\mathcal{O}(\text{poly}(d))$ time because ΠA has $\mathcal{O}(\text{poly}(d))$ rows. Then, by Lemma 3, Step 4 takes $\mathcal{O}(\text{nnz}(A) \cdot \log n)$ time, and Step 5 takes time $\mathcal{T}_1(\epsilon/4; \mathcal{O}(\text{poly}(d) \log(1/\epsilon)/\epsilon^2), d)$. Therefore, the total running time of Algorithm 1 is as stated.

A.4 Proof of Lemma 8

First, we know that $\Pr[|X_p|^p \geq t] = \Pr[|X_p| \geq t^{1/p}] = 2 \cdot \Pr[X_p \geq t^{1/p}]$. Next, we cite the following lemma:

LEMMA 23. (Nolan [27, Thm. 1.12]) *Let $X \sim \mathcal{D}_p$ with $p \in [1, 2)$. Then as $x \rightarrow \infty$,*

$$\Pr[X > x] \sim c_p x^{-p},$$

where $c_p = \sin \frac{\pi p}{2} \cdot \Gamma(p)/\pi$.

By Lemma 23, it follows that, as $t \rightarrow \infty$, $\Pr[|X_p|^p \geq t] \sim 2c_p t^{-1}$. For the Cauchy distribution, we have

$$\Pr[|C| \geq t] = 1 - \frac{2}{\pi} \tan^{-1} t = \frac{2}{\pi} \tan^{-1} \frac{1}{t} \sim \frac{2}{\pi} \cdot t^{-1}.$$

Hence, there exist $\alpha'_p > 0$ and $t_1 > 0$ such that for all $t > t_1$,

$$\Pr[\alpha'_p |C| \geq t] \geq \Pr[|X_p|^p \geq t].$$

Note that all the p -stable distributions with $p \in [1, 2]$ have finite and positive density at $x = 0$. Therefore, there exists $\alpha''_p > 0$ such that for all $0 \leq t \leq t_1$,

$$\Pr[\alpha''_p |C| \geq t] \geq \Pr[|X_p|^p \geq t].$$

Let $\alpha_p = \max\{\alpha'_p, \alpha''_p\}$. We get $\alpha_p |C| \succeq |X_p|^p$. For the Gaussian distribution, we have, as $t \rightarrow \infty$,

$$\Pr[|G|^2 \geq t] \sim 2e^{-t/2} t^{-1/2},$$

which converges to zero much faster than t^{-1} , so we can apply similar arguments to obtain β_p .

A.5 Proof of Lemma 9

Let $C_i = F_c^{-1}(F_p(X_i))$, $i = 1, \dots, m$, where F_c is the CDF of the standard Cauchy distribution and F_p is the CDF of \mathcal{D}_p . C_i follows the standard Cauchy distribution, and, by Lemma 8, we have $\alpha_p |C_i| \geq |X_i|^p$. Therefore, for any $t \geq 1$,

$$\Pr[X \geq t\alpha_p \gamma] \leq \Pr\left[\sum_i \gamma_i |C_i| \geq t\gamma\right] \leq \frac{2 \log(mt)}{t}.$$

The last inequality is from Lemma 5.

A.6 Proof of Lemma 10

Let G_i be independent random variables sampled from the standard Gaussian distribution, $i = 1, \dots, m$. By Lemma 8,

$$\log \Pr[X \leq \beta_p(1-t)\gamma] \leq \log \Pr\left[\sum_i \gamma_i |G_i|^2 \leq (1-t)\gamma\right].$$

The lower tail inequality from Lemma 7 concludes the proof.

A.7 Proof of Theorem 6

The proof is similar to the proof of Sohler and Woodruff [29, Theorem 5], except that the Cauchy tail inequalities are replaced by tail inequalities for the stable distributions. See the technical report version of this paper [23] for details.

A.8 Proof of Theorem 7

Each of Steps 1, 3, and 5 of Algorithm 2 succeeds with a constant probability. Following the corresponding proofs of these steps, we see that we can control the success rate of each by adjusting the constant factor in the embedding dimension, such that all steps succeed with a constant probability. Conditioning on this event, we have $\kappa_p(AR^{-1}) = 6d$ because

$$\begin{aligned} \|AR^{-1}x\|_p &\leq 2\|\tilde{S}AR^{-1}x\|_p \leq 4d\|x\|_2, \\ \|AR^{-1}x\|_p &\geq \frac{2}{3}\|\tilde{S}AR^{-1}x\|_p \geq \frac{2}{3}\|x\|_2, \quad \forall x \in \mathbb{R}^d. \end{aligned}$$

By Lemma 1, $\bar{\kappa}_p(AR^{-1}) \leq 6d^{1/p+1}$, and by Lemma 3, the embedding dimension of S is

$$\mathcal{O}(\bar{\kappa}_p^p(AR^{-1})d^{lp/2-1}d \log(1/\epsilon)/\epsilon^2) = \mathcal{O}(d^{3+p/2} \log(1/\epsilon)/\epsilon^2),$$

which concludes the proof.